HGG
Advances

ARTICLE

TIGAR-V2: Efficient TWAS tool with nonparametric Bayesian eQTL weights of 49 tissue types from GTEx V8

Randy L. Parrish,[1] Greg C. Gibson,[2] Michael P. Epstein,[1] and Jingjing Yang[1,*]

Summary

Standard transcriptome-wide association study (TWAS) methods first train gene expression prediction models using reference transcriptomic data and then test the association between the predicted genetically regulated gene expression and phenotype of interest. Most existing TWAS tools require cumbersome preparation of genotype input files and extra coding to enable parallel computation. To improve the efficiency of TWAS tools, we developed Transcriptome-Integrated Genetic Association Resource V2 (TIGAR-V2), which directly reads Variant Call Format (VCF) files, enables parallel computation, and reduces up to 90% of computation cost (mainly due to loading genotype data) compared to the original version. TIGAR-V2 can train gene expression imputation models using either nonparametric Bayesian Dirichlet process regression (DPR) or Elastic-Net (as used by PrediXcan), perform TWASs using either individual-level or summary-level genome-wide association study (GWAS) data, and implement both burden and variance-component statistics for gene-based association tests. We trained gene expression prediction models by DPR for 49 tissues using Genotype-Tissue Expression (GTEx) V8 by TIGAR-V2 and illustrated the usefulness of these Bayesian *cis*-expression quantitative trait locus (eQTL) weights through TWASs of breast and ovarian cancer utilizing public GWAS summary statistics. We identified 88 and 37 risk genes, respectively, for breast and ovarian cancer, most of which are either known or near previously identified GWAS (~95%) or TWAS (~40%) risk genes and three novel independent TWAS risk genes with known functions in carcinogenesis. These findings suggest that TWASs can provide biological insight into the transcriptional regulation of complex diseases. The TIGAR-V2 tool, trained Bayesian *cis*-eQTL weights, and linkage disequilibrium (LD) information from GTEx V8 are publicly available, providing a useful resource for mapping risk genes of complex diseases.

Introduction

A transcriptome-wide association study (TWAS)[1–5] is a popular technique widely used for integrating reference transcriptomic data with genome-wide association study (GWAS) data to conduct gene-based association studies. TWAS has been shown to improve the power of identifying GWAS risk loci as well as illustrate the underlying biological mechanism of GWAS loci, for example in studies of schizophrenia (MIM: 181500),[6] age-related macular degeneration (MIM: 603075),[7] and broad types of complex traits.[8] In particular, the risk genes identified by TWASs have genetic effects potentially mediated through gene expression.

The standard two-stage TWAS methods[1–3] first fit gene expression prediction models using the reference transcriptomic and genetic data profiled for the same samples and then test the association between the predicted genetically regulated gene expression (GReX) and phenotype of interest for the test GWAS cohort. The TWAS framework enables the advantages of using publicly available reference transcriptomic data such as the Genotype-Tissue Expression (GTEx) project[9,10] and summary-level GWAS data.[11,12]

However, most of the existing tools[1,2,5] require cumbersome preparation of genotype data files and fail to take advantage of parallel computing to improve computational efficiency. These limitations result in difficulties for users who need to train gene expression prediction models using their own reference transcriptomic and genetic data. Here, we develop a new version of the Transcriptome-Integrated Genetic Association Resource (referred to as TIGAR-V2) that takes genotype data of the Variant Call Format (VCF) as input, conducts 5-fold cross-validation[13] to evaluate trained gene expression prediction models, and enables parallel computation to take advantage of high-performance computing clusters.

Additionally, TIGAR-V2 can train gene expression imputation models using either nonparametric Bayesian Dirichlet process regression (DPR)[14] or Elastic-Net penalized regression (as used by PrediXcan[1]). TIGAR-V2 can perform TWASs using either individual-level or summary-level GWAS data. Besides the burden type TWAS test,[1] the software further implements an additional variance-component test[15] for TWAS that retains power under model misspecification.

To make TIGAR-V2 a convenient resource for the public, we trained nonparametric Bayesian DPR gene expression prediction models for 49 tissues from the GTEx V8 reference panel (dbGaP accession number: phs000424.v8.p2).[10] These estimated tissue-specific SNP effect sizes on the expression quantitative traits (eQTs)

[1]Center for Computational and Quantitative Genetics, Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA;
[2]School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA
*Correspondence: jingjing.yang@emory.edu

https://doi.org/10.1016/j.xhgg.2021.100068.

are considered as Bayesian expression quantitative trait locus (eQTL) weights per gene and are provided along with this TIGAR-V2 tool, which can be conveniently used for follow-up gene-based association studies using both individual-level and summary-level GWAS data (i.e., TWAS). In our example application studies, we used eQTL weights obtained from transcriptomic data of breast mammary tissue and ovary tissue from the GTEx V8 reference panel along with publicly available GWAS summary statistics[11,12] to conduct TWASs for studying breast cancer (MIM: 114480) and ovarian cancer (MIM: 167000).

In the following sections, we first outline the TIGAR-V2 framework. We then describe the application of TIGAR-V2 to train gene expression prediction models with the GTEx V8 reference data and TWASs of breast cancer and ovarian cancer. Model training and application results are described. Finally, we conclude with a discussion.

Material and methods

TIGAR-V2 framework

Gene expression prediction model

The standard two-stage TWAS[1–3] first fits gene expression prediction models by taking genotype data (G) of *cis*-SNPs (e.g., within ±1 Mb of the target gene g[2,16,17]) as predictors, assuming the following additive genetic model for the expression quantitative trait (E_g) with respect to a target gene g.

$$\mathbf{E}_g = \mathbf{G}\mathbf{w} + \boldsymbol{\varepsilon}; \quad \boldsymbol{\varepsilon} \sim N\left(0, \sigma_\varepsilon^2 \mathbf{I}\right). \qquad \text{(Equation 1)}$$

The *cis*-eQTL effect size vector w can be estimated by different regression methods from the reference (i.e., training) data. For example, PrediXcan estimates w by a general linear regression model with Elastic-Net penalty;[1] FUSION estimates w by Elastic-Net, LASSO,[18] linear mixed modeling, sum of single effects (SuSiE),[19] and Bayesian sparse linear mixed model (BSLMM);[20] and TIGAR estimates w by a nonparametric Bayesian DPR model (Text S1).[3,14]

TIGAR-V2 implements both nonparametric Bayesian DPR[14] and general linear regression with Elastic-Net penalty as used by PrediXcan[1] to estimate w, which are eQTL effect sizes in a broad sense not considering whether the SNP has a genome-wide significant eQTL p value. Additionally, TIGAR-V2 runs 5-fold cross-validation[13] with the reference data by default to provide an average prediction R^2 per gene across 5 folds of validation data (referred to as 5-fold CV R^2). The 5-fold CV R^2 can be used to evaluate if the trained gene expression prediction model is "valid" for follow-up TWAS (e.g., using the threshold of 5-fold CV $R^2 > 0.005$). Here, we use a more liberal threshold than the threshold 0.01 used by previous studies[2,21,22] to allow more genes to be tested in follow-up TWASs. Because the follow-up gene-based association *Z*-score test statistic is essentially a weighted average of single-variant GWAS *Z*-score statistics with variant weights provided by the eQTL effect sizes (Equations 2 and 3), poorly estimated eQTL weights would only reduce power but will not increase the false-positive rate under the null hypothesis.

Gene-based association study

With the estimates of *cis*-eQTL effect sizes $\widehat{\mathbf{w}}$ and individual-level GWAS data of test samples, TIGAR-V2 predicts GReX values by taking estimates of *cis*-eQTL effect sizes (outputs from the step of training gene expression prediction models) and genotype data (VCF files, G_{test}) of test samples as inputs, and using the formula $\widehat{GReX} = \mathbf{G}_{test}\widehat{\mathbf{w}}$. TIGAR-V2 implements the burden[23,24] type TWAS test by testing the association between $\widehat{GReX}$ and the phenotype of interest (PED format) based on the general linear regression model, with the phenotype as response variable and predicted $\widehat{GReX}$ as a test covariate (Text S2.1). TIGAR-V2 implements the variance-component TWAS test by[15] using the sequence kernel association test (SKAT) framework[25] with variant weights provided by eQTL effect size estimates $\widehat{\mathbf{w}}$. The variance-component TWAS test is recommended if the assumption of the linear relationship between the SNP effect sizes on phenotype and eQTL weights is violated (see Text S2.2). Note that here the eQTL weights $\widehat{\mathbf{w}}$ are specific to the test gene and specific to the tissue type of the reference transcriptomic data.

With summary-level GWAS data (i.e., *Z*-score statistic values from single-variant GWAS tests) of test samples, TIGAR-V2 tests the gene-based association by using both burden[23,24] and variance-component[15] test statistics, where *cis*-eQTL effect size estimates $\widehat{\mathbf{w}}$ are taken as variant weights.

In particular, for burden test, we found that the FUSION *Z*-score statistic[2] as given by Equation 2 will lead to inflated false-positive findings if $\widehat{\mathbf{w}}$ is estimated using non-standardized reference data (i.e., centered gene expression and genotype data as described in Text S1 for the Bayesian DPR model); the S-PrediXcan test statistic[26] as given by Equation 3 should be used in this situation instead. We also show that both FUSION and S-PrediXcan test statistics are equivalent if $\widehat{\mathbf{w}}$ is estimated using standardized reference data (Text S2.2). The S-PrediXcan test statistic is the default test statistic implemented by TIGAR-V2.

$$\tilde{Z}_{g,FUSION} = \frac{\sum_{l=1}^{m}\left(\widehat{w}_l Z_l\right)}{\sqrt{\widehat{\mathbf{w}}'\mathbf{V}\widehat{\mathbf{w}}}}, \qquad \mathbf{V} = Corr(\mathbf{G}_0) \qquad \text{(Equation 2)}$$

$$\tilde{Z}_{g,SprediXcan} = \frac{\sum_{l=1}^{m}\left(\widehat{w}_l \widehat{\sigma}_l Z_l\right)}{\sqrt{\widehat{\mathbf{w}}'\mathbf{V}\widehat{\mathbf{w}}}}, \qquad \widehat{\sigma_l^2} = Var(\mathbf{G}_{0,\mathbf{l}}), \qquad \mathbf{V} = Cov(\mathbf{G}_0). \qquad \text{(Equation 3)}$$

Here, Z_l denotes the *Z*-score statistic value of genetic varaint l by single-variant GWAS test (i.e., summary-level GWAS data). The required linkage disequilibrium (LD) covariance matrix (or correlation matrix for FUSION test statistic) among test *cis*-SNPs (V), and the genotype variance of test *cis*-SNPs $(\widehat{\sigma_l^2} = Var(\mathbf{G}_{0,\mathbf{l}}))$ can be obtained from reference genotype data (G_0) such as 1000 Genomes[27] and GTEx V8.[10]

Tool framework

The tool framework of TIGAR-V2 is shown in Figure 1, where all TWAS steps in TIGAR-V2 are enabled using Python and Bash scripts. Python libraries "pandas,"[28,29] "numpy,"[28,30] "scipy,"[31] "sklearn,"[32,33] and "statsmodels"[34] are used to develop TIGAR-V2. Genotype data in VCF saved as one file per chromosome are input genotype files for TIGAR-V2. TABIX tool[35] is used to extract genotype data per target gene efficiently from VCF genotype files. Parallel computation is enabled by using the "multiprocessing" Python library, allowing users to train gene expression prediction models and test gene-based association of multiple genes in parallel.

This new version uses fewer Python library dependencies for easier setup, speeds up computation by improving genotype

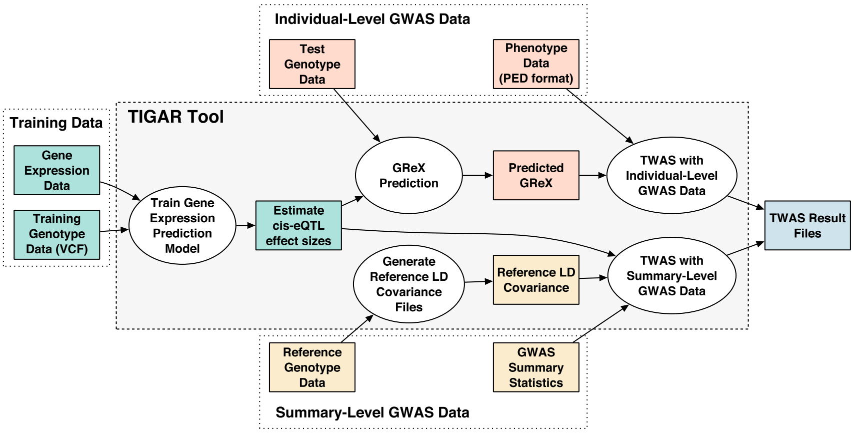
Individual-Level GWAS Data
Test Genotype Data
Phenotype Data (PED format)
Training Data
Gene Expression Data
Training Genotype Data (VCF)
TIGAR Tool
Train Gene Expression Prediction Model
Estimate cis-eQTL effect sizes
GReX Prediction
Predicted GReX
TWAS with Individual-Level GWAS Data
Generate Reference LD Covariance Files
Reference LD Covariance
TWAS with Summary-Level GWAS Data
TWAS Result Files
Reference Genotype Data
GWAS Summary Statistics
Summary-Level GWAS Data

Figure 1. TIGAR-V2 framework
Including TWAS steps of training gene expression prediction models from reference data, predicting GReX with individual-level GWAS data, and testing gene-based association with both individual-level and summary-level GWAS data.

data loading and using functions from the "numpy" Python library, reduces required memory usage by loading genotype data from VCF files with a row-by-row increment, and adds the function to conduct the recently published variance-component gene-based association test.[15] For example, for training gene expression prediction models by Bayesian DPR method with 129 samples, ~1,800 SNPs per gene, four genes, and a single core, the computation time is reduced up to 90% and memory usage up to 50% (mainly due to improved genotype data loading from VCF files), compared to the initial TIGAR tool. The memory usage is linear with respect to the number of *cis*-SNP predictors of the target gene and the training sample size. Training gene expression prediction models using the GTEx V8 reference data requires less than 8 GB of memory per gene, with number of test SNPs up to ~10K per gene and training sample size up to ~600. We would suggest users run one gene per typical computation core in a high-performance computing cluster (e.g., running 4 genes in parallel per chromosome if 4 cores are requested).

Reference resource from GTEx V8

Train Bayesian DPR eQTL weights from GTEx V8

The GTEx project V8 (dbGaP: phs000424.v8.p2) contains comprehensive profiling of whole genome sequencing (WGS) genotype data and RNA sequencing (RNA-seq) transcriptomic data (15,253 normal samples) across 54 tissue types of 838 donors.[9,10,36,37] GTEx V8 provides useful reference data for training tissue-specific gene expression prediction models for diverse tissue types on human bodies. Both PrediXcan and FUSION tools use GTEx V8 data as the reference data and provide estimated *cis*-eQTL weights per gene with respect to 49 tissue types that have >70 samples with profiled WGS genotype and RNA-seq transcriptomic data (Figure 2A) as a public resource for TWASs.

Here, we also train tissue-specific gene expression prediction models for these 49 tissue types using the nonparametric Bayesian DPR method previously implemented in TIGAR. WGS genotype data of *cis*-SNPs within ±1 Mb around gene transcription start sites (TSSs) of the target gene were used as predictors. In particular, variants with missing rate <20%, minor allele frequency >0.01, and Hardy-Weinberg equilibrium p value $>10^{-5}$ were considered for fitting the gene expression prediction models. Gene expression data of transcripts per million (TPM) per sample per tissue were downloaded from the GTEx portal. Genes with >0.1 TPM in ≥10 samples were considered. Raw gene expression data (TPM) were then adjusted for age, body mass index (BMI), top five genotype principal components, and top probabilistic estimation of expression residuals (PEER) factors.[38] The gene expression data of breast mammary tissue were further adjusted for *ESR1* expression following previous TWAS analysis of breast cancer.[39]

Five-fold cross-validation was conducted by default to obtain 5-fold CV R^2 per gene per tissue. Only "significant" gene expression prediction models with 5-fold CV $R^2 > 0.005$ were retained in the output files (see our explanation in the Material and methods and Discussion sections). The estimated Bayesian *cis*-eQTL weights from these "significant" gene expression prediction models can be used to conduct TWASs using both individual-level and summary-level GWAS data and are shared with the public along with our TIGAR-V2 tool.

Further, we compared gene expression prediction models trained from GTEx V8[1,26,40] by nonparametric Bayesian DPR method to the ones (i.e., PredictDB models, see Web resources) trained from the same GTEx V8 reference data by Elastic-Net method using the PrediXcan tool.

Application TWASs of breast and ovarian cancer

We used TIGAR-V2 to conduct TWASs of breast and ovarian cancer by using the Bayesian *cis*-eQTL weights estimated from GTEx V8[10] of breast mammary tissue and ovary tissue and summary-level GWAS data.[11,12] The GWAS summary data of breast and ovarian cancer were respectively obtained from the Breast Cancer Association Consortium (BCAC) with 122,977 cases and 105,974 controls of European ancestry[11] and the Ovarian Cancer Association Consortium (OCAC) with 22,406 cases and 40,941 controls of European ancestry.[12] We also compared with TWAS results using eQTL weights given by Elastic-Net method (i.e., PrediXcan), which were also generated by our TIGAR-V2 tool.

Analyses conducted in this study use de-identified transcriptomic and genetic data from GTEx V8 and summary-level GWAS data of breast and ovarian cancer, which are in accordance with the ethical standards of the Institutional Review Board (IRB) at Emory University.

Results

Bayesian DPR eQTL weights from GTEx V8

From the GTEx V8 reference data as described previously, a total of 1,104,305 "significant" gene expression prediction models with 5-fold CV $R^2 > 0.005$ were successfully trained by TIGAR (using the nonparametric Bayesian DPR method) for genes on the autosomal chromosomes of 49 tissue types. The average and median number of gene expression prediction models obtained per tissue type was ~22.5K. The corresponding Bayesian DPR eQTL weights (i.e., effect sizes of *cis*-SNPs in the fitted gene

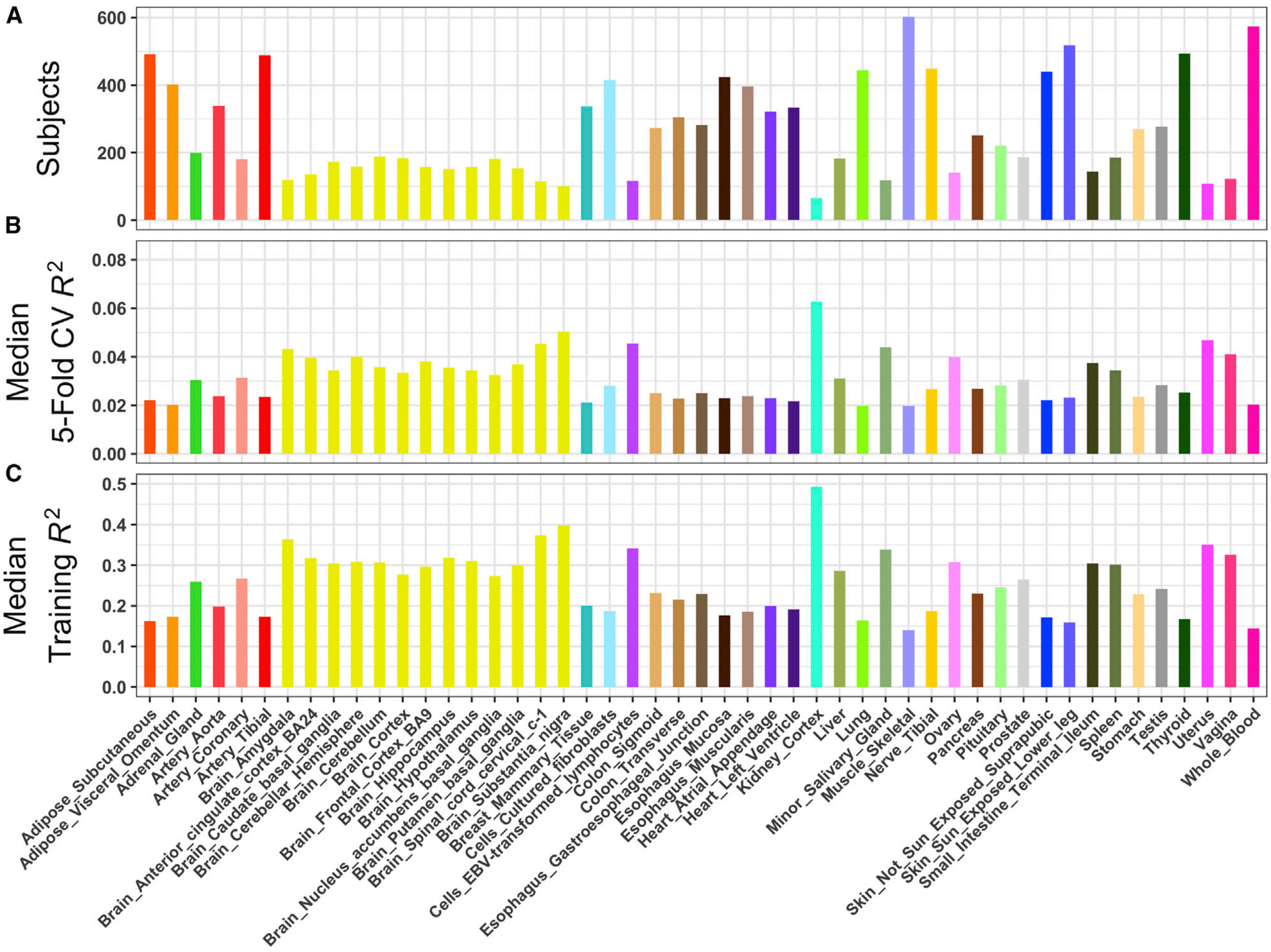

Figure 2. Trained gene expression prediction models of 49 tissue types from GTEx V8 by TIGAR using the nonparametric Bayesian DPR method

(A) Number of training sample size per tissue.
(B) Median 5-fold CV R^2 per tissue.
(C) Median training R^2 per tissue. Colors are coded with respect to groups of tissue types.

expression prediction models by nonparametric Bayesian DPR method) are publicly available along with our TIGAR-V2 tool.

Model over-fitting due to small training sample sizes

We present the median 5-fold CV R^2 and the median training R^2 of genome-wide genes per tissue type by TIGAR in Figures 2B and 2C, respectively. Here, the 5-fold CV R^2 approximates the prediction R^2 in independent data. Surprisingly, we observed that larger median 5-fold CV R^2 and training R^2values were obtained for tissue types with smaller sample size (Figure 2). For example, the top median 5-fold CV R^2 values (~0.04) were obtained for kidney cortex tissue (cyan bar), various brain tissues (yellow bars), and uterus tissue (hot pink bar), which all have sample sizes ~100, whereas tissues that have relatively large sample sizes (400~600; muscle skeletal, skin, and whole blood) have median 5-fold CV $R^2 \approx 0.02$. This trend is further demonstrated in the density plots of 5-fold CV R^2 and training R^2 by TIGAR for all tissues, color-coded with respect to their training sample sizes (Figure S1).

We suspect this controversial trend is mainly due to model over-fitting with small training sample sizes. To further investigate this, we take the gene expression prediction models fitted with breast (n = 337) and ovarian (n = 140) tissue types as examples. First, we down-sampled breast tissue samples to 140 to match with the sample size of ovarian tissue. Second, we trained both PrediXcan Elastic-Net and TIGAR nonparametric Bayesian DPR models on the down-sampled breast tissue data. Third, we made density plot of the 5-fold CV R^2 and training R^2 for genes that have 5-fold CV R^2 greater than various thresholds (0.005, 0.01, 0.05, 0.1, 0.2) (Figures S2 and S3).

We found that the same over-fitting issue existed for both PrediXcan Elastic-Net and TIGAR DPR methods. That is, the down-sampled breast tissue with the same sample size (140) as the ovarian tissue showed similar density distributions with respect to training R^2, which had larger median training R^2 than the breast tissue with sample size 337. As for the 5-fold CV R^2, genes with 5-fold CV $R^2 > 0.2$ had similar distributions between down-sampled and original breast tissues (which is

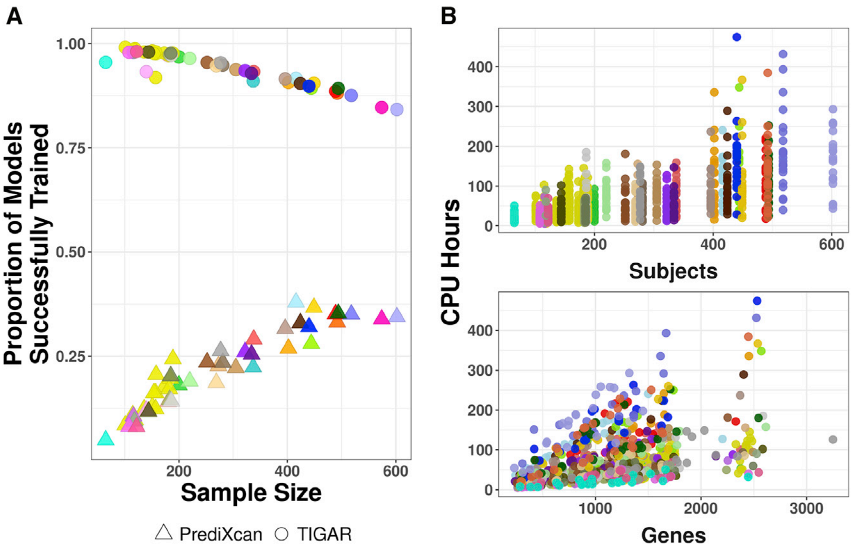

Figure 3. Computation efficiency of TIGAR-V2

Proportion of valid gene expression prediction models by TIGAR versus PrediXcan (A) and computation costs by TIGAR-V2 (B). The same color codes with respect to different tissue types as used in Figure 2 are used here. Computation times are in CPU hours per chromosome per tissue for training gene expression prediction models with GTEx V8 reference data.

expected), whereas other groups of genes had similar distributions between down-sampled breast tissue and ovarian tissue that are of the same training sample size (which is controversial due to overfitting). We think this is mainly driven by genes with relatively small expression heritability that would require a larger sample size to ensure a less over-fitted model. Since the TIGAR DPR method has higher power to fit gene expression precision models for genes with relatively small expression heritability, the TIGAR training results are affected more by this over-fitting issue.

Comparison with PrediXcan eQTL weights

Additionally, we compared the gene expression prediction model training results by TIGAR with the ones by PrediXcan using the same GTEx V8 reference data.[1,26,40] From boxplots of medians (Figure S4) and density plots (Figure S5) of 5-fold CV R^2 and training R^2 by PrediXcan, we observed the similar overfitting trend — relatively larger median 5-fold CV R^2 and median training R^2 values were obtained with relatively smaller training sample sizes. These findings are consistent with our TIGAR training results (Figure 2; Figure S1) as well as our down-sample investigation (Figures S2 and S3).

As shown in Figure S6, more consistent 5-fold CV R^2 and training R^2 were obtained by PrediXcan and TIGAR for genes that were of relatively larger sample sizes (yellowish colors) and relatively higher expression heritability. We found TIGAR had consistently better performance with fitting more valid gene expression prediction models for genes of relatively smaller expression heritability. In particular, the higher median 5-fold CV R^2 shown in Figure S4 by PrediXcan is based on the group of valid genes with 5-fold CV $R^2 > 0.005$ by PrediXcan that is only <50% of the valid genes by TIGAR (Figure 3A; Figure S7). These findings are also consistent with previous studies.[3]

Computation cost by TIGAR-V2

The training computation costs in CPU hours per chromosome per tissue with GTEx V8 reference data by TIGAR-V2 are shown in Figure 3B, with respect to training sample sizes and number of genes in the chromosome. The computation cost per chromosome per tissue ranged from 5 CPU hours to over 474, with a median of 50.6 and mean of 69.1, which is mainly due to various numbers of genes per chromosome and various sample sizes per tissue. That is, with sample size ~300, the average computation time for training a nonparametric Bayesian DPR gene expression prediction model per gene with 5-fold cross-validation is only ~4 min by TIGAR-V2. The computation complexity is linear with respect to training sample sizes. Given the same computation cost for loading VCF genotype data, fitting a Bayesian DPR model costs about 2× computation time than fitting an Elastic-Net model by TIGAR-V2.

TWASs of breast and ovarian cancer

From the gene expression prediction model training results by TIGAR, we respectively obtained 22,781 and 22,823 valid gene expression prediction models with 5-fold CV $R^2 > 0.005$ by using the nonparametric Bayesian DPR method for breast (N = 337) and ovarian (N = 140) tissue types (Figure S7). Using GWAS summary statistics of breast cancer and ovarian cancer[11,12] and Bayesian *cis*-eQTL weights estimated with respect to the corresponding tissue type, TIGAR using our Bayesian eQTL weights respectively detected 88 and 37 significant TWAS genes (p values $< 2.5 \times 10^{-6}$) for breast and ovarian cancer (see Manhattan plot in Figure 4). Of these significant genes, 17 were identified as risk genes of both breast and ovarian cancer (Table S1).

Independently significant TWAS risk genes by TIGAR

Out of these 88 significant TWAS genes for breast cancer by TIGAR, 20 genes are known GWAS risk genes of breast cancer,[11,41–48] 64 are located within a 1 Mb region of a previously identified GWAS locus of breast cancer[11,41–48] (Table S2), and 35 genes are identified by previous TWASs.[21,39,49–52] Similarly, out of these 37 significant TWAS genes for ovarian cancer by TIGAR, 34 genes are located on chromosome 17 including two known GWAS risk genes (*NSF* and *PLEKHM1*),[12,53,54] 33 genes are located

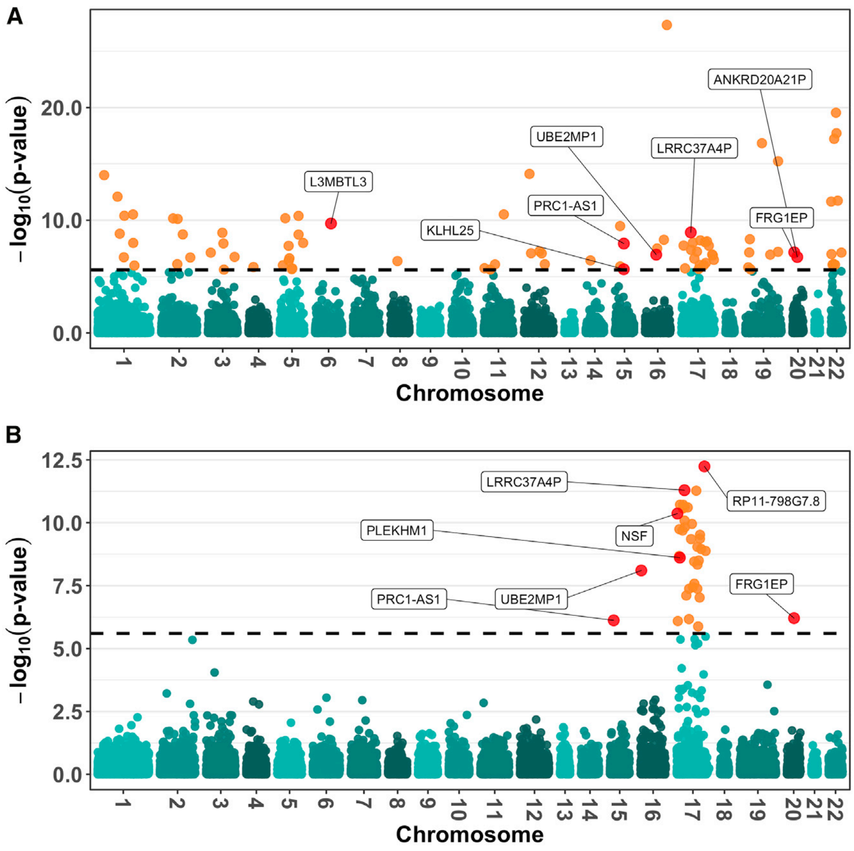

Figure 4. Manhattan plots of TWAS results by TIGAR for studying breast and ovarian cancer. Each dot denotes the -log10(p-value) per gene by TWAS

(A) TWAS results of breast cancer with 88 significant risk genes. Significant gene *FCGR1B* of breast cancer (p value: 4.12 × 10^{-63}) was removed from (A) to reduce the upper limit of the y axis.

(B) TWAS results of ovarian cancer with 37 significant risk genes. Significant genes discussed in the main text are labeled in the plots.

within 1 Mb of these two known GWAS risk genes (Table S3), and 13 genes (including *NSF*[55]) are identified by previous TWASs.[39,55,56] The known GWAS risk genes are curated from GWAS Catalog[57] containing at least one significant SNP within or ±1 Mb around the gene region.

Since the TWAS is conducted using genotype data within a ±1 Mb region of the test gene (i.e., test region), genes with overlapped test regions often have highly correlated GReX values (see locus-zoom plots around the top significant TWAS genes on chromosome 17 for breast and ovarian cancer in Figure 5). Thus, these nearby significant TWAS genes are often not representing independent associations. In Tables 1 and 2, we listed the most significant genes among genes that have shared test regions, which represent the independently significant TWAS risk genes. For breast cancer, 31 out of all 34 independent TWAS risk genes were either identified by a previous GWAS/TWAS or within the ±1 MB region of previously identified risk genes of breast cancer (Table 1). For example, TIGAR identified *L3MBTL3* (previously identified by GWAS[11] and TWAS[21,50–52]) and an additional 6 significant genes within the 1 Mb region of *L3MBTL3*. Of the independent TWAS genes of breast cancer, 17 (54%) have been identified by previous TWASs using PrediXcan and FUSION.[21,39,49–52]

Similarly, as shown in Table 2, TIGAR identified 4 independent significant TWAS genes for ovarian cancer. In particular, TWAS risk gene *RP11-798G7.8* on chromosome 17 was identified by a previous TWAS[39] and lies within 1 Mb of known GWAS risk gene *PLEKHM1*.[12,53] Interestingly, all independent TWAS risk genes of ovarian cancer by TIGAR (*PRC1-AS1*, *UBE2MP1*, *RP11-798G7.8*, and *FRG1EP*) are also TWAS risk genes of breast cancer,[11,39,58] which demonstrates a likely pleiotropy effect for these TWAS risk genes.

Significant TWAS risk genes identified by TIGAR in the 17q21.31 region

In particular, for the cluster of TWAS significant genes on chromosome 17 that were found to be shared by both breast and ovarian cancer, these genes have highly correlated GReX values as shown in Figure 5, including corticotrophin-releasing hormone receptor 1 (*CRHR1*) and microtubule-associated protein tau (*MAPT*). These genes are located in the 17q21.31 region, which contains a common inversion polymorphism of approximately 900 KB in populations with European ancestry,[59,60] where two divergent *MAPT* haplotypes, H1 and H2, are shown to be associated with neurodegenerative diseases. A recent study showed that the expression of several genes in and at the borders of the inversion region were affected by the inversion, where the expression changes were specific to whole blood or different brain regions.[61] Our findings show that the clusters of TWAS significant genes in the 17q21.31 region have differential GReX values in breast and ovary tissues with respect to both breast and ovarian cancers, and these GReX values are likely to be regulated by the *cis*-eQTL that are part of the inversion polymorphism.

Novel findings by TIGAR

TIGAR identified three novel independent TWAS risk genes (*KLHL25*, *UBE2MP1*, and *FRG1EP*) for breast cancer. Gene *KLHL25* has known biological functions involved in carcinogenesis, while genes *UBE2MP1* and *FRG1EP* are near such a gene.[62–66] Interestingly, genes *UBE2MP1* and *FRG1EP* were also identified for ovarian cancer by TIGAR (Table 2), and all three genes are involved with biological functions in carcinogenesis, either directly or indirectly. The protein encoded by *KLHL25* was reported acting as

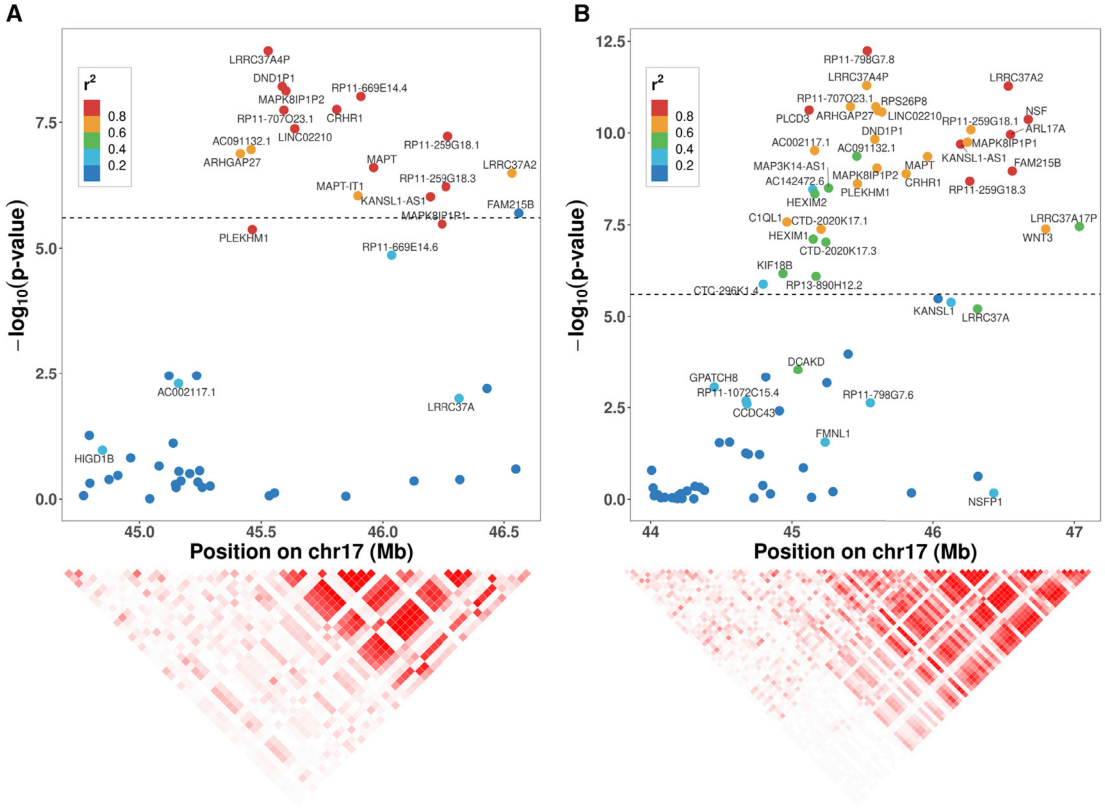

Figure 5. LocuxZoom plots for genes within 1MB around the most signfiicant TWAS genes on chromosome 17
LocusZoom plots for TWAS loci of (A) breast (top significant gene: *LRRC37A4P*) and (B) ovarian (top significant gene: *RP11-789G7.8*) cancer. Each dot denotes the $-\log_{10}$ (TWAS p value) of a gene color-coded with respect to their GReX R^2 with the top significant gene. The bottom heatmap colors denote the pairwise GReX R^2, with bright red denoting GReX R^2 close to 1 and white denoting GReX R^2 close to 0.

an adaptor protein for a suspected lung cancer tumor-suppressing protein *CUL3* to form an enzyme complex that targets ACLY, a protein often overexpressed in cancers, for degradation.[64] Pseudogene *UBE2MP1* was found to have a significant expression-methylation-correlation difference between normal and cancerous breast tissue.[62] *UBE2MP1* was also found to be amplified in gastric cancers (MIM: 613659) with amplified copy number variations in the 16p11.2 region, a mutation found to be associated with shorter overall survival,[66] and was predicted to be a driver of lung adenocarcinoma (MIM: 211980).[65] The test region of *FRG1EP* overlaps with the test region of pseudogene *ANKRD20A21P*, another TWAS risk gene identified by TIGAR, which has been implicated as a potentially important long non-coding RNA (lncRNA) regulator of endometrial carcinogenesis (MIM: 608089).[63]

Comparison with TWAS results by PrediXcan

Additionally, we compared with the TWAS results of breast and ovarian cancer obtained by using *cis*-eQTL weights estimated by Elastic-Net method as used by PrediXcan (Tables S4 and S5), which were generated using the PrediXcan (Elastic-Net) function enabled in our TIGAR-V2 tool. We respectively obtained 11,095 and 12,337 valid gene expression prediction models for breast and ovary tissue types by using the PrediXcan function, about half of the number of valid gene expression prediction models by TIGAR (Figure S7). This is consistent with our above comparison of trained valid gene expression models by TIGAR and PrediXcan (Figure 3A).

As a result, PrediXcan detected 56 significant (32 independent) TWAS genes for breast cancer and 4 significant (2 independent) TWAS genes for ovarian cancer (Figure S8; Tables S4 and S5). Respectively, 30 out of 32 and 2 out of 2 of the independent TWAS risk genes by PrediXcan for breast and ovarian cancer were either identified by previous corresponding GWASs or within the 1 Mb region of a known GWAS risk gene (Tables S6 and S7).

Even though there were 18 (56.25%) and 2 (100%) independent TWAS genes of breast and ovarian cancer, respectively, by PrediXcan also identified by TIGAR (Figure S9; Tables S8–S10), only TIGAR identified the novel TWAS genes *UBE2MP1* and *FRG1EP* shared by both breast and ovarian cancer and the known GWAS risk genes *FGF10*[11,67] and *TOX3*[43,45] of breast cancer. Other exclusive independent TWAS genes identified by TIGAR include lncRNA *RP11-758M4.4*, which was shown to be a potential biomarker of breast cancer;[68] *RPS23*, which was found to be overexpressed in advanced colorectal adenocarcinomas (MIM: 114500);[69] and *ZNF404*, whose

Table 1. Independent TWAS risk genes of breast cancer identified by TIGAR

Gene	MIM	CHR	Start	End	*Z*-score	p value
FCGR1B[a]	601502	1	121087345	121096310	−16.77	4.12e−63
KLHDC7A[b]		1	18480982	18486126	−6.04	1.56e−09
MTX1P1[a]		1	155230975	155234325	5.21	1.92e−07
AC010136.2[a]		2	217978707	217992615	−6.52	6.80e−11
CASP8[b]	601763	2	201233443	201287711	−6.51	7.56e−11
EOMES[a]	604615	3	27715949	27722711	6.07	1.28e−09
PSMD6-AS2[a]		3	64004022	64012148	−5.38	7.50e−08
FAM114A1[a]		4	38867677	38945739	−4.82	1.41e−06
FGF10[b]	602115	5	44303544	44389706	6.60	4.13e−11
SLC22A5[a]	603377	5	132369752	132395614	6.53	6.63e−11
ANKRD55[a]	615189	5	56099678	56233359	−5.63	1.85e−08
RPS23[a]	603683	5	82273358	82278577	4.77	1.86e−06
L3MBTL3[b]	618844	6	130013699	130141451	6.37	1.93e−10
RP11-758M4.4[a]		8	74798784	74866939	5.06	4.17e−07
PIDD1[b]	605247	11	799191	809646	−6.64	3.04e−11
CCDC91[b]	617366	12	28133249	28581511	−7.77	7.76e−15
RP11-116D17.4[a]		12	115318657	115320405	−5.36	8.40e−08
CTD-2325P2.4[a]		14	68627166	68628445	−5.09	3.65e−07
RCCD1[b]	617997	15	90955796	90963125	−6.29	3.26e−10
MAN2C1[a]	154580	15	75358201	75368154	−4.85	1.25e−06
KLHL25		15	85759323	85795030	−4.73	2.22e−06
TOX3[b]	611416	16	52438005	52547802	10.98	4.82e−28
UBE2MP1*		16	35169692	35170241	−5.31	1.13e−07
LRRC37A4P[a]		17	45506741	45550335	6.08	1.20e−09
CBX8[b]	617354	17	79792132	79801683	5.76	8.46e−09
TOM1L1[a]	604701	17	54899387	54960627	4.77	1.84e−06
SSBP4[b]	607391	19	18418864	18434387	8.53	1.47e−17
ZNF404[a]		19	43872363	43884051	5.41	6.31e−08
FRG1EP*		20	29480147	29497179	5.39	6.95e−08
DNAJB7[a]	611336	22	40859549	40861617	−9.22	2.89e−20
TMEM184B[a]		22	38219291	38273034	4.92	8.72e−07

*Novel risk gene.
[a]Genes within 1 Mb of known GWAS risk genes of breast cancer.
[b]Known GWAS risk genes of breast cancer.

dysregulation was linked to breast cancer pathogenesis by eQTL analyses.[70,71] Potentially novel TWAS risk genes by PrediXcan and TIGAR that were not identified by previous GWASs are presented in Table S11.

cis-eQTL weights by PrediXcan and TIGAR

To investigate the reasons that PrediXcan and TIGAR led to different TWAS findings, we took three TWAS risk genes shared by both breast and ovarian cancer as examples. In particular, *FRG1EP* was only identified by TIGAR for both breast and ovarian cancer, while *LRRC37A4P* and *PRC1-AS1* were identified by both PrediXcan and TIGAR for both breast and ovarian cancer. Pseudogene *LRRC37A4P* on chromosome 17 lies within 1 Mb downstream of the known risk gene *PLEKHM1* of breast cancer and ovarian cancer.[12,53] Gene *PRC1-AS1* on chromosome 15 is a lncRNA gene previously identified as being associated with breast carcinoma.[11,58] Regulation of *PRC1-AS1* is known to differ with respect to different types of breast cancers,[72] and increased expression of *PRC1-AS1* lncRNA is associated with hepatocellular carcinoma (MIM: 114550).[73]

Table 2. Independent TWAS risk genes of ovarian cancer identified by TIGAR

Gene	CHR	Start	End	*Z*-score	p value
PRC1-AS1[a]	15	90972860	90988624	4.95	7.56e−07
UBE2MP1*	16	35169692	35170241	5.77	7.88e−09
RP11-798G7.8[a]	17	45531577	45533838	−7.21	5.77e−13
FRG1EP*	20	29480147	29497179	−4.99	6.19e−07

No MIM identifier available for any gene in this table.
*Novel risk gene.
[a]Genes within 1 Mb of known GWAS risk genes of ovarian cancer.

We plotted the *cis*-eQTL weights estimated by Elastic-Net (PrediXcan) and the Bayesian DPR method (TIGAR) from GTEx V8 for these three example TWAS risk genes, color-coded with respect to $-\log_{10}$ (p value) by single-variant GWAS (Figures S10–S12). We observed that Bayesian estimates generally had non-zero values for all SNPs within the test region, while Elastic-Net estimates had non-zero values for <100 SNPs within the test region that had effect sizes (i.e., weights) of relatively larger magnitudes. These results match with the assumptions by the nonparametric Bayesian DPR (TIGAR) and Elastic-Net methods (PrediXcan). We can see that PrediXcan would miss the risk gene if test SNPs with non-zero weights have nonsignificant GWAS p values such as *FRG1EP* (Figure S10). Otherwise, both PrediXcan and TIGAR would have similar power to identify the risk genes such as *LRRC37A4P* and *PRC1-AS1*, whose TWAS association are mainly driven by GWAS significant SNPs (Figures S11 and S12).

Discussion

In this work, we develop a new version of the TIGAR tool[3] with improved computation efficiency, referred to as TIGAR-V2. Compared to the initial TIGAR tool, this new version reduces up to 90% computation time and up to 50% memory usage, mainly due to improved genotype data loading from VCF files and the usage of the Python library “numpy.” TIGAR-V2 can efficiently train gene expression prediction models by using both nonparametric Bayesian DPR and Elastic-Net (as used by PrediXcan) methods, as well as construct gene-based association tests using either individual-level or summary-level GWAS data. Gene-based associated tests implemented in TIGAR-V2 include both burden statistics (based on FUSION[2] and S-PrediXcan[26] *Z*-score test statistics) and variance-component statistics.

We trained gene expression prediction models of 49 tissue types with the GTEx V8 reference data by using the nonparametric Bayesian DPR method. We provide trained eQTL weights of genes that have 5-fold CV $R^2 > 0.005$ in the Synapse database with a link given in Web resources in this paper. Since we used a more liberal threshold than the 0.01 used by previous studies[2,21,22] to allow more genes to be tested in follow-up TWAS, we would suggest users to investigate the 5-fold CV R^2 and test p values of the expression prediction models, as well as the biological functions of significant TWAS risk genes.[26] Along with eQTL weights, we also provide gene information output files (an output file by TIGAR-V2) containing gene annotations (position, ID, name), training sample sizes, numbers of considered *cis*-SNPs, numbers of effective *cis*-SNPs for follow-up TWASs with non-zero eQTL weights, 5-fold CV R^2, training R^2, and a test p value with respect to training R^2. These gene information output files can be used by users to investigate the model training metrics of their TWAS significant genes. A similar approach is also suggested by the recent TWAS paper using GTEx data by PrediXcan,[26] which does not filter out any genes but only investigates the test p value with respect to the training R^2 for significant TWAS genes.

Additionally, a recent power analysis of TWASs suggested useful threshold of expression heritability >0.04 for a causal model where gene expression is directly causal with respect to the phenotype, and a threshold of expression heritability >0.06 for a pleiotropy model where true causal SNPs of the phenotype are also true causal eQTLs with respect to gene expression,[74] which allowed a TWAS that had higher power than a single-variant GWAS for a simulation cohort with sample size 2,504 that was used as both training and test data. We would only suggest TWAS as a secondary analysis to standard single-variant GWAS, instead of as a competing analysis. We want to remind the users that TWASs are essentially gene-based association tests that are not comparable to standard single-variant GWASs, but TWASs can provide extra biological insights with respect to the transcriptome data.

We demonstrated the usefulness of these trained models by performing TWASs of breast and ovarian cancer by integrating the estimated *cis*-eQTL weights of relevant tissue types with the relevant GWAS summary statistics. Compared to the *cis*-eQTL weights estimated by PrediXcan with the GTEx V8 data and TWAS results by PrediXcan, our Bayesian *cis*-eQTL weights led to not only a larger number of significant TWAS risk genes but also interesting novel TWAS risk genes with potential pleiotropy effects for breast and ovarian cancer. With a larger number of “valid” gene expression prediction models trained by the nonparametric Bayesian DPR method, TIGAR is expected to identify a larger number of TWAS risk genes than PrediXcan.

Our TWAS results of breast and ovarian cancer validated our TIGAR-V2 tool with findings consistent with previous GWASs and TWASs, revealed biological insights for known GWAS risk genes (*NSF* and *PLEKHM1*)[12,53,54] in the 17q21.31 region on chromosome 17 with pleiotropy effects for both breast and ovarian cancer, and identified novel risk genes that were shown to be possibly involved in the biological mechanisms of oncogenesis.

The TIGAR-V2 tool still has its limitations, such as considering only *cis*-eQTL data and assuming a two-stage model for TWASs. There are many other alternative TWAS tools available to address these two limits. For example, BGW-TWAS[4] and MOSTWAS[22] use both *cis*- and *trans*- genotype data to train a gene expression prediction model of the target gene, while CoMM[75] and PMR-Egger[5,76] assume a joint model with reference and test data that can achieve higher power when both datasets are homogeneous.

In conclusion, the TIGAR-V2 tool along with Bayesian *cis*-eQTL weights and reference LD covariance data (European ancestry) estimated from the GTEx V8 reference data are freely shared with the public on GitHub and Synapse. Given the convenience of directly loading VCF genotype data saved per chromosome, flexibility of using different training models and TWAS test statistics, and efficient computation enabled by Python source code, we believe our improved TIGAR-V2 tool will provide a useful resource for mapping risk genes of complex diseases by TWAS.

Data and code availability

The TIGAR-V2 tool generated in this study is available at GITHUB (see Web resources). All Bayesian eQTL weights of 49 tissue types from GTEx V8 are available at Synapse (Synapse ID: syn16804296). Individual-level GTEx V8 data are available from dbGaP (Accession number: phs000424.v8.p2). GWAS data for studying breast cancer are available from the Breast Cancer Association Consortium (BCAC). GWAS data for studying ovarian cancer are available from the Ovarian Cancer Association Consortium (OCAC).

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2021.100068.

Acknowledgments

R.L.P. and J.Y. are supported by National Institutes of Health (NIH/NIGMS) grant award R35GM138313. M.P.E. was supported by NIH/NIGMS grant award R01GM117946 and NIH/NIA grant award RF1AG071170.

Declaration of interests

The authors declare no competing interests.

Received: July 20, 2021
Accepted: November 1, 2021

Web resources

Breast Cancer Association Consortium (BCAC), http://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/oncoarray-and-combined-summary-result/gwas-summary-results-breast-cancer-risk-2017/

eQTL weights trained by PrediXcan with GTEx V8, https://predictdb.org

GTEx portal, https://gtexportal.org/home/

OMIM, https://omim.org/

Ovarian Cancer Association Consortium (OCAC), ftp://ftp.ebi.ac.uk/pub/databases/gwas/summary_statistics/PhelanCM_28346442_GCST004415/

Synapse, https://www.synapse.org/#!Synapse:syn16804296

TIGAR-V2, https://github.com/yanglab-emory/TIGAR

References

1. 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., Im, H.K.; and GTEx Consortium (2015). A gene-based association method for mapping traits using reference transcriptome data. Nat. Genet. *47*, 1091–1098.
2. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W.J.H., Jansen, R., de Geus, E.J.C., Boomsma, D.I., Wright, F.A., et al. (2016). Integrative approaches for large-scale transcriptome-wide association studies. Nat. Genet. *48*, 245–252.
3. Nagpal, S., Meng, X., Epstein, M.P., Tsoi, L.C., Patrick, M., Gibson, G., De Jager, P.L., Bennett, D.A., Wingo, A.P., Wingo, T.S., and Yang, J. (2019). TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. Am. J. Hum. Genet. *105*, 258–266.
4. Luningham, J.M., Chen, J., Tang, S., De Jager, P.L., Bennett, D.A., Buchman, A.S., and Yang, J. (2020). Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. Am. J. Hum. Genet. *107*, 714–726.
5. Yuan, Z., Zhu, H., Zeng, P., Yang, S., Sun, S., Yang, C., Liu, J., and Zhou, X. (2020). Testing and controlling for horizontal pleiotropy with probabilistic Mendelian randomization in transcriptome-wide association studies. Nat. Commun. *11*, 3861.
6. Gusev, A., Mancuso, N., Won, H., Kousi, M., Finucane, H.K., Reshef, Y., Song, L., Safi, A., McCarroll, S., Neale, B.M., et al.; Schizophrenia Working Group of the Psychiatric Genomics Consortium (2018). Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. Nat. Genet. *50*, 538–548.
7. Strunz, T., Lauwen, S., Kiel, C., Hollander, A.D., Weber, B.H.F.; and International AMD Genomics Consortium (IAMDGC) (2020). A transcriptome-wide association study based on 27 tissues identifies 106 genes potentially relevant for disease pathology in age-related macular degeneration. Sci. Rep. *10*, 1584.
8. Mancuso, N., Shi, H., Goddard, P., Kichaev, G., Gusev, A., and Pasaniuc, B. (2017). Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. Am. J. Hum. Genet. *100*, 473–487.
9. GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat. Genet. *45*, 580–585.
10. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330.

11. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al.; NBCS Collaborators; ABCTB Investigators; and ConFab/AOCS Investigators (2017). Association analysis identifies 65 new breast cancer risk loci. Nature *551*, 92–94.
12. Phelan, C.M., Kuchenbaecker, K.B., Tyrer, J.P., Kar, S.P., Lawrenson, K., Winham, S.J., Dennis, J., Pirie, A., Riggan, M.J., Chornokur, G., et al.; AOCS study group; EMBRACE Study; GEMO Study Collaborators; HEBON Study; KConFab Investigators; and OPAL study group (2017). Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. Nat. Genet. *49*, 680–691.
13. Hastie, T., Tibshirani, R., and Friedman, J. (2009). 7.10 Cross-Validation. In The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer), pp. 241–249.
14. Zeng, P., and Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. Nat. Commun. *8*, 456.
15. Tang, S., Buchman, A.S., Jager, P.L.D., Bennett, D.A., Epstein, M.P., and Yang, J. (2020). Powerful Variance-Component TWAS method identifies novel and known risk genes for clinical and pathologic Alzheimer's dementia phenotypes. bioRxiv, 2020.05.26.117515.
16. Hu, P., Lan, H., Xu, W., Beyene, J., and Greenwood, C.M. (2007). Identifying cis- and trans-acting single-nucleotide polymorphisms controlling lymphocyte gene expression in humans. BMC Proc. *1* (*Suppl 1*), S7.
17. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D., Ermel, R., Ruusalepp, A., Quertermous, T., Hao, K., et al. (2019). Opportunities and challenges for transcriptome-wide association studies. Nat. Genet. *51*, 592–599.
18. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. J. R. Stat. Soc. B *58*, 267–288.
19. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. B *82*, 1273–1300.
20. Zhou, X., Carbonetto, P., and Stephens, M. (2013). Polygenic modeling with bayesian sparse linear mixed models. PLoS Genet. *9*, e1003264.
21. Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J., Bolla, M.K., Shu, X.-O., Lu, Y., Cai, Q., et al.; NBCS Collaborators; and kConFab/AOCS Investigators (2018). A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat. Genet. *50*, 968–978.
22. Bhattacharya, A., Li, Y., and Love, M.I. (2021). MOSTWAS: Multi-Omic Strategies for Transcriptome-Wide Association Studies. PLoS Genet. *17*, e1009398.
23. 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.
24. Li, B., Liu, D.J., and Leal, S.M. (2013). Identifying Rare Variants Associated with Complex Traits via Sequencing. Curr. Protoc. Hum. Genet *78*, 1.26.1–1.26.22.
25. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet. *89*, 82–93.
26. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J., Wheeler, H.E., Torres, J.M., Torstenson, E.S., Shah, K.P., Garcia, T., Edwards, T.L., et al.; GTEx Consortium (2018). Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat. Commun. *9*, 1825.
27. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. Nature *526*, 68–74.
28. McKinney, W. (2010). Data Structures for Statistical Computing in Python. In Proceedings of the 9th Python in Science Conference (SciPy 2010) (Austin, Texas), pp. 56–61.
29. McKinney, W.; and PyData Development Team (2018). pandas: powerful Python data analysis toolkit, https://pandas.pydata.org/.
30. Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al. (2020). Array programming with NumPy. Nature *585*, 357–362.
31. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al.; SciPy 1.0 Contributors (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272.
32. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. J. Mach. Learn. Res. *12*, 2825–2830.
33. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Muller, A.C., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., et al. (2013). API design for machine learning software: experiences from the scikit-learn project. In Proceedings of the European Conference on Machine Learning and Principles and Practices of Knowledge Discovery in Databases (ECMPKDD'13), p. 122.
34. Seabold, S., and Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with Python. In Proceedings of the 9th Python in Science Conference (SciPy 2010) (Austin, Texas), pp. 92–96.
35. Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics *27*, 718–719.
36. GTEx Consortium (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. Science *348*, 648–660.
37. Battle, A., Brown, C.D., Engelhardt, B.E., Montgomery, S.B.; GTEx Consortium; Laboratory, Data Analysis &Coordinating Center (LDACC)—Analysis Working Group; Statistical Methods groups—Analysis Working Group; Enhancing GTEx (eGTEx) groups; NIH Common Fund; NIH/NCI; NIH/NHGRI; NIH/NIMH; NIH/NIDA; Biospecimen Collection Source Site—NDRI; Biospecimen Collection Source Site—RPCI; Biospecimen Core Resource—VARI; Brain Bank Repository—University of Miami Brain Endowment Bank; Leidos Biomedical—Project Management; ELSI Study; Genome Browser Data Integration &Visualization—EBI; Genome Browser Data Integration &Visualization—UCSC Genomics Institute, University of California Santa Cruz; Lead analysts; Laboratory, Data Analysis &Coordinating Center (LDACC); NIH program management; Biospecimen collection; Pathology; and eQTL manuscript working group (2017). Genetic effects on gene expression across human tissues. Nature *550*, 204–213.

38. 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.

39. Kar, S., Considine, D., Tyrer, J., Plummer, J., Chen, S., Dezem, F., Barbeira, A., Rajagopal, P., Rosenow, W., Anton, F., et al. (2020). Pleiotropy-guided transcriptome imputation from normal and tumor tissues identifies new candidate susceptibility genes for breast and ovarian cancer. bioRxiv, 2020.04.23.043653.

40. Aguet, F., Barbeira, A.N., Bonazzola, R., Brown, A., Castel, S.E., Jo, B., Kasela, S., Kim-Hellmuth, S., Liang, Y., Oliva, M., et al. (2019). The GTEx Consortium atlas of genetic regulatory effects across human tissues. bioRxiv, 787903.

41. Thomas, G., Jacobs, K.B., Kraft, P., Yeager, M., Wacholder, S., Cox, D.G., Hankinson, S.E., Hutchinson, A., Wang, Z., Yu, K., et al. (2009). A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat. Genet. *41*, 579–584.

42. Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K., et al.; Breast and Ovarian Cancer Susceptibility Collaboration; Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON); kConFab Investigators; Australian Ovarian Cancer Study Group; and GENICA (Gene Environment Interaction and Breast Cancer in Germany) Network (2013). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat. Genet. *45*, 353–361, 361e1–2.

43. Ahsan, H., Halpern, J., Kibriya, M.G., Pierce, B.L., Tong, L., Gamazon, E., McGuire, V., Felberg, A., Shi, J., Jasmine, F., et al.; Familial Breast Cancer Study (2014). A genome-wide association study of early-onset breast cancer identifies PFKM as a novel breast cancer gene and supports a common genetic spectrum for breast cancer at any age. Cancer Epidemiol. Biomarkers Prev. *23*, 658–669.

44. Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M.J., Maranian, M.J., Bolla, M.K., Wang, Q., Shah, M., et al.; BOCS; kConFab Investigators; AOCS Group; NBCS; and GENICA Network (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat. Genet. *47*, 373–380.

45. Palomba, G., Loi, A., Porcu, E., Cossu, A., Zara, I., Budroni, M., Dei, M., Lai, S., Mulas, A., Olmeo, N., et al. (2015). Genome-wide association study of susceptibility loci for breast cancer in Sardinian population. BMC Cancer *15*, 383.

46. Couch, F.J., Kuchenbaecker, K.B., Michailidou, K., Mendoza-Fandino, G.A., Nord, S., Lilyquist, J., Olswold, C., Hallberg, E., Agata, S., Ahsan, H., et al. (2016). Identification of four novel susceptibility loci for oestrogen receptor negative breast cancer. Nat. Commun. *7*, 11375.

47. Milne, R.L., Kuchenbaecker, K.B., Michailidou, K., Beesley, J., Kar, S., Lindström, S., Hui, S., Lemaçon, A., Soucy, P., Dennis, J., et al.; ABCTB Investigators; EMBRACE; GEMO Study Collaborators; HEBON; kConFab/AOCS Investigators; and NBSC Collaborators (2017). Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. Nat. Genet. *49*, 1767–1778.

48. Rashkin, S.R., Graff, R.E., Kachuri, L., Thai, K.K., Alexeeff, S.E., Blatchins, M.A., Cavazos, T.B., Corley, D.A., Emami, N.C., Hoffman, J.D., et al. (2020). Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. Nat. Commun. *11*, 4423.

49. Hoffman, J.D., Graff, R.E., Emami, N.C., Tai, C.G., Passarelli, M.N., Hu, D., Huntsman, S., Hadley, D., Leong, L., Majumdar, A., et al. (2017). Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk. PLoS Genet. *13*, e1006690.

50. Ferreira, M.A., Gamazon, E.R., Al-Ejeh, F., Aittomäki, K., Andrulis, I.L., Anton-Culver, H., Arason, A., Arndt, V., Aronson, K.J., Arun, B.K., et al.; EMBRACE Collaborators; GC-HBOC Study Collaborators; GEMO Study Collaborators; ABCTB Investigators; HEBON Investigators; and BCFR Investigators (2019). Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer. Nat. Commun. *10*, 1741.

51. Feng, H., Gusev, A., Pasaniuc, B., Wu, L., Long, J., Abu-Full, Z., Aittomäki, K., Andrulis, I.L., Anton-Culver, H., Antoniou, A.C., et al.; GEMO Study Collaborators; EMBRACE Collaborators; GC-HBOC study Collaborators; ABCTB Investigators; HEBON Investigators; BCFR Investigators; and OCGN Investigators (2020). Transcriptome-wide association study of breast cancer risk by estrogen-receptor status. Genet. Epidemiol. *44*, 442–468.

52. Shu, X., Long, J., Cai, Q., Kweon, S.-S., Choi, J.-Y., Kubo, M., Park, S.K., Bolla, M.K., Dennis, J., Wang, Q., et al. (2020). Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. Nat. Commun. *11*, 1217.

53. Couch, F.J., Wang, X., McGuffog, L., Lee, A., Olswold, C., Kuchenbaecker, K.B., Soucy, P., Fredericksen, Z., Barrowdale, D., Dennis, J., et al.; kConFab Investigators; SWE-BRCA; Ontario Cancer Genetics Network; HEBON; EMBRACE; GEMO Study Collaborators; BCFR; and CIMBA (2013). Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. PLoS Genet. *9*, e1003212.

54. Kuchenbaecker, K.B., Ramus, S.J., Tyrer, J., Lee, A., Shen, H.C., Beesley, J., Lawrenson, K., McGuffog, L., Healey, S., Lee, J.M., et al.; EMBRACE; GEMO Study Collaborators; Breast Cancer Family Registry; HEBON; KConFab Investigators; Australian Cancer Study (Ovarian Cancer Investigators); Australian Ovarian Cancer Study Group; and Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (2015). Identification of six new susceptibility loci for invasive epithelial ovarian cancer. Nat. Genet. *47*, 164–171.

55. Lu, Y., Beeghly-Fadiel, A., Wu, L., Guo, X., Li, B., Schildkraut, J.M., Im, H.K., Chen, Y.A., Permuth, J.B., Reid, B.M., et al. (2018). A transcriptome-wide association study among 97,898 women to identify candidate susceptibility genes for epithelial ovarian cancer risk. Cancer Res. *78*, 5419–5430.

56. Gusev, A., Lawrenson, K., Lin, X., Lyra, P.C., Jr., Kar, S., Vavra, K.C., Segato, F., Fonseca, M.A.S., Lee, J.M., Pejovic, T., et al.; Ovarian Cancer Association Consortium (2019). A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. Nat. Genet. *51*, 815–823.

57. Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. *47* (D1), D1005–D1012.

58. Cai, Q., Zhang, B., Sung, H., Low, S.-K., Kweon, S.-S., Lu, W., Shi, J., Long, J., Wen, W., Choi, J.-Y., et al.; DRIVE GAME-ON Consortium (2014). Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. Nat. Genet. *46*, 886–890.

59. Cruts, M., Rademakers, R., Gijselinck, I., van der Zee, J., Dermaut, B., de Pooter, T., de Rijk, P., Del-Favero, J., and van Broeckhoven, C. (2005). Genomic architecture of human 17q21 linked to frontotemporal dementia uncovers a highly homologous family of low-copy repeats in the tau region. Hum. Mol. Genet. *14*, 1753–1762.

60. Stefansson, H., Helgason, A., Thorleifsson, G., Steinthorsdottir, V., Masson, G., Barnard, J., Baker, A., Jonasdottir, A., Ingason, A., Gudnadottir, V.G., et al. (2005). A common inversion under selection in Europeans. Nat. Genet. *37*, 129–137.

61. de Jong, S., Chepelev, I., Janson, E., Strengman, E., van den Berg, L.H., Veldink, J.H., and Ophoff, R.A. (2012). Common inversion polymorphism at 17q21.31 affects expression of multiple genes in tissue-specific manner. BMC Genomics *13*, 458.

62. Mosquera Orgueira, A. (2015). Hidden among the crowd: differential DNA methylation-expression correlations in cancer occur at important oncogenic pathways. Front. Genet. *6*, 163.

63. Xu, J., Qian, Y., Ye, M., Fu, Z., Jia, X., Li, W., Xu, P., Lv, M., Huang, L., Wang, L., et al. (2016). Distinct expression profile of lncRNA in endometrial carcinoma. Oncol. Rep. *36*, 3405–3412.

64. Zhang, C., Liu, J., Huang, G., Zhao, Y., Yue, X., Wu, H., Li, J., Zhu, J., Shen, Z., Haffty, B.G., et al. (2016). Cullin3-KLHL25 ubiquitin ligase targets ACLY for degradation to inhibit lipid synthesis and tumor progression. Genes Dev. *30*, 1956–1970.

65. Luo, P., Ding, Y., Lei, X., and Wu, F.-X. (2019). deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks. Front. Genet. *10*, 13.

66. Zhu, Z., Fu, H., Wang, S., Yu, X., You, Q., Shi, M., Dai, C., Wang, G., Cha, W., and Wang, W. (2020). Whole-exome sequencing identifies prognostic mutational signatures in gastric cancer. Ann. Transl. Med. *8*, 1484.

67. Zhang, H., Ahearn, T.U., Lecarpentier, J., Barnes, D., Beesley, J., Qi, G., Jiang, X., O'Mara, T.A., Zhao, N., Bolla, M.K., et al.; kConFab Investigators; ABCTB Investigators; EMBRACE Study; and GEMO Study Collaborators (2020). Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. Nat. Genet. *52*, 572–581.

68. Xu, N., Chen, F., Wang, F., Lu, X., Wang, X., Lv, M., and Lu, C. (2015). Clinical significance of high expression of circulating serum lncRNA RP11-445H22.4 in breast cancer patients: a Chinese population-based study. Tumour Biol. *36*, 7659–7665.

69. Lau, T.P., Roslani, A.C., Lian, L.H., Chai, H.C., Lee, P.C., Hilmi, I., Goh, K.L., and Chua, K.H. (2014). Pair-wise comparison analysis of differential expression of mRNAs in early and advanced stage primary colorectal adenocarcinomas. BMJ Open *4*, e004930.

70. Liu, Y., Walavalkar, N.M., Dozmorov, M.G., Rich, S.S., Civelek, M., and Guertin, M.J. (2017). Identification of breast cancer associated variants that modulate transcription factor binding. PLoS Genet. *13*, e1006761.

71. Masoodi, T.A., Banaganapalli, B., Vaidyanathan, V., Talluri, V.R., and Shaik, N.A. (2017). Computational Analysis of Breast Cancer GWAS Loci Identifies the Putative Deleterious Effect of STXBP4 and ZNF404 Gene Variants. J. Cell. Biochem. *118*, 4296–4307.

72. Du, Z., Gao, W., Sun, J., Li, Y., Sun, Y., Chen, T., Ge, S., and Guo, W. (2019). Identification of long non-coding RNA-mediated transcriptional dysregulation triplets reveals global patterns and prognostic biomarkers for ER+/PR+, HER2- and triple negative breast cancer. Int. J. Mol. Med *44*, 1015–1025.

73. Xia, J., Inagaki, Y., Sawakami, T., Song, P., Cai, Y., Hasegawa, K., Sakamoto, Y., Akimitsu, N., Tang, W., and Kokudo, N. (2016). Preliminary investigation of five novel long non-coding RNAs in hepatocellular carcinoma cell lines. Biosci. Trends *10*, 315–319.

74. Cao, C., Ding, B., Li, Q., Kwok, D., Wu, J., and Long, Q. (2021). Power analysis of transcriptome-wide association study: Implications for practical protocol choice. PLoS Genet. *17*, e1009405.

75. 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.

76. Liu, L., Zeng, P., Xue, F., Yuan, Z., and Zhou, X. (2021). Multi-trait transcriptome-wide association studies with probabilistic Mendelian randomization. Am. J. Hum. Genet. *108*, 240–256.