Enhancing Disease Risk Gene Discovery by Integrating Transcription Factor-Linked Trans-located Variants into Transcriptome-Wide Association Analyses

Authors

Jingni He,[1] Wanqing Wen,[2] Jie Ping,[2] Qing Li,[1] Zhishan Chen,[2] Deshan Perera,[1] Xiang Shu,[3] Jirong Long,[2] Qiuyin Cai,[2] Xiao-Ou Shu,[2] Wei Zheng,[2] Quan Long,[1,4,5,6,7]* Xingyi Guo[2,8*]

Affiliations:

[1] Department of Biochemistry & Molecular Biology, University of Calgary, Calgary, Canada.

[2] Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN, USA.

[3] Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, NY, USA

[4] Department of Medical Genetics, University of Calgary, Calgary, Canada

[5] Department of Mathematics & Statistics, University of Calgary, Calgary, Canada

[6] Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Canada

[7] Hotchkiss Brain Institute, University of Calgary, Calgary, Canada

[8] Department of Biomedical Informatics, Vanderbilt University School of Medicine, Nashville, TN, USA

*Corresponding authors:

Dr. Xingyi Guo

Vanderbilt University School of Medicine

Email: xingyi.guo@vumc.org

Dr. Quan Long

University of Calgary

Email: quan.long@ucalgary.ca

Abstract

Transcriptome-wide association studies (TWAS) have been successful in identifying putative disease susceptibility genes by integrating gene expression predictions with genome-wide association studies (GWAS) data. However, current TWAS models only consider cis-located variants to predict gene expression. Here, we introduce transTF-TWAS, which includes transcription factor (TF)-linked trans-located variants for model building. Using data from the Genotype-Tissue Expression project, we predict alternative splicing and gene expression and applied these models to large GWAS datasets for breast, prostate, and lung cancers. Our analysis revealed 887 putative cancer susceptibility genes, including 465 in regions not yet reported by previous GWAS and 137 in known GWAS loci but not yet reported previously, at Bonferroni-corrected $P < 0.05$. We demonstrate that transTF-TWAS surpasses other approaches in both building gene prediction models and identifying disease-associated genes. These results have shed new light on several genetically driven key regulators and their associated regulatory networks underlying disease susceptibility.

Introduction

Approximately 90% of risk variants identified in genome-wide association studies (GWAS) are located in noncoding or intergenic regions, which suggests that they may affect cancer risk by dysregulating gene expression [1-10]. Fine-mapping of genetic risk loci, along with functional experiments provide strong evidence that regulatory variants in linkage disequilibrium (LD) with GWAS-identified risk variants disrupt DNA binding affinities of specific transcription factors (TFs) and modulate expression of susceptibility genes [11-22]. Thus, identifying TFs, whose DNA bindings are altered by risk-associated genetic variations, and their controlling genes can greatly improve the understanding of transcriptional dysregulation in human diseases and cancers[23-26]. A pioneering study analyzed Chromatin immunoprecipitation followed by sequencing (ChIP-seq) data for TFs such as FOXA1 in multiple breast cancer cell lines to investigate the binding of GWAS-identified risk variants [21]. The study suggested that regulatory variants confer to breast cancer risk by mediating their altered binding affinities. Subsequent studies have revealed multiple breast cancer risk-associated TFs such as ESR1, MYC, and KLF4 [20,27] through interrogating data on gene expression, TF ChIP-seq, and GWAS-identified risk variants. We have recently conducted a comprehensive analysis of TF ChIP-seq and GWAS data for breast cancer, and developed an analytical framework to identify TFs that contribute to breast cancer risk. Our study revealed that the genetic variations of 22 TFs were significantly associated with breast cancer risk and highlighted genetic variations of TF-DNA bindings (particularly for FOXA1) underlying breast cancer susceptibility [28].

Transcriptome-wide association studies (TWAS) have successfully uncovered large numbers of putative susceptibility genes for cancers and other diseases, and many of these genes have been further supported by functional experiments [29-33]. In TWAS, a reference with both transcriptome and high-density genotyping data from a small set of subjects, such as the Genotype-Tissue Expression project (GTEx), is used to build prediction models of gene expression for downstream association analyses. However, the accuracy of gene expression prediction predicted by cis-genetic variants could be compromised if these variants are located in non-regulatory elements or if they disrupt binding sites of non-transcribed TFs in target tissues [34-37]. Our recent approach, sTF-TWAS [38], by integrating susceptible TF-occupied cis-regulatory

elements (STFCREs) from risk-associated TFs significantly improve the detection of cancer susceptibility genes compared to the conventional TWAS approaches.

Despite the progress made by TWAS in recent years [37-43], the current models for predicting gene expression were based solely on cis-located genetic variants (<1Mb distance), which generally account for only a modest proportion of disease heritability [44]. In comparison, trans-located genetic variants may have more impact on the disease phenotype due to their advantages in population selection pressure and compensatory post-transcriptional buffering [45,46]. However, including trans-located variants in TWAS analysis is a challenge due to their overwhelming numbers on gene expression than cis-located variants, requiring larger sample sizes for detection [47]. Therefore, an integrative epigenetic data approach is needed to prioritize trans-located variants that may play a regulatory role in gene expression.

In this work, we introduced transTF-TWAS, which included TF-linked trans-located variants, together with cis-regulatory variants for prediction model building in an effort to improve susceptible gene discovery. We showed that transTF-TWAS outperforms other methods by significantly improving prediction models and identifying disease genes. In particular, we conducted transTF-TWAS to analyze both gene expression and alternative splicing with data generated from multiple normal tissues from the Genotype-Tissue Expression (GTEx) and large-scale GWAS data for breast, prostate, and lung cancers and three brain disorders to search for disease susceptibility genes and loci (Table S1).

Results

Overview of transTF-TWAS framework

We introduced our new approach, transTF-TWAS, to build gene expression prediction models by adding TF-linked trans-located variants together with cis-variants under our previous sTF-TWAS framework. As illustrated using data in breast cancer, we firstly identified putative TF cis-regulatory variants that potentially affect expression of a TF (e.g., FOXA1) by conducting cis-eQTL analysis and analyzing epigenetic data generated in breast cancer-related cells. A set of the cis-regulatory variants regulating TF expression (namely TF-cis-regulatory-variants) was

determined based on the significant associations between TF gene expression and genetic variants, as well as regulatory evidence for these variants through interactions with proximal promoters or distal enhancer-promoter regions (Fig. 1A; Materials and Methods).

Secondly, we analyzed TF ChIP-seq data generated in breast cancer-related cells to characterize their genome-wide binding sites for susceptible TFs which have been identified in breast cancer from our prior work [28] (Fig. 1B; Materials and Methods). We next characterized each gene potentially regulated by all possible susceptible TFs based on the evidence of the TF-DNA binding sites that are located in its flanking of transcription start sites (TSS, +/-20K; Fig. 1C). As TF-cis-regulatory-variants have the potential to modulate the expression of the TF protein, which may result in changes in gene expression of downstream targets. Thus, these genetic variations may affect genes regulated by the TF, even if they are located several megabases away or on different chromosomes. To test this, for each TF, we assessed the performance of a prediction model that utilized its TF-cis-regulatory-variants to predict expression of each target gene using Group Lasso method (i.e., number of G TFs; Fig. 1C; Materials and Methods). The Group Lasso's property of encouraging between-group sparsity and within-group retainment aligns to our intention of selecting the actual functioning TFs and then retaining their cis-regulatory-variants. The groups survive the regularization are corresponding to those of TF-cis-regulatory-variants that may affect the expression of the gene. The final set of TF-cis-regulatory variants was identified for downstream gene expression model building by combining the groups from the significant models using standard Elastic Net (Fig. 1C; Table S1; Materials and Methods).

Lastly, by expending our previous sTF-TWAS framework to build gene expression prediction models using the prioritized 50K cis-located variants (Fig. 1B), we included TF-cis-regulatory-variants (as trans-located variants) identified from the above analysis. Here, we only focused on the set with 50K cis-located variants, as the identified genes were highly overlapped among analyses with different number of variants (i.e., 50K vs 500K variants) in our prior work [28,38]. The GTEx reference data were primarily used to build genetically predicted gene expression in four tissues, including breast, prostate, lung and brain (Materials and Methods; Fig. 1D). We conducted TWAS analyses by applying the gene expression prediction models, respectively, to

GWAS summary statistics for breast, prostate, and lung cancers and other diseases to search for their susceptibility genes and loci (Fig. 1D).

transTF-TWAS outperforms existing TWAS approaches

To evaluate the performance of transTF-TWAS, we conducted simulations under an extension of our sTF-TWAS framework[38], sTF-TWAS(R), by adding randomly selected trans-located variants of the equal number (Materials and Methods). We first compared prediction performance of gene expression built from transTF-TWAS with sTF-TWAS(R) and sTF-TWAS. In the analysis using data from breast tissues in GTEx, we observed that transTF-TWAS predicted slightly more genes than sTF-TWAS(R) at a cutoff of $R^2 > 0.01$, while it predicted over 2,000 genes more than sTF-TWAS (Fig. 1E). Using independent datasets generated in breast normal tissues from KOME (n = 181), we further showed that a higher proportion of these predicted genes in GTEx were verified in transTF-TWAS when compared to two alternative approaches (Fig. S1). By further applying the prediction models to GWAS data in breast cancer, we identified 141 putative susceptibility genes using transTF-TWAS, at a Bonferroni-corrected $P < 0.05$, which were more than those identified by sTF-TWAS (62 genes), and sTF-TWAS(R) (41 genes) (Fig. 1E).

We next expanded our comparisons for transTF-TWAS with existing approaches including sTF-TWAS[38], PUMICE[43] and S-PrediXcan[30], in the analysis for breast, prostate and lung cancers. We showed that transTF-TWAS identified more genes than sTF-TWAS under multiple *P*-value cutoffs (Fig. 2A). As described above, we identified 141 putative susceptibility genes from transTF-TWAS, at a Bonferroni-corrected $P < 0.05$, while fewer genes were identified by sTF-TWAS (n=62), S-PrediXcan (n=52), and PUMICE (n=42) (Fig. S2; Fig. S3; Table S2). We conducted similar comparisons for prostate and lung cancers and demonstrated consistent trends of more genes identified by transTF-TWAS compared to the other three approaches (Fig. S2; Fig. S3; Table S3; Table S4).

We further performed functional annotation for the genes identified by transTF-TWAS and sTF-TWAS with known target cancer-related genes of interest (Materials and Methods). We showed that more (i.e., n=61 for transTF-TWAS vs. n=26 for sTF-TWAS) (Fig. 2B) and a

comparable proportion (i.e., 62.2% for transTF-TWAS vs. 61.9% for sTF-TWAS) of breast cancer-related genes were detected by transTF-TWAS than those identified by other approaches (Fig. 2C). We also found an overall higher quantity and higher or comparable proportion of known cancer related genes identified for both prostate and lung cancers by our approach than other approaches (Fig. 2B,C).

Genetically driven key regulators and their associated networks underlying cancer risk

We showed that transTF-TWAS detected more genes than sTF-TWAS in breast, prostate, and lung cancers, whereas a large number of significant genes were uniquely detected by transTF-TWAS (Fig. 3A). To further illustrate how these unique genes contributed by trans-located variants, we examined whether these genes can be predictable by sTF-TWAS. We found that most of the unique genes in breast and prostate cancers failed to be genetically predicted by sTF-TWAS, indicating the trans-located variants significantly contributed risk gene discovery via the improved gene expression prediction performance (Fig. 3A; Table S5; Table S6; Table S7).

Of the identified genes, we next evaluated the lead trans-located variants that present the strongest associations with cancer risk in the prediction model for each of our identified putative susceptibility genes. In breast cancer, we observed that the lead variants are significantly enriched in the TF-cis-regulatory-variants for ESR1 (n= 73 genes), followed by TCF7L2 (n=13), and FOXA1 (n=11) (Fisher's exact test, $P < 0.01$ for all; Fig. 3B, C; Table S8; Table S9). In prostate cancer, we observed that the lead variants are significantly enriched in NKX3-1 (n= 61 genes), followed by GATA2 (n=13) (Fisher's exact test, $P < 0.01$ for all; Table S8; Table S10). These results highlighted these genetically-driven key regulators and their associated regulatory networks that underlie cancer susceptibility.

Charactering novel cancer risk genes identified by transTF-TWAS

To further characterize putative susceptibility genes and loci identified under our transTF-TWAS framework, we additionally analyzed alternative splicing (sp-transTF-TWAS) for breast, prostate and lung cancers (Fig. 4; Fig. S4; Table S11; Table S12; Table S13). We comprehensively compared our findings from both transTF-TWAS and sp-transTF-TWAS, with

those reported from previous TWAS, eQTL, or other genetic studies for breast[32,48; 19; 28,38,49,50], prostate[31,38,51-53] and lung cancers[38,53,54].

For breast cancer, we identified 141 putative susceptibility genes from transTF-TWAS and 239 putative susceptibility genes from sp-transTF-TWAS, at a Bonferroni-corrected $P < 0.05$. Combing the results from both analyses, we identified 374 putative breast cancer susceptibility genes, including 212 genes at 163 novel loci (more than 1Mb away from any previous GWAS-identified risk variant for breast cancer) and 53 previously unreported located in GWAS loci (Fig. 5A, B; Table S14). For prostate cancer, we identified 136 putative susceptibility genes from transTF-TWAS and 318 putative susceptibility genes from sp-transTF-TWAS. Combing the results from both analyses, we identified 443 putative prostate cancer susceptibility genes, including 251 genes at 193 novel loci and 75 genes previously unreported located in GWAS loci (Fig. 5A, B; Table S15). For lung cancer, we identified 36 putative susceptibility genes from transTF-TWAS and 41 putative susceptibility genes from sp-transTF-TWAS. Combing the results from both analyses, we identified 70 putative lung cancer susceptibility genes, including 2 genes at one novel locus and 9 genes previously unreported located in GWAS loci (Fig. 5A,5B; Table S16). Taken together, our analysis revealed total 887 putative susceptibility genes for these three cancer types, including 137 that were previously unreported in GWAS loci and 465 in loci unreported by GWAS (Table S17; Table S18).

Functional evidence of oncogenic roles for the identified putative susceptibility genes

We next examined whether our identified putative cancer susceptibility genes had been reported as predisposition genes[55,56], cancer drivers [57,58], or Cancer Gene Census (CGC) genes [59] (Materials and Methods). We found eight cancer driver genes and five CGC among previously unreported genes for breast cancer, as well as six cancer driver genes and eight CGC among previously reported genes (Fig. 5C). Similarly, for prostate cancer, we found ten cancer driver genes and eight CGC among previously unreported genes, and six cancer driver genes and four CGC among previously reported genes (Fig. 5C). For lung cancer, we identified four cancer driver genes and three CGC among previously reported genes (Fig. 5C). Functional enrichment analysis showed that our identified genes were significantly enriched in those known cancer-

related genes with P = 0.0044 for breast cancer, P = 0.0097 for prostate cancer and P = 0.012 for lung cancer (Materials and Methods).

We also explored the functional roles of the identified putative susceptibility genes using CRISPR-Cas9 screen silencing data to investigate gene essentiality on cell proliferation in breast (n=45), prostate (n=8), and lung (n=130) cancer relevant cell lines (Materials and Methods). Using a cutoff of median CERES Score < −0.5 in the above cells, following the previous literature [60,61], we provided strong evidence of essential roles in cell proliferation for 19 previously unreported genes for breast cancer (Fig. 5D); 36 unreported genes for prostate cancer (Fig. 5E); and two unreported genes for lung cancer (Fig. 5F).

TransTF-TWAS strengthens non-cancer risk gene discovery

To evaluate the generalizability of transTF-TWAS, we conducted additional analysis for brain disorders including schizophrenia (SCZ), Alzheimer's disease (AD), and autism spectrum disorder (ASD). By comparison, we also conducted S-PrediXcan and sTF-TWAS for each of the diseases. We found that transTF-TWAS identified more putative susceptibility genes than both sTF-TWAS and S-PrediXcan for AD and ASD. Using ASD as an example, we identified eight putative susceptibility genes from transTF-TWAS at a Bonferroni-corrected $P < 0.05$, while only one and six genes was identified by S-PrediXcan and sTF-TWAS, respectively. The results suggest that our transTF-TWAS approach has broad applicability for enhancing the discovery of disease susceptibility genes (Fig. S5).

Discussion

In this study, we demonstrated that the new approach, transTF-TWAS, significantly improved the detection of putative cancer susceptibility genes with increased statistical power and accuracy over other existing TWAS approaches (i.e. sTF-TWAS[38], PUMICE[43], and S-PrediXcan[30]). Under transTF-TWAS framework, we predicted alternative splicing and gene expression and applied these models to large GWAS datasets for breast, prostate, and lung cancers. Our analysis revealed total 887 putative cancer susceptibility genes, including 465 in regions not yet reported by previous GWAS and 137 in known GWAS loci but not yet reported previously. Many of the newly identified associations have been supported by their oncogenic

roles in cancer development [57-59], including 88 cancer driver genes, CGC or those with strong evidence of essential roles in target cancer cell proliferation. These findings provide new insights into the genetic susceptibility of the three common cancers.

Previous TWAS mainly use cis-genetic variants in building gene expression models. However, investigation into trans-located variants has been limited due to statistical analysis burden of their large numbers. To address this, transTF-TWAS is to identify TF-linked trans-located variants by comprehensively characterizing TF-cis-regulatory-variants and using Group Lasso to select a set of these variants to significantly contribute prediction models. The use of Group Lasso can be powerful in identifying a set of TF-cis-regulatory-variants, which may affect the expression of the TF regulated target genes. Our analysis identified many TF-linked trans-located variants (i.e., average 10 for each gene for breast cancer) contributed to gene expression prediction, which included several thousand newly predicted genes missed by other approaches. We observed that the newly predicted genes had a similar proportion of verifiability in independent datasets compared to the remaining genes (Fig. S1). Of note, in our prior study of sTF-TWAS, it is demonstrated that the approach improves statistical power compared to existing approaches and the Type-I Error is well under control [38]. By conducting simulations and real data analysis, we further showed that transTF-TWAS predicted a higher proportion of verifiable genes and detected more significant genes with higher accuracy compared to sTF-TWAS, indicating the strong validity of our method (Fig. S1).

Much efforts [20,21,27], including our work[28], have established cancer susceptible TFs, whose DNA binding sites altered by risk genetic variants that regulate cancer susceptibility genes. However, it remains unclearly how susceptibility TF-based transcriptional networks underlying genetic susceptibility to common cancers. In this study, transTF-TWAS can strengthen susceptibility gene discovery through integrating the prior information of TF-cis-regulatory-variants altered regulators and their downstream target genes. For breast cancer, we observed that the variants potentially regulated TFs significantly contribute to expression prediction of their downstream regulated genes. In turn, the putative susceptible genes that were identified appear to be commonly regulated by FOXA1 and ESR1 through the upstream genetically-driven regulatory mechanisms, further highlighting their key roles in driving breast

cancer susceptibility (Fig. 3; Table S9). Similarly, we also highlighted key TFs, NKX3-1 [62-65] and GATA2 [66-69] for prostate cancer (Table S10). Unfortunately, we did not observe significant TFs in lung cancer, likely due to the less genetic effects of the TF on downstream regulated genes [38].

Using the transTF-TWAS framework, we also conducted sp-transTF-TWAS for breast, prostate, and lung cancers. In line with previous work[38], our results also suggested that genetically regulated alternative splicing significantly contribute to cancer risk. We demonstrated that sp-transTF-TWAS improved the detection of cancer susceptibility genes with increased statistical power and accuracy over S-PrediXcan and sTF-TWAS (Table S11; Table S12; Table S13).

Our findings are in line with the evidence that trans-located genetic variants may have more impact on diseases than cis-located genetic variants[70-72]. For our transTF-TWAS, one critical step is to prioritize TF-linked trans-located variants for model building based on identifying TF-cis-regulatory-variants and TF regulated downstream genes. The concept of our approach can also be used to identify trans-located variants based on long distance-based epigenetic signals (>1Mb) such as distal chromatin-chromatin interaction, enhancer-gene link, enhancer-gene correlation as well as trans-eQTLs [73-76]. Thus, our transTF-TWAS will further strengthen disease susceptibility gene discovery with increasing availability of extensive epigenetic datasets in future studies. On the other hand, the prior information of TF-linked trans-located variants can be integrated into other extensions of TWAS, such as multiple-tissue approaches (UTMOST [77] and S-MultiXcan [78]), or variance component (Kernel) TWAS [39,40] or instrumental-variable approaches [41,42].

In conclusion, we demonstrated that our transTF-TWAS, by integrating TF-linked trans-located variants with TWAS, significantly improved disease susceptibility gene discovery and advanced our understanding of complex human diseases, including cancers. Our study also highlighted several genetically driven key regulators and their associated regulatory networks underlying disease susceptibility.

Materials and methods

Data resources

We obtained the individual-level genotype dataset from GTEx (v8)[79,80], which was quality controlled using PLINK[81]. Summary statistics of GWAS data for breast cancer were obtained from the Breast Cancer Association Consortium (BCAC), which has generated GWAS data for 122,977 cases and 105,974 controls from European descendants. GWAS data for prostate cancer were released from the European descendants were released from the Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome (PRACTICAL)[82], with 79,194 cases and 61,112 controls from European descendants. GWAS data for lung cancer were obtained from the websites of the Transdisciplinary Research of Cancer in Lung of the International Lung Cancer Consortium (TRICL-ILCCO) and the Lung Cancer Cohort Consortium (LC3) [83], with 29,266 cases and 56,450 controls from European descendants. GWAS summary statistics for schizophrenia (SCZ, N= 70,100), Alzheimer's disease (AD, N=22,246), and autism spectrum disorder (ASD, N= 10,263) were downloaded from the Psychiatric Genomics Consortium website (PGC) (Fig. 1D).

The TF-occupied regulatory variants for breast, prostate, lung cancers, and three brain disorders were collected based on ChIP-seq data of transcription factors (TFs) generated in diseases related cell lines from the Cistrome database[84]. We evaluated their quality control based on the guidance from the database and selected high-quality datasets for downstream analysis. Detailed ChIP-seq data for breast, prostate, lung cancers and brain disorders were described in our previous work [28,38].

We included germline whole genome sequencing (WGS) and RNA-sequencing (RNA-seq) data from GTEx (release 8) [79,80] for normal breast tissue, prostate tissue, lung tissue, and brain cortex tissue. We selected tissue samples from 151 women for breast tissue, 221 men for prostate tissue, 515 individuals for lung tissue, and 205 individuals (both sexes) for brain cortex tissue. The fully processed, filtered, and normalized gene expression data matrices (in BED format) were downloaded from the GTEx portal. The WGS file and sample attributes were obtained from dbGaP, and the subject phenotypes for sex and age information were obtained from the GTEx portal. The covariates used in eQTL analysis were obtained from

GTEx_Analysis_v8_eQTL_covariates.tar.gz, and the covariates for sQTL analysis were obtained from GTEx_Analysis_v8_sQTL_covariates.tar.gz, both of which were downloaded from the GTEx portal. Normal breast tissue samples for both RNA-sequencing and genotyping from 181 individuals of European were collected through the Susan G. Komen Normal Tissue Bank (KOME). Genotype and gene expression data generation and processing have been described in our previous sTF-TWAS work [38].

We downloaded approximately 3.6 million DNase I hypersensitive sites (DHSs) regions within human genome sequence [85]. The enhancers regions were downloaded from EpiMap repository [86], which contains ~2M non-tissue specific enhancers regions. The CAGE peak regions were downloaded from FANTOM5 [87], and we also included all regions within transcription start site (TSS) +/-2K for each gene as promoter regions. The eQTLs were downloaded from the GTEx portal [79,80] and eQTLGen [88]. The Enhancer to gene link information across 833 cell-types were downloaded from EpiMap repository [86]. We all used cell-type specific chromatin-chromatin interaction data from the 4D genomics and previous literature [89,90].

To analyze cancer-related susceptibility genes, we downloaded a list of gene sets from the Molecular Signatures Database (MGB) on Gene Set Enrichment Analysis (GSEA). Additionally, we downloaded lists of predisposition genes from previous literatures [56,91], cancer-driven genes from two previous literatures [92,93], and CGC [94] from the COSMIC website. To investigate the effect of an individual gene on essentiality for the proliferation and survival of cancer cells, we downloaded two comprehensive datasets, "sample_info.csv" and "CRISPR_gene_effect.csv," from DepMap Public 21Q4.

Identifying TF-cis-regulatory-variants

To determine a set of the cis-regulatory variants that potentially regulate TF expression (namely TF-cis-regulatory-variants), we first prioritized putative regulatory variants by only including TF-occupied variants that are located in DNase I hypersensitive sites (DHSs) [85], enhancer regions[86] and promoter regions [86]. Of them, the significant associations between a TF and its cis-genetic variants were identified at a nominal p-value < 0.05, based on the eQTL analysis in both target tissues and whole blood samples using data from GTEx portal[79,80] and

eQTLGen[88]. Furthermore, we also analyzed epigenetic data to search regulating evidence by these variants through interactions with proximal promoters or distal enhancer-promoter regions. Specifically, we examined if these variants are located in the promoter region of a TF (TSS +/- 2K) or enhance region with an evidence of the enhancer linking to the TF based on expression-enhancer activity correlation across 833 cell-types from the EpiMap repository [86], as well as chromatin-chromatin interaction data from the 4D genomics and previous literature [89,90]. Finally, the TF-cis-regulatory-variants were identified based on the significant associations from eQTL results, and the regulatory evidence from the variants linked to the TF.

Gene expression prediction model building based on trans-located variants

We analyzed TF ChIP-seq data generated in target cancer-related cells to characterize their genome-wide binding sites for susceptible TFs using data from the Cistrome database [84]. We next characterized each gene potentially regulated by all possible susceptible TFs based on the evidence of their TF-DNA binding sites that are located in its flanking 20Kb of TSS (i.e., number of *G* TFs; Fig. 1C). For each TF, we assessed the performance of a prediction model that utilized its TF-cis-regulatory-variants to predict expression of each target gene using Group Lasso method. We trained a Group Lasso to select a group of TF-cis-regulatory-variants from each TF (i.e., 1 to G TF).

$$\boldsymbol{Loss}(\boldsymbol{\beta}^{*}) = \boldsymbol{argmin}||\boldsymbol{y} - \boldsymbol{X\beta}||_2^2 + \boldsymbol{\lambda}\sum_{\boldsymbol{g=1}}^{\boldsymbol{G}}||\boldsymbol{\beta_g}||_2$$

where the coefficient in $\boldsymbol{\beta}$ are divided into *G* groups and $\boldsymbol{\beta_g}$ denotes the coefficients vector of variants in the *g*-th group. $\boldsymbol{X}$ are all trans-located variants from *G* groups. $\boldsymbol{y}$ is normalized gene expression data generated in different tissue samples from GTEx v8. In Group Lasso, the regularizor, $||\boldsymbol{\beta_g}||_2$, also called $l_{2,1}$-norm consists of the intra-group non-sparsity via l_2-norm and inter-group sparsity via l_1-norm. Only significant models were used to determine those groups of TF-cis-regulatory-variants that may affect the expression of the gene. The final set of TF-cis-regulatory variants was identified for downstream gene expression model building by combining the groups from the significant models. We next built gene expression prediction models for the final sets of TF-cis-regulatory variants and cis TF-occupied variants using standard Elastic Net under our sTF-TWAS framework. For each gene, the gene expression level was regressed on the number of effect alleles (0, 1, or 2) for each genetic variant with adjustment

for top 5 genotyping PCs, age, and other potential confounding factors (PEERs). We used 30 PEER factors for our downstream model building based on the recommendation for breast, prostate and brain tissues, and 60 PEER factors for lung tissue. Prediction model performance was assessed using the R^2 via a 10-fold cross-validation.

Simulation study and external verification of gene expression predictions

To evaluate prediction performance of our developed approach, we simulated scenario for each gene that had the equal number of artificial TF groups with our transTF-TWAS. We also randomly generated the same number of trans-located genetic variants (> 1Mb distance) within each TF group with our transTF-TWAS. Similarly, we next used Group Lasso to select significant groups from the artificial TF groups. The final set of trans-located variants was identified for downstream gene expression model building by combining the groups from the significant models. We next built gene expression prediction models for the final sets of and cis TF-occupied variants under sTF-TWAS framework. The models of genetically predicted gene expression were built in breast normal tissues from the GTEx project. To externally verify gene expression prediction performance, we first used the same analytical protocol to build the prediction models using standard Elastic Net based on normalized gene expression data generated in breast tissue from the GTEx (v8), and then we re-calculated the prediction performances in terms of variance explained (R^2) using selected variants trained from the GTEx based on an independent dataset generated in breast normal tissues from KOME, where the genotype and gene expression data were processed following the protocol in GTEx.

Association analyses between predicted gene expression and cancer risk

To evaluate associations of genetic predicted gene expression with cancer risk, we applied the weight matrix obtained from the gene prediction models to the summary statistics implemented in S-PrediXcan [95]**.** The statistical method described in the following equation that was also described elsewhere[32,33], was used for association analyses.

$$Z_g \approx \sum_{l \in \mathrm{Model}_g} w_{lg} \frac{\hat{\sigma}_l}{\hat{\sigma}_g} \frac{\hat{\beta}_l}{\mathrm{se}(\hat{\beta}_l)}$$

Here, Z-score was used to estimate the association between predicted gene expression and cancer risk. Here, w_{lg} is the weight of genetic variant l for predicting the expression of gene

g. $\hat{\beta}_l$and $\mathrm{se}(\hat{\beta}_l)$ are the GWAS-reported regression coefficients, and its standard error for variant l, and $\hat{\sigma}_l$ and $\hat{\sigma}_g$ are the estimated variances of variant l and the predicted expression of gene g, respectively.

By comparison, we also performed TWAS analysis using PUMICE[43] (Prediction Using Models Informed by Chromatin conformations and Epigenomics) with default settings. PUMICE improves the accuracy of transcriptomic imputation through utilizing tissue-specific 3D genomic and epigenomic data to prioritize regions that harbor cis-regulatory variants. The source codes of PUMICE were obtained from https://github.com/ckhunsr1/PUMICE. The precomputed models trained in breast, prostate and lung tissues from GTEx v8 can be found under https://github.com/ckhunsr1/PUMICE/tree/master/models_GTEx_v8.

Genetically-driven key regulators and their associated networks regulating breast cancer susceptibility genes

For each of the identified putative susceptibility genes, we evaluated the lead variant that present the strongest associations with cancer risk in the prediction model. If the lead variant was a trans-located variant, we next identified its potential regulated TF based on the previous analysis of TF-cis-regulatory-variant (see the preceding section). A TF-gene pair was further determined based on the above information of the lead trans-located variant linked to both the gene and TF. Based on the information of TF-gene pairs, a TF-transcriptional network was built using Cytoscape 3.9.1[96]. To evaluate whether our identified susceptibility genes significantly enriched in a TF of interest, we conducted a comparison between this TF and the remaining combined TFs as background, using Fisher's exact test.

Annotation of the identified genes using cancer-related gene database

To verify the evidence whether the TWAS-identified genes are related to cancer susceptibility, we extracted cancer related gene sets from the MGB database. Putative cancer related genes were characterized based on their annotation with the key words 'breast cancer', 'prostate cancer' and 'lung cancer'. We calculated the number and percentage (success rate) of putative cancer related genes that overlapped with those extracted from the MGB database among the identified genes in this study. Previous TWAS or eQTL studies for breast cancer

[19,28,32,38,48] , prostate cancer [31,38,51-53] and lung cancer [38,53,54] reported genes related with these cancers. Genetic variants related with risk of breast cancer [49,50], prostate cancer [97] and lung cancer [83,98] were reported in previous GWAS. We also examined the overlapping between the genes identified in this study with predisposition genes, cancer driver genes and CGC-based gene sets. To evaluate whether our identified genes significantly enriched in these cancer-related genes, we conducted enrichment analysis based on the probability mass function of the hypergeometric distribution. Similar to our previous work[38], the *P*-value is calculated as phyper function implemented in R.

Effect of gene silencing on cell proliferation using data from CRISPR-Cas9 essentiality screens in cancer relevant cells

Gene-dependency levels from CRISPR-Cas9 essentiality screens for a total 17,386 genes using a computational method, CERES, were downloaded from the DepMap portal[60]. CRISPR-Cas9 has enabled genome-scale identification of genes that are important for the proliferation and survival of cancer cells, which have been widely used for genetic studies[28,60,61]. For each gene, we calculated the total count and the median of negative CERES values (for cell proliferation) from 45 breast relevant cells, 8 prostate relevant cells and 130 lung relevant cells. The cutoff of CERES value < -0.5 was used to indicate the essentiality [29,60].

The probability mass function of the hypergeometric distribution is: $P(x) = \frac{\binom{m}{x} * \binom{n}{k-x}}{\binom{N}{k}}$, where m is the total number of genes in all cancer-related gene databases, which includes all predisposition genes, cancer drivers and CGC genes; n is the number of genes that are not included in the cancer-related gene databases (n = N – m, N= 19, 291 protein-coding genes based on the annotation from the Gencode.v26.GRCh38).

References

1. Pickrell, J.K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *Am J Hum Genet* **94**, 559-73 (2014).
2. Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-93 (2007).
3. Cai, Q., Long, J., Lu, W., Qu, S., Wen, W., Kang, D. *et al.* Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium. *Hum Mol Genet* **20**, 4991-9 (2011).
4. Cai, Q., Zhang, B., Sung, H., Low, S.K., Kweon, S.S., Lu, W. *et al.* Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1. *Nat Genet* **46**, 886-90 (2014).
5. Fachal, L., Aschard, H., Beesley, J., Barnes, D.R., Allen, J., Kar, S. *et al.* Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nat Genet* **52**, 56-73 (2020).
6. Zheng, W., Zhang, B., Cai, Q., Sung, H., Michailidou, K., Shi, J. *et al.* Common genetic determinants of breast-cancer risk in East Asian women: a collaborative study of 23 637 breast cancer cases and 25 579 controls. *Hum Mol Genet* **22**, 2539-50 (2013).
7. Law, P.J., Timofeeva, M., Fernandez-Rozadilla, C., Broderick, P., Studd, J., Fernandez-Tajes, J. *et al.* Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat Commun* **10**, 2154 (2019).
8. Huyghe, J.R., Bien, S.A., Harrison, T.A., Kang, H.M., Chen, S., Schmit, S.L. *et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nat Genet* **51**, 76-87 (2019).
9. Zeng, C., Matsuda, K., Jia, W.H., Chang, J., Kweon, S.S., Xiang, Y.B. *et al.* Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. *Gastroenterology* **150**, 1633-1645 (2016).
10. Lu, Y., Kweon, S.S., Tanikawa, C., Jia, W.H., Xiang, Y.B., Cai, Q. *et al.* Large-Scale Genome-Wide Association Study of East Asians Identifies Loci Associated With Risk for Colorectal Cancer. *Gastroenterology* **156**, 1455-1466 (2019).
11. Closa, A., Cordero, D., Sanz-Pamplona, R., Sole, X., Crous-Bou, M., Pare-Brunet, L. *et al.* Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis. *Carcinogenesis* **35**, 2039-46 (2014).
12. Ongen, H., Andersen, C.L., Bramsen, J.B., Oster, B., Rasmussen, M.H., Ferreira, P.G. *et al.* Putative cis-regulatory drivers in colorectal cancer. *Nature* **512**, 87-90 (2014).
13. Spain, S.L., Carvajal-Carmona, L.G., Howarth, K.M., Jones, A.M., Su, Z., Cazier, J.B. *et al.* Refinement of the associations between risk of colorectal cancer and polymorphisms on chromosomes 1q41 and 12q13.13. *Hum Mol Genet* **21**, 934-46 (2012).
14. Li, F.F., Yan, P., Zhao, Z.X., Liu, Z., Song, D.W., Zhao, X.W. *et al.* Polymorphisms in the CHIT1 gene: Associations with colorectal cancer. *Oncotarget* (2016).
15. Ke, J., Lou, J., Zhong, R., Chen, X., Li, J., Liu, C. *et al.* Identification of a Potential Regulatory Variant for Colorectal Cancer Risk Mapping to 3p21.31 in Chinese Population. *Sci Rep* **6**, 25194 (2016).

16. Guo, X., Long, J., Zeng, C., Michailidou, K., Ghoussaini, M., Bolla, M.K. *et al.* Fine-scale mapping of the 4q24 locus identifies two independent loci associated with breast cancer risk. *Cancer Epidemiol Biomarkers Prev* **24**, 1680-91 (2015).
17. Shi, J., Zhang, Y., Zheng, W., Michailidou, K., Ghoussaini, M., Bolla, M.K. *et al.* Fine-scale mapping of 8q24 locus identifies multiple independent risk variants for breast cancer. *Int J Cancer* **139**, 1303-17 (2016).
18. Zeng, C., Guo, X., Long, J., Kuchenbaecker, K.B., Droit, A., Michailidou, K. *et al.* Identification of independent association signals and putative functional variants for breast cancer risk through fine-scale mapping of the 12p11 locus. *Breast Cancer Res* **18**, 64 (2016).
19. Guo, X., Lin, W., Bao, J., Cai, Q., Pan, X., Bai, M. *et al.* A Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci Identified in Genome-wide Association Studies. *Am J Hum Genet* **102**, 890-903 (2018).
20. Castro, M.A., de Santiago, I., Campbell, T.M., Vaughn, C., Hickey, T.E., Ross, E. *et al.* Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat Genet* **48**, 12-21 (2016).
21. Cowper-Sal lari, R., Zhang, X., Wright, J.B., Bailey, S.D., Cole, M.D., Eeckhoute, J. *et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat Genet* **44**, 1191-8 (2012).
22. Dunning, A.M., Michailidou, K., Kuchenbaecker, K.B., Thompson, D., French, J.D., Beesley, J. *et al.* Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nat Genet* **48**, 374-86 (2016).
23. Deplancke, B., Alpern, D. & Gardeux, V. The Genetics of Transcription Factor DNA Binding Variation. *Cell* **166**, 538-554 (2016).
24. Tehranchi, A.K., Myrthil, M., Martin, T., Hie, B.L., Golan, D. & Fraser, H.B. Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell* **165**, 730-41 (2016).
25. Yan, J., Qiu, Y., Ribeiro Dos Santos, A.M., Yin, Y., Li, Y.E., Vinckier, N. *et al.* Systematic analysis of binding of transcription factors to noncoding variants. *Nature* (2021).
26. Choudhuri, A., Trompouki, E., Abraham, B.J., Colli, L.M., Kock, K.H., Mallard, W. *et al.* Common variants in signaling transcription-factor-binding sites drive phenotypic variability in red blood cell traits. *Nat Genet* **52**, 1333-1345 (2020).
27. Li, Q., Seo, J.H., Stranger, B., McKenna, A., Pe'er, I., Laframboise, T. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633-41 (2013).
28. Wen, W., Chen, Z., Bao, J., Long, Q., Shu, X.O., Zheng, W. *et al.* Genetic variations of DNA bindings of FOXA1 and co-factors in breast cancer susceptibility. *Nat Commun* **12**, 5318 (2021).
29. Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B.W. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245-52 (2016).
30. Gamazon, E.R., Wheeler, H.E., Shah, K.P., Mozaffari, S.V., Aquino-Michaels, K., Carroll, R.J. *et al.* A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* **47**, 1091-8 (2015).

31. Wu, L., Wang, J., Cai, Q., Cavazos, T.B., Emami, N.C., Long, J. *et al.* Identification of Novel Susceptibility Loci and Genes for Prostate Cancer Risk: A Transcriptome-Wide Association Study in Over 140,000 European Descendants. *Cancer Res* **79**, 3192-3204 (2019).
32. Wu, L., Shi, W., Long, J., Guo, X., Michailidou, K., Beesley, J. *et al.* A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nat Genet* **50**, 968-978 (2018).
33. Lu, Y., Beeghly-Fadiel, A., Wu, L., Guo, X., Li, B., Schildkraut, J.M. *et al.* A Transcriptome-Wide Association Study Among 97,898 Women to Identify Candidate Susceptibility Genes for Epithelial Ovarian Cancer Risk. *Cancer Res* **78**, 5419-5430 (2018).
34. Mancuso, N., Freund, M.K., Johnson, R., Shi, H., Kichaev, G., Gusev, A. *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. *Nat Genet* **51**, 675-682 (2019).
35. Wainberg, M., Sinnott-Armstrong, N., Mancuso, N., Barbeira, A.N., Knowles, D.A., Golan, D. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* **51**, 592-599 (2019).
36. Amariuta, T., Luo, Y., Gazal, S., Davenport, E.E., van de Geijn, B., Ishigaki, K. *et al.* IMPACT: Genomic Annotation of Cell-State-Specific Regulatory Elements Inferred from the Epigenome of Bound Transcription Factors. *Am J Hum Genet* **104**, 879-895 (2019).
37. Zhang, W., Voloudakis, G., Rajagopal, V.M., Readhead, B., Dudley, J.T., Schadt, E.E. *et al.* Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nat Commun* **10**, 3834 (2019).
38. He, J., Wen, W., Beeghly, A., Chen, Z., Cao, C., Shu, X.O. *et al.* Integrating transcription factor occupancy with transcriptome-wide association analysis identifies susceptibility genes in human cancers. *Nat Commun* **13**, 7118 (2022).
39. Cao, C., Kwok, D., Edie, S., Li, Q., Ding, B., Kossinna, P. *et al.* kTWAS: integrating kernel machine with transcriptome-wide association studies improves statistical power and reveals novel genes. *Brief Bioinform* **22**(2021).
40. Tang, S., Buchman, A.S., De Jager, P.L., Bennett, D.A., Epstein, M.P. & Yang, J. Novel Variance-Component TWAS method for studying complex human diseases with applications to Alzheimer's dementia. *PLoS Genet* **17**, e1009482 (2021).
41. Zhu, Z., Zhang, F., Hu, H., Bakshi, A., Robinson, M.R., Powell, J.E. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).
42. Zhang, Y.H., Quick, C., Yu, K.T., Barbeira, A., Luca, F., Pique-Regi, R. *et al.* PTWAS: investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis. *Genome Biology* **21**(2020).
43. Khunsriraksakul, C., McGuire, D., Sauteraud, R., Chen, F., Yang, L., Wang, L. *et al.* Integrating 3D genomic and epigenomic data to enhance target gene discovery and drug repurposing in transcriptome-wide association studies. *Nat Commun* **13**, 3258 (2022).
44. Yao, D.W., O'Connor, L.J., Price, A.L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nat Genet* **52**, 626-633 (2020).
45. O'Connor, L.J., Schoech, A.P., Hormozdiari, F., Gazal, S., Patterson, N. & Price, A.L. Extreme Polygenicity of Complex Traits Is Explained by Negative Selection. *Am J Hum Genet* **105**, 456-476 (2019).

46. Zeng, J., de Vlaming, R., Wu, Y., Robinson, M.R., Lloyd-Jones, L.R., Yengo, L. *et al.* Signatures of negative selection in the genetic architecture of human complex traits. *Nat Genet* **50**, 746-753 (2018).
47. Stranger, B.E., Montgomery, S.B., Dimas, A.S., Parts, L., Stegle, O., Ingle, C.E. *et al.* Patterns of cis regulatory variation in diverse human populations. *PLoS Genet* **8**, e1002639 (2012).
48. Feng, H.L., Gusev, A., Pasaniuc, B., Wu, L., Long, J.R., Abu-full, Z. *et al.* Transcriptome-wide association study of breast cancer risk by estrogen-receptor status. *Genetic Epidemiology* **44**, 442-468 (2020).
49. Fachal, L., Aschard, H., Beesley, J., Barnes, D.R., Allen, J., Kar, S. *et al.* Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nature Genetics* **52**, 56-73 (2020).
50. Zhan, H.Y., Ahearn, T.U., Lecarpentier, J., Barnes, D., Beesley, J., Qi, G.H. *et al.* Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature Genetics* **52**, 572-+ (2020).
51. Mancuso, N., Gayther, S., Gusev, A., Zheng, W., Penney, K.L., Kote-Jarai, Z. *et al.* Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. *Nat Commun* **9**, 4079 (2018).
52. Wu, L., Shu, X., Bao, J., Guo, X., Kote-Jarai, Z., Haiman, C.A. *et al.* Analysis of Over 140,000 European Descendants Identifies Genetically Predicted Blood Protein Biomarkers Associated with Prostate Cancer Risk. *Cancer Res* **79**, 4592-4598 (2019).
53. Chen, Z., Wen, W., Beeghly-Fadiel, A., Shu, X.O., Diez-Obrero, V., Long, J. *et al.* Identifying Putative Susceptibility Genes and Evaluating Their Associations with Somatic Mutations in Human Cancers. *Am J Hum Genet* **105**, 477-492 (2019).
54. Bosse, Y., Li, Z., Xia, J., Manem, V., Carreras-Torres, R., Gabriel, A. *et al.* Transcriptome-wide association study reveals candidate causal genes for lung cancer. *Int J Cancer* **146**, 1862-1878 (2020).
55. Easton, D.F., Pharoah, P.D., Antoniou, A.C., Tischkowitz, M., Tavtigian, S.V., Nathanson, K.L. *et al.* Gene-panel sequencing and the prediction of breast-cancer risk. *N Engl J Med* **372**, 2243-57 (2015).
56. Hu, C., Hart, S.N., Gnanaolivu, R., Huang, H., Lee, K.Y., Na, J. *et al.* A Population-Based Study of Genes Previously Implicated in Breast Cancer. *N Engl J Med* **384**, 440-451 (2021).
57. Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations (vol 173, 371.e1, 2018). *Cell* **174**, 1034-1035 (2018).
58. Dietlein, F., Weghorn, D., Taylor-Weiner, A., Richters, A., Reardon, B., Liu, D. *et al.* Identification of cancer driver genes based on nucleotide context. *Nature Genetics* **52**, 208-+ (2020).
59. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I. & Forbes, S.A. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nature Reviews Cancer* **18**, 696-705 (2018).

60. Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H. *et al.* Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. *Nature Genetics* **49**, 1779-+ (2017).
61. Gusev, A., Lawrenson, K., Lin, X.Z., Lyra, P.C., Kar, S., Vavra, K.C. *et al.* A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. *Nature Genetics* **51**, 815-+ (2019).
62. Gurel, B., Ali, T.Z., Montgomery, E.A., Begum, S., Hicks, J., Goggins, M. *et al.* NKX3.1 as a marker of prostatic origin in metastatic tumors. *Am J Surg Pathol* **34**, 1097-105 (2010).
63. Le Magnen, C., Virk, R.K., Dutta, A., Kim, J.Y., Panja, S., Lopez-Bujanda, Z.A. *et al.* Cooperation of loss of NKX3.1 and inflammation in prostate cancer initiation. *Dis Model Mech* **11**(2018).
64. Bhatia-Gaur, R., Donjacour, A.A., Sciavolino, P.J., Kim, M., Desai, N., Young, P. *et al.* Roles for Nkx3.1 in prostate development and cancer. *Genes Dev* **13**, 966-77 (1999).
65. Sooreshjani, M.A., Nikhil, K., Kamra, M., Nguyen, D.N., Kumar, D. & Shah, K. LIMK2-NKX3.1 Engagement Promotes Castration-Resistant Prostate Cancer. *Cancers (Basel)* **13**(2021).
66. Kaochar, S., Rusin, A., Foley, C., Rajapakshe, K., Robertson, M., Skapura, D. *et al.* Inhibition of GATA2 in prostate cancer by a clinically available small molecule. *Endocr Relat Cancer* **29**, 15-31 (2021).
67. Rodriguez-Bravo, V., Carceles-Cordon, M., Hoshida, Y., Cordon-Cardo, C., Galsky, M.D. & Domingo-Domenech, J. The role of GATA2 in lethal prostate cancer aggressiveness. *Nat Rev Urol* **14**, 38-48 (2017).
68. Shen, T., Dong, B., Meng, Y., Moore, D.D. & Yang, F. A COP1-GATA2 axis suppresses AR signaling and prostate cancer. *Proc Natl Acad Sci U S A* **119**, e2205350119 (2022).
69. He, B., Lanz, R.B., Fiskus, W., Geng, C., Yi, P., Hartig, S.M. *et al.* GATA2 facilitates steroid receptor coactivator recruitment to the androgen receptor complex. *Proc Natl Acad Sci U S A* **111**, 18261-6 (2014).
70. Vosa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P., Zeng, B. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300-1310 (2021).
71. Liu, X., Li, Y.I. & Pritchard, J.K. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell* **177**, 1022-1034 e6 (2019).
72. Luningham, J.M., Chen, J., Tang, S., De Jager, P.L., Bennett, D.A., Buchman, A.S. *et al.* Bayesian Genome-wide TWAS Method to Leverage both cis- and trans-eQTL Information through Summary Statistics. *Am J Hum Genet* **107**, 714-726 (2020).
73. Wang, Y., Qian, M., Ruan, P., Teschendorff, A.E. & Wang, S. Detection of epigenetic field defects using a weighted epigenetic distance-based method. *Nucleic Acids Res* **47**, e6 (2019).
74. Villicana, S. & Bell, J.T. Genetic impacts on DNA methylation: research findings and future perspectives. *Genome Biol* **22**, 127 (2021).
75. Zhang, Y., Wong, C.H., Birnbaum, R.Y., Li, G., Favaro, R., Ngan, C.Y. *et al.* Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature* **504**, 306-310 (2013).

76. Long, K., Li, X., Su, D., Zeng, S., Li, H., Zhang, Y. *et al.* Exploring high-resolution chromatin interaction changes and functional enhancers of myogenic marker genes during myogenic differentiation. *J Biol Chem* **298**, 102149 (2022).
77. Hu, Y., Li, M., Lu, Q., Weng, H., Wang, J., Zekavat, S.M. *et al.* A statistical framework for cross-tissue transcriptome-wide association analysis. *Nat Genet* **51**, 568-576 (2019).
78. Barbeira, A.N., Pividori, M.D., Zheng, J.M., Wheeler, H.E., Nicolae, D.L. & Im, H.K. Integrating predicted transcriptome from multiple tissues improves association detection. *Plos Genetics* **15**(2019).
79. Consortium, G.T. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **45**, 580-5 (2013).
80. Consortium, G. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).
81. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
82. Schumacher, F.R., Al Olama, A.A., Berndt, S.I., Benlloch, S., Ahmed, M., Saunders, E.J. *et al.* Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet* **50**, 928-936 (2018).
83. McKay, J.D., Hung, R.J., Han, Y., Zong, X., Carreras-Torres, R., Christiani, D.C. *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* **49**, 1126-1132 (2017).
84. Zheng, R., Wan, C., Mei, S., Qin, Q., Wu, Q., Sun, H. *et al.* Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res* **47**, D729-D735 (2019).
85. Meuleman, W., Muratov, A., Rynes, E., Halow, J., Lee, K., Bates, D. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244-251 (2020).
86. Boix, C.A., James, B.T., Park, Y.P., Meuleman, W. & Kellis, M. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature* **590**, 300-307 (2021).
87. Consortium, F., the, R.P., Clst, Forrest, A.R., Kawaji, H., Rehli, M. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462-70 (2014).
88. Vosa, U., Claringbould, A., Westra, H.J., Bonder, M.J., Deelen, P., Zeng, B. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics* **53**, 1300-+ (2021).
89. Rhie, S.K., Perez, A.A., Lay, F.D., Schreiner, S., Shi, J., Polin, J. *et al.* A high-resolution 3D epigenomic map reveals insights into the creation of the prostate cancer transcriptome. *Nat Commun* **10**, 4154 (2019).
90. Teng, L., He, B., Wang, J. & Tan, K. 4DGenome: a comprehensive database of chromatin interactions. *Bioinformatics* **32**, 2727 (2016).
91. da Costa, E.S.C.S., Cury, N.M., Brotto, D.B., de Araujo, L.F., Rosa, R.C.A., Texeira, L.A. *et al.* Germline variants in DNA repair genes associated with hereditary breast and ovarian cancer syndrome: analysis of a 21 gene panel in the Brazilian population. *BMC Med Genomics* **13**, 21 (2020).

92. Bailey, M.H., Tokheim, C., Porta-Pardo, E., Sengupta, S., Bertrand, D., Weerasinghe, A. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385 e18 (2018).
93. Dietlein, F., Weghorn, D., Taylor-Weiner, A., Richters, A., Reardon, B., Liu, D. *et al.* Identification of cancer driver genes based on nucleotide context. *Nat Genet* **52**, 208-218 (2020).
94. Sondka, Z., Bamford, S., Cole, C.G., Ward, S.A., Dunham, I. & Forbes, S.A. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer* **18**, 696-705 (2018).
95. Barbeira, A.N., Dickinson, S.P., Bonazzola, R., Zheng, J.M., Wheeler, H.E., Torres, J.M. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9**(2018).
96. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-504 (2003).
97. Conti, D.V., Darst, B.F., Moss, L.C., Saunders, E.J., Sheng, X., Chou, A. *et al.* Trans-ancestry genome-wide association meta-analysis of prostate cancer identifies new susceptibility loci and informs genetic risk prediction. *Nat Genet* **53**, 65-75 (2021).
98. Seow, W.J., Matsuo, K., Hsiung, C.A., Shiraishi, K., Song, M.S., Kim, H.N. *et al.* Association between GWAS-identified lung adenocarcinoma susceptibility loci and EGFR mutations in never-smoking Asian women, and comparison with findings from Western populations. *Human Molecular Genetics* **26**, 454-465 (2017).

Acknowledgments: We thank GTEx, TCGA, ENCODE, Roadmap and BCAC for providing valuable data resources for this study. The data analyses were conducted using the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University. This research was supported by the grant from US National Institutes of Health grant R37 CA227130 to X.G. and R01 CA235553 to W. Z., a New Frontiers in Research Fund (NFRFE-2018-00748) to Q.L., and a NSERC discovery grant (RGPIN-2017-04860) to Q.L.. J.H. was partly supported by the China Scholarship Council (CSC). D.P. was supported by an Alberta Innovates and an Eyes High scholarship. The computational infrastructure was partly supported by a Canada Foundation for Innovation JELF grant (36605).

Author contributions: X.G. and Q.L. conceived and designed the study. J.H., Q.L. and X.G. performed data collection and processing, bioinformatics and statistical analyses, with additional data preparation and discussion from W.W., J.P., and C.Z.. J.H., Q.L. and X.G. wrote the manuscript with contributions from all other authors. All authors have reviewed and approved the content of the article.

Competing interests: The authors declare that they have no competing interests.

Data and materials availability: Table S1 provides the download information for the summary statistics of GWAS data for breast cancer, prostate cancer, lung cancer, and three brain disorders (SCZ, ASD, and AD); the epigenetic data, including ChIP-seq data of transcription factors, DHSs, enhancer, promoter, 3D genomics informed regions, enhancer gene links, and eQTLs used in this study; and the functional annotation data, including target cancer related genes, CGC, and cancer driven genes. Gene expression and alternative splicing data generated in breast, prostate, lung and brain tissues, were downloaded from GTEx consortium, and the individual-level genotype was downloaded from dbGaP (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2). Gencode annotation (v26.GRCh38) was downloaded from https://www.gencodegenes.org/human/release_26.html. The data from the 1000 Genomes Project data was downloaded through the website, https://www.genome.gov/27528684/1000-genomes-project. For data of essentiality for proliferation and survival of cancer cells, we downloaded the comprehensive datasets including “sample_info.csv” and

"Achilles_gene_effect.csv" from the DepMap portal. The developed pipeline and main source R codes used in this work are available from Github website: https://github.com/theLongLab/transTF-TWAS or https://github.com/XingyiGuo/transTF-TWAS/.

Supplementary materials

This PDF file includes:

Figs. S1 to S5

Legends for tables S1 to S18

Other Supplementary Materials for this manuscript include the following:

Tables S1 to S18

FIGURES

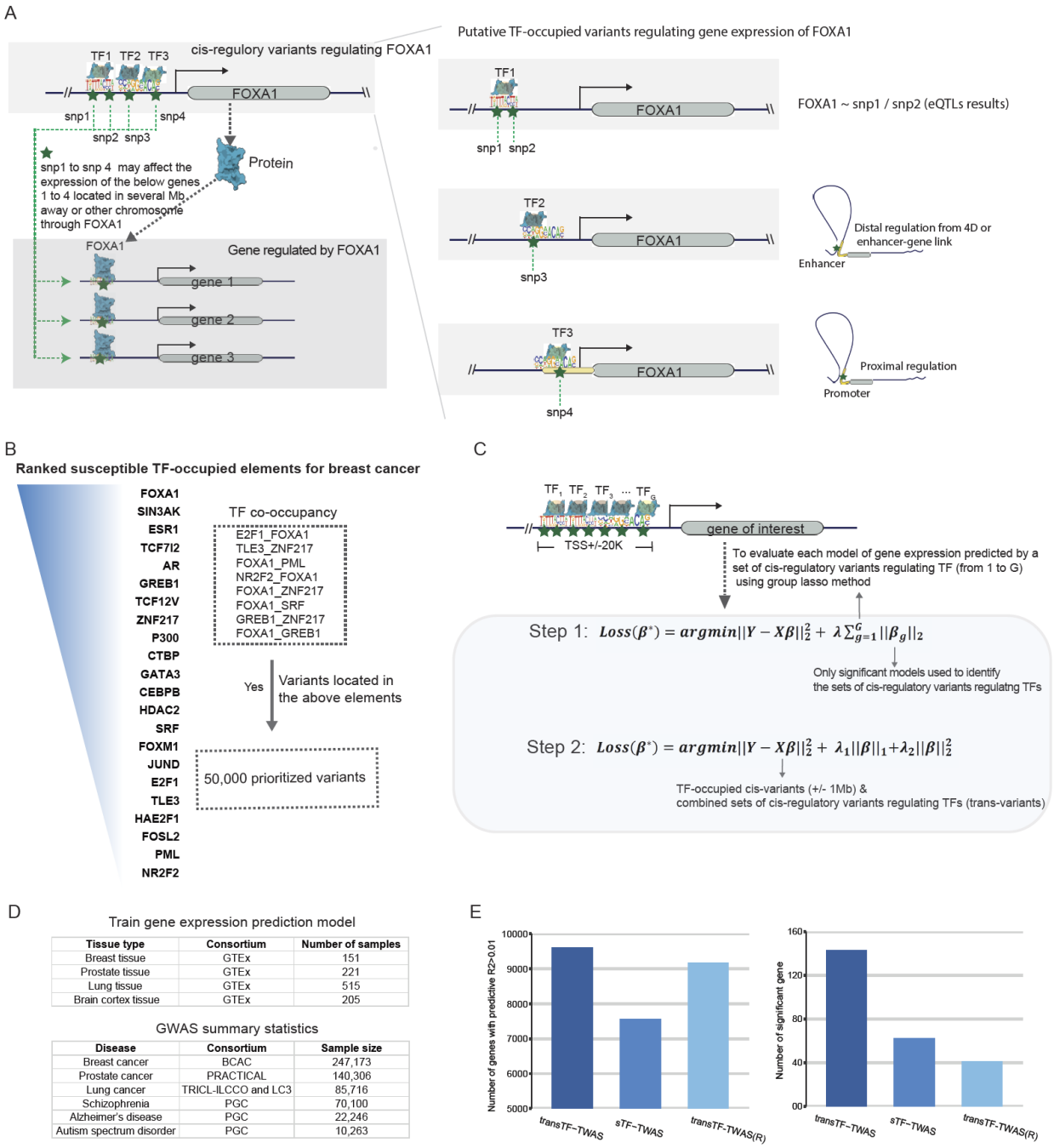

Figure 1. Overview of the Developed Analytical Framework. A. An illustration of how to prioritize TF-linked trans-located variants for prediction model building (using FOXA1 as an example). B. Flow chart showing prioritized TF-occupied regulatory variants (50K), which were ranked based on established TF-occupied elements associated with breast cancer risk. C. An

illustration of the two-step gene expression prediction model building in transTF-TWAS. D. The top table showed the sample size of the data that used for training gene expression prediction model. The bottom table showed the sample size for GWAS cohort. E. The left bar chart showed the number of gene with predictive $R^2 > 0.01$ among transTF-TWAS, sTF-TWAS and simulated model. The right bar chart showed the number of significant gene among transTF-TWAS, sTF-TWAS and simulated model. The number of significantly identified genes was indicated at a Bonferroni-corrected $P < 0.05$.

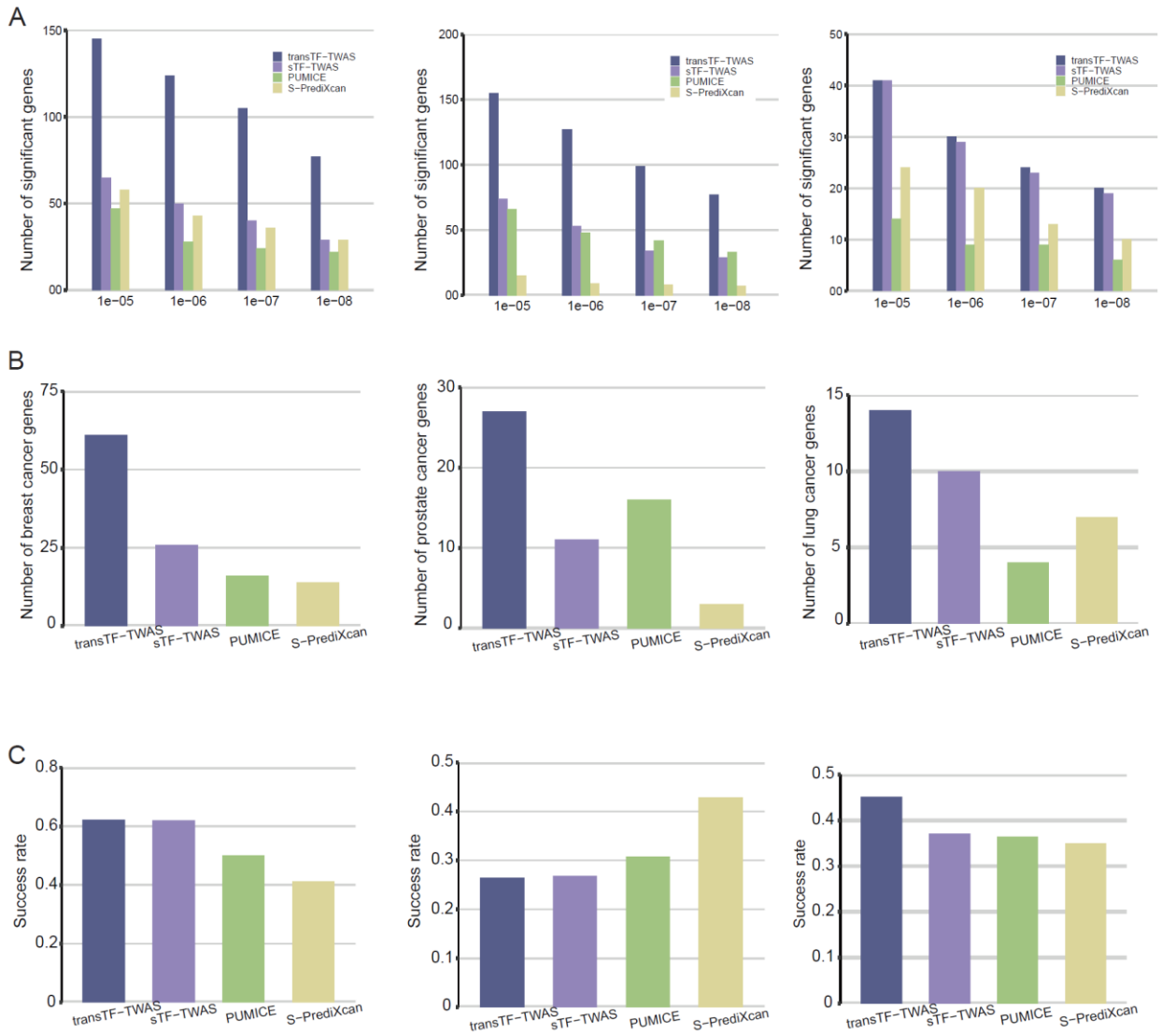

Figure 2. Comparison of gene-trait associations between transTF-TWAS with other TWAS approaches (sTF-TWAS, PUMICE, and S-PrediXcan) for breast, prostate and lung cancer. A. Bar chart showing the number of genes identified from transTF-TWAS and other TWAS approaches under various *P*-value cutoffs (i.e., P < 1e-05, 1e-06, 1e-07, and 1e-08). The P-values are the nominal P-values from the Z score test from TWAS. B. Bar chart showing a comparison between the total number of target cancer related genes among transTF-TWAS and other TWAS approaches. C. Bar chart showing a comparison of the proportion (success rate) of target cancer related gene among transTF-TWAS and other TWAS approaches, relative to the total number of genes identified from the set.

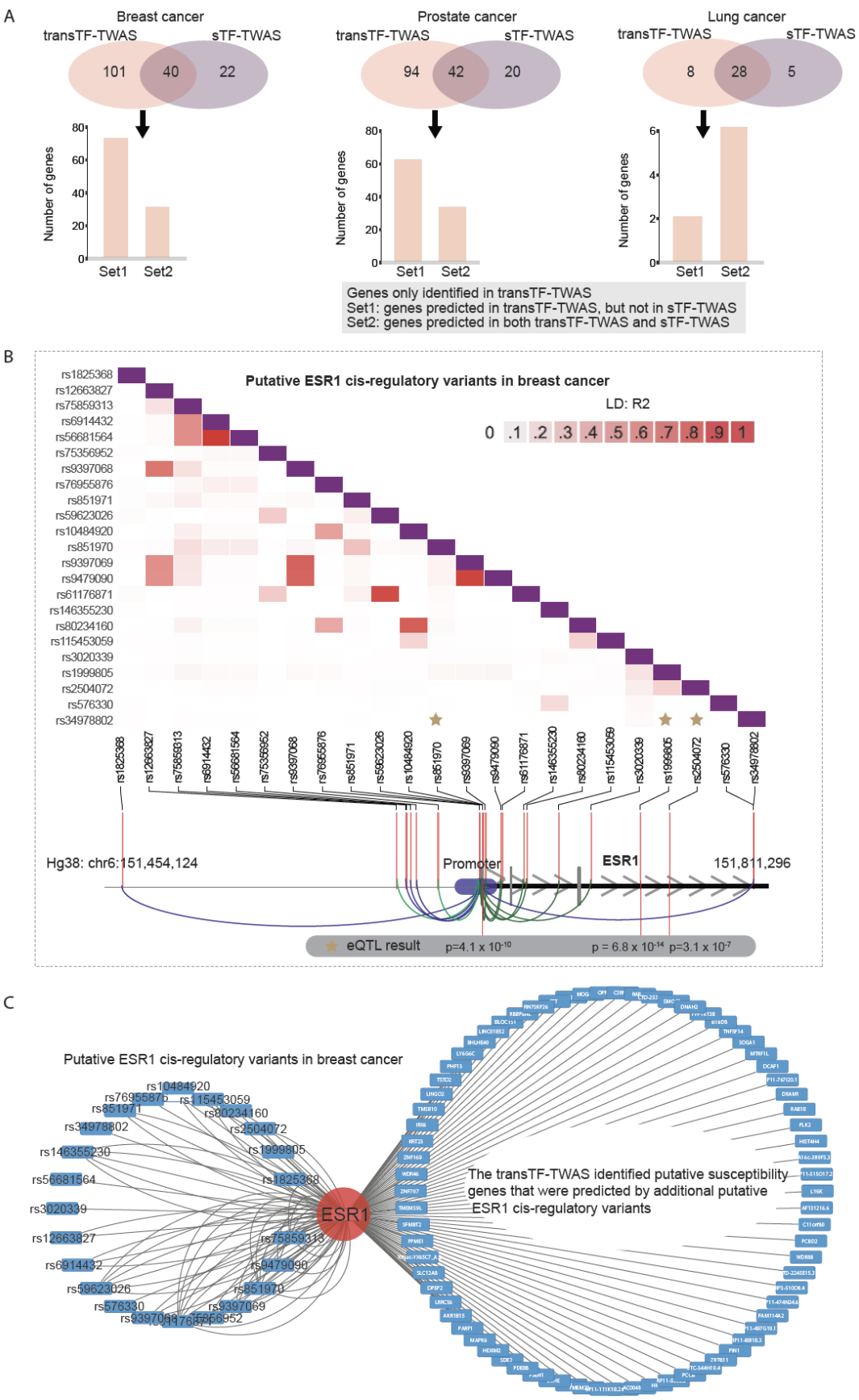
A
Breast cancer
transTF-TWAS
sTF-TWAS
101
40
22
Number of genes
Set1
Set2
Prostate cancer
transTF-TWAS
sTF-TWAS
94
42
20
Number of genes
Set1
Set2
Lung cancer
transTF-TWAS
sTF-TWAS
8
28
5
Number of genes
Set1
Set2
Genes only identified in transTF-TWAS
Set1: genes predicted in transTF-TWAS, but not in sTF-TWAS
Set2: genes predicted in both transTF-TWAS and sTF-TWAS
B
Putative ESR1 cis-regulatory variants in breast cancer
LD: R2
0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1
rs1825368
rs12663827
rs75859313
rs6914432
rs56681564
rs75356952
rs9397068
rs76955876
rs851971
rs59623026
rs10484920
rs851970
rs9397069
rs9479090
rs61176871
rs146355230
rs80234160
rs115453059
rs3020339
rs1999805
rs2504072
rs576330
rs34978802
Hg38: chr6:151,454,124
Promoter
ESR1
151,811,296
eQTL result
p=4.1 x 10-10
p = 6.8 x 10-14
p=3.1 x 10-7
C
Putative ESR1 cis-regulatory variants in breast cancer
ESR1
The transTF-TWAS identified putative susceptibility genes that were predicted by additional putative ESR1 cis-regulatory variants

Figure 3. The gene regulatory network underlying cancer risk driven by master regulators. A. Venn diagrams showing the number of putative susceptibility genes commonly or uniquely identified by transTF-TWAS and sTF-TWAS. An arrow points from the uniquely identified genes by transTF-TWAS to a bar chart showing: Set 1: genes predicted in transTF-TWAS, but not in sTF-TWAS. Set 2: genes predicted in both transTF-TWAS and sTF-TWAS. B. A heatmap showing the LD structure among putative ESR1 cis-regulatory variants in breast cancer. The ESR1 cis-regulatory variants are trans-located variants that present the strongest associations with cancer risk in the prediction model. C. A network showing the connections between the putative ESR1 cis-regulatory variants in breast cancer and putative susceptibility genes identified by transTF-TWAS that were contributed by putative ESR1 cis-regulatory variants.

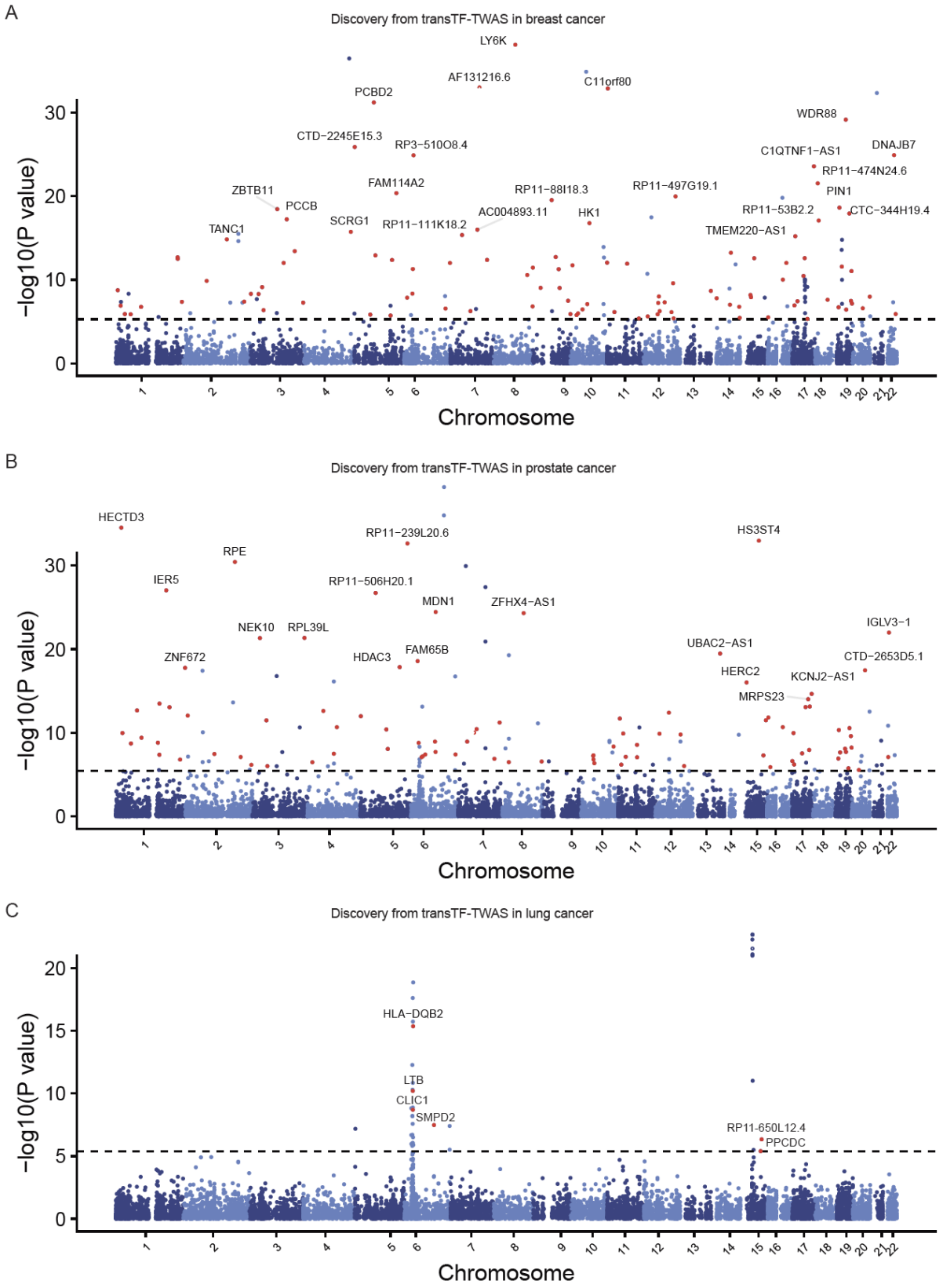
A
Discovery from transTF-TWAS in breast cancer
LY6K
AF131216.6
C11orf80
PCBD2
WDR88
CTD−2245E15.3
RP3−510O8.4
C1QTNF1−AS1
DNAJB7
RP11−474N24.6
ZBTB11
FAM114A2
RP11−88I18.3
RP11−497G19.1
PIN1
PCCB
AC004893.11
HK1
RP11−53B2.2
CTC−344H19.4
TANC1
SCRG1
RP11−111K18.2
TMEM220−AS1
−log10(P value)
Chromosome
B
Discovery from transTF-TWAS in prostate cancer
HECTD3
RP11−239L20.6
HS3ST4
RPE
IER5
RP11−506H20.1
MDN1
ZFHX4−AS1
NEK10
RPL39L
IGLV3−1
UBAC2−AS1
ZNF672
HDAC3
FAM65B
CTD−2653D5.1
HERC2
KCNJ2−AS1
MRPS23
−log10(P value)
Chromosome
C
Discovery from transTF-TWAS in lung cancer
HLA−DQB2
LTB
CLIC1
SMPD2
RP11-650L12.4
PPCDC
−log10(P value)
Chromosome

Figure 4. Putative susceptibility genes identified by transTF-TWAS. Manhattan plots showing the associations identified from transTF-TWAS. Red dots indicated all newly identified susceptibility genes, and the grey dashed line refers to Bonferroni-corrected $P < 0.05$. The newly identified putative susceptibility genes with $P < 10^{-15}$ were highlighted. The *P*-values are the raw *P*-values from the Z score test from TWAS (two-sided). A) breast cancer. B) prostate cancer. C) lung cancer.

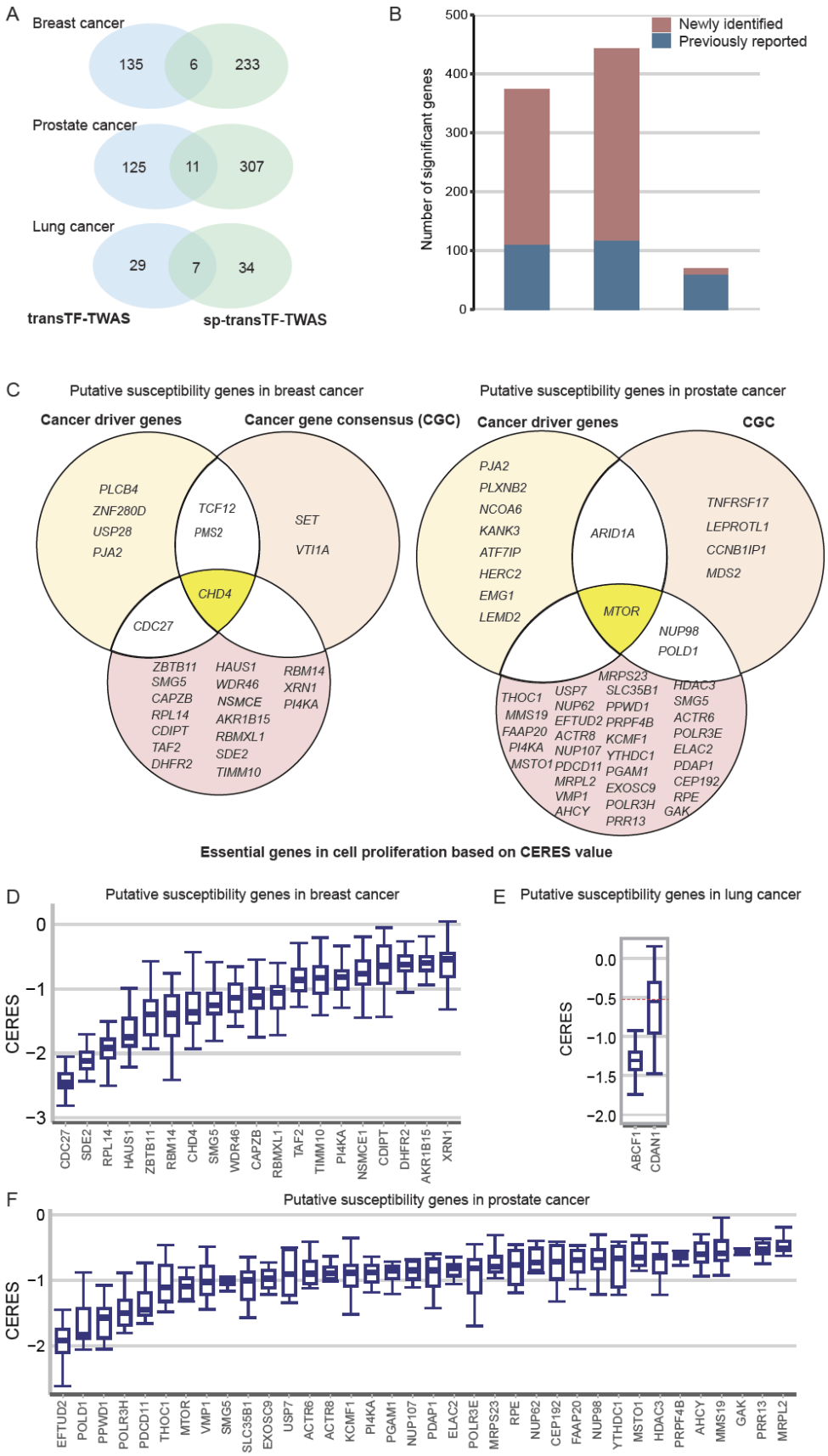
A
Breast cancer
135
6
233
Prostate cancer
125
11
307
Lung cancer
29
7
34
transTF-TWAS
sp-transTF-TWAS
B
Newly identified
Previously reported
Number of significant genes
500
400
300
200
100
0
C
Putative susceptibility genes in breast cancer
Cancer driver genes
Cancer gene consensus (CGC)
PLCB4
ZNF280D
USP28
PJA2
TCF12
PMS2
SET
VTI1A
CHD4
CDC27
ZBTB11
SMG5
CAPZB
RPL14
CDIPT
TAF2
DHFR2
HAUS1
WDR46
NSMCE
AKR1B15
RBMXL1
SDE2
TIMM10
RBM14
XRN1
PI4KA
Putative susceptibility genes in prostate cancer
Cancer driver genes
CGC
PJA2
PLXNB2
NCOA6
KANK3
ATF7IP
HERC2
EMG1
LEMD2
ARID1A
TNFRSF17
LEPROTL1
CCNB1IP1
MDS2
MTOR
NUP98
POLD1
THOC1
MMS19
FAAP20
PI4KA
MSTO1
USP7
NUP62
EFTUD2
ACTR8
NUP107
PDCD11
MRPL2
VMP1
AHCY
MRPS23
SLC35B1
PPWD1
PRPF4B
KCMF1
YTHDC1
PGAM1
EXOSC9
POLR3H
PRR13
HDAC3
SMG5
ACTR6
POLR3E
ELAC2
PDAP1
CEP192
RPE
GAK
Essential genes in cell proliferation based on CERES value
D
Putative susceptibility genes in breast cancer
CERES
0
−1
−2
−3
CDC27
SDE2
RPL14
HAUS1
ZBTB11
RBM14
CHD4
SMG5
WDR46
CAPZB
RBMXL1
TAF2
TIMM10
PI4KA
NSMCE1
CDIPT
DHFR2
AKR1B15
XRN1
E
Putative susceptibility genes in lung cancer
CERES
0.0
−0.5
−1.0
−1.5
−2.0
ABCF1
CDAN1
F
Putative susceptibility genes in prostate cancer
CERES
0
−1
−2
EFTUD2
POLD1
PPWD1
POLR3H
PDCD11
THOC1
MTOR
VMP1
SMG5
SLC35B1
EXOSC9
USP7
ACTR6
ACTR8
KCMF1
PI4KA
PGAM1
NUP107
PDAP1
ELAC2
POLR3E
MRPS23
RPE
NUP62
CEP192
FAAP20
NUP98
YTHDC1
MSTO1
HDAC3
PRPF4B
AHCY
MMS19
GAK
PRR13
MRPL2

Figure 5. Putative susceptibility genes identified by transTF-TWAS and sp-transTF-TWAS. A. Venn diagrams showing the number of putative susceptibility genes commonly or uniquely identified by transTF-TWAS and sp-transTF-TWAS. B. Bar chart showing the total identified putative susceptibility genes combined from transTF-TWAS and sp-transTF-TWAS for breast, prostate and lung cancer. C. Venn diagrams showing all newly identified genes that were cancer driven genes, Cancer Gene Census (CGC), or genes with CERES<-0.5 for breast cancer and prostate cancer. D-F. Boxplot showing all newly identified genes with evidence of essential roles in cell proliferation based on a cutoff of median CERES values < -0.5 for D) breast cancer (sample size: 45 cell lines), E) lung cancer (sample size: 130 cell lines), and F) prostate cancer (sample size: 8 cell lines). In the boxplots shown in these figures, the whiskers denote the range; the boxes denote the interquartile range; the middle bars in denote the median.