







# MiXcan: a framework for cell-type-aware transcriptome-wide association studies with an application to breast cancer

Received: 16 March 2022

Accepted: 5 January 2023

Published online: 23 January 2023

 Check for updates

Xiaoyu Song <sup>1,2</sup> ✉, Jiayi Ji<sup>1,2</sup>, Joseph H. Rothstein<sup>2,3</sup>, Stacey E. Alexeeff<sup>4</sup>, Lori C. Sakoda <sup>4</sup>, Adriana Sistig<sup>3</sup>, Ninah Achacoso<sup>4</sup>, Eric Jorgenson <sup>4,5</sup>, Alice S. Whittemore<sup>6,7</sup>, Robert J. Klein <sup>1,3</sup>, Laurel A. Habel<sup>4</sup>, Pei Wang <sup>1,3,8</sup> ✉ & Weiva Sieh <sup>1,2,3,8</sup> ✉

Human bulk tissue samples comprise multiple cell types with diverse roles in disease etiology. Conventional transcriptome-wide association study approaches predict genetically regulated gene expression at the tissue level, without considering cell-type heterogeneity, and test associations of predicted tissue-level expression with disease. Here we develop MiXcan, a cell-type-aware transcriptome-wide association study approach that predicts cell-type-level expression, identifies disease-associated genes via combination of cell-type-level association signals for multiple cell types, and provides insight into the disease-critical cell type. As a proof of concept, we conducted cell-type-aware analyses of breast cancer in 58,648 women and identified 12 transcriptome-wide significant genes using MiXcan compared with only eight genes using conventional approaches. Importantly, MiXcan identified genes with distinct associations in mammary epithelial versus stromal cells, including three new breast cancer susceptibility genes. These findings demonstrate that cell-type-aware transcriptome-wide analyses can reveal new insights into the genetic and cellular etiology of breast cancer and other diseases.

Transcriptome-wide association studies (TWAS) aim to identify genes that are associated with disease through their genetically regulated gene expression (GR<sub>EX</sub>) levels<sup>1,2</sup>. Conventional TWAS approaches such as PrediXcan<sup>1</sup> predict tissue-level GR<sub>EX</sub> using models trained on transcriptomic and genomic data from bulk tissue samples, and test associations between the predicted tissue-level GR<sub>EX</sub> and disease. By reducing the multiple testing burden from millions of variants to thousands of genes, TWAS can improve the power of genome-wide association studies (GWAS) while providing biological insights into the genes and regulatory mechanisms underlying disease. However,

conventional TWAS approaches do not account for cell-type heterogeneity of bulk tissue samples, which can reduce the accuracy of GR<sub>EX</sub> prediction models and obscure disease associations, particularly when the most mechanistically relevant cell type for the disease is a minor cell type in the tissue<sup>3</sup>.

Breast carcinoma is a common and highly heritable cancer that arises from epithelial cells, which line the ducts and lobules that produce milk during lactation<sup>4,5</sup>. Human mammary tissue has highly variable cell composition. Visualized on mammography, breast composition can range from extremely dense (light), reflecting a high

<sup>1</sup>Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>2</sup>Department of Population Health Science and Policy, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>3</sup>Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. <sup>4</sup>Division of Research, Kaiser Permanente Northern California, Oakland, CA, USA. <sup>5</sup>Regeneron Genetics Center, Tarrytown, NY, USA. <sup>6</sup>Department of Epidemiology and Population Health, Stanford University School of Medicine, Stanford, CA, USA. <sup>7</sup>Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA. <sup>8</sup>These authors jointly supervised this work: Pei Wang, Weiva Sieh. ✉e-mail: [xiaoyu.song@mountsinai.org](mailto:xiaoyu.song@mountsinai.org); [pei.wang@mssm.edu](mailto:pei.wang@mssm.edu); [weiva.sieh@mssm.edu](mailto:weiva.sieh@mssm.edu)

proportion of fibroglandular tissue, to almost entirely fatty (dark), reflecting a high proportion of adipose tissue<sup>6,7</sup>. Whereas higher mammographic density is associated with increased risk of breast cancer, a higher amount of nondense fatty tissue is associated with decreased risk, indicating disparate roles of the different cellular components of mammary tissue in carcinogenesis<sup>8–10</sup>. Breast cancer susceptibility loci identified by prior GWAS<sup>11–13</sup> and TWAS<sup>14–16</sup> approaches that do not account for cell-type heterogeneity explain only a fraction of the familial relative risk. Disentangling the distinct effects of gene expression in mammary epithelial cells from other cell types through cell-type-aware analysis could lead to new gene discoveries and biological insights.

To our knowledge, no statistical methods currently exist for conducting cell-type-aware TWAS using GWAS data. Single-cell sorting and transcriptome profiling are costly, and large reference panels with both single-cell transcriptomic and genomic data are not yet widely available for training robust GRex prediction models. Recent studies of bulk tissue transcriptomic data have used computational estimates of cell-type enrichment, which are correlated with their proportions, to evaluate cell-type-specific effects. The Genotype-Tissue Expression (GTEx<sup>17</sup>) consortium estimated cell-type enrichment scores in bulk tissue samples using xCell<sup>18</sup> and tested for interactions between genotype and xCell scores in linear regression models of gene expression to identify interaction expression quantitative trait loci (ieQTL)<sup>19</sup>. The breast was among the human tissues with the most ieQTLs, specifically involving mammary epithelial cells and adipocytes<sup>19</sup>, highlighting the potential for new methods that harness cell-type-specific genetic regulation of expression to improve the power of breast cancer TWAS. Methods that integrate bulk tissue data with single-cell reference profiles to estimate cell-type-level gene expression have also been proposed to study cell-type-specific disease associations<sup>20,21</sup>. However, these methods all require transcriptomic data from the disease-relevant tissue and cannot be applied to existing GWAS datasets to perform TWAS in large populations.

Here we present MiXcan, a new statistical framework for conducting cell-type-aware TWAS using GWAS data. MiXcan builds cell-type-level GRex prediction models through decomposition of bulk tissue data, identifies disease-associated genes via combination of signals from cell-type-level association analyses of multiple cell types, and provides insight into the cell type responsible for the disease association. We show that MiXcan improves the tissue-level GRex prediction accuracy compared with conventional approaches in an independent bulk-tissue validation set, and reliably predicts epithelial cell GRex in a single-nucleus RNA sequencing (snRNAseq) dataset. Simulation studies show that MiXcan controls the type I error, and provides higher power than conventional TWAS approaches when disease associations are driven by a minor cell type (e.g. mammary epithelial cells) rather than the predominant cell type in a tissue, or have opposite directions in different cell types. We apply MiXcan to

conduct the first cell-type-aware TWAS of breast cancer risk in 31,716 cases and 26,932 controls, and report three new susceptibility genes (*ZNF703*, *TMEM245*, and *PSG4*) with evidence of distinct associations in mammary epithelial versus stromal cells that were not detected by prior TWAS nor GWAS. These findings provide a proof a concept that cell-type-aware TWAS can reveal new insights into the genetic and cellular etiology of breast cancer and other diseases.

## Results

### MiXcan framework

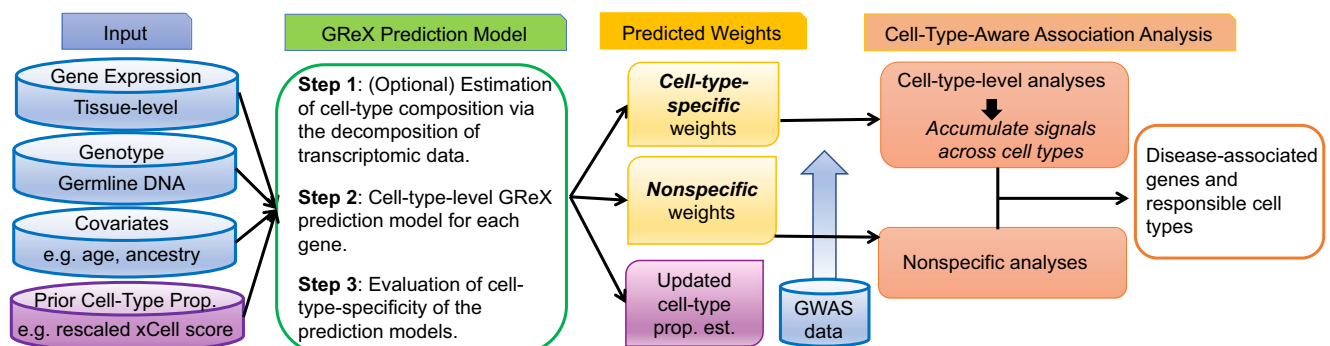
We developed the MiXcan framework for conducting cell-type-aware TWAS (Fig. 1). To build GRex prediction models, MiXcan requires specification of the cell type of interest and a prior estimate of its proportion in bulk tissue training samples with transcriptomic and genomic data. The cell type of interest for a given disease may be selected based on prior biologic knowledge, and its proportion estimated from the transcriptomic data using existing deconvolution methods and reference panels<sup>21–23</sup>. For cell types without large reference panels or direct proportion estimates, a cell-type enrichment score can be estimated from the bulk tissue transcriptomic data using xCell<sup>18</sup>. MiXcan can utilize xCell or other enrichment scores as a prior to estimate the cell-type proportion (see “Methods”). MiXcan then decomposes the bulk tissue gene expression level into its cell-type levels and uses joint penalized regression to model the association of genetic variants (SNPs) with gene expression for each cell type. The regression coefficients (SNP weights) are compared to determine whether the GRex prediction models for each gene are cell-type-specific (different weights in different cell types) or nonspecific (same weights across cell types). Simulation studies (below) show that MiXcan prediction models are robust to misspecification of the cell-type proportion, which can result from inaccurate estimates<sup>24–26</sup>.

To conduct cell-type-aware TWAS, MiXcan uses the predicted GRex to test the following composite null and alternative hypotheses:

$H_0$ : There is no association between the predicted GRex and the disease in any cell type.

$H_A$ : There is an association between the predicted GRex and disease in at least one cell type.

Genes with cell-type-specific GRex prediction models are first associated with disease within each cell type, and then the signals are combined across cell types using the Cauchy-based *p*-value combination method<sup>27</sup>. Genes with nonspecific GRex prediction models are tested for their association with disease in one step. Significant associations are identified using an appropriate threshold to control the family-wise error rate (FWER) or false discovery rate (FDR). For significant genes with cell-type-specific GRex models, the cell-type-level results are compared to provide further insight into the cell type(s) likely to be responsible for the disease association.



**Fig. 1 | MiXcan framework.** MiXcan estimates cell-type composition using transcriptomic data, builds cell-type-specific and nonspecific GRex prediction models, identifies disease-associated genes in any cell types, and provides insight into the cell type responsible for the disease association.

While the MiXcan framework is general, its performance depends on the cell types under consideration and the available training data. As the number of cell types increases, the number of parameters increases and the accuracy of the model decreases. At present, given the limited sample sizes of transcriptomic and genomic datasets available for most human tissues through public repositories such as GTEx, it is practical to consider only two categories of cells using MiXcan and to focus on the cell type of greatest interest for the disease under investigation versus the other cell types. As breast carcinoma is known to arise from epithelial cells, we developed MiXcan epithelial and stromal (nonepithelial) cell models using bulk mammary tissue transcriptomic and genomic data available for 125 European ancestry (EA) women in GTEx v8.

### Prediction performance

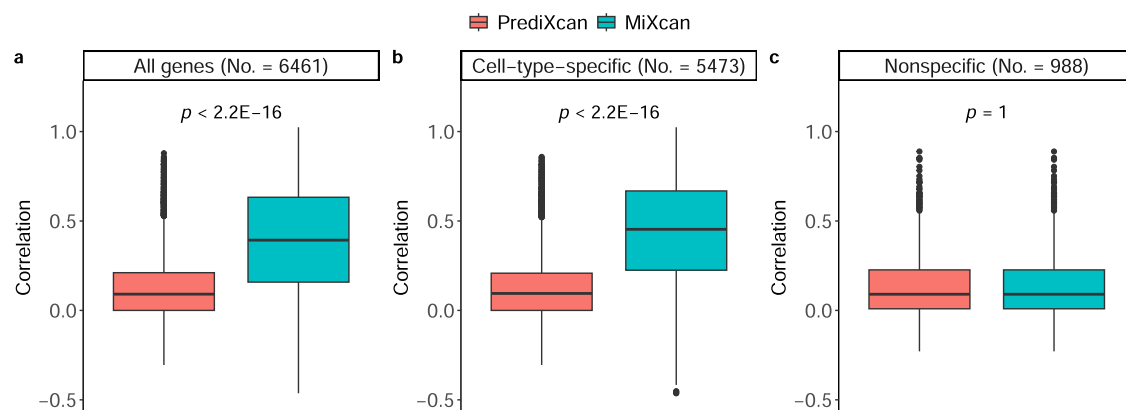
The accuracy of MiXcan and PrediXcan GrEx prediction models trained using GTEx v8 data for 125 mammary tissue samples from EA women was initially evaluated in an independent dataset of 103 tumor-adjacent normal mammary tissue samples from EA women in The Cancer Genome Atlas (TCGA). MiXcan estimates of the epithelial cell proportions were highly correlated with the xCell<sup>18</sup> epithelial cell enrichment scores (used as a prior), with Pearson correlations ( $r$ ) of 0.90 and 0.89 in normal mammary tissue samples from EA women in GTEx ( $N = 125$ ) and TCGA ( $N = 103$ ), respectively (Supplementary Fig. 1). However, MiXcan estimates of the epithelial cell proportion were more highly correlated with the expression levels of 126 genes included in the xCell epithelial cell gene signature (median  $r$  of 0.54 in GTEx and 0.60 in TCGA samples) than were the xCell enrichment scores themselves (median  $r$  of 0.36 in GTEx and 0.39 in TCGA samples) indicating that MiXcan can improve cell proportion estimation from its prior (Supplementary Fig. 1).

MiXcan estimated cell-type-specific prediction models for 5473 (84.7%) and nonspecific prediction models for 988 (15.3%) of 6461 genes that had mammary tissue-level prediction models available in PredictDB<sup>28</sup> (Fig. 2). The tissue-level GrEx was computed using MiXcan estimates of the cell-type proportion and predicted cell-type-level GrEx values. The median correlation of predicted GrEx and measured mammary tissue expression levels for the 6461 genes in the TCGA validation set was significantly higher for MiXcan compared with PrediXcan (median  $r$  of 0.41 vs. 0.10;  $p$  value  $< 2.2 \times 10^{-16}$ ) models trained using the same dataset of 125 GTEx EA women. The prediction accuracy for the 5473 genes with cell-type-specific models in MiXcan was

significantly better than PrediXcan (median  $r$  of 0.43 vs. 0.12;  $p < 2.2 \times 10^{-16}$ ), whereas the prediction accuracy for the remaining 988 genes with nonspecific models in MiXcan was the same as PrediXcan (median  $r$  of 0.08 vs. 0.08;  $p$  value=1). These results indicate that allowing for cell-type-level GrEx prediction models increases the prediction accuracy for genes with evidence of cell-specific genetic regulation, and does not decrease the prediction accuracy for other genes compared with standard approaches for predicting tissue-level GrEx.

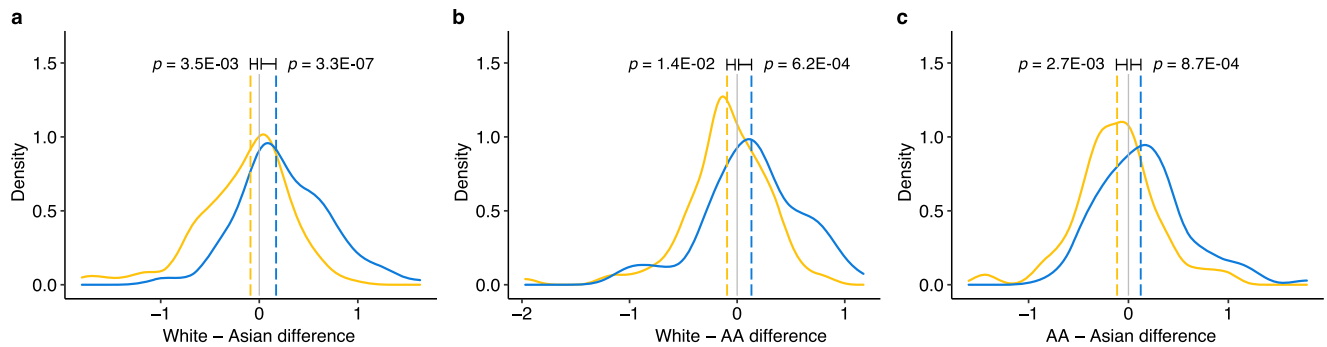
To examine potential sources of the gain in prediction accuracy, three additional approaches were compared with MiXcan and PrediXcan (Supplementary Fig. 2). The median correlation of predicted GrEx with measured mammary tissue-level expression for all 6461 genes in the TCGA validation set was slightly higher for PredictDB ( $r = 0.12$ ) elastic-net models trained using 337 GTEx EA men and women on the entire genome compared with PrediXcan ( $r = 0.10$ ) trained using 125 EA women indicating modest gains from the inclusion of 212 EA men in the training dataset. Accounting for cell composition using penalized regression models including interactions of SNPs with the xCell epithelial cell score (xCell Interaction;  $r = 0.20$ ) or MiXcan cell proportion (MiXcan<sub>0</sub>;  $r = 0.38$ ) led to substantial gains in prediction accuracy. Symmetric estimation of cell-type-level prediction models employed in MiXcan ( $r = 0.41$ ) further improved performance compared with standard interaction models that employ asymmetric penalization for the two cell types. Importantly, whereas standard interaction models require estimates of cell-type composition, which often are unavailable for the tissue of interest in GWAS of human diseases, MiXcan prediction models can be applied directly to GWAS genotype data to perform cell-type-aware TWAS.

Finally, to evaluate the prediction accuracy of MiXcan at the cell-type level, we compared the predicted epithelial cell GrEx with measured mammary epithelial cell snRNAseq data available for three GTEx women of European, Asian, and African ancestry<sup>29</sup>. We found that genes ( $n = 100$ ) predicted to have the largest GrEx differences based on the SNP genotypes in each pair of women also had significantly different measured snRNAseq levels in their mammary epithelial cells ( $p$  value range: 0.01 to  $3.3 \times 10^{-7}$ ), as expected (Fig. 3). The observed snRNAseq differences were significant despite the potentially poorer prediction accuracy of MiXcan models in women of Asian and African ancestry who were not represented in the training dataset. These snRNAseq results support the robustness of the cell-type level GrEx predictions obtained using the MiXcan approach.



**Fig. 2 | Validation of tissue-level GrEx predictions in an independent bulk mammary tissue dataset.** The correlation of tissue-level GrEx predictions using MiXcan or PrediXcan with measured gene expression levels in adjacent normal mammary tissue samples from 103 European ancestry women with breast cancer in TCGA were computed for (a) all 6461 genes with MiXcan and PrediXcan models trained using mammary tissue samples from 125 European ancestry women in

GTEx, (b) 5473 genes with cell-type-specific MiXcan models, and (c) 988 genes with nonspecific MiXcan models. Differences between the correlations for MiXcan and PrediXcan were compared using the two-sided Wilcoxon signed-rank test. Boxplot bounds show the lower, median, and upper quartiles; whisker lengths are 1.5 times the interquartile range; and points beyond the whiskers are outliers. Source data are provided as a Source Data file.



**Fig. 3 | Validation of MiXcan epithelial cell GRex predictions using mammary epithelial cell snRNAseq data.** Measured mammary epithelial cell snRNAseq levels for three GTEx women of White, Asian, and African-American (AA) ancestry were compared for six sets of 100 genes predicted to have the largest GRex differences in each pair of women. Distributions of the observed differences in the

measured snRNAseq levels are shown for the 100 genes predicted to have the largest positive (blue) and negative (yellow) GRex differences for the (a) White – Asian, (b) White – AA, and (c) AA – Asian women. Dashed lines show the median of each distribution, and departures from zero were evaluated using the one-sided Wilcoxon signed-rank test. Source data are provided as a Source Data file.

### Simulation studies

**Type I error and power.** To evaluate type I error and power of MiXcan association tests, datasets were simulated (see “Methods”) under a broad range of realistic settings for the associations of genetic variants with gene expression (SNP-Exp) and gene expression with disease (Exp-Disease). MiXcan predicted GRex with higher accuracy than PrediXcan in the presence of cell-type heterogeneity of SNP-Exp associations, while maintaining comparable accuracy in the absence of cell-specific effects (Supplementary Fig. 3), consistent with results in the independent TCGA validation dataset (Fig. 2).

The type I error was well controlled for MiXcan and PrediXcan under all simulated data scenarios (Fig. 4 col. 1). When SNP-Exp associations were homogeneous in the two cell types (Fig. 4a), the power was similar for MiXcan and PrediXcan whether the Exp-Disease associations were homogeneous or heterogeneous across cell types. Differences in the mean gene expression level between the two cell types that were not determined by SNP-Exp associations (Fig. 4a, b) did not impact the power of MiXcan and PrediXcan indicating robustness to differential expression that is not regulated by genetic variants.

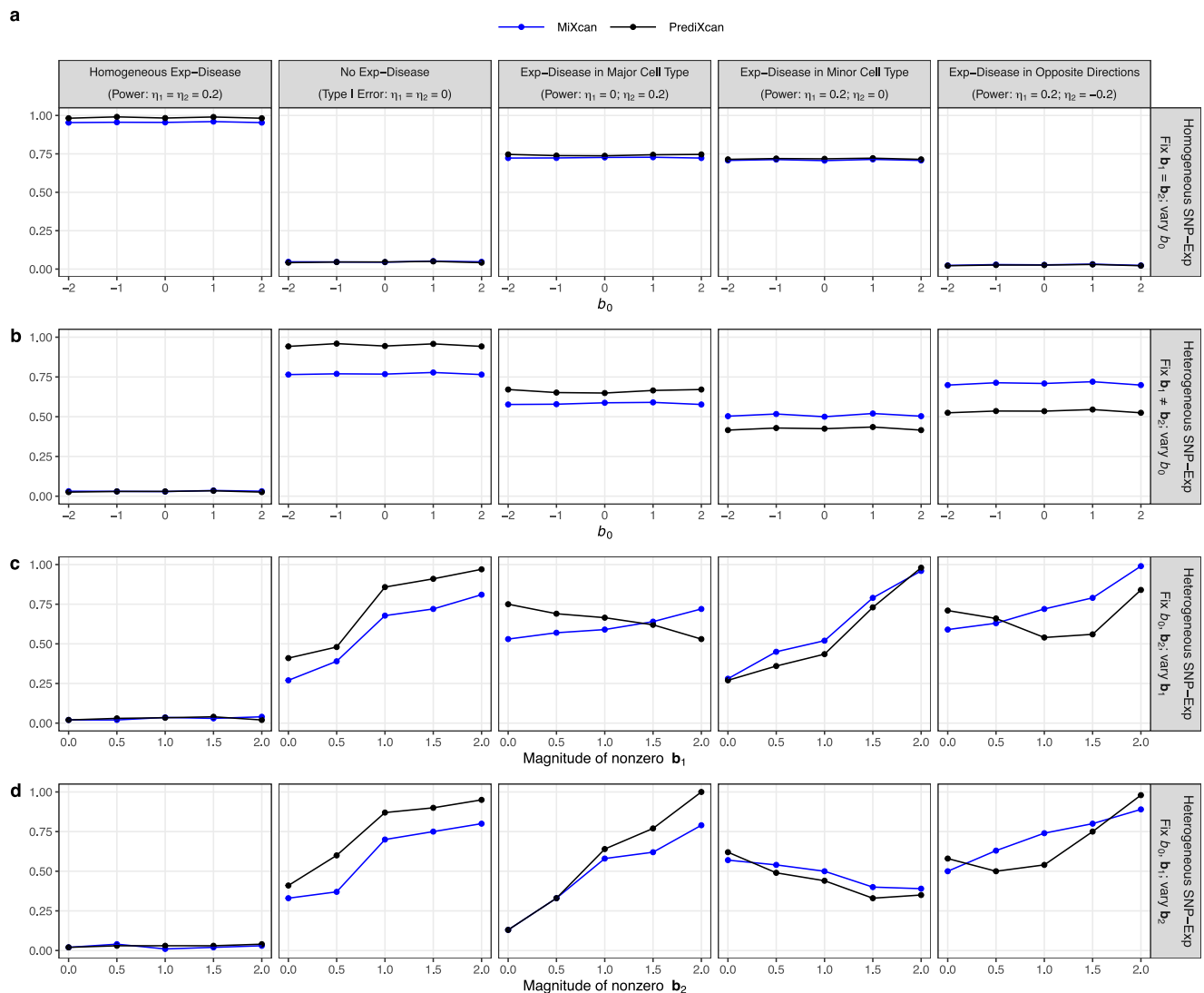
When SNP-Exp associations were heterogeneous in the two cell types (Fig. 4b–d), the relative power of MiXcan and PrediXcan depended on the mechanisms of the Exp-Disease and SNP-Exp associations. PrediXcan was generally more powerful than MiXcan when the Exp-Disease association was either homogeneous across cell types (Fig. 4 col. 2) or present only in the major cell type (Fig. 4 col. 3). However, MiXcan was generally more powerful than PrediXcan when the Exp-Disease association was present only in the minor cell type (Fig. 4 col. 4) or had opposite directions in the two cell types (Fig. 4 col. 5).

As the strength of the SNP-Exp association increased in the same cell type as the Exp-Disease association, the power increased for both PrediXcan and MiXcan (Fig. 4c col. 4; Fig. 4d col. 3). However, as the strength of the SNP-Exp association increased in a different cell type from the Exp-Disease association, the power decreased for PrediXcan but not MiXcan (Fig. 4c col. 3; Fig. 4d col. 4). When the Exp-Disease association had opposite directions in the two cell types, the power was U-shaped for PrediXcan but increased for MiXcan as the strength of the SNP-Exp association increased in either cell type (Fig. 4c–d col. 5). Similar patterns were observed when type I error and power were evaluated in relation to the expression heritability in the two cell types instead of the strength of SNP-Exp associations (Supplementary Fig. 4). These patterns show that different association signals in the two cell types can cancel each other out in PrediXcan, which averages their effects, but are aggregated across cell types in MiXcan thereby preserving power to detect associations due to the minor cell type or that differ across cell types.

In addition to providing valid tissue-level association tests, MiXcan provides information for each cell type separately. Simulation studies showed that the type I error was well controlled for MiXcan cell-type-level tests when no Exp-Disease association was present in any cell type (Supplementary Fig. 5 col. 1). When SNP-Exp associations were homogeneous across cell types (Supplementary Fig. 5a), MiXcan generally estimated nonspecific GRex prediction models which yield the same disease-association test results for all cell types. Thus, cell-type-level inferences can only be made in the heterogeneous SNP-Exp setting (Supplementary Fig. 5b–d), when MiXcan estimates cell-type-specific GRex prediction models. When the Exp-Disease association was present in both cell types in the same or opposite directions (Supplementary Fig. 5 cols. 2 & 5), the power of the cell-type-level tests was similar when the SNP-Exp associations had similar magnitude (regardless of direction) and increased as the magnitude of the SNP-Exp association increased. When the Exp-Disease association was present in only one cell type (Supplementary Fig. 5 cols. 3–4), the power was always highest in this cell type, but the association signal was shared to some degree with the uninvolved cell type. This correlation of the cell-type-level results arises from the joint estimation of the SNP weights for the cell-type-level GRex prediction models in MiXcan. Therefore, we recommend using the combined p-value for all cell types to make inferences regarding whether GRex is significantly associated with disease in any cell type in the tissue, and the cell-type-level results to compare the evidence that different cell types are involved for significant genes.

Finally, we evaluated the impact of the sample size of the training dataset on the type I error and power of MiXcan association tests in simulation studies (Supplementary Fig. 6). As the training dataset increased from 100 to 300 samples, the power of association studies with 3000 samples increased while the type I error remained well controlled. Prediction models trained using only 100–150 samples provided reasonable power for gene identification.

**Performance under model misspecification.** MiXcan decomposes bulk tissue expression levels into two components using an estimate of the cell-type proportion  $\hat{\pi}$  for the cell of interest. First, we evaluated the performance of MiXcan under misspecification of  $\hat{\pi}$  (Fig. 5a). In simulation studies using a broad range of biased and noisy estimates of  $\hat{\pi}$ , the type I error was consistently well controlled. The power of MiXcan also was generally maintained when  $\hat{\pi}$  was misspecified, and compared favorably with PrediXcan when the Exp-Disease association was in the minor cell type or had opposite directions in the two cell types. Second, we evaluated the performance of MiXcan when a latent third cell type was present that had different SNP-Exp associations from the



**Fig. 4 | Simulation studies to evaluate the type I error and power of MiXcan and PrediXcan to detect associations of GREx with disease at the tissue level.** Bulk tissue samples ( $N=300$ ) for training GREx prediction models and independent studies ( $N=3000$ ) for testing disease associations were simulated under a range of realistic data scenarios. Gene expression levels were modeled by  $u = b_0 + \mathbf{b}_1x + e_u$  in the minor cell type,  $v = \mathbf{b}_2x + e_v$  in the major cell type, and  $y = \pi u + (1 - \pi)v$  at the tissue level, where  $\pi$  denotes the minor cell-type proportion,  $b_0$  denotes the mean difference of the gene expression levels in the two cell types, and  $\mathbf{b}_1$  and  $\mathbf{b}_2$  denote the weights for the association of SNPs X with gene expression levels in the

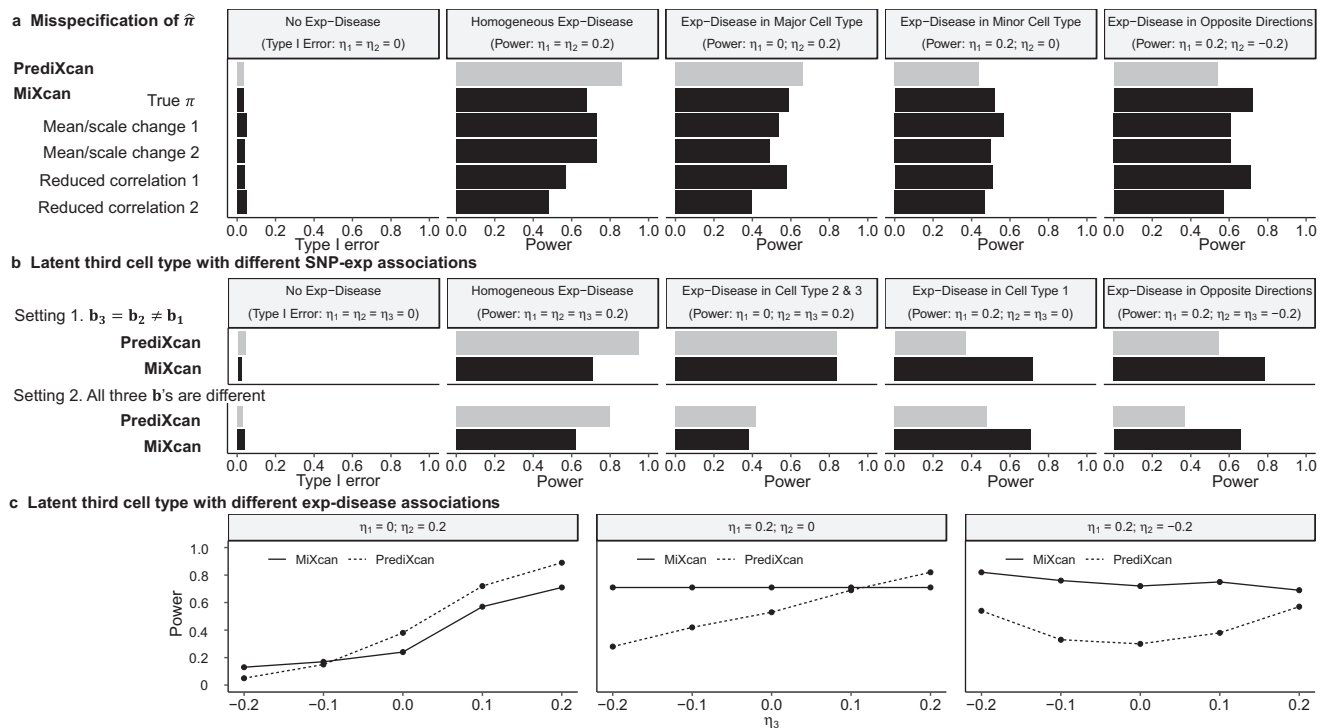
and major cell types, respectively. The disease D was modeled by logit  $P(D=1) = \eta_0 + \eta_1u + \eta_2v$  where  $\eta_1$  and  $\eta_2$  denote the associations of the gene expression levels with disease in the two cell types, respectively. **(a)** Homogeneous SNP-Exp associations ( $\mathbf{b}_1 = \mathbf{b}_2$ ) in the two cell types, varying the mean difference in gene expression levels between the two cell types ( $b_0$ ). Heterogeneous SNP-Exp associations ( $\mathbf{b}_1 \neq \mathbf{b}_2$ ) in the two cell types, varying the: **(b)** mean difference in gene expression levels between the two cell types ( $b_0$ ); **(c)** magnitude of the SNP-Exp association in the minor cell type ( $\mathbf{b}_1$ ); and **(d)** magnitude of the SNP-Exp association in the major cell type ( $\mathbf{b}_2$ ). Source data are provided as a Source Data file.

other cell types (Fig. 5b). We simulated a tissue with three cell types comprising 40%, 50% and 10% of the tissue, respectively, and assumed that MiXcan decomposed the tissue into cell type 1 versus a mixture of cell types 2 and 3. The type I error was consistently well controlled in the presence of a latent third cell type, and the power of MiXcan remained higher than PrediXcan when the Exp-Disease association was in cell type 1 (corresponding to the minor cell type in correctly specified models) or in opposite directions in cell types 1 vs. 2 and 3. The latent third cell type reduced the power of both PrediXcan and MiXcan when the Exp-Disease association was present in the most common cell type or homogeneous across all cell types. Third, we evaluated the impact of Exp-Disease associations in a latent third cell type on study power (Fig. 5c). Similar to the performance under correctly specified models, PrediXcan was more powerful mostly when Exp-Disease associations exist in cell type 2 (the most common cell

type), and MiXcan was more powerful mostly when Exp-Disease associations exist in cell type 1 (the minor cell type of interest) or in opposite directions in cell types 1 and 2.

### Cell-type-aware TWAS of breast cancer

As a proof of concept, we applied MiXcan to conduct the first cell-type-aware TWAS of breast cancer in a publicly available dataset of 58,648 EA women (31,716 cases and 26,932 controls) from the DRIVE GWAS (Discovery, Biology, and Risk of Inherited Variants in Breast Cancer) who were genotyped using the OncoArray<sup>30</sup>. Transcriptome-wide significance was determined using the Bonferroni-corrected threshold of  $7.7 \times 10^{-6}$  to account for the 6461 genes tested, and suggestive associations were determined using the Benjamini-Hochberg false discovery rate (FDR) of 0.10. MiXcan identified 12 significant genes ( $p$  value  $< 7.7 \times 10^{-6}$ ) (Table 1, Fig. 6) and 82 suggestive genes (FDR  $< 0.10$ ) (Supplementary Data 1) whose predicted GREx in



**Fig. 5 | Simulation studies to assess the performance of MiXcan under model misspecification.** **a** Type I error (column 1) and power (columns 2-5) when the estimated cell-type proportion  $\hat{\pi}$  used by MiXcan equals the true  $\pi$  of 0.4;  $0.8\pi$  biasing the mean to 0.32 and changing the scale from (0, 1) to (0, 0.8) (mean/scale change 1);  $0.7\pi + 0.2$  biasing the mean to 0.48 and changing the scale to (0.2, 1) (mean/scale change 2);  $Beta(50\pi, 50(1-\pi))$  reducing the correlation with  $\pi$  to 0.9 (reduced correlation 1); and  $Beta(5.5\pi, 5.5(1-\pi))$  reducing the correlation with  $\pi$  to 0.6 (reduced correlation 2). **b** Type I error (column 1) and power (columns 2-5) when

a latent third cell type contributes to SNP-Exp associations. MiXcan was assumed to decompose the tissue into two components, cell type 1 ( $\pi_1 = 40\%$ ) vs. a mixture of cell types 2 ( $\pi_2 = 50\%$ ) and 3 ( $\pi_3 = 10\%$ ). In Setting 1, the SNP-Exp associations were the same in cell types 3 and 2 and different in cell type 1 ( $b_3 = b_2 \neq b_1$ ). In Setting 2, all three cell types had different SNP-Exp associations ( $b_3 \neq b_2 \neq b_1$ ). **c** Study power for a tissue with a latent third cell type that contributes to Exp-Disease associations under three heterogeneous Exp-Disease settings where  $\eta_1 \neq \eta_2$  and  $\eta_3$  varied from  $-0.2$  to  $0.2$ . Source data are provided as a Source Data file.

mammary tissue were associated with breast cancer risk. In comparison, PrediXcan (trained using the same 125 EA female mammary tissue samples as MiXcan) identified only 8 significant genes ( $p$  value  $< 7.7 \times 10^{-6}$ ), and 31 suggestive genes ( $FDR < 0.10$ ). The inflation factor obtained using the BACON v1.26 Bayesian method to estimate the empirical null distribution was 1.01 for MiXcan and 1.03 for PrediXcan, indicating well-controlled type I error<sup>31</sup>.

Four significant genes (*CH17-437K3.1*, *SLC4A7*, *L3MBTL3*, and *RCCD1*) were identified by both MiXcan and PrediXcan (Table 1). MiXcan estimated nonspecific prediction models for these genes, yielding the same results as PrediXcan. All four genes were near ( $< 500$  kb) breast cancer SNPs previously identified by GWAS<sup>11</sup>, and two genes (*L3MBTL3* and *RCCD1*) also were reported by prior breast cancer TWAS (Supplementary Data 1)<sup>14,15,32</sup>. Follow-up analyses in a larger sample of 228,951 EA women (122,977 cases and 105,974 controls) in the Breast Cancer Association Consortium (BCAC) and DRIVE studies using GWAS summary statistics<sup>11</sup> and S-PrediXcan<sup>33</sup> models for mammary tissue confirmed that the tissue-level GrEx for all four genes were significantly ( $p$  value  $< 7.7 \times 10^{-6}$ ) associated with breast cancer risk.

Eight genes were identified by MiXcan but not PrediXcan (Table 1). MiXcan estimated cell-type-specific GrEx prediction models for all eight of these genes (Supplementary Data 2). Six of these genes (*MRPS30*, *SETD9*, *ADGRV1*, *ZNF703*, *PRR33*, and *PSG4*) showed different directions of association with breast cancer in epithelial vs. stromal (nonepithelial) cells. In MiXcan the signals from the two cell types were aggregated, whereas in PrediXcan they canceled each other out reducing the tissue-level signal. Notably, for *ADGRV1* and *PRR33*, no SNPs in the training dataset were predictive of tissue-level GrEx because of the mixture of the different cell-type effects, and

consequently the PrediXcan association analysis could not be performed. Two genes (*TMEM245* and *CDYL2*) showed the same directions of association, with stronger effects in epithelial vs. stromal cells. These results indicate that MiXcan may be more powerful than PrediXcan in the presence of cell-type heterogeneity of GrEx and when the disease association is present in a minor cell type, e.g. mammary epithelial cells, rather than the predominant cell type in the tissue.

Importantly, MiXcan uniquely identified three novel breast cancer susceptibility genes (*ZNF703*, *TMEM245*, and *PSG4*) that were not previously implicated by breast cancer GWAS<sup>11-13</sup> nor TWAS<sup>14-16,32</sup> (Table 1). For *ZNF703*, GrEx was associated with increased breast cancer risk in stromal cells ( $p$  value= $2.4 \times 10^{-9}$ ) and decreased risk in epithelial cells ( $p$  value= $2.7 \times 10^{-7}$ ). Because adipocytes are the predominant stromal cell type in mammary tissue, we also performed follow-up analyses using the S-PrediXcan subcutaneous fat model and discovered a highly significant ( $p$  value= $2.1 \times 10^{-20}$ ) association of *ZNF703* with increased breast cancer risk in the four-fold larger BCAC/DRIVE dataset, consistent with the MiXcan stromal cell results in the DRIVE data only. For *TMEM245* and *PSG4* the signal was stronger in epithelial cells, which are a minor cell type in mammary tissue and may explain why their tissue-level GrEx was not significantly associated with breast cancer risk. MiXcan also identified two breast cancer genes (*ADGRV1* and *CDYL2*) at previously reported GWAS loci<sup>11</sup> that had different associations in epithelial and stromal cells and were not detected in prior TWAS.

Four genes (*SRGAP2C*, *CASP8*, *ALS2CR12*, and *STXBP4*) were identified by PrediXcan but not MiXcan (Table 1). There was high correlation between the predicted tissue-level GrEx of *CH17-437K3.1* (also identified by MiXcan) and *SRGAP2C* at 1p11.2 ( $r = 0.95$ ) and *CASP8* and *ALS2CR12* at 2q33.1 ( $r = -0.97$ ) indicating that these associations may

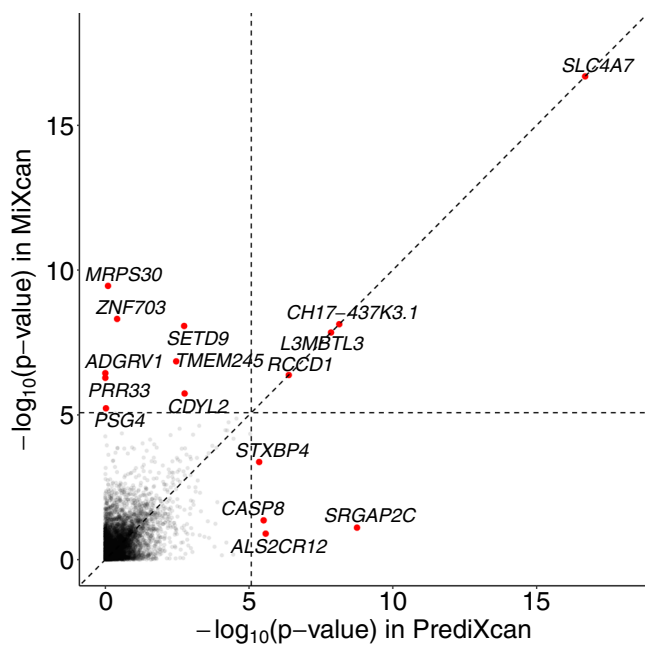
**Table 1 | Genes significantly associated with breast cancer risk using the MiXcan or PrediXcan approaches in 58,648 women, and the corresponding S-PrediXcan and GWAS results in a larger sample of 228,951 women of European ancestry**

Gene Name	Cytoband	MiXcan <sup>a</sup>		Stromal		Combined		PrediXcan <sup>b</sup>		S-PrediXcan <sup>c</sup>		GWAS <sup>d</sup> SNPs within 500 kb			
		Effect	SE	P value	Effect	SE	P value	Effect	SE	Effect	SE		P value		
		Genes identified by both MiXcan and PrediXcan													
Genes identified by both MiXcan and PrediXcan															
CH17-437K3.1	1p11.2	-1.62	0.28	7.3E-09	-1.62	0.28	7.3E-09	7.3E-09	-1.62	0.28	7.3E-09	-0.56	0.05	3.1E-28***	rs11249433
SLC4A7	3p24.1	2.30	0.27	2.0E-17	2.30	0.27	2.0E-17	2.0E-17	2.30	0.27	2.0E-17	1.11	0.09	3.5E-38***	rs4973768
L3MBTL3*	6q23.1	-0.12	0.02	1.4E-08	-0.12	0.02	1.4E-08	1.4E-08	-0.12	0.02	1.4E-08	-0.11	0.02	2.1E-11***	rs6569648
RCCD1*	15q26.1	-0.16	0.03	4.2E-07	-0.16	0.03	4.2E-07	4.2E-07	-0.16	0.03	4.2E-07	-0.16	0.02	3.5E-11***	rs2290203
Genes identified by MiXcan only															
MRPS30	5p12	3.00	0.67	6.8E-06	-4.82	0.75	1.8E-10	3.5E-10	-0.08	0.31	8.1E-01	0.50	0.04	3.3E-30***	rs10941679
SETD9	5q11.2	0.15	0.04	5.2E-04	-0.57	0.10	4.2E-09	8.4E-09	-0.07	0.02	1.8E-03	-0.10	0.02	1.9E-08***	rs62355902
ADGRV1	5q14.3	-0.27	0.05	3.6E-07	0.27	0.05	3.6E-07	3.6E-07	NA	NA	NA	-0.09	0.15	5.5E-01	rs10474352
ZNF703	8p11.23	-1.96	0.38	2.7E-07	1.59	0.27	2.4E-09	4.8E-09	0.07	0.08	3.9E-01	0.17	0.06	4.3E-03*	-
TMEM245	9q31.3	-0.64	0.12	7.0E-08	-0.30	0.10	2.2E-03	1.4E-07	-0.42	0.14	3.4E-03	0.04	0.07	5.4E-01	-
PRR33	11p15.5	-1.41	0.28	5.2E-07	1.41	0.28	5.2E-07	5.2E-07	NA	NA	NA	-1.39	0.14	1.4E-23***	rs3817198
CDYL2	16q23.2	-0.13	0.03	9.1E-07	-0.08	0.03	1.0E-03	1.8E-06	-2.67	0.85	1.7E-03	-0.10	0.04	3.7E-03*	rs13329835
PSG4	19q13.31	-0.16	0.04	3.0E-06	0.14	0.04	1.1E-04	5.9E-06	0.00	0.03	9.5E-01	0.01	0.01	5.3E-01	-
Genes identified by PrediXcan only															
SRGAP2C	1p11.2	-0.19	0.25	4.3E-01	-0.58	0.28	4.1E-02	7.8E-02	-0.65	0.11	1.8E-09	-0.56	0.05	1.2E-28***	rs11249433
CASP8	2q33.1	-0.65	0.30	2.7E-02	0.43	0.27	1.1E-01	4.4E-02	-0.13	0.03	3.1E-06	-0.11	0.02	3.1E-11***	rs1830298
ALS2CR12	2q33.1	-0.10	0.12	4.0E-01	0.40	0.22	7.0E-02	1.3E-01	0.18	0.04	2.6E-06	0.15	0.02	4.2E-12***	rs1830298
STXBP4	17q22	-0.06	0.07	4.1E-01	0.18	0.05	2.1E-04	4.2E-04	0.16	0.03	4.5E-06	0.32	0.03	3.4E-23***	rs2787486

<sup>a</sup>Genes with a MiXcan combined  $p$  value  $< 7.7 \times 10^{-6}$  (0.05/6461 genes tested) aggregating across epithelial and stromal (non-epithelial) cell types were considered statistically significant. <sup>b</sup>denotes same model for epithelial and stromal cell types in MiXcan. <sup>c</sup>Elastic net models were built using GTEx v8 mammary tissue data for 125 European ancestry women and all SNPs in the PredictDB database <https://predictdb.org/>. NA indicates that none of the SNPs evaluated were selected as predictors in the elastic net models for ADGRV1 and PRR33. Genes with a PrediXcan  $p$  value  $< 7.7 \times 10^{-6}$  were considered statistically significant.

<sup>d</sup>S-PrediXcan<sup>35</sup> was used to test associations of predicted GRex with breast cancer risk using GWAS summary statistics for 122,977 cases and 105,974 controls. <sup>†</sup> \*\*\* denotes  $p$  value  $< 7.7 \times 10^{-6}$  and \* denotes  $p$  value  $< 0.05$ .

<sup>e</sup>Breast cancer SNPs reported in a GWAS of 122,977 breast cancer cases and 105,974 controls<sup>31</sup>.



**Fig. 6 | Transcriptome-wide association studies of breast cancer.** MiXcan identified 12 genes and PrediXcan identified 8 genes that were significantly associated with breast cancer risk at  $p$  value  $< 7.7 \times 10^{-6}$ , applying a Bonferroni correction for the 6461 genes tested in 31,716 breast cancer cases and 26,932 controls of European ancestry from the DRIVE study. Source data are provided as a Source Data file.

represent only two independent loci. All four genes identified by PrediXcan only were located near breast cancer SNPs previously identified by GWAS<sup>11</sup>, and three genes (*CASP8*, *ALS2CR12*, and *STXBP4*) also were reported by prior breast cancer TWAS (Supplementary Data 1)<sup>15,32</sup>. Associations of the mammary tissue-level GR<sub>EX</sub> with breast cancer risk were found for all four genes in S-PrediXcan analyses of the BCAC/DRIVE data, as expected. MiXcan estimated cell-type-specific GR<sub>EX</sub> prediction models for these four genes (Supplementary Data 2), and different associations with breast cancer risk in epithelial and stromal cells that did not reach statistical significance in part because of the larger number of model parameters compared with PrediXcan. However, the cell-type-specific MiXcan results for *STXBP4* suggest that stromal cells (estimated effect=0.18;  $p$  value =  $2.1 \times 10^{-4}$ ) may play a more important role than epithelial cells (estimated effect = -0.06;  $p$  value=0.41) in driving the positive association of tissue-level GR<sub>EX</sub> with breast cancer risk.

Finally, we compared the TWAS results using MiXcan and PrediXcan with publicly available PredictDB<sup>28</sup> elastic-net models trained using GTEx mammary tissue data for 212 EA men in addition to 125 EA women (Supplementary Fig. 7). There was substantial overlap of the genes detected by PrediXcan and PredictDB as expected, although PredictDB detected a larger number of genes perhaps because the larger training dataset enabled more accurate prediction models for genes that have similar expression patterns in male and female mammary tissue. However, the etiology and pathobiology of male breast cancer is distinct from female breast cancer<sup>34</sup>. Thus, prior TWAS of breast cancer in EA women<sup>15,32</sup> also trained PrediXcan models using mammary tissue samples ( $n=67$ ) from EA women only, as we did here using a larger number of EA women from GTEx. Cell-type-specific TWAS using MiXcan identified five genes (*ADGRV1*, *ZNF703*, *TMEM245*, *CDYL2*, and *PSG4*), including three novel breast cancer susceptibility genes, that were not identified by either PrediXcan or PredictDB.

## Discussion

MiXcan is a new statistical framework for conducting cell-type-aware TWAS using GWAS data. In contrast to standard TWAS methods,

MiXcan builds cell-type-level prediction models for the genetically regulated component of gene expression and performs association tests taking into consideration the signals from multiple cell types. We have shown that MiXcan improves the prediction accuracy of GR<sub>EX</sub> at both the tissue and cellular levels in independent validation datasets, and improves the power to detect disease associations that are driven by a minor cell type or are heterogeneous between cell types compared with standard approaches. We applied MiXcan to perform the first cell-type-aware TWAS of breast cancer risk and identified three new susceptibility genes (*ZNF703*, *TMEM245*, and *PSG4*) with evidence of distinct associations in mammary epithelial versus stromal cells that were not detected by prior GWAS<sup>11–13</sup> nor TWAS<sup>14–16,32</sup>. These findings provide a proof of concept that cell-type-aware TWAS can reveal novel insights into the genetic and cellular etiology of human diseases.

Several recent studies have explored context-specific TWAS<sup>35–37</sup>. Li et al. proposed a tissue-specificity-aware TWAS framework that uses prior knowledge of trait-related tissue types for accurate detection of single-tissue and cross-tissue TWAS<sup>35</sup>. Feng et al. proposed to derive cross-tissue expression features using sparse canonical correlation analysis, and then combine expression-outcome associations across single- and cross-tissue features for powerful detection<sup>36</sup>. Thompson et al. proposed CONTENT to go one step further and model both shared and tissue-specific components of gene expression in bulk multi-tissue data for model construction<sup>37</sup>. This approach can also be used for modeling shared and cell-type-specific components in single-cell RNAseq data<sup>37</sup>. These recently developed TWAS methods model associations at the same resolution of the data, such as modeling tissue-level associations for bulk profiling data and cell-type-level associations for single-cell data, and thus do not provide higher-resolution (single cell) understanding of disease using lower-resolution (bulk tissue) data as does MiXcan.

Recent studies have also explored methods for performing cell-type-level association analyses when the tissue-level data are available for all study participants<sup>19,38–42</sup>. Luo et al. evaluated cell-type-specific associations between DNA methylation and traits<sup>40</sup>, but this method did not involve prediction models and methylation data are required for all subjects. Liu et al. built tissue-level GR<sub>EX</sub> prediction models, inferred cell types from the predicted GR<sub>EX</sub>, and looked for associations of the inferred cell-type proportions with disease rather than constructing a TWAS framework for identifying genes<sup>42</sup>. In this study, MiXcan enables cell-type-aware TWAS in large populations using existing GWAS datasets that do not have transcriptomic and cell composition data from the disease-relevant tissue. MiXcan evaluates the composite null hypothesis that there is no association between the GR<sub>EX</sub> in any cell type with the disease, which tolerates decomposition uncertainty to provide robust cell-type-aware analysis using bulk tissue samples. By carefully modeling cell-type-level expression, MiXcan is more powerful than PrediXcan when disease associations are driven by a minor cell type or have opposite directions in different cell types. However, when the association of GR<sub>EX</sub> with disease is similar in all cell types or driven by the major cell type, then conventional TWAS approaches using more parsimonious tissue-level GR<sub>EX</sub> prediction models that assume cell-type homogeneity can be more powerful. Thus, these two TWAS approaches are complementary, and additional cell-type-aware analyses are especially valuable for diseases where cell-type heterogeneity and a minor cell of origin are hypothesized, as for breast carcinoma and many other human diseases.

To construct cell-type-level GR<sub>EX</sub> prediction models, MiXcan uses a scaled xCell<sup>18</sup> cell-type enrichment score in the training data as prior information. While estimates from other approaches<sup>20,21,23,43–45</sup> can also be used as priors, xCell is among the most widely used. Building upon the priors, MiXcan fits mixture models for the expression levels of the epithelial cell signature genes in the training data to improve estimation of the cell-type proportion. By incorporating better estimates of the cell-type proportion, and penalizing all cell types equally, MiXcan



improves the accuracy of the GRex prediction models, as well as the power and type I error of the downstream association tests. Compared with standard interaction models that include interaction effects between cell-type enrichment scores and genetic variants<sup>19</sup>, an important advantage of MiXcan is that the predicted values correspond to the cell-type-specific GRex, which are directly interpretable and biologically meaningful. Moreover, standard interaction models require cell composition information, which is often unavailable for the tissue of interest, whereas MiXcan prediction models can be applied directly to GWAS data without transcriptomic or cell composition data to interrogate genetic associations within the disease-relevant tissue and cell type context.

MiXcan has several limitations. First, although the MiXcan framework is generalizable, the model building procedure requires hypotheses regarding the disease-related cell types and tissue. The prediction models for this cell-type-aware TWAS of breast cancer were developed specifically for human mammary tissue with a focus on distinguishing epithelial and stromal (nonepithelial) cells, which have distinct roles in breast carcinogenesis<sup>8–10</sup>. Second, the current MiXcan framework decomposes a tissue into two cell-type components only. The MiXcan framework can be extended naturally to model more than two cell types. However, given limited sample sizes of the available training datasets, the model performance becomes quite variable with additional cell types due to the rapidly increasing dimension of the parameter space and complexity of the numerical optimization. As larger training sets with bulk tissue transcriptomic and genomic data become available, a careful evaluation of its analytical performance can be performed for additional less common cell types. Third, paired single-cell RNAseq and genomic datasets are presently very limited, and the validation of MiXcan epithelial cell predictions was performed for a small set of genes using epithelial cell snRNAseq data available for only three women in GTEx. Human single-cell transcriptome profiling efforts<sup>46,47</sup> are currently underway, and will enable further evaluation of the performance of MiXcan in larger datasets. Future studies can also investigate integrating single-cell transcriptome profiles into MiXcan, for example to improve estimation of cell composition in bulk tissue samples<sup>23,45</sup> or to provide an initial estimate of SNP weights, which could be used to tune separate penalty terms for different SNPs in adaptive elastic-net models<sup>48</sup>.

Human mammary tissue has variable cell composition and numerous eQTLs with distinct effects in epithelial cells and adipocytes, which are a major stromal cell type in the breast<sup>19</sup>. Cell-type-aware TWAS using MiXcan mammary tissue models applied to publicly available GWAS data identified three new breast cancer susceptibility genes that were associated with disease risk through their GRex in normal mammary epithelial or stromal cells. *ZNF703* (zinc finger protein 703) is an oncogene that is commonly amplified in luminal B breast tumors, and has been shown to regulate genes involved in proliferation, invasion, and an altered balance of progenitor stem cells<sup>49–51</sup>. To our knowledge, common germline variants in *ZNF703* have not previously been implicated in breast cancer risk. Our finding that genetic upregulation of *ZNF703* in normal mammary stromal cells (predominantly adipocytes) was associated with increased breast cancer risk in 58,648 women was confirmed by a highly significant ( $p$ -value =  $2.1 \times 10^{-20}$ ) association of tissue-level GRex predicted using S-PrediXcan subcutaneous fat models in 228,951 EA women, which has not previously been reported to our knowledge. Notably, the mammary tissue-level results for *ZNF703* did not reach transcriptome-wide significance, underscoring the importance of accounting for cell-type heterogeneity to elucidate disease etiology. *TMEM245* (transmembrane protein 245) is the host gene for microRNA 32, which has been shown to promote proliferation and suppress apoptosis of breast cancer cells<sup>32</sup>. Relatively little is known about *PSG4* (pregnancy specific beta-1-glycoprotein 4), a member of the carcinoembryonic antigen gene family that may

play a role in regulation of the innate immune system<sup>53</sup>. Future studies are needed to provide experimental validation of the breast cancer genes identified in this study and better understand the cellular mechanisms underlying the associations.

In conclusion, the MiXcan framework enables cell-type-aware TWAS using prediction models that allow for differences across cell types in the disease-relevant tissue. MiXcan mammary tissue models are available at <https://github.com/songxiaoyu/MiXcan><sup>54</sup> and can be applied to GWAS genotype data to identify genes associated with complex traits through their GRex in epithelial or stromal cells. MiXcan software is also freely available to facilitate training prediction models for other tissues and cell types, and conducting cell-type-aware TWAS. MiXcan prediction models had excellent performance in independent validation datasets, and identified new breast cancer susceptibility genes in the first cell-type-aware TWAS of breast cancer. These findings provide a proof of concept that cell-type-aware TWAS are feasible using existing bulk tissue training datasets and GWAS data, and can lead to the discovery of new disease genes and cellular mechanisms. Future research is needed to develop MiXcan models for other human tissues and cell types, extend the MiXcan framework to GWAS summary statistics, and explore alternative modeling and inference strategies<sup>2,55,56</sup>. These research directions will enable the broad application of cell-type-aware TWAS to improve our understanding of the genetic and cellular mechanisms underlying human diseases.

## Methods

### MiXcan framework

We first summarize PrediXcan<sup>1</sup>, an established tissue-level TWAS framework, and then present the MiXcan cell-type-aware TWAS framework. Let  $y_i$  denote the measured expression level of a gene in the bulk tissue sample  $i \in (1, \dots, N)$ ,  $\mathbf{x}_i$  denote the genetic variants (e.g. SNPs) used in model training, and  $\mathbf{z}_i$  denote the non-genetic covariates (e.g. age).

**PrediXcan tissue-level TWAS framework.** PrediXcan uses a linear additive model to characterize the gene expression level:

$$y_i = a + \mathbf{x}_i^T \mathbf{b} + \mathbf{z}_i^T \mathbf{c} + e_i, \quad (1)$$

where  $e_i \sim N(0, \sigma^2)$  is the model error,  $\mathbf{x}_i^T \mathbf{b}$  is the genetically determined component of gene expression, and  $\mathbf{z}_i^T \mathbf{c}$  is the non-genetically determined component. This model can be estimated with the elastic-net method<sup>57</sup>, which maximizes the following penalized log-likelihood function:

$$(\hat{a}, \hat{\mathbf{b}}, \hat{\mathbf{c}}, \hat{\sigma}) = \arg \max_{\mathbf{a}, \mathbf{b}, \mathbf{c}, \sigma} \sum_{i=1}^n \left( -\frac{(y_i - a - \mathbf{x}_i^T \mathbf{b} - \mathbf{z}_i^T \mathbf{c})^2}{2\sigma^2} - \frac{1}{2} \log(\sigma^2) \right) - \lambda(P(\mathbf{b}) + P(\mathbf{c})), \quad (2)$$

where  $P(\cdot) = \alpha \|\cdot\|_2^2 + (1 - \alpha) \|\cdot\|_1$  is the elastic-net penalty function with the mixing parameter  $\alpha \in (0, 1)$ . In PrediXcan,  $\alpha$  is set at 0.5 and  $\lambda$  is selected via 10-fold cross validation (CV)<sup>58</sup>. The estimated SNP weights  $\hat{\mathbf{b}}$  can be used to predict the GRex by  $\hat{y}_j = \hat{\mathbf{x}}_j^T \hat{\mathbf{b}}$  where  $\hat{\mathbf{x}}_j$  denotes the SNP genotypes of the GWAS subject  $j \in (1, \dots, M)$ . Then, the association of  $\hat{y}_j$  with the phenotype (e.g. disease status)  $d_j$  can be evaluated using a generalized linear model  $g(d_j) = \eta_0 + \hat{y}_j \eta_1$ , where  $g(\cdot)$  is a link function. The null and alternative hypotheses,  $H_0: \eta_1 = 0$  vs.  $H_A: \eta_1 \neq 0$ , test whether the GRex at the tissue level is associated with the phenotype.

**MiXcan cell-type-level GRex prediction model.** MiXcan extends upon PrediXcan to enable cell-type-aware TWAS. In this section, we build the prediction models for the cell-type-level GRex. In the next

section, we develop strategies for applying these prediction models in disease-association studies.

Human bulk tissue samples from solid tissue (e.g. mammary tissue) comprise a mixture of cells of different types. Let  $\pi_i$  and  $1 - \pi_i$  be the proportions of the cell type of interest (e.g. epithelial cells) and all other cell types (e.g. stromal cells) in the  $i^{\text{th}}$  tissue sample, respectively. We assume the observed bulk tissue expression level of a given gene  $y_i$  is a linear combination of expression levels in the cell types. Introducing two latent variables  $u_i$  and  $v_i$  to denote the unobserved average gene expression levels in the epithelial and stromal cells, respectively, we have:

$$y_i = \pi_i u_i + (1 - \pi_i) v_i. \tag{3}$$

We model both  $u_i$  and  $v_i$  using the linear additive models:

$$u_i = a_u + \mathbf{x}_i^T \mathbf{b}_u + \mathbf{z}_i^T \mathbf{c} + e_{ui}, \quad \text{and} \quad v_i = a_v + \mathbf{x}_i^T \mathbf{b}_v + \mathbf{z}_i^T \mathbf{c} + e_{vi}, \tag{4}$$

where  $e_{ui} \sim N(0, \sigma_u^2)$  and  $e_{vi} \sim N(0, \sigma_v^2)$  are the model errors. In Eq. (4), the intercepts  $a_u$  and  $a_v$  and genetic parameters  $\mathbf{b}_u$  and  $\mathbf{b}_v$  differ between cell types, allowing for different mean expression levels and cell-type-specific effects of genetic variants on gene expression. A shared parameter  $\mathbf{c}$  is used for the non-genetic component  $\mathbf{z}_i$  to simplify the model as the non-genetic variables are not necessarily used in downstream analyses.

The proportion of epithelial cells  $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_N\}$  is a feature of the tissue samples, which can be jointly estimated using multiple genes, whereas  $\boldsymbol{\theta} = (a_u, a_v, \mathbf{b}_u, \mathbf{b}_v, \mathbf{c}, \sigma_u^2, \sigma_v^2)$  are features of each gene for investigation. Therefore, we present a step-wise procedure to first estimate  $\boldsymbol{\pi}$  using multiple epithelial signature genes, and then estimate the cell-type-level effects  $\boldsymbol{\theta}$  for each gene.

**Estimation of  $\boldsymbol{\pi}$ .** The notation in the sections above focuses on individual genes, and here we introduce additional notation to describe the joint modeling of multiple genes. Let  $W_{N \times G} = \{w_i^g\}$  be the observed expression of  $G$  epithelial signature genes in  $N$  tissue samples; and  $S_{N \times G} = \{s_i^g\}$  and  $T_{N \times G} = \{t_i^g\}$  be the unobserved gene expression levels in epithelial and stromal cells, respectively. Similarly as in Eq. (3), we have:

$$w_i^g = \pi_i s_i^g + (1 - \pi_i) t_i^g, \quad g = 1, \dots, G, \quad i = 1, \dots, N. \tag{5}$$

Leveraging primarily the mean differences of signature genes in epithelial and stromal cells for estimating  $\boldsymbol{\pi}$ , we model the marginal distributions of individual genes and omit the complex gene-gene correlations for computationally efficient estimation, as supported by our previous work<sup>59</sup>. Specifically, we assume:

$$s_i^g \sim N(\mu_{Si}^g, \sigma_{Si}^g), \quad \text{and} \quad t_i^g \sim N(\mu_{Ti}^g, \sigma_{Ti}^g).$$

Across all  $G$  genes, the parameters include  $\boldsymbol{\Gamma} = \{\{\boldsymbol{\Gamma}^g\}\}_{g=1}^G$ , where  $\boldsymbol{\Gamma}^g = (\boldsymbol{\mu}_S^g, \boldsymbol{\mu}_T^g, \sigma_S^g, \sigma_T^g)$ .

In parallel, we also take advantage of a prior cell-type proportion estimate  $h_i$  based on existing tools (i.e. rescaled xCell<sup>18</sup> enrichment scores). We link the prior estimates  $h_i$  to the true  $\pi_i$  using a *Beta* distribution such that  $h_i \sim \text{Beta}(\pi_i \delta, (1 - \pi_i) \delta)$  for some positive parameter  $\delta$ . We have  $E(h_i) = \pi_i$  and  $\text{var}(h_i) = \pi_i(1 - \pi_i)/(\delta + 1)$  such that  $h_i$  is an unbiased estimator of the true  $\pi_i$  with variation.

We then join these two models for parameter estimation and solve the following maximization problem:

$$(\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\Gamma}}, \hat{\delta}) = \arg \max_{\boldsymbol{\Gamma}, \boldsymbol{\pi}, \delta} \sum_{i=1}^N \left[ \sum_{g=1}^G l(\boldsymbol{\Gamma}, \pi_i | w_i^g) + l(\pi_i, \delta | h_i) \right], \tag{6}$$

where  $\sum_{g=1}^G l(\boldsymbol{\Gamma}, \pi_i | w_i^g)$  and  $l(\pi_i, \delta | h_i)$  are the log-likelihood of the observed gene expression profile and cell proportion estimate of

the  $i^{\text{th}}$  sample, respectively. This optimization problem is solved using an Expectation-Maximization (EM) algorithm similar to that in Petralia et al.<sup>59</sup>

To enhance the robustness of the estimation, we implement a bagging strategy to estimate the parameters with randomly selected bootstrap samples, and aggregate multiple estimates by calculating a tail truncated mean. This bagging strategy further stabilizes the estimates, and may also be used to investigate the consistency of  $\boldsymbol{\pi}$  estimation.

**Estimation of  $\mathbf{b}_u$  and  $\mathbf{b}_v$ .** Given  $\hat{\boldsymbol{\pi}}$ , we next estimate  $\mathbf{b}_u$  and  $\mathbf{b}_v$  in Eq. (4). Since  $u_i$  and  $v_i$  are unobserved, we integrate Eqs. (3) and (4) to have:

$$y_i = \hat{\pi}_i (a_u + \mathbf{x}_i^T \mathbf{b}_u + \mathbf{z}_i^T \mathbf{c} + e_{ui}) + (1 - \hat{\pi}_i) (a_v + \mathbf{x}_i^T \mathbf{b}_v + \mathbf{z}_i^T \mathbf{c} + e_{vi}). \tag{7}$$

This equation can be rearranged as:

$$y_i = a_v + \hat{\pi}_i (a_u - a_v) + \mathbf{x}_i^T \mathbf{b}_v + \hat{\pi}_i \mathbf{x}_i^T (\mathbf{b}_u - \mathbf{b}_v) + \mathbf{z}_i^T \mathbf{c} + \epsilon_i, \tag{8}$$

where  $\epsilon_i = \hat{\pi}_i e_{ui} + (1 - \hat{\pi}_i) e_{vi}$ . A simple strategy for estimating  $\mathbf{b}_u$  and  $\mathbf{b}_v$  is to apply elastic-net regression to Eq. (8). Specifically, the elastic-net regularization is put on  $(\mathbf{b}_v, \mathbf{b}_u - \mathbf{b}_v, \mathbf{c})$ —the dependence of expression levels on genetic variants and covariates—but not on  $(a_v, a_u - a_v)$ —the mean expression levels in the two cell types. We refer to this strategy as MiXcan<sub>0</sub>.

One issue with MiXcan<sub>0</sub> is that the two cell components in the mixture model are not treated in a symmetric manner. In other words, the penalization on  $\mathbf{b}_u$  and  $\mathbf{b}_v$  differs:  $\mathbf{b}_v$  is shrunk towards zero, while  $\mathbf{b}_u$  is shrunk towards  $\mathbf{b}_v$ . This asymmetric penalization results in different models if the order of the two components is switched. To address this issue, we introduce  $\hat{c}_i = \hat{\pi}_i - 0.5$  and rewrite Eq. (8) as:

$$y_i = \frac{a_u + a_v}{2} + \hat{c}_i (a_u - a_v) + \mathbf{x}_i^T \frac{\mathbf{b}_u + \mathbf{b}_v}{2} + \hat{c}_i \mathbf{x}_i^T (\mathbf{b}_u - \mathbf{b}_v) + \mathbf{z}_i^T \mathbf{c} + \epsilon_i. \tag{9}$$

When fitting elastic-net regression to Eq. (9), we include penalties on  $(\frac{\mathbf{b}_u + \mathbf{b}_v}{2}, (\mathbf{b}_u - \mathbf{b}_v), \mathbf{c})$ , which impose the same degree of regularization on  $\mathbf{b}_u$  and  $\mathbf{b}_v$ ; the penalty on  $\frac{\mathbf{b}_u + \mathbf{b}_v}{2}$  regularizes the overall sparsity of the genetic effects, while the penalty on  $\mathbf{b}_u - \mathbf{b}_v$  encourages similarities between the two components. We refer to this strategy as MiXcan.

Note, when fitting elastic-net regressions in MiXcan<sub>0</sub> and MiXcan, we do not consider the varying variances of  $\epsilon_i$  as  $\hat{\pi}_i$  takes different values. This is because the residual variance structure has limited impact on the coefficient estimates, especially for regularized regression. In a trade-off between extensive computational costs (allowing varying residual variances) and minimal sacrifice of estimation accuracy (assuming constant variance), we chose the latter and take advantage of the fast implementation of elastic-net regression in the *glmnet* package.

**Model Aggregation.** In the prediction models, the term  $\hat{\mathbf{b}}_u - \hat{\mathbf{b}}_v$  is of particular importance: a non-zero value suggests that the dependence structure between genetic variants and expression levels is cell-type-specific. Therefore, it is critical to know the selection robustness of  $\hat{\mathbf{b}}_u - \hat{\mathbf{b}}_v$ . We employ a procedure similar to stability selection<sup>60</sup> for its evaluation. Specifically, for models that select non-zero  $\hat{\mathbf{b}}_u - \hat{\mathbf{b}}_v$ , we generate  $B$  bootstrap samples (e.g.  $B = 200$ ), perform ordinary least square analysis on the pre-selected variables, and record  $\widehat{\text{diff}}_b^{(b)} = \hat{\mathbf{b}}_u^{(b)} - \hat{\mathbf{b}}_v^{(b)}$  for  $b = 1, \dots, B$ . Only when the 95% confidence interval (CI) for  $\widehat{\text{diff}}_b$  excludes 0 do we employ cell-type-specific prediction models (inferred using the complete data set). Otherwise, nonspecific models that have the same prediction weights for the two cell types will be used, as in Eq. (2) of PrediXcan.

**Association analysis with cell-type-level prediction models.** The model building procedure in MiXcan selects cell-type-specific prediction models for some genes and nonspecific models for other genes. Cell-type-specific prediction models estimate different SNP weights in the two cell types ( $\hat{\mathbf{b}}_u \neq \hat{\mathbf{b}}_v$ ) resulting in different predicted GREX from the same genotype data  $\tilde{x}_j$  ( $j \in (1, \dots, M)$ ), such that  $\tilde{y}_{uj} = \tilde{x}_j^T \hat{\mathbf{b}}_u$  and  $\tilde{y}_{vj} = \tilde{x}_j^T \hat{\mathbf{b}}_v$ . These cell-type-specific GREX levels cannot be combined into tissue levels in GWAS datasets that lack cell-type proportion estimates, requiring a novel statistical framework for association analysis. One natural idea is to infer cell-type-specific associations by directly associating the phenotype  $d_j$  with  $\tilde{y}_{uj}$  and  $\tilde{y}_{vj}$ , either separately, such that  $g(d_j) = \eta_0 + \eta_u \tilde{y}_{uj}$  and  $g(d_j) = \eta'_0 + \eta_v \tilde{y}_{vj}$ , or jointly, such that  $g(d_j) = \eta_0 + \eta_u \tilde{y}_{uj} + \eta_v \tilde{y}_{vj}$ . As  $\tilde{y}_{uj}$  and  $\tilde{y}_{vj}$  are predicted from the same genotype data with prediction weights jointly learned using the bulk tissue data in MiXcan, they may capture leaked information from each other. As a result, if an association exists in one cell type, analysis in the other cell type may also capture this association, resulting in an inflated type I error for inferring associations in each cell type. To avoid this inflation, we propose a composite hypothesis test to test whether association exists in any cell types in the tissue:

$H_0$ : There is no association in any cell type in the tissue vs.

$H_A$ : There is an association in at least one cell type in the tissue.

This composite null is robust against information leakage, as the leaked values under the null are not associated with the phenotype. To perform the test, we first associate  $d_j$  with  $\tilde{y}_{uj}$  and  $\tilde{y}_{vj}$ , separately if the  $(\tilde{y}_{uj}, \tilde{y}_{vj})$  are highly correlated (e.g.  $r = \pm 1$ ), or jointly otherwise. Then, we propose to aggregate the resulting  $p$  values  $p_u$  and  $p_v$  for  $\tilde{y}_{uj}$  and  $\tilde{y}_{vj}$  using Cauchy combination<sup>27</sup>. The Cauchy combination provides valid test for correlated  $p$ -values, and in this setting the test statistic can be written as:

$$T_{Cauchy} = \tan\{\pi(0.5 - p_u)\} + \tan\{\pi(0.5 - p_v)\}, \quad (10)$$

where  $\pi$  is the mathematical constant approximately equal to 3.14159. The combined  $p$ -value for the tissue is approximated by:

$$p_{tissue} \approx 1/2 - \{\arctan(T_{Cauchy})\}\pi. \quad (11)$$

The  $p_{tissue}$  tests whether association exists in any cell type in the tissue. Unlike PrediXcan that tests associations averaged across all cell types, the  $p_{tissue}$  in MiXcan accumulates signals from different cell types. Note that  $p_u$  and  $p_v$  are building blocks of  $T_{Cauchy}$  and the resulting  $p_{tissue}$  is between  $p_u$  and  $p_v$ . After the tissue-level hypothesis test,  $p_u$  and  $p_v$  can provide information on the cell type(s) driving the association. For example,  $p_u \ll p_v$  indicates that a significant  $p_{tissue}$  is primarily driven by epithelial cells.

Some genes have nonspecific models with the same estimated SNP weights and predicted GREX in the two cell types ( $\hat{\mathbf{b}}_u = \hat{\mathbf{b}}_v$ ). While association analyses can follow the same strategy described above for cell-type-specific models, it is equivalent to performing a single tissue-level association analysis as in PrediXcan. Finally, in transcriptome-wide studies,  $p_{tissue}$  from genes with cell-type-specific prediction models and  $p$ -values from genes with nonspecific models can be jointly used to adjust for multiple testing, and infer transcriptome-wide significant discoveries.

### Build MiXcan prediction models using GTEx mammary tissue data

MiXcan gene expression prediction models were developed using the GTEx v8 genotype and gene expression data for mammary tissue samples from 125 European ancestry women (dbGaP accession number phs000424.v8.p2 <[https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000424.v8.p2](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2)>). PredictDB (<http://predictdb.org>) provides tissue-level expression prediction models

trained on 337 men and women of European ancestry with mammary tissue data available in GTEx v8. PredictDB included mammary tissue elastic-net models for a total of 6461 genes that were well predicted by genetic variants (176,983 SNPs). An additional 1,715 SNPs were included in PredictDB mammary tissue MASHR models<sup>61</sup>. For the purpose of comparison to PrediXcan as a proof of concept, we developed cell-type-level prediction models for these 6461 genes using all 178,698 SNPs in the PredictDB mammary tissue database. MiXcan cell-type-level prediction models were developed for mammary epithelial cells, the cell of origin for breast carcinoma, and stromal (non-epithelial) cells.

**$\pi$  estimation.** The epithelial cell proportion  $\pi_i$  was estimated using 126 epithelial cell signature genes<sup>18</sup> available in the training dataset of 125 GTEx female mammary tissue samples. We first computed xCell gene set enrichment scores for epithelial cells and 63 other cell types using the curated set of cell signature genes for each cell type<sup>18</sup> and the bulk tissue transcriptomic data for each sample. We then re-scaled the xCell epithelial cell enrichment score to range from 0 to 1 for use as a prior estimate of the cell-type proportion in MiXcan. The  $\pi_i$  estimation was performed using 100 bootstrap samples (80% random draw with replacement). The final estimate was computed by excluding the most extreme 5% of bootstrap estimates in each of the two tails and averaging the remaining estimates.

**Prediction model.** Using  $\hat{\pi}$ , we modeled the cell-type-level expression levels for each of the 6461 genes using MiXcan with tuning parameter  $\lambda$  selected by tenfold CV. We adjusted for covariates that were used in GTEx eQTL analyses including age, platform, PCR, genomic principal components (PC) 1-5, and PEER factors 1-15<sup>62</sup>. For genes with  $\hat{\mathbf{b}}_u \neq \hat{\mathbf{b}}_v$ , we performed ordinary least squares regression on the pre-selected variables for 200 bootstrap samples, and calculated the 95% bootstrap CI of  $\hat{\mathbf{b}}_u - \hat{\mathbf{b}}_v$ . If the 95% CI excluded 0, we used cell-type-specific models with parameters estimated using the full data; otherwise, we used nonspecific models that were the same as PrediXcan. The average computation time for training models for 1000 genes using 125 samples on a single CPU core was 11 min (standard deviation, 2.8 minutes).

### Evaluate MiXcan prediction accuracy in independent TCGA data

We evaluated the prediction performance of MiXcan in an independent dataset of 103 European ancestry female breast cancer patients with adjacent normal tissue samples from TCGA<sup>63,64</sup>. To minimize the study effect, we re-processed the TCGA gene expression data using methods analogous to those used to process the GTEx expression data (<https://gtexportal.org/home/documentationPage>). Briefly, we required genes to have Transcripts Per Kilobase Million (TPM) >0.1 in at least 20% of samples, and at least six reads in at least 20% of samples, resulting in a set of 25,702 out of 25,849 total genes that met these quality control (QC) requirements. Expression data were then normalized using the trimmed mean of M values method (TMM)<sup>65</sup> as implemented in the R package edgeR v3.16.5<sup>66</sup>, and the results were quantile-normalized to a standard normal distribution with mean=0 and variance = 1. Comparison of these normalized gene expression levels showed no systematic differences between the GTEx and TCGA data (Supplementary Fig. 7). To process the genotype data, we removed all indels, monomorphisms, and ambiguous pairs (e.g. A/T, C/G). SNPs with >5% missing genotypes or Hardy-Weinberg equilibrium (HWE) test  $p$  value <  $1e-05$  were also removed. The remaining SNPs were aligned to build 37 coordinates, and imputation was performed on the TOPMed imputation server<sup>67</sup>. A total of 97% (54,663 out of 56,531) and 97% (52,031 out of 53,876) SNPs used in MiXcan and PrediXcan prediction models, respectively, were available for analysis.

We estimated the epithelial cell proportion in the TCGA samples as described above, and used this estimate to combine

the predicted cell-type-level GRex from epithelial and stromal components into the tissue level. To evaluate predication accuracy, we computed the Pearson correlation between the predicted and observed tissue-level gene expression, and compared the results with the predicted tissue-level expression using PrediXcan. The observed bulk tissue expression levels showed significantly higher correlation with the tissue-level GRex predicted by MiXcan compared with PrediXcan. To investigate the sources of the improved performance, we compared five approaches for predicting tissue-level GRex:

- Existing PredictDB elastic-net models (PredictDB)
- PrediXcan trained on the same dataset as MiXcan (PrediXcan)
- Prediction model including interactions between SNPs and the xCell enrichment score (xCell interaction)
- Prediction model including interactions between SNPs and the MiXcan cell-type proportion (MiXcan<sub>0</sub>)
- Cell-type-level prediction models with symmetric penalization (MiXcan).

These comparisons evaluated incorporating cell-type composition, use of the MiXcan cell-type proportion estimate, and symmetric penalization of the two cell types in prediction models, as well as use of a larger training set including both men and women in PredictDB. It is worth noting that “MiXcan<sub>0</sub>” and “xCell interaction” models are not applicable to GWAS datasets that lack cell-type composition information for the tissue of interest, and are included here only for the purpose of understanding the sources of improved prediction performance for MiXcan.

### Evaluate MiXcan epithelial cell prediction accuracy in snRNAseq data

We evaluated the performance of MiXcan epithelial cell prediction models using snRNAseq data for normal mammary epithelial cells and paired genomic data available for three women of European, Asian and African ancestry from GTEx v9<sup>29</sup>. Details of the snRNAseq data generation and processing were provided in<sup>29</sup>. The preprocessed log count expression profiles for a total of 5990, 2324 and 1456 nuclei, including 2292 (38%), 2180 (94%) and 1327 (91%) epithelial cell nuclei, from the European, Asian and African ancestry woman, respectively, were downloaded from <https://gtexportal.org/home/datasets>. Mammary epithelial cell snRNAseq data were available in all three women for 4751 genes with MiXcan prediction models. To enable comparisons between women, the snRNAseq levels were averaged for each gene and quantile normalized to a standard normal distribution within each woman to reduce the impact of noise, skewness and outliers.

To evaluate prediction accuracy, MiXcan epithelial cell prediction models were applied to the genotype data for each woman, and the between-woman difference in the predicted GRex for each gene was computed to identify the two sets of 100 genes predicted to have the largest positive or negative differences in mammary epithelial cell expression in each pair of women. The Wilcoxon signed-rank test was used to test whether the observed snRNAseq differences for genes predicted to have the largest GRex differences between women were significantly different from zero, as expected.

### Simulation studies

To evaluate the type I error and power of MiXcan association tests, we performed extensive simulation studies under a broad range of realistic models for the associations of genetic variants with gene expression (SNP-Exp) and gene expression with disease (Exp-Disease). Mimicking real data, in each simulation, we generated a training dataset for building the GRex prediction models, and a GWAS dataset for testing the associations of GRex with disease. Without loss of generality, non-genetic covariates were excluded from simulations to allow direct evaluation of the predicted GRex.

For the training dataset, we simulated 300 bulk tissue samples with observed SNP genotypes and tissue-level gene expression. We assumed each tissue  $i \in (1, \dots, 300)$  was a mixture of two cell types and that the minor cell type (cell type 1) comprised an average of 40% of the tissue, with proportion  $\pi_i \sim \text{Beta}(\alpha = 2, \beta = 3)$ . We further simulated the genotypes of 50 neighboring SNPs  $\mathbf{x}_i = \{x_{1i}, \dots, x_{50i}\}$  using the genome simulator R package *sim1000G*, with its default reference genome region (chromosome 4) and minor allele frequency (MAF) range 0.05–0.50. For the expression levels in the two cell types  $u_i$  and  $v_i$ , we considered a linear additive model such that  $u_i = b_0 + \mathbf{b}_1 \mathbf{x}_i + e_{ui}$  and  $v_i = \mathbf{b}_2 \mathbf{x}_i + e_{vi}$  where  $e_{ui}, e_{vi} \sim \mathcal{N}(0, 1)$ . The parameter  $b_0$  determined the mean expression difference in the two cell types under  $\mathbf{x}_i = 0$ , and  $\mathbf{b}_1$  and  $\mathbf{b}_2$  determined the association patterns between the SNPs  $X$  and gene expression level  $Y$  in the minor and major cell types, respectively. Then, the tissue-level gene expression was a weighted average of the expression levels in the two cell types:  $y_i = \pi_i u_i + (1 - \pi_i) v_i$ .

We considered two SNP-Exp settings: *Homogeneous SNP-Exp Association* ( $\mathbf{b}_1 = \mathbf{b}_2$ ) and *Heterogeneous SNP-Exp Association* ( $\mathbf{b}_1 \neq \mathbf{b}_2$ ). Under the *Homogeneous SNP-Exp* setting, we randomly selected one genetic variant  $p$  to be associated with expression levels in the two cell types and let  $b_{1p} = b_{2p} = 1$  or  $-1$  with equal chance, corresponding to a median heritability of 0.27 and interquartile range (IQR) of 0.10. We varied  $b_0$  from  $-2$  to  $2$  to evaluate the impact of the intercept (mean expression difference in two cell types under  $\mathbf{x}_i = 0$ ) on TWAS. Under the *Heterogeneous SNP-Exp* setting, we randomly selected two SNPs  $p_1 \neq p_2 \in (1, \dots, 50)$  with SNPs  $p_1$  and  $p_2$  associated with expression levels in the minor and major cell types, respectively, corresponding to the same heritability in both cell types (median 0.27; IQR 0.10). Similar to the *Homogeneous SNP-Exp* setting, we first evaluated the impact of the intercept by varying  $b_0$  from  $-2$  to  $2$  while fixing  $b_{1p_1}$  at 1 or  $-1$  and  $b_{2p_2}$  at 1 or  $-1$  with equal chance. Second, we varied the magnitude of  $b_{1p}$  from 0 to 2 (median heritability 0 to 0.59) with equal chance of a positive or negative sign, while fixing  $b_0$  at 1 and  $b_{2p} = \pm 1$  to understand the impact of the SNP-Exp association strength in the minor cell type. Third, we varied the magnitude of  $b_{2p}$  from 0 to 2 (allowing a random sign with equal chance), while fixing  $b_0$  at 1 and  $b_{1p} = \pm 1$  to understand the impact of the SNP-Exp association strength in the major cell type. Finally, to evaluate the impact of the training data sample size, we assessed sample sizes ranging from 100 to 300, while fixing  $b_0$  at 1 and the magnitude of non-zero components of  $\mathbf{b}_1, \mathbf{b}_2$  at 1.

For the GWAS dataset, we simulated a case-control study of 3000 participants with observed genotypes and disease status. We assumed the unobserved cell-type composition and cell-type-level gene expression in this dataset followed the same distributions as in the training dataset. Disease risk was simulated using a logistic model, such that  $\text{logit}P(d_j = 1) = \eta_0 + \eta_1 u_j + \eta_2 v_j$  for  $j \in (1, \dots, 3000)$ . The intercept  $\eta_0$  was set to reflect a 1:1 ratio of cases and controls. We considered five different settings for  $\eta_1, \eta_2$  to capture the dynamic relationship between gene expression levels in the two cell types and disease risk:

1. No Exp-Disease Association:  $\eta_1 = \eta_2 = 0$ , i.e. disease is not associated with the gene expression in either cell type.
2. Homogeneous Exp-Disease Association:  $\eta_1 = \eta_2 = 0.2$ , i.e. disease is associated with the gene expression in both cell types in the same way.
3. Exp-Disease Association in Major Cell:  $\eta_1 = 0$  and  $\eta_2 = 0.2$ , i.e. disease is associated with the gene expression in the major cell type (cell type 2).
4. Exp-Disease Association in Minor Cell:  $\eta_1 = 0.2$  and  $\eta_2 = 0$ , i.e. disease is associated with the gene expression in the minor cell type (cell type 1).
5. Exp-Disease Association in Opposite Directions:  $\eta_1 = -0.2$  and  $\eta_2 = 0.2$ , i.e. disease is associated with the gene expression in the two cell types in opposite directions.

We compared the prediction accuracy, type I error, and power of MiXcan with PrediXcan, which ignores cell-type heterogeneity in 500 Monte Carlo simulations.

**MiXcan performance under misspecified models.** MiXcan decomposes tissues into two cell-type components. To evaluate the robustness of MiXcan to misspecification of the cell-type proportion, where a noisy estimate of  $\pi$  is used instead of the true  $\pi$ , we simulated: (a)  $\hat{\pi} = 0.8\pi$  to shift the mean from 0.4 to 0.32 and scale from (0–1) to (0–0.8); (b)  $\hat{\pi} = 0.7\pi + 0.2$  to shift the mean from 0.4 to 0.48 and scale from (0–1) to (0.2–1); (c)  $\hat{\pi}_i \sim \text{Beta}(50\pi_i, 50(1 - \pi_i))$  to reduce the correlation with the true value  $\text{cor}(\hat{\pi}_i, \pi_i)$  to 0.9; and (d)  $\hat{\pi}_i \sim \text{Beta}(5.5\pi_i, 5.5(1 - \pi_i))$  to further reduce  $\text{cor}(\hat{\pi}_i, \pi_i)$  to 0.6. We compared the performance of MiXcan using the misspecified  $\hat{\pi}$  with MiXcan using the true  $\pi$  and PrediXcan.

We also evaluated the robustness of MiXcan to the presence of a latent third cell type. We simulated a tissue with three cell types that have different SNP-Exp or Exp-Disease associations, and evaluated the performance of MiXcan by decomposing the tissue into cell type 1 vs. a mixture of cell types 2 and 3. Specifically, we simulated 300 bulk tissue samples comprised of three cell types with proportions  $\pi_{1i} = 40\%$ ,  $\pi_{2i} = 50\%$  and  $\pi_{3i} = 10\%$ . As in the simulations above, gene expression levels in three cell types were linearly dependent on 50 neighboring SNPs as determined by  $\mathbf{b}_1$ ,  $\mathbf{b}_2$  and  $\mathbf{b}_3$ , and logit transformed disease risk was linearly dependent on the expression levels in the three cell types as determined by  $\eta_1$ ,  $\eta_2$  and  $\eta_3$ . To evaluate the impact of a latent third cell type contributing to the SNP-Exp association, we compared the performance of MiXcan and PrediXcan for  $\mathbf{b}_2 = \mathbf{b}_3$  vs.  $\mathbf{b}_2 \neq \mathbf{b}_3$  assuming  $\eta_2 = \eta_3$  under the Heterogeneous SNP-Exp Association ( $\mathbf{b}_1 \neq \mathbf{b}_2$ ) setting. In detail, we randomly selected three different SNPs  $p_1, p_2, p_3 \in (1, \dots, 50)$  to be associated with expression levels in the three cell types, respectively, and fixed the magnitude of these non-zero  $\mathbf{b}$  ( $b_{1p_1}, b_{2p_2}, b_{3p_3}$ ) at 1, with equal chance of a positive or negative sign. We evaluated type I error and power in simulated GWAS datasets ( $N = 3000$ ) under five Exp-Disease patterns as described for  $\eta_1, \eta_2$  above. As we observed that type I error was well controlled under various SNP-Exp associations in a latent third cell type, we next evaluated the impact of a latent third cell type contributing to the Exp-Disease association on the study power of MiXcan. In this simulation, we fixed  $\mathbf{b}_3 = \mathbf{b}_2$  but simulated  $\eta_3 \neq \eta_2$  with  $\eta_3$  values ranging from -0.2 to 0.2.

### Apply MiXcan to perform cell-type-aware TWAS of breast cancer Cell-type-aware TWAS of breast cancer.

We performed cell-type-aware TWAS of breast cancer risk using GWAS data from the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) study. Genotype data for 60,014 women (32,438 cases and 27,576 controls) assayed using the Oncoarray<sup>30</sup>, which includes >500,000 variants and provides excellent coverage of most common variants, were downloaded from dbGaP (phs001265.v1.p1 <[https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs001265.v1.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001265.v1.p1)>). After imputation (as described above for TCGA data), 95% (53,528 out of 56,531) and 95% (51,049 out of 53,876) SNPs used in MiXcan and PrediXcan prediction models, respectively, were available for analysis.

Principle component analysis (PCA) was performed using 20,629 SNPs, after excluding SNPs with a missing rate above 0.01% and selecting SNPs in approximate linkage equilibrium using PLINK v1.90 (indep-pairwise option with window size=50kb, step size=5,  $r^2$  threshold=0.05)<sup>68</sup>. EIGENSOFT v6.1.4 was used to compute PCs with the fast mode option enabled, which implements the FastPCA approximation<sup>69</sup>. The first PC separated individuals of African (e.g. from Nigeria, Uganda and Cameroon) vs. European (e.g. from Australia) ancestry. In total, 58,648 women (31,716 cases and 26,932 controls) of European ancestry determined by PCs were included in TWAS analyses.

MiXcan, PrediXcan and PredictDB elastic-net mammary tissue models were applied to the individual-level genotype data to perform

cell-type-aware or tissue-level TWAS, as described above. All three models were adjusted for the same covariates, including age, country of origin and the top 10 PCs<sup>15</sup>.

**Evaluation of TWAS findings.** We evaluated significant TWAS genes identified by MiXcan and PrediXcan in a substantially larger study of 228,951 European ancestry women (122,977 cases and 105,974 controls) from the combined DRIVE and Breast Cancer Association Consortium (BCAC) GWAS meta-analysis of breast cancer<sup>11</sup>. The summary statistics for the “Combined Oncoarray, iCOGS GWAS meta analysis” were downloaded from <https://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/oncoarray-and-combined-summary-result/gwas-summary-results-breast-cancer-risk-2017/>. Associations of predicted tissue-level GRx with breast cancer risk were evaluated using S-PrediXcan<sup>33</sup> with PredictDB<sup>28</sup> elastic-net models derived from GTEx v8 mammary tissue data for 337 men and women of European ancestry. We also determined whether TWAS genes identified by MiXcan and PrediXcan were located within 500 kb of 214 previously reported genome-wide significant breast cancer susceptibility loci<sup>11–13</sup>.

### MiXcan software

We developed a computationally efficient R package *MiXcan* to facilitate estimation of cell-type-level GRx prediction models in the two cell components of bulk tissue data, and to perform cell-type-aware TWAS. The *MiXcan* R package, and pre-trained models for the epithelial and stromal (non-epithelial) cell components of mammary tissue derived from 125 European ancestry women in GTEx v8 are freely available at <https://github.com/songxiaoyu/MiXcan><sup>54</sup>.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

### Data availability

The data used in this study are publicly available from the following sources: GTEx v8 (dbGaP accession number phs000424.v8.p2 <[https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000424.v8.p2](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000424.v8.p2)>); GTEx v9 (<https://gtexportal.org/home/datasets>); TCGA (dbGaP accession number phs000178.v8.p7 <[https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000178.v8.p7](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v8.p7)>); Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) (dbGaP accession number phs001265.v1.p1 <[https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs001265.v1.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001265.v1.p1)>); PredictDB (<http://predictdb.org>); and the Breast Cancer Association Consortium (BCAC) (<https://bcac.ccge.medschl.cam.ac.uk/bcacdata/>). Source data are provided with this paper.

### Code availability

The *MiXcan* R package is publicly available at <https://github.com/songxiaoyu/MiXcan><sup>54</sup>.

### References

- Gamazon, E. R. et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat. Genet.* **47**, 1091 (2015).
- Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).
- Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
- National Cancer Institute. SEER Cancer Statistics Factsheets. Breast Cancer. <http://seer.cancer.gov/statfacts/html/breast.html>.
- Mucci, L. A. et al. Familial risk and heritability of cancer among twins in Nordic countries. *J. Am. Med. Assoc.* **315**, 68–76 (2016).
- Boyd, N. F. et al. Mammographic breast density as an intermediate phenotype for breast cancer. *Lancet Oncology* **6**, 798–808 (2005).

7. Sickles, E. A. et al. ACR BI-RADS Mammography. In *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System* (American College of Radiology, Reston, VA, 2013).
8. Pettersson, A. et al. Mammographic density phenotypes and risk of breast cancer: a meta-analysis. *J. Natl. Cancer Inst.* **106**, dju078 (2014).
9. Arendt, L. M., Rudnick, J. A., Keller, P. J. & Kuperwasser, C. Stroma in breast development and disease. *Seminars Cell Dev. Biol.* **21**, 11–18 (2010).
10. Sieh, W. et al. Identification of 31 loci for mammographic density phenotypes and their associations with breast cancer risk. *Nat. Commun.* **11**, 5116 (2020).
11. Michailidou, K. et al. Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94 (2017).
12. Milne, R. L. et al. Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet.* **49**, 1767–1778 (2017).
13. Zhang, H. et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* **52**, 572–581 (2020).
14. Hoffman, J. D. et al. Cis-eQTL-based trans-ethnic meta-analysis reveals novel genes associated with breast cancer risk. *PLoS Genet.* **13**, e1006690 (2017).
15. Wu, L. 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).
16. Bhattacharya, A. et al. A framework for transcriptome-wide association studies in breast cancer in diverse study populations. *Genome Biology* **21**, 1–18 (2020).
17. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
18. Aran, D., Hu, Z. & Butte, A. J. xCell: digitally portraying the tissue cellular heterogeneity landscape. *Genome Biology* **18**, 1–14 (2017).
19. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
20. Wang, J., Roeder, K. & Devlin, B. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. *Genome Res.* **31**, 1807–1818 (2021).
21. Newman, A. M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
22. Hunt, G. J., Freytag, S., Bahlo, M. & Gagnon-Bartsch, J. A. dtangle: accurate and robust cell type deconvolution. *Bioinformatics* **35**, 2093–2099 (2019).
23. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* **10**, 380 (2019).
24. Sturm, G. et al. Comprehensive evaluation of transcriptome-based cell-type quantification methods for immuno-oncology. *Bioinformatics* **35**, i436–i445 (2019).
25. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **11**, 1–14 (2020).
26. Sutton, G. J. et al. Comprehensive evaluation of deconvolution methods for human brain gene expression. *Nat. Commun.* **13**, 1–18 (2022).
27. Liu, Y. & Xie, J. Cauchy combination test: a powerful test with analytic *p* value calculation under arbitrary dependency structures. *J. Am. Stat. Assoc.* **115**, 393–402 (2020).
28. PredictDB Data Repository—GTEx V8 Model Release (2019). <https://predictdb.org/>.
29. Eraslan, G. et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eabl4290 (2022).
30. Amos, C. I. et al. The OncoArray consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiol. Biomarkers Prevent.* **26**, 126–135 (2017).
31. van Iterson, M., van Zwet, E. W. & Heijmans, B. T. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology* **18**, 1–13 (2017).
32. Feng, H. et al. Transcriptome-wide association study of breast cancer risk by estrogen-receptor status. *Genet. Epidemiol.* **44**, 442–468 (2020).
33. Barbeira, A. N. et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature Communications* **9**, 1–20 (2018).
34. Gucalp, A. et al. Male breast cancer: a disease distinct from female breast cancer. *Breast Cancer Res. Treat.* **173**, 37–48 (2019).
35. Li, B. et al. Tissue specificity-aware TWAS (TSA-TWAS) framework identifies novel associations with metabolic, immunologic, and virologic traits in hiv-positive adults. *PLoS Genet.* **17**, e1009464 (2021).
36. Feng, H. et al. Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. *PLoS Genet.* **17**, e1008973 (2021).
37. Thompson, M. et al. Multi-context genetic modeling of transcriptional regulation resolves novel disease loci. *Nat. Commun.* **13**, 5704 (2022).
38. Donovan, M. K., D’Antonio-Chronowska, A., D’Antonio, M. & Frazer, K. A. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nat. Commun.* **11**, 1–14 (2020).
39. Chen, J. et al. Fast and robust adjustment of cell mixtures in epigenome-wide association studies with SmartSVA. *BMC Genomics* **18**, 413 (2017).
40. Luo, X., Yang, C. & Wei, Y. Detection of cell-type-specific risk-CpG sites in epigenome-wide association studies. *Nat. Commun.* **10**, 3113 (2019).
41. Rahmani, E. et al. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nat. Commun.* **10**, 3417 (2019).
42. Liu, W. et al. A statistical framework to identify cell types whose genetically regulated proportions are associated with complex diseases. medRxiv (2021).
43. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
44. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).
45. Tsoucas, D. et al. Accurate estimation of cell-type composition from gene expression data. *Nat. Commun.* **10**, 2975 (2019).
46. Nguyen, Q. H. et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.* **9**, 2028 (2018).
47. Han, X. et al. Construction of a human cell landscape at single-cell level. *Nature* **581**, 303–309 (2020).
48. Zou, H. & Zhang, H. H. On the adaptive elastic-net with a diverging number of parameters. *Ann. Stat.* **37**, 1733 (2009).
49. Holland, D. G. et al. ZNF703 is a common Luminal B breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium. *EMBO Mol. Med.* **3**, 167–180 (2011).
50. Sircoulomb, F. et al. ZNF703 gene amplification at 8p12 specifies luminal B breast cancer. *EMBO Mol. Med.* **3**, 153–166 (2011).
51. Slorach, E. M., Chou, J. & Werb, Z. Zeppo1 is a novel metastasis promoter that represses E-cadherin expression and regulates p120-

- catenin isoform expression and localization. *Genes Dev.* **25**, 471–484 (2011).
52. Xia, W. et al. MicroRNA-32 promotes cell proliferation, migration and suppresses apoptosis in breast cancer cells by targeting FBXW7. *Cancer Cell Int.* **17**, 14 (2017).
  53. Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information (2021). <https://www.ncbi.nlm.nih.gov/gene/>.
  54. Song, X. et al. MiXcan: a framework for cell-type-aware transcriptome-wide association studies with an application to breast cancer. *Zenodo* (2022). <https://doi.org/10.5281/zenodo.7350463>.
  55. Nagpal, S. et al. TIGAR: an improved bayesian tool for transcriptomic data imputation enhances gene mapping of complex traits. *Am. J. Hum. Genet.* **105**, 258–266 (2019).
  56. Mancuso, N. et al. Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* **51**, 675–682 (2019).
  57. Friedman, J., Hastie, T. & Tibshirani, R. *The Elements of Statistical Learning*, vol. 1 (Springer series in statistics New York, 2001).
  58. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1 (2010).
  59. Petralia, F. et al. A new method for constructing tumor specific gene co-expression networks based on samples with tumor purity heterogeneity. *Bioinformatics* **34**, i528–i536 (2018).
  60. Meinshausen, N. & Bühlmann, P. Stability selection. *J. R. Stat. Soc.: Series B (Stat. Methodol.)* **72**, 417–473 (2010).
  61. Urbut, S. M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat. Genet.* **51**, 187–195 (2019).
  62. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (peer) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
  63. Network, T. C. G. A. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70 (2012).
  64. Grossman, R. L. et al. Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
  65. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology* **11**, 1–9 (2010).
  66. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
  67. Das, S. et al. Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).
  68. Purcell, S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
  69. Galinsky, K. J. et al. Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.* **98**, 456–472 (2016).

## Acknowledgements

This study was supported by the National Institutes of Health R01CA237541 (W.S., L.A.H., P.W.), R01CA244948 (R.J.K.), R01CA264987 (W.S., L.A.H.), R03AG075567 (X.S.), U24CA210993 (P.W.), U24CA271114 (P.W.), and P30CA196521 (X.S.). The authors acknowledge the compu-

tational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. The BCAC breast cancer genome-wide association analyses were supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research, the ‘Ministère de l’Économie, de la Science et de l’Innovation du Québec’ through Genome Québec and grant PSR-SIIRI-701, the National Institutes of Health (U19CA148065, X01HG007492), Cancer Research UK (C1287/A10118, C1287/A16563, C1287/A10710) and the European Union (HEALTH-F2-2009-223175 and H2020 633784 and 634935); all studies and funders are listed<sup>11</sup>. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The results published here are in whole or part based upon data generated by the TCGA Research Network.

## Author contributions

Study conception and design (X.S., J.H.R., R.J.K., L.A.H., P.W., W.S.); data curation and processing (J.H.R., R.J.K., W.S.); software development (X.S., J.J.); statistical analysis (X.S., J.J., J.H.R., R.J.K., P.W., W.S.); interpretation of results and critical review of the manuscript (X.S., J.J., J.H.R., S.E.A., L.C.S., A.S., N.A., E.J., A.S.W., R.J.K., L.A.H., P.W., W.S.).

## Competing interests

E.J. is an employee at Regeneron. The remaining authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-023-35888-4>.

**Correspondence** and requests for materials should be addressed to Xiaoyu Song, Pei Wang or Weiva Sieh.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023