A FRAMEWORK FOR TRANSCRIPTOME-WIDE ASSOCIATION STUDIES IN BREAST CANCER IN DIVERSE STUDY POPULATIONS

Arjun Bhattacharya (abhattac@live.unc.edu)[1]
Montserrat García-Closas (montserrat.garcia-closas@nih.gov)[2,3]
Andrew F. Olshan (andy_olshan@unc.edu)[4,5]
Charles M. Perou (cperou@med.unc.edu)[5,6,7]
Melissa A. Troester (troester@email.unc.edu)[4,7]
Michael I. Love (milove@email.unc.edu)[1,6]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill
[2]Division of Cancer Epidemiology and Genetics, National Cancer Institute
[3]Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK
[4]Department of Epidemiology, University of North Carolina at Chapel Hill
[5]Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill
[6]Department of Genetics, University of North Carolina at Chapel Hill
[7]Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill

Corresponding author: Michael I. Love (milove@email.unc.edu)

Abstract

Background: The relationship between germline genetic variation and breast cancer survival is largely unknown, especially in understudied minority populations who often have poorer survival. Genome-wide association studies (GWAS) have interrogated breast cancer survival but often are underpowered due to subtype heterogeneity and many clinical covariates and detect loci in non-coding regions that are difficult to interpret. Transcriptome-wide association studies (TWAS) show increased power in detecting functionally-relevant loci by leveraging expression quantitative trait loci (eQTLs) from external reference panels in relevant tissues. However, ancestry- or race-specific reference panels may be needed to draw correct inference in ancestrally-diverse cohorts. Such panels for breast cancer are lacking.

Results: We provide a framework for TWAS for breast cancer in diverse populations, using data from the Carolina Breast Cancer Study (CBCS), a North Carolina population-based cohort that oversampled black women. We perform eQTL analysis for 406 breast cancer-related genes to train race-stratified predictive models of tumor expression from germline genotypes. Using these models, we impute expression in independent data from CBCS and TCGA, accounting for sampling variability in assessing performance. These models are not applicable across race, and their predictive performance varies across tumor subtype. Within CBCS ($N = 3{,}828$), at a false discovery-adjusted significance of 0.10 and stratifying for race, we identify associations in black women near *AURKA, CAPN13, PIK3CA, and SERPINB5* via TWAS that are underpowered in GWAS.

Conclusions: We show that carefully implemented and thoroughly validated TWAS is an efficient approach for understanding the genetics underpinning breast cancer outcomes in diverse populations.

Keywords: transcriptome-wide analysis (TWAS); breast cancer; expression quantitative trait loci (eQTL); survival; polygenic traits

Background

Breast cancer remains the most common cancer among women in the world [1]. Breast cancer tends to be more aggressive in young women and African American women, though underlying germline determinants of poor outcomes are not well-studied. Cohorts that represent understudied minority populations, like the Carolina Breast Cancer Study (CBCS), have identified differences in healthcare access, socioeconomics, and environmental exposures associated with disparities in outcome [2–4], but more targeted genomic studies are necessary to interrogate these disparities from a biologic and genetic perspective.

Few genome-wide association studies (GWAS) have studied the relationship between germline variation and survival outcomes in breast cancer, with most focusing instead on genetic predictors of risk [5,6]. Recently, GWAS have shown evidence of association between candidate common germline variants and breast cancer survival, but these studies are often underpowered [7,8]. Furthermore, the most significant germline variants identified by GWAS, in either risk or survival, are often located in non-coding regions of the genome, requiring *in vitro* follow-up experiments and co-localization analyses to interpret functionally [9]. It is important to seek strategies for overcoming these challenges in GWAS, especially because several studies in complex traits and breast cancer risk have shown that regulatory variants not significant in GWAS account for a large proportion of trait heritability [10–12].

Novel methodologic approaches that integrate multiple data types offer advantages in interpretability and statistical efficiency. Escala-García et al. has suggested that aggregating variants by integrating gene expression or other omics may better explain underlying biological mechanisms while increasing the power of association studies beyond GWAS [7]. To alleviate problems with statistical power and interpretability, a recent trend in large-scale association studies is the transcriptome-wide association study (TWAS). TWAS aggregates genomic information into functionally-relevant units that map to genes and their expression. This gene-based approach combines the effects of many regulatory variants into a single testing unit that increases study power and provides more interpretable trait-associated genomic loci [13–15]. Hoffman et al. and Wu et al. have recently conducted TWAS for breast cancer risk and have reported several significant associations for genes with breast cancer susceptibility, showing increased power over GWAS [15,16]. However, these studies either draw from ancestrally-homogeneous reference panels like subsets of women of European ancestry from the Genotype-Tissue Expression (GTEx) project [16] or study populations of European descent from the Breast Cancer Association Consortium (BCAC) [15]. It is not known whether these models can be informative in African American women and other groups. Recent findings have suggested that stratification by race or ancestry may be necessary to construct proper tests of association across race or ancestry [17,18]. However, many cohorts, especially large-scale genetic cohorts, may not have a sufficient sample size in minority populations to power these tests.

Here, we provide a framework for TWAS for complex disease outcomes in diverse study populations using transcriptomic reference data from the Carolina Breast Cancer Study (CBCS), a multi-phase cohort that includes an over-representation of African American women [19]. We train race-stratified predictive models of tumor expression from germline variation and carefully validate their performance, accounting for sampling variability and disease heterogeneity, two aspects that previous TWAS in breast cancer have not considered. This framework shows promise for scaling up into larger GWAS cohorts for further detection of risk- or outcome-associated loci.

Results

Race specific germline eQTL analysis

To assess the association between germline genomic variation and tumor expression of 406 autosomal genes, targeted by the CBCS because of their association with breast cancer progression, we first conducted a full cis-trans expression quantitative trait loci (eQTL) analysis, stratifying on race and controlling for key biological covariates and population stratification (see **Methods**). We discuss the relationship between self-reported race and ancestry in CBCS in **Supplemental Results**.

We evaluated associations between the tumor expression levels of 406 autosomal genes and 5,989,134 germline SNPs. SNPs and genes found in association in an eQTL will be called eSNPs and eGenes, respectively. At a Benjamini-Bogomolov [20] FDR-corrected P-value ($BBFDR < 0.05$), we identified 266 cis-eQTLs and 77 trans-eQTLs in the AA sample across 32 eGenes, and 691 cis-eQTLs and 15 trans-eQTLs in the WW sample

97 across 24 eGenes, shown in **Supplemental Figure 2**. Of these eGenes, 4 are in common
98 across race: *PSPHL*, *GSTT2*, *EFHD1*, and *SLC16A3*. Expressions of *PSPHL* and *GSTT2*
99 have been previously reported to be governed by respective cis-deletions and serve as
100 distinguishing biomarkers for race [21–24]. The majority of significant eQTLs in both the
101 AA and WW samples were found in cis-association with respective eGenes. However,
102 we saw a higher proportion of significant trans-eQTLs in the AA sample (**Supplemental**
103 **Figure 2**). The locations and strengths of top eQTLs for all 406 autosomal genes are
104 shown in **Figure 1A**. All significant eQTLs are plotted in **Supplemental Figure 2**.
105
106 We further adjusted our eQTL models for a computationally-derived estimate of tumor
107 purity, which showed little effect on the strength and location of top eQTLs by eGene
108 (**Supplemental Results**). We do not consider tumor purity in any downstream analyses
109 and train predictive models on bulk tumor expression.
110
111 We lastly sought to evaluate the source of the significant eQTLs we detect in CBCS.
112 Similar to previous pan-cancer germline eQTL analyses [25], we cross-referenced
113 eGenes found in CBCS with eGenes detected in relevant healthy tissues from Genotype-
114 Tissue Expression (GTEx) Project. We attributed all but 7 of the cis-eGenes from CBCS
115 across both AA and WW women found in GTEx to one of these three tissue types (**Figure**
116 **1B**), with the effect sizes of the top eQTLs for these eGenes correlating very well between
117 CBCS and GTEx (see **Supplemental Figure 5**).
118
119 *Race-specific predictive models of tumor expression*

Using the significant germline eQTLs of tumor expression as motivation, we used tumor expression and genotyping data from 628 AA women and 571 WW women from CBCS to build predictive models of tumor RNA expression levels for each gene's breast tumor expression (see **Methods**). Mean cis-heritability (cis-h^2) of the 417 genes is 0.016 ($SE = 0.019$) in AA women and 0.015 ($SE = 0.019$), as estimated by GREML-LDMS analysis [26]. For downstream analysis, we only consider genes with cis-h^2 significantly greater than 0 at a nominal P-value less than 0.10 from the relevant likelihood ratio test. Considering only these genes, the mean cis-h^2 of genes is 0.049 ($SE = 0.016$) in AA models and 0.052 ($SE = 0.016$) in WW models. Of the predictive models built for these genes, 125 showed a five-fold cross-validation prediction performance (CV R^2) of at least 0.01 (10% Pearson correlation between predicted and observed expression with $P < 0.05$) in one of the two predictive models. **Figure 2A** shows the CV R^2 of these 153 genes across race. The median CV R^2 for the 153 genes was 0.011 in both AA and WW women. Cis-h^2 and CV R^2 are compared in **Supplemental Figure 6.**

Based on model performance in CBCS, we selected 46 genes in AA women and 57 genes in WW women for association analyses between predicted tumor gene expression and breast cancer survival, using data from all patients from CBCS with genotype data. These genes were selected because they showed an CV $R^2 > 0.01$ (10% correlation between observed and predicted expression in the CBCS training set) and cis-$h^2 \geq 0$ with nominal $P < 0.10$ in a given race strata.

Evaluation of predictive models in independent data

Predictive performance was strong across race and biological and molecular subtype in two external samples: The Cancer Genome Atlas (TCGA) and a held-out CBCS sample set. We defined the imputed expression of a given gene in an external cohort as the GReX, or the germline-genetically regulated tumor expression, of that gene.

The first sample is derived from TCGA breast tumor tissues with 179 AA and 735 WW women. We compared predictive performance by calculating an external validation R^2 (EV R^2) with squared Spearman correlations. Of the 151 genes modeled in CBCS training data with significant cis-h^2, 149 genes were measured via RNA-seq in TCGA. A comparison of predictive performance in TCGA for these 149 genes is shown in **Figure 2B**, showing adequate performance in AA women (33 genes with EV $R^2 > 0.01$) and poor performance in WW women (7 genes with EV $R^2 > 0.01$). The top predicted gene in cross-validation from CBCS for both races, *PSPHL*, was not present in the TCGA normalized expression data and could not be validated. Another top cross-validated gene, *GSTT2*, was present in TCGA expression data and was validated as the top genetically predicted gene in TCGA by EV R^2.

We also imputed expression into entirely held-out samples from CBCS data (1,121 AA and 1,070 WW women) that have gene expression for a subset of the genes (166 of 417 genes) in the CBCS training set. These samples were largely derived from Phases I and II of CBCS (see **Methods**). A comparison of imputation performance in CBCS for 51 genes is shown in **Figure 2C**, showing adequate performance in both AA and WW women (18 and 15 genes with EV $R^2 > 0.01$ in AA and WW women).

166

167 *Predictive models are not applicable across race*

168 We find that the predictive accuracy of most genes was lower when expression was
169 imputed in AA women using models trained in the WW sample. We employed the WW
170 predictive models to impute expression into AA samples from TCGA and held-out CBCS
171 data. We compare the performances of the WW model and AA model in the AA sample
172 in **Figure 2D** (TCGA) and **2E** (CBCS). In held-out CBCS samples, with the WW model,
173 we could only predict *PSPHL* and *GSTT2* at $R^2 > 0.01$ in the AA sample, as the
174 expression of these genes is modulated mostly by strongly associated cis-eSNPs. In
175 TCGA, our WW models performed adequately in AA women, though the WW models
176 predicted fewer genes at $R^2 > 0.01$ than the AA models.

177

178 *Evaluation of predictive performance across subtype*

179 While predictive accuracy of expression models was stable across datasets, there was
180 greater heterogeneity across biological and molecular subtype. In part, this is due to small
181 sample sizes within race and subtype-specific strata. Upon first inspection, we see vast
182 differences in the performance of our models across subtype (**Supplemental Figure 7**),
183 with a large majority of genes performing at EV $R^2 > 0.01$ in rarer subtypes, like HER2-
184 enriched breast cancers. However, we recognized sample sizes in the TCGA validation
185 set were relatively small, especially when considering AA women and women of certain
186 subtype, e.g. as low as 16 AA women with HER2-enriched breast cancer. As overall
187 correlation between observed and imputed expressions are near 0, we sought to account
188 for sampling variability when imputing into groups of women with such small sample sizes.

We employed a permutation scheme: permuting observed expression values among samples 10,000 times to generate a null distribution for EV R^2. We then tested for the null hypothesis $R^2 = 0$, controlling for false discovery, according to this null distribution. Supplemental Figure 9 displays q-values in Manhattan form [27], showing that the proportion of genes with EV R^2 significantly different from 0 is similar across subtypes. We inverted this permutation test [28] to construct a confidence interval for EV R^2. We find that the EV R^2 of several genes are highly variable across subtypes, even when accounting for differences in sample size and therefore sampling variation. Key examples of such genes with variable EV R^2 across subtypes are shown in **Figure 3**.

Predicted expression associated with breast cancer-specific survival

To assess association between imputed gene expression and breast cancer-specific survival, we constructed race-stratified cause-specific proportional hazard models for 3,828 samples from CBCS (1,865 AA and 1,963 WW), where we model time to mortality due to breast cancer. Of the genes evaluated, we detected 4 whose GReX were associated with breast-cancer specific survival at FDR-adjusted $P < 0.10$ in AA women, shown in **Table 1** and **Figure 4**. We did not identify any genes with GReX associated with survival in WW women.

An association between increased GReX and increased risk of breast cancer-specific mortality was identified for *CAPN13* (*2p23.1*). We also found protective associations between higher GReX of *AURKA (20q13.2)*, *PIK3CA (3q26.32)*, *SERPINB5* (*18q21.33*)

and lower risk of breast cancer-mortality (**Figure 4C**). Of these 4 loci, associations with survival have been reported with SNPs in the same chromosomal region as *AURKA*, *PIK3CA,* and *SERPINB5* [8,29–33], though none of these reported SNPs were utilized in constructing the GReX of this gene. Furthermore, the GReX of these four genes were not significantly correlated ($P > 0.05$ for all pairwise Spearman correlation tests), and the sets of SNPs used in constructing the GReX of these four genes had no pairwise intersections, providing evidence that their independent association with breast cancer-specific survival was not a pleiotropic effect from shared or correlated SNPs.

To determine whether the associations between predicted gene expression and breast cancer-specific survival were independent of GWAS-identified association signals, we performed conditional analyses adjusted for the most significant GWAS-identified survival-associated SNPs closest to the TWAS-identified gene by adjusting the cause-specific proportional hazards model for the genotype from this SNP. We found that the association for *PIK3CA* had a small change in effect size after adjustment for its adjacent survival-associated SNP, and its SNP-adjusted association was insignificant, while the other genes' associations remained significant after adjustment (**Table 2**). This conditional analysis suggests that the GReX of *AURKA*, *CAPN13*, and *SERPINB5* may be associated with breast cancer-specific survival independent of the GWAS-identified variant. No previously reported survival-associated SNPs were found significant at the genome-wide significance level in our dataset, and none of the closest survival-associated SNPs used in conditional adjustment were significant (Figure 4A). This

supports our observation that correctly analyzed TWAS using relevant tissue gene expression may increase power for association testing.

As we deal with case-only data, we wished to inspect any collider bias that arises from unmeasured confounders that are associated with both breast cancer incidence and survival (see **Supplemental Figure 13**) [34]. Since a case-control dataset was not readily available to us to test associations between the GReX of genes with breast cancer risk, we construct the weighted burden test, as in FUSION [14], for the GReX of *AURKA*, *CAPN13, PIK3CA*, and *SERPINB5* in the GWAS summary statistics for breast cancer risk in AA women available from BCAC using the iCOGs dataset and additional GWAS [35–37]. We find that none of the GReX of these genes are significantly associated with breast cancer incidence ($Z > 1.96,\ P < 0.05$), suggesting minimal presence of collider bias in our estimates of association with survival for the GReX of these four genes.

Lastly, we examined the association of the GReX of these four genes with breast cancer-specific survival in AA women, stratified by estrogen receptor (ER) subtype. We find that overall associations with survival are often driven by significant associations in a single subtype, though there is evidence of significant hazardous association in both ER subtypes for *CAPN13* (**Supplementary Figure 10**). We also did not detect a survival association with the total expression of these 4 genes, as estimated from breast cancer-specific Cox models (**Supplementary Figure 11**).

Discussion

In this paper, we studied the relationship between breast cancer-specific survival and germline genetics using a TWAS framework, wherein we aggregate the germline genome into testing units that map to the transcriptome to greatly mitigate the multiple testing burden found in GWAS. This study is the first systematic TWAS for breast cancer-specific survival, motivated by a full cis-trans eQTL analysis with one of the largest sample sizes for breast tumor gene expression in African American women. Our analyses underscore the importance of accounting for sampling variability when validating predictive models for TWAS and incorporating race or ancestry in these models, an aspect which confounds naïve comparisons involving imputed GReX across validation sub-groups of different sample size.

Using a training set from CBCS, we leveraged race-stratified germline eQTLs of tumor expression to train race-stratified models of tumor expression from germline variation. Our eQTL analysis reveals a strong cis-signal between germline variants and tumor expression of several genes, that is both differential across race and not exclusively attributable to healthy breast tissue. Our models showed strong cross-validation predictive performance in genes with significant cis-heritability. We also show strong predictive performance in a held-out test set from CBCS and adequate performance of our WW models in TCGA-BRCA data. We suspect that this discrepancy in validation performance between CBCS and TCGA may be attributed to a poor intersection of SNPs in the genotyping data from TCGA and CBCS (only approximately 85% of SNPs from CBCS represented in TCGA imputed genotype data). There could also be a lack of cis-heritability of the tumor expression of a majority of genes assayed in TCGA. For example,

Gusev et al. has trained models for gene expression in breast tumors in TCGA; only 8 of the 417 genes in the CBCS Nanostring panel showed significant cis-heritability in their models [14], which we downloaded from the Gusev Lab's TWAS/FUSION repository. We believe that predictive performance in TCGA data consistent with CBCS data is a high bar for validation due to both genotyping and RNA expression platform differences between CBCS (Oncoarray and Nanostring) and TCGA (Affymetrix 6.0 and RNAseq). Reproducible performance in both AA and WW women in our independent test set from CBCS data suggests that our models are quite robust. Follow-up studies, in which models of tumor expression are trained in TCGA RNA-seq data and validated in CBCS Nanostring data, could elucidate any discrepancies in predictive performance across platform.

An important implication of our work is the race-specificity of TWAS methods. In our validation scheme, we assessed the applicability of imputing expression in AA samples using the WW predictive models, as publicly available tumor expression data is often measured in predominantly WW cohorts. We find that WW models generally have poor performance in AA women. Epidemiological studies have stressed accounting for differences in race by stratification or adjustment for admixture estimates when constructing polygenic scores [38]. Our key finding of poor predictive performance across race suggests that this epidemiological note of caution extends to creating predictive models for RNA expression. Previous TWAS studies of breast cancer risk have either used models trained in a sample of predominantly European ancestries [16] or imputed into large cohorts of strictly patients of European descent [15]. Hoffman et al. does

exclude SNPs that were monomorphic in any of the 14 different ancestral populations they analyze [16], though this may not capture all effects of ancestry on genetic regulation of expression, including the possibility for interactions. We contend that accounting for ancestry or stratifying by race may be necessary to draw correct inference in large, ancestrally-heterogeneous cohorts.

Our data also suggests that predictive performance may vary by molecular subtype. Previous groups have shown the predictive utility of catering polygenic risk scores to breast cancer subtype [39,40], a phenomenon we investigated in our predictive models of tumor expression. As the estimates of sample correlations between observed and predicted expression were small and the sample sizes per subtype were small, we recognized the need to employ a permutation method to assess the precision of our prediction R^2. We found that a significant portion of the variability of predictive performance across subtype was explained by sampling variability. Nevertheless, even after accounting for sampling variability, we noticed that several genes have varied predictive performance across subtype and race. This finding suggests that TWAS predictive models of expression may need to account of biological heterogeneity. We also reinforce the importance of sampling variability in the validation of predictive models in external cohorts prior to generalized imputation and association testing. For example, Wu et al. trained their models in a relatively small set of 67 women from GTEx and validated their 12,824 models in a validation set of 86 women from TCGA without accounting for sampling variability of predictive performance [15]. A recent multi-tissue TWAS in ovarian cancer from Gusev et al. considered a more thorough validation of predictive models by

leveraging multiple independent cohorts to assess replication rates for their models [41]. We recommend such an approach if multiple independent cohorts are accessible. But, in TWAS evaluation in a single tissue, studies should place a strong emphasis on validation, accounting for sampling variability of prediction R^2, ideally prior to imputation in larger cohorts.

While many of the most significant findings here are methodological in nature, we also have data to suggest that four genomic loci may merit further investigation relative to breast cancer survival. We identified 4 genomic loci associated with breast cancer survival at an FDR-adjusted significance level of 0.10 in AA women. After adjustment for genetics at the most significantly survival-associated SNP close to the gene in question, survival associations at 3 of these 4 locations remained marginally significant. We did not observe any significant association between the total expression of these 4 genes and breast cancer-specific survival. This suggests that the germline-regulated component of the tumor expression of these genes – a small fraction of the total expression variation – may be associated with survival outcomes. Numerous factors, including copy number alterations, epigenetic or post-transcriptional regulation, and exposures and technical artifacts in measurement contributed to the total expression measured in the tumor. Thus, we do not expect that significant GReX association implies total expression association, or vice versa.

While nearly all of the genes on the CBCS Nanostring panel are relevant to breast cancer research, many have not been shown to be associated with survival. Two of these 4

TWAS-identified genes have strong functional evidence in breast cancer survival literature. Mutations in *AURKA* and *PIK3CA* have previously been shown to be significantly associated with breast cancer survival rates [29–31]. Less is known about the involvement of *SERPINB5* and *CAPN13* in breast cancer survival. *SERPINB5* is a tumor-suppressor gene that has been shown to promote development of breast cancers in humans [42]. The calpain family, which contain *CAPN13*, is a group of proteases that is involved in apoptosis and the progression and proliferation of breast cancer cells and has been suggested as therapy targets for various cancers [43–45]. These four loci merit further studies for validation and functional characterization, both in large GWAS cohorts and using *in vitro* studies.

We also observed that 3 of the 4 associations were driven by very strong effect sizes within a single subtype (Supplementary Figure 11). Though we cannot contextualize this result, it highlights an often-overlooked modeling consideration. In a cohort that is both biologically and ancestrally-heterogenous, as in CBCS, investigators should consider modeling choices beyond simple linear adjustments for subtype and race. Given a large enough sample size, it may be prudent in future TWAS to stratify predictive models on both race and biological subtype to increase power to detect outcome-associated loci that are strongly present within only one such strata or have heterogeneous effects across strata. This idea is akin to the logic of Begg et al and Martínez et al in detecting etiological risk factors for ER-positive and -negative tumors [46,47].

Since the CBCS analysis was a case-only study, we were wary of potential collider bias by unmeasured confounders associated with both breast cancer risk and progression [34,48–50]. These colliders may affect the magnitude and direction of effect sizes on association between survival and GReX of genes (Supplemental Figure 14). We find that, using summary statistics for breast cancer risk GWAS from iCOGs [35–37], none of the GReX of these four genes showed significant transcriptome-wide associations with breast cancer risk in this iCOGs data. This suggests that our estimates of association may be free of the collider bias, outlined in Supplemental Figure 14. As Escala-García et al. highlights, germline variation can affect breast cancer prognosis via tumor etiology (risk of developing a tumor of a certain subtype), or via mechanisms that are relevant post-tumorigenesis, such as the cellular response to therapy, or the host-tumor micro-environment, including immune response and stroma-tumor interactions [7]. Ideally, in future TWAS and integrated omic analyses of breast cancer survival, it is prudent to consider joint models of breast cancer risk and survival to account for the many effects of germline genotype and any associations with unmeasurable confounders [49].

One limitation of our study is that data on somatic amplifications and deletions were not yet available for the CBCS cohort we analyzed. Removing the somatic copy number variation signal from tumor expression profiles may improve our estimates of cis-heritability and perhaps the predictive performance of our models, as previous TWAS have shown [41]. Furthermore, not all genes in the CBCS Nanostring panel have a significant heritable component in expression regulation. These genes, like *ESR1*, which have a significant role in breast cancer etiology [51], could not be investigated in our

394 study. Lastly, since CBCS mRNA expression is assayed by the Nanostring nCounter
395 system, we could only analyze 94 aggregated locations on the human transcriptome
396 across race. However, the Nanostring platform allows the CBCS to robustly measure
397 expression from FFPE samples on a targeted panel of breast cancer and race-related
398 genes, allowing us to leverage the large sample size from all three phases of the CBCS.
399 One of the greatest strengths of our study is that the CBCS affords us both a large training
400 and test set of AA and WW women for race-stratified predictive models. Such data is
401 important in drawing inference in more ancestrally-heterogeneous populations.
402 Accordingly, the statistical power of our study is high to detect associations for genes with
403 relatively high cis-heritability. Nonetheless, the specific survival-associated loci merit
404 further investigation in external datasets. Future studies in large GWAS cohorts, such as
405 those within the Breast Cancer Association Consortium, will elucidate how to account for
406 ancestral and biological heterogeneity in detecting survival-associated loci.
407

408 **Conclusion**

409 We have provided a framework of transcriptome-wide association studies (TWAS) for
410 breast cancer outcomes in diverse study populations, considering both ancestral and
411 subtype-dependent biological heterogeneity in our predictive models. From a more
412 theoretical perspective, this work will inform the utilization of TWAS methods in polygenic
413 traits and diverse study populations, stressing rigorous validation of predictive models
414 prior to imputation and careful modeling to capture associations with outcomes of interest
415 in diverse populations.
416

Methods

Data collection

Study population

The Carolina Breast Cancer Study (CBCS) is a population-based study conducted in North Carolina (NC) that began in 1993; study details and sampling schemes have been described in previous CBCS work [19,52]. Patients of breast cancer aged between 20 and 74 years were identified using rapid case ascertainment in cooperation with the NC Central Cancer Registry, with self-identified African American and young women (ages 20-49) oversampled using randomized recruitment [19]. Randomized recruitment allows sample weighting to make inferences about the frequency of subtype in the NC source population. Details regarding patient recruitment and clinical data collections are described in Troester et al [2].

Date of death and cause of death were identified by linkage to the National Death Index. All diagnosed with breast cancer have been followed for vital status from diagnosis until date of death or date of last contact. Breast cancer-related deaths were classified as those that listed breast cancer (International Statistical Classification of Disease codes 174.9 and C-50.9) as the underlying cause of death on the death certificate. By the end of follow-up, we identified 674 deaths, 348 of which were due to breast cancer. In total, we compiled 3,828 samples (1,865 AA and 1,963 WW) from all phases of CBCS with relevant survival and clinical variables.

CBCS genotype data

Approximately 50% of the SNPs for the OncoArray were selected as a "GWAS backbone" (Illumina HumanCore), which aimed to provide high coverage for the majority of common variants through imputation. The remaining SNPs were selected from lists supplied by six disease-based consortia, together with a seventh list of SNPs of interest to multiple disease-focused groups. Approximately 72,000 SNPs were selected specifically for their relevance to breast cancer. The sources for the SNPs included in this backbone, as well as backbone manufacturing, calling, and quality control, are discussed in depth by the OncoArray Consortium [53]. All samples were imputed using the October 2014 (v.3) release of the 1000 Genomes Project dataset as a reference panel in the standard two-stage imputation approach, using *SHAPEIT2* for phasing and *IMPUTEv2* for imputation [54–56]. All genotyping, genotype calling, quality control, and imputation was done at the DCEG Cancer Genomics Research Laboratory [53].

From the provided genotype data, we excluded variants (1) with a minor frequency less than 5% and (2) that deviated significantly from Hardy-Weinberg equilibrium at $P < 10^{-8}$ using the appropriate functions in *PLINK v1.90b3* [57,58]. Finally, we intersected genotyping panels for the AA and WW samples, resulting in 5,989,134 autosomal variants and 334,391 variants of the X chromosome. CBCS genotype data was coded as dosages, with reference and alternative allele coding as in the National Center for Biotechnology Information's Single Nucleotide Polymorphism Database (dbSNP).

CBCS gene expression data

462 Paraffin-embedded tumor blocks were requested from participating pathology
463 laboratories for each sample, reviewed, and assayed for gene expression using
464 Nanostring as discussed previously [2]. In total, 1,388 samples with invasive breast
465 cancer from the CBCS were analyzed for a total of 406 autosomal genes and 11 genes
466 on the X chromosome. All assays were performed in the Translational Genomics
467 Laboratory at the University of North Carolina at Chapel Hill.
468
469 We used the *NanoStringQCPro* package in Bioconductor to first eliminate samples that
470 did not have sufficient Nanostring data quality [59]. Next, we normalized distributional
471 differences between lanes with upper-quartile normalization [60]. Unwanted technical and
472 biological variation (i.e. tissue heterogeneity) was estimated in the resulting gene
473 expression data with techniques from the *RUVSeq* package from Bioconductor [61].
474 Unwanted variation was controlled using the distribution of 11 endogenous housekeeping
475 genes on the Nanostring gene expression panel. Ultimately, we removed 2 dimensions
476 of unwanted variation from the variance-stabilized transformation of the gene expression
477 data [62,63]. We lastly used principal component analysis to detect and remove any
478 significant, potential outliers. A final intersection of samples that had both genotype and
479 gene expression data gave us a final sample of 1,199 subjects (628 AA women and 571
480 WW women).
481
482 *TCGA genotype data*
483 Birdseed genotype files of 914 of WW and AA women were downloaded from the Genome
484 Data Commons (GDC) legacy (GRCh37/hg19) archive. Genotype files were merged into

a single binary PLINK file format (BED/FAM/BIM) and imputed using the October 2014 (v.3) release of the 1000 Genomes Project dataset as a reference panel in the standard two-stage imputation approach, using SHAPEIT v2.837 for phasing and IMPUTE v2.3.2 for imputation [54–56]. We excluded variants (1) with a minor allele frequency of less than 1%, (2) that deviated significantly from Hardy-Weinberg equilibrium ($P < 10^{-8}$) using appropriate functions in PLINK v1.90b3 [57,58], and (3) located on sex chromosomes. We further excluded any SNPs not found on the final, quality-controlled CBCS genotype data. Final TCGA genotype data was coded as dosages, with reference and alternative allele coding as in dbSNP.

TCGA expression data

TCGA level-3 normalized RNA expression data were downloaded from the Broad Institute's GDAC Firehose (2016/1/28 analysis archive) and subsetted to the 417 genes analyzed in CBCS. A total of 412 of these 417 were available in TCGA expression data.

Computational methods

Deconvolution of bulk tumor RNA

A study pathologist analyzed tumor microarrays (TMAs) from 176 of the 1,199 subjects to estimate area of dissections originating from epithelial tumor, assumed here as a proxy for the proportion of the bulk RNA expression attributed to the tumor. Using these 176 observations as a training set and the normalized gene expressions as the design matrix, we trained a support vector machine model tuned over a 10-fold cross-validation [64,65]. The cross-validated model was then used to estimate tumor purities for the remaining

1,023 samples from their gene expressions. We do not consider tumor purity in final eQTL models and all downstream analyses.

eQTL analysis

We assessed the additive relationship between the gene expression values and genotypes with linear regression analysis using *MatrixeQTL* [66], in the following model:

$$E_g = X_s\beta_s + X_C\beta_C + \epsilon_g,$$

where E_g is the gene expression of gene g, X_s is the vector of genotype dosages for a given SNP s, C is a matrix of covariates, β_s and β_C are the effect-sizes on gene expression for the SNP s and the covariates C, respectively, and ϵ is assumed to be Gaussian random error with mean 0 and common variance σ^2 for all genes g.

We calculated both cis- (variant-gene distance less than 500 kb) and trans-associations between variants and genes. Classical P-values were calculated for Wald-type tests of H_0: $\beta_s = 0$ and were adjusted post-hoc via the Benjamini-Bogomolov hierarchical error control procedure, *TreeQTL* [20]. We conducted all eQTL analyses stratified by race. Age, BMI, postmenopausal status, and the first 5 principal components of the joint AA and WW genotype matrix were included in the models as covariates in C. Estimated tumor purity was also included as a covariate to assess its impact on strength and location of eQTLs. Any SNP found in an eQTL with Benajmini-Bogomolov adjust P-value $BBFDR < 0.05$ is defined as an eSNP using *TreeQTL* [20]. The corresponding gene in that eQTL is defined as an eGene. We exclude samples with Normal-like subtype, as classified by the PAM50 classifier, due to generally low tumor content.

We downloaded healthy tissue eQTLs from the Genotype-Tissue Expression (GTEx) Project and cross-referenced eGenes and corresponding eSNPs between CBCS and GTEx in healthy breast mammary tissue, EBV-transformed lymphocytes, and subcutaneous adipose tissue. The Genotype-Tissue Expression (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 data used for the analyses described in this manuscript were obtained from the GTEx Portal on 05/12/19.

Estimation of cis-heritability

Cis-heritability (cis-h^2) was estimated using the GREML-LDMS method, proposed to estimate heritability by correction for bias in linkage disequilibrium (LD) in estimated SNP-based heritability [26]. Analysis was conducted using *GCTA* v.1.92 [67]. For downstream analysis, we only consider the 151 genes (81 in AA women and 100 in WW women) with cis-h^2 that can be estimated with nominal P-value < 0.10.

Predictive tumor expression models

We adopt general techniques from PrediXcan and FUSION to estimate eQTL-effect sizes for predictive models of tumor expression from germline variants [13,14]. First, gene expressions were residualized for the covariates C included in the eQTL models (age, BMI, postmenopausal status, and genotype PCs) given the following ordinary least squares model:

$$E_g = X_C\beta_C + \epsilon_g.$$

We then consider downstream analysis on $\tilde{E}_g \equiv E_g - X_C\hat{\beta}_C$.

For a given gene g, we consider the following linear predictive model:

$$\tilde{E}_g = X_g w_g + \epsilon_g,$$

where $\tilde{E}_g$ is the gene expression of gene g, residualized for the covariate matrix X_C, X_g is the genotype matrix for gene g that includes all cis-SNPs for gene g (within 500 kb of either the 5' or 3' end of the gene) and all trans-eQTLs with $BBFDR < 0.01$, w_g is a vector of effect-sizes for eQTLs in X_g, and ϵ_g is Gaussian random error with mean 0 and common variance for all g.

We estimate w_g with the best predictive of three schemes: (1) elastic-net regularized regression with mixing parameter $\alpha = 0.5$ and λ penalty parameter tuned over 5-fold cross-validation [13,68], (2) linear mixed modeling where the genotype matrix X_g is treated as a matrix of random effects and $\hat{w}_g$ is taken as the best linear unbiased predictor (BLUP) of w_g, using *rrBLUP* [69], and (3) multivariate linear mixed modeling as described above, estimated using *GEMMA* v.0.97 [70].

In these models, the genotype matrix X_g is pruned for linkage disequilibrium (LD) prior to modeling using a window size of 50, step size of 5, and LD threshold of 0.5 using *PLINK* v.1.90b3 [58] to account for redundancy in signal. The final vectors $\hat{w}_g$ of effect-sizes for each gene g are estimated by the estimation scheme with the best 5-fold cross-validation

performance. All predicted models are stratified by race, i.e. an individual model of tumor expression for AA women and WW women for each gene g.

To impute expression into external cohorts, we then construct the germline genetically-regulated tumor expression $GReX_g$ of gene g given $\widehat{w}_g$ in the predictive model as follows:

$$GReX_g = X_{g,new}\widehat{w}_g,$$

where $X_{g,new}$ is the genotype matrix of all available SNPs in the feature set of $\widehat{w}_g$ in a GWAS cohort.

All final models are available here: https://github.com/bhattacharya-a-bt/CBCS_TWAS_Paper.

Validation in TCGA

Using our stratified predictive models of tumor expression, we imputed expression in TCGA and measured predictive accuracy of each gene through prediction R^2, defined here as the squared Spearman correlation between observed and imputed expression. It is important to note that all variants in the CBCS-trained predictive models are not represented in the TCGA genotype data. Predictive performance in TCGA was also assessed stratified by PAM50 intrinsic subtype and estrogen receptor status.

To account for sampling variability in calculating correlations in validation cohorts of smaller sample sizes, we calculated a permutation null distribution for each gene by permuting observed expressions 10,000 times and calculating a “null” prediction R^2 at

each permutation. The sample validation prediction R^2 was compared to this permutation null distribution to generate an empirical P-value for the sample R^2, using Storey's *qvalue* package. We then calculated q-values from these empirical P-values, controlling for a false discovery rate of 0.05 [27]. Lastly, we constructed confidence intervals for R^2 by inverting the acceptance region from the permutation test [28].

Validation in CBCS

We used an entirely held-out sample of 2,308 women from CBCS as a validation set of Nanostring nCounter data on a codeset of 166 genes. These samples were normalized as outlined before. We used the same validation methods as in TCGA, as well using a permutation method to assess the statistical significance of predictive performance, stratified by PAM50 subtype and estrogen receptor status.

PAM50 subtyping

GReX in CBCS were first estimated as outlined above. We residualized the original tumor expression E for these imputed expression values to form a matrix of tumor expression adjusted for GReX ($\tilde{E}$). We then classified each subject into PAM50 subtypes based on both E and $\tilde{E}$, using the procedure summarized by Parker et al [71,72].

Survival modeling

Here, we defined a relevant event as a death due to breast cancer. We aggregated all deaths not due to breast cancer as a competing risk. Any subjects lost to follow-up were treated as right-censored observations. We estimated the association of GReX with

breast cancer survival by modeling the race-stratified cause-specific hazard function of breast cancer-specific mortality, stratifying on race [73]. For a given gene g, the model has form

$$\lambda_k(t) = \lambda_{0k}(t)e^{GReX_g\beta_g + Z_C\beta_C},$$

where β_g is the effect size of $GReX_g$ on the hazard of breast cancer-specific mortality, Z_C represents the matrix of covariates (age at diagnosis, estrogen-receptor status at diagnosis, tumor stage at diagnosis, and study phase), and β_C are the effect sizes of these covariates on survival. $\lambda_k(t)$ is the hazard function specific to breast cancer mortality, and $\lambda_{0k}(t)$ is the baseline hazard function. We test H_0: $\beta_g = 0$ for each gene g with Wald-type tests, as in a traditional Cox proportional hazards model. We correct for genomic inflation and bias using *bacon*, a method that constructs an empirical null distribution using a Gibbs sampling algorithm by fitting a three-component normal mixture on Z-statistics from TWAS tests of association [74].

Here, we consider only the 46 genes that have CV $R^2 > 0.01$ in AA women and the 57 genes that have CV $R^2 > 0.01$ in WW women for race-stratified survival modeling. We adjust tests for β_g via the Benjamini-Hochberg procedure at a false discovery rate of 0.10.

For comparison, we run a GWAS to analyze the association between germline SNPs and breast cancer-specific survival using *GWASTools* [75]. We use a similar cause-specific hazards model with the same covariates as in the TWAS models of association, correcting for false discovery with the Benjamini-Hochberg procedure.

644 *Inspection of collider bias*

645 To assess collider bias when conditioning for breast cancer incidence in case-only
646 studies, such as CBCS, we test for association for the GReX of genes with breast cancer
647 risk using iCOGs summary statistics from BCAC [35–37], using the weighted burden test
648 identified by FUSION [14]. In summary, we compose a weighted Z test statistic as follows:

649 $$\tilde{Z} = \frac{\boldsymbol{W}Z}{\boldsymbol{W}\boldsymbol{\Sigma}_{S,S}\boldsymbol{W}'},$$

650 where Z is the vector of Z-statistics from iCOGs and $\boldsymbol{W} = \boldsymbol{\Sigma}_{\boldsymbol{e,s}}^{-\mathbf{1}}\boldsymbol{\Sigma}_{\boldsymbol{s,s}}$ with $\boldsymbol{\Sigma}_{\boldsymbol{e,s}}$ is the
651 covariance matrix between all SNPs represented in Z and the gene expression of the
652 given gene and $\boldsymbol{\Sigma}_{s,s}$ is the covariance among all SNPs.

653

654 *Power analysis*

655 Using *survSNP* [76], we generated the empirical power of a GWAS to detect various
656 hazard ratios with 3,828 samples with 1,000 simulation replicates at a significance level
657 of $P = 1.70 \times 10^{-8}$, corresponding to an FDR-adjusted $P = 0.10$. We assume an event
658 rate of 10%, a relative allelic frequency of the risk allele of 0.1 and estimate the 90th
659 percentile of times-to-event as a landmark time. Similarly, for genes of various cis-h^2, we
660 assessed the power of TWAS to detect various hazard ratios at $P = 0.0096$
661 (corresponding to FDR-adjusted $P = 0.10$) over 1,000 simulation replications from the
662 empirical distribution function of the GReX of the given gene.

663

664 **Abbreviations**

665 CBCS: Carolina Breast Cancer Study

666 GWAS: Genome-wide association study

667 LD: Linkage disequilibrium

668 SNP/V: Single nucleotide polymorphism/variant

669 TWAS: Transcriptome-wide association study

670 GTEx: The Genoype-Tissue Expression Project

671 BCAC: Breast Cancer Association Consortium

672 PRS: Polygenic risk score

673 WW: self-identified white women

674 AA: self-identified African American women

675 ER: estrogen receptor

676 eQTL: expression quantitative trait loci

677 AMBER: Alberta Moving Beyond Breast Cancer

678 eGene: eQTL-associated gene

679 eSNP: SNP found in an eQTL

680 FDR: false discovery rate

681 BBFDR: Benjamini-Bogomolov adjusted false discovery rate

682 h^2: heritability

683 TCGA: The Cancer Genome Atlas

684 BRCA: breast cancer

685 GReX: germline-genetically regulated tumor expression

686

687 **Declarations**

688 *Ethics approval and consent to participate*

689 This study was approved by the Office of Human Research Ethics at the University of
690 North Carolina at Chapel Hill, and written informed consent was obtained from each
691 participant.
692
693 *Consent for publication*
694 Not applicable
695
696 *Availability of data and materials*
697 Summary statistics eQTL results, tumor expression models, and relevant R code for
698 training expression models in CBCS are freely available at
699 https://github.com/bhattacharya-a-bt/CBCS_TWAS_Paper/.
700
701 *Competing interests*
702 C.M.P. is an equity stockholder in and consultant for BioClassifier LLC; C.M.P. is also
703 listed as an inventor on patent applications on the Breast PAM50 Subtyping assay. The
704 other authors declare that they have no competing interests.
705
706 *Funding*
707 Susan G. Komen® provided financial support for CBCS study infrastructure. Funding
708 was provided by the National Institutes of Health, National Cancer Institute P01-
709 CA151135, P50-CA05822, and U01-CA179715 to A.F.O, C.M.P. and M.A.T. M.I.L. is
710 supported by R01-HG009937, R01-MH118349, P01-CA142538, and P30-ES010126.
711 The Translational Genomics Laboratory is supported in part by grants from the National

712 Cancer Institute (3P30CA016086) and the University of North Carolina at Chapel Hill
713 University Cancer Research Fund. Genotyping was done at the DCEG Cancer Genomics
714 Research Laboratory using funds from the NCI Intramural Research Program.

715 The funders had no role in the design of the study, the collection, analysis, or
716 interpretation of the data, the writing of the manuscript, or the decision to submit the
717 manuscript for publication.

718

719 *Authors' contributions*

720 A.B., M.G., A.F.O., M.A.T., and M.I.L. conceived the study. A.B. developed the statistical
721 approaches, performed the analysis, and drafted the paper. A.B., M.A.T., and M.I.L.
722 performed initial exploratory analysis. C.M.P., M.A.T., and M.I.L. provided insight in
723 methodological approaches and analysis. M.G., A.F.O., C.M.P., and M.A.T. provided data
724 resources. M.A.T. and M.I.L. supervised the study. All authors approved and edited the
725 final manuscript.

726

727 *Acknowledgements*

728 We thank the Carolina Breast Cancer Study participants and volunteers. We also thank
729 Colin Begg, Jianwei Cai, Nilanjan Chatterjee, Alexander Gusev, Katherine Hoadley, Yun
730 Li, John Witte, and Emily Zabor for valuable discussion during the research process. We
731 thank Erin Kirk and Jessica Tse for their invaluable support during the research process.
732 We thank the DCEG Cancer Genomics Research Laboratory and acknowledge the
733 support from Stephen Chanock, Rose Yang, Meredith Yeager, Belynda Hicks, and Bin
734 Zhu.

References

1. Centers for Disease Control and Prevention. National Death Index [Internet]. 2018 [cited 2019 Apr 24]. Available from: https://www.cdc.gov/nchs/ndi/index.htm

2. Troester MA, Sun X, Allott EH, Geradts J, Cohen SM, Tse C-K, et al. Racial Differences in PAM50 Subtypes in the Carolina Breast Cancer Study. J Natl Cancer Inst [Internet]. Oxford University Press; 2018 [cited 2019 May 29];110:176. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28859290

3. Carey LA, Perou CM, Livasy CA, Dressler LG, Cowan D, Conway K, et al. Race, Breast Cancer Subtypes, and Survival in the Carolina Breast Cancer Study. JAMA [Internet]. 2006 [cited 2019 Jul 2];295:2492. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16757721

4. Hall IJ, Moorman PG, Millikan RC, Newman B. Comparative Analysis of Breast Cancer Risk Factors among African-American Women and White Women. Am J Epidemiol [Internet]. 2005 [cited 2019 Jul 2];161:40–51. Available from: https://academic.oup.com/aje/article-lookup/doi/10.1093/aje/kwh331

5. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature [Internet]. Europe PMC Funders; 2017 [cited 2019 Mar 22];551:92–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29059683

6. Lilyquist J, Ruddy KJ, Vachon CM, Couch FJ. Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. Cancer Epidemiol Biomarkers Prev [Internet]. American Association for Cancer Research; 2018 [cited 2019 Jun 7];27:380–94. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29382703

758 7. Escala-Garcia M, Guo Q, Dörk T, Canisius S, Keeman R, Dennis J, et al. Genome-
759 wide association study of germline variants and breast cancer-specific mortality. Br J
760 Cancer [Internet]. Nature Publishing Group; 2019 [cited 2019 Apr 25];120:647–57.
761 Available from: http://www.nature.com/articles/s41416-019-0393-x
762 8. Pirie A, Guo Q, Kraft P, Canisius S, Eccles DM, Rahman N, et al. Common germline
763 polymorphisms associated with breast cancer-specific survival. Breast Cancer Res
764 [Internet]. BioMed Central; 2015 [cited 2019 Apr 25];17:58. Available from:
765 http://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-015-0570-
766 7
767 9. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al.
768 Potential etiologic and functional implications of genome-wide association loci for
769 human diseases and traits. Proc Natl Acad Sci U S A [Internet]. National Academy of
770 Sciences; 2009 [cited 2019 Apr 25];106:9362–7. Available from:
771 http://www.ncbi.nlm.nih.gov/pubmed/19474294
772 10. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjálmsson BJ, Xu H, et al. Partitioning
773 Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common
774 Diseases. Am J Hum Genet [Internet]. Elsevier; 2014 [cited 2019 Apr 25];95:535.
775 Available from: http://www.ncbi.nlm.nih.gov/pubmed/25439723
776 11. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, et al. A method to
777 predict the impact of regulatory variants from DNA sequence. Nat Genet [Internet].
778 2015 [cited 2019 Apr 25];47:955–61. Available from:
779 http://www.ncbi.nlm.nih.gov/pubmed/26075791
780 12. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic

781 to Omnigenic. Cell [Internet]. Elsevier; 2017 [cited 2019 Jun 7];169:1177–86.
782 Available from: http://www.ncbi.nlm.nih.gov/pubmed/28622505
783 13. Gamazon ER, Wheeler HE, Shah KP, Mozaffari S V, Aquino-Michaels K, Carroll RJ,
784 et al. A gene-based association method for mapping traits using reference
785 transcriptome data. Nat Genet [Internet]. Nature Publishing Group; 2015 [cited 2019
786 Apr 6];47:1091–8. Available from: http://www.nature.com/articles/ng.3367
787 14. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. Integrative
788 approaches for large-scale transcriptome-wide association studies. Nat Genet
789 [Internet]. 2016 [cited 2019 Apr 6];48:245–52. Available from:
790 http://www.ncbi.nlm.nih.gov/pubmed/26854917
791 15. Wu L, Shi W, Long J, Guo X, Michailidou K, Beesley J, et al. A transcriptome-wide
792 association study of 229,000 women identifies new candidate susceptibility genes
793 for breast cancer. Nat Genet [Internet]. NIH Public Access; 2018 [cited 2019 Apr
794 25];50:968–78. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29915430
795 16. Hoffman JD, Graff RE, Emami NC, Tai CG, Passarelli MN, Hu D, et al. Cis-eQTL-
796 based trans-ethnic meta-analysis reveals novel genes associated with breast cancer
797 risk. PLoS Genet [Internet]. Public Library of Science; 2017 [cited 2019 Apr 25];13.
798 Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5391966/
799 17. Mogil LS, Andaleon A, Badalamenti A, Dickinson SP, Guo X, Rotter JI, et al.
800 Genetic architecture of gene expression traits across diverse populations. Epstein
801 MP, editor. PLOS Genet [Internet]. Public Library of Science; 2018 [cited 2019 Jun
802 13];14:e1007586. Available from: https://dx.plos.org/10.1371/journal.pgen.1007586
803 18. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. Genetic

804 analyses of diverse populations improves discovery for complex traits. Nature
805 [Internet]. Nature Publishing Group; 2019 [cited 2019 Jun 26];1. Available from:
806 http://www.nature.com/articles/s41586-019-1310-4
807 19. Newman B, Moorman PG, Millikan R, Qaqish BF, Geradts J, Aldrich TE, et al. The
808 Carolina Breast Cancer Study: integrating population-based epidemiology and
809 molecular biology. Breast Cancer Res Treat [Internet]. Kluwer Academic Publishers;
810 1995 [cited 2019 Mar 21];35:51–60. Available from:
811 http://link.springer.com/10.1007/BF00694745
812 20. Peterson CB, Bogomolov M, Benjamini Y, Sabatti C, Stegle O. Genetics and
813 population analysis TreeQTL: hierarchical error control for eQTL findings. [cited
814 2019 May 20]; Available from: http://bioinformatics.oxfordjournals.org/
815 21. Rummel S, Penatzer CE, Shriver CD, Ellsworth RE. PSPHL and breast cancer in
816 African American women: causative gene or population stratification? BMC Genet
817 [Internet]. BioMed Central; 2014 [cited 2019 May 29];15:38. Available from:
818 http://www.ncbi.nlm.nih.gov/pubmed/24650299
819 22. D'Arcy M, Fleming J, Robinson WR, Kirk EL, Perou CM, Troester MA. Race-
820 associated biological differences among Luminal A breast tumors. Breast Cancer
821 Res Treat [Internet]. 2015 [cited 2019 May 29];152:437–48. Available from:
822 http://www.ncbi.nlm.nih.gov/pubmed/26109344
823 23. Lu L, Pandey AK, Houseal MT, Mulligan MK. The Genetic Architecture of Murine
824 Glutathione Transferases. Sanyal S, editor. PLoS One [Internet]. Public Library of
825 Science; 2016 [cited 2019 May 29];11:e0148230. Available from:
826 https://dx.plos.org/10.1371/journal.pone.0148230

827 24. Zhang M, Lykke-Andersen S, Zhu B, Xiao W, Hoskins JW, Zhang X, et al.
828 Characterising cis-regulatory variation in the transcriptome of histologically normal
829 and tumour-derived pancreatic tissues. Gut [Internet]. NIH Public Access; 2018
830 [cited 2019 May 29];67:521–33. Available from:
831 http://www.ncbi.nlm.nih.gov/pubmed/28634199
832 25. Calabrese C, Lehmann K, Urban L, Liu F, Erkek S, Fonseca N, et al. Assessing the
833 Gene Regulatory Landscape in 1,188 Human Tumors. bioRxiv [Internet]. Cold
834 Spring Harbor Laboratory; 2017 [cited 2019 Jun 7];225441. Available from:
835 https://www.biorxiv.org/content/10.1101/225441v1.article-info
836 26. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, Lee SH, et al. Genetic
837 variance estimation with imputed variants finds negligible missing heritability for
838 human height and body mass index. Nat Genet [Internet]. NIH Public Access; 2015
839 [cited 2019 May 29];47:1114–20. Available from:
840 http://www.ncbi.nlm.nih.gov/pubmed/26323059
841 27. Storey JD, Bass AJ, Dabney A, Robinson D, Warnes G. qvalue: Q-value estimation
842 for false discovery rate control [Internet]. 2019. Available from:
843 http://github.com/jdstorey/qvalue
844 28. Tritchler D. On inverting permutation tests. J Am Stat Assoc [Internet]. 1984 [cited
845 2019 Jun 10];79:200–7. Available from: https://about.jstor.org/terms
846 29. Liao Y, Liao Y, Li J, Li J, Fan Y, Xu B. Polymorphisms in AURKA and AURKB are
847 associated with the survival of triple-negative breast cancer patients treated with
848 taxane-based adjuvant chemotherapy. Cancer Manag Res [Internet]. Dove Press;
849 2018 [cited 2019 Aug 20];10:3801–8. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/30288111

30. Shimoi T, Hamada A, Yamagishi M, Hirai M, Yoshida M, Nishikawa T, et al. *PIK3CA* mutation profiling in patients with breast cancer, using a highly sensitive detection system. Cancer Sci [Internet]. Wiley-Blackwell; 2018 [cited 2019 Aug 20];109:2558–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29906308

31. Cizkova M, Susini A, Vacher S, Cizeron-Clairac G, Andrieu C, Driouch K, et al. PIK3CA mutation impact on survival in breast cancer patients and in ERα, PR and ERBB2-based subgroups. Breast Cancer Res [Internet]. BioMed Central; 2012 [cited 2019 Aug 20];14:R28. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22330809

32. Rafiq S, Khan S, Tapper W, Collins A, Upstill-Goddard R, Gerty S, et al. A Genome Wide Meta-Analysis Study for Identification of Common Variation Associated with Breast Cancer Prognosis. Miao X, editor. PLoS One [Internet]. Public Library of Science; 2014 [cited 2019 Jun 3];9:e101488. Available from: https://dx.plos.org/10.1371/journal.pone.0101488

33. Khan S, Fagerholm R, Kadalayil L, Tapper W, Aittomäki K, Liu J, et al. Meta-analysis of three genome-wide association studies identifies two loci that predict survival and treatment outcome in breast cancer. Oncotarget [Internet]. Impact Journals, LLC; 2018 [cited 2019 Jun 3];9:4249–57. Available from: http://www.ncbi.nlm.nih.gov/pubmed/29423119

34. Paternoster L, Tilling K, Davey Smith G. Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges. Barsh GS, editor. PLOS Genet [Internet]. Public Library of Science;

2017 [cited 2019 Jul 15];13:e1006944. Available from: https://dx.plos.org/10.1371/journal.pgen.1006944

35. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet [Internet]. 2013 [cited 2019 Jul 15];45:353–61. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23535729

36. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet [Internet]. 2015 [cited 2019 Jul 15];47:373–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25751625

37. Guo Q, Schmidt MK, Kraft P, Canisius S, Chen C, Khan S, et al. Identification of Novel Genetic Markers of Breast Cancer Survival. JNCI J Natl Cancer Inst [Internet]. 2015 [cited 2019 Jul 15];107. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25890600

38. Martens FK, Janssens ACJW. How the Intended Use of Polygenic Risk Scores Guides the Design and Evaluation of Prediction Studies. Curr Epidemiol Reports [Internet]. Springer International Publishing; 2019 [cited 2019 May 29];1–7. Available from: http://link.springer.com/10.1007/s40471-019-00203-7

39. Mavaddat N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. Am J Hum Genet [Internet]. 2019 [cited 2019 May 29];104:21–34. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30554720

40. Ahearn TU, Zhang H, Michailidou K, Milne RL, Bolla MK, Dennis J, et al. Common

breast cancer risk loci predispose to distinct tumor subtypes [Internet]. bioRxiv. Peter A. Fasching; Available from: https://www.biorxiv.org/content/biorxiv/early/2019/08/15/733402.full.pdf?%3Fcollection=

41. Gusev A, Lawrenson K, Lin X, Lyra PC, Kar S, Vavra KC, et al. A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. Nat Genet [Internet]. Nature Publishing Group; 2019 [cited 2019 Jun 5];51:815–23. Available from: http://www.nature.com/articles/s41588-019-0395-x

42. Vecchi M, Confalonieri S, Nuciforo P, Viganò MA, Capra M, Bianchi M, et al. Breast cancer metastases are molecularly distinct from their primary tumors. Oncogene [Internet]. 2008 [cited 2019 Jun 5];27:2148–58. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17952122

43. Storr SJ, Thompson N, Pu X, Zhang Y, Martin SG. Calpain in Breast Cancer: Role in Disease Progression and Treatment Response. Pathobiology [Internet]. Karger Publishers; 2015 [cited 2019 Jun 5];82:133–41. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26330354

44. Storr SJ, Zhang S, Perren T, Lansdown M, Fatayer H, Sharma N, et al. The calpain system is associated with survival of breast cancer patients with large but operable inflammatory and non-inflammatory tumours treated with neoadjuvant chemotherapy. Oncotarget [Internet]. 2016 [cited 2019 Jun 5];7:47927–37. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27323818

45. Leloup L, Wells A. Calpains as potential anti-cancer targets. Expert Opin Ther

919 Targets [Internet]. NIH Public Access; 2011 [cited 2019 Jun 5];15:309–23. Available
920 from: http://www.ncbi.nlm.nih.gov/pubmed/21244345
921 46. Begg CB, Zabor EC. Detecting and Exploiting Etiologic Heterogeneity in
922 Epidemiologic Studies. Am J Epidemiol [Internet]. 2012 [cited 2019 Aug
923 20];176:512–8. Available from: https://academic.oup.com/aje/article-
924 lookup/doi/10.1093/aje/kws128
925 47. Martínez ME, Cruz GI, Brewster AM, Bondy ML, Thompson PA. What can we learn
926 about disease etiology from case-case analyses? Lessons from breast cancer.
927 Cancer Epidemiol Biomarkers Prev [Internet]. NIH Public Access; 2010 [cited 2019
928 Aug 20];19:2710–4. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20870734
929 48. Mitchell RE, Paternoster L, Davey Smith G. Mendelian Randomization in Case Only
930 Studies: A Promising Approach to be Applied With Caution. Am J Cardiol [Internet].
931 Excerpta Medica; 2018 [cited 2019 Jul 15];122:2169–71. Available from:
932 https://www.sciencedirect.com/science/article/pii/S0002914918318563
933 49. Dudbridge F, Allen RJ, Sheehan NA, Schmidt AF, Lee JC, Jenkins RG, et al.
934 Adjustment for index event bias in genome-wide association studies of subsequent
935 events. Nat Commun [Internet]. Nature Publishing Group; 2019 [cited 2019 Jul
936 15];10:1561. Available from: http://www.nature.com/articles/s41467-019-09381-w
937 50. Guo Q, Burgess S, Turman C, Bolla MK, Wang Q, Lush M, et al. Body mass index
938 and breast cancer survival: a Mendelian randomization analysis. Int J Epidemiol
939 [Internet]. Oxford University Press; 2017 [cited 2019 Jul 17];46:1814–22. Available
940 from: http://www.ncbi.nlm.nih.gov/pubmed/29232439
941 51. Conway K, Parrish E, Edmiston SN, Tolbert D, Tse C-K, Moorman P, et al. Risk

942 factors for breast cancer characterized by the estrogen receptor alpha A908G
943 (K303R) mutation. Breast Cancer Res [Internet]. BioMed Central; 2007 [cited 2019
944 Jun 6];9:R36. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17553133
945 52. Hair BY, Hayes S, Tse C-K, Bell MB, Olshan AF. Racial differences in physical
946 activity among breast cancer survivors: implications for breast cancer care. Cancer
947 [Internet]. NIH Public Access; 2014 [cited 2019 Mar 21];120:2174–82. Available
948 from: http://www.ncbi.nlm.nih.gov/pubmed/24911404
949 53. Amos CI, Dennis J, Wang Z, Byun J, Schumacher FR, Gayther SA, et al. The
950 OncoArray Consortium: A Network for Understanding the Genetic Architecture of
951 Common Cancers. Cancer Epidemiol Biomarkers Prev [Internet]. NIH Public Access;
952 2017 [cited 2019 Apr 1];26:126–35. Available from:
953 http://www.ncbi.nlm.nih.gov/pubmed/27697780
954 54. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A General
955 Approach for Haplotype Phasing across the Full Spectrum of Relatedness. Gibson
956 G, editor. PLoS Genet [Internet]. Public Library of Science; 2014 [cited 2019 Jun
957 26];10:e1004234. Available from: https://dx.plos.org/10.1371/journal.pgen.1004234
958 55. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for
959 thousands of genomes. Nat Methods [Internet]. Nature Publishing Group; 2012
960 [cited 2019 Jun 26];9:179–81. Available from:
961 http://www.nature.com/articles/nmeth.1785
962 56. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation
963 Method for the Next Generation of Genome-Wide Association Studies. Schork NJ,
964 editor. PLoS Genet [Internet]. Public Library of Science; 2009 [cited 2019 Jun

965 26];5:e1000529. Available from: https://dx.plos.org/10.1371/journal.pgen.1000529
966 57. Wigginton JE, Cutler DJ, Abecasis GR. A Note on Exact Tests of Hardy-Weinberg
967 Equilibrium [Internet]. Am. J. Hum. Genet. 2005. Available from:
968 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1199378/pdf/AJHGv76p887.pdf
969 58. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al.
970 PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage
971 Analyses. Am J Hum Genet [Internet]. 2007 [cited 2019 Mar 21];81:559–75.
972 Available from: www.ajhg.org
973 59. Nickles D, Sandmann T, Ziman R, Bourgon R. NacoStringQCPro [Internet].
974 Available from:
975 https://www.bioconductor.org/packages/release/bioc/vignettes/NanoStringQCPro/ins
976 t/doc/vignetteNanoStringQCPro.pdf
977 60. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for
978 normalization and differential expression in mRNA-Seq experiments. BMC
979 Bioinformatics [Internet]. BioMed Central; 2010 [cited 2019 Mar 22];11:94. Available
980 from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-
981 94
982 61. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor
983 analysis of control genes or samples. Nat Biotechnol [Internet]. Nature Publishing
984 Group; 2014 [cited 2019 Mar 22];32:896–902. Available from:
985 http://www.nature.com/articles/nbt.2931
986 62. Anders S, Huber W. Differential expression analysis for sequence count data.
987 Genome Biol [Internet]. BioMed Central; 2010 [cited 2019 Mar 22];11:R106.

988 Available from: http://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-
989 11-10-r106
990 63. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion
991 for RNA-seq data with DESeq2. Genome Biol [Internet]. BioMed Central; 2014 [cited
992 2019 Mar 22];15:550. Available from:
993 http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8
994 64. Cortes C. Support-Vector Networks [Internet]. 1995. Available from:
995 http://image.diku.dk/imagecanon/material/cortes_vapnik95.pdf
996 65. Misc Functions of the Department of Statistics, ProbabilityTheory Group (Formerly:
997 E1071), TU Wien [R package e1071 version 1.7-1]. Comprehensive R Archive
998 Network (CRAN); [cited 2019 Apr 16]; Available from: https://cran.r-
999 project.org/web/packages/e1071/index.html
1000 66. Shabalin AA. Gene expression Matrix eQTL: ultra fast eQTL analysis via large
1001 matrix operations. 2012 [cited 2019 Mar 21];28:1353–8. Available from:
1002 http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL
1003 67. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide
1004 complex trait analysis. Am J Hum Genet [Internet]. Elsevier; 2011 [cited 2019 May
1005 29];88:76–82. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21167468
1006 68. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear
1007 Models via Coordinate Descent. J Stat Softw [Internet]. 2010 [cited 2019 Apr
1008 6];33:1–22. Available from: http://www.jstatsoft.org/v33/i01/
1009 69. Endelman JB. Ridge Regression and Other Kernels for Genomic Selection with R
1010 Package rrBLUP. Plant Genome. 2011;4:250–5.

1011 70. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for
1012 genome-wide association studies. Nat Methods [Internet]. NIH Public Access; 2014
1013 [cited 2019 May 29];11:407–9. Available from:
1014 http://www.ncbi.nlm.nih.gov/pubmed/24531419
1015 71. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised
1016 risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol [Internet].
1017 American Society of Clinical Oncology; 2009 [cited 2019 May 20];27:1160–7.
1018 Available from: http://www.ncbi.nlm.nih.gov/pubmed/19204204
1019 72. Gendoo DMA, Ratanasirigulchai N, Schröder M, Pare L, Parker JS, Prat A, et al.
1020 genefu: a package for breast cancer gene expression analysis [Internet]. 2018.
1021 Available from:
1022 http://www.bioconductor.org/packages/release/data/experiment/html/breastCancerV
1023 DX.html
1024 73. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model
1025 analyses for competing risk data. Stat Med [Internet]. Wiley-Blackwell; 2017 [cited
1026 2019 Apr 29];36:4391–400. Available from:
1027 http://www.ncbi.nlm.nih.gov/pubmed/28913837
1028 74. van Iterson M, van Zwet EW, Heijmans BT, Heijmans BT. Controlling bias and
1029 inflation in epigenome- and transcriptome-wide association studies using the
1030 empirical null distribution. Genome Biol [Internet]. BioMed Central; 2017 [cited 2019
1031 Jun 3];18:19. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28129774
1032 75. Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I, et al.
1033 GWASTools: an R/Bioconductor package for quality control and analysis of genome-

wide association studies. Bioinformatics [Internet]. 2012 [cited 2019 Jun 7];28:3329–31. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts610

76. Owzar K, Li Z, Cox N, Jung S-H. Power and Sample Size Calculations for SNP Association Studies With Censored Time-to-Event Outcomes. Genet Epidemiol [Internet]. John Wiley & Sons, Ltd; 2012 [cited 2019 Jun 4];36:538–48. Available from: http://doi.wiley.com/10.1002/gepi.21645

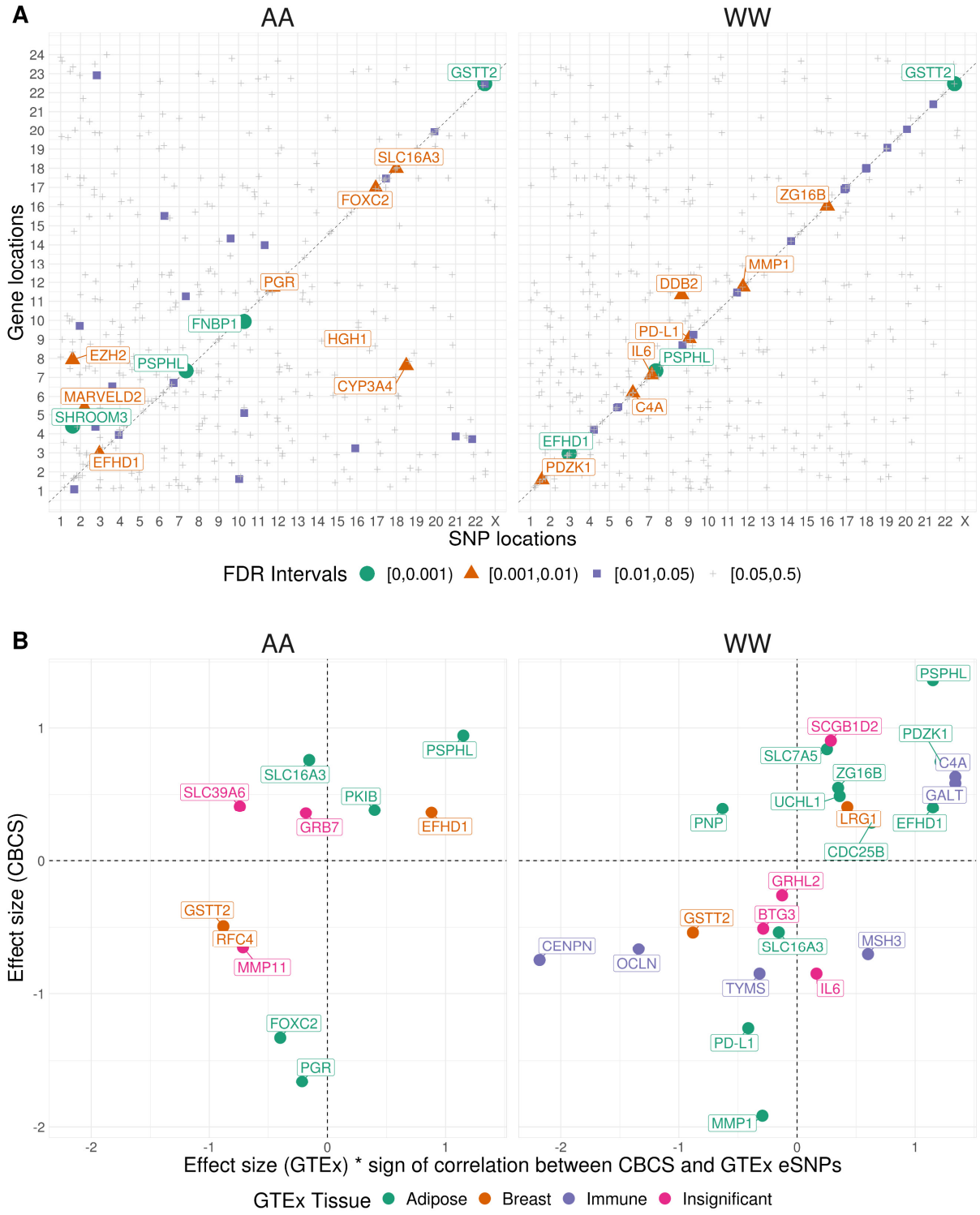

Figure 1: (A) Cis-trans plot of top eQTL by gene stratified by SRR. Each point represents the top eQTL for a given gene. The color and size of each point reflects the Benjamini-Bogomolov FDR-adjusted P-value (BBFDR) for that eQTL. eGenes with $BBFDR < 0.01$ are labelled. (B) Comparison of effect sizes of eGenes with significant cis-eQTLs in CBCS (Y -axis) and GTEx (X-axis) over tissue type, stratified by race. eGenes are colored by the GTEx tissue that shows the largest effect size. GTEx effect sizes on the X-axis are multiplied by the sign of the correlation between the genotypes of the GTEx and CBCS eSNPs.

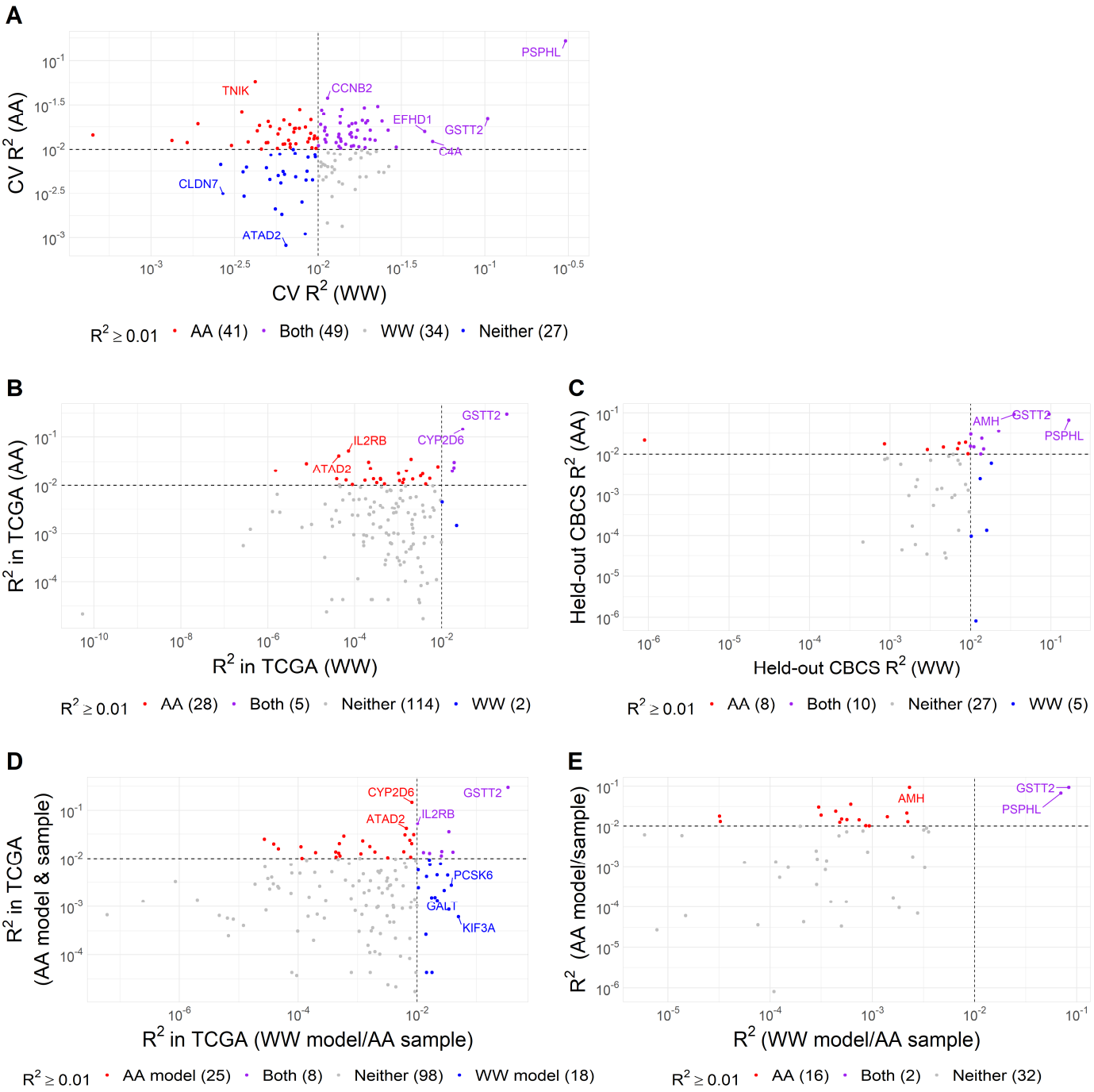

Figure 2: (A) Comparison of cross-validation R^2 across race in CBCS. Cross-validation R^2 in CBCS WW women (X-axis) and CBCS AA women (Y-axis) for each of the 151 analyzed genes. Scales are logarithmic. Dotted lines represent $R^2 = 0.01$. Colors represent the model with which a given gene can be predicted at $R^2 > 0.01$. (B) Cross-validation R^2 in CBCS (X-axis) and square Spearman correlation between observed expression and GReX in TCGA-BRCA (Y-axis) in AA sample (left) and WW sample (right). Pearson correlations between R^2calculated on the raw scale. R^2 are plotted on the log-scale. (C) Comparison of validation R^2 across race in TCGA for 149 analyzed genes found in TCGA expression data. (D) Comparison of validation R^2 across race in held-out CBCS samples for 51 analyzed genes. (E) Comparison of R^2 of genes in TCGA AA sample imputed from WW models (X-axis) and the AA models (Y-axis). (F) Comparison of R^2 of genes in held-out CBCS AA sample imputed from WW models (X-axis) and the AA models (Y-axis)

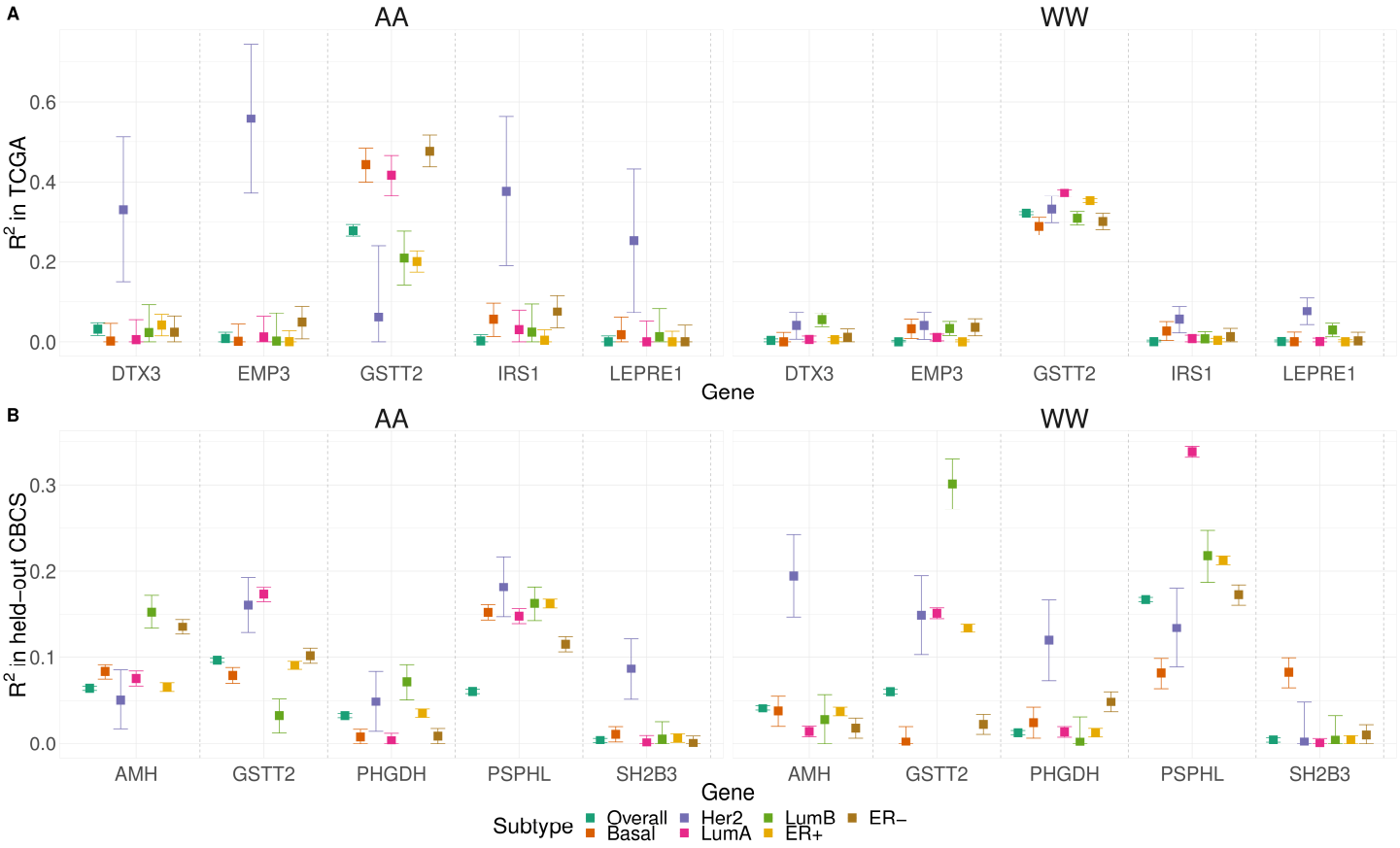

Figure 3: Validation R^2 across PAM50 molecular subtype and estrogen receptor status, stratified by race, for example genes with highly variable R^2 in TCGA (A) and held-out CBCS (B). Squared Spearman correlation (Y-axis), denoted R^2, between observed and predicted gene expression is plotted for different genes (X-axis), stratfied by PAM50 subtype and estrogen receptor status. Points are colored and shaped according to subtype. Error bars provide 90% confidence intervals inverted from the corresponding permutation test.

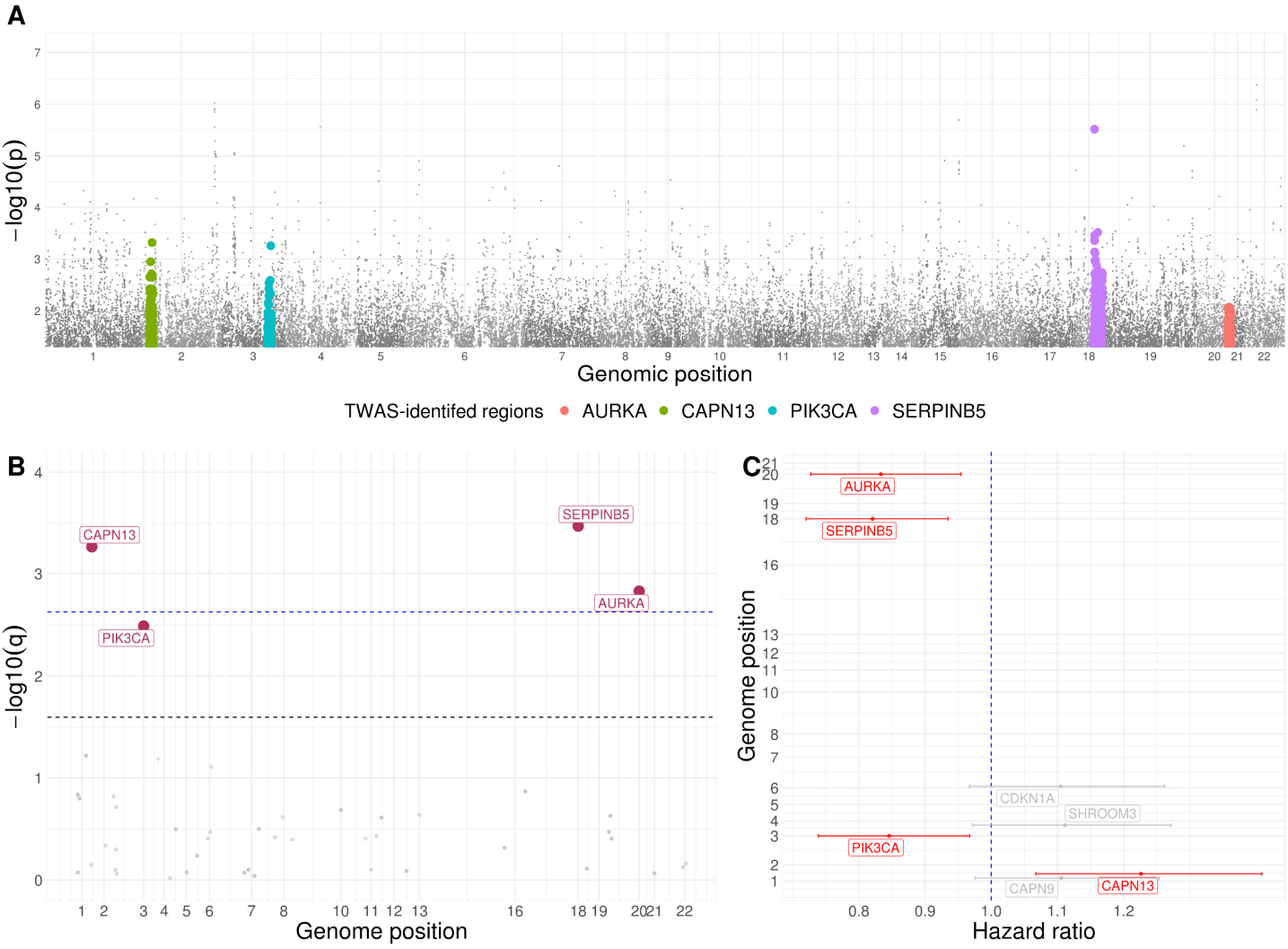

Figure 4: GWAS and TWAS results in AA women. (A) Manhattan plot of traditional GWAS on breast cancer survival. Genomic regions found to be significantly associated with survival in TWAS are represented in various colors. No SNVs reach Benjamini-Hochberg FDR-adjusted genome-wide significance. (B) Manhattan plot of TWAS on breast cancer survival. Genomic regions found to be significant at FDR-adjusted $P < 0.10$ are highlighted in red. The blue line represents a cutoff of FDR-adjusted $\alpha = 0.05$ and the dotted black line represents a cutoff of FDR-adjusted $\alpha = 0.10$. (C) Caterpillar plot of log-hazard rates with FDR-adjusted 90% confidence levels (X-axis) and genomic position (Y-axis). Results shown are significant at nominal $P < 0.10$. Genes highlighted in red represent genes with GReX significantly associated with survival at FDR-adjusted $P < 0.10$.

Region	Gene	Hazard Ratio (90% CI)[a]	Z-Statistic[a]	P-value[a]	GReX R^2[b]
20q13.2	*AURKA*	0.83 (0.73, 0.95)	-2.52	1.5×10^{-3}	0.021
2p23.1	*CAPN13*	1.22 (1.07, 1.41)	2.76	5.4×10^{-4}	0.011
3q26.32	*PIK3CA*	0.85 (0.74, 0.97)	-2.34	3.2×10^{-3}	0.013
18q21.33	*SERPINB5*	0.82 (0.72, 0.93)	-2.85	3.4×10^{-4}	0.010

Table 1: Genes with GReX found in association with breast cancer-specific survival in AA women. (a) Hazard ratio and FDR-adjusted 90% confidence intervals, Z-statistic, and P-value of association of GReX with breast cancer-specific survival. (b) Cross-validation R^2 of gene expression in AA models.

Gene	Closest survival-associated SNP[a]	Distance to closest survival-associated SNP[a]	Hazard ratio, adjusting for adjacent GWAS-SNP (90% CI)[b]	P-value, adjusting for adjacent risk SNPs[b]
AURKA	rs202100873	87.1 kb	0.84 (0.74, 0.94)	0.027
CAPN13	rs72068647	266.9 kb	1.18 (1.04, 1.33)	0.046
PIK3CA	rs66487567	271.9 kb	0.88 (0.78, 1.00)	0.096
SERPINB5	rs376302305	89.4 kb	0.84 (0.75, 0.94)	0.028

Table 2: Genes with GReX found in association with breast cancer-specific survival. (a) Top survival-associated SNP in cis-region of the given gene from GWAS for survival and distance of top cis-SNP from gene. (b) FDR-adjusted hazard ratio, 90% confidence interval, and P-value for association of GReX and breast cancer-specific survival, adjusting for adjacent survival-associated SNPs.