Transcriptome-wide association study of risk of recurrence in Black and White breast cancer patients

Achal Patel[1], Montserrat García-Closas[2,3], Andrew F. Olshan[1,4], Charles M. Perou[4,5,6], Melissa A. Troester[1,6], Michael I. Love[5,7*], Arjun Bhattacharya[8*]

1. Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA
2. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, USA
3. Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK
4. Lineberger Comprehensive Cancer Center, University of North Carolina-Chapel Hill, Chapel Hill, USA
5. Department of Genetics, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA
6. Department of Pathology and Laboratory Medicine, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA
7. Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA
8. Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California-Los Angeles, Los Angeles, CA, USA

*Correspondence can be directed to AB (abtbhatt@ucla.edu) and MIL (milove@email.unc.edu)

ABSTRACT

Background: Continuous risk of recurrence scores (CRS) based on PAM50 gene expression are vital prognostic tools for breast cancer (BC). Studies have shown that Black women (BW) have higher CRS than White women (WW). Although systemic injustices contribute substantially to BC disparities, evidence for biological and germline contributions is emerging. We investigated germline genetic associations with CRS and CRS disparity through a Transcriptome-Wide Association Study (TWAS).

Methods: In the Carolina Breast Cancer Study, using race-specific predictive models of tumor expression from germline genetics, we performed race-stratified (N=1,043 WW, 1083 BW) linear regressions of three CRS (ROR-S: PAM50 subtype score; Proliferation Score; ROR-P: ROR-S plus Proliferation Score) on imputed Genetically-Regulated tumor eXpression (GReX). Using Bayesian multivariate regression and adaptive shrinkage, we tested TWAS-significant genes for associations with PAM50 tumor expression and subtype to elucidate patterns of germline regulation underlying TWAS-gene and CRS associations.

Results: At FDR-adjusted $P < 0.10$, we detected 7 TWAS-genes among WW and 1 TWAS-gene among BW. Among WW, CRS showed positive associations with *MCM10, FAM64A, CCNB2*, and *MMP1* GReX and negative associations with *VAV3, PCSK6*, and *GNG11* GReX. Among BW, higher *MMP1* GReX predicted lower Proliferation score and ROR-P. TWAS-gene and PAM50 tumor expression associations highlighted potential mechanisms for TWAS-gene to CRS associations.

Conclusions: Among BC patients, we find differential germline associations with three CRS by race, underscoring the need for larger, more diverse datasets in molecular studies of BC. Our findings also suggest possible germline *trans*-regulation of PAM50 tumor expression, with potential implications for interpreting CRS in clinical settings.

Keywords: breast cancer recurrence, risk of recurrence, transcriptome-wide association study, molecular subtype, trans-eQTL mapping

ABBREVIATIONS

BC	Breast Cancer
BW	Black Women
CBCS	Carolina Breast Cancer Study
CRS	Continuous Risk of recurrence Score
eQTL	expression Quantitative Trait Locus
ER	Estrogen Receptor
GReX	Genetically-Regulated tumor eXpression
GWAS	Genome-Wide Association Study
HR	Hormone Receptor
LumA	Luminal A
LumB	Luminal B
ROR	Risk of Recurrence
SCC	Subtype-Centroid Correlations
SNP	Single Nucleotide Polymorphism
TCGA	The Cancer Genome Atlas
TWAS	Transcriptome-Wide Association Study
WW	White Women

INTRODUCTION (Manuscript Word Count = 3000)

Tumor expression-based molecular profiling has improved clinical classification of breast cancer (BC) [1-3]. One tool is the PAM50 assay which integrates tumor expression of 50 genes (from approximately 1,900 “intrinsic” genes identified through microarray) to determine intrinsic molecular subtypes: Luminal A (LumA), Luminal B (LumB), Human epidermal growth factor 2-enriched (HER2-enriched), Basal-like, Normal-like [1, 4]. Continuous risk of recurrence scores (CRS) generated from PAM50 tumor expression have prognostic value in clinical settings. [5-7]. For node negative, hormone receptor (HR) positive/HER2 negative BC, ROR-PT (a CRS determined by PAM50-subtype score, PAM50-based Proliferation score, and tumor size) offers overall and late distant recurrence information; other multigene signatures (OncotypeDx and EPclin) provide similar prognostic information for clinical decision-making [7, 8].

In the Carolina Breast Cancer Study (CBCS), Black women (BW) with breast cancer have disproportionately higher CRS than White Women [9], with similar disparities in Oncotype Dx recurrence score [9, 10]. Systemic injustices, like disparities in healthcare access, explain a substantial proportion of breast cancer outcome disparities [11-14], but recent studies suggest germline genetic variation may also play a role in outcome disparity. In The Cancer Genome Atlas (TCGA), BW had substantially higher polygenic risk scores for the more aggressive ER-negative subtype than WW, suggesting differential genetic contributions towards BC and especially ER-negative BC incidence [15]. In a transcriptome-wide association study (TWAS) of BC mortality, germline-regulated gene expression of four genes was associated with mortality among BW andnone associated among WW [16]. However, the role of germline genetic variation in relation to CRS and CRS disparity remains an important knowledge gap.

As racially-diverse genetic datasets typically have small samples of BW, gene-level association tests can be used to increase study power. These approaches include TWAS, which integrates relationships between single nucleotide polymorphisms (SNP) and gene expression with genome-wide association studies (GWAS) to prioritize gene-trait associations [17, 18]. TWAS has identified cancer susceptibility genes at loci previously undetected through GWAS, highlighting its improved power and interpretability [19-21]. Previous studies show that stratification of the entire TWAS (model training,

imputation, and association testing) is preferable in diverse populations, as models may perform poorly across ancestry groups and methods for TWAS in admixed populations are unavailable [16, 22].

Here, using data from the CBCS, which includes a large sample of Black BC patients with tumor gene expression data, we study race-specific germline genetic associations for CRS using TWAS. CRS included in this study are ROR-S (PAM50 subtype score), PAM50-based Proliferation score, and ROR-P (ROR-S + Proliferation score). Using race-specific predictive models for tumor expression from germline genetics, we identify sets of TWAS-genes associated with these CRS across BW and WW. We additionally investigate TWAS-genes for ROR-P for associations with PAM50 subtype and subtype-specific tumor gene expressions to elucidate germline contributions to PAM50 subtype, and how these mediate TWAS-gene and CRS associations. Unlike previous studies that correlated tumor gene expression (as opposed to germline-regulated tumor gene expression) with subtype or subtype-specific tumor gene expressions, TWAS enables directional interpretation of observed associations by ruling out reverse causality [17, 18].

METHODS

Data collection

Study population

The CBCS is a population-based study of North Carolina BC patients with three phases; study details have been previously described [23, 24]. Patients aged 20 to 74 were identified using rapid case ascertainment with the NC Central Cancer Registry with randomized recruitment to oversample self-identified Black and young women (ages 20-49) [9, 24]. Demographic and clinical data (age, menopausal status, body mass index, hormone receptor status, tumor stage, study phase, recurrence) were obtained through questionnaires and medical records. Recurrence data were available for CBCS Phase 3. The study was approved by the Office of Human Research Ethics at the University of North Carolina at Chapel Hill, and informed consent was obtained from each participant.

CBCS genotype data

Genotypes were assayed on the OncoArray Consortium's custom SNP array (Illumina Infinium OncoArray) [25] and imputed using the 1000 Genomes Project (v3) as a reference panel for two-step phasing and imputation using SHAPEIT2 and IMPUTEv2 [26-29]. The DCEG Cancer Genomics Research Laboratory conducted genotype calling, quality control, and imputation [25]. We excluded variants with less than 1% minor allele frequency and deviations from Hardy-Weinberg equilibrium at $P < 10^{-8}$ [30, 31]. We intersected genotyping panels for BW and WW samples, resulting in 5,989,134 autosomal variants and 334,391 variants on the X chromosome [32].

CBCS gene expression data

Paraffin-embedded tumor blocks were assayed for gene expression of 406 BC-related and 11 housekeeping genes using NanoString nCounter at the Translational Genomics Laboratory at UNC-Chapel Hill [4, 9]. As described previously, we eliminated samples with insufficient data quality using NanoStringQCPro [16, 33], scaled distributional difference between lanes with upper-quartile normalization [34], and removed two dimensions of unwanted technical and biological variation, estimated from housekeeping genes using RUVSeq [34, 35]. The current analysis included 1,199 samples with both genotype and gene expression data (628 BW, 571 WW).

Statistical analysis

Overview of TWAS

TWAS integrates expression data with GWAS to prioritize gene-trait associations through a two-step analysis (**Figure 1A-B**). First, using genetic and transcriptomic data, we trained predictive models of tumor gene expression using all SNPs within 0.5 Megabase of the gene [16, 18]. Second, we used these models to impute expression into an external GWAS panel to generate the Genetically-Regulated tumor eXpression (GReX) of a gene. This quantity represents the portion of tumor expression explained by *cis*-genetic regulation and is used to test for gene-trait associations with an outcome. By focusing on genetically regulated expression, TWAS avoids instances of expression-trait association that are not consequences of genetic variation but are driven by the effect of traits on expression. If sufficiently

heritable genes are assayed in the correct tissue, TWAS increases power to detect gene-trait associations and aids interpretability of results, as associations are mapped to individual genes [18, 36].

CRS TWAS in CBCS

We adopted techniques from FUSION to train predictive models of tumor expression from *cis*-germline genotypes, as discussed previously [16, 18]. Motivated by strong associations between germline genetics and tumor expression in CBCS [16], for genes with non-zero *cis*-heritability at nominal $P < 0.10$, we trained predictive models for covariate-residualized tumor expression with all *cis*-SNPs within 0.5 Megabase using linear mixed modeling or elastic net regression (**Supplementary Methods, Supplementary Materials**). We selected models with five-fold cross-validation adjusted $R^2 > 0.01$ between predicted and observed expression values, resulting in 59 and 45 models for WW and BW, respectively (**Supplementary Data**). Using only germline genetics as an input, we imputed GReX in 1,043 WW and 1,083 BW, respectively, in CBCS; for samples in both the training and imputation samples, GReX was imputed via cross-validation to minimize data leakage. We tested GReX for associations with ROR-S, Proliferation Score, and ROR-P using multiple linear regression adjusted for age, estrogen receptor (ER) status, tumor stage, and study phase [1]. We corrected for test-statistic bias and inflation using *bacon* and adjusted for multiple testing using the Benjamini-Hochberg procedure [37, 38]. To compare germline effects with total (germline and post-transcriptional) effects on ROR, we assessed relationships between tumor expression of TWAS genes and CRS using similar linear models. We were underpowered to study time-to-recurrence due to small sample size, as recurrence data was collected only in CBCS Phase 3 (635 WW, 742 BW with GReX and recurrence data; 183 WW, 283 BW with tumor expression and recurrence data).

PAM50 assay and ROR-S, Proliferation score, and ROR-P calculation

Using partition-around-medoid clustering, we calculated correlation with each subtype's centroid for study individuals based on PAM50 expressions (10 PAM50 genes per subtype); the largest subtype-centroid correlation defined the individual's molecular subtype [1]. ROR-S was determined via linear combination of the PAM50 subtype-centroid correlations (SCCs) [1]. Proliferation score was computed

using log-scale expression of 11 PAM50 genes while ROR-P was computed by combining ROR-S and Proliferation score.

Bayesian multivariate regressions and multivariate adaptive shrinkage

To better understand germline *trans*-regulation of PAM50 tumor gene expression and germline contribution to subtype, and to understand how these mediate TWAS-gene and CRS associations, we assessed TWAS-genes (for ROR-P) in relation to SCCs and PAM50 tumor gene expressions (**Figure 1C**). We found that none of our TWAS-genes were within 1 Megabase of PAM50 genes and that most TWAS-genes were not on the same chromosome as PAM50 genes (**Supplementary Table S1**). Existing gene-based mapping techniques for *trans*-expression quantitative trait loci (eQTL) (SNP and gene are separated by more than 1 Megabase) mapping include *trans*-PrediXcan and GBAT [39, 40]. We employed Bayesian multivariate linear regression (BtQTL) to account for correlation in multivariate outcomes (SCCs and PAM50 gene expression) in association testing. BtQTL improves power to detect significant *trans*-associations, especially when considering multiple genes with highly correlated (>0.5) expression (**Supplementary Methods**, **Supplementary Figures S1-S2, Supplementary Materials**). Lastly, we conducted adaptive shrinkage on BtQTL estimates using mashr, an empirical Bayes method to estimate patterns of similarity and improve accuracy in associations tests across multiple outcomes [41]. mashr outputs revised posterior means, standard deviations, and corresponding measures of significance (local false sign rates).

RESULTS

Association between GReX and risk of recurrence scores

We performed race-specific TWAS for CRS to investigate the role of germline genetic variation in CRS and CRS racial disparity. We identified 8 genes (*MCM10, FAM64A, CCNB2, MMP1, VAV3, PCSK6, NDC80, MLPH),* 8 genes (*MCM10, FAM64A, CCNB2, MMP1, VAV3, NDC80, MLPH, EXO1)*, and 10 genes (*MCM10, FAM64A, CCNB2, MMP1, VAV3, PCSK6, GNG11, NDC80, MLPH, EXO1)* whose GReX was associated with ROR-S, proliferation, and ROR-P, respectively, in WW, and 1 gene (*MMP1*) whose GReX was associated with proliferation and ROR-P in BW at FDR-adjusted $P < 0.10$ (**Figure 2A, 2B**). No

associations were detected between GReX and ROR-S among BW. We refer to genes with statistically significant TWAS associations (FDR-adjusted P < 0.10) as TWAS-genes. Among these identified genes, only genes that are not part of the PAM50 panel (i.e., excluding *NDC80, MLPH, EXO1*) were considered in downstream permutation and TWAS-gene follow up analyses (**Figure 1C**), as we wished to focus investigation on relationship between non-PAM50 TWAS-genes and PAM50 (tumor) genes.

Among WW, increased GReX of *MCM10*, *FAM64A*, *CCNB2*, and *MMP1* were associated with higher CRS while increased GReX of *VAV3*, *PCSK6*, and *GNG11* were associated with lower CRS (**Figure 2A**). Among BW, increased GReX of *MMP1* was associated with lower CRS (Proliferation, ROR-P, but not ROR-S) (**Figure 2A**). To provide statistical context for variance in CRS explained by significant TWAS-genes, we permuted covariate-residualized CRS to generate a null distribution for adjusted R^2 between TWAS-genes and CRS. Across WW and BW, the observed R^2 of TWAS-genes against CRS (7-10% among WW and 1% among BW) were statistically significant against the respective null distributions ($P < 0.001$ among WW and $P < 0.05$ among BW) (**Figure 2B**).

Associations between tumor expression of TWAS-genes and CRS were concordant, in terms of direction of association to germline-only effects among WW; findings were discordant among BW (**Supplementary Table S2-S3).** Permutation tests for analyses of tumor expression of TWAS-genes and CRS are available in **Supplementary Figure S3**.

Associations between TWAS-genes and breast cancer molecular subtype

Among WW, a one standard deviation increase in *FAM64A* and *CCNB2* GReX resulted in significantly increased Basal-like SCC while an identical increase in *VAV3*, *PCSK6*, and *GNG11* resulted in significantly increased Luminal A SCC. The magnitude of increase in correlation for respective subtypes per GReX gene was approximately 0.05, and most estimates had credible intervals that did not intersect the null**.** Among WW, associations between HER2-like SCC and GReX followed similar patterns to associations for the Basal-like subtype, although associations for HER2 were more precise (**Figure 3A**). We found predominantly null associations for GReX for Luminal B SCC among WW (**Figure 3A**). Unlike in WW, for BW, an increase in *MMP1* GReX was not associated with Luminal A, HER2 or Basal-

like SCCs. Instead, among BW, *MMP1* GReX was significantly negatively associated with Luminal B SCC. Estimates from univariate regressions are provided in **Supplementary Tables S4-S7.**

Association between TWAS-genes and PAM50 gene expression

For both WW and BW, the pattern of associations between significant GReX and PAM50 tumor expression were predominantly congruent with observed associations for SCCs and CRS (**Figure 4**). In WW, a one standard deviation increase in *CCNB2* GReX was associated with significantly increased *ORC6L*, *PTTG1*, and *KIF2C* (Basal-like genes) expression and *UBE2T, MYBL2* (LumB genes) expression. By contrast, a one standard deviation increase in *PCSK6* GReX significantly increased *BAG1*, *FOXA1*, *MAPT,* and *NAT1* (LumA genes) expression (**Figure 4**). While increased *MMP1* GReX was associated with significantly increased expression of *ORC6L* (basal-like gene), *MYBL2,* and *BIRC5* (LumB genes) among WW, this was not the case among BW. Instead, increased *MMP1* GReX among BW was significantly associated with increased expression of *SLC39A6* (LumA gene) and decreased expression of *ACTR3B, PTTG1*, and *EXO1* (Basal-like genes) (**Figure 4**). **Supplementary Tables S8-S11** and **Figure 4** provide all TWAS-gene and PAM50 gene expression associations across WW and BW.

DISCUSSION

Through TWAS, we identified 7 and 1 genes among WW and BW, respectively, for which GReX was associated with CRS and underlying PAM50 expressions and subtype. Among WW, these 7 TWAS-genes explained between 7-10% of the variation in CRS, a large and statistically significant proportion of variance. Among BW, the singular TWAS-gene explained ~1% of the variation in Proliferation score and ROR-P. Differences in the number and effect of identified TWAS-genes by race may point to factors that warrant further investigation: (1) potentially greater contribution of *trans*-regulation in tumor gene expression in BW, as shown previously, and (2) potential racial differences in tumor methylation and somatic alternations, which could not be accounted for in CBCS[16, 42-47].

There are two key novel aspects to this study. First, existing literature on associations between tumor gene expression and recurrence (for which CRS are a proxy) cannot distinguish between genetic

and non-genetic component of effects [48]. Second, TWAS allows causal interpretation of observed associations. For instance, prior studies report *CCNB2* is upregulated in triple-negative breast cancers (TNBC) but were unable to determine whether increased *CCNB2* expression contributes to development or maintenance of TNBC or is part of the molecular response to cancer progression [49, 50]. By contrast, GReX is a function of only genetic variation. Thus, TWAS allows causal interpretation, subject to effective control for population stratification and minimal horizontal pleiotropy [17, 18].

Our WW-specific finding that prioritizes *MCM10*, *FAM64A*, and *CCNB2* associations with Basal-like and HER2-enriched subtypes and subtype-specific gene expressions are consistent with literature. Prior investigations in cohorts of primarily European ancestry have reported that *MCM10*, *FAM64A*, and *CCNB2* expression is higher in ER-negative than ER-positive tumors [49-51]. In studies that compared triple-negative and non-triple negative subtypes, higher *MCM10*, *FAM64A*, and *CCNB2* expression was detected in triple-negative BC [49, 50]. Histologically, HER2-enriched and Basal-like subtypes are typically ER-negative, and triple-negatives are similar to Basal-like subtypes [9, 52]. *MCM10*, *FAM64A*, and *CCNB2* are all implicated in cell cycle processes, including DNA replication [51, 53, 54]. Our WW-specific findings that GReX of *PCSK6* and *VAV3* associated with Luminal A and Luminal A specific gene expressions are also consistent with previous results of *PCSK6* and *VAV3* upregulation in ER-positive subtypes [55, 56].

Presently, little is known about germline genetic regulation of PAM50 tumor expression. In CBCS, we found that tumor expression of most PAM50 genes is not *cis*-heritable. Instead, observed TWAS-gene and PAM50 gene expression associations may implicate *trans*-gene regulation of the PAM50 signature. For instance, we found that *VAV3* GReX is significantly positively associated with tumor expression of *BAG1*, *FOXA1*, *MAPT*, and *NAT1* and nominally with increased tumor *ESR1* expression, all of which are Luminal A-specific genes. Such *trans*-genic regulation signals, especially in the case of *ESR1*, pose significant clinical and therapeutic implication if confirmed under experimental conditions. For example, *VAV3* activates *RAC1* which upregulates *ESR1* but such mechanistic evidence is sparse for other putative TWAS-gene to PAM50 gene associations [57, 58]. More generally, two of the TWAS-genes among WW (*FAM64A*, *PCSK6*) have been found to activate the oncogenic *STAT3* signaling pathway, housing many purported anti-cancer drug targets [59, 60].

Interestingly, we found *MMP1* GReX has divergent associations with ROR across race. There are a few potential explanations. First, the range of *MMP1* GReX was manifold among WW than BW, suggesting sparser *cis*-eQTL architecture of *MMP1* in BW and more influence from *trans*-acting signals. Potential differences in influence of germline genetics on tumor expression and ROR by race could be an artifact of divergent somatic or epigenetic factors that CBCS has not assayed [44-47]. Second, while studies generally report that *MMP1* tumor expression is higher in triple-negative and Basal-like breast cancer, one study reported that *MMP1* expression in tumor cells does not significantly differ by subtype [61-63]. Instead, Bostrom *et al*. reported that *MMP1* expression differs in stromal cells of patients with different subtypes [63]. There is evidence to suggest that tumor composition, including stromal and immune components, may influence BC progression in a subtype-specific manner and future studies should consider expression predictive models that integrate greater detail on tumor cell-type composition [64, 65].

There are a few limitations to this study. First, as CBCS used a custom Nanostring nCounter probeset for mRNA expression quantification, we could not analyze the whole human transcriptome. While this probeset may exclude several *cis*-heritable genes, CBCS contains one of the largest breast tumor transcriptomic datasets for Black women, allowing us to build well-powered race-specific predictive models, a pivotal step in transethnic TWAS. Second, CBCS lacked data on somatic amplifications and deletions, inclusion of which could enhance the performance of predictive models [66]. Third, as recurrence data was collected in a small subset with few recurrence events, we were unable to make a direct comparison between CRS and recurrence results, which may affect clinical generalizability. However, to our knowledge, CBCS is the largest resource of PAM50-based CRS data.

Our analysis provides evidence of putative CRS and germline variation associations in breast tumors across race, motivating larger diverse cohorts for genetic epidemiology studies of breast cancer. Future studies should consider subtype-specific TWAS (i.e., stratification by subtype in predictive model training and association analyses) to elucidate heritable gene expression effects on breast cancer outcomes both across and within subtype, which may yield further hypotheses for more fine-tuned clinical intervention.

FUNDING

This work was supported by Susan G. Komen® for the Cure for CBCS study infrastructure. Funding was provided by the National Institutes of Health, National Cancer Institute P01-CA151135, P50-CA05822, and U01-CA179715 to AFO, CMP, and MAT. AP is supported by T32ES007018. MIL is supported by R01-HG009937, R01-MH118349, P01-CA142538, and P30-ES010126. The Translational Genomics Laboratory is supported in part by grants from the National Cancer Institute (3P30CA016086) and the University of North Carolina at Chapel Hill University Cancer Research Fund. Genotyping was done at the DCEG Cancer Genomics Research Laboratory using funds from the NCI Intramural Research Program.

Funding for BCAC and iCOGS came from: Cancer Research UK [grant numbers C1287/A16563, C1287/A10118, C1287/A10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565], the European Union's Horizon 2020 Research and Innovation Programme (grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), the European Community's Seventh Framework Programme under grant agreement n° 223175 [HEALTHF2-2009-223175] (COGS), the National Institutes of Health [CA128978] and Post-Cancer GWAS initiative [1U19 CA148537, 1U19 CA148065-01 (DRIVE) and 1U19 CA148112 - the GAME-ON initiative], the Department of Defence [W81XWH-10-1-0341], and the Canadian Institutes of Health Research CIHR) for the CIHR Team in Familial Risks of Breast Cancer [grant PSR-SIIRI-701]. All studies and funders as listed in Michailidou K *et al* (2013 and 2015) and in Guo Q et al (2015) are acknowledged for their contributions.

NOTES

Affiliations of authors: Department of Epidemiology, Gillings School of Global Public Health (AP, AFO, MAT), Department of Biostatistics (MIL), Department of Genetics (MIL, CMP), Department of Pathology and Laboratory Medicine (MAT, CMP), Lineberger Comprehensive Cancer Center (AFO, CMP), University of North Carolina at Chapel Hill; Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, University of California, Los Angles (AB); Division of Cancer Epidemiology and Genetics, National Cancer Institute (MG); Division of Genetics and Epidemiology, Institute of Cancer Research (MG)

Prior presentation: This work has been presented in poster sessions at the 2020 American Association for Cancer Research: The Science of Cancer Health Disparities and the 2020 Harvard Population Quantitative Genetics conferences.

Disclaimers: CMP is an equity stock holder, consultant, and board of directors member of BioClassifier LLC and GeneCentric Diagnostics. CMP is also listed as an inventor on patent applications on the Breast PAM50 assay.

Role of the funder: This content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funder had no role in study design, data collection, analysis or interpretation, or writing of the manuscript.

Acknowledgements: We thank the Carolina Breast Cancer Study participants and volunteers. We also thank Colin Begg, Jianwen Cai, Katherine Hoadley, Yun Li, and Bogdan Pasaniuc for valuable discussion during the research process. We thank Erin Kirk and Jessica Tse for their invaluable support during the research process. We thank the DCEG Cancer Genomics Research Laboratory and acknowledge the support from Stephen Chanock, Rose Yang, Meredith Yeager, Belynda Hicks, and Bin Zhu. We also acknowledge the iCOGs Consortium for their publicly available GWAS summary statistics.

AVAILABILITY OF DATA AND MATERIALS

Expression data from CBCS is available on NCBI GEO with accession number GSE148426. CBCS genotype datasets analyzed in this study are not publicly available as many CBCS patients are still being followed and accordingly CBCS data is considered sensitive; the data is available from M.A.T upon reasonable request. Supplementary Data includes summary statistics for eQTL results, tumor expression models, and relevant R code for training expression models in CBCS and are freely available at https://github.com/bhattacharya-a-bt/CBCS_TWAS_Paper/. iCOGs summary statistics are available online at http://bcac.ccge.medschl.cam.ac.uk/bcacdata/icogs-complete-summary-results.

REFERENCES

1. Parker JS, Mullins M, Cheang MC*, et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol 2009;27(8):1160-7.
2. Wallden B, Storhoff J, Nielsen T*, et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med Genomics 2015;8:54.
3. Paik S, Shak S, Tang G*, et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351(27):2817-26.
4. Geiss GK, Bumgarner RE, Birditt B*, et al.* Direct multiplexed measurement of gene expression with color-coded probe pairs. Nat Biotechnol 2008;26(3):317-25.
5. Harris LN, Ismaila N, McShane LM*, et al.* Use of Biomarkers to Guide Decisions on Adjuvant Systemic Therapy for Women With Early-Stage Invasive Breast Cancer: American Society of Clinical Oncology Clinical Practice Guideline. Journal of Clinical Oncology 2016;34(10):1134-1150.
6. Coates AS, Winer EP, Goldhirsch A*, et al.* Tailoring therapies--improving the management of early breast cancer: St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2015. Ann Oncol 2015;26(8):1533-46.
7. Dowsett M, Sestak I, Lopez-Knowles E*, et al.* Comparison of PAM50 Risk of Recurrence Score With Oncotype DX and IHC4 for Predicting Risk of Distant Recurrence After Endocrine Therapy. Journal of Clinical Oncology 2013;31(22):2783-2790.
8. Sestak I, Buus R, Cuzick J*, et al.* Comparison of the Performance of 6 Prognostic Signatures for Estrogen Receptor-Positive Breast Cancer: A Secondary Analysis of a Randomized Clinical Trial. JAMA Oncol 2018;4(4):545-553.
9. Troester MA, Sun X, Allott EH*, et al.* Racial Differences in PAM50 Subtypes in the Carolina Breast Cancer Study. J Natl Cancer Inst 2018;110(2):176-82.
10. Albain KS, Gray RJ, Makower DF*, et al.* Race, ethnicity and clinical outcomes in hormone receptor-positive, HER2-negative, node-negative breast cancer in the randomized TAILORx trial. J Natl Cancer Inst 2020; 10.1093/jnci/djaa148.
11. Reeder-Hayes KE, Anderson BO. Breast Cancer Disparities at Home and Abroad: A Review of the Challenges and Opportunities for System-Level Change. Clin Cancer Res 2017;23(11):2655-2664.

12. Durham DD, Robinson WR, Lee SS*, et al.* Insurance-Based Differences in Time to Diagnostic Follow-up after Positive Screening Mammography. Cancer Epidemiol Biomarkers Prev 2016;25(11):1474-1482.

13. Wheeler SB, Reeder-Hayes KE, Carey LA. Disparities in breast cancer treatment and outcomes: biological, social, and health system determinants and opportunities for research. Oncologist 2013;18(9):986-93.

14. Ko NY, Hong S, Winn RA*, et al.* Association of Insurance Status and Racial Disparities With the Detection of Early-Stage Breast Cancer. JAMA Oncology 2020;6(3):385-392.

15. Huo D, Hu H, Rhie SK*, et al.* Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. JAMA Oncol 2017;3(12):1654-1662.

16. Bhattacharya A, García-Closas M, Olshan AF*, et al.* A framework for transcriptome-wide association studies in breast cancer in diverse study populations. Genome Biol 2020;21(1):42.

17. Gamazon ER, Wheeler HE, Shah KP*, et al.* A gene-based association method for mapping traits using reference transcriptome data. Nat Genet 2015;47(9):1091-8.

18. Gusev A, Ko A, Shi H*, et al.* Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet 2016;48(3):245-52.

19. Zhong J, Jermusyk A, Wu L*, et al.* A Transcriptome-Wide Association Study Identifies Novel Candidate Susceptibility Genes for Pancreatic Cancer. J Natl Cancer Inst 2020;112(10):1003-1012.

20. Wu L, Shi W, Long J*, et al.* A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. Nat Genet 2018;50(7):968-978.

21. Mancuso N, Gayther S, Gusev A*, et al.* Large-scale transcriptome-wide association study identifies new prostate cancer risk regions. Nat Commun 2018;9(1):4079.

22. Keys KL, Mak ACY, White MJ*, et al.* On the cross-population generalizability of gene expression prediction models. PLoS Genet 2020;16(8):e1008927.

23. Hair BY, Hayes S, Tse CK*, et al.* Racial differences in physical activity among breast cancer survivors: implications for breast cancer care. Cancer 2014;120(14):2174-82.

24. Newman B, Moorman PG, Millikan R*, et al.* The Carolina Breast Cancer Study: integrating population-based epidemiology and molecular biology. Breast Cancer Res Treat 1995;35(1):51-60.

25. Amos CI, Dennis J, Wang Z*, et al.* The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev 2017;26(1):126-135.

26. Auton A, Brooks LD, Durbin RM*, et al.* A global reference for human genetic variation. Nature 2015;526(7571):68-74.

27. O'Connell J, Gurdasani D, Delaneau O*, et al.* A general approach for haplotype phasing across the full spectrum of relatedness. PLoS Genet 2014;10(4):e1004234.

28. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. Nat Methods 2011;9(2):179-81.

29. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009;5(6):e1000529.

30. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet 2005;76(5):887-93.

31. Purcell S, Neale B, Todd-Brown K*, et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81(3):559-75.

32. Sherry ST, Ward MH, Kholodov M*, et al.* dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29(1):308-11.

33. Bhattacharya A, Hamilton AM, Furberg H*, et al.* An approach for normalization and quality control for NanoString RNA expression data. Brief Bioinform 2020; 10.1093/bib/bbaa163.

34. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11(10):R106.

35. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550.

36. Ding B, Cao C, Li Q*, et al.* Power analysis of transcriptome-wide association study. bioRxiv 2020; 10.1101/2020.07.19.211151:2020.07.19.211151.

37. van Iterson M, van Zwet EW, Heijmans BT. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. Genome Biol 2017;18(1):19.

38. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) 1995;57(1):289-300.

39. Wheeler HE, Ploch S, Barbeira AN*, et al.* Imputed gene associations identify replicable trans-acting genes enriched in transcription pathways and complex traits. Genetic Epidemiology 2019;43(6):596-608.

40. Liu X, Mefford JA, Dahl A*, et al.* GBAT: a gene-based association test for robust detection of trans-gene regulation. Genome Biology 2020;21(1):211.

41. Urbut SM, Wang G, Carbonetto P*, et al.* Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. Nat Genet 2019;51(1):187-195.

42. Gravel S. Population genetics models of local ancestry. Genetics 2012;191(2):607-19.

43. Nelson D, Kelleher J, Ragsdale AP*, et al.* Accounting for long-range correlations in genome-wide simulations of large cohorts. PLoS Genet 2020;16(5):e1008619.

44. Shang L, Smith JA, Zhao W*, et al.* Genetic Architecture of Gene Expression in European and African Americans: An eQTL Mapping Study in GENOA. Am J Hum Genet 2020;106(4):496-512.

45. Wang S, Dorsey TH, Terunuma A*, et al.* Relationship between tumor DNA methylation status and patient characteristics in African-American and European-American women with breast cancer. PLoS One 2012;7(5):e37928.

46. Conway K, Edmiston SN, Tse CK*, et al.* Racial variation in breast tumor promoter methylation in the Carolina Breast Cancer Study. Cancer Epidemiol Biomarkers Prev 2015;24(6):921-30.

47. Chen Y, Sadasivan SM, She R*, et al.* Breast and prostate cancers harbor common somatic copy number alterations that consistently differ by race and are associated with survival. BMC Med Genomics 2020;13(1):116.

48. Parada H, Jr., Sun X, Fleming JM*, et al.* Race-associated biological differences among luminal A and basal-like breast cancers in the Carolina Breast Cancer Study. Breast Cancer Res 2017;19(1):131.

49. Prat A, Adamo B, Cheang MC*, et al.* Molecular characterization of basal-like and non-basal-like triple-negative breast cancer. Oncologist 2013;18(2):123-33.

50. Zhang C, Han Y, Huang H*, et al.* Integrated analysis of expression profiling data identifies three genes in correlation with poor prognosis of triple-negative breast cancer. Int J Oncol 2014;44(6):2025-33.

51. Mahadevappa R, Neves H, Yuen SM*, et al.* DNA Replication Licensing Protein MCM10 Promotes Tumor Progression and Is a Novel Prognostic Biomarker and Potential Therapeutic Target in Breast Cancer. Cancers (Basel) 2018;10(9).

52. Hagemann IS. Molecular Testing in Breast Cancer: A Guide to Current Practices. Arch Pathol Lab Med 2016;140(8):815-24.

53. Yao Z, Zheng X, Lu S*, et al.* Knockdown of FAM64A suppresses proliferation and migration of breast cancer cells. Breast Cancer 2019;26(6):835-845.

54. Gong D, Ferrell JE, Jr. The roles of cyclin A2, B1, and B2 in early and late mitotic events. Mol Biol Cell 2010;21(18):3149-61.

55. Thakkar AD, Raj H, Chakrabarti D*, et al.* Identification of gene expression signature in estrogen receptor positive breast carcinoma. Biomark Cancer 2010;2:1-15.

56. Aguilar H, Urruticoechea A, Halonen P*, et al.* VAV3 mediates resistance to breast cancer endocrine therapy. Breast Cancer Res 2014;16(3):R53.

57. Zeng L, Sachdev P, Yan L*, et al.* Vav3 mediates receptor protein tyrosine kinase signaling, regulates GTPase activity, modulates cell morphology, and induces cell transformation. Mol Cell Biol 2000;20(24):9212-24.

58. Rosenblatt AE, Garcia MI, Lyons L*, et al.* Inhibition of the Rho GTPase, Rac1, decreases estrogen receptor levels and is a novel therapeutic strategy in breast cancer. Endocr Relat Cancer 2011;18(2):207-19.

59. Xu Z-S, Zhang H-X, Li W-W*, et al.* FAM64A positively regulates STAT3 activity to promote Th17 differentiation and colitis-associated carcinogenesis. Proceedings of the National Academy of Sciences 2019;116(21):10447-10452.

60. Jiang H, Wang L, Wang F*, et al.* Proprotein convertase subtilisin/kexin type 6 promotes in vitro proliferation, migration and inflammatory cytokine secretion of synovial fibroblast⍰like cells from rheumatoid arthritis via nuclear⍰κB, signal transducer and activator of transcription 3 and extracellular signal regulated 1/2 pathways. Mol Med Rep 2017;16(6):8477-8484.

61. Wang QM, Lv L, Tang Y*, et al.* MMP-1 is overexpressed in triple-negative breast cancer tissues and the knockdown of MMP-1 expression inhibits tumor cell malignant behaviors in vitro. Oncol Lett 2019;17(2):1732-1740.

62. McGowan PM, Duffy MJ. Matrix metalloproteinase expression and outcome in patients with breast cancer: analysis of a published database. Ann Oncol 2008;19(9):1566-72.

63. Boström P, Söderström M, Vahlberg T*, et al.* MMP-1 expression has an independent prognostic value in breast cancer. BMC Cancer 2011;11:348.

64. Acerbi I, Cassereau L, Dean I*, et al.* Human breast cancer invasion and aggression correlates with ECM stiffening and immune cell infiltration. Integr Biol (Camb) 2015;7(10):1120-34.

65. González LO, Corte MD, Junquera S*, et al.* Expression and prognostic significance of metalloproteases and their inhibitors in luminal A and basal-like phenotypes of breast carcinoma. Hum Pathol 2009;40(9):1224-33.

66. Xia Y, Fan C, Hoadley KA*, et al.* Genetic determinants of the molecular portraits of epithelial cancers. Nat Commun 2019;10(1):5666.

FIGURE LEGENDS

Figure 1. *Schematic of study analytic approach.* A) In CBCS, constructed race-stratified predictive models of tumor gene expression from *cis*-SNPs. B) In CBCS, imputed GReX at individual-level using genotypes and tested for associations between GReX and CRS in race-stratified linear models; only GReX of genes with significant *cis*-h^2 and high cross validation performance ($R^2 > 0.01$ between observed and predicted expression) considered for race-stratified association analyses. C) Follow-up analyses on TWAS-genes (i.e., genes whose GReX were significantly associated with CRS at FDR <0.10). In race-stratified models, PAM50 SCCs and PAM50 tumor expressions were regressed against TWAS-genes under a Bayesian multivariate regression and multivariate adaptive shrinkage approach.

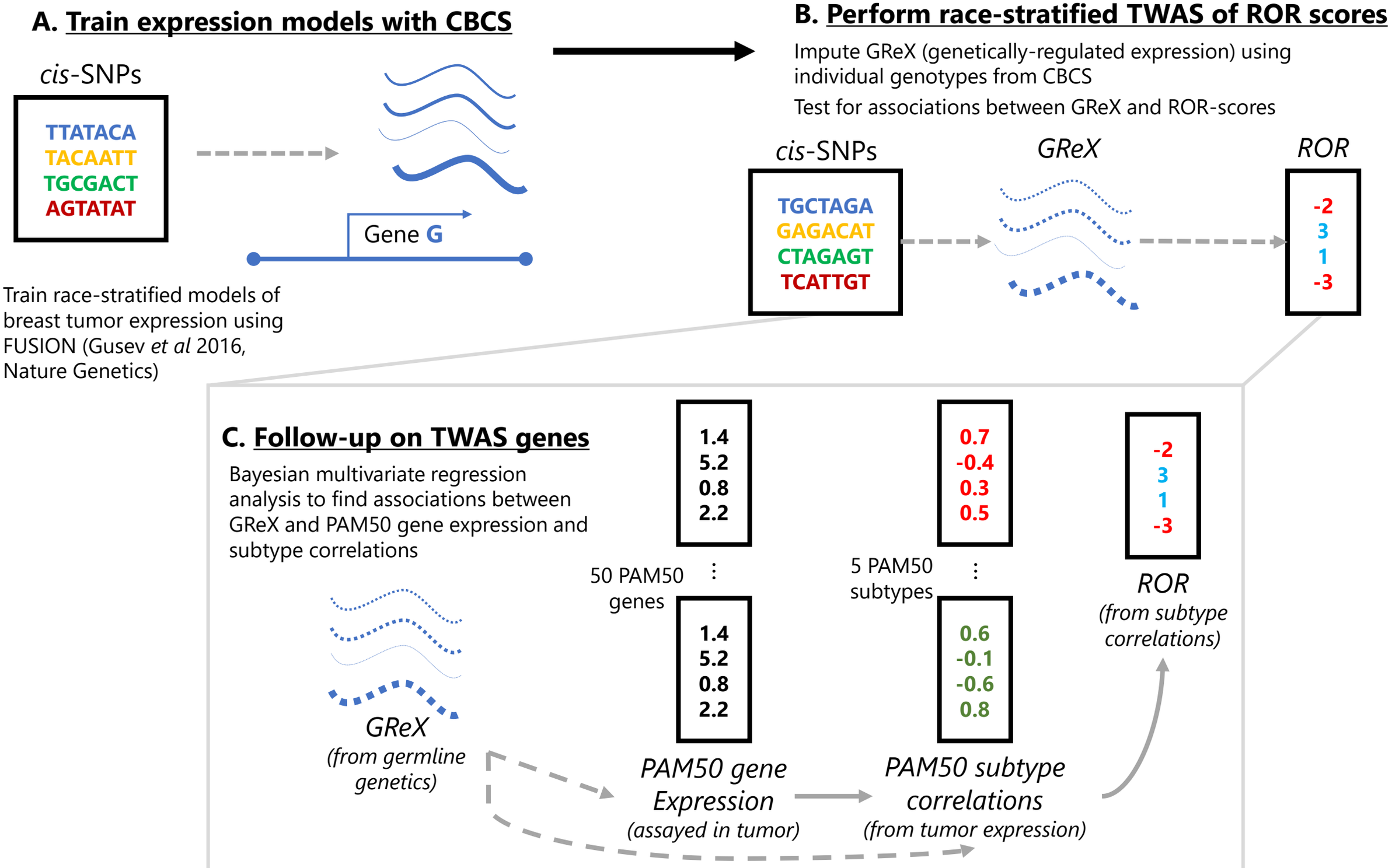
Figure 2. *Permutation tests and associations between TWAS-genes and CRS for WW and BW.* A) Effect estimates correspond to change in ROR-S, Proliferation score, and ROR-P per one standard deviation increase in TWAS-gene expression (i.e., one standard deviation increase in GReX of gene). Circle

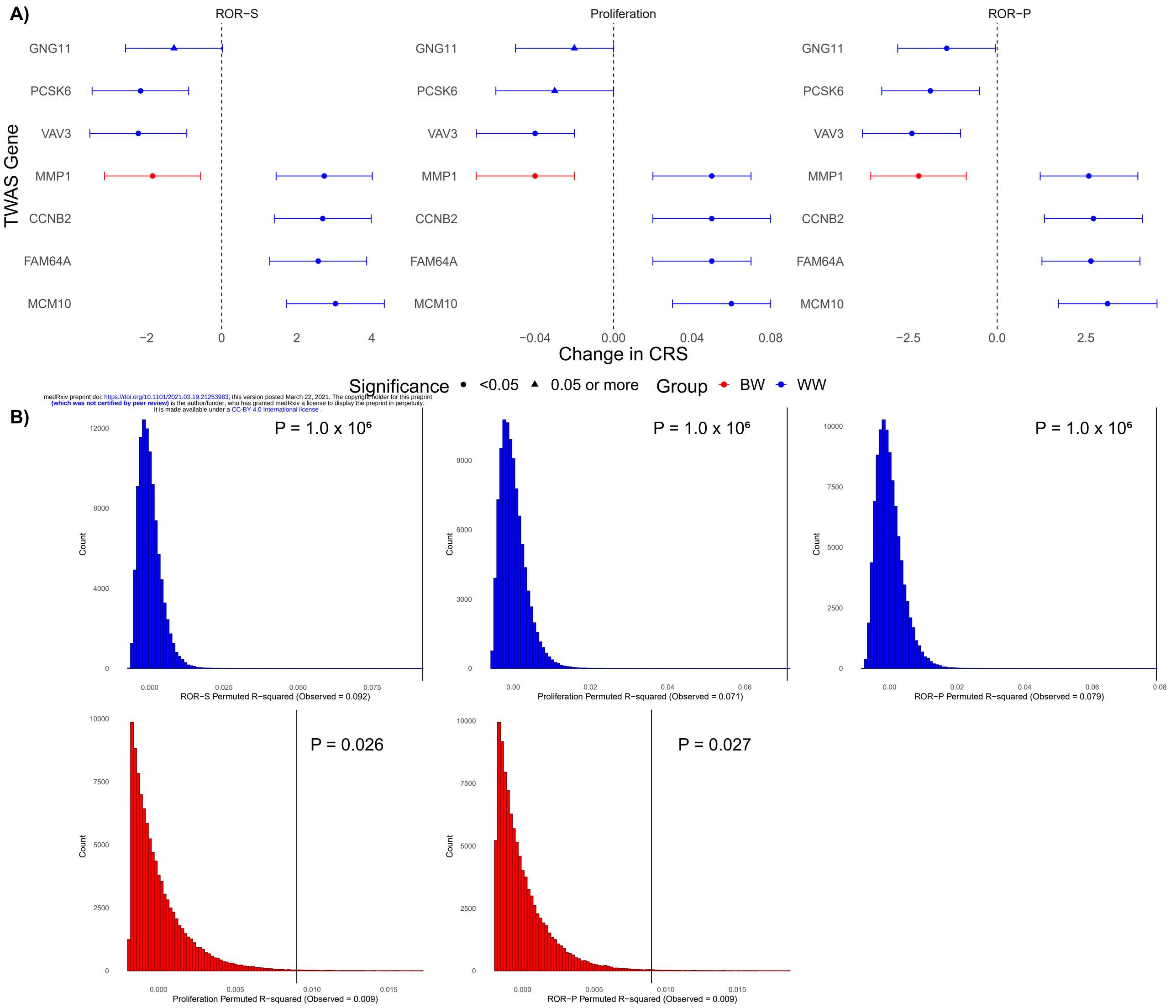
denotes a statistically significant association while triangle denotes a non-significant association at significance threshold of *p*-value <0.05. Blue denotes WW and red denotes BW. B) Histograms correspond to null distributions of covariates (age at selection, estrogen receptor status, study phase, tumor stage) residualized-R^2 for regressions of CRS on TWAS-genes. Dashed vertical lines correspond to observed covariates residualized-R^2. Blue denotes WW and red denotes BW.

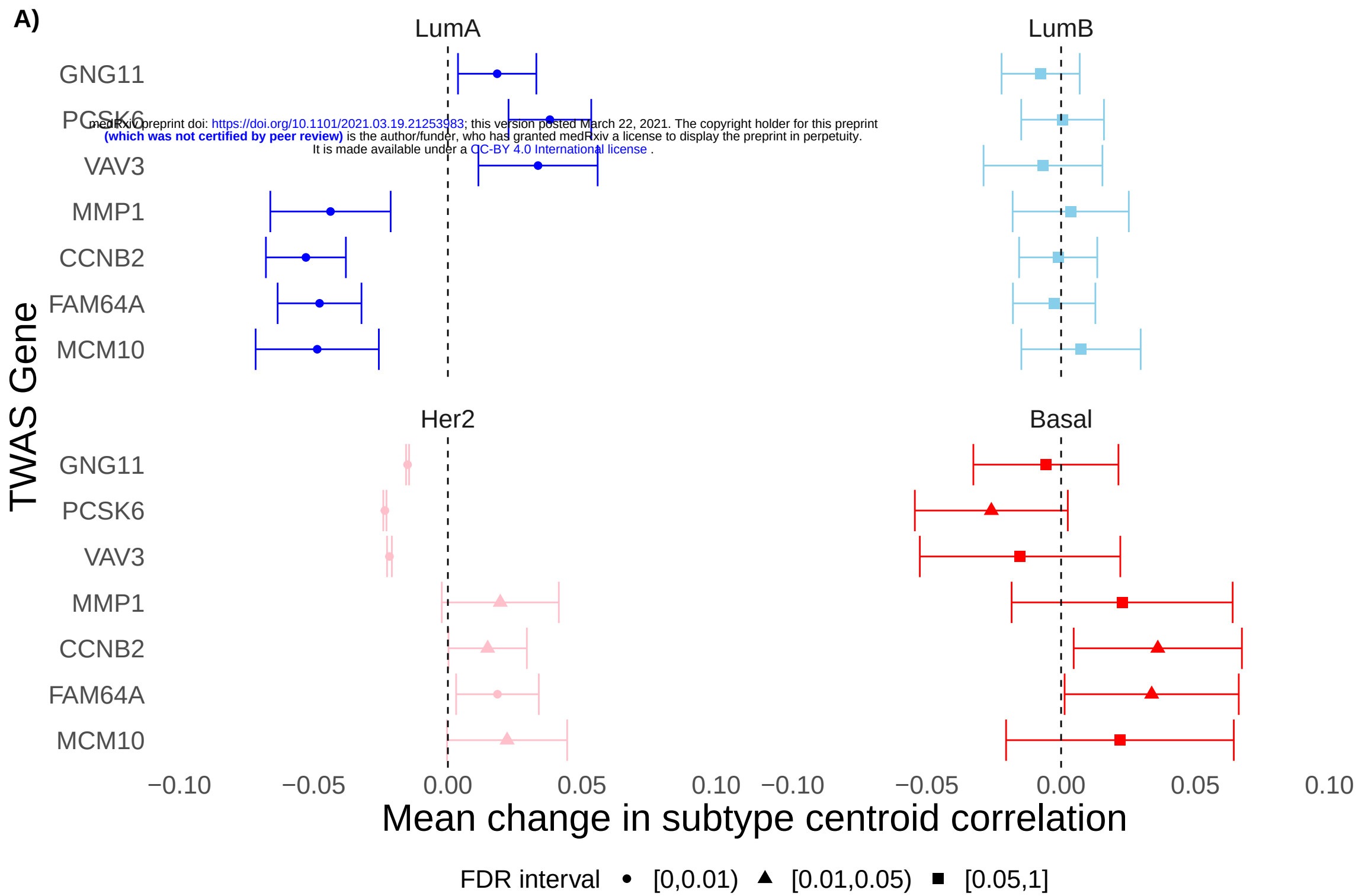
Figure 3. *Associations between TWAS-genes and PAM50 SCCs.* A) Among WW, associations between TWAS-genes (genes whose GReX was significantly associated with CRS at FDR <0.10) and PAM50 SCCs using Bayesian multivariate regression and multivariate adaptive shrinkage. Effect estimates correspond to change in subtype centroid correlations (range -1 to 1) for one standard deviation increase in TWAS-gene expression (i.e., one standard deviation increase in GReX of gene). Circle, triangle, and square denote corresponding FDR intervals for effect sizes. B) Among BW, associations between TWAS-genes and PAM50 SCCs using Bayesian multivariate regression and multivariate adaptive shrinkage. Effect estimates correspond to change in SCCs (range -1 to 1) for one standard deviation increase in TWAS-gene expression (i.e., one standard deviation increase in GReX of gene). Circle, triangle, and square denote corresponding FDR intervals for effect sizes.

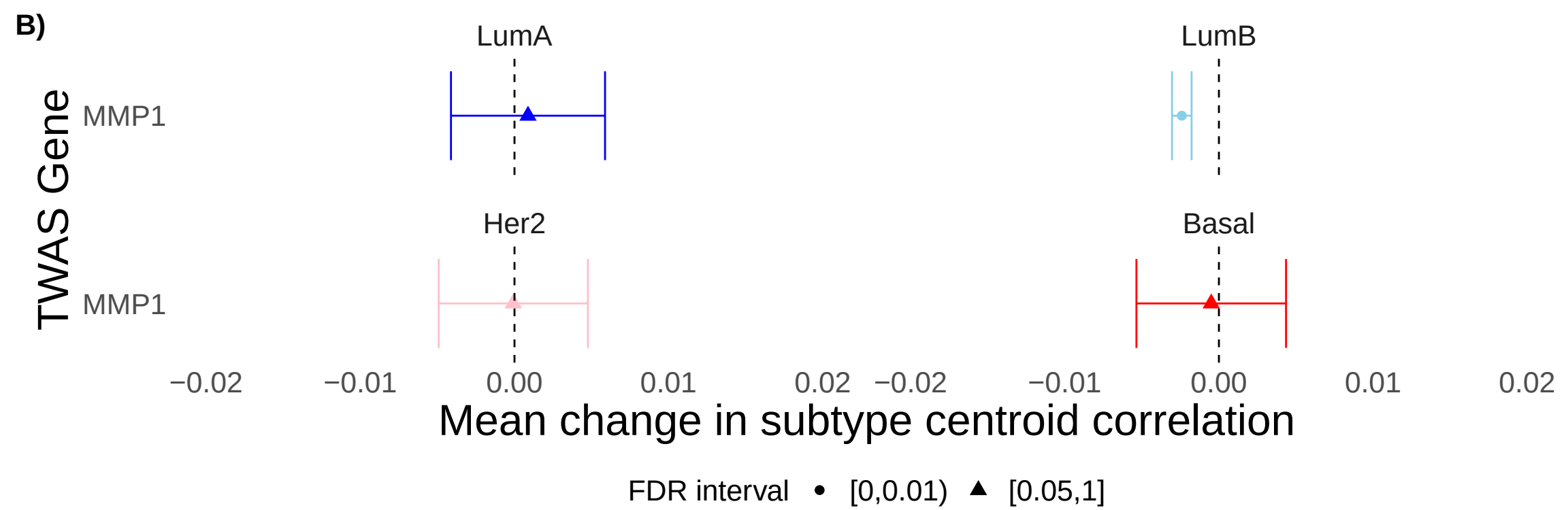
Figure 4. *Heatmap of associations between TWAS-genes and PAM50 tumor gene expressions using Bayesian multivariate regression and multivariate adaptive shrinkage.* There were 7 TWAS-genes among WW and 1 TWAS-gene among BW. Effect estimates correspond to change in $\log_2$ normalized PAM50 tumor expression for one standard deviation increase in TWAS-gene expression (i.e., one standard deviation increase in GReX of gene). Red denotes positive change in $\log_2$ normalized tumor expression and blue denotes negative mean change in $\log_2$ normalized tumor expression. *, **, *** denote FDR intervals for effect sizes. Assignment of PAM50 gene to subtype was based on PAM50 gene centroid values for each subtype; the subtype assigned to a PAM50 gene corresponded to the largest positive centroid value across subtypes for that gene. Importantly, subtype assignment through this “greedy algorithm” are specific to this study and represent a simplified reality (e.g., ESR1 classified as part of Luminal A subtype only even though ESR1 expression correlates with both Luminal A and to a slightly

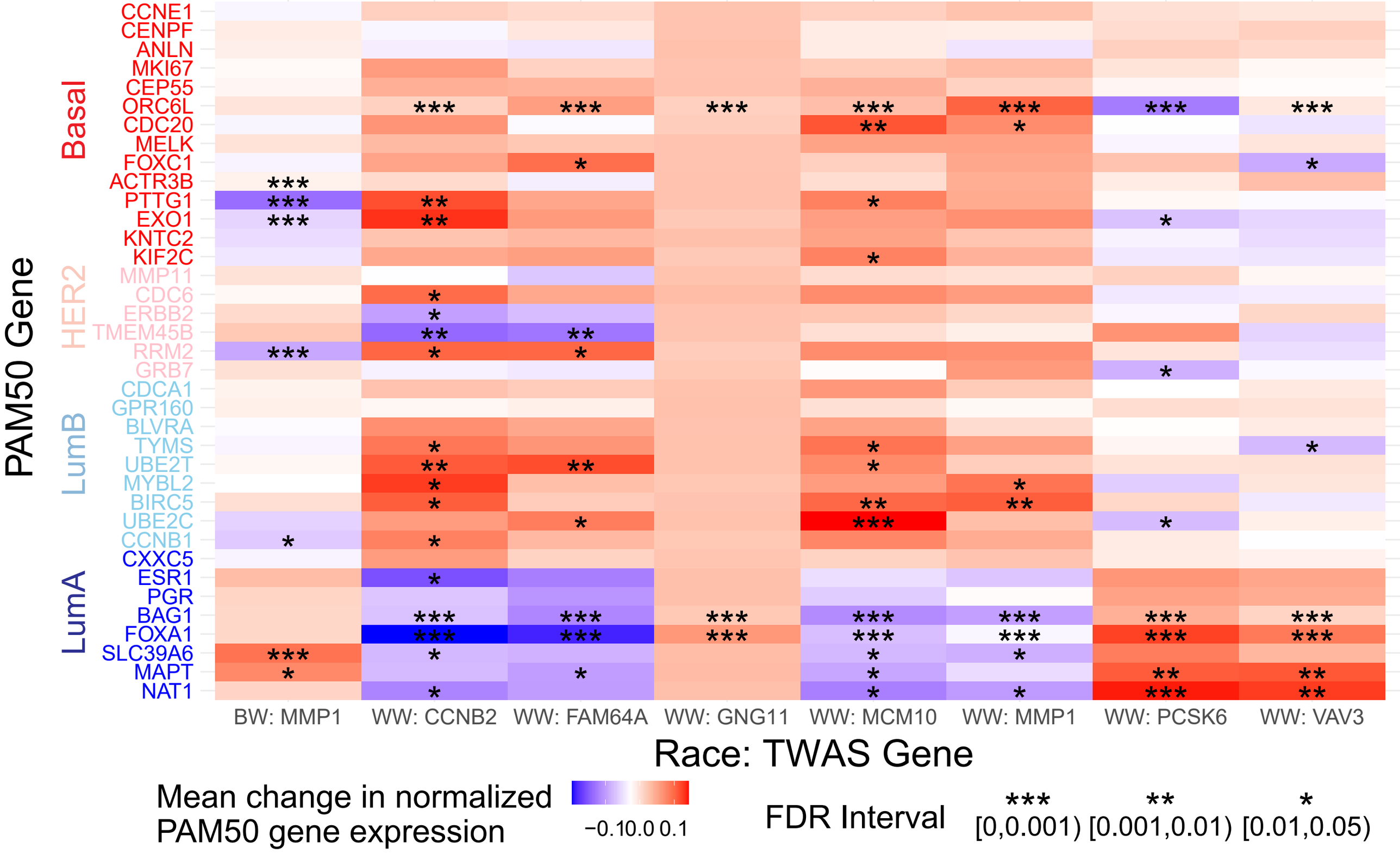
lesser degree Luminal B subtype). Moreover, subtype assignment for this portion of analyses was conducted only for visual comparison of patterns of associations between TWAS-genes and PAM50 tumor gene expressions (i.e., subtype assignment in this portion of analyses had no bearing on continuous ROR score calculations or subtype-centroid correlations).


A. Train expression models with CBCS
cis-SNPs
TTATACA
TACAATT
TGCGACT
AGTATAT
Gene G
Train race-stratified models of breast tumor expression using FUSION (Gusev et al 2016, Nature Genetics)
B. Perform race-stratified TWAS of ROR scores
Impute GReX (genetically-regulated expression) using individual genotypes from CBCS
Test for associations between GReX and ROR-scores
cis-SNPs
TGCTAGA
GAGACAT
CTAGAGT
TCATTGT
GReX
ROR
-2
3
1
-3
C. Follow-up on TWAS genes
Bayesian multivariate regression analysis to find associations between GReX and PAM50 gene expression and subtype correlations
1.4
5.2
0.8
2.2
50 PAM50 genes
1.4
5.2
0.8
2.2
0.7
-0.4
0.3
0.5
5 PAM50 subtypes
0.6
-0.1
-0.6
0.8
-2
3
1
-3
ROR
(from subtype correlations)
GReX
(from germline genetics)
PAM50 gene Expression
(assayed in tumor)
PAM50 subtype correlations
(from tumor expression)








PAM50 Gene
Basal
HER2
LumB
LumA
CCNE1
CENPF
ANLN
MKI67
CEP55
ORC6L
CDC20
MELK
FOXC1
ACTR3B
PTTG1
EXO1
KNTC2
KIF2C
MMP11
CDC6
ERBB2
TMEM45B
RRM2
GRB7
CDCA1
GPR160
BLVRA
TYMS
UBE2T
MYBL2
BIRC5
UBE2C
CCNB1
CXXC5
ESR1
PGR
BAG1
FOXA1
SLC39A6
MAPT
NAT1
BW: MMP1
WW: CCNB2
WW: FAM64A
WW: GNG11
WW: MCM10
WW: MMP1
WW: PCSK6
WW: VAV3
Race: TWAS Gene
Mean change in normalized PAM50 gene expression
−0.1 0.0 0.1
FDR Interval
*** [0,0.001)
** [0.001,0.01)
* [0.01,0.05)