GENE-LEVEL GERMLINE INVESTIGATION OF BREAST CANCER SUBTYPES AND MORTALITY AMONG WOMEN OF EUROPEAN AND AFRICAN ANCESTRY

Achal Pareshkumar Patel

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements of the Doctor of Philosophy in the Department of Epidemiology in the Gillings School of Global Public Health.

Chapel Hill
2022

Approved by:

Melissa Troester

Michael Love

Charles Perou

Kari North

Hazel Nichols

© 2022
Achal Pareshkumar Patel
ALL RIGHTS RESERVED

ABSTRACT

Achal Pareshkumar Patel: Gene-Level Germline Investigation of Breast Cancer Subtypes and Mortality among Women of European and African Ancestry
(Under the direction of Melissa Troester and Michael Love)

Genome-wide investigations to date have uncovered over 210 germline variants associated with BC incidence, and only a handful of mostly non-replicable variants associated with BC mortality. There remain a few key gaps in knowledge across genetic investigations of BC incidence and mortality. First, BC is a heterogeneous disease, with subtype-specific outcomes. Understanding of the germline genetic underpinnings of BC subtypes is sparse among individuals of European and even sparser for individuals of African ancestry. It is also thought that integration of BC subtype (i.e., stratification by) into germline investigations of BC mortality may a key step towards bettering understanding on this front. In this dissertation work, we address these gaps in knowledge by leveraging a statistically efficient and highly interpretable framework for genetic association testing in Transcriptome-Wide Association Study (TWAS).

In Aim 1 (Chapter 4) multi-tissue (normal breast, breast tumor), multi-ancestry TWAS-based germline-regulated gene expression (GReX) analysis of 396 BC-related genes in relation to BC subtype, we find 40 GReX-prioritized genes for BC subtype, including ten shared, and 34 and six unique to Luminal-like (LL) and Basal-like (BL) subtype, respectively (after Bayesian correction for potential test statistic inflation and at global False Discovery Rate (FDR) <0.05). Among individuals of African ancestry, we see suggestion of association (global FDR = 0.06 to 0.18) for five genes. By comparison, the largest genome-wide study to date among African ancestry individuals has reported no loci at or near genome-wide significance.

In Aim 2 (Chapter 5) ancestry and subtype specific GReX analysis of BC mortality, we find no associations at global FDR <0.05. However, we uncover potential differential germline-regulation of tumor expression across LL and BL subtypes, within each group of European and African ancestry individuals (321 loci across 37 genes for European ancestry, 23 loci across 4 genes for African ancestry). That,

along with our methodological work on correction of collider stratification bias represents a valuable platform for future work.

The spectrum of findings across the two studies are important towards more targeted risk and prognosis stratification and potentially therapeutic efforts to reduce burden of BC outcomes.

Dedicated to my father, Pareshkumar Babubhai Patel

ACKNOWLEDGMENTS

I would like to begin by thanking my family, in particular my father and my mother. You two have been a source of endless support and wisdom throughout my educational career.

I would like to thank my committee, namely my advisor and dissertation chair Melissa Troester, my co-advisor Michael Love, Charles Perou, Kari North, and Hazel Nichols. Melissa and Mike, the two of you have been instrumental in cultivating and shaping my path into the intersection of cancer epidemiology and genetic epidemiology. Both of you are excellent mentors, and despite your tremendously busy schedule, I always felt I had your expertise and support to fall back on as I navigated this dissertation work.

Melissa, I especially enjoy the structure of your mentorship, and I think your setup of individualized feedback coupled with group feedback (Works in Progress Meetings) was key to my growth as a scientist. In fact, participating at those WIP meetings taught me the art of scientific discourse/communication, which I believe is one of the most important takeaways from my PhD training.

Mike, I especially appreciate your seemingly endless knowledge of the nuances of statistical genetics research and your openness and help towards crafting the next phase of my academic training.

Chuck, I truly appreciate your expertise on the nuances across the spectrum of breast cancer research. In particular, I enjoyed the challenge of the questions you raised, questions that I felt probed at the essence of the research work.

Kari, you were my introduction to genetic epidemiology in my very first year at UNC, and I believe your enthusiasm for the promises of this field is contagious, and was pivotal in shaping my desire to pursue a similar career path.

Hazel, you were my introduction to the challenges of breast cancer survivorship research. In addition to equipping me better to deal with methodological challenges, you brought a much needed perspective in your coursework and feedback, which is that cancer survivors are actual people, with lives and stories, beyond just the 1s and 0s we encounter in the data.

I would like to especially thank Arjun Bhattacharya. The genesis of this dissertation work came out of a meeting I had with him in March 2020, just as the pandemic took over and reshaped lives. You are a brilliant researcher, and like Mike, have a seemingly endless knowledge of the nuances of statistical genetics research. You have been key to my professional growth and towards me coming into my own as a scientist.

I would like to thank individuals I have collaborated with over the years, including Suril Mehta, Ruth Lunn, Montserrat Garcia Closas, Christine Gray, Danelle Lobdell, Jyotsna Jagai, and Lynne Messer.

Lastly, I would like to thank the individuals who have contributed to the Carolina Breast Cancer Study, Normal Breast Study, Genotype Tissue Expression Project, The Cancer Genome Atlas, and the Breast Cancer Association Consortium. Without the invaluable data contributions of these individuals, none of this dissertation work is possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GWAS – Genome-Wide Association Study

TWAS – Transcriptome-Wide Association Study

GReX – Germline regulated gene expression

eQTL – Expression quantitative trait loci

BC – Breast cancer

ER – Estrogen receptor

PR – Progesterone receptor

WW – White women

BW – Black women

GTEx – Genotype Tissue Expression Consortium

CBCS – Carolina Breast Cancer Study

NBS – Normal Breast Study

TCGA – The Cancer Genome Atlas

LumA – Luminal A

LumB – Luminal B

HER2 – HER2-enriched

TN – Triple negative

HR – Hormone receptor

BMI – Body mass index

RR – Risk Ratio

MR – Mendelian Randomization

IARC – International Agency for Research on Cancer

FHS – Family history score

BCAC – Breast Cancer Association Consortium

SNP – Single nucleotide polymorphism

LDSR – LD Score Regression

LD – Linkage Disequilibrium

SCCS – Southern Community Cohort Study

NBHS – Nashville Breast Health Study

FDR – False discovery rate

NC – North Carolina

LRT – Likelihood Ratio Test

GCTA – Genome-wide complex trait analysis

HR – Hazard Ratio

EA – European ancestry

AA – African ancestry

CHAPTER 1. SPECIFIC AIMS

Breast cancer (BC) poses a significant public health burden, both globally and within the United States (US). In the US, it is the second most common cancer among women, with an estimated 268,600 invasive cases diagnosed in 2019, and also the second leading cause of cancer mortality with an estimated 41,760 deaths in 2019 [1]. As part of efforts to reduce public health burden of BC and BC outcomes, genome-wide investigations to date have uncovered over 210 common BC susceptibility variants, which explain roughly 20% of the twofold familial relative risk [2-6]. There remain, however, several key gaps in knowledge across the spectrum of genetic investigations of BC incidence and mortality.

First and foremost, BC is a heterogeneous disease, with multiple subtypes with distinct outcome trajectories. Recent GWAS of BC suggest that germline associations may differ across BC subtypes (and at times these subtype-specific associations can be masked at the level of BC vs. control investigations if effects are in opposite directions) [7,8]. A more robust understanding of the germline genetic basis of BC subtypes can offer more targeted risk stratification and potentially even inform more targeted clinical efforts to reduce BC burden. This is especially true for diverse populations (e.g., individuals of African ancestry (AA)), who are vastly underrepresented in genetic investigations, despite, in the case of BC, standing among the most to benefit given the higher proportion of more aggressive BC subtypes and poorer mortality outcomes seen in this population group [1,9]. In fact, in the largest investigation of BC among individuals of African ancestry to date, no loci have been identified at or even near genome-wide significance [10]. To bridge these gaps in BC susceptibility research, in Aim 1 (Aim 1b), we propose a multi-ancestry (European ancestry (EA), African ancestry) Transcriptome-Wide Association Study (TWAS) based analysis (i.e. germline-regulated gene expression (GReX) analysis). TWAS-based GReX analysis, as a gene based association test, offers increased statistical efficiency, enabling discovery at both GWAS loci and loci where small effects aggregate at gene level (these would be missed in even large scale GWAS) [11-13]. GReX analysis offers the added benefit of increased interpretability, as GReX-prioritized genes can be subject to interrogation of biological plausibility and functional follow-up. Innovative aspects of our GReX analysis in Aim 1 include evaluation of the two most pertinent tissue contexts (normal breast, breast tumor tissue) across multiple genotype-expression panels in construction of the predictive models of gene expression from germline variation underlying GWAS (Aim 1a). We additionally investigate different definitions of BC subtype, including etiologic subtype, which has been shown to capture greater etiologic heterogeneity compared to molecular subtype [14].

Some of the challenges facing BC genetic etiologic research have implications for genetics of BC survivorship, a less developed research area (Aim 2). BC subtype is a well-established predictor of BC mortality, with markedly poorer mortality for more aggressive subtypes such as Basal-like [15]. Genetic investigations to date for BC mortality have not been fruitful, yielding only a few, mostly non-replicable loci. More recently, a GReX analysis of BC mortality suggest a complex interplay between BC molecular subtype and ancestry in BC mortality [16]. Specifically, in that GReX analysis of BC survival (in breast tumor tissue), race-specific predictive models uncovered no associations among White women (WW) and four associations among Black women (BW). Importantly, associations among BW appeared to be driven by very strong effect sizes within ER+ subtype. This suggests that BC biological (i.e. subtype) heterogeneity may be an important determinant of the relative lack of evidence on genetic underpinnings of BC mortality. Aim 2 of this dissertation will address this gap through a TWAS-based GReX analysis that leverages ancestry and subtype-specific predictive models.

Aim 1: Conduct a Transcriptome-Wide Association Study based GReX analysis of BC molecular subtype and etiologic subtype in normal breast and breast tumor tissue, for individuals of European and African ancestry.

Hypothesis: We hypothesize eQTLs/predictive models will differ by choice of breast tissue (normal, tumor) and that different putative causal genes will be identified by subtype and across subtype classification schemes (molecular, etiologic). We also hypothesize that eQTLs will differ across ancestries and that this will contribute to ancestry-specific differences in effect of putative causal genes. The approach will be:

1a. Gene expression predictive models for a suite of 396 BC-related genes (using the Carolina Breast Cancer Study (CBCS) as anchoring data source) will be constructed using paired genotype and expression data. Two sets of predictive models will be in normal breast tissue (Genotype Tissue

Expression Consortium (GTEx): N = 337 EA, N = 47 AA; UNC Normal Breast Study (NBS): N = 93 EA, N = 37 AA). Two sets of predictive models will be in breast tumor tissue from the Carolina Breast Cancer Study (CBCS, N=571 EA, N=628 AA) and The Cancer Genome Atlas (TCGA, N=715 AA, 170 EA). Predictive models will be mutually validated by tissue type. Predictive models will then be compared across normal-breast and breast tumor tissue in terms of model performance (heritability, correlation between GReX and observed expression).

1b. Tissue-specific (normal breast, breast tumor) predictive models will be used to impute gene expression in Breast Cancer Association Consortium (BCAC, N = 146,177 EA, N= 5,092 AA) for association testing with BC subtype (ER or PR/HER2 based; ER/TP53 based) in a case-control analysis.

Aim 2: Conduct a transcriptome-wide association study based GReX-analysis of BC mortality (all-cause, BC-specific) using predictive models that integrate BC molecular subtype and ancestry

Hypothesis: We hypothesize that increased resolution from integration of BC molecular subtype and ancestry will identify novel genetic loci for BC mortality. The approach will be:

Predictive models based on ancestry and BC molecular subtype stratification will be constructed in tumor tissue (CBCS, N=358 AA-Luminal-like, N=224 AA-Basal like, N=410 EA-Luminal-like, N=116 EA-TN/Basal-like). Predictive models will be used to impute gene expression in BCAC for association testing with BC mortality in case-only analyses (N= 89,992 across EA and AA).

CHAPTER 2. LITERATURE REVIEW AND RATIONALE

2.1 Epidemiology of Breast Cancer Incidence and Mortality

Breast cancer (BC) is the second most common cancer among women in the United States (U.S.), with an estimated 48,100 ductal carcinoma *in situ* (DCIS) and 268,600 invasive cases diagnosed in 2019 [1]. Incidence rates of BC across the U.S. population increased until 2000, declined between 2000 and 2005, and have remained relatively stable since [17-20] Across racial/ethnic groups, White women (WW) have historically had the highest incidence rate (~130 cases per 100,000 individuals per year), although since 2012, incidence rate among Black women (BW) has been nearly equivalent to those among WW [21,22].

BC can be classified into molecular subtypes based on combination of estrogen (ER) and human epidermal growth factor 2 (HER2) receptor status (determined through immunohistochemistry) or into intrinsic subtypes through tumor gene expression [23]. Molecular subtypes of BC determined through immunohistochemistry include Luminal A-like (Lum-A, ER+ or PR+, HER2-), Luminal B-like (LumB, ER+ or PR +, HER2+), HER2-enriched (HER2, ER and PR - , HER2+), TN/Basal-like (TN, ER-, PR-, HER2-) [23]. In the U.S., the distribution of BC molecular subtypes in the general population is as follows: LumA – 68%; LumB – 10%; HER2 – 4%; TN/Basal-like – 10%; and 7% unknown [1,23]. Across all-groups (i.e., 25-39 years, 40-54 years, 55-69 years, and 70+ years), TN BC is more prevalent among BW compared to WW. For WW and BW, TN is more prevalent among younger age groups (i.e., most prevalent among women 25-39 years)[1,9].

BC is the second leading cause of cancer mortality in the U.S. with an estimated 41,760 deaths in 2019 [1]. There is significant disparity in BC mortality; BW have roughly 40% greater mortality rate compared to WW (28.4 per 100,000 individuals per year in BW compared to 20.3 per 100,000 individuals per year in WW)[1]. Mortality rate differences between WW and BW also differ by age-groups. Among 20-29-year-old women, the mortality rate is 2.62 times that of WW in BW, while among 70-79-year-old women, the mortality rate is 1.11 times that of WW in BW[1]. Mortality rate differences for WW and BW across age groups partly reflect differences in BC molecular subtype distribution for WW and BW across age groups as BC molecular is a strong prognostic factor for BC mortality[1,23]. 5-year relative survival for BC is 90.3% while the 5-year relative survival for BC molecular subtypes is as follows: LumA – 94.3%; LumB – 90.5%; HER2 – 84%; TN/basal-like – 76.9% [23].

2.2 Genetic factors for BC Incidence

Family History

Early investigations of genetic factors in relation to BC were focused on family history and study designs that exploit different levels of shared genetic architecture across monozygotic and dizygotic twins. Women whose mothers or sisters have experienced BC are 2 times and 2.3 times as likely, respectively, to develop BC; women with both mothers and sisters that have experienced BC are 3.6 times as likely to develop BC [24]. An analysis leveraging the Generations Study, which is a cohort of~113,000 women from the UK general population, reported a 3.5-fold increase (95% CI: 2.56, 4.79) in BC risk comparing highest family history score (FHS) and lowest FSH groups, adjusted for established reproductive and lifestyle risk factors [25]. The FHS used in Brewer et al. compared observed BC cases in the family in relation to expected based on population incidence rates by age and calendar period, which offered an improved account for family structure compared to conventional approach of enumerating cases across relatives [25]. Furthermore, in one of the largest twin studies to date, including 80,309 monozygotic twins and 123,382 dizygotic twins, the probability of developing BC was 28.1% (95% CI: 23.9, 32.8) if a monozygotic twin sister had breast cancer[26]. The probability of developing BC was lower at 19.9% if a dizygotic twin sister had breast cancer[26].

Inherited mutations (rare variants)

With the completion of the Human Genome Project and proliferation of genotyping technology and statistical methods for analysis of genetic data, focus in understanding of genetic architecture of BC risk shifted to genetic variants. Inherited mutations (rare variants) in several genes have been implicated in BC risk, foremost of which are *BRCA1* and *BRCA2*. *BRCA1* and *BRCA2* are involved in DNA repair and function as tumor suppressor genes in healthy cells [27,28]. Women with mutation in these genes have a 70% chance of developing BC by age 80 [29]. It has been estimated that approximately 1 in 400 to 800 individuals in the general population carry pathogenic *BRCA* variants [30-32].

Compared to the general population, women who develop BC at any age have a 2% chance (1 in 50 women) of carrying pathogenic *BRCA* variants while women who develop BC at age 40 or lower have a 10% chance (1 in 10 women) of carrying pathogenic *BRCA* variants [33-35]. Prevalence of *BRCA1* pathogenic variants among BC patients differs by racial/ethnic groups and is highest among Hispanics at 3.5%, followed by 2.2 - 2.9 % in non-Ashkenazi WW and 1.4% in BW [36]. Prevalence of *BRCA2* pathogenic variant among BC patients also varies by racial/ethnic groups and is highest among BW at 2.6% and slightly lower among WW at 2.1% [36]. Prevalence of *BRCA* pathogenic variants also differs by BC molecular subtype, and has been reported to be highest among TN BC. In a pooled analysis of 12 studies (from the Triple Negative Breast Cancer Consortium) of TN BC patients unselected for family history of breast or ovarian cancer, 8.5% had *BRCA1* pathogenic variants and 2.7% had *BRCA2* pathogenic variants [37].

BRCA1 and *BRCA2* are considered high penetrance genes; penetrance refers to proportion of individuals with pathogenic variants that demonstrate symptoms of the disease/disorder. Other high penetrance genes implicated in relation to BC risk are *CDH1*, *PALB2, PTEN*, *STK11*, and *TP53* [38].Moderate penetrance genes associated with BC risk include *ATM*, *BRIP1*, *CHEK2*, *FANCD2*, and *RAD51C* [38]. TP53 germline (inherited) mutations are relatively uncommon at an estimated 3% in the general population; TP53 mutations, by contrast, are found in approximately 30% of breast cancer patients [39,40]. Within BC subtypes, TP53 mutations are most common among TN/basal-like BC [40]. Many of aforementioned genes are involved in repair of damaged DNA and regulation of cell growth, including in the presence of DNA damage [34]. A 2021 population-based case-control study of 48,826 cases and 50,703 controls found the following BC associations for protein-truncating (pathogenic) variants for some of the aforementioned genes (Table 1) [3].

Table 2.1 Risk of breast cancer associated with pathogenic protein-truncating variants in select genes in population-based studies in the Breast Cancer Association Consortium

Gene	Odds Ratio (95% CI)	P-value
BRCA1	10.57 (8.02, 13.93)	1.1 x 10^{-62}
BRCA2	5.85 (4.85, 7.06)	2.2 x 10^{-75}
CDH1	0.86 (0.37, 1.98)	0.72
PALB2	5.02 (3.73, 6.76)	1.6 x 10^{-26}
PTEN	2.25 (0.85, 6.00)	0.10
STK11	1.60 (0.48, 5.28)	0.44
TP53	3.06 (0.63, 14.91)	0.17
ATM	2.10 (1.71, 2.57)	9.2 x 10^{-13}
BRIP1	1.11 (0.80, 1.53)	0.54
CHEK2	2.54 (2.21, 2.91)	3.1 x 10^{-39}
FANCD2	-	-
RAD51C	1.93 (1.20, 3.11)	0.0070

Adapted from Dorling et al. 2021.

Genome wide studies of BC risk

BC is a complex trait, meaning that it does not follow classical Mendelian inheritance patterns. This recognition, along with recognition that common variants (single nucleotide polymorphisms (SNPs)) with lower penetrance may explain a substantial portion of BC heritability has prompted numerous genome-wide association study (GWAS) over the years [38]. By convention, variants with greater than 1% frequency in the population are termed SNPs while those with less than 1% are termed mutations [41]. More generally, there is thought to be an inverse relationship between penetrance and variant frequency; common variants tend to have lower risk while rare variants (*BRCA1* and *BRCA2* mutations for instance) tend to have higher risk [38].

The first GWAS of BC were conducted in East Asian and European populations. In a discovery sample of 3,027 Chinese women and a replication sample of 7,502 East Asian and 3,057 European women, Zheng et al. identified rs2046210 at 6q25.1, located upstream of the estrogen receptor alpha (ESR1) gene, as conferring increased risk of BC [42]. Thomas et al. conducted a three-stage GWAS in

9,770 cases and 10,799 controls in the Cancer Genetic Markers of Susceptibility (CGEMS) initiative, where top SNP associations at each stage were prioritized for further association testing [43]. Thomas et al. identified two loci at genome-wide significance, namely rs11249433 and rs999737, the latter localizing to the *RAD51L1* gene that is involved in the homologous recombination DNA repair pathway [43]. Thomas et al. further investigated heterogeneity in effect by ER status, finding that the association was stronger for rs11249433for ER+ compared to ER-, while the association for rs999737 was roughly similar across ER tumor types[43].

In 2013, Garcia-Closas et al. conducted the first GWAS specific to ER- BC in an effort to better understand the genetic susceptibility specific to this tumor type, which differs in its etiology and clinical prognosis compared to ER+ BC, as mentioned previously. In a meta-analysis of 3 GWAS of 4,193 ER-BC and 35,194 controls (and a replication sample of 47,969 women, all of European ancestry), Garcia-Closas et al. identified four susceptibility loci, two of which localized to *MDM4* and *LGR6* genes [44]. Importantly, all four susceptibility loci were associated with ER- but not ER+ BC, indicating potentially different genetic architecture of ER+ and ER- BC risk [44].

Present understanding of the genetic architecture for BC risk is significantly better among women of European ancestry compared to women of African ancestry. Among women of European ancestry, the bulk of understanding stems from four GWAS studies from 2013, 2015, 2017, and 2020. Building on the 27 loci that had been identified till 2013, Michailidou et al. performed a meta-analysis of 9 GWAS studies among women of European ancestry, and identified 29,807 SNPs for further genotyping in 45,290 cases and 41,880 controls in the Breast Cancer Association Consortium (BCAC) [6]. These genotyped SNPs were part of a custom Illumina iSelect genotyping array (iCOGS – 200K variants), and the iCOGS based GWAS in BCAC identified 41 loci at genome-wide significance ($P < 5 \times 10^{-8}$) [6]. Michailidou et al.'s work in 2015 built upon the 2013 GWAS and leveraged genotype imputation (using the 1000 Genomes Project March 2012 release) to perform GWAS on 11 million SNPs [5]. The 2015 work identified 15 novel susceptibility loci [5]. In the 2017 GWAS, Michailidou et al. leveraged another custom genotyping array (OncoArray – 550K variants and 21 million imputed SNPs) and performed a GWAS in 61,282 cases and 45,494 controls of European ancestry in BCAC, results of which were combined with the prior iCOGS analyses and 11 other BC GWAS using fixed-effect meta-analysis to yield 65 novel loci for BC risk [4]. In the OncoArray, 72K SNPs were targeted towards BC, including SNPs that showed evidence in prior GWAS (for overall or ER-specific BC), were potentially associated with BC survival, and were loci of interest from GWAS in other (e.g., Asian) populations [4].

The most recent (2020) GWAS in BCAC by Zhang et al. was motivated by observations of heterogeneity by tumor subtype. In this GWAS, Zhang et al. accounted for tumor heterogeneity (i.e., ER, progesterone receptor (PR), HER2) as well as tumor grade [2]. Zhang et al. identified 32 new loci at genome-wide significance ($P < 5 \times 10^{-8}$) and several (15) of these loci demonstrated evidence of association with at least one tumor heterogeneity feature [2]. Moreover, five loci showed associations in opposite directions for luminal and non-luminal subtypes, respectively, across the spectrum of BC GWAS signa. Follow-up analyses for these five loci uncovered that these loci contained cell-specific enhancers which differ between normal luminal and basal cells in mammary tissue. Zhang et al. also estimated the proportion of genetic variance explained by the set of 210 loci (178 previously identified + 32 novel in their study) compared to the genetic variance explained by all loci and found differences in this proportion by BC subtype [2]. Together, the 210 variants explained approximately 18% of the twofold familial relative risk for invasive BC while the set of all variants (including imputed variants) explained twice that at 37.1% [2]. Lastly, Zhang et al. estimated the genetic correlation between BC molecular subtypes through LD score regression (LDSR). In LDSR, relationships between linkage disequilibrium (LD) scores for a given variant and variant test-statistics (i.e., statistic computed in the study or summary statistics from external GWAS) are assessed, where LD score for a given variant is the sum of LD r^2 for that variant with variants within 1 cM [45]. The considered subtypes were moderately to highly correlated. In particular, genetic correlation for LumA BC and TN BC was 0.46 (SE=0.05) while genetic correlation for TN BC and BC among carriers of BRCA1 was 0.83 (SE=0.08) [2]. The pattern of genetic correlations across subtypes indicate shared genetic architecture for BC risk across subtypes but also point to subtype-specific genetic architectures for BC risk [2]. Importantly, as with previous GWAS by Michailidou et al, this GWAS was limited to women of European ancestry.

Parallel to the Zhang et al.'s work on identifying BC susceptibility loci while accounting for tumor heterogeneity, Ahearn et al. interrogated the relationship between known susceptibility variants and: 1) tumor markers (ER, PR, HER2, grade) that define BC subtype; 2) BC subtypes (LumA, LumB, HER2, TN/Basal-like) [8]. Ahearn et al.'s work is the most in-depth investigation of BC susceptibility variants in relation to tumor heterogeneity, and in particular BC molecular subtypes; it builds upon less systematic evidence from prior studies that demonstrate heterogeneity in effect of susceptibility variants by ER status [8]. Ahearn et al. developed a two-stage polytomous logistic regression that addressed several challenges involved in investigation of susceptibility variants in relation to BC molecular subtypes: (1) tumor heterogeneity marker (ER, PR, HER2) correlations; (2) missing data on tumor markers; (3) lower sample size among less common subtypes [8]. Ahearn et al. reported that 85 of 173 known susceptibility variants were associated with at least one tumor heterogeneity (ER, PR, HER2, grade) feature (at false discovery rate (FDR) < 0.05). Combined with the 15 variants (among the 32 novel variants in Zhang et al.) that showed similar association with at least one tumor heterogeneity marker (ER, PR, HER2, grade), there is substantial evidence of etiologic heterogeneity for BC molecular subtypes across genome-wide significant susceptibility variants [2,8]. More specifically, Ahearn et al. report that while most (N=83) of the 85 susceptibility variants associated with at least tumor heterogeneity marker were associated with at least one luminal subtype, only 41 of the 85 were associated with at least one non-luminal subtype [8]. Moreover, 32 of these 85 susceptibility variants showing evidence of association with at least one tumor heterogeneity marker were associated with TN BC [8]. Lastly, 10 of the 85 were associated with risk across all BC molecular subtypes, with differing magnitudes of effect [8]. Given that BC is a heterogeneous disease where BC molecular subtypes have dramatically different therapeutic options and prognoses, understanding of genetic susceptibility across subtypes, especially in terms of genetic mechanisms (i.e., SNP -> gene expression -> trait) has clinical implications.

As previously mentioned, there is comparatively limited understanding of genetic architecture of BC risk among women of African ancestry, in large part due to the smaller sample sizes of studies and consortia with genotype information. A 2013 study by Long et al. evaluated 67 SNPs that had been identified in relation to BC risk (at the time of study initiation) in 1,231 cases and 2,069 African-American controls recruited within the Southern Community Study (SCCS) and the Nashville Breast Health Study (NBHS) [46]. Long et al. reported 7 of the67 evaluated SNPs were nominally significant (P <0.05) and demonstrated effect in the same direction as identified in women of European ancestry [46]. In 2016, Huo et al. performed a two-stage GWAS, first identifying 18,376 SNPs in a meta-analysis of roughly 5,000 cases and 5,000 controls before replication follow-up among 1,984 cases and 2,939 controls; Huo et al. strongly linked rs13074711 and rs10069690 with ER- BC and rs12998806 with ER+ BC [47]. An interesting finding of this study was that all three variants were all highly heterogeneous with regards to ER status, suggesting that the genetic architecture of BC risk among women of African ancestry may have even more subtype-specificity compared to the genetic architecture of BC risk among women of European ancestry [47]. In the largest genetic study (6,522 cases and 7,643 controls) among women of African ancestry to date, Feng et al. assessed 74 SNPs that had been identified among women of European ancestry, and reported 12 with nominally significant associations, all of which showed directional consistency in effect [48]. In follow-up fine-mapping of susceptibility regions, Feng et al. also identified variants that better characterized the risk signal compared to the variants reported among women of European ancestry [48]. The sum of these studies highlight the central limitation towards understanding of genetic architecture of BC risk among women of African ancestry, namely, that the restrictive sample sizes do not allow for robust genome-wide significance testing, limiting discovery of variants specific to BC risk and BC subtypes among women of African ancestry.

Transcriptome Wide Association Study of BC risk

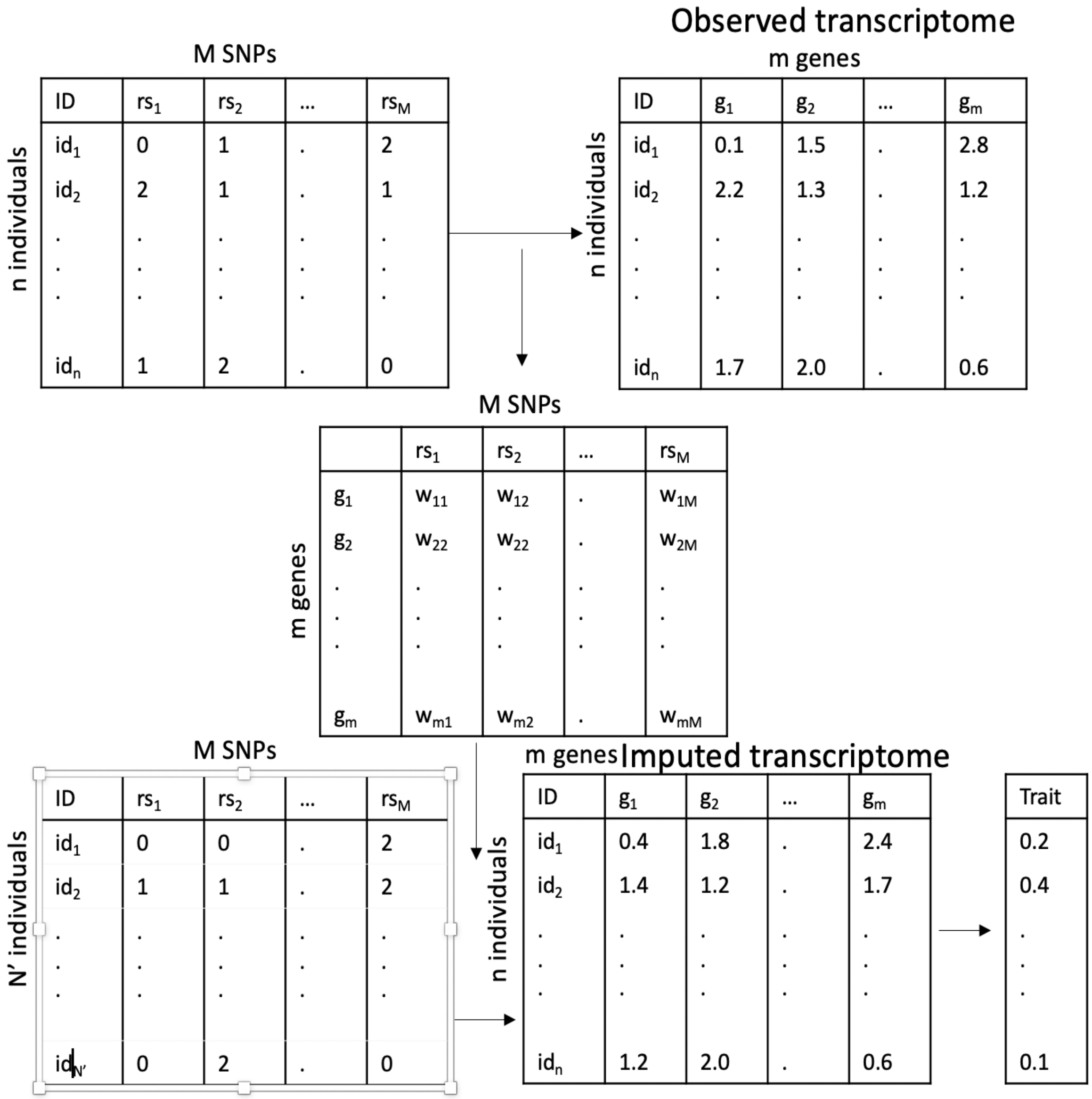
The combination of family studies, studies involving rare variants (mutations), and especially GWAS of common variants (SNPs) have established that germline genetics play an essential role in BC risk. A key finding in GWAS studies of BC risk is that many susceptibility variants are located in non-coding regions; moreover, these variants are not in LD with SNPs in coding regions [49]. Studies show that many susceptibility variants are located in regulatory regions, wherein they are thought to drive risk of BC through regulatory influence on the expression of nearby and distal genes [50-56]. While target genes for several BC susceptibility variants with hypothesized regulatory influence on gene expression have been identified, these approaches are limited in the sense that they can only identify target genes for variants that meet genome-wide significance in GWAS.

Numerous studies have reported that regulatory variants can account for a large proportion of disease heritability that has not yet been uncovered through GWAS [57-59]. Specifically, there may be numerous regulatory variants for a given gene; such variants individually may not necessarily meet genome-wide significance in GWAS but (small) effects from these variants can aggregate and influence risk through genetic regulation of expression of local (<1 Mb) and distal genes. Discovery of expression quantitative trait loci (eQTLs) over the past several years have provided considerable support to this line of thought [54,60]. An eQTL is a genomic locus that explains some portion of the variance in expression of nearby or distal genes [61]. eQTLs can be classified as: 1) *cis* or *trans*, depending on whether they are thought to influence gene expression directly or indirectly; 2) local or distal, depending on genetic distance to the gene whose variance in expression they explain [61]. The focus on uncovering mechanisms for variant and risk associations (in BC and other diseases) along with recognition that regulatory variants with small effect sizes may aggregate and be consequential towards disease heritability has prompted intense interest in Transcriptome-Wide Association Study (TWAS).

TWAS is a gene-based association testing approach where the relationship between germline variation and gene expression is first determined; among genes that show sufficient heritability, the germline-regulated component of gene expression (GReX; i.e., the expression of the gene predicted through germline variation) can then be used as a variable in regression analyses of continuous or binary traits. TWAS offers several advantages to GWAS, namely: 1) as a gene-based association test, the multiple testing burden is greatly reduced, allowing greater statistical efficiency and power; 2) TWAS-significance identifies potential mechanisms for the relationship between variants and disease risk (i.e., it identifies SNPs -> gene expression -> disease risk) that can be immediately subject to functional validation [13,62].

There are two main frameworks for conducting TWAS: 1) PrediXcan; 2) FUSION [13,62]. PrediXcan was developed by Gamazon et al. and in essence, the approach decomposes total gene expression into three compartments: 1) GReX; 2) environmental and other, perhaps epigenetic factors; 3) expression that is influenced by the trait/outcome of interest itself. PrediXcan isolates GReX through the following algorithm: 1) Datasets with paired genotype and reference transcriptome data (reference can be single tissue transcriptome; multiple references corresponding to multiple tissue are possible) are identified; 2) Additive models of gene expression in the reference transcriptome data based on germline variants are constructed, typically, through Elastic Net Regression. Weights (regression betas) corresponding to the effect of 0, 1, 2 dosages of the alternative allele on gene expression are stored at the gene-level; 3) In an external, usually considerably larger dataset with genotyping information, gene-level models can be used to impute the GReX for that gene (for a given reference tissue); 4) Imputed GReX of genes can be tested against trait of interest [13]. Several variants of PrediXcan exist. For example, PrediXcan can also be performed using summary statistics from GWAS in the absence of individual level genotype datasets for GReX imputation. Another variant of PrediXcan termed MultiXcan can combine information across prediction models of gene expression trained in multiple tissues (to take advantage of common and different eQTL structure across tissues), effectively mimicking a meta-analysis of effects across tissue [63]. The conceptual framework for PrediXcan is as follows (Figure 1).

Figure 2.1 Conceptual framework of PrediXcan, one of primary approaches for TWAS



ID	rs_1	rs_2	...	rs_M
id_1	0	1	.	2
id_2	2	1	.	1
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
id_n	1	2	.	0

ID	g_1	g_2	...	g_m
id_1	0.1	1.5	.	2.8
id_2	2.2	1.3	.	1.2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
id_n	1.7	2.0	.	0.6

	rs_1	rs_2	...	rs_M
g_1	w_{11}	w_{12}	.	w_{1M}
g_2	w_{22}	w_{22}	.	w_{2M}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
g_m	w_{m1}	w_{m2}	.	w_{mM}

ID	rs_1	rs_2	...	rs_M
id_1	0	0	.	2
id_2	1	1	.	2
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$id_{N'}$	0	2	.	0

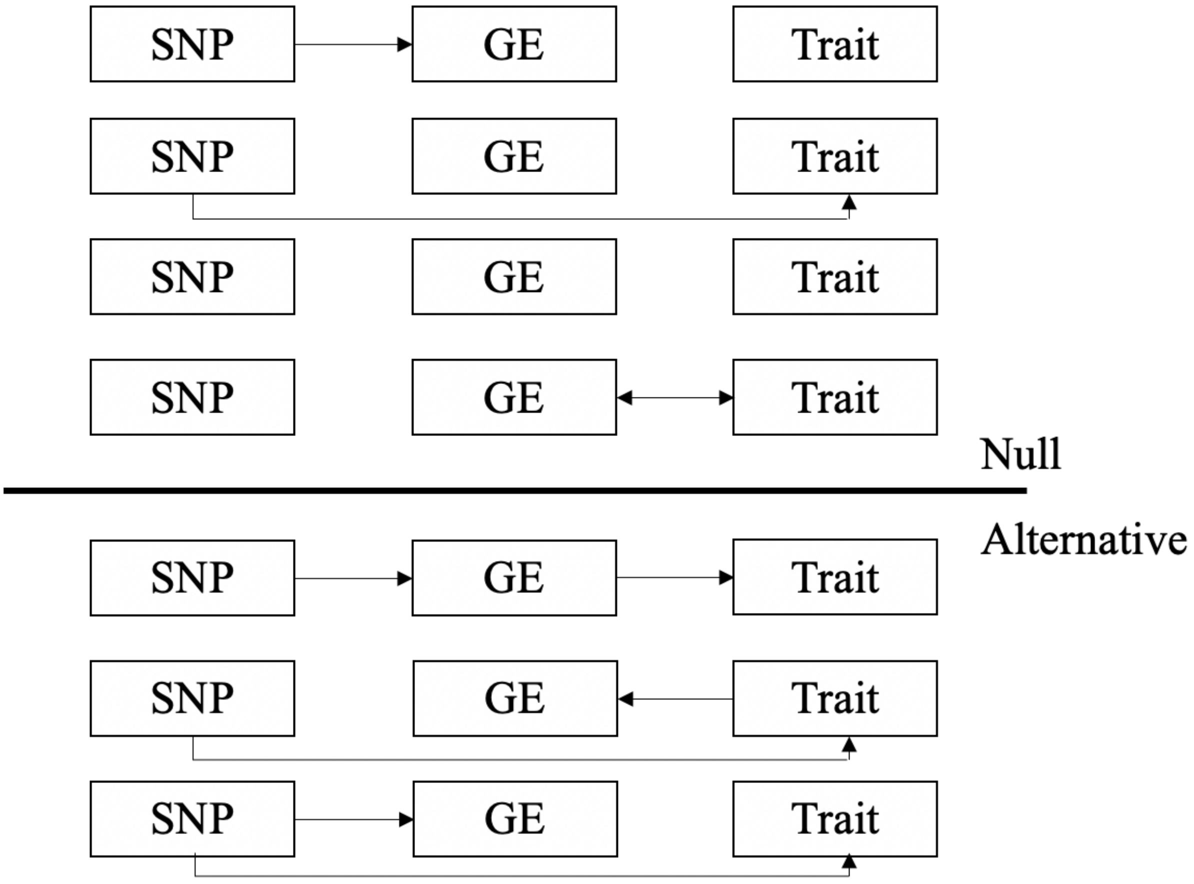
ID	g_1	g_2	...	g_m
id_1	0.4	1.8	.	2.4
id_2	1.4	1.2	.	1.7
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
id_n	1.2	2.0	.	0.6

Trait
0.2
0.4
.
.
.
0.1

Adapted from Gamazon et al. 2015.

Around the same time as Gamazon et al. proposed PrediXcan, Gusev et al. proposed an alternative TWAS approach – FUSION [62]. As with PrediXcan, FUSION can be implemented using either individual level genotype data (for imputation) or summary statistics from variant –trait associations in GWAS. For both implementations, FUSION requires predictive models of gene expression based on germline variation, constructed in a single tissue or multiple tissues. Under the summary statistics approach in FUSION, associations between GReX and trait are indirectly estimated as a weighted linear combination of standardized variant-trait associations (Z-scores from GWAS); FUSION explicitly accounts for LD between variants used in the weighted burden test [62]. A key difference between FUSION and PrediXcan is that under FUSION, several approaches (*cis*-eQTL, elastic net, best unbiased linear predictor (BLUP) or Bayesian linear mixed model (BSLMM)) are utilized to train predictive models of gene expression [62]. For a given gene, a single model that perform best on heritability and cross-validation R^2 (GReX vs. total expression) is selected for downstream association analyses. Gusev et al. identify scenarios under which TWAS-significant genes can be identified using FUSION (Figure 2).

Figure 2.2 Scenarios under which genes can be identified as TWAS-significant (Gusev et al. 2016)



To date, there have been two notable TWAS of BC risk, both in women of European ancestry and in the BCAC. Wu et al.'s TWAS comprised 122,977 cases and 105,984 controls and they constructed predictive models of gene expression in breast mammary tissue using the Genotype Tissue Expression Consortium (GTEx, N=67 for breast mammary tissue transcriptome and genotype panel) [64]. Among 8,597 genes that showed sufficient cross-validation R^2 (GReX versus total expression performance within GTEx and in external comparisons using The Cancer Genome Atlas (TCGA), Wu et al. identified 48 genes at transcriptome-wide significance ($P < 5.82 \times 10^{-6}$), including 14 genes at loci that had not yet been reported in BC GWAS, underscoring the power of TWAS (and TWAS based GReX association analysis) in identifying regulatory variants with small effects on gene expression that can aggregate to influence BC risk (such variants would miss discovery even under large scale GWAS). Wu et al. performed functional follow-up on 13 of the identified TWAS-significant genes, silencing them and assessing effects on cell proliferation and/or colony-forming efficiency; 11 of the 13 silenced genes showed an effect consistent with that predicted by TWAS, highlighting TWAS's ability to accurately identify putative causal genes under most scenarios of statistical significance [64]. An important limitation to Wu et al.'s TWAS is the relatively sparse dataset (N=67 European ancestry) of breast mammary tissue transcriptome and genotypes for construction of predictive models of gene expression.

Feng et al.'s TWAS comprised of the 122,977 cases and 105,974 controls in Wu et al. and also included 11 other GWAS that spanned 14,910 cases and 17,588 controls [65]. As with Wu et al., Feng et al. constructed their predictive models using gene expression in breast mammary tissue (and corresponding genotype information) in 67 women of European ancestry in GTEx. After narrowing the candidate gene set to 901 genes based on sufficient heritability (nominal p <0.01) and cross-validation R^2 of 0.01 (10% correlation between GReX and total gene expression), Feng et al. identified 30 TWAS-significant genes, including 4 not identified by Wu et al. Further, in ER (ER+ / ER-) specific TWAS, Feng et al. identified two genes, namely *STXBP4* and *HIST2H2BA*, specific to ER+; these genes showed no associations with ER-BC [65]. Although Feng et al. improved upon the work in Wu et al. by only testing genes with sufficient heritability, the study was limited by the same sample size (N=67 European ancestry) of the reference expression panel for predictive model construction. Importantly, both Wu et al. and Feng et al. were performed among individuals of European ancestry. To date, there have been no TWAS or TWAS-like gene based association analyses among individuals of African ancestry for BC.

2.3 Genetic and Clinical Risk Factors for Breast Cancer Mortality

Clinical and Healthcare Access Factors

Tumor size and stage are strong prognostic factors for BC mortality [66,67]. Women with larger tumors are more likely to be diagnosed later stage, more likely to have metastasized, and more likely to die within 5 years post diagnosis compared to women with smaller tumors [68]. Women diagnosed at the localized stage have a 99% 5-year relative survival, and those diagnosed at regional and distant have 86% and 28%, respectively [69]. Tumor grade is also a prognostic factor; generally, poorly differentiated tumors (higher grade) tumors have worse prognosis compared to well differentiated (lower grade) tumors [70]. BC subtype is among the strongest prognostic factors. Compared to LumA, HR of 1.7 (1.0, 2.9) and 1.4 (0.9, 2.1) have been reported for Basal-like BC for WW and BW, respectively [15]. For HER2 compared to LumA, HR of 1.4 (0.7, 2.9) and 1.8 (1.0, 3.1) have been reported for WW and BW, respectively [15]. BC subtype is a strong predictor of treatment. For instance, LumA and LumB (ER+ BC) are primarily treated with endocrine therapy while chemotherapy alternatives are preferred for Basal-like and HER2 BC. For HER2 BC, immunotherapy with Herceptin complements chemotherapy [71]. There appears to be a complex interplay between BC subtype, race, and BC mortality, where race (WW, BW) and BC subtype independently affect BC mortality, and race additionally affects subtype distribution (see 2.1) [15,72]. Lastly, health-care access is an independent risk factor for BC mortality, where increased geographic distance to health-care services, increased travel time, and quality of health-care services are associated with poorer prognosis [73,74].

Family history and mutations (inherited, somatic)

Although family history is one of the strongest risk factors for BC, investigations of prognostic value of family have yield mixed conclusions. Some studies suggest family history is associated with higher mortality while other studies report association with lower mortality or lack of an association [75-78]. Reasons for this heterogeneity in findings are unknown and may be an artifact of collider stratification bias, given that family history is a strong risk factor for BC, and conditioning on BC incidence in an analysis of family history and BC mortality induces confounding from factors related to BC incidence and mortality [79]. *BRCA1*, *BRCA2*, and *TP53* are strong prognostic factors. As previously mentioned, mutations in all three genes are found in higher frequencies among more aggressive subtypes such as HER2 and especially TN. Huszno et al. reported that the 10-year overall survival for BRCA mutation carriers was roughly20% lower compared to non BRCA mutation carriers; *TP53* mutation carriers, on the other hand, have been reported to have both poorer and better survival compared to *TP53* wild-type carriers, depending on treatment modalities [80,81].

Genome-Wide Association Study of Breast Cancer Mortality

Compared to BC incidence, there have been relatively fewer GWAS of BC mortality, and most of them report suggestive findings or findings that are not replicable [82-92]. Among the two notable, Shu et al. identified two variants (rs3784099, rs9934948), the former localizing to *RAD51L1* which is involved in DNA repair pathways; rs9934948 replicated in a sample of women of European ancestry from the Nurses' Health Study [86]. The largest GWAS of BC mortality to date among 96,661 women of European ancestry in BCAC reported no associations between any variants and BC mortality risk at genome-wide significance; however, there was suggestion of association for one variant each for ER+ and ER- BC and these variants were located proximal to genes with biological relevance to BC outcomes [93].

TWAS of BC mortality

To date, there has only been one TWAS of BC mortality. Bhattacharya et al. assessed the relationship between GReX of 406-BC related genes (396 autosomal) in breast tumor tissue and BC mortality in the Carolina Breast Cancer Study (CBCS) [16]. CBCS is comprised of an ancestrally diverse population, with roughly half being women of European ancestry and the other half being women of African ancestry. As such, Bhattacharya et al. structured their TWAS in this ancestrally diverse study population as a race-specific TWAS after finding that predictive models of tumor gene expression from germline variation were not transportable across race. Bhattacharya et al.'s work shows that in performing TWAS across ancestrally diverse populations, race (as a proxy for ancestry) stratification of predictive model training, imputation, and association testing may be necessary to draw correct inferences, especially in the absence of more robust methods for ancestry-specific TWAS, especially in ancestrally heterogeneous or admixed populations [16]. In their TWAS, Bhattacharya et al. found associations for 4 genes among BW and associations for 0 genes among WW. Interestingly, 3 of the 4 associations among BW were driven by strong effects in one BC subtype, offering indication that a TWAS of BC mortality where predictive models are trained on ancestry and BC subtype may offer significantly improved account of biological heterogeneity and potentially improved power for identification of putative causal genes for BC mortality [16]. Analogous to the observation in Bhattacharya et al, heterogeneity in effect of variants by subtype in the GWAS setting has been speculated as a possible reason for reduced power and failure to detect genome-wide significant variants [93].

Summary

BC is a disease that impacts millions on individuals around the world and has a high mortality burden. In efforts to reduce public health burden of BC and BC outcomes, genome-wide investigations to date have greatly furthered our understanding of the germline genetic basis of BC; however, there remain a few key gaps in knowledge.

BC is a heterogeneous disease, with multiple subtypes with distinct outcome trajectories. Evidence suggests that germline associations may differ across BC subtypes (and at times these subtype-specific associations can be masked at the level of BC vs. control investigations), which means that a more robust understanding of the germline genetic basis of BC subtypes can offer more targeted risk stratification and potentially even inform more targeted clinical efforts to reduce BC burden. This is especially true for individuals of African ancestry, who are vastly underrepresented in genetic investigations, despite being the population standing most to benefit given the higher proportion of more aggressive BC subtypes and poorer mortality outcomes seen in this population group.

Leveraging biological heterogeneity in BC (i.e., subtypes) may also be key towards bridging another gap in knowledge in the etiology of BC outcomes, which is the large gap in number of genetic studies and findings between genetic studies of BC incidence versus mortality (again, thus far, limited to European ancestry populations).

In this dissertation, we address some of these key gaps in knowledge, using a multitude of publicly available (Genotype Tissue Expression Project, The Cancer Genome Atlas) and protected data (Genotype Tissue Expression Project (genotype), The Cancer Genome Atlas (genotype), Carolina Breast Cancer Study, Normal Breast Study, Breast Cancer Association Consortium). As a point of emphasis, the Carolina Breast Cancer Study is among the largest sources of genetic data for diverse populations, and for the Breast Cancer Association Consortium, we worked to gain approval for use of individual level genotype data, when summary level data would have sufficed if our work was limited to individuals of European ancestry. Our work is a small step towards bridging the disparity in genetic investigations across European and African ancestries across BC outcomes, and it represents our best possible effort with available data.

CHAPTER 3. METHODS

3.1 Data Sources

Genotype-expression reference panels for predictive model construction

Carolina Breast Cancer Study (CBCS)

CBCS is a population-based study of North Carolina (NC) breast cancer patients enrolled over 3 phases; study details have been previously described [94,95]. Patients aged between 20 to 74 years were identified using rapid case ascertainment with the NC Central Cancer Registry. CBCS oversampled for self-identified Black and young women (ages 20-49) [95,96]. Data on demographic and clinical factors (e.g., age, menopausal status, body mass index, hormone receptor status, tumor stage, study phase, mortality) were obtained through a combination of questionnaires and medical records.

Genotypes in CBCS were assayed using a custom SNP array designed for the OncoArray Consortium (Illumina Infinium OncoArray) and imputed using the 1000 Genomes Project Phase 3 as a reference panel for two-step phasing and imputation (SHAPEIT2 and IMPUTEv2) [97-100]. Genotype calling, quality control, and imputation was conducted by the DCEG Cancer Genomics Research Laboratory [101]. For Aim 2, which only used CBCS data for predictive model construction, we used genotypes that had been imputed to the 1000 Genome Project Phase 3. For Aim 1, to ensure consistency in genotype imputation underlying predictive models constructed across various data sources, we imputed CBCS genotype data to the TOPMed reference panel [102]. In TOPMed, phasing was performed using eagle and imputation using minimac4 [103,104]. For TOPMed imputed CBCS genotype data for Aim 1, we retained variants with high imputation quality ($r^2 > 0.8$). Additionally, in further quality control of genotype data, we excluded variants significantly deviating from Hardy-Weinberg Equilibrium (at $P < 1.0 \times 10^{-8}$).

Gene expression in CBCS from paraffin-embedded tumor blocks was assayed for a panel of 396 autosomal BC-related genes (along with 11 housekeeping genes) using the NanoString nCounter platform; assays were performed at the Translational Genomics Laboratory at UNC-Chapel Hill [96,105]. These 396 BC-related genes include genes part of the *PAM50*, *TP53*, *E2*, *IGF*, and *EGFR* signatures, among others (Appendix 1). As part of the quality control process, samples with insufficient data quality were eliminated using NanoStringQCPro [106,107]. Distributional differences between lanes were scaled with upper-quartile normalization [108], and two dimensions of unwanted technical and biological variation, estimated using the 11 housekeeping genes in RUVSeq, were removed [108,109]. Final expression data was in $\log_2$.The sample sizes of matched (i.e., data available for both genotype and expression and covariates for the same individual) genotype-expression reference panels for Aim 1 were 571 EA, 628 AA women. Sample sizes of the reference panel for Aim 2 were 410 EA – LL, 116 EA – BL, 358 AA – LL, 224 AA – BL, where LL ((ER+ / PR+) and HER2-, (ER+ / PR+) and HER2+), (ER-, PR-, HER2+)) BL (ER-, PR-, HER2-) denote Luminal-like and Basal-like subtype, respectively.

Normal Breast Study (NBS)

NBS is a study of normal breast tissue and the breast cancer microenvironment, participants of which were women ≥ 18 years undergoing breast surgery between 2009 and 2013 [110]. The NBS included 399 women with BC and 75 women without malignant disease, all of whom donated at least one histologically normal breast tissue specimen as determined by pathologists at UNC hospitals [110]. Demographic, risk factor, and clinical data were collected through a combination of telephone interview and medical records [110]. NBS data was used only for Aim 1 of this dissertation.

Genotypes in NBS were assayed using the Illumina Infinium OncoArray. We excluded variants with a genotype call rate below 98%, and then imputed to the TOPMed reference panel [102]. In TOPMed, phasing was performed using eagle and imputation using minimac4 [103,104]. From the TOPMed imputed data, we retained variants with high imputation quality ($r^2 > 0.8$) and variants which were in Hardy Weinberg Equilibrium (at $P < 1.0 \times 10^{-8}$).

Gene expression data in NBS was determined through microarray [110]. Details regarding the assay are described elsewhere [110]. Several quality control and processing steps were performed for NBS expression data gene expression data. NBS included individuals who provided multiple samples, including samples from tumor tissue in addition to histologically normal breast. Moreover, there were technical replicates for a given sample. We excluded samples with more than 30% missing data, averaged over technical replicates, and for a given individual chose the sample that corresponded to the most variation across the gene set. Missing data was imputed and 319 of the 396 CBCS genes were available for further analysis. Final expression data was in $\log_2$. The sample size of the genotype-expression reference panel for NBS was 93 EA and 37 AA.

Genotype Tissue Expression Consortium (GTEx)

GTEx is a collaborative effort started in 2010 with the objective of cataloging genetic effects on gene expression across a range of tissues [111]. GTEx gene expression was obtained from non-diseased tissues from rapid autopsy[111]. The latest release of GTEx (v8) consists of 838 postmortem donors and 17,382 expression samples (RNA-seq) from 52 tissues [111]. Whole-genome sequencing was conducted for each donor to a median depth of 32x and roughly 43 million variants have been catalogued after quality control [111]. Further details on quality control of the whole-genome sequence data and the RNA-seq data are available elsewhere [111].

For breast mammary tissue, there were 396 individuals with both genotype and expression data (337 EA; 47 AA). Although GTEx assays the whole transcriptome with RNA-seq, we limited our analysis to 365 of the 396 autosomal BC-related genes from CBCS that were available in GTEx. TPM normalized RNA-seq data were obtained, and final expression data was in $\log_2$. GTEx data was only used for predictive model construction in Aim 1 of this dissertation. Access to controlled GTEx data was under dbGaP project # 28275.

The Cancer Genome Atlas (TCGA)

TCGA is a collaborative effort started in 2005 with the objective of characterizing the genomic epigenomic, and molecular changes associated with cancer [112]. TCGA has genomic sequencing, expression, methylation, and copy number variation data on over 10,000 individuals across 30 plus cancers [112,113].

For BC, genotypes in TCGA were assayed using the Affymetrix 6.0 SNP array and RNA using Illumina RNA-seq platform. Further details on quality control of genotyping array and RNA-seq are available elsewhere [114,115]. As with NBS genotype data, imputation was performed against the TOPMed reference panel [102]. where phasing was performed using eagle and imputation using minimac4 [103,104]. From the TOPMed imputed data, we retained variants with high imputation quality ($r^2 > 0.8$) and variants which were in Hardy Weinberg Equilibrium (at $P < 1.0 \times 10^{-8}$). TCGA RNA-seq level 3 normalized data (normalized using upper quartile normalization) were downloaded from the Broad Institute GDAC

Firehose via FireBrowse [116]. While TCGA assays the whole transcriptome with RNA-seq, we limited our analysis to 378 of the 396 autosomal BC-related genes from CBCS that were available in TCGA. Final expression data was in $\log_2$ and there were 715 EA and 170 AA individuals with genotype and expression data available. TCGA data was used both for predictive model construction in Aim 1 of this dissertation and also for external validation of CBCS predictive models in Aim 2. Access to controlled TCGA data was under dbGaP project # 19922.

Genotype panel for imputation and association testing

Breast Cancer Association Consortium (BCAC)

Breast Cancer Association Consortium is a consortium with the aim of bettering understanding of the inherited risk of BC. There are ~ 100 studies that comprise BCAC and member studies submit information on study subjects, including demographics, clinical factors, risk factors, and genetics, which is then harmonized under the BCAC umbrella. Individuals in the BCAC have been genotyped on two custom platforms – the Illumina iSelect (iCOGS, ~200K SNPs) and the Illumina Infinium OncoArray (~550K SNPs) [2]. Imputation for both arrays was performed using the 1000 Genomes Project Phase 3 as a reference panel for two-step phasing and imputation (SHAPEIT2 and IMPUTEv2) [2]. In the BCAC, there are 217,413 BC cases and controls with genotyping data across the iCOGS or OncoArray platforms. We secured approval for use of BCAC genotyping data through the BCAC data access coordinating committee (BCAC approved project ID: 716).

3.2 Exposure

Across Aim 1 and Aim 2 of this study, the exposure variable was germline-regulated gene expression (GReX). To construct GReX, we adopted techniques from FUSION and PrediXcan [12,13]; specifically construction of GReX was a two-step process involving: 1) Construction of predictive models of gene expression from germline variants within 1 Mb of gene boundaries, followed by assessment of predictive accuracy by comparison of the sample size adjusted R^2 between predicted and observed expression of a gene; genes with predictive accuracy $R^2 > 0.01$ were selected for downstream GReX imputation and association testing with outcomes; 2) imputation (construction) of GReX involved a linear combination of a multiplication of SNP-gene weights from predictive models (step 1) and genotype dosages in BCAC (the imputation panel and also where association testing was subsequently performed).

The algorithm for predictive model training in Step 1, in more detail, is as follows:

(a) Gene expressions were first residualized for important covariates: Five genotype principal components computed in the full data (i.e., both EA and AA), age, BMI (where available), sex (GTEx only), and menopausal status (where available)

(b) For a given gene g, the predictive model was represented as

$$Y_g = X_g w_g + \varepsilon_g$$

where Y_g is the gene expression of gene g residualized on covariates, $X_{g.}$ is the genotype matrix of cis-SNPs (i.e., SNPs within 1 Mb of gene start and end position), w_g is the vector of effect sizes estimated, and ε_g is random error with mean 0 and common variance for all genes

(c) Estimation of w_g was done using one of two approaches (Elastic Net Regression [117], Best Linear Unbiased Predictor [118]) that yielded the best five-fold cross-validation R^2. Prior to estimation, we pruned the genotype matrix for LD using a LD threshold of 0.5 to avoid redundancy in estimated weights. Owing to the small sample size of some reference panels (e.g., AA GTEx = 48 individuals, AA NBS = 37 individuals), we applied a 0.05 minor allele frequency filter for these comparatively sparse reference panel. For all other reference panels, a 0.01 minor allele frequency was applied. For Aim 1, we constructed GReX (of genes with high predictive accuracy) for 8 reference panels across two ancestry groups and breast tissue types (two most pertinent tissue contexts for BC risk in normal mammary tissue and breast tumor tissue): GTEx –EA (n=337), GTEx – AA (n=47), NBS – EA (n=93), NBS – AA (n=37), CBCS – EA (n=571), CBCS – AA (n=628), TCGA – EA (n=715), TCGA – AA (n=170). For Aim 2, we constructed GReX for 4 reference panels (across ancestry and subtype groups) in CBCS: EA – LL (n=410), EA – BL (n=116), AA – LL (n=358), AA – BL (n=224).

3.3 Outcome

In Aim 1, the outcome of interest was breast cancer subtypes, defined two ways. Molecular subtype was defined as Lumina-like ((ER+ / PR+) and HER2-, (ER+ / PR+) and HER2+), (ER-, PR-, HER2+)) or Basal-like (ER-, PR-, HER2-). Molecular subtype was the primary outcome of interest in Aim 1 GReX association analyses and association analyses for molecular subtype was carried out in individuals of European and African ancestry (see Table 4.1 for full details on the ancestry-specific sample sizes). Etiologic subtype was a secondary outcome of interest in supplementary Aim 1 GReX association analyses. Etiology subtype was defined as ER- and TP53-, ER- and TP53+, ER+ and TP53-, and ER+ and TP53+. TP53 positivity was determined by a score of 2 or more on a 0-3 scale (where 0 indicated no staining and 1 indicated less than 10% staining). Analyses for etiologic subtype were limited to individuals of European ancestry due to insufficient sample sizes (n=0) for such analyses among the African ancestry population (see Table 4.3 for full details on sample sizes).

In Aim 2, the outcome of interest was BC mortality, and specifically both all-cause and BC-specific mortality. Deaths were coded according to the 10th revision of the International Classification of Diseases (ICD-10-WHO). For BC-specific mortality, individuals who died of any cause other than BC were considered right censored. Detailed information on number of mortality events across the analytic sample are available in Figure 5.4 and Figure 5.5.

3.4 Statistical Analysis

As mentioned in 3.2, GReX (exposure variable) was computed only for genes that showed a sample-size adjusted R^2 of > 0.01 between observed and predicted expression. Although some gene-based association testing studies further select genes based on heritability we found heritability estimation challenging and imprecise for some of our reference panels where n < 100 for the reference panel. We note here that most TWAS and TWAS based studies select genes for downstream analyses based on predictive accuracy alone [65,119-122].

In Aim 1, we performed multinomial logistic regression for GReX against molecular and etiologic subtype definitions (ref. control status). GReX imputation and association testing were aligned by ancestry, meaning that association analyses were performed among individuals of African ancestry for predictive models trained among individuals of African ancestry. Analyses were also conducted separately for the iCOGS and OncoArray sub-samples within each ancestry group and array specific effect estimates (corresponding to 1 standard deviation increase in GReX of a given gene) were pooled using fixed effects, inverse variance weighted meta-analysis, in line with other genomic investigations in BCAC data [4,7,64]. We adjusted for age in the multinomial models; adjustment for genetic ancestry, age, and other relevant variables such as BMI, sex (GTEx), and menopausal status had already been performed at the level of predictive model construction. We employed *bacon*, a Bayesian approach to control for test statistic inflation, and then applied a global FDR threshold of 0.05 on the *bacon* corrected p-values to determine statistical significance (statistically significant genes were termed GReX-prioritized genes) [123]. We performed follow-up analyses where we assessed for: 1) genomic overlap among GReX-prioritized genes (if overlap was present, we assessed GReX correlations and subsequent conditional analyses); 2) genomic overlap with established GWAS signal (for overall BC vs. controls as GWAS signals for BC subtype have not been established (i.e., determined as the putative causal signal); 3) PANTHER pathway overrepresentation analyses (for both subtype-specific GReX-prioritized genes as well as GReX prioritized genes by normal breast vs. breast tumor tissue). We note here that we performed mutual external validation of the predictive models used for association analyses in Aim 1, by ancestry and tissue type groupings (e.g., predictive models constructed in NBS (normal breast tissue) for EA were assessed in GTEx (normal breast tissue) EA individuals.

In Aim 2, we performed Cox Proportional Hazards Regression for GReX against time to all-cause and BC-specific mortality. Age was chosen as the time scale for the study due to different entry times into analysis across component BCAC studies (left truncation) and also because of variable follow-up lengths across studies. We allowed baseline hazards to vary by BCAC study in Cox models, both to adjust for potential confounding and also effect heterogeneity across BCAC studies. Adjustment for genetic ancestry, age, BMI, and menopausal status had already been performed at the level of predictive model construction. Effects were estimated for 1 standard deviation increase in GReX. In main GReX analyses, predictive models and imputation/association testing were aligned by ancestry and subtype, whereas in supplementary GReX analyses, predictive models and imputation/association testing were aligned by ancestry alone. Sample sizes across both the main approach and the approach in supplementary analyses are provided in Figures 5.4 and 5.5. As with Aim 1, analyses were conducted separately for the iCOGS and OncoArray sub-samples (within the European ancestry – Luminal-Like subtype strata of BCAC for example) and array specific effect estimates (corresponding to 1 standard deviation increase in GReX of a given gene) were pooled using fixed effects, inverse variance weighted meta-analysis. Significance was set at a global (all association tests performed for main GReX analyses, for example) FDR <0.05. For Aim 2, we also performed a slew of analyses to afford greater context to GReX analyses. We : 1) assessed ancestry and subtype-specific eQTLs; 2) conducted a formal heterogeneity test for differential germline-regulated tumor expression across subtypes, within each ancestry group; 3) assessed portability of predictive models across subtype, within ancestry group; 4) performed external validation in TCGA for ancestry and subtype-specific predictive models.

CHAPTER 4: AIM1. EVIDENCE FOR GENE-LEVEL GERMLINE ASSOCIATIONS FOR BREAST CANCER MOLECULAR SUBTYPES AMONG WOMEN OF EUROPEAN AND AFRICAN ANCESTRY IN THE BREAST CANCER ASSOCIATION CONSORTIUM (BCAC)

4.1 Introduction

Breast cancer (BC) is the second most common cancer among women in the United States (U.S.) with an estimated 268,600 invasive cases diagnosed in 2019 [1]. BC is a heterogeneous disease, spanning multiple subtypes defined by either receptor status (Estrogen Receptor (ER) positivity, Progesterone Receptor (PR) positivity, HER2 positivity, TP53 positivity) or the PAM50 gene expression based classifier [124]. Studies have shown marked differences in recurrence and mortality risk by BC subtype, making BC subtype one of the most critical prognostic markers for BC patients [15,125]. In the most recent GWAS of BC among individuals of European ancestry (which reported 210 genome-wide significant loci), several identified loci showed heterogeneity in effect across BC subtypes (including some loci where direction of effect was reversed across subtypes); moreover, 15 of the 32 identified novel loci showed significant associations with subtype defining markers such as ER, PR, and HER2 [2]. This suggests that important subtype-specific associations may be masked in aggregation of association testing of BC cases overall vs. controls. Knowledge of overall- and subtype-specific germline underpinnings of BC can inform targeted public health initiatives towards mitigation, an especially important challenge in prevention of aggressive subtypes such as Basal-like (BL).

While GWAS offers a robust approach for investigation of the role of germline genetics in BC subtype etiology, two limitations to this approach are sample size requirement and often, lack of biological interpretability of findings. The sample size limitation, in particular, precludes investigations across ancestrally diverse populations, and in among particular subpopulations (to date, the largest GWAS among women of African ancestry has found no genome-wide significant loci [4]), who have a disproportionately higher burden of more aggressive BC subtypes [15,126-128]. Gene-based association testing approaches such as Transcriptome-Wide Association Study (TWAS) offer an alternative to GWAS that may overcome some of these limitations [12,13]. By aggregating germline effects to the gene level, TWAS allows for reduced multiple testing burden and improved statistical efficiency, allowing discovery at loci that might otherwise be missed in a GWAS. In other cases, TWAS can also contextualize GWAS findings by prioritizing genes that serve as a potential biological mechanism for variant to phenotype associations [119,120].

In this study of the germline-regulated gene expression (GReX; TWAS methodology) of genes in relation to BC subtype across European and African ancestry individuals in the Breast Cancer Association Consortium (BCAC), we address many of the aforementioned gaps in knowledge. We leverage four genotype-expression panels across normal (Genotype Tissue Expression Project (GTEx), Normal Breast Study (NBS)) and breast tumor tissue (Carolina Breast Cancer Study (CBCS), The Cancer Genome Atlas (TCGA)) in an effort to capture the spectrum of potential germline effects across healthy and diseased mammary tissue. Across the four reference panels, which includes among the largest reference panels for diverse individuals in CBCS, we generate ancestry-predictive models of gene expression, which we then used for ancestry-aligned imputation and association testing in BCAC for molecular subtype (ER/PR/HER2 based; both ancestries) and etiologic subtype (ER/TP53 based; exploratory analyses among individuals of European ancestry only).

4.2 Methods

Study Population – Genotype and gene expression reference panels

In this study, we used four genotype and gene expression reference panels for construction of predictive models of gene expression from germline genetics; two reference panels (GTEx, NBS) were in normal/healthy breast tissue and the other two (CBCS, TCGA) were in breast tumor tissue. Details about the studies underlying the genotype and gene expression reference panels have been previously published [34,94,110-115,126].

Genotypes across GTEx, NBS, CBCS, and TCGA were assayed using WGS, Illumina Infinium OncoArray, Illumina Infinium OncoArray, and Affymetrix 6.0 array, respectively. To ensure consistency across genotype imputation (with the exception of GTEx, which was WGS), we imputed NBS, CBCS, and TCGA genotype data to the TOPMed reference panel [102]. Briefly, in TOPMed, phasing was performed using eagle and imputation using minimac4 [103,104]. For imputed genotype data, we retained variants with high imputation quality ($r^2 > 0.8$) for predictive model construction. In further quality control of genotype data, we excluded variants significantly deviating from Hardy-Weinberg Equilibrium (at $P < 1.0 \times 10^{-8}$) and also excluded variants with MAF <0.01 (for GTEx and NBS genotype data among BW, MAF inclusion threshold was increased to 0.05 due to limited sample sizes for downstream predictive model construction).

Gene expression across GTEx, NBS, CBCS, and TCGA was assayed using RNA-seq, microarray, NanoString nCounter, and RNA-seq, respectively. For each expression panel, data obtained had either already been normalized and ascertained for quality control or normalization and quality control steps were performed, as appropriate. Downloaded GTEx and TCGA RNA-seq data had been normalized using the TPM and upper-quartile normalization, respectively [116]. For CBCS, upper-quartile normalization was applied, and two further dimensions of unwarranted biological and technical variation were removed using RUV-seq [129,130]. For NBS, data were normalized to sample over control. The sample sizes for the reference panels were as follows: GTEx (n=337 EA, 47 AA), NBS (n=93 EA, n=37 AA), CBCS (n=571 EA, n=628 AA), and TCGA (n=715 EA, n=170 AA). For the 396 BC-related genes available in CBCS that were assessed in this gene-level germline investigation, 378, 365, and 319 of those 396 genes had expression data available across TCGA, GTEx, and NBS, respectively. While our use of CBCS data limited our investigation across other reference panels to the set of CBCS genes, CBCS offers among the largest resource of genotype-expression data for individuals of African ancestry (n=628 AA). At the same time an advantage of the test gene set of 396 was reduced multiple testing burden and increased biological plausibility and interpretability (all 396 genes are protein coding genes that have been implicated in key signatures such as *PAM50*, *P53*, *IGF*, and *EGFR* (Appendix 1).

Study Population – Breast Cancer Association Consortium

Genetic association testing was performed in BCAC, which is a consortium comprising roughly 100 studies; the consortium offers harmonized demographic, clinical, and genetic data. Individuals in BCAC have been genotyped on the Illumina iSelect (iCOGS) and Illumina Infinium OncoArray (OncoArray) platforms. Phasing and imputation across both genotyping arrays was performed using SHAPEIT2 and IMPUTEv2 with the 1000 Genomes Project (v3) as a reference panel [97,99,101,131,132]. In this study, we examined only individuals of European and African ancestry in BCAC.

Construction of predictive models of gene expression from cis-germline variation

In this study, GReX was the exposure of interest. For a given gene, GReX represents the portion of tumor expression explained by *cis*-germline variation (where *cis*-regions to gene span the 1 Megabase surrounding the gene's start and end position). We constructed GReX using TWAS methodology, where we first trained predictive models of gene expression from *cis*-germline variation using Elastic Net Regression (EN) with five-fold cross validation and Best Linear Unbiased Predictor (BLUP) [12,13,129]. The approach yielding the better predictive accuracy, defined as sample size adjusted squared correlation between predicted and observed expression was chosen as a given gene's predictive model. We pruned variants prior to model training using PLINK v1.9 (pruning parameters: window size of 50 base pairs, window shift of 5, and LD threshold of 0.5) to avoid redundancy in the predictive model. Importantly, we constructed predictive models separately for individuals of European and African ancestry across the four reference panels (GTEx, NBS, CBCS, TCGA), in line with prior findings that predictive models of gene expression from germline variation have poor portability across ancestry groups in both breast tumor and healthy tissues [16,133]. In predictive model construction (not, however, in GReX imputation and association testing) race was used as a proxy for genetic ancestry, in line with previous genetic investigations that leverage a TWAS methodology [16,119,120,133]. Methods that allow construction of predictive models across the spectrum (continuous range) of genetic ancestry are not yet developed. Lastly, we residualized expression based on the first five principal components of the combined (i.e., across White plus Black individuals for each reference panel) genotype matrix, age, and other demographic factors (e.g., body mass index (BMI)), where available.

The sample sizes for the genotype-expression reference panels for predictive model construction were as follows: GTEx (n=337 EA, 47 AA), NBS (n=93 EA, n=37 AA), CBCS (n=571 EA, n=628 AA), and TCGA (n=715 EA, n=170 AA). Importantly, for both ancestry groups, we constructed predictive models across normal breast and breast tumor tissue as we hypothesize these to be the two most relevant tissue types for investigation of germline effects on expression for BC (including BC subtypes)[134,135]. Additionally, we used two reference panels of the same ancestry and tissue type to enable aspects such as mutual external validation of predictive models, a feature lacking in most TWAS-based investigations. For downstream association testing, genes with sample size adjusted cross-validation squared correlation (i.e., predictive accuracy) greater than 0.01 between observed and predicted expression were selected. In some studies, an additional filter based on gene expression heritability is applied for further selection [16,119,133,136]; however, estimation of heritability was difficult and imprecise across some of our reference samples (e.g., n=47 AA in GTEx, n=37 AA in NBS, and n=93 EA in NBS, and N=170 AA in TCGA). Notably, most TWAS and TWAS-based studies select genes for downstream analyses based on predictive accuracy alone [65,119-122].

Mutual (external) validation of predictive models, by tissue type

A strength of this study is that we leveraged multiple expression panels across the two tissue types of interest (normal breast, breast tumor). This enabled us to determine the predictive accuracy of our models in the corresponding reference panel, by tissue type and ancestry group (i.e., predictive accuracy of EA GTEx predictive models against EA gene expression in NBS, predictive accuracy of EA CBCS predictive models against EA gene expression in TCGA etc.). Predictive accuracy, as example, for EA GTEx validation in NBS, was defined as squared correlation between predicted NBS expression using GTEx predictive model (SNP-gene weights) and NBS gene expression, for a given gene. For each of the eight reference panels, validation of predictive models was performed, aligned by ancestry and tissue type. For example, predictive models created in normal breast tissue (GTEx, NBS) for EA individuals were mutually validated (i.e., models created in GTEx validated in NBS and vice-versa).

GReX imputation and association testing

Imputation of GReX (for genes with high predictive accuracy across reference panels) was performed in the Breast Cancer Association Consortium (BCAC). Ancestry and tissue-specific GReX of a given gene was imputed by multiplying the genotype dosage in BCAC with the SNP-gene expression weights from that gene's ancestry and tissue-specific predictive model. Imputation was performed separately for the iCOGS and OncoArray genotyped samples in BCAC. Once GReX was imputed, we employed multinomial logistic regression to evaluate associations between 1 standard deviation increase in GReX and odds of breast cancer subtype (ref. controls). Main association analyses were for molecular subtype definition (Luminal-like (LL), Basal-like (BL)); we additionally performed exploratory analyses for etiologic subtype definition (ER-/TP53-, ER-/TP53+, ER+/TP53-, ER+/TP53+), which were limited to individuals of European ancestry due to sample size constraints. The rationale for investigation of etiologic subtype definition as a secondary, exploratory analysis was that a prior study in CBCS found that etiologic subtype definition had higher etiologic heterogeneity compared to the traditional, molecular subtype definition[14]. We note, however, that we were underpowered (for etiologic subtype definition) for both discovery of GReX-prioritized genes and a formal statistical comparison against findings for molecular subtype. The multinomial models were adjusted for age (adjustment for principal components of the genotype matrix had already been performed at the predictive model construction stage). As with GReX imputation, GReX association testing was conducted separately across the iCOGS and OncoArray genotyped samples; array specific estimates were then combined using fixed effects, inverse variance weighted meta-analysis. As TWAS-based GReX analyses can be prone to inflation of test statistics, we employed a Bayesian bias and inflation adjustment method (bacon) across the meta-analyzed effect estimates. We then adjusted for multiple testing using the Benjamini-Hochberg procedure. We defined genes passing a global (n = all association tests performed across ancestry and tissue type) false discovery rate (FDR) threshold of 0.05 as GReX-prioritized genes [123,137]. In follow-up analyses to GReX-prioritized genes, we assessed: 1) Genomic overlap between GReX-prioritized genes and established GWAS-significant variants for BC (established GWAS-significant variants for BC subtype definitions used in this study were not available); 2) PANTHER pathways statistical overrepresentation (through binomial test), comparing LL and BL specific GReX-prioritized genes, as well as normal breast and breast tumor identified GReX-prioritized genes [138].

4.3 Results

Across the ancestry and tissue-specific reference panels, the following number of genes had high predictive accuracy and were chosen for further GReX association testing: GTEx – EA (n=129 genes), GTeX – AA (n=261 genes), NBS – EA (n=146 genes), NBS – AA (n=185 genes), CBCS – EA (n=86 genes), CBCS – AA (n=65 genes), TCGA – EA (n=109 genes), and TCGA – AA (n=236 genes). The mean and interquartile range of the CV R^2 for each ancestry, tissue-specific reference panel are provided in Table 4.1. The mean predictive accuracy was higher for predictive models constructed in normal breast tissue as opposed to tumor tissue, although some of this is likely due to overfitting among the comparatively smaller sample sizes of reference panels for normal breast versus tumor size, especially among AA individuals. In further assessment of model performance constructed across the same ancestry group and tissue type, we found that the use of complementary data sources added value beyond the use of a singular data source, as seen in other TWAS based studies involving multiple reference panels [12,119]. For instance, for models trained among European ancestry individuals in normal breast tissue, there were 82 genes with high predictive accuracy unique to GTEx and 99 unique to NBS; 47 genes had high predictive accuracy across both reference panels, where we additionally observed a statistically significant correlation in the magnitude of predictive accuracy, despite differences in expression quantification method (RNA-seq vs. Microarray) (Pearson's r = 0.46, p=0.001). A similar trend was observed for models trained among tumors from European ancestry individuals (Table 4.2). However, for models trained among African ancestry individuals, we did not observe a correlation in predictive accuracy among genes that were highly correlated, likely due to differences in expression quantification method and higher burden of somatic alterations (for models constructed in tumor tissue), which were not accounted for in predictive models [139].

For molecular subtype association analyses, our analytic sample (after exclusions based on non-invasive case status and missingness of tumor markers (ER, PR, and HER2)) comprised 146,177 EA (91,101 controls, 47,678 LL, 7,398 BL) and 5,092 AA (2,885 controls, 1,505 LL, 702 BL) individuals across iCOGS and OncoArray samples (Table 4.1). Analytic sample sizes for etiologic subtype association analyses among EA individuals are provided in Table 4.3.

External Validation

We performed mutual external validation for the ancestry and tissue-specific predictive models (i.e., models constructed for EA in GTEx (normal breast) tissue were assessed against gene expression in NBS (normal breast) EA samples). We found moderate to good performance for our predictive models (Table 4.4). External predictive performance was higher for models constructed in normal breast tissue compared to tumor tissue (roughly ~ 50% of genes had high predictive accuracy (R^2 > 0.01) in normal breast tissue compared to roughly ~45-46% of genes having the same in breast tumor tissue). External predictive accuracy between EA and AA models was comparable across tissue types, although external predictive performance was slightly better for EA compared to AA models (Table 4.4).

Associations between ancestry, tissue-specific GReX and BC molecular subtype

Across the ancestry and tissue reference panels, we observed 53 GReX – BC molecular subtype associations spanning 40 unique GReX-prioritized genes at global FDR <0.05 (Table 4.5, Figure 5.1). All 53 of the GReX – BC molecular subtype associations were among EA individuals, although we did note suggestive associations (FDR <0.20) among AA individuals for several genes such as *GPR44* (LL), *AURKA* (LL), *PSPHL* (BL), *IL6* (LL), and *CRYAB* (BL) (Table 4.6) Of the 40 unique GReX-prioritized genes at global FDR <0.05, 34 were unique to LL and 6 unique to BL (Table 4.5, Figure 5.1). We noted 10 GReX-prioritized genes (*ABAT, DDR1, PDSS1, VAV3, ZG16B, SLC16A3, TUBA4A, C14orf45, FAM214A, and ZAP70*) with a statistically significant association for both LL and BL subtype (e.g., *DDR1*: LL Z-statistic (6.61), BL Z-statistic (5.61) ; in all instances, the direction of effect for the subtypes was similar, although there were differences in magnitude of association (Table 4.5, Figure 5.1).

Three GReX-prioritized genes (*C4A, KCNN4, UGT2B7*) were significant for the same subtype (LL) across different reference panels (Table 4.7). For *C4A*, the Z-statistics were 5.09 and 4.96 across GTEx and CBCS while for *UGT2B7*, the Z-statistics were -6.04 and -4.07 for GTEx and TCGA (Table 4.7). Given that the biological function of a gene as it relates to conferring or mitigating risk to a subtype is expected to be consistent across reference panels (especially so for reference panels of the same tissue type), the consistency in signal for LL risk for *C4A* and *UGT2B7*, despite differences in expression assay, makes these two genes particular candidates for further functional follow-up. We do note, however, that for *KCNN4*, we observed effects in opposing directions across reference panels (Z-statistics for CBCS and NBS were 4.98 and -6.00, respectively). We followed up on this counter-intuitive finding by assessing the distribution of the GReX of the gene across the reference panels, finding that the IQR for KCNN4 GReX was 4.62 x 10^{-7} and 7.86 x 10^{-8} across CBCS and NBS, respectively. By comparison, the IQR for *C4A* GReX was 0.23 and 0.21 across GTEx and CBCS while the IQR of *UGT2B7* GReX was 0.28 and 0.18 across GTEx and TCGA. It is possible that the lack of range for *KCNN4* compared to that for *C4A* and *UGT2B7* could have contributed to conflicting associations across reference panels. Therefore, as further context for GReX associations observed in this study, we provide GReX mean and IQR for all genes that were tested for an association across ancestry and tissue groups.

We observed that 2 of our GReX-prioritized genes (*C4A, HLA-DOB*) were concentrated at a genomic locus; however; the GReX of these genes did not demonstrate significant correlation. For the remainder of the GReX-prioritized genes, there was no genomic overlap (defined as one gene being within 1 Mb of another gene's start or end site). In analyses of genomic overlap between established GWAS (for BC case vs. control) loci and GReX-prioritized genes, we found that 7 of the 40 GReX-prioritized genes were within 1 Mb of an established GWAS loci; for 4 of these 7 GReX-prioritized genes (*MUC1, CCNE1, HLA-DOB,* and *KCNN4* we found concordance in the direction of effect between proximal GWAS signal and GReX-prioritized genes) (Table 4.8). 3 of these genes (*MUC1*, *CCNE1*, and *KCNN4*) have been hypothesized as predicted target genes in prior BC GWAS [2,8]. In PANTHER pathway overrepresentation analysis of LL and BL GReX-prioritized genes for the EA individuals, we found no significantly overrepresented pathways at FDR <0.05; however, we did observe that the P53 pathway was the most overrepresented (FDR = 0.12) among BL GReX-prioritized genes while this was not the case for LL GReX-prioritized genes (FDR for p53 pathway = 1.00). Similarly, we found that GReX-prioritized genes among breast tumor tissue showed suggestion of overrepresentation for the P53 pathway (FDR = 0.15), while this was not the case for GReX-prioritized genes among normal breast tissue. Lastly, we did not find any genes whose GReX was associated with etiologic subtypes at global FDR <0.05; moreover none of the genes showed suggestive associations (defined as FDR <0.2, largest association was for CDH3 in relation to ER-/TP53+ subtype, FDR = 0.24).

4.4 Discussion

In this study, we leveraged TWAS methodology and performed a GReX analysis for BC subtypes across healthy and breast tumor tissue in the largest available sample of BC subtypes and controls in BCAC. To our knowledge, this is the first GReX analysis for a BC phenotype (in this case, BC subtype) among individuals of African ancestry, enabled by our use of non-publicly available reference panels with a large number of individuals with African ancestry . We initially employed a Bayesian bias and inflation adjustment method in *bacon* to correct for inflation of test statistic, following which, we employed a conservative correction for multiple comparisons with a global (all association tests performed across ancestry and tissue types) FDR threshold for GReX-prioritization of 0.05. Out of the 396 genes tested across multiple reference panels, global FDR <0.05, we found 53 GReX – BC molecular subtype associations spanning 40 unique GReX-prioritized genes, of which 34 genes were unique to LL, 6 unique to BL, and 10 common across LL and BL. All 40 GReX-prioritized genes were found among EA individuals. For AA individuals, we found 5 suggestive associations (FDR <0.20, range 0.06 – 0.18). In assessment of genomic overlap between GReX-prioritized genes and GWAS signals (for BC), 7 of the 40 GReX-prioritized genes overlapped with the GWAS signal, 4 (*MUC1, CCNE1, HLA-DOB,* and *KCNN4*) of which showed the same direction of association as the GReX signal; *MUC1*, *CCNE1* and *KCNN4* had previously been hypothesized to be target genes for the GWAS signals.

Identified effects of many of our GReX-prioritized genes are in line with literature, although we do note that the literature spans the effect of the total expression whereas our findings are particular to the germline-regulated portion of gene expression. Effects for one do not necessarily imply the same effect for the other [12,140]. We first focused on GReX-prioritized genes that are shared across LL and BL subtypes, as we hypothesized these genes likely impact global carcinogenic processes that impart BC risk, irrespective of subtype. In this study, we find that increased germline-regulated expression of *DDR1* (discoidin domain receptor 1) is associated with increased risk of both LL (Z=6.61) and BL (Z=5.61) subtypes. *DDR1* is a collagen receptor with tyrosine kinase activity that has been shown to instigate immune exclusion via collagen fiber alignment in the extra-cellular matrix (ECM) [141]. In BL BC expression of *DDR1* negatively correlates with intratumoral concentration of anti-tumor T cells; moreover, ablation of *DDR1* in tumors has been shown to enhance intratumoral T cell penetration and destruction of tumor in mouse models [141]. We find similar strong evidence for *PDSS1* (Z = 4.45 LL, Z = 4.35 BL), knockdown of which among BL cells has been shown to inhibit BL cell migration, proliferation, and metastasis [142]. *PDSS1* has been shown as a key activator of *CAMK2A* and *STAT3* in the *PDSS1/CAMK2A/STAT3* oncogenic signaling axis [142]. Another study showed that silencing of *ZG16B* (Z = 4.34 LL, 4.29 BL) (*PAUF*; pancreatic adenocarcinoma upregulated factor) inhibited proliferation, and induced apoptosis and G0/G1 cell cycle arrest among colorectal cancers [143]. None of *DDR1*, *PDSS1*, and *ZG16B* have been implicated as putative causal genes for BC or BC subtype in prior genetic investigations [120].

We found four GReX-prioritized genes (*MUC1, CCNE1, HLA-DOB,* and *KCNN4*) with concordant effect direction and overlap with BC GWAS loci. Although *MUC1*, *CCNE1*, and *KCNN4* were predicted target genes for the BC GWAS loci in prior investigations, this study adds to that existing knowledge by offering statistical evidence for these genes as a potential genetic mechanism (i.e., BC GWAS loci → Gene Expression -> BC (BC subtype)) for those BC GWAS loci. There is extensive support for MUC1 as an oncogene across multiple tumors, including BC [144-146]. Comparatively, evidence for *CCNE1* and *KCNN4* is sparse. Two GReX-prioritized genes (*C4A*, *UGT2B7*) were identified for LL risk across multiple reference panels. C4A encodes a protein in the complement pathway, and there is conflicting evidence (albeit more recent studies are in favor of a deleterious role) for *C4A* across disease phenotypes, including tumors, degenerative, and inflammatory disease [147,148]. Studies indicate that higher levels of *C4A* interfere with normal metabolic processes/ insulin signaling and are associated with significantly increased risk of developing metabolic syndrome [149,150]. *UGT2B7* is a phase II metabolism protein that is involved in removal of potentially toxic xenobiotic as well as endogenous compounds, a function consistent with the significantly decreased risk of LL that we observed for increased GReX OF *UGT2B7* [151].

As one of the motivating factors behind this study was heterogeneity in effect across subtypes for certain GWAS variants, we additional assessed the direction and magnitude of effect across the corresponding subtype (i.e., the effect for BL subtype for a GReX-prioritized gene for LL subtype in a given reference panel) (Figure 4.2). Here we find that most GReX-prioritized genes have a similar direction of effect across subtypes, although there are often stark differences in magnitude (e.g., *KNCMA1*,*HLA-DOB*, *DDIT4*, *LHPF* etc.). For genes such as *PPBP*, however, we note differences in direction of effect. There are a few possible reasons for the lesser heterogeneity in subtype-specific estimates observed in our study compared to the BC GWAS. First, the BC GWAS spanned the human genome while we assess the GReX of a targeted panel of BC-related genes, and in particular, our panel emphasized genes with important roles in survivorship rather than etiology. Secondly, our outcome comparison groups were less granular (i.e., LL and BL subtype vs. controls) compared to the BC GWAS which assessed effects across the full spectrum of subtypes. Our choice of outcome groupings, as detailed in the Methods, was because our study assessed subtype-specific effects across ancestry groups (subtypes such as HER2 are sparse in the African ancestry population); moreover, we were specifically interested in the BL versus non BL (LL subtype) as BL is the most aggressive and most challenging BC subtype to treat. We considered the possibility that the differences in magnitude of subtype-specific effects may be an artifact of differences in sample size across LL and BL groups. While this is a possibility and potentially a reason behind observed differences in effect magnitude, we note that the effect was higher for BL versus LL for several GReX-prioritized genes (e.g., *EPCAM, CD84, CDKN3, CCNE1, CDKN1A, ZAP70*), which suggests that differences in sample size alone are not responsible for differences in effect magnitude across subtypes.

Although we did not find any GReX-prioritized genes at global FDR <0.05 for individuals of AA, we found several (5) that showed suggestion of an association (global FDR <0.20), namely *AURKA*, *CRYAB*, *GPR44*, *IL6*, and *PSPHL*. *GPR44* and *AURKA*, in particular showed moderate evidence at global FDR of 0.06 and 0.11, respectively. For *AURKA* (Aurora kinase A), we observed increased risk (Z = 3.63, P = 2.8 x 10^{-4}), consistent with *AURKA*'s involvement/ interaction with proteins in the KEGG pathway, many of which are in turn involved in oncogenic pathways [152]. Here, we demonstrate the utility of gene-based association testing approaches in helping bridge some of the gaps in genetic findings across diverse ancestries. Despite a 30-fold difference in association analyses sample size (146,177 EA, 5,092 AA) and lesser sample size of the reference panels across all data sources except CBCS, TWAS enabled findings with moderate evidence for further follow-up. In comparison, the most recent and largest GWAS (~ 20,000) among individuals of AA found no loci near genome-wide significance (largest hit P-value = 5.2 x 10^{-4}) [10].

We note several limitations to our study. First, CBCS used a custom NanoString nCounter probeset for RNA expression quantification of BC-related genes, and therefore we were not able to analyze the whole transcriptome. However, this panel had high success rates and very few sample failures (<5%) and therefore allowed for expression quantification that was relatively free of selection bias due to sample drop out. Similarly in microarray based NBS expression data, we were not able to assay the whole transcriptome. This limited our investigation to 396 BC-related genes available in CBCS; however, we note that inclusion of CBCS and NBS data was pivotal towards the aims of this investigation for several reasons. The use of CBCS data was key towards building African ancestry specific models of breast tumor gene expression, as CBCS offers, to our knowledge, the largest available resource (in terms of sample size) of tumor transcriptomic data in this population; CBCS reference panel for AA was 628 individuals, compared to 170 in TCGA, 47 in GTEx, and 37 in NBS. A second limitation is that like most GWAS analyses, CBCS lacked data on somatic alterations and epigenetic changes (something the TCGA offers), which could improve predictive model performance. Third, although both NBS and GTEx offer expression data in histologically normal breast tissue, individuals in NBS were mostly individuals with tumors (donating histologically normal breast tissue adjacent to the cancer), while individuals in GTEx donated non-diseased tissue and were identified through rapid autopsy programs. However, our finding that P53 pathway genes were identified from tumor-based GReX and not normal-based suggests that some detection may be context dependent and using both tumor and normal to understand the range of pathway factors may have value. Fourth, although care was taken to ensure analytic consistency across the reference panels (e.g., we re-imputed CBCS, TCGA, and NBS data to the TOPMed panel), there were differences in expression quantification platforms (e.g., RNA-seq, microarray, NanoString) that make full comparability of findings across reference panels difficult. However, our normalization methods were standard for the field. Fifth, although our subtype categorization is aimed at elucidating differences in potential germline genetic basis of BL (most aggressive) versus LL subtypes (less aggressive), this classification is not fine-tuned enough for genes such as *ERBB2*, whose GReX we expect to be positively associated with HER2 subtype. However, since HER2 subtype was classed among the non-BL (LL) category along with LumA and LumB subtypes, this specific association can be masked by LumA and LumB specific-associations in estimation of the LL association. We note here that our choice of BL v LL comparison was in part, also motivated by lack of sample size for the rarer subtype categories (e.g., HER2) among AA individuals. Lastly, we note the lesser sample sizes we had for the GReX-analysis among AA compared to EA, both for predictive model construction (with the exception of CBCS) and for association testing in BCAC. This could lead to more uncertainty in the GReX and could contribute to the smaller number of hits in AA participants. Interpretations should be made with caution for the suggestive associations uncovered for AA, and we additionally caution against a direct, one to one comparison of findings across ancestry groups in light of the sample size discrepancies. However, these results also underscore the importance of increasing the size of reference gene expression data in AA ancestry.

In conclusion, we find several genes whose germline-regulated expression across normal breast and breast tumor tissues are associated with risk of BC subtypes among EA individuals, and a few genes with suggestion of association with risk among AA individuals. For EA individuals, we demonstrate shared (i.e., gene associated with both LL and BL) as well as potentially divergent germline genetic basis for BC subtypes, which has relevance for more targeted risk stratification and potentially even targeted clinical efforts to reduce burden of BC outcomes. Future studies should further this line of investigation across other etiologic tissue types (i.e., fibroblasts, adipose tissue, immune tissue etc.) and for the full spectrum of BC subtypes (LumA, LumB/HER2-, LumB, HER2, Basal). Our work also demonstrates potential utility of TWAS based approaches in helping bridge some of the gaps in genetic investigations across ancestries, serving as a framework for future investigations in this mold, and in doing so, elucidating the need for larger, diverse datasets such as the CBCS.

4.5 Tables and Figures

Table 4.1 Number of genes tested for associations by reference panel and ancestry group, along with sample sizes for GReX association analyses (overall), and by genotyping array

Reference Panel	Ancestry Group	n models tested	Mean, IQR of CV R2	iCOGS	OncoArray	iCOGS Cont.	iCOGS LL	iCOGS BL	OncoArray Cont.	OncoArray LL	OncoArray BL
GTeX	EA	129	0.031, 0.015	53775	92402	37059	14105	2611	54042	33573	4787
GTeX	AA	261	0.068, 0.061	942	4150	817	122	3	2068	1383	699
NBS	EA	146	0.048, 0.038	53775	92402	37059	14105	2611	54042	33573	4787
NBS	AA	185	0.097, 0.096	942	4150	817	122	3	2068	1383	699
CBCS	EA	86	0.019, 0.008	53775	92402	37059	14105	2611	54042	33573	4787
CBCS	AA	65	0.018, 0.007	942	4150	817	122	3	2068	1383	699
TCGA	EA	109	0.025, 0.012	53775	92402	37059	14105	2611	54042	33573	4787
TCGA	AA	236	0.031, 0.023	942	4150	817	122	3	2068	1383	699

Abbreviations: GTEx – Genotype Tissue Expression Project, NBS – Normal Breast Study, CBCS – Carolina Breast Cancer Study, TCGA – The Cancer Genome Atlas, EA – European ancestry, AA – African ancestry, IQR – Interquartile Range, CV – Cross Validation, LL – Luminal-like, BL – Basal-like, GReX – Germline-regulated gene expression

Table 4.2 Number of shared predictive models (i.e., models for genes with high predictive accuracy in a given reference panel) and corresponding correlation coefficients for predictive performance, by ancestry and tissue groups

Ancestry Group	Tissue Type	Data 1	Data 2	Predictive Models Data 1	Predictive Models Data 2	n Shared Predictive Models	Pearson's r (Shared)	P-value Pearson's r for shared genes
EA	Normal	GTEx	NBS	129	146	47	0.46	0.0011
AA	Normal	GTEx	NBS	261	186	132	0.02	0.82
EA	Tumor	CBCS	TCGA	86	109	24	0.45	0.02
AA	Tumor	CBCS	TCGA	65	236	39	0.03	0.88

Abbreviations: GTEx – Genotype Tissue Expression Project, NBS – Normal Breast Study, CBCS – Carolina Breast Cancer Study, TCGA – The Cancer Genome Atlas, EA – European ancestry, AA – African ancestry

Table 4.3 Sample sizes for GReX association analyses* for etiologic subtype

Data	Ancestry	iCOGS	iCOGS Controls	iCOGS ERNeg/P53 Neg	iCOGS ERNeg/P53 Pos	iCOGS ERPos/P53 Neg	iCOGS ERPos/P53 Pos
GTeX	EA	39392	37059	283	180	1702	168
NBS	EA	39392	37059	283	180	1702	168
CBCS	EA	39392	37059	283	180	1702	168
TCGA	EA	39392	37059	283	180	1702	168

Abbreviations: GTEx – Genotype Tissue Expression Project, NBS – Normal Breast Study, CBCS – Carolina Breast Cancer Study, TCGA – The Cancer Genome Atlas, EA – European ancestry, GReX – Germline-regulated gene expression, ER – Estrogen Receptor

* insufficient sample size for association analyses among African ancestry individuals and OncoArray samples

Table 4.4 Predictive performance of ancestry and tissue-specific models in external data (mutual external validation)

Training	Imputation	N genes	N genes with EV*R*2 > 0.01
GTEx - AA	NBS - AA	156	79
GTEx - EA	NBS - EA	63	33
NBS - AA	GTEx - AA	177	83
NBS - EA	GTEx - EA	141	74
CBCS - AA	TCGA - AA	57	28
CBCS - EA	TCGA - EA	74	35
TCGA - AA	CBCS - AA	233	99
TCGA - EA	CBCS - EA	107	50

Abbreviations: GTEx – Genotype Tissue Expression Project, NBS – Normal Breast Study, CBCS – Carolina Breast Cancer Study, TCGA – The Cancer Genome Atlas, EA – European ancestry, EV – External Validation

Table 4.5 Effects and gene locations for GReX-prioritized genes

Gene	Chr	Start*	End*	Z	Data	Ancestry	Subtype	FDR
ABAT	Chr16	8,674,588	8,784,575	6.76886	GTEx	EA	LL	4.55E-05
ABAT	Chr16	8,674,588	8,784,575	6.143925	GTEx	EA	BL	0.00036
ABCC8	Chr11	17,392,886	17,476,902	4.216681	GTEx	EA	LL	0.041593
BLR1(CXCR5)	Chr11	118,883,767	118,897,799	4.341383	TCGA	EA	LL	0.033352
C14orf45	Chr14	74,019,357	74,066,093	-4.88202	TCGA	EA	LL	0.008615
C14orf45	Chr14	74,019,357	74,066,093	-4.07993	TCGA	EA	BL	0.048356
C4A	Chr6	31,982,719	31,984,691	5.093764	GTEx	EA	LL	0.006719
C4A	Chr6	31,982,719	31,984,691	4.959555	CBCS	EA	LL	0.008349
CAPN9	Chr1	230,747,385	230,802,003	-5.21618	TCGA	EA	LL	0.003833
CCNE1	Chr19	29,811,995	29,824,308	4.294771	CBCS	EA	BL	0.036147
CD6	Chr11	60,971,642	61,013,563	-5.39181	CBCS	EA	LL	0.002461
CD84	Chr1	160,541,095	160,579,516	4.881346	GTEx	EA	BL	0.009216
CDKN1A	Chr6	36,676,461	36,687,339	4.313248	CBCS	EA	BL	0.035087
CDKN3	Chr14	54,396,956	54,420,216	-4.22423	NBS	EA	BL	0.036147
CKS1B	Chr1	154,974,643	154,979,249	-4.82194	CBCS	EA	LL	0.009216
CYP7B1	Chr8	64,595,973	64,798,791	-4.2828	NBS	EA	LL	0.033352
DDIT4	Chr10	72,273,920	72,276,039	5.285804	NBS	EA	LL	0.003833
DDR1	Chr6	30,882,918	30,900,156	6.6148	GTEx	EA	LL	7.15E-05
DDR1	Chr6	30,882,918	30,900,156	5.611111	GTEx	EA	BL	0.001448
EPCAM	Chr2	47,369,149	47,387,028	-4.53707	GTEx	EA	BL	0.020306
ERBB2	Chr17	39,688,141	39,695,181	-4.90621	NBS	EA	LL	0.008349
EVI2A	Chr17	31,316,411	31,321,749	4.466954	NBS	EA	LL	0.028939
FAM214A/KIAA1370	Chr15	52,289,121	52,417,620	-4.38271	TCGA	EA	BL	0.029128
FAM214A/KIAA1370	Chr15	52,289,121	52,417,620	-4.19043	TCGA	EA	LL	0.039019
FAM54A	Chr6	136,231,031	136,250,311	-4.99748	NBS	EA	LL	0.007323
HLA-DOB	Chr6	32,812,764	32,817,048	-6.73291	GTEx	EA	LL	4.55E-05
KCNMA1	Chr10	76,869,602	77,637,819	-7.16039	GTEx	EA	LL	1.72E-05
KCNN4	Chr19	43,766,534	43,781,257	4.985563	CBCS	EA	LL	0.008349
KCNN4	Chr19	43,766,534	43,781,257	-6.00395	NBS	EA	LL	0.00036
KIFC1	Chr6	33,391,537	33,409,922	4.435902	TCGA	EA	LL	0.029128
KRT8	Chr12	52,897,188	52,905,084	-4.14027	GTEx	EA	LL	0.042887
LHFP	Chr13	39,342,893	39,603,219	-4.81633	NBS	EA	BL	0.009216
MUC1	Chr1	155,185,825	155,189,347	5.977081	GTEx	EA	LL	0.000416
PDSS1	Chr10	26,697,667	26,746,797	4.453409	NBS	EA	LL	0.029128
PDSS1	Chr10	26,697,667	26,746,797	4.35517	NBS	EA	BL	0.033352
PPBP	Chr4	73,986,440	73,988,190	4.373034	TCGA	EA	LL	0.03329
SLC16A3	Chr17	82,228,407	82,239,499	-5.70213	CBCS	EA	LL	0.000898
SLC16A3	Chr17	82,228,407	82,239,499	-4.86715	CBCS	EA	BL	0.008701
SPINT2	Chr19	38,264,459	38,292,614	-6.00471	CBCS	EA	LL	0.00036

SYBU	Chr8	109,574,177	109,580,537	4.820178	GTEx	EA	LL	0.010205
TNFRSF17	Chr16	11,965,108	11,968,068	-5.27404	GTEx	EA	LL	0.003516
TRPM7	Chr15	50,557,156	50,582,469	4.182548	CBCS	EA	LL	0.043163
TUBA4A	Chr2	219,250,280	219,253,916	-6.23169	CBCS	EA	LL	0.000238
TUBA4A	Chr2	219,250,280	219,253,916	-4.91061	CBCS	EA	BL	0.008349
UGT1A10	Chr2	233,636,478	233,771,615	4.402316	GTEx	EA	LL	0.031338
UGT2B7	Chr4	69,096,476	69,112,987	-6.03938	GTEx	EA	LL	0.00036
UGT2B7	Chr4	69,096,476	69,112,987	-4.06618	TCGA	EA	LL	0.049318
VAV3	Chr1	107,571,161	107,678,009	-4.136	NBS	EA	BL	0.042887
VAV3	Chr1	107,571,161	107,678,009	-7.08073	NBS	EA	LL	1.72E-05
ZAP70	Chr2	97,713,569	97,739,860	5.997376	TCGA	EA	BL	0.000416
ZAP70	Chr2	97,713,569	97,739,860	4.823936	TCGA	EA	LL	0.010205
ZG16B	Chr16	2,830,173	2,832,284	4.287282	NBS	EA	BL	0.036147
ZG16B	Chr16	2,830,173	2,832,284	4.340846	NBS	EA	LL	0.033352

Abbreviations: GTEx – Genotype Tissue Expression Project, NBS – Normal Breast Study, CBCS – Carolina Breast Cancer Study, TCGA – The Cancer Genome Atlas, EA – European ancestry, AA – African ancestry

* hg38 genomic build

Figure 4.1 Effects for GReX-prioritized genes for Luminal-like (LL) and Basal-like (BL) breast cancer subtypes, by reference panel (all GReX-prioritized genes found among individuals of European ancestry)

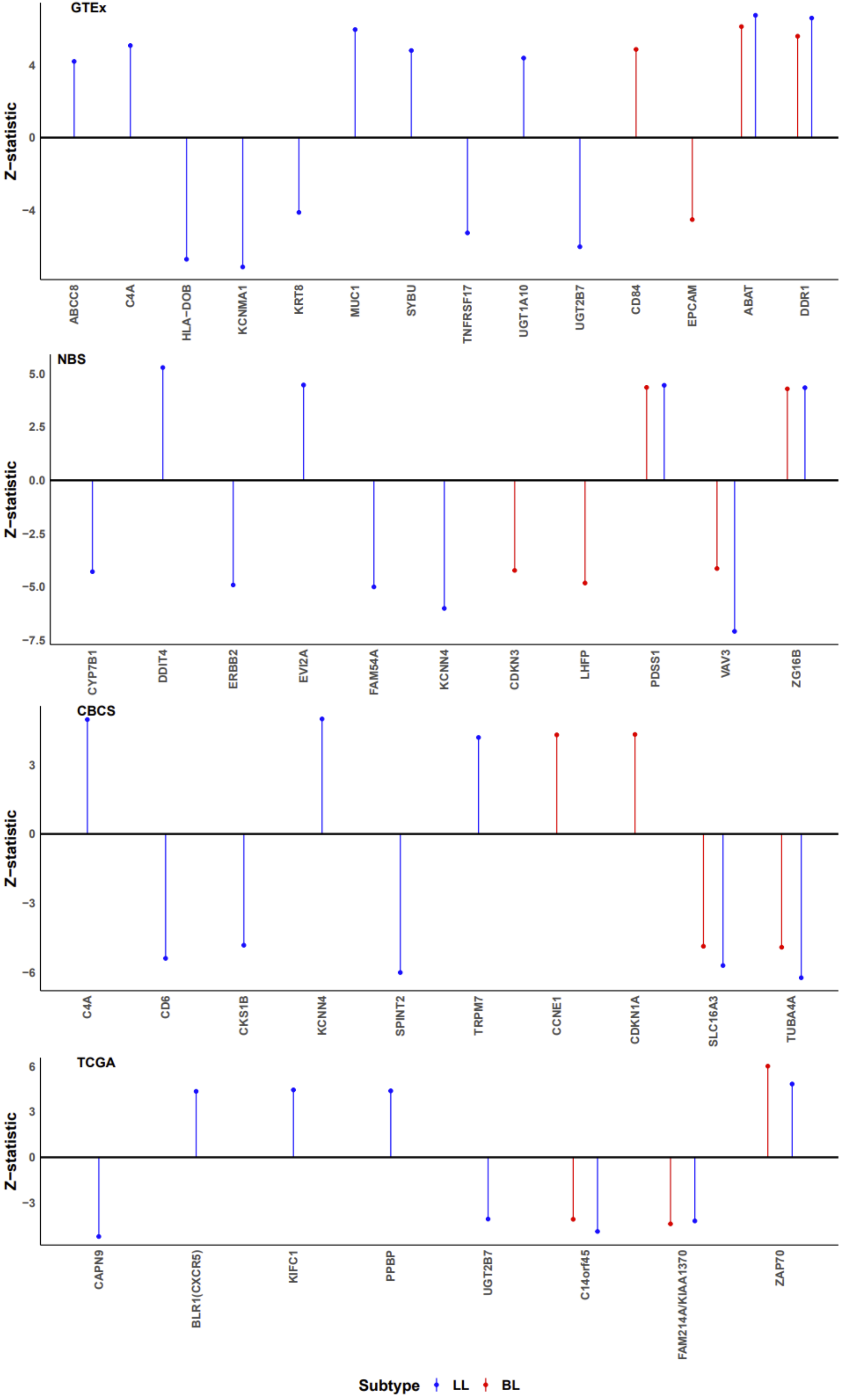


Table 4.6 Effects and gene locations for suggestive associations in GReX analysis among individuals of African ancestry

Gene	Chr	Start*	End*	Ancestry	Subtype	Data	Z	FDR
GPR44	11	60,850,924	60,855,971	AA	LL	TCGA	-3.92	0.06
AURKA	20	56,369,388	56,392,295	AA	LL	TCGA	3.63	0.11
PSPHL	7	55,697,103	55,705,595	AA	BL	CBCS	3.39	0.16
IL6	7	22,727,146	22,730,538	AA	LL	GTEx	3.33	0.17
CRYAB	11	111,908,625	111,911,749	AA	BL	CBCS	3.31	0.18

Abbreviations: GTEx – Genotype Tissue Expression Project, CBCS – Carolina Breast Cancer Study, TCGA – The Cancer Genome Atlas, GReX – Germline-regulated gene expression

* hg38 genomic build

Table 4.7 Effect and GReX IQR for GReX-prioritized genes that were identified across multiple reference panels

Gene	Subtype	Data	Ancestry	Z	GReX IQR
C4A	LL	GTEx	EA	5.093764	0.232
C4A	LL	CBCS	EA	4.959555	0.209
KCNN4	LL	CBCS	EA	4.985563	4.62E-07
KCNN4	LL	NBS	EA	-6.00395	7.86E-08
UGT2B7	LL	GTEx	EA	-6.03938	0.28
UGT2B7	LL	TCGA	EA	-4.06618	0.175

Abbreviations: GTEx – Genotype Tissue Expression Project, NBS – Normal Breast Study, CBCS – Carolina Breast Cancer Study, TCGA – The Cancer Genome Atlas, EA – European ancestry, IQR – Interquartile Range, LL – Luminal-like

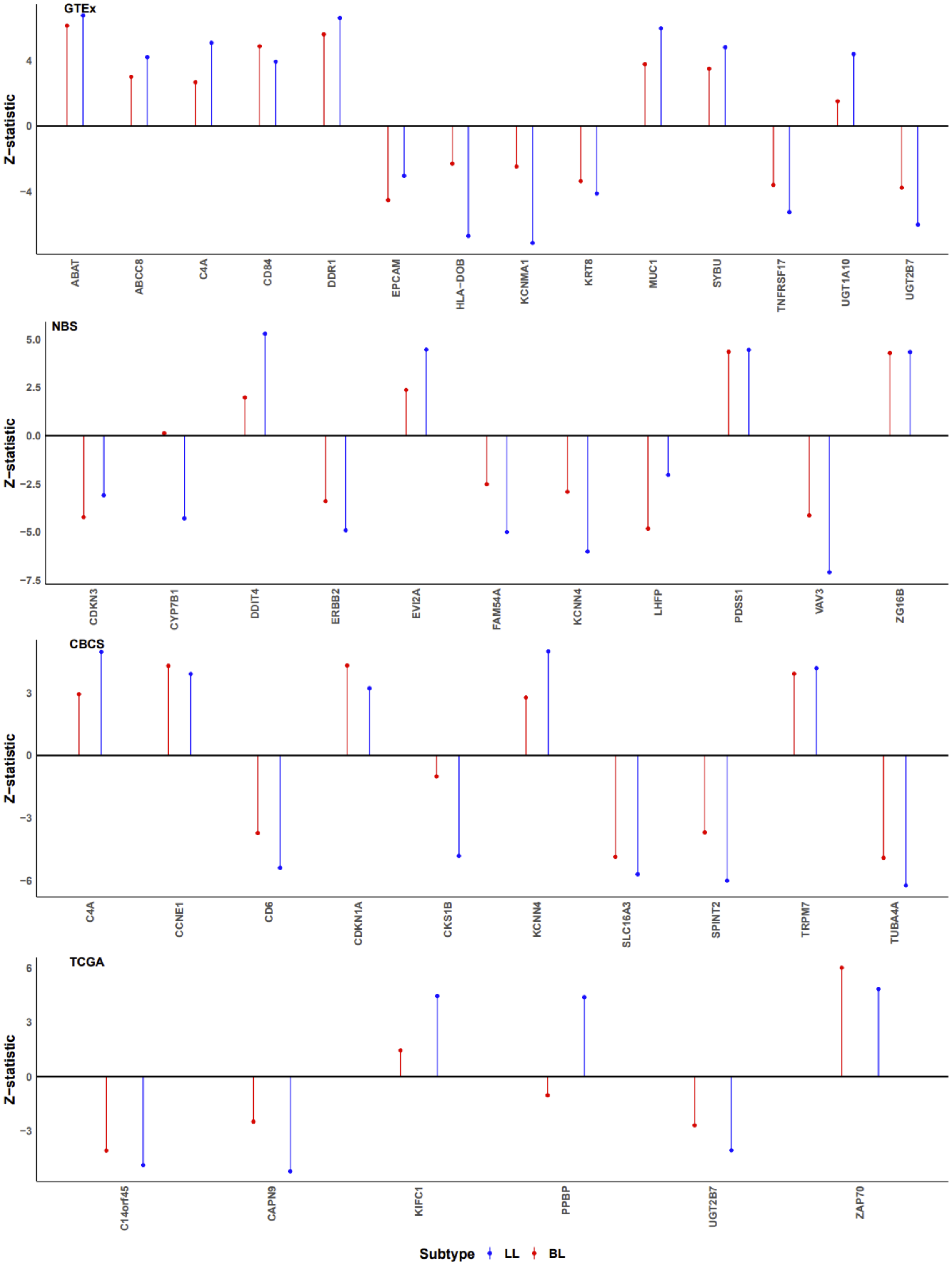
Table 4.8 Genomic overlap (+/- 1 Mb) between GWAS signal (BC cases vs. controls) and GReX-prioritized genes

Gene	Chr	GWAS pos*	Gene start*	Gene end	Z	Subtype	Data	Ancestry	FDR
CKS1B	Chr1	155,176,305	154,974,643	154,176,305	-4.82	LL	CBCS	EA	0.00922
MUC1	Chr1	155,176,305	155,185,825	154,176,305	5.98	LL	GTEx	EA	0.00042
EVI2A	Chr17	30,903,502	31,316,411	29,903,502	4.47	LL	NBS	EA	0.02894
CCNE1	Chr19	29,786,822	29,811,995	28,786,822	4.29	BL	CBCS	EA	0.03615
KCNN4	Chr19	43,782,361	43,766,534	42,782,361	4.99	LL	CBCS	EA	0.00835
KCNN4	Chr19	43,782,361	43,766,534	42,782,361	-6.00	LL	NBS	EA	0.00036
HLA-DOB	Chr6	33,272,092	32,812,764	32,272,092	-6.73	LL	GTEx	EA	0.00005
KIFC1	Chr6	33,272,092	33,391,537	32,272,092	4.44	LL	TCGA	EA	0.02913

Abbreviations: GTEx – Genotype Tissue Expression Project, NBS – Normal Breast Study, CBCS – Carolina Breast Cancer Study, TCGA – The Cancer Genome Atlas, EA – European ancestry, LL – Lumina-like, BL – Basal-like, Mb – Megabase, BC – Breast Cancer

* hg38 genomic build

Figure 4.2 Effects for GReX-prioritized genes for Luminal-like (LL) and Basal-like (BL) breast cancer subtypes, by reference panel, where non-significant associations are also provided for each gene



CHAPTER 5: AIM2. ASSESSMENT OF GENE-LEVEL GERMLINE ASSOCIATIONS FOR BREAST CANCER MORTALITY AMONG EUROPEAN AND AFRICAN ANCESTRY INDIVIDUALS IN THE BREAST CANCER ASSOCIATION CONSORTIUM (BCAC)

5.1 Introduction

Breast cancer (BC) is the most common cancer among women in the world and in the United States, it is the second leading cause of cancer mortality with an estimated 44,130 deaths in 2021 [153,154]. Moreover, there is racial disparity in BC mortality, where African ancestry (AA) individuals have roughly 40% greater mortality rate compared to European ancestry (EA) individuals [153]. In contrast to breast cancer incidence where studies have uncovered a host of demographic, lifestyle, clinical, and genetic risk factors, identification of risk factors for BC mortality has been sparse; in particular, compared to the 210 germline variants identified in relation to BC incidence, genome-wide association study (GWAS) to date have not identified more than a handful of suggestive associations, with replication of these associations proving particularly challenging [7,82-92]. Importantly, only one of the prior genetic investigations of BC mortality has been among individuals of non-European ancestry due to limited sample size availability (for GWAS) for this population [16]. Both the paucity and lack of replicability of findings among individuals of European ancestry as well as the lack of investigation of germline underpinnings of BC mortality among non-European ancestry individuals represents a critical gap in knowledge, as better understanding of potential genetic underpinnings of BC mortality across diverse populations has potential to inform clinical decision making and bridge racial disparities in BC.

Given the need for larger sample sizes in traditional GWAS of BC mortality across diverse populations, gene-based association tests have emerged as an alternative with higher statistical efficiency [12,13]. Bhattacharya et al. applied a Transcriptome-Wide Association Study (TWAS) approach and investigated a panel of 406 BC-related genes in relation to BC-specific mortality in the Carolina Breast Cancer Study (CBCS), uncovering four associations among AA and zero among EA individuals [16]. Importantly, predictive models of tumor gene expression based on germline genetics underlying the association testing were found not to be transportable across ancestry [16]. Moreover, a key finding was that the associations among AA individuals appeared to be driven by associations within the Estrogen Receptor positive (ER+) strata of AA [16]. These findings suggest that both population heterogeneity and tumor heterogeneity are important considerations for TWAS. BC is a heterogeneous disease with distinct subtypes and studies have highlighted that accounting for such biological heterogeneity in the assessment of germline genetic associations with BC mortality may elucidate subtype-specific associations that may otherwise be masked in aggregate (i.e., all BC cases) association testing [16,92]. Another key limitation to previous genetic investigations (GWAS or TWAS-based) is that they do not correct for potential collider bias [155-157]. Collider bias in the context of genetic studies of BC mortality occurs when genetic factors are associated with both incidence and mortality (Figure 5.1). Because studies of BC mortality are by design case-only analyses, this inherent conditioning on BC incidence induces an association between the genetic factor under study and other factors which affect both incidence and mortality, thereby confounding the association of interest between genetic factor under study and BC mortality (Figure 5.1).

In this study, we address some key limitations of prior investigations. We adopt a TWAS approach to enable gene-based genetic investigations across individuals of Europeans and African ancestry. In our TWAS-based analyses of GReX (constructed under TWAS approach) of 396 BC-related genes in relation to BC mortality endpoints (all-cause, BC-specific), we construct predictive models of gene expression stratified by combinations of ancestry and subtype (i.e., among a sample of EA individuals, predictive models constructed separately for EA individuals with Luminal-like and Basal-like BC). Towards this end, we leverage among the largest resources for matched genotype-expression data in diverse populations in the CBCS (N = 526 EA, 582 AA) for predictive model construction and the largest breast cancer genetic data resource for association testing in the Breast Cancer Association Consortium (BCAC; N= 217,314 cases and controls). Moreover, we apply a robust method for collider bias correction called Slope-Hunter and present both “corrected” and Slope-Hunter naïve estimates [158] .

5.2 Methods

Study Population – Carolina Breast Cancer Study

CBCS is a population-based study of North Carolina (NC) breast cancer patients, comprising of three study phases; study details have been described previously [94,95]. Briefly, patients aged 20 to 74 were identified using rapid case ascertainment with the NC Central Cancer Registry; self-identified Black and young women (20-49 years) were oversampled [94,95]. Variants in CBCS were assayed using the OncoArray custom SNP array and imputed using the 1000 Genomes Project (v3) as a reference panel in two-step phasing and imputation using SHAPEIT2 and IMPUTEv2 [97,99,101,131,132]. Genotype calling, imputation, and quality control were performed using the Division of Cancer Epidemiology and Genetic Cancer Genomics Research Laboratory. We removed variants with minor allele frequency <1% and deviation from Hardy-Weinberg equilibrium at $P < 10^{-8}$ [159]. Genotypes were intersected across the race and subtype combination samples (i.e., EA – LL, EA – BL, AA – LL, AA – BL) for a total of 5,158,798 SNPs analyzed.

Tumor gene expression of 396 BC-related autosomal genes plus 11 housekeeping genes was assayed using NanoString nCounter at the Translational Genomics Laboratory at UNC-Chapel Hill. These 396 BC-related genes include genes that are part of the *PAM50, TP53, E2, EGFR,* and *IGF* signatures, among others (Appendix 1). QC was performed using NanonStringQCPro; distributional difference across lanes was scaled using upper-quartile normalization and two dimensions of unwanted technical and biological variation were removed using RUVSeq [129,130]. The sample sizes (of matched genotype and expression) for downstream analyses were as follows: 410 (EA – LL), 116 (EA – BL), 358 (AA – LL), and 224 (AA – BL).

Study Population – Breast Cancer Association Consortium

BCAC is aimed at bettering understanding of the inherited risk of BC and includes roughly 100 studies, each of which has contributed demographic, clinical, and genetic data on study participants. Individuals in BCAC have been genotyped on the Illumina iSelect (iCOGS) and Illumina Infinium OncoArray (OncoArray) platforms. Phasing and imputation across both genotyping arrays was performed using SHAPEIT2 and IMPUTEv2 with the 1000 Genomes Project (v3) as a reference panel [97,99,101,131,132]. In this study, we examined only individuals of European and African ancestry in BCAC.

Assessment of race and subtype-specific expression quantitative trait loci (eQTL)

To provide added context for downstream race and subtype-specific predictive models of gene expression from tumor germline variation, we performed two eQTL analyses using matrixeQTL[160]. First, we assessed eQTLs within strata defined by ancestry and subtype (i.e., eQTLs among EA-LL, EA-BL, AA-LL, and AA-BL samples, respectively), mirroring the intended ancestry and subtype-stratification for downstream predictive models for GReX analysis. Ancestry strata in CBCS were defined based on self-identified race; in CBCS data there is strong concordance between self-reported race and genetic ancestry (first principal component of combined genotype matrix) as well as concordance between race and genetic ancestry as they pertain to GReX and tumor gene expression [16,136]. We emphasize that genetic ancestry exists along a continuum; however, in the absence of eQTL and TWAS approaches that allow consideration of the continuum of genetic ancestry, we chose stratification as the most practical solution for ancestry-specific analyses. Second, we assessed heterogeneity of eQTL effects across subtype through a Wald t-test on an interaction term for genotype dosage times subtype (ref. subtype : LL). The second eQTL analysis was performed separately for the two ancestry strata since a formal assessment of eQTL heterogeneity across ancestry groups is challenging due to differences in allele frequencies and linkage-disequilibrium (LD) patterns. We do not expect differences in allele frequencies and LD patterns across subtype samples of the same ancestry. To control for potential confounding of the genotype-gene expression relationship across both eQTL analyses, we adjusted for the first five principal components from the combined (EA, AA) genotype matrix, age, menopausal status, and body mass index (BMI).

In follow-up analyses to the eQTL analyses, we first assessed concordance of eQTL effect direction across subtypes, within each ancestry strata, as a function of significance thresholding (i.e., at false discovery rate (FDR) threshold of 0.05, 0.10, and 0.50); the goal of this analysis was to examine the influence of potential differences in statistical power across subtypes in discovery of subtype-specific significant eQTLs (within each ancestry strata). In the second follow-up analysis, we compared, across ancestry, the minor allele frequency (MAF) of eQTLs that were found to be heterogeneous across subtypes within a given ancestry stratum (e.g., comparison of MAF across EA and AA for a loci that was found to be significant in formal eQTL heterogeneity assessment across subtypes among EA).

Construction of race and subtype-specific predictive models of tumor gene expression from germline variation in CBCS and imputation (construction of GReX) in BCAC

GReX was the exposure of interest in this study, and for a given gene, GReX represents the portion of tumor expression explained by *cis*-genetic variation (where *cis*-regions to gene span the 1 Megabase surrounding the gene’s start and end position). To construct GReX, we adopted TWAS methodology, where we first trained predictive models of tumor gene expression from *cis*-germline variation using Elastic Net Regression with five-fold cross validation (EN) and Best Linear Unbiased Predictor (BLUP) [12,13,16]. To mitigate model redundancy, we pruned variants prior to model training using PLINK v1.9 with the following pruning parameters: window size of 50 base pairs, window shift of 5, and LD threshold of 0.5 [161]. We included the first five principal components of the full genotype matrix (N=1,108), as well as age, menopausal status, and body mass index as adjustment variables during predictive model construction. The approach yielding the better predictive accuracy, defined as McNemar's R^2 (sample-size adjusted R^2) between observed and predicted expression, was selected as the gene's predictive model. The predictive model for a gene contains SNP-gene expression weights corresponding to the magnitude and direction of the effect of a *cis*-variant on the gene's tumor expression.

Importantly, predictive models were constructed by strata corresponding to combinations of ancestry and subtype (i.e., EA – LL, EA – BL, AA – LL, AA – BL) for two reasons. First, prior investigations show that predictive models of (tumor) gene expression are not generally transportable across ancestry [16,133]. Second, we aimed to additionally investigate the role of breast cancer heterogeneity (i.e., subtypes) on relationships between germline variation and BC mortality [16]. The sample sizes (of matched genotype and expression) underlying the predictive models were as follows: 410 (EA – LL), 116 (EA – BL), 358 (AA – LL), and 224 (AA – BL). Genes with good predictive accuracy, defined as adjusted-R^2 > 0.01 between observed and predicted expression, were selected for imputation and association testing. Some GReX (TWAS) analyses include further feature selection of genes for imputation (GReX construction) and association testing based on cis-h^2; in line with these analyses we computed cis-h^2 using GCTA [12,16,162] (Table 5.1). However, due to the relatively small sample size for computation of cis-h^2 across the race and subtype strata, GCTA analyses yielded imprecise heritability estimates and cis-h^2 was not used in addition to predictive accuracy to select genes for further imputation and association testing, as in some prior investigations [120,121].

We imputed GReX into BCAC data and tested for associations between GReX and all-cause and BC-specific mortality. We aligned predictive models and imputation samples in BCAC based on: 1) combination of genetic ancestry and subtype (main analysis) (Figure 5.2); 2) genetic ancestry alone (supplementary analysis) (Figure 5.3) [16]. For the former, for example, imputation of GReX was performed among women of African ancestry and BL subtype in BCAC for predictive models constructed in the AA – BL stratum in CBCS. Ancestry and subtype-specific GReX was imputed by multiplying the genotype dosage in BCAC with the SNP-gene expression weights derived from the ancestry and subtype-specific predictive models, based on aforementioned alignment schemes between imputation panel and predictive models. Imputation was performed separately for the iCOGS and OncoArray genotyped samples in BCAC.

Assessment of predictive model transportability across subtypes

To assess whether predictive models are transportable across subtype, we used predictive models constructed in LL tumor tissue to impute expression in BL tumor samples; ancestry was held constant in this assessment of model portability (i.e., models trained in WW – LL samples were imputed into WW – BL samples while models trained in AA – LL samples were imputed into AA – BL samples). In the imputation samples (AA – BL and AA – BL), adjusted-R^2 for predicted vs. observed expression were computed.

External validation of race and subtype-specific models in The Cancer Genome Atlas (TCGA)

We obtained TCGA genotype data using the Genomic Data Commons (GDC) legacy archive. Genotype files were merged across individuals and phasing and imputation were performed using eagle v2.4 and minimac4 at the TOPMed imputation server [102,104,163]. RNA expression data were obtained using the Broad Institute's GDAC Firehose via FireBrowse [116]; RNA expression was quantified using RNA-seq and downloaded data were TCGA level-3 normalized. We further $\log_2$ transformed expression data. The sample sizes of the matched genotype-expression datasets by strata of race and subtype combinations were: 606 (EA – LL), 111 (EA – BL), 107 (AA – LL), and 65 (AA – BL). Of the 124, 248, 157, and 157 well-predicted genes in CBCS across EA – LL, EA – BL, AA – LL, and AA – BL, respectively, we were able to impute GReX in TCGA and compute adjusted-R^2 (observed vs. predicted) for 106, 214, 136, and 143 genes respectively.

Association testing between GReX and BC mortality (all-cause, BC-specific) in BCAC

We estimated associations between race and subtype-specific GReX of genes and BC mortality (all-cause, BC-specific) using Cox Proportional Hazards Regression. Association testing was carried out separately for iCOGS and OncoArray genotyped samples in BCAC [89]. We used age as the time scale, where time of entry into the study corresponded to the age at BC diagnosis and time at study exit was the age at last follow-up, where last follow-up was either mortality or censoring event. We opted to use age as the time scale to mitigate potential bias by left truncation (as component studies in BCAC have variable study periods and lengths of follow-up) [164]; age as time scale also enables confounding control for age in assessment of GReX – mortality associations. We additionally allowed distinct baseline hazards per BCAC study in our Cox Proportional Hazards models to both control for potential confounding by and to incorporate potential heterogeneity across BCAC studies.

Effect estimates (log hazard ratio (HR)) for one standard deviation increase in GReX and corresponding standard errors for iCOGS and OncoArray samples were pooled using fixed-effects, inverse variance weighted meta-analysis. We adjusted the nominal significance level for multiple comparisons using the Benjamini-Hochberg (BH) procedure, where statistical significance was set at false discovery rate (FDR) <0.05 across all association tests performed (N=1,362 across ancestry – subtype and mortality combinations).

A major concern in case-only analyses, as previously mentioned, is collider bias, where the exposure of interest (GReX in this study) is associated with incidence and mortality [155-157]. Unlike prior investigations (either GWAS or TWAS) of germline genetics in relation to BC mortality, we explicitly corrected for potential collider bias, using Slope-Hunter [158]. Briefly, Slope-Hunter uses model-based clustering and identifies exposures (in this study, GReX of genes) that affect only BC incidence (but not BC mortality); this set of GReX of genes is then used to compute an adjustment factor for the “raw” effect estimates [158]. Slope-Hunter correction requires effect estimates for GReX of genes in relation to both BC incidence and mortality [158]. To generate the effect estimates for BC incidence, we imputed GReX across the set of cases and controls in BCAC using our ancestry and subtype-specific models, followed by association testing between the GReX and case versus control status using logistic regression.

The sample sizes, including mean ages, for main and supplementary GReX association analyses, respectively after exclusions based on case invasiveness status, metastatic status, and missing data on confounding and follow-up variables are provided in Figure 5.4 and Figure 5.5. We note that in the main GReX analysis, the mean age across the ancestry and subtype specific cohorts for the endpoints were generally in line with population based studies of breast cancer patients such as Surveillance, Epidemiology, and End Results (SEER) and the CBCS [126,165,166].

5.3 Results

Assessment of race and subtype-specific expression quantitative trait loci (eQTL)

Among EA individuals, we found 715 and 0 significant SNPs (eSNPs) across 17 and 0 genes (eGenes) for LL and BL subtypes, respectively, at Benjamini-Hochberg FDR threshold of 0.05 (Figure 5.6A). Among AA individuals, we found 229 and 49 eSNPs across 15 and 3 eGenes, respectively (Figure 5.6B). In formal assessment of eQTL heterogeneity across subtype (within ancestry strata), we found 321 SNPs across 37 genes with significantly different germline regulation of tumor expression by subtype among EA individuals (Figure 5.7A). Among AA individuals, we found 23 SNPs across 4 genes with significantly different germline regulation of tumor expression by subtype (Figure 5.7B). In follow-up analyses, we found that the level of concordance in eQTL effect direction across subtypes was not sensitive to eQTL significance thresholding (FDR thresholds of 0.05, 0.10, and 0.50) in both EA and AA individuals (Figure 5.8, Figure 5.9). In additional follow-up analyses, we found that the MAF differed across EA and AA samples for subtype-heterogeneity eSNPs discovered among EA or AA samples (Figure 5.10).

Overlap in genes with good predictive accuracy across race and subtype strata

For the 396 BC-related genes, we were able to build predictive models with good predictive accuracy (adjusted-R^2 of >0.01 for observed and predicted expression) for 124, 248, 157, and 157 genes for EA – LL, EA – BL, AA – LL, and AA – BL, respectively (Figure 5.11). For WW, 24 genes were well-predicted across both LL and BL samples, while among BW, only 8 genes were well-predicted across both LL and BL samples **(**Figure 5.11**)**. 18 genes were well-predicted across all of EA – LL, EA – BL, AA – LL, and AA – BL samples, including genes such as *AURKA*, *EGFR*, *PSPHL*, and *GPR160*.

Assessment of predictive model transportability across subtype and external validation of models in TCGA

In prior investigations in CBCS, predictive models of tumor gene expression constructed among EA individuals were not transportable to AA individuals [16]. In this study, in addition to assessment of eQTL heterogeneity by subtype, we further assessed whether predictive models of tumor gene expression constructed for a given subtype are transportable to another subtype, within ancestry strata.

We found that predictive accuracy was roughly 3 fold and 5 fold higher, for EA and AA, respectively, when there was alignment between training and imputation sample (i.e., EA models trained in LL were used to impute in the LL sample, as opposed to BL sample) (Figure 5.12A).

We also assessed external validity of our ancestry and subtype-specific models in TCGA. In CBCS, the mean (standard deviation) cross-validation R^2 of well-predicted genes was 0.021 (0.020), 0.043 (0.030), 0.020 (0.013), and 0.025 (0.0140) for EA – LL, EA – BL, AA – LL, and AA – BL, respectively (Figure 5.12B). By comparison, the adjusted-R^2 in TCGA was 0.001 (0.003), 0.005 (0.015), 0.006 (0.014), and 0.007 (0.016) for EA – LL, EA – BL, EA – LL, and EA – BL, respectively (Figure 5.12B).). While the mean adjusted-R^2 in TCGA may be low, we can also examine performance with regards to number of genes with adjusted-R^2 above the 0.01 threshold. Here, we found that predictive models constructed in EA – BL, AA – LL, and AA – BL samples showed sufficiently strong performance in TCGA, with 32, 30, and 26 genes showing predictive accuracy of >0.01 in TCGA (Table 5.2). Predictive performance was poor in the EA – LL sample, with only 2 genes showing predictive accuracy of >0.01 (Table 5.2). Stratification by subtype (e.g., EA – BL) may improve model performance among WW, as in prior CBCS investigations, only 7 of the 151 cis-heritable genes (out of 416) among EA individuals showed predictive accuracy of > 0.01 [16].

Associations between GReX and BC mortality (all-cause, BC-specific)

At global FDR <0.05 across all association tests performed (N=1,362 across race – subtype and mortality combinations), we found no statistically significant associations with either all-cause or BC-specific mortality in either the main (predictive models and GReX imputation sample aligned on genetic ancestry and subtype) or supplementary (predictive models and GReX imputation sample aligned on genetic ancestry) GReX analysis (Table 5.3**).** In the main GReX analysis, we observed a suggestive association for *PTGER3* among the EA – BL sample for all-cause mortality (FDR = 0.14, corresponding HR [95% CI] : 0.88 [0.82, 0.94]) (Table 5.3). The next four nominally-significant genes with the highest significance level included *CRYBB2* (AA – BL; BC-specific), *ZEB2* (EA – BL, BC-specific), *RAD17* (EA – LL; all-cause), and *RAB25* (AA – BL; all-cause); the corresponding HR [95% CI] for these genes were 1.97 [1.26, 3.07], 0.88 [0.81, 0.96], 0.95 [0.92, 0.99], and 1.37 [1.10, 1.71], respectively (Table 5.3).

5.4 Discussion

In this study, we leveraged a large resource of matched genotype-expression data (N = 526 EA, N=582 AA) to perform a GReX analysis (for a panel of 396 BC-related genes) of BC mortality (all-cause, BC-specific) in the largest available data resource of BC cases in the BCAC. Across the spectrum of eQTL and predictive model transportability analyses to contextualize our GReX analyses, we found: 1) extensive heterogeneity in germline regulation of tumor expression across subtypes, within ancestry strata (e.g., 321 eSNPs across 37 eGenes among EA); 2) predictive models of tumor gene expression from germline variation are not transportable across subtype. Both insights demonstrate the need for consideration of biological heterogeneity (e.g., subtype) in future genetic investigations of BC progression. Moreover, our results on differential germline regulation of tumor expression are pertinent to further explorations of the genetic basis of treatment resistance based on tumor gene expression patterns. In GReX analyses, our findings of no statistically significant associations at a conservative global FDR < 0.05 are mostly in line with existing GWAS of BC survival, although we do note a suggestive association (FDR = 0.14) for *PTGER3.* Importantly, in our GReX analyses, we demonstrate the need for formal collider bias correction, as doing so can correct naïve (biased) estimates that are, in some cases, even contrary to a gene's known biological function/role within BC progression and mortality.

As previously mentioned, we found a suggestive association for *PTGER3* among EA – BL patients (FDR = 0.14), where increased *PTGER3* germline-regulated expression was associated with reduced all-cause mortality (**Table 5.3**). This finding is in line with a prior study which indicated significantly improved overall and progression free survival with increased *PTGER3* expression among breast cancer individuals [167]. Our finding of no statistically significant associations at global FDR <0.05 is in line with many prior investigations of germline variation in relation to BC mortality [82-92]. Across the previous germline investigations of BC mortality across mostly individuals of European ancestry, a total of five loci have been reported at or near genome-wide significance [82-92]; a common limitation to these investigations is that findings are not replicable across studies. In fact, the most recent GWAS in the PATHWAYS study was not able to replicate any of the previously identified at or near genome-wide significance loci [92]. One likely explanation for such spurious and non-replicable associations is collider bias, as we, for example, observed in supplementary GReX analysis for *CCNA2* in relation to all-cause mortality among AA individuals (non SH FDR *p*-value = 0.17). *CCNA2* belongs to the highly conserved cyclin family and plays a critical role in cell cycle control at the G1/S and G2/M transitions [168,169]; *CCNA2* expression has been implicated with poor mortality outcomes among BC patients [168,169], whereas in our analyses without correction for potential collider bias, we initially observed a confounded protective effect with increased *CCNA2* GReX among AA – LL imputation sample. This underscores the need for collider bias correction in genetic investigations of BC mortality, and interpretation with caution for already reported associations. Presently, Slope-Hunter and Dudbridge et al. represent the most recent and 'most robust' approaches towards collider bias correction, although between the two, we chose to use Slope-Hunter because it offers better type I error rate at comparable power [158,170]. Other approaches such as inverse-probability weighting (IPW) of cases based on risk factors should also be considered in future investigations [155].

An interesting suggestion in the most recent GWAS of BC mortality in the PATHWAYS study is that the sheer magnitude of treatment heterogeneity may overwhelm (modest) genetic associations [92]. In that study, there was one variant identified at genome-wide significance, but this association was statistically significant only among individuals with Par-4 dependent chemotherapy [92]. We lacked the level of granular information on treatment in BCAC that the PATHWAYS study offers to perform similar follow-up analyses. We note that this study did not correct for collider bias, which remains a concern in stratified analyses by treatment as treatment is determined by BC subtype (incidence) (Figure 5.1). Nevertheless, future studies should explore the possibility of treatment-specific germline genetic effects on BC mortality while correcting for collider bias through approaches such as Slope-Hunter or IPW.

A novel aspect of this study is that we account for potential biological heterogeneity in germline associations with BC mortality by constructing subtype-specific (within each ancestry strata) predictive models of tumor gene expression from germline variation. In eQTL heterogeneity analyses across subtypes (within each ancestry strata), we found extensive evidence (321 SNPs across 37 genes) of differential germline regulation across subtypes among EA individuals and modest evidence (23 SNPs across 4 genes) of the same among AA individuals. Beyond this current investigation of germline associations for BC mortality, our findings of differential germline regulation of tumor gene expression across subtypes in EA and AA populations has clinical implications. As example, we find that among AA individuals, the presence of the reference (minor allele) versus the alternative allele for rs4656930 is associated with significantly higher expression of *KLHDC9* for BL subtype compared to LL subtype (Figure 5.7B). Expression of *KLHDC9* has been shown to be associated with decreased sensitivity to paclitaxel, a commonly used chemotherapeutic agent among triple negative breast cancer (TNBC; molecular subtype TNBC is closely related to PAM50 based BL subtype) [171,172]. Therefore, a higher germline regulated expression of *KLHDC9* for BL subtype among AA individuals may point to germline contribution towards treatment resistance in the AA-BL population that warrants further investigation. In this mold, our reporting on differential regulation of BC-related genes across subtypes for EA and AA individuals may be beneficial towards efforts to understand genetic underpinnings of BC treatment resistance.

There are a few limitations to this study. First, CBCS used a NanoString nCounter probeset for RNA expression quantification of BC-related genes, and therefore we were not able to analyze the whole transcriptome. However, the use of CBCS data was key towards building ancestry and subtype-specific models of tumor gene expression underlying our GReX analyses of BC mortality, as CBCS contains one of the largest tumor transcriptomic datasets for AA. Second, CBCS lacks data on somatic alternations in the genome; inclusion of these elements in predictive mode construction could enhance model performance. Third, we were not able to stratify predictive models beyond LL and BL, because the sample sizes of matched genotype-expression data for subtypes such as Her2-like were below a 100, which is a challenging sample size for predictive model construction in tumor tissue.

In conclusion, we find no significant associations at global FDR <0.05 between the GReX of a panel of BC-related genes and BC mortality (all-cause, BC-specific), in line with many prior investigations of germline variation in relation to BC mortality. We demonstrate how collider bias can potentially confound findings via comparison of collider-bias naïve and collider bias corrected analyses, and demonstrate both differential germline regulation of tumor gene expression across subtypes (within ancestry strata) and the lack of transportability of predictive models of tumor gene expression across subtypes (within ancestry strata), a finding which underscores the need for larger and more racially and biologically diverse cohorts for future investigations.

5.5 Tables and Figures

Figure 5.1 Directed acyclic graph (DAG) demonstrating potential for collider bias in germline genetic investigations of breast cancer mortality

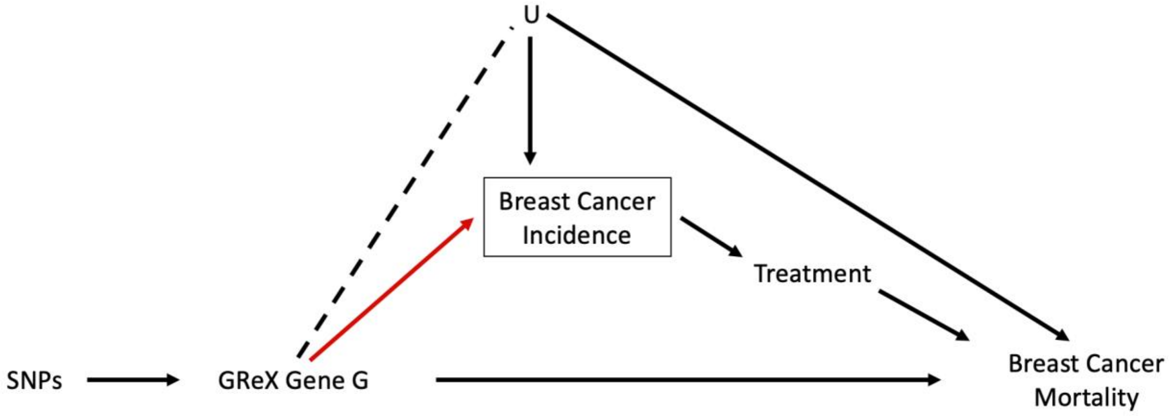
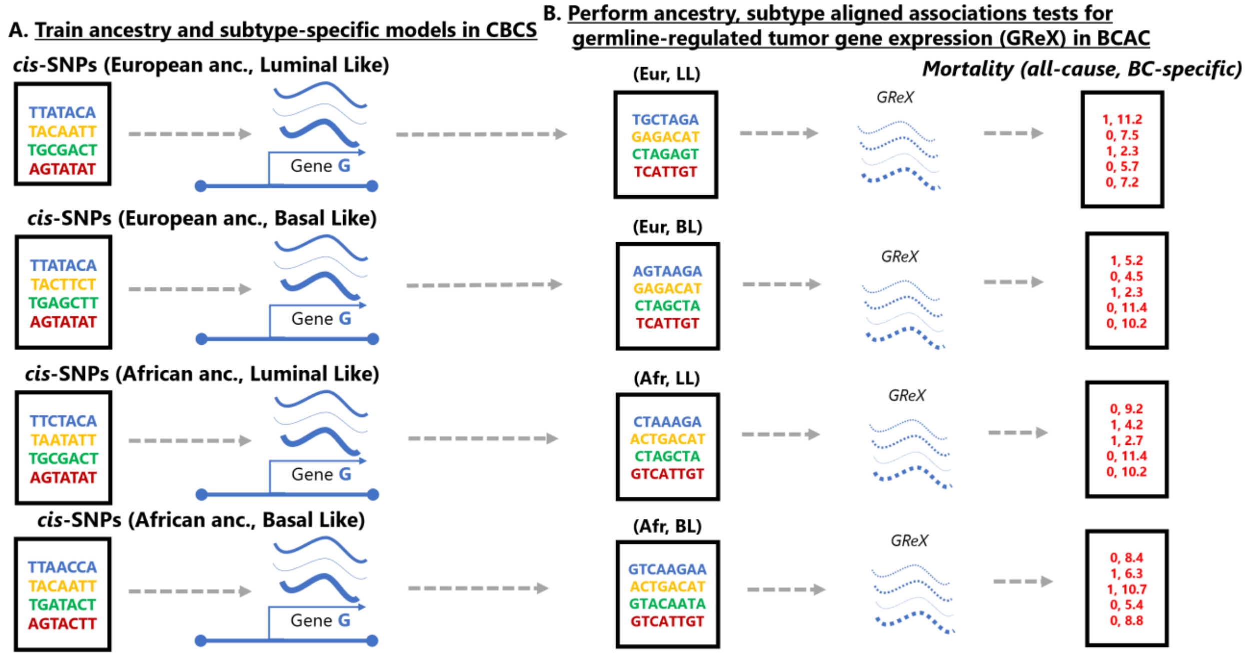


Table 5.1 Tumor gene expression heritability for race and subtype sample combinations, computed using GCTA

	Mean(SD) of cis-h^2
EA - LL	0.004 (0.036)
EA - BL	0.018 (0.14)
AA - LL	0.007 (0.05)
AA - BL	0.009 (0.08)

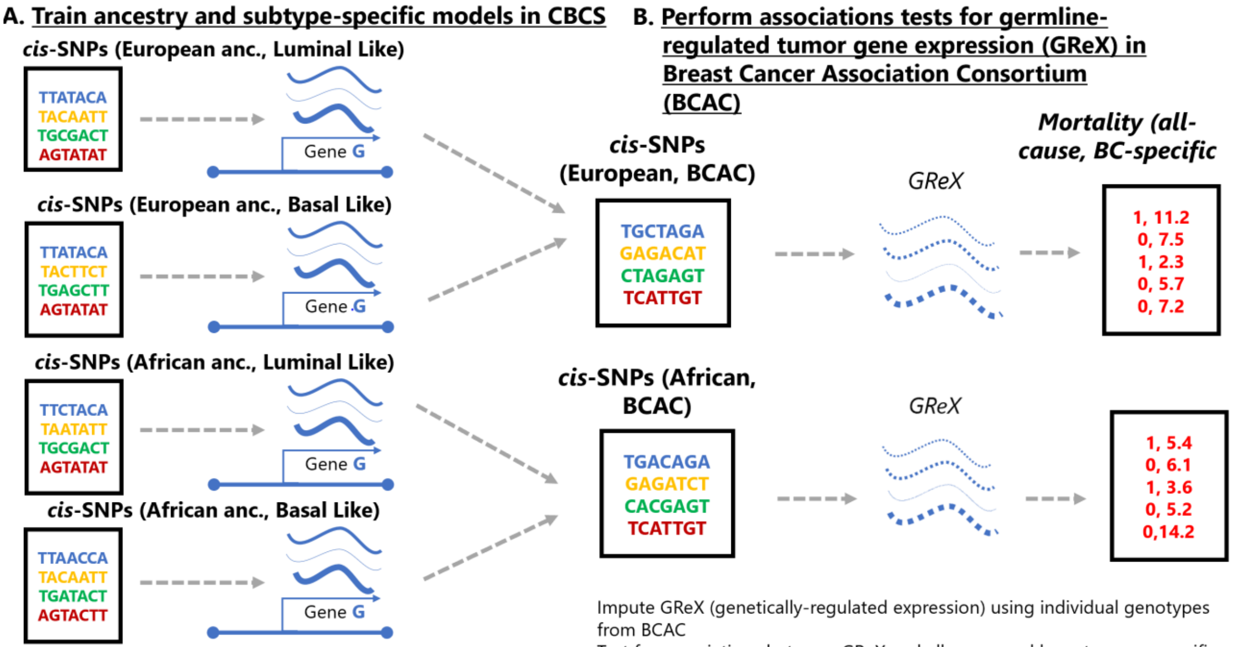
Abbreviations: GCTA – Genome-Wide Complex Trait Analysis, SD – Standard Deviation, EA – European ancestry, AA – African ancestry, LL – Luminal-like, BL – Basal-like

Figure 5.2 Study schematic A) In the Carolina Breast Cancer Study (CBCS), we leveraged Transcriptome-Wide Association Study (TWAS) methodology and trained ancestry and subtype-specific predictive models of tumor gene expression from *cis*-germline variation (defined as <1 Megabase to gene start and end sites). B) We integrated CBCS predictive models with genotypes from the Breast Cancer Association Consortium (BCAC), and performed imputation and association testing for breast cancer mortality in BCAC. Imputation and association testing were aligned with predictive models based on genetic ancestry and subtype



Train ancestry and subtype-specific models of breast tumor expression using FUSION (Gusev *et al* 2016, Nature Genetics)

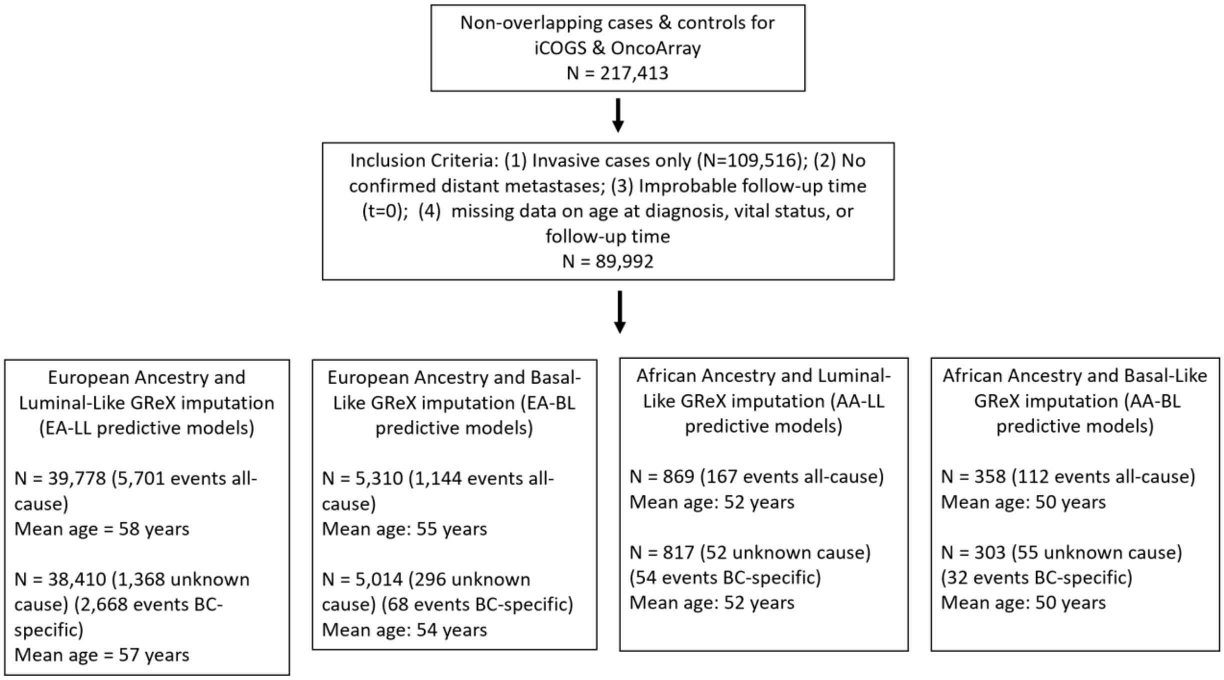
Figure 5.3 Study schematic for supplementary GReX analysis. Imputation and association testing in BCAC were aligned with predictive models from CBCS based on genetic ancestry alone (in CBCS, race served as proxy for genetic ancestry)


A. Train ancestry and subtype-specific models in CBCS
cis-SNPs (European anc., Luminal Like)
TTATACA
TACAATT
TGCGACT
AGTATAT
Gene G
cis-SNPs (European anc., Basal Like)
TTATACA
TACTTCT
TGAGCTT
AGTATAT
Gene G
cis-SNPs (African anc., Luminal Like)
TTCTACA
TAATATT
TGCGACT
AGTATAT
Gene G
cis-SNPs (African anc., Basal Like)
TTAACCA
TACAATT
TGATACT
AGTACTT
Gene G
B. Perform associations tests for germline-regulated tumor gene expression (GReX) in Breast Cancer Association Consortium (BCAC)
cis-SNPs (European, BCAC)
TGCTAGA
GAGACAT
CTAGAGT
TCATTGT
GReX
Mortality (all-cause, BC-specific
1, 11.2
0, 7.5
1, 2.3
0, 5.7
0, 7.2
cis-SNPs (African, BCAC)
TGACAGA
GAGATCT
CACGAGT
TCATTGT
GReX
1, 5.4
0, 6.1
1, 3.6
0, 5.2
0,14.2

Train ancestry and subtype-specific models of breast tumor expression using FUSION (Gusev *et al* 2016, Nature Genetics)

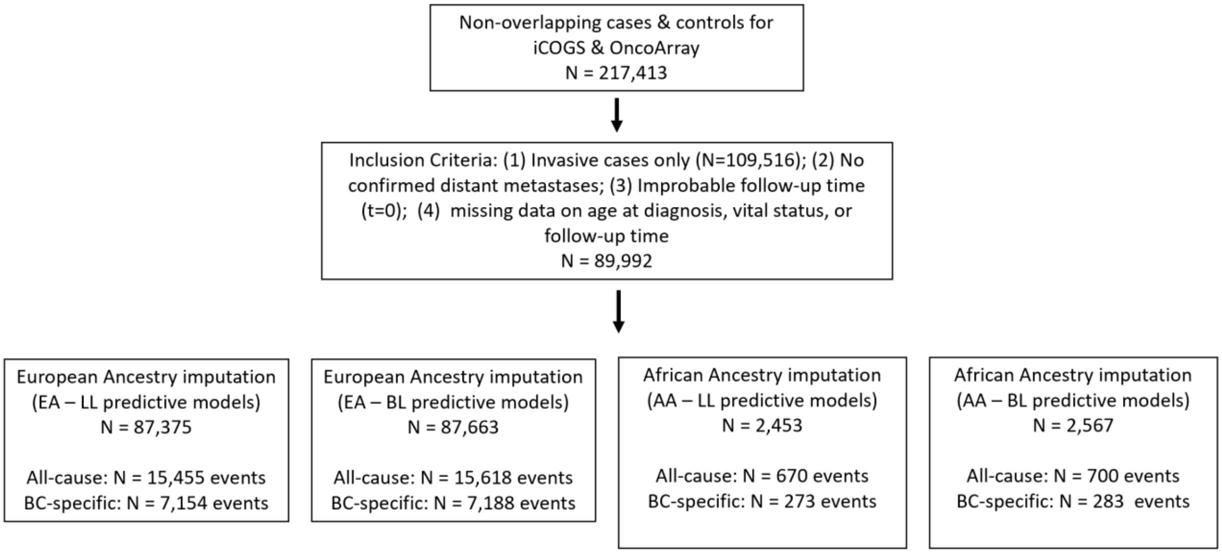
Impute GReX (genetically-regulated expression) using individual genotypes from BCAC
Test for associations between GReX and all-cause and breast cancer-specific mortality

Figure 5.4 Flow diagram indicating sample sizes for main GReX analysis based on inclusion/exclusion criteria


Non-overlapping cases & controls for iCOGS & OncoArray
N = 217,413
Inclusion Criteria: (1) Invasive cases only (N=109,516); (2) No confirmed distant metastases; (3) Improbable follow-up time (t=0); (4) missing data on age at diagnosis, vital status, or follow-up time
N = 89,992
European Ancestry and Luminal-Like GReX imputation (EA-LL predictive models)
N = 39,778 (5,701 events all-cause)
Mean age = 58 years
N = 38,410 (1,368 unknown cause) (2,668 events BC-specific)
Mean age = 57 years
European Ancestry and Basal-Like GReX imputation (EA-BL predictive models)
N = 5,310 (1,144 events all-cause)
Mean age: 55 years
N = 5,014 (296 unknown cause) (68 events BC-specific)
Mean age: 54 years
African Ancestry and Luminal-Like GReX imputation (AA-LL predictive models)
N = 869 (167 events all-cause)
Mean age: 52 years
N = 817 (52 unknown cause) (54 events BC-specific)
Mean age: 52 years
African Ancestry and Basal-Like GReX imputation (AA-BL predictive models)
N = 358 (112 events all-cause)
Mean age: 50 years
N = 303 (55 unknown cause) (32 events BC-specific)
Mean age: 50 years

Abbreviations: EA – European ancestry, AA – African ancestry, LL – Luminal-like, BL – Basal-like

Figure 5.5 Flow diagram indicating sample sizes for supplementary GReX analysis based on inclusion/exclusion criteria


Non-overlapping cases & controls for iCOGS & OncoArray N = 217,413
Inclusion Criteria: (1) Invasive cases only (N=109,516); (2) No confirmed distant metastases; (3) Improbable follow-up time (t=0); (4) missing data on age at diagnosis, vital status, or follow-up time N = 89,992
European Ancestry imputation (EA – LL predictive models) N = 87,375 All-cause: N = 15,455 events BC-specific: N = 7,154 events
European Ancestry imputation (EA – BL predictive models) N = 87,663 All-cause: N = 15,618 events BC-specific: N = 7,188 events
African Ancestry imputation (AA – LL predictive models) N = 2,453 All-cause: N = 670 events BC-specific: N = 273 events
African Ancestry imputation (AA – BL predictive models) N = 2,567 All-cause: N = 700 events BC-specific: N = 283 events

Abbreviations: EA – European ancestry, AA – African ancestry, LL – Luminal-like, BL – Basal-like

Figure 5.6 T-statistic for significant expression quantitative trait loci (eQTL) genes (top eQTL per gene shown) across Luminal-like (LL) and Basal-like (BL) samples among European ancestry individuals (EA); B) T-statistic for significant expression quantitative trait loci (eQTL) genes (top eQTL per gene shown) across Luminal-like (LL) and Basal-like (BL) samples among African ancestry individuals (AA). Across A and B), darkgreen, red, and blue denote significance among LL, BL, and both samples, respectively

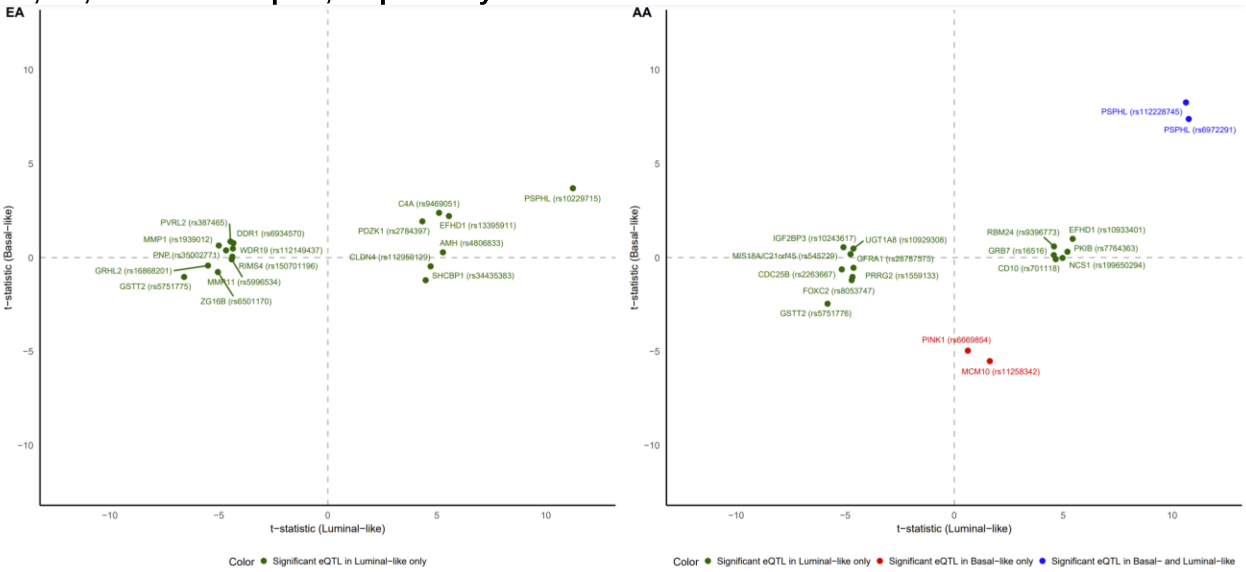
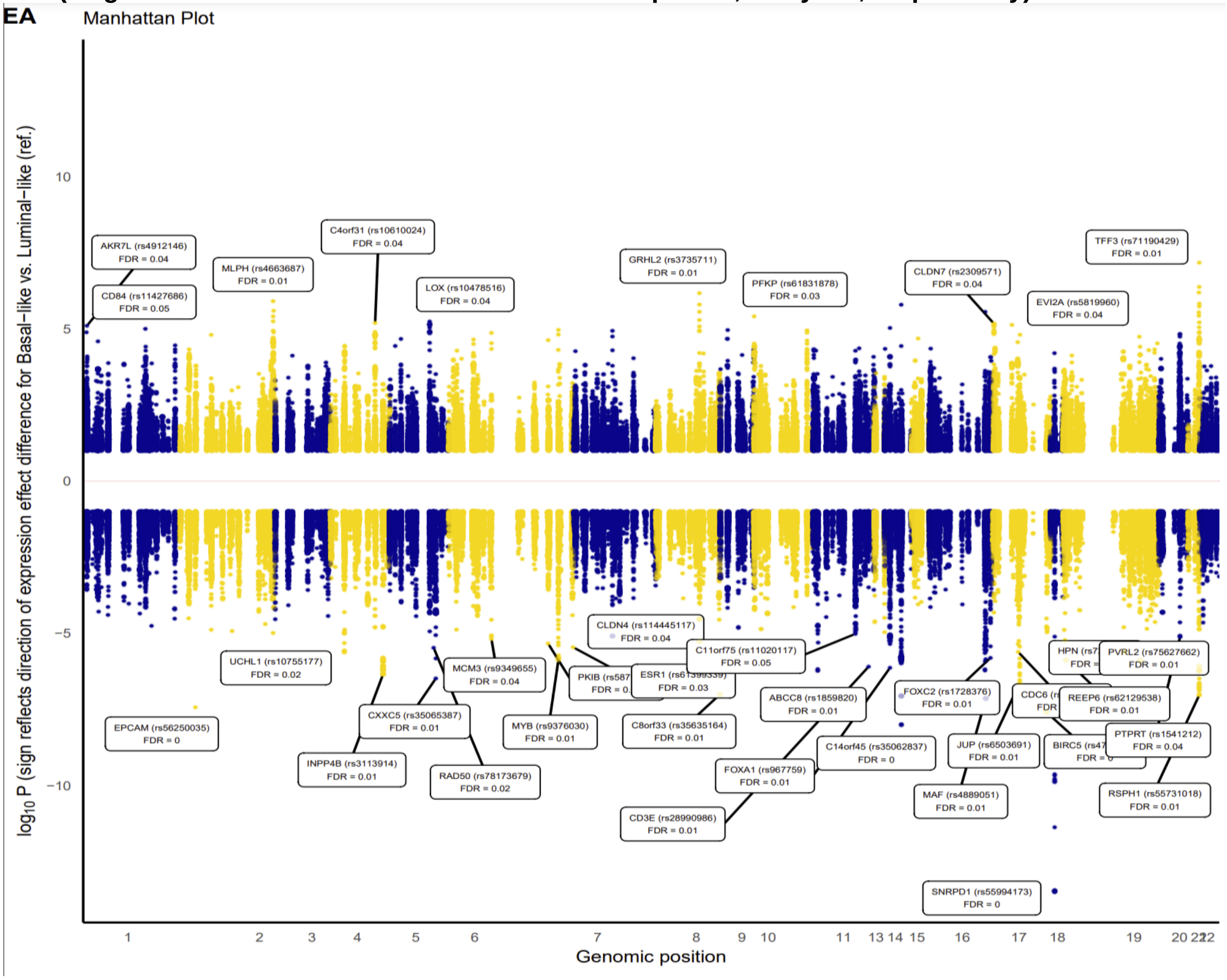


Figure 5.7 Manhattan plot indicating top eQTL for genes with significantly different germline-regulated expression across Luminal-like (LL) and Basal-like (BL) samples among European ancestry (EA) individuals (top panel); Manhattan plot indicating top eQTL for genes with significantly different germline-regulated expression across Luminal-like (LL) and Basal-like (BL) samples among African ancestry (AA) individuals (bottom panel). FDR denotes the false discovery rate (all genes shown had FDR <0.05 in EA and AA-specific, analyses, respectively)



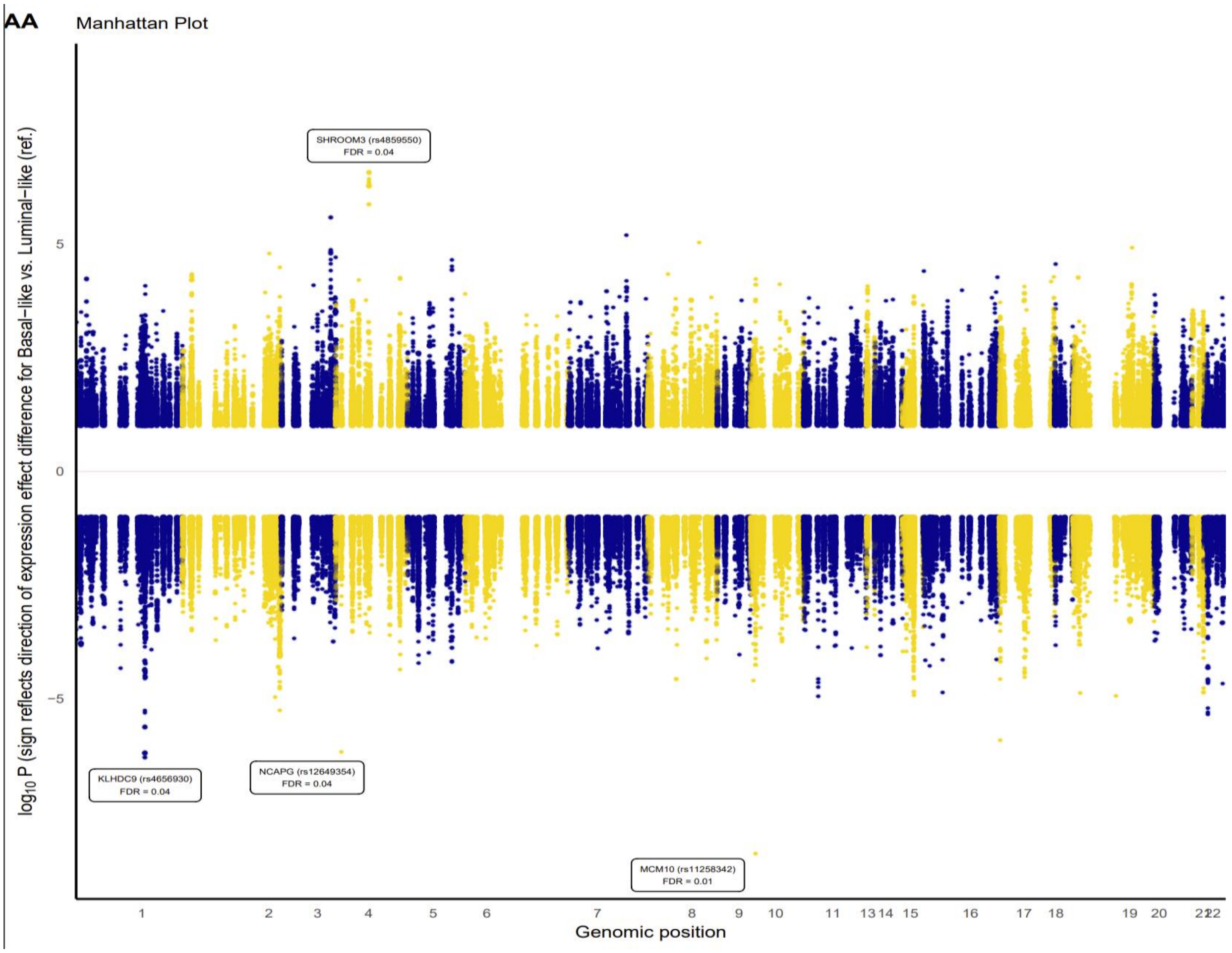
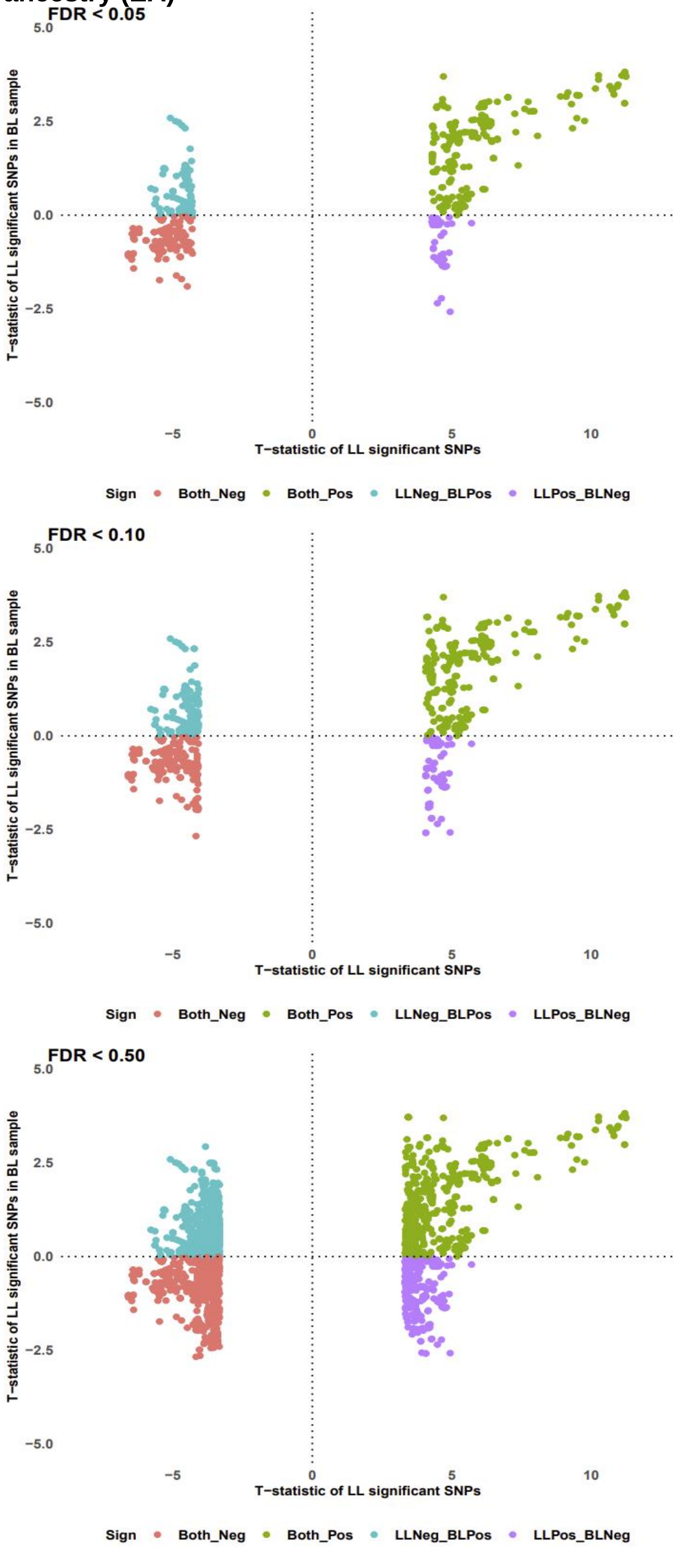

AA
Manhattan Plot
log$_{10}$ P (sign reflects direction of expression effect difference for Basal-like vs. Luminal-like (ref.)
5
0
−5
SHROOM3 (rs4859550)
FDR = 0.04
KLHDC9 (rs4656930)
FDR = 0.04
NCAPG (rs12649354)
FDR = 0.04
MCM10 (rs11258342)
FDR = 0.01
1
2
3
4
5
6
7
8
9
10
11
13
14
15
16
17
18
19
20
21
22
Genomic position

Figure 5.8 Concordance in eQTL effect direction across Luminal-like (LL) and Basal-like (BL) samples for LL significant SNPs, compared across varying significance (false discovery rate (FDR)) thresholds of 0.05, 0.10, and 0.50. Analysis conducted among individuals of European ancestry (EA)



FDR < 0.05

FDR < 0.10

FDR < 0.50

Figure 5.9 Concordance in eQTL effect direction across Luminal-like (LL) and Basal-like (BL) samples for LL significant SNPs, compared across varying significance (false discovery rate (FDR)) thresholds of 0.05, 0.10, and 0.50. Analysis conducted among individuals of African ancestry (AA)

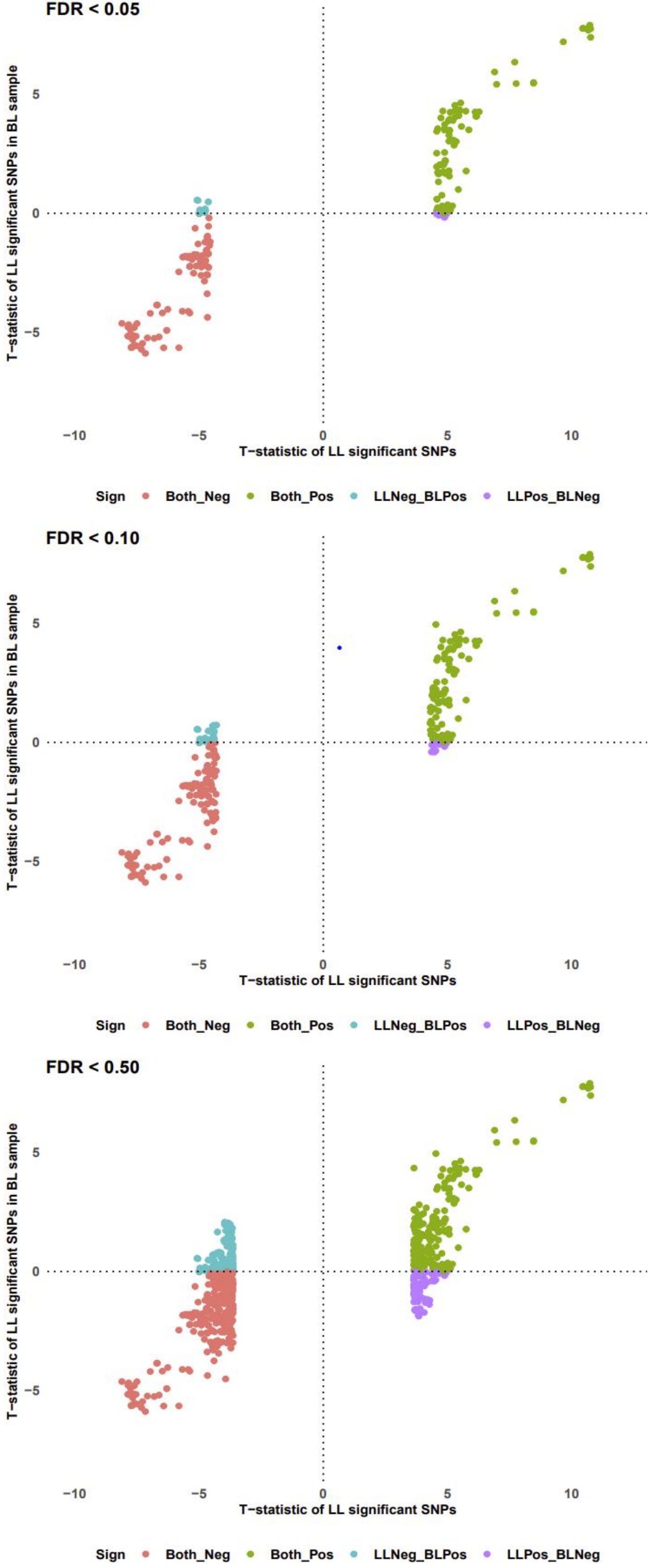
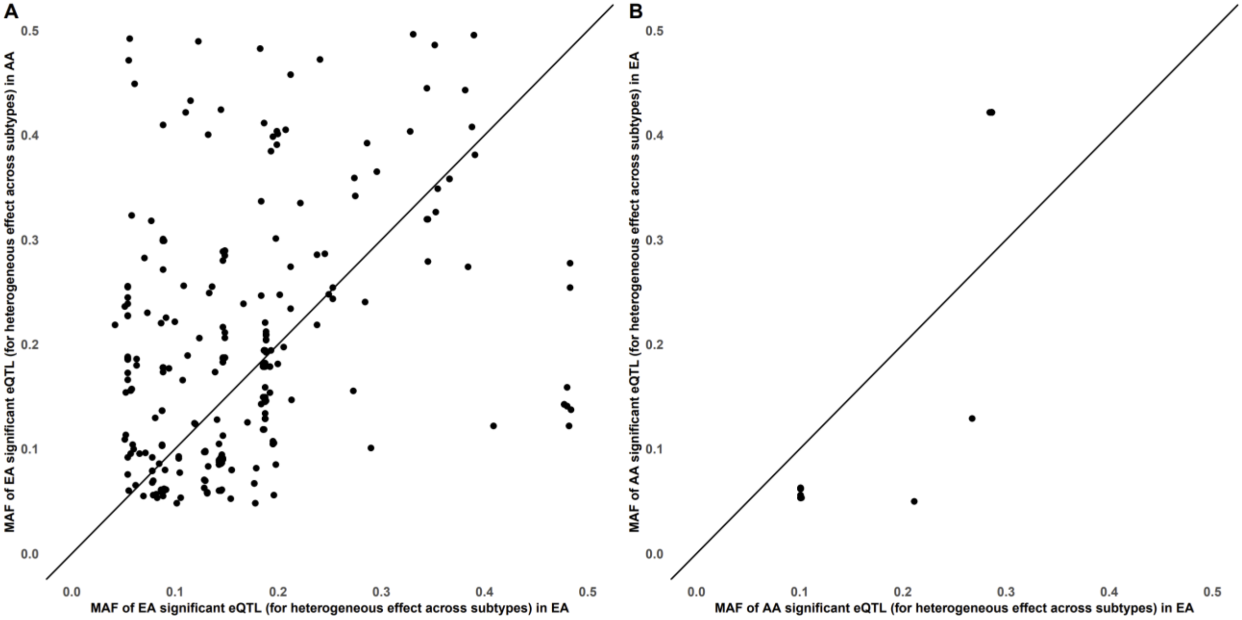
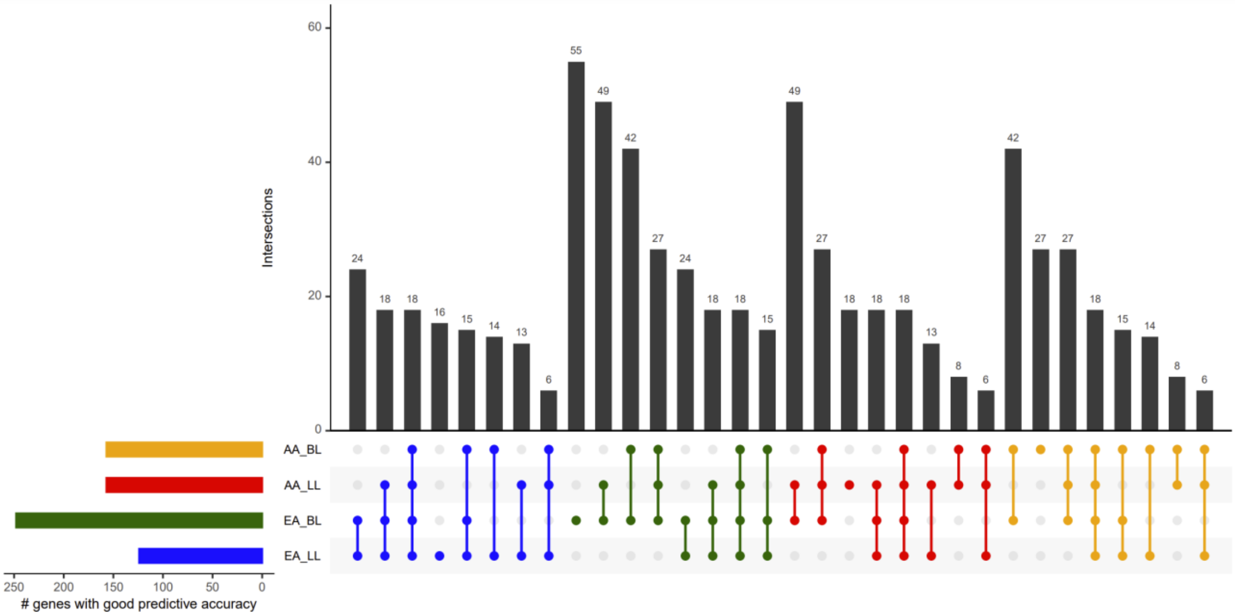


Figure 5.10 A) Comparison of minor allele frequency (MAF) of EA significant eQTL (for heterogeneous effect on tumor expression across subtypes) across EA and AA samples. B) Comparison of minor allele frequency (MAF) of AA significant eQTL (for heterogeneous effect on tumor expression across subtypes) across EA and AA samples



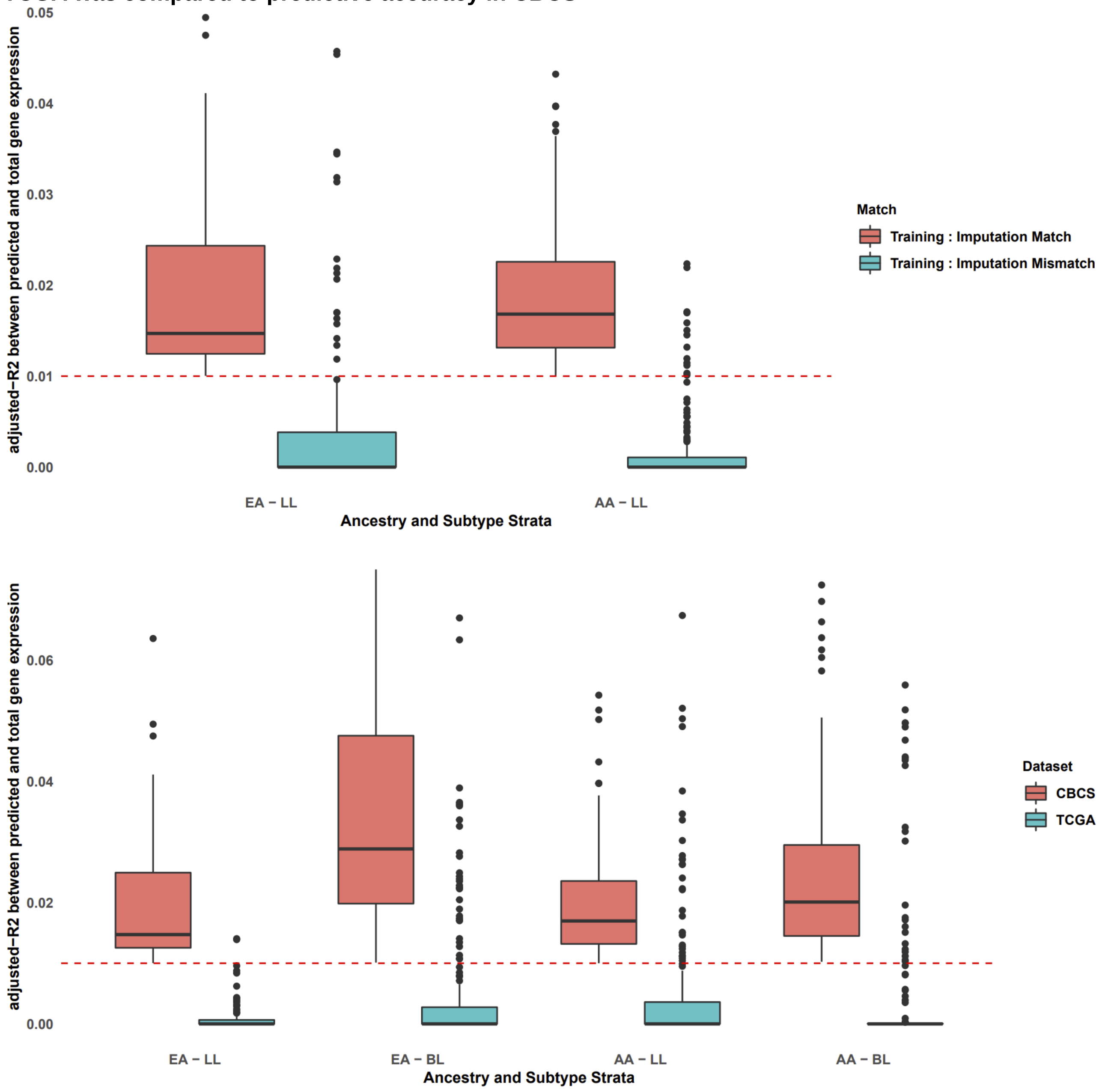
Abbreviations: EA – European ancestry, AA – African ancestry, LL – Luminal-like, BL – Basal-like, MAF – Minor Allele Frequency

Figure 5.11 UpSet plot of intersections between well-predicted genes (defined as adjusted-R^2 > 0.01 between observed and predicted expression) for ancestry and subtype strata investigated in this study. Well-predicted genes per ancestry and subtype strata were selected for association testing with breast cancer mortality



Abbreviations: EA – European ancestry, AA – African ancestry, LL – Luminal-like, BL – Basal-like

Figure 5.12 A) Assessment of predictive model transportability across subtypes, per ancestry stratum. 'Training: Imputation Match' denotes that models were trained and imputed in the same sample (Models trained in European ancestry – Luminal-like sample were imputed in European ancestry – Luminal-like sample while models trained in African ancestry – Luminal-like sample were imputed in African ancestry – Luminal-like sample). Mismatch denotes that models were trained in Luminal-like samples but imputed in Basal-like samples. Predictive accuracy (adjusted-*R*^{2} between observed and predicted (imputed) expression) were compared across matched and mismatched imputations to determine model transportability. B) External validation of Carolina Breast Cancer Study (CBCS) trained predictive models in The Cancer Genome Atlas (TCGA) data. Imputation was performed in TCGA data using CBCS trained models, and predictive accuracy in TCGA was compared to predictive accuracy in CBCS



Abbreviations: EA – European ancestry, AA – African ancestry, LL – Luminal-like, BL – Basal-like

Table 5.2 Number of genes meeting various adjusted-R2 (observed vs. predicted expression) cutoffs in TCGA data, for imputation performed using CBCS trained models (only CBCS models with good predictive accuracy tested)

	Genes Imputed	0.01	0.0025	0.0001
EA - LL	106	2	14	36
AA - LL	136	30	36	41
AA - BL	143	26	34	37
EA - BL	214	32	57	66

Abbreviations: EA – European ancestry, AA – African ancestry, LL – Luminal-like, BL – Basal-like, CBCS – Carolina Breast Cancer Study, TCGA – The Cancer Genome Atlas

Table 5.3. Five most nominally significant associations between germline-regulated gene expression (GReX; constructed across race-subtype strata) and breast cancer (BC) mortality endpoints (all-cause, BC-specific) – Main (predictive models in CBCS and imputation in BCAC aligned on ancestry and subtype) GReX analysis

Gene	Chr	Start *	End	Ancestry-Subtype	Endpoint	SH HR^	SH LCI	SH UCI	SH *p*-val	SH FDR	nonSH HR	non SH LCI	non SH UCI	nonSH *p*-val	non SH FDR	Incidence *p*-val
PTGER3	1	7.13E+07	7.15E+07	EA - BL	All Cause	0.88	0.82	0.94	9.97E-05	0.14	0.89	0.84	0.95	2.60E-04	0.18	7.36E-02
CRYBB2	22	2.56E+07	2.56E+07	AA - BL	BC-specific	1.97	1.26	3.07	2.76E-03	0.93	1.80	1.24	2.61	1.95E-03	0.46	4.50E-01
ZEB2	2	1.45E+08	1.45E+08	EA - BL	BC-specific	0.88	0.81	0.96	2.78E-03	0.93	0.88	0.81	0.95	2.04E-03	0.46	9.66E-01
RAD17	5	6.87E+07	6.87E+07	EA - LL	All Cause	0.95	0.92	0.99	4.09E-03	0.93	0.98	0.95	1.00	1.02E-01	0.93	1.67E-04
RAB25	1	1.56E+08	1.56E+08	AA - BL	All Cause	1.37	1.10	1.71	4.68E-03	0.93	1.44	1.18	1.75	2.63E-04	0.18	1.46E-01

Abbreviations: EA – European ancestry, AA – African ancestry, LL – Luminal-like, BL – Basal-like, CBCS – Carolina Breast Cancer Study, GReX – Germline-regulated gene expression, BCAC – Breast Cancer Association Consortium

*Start and end genomic coordinates for gene are build 37

^ HR – Hazard Ratio, from fixed effect, inverse-variance weighted Cox Proportional Hazards Regression

CHAPTER 6. DISCUSSION

6.1 Summary of findings

In Aim 1 of this dissertation, we assessed germline-regulated gene expression (GReX) for genes in relation to BC subtype for individuals of European and African ancestries. In doing so, we leveraged patterns of germline-regulation of gene expression across the two most relevant tissue contexts (normal breast, breast tumor) for BC (and by extension, BC subtypes). We find 53 GReX-subtype associations after Bayesian correction for potential test inflation bias [123] and a conservative global FDR <0.05, which spans 40 unique GReX-prioritized genes. All 40 GReX-prioritized genes were among EA individuals, though we do find suggestive associations for 5 genes among AA individuals (FDR = 0.06 – 0.18). Of the 40 GReX-prioritized genes for EA individuals, 10 were significant for both subtypes (LL, BL), 34 were unique to LL, and 6 unique to BL. We identify 7 genes with genomic overlap with established GWAS signal (for BC vs. controls), and of these, 4 genes (*MUC1, CCNE1, HLA-DOB,* and *KCNN4)* show concordance in effect direction with GWAS effect at the genomic locus [2], making these 4 genes prime candidates as mediators of previously identified GWAS signals. We find moderate to strong literature support for many of our GReX-prioritized genes (e.g., *DDR1, PDSS1, MUC1, AURKA*) in terms of the direction of effect and known aspects of that gene's biological function and/or relation to BC disease phenotypes [141,142,144-146,152]. This work marks an important contribution, both from the standpoint of findings we uncover and also in terms of the interpretability (i.e., gene level) of those findings, towards the understanding of the germline etiology of BC subtypes.

In Aim 2 of this dissertation, we leveraged biological heterogeneity (i.e., subtype) within the context of ancestry-specific germline investigation of BC mortality (all-cause, BC-specific). This work built off long standing conjecture that accounting for biological heterogeneity may be key towards better understanding of the germline basis of BC mortality. Compared to all previous genetic investigations of BC mortality [82-92], we included a formal correction for potential collider bias, an important and increasingly recognized bias that impacts studies of mortality (due to inherent conditioning on case-only status in mortality analyses) [79,155-158]. In this work we find differences in germline-regulation of tumor gene expression across subtypes, within ancestry groups (321 loci across 37 genes among EA, 23 loci across 4 genes among AA). However, despite accounting for biological heterogeneity and additional rigor for collider bias correction, we did not find any associations for BC mortality at a conservative, multiple-testing correction threshold (*PTGER3* had a suggestive association among EA – BL sample for all-cause mortality at FDR = 0.14), which is in line with previous genetic investigations [82-92] of BC mortality (although these studies did not correct for collider bias). Our findings suggest that future work should turn towards account for granular treatment information (importantly, treatment is a descendent of a collider in causal frameworks, therefore accounting for treatment should always first include correction for potential collider bias), as treatment effects may overwhelm more modest genetic contributions towards survival [92].

6.2 Limitations and Strengths

We note several limitations across the two studies performed as part of this dissertation. Across both study aims, a key data source (in relation to our focus on cross-ancestry analyses) for construction of predictive models of breast tumor gene expression from germline variation was the CBCS. In Aim 1, CBCS was used to construct predictive models for European and African ancestries, respectively, while in Aim 2, CBCS was used to construct ancestry and subtype-specific predictive models (i.e., AA – LL, AA – BL etc.). CBCS is the ideal data source for construction of such predictive models because it offers matched genotype – gene expression data on a large number of AA individuals (n = 628 AA, n = 571 EA in CBCS). In comparison, matched genotype-expression data for other data sources such as TCGA and GTEx is 170 and 47, respectively, for AA individuals. While CBCS offers unique analytic opportunities, it also poses some limitations. First, CBCS used a custom NanoString nCounter probeset for RNA expression quantification of BC-related genes, and therefore we were not able to analyze the whole transcriptome. We were limited to analysis of 396 BC-related genes (and by extension, however many of these 396 were available across other reference panels to ensure comparability). While this is a limitation, it also reduces multiple testing burden and raises biological plausibility of findings, as the 396 genes have all been implicated in key BC pathways (Appendix 1) and are protein coding genes.

A key limitation in Aim 1 is that although our subtype categorization is aimed at uncovering differences in potential germline genetic underpinning along a key demarcation in BC (BL (most aggressive) versus LL (less aggressive)), this classification may not be fine-tuned enough for certain genes (e.g., HER2). We emphasize that our choice of BL v LL comparison was in part, also motivated by lack of sample size for the fine-tuned subtype categories (e.g., HER2) among AA individuals. Therefore, as sample sizes for more fine-tuned classifications become available across ancestry groups, cross-ancestry analysis focus should shift along these more granular subtype classifications.

Third, although we took care ensure analytic consistency across the reference panels (e.g., we re-imputed CBCS, TCGA, and NBS data to the TOPMed panel to mitigate differences in genotyping quality), there were differences in expression quantification platforms (e.g., RNA-seq, microarray, NanoString) that make full comparability of findings across reference panels challenging.

Finally, although we leveraged among the largest available genetic resources for AA individuals, we note the lesser sample sizes we had for the GReX-analysis among AA compared to EA, both for predictive model construction for Aim 1 (with the exception of CBCS) and for association testing in BCAC (for both Aims). As such, interpretations of ancestry-specific findings along both aims should be made with caution. More progress to increase sample sizes of genetic data resources for diverse populations is needed, especially for publicly available data resources such as the GTEx and TCGA.

There are several strengths to this dissertation work. In terms of innovation, this is the first study to systematically evaluate germline-genetic basis for BC subtypes across individuals of European and African ancestry. As a gene-level, TWAS-based analysis with reduced multiple testing burden, we were able to prioritize potential causal genes for BC subtype at loci where there might be aggregation of small effects (which would be missed under GWAS), and in conjunction, we prioritized several genes which overlap with known GWAS loci (for BC), offering insight into a potential genetic mechanism for those loci (notably, most GWAS signals are found in regulatory regions and an immediate genetic mechanism for these loci is lacking [11,62,102,173]). Since our analyses were gene level, there is also increased interpretability of findings, as prioritized genes can be subject to functional follow-up testing/experiments. We employ conservative assessment of GReX signals (i.e., where applicable correction for test statistic inflation through a Bayesian approach called *bacon* and a global FDR threshold of 0.05 on *bacon* corrected test statistics [123]). We also performed external validation for all predictive models constructed across the two aims, a feature absent from many TWAS-based analyses [119]. In Aim 2, ours is the first study to formally correct for potential collider stratification bias, which is an important and increasingly recognized bias that impacts studies of mortality [79,155-158]. Lastly, our work leverages among the largest available data resources for individuals of African ancestry, which is a small step towards bridging the gap in genetic investigations across diverse populations.

6.3 Future directions

Our work here provides a platform for a multitude of ideas for future investigations. In Aim 1, we analyzed what we hypothesized to be the two most important tissue contexts (normal breast, breast tumor) for understanding of the germline genetic basis (as mediated through gene expression) for BC subtypes. Other tissue contexts that may be of etiologic relevance include adipose tissue, fibroblasts, and immune tissue, all cell types found within bulk tumor tissue and some with direct relevance to BC etiology [174-176]. Additionally, assessment of more granular subtype definitions (LumA, LumB/HER2-, LumB, HER2, Basal) might reveal further shared and potentially divergent germline etiologies for these subtypes. In fact, in another study, we have performed a MOSTWAS[177] (a multiomic extension TWAS where we include tissue-specific distal regulatory variants) of granular BC subtypes across etiologic tissue (albeit for individuals of European ancestry only due to limited availability of multiomic data for individuals of non-European ancestry). Presently, the limited sample sizes of genotype-expression (and multiomic) data for non-European ancestry individuals in publicly available data resources such as GTEx and TCGA represents a major public health gap. Along with larger sample sizes, more biologically rich (e.g., breast tumor, tumor-adjacent, and normal tissue) data has the potential to facilitate discovery of potentially differential and disease-continuum specific germline-regulation of gene expression pathways relevant to breast carcinogenesis. Lastly, our findings for Aim 2 point towards increased emphasis on treatment as part of the analytic framework, and as part of inclusion of treatment within the analytic framework, careful correction for potential collider bias (as treatment in BC mortality causal frameworks is a descendant of a collider (BC)).

APPENDIX 1. THE 396 BREAST-CANCER RELATED GENES ANALYZED

Gene	Kinome	PAM50	P53	Claudin-low	Hypoxia	Methylation	Race	E2	IGF	EMT	Tamoxifen	Immune	HGF	EGFR	DNA Repair	Others	Housekeeping
TMEM158	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
VIM	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
RAD50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
MRPL19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
PSMC4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
SF3A1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
VEGFA	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0
TNIK	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ERBB2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BRCA1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0
ACTB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
RPLP0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
TCEAL1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TRIP13	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AURKA	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCNE1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CDC20	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NDC80	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NUF2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RRM2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TMEM45B	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GRB7	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
BAG1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BCL2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BIRC5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCNB1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ESR1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MMP11	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PGR	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CEP55	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PTTG1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
UBE2C	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MKI67	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYBL2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

TYMS	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
MSH3	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
STK38	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AXL	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
CAV1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
CD24	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
CLDN4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
DSP	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
EMP3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ESRP1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
EVI2A	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
F11R	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
FBN1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
GNG11	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
GRHL2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
JUP	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT19	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT8	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
LEPRE1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
LHFP	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
MPP1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
NT5E	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
PVRL3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
RAB25	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
SH2B3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
SPINT1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
SPINT2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
ZEB1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
EPCAM	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
ADM	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ANGPTL4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
DDIT4	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
FABP5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
FLVCR2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
GAL	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
NDRG1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
PLOD1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
PNP	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
RRAGD	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
SLC16A3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
UCHL1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

ERBB4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ULK1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DDR1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
DAPK1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
KIT	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
CRMP1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
GSTP1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ADHFE1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
AMH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
AMHR2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
CLDN3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
CLDN7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
CRYAB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
ERBB3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
MET	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
PIK3CA	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
PTEN	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
RAD17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
RB1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
APH1B	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ATAD2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BTG2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCNA2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CCND1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CDC25B	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CDC25C	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CDCA7L	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CDK1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CDKN1A	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CDKN3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CKS1B	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DDB2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FAM198B	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FAM214A/KIAA1370	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FNBP1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FOXM1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GGH	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KIAA0040	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KIF23	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KIFC1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LOC400043	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

MAD2L1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MAP2K4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MCM3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MIS18A/C21orf45	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NCAPH2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NEO1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NPEPPS	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NUDT1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
POLD1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PREP	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RFC4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
RNF103	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TOP2A	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TUBA4A	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GATA3	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ACTR3B	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ANLN	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BLVRA	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CDC6	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CDH3	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CXXC5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EXO1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FOXA1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FOXC1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GPR160	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KIF2C	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT14	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT17	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KRT5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MAPT	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MDM2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MELK	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MIA	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MLPH	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
MYC	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NAT1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ORC6L	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
UBE2T	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EGFR	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
FGFR4	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
PHGDH	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0

SFRP1	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
CENPF	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SLC39A6	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ACOX2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
CRYBB2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
FAM177A1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
GSTT2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
MUC1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
PSPH	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
PSPHL	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
SQLE	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
TRPC1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ABAT	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ABCC8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
AR	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
AURKB	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
BUB1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
C10orf116	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
C16orf45	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
C1orf106	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
C4A	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CCDC103	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CCNB2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CDC45	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CDCA5	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CDCA8	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CELSR1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CENPA	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CENPN	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
CYP4B1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
DEPDC1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
DLGAP5	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
FAM54A	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
FAM64A	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
GTSE1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
HJURP	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
HPN	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
IDO1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
IGF2BP3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
KDM4B	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
KLHDC9	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

LRG1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
MMP1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
MND1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
NCAPG	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
NTN4	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
NXNL2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
PDZK1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
PLK1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
PTPRT	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
RAD54L	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
RBM24	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
RSPH1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
SCGB1D2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
SEC14L2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
SEMA3B	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
SHCBP1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
STC2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
SYT1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
TRAT1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
XCL1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
ZG16B	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
LRP8	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
RAI2	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
TWIEST2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ZEB2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ALDH1A1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
CD10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
FOXC2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
OCLN	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ACADSB	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ADCY1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
APBB2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
BTG3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
C11orf75	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
C14orf45	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
CXCR4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ELOVL2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
EZH2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
FAM63A	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
FLJ20152	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
FMO5	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

FSCN1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
IFRD1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
IRS1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ITGB5	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
LRRC50	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
MAGED2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
NXPH4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
PCSK6	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
PFKP	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
PINK1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
PTGER3	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
REPS2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
RNASE4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
RPS6KB2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
SLC1A2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
SLC7A5	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
XBP1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
TPX2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
UGT1A10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
VAV3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
WDR12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
BMP2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
CYP19A1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
CYP27A1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
CYP7B1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
IL12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
IL1B	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
IL6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
LOX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
NR1H3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
PGE3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
PTGS2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
SERPINB5(MASPIN)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
TBC1D9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
TRPM7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
ABCB1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CYP2D6	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CYP3A4	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
CYP3A5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
SULT1E1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
SULT2A1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0

UGT1A4	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
UGT1A8	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
UGT2B7	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
FANCA	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0
CDCA7	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
REEP6	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
MYB	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
SCUBE2	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0
CTSL2	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
SYBU	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0
ACTG1P3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
BOP1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
C8orf33	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
CACNB3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
CALCP	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
CMC2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
DNM2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
ECE2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
FBXL6	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
GALT	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
GUCA1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
HGH1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
KLHL7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
MRPS17	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
NCS1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
NLN	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
PGAM5	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
PTDSS1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
PUF60	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
SDCBP	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
SLC52A2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
SNRPD1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
WDR19	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
CDH1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
MMP2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
MMP3	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
SNAI1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
SNAI2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
SOX10	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
TWIST1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0
FN1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0

TMSB15B	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
AKR7L	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
AQP5	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
C1QTNF3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
C2orf27A	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
C4orf31	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
C9orf98	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
CAPN13	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
CASKIN1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
CMYA5	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
DOCK3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
DTX3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
EFHD1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
F7	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
FMNL2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
FUT8	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
GCNT2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
HRC	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
INPP4B	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ISLR2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
KCNMA1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
KCNN4	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
KIF3A	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
MAGI2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
MARVELD2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
PKIB	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
PRRG2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
PRRT2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
PVRL2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
RIMS4	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
SHROOM3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
SKAP1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
CLTC	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
HPRT1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
PGK1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
TUBB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
GFRA1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0
CAPN9	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
IGF2BP2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
IL6ST	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
MCM10	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0

PDSS1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
S100A8	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
NME5	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
TFF3	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
TNFRSF17	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
BLK	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CCL7	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CCR3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CD19	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CD28	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CD3E	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CD3G	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CD4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CD6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CD84	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CD8A	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CD96	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CXCL13	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CXCL5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CYBB	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
FCRL2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
FOXP3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
FPRL1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
GPR44	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
GZMM	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
HLA-DOB	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ICOS	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
IL2RB	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
IL5RA	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
IL8RA	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
IL8RB	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
KIAA0125	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
LAG-3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
LCK	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
LILRB2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
MAF	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
MS4A1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
MSR1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
NFKB1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
PDCD1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
PD-L1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

PPBP	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
PRF1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
SH2D1A	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
SIRPG	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
TIM-3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
TRAF1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
ZAP70	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CD2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CD68	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
GAPDH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
GUSB	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

REFERENCES

1. DeSantis CE, Ma J, Gaudet MM, et al. Breast cancer statistics, 2019. *CA: A Cancer Journal for Clinicians.* 2019;69(6):438-451.

2. Zhang H, Ahearn TU, Lecarpentier J, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat Genet.* 2020;52(6):572-581.

3. Dorling L, Carvalho S, Allen J, et al. Breast Cancer Risk Genes - Association Analysis in More than 113,000 Women. *N Engl J Med.* 2021;384(5):428-439.

4. Michailidou K, Lindström S, Dennis J, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature.* 2017;551(7678):92-94.

5. Michailidou K, Beesley J, Lindstrom S, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet.* 2015;47(4):373-380.

6. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet.* 2013;45(4):353-361, 361e351-352.

7. Zhang H, Ahearn TU, Lecarpentier J, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nature Genetics.* 2020;52(6):572-581.

8. Ahearn TU, Zhang H, Michailidou K, et al. Common breast cancer risk loci predispose to distinct tumor subtypes. *bioRxiv.* 2019:733402.

9. Acheampong T, Kehm RD, Terry MB, Argov EL, Tehranifar P. Incidence Trends of Breast Cancer Molecular Subtypes by Age and Race/Ethnicity in the US From 2010 to 2016. *JAMA Network Open.* 2020;3(8):e2013226-e2013226.

10. Adedokun B, Du Z, Gao G, et al. Cross-ancestry GWAS meta-analysis identifies six breast cancer loci in African and European ancestry women. *Nature Communications.* 2021;12(1):4198.

11. Gusev A, Mancuso N, Won H, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature Genetics.* 2018;50(4):538-548.

12. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet.* 2016;48(3):245-252.

13. 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):1091-1098.

14. Benefield HC, Zabor EC, Shan Y, Allott EH, Begg CB, Troester MA. Evidence for Etiologic Subtypes of Breast Cancer in the Carolina Breast Cancer Study. *Cancer Epidemiol Biomarkers Prev.* 2019;28(11):1784-1791.

15. O'Brien KM, Cole SR, Tse CK, et al. Intrinsic breast tumor subtypes, race, and long-term survival in the Carolina Breast Cancer Study. *Clin Cancer Res.* 2010;16(24):6100-6110.

16. Bhattacharya A, García-Closas M, Olshan AF, Perou CM, Troester MA, Love MI. A framework for transcriptome-wide association studies in breast cancer in diverse study populations. *Genome Biology.* 2020;21(1):42.

17. Ravdin PM, Cronin KA, Howlader N, et al. The decrease in breast-cancer incidence in 2003 in the United States. *N Engl J Med.* 2007;356(16):1670-1674.

18. Anderson WF, Katki HA, Rosenberg PS. Incidence of breast cancer in the United States: current and future trends. *J Natl Cancer Inst.* 2011;103(18):1397-1402.

19. Cronin KA, Ravdin PM, Edwards BK. Sustained lower rates of breast cancer in the United States. *Breast Cancer Res Treat.* 2009;117(1):223-224.

20. Glass AG, Lacey JV, Jr., Carreon JD, Hoover RN. Breast cancer incidence, 1980-2006: combined roles of menopausal hormone therapy, screening mammography, and estrogen receptor status. *J Natl Cancer Inst.* 2007;99(15):1152-1161.

21. Davis Lynn BC, Rosenberg PS, Anderson WF, Gierach GL. Black-White Breast Cancer Incidence Trends: Effects of Ethnicity. *J Natl Cancer Inst.* 2018;110(11):1270-1272.

22. DeSantis CE, Fedewa SA, Goding Sauer A, Kramer JL, Smith RA, Jemal A. Breast cancer statistics, 2015: Convergence of incidence rates between black and white women. *CA Cancer J Clin.* 2016;66(1):31-42.

23. Institute NC. Cancer Stat Facts: Female Breast Cancer Subtypes. 2021.

24. Pharoah PD, Day NE, Duffy S, Easton DF, Ponder BA. Family history and the risk of breast cancer: a systematic review and meta-analysis. *Int J Cancer.* 1997;71(5):800-809.

25. Brewer HR, Jones ME, Schoemaker MJ, Ashworth A, Swerdlow AJ. Family history and risk of breast cancer: an analysis accounting for family structure. *Breast Cancer Res Treat.* 2017;165(1):193-200.

26. Mucci LA, Hjelmborg JB, Harris JR, et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *Jama.* 2016;315(1):68-76.

27. Takaoka M, Miki Y. BRCA1 gene: function and deficiency. *Int J Clin Oncol.* 2018;23(1):36-44.

28. Mehrgou A, Akouchekian M. The importance of BRCA1 and BRCA2 genes mutations in breast cancer development. *Med J Islam Repub Iran.* 2016;30:369.

29. Kuchenbaecker KB, Hopper JL, Barnes DR, et al. Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers. *Jama.* 2017;317(23):2402-2416.

30. Ford D, Easton DF, Peto J. Estimates of the gene frequency of BRCA1 and its contribution to breast and ovarian cancer incidence. *Am J Hum Genet.* 1995;57(6):1457-1462.

31. Whittemore AS, Gong G, John EM, et al. Prevalence of BRCA1 mutation carriers among U.S. non-Hispanic Whites. *Cancer Epidemiol Biomarkers Prev.* 2004;13(12):2078-2083.

32. Group ABCS. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. *Br J Cancer.* 2000;83(10):1301-1308.

33. Papelard H, de Bock GH, van Eijk R, et al. Prevalence of BRCA1 in a hospital-based population of Dutch breast cancer patients. *Br J Cancer.* 2000;83(6):719-724.

34. Newman B, Mu H, Butler LM, Millikan RC, Moorman PG, King MC. Frequency of breast cancer attributable to BRCA1 in a population-based series of American women. *Jama.* 1998;279(12):915-921.

35. Loman N, Johannsson O, Kristoffersson U, Olsson H, Borg A. Family history of breast and ovarian cancers and BRCA1 and BRCA2 mutations in a population-based series of early-onset breast cancer. *J Natl Cancer Inst.* 2001;93(16):1215-1223.

36. Malone KE, Daling JR, Doody DR, et al. Prevalence and predictors of BRCA1 and BRCA2 mutations in a population-based study of breast cancer in white and black American women ages 35 to 64 years. *Cancer Res.* 2006;66(16):8297-8308.

37. Couch FJ, Hart SN, Sharma P, et al. Inherited mutations in 17 breast cancer susceptibility genes among a large triple-negative breast cancer cohort unselected for family history of breast cancer. *J Clin Oncol.* 2015;33(4):304-311.

38. Institute NC. Genetics of Breast and Gynecologic Cancers. 2021.

39. Mitchell G, Ballinger ML, Wong S, et al. High frequency of germline TP53 mutations in a prospective adult-onset sarcoma cohort. *PLoS One.* 2013;8(7):e69026.

40. Bertheau P, Lehmann-Che J, Varna M, et al. p53 in breast cancer subtypes and new insights into response to chemotherapy. *Breast.* 2013;22 Suppl 2:S27-29.

41. Brookes AJ. The essence of SNPs. *Gene.* 1999;234(2):177-186.

42. Zheng W, Long J, Gao YT, et al. Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat Genet.* 2009;41(3):324-328.

43. Thomas G, Jacobs KB, Kraft P, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat Genet.* 2009;41(5):579-584.

44. Garcia-Closas M, Couch FJ, Lindstrom S, et al. Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat Genet.* 2013;45(4):392-398, 398e391-392.

45. Bulik-Sullivan B, Finucane HK, Anttila V, et al. An atlas of genetic correlations across human diseases and traits. *Nature Genetics.* 2015;47(11):1236-1241.

46. Long J, Zhang B, Signorello LB, et al. Evaluating genome-wide association study-identified breast cancer risk variants in African-American women. *PLoS One.* 2013;8(4):e58350.

47. Huo D, Feng Y, Haddad S, et al. Genome-wide association studies in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer. *Hum Mol Genet.* 2016;25(21):4835-4846.

48. Feng Y, Rhie SK, Huo D, et al. Characterizing Genetic Susceptibility to Breast Cancer in Women of African Ancestry. *Cancer Epidemiol Biomarkers Prev.* 2017;26(7):1016-1026.

49. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences.* 2009;106(23):9362-9367.

50. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57-74.

51. Kundaje A, Meuleman W, Ernst J, et al. Integrative analysis of 111 reference human epigenomes. *Nature.* 2015;518(7539):317-330.

52. Dunning AM, Michailidou K, Kuchenbaecker KB, et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nat Genet.* 2016;48(4):374-386.

53. Ghoussaini M, Edwards SL, Michailidou K, et al. Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nature Communications.* 2014;5(1):4999.

54. Li Q, Seo JH, Stranger B, et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell.* 2013;152(3):633-641.

55. Darabi H, McCue K, Beesley J, et al. Polymorphisms in a Putative Enhancer at the 10q21.2 Breast Cancer Risk Locus Regulate NRBF2 Expression. *Am J Hum Genet.* 2015;97(1):22-34.

56. Lawrenson K, Kar S, McCue K, et al. Functional mechanisms underlying pleiotropic risk alleles at the 19p13.1 breast–ovarian cancer susceptibility locus. *Nature Communications.* 2016;7(1):12675.

57. Lee D, Gorkin DU, Baker M, et al. A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics.* 2015;47(8):955-961.

58. Finucane HK, Bulik-Sullivan B, Gusev A, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics.* 2015;47(11):1228-1235.

59. Gusev A, Lee SH, Trynka G, et al. Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet.* 2014;95(5):535-552.

60. Jansen R, Hottenga JJ, Nivard MG, et al. Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Hum Mol Genet.* 2017;26(8):1444-1451.

61. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci.* 2013;368(1620):20120362.

62. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics.* 2016;48(3):245-252.

63. Barbeira AN, Pividori M, Zheng J, Wheeler HE, Nicolae DL, Im HK. Integrating predicted transcriptome from multiple tissues improves association detection. *PLoS Genet.* 2019;15(1):e1007889.

64. 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. *Nature Genetics.* 2018;50(7):968-978.

65. Feng H, Gusev A, Pasaniuc B, et al. Transcriptome-wide association study of breast cancer risk by estrogen-receptor status. *Genet Epidemiol.* 2020;44(5):442-468.

66. Narod SA. Tumour size predicts long-term survival among women with lymph node-positive breast cancer. *Curr Oncol.* 2012;19(5):249-253.

67. Saadatmand S, Bretveld R, Siesling S, Tilanus-Linthorst MM. [Influence of tumour stage at breast cancer detection on survival in modern times: population based study in 173,797 patients]. *Ned Tijdschr Geneeskd.* 2016;160:A9800.

68. Schairer C, Mink PJ, Carroll L, Devesa SS. Probabilities of Death From Breast Cancer and Other Causes Among Female Breast Cancer Patients. *JNCI: Journal of the National Cancer Institute.* 2004;96(17):1311-1321.

69. Society AC. Survival Rates for Breast Cancer. 2021.

70. Schwartz AM, Henson DE, Chen D, Rajamarthandan S. Histologic grade remains a prognostic factor for breast cancer regardless of the number of positive lymph nodes and tumor size: a study of 161 708 cases of breast cancer from the SEER Program. *Arch Pathol Lab Med.* 2014;138(8):1048-1052.

71. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J Clin Oncol.* 2014;5(3):412-424.

72. Carey LA, Perou CM, Livasy CA, et al. Race, Breast Cancer Subtypes, and Survival in the Carolina Breast Cancer Study. *JAMA.* 2006;295(21):2492-2502.

73. Guerra MR, Silva GA, Nogueira MC, et al. Breast cancer survival and health iniquities. *Cad Saude Publica.* 2015;31(8):1673-1684.

74. Panagopoulou P, Gogas H, Dessypris N, Maniadakis N, Fountzilas G, Petridou ET. Survival from breast cancer in relation to access to tertiary healthcare, body mass index, tumor characteristics and treatment: a Hellenic Cooperative Oncology Group (HeCOG) study. *Eur J Epidemiol.* 2012;27(11):857-866.

75. Mohammed SN, Smith P, Hodgson SV, et al. Family history and survival in premenopausal breast cancer. *Br J Cancer.* 1998;77(12):2252-2256.

76. Slattery ML, Berry TD, Kerber RA. Is survival among women diagnosed with breast cancer influenced by family history of breast cancer? *Epidemiology.* 1993;4(6):543-548.

77. Gajalakshmi CK, Shanta V, Hakama M. Survival from contralateral breast cancer. *Breast Cancer Res Treat.* 1999;58(2):115-122.

78. Russo A, Herd-Smith A, Gestri D, et al. Does family history influence survival in breast cancer cases? *Int J Cancer.* 2002;99(3):427-430.

79. Cole SR, Platt RW, Schisterman EF, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol.* 2010;39(2):417-420.

80. Shahbandi A, Nguyen HD, Jackson JG. TP53 Mutations and Outcomes in Breast Cancer: Reading beyond the Headlines. *Trends Cancer.* 2020;6(2):98-110.

81. Huszno J, Kołosza Z, Grzybowska E. BRCA1 mutation in breast cancer patients: Analysis of prognostic factors and survival. *Oncol Lett.* 2019;17(2):1986-1995.

82. Khan S, Fagerholm R, Rafiq S, et al. Polymorphism at 19q13.41 Predicts Breast Cancer Survival Specifically after Endocrine Therapy. *Clin Cancer Res.* 2015;21(18):4086-4096.

83. Guo Q, Schmidt MK, Kraft P, et al. Identification of novel genetic markers of breast cancer survival. *J Natl Cancer Inst.* 2015;107(5).

84. Rafiq S, Khan S, Tapper W, et al. A genome wide meta-analysis study for identification of common variation associated with breast cancer prognosis. *PLoS One.* 2014;9(12):e101488.

85. Rafiq S, Tapper W, Collins A, et al. Identification of inherited genetic variations influencing prognosis in early-onset breast cancer. *Cancer Res.* 2013;73(6):1883-1891.

86. Shu XO, Long J, Lu W, et al. Novel genetic markers of breast cancer survival identified by a genome-wide association study. *Cancer Res.* 2012;72(5):1182-1189.

87. Azzato EM, Pharoah PD, Harrington P, et al. A genome-wide association study of prognosis in breast cancer. *Cancer Epidemiol Biomarkers Prev.* 2010;19(4):1140-1143.

88. Chou WC, Hsiung CN, Chen WT, et al. A functional variant near XCL1 gene improves breast cancer survival via promoting cancer immunity. *Int J Cancer.* 2020;146(8):2182-2193.

89. Escala-Garcia M, Guo Q, Dörk T, et al. Genome-wide association study of germline variants and breast cancer-specific mortality. *Br J Cancer.* 2019;120(6):647-657.

90. Escala-Garcia M, Abraham J, Andrulis IL, et al. A network analysis to identify mediators of germline-driven differences in breast cancer prognosis. *Nature Communications.* 2020;11(1):312.

91. Kadalayil L, Khan S, Nevanlinna H, et al. Germline variation in ADAMTSL1 is associated with prognosis following breast cancer treatment in young women. *Nat Commun.* 2017;8(1):1632.

92. Zhu Q, Schultz E, Long J, et al. UACA locus is associated with breast cancer chemoresistance and survival. *npj Breast Cancer.* 2022;8(1):39.

93. Escala-Garcia M, Guo Q, Dörk T, et al. Genome-wide association study of germline variants and breast cancer-specific mortality. *British Journal of Cancer.* 2019;120(6):647-657.

94. Hair BY, Hayes S, Tse CK, Bell MB, Olshan AF. Racial differences in physical activity among breast cancer survivors: implications for breast cancer care. *Cancer.* 2014;120(14):2174-2182.

95. 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.

96. 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):176-182.

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

98. 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.

99. Delaneau O, Marchini J, Zagury JF. A linear complexity phasing method for thousands of genomes. *Nat Methods.* 2011;9(2):179-181.

100. 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.

101. 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.

102. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature.* 2021;590(7845):290-299.

103. Loh P-R, Danecek P, Palamara PF, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics.* 2016;48(11):1443-1448.

104. Das S, Forer L, Schönherr S, et al. Next-generation genotype imputation service and methods. *Nature Genetics.* 2016;48(10):1284-1287.

105. Geiss GK, Bumgarner RE, Birditt B, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat Biotechnol.* 2008;26(3):317-325.

106. Bhattacharya A, García-Closas M, Olshan AF, Perou CM, Troester MA, Love MI. A framework for transcriptome-wide association studies in breast cancer in diverse study populations. *Genome Biol.* 2020;21(1):42.

107. Bhattacharya A, Hamilton AM, Furberg H, et al. An approach for normalization and quality control for NanoString RNA expression data. *Brief Bioinform.* 2020.

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

109. 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.

110. Sun X, Shan Y, Li Q, et al. Intra-individual Gene Expression Variability of Histologically Normal Breast Tissue. *Scientific Reports.* 2018;8(1):9137.

111. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science.* 2020;369(6509):1318-1330.

112. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet.* 2013;45(10):1113-1120.

113. Gao GF, Parker JS, Reynolds SM, et al. Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Syst.* 2019;9(1):24-34.e10.

114. Guo Y, Sheng Q, Li J, Ye F, Samuels DC, Shyr Y. Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PLoS One.* 2013;8(8):e71462.

115. Guo M, Yue W, Samuels DC, et al. Quality and concordance of genotyping array data of 12,064 samples from 5840 cancer patients. *Genomics.* 2019;111(4):950-957.

116. Deng M, Brägelmann J, Kryukov I, Saraiva-Agostinho N, Perner S. FirebrowseR: an R client to the Broad Institute's Firehose Pipeline. *Database (Oxford).* 2017;2017.

117. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw.* 2010;33(1):1-22.

118. Endelman JB. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome.* 2011;4(3).

119. Zhong J, Jermusyk A, Wu L, et al. A Transcriptome-Wide Association Study Identifies Novel Candidate Susceptibility Genes for Pancreatic Cancer. *Journal of the National Cancer Institute.* 2020;112(10):1003-1012.

120. 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.

121. Rowland B, Venkatesh S, Tardaguila M, et al. Transcriptome-wide association study in UK Biobank Europeans identifies associations with blood cell traits. *Human Molecular Genetics.* 2022:ddac011.

122. Tapia AL, Rowland BT, Rosen JD, et al. A large-scale transcriptome-wide association study (TWAS) of 10 blood cell phenotypes reveals complexities of TWAS fine-mapping. *Genet Epidemiol.* 2022;46(1):3-16.

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

124. Parker JS, Mullins M, Cheang MC, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol.* 2009;27(8):1160-1167.

125. van Maaren MC, de Munck L, Strobbe LJA, et al. Ten-year recurrence rates for breast cancer subtypes in the Netherlands: A large population-based study. *International Journal of Cancer.* 2019;144(2):263-272.

126. Troester MA, Sun X, Allott EH, et al. Racial Differences in PAM50 Subtypes in the Carolina Breast Cancer Study. *JNCI: Journal of the National Cancer Institute.* 2017;110(2):176-182.

127. Seshie B, Adu-Aryee NA, Dedey F, Calys-Tagoe B, Clegg-Lamptey JN. A retrospective analysis of breast cancer subtype based on ER/PR and HER2 status in Ghanaian patients at the Korle Bu Teaching Hospital, Ghana. *BMC Clin Pathol.* 2015;15:14.

128. Der EM, Gyasi RK, Tettey Y, et al. Triple-Negative Breast Cancer in Ghanaian Women: The Korle Bu Teaching Hospital Experience. *Breast J.* 2015;21(6):627-633.

129. Bhattacharya A, Hamilton AM, Furberg H, et al. An approach for normalization and quality control for NanoString RNA expression data. *Brief Bioinform.* 2021;22(3).

130. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol.* 2014;32(9):896-902.

131. O'Connell J, Gurdasani D, Delaneau O, et al. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLOS Genetics.* 2014;10(4):e1004234.

132. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLOS Genetics.* 2009;5(6):e1000529.

133. Bhattacharya A, Hirbo JB, Zhou D, et al. Best practices for multi-ancestry, meta-analytic transcriptome-wide association studies: lessons from the Global Biobank Meta-analysis Initiative. *medRxiv.* 2021:2021.2011.2024.21266825.

134. Kar SP, Considine DPC, Tyrer JP, et al. Pleiotropy-guided transcriptome imputation from normal and tumor tissues identifies candidate susceptibility genes for breast and ovarian cancer. *HGG Adv.* 2021;2(3).

135. Guo X, Lin W, Bao J, et al. A Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci Identified in Genome-wide Association Studies. *Am J Hum Genet.* 2018;102(5):890-903.

136. Patel A, García-Closas M, Olshan AF, et al. Gene-Level Germline Contributions to Clinical Risk of Recurrence Scores in Black and White Patients with Breast Cancer. *Cancer Res.* 2022;82(1):25-35.

137. 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.

138. Mi H, Muruganujan A, Huang X, et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). *Nature Protocols.* 2019;14(3):703-721.

139. 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 Medical Genomics.* 2020;13(1):116.

140. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol.* 2017;32(5):377-389.

141. Sun X, Wu B, Chiang H-C, et al. Tumour DDR1 promotes collagen fibre alignment to instigate immune exclusion. *Nature.* 2021;599(7886):673-678.

142. Yu TJ, Liu YY, Li XG, et al. PDSS1-Mediated Activation of CAMK2A-STAT3 Signaling Promotes Metastasis in Triple-Negative Breast Cancer. *Cancer Res.* 2021;81(21):5491-5505.

143. Liu PF, Wu YY, Hu Y, et al. Silencing of pancreatic adenocarcinoma upregulated factor by RNA interference inhibits the malignant phenotypes of human colorectal cancer cells. *Oncol Rep.* 2013;30(1):213-220.

144. Jing X, Liang H, Hao C, Yang X, Cui X. Overexpression of MUC1 predicts poor prognosis in patients with breast cancer. *Oncol Rep.* 2019;41(2):801-810.

145. Nath S, Mukherjee P. MUC1: a multifaceted oncoprotein with a key role in cancer progression. *Trends Mol Med.* 2014;20(6):332-342.

146. Kufe DW. MUC1-C oncoprotein as a target in breast cancer: activation of signaling pathways and therapeutic approaches. *Oncogene.* 2013;32(9):1073-1081.

147. Zafar GI, Grimm EA, Wei W, Johnson MM, Ellerhorst JA. Genetic deficiency of complement isoforms C4A or C4B predicts improved survival of metastatic renal cell carcinoma. *J Urol.* 2009;181(3):1028-1034.

148. Holers VM. Complement and Its Receptors: New Insights into Human Disease. *Annual Review of Immunology.* 2014;32(1):433-459.

149. Liu Z, Tang Q, Wen J, et al. Elevated serum complement factors 3 and 4 are strong inflammatory markers of the metabolic syndrome development: a longitudinal cohort study. *Scientific Reports.* 2016;6(1):18713.

150. Dandona P, Aljada A, Bandyopadhyay A. Inflammation: the link between insulin resistance, obesity and diabetes. *Trends in Immunology.* 2004;25(1):4-7.

151. Ménard V, Collin P, Margaillan G, Guillemette C. Modulation of the UGT2B7 enzyme activity by C-terminally truncated proteins derived from alternative splicing. *Drug Metab Dispos.* 2013;41(12):2197-2205.

152. Du R, Huang C, Liu K, Li X, Dong Z. Targeting AURKA in Cancer: molecular mechanisms and opportunities for Cancer therapy. *Molecular Cancer.* 2021;20(1):15.

153. Siegel RL, Miller KD, Fuchs HE, Jemal A. Cancer Statistics, 2021. *CA Cancer J Clin.* 2021;71(1):7-33.

154. Tao Z, Shi A, Lu C, Song T, Zhang Z, Zhao J. Breast Cancer: Epidemiology and Etiology. *Cell Biochem Biophys.* 2015;72(2):333-338.

155. Mitchell RE, Hartley A, Walker VM, et al. Strategies to investigate and mitigate collider bias in genetic and Mendelian randomization studies of disease progression. *medRxiv.* 2022:2022.2004.2022.22274166.

156. Munafò MR, Tilling K, Taylor AE, Evans DM, Davey Smith G. Collider scope: when selection bias can substantially influence observed associations. *International Journal of Epidemiology.* 2017;47(1):226-235.

157. Griffith GJ, Morris TT, Tudball MJ, et al. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nat Commun.* 2020;11(1):5749.

158. Mahmoud O, Dudbridge F, Davey Smith G, Munafo M, Tilling K. A robust method for collider bias correction in conditional genome-wide association studies. *Nature Communications.* 2022;13(1):619.

159. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet.* 2005;76(5):887-893.

160. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics.* 2012;28(10):1353-1358.

161. 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):559-575.

162. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet.* 2011;88(1):76-82.

163. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics.* 2015;31(5):782-784.

164. Cain KC, Harlow SD, Little RJ, et al. Bias due to left truncation and left censoring in longitudinal studies of developmental and disease processes. *Am J Epidemiol.* 2011;173(9):1078-1084.

165. Kong X, Liu Z, Cheng R, et al. Variation in Breast Cancer Subtype Incidence and Distribution by Race/Ethnicity in the United States From 2010 to 2015. *JAMA Network Open.* 2020;3(10):e2020303-e2020303.

166. Benefield HC, Zabor EC, Shan Y, Allott EH, Begg CB, Troester MA. Evidence for Etiologic Subtypes of Breast Cancer in the Carolina Breast Cancer Study. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology.* 2019;28(11):1784-1791.

167. Semmlinger A, von Schoenfeldt V, Wolf V, et al. EP3 (prostaglandin E2 receptor 3) expression is a prognostic factor for progression-free and overall survival in sporadic breast cancer. *BMC Cancer.* 2018;18(1):431-431.

168. Gao T, Han Y, Yu L, Ao S, Li Z, Ji J. CCNA2 is a prognostic biomarker for ER+ breast cancer and tamoxifen resistance. *PLoS One.* 2014;9(3):e91771.

169. Xing Z, Wang X, Liu J, Zhang M, Feng K, Wang X. Expression and prognostic value of CDK1, CCNA2, and CCNB1 gene clusters in human breast cancer. *J Int Med Res.* 2021;49(4):300060520980647.

170. Dudbridge F, Allen RJ, Sheehan NA, et al. Adjustment for index event bias in genome-wide association studies of subsequent events. *Nature Communications.* 2019;10(1):1561.

171. Yin L, Duan J-J, Bian X-W, Yu S-c. Triple-negative breast cancer molecular subtyping and treatment progress. *Breast Cancer Research.* 2020;22(1):61.

172. Safikhani Z, Smirnov P, Thu KL, et al. Gene isoforms as expression-based biomarkers predictive of drug response in vitro. *Nature Communications.* 2017;8(1):1126.

173. Grishin D, Gusev A. Allelic imbalance of chromatin accessibility in cancer identifies candidate causal risk variants and their mechanisms. *Nature Genetics.* 2022;54(6):837-849.

174. Soguel L, Durocher F, Tchernof A, Diorio C. Adiposity, breast density, and breast cancer risk: epidemiological and biological considerations. *Eur J Cancer Prev.* 2017;26(6):511-520.

175. Kothari C, Diorio C, Durocher F. The Importance of Breast Adipose Tissue in Breast Cancer. *Int J Mol Sci.* 2020;21(16).

176. Place AE, Jin Huh S, Polyak K. The microenvironment in breast cancer progression: biology and implications for treatment. *Breast Cancer Research.* 2011;13(6):227.

177. Bhattacharya A, Li Y, Love MI. MOSTWAS: Multi-Omic Strategies for Transcriptome-Wide Association Studies. *PLoS Genet.* 2021;17(3):e1009398.