# STATISTICAL METHODS FOR INFERRING GENETIC REGULATION ACROSS HETEROGENEOUS SAMPLES AND MULTIMODAL DATA

Arjun Bhattacharya

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

Chapel Hill
2020

Approved by:

Michael I. Love

Melissa A. Troester

Yun Li

Naim U. Rashid

Katherine A. Hoadley

## ABSTRACT

Arjun Bhattacharya: Statistical methods for inferring genetic regulation across heterogeneous samples and multimodal data
(Under the direction of Dr. Michael I. Love and Dr. Melissa A. Troester)

As clinical datasets have increased in size and a wider range of molecular profiles can be credibly measured, understanding sources of heterogeneity has become critical in studying complex phenotypes. Here, we investigate and develop statistical approaches to address and analyze technical variation, genetic diversity, and tissue heterogeneity in large biological datasets.

Commercially available methods for normalization of NanoString nCounter RNA expression data are suboptimal in fully addressing unwanted technical variation. First, we develop a more comprehensive quality control, normalization, and validation framework for nCounter data, benchmark it against existing normalization methods for nCounter, and show its advantages on four datasets of differing sample sizes. We then develop race-specific and genetic ancestry-adjusted tumor transcriptomic prediction models from germline genetics in the Carolina Breast Cancer Study (CBCS) and study the performance of these models across ancestral groups and molecular subtypes. These models are employed in a transcriptome-wide association study (TWAS) to identify four novel genetic loci associated with breast-cancer specific survival.

Next, we extend TWAS to a novel suite of tools, MOSTWAS, to prioritize distal genetic variation in transcriptomic predictive models with two multi-omic approaches that draw from mediation analysis. We empirically show the utility of these extensions in simulation analyses, TCGA breast cancer data, and ROS/MAP brain tissue data. We develop a novel distal-SNPs added-last test, to be used with MOSTWAS models, to prioritize distal loci that give added information, beyond the association in the local locus around a gene. Lastly, we develop DeCompress, a deconvolution method from gene expression from targeted RNA panels such as NanoString, which have a much smaller feature space than traditional RNA expression assays. We propose an ensemble approach that leverages compressed sensing to expand the feature space and validate it on data from the CBCS. We conduct extensive benchmarking of existing deconvolution methods using simulated *in-silico* experiments, pseudo-targeted panels from published mixing experiments, and data from the CBCS to show the advantage of DeCompress over reference-free methods. We lastly show the utility of *in-silico* cell-type proportion estimation in outcome prediction and eQTL mapping.

## ACKNOWLEDGEMENTS

I've never been accused of being concise, and I definitely won't start here. I'd like to thank:

My parents, first and foremost, for their love, support, and patience over the past twenty-six years. My mom taught me hard work and perseverance, and my dad taught me how to not take all of it so seriously. Without their presence in my life, none of this would have been possible;

My brother for being the stern and ever-present role model every younger brother needs but doesn't always appreciate in the moment;

My advisors, Drs. Mike Love and Melissa Troester, for giving me the opportunity and space to grow and learn but the guidance to keep me grounded - not only for their scientific mentorship but also teaching me the little things about academia;

My committee members, Drs. Yun Li, Katie Hoadley, and Naim Rashid, for their insight and wisdom during the research process;

Dr. Hudson Santos for being a mentor, colleague, and a friend;

My fellow graduate students who have made the ride that much more enjoyable: Alina Hamilton, Linnea Olsson, John Kidd, Sean McCabe, and Anqi Zhu;

My friends for always asking how I'm doing but never asking how many more years I have left: Renu Gharpure Kohlmann, CJ Norsigian, Karthik Ardhanareeswaran, BT Wilkins, Anagha Gogate, and Jacob Rohde;

Kanishka Patel for her patience and encouragement through iteration after iteration of papers and presentations;

And Spencer for just being such a good little boy.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| **BBFDR** | Benjamini-Bogomolov false discovery rate |
| **BLUP** | Best linear unbiased predictor |
| **BSLMM** | Bayesian sparse linear mixed model |
| **CBCS** | Carolina Breast Cancer Study |
| **CEU** | Utah residents with Northern and Western European ancestry |
| **CV** | Cross-validation |
| **DePMA** | Distal-eQTL prioritization via mediation analyis |
| **DSA** | Digital Sorting Algorithm |
| **eQTL** | Expression quantitative trait loci |
| **FDR** | False discovery rate |
| **FFPE** | Formalin-fixed, paraffin-embedded |
| **GCTA** | Genome-wide Complex Trait Analysis |
| **GDC** | Genome Data Commons |
| **GReX** | Genetically regulated expression |
| **GRM** | Genetic relationship matrix |
| **GTEx** | The Genotype-Tissue Expression Project |
| **GWAS** | Genome-wide association study |
| **LARS** | Least angles regression |
| **LD** | Linkage disequilibrium |
| **LOD** | Limit of detection |
| **MAF** | Minor allele frequency |
| **MeTWAS** | Mediator-enriched TWAS |
| **MLE** | Maximum likelihood estimate |
| **MOSTWAS** | Multi-omic strategies for TWAS |
| **MSE** | Mean square error |
| **NMF** | Non-negative matrix factorization |
| **NNLS** | Non-negative least squares |
| **QC** | Quality control |
| **RLE** | Relative log-expression |
| **ROS/MAP** | The Religious Orders Study and the Rush Memory and Aging Project |

| | |
|---|---|
| **RUV** | Remove Unwanted Variation |
| **SNP** | Single nucleotide polymorphism |
| **TCGA** | The Cancer Genome Atlas |
| **TCGA-BRCA** | TCGA breast cancer |
| **TCGA-LUAD** | TCGA lung adenocarcinoma |
| **TCGA-PRAD** | TCGA prostate adenocarcinoma |
| **TGL** | Translational Genomics Laboratory |
| **TME** | Total mediation effect |
| **TWAS** | Transcriptome-wide association study |
| **UTMOST** | Unified Test of Molecular Signatures |
| **YRI** | Yoruban ancestry |

# CHAPTER 1: LITERATURE REVIEW

## 1.1 Normalization of NanoString nCounter expression

The NanoString nCounter platform offers a comparatively inexpensive alternative for gene expression measurement of a panel of pre-specified genes due to its ability to measure mRNA expression without requiring cDNA synthesis or any amplification steps[18]. The technology offers key advantages in sensitivity, technical reproducibility, and strong robustness for analysis of formalin-fixed, paraffin-embedded (FFPE) samples[19]. Given these advantages, the NanoString nCounter platform is increasingly being used in academic settings globally to study differential gene expression, despite the admitted limitation of requiring pre-specification of genes to measure[20–23]. nCounter is especially attractive for longitudinal studies involving FFPE samples carried out over several years[24] and diagnostic assays in clinical settings, as shown by the Clinical Laboratory Improvement Amendments-approved PAM50-based breast cancer signature assay developed by Prosigna[25,26]. The following section gives a quick discussion of the importance of proper quality control and normalization for mRNA expression panels and overviews existing methods specific to nCounter.

### 1.1.1 Importance of proper normalization

Proper normalization and quality control (QC) of mRNA expression is necessary prior to statistical analysis to mitigate any confounding noise from unwanted biological and technical variables that are associated with potentially important covariates of interest, such as batch effects or degradation of groups of samples that have been stored over time[27,28]. Often times, all sources of unwanted noise cannot be enumerated *a priori* or measured, beyond those that are easily catalogued in a sample table, such as different research centers, technicians, or storage units for samples. In all cases, it is advised to use a proper quality control and normalization pipeline to address any degraded samples and estimate any such technical noise. All normalization methods deal with a trade-off between any bias that needs to be corrected and the variance that may be introduced to the data due to estimation of bias effects[29]. Naïve normalization methods may err too heavily on the side of bias correction and

result in adding excessive variance to the expression measurements. Molania et al have recently suggested an iterative process to normalization, wherein several parameters (i.e. number of housekeeping genes, number of detected outliers, number of dimensions of technical noise) are tuned over several iterations with several relevant biological checks used as validation, in nCounter datasets with technical replicates[28].

### 1.1.2 Normalization using the nSolver software

NanoString provides nSolver 4.0, a graphical user interface software, that aids in QC and normalization. After imaging, binding density, positive control, and limit of detection quality controls, NanoString provides two forms of normalization in its nSolver Analysis Software[30]: (1) a more user-friendly procedure with optional background correction, followed by positive control and housekeeping gene normalization and (2) the Advanced Analysis tool, a wizard-based add-on that draws from the NormqPCR R package[31,32].

Briefly, the nSolver normalization procedure is as follows: the arithmetic mean of the geometric means of the positive controls for each lane is computed and then divided by the geometric mean of each lane to generate a lane-specific positive control normalization factor[30,31]. The counts for every gene are then multiplied by its lane-specific normalization factors. To account for any noise introduced into the nCounter assay by positive normalization, the housekeeping genes are used similarly as the positive control genes to compute housekeeping normalization factors used to scale the expression values[30,31].

### 1.1.3 NanoStringDiff

Wang et al generated a normalization method for NanoString using negative binomial linear modelling with an empirical Bayes approach[33]. This method introduces three normalization parameters to quantify variation and noise across different experimental conditions: (1) the positive control size factor ($c_i$), accounting for lane-by-lane variation; (2) the background noise parameter ($\theta_i$), quantifying the non-specific background level; and (3) the housekeeping size factor ($d_i$), adjusting for the variation in the amount of input sample material.

Here, we denote the observed count from gene $g$ in sample $i$ with $Y_{gi}$, and the unobserved expression rate by $\lambda_{gi}$. The data is assumed to be generated from the following hierarchical model:

$$
\begin{aligned}
Y_{gi}|\lambda_{gi} &\sim \text{Poisson}(c_i d_i \lambda_{gi} + \theta_i) \\
\lambda_{gi}|u_{gi}, \eta_g &\sim \text{Gamma}(u_{gi}, \eta_g) \\
\eta_g &\sim \text{Normal}(m_0, \tau^2) \\
\log u_{gi} &= X_i \beta_g^T,
\end{aligned}
$$

where $u_{gi}$ and $\eta_g$ denote the mean and log-dispersion of the expression rate $\lambda_{gi}$ to deal with overdispersion. The mean parameter $u_{gi}$ is specified based on a generalized linear model with logarithmic link function, where $X_i$ gives the $i$th row of the design matrix of covariates $X$ and $\beta_g$ is a vector of regression coefficients.

Hyper-parameters are empirically estimated from the expression data for endogenous genes. For each endogenous gene, the maximum likelihood estimate (MLE) of the log-dispersion parameter $\hat{\eta}_g$ is calculated. These estimates are only used from endogenous genes with read counts larger than the maximum value of negative controls to estimate further hyperparameters due to background noise in endogenous genes with low read counts. The median of $\hat{\eta}_g$ for endogenous counts is used to find $\hat{m}_o = \text{median}_g\,(\hat{\eta}_g)$. As Wu et al points out[34], the sample variance of $\hat{\eta}_g$ overestimates $\tau$ and accordingly, Wang et al apply an *ad hoc* method to compute pseudo datasets with $\tau^2 = 0$ to estimate $\text{var}(\hat{\eta}_g|\eta_g)$ and subtract it from the sample variance of $\hat{\eta}_g$ to obtain an estimate of $\tau^2$. Model parameters $\beta_g$ and $\eta_g$ are then estimated through an iterative process that maximizes the conditional likelihoods of $\beta_g|\eta_g$ and $\eta_g|\beta_g$ until convergence.

#### 1.1.4 Remove Unwanted Variation III (RUV-III)

Molania et al proposed a method in the line of Remove Unwanted Variation (RUV) methods[27,29] that is catered to a NanoString nCounter panel with technical replicates. RUV-III estimates a user-defined $k$ dimensions of unwanted variation from differences between expression values of technical replicates and and the distribution of expression of negative control transcripts.

Here, we assume that we have data from $m$ nCounter assays on $m' < m$ distinct samples. Let $M$ be the alliteratively-named $m \times m'$ mapping matrix that maps assays to samples, such that the $i, j$-th element of $M$ $m_{i,j} = 1$ if the $i$-th assay is an assay of sample $j$. We assume that all assays have $n$ probes. Let $Y$ be the $m \times n$ matrix of observed log-transformed expression values and we model

$$Y_{m\times n} = X_{m\times p}\beta_{p\times n} + W_{m\times k}\alpha_{k\times n} + \epsilon_{m\times n} \tag{1.1}$$

where $X\beta$ is the biological variation, $W\alpha$ is the unwanted variation, $\epsilon$ is random error. We assume $p + k < m$ and $k < m - m'$. Here, $X$ corresponds to biological factors of interest and not technical factors, like batch. In the most general case of RUV-III, both $X$ and $W$ are unobserved. Lastly, assume that $n_c < n$ of the probes are negative controls. We herein indicate sub-matrices of the matrices identified in Equation 1.1 with a subscript of $c$. These negative control probes are assumed to unaffected by factors in $X$.

Note that if two assays are technical replicates of sample $j$, then the rows of $X$ corresponding to this assay are identical. Thus, we have $X = M\mathbf{X}$, where $\mathbf{X}_{m'\times p}$ as the biological factors of interest in terms of samples. The goal is to estimate $W\alpha$ in Equation 1.1 and regress it out of $Y$, leaving $\hat{Y}$ that is used in downstream analysis. Let

$$R_M = I - M(M'M)M^{-1}M,$$

be the residual operator of $M$. Thus,

$$\begin{aligned} R_M Y &= R_M(X\beta + W\alpha + \epsilon) \\ &= R_M M\mathbf{X}\beta + R_M W\alpha + R_M\epsilon \\ &= R_M W\alpha + R_M\epsilon. \end{aligned}$$

$\alpha$ may be estimated with a form of factor analysis on $R_M Y$, as long as $R_M W$ is full rank. Let $\hat{\alpha}$ be the first $k$ singular vectors of $R_M Y$, and accordingly $\hat{W} = Y_c\hat{\alpha}_c'(\hat{\alpha}_c\hat{\alpha}_c')^{-1}$. It is easy to show that $\hat{W} \approx W$.

### 1.1.5 Summary

It is becoming increasingly popular in both clinical and academic settings to use mRNA expression measurements from the NanoString nCounter platform. Even though groups have addressed normalization in this setting previously, the most popular method for normalization is the NanoString-provided nSolver software. Proper evaluation of this normalization method has not been conducted before, especially in large cohorts without technical replicates.

## 1.2 Transcriptome-wide association studies (TWAS)

### 1.2.1 Applications of TWAS in breast cancer

Few genome-wide association studies (GWAS) have studied the relationship between germline variation and survival outcomes in breast cancer, with most focusing instead on genetic predictors of risk[11,35]. Recently, GWAS have shown evidence of association between candidate common germline variants and breast cancer survival, but these studies are often underpowered[36,37]. Furthermore, the most significant germline variants identified by GWAS, in either risk or survival, are often located in non-coding regions of the genome, requiring *in vitro* follow-up experiments and co-localization analyses to interpret functionally[38]. It is important to seek strategies for overcoming these challenges in GWAS, especially because several studies in complex traits and breast cancer risk have shown that regulatory variants not significant in GWAS account for a large proportion of trait heritability[39–41].

Novel methodological approaches that integrate multiple data types offer advantages in interpretability and statistical efficiency. Escala-García et al has suggested that aggregating variants by integrating gene expression or other omics may better explain underlying biological mechanisms while increasing the power of association studies beyond GWAS[36]. To alleviate problems with statistical power and interpretability, a recent trend in large-scale association studies is the transcriptome-wide association study (TWAS). TWAS aggregates genomic information into functionally-relevant units that map to genes and their expression. This gene-based approach combines the effects of many regulatory variants into a single testing unit that increases study power and provides more interpretable trait-associated genomic loci[42,3,43]. Here, we describe a few current approaches to transcriptomic imputation and subsequent downstream tests of associations.

### 1.2.2 PrediXcan

Gamazon et al's PrediXcan[42] identifies trait-associated genes by estimating the genetic control of of phenotype through the mechanism of genetic control. Gene expression levels are decomposed into (1) the genetically regulated expression (GReX) components, (2) a component altered by the trait itself, and (3) a remaining component attributed to environmental or other factors. PrediXcan tests the mediated effect of gene expression by quantifying the association between GReX and the phenotype of interest.

Reference transcriptome datasets from GTEx[44], GEUVADIS[45], and DGN[46] were used to train additive models of gene expression from genetic variation as follows:

$$Y_g = \sum_k w_{k,g} X_k + \epsilon, \tag{1.2}$$

where $Y_g$ is the expression trait of gene $g$, $w_{k,g}$ is the effect size of marker $k$ for gene $g$, $X_k$ is the number of reference alleles of marker $k$, and $\epsilon$ is the contribution of other factors that determine the expression trait assumed to independent of the genetic component. Gamazon et al have built PredictDB, a database of predictive models using DGN using LASSO[47], elastic net[48], and/or the polygenic score at various $P$-value thresholds.

The genetic heritability of gene expression serves as an upper bound for the prediction of the GREx of a given gene. Here, the cis-heritability (cis-$h^2$) was estimated for each gene using a variance component model with a genetic relationship matrix (GRM) estimated from genotype data within 1 Megabase (Mb) of the gene with minor allele frequency (MAF) $> 0.05$ and in Hardy-Weinberg Equilibrium ($P > 0.05$). Gamazon et al calculated the proportion of the variance of gene expression explained by local single nucleotide polymorphisms (SNPs) using linear mixed modelling in GCTA[49]:

$$Y = Xb + G_{local} + e, \text{ var}(Y) = A_{local}\sigma^2_{local} + I\sigma^2_e,$$

where $Y$ is a gene expression trait, $b$ is a vector of fixed effects, $A_{local}$ is the GRM from local SNPs, and the random effect $G_{local}$ is the genetic effect attributable to the set of local SNPs with var$(G_{local}) = A_{local}\sigma^2_{local}$..

Given an optimal vector $\hat{w}_{k,g}$ fitted with Elastic Net[48] with $\alpha = 0.5$ that best predicts $Y_g$ as assessed by 10-fold cross-validation $R^2$, the PrediXcan framework imputes the GREx of each gene in external GWAS panels using the same genotypes $X_k$:

$$\hat{GReX}_g = \sum_k \hat{w}_{k,g} X_k.$$

These imputed $\hat{GReX}_g$ values are then employed in downstream tests of association to identify gene-trait associations.

PrediXcan has several advantages in identifying gene-trait associations. It has a much smaller multiple-testing burden than GWAS, with approximately 10,000 gene-based tests as opposed to 5-10 million single variant tests in GWAS. No transcriptome data is needed since the predicted expression levels are a function of genetic variation alone and thus can be applied to any existing GWAS panel. Reverse causality is not a concern since disease status or drug treatment cannot alter germline

genetic variation. PrediXcan also lends itself seamlessly to meta-analysis as less stringent harmonization between studies is required.

### 1.2.3 FUSION

Gusev et al proposed a similar transcriptome-wide association study approach, called FUSION, nearly concurrently with Gamazon et al's PrediXcan method[3]. Again, FUSION uses a reference panel in relevant tissue to train predictive models of mRNA expression from cis-genotypes. Then, using these optimally trained models, FUSION either (1) directly predicts expression in genotype samples using effect-sizes from the predictive models and measures association between predicted expression and trait or (2) indirectly estimates association between predicted expression and trait as a weighted linear combination (weighted burden test) of SNP-trait standardized effect sizes while accounting for linkage disequilibrium among SNPs, as first proposed in Pasaniuc et al[50]. Similar to PrediXcan, by focusing on the genetic-component of expression, FUSION avoids instances of expression-trait associations that are not a consequence of genetic variation but are driven by variation in trait. **Figure 1.1**, adapted from Gusev et al[3], summarizes the possible models of causality for the relationship between genetic markers, gene expression, and trait.

Model fitting in FUSION is very similar to that in PrediXcan. For genes that are cis-heritable at $P < 0.05$, the same additive model for gene expression as in Equation 1.2 is fit using one of the following schemes:

- the cis-eQTL, the single most significantly associated cis-eSNP (SNP in an eQTL) in the training set was used as the only predictor;
- LASSO or elastic net[47,48] with mixing parameter $\alpha \in \{0, 0.5\}$ and $\lambda$ tuned over 5 folds;
- the best linear predictor (BLUP)[51] which estimates the causal effect-sizes of all SNPs in the cis-locus jointed using a single-variance component;
- the Bayesian linear mixed model (BSLMM) which estimates the underlying effect-size distribution and then fits all SNPs in the locus jointly.

The BLUP and BSLMM are fit using all post-QC SNPs using GEMMA[52] and perform shrinkage of the SNP weights, but not variable selection. Predictive accuracy was measured by five-fold cross-validation in a random sampling of 1,000 of the highly heritable genes using the predictive $R^2$ between predicted and true expression across all predicted folds.

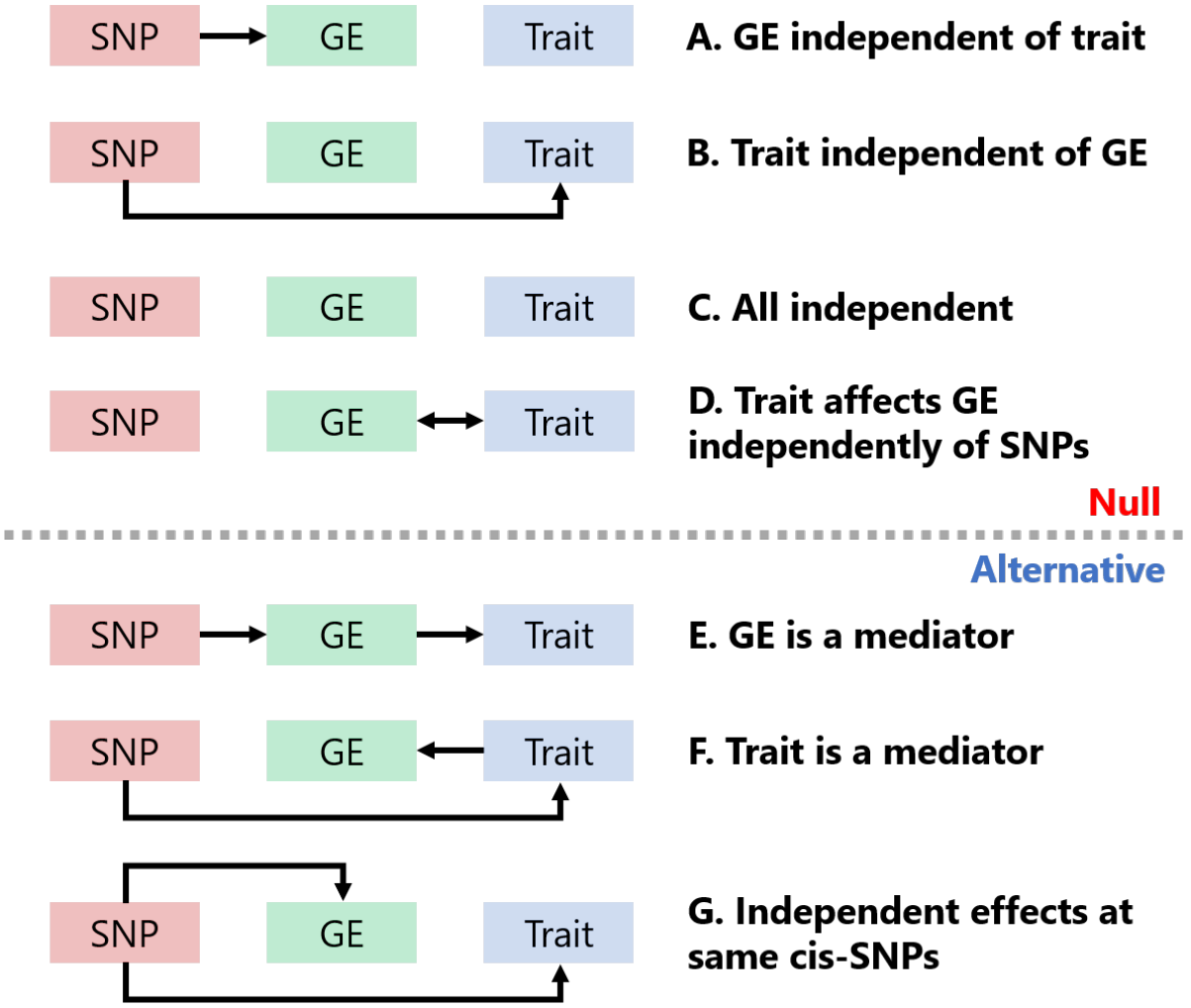


**Figure 1.1:** *Modes of expression causality using FUSION.* Diagrams here shown the possible modes of causality for the relationship between genetics markers (labelled SNP in blue), gene expression (GE, green), and trait (red). Models A-D describes scenarios that are considered null models by the TWAS framework. E-G shows scenarios that can be identified as significant and can be further studied functionally. Modified from Gusev et al[3].

FUSION's novelty comes in its summary-based imputation that extends the ImpG-Summary algorithm[50] to train on the cis-genetic component of expression. Let $Z$ be a vector of standardized effect sizes ($z$-scores) of SNP on trait at a given cis-locus. Here, the Wald-type test statistics (i.e. $\frac{\beta}{SE(\beta)}$) are considered. The $z$-score of the expression and trait is imputed as the linear combination of elements of $Z$ with weights $W$. With $\Sigma_{e,s}$ as the covariance matrix between all SNPs at the locus and gene expression and $\Sigma_{s,s}$ is the linkage disequilibrium (LD) among all SNPs,

$$W = \Sigma_{e,s}\Sigma_{s,s}^{-1}.$$

Under the null of no association and a multivariate Gaussian assumption $Z \sim N(0, \Sigma_{s,s})$, it can be shown that the imputed $z$-score of expression and trait ($WZ$) has variance $W\Sigma_{s,s}W'$. Thus, the imputation $z$-score of cis-genetic effect on trait is computed as

$$\frac{WZ}{(W\Sigma_{s,s}W')^{\frac{1}{2}}}.$$

FUSION optionally further subjects significant TWAS-identified loci to a highly conservation permutation test that tests the loci conditional on high GWAS effects. The eQTL weights are reshuffled 1,000 times to construct a null distribution for the TWAS $z$-score. This permutation test assesses if the same distribution of eQTL effect sizes could yield a significant association by chance. The test is implemented adaptively, so permutation will stop after a sufficient number of significant observations (or at the maximum specified). This statistic is highly conservative as truly causal genes can fail the test if their eQTLs are in high LD with many other SNPs, and intended to prioritize associations that are already significant in the standard test for follow-up[3].

### 1.2.4 Alternative methods for TWAS

Here, we briefly survey three further methods relevant to TWAS: (1) UTMOST[53], (2) TIGAR[54], and (3) FOCUS[55]:

The UTMOST (Unifed Test for MOlecular SignaTures) method performs cross-tissue expression imputation and gene-level association analyses to unify complex traits that are modulated across various tissues in the human body. Cross-tissue expression imputation is formulated as a penalized multivariate regression problem:

$$Y_{N\times P} = X_{N\times M}B_{M\times P} + \epsilon_{N\times P},$$

where $N$ is the sample size of the training data, $M$ is the number of SNPs in the imputation models, and $P$ is the number of tissues. Under the assumption that only a subset of tissues was collected from each individual, expression data in $Y$ is incomplete and the sample sizes for different tissues are unbalanced. UTMOST estimates $B$ by minimizing the squared loss function with a LASSO penalty on the within-tissue effects (columns) and a grouped-LASSO penalty on the cross-tissue effects (rows):

$$\hat{B} = \text{argmin}_B \sum_{i=1}^{P} \frac{\|Y_i - X_i B_{.i}\|_2^2}{2N_i} + \lambda_1 \sum_{i=1}^{P} \frac{\|B_{.i}\|_1}{N_i} + \lambda_2 \sum_{i=1}^{M} \|B_{j.}\|_2,$$

where $Y_i, X_i, N_i$ are the observed expressions, genotypes, and sample sizes of the $i$th tissue, respectively.

Per-tissue tests of association are computed similarly to PrediXcan and FUSION. Imputed gene expression in the $i$th tissue is computed as $E_i = X_i \hat{B}_{.i}$ and is tested for associated with the trait using a univariate regression model. The $z$-scores for gene-traited associated in the $i$th tissue is denoted as

$$Z_i = \frac{\hat{\gamma}_i}{SE(\hat{\gamma}_i)} \approx \hat{B}'_{.i} \Gamma_i \tilde{Z},$$

where $\tilde{Z}$ denotes the SNP-trait $z$-scores and $\Gamma_i$ is a diagonal matrix wit the ratio between the standard deviations of the $j$th SNP and the imputed expression in the $i$th tissue. Under the null of no SNP-trait association, $\tilde{Z} \sim N(0, D)$, where $D$ is the LD matrix for the SNPs and accordingly, $\text{Cov}(Z) = \Lambda' D \Lambda$, where $\Lambda = (\hat{B}_{.1}\Gamma_1, \ldots, \hat{B}_{.P}\Gamma_P)$. Lastly, the per-tissue gene-trait association results are combined using a generalized Berk-Jones test, taking into account the covariance among single-tissue test statistics[56].

TIGAR adds to the transcriptomic imputation methods employed by PrediXcan and FUSION by introducing a non-parametric Bayesian model using a latent Dirichlet process regression (DPR) model[57]. The cis-eQTL effect sizes $w$ are given a Normal prior $N(0, \sigma_w^2)$ and Dirichlet process (DP) prior[58] for the effect-size variance $\sigma_w^2$ such that

$$w_i \sim N(0, \sigma_w^2),\ \sigma_w^2 \sim D,\ D \sim DP(IG(a, b), \xi).$$

$\sigma_w^2$ is a latent variable and integrating it out induces a non-parametric prior distribution on $w_i$ that is equivalent to a DP mixture model of an infinite number of Normal distributions[57,59]:

$$w_i \sim \sum_{k=0}^{\infty} \pi_k N(0, \sigma_k^2),\ \sigma_k^2 \sim IG(a_k, b_k),\ \pi_k = \nu_k \prod_{l=0}^{k-1} (1 - \nu_l),\ \nu_k \sim Beta(1, \xi).$$

Conjugate hyperpriors $\xi \sim$ Gamma$(a_\xi, b_\xi)$ and $\sigma_\epsilon^2 \sim IG(a_\epsilon, b_\epsilon)$ are assumed and are generally set as non-informative. The posterior estimates for $w$ are then obtained by either Markov Chain Monte Carlo or variational Bayesian algorithms[59,60].

FOCUS, a fine-mapping method proposed by Mancuso et al, extends the TWAS testing framework outlined in Pasaniuc et al and Gusev et al[50,3] that models correlation among TWAS signals to assign a probability for every gene in the risk region to explain the observed association signal. Here, a quantitative trait $y$ is modelled by a linear combination of expression levels for $m$ genes $G \in \mathbb{R}^{n \times m}$ as

$$Y = X\beta + G\alpha + \epsilon,$$

where $X \in \mathbb{R}^{n \times p}$ is the genome-wide genotype matrix at $p$ SNPs, $\beta$ is the $p$ pleiotropic effects of $X$ on $y$, $\alpha$ is the vector of causal effects for the $m$ genes, and $\epsilon$ is the random environmental noise. As in TWAS predictive models, $G = XW$, where $W$ is the eQTL effect-size matrix. The marginal TWAS tests on $y$ using predicted expression $\hat{G} = X\Omega$ are modeled with

$$Z_{TWAS} = \frac{1}{\sigma_e \sqrt{n}} \hat{G}^T y,$$

where $\Omega$ is an estimate of $W$ from an independent reference panel and $\sigma_e$ is the diagonal variance parameter for $\epsilon$.

Marginalizing out unknown causal gene effects $\alpha$, the sampling distribution for the marginal TWAS test statistics is

$$Z_{TWAS} | \lambda_{snp}, \Omega, V, c, n\sigma_c^2 \sim N(\Omega^T V \lambda_{snp}, \mathcal{V} D_c \mathcal{V} + \mathcal{V}),$$

where $V = n^{-1} X^T X$ is the LD matrix, $\lambda_{snp}$ is the pleiotropic SNP non-centrality parameter, $\mathcal{V} = \Omega^T V \Omega$ is the predicted expression covariance, and $D_c$ is the prior variance for effects at causal genes ($n\sigma_c^2$) as indicated by a binary status vector $c$. Inference for which genes are causal given the TWAS statistics is performed by computing the posterior distribution of any set of causal genes $c$, assuming a Bernoulli prior (with default probability $10^{-3}$) for the causal status of a given gene. The posterior inference probability offers a flexible mechanism to generate a credible gene sets, as in previous fine-mapping approaches[61,62].

### 1.2.5 Summary

Transcriptome-wide association studies are quickly being used to increase power in detection of SNP-trait associations over traditional genome-wide association studies. Questions of predictive performance of predictive gene expression models across ancestrally-different populations and in understudied tissues is still open. Furthermore, it is also important to assess how TWAS performs to granularity introduced by sample-specific, biological, and disease subtype heterogeneity.

## 1.3 Mediation analysis in gene regulation

### 1.3.1 Implications of the omnigenic model

The omnigenic model of the genetics of complex traits advanced that human gene regulatory networks are so interconnected that thousands of individual genes contribute at least slightly to the phenotype through expression in relevant cells[41]. This model extended upon the infinitesimal model, first proposed in 1918, that quantitative phenotypes are the sum of a genetic and non-genetic component, such that the genetic component is distributed within families as a Normal random variable with variance independent of parental traits[63]. The omnigenic model also includes the concept of universal pleiotropy, wherein genetic variation in one region of the genome potentially has an indirect effects on many traits[64]. If the omnigenic model holds true, then many complex traits are driven by large numbers of genetic variants with small effects on a phenotype of interest, and thus implicating most regulatory variants that are active in disease-relevant tissues[41]. Boyle et al hypothesized further that disease risk is largely driven by gene with no direct relevance to disease and is propagated through multi-level regulatory networks with a small number of core genes with direct effects and a much larger set of peripheral genes with indirect effects[41].

These ideas of core and peripheral gene effects on phenotype are similar to ideas of genetic regulation of genes. Identification of expression trait quantitative loci (eQTLs) is one of the most important methods of discovering potential genetic regulators of the mRNA transcription of a gene. An eQTL is a genomic locus that explains a portion of variance in the expression level of the mRNA transcript of a given gene. eQTLs can be classified by their relative distance to the gene of interest (local or distal based on a defined window around a gene) or the mechanism of action on the transcription of a gene (cis- and trans-eQTLs act directly and indirectly, respectively)[65].

Liu et al models the contribution of core and peripheral genes to complex trait heritability[66]:

$$Y_i = \bar{Y} + \sum_{j=1}^{M} \gamma_j (x_{i,j} - \bar{x}_j) + \sum_{j-M+1}^{N} 0 \times (x_{i,j} - \bar{x_j}) + \epsilon_{Y_i},$$

where $Y_i$ is the phenotype value of individual $i$, $\bar{Y}$ is the population mean of the phenotype, $\gamma_j$ represents the direct effect of a unit change in expression of core gene $j$ on $E(Y_i)$, and $x_{i,j}$ is the expression of gene $j$ in individual $i$ (with population mean $\bar{x}_j$). We assume that there are $M$ core genes out of $N$ total expressed genes in a tissue, and the random error $\epsilon_{Y_i}$ has mean 0 and is independent of genotype and gen expression. Although the $N - M$ peripheral genes have no direct effects on the phenotype, they may modify the expression of core genes as trans-eQTLs. Phenotypic variation can thus be broken down as:

$$\begin{aligned} \mathsf{Var}(Y_i) &= \sum_{j=1}^{M} \gamma_j^2 \mathsf{Var}(x_{i,j}) + \sum_{j=1}^{M} \sum_{k=1}^{j-1} 2\gamma_j \gamma_k \mathsf{Cov}(x_{i,j}, x_{i,k}) + \mathsf{Var}(\epsilon_{Y_i}) \\ &= \sum_{j=1}^{M} \gamma_j^2 V_{i,\mathsf{cis}} \sum_{j=1}^{M} \gamma_j^2 V_{i,\mathsf{trans}} + \sum_{j=1}^{M} \sum_{k=1}^{j-1} 2\gamma_j \gamma_k C_{j,k} \end{aligned}$$

Here, $V_{j,\mathsf{cis}}$ and $V_{j,\mathsf{trans}}$ are the genetic variances of of core gene $j$ determined by cis and trans effects, respectively. $C_{j,k}$ represents the genetic covariance of expression of genes $j$ and $k$. The first pair of terms on the right-hand side of this variance decomposition depend on the relative importance of cis and trans effects in determining expression heritability of core genes. Liu et al estimates that, in general, about 70% of expression heritability is caused by these trans effects. The last term depends on covariances between core genes. As core genes are seldom adjacent in the genome, genetic covariances arise from trans effects. As there are more core gene pairs ($M^2$) than singleton core genes ($M$), these trans effects dominate the heritability for most traits[66].

The effects of a single SNP may potentially fan through multiple core genes to affect the phenotype. Suppose SNP $s$ is an eQTL for a core gene $j$. We let $\alpha_{s,j}$ be the effect size of SNP $s$ on the expression of gene $j$ and the change in phenotype $Y$ due to one additional copy of the alternative allele as $\delta_s$. In the case that $s$ is a trans-eQTLs for multiple core genes, the total phenotypic effect of $s$ is a sum of trans-effects mediated through each core gene $j$, and $\delta_s = \sum_{j=1}^{M} \alpha_{s,j}\gamma_j = M\alpha_{s,\bar{j}},\gamma_j$. If we assume that the effects of SNP $s$ has expectation 0 and are uncorrelated across $j$, then the effects cancel out on average, the variance scales multiplicatively with $M$, and the total effect is not large. Alternatively, if there exist peripheral master regulators that drive coordinated effects on many downstream target core genes, these trans-effects can be considerable[66].

### 1.3.2 Inference on trans-eQTLs

Many groups have cast the omnigenic model directly onto the eQTL framework, treating the problem of identifying trans-eQTLs for genes as a mediation problem[67–69]. Distal or trans-eQTLs are far more difficult to detect that local or cis-eQTLs due to the significant multiple testing burden of comparing millions of SNPs to thousands of transcripts. Trans-eQTLs are especially important in identifying to understand tissue-specific gene regulatory mechanism[70]. Here, we review a few methods for trans-eQTL prioritization based extensions on mediation analysis.

Brynedal et al demonstrated an approach to look for SNPs associated with the expression of many genes simultaneously, finding that hundreds of trans-eQTLs each affect hundreds of transcripts[68]. At each marker, they tested for overdispersion of association $-\log_{10}(P)$-values across all probe sets with a null hypothesis that $-\log_{10}(P)$ values are exponentially distributed with $\lambda = 1$ against the joint alternative hypothesis that a subset of association statistics are non-null (i.e. $\lambda \neq 1$). Evidence for these hypotheses were compared as a likelihood ratio test for the cross-phenotype meta analysis, where the test statistic is defined as

$$S_{CPMA} = -2 \times \log\left(\frac{P(Data|\lambda = 1)}{P(Data|\lambda = \hat{\lambda})}\right) \sim \chi^2_1,$$

where $\hat{\lambda}$ is the observed exponential decay in the data. Correlation between the probe set levels across individuals was accounted for using empirical significant testing by simulating eQTL association studies under the null expection of no association to any marker given the observed correlation between probe sets[68]. In summary, Brynedal et al discovered that target transcripts of a high-confidence set of trans-eQTLs encode proteins that interact more frequently than expected by chance and are bound by the sample transcription factors[68].

A more recent paper by Shan et al establishes a simple mediation framework to identify trans-eQTLs that are mediated by multiple mediating cis-eGenes. First, a candidate trio composed of a SNP, one or more cis-genes to the SNP, and the trans-gene of interest were selected. Trans SNP-gene pairs were included if the association has $P \leq 10^{-6}$[69]. Mediating cis-genes were selected if they were associated with the SNP at false discovery rate- (FDR) adjusted $P \leq 0.05$. The following set of linear models are chosen to assess the mediation effect. For the $i$th subject, let $Y_i$ be the expression level of the trans-gene, $X_i$ by the SNP dosage, $\mathbf{M}_i = (M_{i1}, \ldots, M_{ip})^T$ be the expression levels of the $p$ cis-genes, and $\mathbf{C}_i$ represents the $q$ covariates. Consider:

$$Y_i = \beta_0 + X_i\beta_X + \mathbf{M}_i^T\beta_M + \mathbf{C}_i^T\beta_C + \epsilon_{Y_i}$$
$$M_{ij} = \alpha_{0j} + X_i\alpha_{X_j} + \mathbf{C}_i^T\alpha_{C_j} + \epsilon_{M_{ij}},$$

where $\beta_M = \left(\beta_{M_1}, \ldots, \beta_{M_p}\right)^T$ is the effect of the $p$ cis-genes on the trans-gene, adjusting for the SNP and covariates, $\alpha_X = (\alpha_{X_1}, \ldots, \alpha_{X_p})^T$ is the effect of the SNP on the $p$ cis-genes, adjusted for covariates. $\epsilon_{Y_i}$ and $\epsilon_{M_{ij}}$ are the measurement errors, independent and distributed normally such that dependence is allowed among the $p$ cis-genes. Two quantities are estimated and tested for equality to 0 via bootstrapping[71]: the total mediation effect (TME) $\Delta = \alpha_X^T\beta_M$ and the component-wise effects $\delta = (\delta_1, \ldots, \delta_p)^T$, where $\delta_j = \alpha_{X_j}\beta_{M_j}$. The test of TME is a broader class of null than the test of CME. In the case of positive mediation effect through one cis-gene and negative mediation through another, the test of CME can be more powerful than the TME test[72].

Lastly, we consider a pair of cross-condition mediation methods from Yang et al: CCmed$_{\text{gene}}$ and CCmed$_{\text{GWAS}}$. CCmed takes in summary statistics from multiple studies, tissue types, or conditions and aims to detect robust mediation and trans-association effects shared across conditions. To validate the trait-associations of the identified trans-genes for GWAS SNPs, a two sample Mendelian randomization method robust to correlated and some invalid instruments (MR-Robin) was developed. CCmed$_{\text{gene}}$ detects candidate trios of eQTL set, cis-gene, and trans-gene that show evidence of cross-tissue trans-association and mediation effects by quantifying the joint probability of the following two conditions being satisfied in at least $K_1$ out of $K$ tissue types: (1) gene-level cis-associations and (2) non-zero correlations between the expression levels of the cis- and trans-genes conditioning on the eQTL genotypes. For each trio $(\mathbf{L}_i, C_i, T_j)$, where $\mathbf{L}_i$ is a set of eQTL genotypes for a cis-gene $i$, $C_i$ is the cis-gene expression, and $T_j$ is the expression level of a trans-gene $j$, $P_{\text{med},ij}$, the probability that $C_i$ mediates the effects of $\mathbf{L}_i$ on $T_j$ in at least $K_1$ tissue types, is computed as follows:

$$\begin{aligned} P_{\text{med},ij} &= P(\mathbf{L}_i \to C_i \to T_j \text{ in at least } K_1 \text{ out of } K \text{ tissues}) \\ &= P(\alpha_C \neq 0 \text{ in all } K \text{ tissues}) \times P(\beta_1 \neq 0 \text{ in at least } K_1 \text{ tissues}), \end{aligned}$$

where $\alpha_c$ is a vector of cis-association effects for the set of eQTLs in a single tissue type, and $\beta_1$ is the conditional correlation of cis- and trans-gene expression levels in a single tissue type.

To quantify the cross-tissue cis-association probability $P(\alpha_c \neq 0 \text{ in all } K \text{ tissues})$, the gene-level cis-association statistics are obtains for $M$ cis-genes by $F$-tests. Using Gleason et al's integrative

association analysis approach Primo[73], the estimated $\hat{P}(\alpha_c \neq 0$ in all $K$ tissues$)$ for genes $1 \leq i \leq M$. Similarly, Primo is applied to the conditional correlation statistics from $K$ tissue types for the $M_i'$ trans-genes for each cis-gene $i$ to estimate the probability of non-zero conditional correlation in at least $K_1$ tissue types for all trans-genes of a cis-gene. The product of these two probabilities gives a lower bound on the probability of gene-level cis-mediated trans-associations for each trio. CCmed$_{\text{GWAS}}$ detects trans-genes associated with GWAS SNPs similarly to CCmed$_{gene}$ using Primo to estimate the probability that (1) the GWAS SNP is also a cis-eQTL for the cis-gene conditional on other cis-eQTLs and (2) there is a non-zero correlation between the cis- and trans-gene expression levels conditioning on the genotypes of eQTL and GWAS SNPs.

### 1.3.3 Incorporation of regulatory information in TWAS

Here, we give a brief review of an extension of traditional cis-only TWAS that incorporates information from regulatory elements into the prediction framework. Traditional cis-only prediction models treat all local genotypes as equally predictive of expression, though variants that lie with cis-regulatory elements like promoters or enhancers are more likely to affect expression[74–76]. To this end, Zhan et al proposes EpiXcan, a simple extension of the PrediXcan, that prioritizes local SNPs around the gene of interest if they are involved in a cis-regulatory element[77]. Here, epigenomic annotations for a given tissue are obtained from the Reference Epigenome Mapping Centers[78] to estimate a posterior probability that a given eQTL is causal for the regulation of the given gene based on the annotations. This posterior probability is estimated using qtlBHM[79], a Bayesian framework that uses eQTL summary statistics and functional annotations. These posterior probabilities are then rescaled to penalty factors using Bézier curves employing a shifting-window strategy to approximate the data-driven function. Finally, these penalty factors are included into a weighted elastic net prediction model that minimizes a modification of the elastic net objective function:

$$f(\beta, \lambda, \alpha) = \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda\alpha \sum_{j=1}^{m} \omega_j |\beta_j| + \lambda(1-\alpha)\beta^T \Omega \beta,$$

where $\beta$ is the SNP effect-sizes on gene expression, $y_i$ is the gene expression for the $i$th sample, $X_i$ is the dosages for cis-genotypes of the $i$th sample, $\omega$ is the vector of SNP penalties, $\Omega$ is a matrix with diagonal $\omega$ and 0 for off-diagonal elements, and $\lambda$ and $\alpha$ are penalization parameters as in elastic net[47,48]. The optimal $\beta$ gives the SNP weights for the predictive model of the gene of interest. In

general, this prioritization of cis-SNPs generates a small, yet considerable, gain in prediction of gene expression and increases the power to detect significant gene-trait associations[77].

### 1.3.4 Summary

In general, several studies have shown that variation in both phenotypes and gene expression is attributed to the aggregation of countless distal variants with small effects on the trait or gene expression. Groups have shown that a mediation framework is powerful to identify these most important trans variants, usually associating trans-eQTLs with cis-regulators that cascade effects to many distal genes that are important in a given tissue. TWAS extensions have also shown the utility of included regulatory information into the TWAS predictive framework. However, there are gaps in current transcriptomic prediction in identifying, prioritizing, and leveraging these distal or trans-SNPs for increase predictive power and power to detect gene-trait associations.

## 1.4 mRNA expression-based cell-type deconvolution

Here, we discuss another source of biological heterogeneity: cell-type composition in bulk tissue. Bulk tissue, especially in cancerous tumors, comprise of many different cell types, many rare, and each contributing a different amount to the assay of interest (i.e. mRNA expression, DNA methylation, etc)[80,81]. This cell-type heterogeneity makes it difficult to distinguish gene expression variability that reflects shifts in cell populations from variability that reflects changes of cell-type-specific expression[82]. Since the advent of RNA-seq technology, cell-type deconvolution from mRNA expression has become important in genetic and genomic association testing, either using compositions in regression models as covariates to adjust for the association between cell-type proportions and phenotype[83–85] or use them as inputs to solve for cell-type specific quantities[6].

mRNA expression-based cell-type deconvolution can be formulated as a matrix factorization of $X$, a $n \times p$ matrix of raw scale mRNA expression from $p$ genes and $n$ samples:

$$X = PS \tag{1.3}$$

where $P$ is the $n \times k$ proportion matrix for $n$ samples across $k$ cell types, and $S$ is the $k \times p$ expression signature matrix for $k$ cell-types across $p$ genes. In all cases of deconvolution, $X$ is known and is a required input. In many studies, references for cell-type gene signatures are known; in these

*reference-based* methods, estimation of $S$ is direct from the references, and $P$ can be estimated via some modification of regression. Alternatively, when reference gene signatures are not readily available for a given sample, there are multiple *reference-free* methods. Here, we outline several reference-based and reference-free methods for cell-type deconvolution.

### 1.4.1 Reference-based deconvolution methods

Early reference-based methods have approached reference-based deconvolution using some form of non-negative or constrained least squares method. After filtering out low and high variance genes, Gong et al's DeconRNASeq optimizes the following objective function with quadratic programming[86]:

$$\min_{P} \left( \|PS - X\|^2 \right), \text{ such that } \sum_{i=1}^{k} p_{ki} = 1,\ p_{ki} \geq 0,\ \forall i.$$

The unmix method in Love et al's DESeq2 package modifies this objective function to include low and high variance genes by employing a variance stabilizing transformation $VST(\cdot)$[16] and solves it with a limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm:

$$\sum_{i=1}^{k} \left( |VST(X_{.i}) - VST(PS_{.i})|^q \right), \text{ such that } \sum_{i=1}^{k} a_{ki} = 1,\ a_{ki} \geq 0,\ \forall i.$$

The Digital Sorting Algorithm (DSA) is an extension of these regression based methods that use gene signatures, rather than reference cell-type expression profiles, to first estimate $S$[87]. DSA requires an input of cell-type specific gene signatures. We encode the average of all genes highly expressed in a single cell-type in the rows of $\tilde{S}$, which is not observed. Instead, let $\tilde{X}_s$ be the average of all genes highly expressed in the observed mixed sample. We then consider the rearrangement of Equation 1.3 as $\tilde{S}_s^{-1}\tilde{X}_s = W$, and since each column of $W$ sums to 1, the elements of $\tilde{S}_s$ can be determined with least squares. Accordingly, $P$ can be estimated via non-negative least squares from this estimated $\tilde{S}_s$.

There are other reference-based approaches that employ other statistical techniques to estimate cell-type proportions. Quon et al builds upon their own ISOLATE computational strategy in the ISOpure algorithm by maximum *a posteriori* estimation of tumor proportions[88]. Given $R$ healthy (or non-tumor) profiles in the data denoted $b_1, \ldots, b_R$, the $n$th sample's total tumor profile $t_n$ can be expressed as

$$t_n = \alpha_n c_n + \sum_{r=1}^{R} \theta_{n,r} b_r + e_n,$$

where $\theta_{n,1}, \ldots, \theta_{n,R}$ are parameters to be estimated by ISOpure with the assumptions that these parameters are non-negative and satisfy $\alpha_n + \sum_{r=1}^{R} \theta_{n,r} = 1$. The ISOpure algorithm reduces to the maximization of a count vector $x_n$ under a multionomial distribution whose probability vector over transcripts is $\hat{x}_n = \alpha_n c_n + \sum_{r=1}^{R} \theta_{n,r} b_r$. The score of a given parameter setting is the product of the score of the parameters under the Dirichlet prior distributions and the probability of the discretized tumor profiles under the multinomial distribution for $\hat{x}_n$.

Another reference-based method, DeMixT, is a semi-supervised approach to tumor deconvolution that assumes that bulk tumor tissue expression $Y_{ig}$ for a single gene $g$ is a mixture of a single tumor component $T_{ig}$ and two non-tumor components $N_{1,ig}$ and $N_{2,ig}$[89]:

$$Y_{ig} = \pi_{1,i} N_{1,ig} + \pi_{2,i} N_{2,ig} + (1 - \pi_{1,i} + \pi_{2,i}) T_{ig}.$$

Here, only two of $N_{1,ig}, N_{2,ig}$ Each of the component expressions are assumed to be $\log_2$-normally distributed and all model parameters are estimated via maximum likelihood using iterated conditional modes[90]. DeMixT is a powerful method, however the assumption of three total tissue components for tumors may be untenable.

More specifically for the deconvolution of immune infiltrate, CIBERSORT uses a $\nu$-support vector regression to solve for the $P$ matrix, given inputs for $X$ and $S$[91]. Briefly, the method defines a hyperplane that captures as many data point as possible given defined constraints and reduces overfitting by only penalizing data point outside an error radius using a linear epsilon-insensitive loss function. The orientation of the hyperplane determines the estimated $P$.

In many cases, microdissection or pure samples of cell-types cannot be obtained. In this case, external reference panels can also be applied, with several methods addressing this approach. In particular, Dong et al proposed SCDC that deconvolves bulk gene expression using multiple single-cell RNA sequencing (scRNA-seq) references[92] in an ensemble approach. Based on every single-cell reference $i$ obtained, we can obtain an estimated proportion matrix $\hat{P}_i$ and integrate these estimated proportion matrices with weights $w_i$, such that $\hat{P} = \sum_{i=1}^{R} w_i \hat{P}_i$. The weights can be optimized by minimizing the difference $P - \hat{P}$. Since $P$ is unknown, the surrogate $X$ and $X_i$ can be used to find

$$(\hat{w}_1, \ldots, \hat{w}_R) = \text{argmin}_{(w_1,\ldots,w_R)} \| X - \sum_{i=1}^{R} w_i X_i \|$$

using numerical method based on grid search to maximize the Spearman correlation between $X$ and $\hat{X}$.

Concurrently, Wang et al constructed MuSiC, a method that also utilizes cell-type specific gene expression from scRNA-seq data to characterize cell type compositions from bulk RNA-seq data[93]. A key concept in MuSiC is marker gene consistency - that, when using scRNA-seq data as a reference for cell type deconvolution, cross-subject and cross-cell consistency must be considers to guard against bias in subject selection and cell capture in scRNA-seq, respectively. Rather than pre-selecting marker genes from scRNA-seq based only on mean expression, MuSiC gives weights to each gene to allow for the use of more genes. Genes with low cross-subject variances are down-weighted, whereas genes with high cross-subject variances are up-weighted. To deal with collinearity from correlated genes, MuSiC uses a tree-guided process that recursively finds closely related cell types.

### 1.4.2 Reference-free deconvolution methods

The precursor to reference-free deconvolution methods was deconf, a algorithm based on non-negative matrix factorization and iteration of estimation of $S$ and $P$ until $\|X - PS\|$ reaches convergence[12]. Since then, a population form of reference-free methods uses a geometric approach. UNDO, a method from Wang et al, assumes tumor and stroma compartments for bulk cancerous tissue and attempts to discern relative proportions with the assumption that there exist genes that are significantly expressed in one compartment over the other[94]. Here, for genes $i = 1, \ldots, p$,

$$\begin{bmatrix} x_{\text{sample1}}(i) \\ x_{\text{sample2}}(i) \end{bmatrix} = \begin{bmatrix} a_{12} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} s_{\text{tumor}}(i) \\ s_{\text{stroma}}(i) \end{bmatrix} \rightarrow \mathbf{x}(i),$$

where $s_{\text{tumor}}(i)$ and $s_{\text{stroma}}(i)$ are the gene expressions in pure cells and $\mathbf{x}(i)$ are the gene expression values in heterogeneous samples, and $a_{jk}$ are the mixing proportions, such that $a_{11} + a_{12} = a_{21} + a_{22}$. Given this assumption of cell-specific marker genes, the linear latent variable model above is identifiable. Marker genes are located by searching along the two radii of a scatter sector that correspond to genes with the minimum and maximum ratio between the two mixed samples. From here, the tumor-stroma proportions are estimated using the marker gene expression and the cell-specific expression profiles are obtained via matrix inversion[94].

Zaitsev et al extends this simplex hunting formulation for multiple cell-types in LINSEED motivated by a mutual linearity assumption that proposes that genes that are highly co-expressed in a single cell type are directly proportionally expressed[13]. Given the inputted matrix of observed mixed expressions $X$, each row is normalized by its sum aligns this proportion between co-expressed genes in a cell-type to 1 and $X$ is transposed. Using the SISAL algorithm[95], the vertices of the geometric simplex can be

identified and the top genes per corner can be obtained as a gene signature for a given cell type. These gene signatures can then be inputted to the DSA framework to deconvolve the bulk expression signal[87]. Singular value decomposition on $X$ to find the number of linearly independent components that contribute to variation can be used to obtain the number of cell types in the dataset.

CDSeq, a Bayesian framework for deconvolution from raw RNA-seq data, provides an alternative to full reference-free deconvolution using a Latent Dirichlet Allocation model[5]. Here, the random variable that models the cell-type specific gene expression profiles depends on gene length. Next, the probability of having a read from a cell type depends on both the proportion of the cell type in the sample and the typical amount of RNA produced by cells of that type. Together, these assumptions account for the ratio of mRNA expression contributed by a specific cell type to the size of the cell.

Lastly, an iterative algorithm by Li and Wu called TOAST that better selects features by identifying features showing distinct profiles among difference cell types, without known the pure cell type profiles or mixing proportions *a priori*[14]. The general rule of thumb for selecting informative features are genes with low within-cell type variation and high cross-cell type variation. Assume, for the $p$-th feature, we have $Y_p = [Y_{p1}, \ldots, Y_{pn}]^T$. The proportions obtained for the $i$th sample are denoted as $\theta_i = (\theta_{i1}, \ldots, \theta_{sk})$. With known proportions, the observed data can be modeled by a linear model: $E(Y_p) = V\beta_p$, where $V$ is the matrix of $\theta$ proportions and $\beta_p$ is the mean levels for the $p$th feature in the $j$th cell type. This model allows for the testing of the null hypothesis

$$H_0 : \ \mu_{pj} - (k-1)\sum_{i \neq j} \mu_{pi} = 0, \ j = 1, \ldots, k.$$

Features with significant test results are cell-type specific features. TOAST can be used with any form of reference-free deconvolution method to improve estimation of $S$ and $P$ iteratively[14].

### 1.4.3 Recapitulation of cell-type specific expression

A moonshot goal for many deconvolution methods is recapitulating cell-type expression profile per sample. DeMixT addresses this goal by successive parabolic interpolations to find the maximum of the joint density function with respect to the expressions specific to the normal compartments[89], with positive constraints such that sum of the normal components cannot exceed the total mixed expression. The tumor-specific compartment can then be easily estimated from there.

Wang et al also develop a method based on linear mixed modeling for the purpose of estimating the cell-type specific expression profiles per sample[93]. MIND extends single-measure deconvolution

by borrowing information across multiple measurements $t = 1, \ldots, T_i$ from the same tissue for subject $i$ to estimate subject-specific and cell-type specific gene expression. First, cell-type fractions for subject $i$ and measure $t$ (denoted $W_{it}$) can be estimated and combined across measures to yield $W_i$ a $T_i \times k$ matrix. Next, treating $W_i$ as known, the problem is reversed to estimated the cell-type specific expression. For gene $j$ in subject $i$, the observed gene expression $X_{ij}$ is a $T_i$-dimensional vectors that represents $T_i$ measurements, rather than a scalar and can be modelled as a product of $W_i$ and the cell-type specific expressions $A_{ij}$, such that $X_{ij} = W_i A_{ij} + e_{ij}$. It is assumed that $A_{ij} \sim N(a_j, \Sigma_c)$ and $e_{ij} \sim N(0, \sigma_e^2 I_T)$. Parameters $a_j$ and $\Sigma_c$ are estimated with an Expectation-Maximization and $A_{ij}$ is estimated via an empirical Bayes procedure.

### 1.4.4 Summary

We have outlined several deconvolution methods, both reference-based and reference-free. Each method hinges on identifying genes whose distributions can distinguish different, often rare, cell types. This is already a challenging problem in many RNA-seq datasets with thousands of genes. In targeted panels that assay gene expression with only hundreds of genes, the limited feature space casts a considerable statistical challenge in inferring cell-type proportions and recapitulating cell-type specific expression from bulk mRNA expression.

# CHAPTER 2: AN APPROACH FOR NORMALIZATION AND QUALITY CONTROL FOR NANOSTRING RNA EXPRESSION DATA

In this chapter, we provide a framework for the quality control and normalization of mRNA expression count data from the NanoString nCounter platform, using a large dataset of breast tumor expression from the Carolina Breast Cancer Study (CBCS) and various other cohorts of differing sample size. We illustrate some of the pitfalls in the popularly-used nSolver method of background correction and positive control normalization and provide an alternative approach that uses RUVSeq[27], which efficiently estimates unwanted variation from endogenous housekeeping genes. Lastly, we provide various quality checks for normalization and outline the impact of proper normalization on inference for endogenous genetic associations and expression-based disease subtyping.

## 2.1 Overview of quality control and normalization process

The full quality control and normalization process using nSolver and RUVSeq is summarized in **Figure 2.1**, starting with familiarization of the raw data (**Figure 2.1.1**), technical quality control (**Figure 2.1.2**), pre-normalization assessment of housekeeping genes (**Figure 2.1.3**) and data visualization to detect problematic samples and assess whether flagged samples should be removed (**Figure 2.1.4**). Normalization is performed with either nSolver or RUVSeq (**Figure 2.1.5**), and the processed expression data is assessed for validity through relevant visualization and biological checks (**Figure 2.1.6**). If validation is unsatisfactory and technical variation is still present, this process is iterated.

### 2.1.1 Technical quality control flags

The first step in quality control is an assessment of the assay quality. nSolver provides several quality control (QC) flags to assess the quality of the data for imaging, binding density, linearity of the positive controls, and limit of detection. The definition and implementation of thess QC flags are summarized in detail in the nSolver[30] and NanoStringNorm[31] documentation. Here, we mark any sample that is flagged in at least one of these four QC assessments as *technical quality control*. We

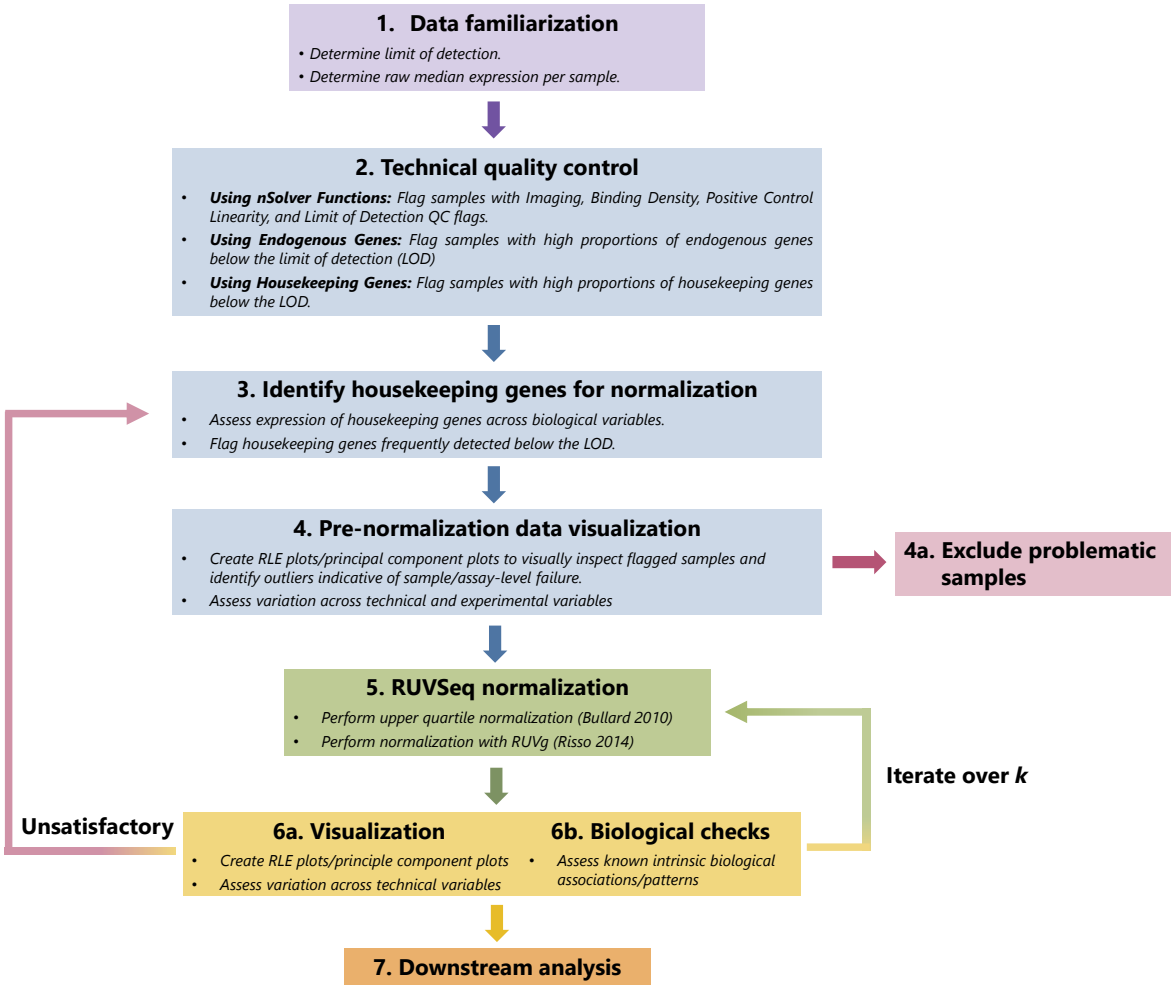


**Figure 2.1:** *Graphical summary of both nSolver and RUVSeq normalization pipelines.* The quality control and normalization process starts with familiarization with the data (**Step 1**) and technical quality control to flag samples with potentially poor quality (**Step 2**). After a set of housekeeping genes are selected (**Step 3**), important unwanted technical variables are also investigated through visualization techniques (**Step 4**). Problematic samples (e.g. those that are flagged multiple times in technical quality control checks) are excluded. Next, the data is normalized using upper quartile normalization and RUVSeq (**Step 5**), and the normalized data is visualized to assess the removal of unwanted technical variation and retention of important biological variation (**Step 6**). Steps 3—6 are iterated until technical variation is satisfactorily removed, changing the set of housekeeping genes or the number of dimensions of unwanted technical variation ($k$) estimated using RUVSeq. This data can then be used for downstream analysis (**Step 7**).

use these QC flags in both nSolver normalization and RUVSeq normalization to indicate samples that may be of poor quality.

### 2.1.2 Housekeeping gene assessment

Next, we consider the genes on the nCounter assay that can potentially serve as housekeepers in nSolver- or RUVSeq-normalization. Housekeeping genes serve two purposes: (1) for QC purposes to remove samples with overall poor quality and (2) to use for assessing the amount of technical variation and further normalization. There are differences in the definition of housekeeping genes (or negative control probes) between the nSolver and RUVSeq-based processes[27–30]. We define a good housekeeping gene for nCounter expression as one that has little-to-no variability across all treatment conditions and is not expressed below the limit of detection in samples that pass QC. NanoString further suggests that ideal housekeeping genes are highly expressed, have similar coefficients of variation, and have expression values that correlate well with other housekeeping genes across all samples. Because of these definitions, these targets will ideally vary only due to the level of technical variation present. To assess the potential for housekeeping correction to introduce bias, housekeeping genes were assessed for differential expression across a primary biological covariate of interest (estrogen receptor status in CBCS, tumor stage in the kidney and bladder cancer data, and treatment groups in Sabry et al) using negative binomial regression on the raw counts from the MASS package[96]. Genes with Benjamini-Hochberg[1] FDR-adjusted $P < 0.05$ were flagged, as their association with the outcome of interest may lead to removal of biological variance due to the primary outcome of interest.

### 2.1.3 Below limit of detection (LOD) quality control

Lastly, as samples with high proportions of both endogenous and housekeeping genes below the limit of detection (LOD) may be indicative of reduced assay or sample quality, we define another QC flag: the number of genes measured at below the LOD per sample. Here, we define the per-sample limit of detection as the mean of the counts of negative control probes for that given sample. We further assessed the percent of counts below the LOD in the housekeeping genes per sample as an added QC step to flag both poor quality samples and poor housekeeping genes. Samples were flagged if they have more than one housekeeping gene missing and a median percent below LOD in the endogenous genes greater than the 75th percentile of the samples with all housekeeping genes present. Based on the age of samples and the level of degradation detected, this per-sample LOD can be tuned. For example, the LOD can be shifted by multiples of the standard deviation to allow for more

liberal or conservative cutoffs. We further assessed the percent of counts below the LOD in the housekeeping genes per sample as an added QC step to flag both housekeeping genes and samples with high proportions of below the LOD. For a given sample, an inflated proportion below the LOD in housekeeping genes may indicate degradation of the sample, especially when it correlates with a high proportion below the LOD in endogenous genes.

## 2.2 Normalization of mRNA expression

### 2.2.1 Background correction with nSolver

NanoString whitepapers and guidelines suggest background corrections[30,31] by either subtraction or thresholding for an estimated background noise level for experiments in which low expressing targets are common, or when the presence or absence of a transcript has an important research implication[29,31]. We believe that the datasets we consider in this work do not fall under this criterion, and accordingly, we do not background correct by either thresholding or subtraction. However, we contend that this step may introduce bias in most analyses conducted on NanoString data and should be generally avoided, as Freytag et al and Irizarry et al point out[97,98].

Background thresholding led to increased per-sample variance while per-sample medians remained relatively similar (**Supplemental Figure S1A**). The distributions of per-sample median expression values were more right-skewed (greater mean than median) when using background thresholding prior to normalization compared to not using background thresholding (**Supplemental Figure S1B**). Based on this analysis, we did not perform background correction prior to normalization for all cohorts analyzed.

### 2.2.2 Positive control and housekeeping gene-based normalization with nSolver

For nSolver normalization, the arithmetic mean of the geometric means of the positive controls for each lane is computed and then divided by the geometric mean of each lane to generate a lane-specific positive control normalization factor[30,31]. The counts for every gene are then multiplied by its lane-specific normalization factors. To account for any noise introduced into the nCounter assay by positive normalization, the housekeeping genes are used similarly as the positive control genes to compute housekeeping normalization factors used to scale the expression values[30,31]. NanoString

also flags samples with large housekeeping gene scaling factors (we call this a *housekeeping QC flag*) and large positive control spike-in scaling factors (*positive QC flag*).

### 2.2.3 RUVSeq normalization pipeline

After quality control and housekeeping assessment, we alternatively also started the RUVSeq-based normalization process, an alternative approach to nSolver normalization (see Figure 2.1). We rescaled distributional differences between lanes with upper-quartile normalization[99]. Unwanted technical factors were estimated in the resulting gene expression data with the RUVg function from the RUVSeq Bioconductor package[27,29]. RUV-III has been created specifically for normalization of NanoString data with technical replicates[28]; however, as the datasets we discuss here do not have technical replicates, we proceeded with RUVg. Unwanted variation was estimated using the distribution of the endogenous housekeeping genes not associated with the outcome of interest on the NanoString gene expression panel. We removed $k$ dimensions of unwanted variation (varied by dataset) from the variance-stabilized transformed-scaled counts of gene expression data[100,16]. We lastly used relative log-expression (RLE) plots and principal component analysis to detect systemic deviation across various technical and biological groups and any potential outliers.

### 2.2.4 Alternative normalization methods for benchmarking

Using CBCS data, we compared the normalized datasets from nSolver, RUVSeq[29], NanoStringDiff[33], and RCRnorm[101] with the raw data through visualization methods outlined above (**Figure 2.1.1 to 2.1.4**, RLE plots and scatter plots of principal components over important technical and biological sources of variation). Details about these methods are provided in **Supplemental Table S1**.

## 2.3 Results

We evaluated the ability of normalization methods to remove technical variation while retaining biologically meaningful variation across four cohorts of differing sample size and varying sources of technical bias. Known sources of technical variation included age of sample (study phase) and different study sites. The cohorts varied in preservation methods; two cohorts used fresh-frozen specimens, while two used archival FFPE specimens. The number of genes measured for both

endogenous genes and housekeeping genes also varied by study. In addition, some studies used validated and optimized code sets for specific gene signatures versus a more general code set.

In cohorts with large technical biases, RUVSeq provided superior normalization with more robust removal of technical variation and provided stronger biological associations compared to other normalization methods. In two of the datasets, we found that downstream analyses performed on data normalized with nSolver and RUVSeq detected substantially different biological associations. However, when few strong technical biases were present or if a validated and optimized code set (e.g. PAM50 genes) was used, nSolver and RUVSeq performed comparably.

### 2.3.1 Case study: targeted panel from the Carolina Breast Cancer Study (CBCS)

The Carolina Breast Cancer Study (CBCS) is a multi-phase cohort of women with breast cancer in North Carolina. Samples were collected during three study phases: Phase 1 (1993-1996), Phase 2 (1996-2001), and Phase 3 (2008-2013). Paraffin-embedded tumor blocks were reviewed and assayed for gene expression using the NanoString nCounter system as discussed previously[24,102]. Study phase gives the relative age of the tumor block. In total, 1,649 samples from patients with invasive breast cancer from CBCS, across all three study phases, were analyzed on a custom panel of 417 genes. All assays were performed in the Translational Genomics Laboratory (TGL) at the University of North Carolina at Chapel Hill (UNC). After quality control and normalization, 1,264 samples remained in the nSolver-normalized data, and 1,219 samples remained in the RUVSeq-normalized data. This dataset was also used to benchmark against NanoStringDiff[33] and RCRnorm[101], using the same 1,264 samples in the nSolver-normalized set.

#### 2.3.1.1 Quality assessment of expression levels using LOD of housekeeping genes

We used the housekeeping genes to assess if the lack of expression of endogenous genes was due to biology or due to technical failures. We compared the level of missing endogenous genes in samples with all housekeeping genes present to those with increasing number of housekeeping genes below LOD. There was a strong positive correlation for increasing proportions of genes below the LOD in both the endogenous and housekeeping genes (**Figure 2.2A** and **Supplemental Figure S1**). Samples with higher numbers of genes below the LOD were from earlier phases of CBCS (i.e. Phase 1 from 1993-1996 and Phase 2 from 1996-2001), and thus associated with sample age (**Figure 2.2A** and **Supplemental Figure S2**). Samples with a higher proportion of endogenous genes below the LOD had increased numbers of QC flags as well (**Supplemental Figure S1**).

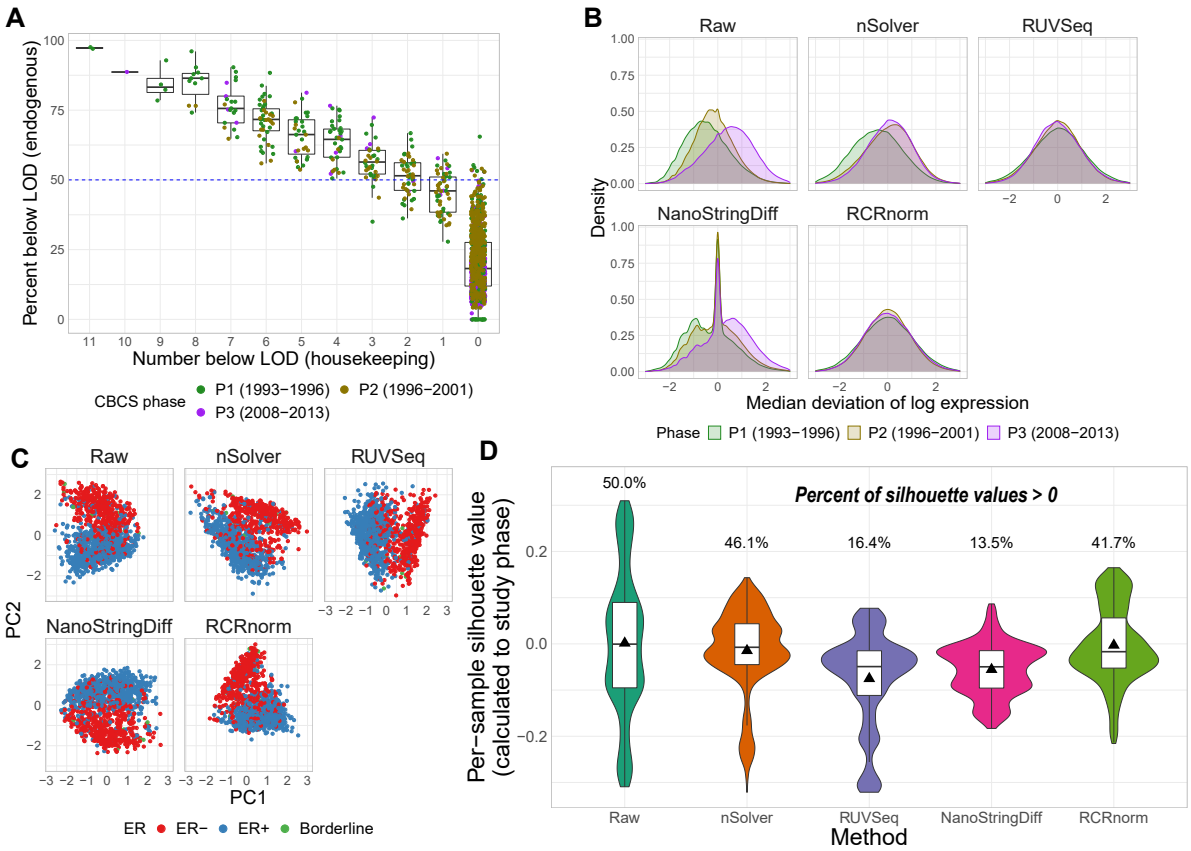


**Figure 2.2:** *Quality control and normalization validation in CBCS.* (A) Boxplot of percent of endogenous genes below the limit of detection (LOD) ($Y$-axis) over varying numbers of the 11 housekeeping genes below LOD ($X$-axis), colored by CBCS study phase. Note that the $X$-axis scale is decreasing. (B) Kernel density plots of deviations from median per-sample $\log_2$-expression from the raw, nSolver-, RUVSeq-, NanoStringDiff-, and RCRnorm-normalized expression matrices, colored by CBCS study phase. (C) Plots of the first principal component ($X$-axis) vs. second principal component ($Y$-axis) colored by estrogen receptor subtype of the raw, nSolver-, RUVSeq-, NanoStringDiff-, and RCRnorm-normalized expression data. (D) Violin plots of the distribution of per-sample silhouette values, as calculated to study phase, using raw, nSolver-, RUVSeq-, NanoStringDiff-, and RCRnorm-normalized expression. The boxplot shows the 25% quartile, median, and 75% quartile of the distribution, and the plotted triangle shows the mean of the distribution.

### 2.3.2 Evaluation of normalization methods

We benchmarked RUVSeq and nSolver with two other normalization methods, NanoStringDiff[33] and RCRnorm[101]. We observed differences across the four normalization strategies (described in **Supplemental Table S1**), namely greater remaining technical variation using nSolver and NanoStringDiff than RCRnorm and RUVSeq (**Figure 2.2B-D**). A large portion of the variation in the raw expression could be attributed to study phase (**Supplemental Figure S4A**). While all methods reduced study phase associated variation compared to the raw data, there were considerable differences in the deviations from the median log-expressions in the nSolver- and NanoStringDiff-normalized expression that are not present in the RUVSeq- and RCRnorm-normalized data (**Figure 2.2B**). The nSolver and NanoStringDiff methods retained technical variation, either not fully corrected or re-introduced during the nSolver normalization process.

We examined the ability of each normalization method to retain biological variation. Estrogen Receptor (ER) status is one of the most important clinical and biological features in breast cancer and is used for determining course of treatment[103,104]. ER status drives many of the molecular classification[105–107] and even drives separate classification of breast tumors in TCGA's pan-cancer analysis of 10,000 tumors[108]. In the raw expression, variation due to ER status was captured in PC2 rather than PC1 (study age); however, after RUVSeq-normalization, ER status was reflected predominantly in PC1 (**Figure 2.2C**). In the nSolver-, NanoStringDiff-, and RCRnorm-normalized data, ER status was shared between PC1 and PC2, suggesting that unresolved technical variation was still present. RUVSeq demonstrated effective removal of technical variation and boosting of the true biological signal. The PAM50 molecular subtypes[109], which are also linked with ER status, were also clearly separated by PC1 for RUVSeq-normalized data, but this was not the case for nSolver-, NanoStringDiff-, or RCRnorm-normalization (**Supplemental Figure S4B**). These results suggest that RUVSeq-normalization best balances the removal of technical variation with the retention of important axes of biological variation, with RCRnorm showing better performance than nSolver and NanoStringDiff, but not superior to RUVSeq. A significant disadvantage of RCRnorm is its computational cost: RCRnorm was unable to run on the CBCS dataset ($N = 1278$ after QC) on a 64-bit operating system with 8 GB of installed RAM, requiring RCRnorm-normalization to be performed on a high-performance cluster. We summarize the maximum memory used by method in CBCS in **Supplemental Table S1**.

We used silhouette width to assess extent of unwanted technical variation from study phase remaining by the normalization methods. We consider the silhouette value $s_i$ for sample $i$[110], defined as

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)},$$

where $a_i$ is the mean distance between sample $i$ and all other samples in the same study phase and $b_i$ is the smallest mean distance of sample $i$ to all points in any other study phase. Larger positive silhouette values indicate within-group similarity (i.e. samples clustering by study phase). Per-sample silhouettes across the alternatively normalized datasets showed that RUVSeq best addressed the largest source of technical variation identified in the raw data (**Figure 2.2D**; **Supplemental Figure S5A**) while also not removing a significant portion of biological variation (**Supplemental Figure S5B**). NanoStringDiff also demonstrated less similarity of samples across study phase similar to RUVSeq but removed biologically relevant similarity of samples grouped by ER status. Due to the performance of NanoStringDiff and computational limitations of RCRnorm, for subsequent analyses and datasets, we only illustrate differences between nSolver- and RUVSeq-normalized data.

### 2.3.3 Genomic analyses and expression profiles across normalization methods

We evaluated the impact of normalization choice on downstream analyses including eQTLs, PAM50 molecular subtyping, known expression patterns, and similarity to RNA-seq data. In a full cis-trans eQTL analysis accounting for race and genetic-based ancestry, we found considerably more eQTLs using nSolver as opposed to RUVSeq, thresholding at nominal $P < 10^{-3}$ (2,050 vs. 1,143). We identified strong cis-eQTL signals in both normalized datasets; however, stronger FDR values were identified with RUVSeq (**Figure 2.3A**, densely populated around the 45-degree line). We observed considerably more trans-eQTLs using nSolver, including a higher proportion of trans-eQTLs across various FDR-adjusted significance levels (**Figure 2.3B**; **Supplemental Figure S8**). We suspected that spurious trans-eQTLs may have resulted from residual technical variation in expression data that was confounded with study phase, subsequently being identified as a QTL due to ancestry differences across study phase. In cross-chromosomal trans-eQTL analysis, distributions of absolute differences in minor allele frequency (MAF) for trans-eSNPs across women of African and European ancestry were wide for both methods (**Supplemental Figure S8**). However, we observed substantially more trans-eSNPs with moderate absolute MAF differences across study phase with nSolver, compared to RUVSeq. This provides some evidence for the presence of residual confounding

technical variation in the nSolver-normalized expression data leading to spurious trans-eQTL results (with a directed acyclic graph for this hypothesis in **Supplemental Figure S9**), though we cannot confirm this with eQTL analysis alone.

We compared each normalization method for the ability to classify breast cancer samples into PAM50 intrinsic molecular subtype using the classification scheme outlined by Parker et al[109]. Our PAM50 subtyping calls were robust across normalization methods with 91% agreement and a Kappa of 0.87 (95% CI (0.85, 0.90)) (**Supplemental Figure S6**). Among discordant calls, approximately half had low confidence values from the subtyping algorithm, and half had differences in correlations to centroids less than 0.1 between the discordant calls. Most of these discordant calls were among HER2-enriched, luminal B and luminal A subtypes, which are molecularly similar[111].

We observed noticeable differences between the RUVSeq- and nSolver-normalized gene expression when visualized after hierarchical clustering via heatmaps, similar to the principal component analysis. Using this method, we identified 14 additional samples with strong technical errors in the nSolver-normalized data not previously marked by QC flags (**Supplemental Figure S10**), emphasizing the need for post-normalization data visualization. In early breast cancer clustering papers, the first major division was by ER status separating basal-like and HER2-enriched molecular subtypes (predominantly ER-negative) from luminal A and B molecular subtypes (predominantly ER-positive)[109]. This pattern was observed in RUVSeq-data but only partially preserved with nSolver normalization (**Supplemental Figure S10**). Rather, nSolver data clustering was driven by a combination of ER status and study phase. Study phase dominated two of the groups and were formed by Phase 1 and Phase 3 samples, respectively—samples with a 10+ year difference in age.

Lastly, we compared normalization choices for NanoString data to RNA-seq data performed on the same samples. CBCS collected RNA-seq measurements for 70 samples that have data on a different nCounter codeset (162 genes instead of 417) and RNA-seq normalized using standard procedures. A permutation-based test of independence using the distance correlation[112,113] revealed that the distance correlation between the RNA-seq and nSolver data was small and near 0 (distance correlation = 0.051, $P = 0.24$) while the distance correlation between the RNA-seq and RUVSeq- data was larger (distance correlation = 0.36, $P = 0.02$). The permutation-based test rejected the null hypothesis of independence (distance correlation of zero for unrelated datasets) between RUVSeq-normalized nCounter data and RNA-seq data but fails to reject the null hypothesis for nSolver-normalization nCounter and RNA-seq data. We conclude that RUVSeq produced normalized data with closer relation to the RNA-seq, in terms of distance correlation and test of independence, compared to nSolver.

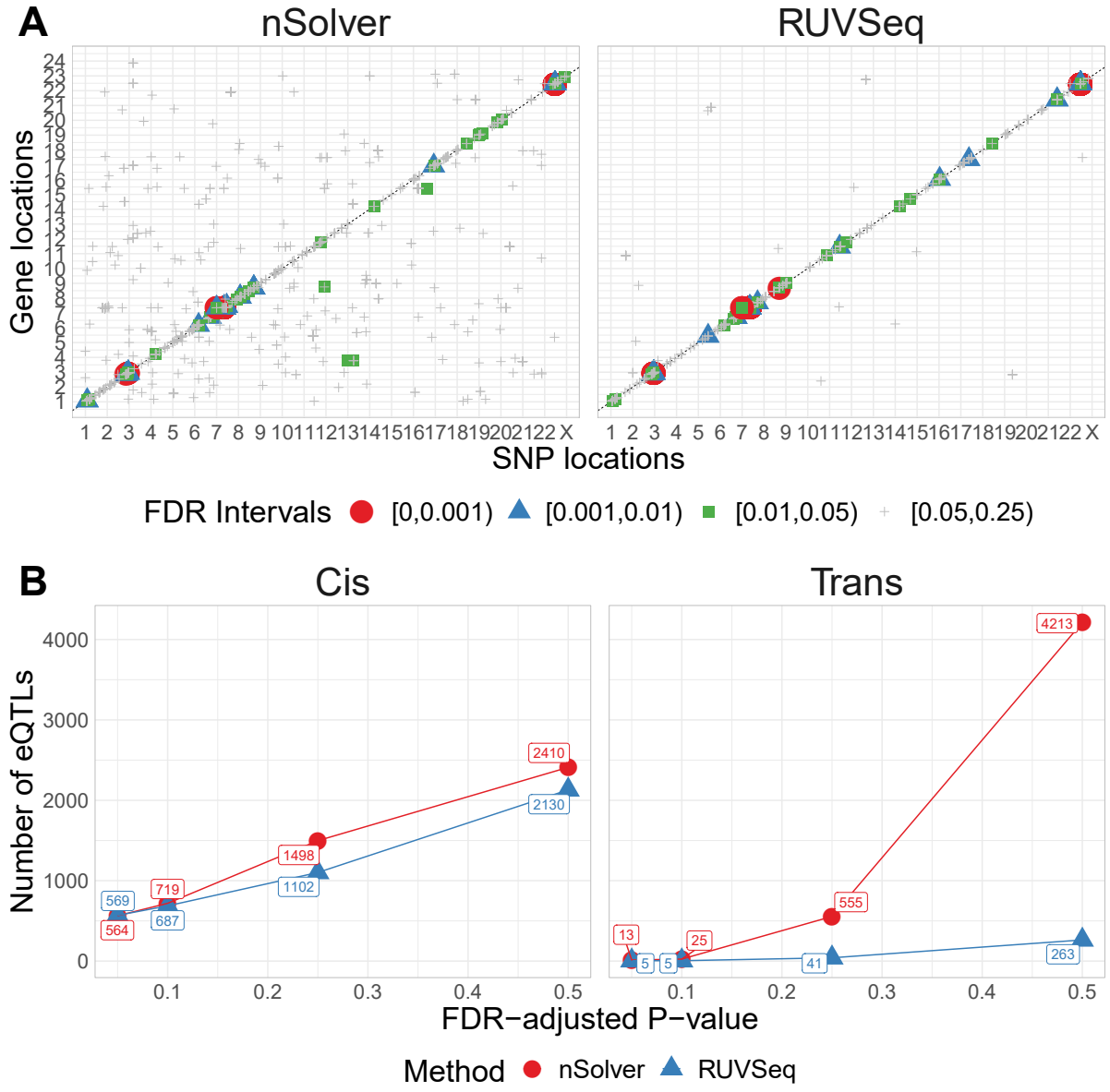


**Figure 2.3:** *eQTL analysis in CBCS.* (A) Cis-trans plots of eQTL results from nSolver-normalized (left) and RUVSeq-normalized data with chromosomal position of eSNP on the $X$-axis and the transcription start site of eGene on the $Y$-axis. Points for eQTLs are colored by FDR-adjusted $P$-value of the association. The dotted line provides a 45-degree reference line for cis-eQTLs. (B) Number of cis- (left) and trans-eQTLs (right) across various FDR-adjusted significance levels. The number of eQTLs identified in nSolver-normalized data is shown in red and the number of eQTLs identified in RUVSeq-normalized data is shown in blue.

### 2.3.4 Case study: differential expression analysis in natural killer cells

We looked at the impact of the two normalization methods in a small cohort ($N = 12$) on DE analysis across natural killer (NK) cells primed for tumor-specific cells and cytokines from Sabry et al[4]. RLE plots before and after normalization showed minor differences between the two normalization methods (**Supplemental Figure S10**).

Using DESeq2[16], we identified genes differentially expressed in NK cells primed by CTV-1 or IL-2 cytokines compared to unprimed NK cells at FDR-adjusted $P < 0.05$. The two normalization methods led to a different number of differentially expressed genes with a limited overlap of significant genes by both methods (**Figure 2.4A**). The raw $P$-value histograms from differential expression analysis using nSolver-normalized expression exhibited a slope toward 0 for $P$-values under 0.3, which can indicate issues with unaccounted-for correlations among samples[114], such as residual technical variation. The distributions of $P$-values using the RUVSeq-normalized data were closer to uniform throughout the range $[0, 1]$ for most genes (**Figure 2.4B**). While the $\log_2$-fold changes were correlated between the two normalization procedures, the genes found to be differentially expressed only with nSolver-normalized data tended to have large standard errors with RUVSeq-normalized data and therefore not statistically significant using RUVSeq (**Figure 2.4C**). These differences in DE results emphasize the importance of properly validating normalization prior to downstream genomic analyses.

### 2.3.5 Further case studies

#### 2.3.5.1 Case study: bladder cancer gene expression

RUVSeq reduced technical variation (study site) while maintaining the biological variation (tumor grade). RUVSeq data showed the most homogeneity in per-sample median deviation of log-expressions compared to raw and nSolver data (**Figure 2.5A**). The first principal component of nSolver data had significant differences by study sites, which was not present in RUVSeq data (**Figure 2.5B**). In addition, there was a stronger biological association with tumor grade in the first principal component of expression using RUVSeq data (**Figure 2.5C**).

#### 2.3.5.2 Case study: kidney cancer gene expression

We only found subtle differences in the deviations from the median expression between the normalization procedures for the kidney cancer dataset (**Figure 2.6A**). This cohort did not have the same known technical variables observed in the other cohorts such as study site or sample age, and

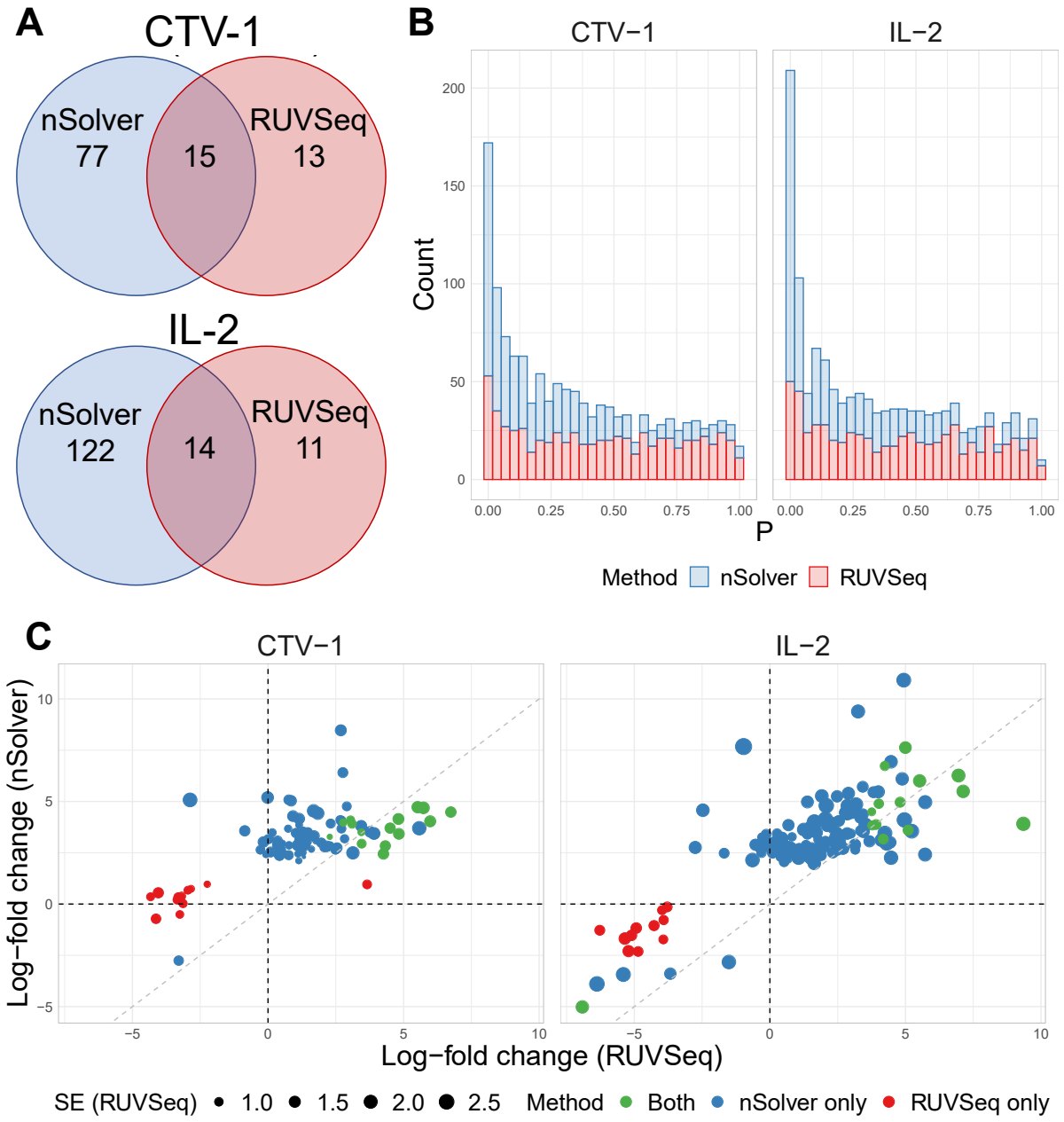


**Figure 2.4:** *Differential expression analysis from Sabry et al*[4] (A) Venn diagram of the number of differentially expressed genes using nSolver-normalized (blue) and RUVSeq-normalized data (red) across comparisons for IL-2-primed (top) and CTV-1-primed NK cells (bottom). (B) Raw P-value histograms for differential expression analysis using nSolver-normalized (blue) and RUVSeq-normalized (red) data across the two comparisons. (C) Scatterplots of $\log_2$-fold changes from differential expression analysis using RUVSeq-normalized data (X-axis) and nSolver-normalized data (Y-axis) for any gene identified as differentially expressed in either one of the two datasets. Points are colored by the datasets in which that given gene was classified as differentially expressed. The size of point reflects the standard error of the effect size as estimated in the RUVSeq-normalized data. $X = 0, Y = 0$, and the 45-degree lines are provided for reference.

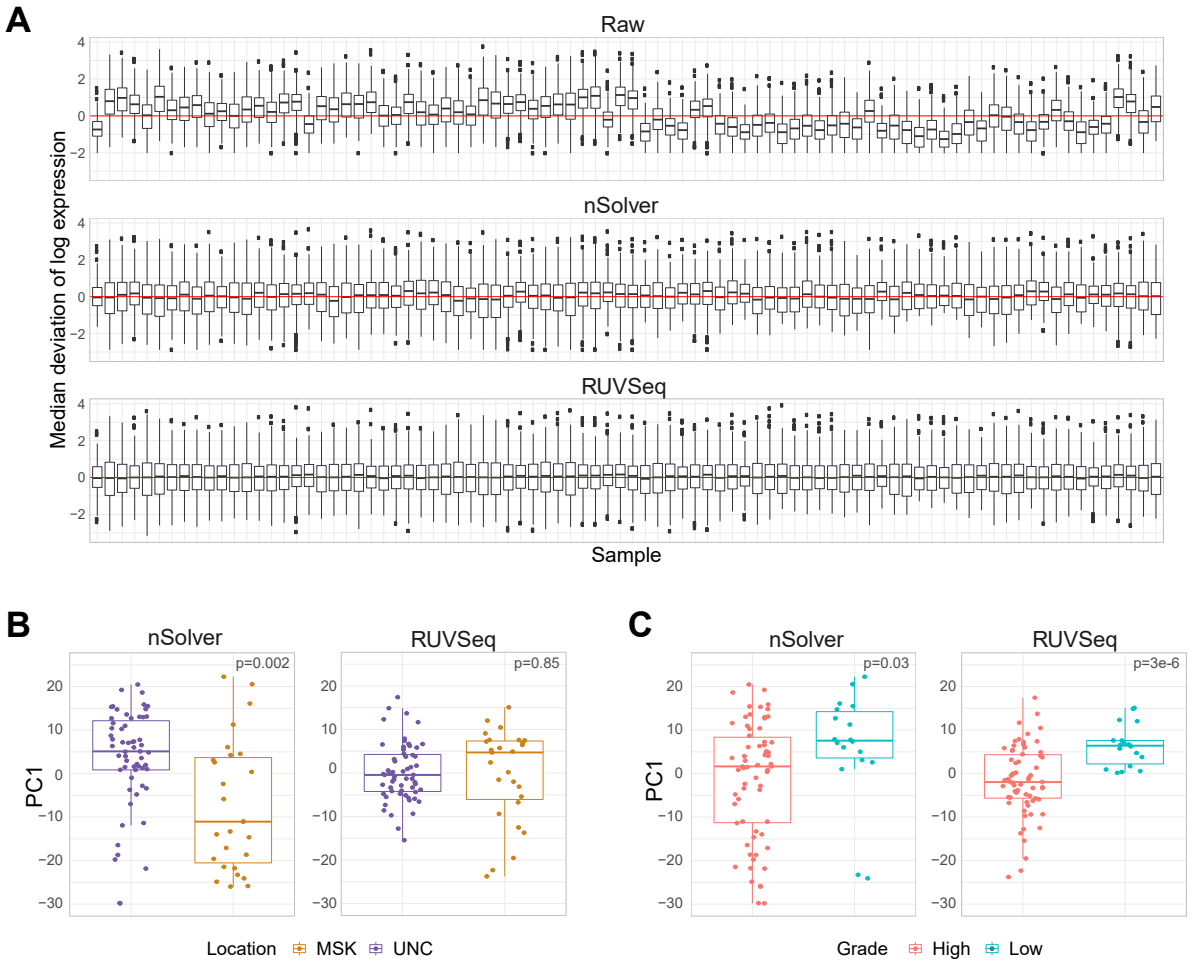


**Figure 2.5:** *Normalization differences in bladder cancer dataset.* (A) RLE plot from bladder cancer dataset, colored by assay month. (B) Boxplot of first principal component of expression by tumor collection site (location) across nSolver- (top) and RUVSeq-normalized (bottom) data. (C) Boxplot of first principal component of expression by tumor grade across nSolver- (top) and RUVSeq-normalized (bottom) data.

the RNA came from fresh-frozen material. We evaluated normalization methods on a source of technical variation, DV300, the proportion of RNA fragments detected at greater than 300 base pairs as a source of technical variation, and tumor stage as a biological variable of interest. The first two principal components colored by level of DV300 (**Figure 2.6B**) and tumor stage (**Figure 2.6C**) showed little difference across the two normalization methods. When there were limited sources of technical variation and a robust, high quality dataset, we found both normalization methods performed equally well.

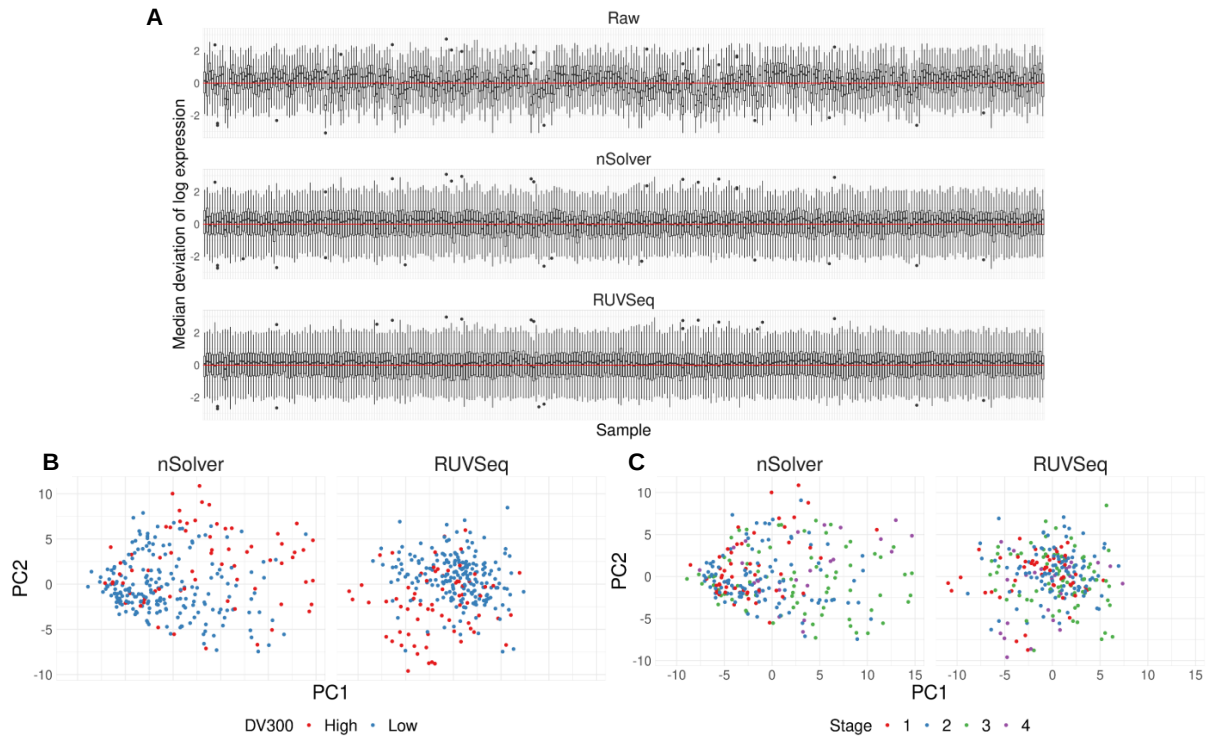


**Figure 2.6:** *Equal performance of normalization procedures in kidney cancer dataset.* (A) RLE plot of per-sample deviations from the median for raw, nSolver, and RUVSeq-normalized data. (B) Scatter plot of the first and second principal component of nSolver- (left) and RUVSeq-normalized (right) expression, colored by high and low DV300. (C) Scatter plot of the first and second principal component of nSolver- (left) and RUVSeq-normalized (right) expression, colored by tumor stage.

## 2.4 Discussion

Proper normalization is imperative in performing correct statistical inference from complex gene expression data. Here, we outline a sequential framework for NanoString nCounter RNA expression

data that provides both quality control checks, considerations for choosing housekeeping genes, and iterative normalization with biological validation using both NanoString's nSolver software[30,31] and RUVSeq[29]. We show that RUVSeq provided a superior normalization to nSolver on three out of four datasets by more efficiently removing sources of technical variation, while retaining robust biological associations. We also benchmark RUVSeq-normalization with two other normalization methods implemented in R and show that RUVSeq outperformed all methods in reducing technical variation.

We observed that normalization methods were sensitive to the quality and the set of housekeeping genes. Several genes thought to behave exclusively in a "housekeeping" fashion in fact associate with biological variables under certain conditions[115] or across different tissue types[116]. A careful validation of housekeeping gene stability on a case-by-case basis and separately for new studies, considering both technical and biological sources of variation in each dataset, is therefore imperative for an optimized normalization procedure.

We developed a quality metric to assess sample quality: samples with high proportions of genes detected below the LOD in both endogenous genes and housekeepers were indicative of either low-quality samples or reduced assay efficiency. Sample age was correlated with higher proportions of genes below the LOD in both endogenous and housekeeping genes, which was likely due to RNA degradation over time. We stress that missing counts in endogenous genes alone does not suggest poor sample quality in the absence of additional QC flags but could represent genes not expressed and therefore not detected under certain biological conditions or cell types. An example includes using an immuno-oncology gene panel in a tumor sample with little to no immune cell infiltration. Conversely, many samples with counts below the LOD in both endogenous genes and housekeepers had additional quality control flags including those derived from nSolver's assessment of data quality. We excluded these samples for analysis in both the nSolver- and RUVSeq-based procedures.

nSolver-normalized data was prone to residual unwanted technical variation when there were known technical biases, such as in CBCS and the bladder example. We checked for known biological associations that are intrinsic to the sample, as in eQTL analysis, to judge the performance of the normalization process[117,70]. A full cis-trans eQTL analysis using nSolver- and RUVSeq-normalized data showed a strong cis-eQTL signal in data from both normalization methods. We found significantly more trans-eQTLs with the nSolver-normalized data. However, many of the trans-eSNPs for the loci found with nSolver-normalized data tended to have moderate MAF differences across phase, leading us to suspect they were spurious associations driven by residual technical variation in gene expression. Such spurious associations from population stratification have been described in many previous studies of eQTL analysis[118–121].

The choice of normalization procedure is less of a concern in cohorts with minimal sources of technical variation or in nCounter targeted gene panels that have been optimized for robust measurement across preservation methods. In the CBCS breast cancer cohort, we identified significant differences in gene expression between normalization methods across the entire gene set (417 total genes). However, PAM50 subtyping was robust across the two normalization procedures. The genes in the PAM50 classifier were selected due to their consistent measurement in both FFPE and fresh frozen breast tissues[109], suggesting that robustly measured genes may be less affected by different normalization procedures. Furthermore, we see minimal differences in residual technical variation in the kidney cancer dataset and the Sabry et al dataset, both of which were measured on either robustly validated genes or nCounter panels. The kidney cancer example had newer, fresh-frozen specimens that were profiled using a small and well-validated set of genes important in that cancer type. This dataset gives an opportunity to stress the importance of the general principles of normalization: as Gagnon-Bartsch et al and Molania et al recommend[27,29], normalization should be a part of scientific process and should be approached iteratively with visual inspection and biological validation to tune the process. One normalization procedure is not necessarily applicable to all datasets and must be re-evaluated on each dataset.

In conclusion, we outline a systematic and iterative framework for the normalization of NanoString nCounter expression data. Even without background correction, a technique which has been shown to impair normalization of microarray expression data[98,97], we believe that relying solely on positive control and housekeeping gene-based normalization may result in residual technical variation after normalization. Here, we show the merits of a comprehensive procedure that includes sample quality control checks including the addition of new checks, assessments of housekeeping genes, normalization with RUVSeq[29] and data analysis with popular count-based R/Bioconductor packages, as well as iterative data visualization and biological validation to assess normalization. Researchers must pay close attention to the normalization process and systematically assess pipelines that best suit each dataset.

# CHAPTER 3: A FRAMEWORK FOR TRANSCRIPTOME-WIDE ASSOCIATION STUDIES IN BREAST CANCER IN DIVERSE STUDY POPULATIONS

This chapter provides a framework for transcriptome-wide association studies for complex disease outcomes in diverse study populations using transcriptomic reference data from the Carolina Breast Cancer Study, a multi-phase cohort that includes an over-representation of African American women[122]. We train race-stratified predictive models of tumor expression from germline variation and carefully validate their performance, accounting for sampling variability and disease heterogeneity, two aspects that previous TWAS in breast cancer have not considered. This framework shows promise for scaling up into larger GWAS cohorts for further detection of risk- or outcome-associated loci.

## 3.1 The Carolina Breast Cancer Study

The Carolina Breast Cancer Study (CBCS) is a population-based study conducted in North Carolina that began in 1993. Study details and sampling schemes are described in previous CBCS work[122,123]. Patients of breast cancer aged between 20 and 74 years were identified using rapid case ascertainment in cooperation with the NC Central Cancer Registry, with self-identified African American and young women (ages 20-49) oversampled using randomized recruitment[122]. Randomized recruitment allows sample weighting to make inferences about the frequency of subtype in the NC source population. Details regarding patient recruitment and clinical data collections are described in Troester et al[24].

Date of death and cause of death were identified by linkage to the National Death Index. All diagnosed with breast cancer have been followed for vital status from diagnosis until date of death or date of last contact. Breast cancer-related deaths were classified as those that listed breast cancer (International Statistical Classification of Disease codes 174.9 and C-50.9) as the underlying cause of death on the death certificate. By the end of follow-up, we identified 674 deaths, 348 of which were due to breast cancer.

In total, we compiled 3,828 samples with 1,865 self-identified African American (AA) women and 1,963 self-identified white (WW) women from all phases of CBCS with relevant survival and clinical

variables. All 3,828 samples have associated germline genotype data, measured using the OncoArray genotyping assay developed by Illumina and the OncoArray Consortium[124]. This data was imputed using the October 2014 (v.3) release of the 1000 Genomes Project dataset as a reference panel using SHAPEIT2 for phasing and IMPUTEv2 for imputation[125–128]. SNPs with minor allele frequency (MAF) less than 1% and significant deviations from Hardy-Weinberg equilibrium at $P < 10^{-8}$ were excluded.

Of these 3,828 samples, we consider 1,199 (621 AA and 578 WW) samples with NanoString nCounter expression data for subsequent eQTL analysis and training of predictive expression models. Quality control and normalization was conducted as detailed in Sections 2.1 and 2.2.

## 3.2 eQTL analysis

Using the 1,199 samples (621 AA, 578 WW) with expression data, we assessed the additive relationship between the gene expression values and genotypes with linear regression analysis using MatrixeQTL[129], in the following model:

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

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

We calculated both cis- (variant-gene distance less than 500 kb) and trans-associations between variants and genes. Classical $P$-values were calculated for Wald-type tests of $H_0 : \beta_s = 0$ and were adjusted post-hoc via the Benjamini-Bogomolov hierarchical error control procedure, TreeQTL[130]. We conducted all eQTL analyses stratified by race. Age, BMI, postmenopausal status, and the first 5 principal components of the joint AA and WW genotype matrix were included in the models as covariates in $C$. Estimated tumor purity was also included as a covariate to assess its impact on strength and location of eQTLs. Any SNP found in an eQTL with Benajmini-Bogomolov adjust $P$-value $BBFDR < 0.05$ is defined as an eSNP. The corresponding gene in that eQTL is defined as an eGene. We exclude samples with Normal-like subtype, as classified by the PAM50 classifier, due to generally low tumor content. We developed a formal quality control procedure to follow-up on significant eQTLs by define a further MAF cutoff based on additive genotypes (i.e. 0,1, and 2 copies of the minor allele) and rigorous visual inspection.

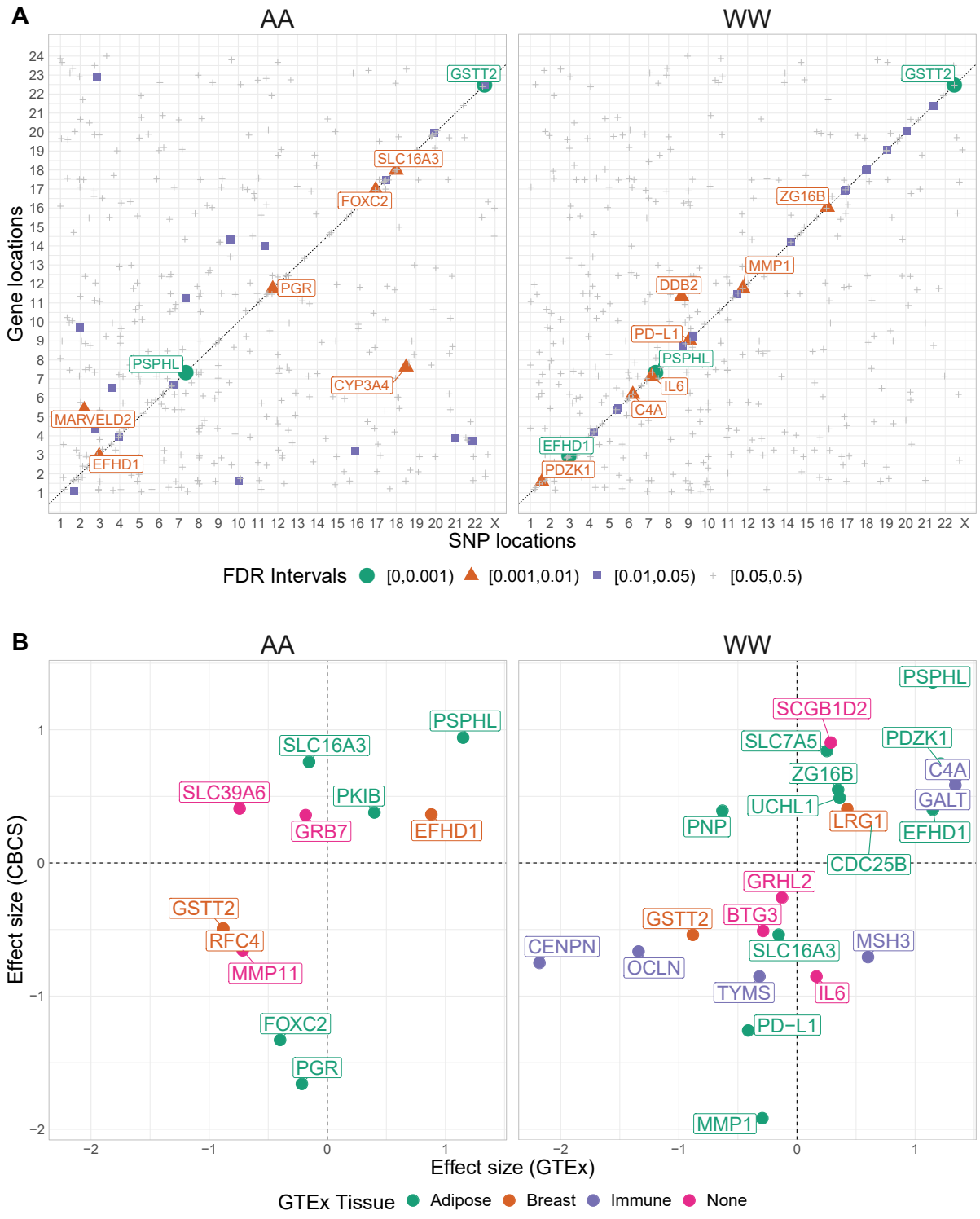


**Figure 3.1:** *CBCS eQTL results across race and compared with GTEx.* (A) Cis-trans plot of top eQTL by gene stratified by self-reported race. Each point represents the top eQTL for a given gene. The color and size of each point reflects the Benjamini-Bogomolov FDR-adjusted $P$-value (BBFDR) for that eQTL. eGenes with $BBFDR < 0.01$ are labelled. (B) Comparison of effect sizes of eGenes with significant cis-eQTLs in CBCS (Y-axis) and GTEx (X-axis) over tissue type, stratified by race. eGenes are colored by the GTEx tissue that shows the largest effect size. GTEx effect sizes on the X-axis are multiplied by the sign of the correlation between the genotypes of the GTEx and CBCS eSNPs.

At a Benjamini-Bogomolov[130] FDR-corrected $P$-value ($BBFDR < 0.05$) and after quality control, we identified 266 cis-eQTLs and 71 trans-eQTLs in the AA sample across 32 eGenes, and 691 cis-eQTLs and 15 trans-eQTLs in the WW sample across 24 eGenes. Of these eGenes, 4 are in common across race: *PSPHL*, *GSTT2*, *EFHD1*, and *SLC16A3*. Expressions of *PSPHL* and *GSTT2* have been previously reported to be governed by respective cis-deletions and serve as distinguishing biomarkers for race[131–134]. The majority of significant eQTLs in both the AA and WW samples were found in cis-association with respective eGenes. However, we saw a higher proportion of significant trans-eQTLs in the AA sample (**Supplemental Figure S12**). The locations and strengths of top eQTLs for all 406 autosomal genes are shown in **Figure 3.1A**, with minor allele frequencies of significant eSNPs plotted in **Supplemental Figure S13**.

### 3.2.1 Adjustment for tumor purity

Geeleher et al. show that only a third of conventional eQTLs in bulk breast cancer tumor expression could be attributed to cancer cells in TCGA[85]. We wished to assess the extent to which this observation bore out in CBCS. A study pathologist analyzed tumor microarrays (TMAs) from 176 of the 1,199 subjects to estimate area of dissections originating from epithelial tumor, assumed here as a proxy for the proportion of the bulk RNA expression attributed to the tumor. Using these 176 observations as a training set and the normalized gene expressions as the design matrix, we trained a support vector machine model tuned over a 10-fold cross-validation[135,136]. The cross-validated model was then used to estimate tumor purities for the remaining 1,023 samples from their gene expressions.

In general, we do not see significant differences in the strength and location of significant eQTLs, as shown in comparative cis-trans plots of all eQTLs across race and adjustment for tumor purity (**Supplemental Figures S13 and S14**). For most genes, top eQTL regions with linkage disequilibrium (LD) support had small differences in strength of effect size (Manhattan plot for a representative gene shown in **Supplemental Figure S13**). Adjusting for tumor purity, at $BBFDR < 0.05$ and after quality control, we identified 266 cis-eQTLs and 84 trans-eQTLs in the AA sample across 36 eGenes, and 634 cis-eQTLs and 14 trans-eQTLs in the WW sample across 23 eGenes, shown in **Supplemental Figure S14**. All WW eGenes, adjusting for tumor purity, are in common with WW eGenes from bulk tumor expression, and 32 of 36 AA eGenes, adjusting for tumor purity, are in common with AA eGenes from bulk tumor expression. Top eQTLs for eGenes remain largely the same across adjustment for tumor purity. Due to limited differences when we adjust for tumor purity, all downstream analyses do not involve our computational estimate of tumor purity.

We do not observe the same difference in eQTLs across adjustment for tumor purity as in Geeleher et al[85]. The NanoString expression data from CBCS includes only 417 genes, all of which were selected for the panel because of their involvement in breast cancer tumorigenesis, biology, or outcome disparities due to race. Furthermore, our normalization procedure involves the RUV method, which accounts for unwanted technical and biological variation, estimated from the distributions of housekeeping negative controls with an unsupervised method[27,29]. We hypothesize that the RUV method accounts for a significant percentage of the variability from cell-type heterogeneity that may confound traditional eQTL analysis in bulk tumor RNA expression. Further implementations of deconvolution algorithms specialized for expression measured for targeted panels of genes, as in NanoString, would aid in distinguishing the source cell types or tissues for various breast tumor eQTLs. Accurate bulk expression deconvolution may also be important in future TWAS to consider sources of variation in tumor expression due to tissue heterogeneity and how deconvoluted tumor expression signals contribute to outcomes of interest.

We lastly sought to evaluate the source of the significant eQTLs we detect in CBCS. Similarly to previous pan-cancer gerrmline eQTL analyses[137], we cross-referenced eGenes found in CBCS with eGenes detected in relevant healthy tissues from Genotype-Tissue Expression (GTEx) Project: mammary tissue (breast), subcutaneous adipose, and EBV-transformed lymphocytes (immune). We attributed all but 7 of the cis-eGenes from CBCS across both AA and WW women found in GTEx to one of these three tissue types (**Figure 3.1B**), with the effect sizes of the top eQTLs for these eGenes correlating very well between CBCS and GTEx (see **Supplemental Figure S17**). We also found adequate overlap of cis-eSNPs in these GTEx tissues and TCGA-BRCA based on the P-value of SNP-gene association (see **Supplemental Figure S18**). Note that, in GTEx v7, adipose ($N = 298$) has a larger sample size that mammary tissue ($N = 183$) and lymphocytes ($N = 114$). We were unable to replicate CBCS trans-eQTLs in GTEx and TCGA-BRCA[138]. The majority of CBCS trans-eQTLs were identified in AA women, and the sample sizes of individuals of African descent is low in GTEx version 7 and TCGA-BRCA.

### 3.2.2 Local ancestry adjustment of eQTLs

For cis-eGenes that were identified in only one of AA or WW women, we followed up with a cis-eQTL analysis adjusted for inferred local ancestry. Reference genotypes were downloaded from the 1000 Genomes Project version 3 for Utah residents with Northern and Western European ancestry (CEU) and Yoruban individuals from Ibadan, Nigeria (YRI)[125]. Phased genotypes from the assumed

admixed samples from CBCS were then compared to reference genotypes using RFMix v1.5.4 to estimate the posterior probability of CEU and YRI ancestry at a given haplotype, which is converted to an estimated dosage of inherited YRI alleles[139,140]. We then follow Zhong et al's framework for adjusting eQTLs by estimated local ancestry[141]. Briefly, for gene expression $g$, dosage of SNP of interest $s$, covariates $X_C$, and estimated local ancestry $l$ for the given SNP, we first residualize and scale to zero mean and unit variance $g$,$s$, and $l$ by $X_C$. We then fit the following linear model to estimate the local ancestry-adjusted eQTL effects:

$$\tilde{g} = \tilde{s} + \tilde{l} + \epsilon,$$

where $\tilde{g}$, $\tilde{s}$, and $\tilde{l}$ are the residualized and scaled gene expression, SNP dosage, and estimated local ancestry, respectively[141].

Overall, we find marginal increase in the strength of association between lead SNP and cis-eGene using an estimated local ancestry-adjustment over the association measured with a genome-wide ancestry adjustment. However, we did not observe considerable harmonization of stratified cis-eQTLs across populations; in general, race-specific, local ancestry-adjusted lead cis-eQTLs in a given race-stratified sample did not show similar association in the other (**Supplemental Figure S16**).

It has been shown that, due to allele frequency differences between populations, the underlying genetic and eQTL architecture for complex traits may not be well-correlated across diverse populations[142,143]. Zhong et al shows that incorporating local ancestry helps to better characterize the heritability of gene expression and complex traits and accurately map genetics associations[141]. However, our local ancestry-adjusted cis-eQTLs were not well-correlated across AA and WW women. Perhaps, the persistence of this difference can be due to the simplicity of the commonly used assumption that t here are two major source populations of admixture in CBCS samples (i.e. CEU and YRI). Several genetic studies into the genome-wide and local hereditary of admixed populations in the United States have shown that migratory patterns greatly inform these patterns of genetic ancestry[144,145]. Though this follow-up analysis is beyond the scope of this work, a full cis-trans eQTL ancestry incorporating local ancestry estimates, as well as an assessment of the impact of local ancestry adjustment on the portability of our eventual predictive models of tumor expression across ancestral populations, could reveal insights into the genetic architecture of breast tumor expression heritability in admixed populations.

## 3.3 Predictive models of tumor expression

### 3.3.1 Race-specific predictive models of tumor expression

Cis-heritability (cis-$h^2$) using genotypes within 500 kb of the gene of interest was estimated using the GREML-LDMS method, proposed to estimate heritability by correction for bias in linkage disequilibrium (LD) in estimated SNP-based heritability[49]. We do not consider the trans components in heritability estimation. Analysis was conducted using GCTA v.1.92[146]. Briefly, Yang et al shows that estimates of heritability are often biased if causal variants have a different minor allele frequency (MAF) spectrums or LD structures from variants used in analysis. They proposed an LD and MAF-stratified GREML analysis, where variants are stratified into groups by MAF and LD, and genetic relationship matrices (GRMs) from these variants in each group are jointly fit in a multi-component GREML analysis. Mean cis-$h^2$ of the 406 genes is 0.016 ($SE = 0.019$) in AA women and 0.015 ($SE = 0.019$) in WW women, as estimated by GREML-LDMS analysis[49]. For downstream analysis, we only consider genes with cis-$h^2$ significantly greater than 0 at a nominal $P$-value less than 0.10 from the relevant likelihood ratio test. Considering only these genes, the mean cis-$h^2$ of genes is 0.049 ($SE = 0.016$) in AA models and 0.052 ($SE = 0.016$) in WW models.

We adopt general techniques from PrediXcan and FUSION to estimate eQTL-effect sizes for predictive models of tumor expression from germline variants[42,3]. First, gene expressions were residualized for the covariates C included in the eQTL models (age, BMI, postmenopausal status, and genotype PCs) given the following ordinary least squares model:

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

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

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

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

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

We estimate $w_g$with the best predictive of three schemes: (1) elastic-net regularized regression with mixing parameter $\alpha = 0.5$ and $\lambda$ penalty parameter tuned over 5-fold cross-validation[42,3,48], (2) linear mixed modeling where the genotype matrix $X_g$ is treated as a matrix of random effects and $\hat{w}_g$ is taken as the best linear unbiased predictor (BLUP) of $w_g$, using rrBLUP[147], and (3) multivariate linear mixed modeling as described above, estimated using GEMMA v.0.97[148].

In these models, the genotype matrix $X_g$ is pruned for LD, prior to modeling using a window size of 50, step size of 5, and LD threshold of 0.5 using PLINK v.1.90b3[149] to account for redundancy in signal. We believe that our LD-pruning thresholds and window sizes are not stringent[150] and noticed that LD-pruning the design matrix of genotypes lead to greater cross-validation $R^2$ (**Supplemental Figure S19**). The final vectors $\hat{w}_g$ of effect-sizes for each gene $g$ are estimated by the estimation scheme with the best 5-fold cross-validation performance. All predicted models are stratified by race, i.e. an individual model of tumor expression for AA women and WW women for each gene g.

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

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

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

Of the predictive models built for these genes, 125 showed a five-fold cross-validation prediction performance (CV $R^2$) of at least 0.01 (10% Pearson correlation between predicted and observed expression with $P < 0.05$) in one of the two predictive models. **Figure 2.2A** shows the CV $R^2$ of these 153 genes across race. The median CV $R^2$ for the 153 genes was 0.011 in both AA and WW women. Cis-$h^2$ and CV $R^2$ are compared in **Supplemental Figure S20**. We also show mean CV and external validation (EV) $R^2$ with quantiles for prioritized genes across the training set and both external test sets in **Supplemental Table S2**.

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

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

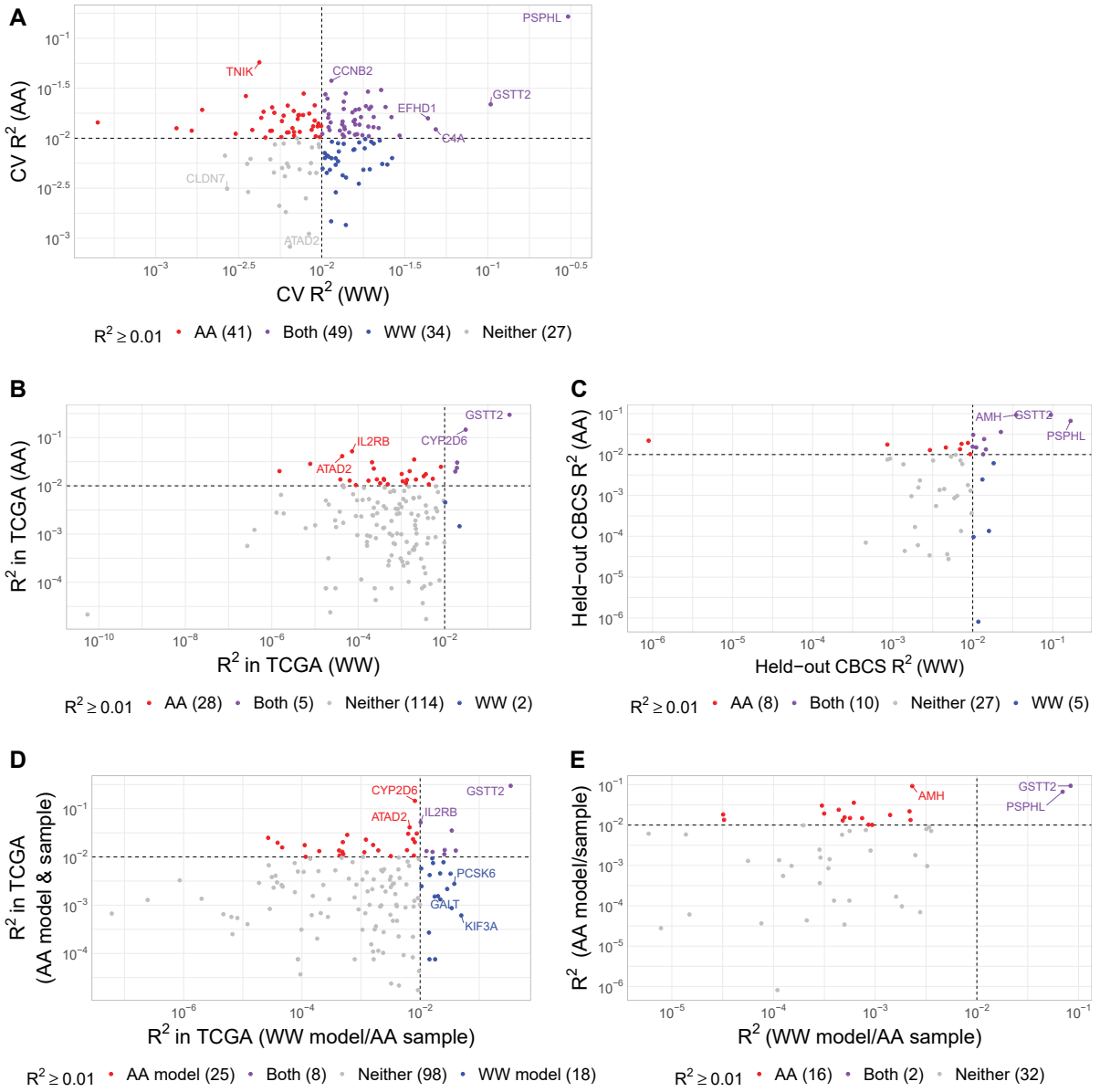


**Figure 3.2:** *Predictive performance of models in cross-validation, external validation, and across race.* (A) Comparison of cross-validation $R^2$ across race in CBCS. Cross-validation $R^2$ in CBCS WW women (X-axis) and CBCS AA women (Y-axis) for each of the 151 analyzed genes. Scales are logarithmic. Dotted lines represent $R^2 = 0.01$. Colors represent the model with which a given gene can be predicted at $R^2 > 0.01$. (B) Cross-validation $R^2$ in CBCS (X-axis) and square Spearman correlation between observed expression and GReX in TCGA-BRCA (Y-axis) in AA sample (left) and WW sample (right). Pearson correlations between $R^2$ calculated on the raw scale. $R^2$ are plotted on the log-scale. (C) Comparison of validation $R^2$ across race in TCGA for 149 analyzed genes found in TCGA expression data. (D) Comparison of validation $R^2$ across race in held-out CBCS samples for 50 analyzed genes. (E) Comparison of $R^2$ of genes in TCGA AA sample imputed from WW models (X-axis) and the AA models (Y-axis). (F) Comparison of $R^2$ of genes in held-out CBCS AA sample imputed from WW models (X-axis) and the AA models (Y-axis)

### 3.3.2 Evaluation of predictive models in independent data

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

The first sample is derived from TCGA breast tumor tissues with 179 AA and 735 WW women. We compared predictive performance by calculating an external validation $R^2$ (EV $R^2$) with squared Spearman correlations. Of the 151 genes modeled in CBCS training data with significant cis-$h^2$, 149 genes were measured via RNA-seq in TCGA. A comparison of predictive performance in TCGA for these 149 genes is shown in **Figure 3.2**, showing adequate performance in AA women (33 genes with EV $R^2 > 0.01$) and poor performance in WW women (7 genes with EV $R^2 > 0.01$). The top predicted gene in cross-validation from CBCS for both races, *PSPHL*, was not present in the TCGA normalized expression data and could not be validated. Another top cross-validated gene, *GSTT2*, was present in TCGA expression data and was validated as the top genetically predicted gene in TCGA by EV $R^2$.

We also imputed expression into entirely held-out samples from CBCS data (1,121 AA and 1,070 WW women) that have gene expression for a subset of the genes (166 of 417 genes) in the CBCS training set. These samples were largely derived from Phases I and II of CBCS. A comparison of imputation performance in CBCS for 50 genes (genes with significant cis-$h^2$ in CBCS training set) is shown in **Figure 3.2C**, showing adequate performance in both AA and WW women (18 and 15 genes with EV $R^2 > 0.01$ in AA and WW women).

Predictive models are not applicable across race We find that the predictive accuracy of most genes was lower when expression was imputed in AA women using models trained in the WW sample. We employed the WW predictive models to impute expression into AA samples from TCGA and held-out CBCS data. We compare the performances of the WW model and AA model in the AA sample in **Figure 3.2D** (TCGA) and 3.2E (CBCS). In held-out CBCS samples, with the WW model, we could only predict *PSPHL* and *GSTT2* at $R^2 > 0.01$ in the AA sample, as the expression of these genes is modulated mostly by strongly associated cis-eSNPs. In TCGA, our WW models performed adequately in AA women, though the WW models predicted fewer genes at $R^2 > 0.01$ than the AA models.

### 3.3.3 Evaluation of predictive performance across subtype

While predictive accuracy of expression models was stable across datasets, there was greater heterogeneity across biological and molecular subtype. In part, this is due to small sample sizes within

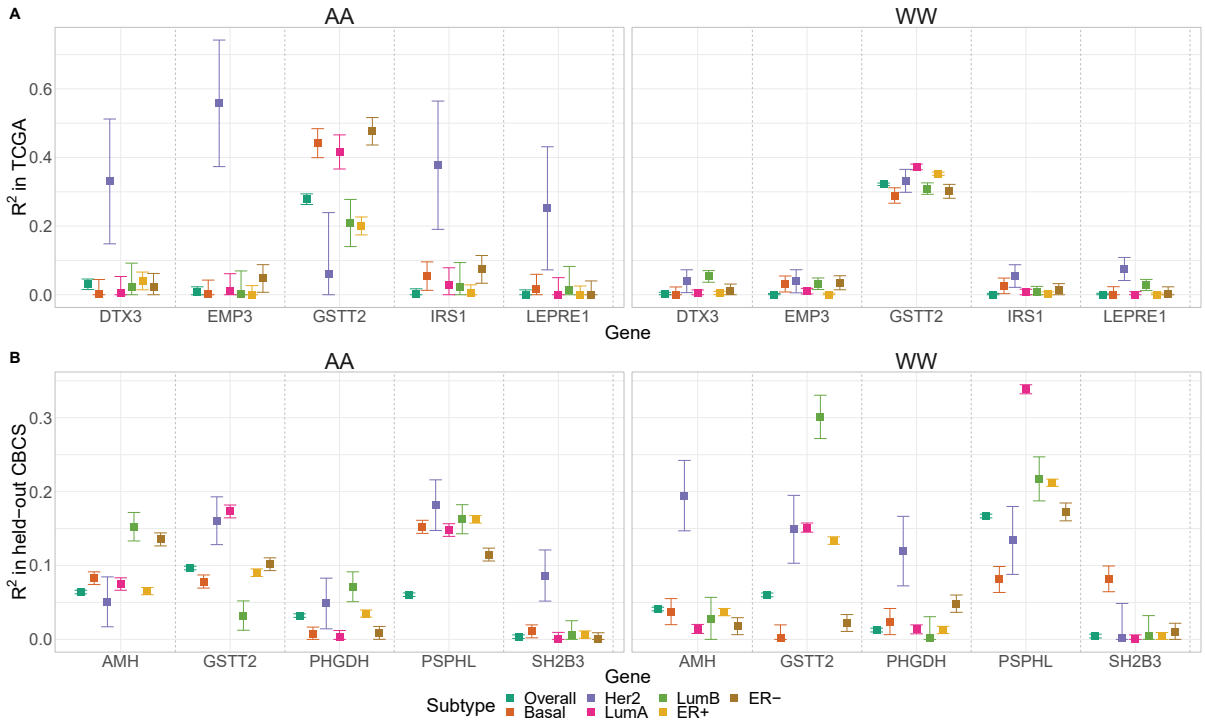


**Figure 3.3:** *Predictive performance of key genes, accounting for sampling variability.* Validation $R^2$ across PAM50 molecular subtype and estrogen receptor status, stratified by race, for example genes with highly variable $R^2$ in TCGA (A) and held-out CBCS (B). Squared Spearman correlation (Y-axis), denoted $R^2$, between observed and predicted gene expression is plotted for different genes (X-axis), stratfied by PAM50 subtype and estrogen receptor status. Points are colored and shaped according to subtype. Error bars provide 90% confidence intervals inverted from the corresponding permutation test.

race and subtype-specific strata. Upon first inspection, we see vast differences in the performance of our models across subtype (**Supplemental Figure S21**), with a large majority of genes performing at EV $R^2 > 0.01$ in rarer subtypes, like HER2-enriched breast cancers. However, we recognized sample sizes in the TCGA validation set were relatively small, especially when considering AA women and women of certain subtype, e.g. as low as 16 AA women with HER2-enriched breast cancer. As overall correlation between observed and imputed expressions are near 0, we sought to account for sampling variability when imputing into groups of women with such small sample sizes.

To account for sampling variability in calculating correlations in validation cohorts of smaller sample sizes, we calculated a permutation null distribution for each gene by permuting observed expressions 10,000 times and calculating a null prediction $R^2$ at each permutation. The sample validation prediction $R^2$ was compared to this permutation null distribution to generate an empirical $P$-value for the sample $R^2$, using Storey's qvalue package[151]. We then calculated $q$-values from these empirical $P$-values, controlling for a false discovery rate of 0.05[1,151]. Lastly, we constructed confidence intervals for $R^2$ by inverting the acceptance region from the permutation test[152].

**Supplemental Figure S22** displays $q$-values in Manhattan form[151], showing that the proportion of genes with EV $R^2$ significantly different from 0 is similar across subtypes. After inverting this permutation test to construct a confidence interval for EV $R^2$, we find that the EV $R^2$ of several genes are highly variable across subtypes, even when accounting for differences in sample size and therefore sampling variation. Key examples of such genes with variable EV $R^2$ across subtypes are shown in **Figure 3.3**.

## 3.4 Association with breast cancer-specific survival

### 3.4.1 Power analysis of detecting survival associations

Using survSNP[153], we generated the empirical power of a GWAS to detect various hazard ratios with 3,828 samples with 1,000 simulation replicates at a significance level of $P = 1.70 \times 10^{-8}$, corresponding to an FDR-adjusted $P = 0.10$. We assume an event rate of 10%, a relative allelic frequency of the risk allele of 0.1 and estimate the 90th percentile of times-to-event as a landmark time. Similarly, for genes of various cis-$h^2$, we assessed the power of TWAS to detect various hazard ratios at $P = 0.0096$ (corresponding to FDR-adjusted P = 0.10) over 1,000 simulation replications from the empirical distribution function of the GReX of the given gene. It is important to note that the detectable hazard ratios at 80% for GWAS and TWAS are incomparable due to differences in units of measure. At 80% power, a GWAS with CBCS data with N = 3,828 is powered to detect a hazard ratio of breast cancer-specific survival of 1.88 with an addition of one alternative allele in a given SNP. At 80% power, in our study, TWAS can detect hazard ratios 1.186, 1.203, and 1.216 with the GReX of a gene with cis-$h^2 \approx 0.100,\ 0.055,$ and $0.030$, with respect to an increase of one standard deviation, respectively (**Supplemental Figure S23**).

### 3.4.2 Predicted expression associated with breast cancer-specific survival

Here, we defined a relevant event as a death due to breast cancer. We aggregated all deaths not due to breast cancer as a competing risk. Any subjects lost to follow-up were treated as right-censored observations. We estimated the association of GReX with breast cancer survival by modeling the race-stratified cause-specific hazard function of breast cancer-specific mortality, stratifying on race[154]. For a given gene $g$, the model has form

| Region | Gene | Hazard Ratio (90% CI)[a] | $Z$-statistic[a] | $P$-value[a] | GReX $R^2$ ($h^2$)[b] |
|---|---|---|---|---|---|
| *20q13.2* | *AURKA* | 0.83 (0.73, 0.95) | -2.52 | $1.5 \times 10^{-3}$ | 0.021 (0.055) |
| *2p23.1* | *CAPN13* | 1.22 (1.07, 1.41) | 2.76 | $5.4 \times 10^{-4}$ | 0.011 (0.047) |
| *3q26.32* | *PIK3CA* | 0.85 (0.74, 0.97) | -2.34 | $3.2 \times 10^{-3}$ | 0.020 (0.033) |
| *18q21.33* | *SERPINB5* | 0.82 (0.72, 0.93) | -2.85 | $3.4 \times 10^{-4}$ | 0.010 (0.026) |

**Table 3.1:** Genes with GReX found in association with breast cancer-specific survival in AA women. (a) Hazard ratio and FDR-adjusted 90% confidence intervals, $Z$-statistic, and $P$-value of association of GReX with breast cancer-specific survival. (b) Cross-validation $R^2$ of gene expression in AA models.

$$\lambda_k(t) = \lambda_0(t) \exp\left\{GReX_g\beta_g + Z_C\beta_C\right\},$$

where $\beta_g$ is the effect size of $GReX_g$ on the hazard of breast cancer-specific mortality, $Z_C$ represents the matrix of covariates (age at diagnosis, estrogen-receptor status at diagnosis, tumor stage at diagnosis, and study phase), and $\beta_C$ are the effect sizes of these covariates on survival. $\lambda_k(t)$ is the hazard function specific to breast cancer mortality, and $\lambda_{0k}(t)$ is the baseline hazard function. We test $H_0 : \beta_g = 0$ for each gene $g$ with Wald-type tests, as in a traditional Cox proportional hazards model. We correct for genomic inflation and bias using bacon, a method that constructs an empirical null distribution using a Gibbs sampling algorithm by fitting a three-component normal mixture on $Z$-statistics from TWAS tests of association[155]. We control of multiple testing burden using the Benjamini-Hochberg procedure[1]. For comparison, we run a GWAS to analyze the association between germline SNPs and breast cancer-specific survival using GWASTools[156]. We use a similar cause-specific hazards model with the same covariates as in the TWAS models of association, correcting for false discovery with the Benjamini-Hochberg procedure.

Of the genes evaluated, we detected 4 whose GReX were associated with breast-cancer specific survival at FDR-adjusted $P < 0.10$ in AA women, shown in **Table 3.1** and **Figure 3.4**. We did not identify any genes with GReX associated with survival in WW women.

An association between increased GReX and increased risk of breast cancer-specific mortality was identified for *CAPN13* (*2p23.1*). We also found protective associations between higher GReX of *AURKA* (*20q13.2*), *PIK3CA* (*3q26.32*), *SERPINB5* (*18q21.33*) and lower risk of breast cancer-mortality (**Figure 3.4C**). Of these 4 loci, associations with survival have been reported with SNPs in near the same chromosomal region as *AURKA*, *PIK3CA*, and *SERPINB5*[37,157–161], though none of these

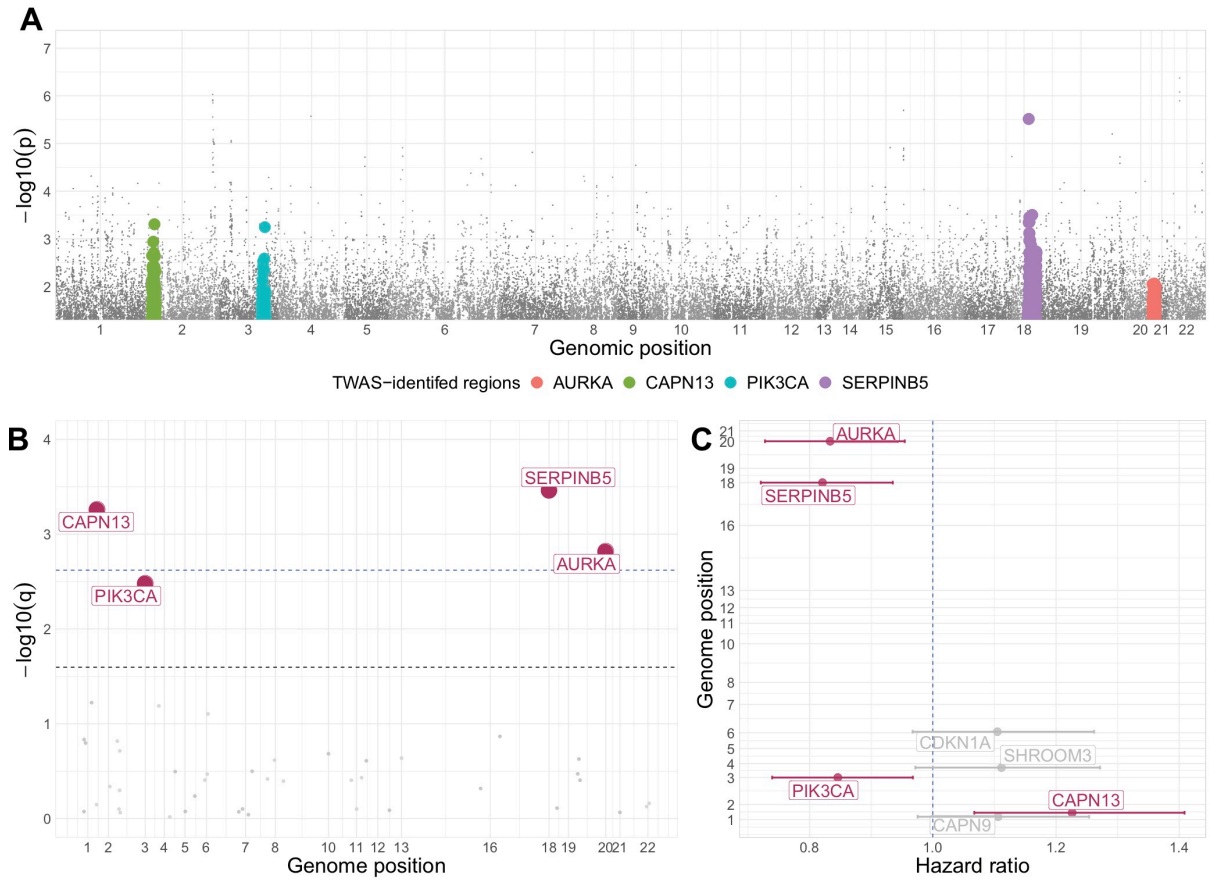


**Figure 3.4:** *GWAS and TWAS results in AA women.* (A) Manhattan plot of traditional GWAS on breast cancer survival. Genomic regions found to be significantly associated with survival in TWAS are represented in various colors. No SNVs reach Benjamini-Hochberg FDR-adjusted genome-wide significance. (B) Manhattan plot of TWAS on breast cancer survival. Genomic regions found to be significant at FDR-adjusted $P < 0.10$ are highlighted in red. The blue line represents a cutoff of FDR-adjusted $\alpha = 0.05$ and the dotted black line represents a cutoff of FDR-adjusted $\alpha = 0.10$. (C) Caterpillar plot of log-hazard rates with FDR-adjusted 90% confidence levels (X-axis) and genomic position (Y-axis). Results shown are significant at nominal $P < 0.10$. Genes highlighted in red represent genes with GReX significantly associated with survival at FDR-adjusted $P < 0.10$.

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

**Table 3.2:** Genes with GReX found in association with breast cancer-specific survival. (a) Top survival-associated SNP in cis-region of the given gene from GWAS for survival and distance of top cis-SNP from gene. (b) FDR-adjusted hazard ratio, 90% confidence interval, and $P$-value for association of GReX and breast cancer-specific survival, adjusting for adjacent survival-associated SNPs.

reported SNPs were utilized in constructing the GReX of this gene. Furthermore, the GReX of these four genes were not significantly correlated ($P > 0.05$ for all pairwise Spearman correlation tests), and the sets of SNPs used in constructing the GReX of these four genes had no pairwise intersections, providing evidence that their independent association with breast cancer-specific survival was not a pleiotropic effect from shared or correlated SNPs.

To determine whether the associations between predicted gene expression and breast cancer-specific survival were independent of GWAS-identified association signals, we performed conditional analyses adjusted for the most significant GWAS-identified survival-associated SNPs closest to the TWAS-identified gene by adjusting the cause-specific proportional hazards model for the genotype from this SNP. We found that the association for *PIK3CA* had a small change in effect size after adjustment for its adjacent survival-associated SNP, and its SNP-adjusted association was insignificant, while the other genes' associations remained significant after adjustment (**Table 3.2**). This conditional analysis suggests that the GReX of *AURKA*, *CAPN13*, and *SERPINB5* may be associated with breast cancer-specific survival independent of the GWAS-identified variant. No previously reported survival-associated SNPs were found significant at the genome-wide significance level in our dataset, and none of the closest survival-associated SNPs used in conditional adjustment were significant (**Figure 3.4A**). This supports our observation that correctly analyzed TWAS using relevant tissue gene expression may increase power for association testing.

As we deal with case-only data, we wished to inspect any collider bias that arises from unmeasured confounders that are associated with both breast cancer incidence and survival (see **Supplemental Figure S24**). Since a case-control dataset was not readily available to us to test associations between the GReX of genes with breast cancer risk, we construct the weighted burden test, as in FUSION[50,3], for the GReX of *AURKA*, *CAPN13*, *PIK3CA*, and *SERPINB5* in the GWAS

summary statistics for breast cancer risk in AA women available from BCAC using the iCOGs dataset and additional GWAS[9,10,162].

In summary, we compose a weighted Z test statistic[50,3] as follows:

$$\tilde{Z} = \frac{WZ}{(W\Sigma_{s,s}W')^{1/2}},$$

where $Z$ is the vector of $Z$-statistics from iCOGs and $W = \Sigma_{e,s}\Sigma_{s,s}^{-1}$ such that $\Sigma_{e,s}$ is the covariance matrix between all SNPs represented in $Z$ and the gene expression of the given gene and $\Sigma_{s,s}$ is the covariance among all SNPs. We find that none of the GReX of these genes are significantly associated with breast cancer incidence ($\tilde{Z} > 1.96, P < 0.05$), suggesting minimal presence of collider bias in our estimates of association with survival for the GReX of these four genes.

Lastly, we examined the association of the GReX of these four genes with breast cancer-specific survival in AA women, stratified by estrogen receptor (ER) subtype. We find that overall associations with survival are often driven by significant associations in a single subtype, though there is evidence of significant hazardous association in both ER subtypes for *CAPN13* (**Supplemental Figure S25**). We also did not detect a survival association with the total expression of these 4 genes, as estimated from breast cancer-specific Cox models (**Supplemental Figure S26**).

## 3.5 Discussion

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

Our race-stratified eQTL analysis reveals a strong cis-signal between germline variants and tumor expression of several genes, that is both differential across race and not exclusively attributable to healthy breast tissue. We also identified considerably more trans-eQTLs in the AA sample. This result may reinforce race differences in eQTL architecture as the ratio of detected trans-eQTLs to cis-eQTLs is not directly linked to sample size[44]. Differences in allele frequencies and linkage disequilibrium may contribute to observed differences in cis-eQTLs, as reported by Mogil et al[143], and we hypothesize that

such differences may likewise affect trans-eQTLs. Alternatively, there is a prevailing thought in literature about trans genetic regulation in admixed populations that the genetic diversity in individuals of African ancestry leads to added power of eQTL detection[68,72].

These race differences in eQTLs motivated the racial stratification of our predictive expression models[72,163]. Our models showed strong cross-validation predictive performance in genes with significant cis-heritability. We also show strong predictive performance in a held-out test set from CBCS and adequate performance of our WW models in TCGA-BRCA data. We noticed a difference in EV $R^2$ of our predictive expression models in held-out CBCS samples and TCGA-BRCA. We believe that this difference can be attributed partly to the difference in genotyping platform between the two samples (only approximately 85% of SNPs from CBCS represented in TCGA imputed genotype data). There could also be a lack of cis-heritability of the tumor expression of a majority of genes assayed in TCGA. For example, Gusev et al. has trained models for gene expression in breast tumors in TCGA; only 8 of the 417 genes in the CBCS NanoString panel showed significant cis-heritability in their models[3], which we downloaded from the Gusev Lab's TWAS/FUSION repository. We believe that predictive performance in TCGA data consistent with CBCS data is a high bar for validation due to both genotyping and RNA expression platform differences between CBCS (Oncoarray and NanoString) and TCGA (Affymetrix 6.0 and RNAseq). Reproducible performance in both AA and WW women in our independent test set from CBCS data suggests that our models are quite robust. Follow-up studies, in which models of tumor expression are trained in TCGA RNA-seq data and validated in CBCS NanoString data, could elucidate any discrepancies in predictive performance across platform.

An important implication of our work is the race-specificity of TWAS methods. We find that expression models trained in WW women generally have poor performance in AA women. Epidemiological studies have stressed accounting for differences in race by stratification or adjustment for admixture estimates when constructing polygenic scores[164]. Our observations suggest that this epidemiological note of caution extends to creating predictive models for RNA expression. Previous TWAS studies of breast cancer risk have either used models trained in a sample of predominantly European ancestries[165] or imputed into large cohorts of strictly patients of European descent[43]. Hoffman et al. excludes SNPs that were monomorphic in any of the 14 different ancestral populations they analyze[165], though this may not capture all effects of ancestry on genetic regulation of expression, including the possibility for interactions. We contend that accounting for ancestry or stratifying by race may be necessary to draw correct inference in large, ancestrally-heterogeneous cohorts.

Our data also suggests that predictive performance may vary by molecular subtype. Previous groups have shown the predictive utility of catering polygenic risk scores to breast cancer

subtype[166,167], a phenomenon we investigated in our predictive models of tumor expression. Even after accounting for sampling variability in prediction, we found that several genes have varied degrees of GReX across subtype and race. Not only does this finding suggest that TWAS predictive models may need to account for subtype heterogeneity, we reinforce the importance of sampling variability in validation of predictive models in external cohorts. For example, Wu et al. trained their models in a relatively small set of 67 women from GTEx and validated their 12,824 models in a validation set of 86 women from TCGA without accounting for sampling variability of predictive performance[43]. A recent multi-tissue TWAS in ovarian cancer from Gusev et al. considered validation of their predictive models by leveraging multiple independent cohorts to assess replication rates[168]. We recommend such an approach if multiple independent cohorts are accessible. But, in TWAS evaluation in a single tissue, studies should place a strong emphasis on validation, accounting for sampling variability of prediction $R^2$ prior to imputation in larger cohorts.

While many of the most significant findings here are methodological in nature, we also have data to suggest that four genomic loci in AA women may merit further investigation relative to breast cancer survival. Two of these 4 TWAS-identified genes have strong functional evidence in breast cancer survival literature. Mutations in *AURKA* and *PIK3CA* have previously been shown to be significantly associated with breast cancer survival rates[157–159]. Less is known about the involvement of *SERPINB5* and *CAPN13* in breast cancer survival, though they have been identified in studies into breast cancer progression[169–173]. These four loci merit further studies for validation and functional characterization, both in large GWAS cohorts and using *in vitro* studies. We did not observe any significant association between the total expression of these 4 genes and breast cancer-specific survival. This suggests that the germline-regulated component of the tumor expression of these genes – a small fraction of the total expression variation – may be associated with survival outcomes. Numerous factors, including copy number alterations, epigenetic or post-transcriptional regulation, and exposures and technical artifacts in measurement contributed to the total expression measured in the tumor. Thus, we do not expect that significant GReX association implies total expression association, or vice versa.

We also observed that 3 of the 4 associations were driven by very strong effect sizes within a single subtype. Though we cannot contextualize this result, it highlights an often-overlooked modeling consideration. In a cohort that is both biologically and ancestrally-heterogeneous, as in CBCS, investigators should consider modeling choices beyond simple linear adjustments for subtype and race. Akin to the logic of Begg et al and Martínez et al[174,175], it may be prudent in future TWAS to stratify predictive models on both race and biological subtype to increase power to detect outcome-associated loci that are strongly present within only one such strata or have heterogeneous effects across strata.

Since the CBCS analysis was a case-only study, we were wary of potential collider bias by unmeasured confounders associated with both breast cancer risk and progression[176,174,175,177], which may affect the effect sizes of association between survival and GReX of genes. None of the GReX of these four genes showed significant transcriptome-wide associations with breast cancer risk in iCOGs data[9,10,162], suggesting that our estimates of association may be free of the collider bias. As Escala-García et al. highlights, germline variation can affect breast cancer prognosis via tumor etiology (risk of developing a tumor of a certain subtype), or via mechanisms that are relevant post-tumorigenesis, such as the cellular response to therapy or the host-tumor micro-environment[36]. Ideally, in future TWAS and integrated omic analyses of breast cancer survival, it is prudent to consider joint models of breast cancer risk and survival to account for pleiotropic effects of germline genotype and any associations with unmeasurable confounders[178].

One limitation of our study is that data on somatic amplifications and deletions were not yet available for the CBCS cohort we analyzed. Removing the somatic copy number variation signal from tumor expression profiles may improve our estimates of cis-heritability and perhaps the predictive performance of our models, though previous TWAS in ovarian cancer shows the effect to be qualitatively small (approximately less than 2% change in heritability)[168]. Furthermore, not all genes in the CBCS NanoString panel have a significant heritable component in expression regulation. These genes, like *ESR1*, which have a significant role in breast cancer etiology[179], could not be investigated in our study. Lastly, since CBCS mRNA expression is assayed by the NanoString nCounter system, we could only analyze 94 aggregated locations on the human transcriptome across race. However, the NanoString platform allows the CBCS to robustly measure expression from FFPE samples on a targeted panel of breast cancer and race-related genes, allowing us to leverage the large sample size from all three phases of the CBCS. One of the greatest strengths of our study is that the CBCS affords us both a large training and test set of AA and WW women for race-stratified predictive models. Such data is important in drawing inference in more ancestrally-heterogeneous populations. Accordingly, the statistical power of our study is high to detect associations for genes with relatively high cis-heritability. Future studies in large GWAS cohorts, such as those within the Breast Cancer Association Consortium, will elucidate how to account for ancestral and biological heterogeneity in detecting survival-associated loci.

We have provided a framework of transcriptome-wide association studies (TWAS) for breast cancer outcomes in diverse study populations, considering both ancestral and subtype-dependent biological heterogeneity in our predictive models. From a more theoretical perspective, this work will inform the utilization of TWAS methods in polygenic traits and diverse study populations, stressing

rigorous validation of predictive models prior to imputation and careful modeling to capture associations with outcomes of interest in diverse populations.

# CHAPTER 4: MULTI-OMIC STRATEGIES FOR TRANSCRIPTOME-WIDE ASSOCIATION STUDIES

In this chapter, we outline two extensions to TWAS that draw from ideas of eQTL mediation and borrowing information from other omics assays. The first extension works backwards from gene expression by identifying associated, mediating biomarkers (e.g. DNA methylation at relevant loci, or expression levels of microRNAs and transcription factors) to the gene of interest. We train prediction models for these mediators using their local SNPs and incorporate their predicted values as fixed effects in the eventual model of gene expression. The second extension uses mediation analysis to identify distal eQTLs that show large total mediation effects through local mediators. These prioritized distal SNPs are upweighted in the eventual gene expression model. Using simulations and data from the The Cancer Genome Atlas (TCGA)[180] and Religious Orders Study and the Rush Memory and Aging Project (ROS/MAP)[181], we show improvements in both in-sample and out-of-sample predictive performance and power to detect gene-trait associations over local-only models. These **M**ulti-**O**mic **S**trategies for **T**ranscriptome-**W**ide **A**ssociation **S**tudies are made available in the R package MOSTWAS, available freely at `www.github.com/bhattacharya-a-bt/MOSTWAS`.

## 4.1 Overview of MOSTWAS

We first outline the two methods proposed in MOSTWAS: (1) mediator-enriched transcriptome-wide prediction (MeTWAS) and (2) distal eQTL prioritization via mediation analysis (DePMA). In MOSTWAS, we define that two biological objects (e.g. genetic variant, gene, microRNA, or CpG site) are *local* to one another if the genomic distance between them is less than or equal to 0.5 Megabases (Mb). Otherwise, we define the two objects as *distal*. We adopt *local* and *distal*, rather than *cis* and *trans*, to avoid any confusion with biological mechanism.

### 4.1.1 Heritability estimation

Prior to any predictive modeling, we estimate the heritability of a gene of interest using GCTA v.1.92[146] using all local and distal SNPs considered in either MeTWAS or DePMA. MOSTWAS allows

the user the capability to employ the GREML-LDMS method[49] to estimate heritability in imputed genotype panels. MOSTWAS will only proceed to predictive modeling if the gene is heritable from the specified local and distal SNPs at a user-defined $P$-value threshold (default $P < 0.10$ for the relevant likelihood ratio test).

### 4.1.2 Mediator-enriched TWAS (MeTWAS)

#### 4.1.2.1 Transcriptomic prediction using MeTWAS

We first describe mediator-enriched TWAS, or MeTWAS, one of the two tools available in the MOSTWAS R package. Across $n$ individuals, consider the vector $Y_G$ of expression a gene $G$ of interest, the matrix $\mathbf{X}_G$ of local-SNP dosages in a 0.5 Mb window around gene $G$, and $m_G$ mediating biomarkers that are estimated to be significantly associated with the expression of gene $G$ via a relevant one-way test of association. These mediating biomarkers could be, for example, DNA methylation sites, microRNAs, or transcription factors. Accordingly, let the matrix $\mathbf{X}_{M_j}$ be the local-SNP dosages in a 500 kilobase (kb) window around mediator $j,\ 1 \leq j \leq m_G$. Furthermore, let $M_j$ be the intensity of mediator $j$ (i.e. methylation $M$-value if $j$ is a CpG site or log scale expression if $j$ is an miRNA or a gene). Prior to any modelling, we scale $Y_G$ and all $M_j,\ 1 \leq j \leq m_G$ to zero mean and unit variance. We also residualize $M_j,\ 1 \leq j \leq m_G$ and $Y_G$ with the covariate matrix $\mathbf{X}_C$ to account for population stratification using principal components of the global genotype matrix and relevant clinical covariates to obtain $\tilde{M}_j\ 1 \leq j \leq m_G$ and $\tilde{E}_G$. The number of genotype principal components included is user-defined and dependent on the dataset.

Transcriptome prediction in MeTWAS draws from two-step regression, as summarized in **Figure 4.1A**. First, in the training set for a given training-test split, for $1 \leq j \leq m_G$, we model the residualized intensity $\tilde{M}_j$ of training-set specific mediator $j$ with the following additive model:

$$\tilde{M}_j = \mathbf{X}_{M_j,\text{train}} w_j + \epsilon_m, \tag{4.1}$$

where $w_j$ is the effect-sizes of the SNPs in $\mathbf{X}_{M_j,\text{train}}$ on $\tilde{M}_j$ in the training set. As in traditional transcriptomic imputation models[42,3], we find $\hat{w}_m$ using the method that best predicts expression out of the following methods: (1) elastic net regression with mixing parameter $\alpha = 0.5$ and $\lambda$ tuned over 5-fold cross validation using glmnet[48], or (2) linear mixed modelling assuming random effects for $\mathbf{X}_{M_j}$ using rrBLUP[147].

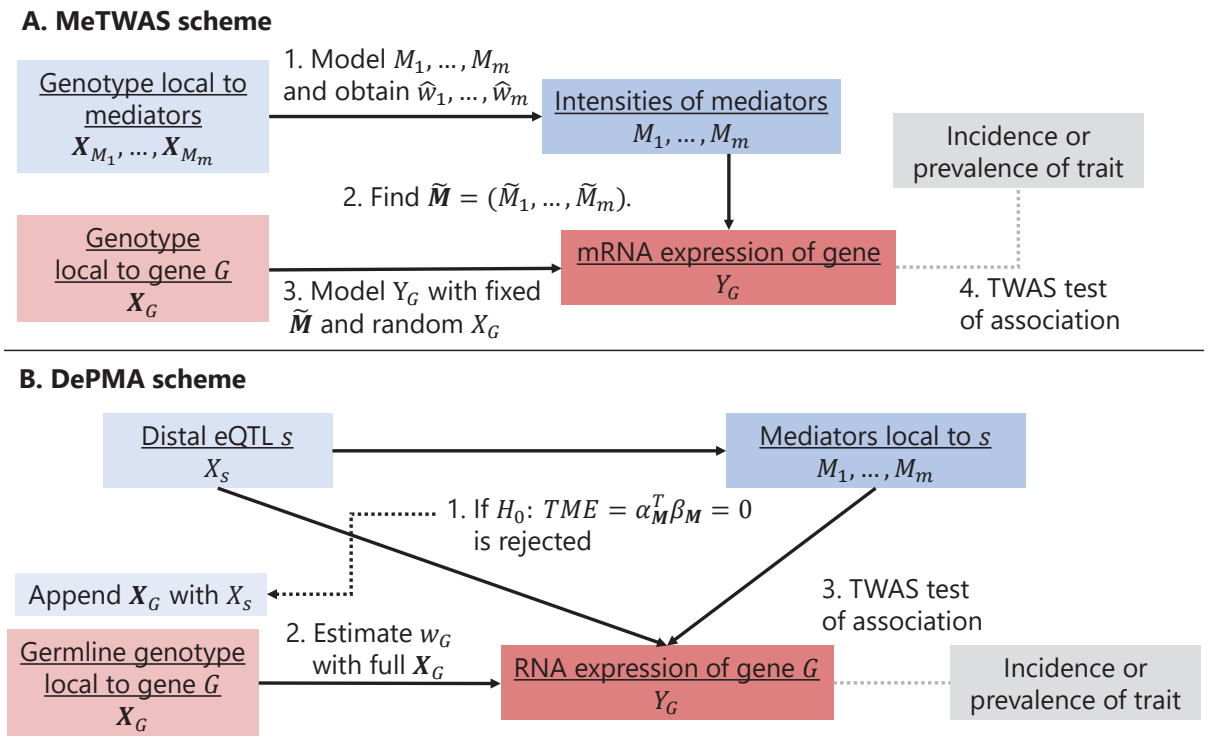


**Figure 4.1:** *Modeling schemes for MOSTWAS.* (A) Two-step regression scheme in MeTWAS that enriches transcriptomic prediction with mediating-biomarkers. (B) Mediation analysis based DePMA procedure to prioritize distal-eQTLs with large total mediation effects for transcriptomic prediction.

For all $j$, using these optimized predictive models for $M_j$ as denoted by $\hat{w}_{M_j}$, we estimate the genetically regulated intensity (GRIn) of the mediator $m_j$, denoted $M_{m_j}$, in the test set. Denote $\hat{\mathbf{M}}_{n\times m}$ as the matrix of estimated GRIn, such that the $j$th column of $\hat{\mathbf{M}}_j$ is $M_{m_j}$ across all $n$ samples.

Next, we consider the following additive model for the residualized expression of gene $G$:

$$\tilde{Y}_G = \hat{\mathbf{M}}\beta_M + \mathbf{X}_G w_G + \epsilon_{Y_G},$$

where $\beta_M$ is the fixed effect-sizes of $M_{m_j}$ on $\tilde{Y}_G$, $\hat{\mathbf{M}}$ is the matrix of estimated GRIn for all $m_j$ mediators, $\mathbf{X}_G$ are the local-genotypes to gene $G$, and $w_G$ are the "random" or regularized effect sizes of the local-genotypes. We estimate $\beta_M$ by traditional ordinary least squares, where $\hat{\beta}_M = \left(\hat{\mathbf{M}}^T\hat{\mathbf{M}}\right)^{-1}\hat{\mathbf{M}}^T\tilde{Y}_G$. Next, using one of the methods outlined above when estimating $\hat{w}_{M_j}$, we can generate estimated effect sizes $\hat{w}_G$ of the local-genotypes on $\tilde{Y}_G$, residualized with $\hat{\mathbf{M}}$.

#### 4.1.2.2 Transcriptomic imputation with MeTWAS

In an external GWAS panel, if individual genotypes are available, we construct the genetically regulated expression (GReX) of gene $G$ directly using $\hat{w}_G$ and $(\hat{w}_j, \hat{\beta}_j),\ 1 \leq j \leq m_G$:

$$\text{GReX}_G = \sum_{j=two1}^{m_G} \mathbf{X}_{M_j,\text{GWAS}} \hat{w}_{M_j} \hat{\beta}_{M,j} + \mathbf{X}_{G,\text{GWAS}} \hat{w}_G,$$

where $\mathbf{X}_{M_j,\text{GWAS}}$ and $\mathbf{X}_{G,\text{GWAS}}$ are the genotypes in the GWAS panel local to mediator $j$ and gene $G$, respectively. $\text{GReX}_G$ can be used in downstream tests of association.

If individual genotypes are not available, then the weighted burden $Z$-test proposed by Pasaniuc et al and Gusev et al can be employed[50,3] using summary statistics. Briefly, we compute

$$\tilde{Z} = \frac{\mathbf{W}Z}{(\mathbf{W}\Sigma_{s,s}\mathbf{W}^T)^{1/2}} \tag{4.2}$$

Here, $Z$ is the vector of $Z$-scores of SNP-trait associations for SNPs used in estimating $\hat{w}_{M_j}$ and $\hat{w}_G$. The matrix $W$ is defined as $\Sigma_{e,s}\Sigma_{s,s}^{-1}$, the product of the covariance matrix between all SNPs and the expression of gene $G$ and the covariance matrix among all SNPs. These covariance matrices are estimated from the reference panel used to estimate $\hat{w}_{M_j}$ and $\hat{w}_G$. The test statistic $\tilde{Z}$ can be compared to the standard Normal distribution for inference.

MOSTWAS also implements permutation testing to quantify the significance of the expression-trait association conditioning on the SNP-trait effects at the locus[3]. Here, we perform 1,000 permutations of the SNP-expression weights in the predictive model and compute the $Z$-test statistic at each permutation. A permutation $P$-value is calculated by comparing $\tilde{Z}$ to the distribution of permuted $Z$-test statistics.

### 4.1.3 Distal eQTL prioritization via mediation analysis (DePMA)

#### 4.1.3.1 Transcriptomic prediction using DePMA

We now describe Distal eQTL Prioritization via Mediation Analysis (DePMA), the second of two tools available in MOSTWAS. Expression prediction in DePMA hinges on up-weighting distal eQTLs to the gene of interest via mediation analysis, adopting methods from previous studies[67,69,72]. This process is summarized in **Figure 4.1B**. We first split data for gene expression, SNP dosages, and any

potential mediators into $k$ training-testing splits. Depending on the minor allele frequencies of SNPs and sample size, we generally recommend a low number of splits (i.e. $k \leq 5$).

In the training set, we identify mediation test triplets that consist of (1) a gene of interest $G$ with expression $Y_G$ (scaled to zero mean and unit variance), (2) a distal eSNP $s$ in association with $G$ at a user-defined $P$-value threshold (default of $P = 10^{-6}$) with dosages $X_s$, and (3) a set of $m$ biomarkers local to $s$ that are associated with $s$ at a user-defined $P$-value threshold (default of FDR-adjusted $P = 0.05$) with intensities as $m$ columns of $\mathbf{M}_{n\times m}$. The columns of **M** are scaled to zero mean and unit variance. Consider the following mediation model for $1 \leq j \leq m$:

$$\begin{aligned} Y_G &= X_s\beta_s + \mathbf{M}\beta_{\mathbf{M}} + \mathbf{X}_C\beta_C + \epsilon_{Y_G} \\ M_j &= X_s\alpha_{M_j} + \mathbf{X}_C\alpha_{C,j} + \epsilon_{M_j} \end{aligned} \tag{4.3}$$

Here, we have $\beta_{\mathbf{M}}$ as the effects of the $M$ mediators local to $s$ on $Y_G$ adjusting for the effects from $s$ and the covariates and $\alpha_{\mathbf{M}} = (\alpha_{M_1}, \ldots, \alpha_{M_m})^T$ as the effects of $s$ on mediators $M_j$, for $1 \leq j \leq m$. We assume that $\epsilon_{Y_G} \sim N(0, \sigma^2)$ and $\epsilon_{\mathbf{M}} \sim \mathbf{N}_m\,(0, \square_M)$, where $\square_M$ may have non-zero off-diagonal elements that represent covariance between mediator intensities. Further, we assume that $\epsilon_{Y_G}$ and $\epsilon_{\mathbf{M}}$ are independent. We define the total mediation effect (TME)[182] of SNP $s$ as

$$\text{TME} = \alpha_{\mathbf{M}}^T\beta_{\mathbf{M}}.$$

We are interested in SNPs with large TME, which we prioritize with the test of $H_0$ : TME $= 0$. We assess this hypothesis with a permutation test, as more direct methods of computing standard errors for the estimated TME are often biased[71,72], obtaining a permutation $P$-value. We also provide an option to estimate an asymptotic approximation to the standard error of TME and conduct a Wald-type test for TME $= 0$. This asymptotic option is significantly faster at the cost of inflated false positives. Corresponding to the $t$ testing triplets identified, we obtain vectors of length $t$ of TMEs and $P$-values for each distal eSNP to $G$. For the predictive model, we select distal SNPs with evidence of TME $\neq 0$ at a given $q$-value threshold ($q < 0.10$ as a default) and include them with all local genotypes in a design matrix. We then find estimated SNP weights using either elastic net or weighted least squared regression.

#### 4.1.3.2 Asymptotic test of total mediation effect

In DePMA, a distal-eQTL $s$ is tested for its total mediation effect on gene $G$ through $m$ mediators that are local to $s$. Consider the following mediation model for $1 \leq j \leq m$:

$$\begin{aligned} Y_G &= X_s\beta_s + \mathbf{M}\beta_{\mathbf{M}} + \mathbf{X}_C\beta_C + \epsilon_{Y_G} \\ M_j &= X_s\alpha_{M_j} + \mathbf{X}_C\alpha_{C,j} + \epsilon_{M_j} \end{aligned} \tag{4.4}$$

.

Here, we construct the total mediation effect

$$\text{TME} = \alpha_{\mathbf{M}}^T\beta_{\mathbf{M}} = \sum_{i=1}^{m} \alpha_{M_i}\beta_{M_i}.$$

Note that TME is distributed as the product of two multivariate Normal distributions. By the multivariate Delta method[183], we can obtain the standard error for the estimated TME. Let $\boldsymbol{\theta} = (\alpha_{\mathbf{M}}, \beta_{\mathbf{M}})$ and define $f(\boldsymbol{\theta}) = TME = \sum_{i=1}^{m} \alpha_{M_i}\beta_{M_i}$.

The first order partial derivative of $f(\hat{\boldsymbol{\theta}})$ is

$$d_{\hat{\boldsymbol{\theta}}} = \frac{\partial(\sum_{i=1}^{m} \alpha_{M_i}\beta_{M_i})}{\partial\hat{\boldsymbol{\theta}}} = [\beta_{\mathbf{M}}\ \alpha_{\mathbf{M}}]^T.$$

We also obtain the estimated variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$ of $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{\alpha_{\mathbf{M}}} & \hat{\boldsymbol{\Sigma}}_{\alpha_{\mathbf{M}}\beta_{\mathbf{M}}} \\ \hat{\boldsymbol{\Sigma}}_{\alpha_{\mathbf{M}}\beta_{\mathbf{M}}} & \hat{\boldsymbol{\Sigma}}_{\beta_{\mathbf{M}}} \end{bmatrix},$$

where $\hat{\boldsymbol{\Sigma}}_{\alpha_{\mathbf{M}}}, \hat{\boldsymbol{\Sigma}}_{\beta_{\mathbf{M}}},$ and $\hat{\boldsymbol{\Sigma}}_{\alpha_{\mathbf{M}}\beta_{\mathbf{M}}}$ are the variances and covariance of $\hat{\alpha}_{\mathbf{M}}$, $\hat{\beta}_{\mathbf{M}}$, and between $\hat{\alpha}_{\mathbf{M}}$ and $\hat{\beta}_{\mathbf{M}}$, respectively. Sobel previously has shown, that with sufficient sample size, $\hat{\boldsymbol{\Sigma}}_{\alpha_{\mathbf{M}}\beta_{\mathbf{M}}} \approx 0$[182,184]. Thus, the standard error of $\hat{\boldsymbol{\theta}}$ is given by

$$\hat{\sigma}^2_{\hat{\boldsymbol{\theta}}} = d^T_{\hat{\boldsymbol{\theta}}}\hat{\boldsymbol{\Sigma}}d_{\hat{\boldsymbol{\theta}}}.$$

We then test $H_0$ : TME $= 0$ against $H_1$ : TME $\neq 0$ with the two-sided Wald-type test with the test statistic $Z = \frac{\alpha_{\mathbf{M}}^T\beta_{\mathbf{M}}}{\sqrt{\hat{\sigma}^2_{\hat{\boldsymbol{\theta}}}}}$ and comparing to the null standard Normal distribution.

We illustrate the trade-off between power and computational speed using the asymptotic Sobel test and the permutation speed. Consider the following simulation framework with $m = 5$ mediators, 3 covariates and a sample size of $n \in \{200, 500, 700, 1000\}$ for the model in Equations 4.4:

- an $n$-length genotype vector for SNP $s$ is drawn from Binomial$(2, MAF)$, where the minor allele frequency $MAF$ is set at 0.1 in **Figure 4.2** below;
- Under the alternative, we simulated $\beta_X \sim N(0, 1)$, $\beta_{\mathbf{M}} \sim \mathbf{N}_5(\mathbf{0}, \mathbf{I}_5)$, $\beta_C \sim \mathbf{N}_3(\mathbf{0}, \mathbf{I}_3)$, $\alpha_{M_j}|_{j=1}^{m=5} \sim N(0, 1)$, $\alpha_C \sim \mathbf{N}_5(\mathbf{0}, \mathbf{I}_5)$.
- Under the null, all regression parameters were simulated as in the alternative case. However, we set $\alpha_{M_j} = 0|_{j=1}^{m}$ and $\beta_{\mathbf{M}} = \mathbf{0}$.
- Lastly, $\epsilon_{Y_G} \sim N(0, 1 - h^2)$ and $\epsilon_{M_j} \sim N(0, 1 - h_M^2)$, where $h^2 = h_M^2 = 0.1$ in **Figure 4.2** below.
- We then constructed $Y_G$ and **M** using Equations 4.4.

We found, that over 10,000 simulations, the permutation test was considerably more powerful, albeit considerably slower. However, in most cases of implementing DePMA, the number of tests of mediation are usually on the order of $10^1$ to $10^2$. We recommend the permutation test in most cases, unless gene $G$ has thousands of identified distal-eQTLs. Parallel implementations have been offered as options in the MOSTWAS package.

#### 4.1.3.3 Transcriptomic imputation with DePMA

In an external GWAS panel, if individual genotypes are available, we construct the genetically regulated expression (GReX) of gene $G$ directly using $\hat{w}_G$ and $\hat{w}_t$:

$$\text{GReX}_G = \mathbf{X}_{t,\text{GWAS}}\hat{w}_t + \mathbf{X}_{G,\text{GWAS}}\hat{w}_G,$$

where $\mathbf{X}_{t,\text{GWAS}}$ is the matrix of dosages of the $t$ distal SNPs and $\mathbf{X}_{G,\text{GWAS}}$ is the matrix of dosages of the local SNPs to gene $G$ in the external GWAS panel. $\text{GREX}_G$ can be used in downstream tests of association. If individual genotypes are not available, the weighted burden test can be employed using summary statistics[3].

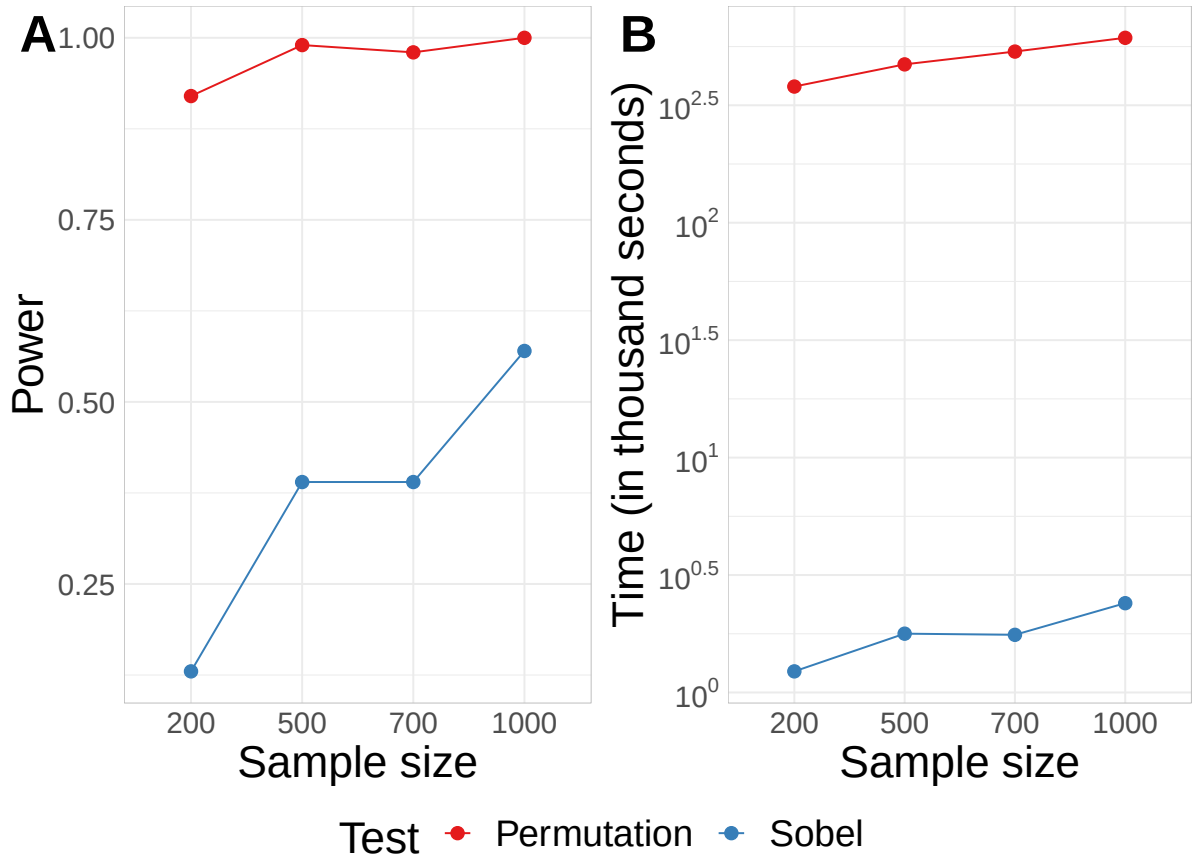


**Figure 4.2:** *Comparison of power and computational speed comparison of permutation and Sobel test.* Power (A) and computational speed (B) of permutation test (red) and asymptotic Sobel test (blue) in simulation framework

#### 4.1.4 Added-last test of association from distal variants

In addition to the overall TWAS tests of association and permutation testing, as implemented by Pasaniuc et al and Gusev et al[50,3], we develop here a technique to assess whether the distal loci included in the predictive models are significantly associated with the phenotype of interest, given the association at the local locus. In the scenario that individual genotype data is available, we can simply run a group added-last test in the linear or survival model employed to assess the TWAS association with phenotype. We use similar logic to develop an added-last test for distal variants conditional on the local association, when only GWAS summary statistics are available.

Let $\mathbf{Z}_l$ (an $n_l$-vector) and $\mathbf{Z}_d$ (an $n_d$-vector) be the $Z$-scores local and distal SNPs identified by a MOSTWAS model, with $\mathbf{Z} = [\mathbf{Z}_l\ \mathbf{Z}_d]^T$ (an $n$ vector). The local and distal SNP effects from the MOSTWAS model are represented in $\mathbf{w}_l$ (an $n_l$-vector) and $\mathbf{w}_d$ (an $n_d$-vector), with $\mathbf{w} = [\mathbf{w}_l\ \mathbf{w}_d]^T$ (an $n$ vector). Here, we are interested in testing

$$H_0 : \mathbf{w}_d^T \mathbf{Z}_d | \mathbf{w}_l^T \mathbf{Z}_l = \tilde{Z}_{l,\text{obs}} = 0,$$

where $\tilde{Z}_{l,\text{obs}}$ is the observed weighted $Z$-score from local SNPs.

Under the null distribution, as proposed by Pasaniuc et al and Gusev et al in the Imp-G framework[50,3], we assume that $Z \sim N_n(\mathbf{0}, \square)$, where

$$\square = \begin{bmatrix} \square_l & \square_{l,d} \\ \square_{l,d}^T & \square_d \end{bmatrix}$$

is the LD matrix for the SNPs, as estimated from the reference panel. $\square_l$ and $\square_d$ represent the LD matrices for local and distal SNPs, respectively. The LD matrix between local and distal SNPs $\square_{l,d}$ can be assumed to be zero, though recent studies have showed long-range LD in the human genome[185,186]. We allow the user to set cross-chromosomal LD to 0, though by default, we estimate LD from the reference panel.

Now, we see that, under this null hypothesis, the joint distribution of $(\tilde{Z}_l, \tilde{Z}_d) = (w_l^T Z_l, w_d^T Z_d)$ is given by:

$$\begin{pmatrix} \tilde{Z}_l \\ \tilde{Z}_d \end{pmatrix} \sim N_2 \left( \mathbf{0}, \begin{bmatrix} w_l^T \square_l w_l & w_l^T \square_{l,d} w_d \\ w_d^T \square_{l,d}^T w_l & w_d^T \square_d w_d \end{bmatrix} \right).$$

It follows that, under the null hypothesis and given $\tilde{Z}_l = \tilde{Z}_{l,\text{obs}}$,

$$\tilde{Z}_d | \tilde{Z}_l = \tilde{Z}_{l,\text{obs}} \sim N \left( \frac{w_l^T \square_{l,d} w_d}{w_l^T \square_l w_l} \tilde{Z}_{l,\text{obs}}, w_d^T \square_d w_d - \frac{[w_l^T \square_{l,d} w_d]^2}{w_l^T \square_l w_l} \right).$$

We can use this null distribution for the one-sided test of $H_0 : \mathbf{w}_d^T \mathbf{Z}_d | \mathbf{w}_l^T \mathbf{Z}_l = \tilde{Z}_{l,\text{obs}} = 0$ against $H_1 : \mathbf{w}_d^T \mathbf{Z}_d | \mathbf{w}_l^T \mathbf{Z}_l = \tilde{Z}_{l,\text{obs}} > 0$. This test is implemented in MOSTWAS as a follow-up to the weighted-burden test.

#### 4.1.5 Data acquisition for TCGA-BRCA and iCOGs

We retrieved genotype, RNA expression, miRNA expression, and DNA methylation data for breast cancer indications in The Cancer Genome Atlas[180]. Birdseed genotype files of 914 subject were downloaded from the Genome Data Commons (GDC) legacy (GRCh37/hg19) archive. Genotype files were merged into a single binary PLINK file format (BED/FAM/BIM) and imputed using the October 2014 (v.3) release of the 1000 Genomes Project dataset as a reference panel in the standard

two-stage imputation approach, using SHAPEIT v2.87 for phasing and IMPUTE v2.3.2 for imputation[126–128]. We excluded variants (1) with a minor allele frequency of less that 1% based on genotype dosage, (2) that deviated significantly from Hardy-Weinberg equilibrium ($P < 10^{-8}$) using appropriate functions in PLINK v1.90b3[187,149], and (3) located on sex chromosomes. Final TCGA genotype data was coded as dosages, with reference and alternative allele coding as in dbSNP.

TCGA level-3 normalized RNA-seq expression data, miRNA-seq expression data, and DNA methylation data collected on Illumina Infinium HumanMethylation450 BeadChip were downloaded from the Broad Institute's GDAC Firehose (2016/1/28 analysis archive). We intersectted to the subset of samples assayed for genotype (4,564,962 variants), RNA-seq (15,568 genes), miRNA-seq (1,046 miRNAs), and DNA methylation (485,578 CpG sites), resulting in a total of 563 samples. We only consider the autosome in our analyses. We adjusted gene and miRNA expression and DNA methylation by relevant covariates (5 principal components of the genotype matrix, tumor stage at diagnosis, and age).

For association testing, we downloaded iCOGs GWAS summary statistics for breast cancer-specific survival for women of European ancestry[162]. Funding for BCAC and iCOGS came from: Cancer Research UK [grant numbers C1287/A16563, C1287/A10118, C1287/A10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565], the European Union's Horizon 2020 Research and Innovation Programme (grant numbers 634935 and 633784 for BRIDGES and B-CAST respectively), the European Community's Seventh Framework Programme under grant agreement n° 223175 [HEALTHF2-2009-223175] (COGS), the National Institutes of Health [CA128978] and Post-Cancer GWAS initiative [1U19 CA148537, 1U19 CA148065-01 (DRIVE) and 1U19 CA148112 - the GAME-ON initiative], the Department of Defence [W81XWH-10-1-0341], and the Canadian Institutes of Health Research CIHR) for the CIHR Team in Familial Risks of Breast Cancer [grant PSR-SIIRI-701]. All studies and funders as listed in Michailidou et al[9,10] and in Guo et al[162] are acknowledged for their contributions.

#### 4.1.6 Data acquisition from ROS/MAP, IGAP, and PGC

We retrieved imputed genotype, RNA expression, miRNA expression, and DNA methylation data from The Religious Orders Study and Memory and Aging Project (ROS/MAP) Study for samples derived from human pre-frontal cortex[188,189]. We excluded variants (1) with a minor allele frequency of less that 1% based on genotype dosage, (2) that deviated significantly from Hardy-Weinberg equilibrium ($P < 10^{-8}$) using appropriate functions in PLINK v1.90b3[187,149], and (3) located on sex

chromosomes. Final ROS/MAP genotype data was coded as dosages, with reference and alternative allele coding as in dbSNP. We intersectted to the subset of samples assayed for genotype (4,141,537 variants), RNA-seq (15,857 genes), miRNA-seq (247 miRNAs), and DNA methylation (391,626 CpG sites), resulting in a total of 370 samples. We only consider the autosome in our analyses. We adjusted gene and miRNA expression and DNA methylation by relevant covariates (20 principal components of the genotype age at death, and sex).

For association testing, we downloaded GWAS summary statistics for risk of late-onset Alzheimer's disease from the International Genomics of Alzheimer's Project (IGAP)[2]. We also downloaded GWAS and genome-wide association by proxy (GWAX) summary statistics for risk of major depressive disorder (MDD) from the Psychiatric Genomics Consortium[190] and the UK Biobank[191], respectively.

IGAP is a large two-stage study based on GWAS on individuals of European ancestry. In stage 1, IGAP used genotyped and imputed data on 7,055,881 single nucleotide polymorphisms (SNPs) to meta-analyse four previously-published GWAS datasets consisting of 17,008 Alzheimer's disease cases and 37,154 controls (The European Alzheimer's disease Initiative – EADI the Alzheimer Disease Genetics Consortium – ADGC The Cohorts for Heart and Aging Research in Genomic Epidemiology consortium – CHARGE The Genetic and Environmental Risk in AD consortium – GERAD). In stage 2, 11,632 SNPs were genotyped and tested for association in an independent set of 8,572 Alzheimer's disease cases and 11,312 controls. Finally, a meta-analysis was performed combining results from stages 1 and 2.

## 4.2 Simulation analysis

We conducted simulations to assess the predictive capability and power to detect gene-trait associations under various phenotype ($h^2_p$), local heritability of expression ($h^2_{e,l}$), distal heritability of expression ($h^2_{e,d}$), and proportion of causal local ($p_{c,l}$) and distal ($p_{c,e}$) SNPs for MeTWAS and DePMA. We considered two scenarios for each combination of $(h^2_p, h^2_{e,l}, h^2_{e,d}, p_{c,l}, p_{c,e})$: (1) the simulated distal-eQTL (association between SNP and gene of interest) exists in both the reference and imputation panel, and (2) the distal-eQTL exists in the reference panel, but the distal SNP does not affect gene expression in the imputation panel (i.e. $h^2_{e,d} \equiv 0$ in the imputation panel regardless of $h^2_{e,d}$ in the reference panel).

Using TCGA data in breast cancer, we extracted 2,592 genotypes local to the gene *ESR1* on Chromosome 6 and 1,431 genotypes local to the gene *FOXA1*. Though the choice of these loci were arbitrary for simulations, there is evidence that *ESR1* and *FOXA1* are highly co-expressed in breast tumors and local-eQTLs of *FOXA1* have been shown to be distal-eQTLs of *ESR1*[192]. We believe these loci served as a strong reference for these simulations. We generated (1) a reference panel with sample size 400 with simulated genotypes, expressions, and one mediators and (2) a GWAS panel of 1,500 samples with simulated genotypes and phenotypes using the following data generating process, modified from Mancuso et al's framework[55]:

- We estimated the linkage disequilibrium $LD$ matrix of the genotypes $X_G$ with $n$ samples and $p$ SNPs, as follows with regularization to ensure $LD$ is positive semi-definite:

  $$LD = \frac{1}{n} X_G^T X_G + \frac{1}{10} I_p.$$

  We computed the Cholesky decomposition of $LD$ for faster sampling. We simulated genotypes for a 400-sample reference panel $X_{g,\text{ref}}$ and 1,500-sample GWAS panel $X_{g,\text{GWAS}}$.

- We then simulated effect sizes for $p_{c,l}$ of the 2,592 local genotypes $w_{g,l}$ from a standard Normal distribution. We generated locally heritable expression

  $$E_{g,l} = X_{G,\text{ref}}^T w_{g,l} + \epsilon_l,$$

  with $\epsilon_l \sim N(0, 1 - h^2_{e,l})$ and $w_{g,l}$ scaled to ensure the given $h^2_{e,l}$.

  Similarly, we simulated effect sizes for $p_{c,d}$ of the 1,431 distal genotypes $w_{g,d}$ and generated the distally heritable intensity of the mediator $M_{g,d}$. We constructed the distally heritable expression $E_{g,d}$ by scaling $M_{g,d}$ by $\beta \sim N(0, 1)$ and adding random noise that scaled distal heritability to $h^2_{e,d}$. We lastly formed the total expression $E_g = E_{g,l} + Eg, d$.

- Next, we simulated the phenotype in the GWAS panel such that the variance explained in the phenotype reflects only that due to genetics. We drew a causal effect size from gene expression $\alpha \sim N(0, 1)$. We computed the "unobserved" gene expression in the GWAS panel as

  $$E_{g,\text{GWAS}} = X^T_{g,\text{GWAS,local}} w_{g,l} + X^T_{g,\text{GWAS,distal}} w_{g,d} \beta.$$

Here, we also considered a "null" case as well, where the distal eQTLs were not present in the GWAS panel (i.e. $w_{g,d} = 0$ for all distal SNPs). GWAS summary statistics were computed in this step for downstream weighted burden testing.

- We then fitted predictive models using MeTWAS, DePMA, and local-only models (i.e. FUSION[3]), computed the adjusted predictive $R^2$ in the reference panel, and tested the gene-trait association in the GWAS panel using a weighted burden test.

The association study power was defined as the proportion of gene-trait association tests with $P < 2.5 \times 10^{-6}$, the Bonferroni-corrected significance threshold for testing 20,000 independent genes.

In these simulation studies, we found that MOSTWAS methods performed well in prediction across different causal proportions and local and distal mRNA expression heritabilities. Furthermore, across all simulation settings, we observed that MOSTWAS showed greater or equal power to detect gene-trait associations as local-only models. We saw that, as the proportion of total expression heritability that is attributed to distal genetic variation, the positive difference in predictive performance between the best MOSTWAS model and the local-only model increased (**Supplemental Figure S28**). Similarly, we found that, under the setting that distal variation contributes to trait heritability, the best MOSTWAS model has greater power to detect gene-trait associations than the local-only model, with the advantage in power over local-only models increasing with increased distal expression heritability (**Figure 4.3A**). Under the null case that distal variation influences expression in the reference panel but does not affect the trait in the GWAS panel, we find that local-only and MOSTWAS models perform similarly. At low causal proportion ($p_c = 0.01$) and low trait heritability ($h_p^2 = 0.2$), local-only models have a modest advantage in TWAS power over MOSTWAS models. This difference is mitigated at larger causal proportions and trait heritabilities (**Figure 4.3**). Overall, these results demonstrated the advantages of MOSTWAS methods for modeling the complex genetic architecture of transcriptomes, especially when distal variation has a discernibly large effect on the heritability of both the gene and trait of interest.

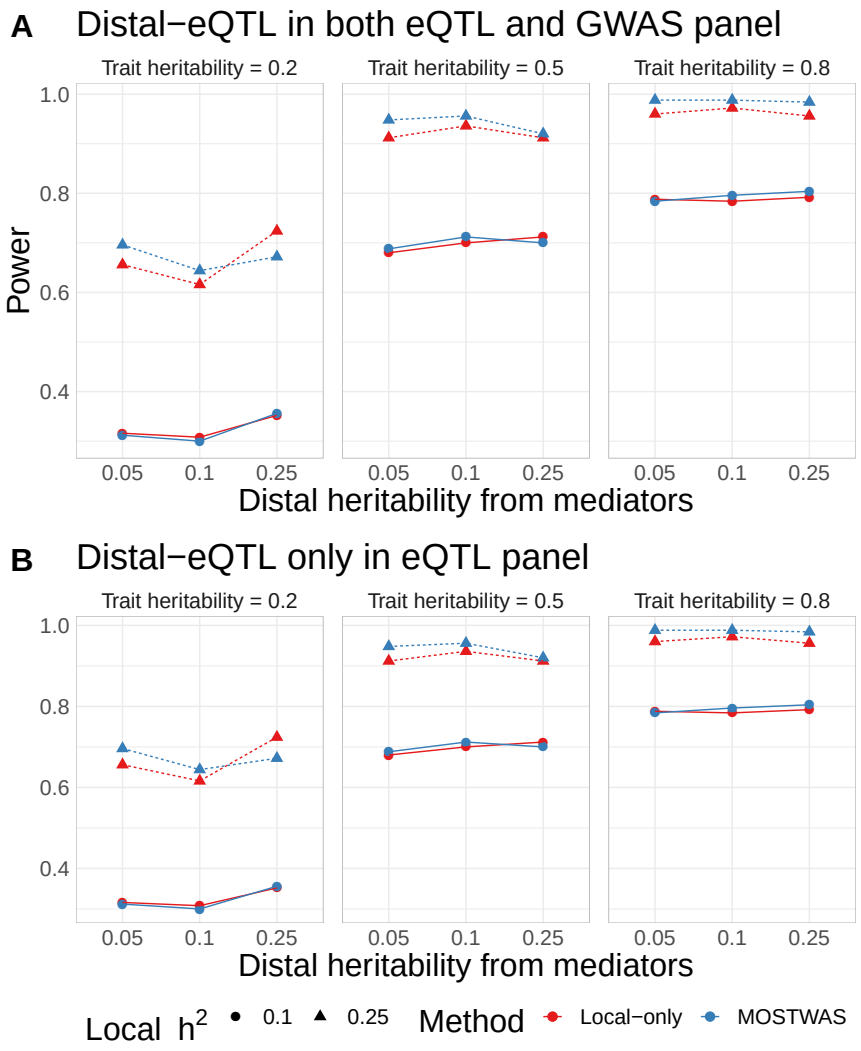


**Figure 4.3:** *Comparison of TWAS power via simulations using MOSTWAS and local-only models.* (A) Proportion of gene-trait associations at $P < 2.5 \times 10^{-6}$ using local-only (red) and the most predictive MOSTWAS (blue) models across various local and distal expression heritabilities, trait heritability, and causal proportions. (B) Proportion of significant gene-trait associations across the same simulation parameters with no distal effect on the trait in the simulated external GWAS panel.

## 4.3 Applications of MOSTWAS in real data

### 4.3.1 Breast cancer expression and survival outcomes

We wished to apply MOSTWAS in the context of breast tumor multi-omics and disease outcomes, motivated by recent GWAS and TWAS into breast cancer-specific survival[161,9,10,162,102]. Breast tumor eQTL studies have also revealed several signficant distal-eQTLs in trait-associated loci, many of which are in regulatory or epigenetic hotspots[102,193], making breast tumors a natural setting for MOSTWAS application. Using TCGA-BRCA data on germline SNPs, tumor mRNA expression, DNA methylation, and miRNA expression, we trained MeTWAS, DePMA, and traditional local-only predictive models for the mRNA expression of all genes with germline heritability $h^2 > 0$ at $P < 0.05$. Estimates of heritability for genes were considerably larger when we considered distal variation using MOSTWAS methods (mean heritabilities in **Supplemental Table S3**). We also found that MeTWAS and DePMA perform better in cross-validation $R^2$ in cross-validation, with larger numbers of models at $R^2 \geq 0.01$ using MOSTWAS methods than local-only models (**Figures 4.3A-C**). Mean predictive $R^2$ for local-only models was 0.011 (25% to 75% inter-quartile interval (0.0,0.013)), for MeTWAS models was 0.028 (0.013, 0.032), and for DePMA models was 0.051 (0.019, 0.068).

In addition to cross-validation, we used 351 paired samples in TCGA-BRCA with genotype and mRNA expression data that were not used in model training to test the portability of MOSTWAS models in independent external cohorts. As shown in **Figure 4.4A**, DePMA models obtain the highest predictive adjusted $R^2$ in the external cohort (mean 0.016, 25% to 75% inter-quartile interval (0.003, 0.018)), with local-only models (0.013, (0.00,0.013)) outperforming MeTWAS models (0.011, (0.002, 0.012)), considering only genes that attained cross-validation adjusted $R^2 \geq 0.01$ using a given method. Overall, among genes with cross-validation adjusted $R^2 \geq 0.01$, 37 out of 280 genes achieved external predictive $R^2 \geq 0.01$ using local-only models, 89 out of 709 using MeTWAS, and 787 out of 1,185 using DePMA (**Figure 4.3A-C**).

Lastly, we conducted association studies for breast cancer-specific survival using local-only and the most predictive (in cross-validation) MOSTWAS model trained in TCGA-BRCA and summary-level GWAS data from iCOGs. Here, we constructed the weighted burden test, as described above and in Pasaniuc et al and Gusev et al[50,3]. We prioritized genes with Benjamini-Hochberg adjusted $P < 0.05$ for permutation testing. Of the 122 genes that had cross-validation $R^2 \geq 0.01$ in TCGA-BRCA using both local-only and MOSTWAS models, we found 2 survival associations at Benjamini-Hochberg FDR-adjusted $0.05$ using both local-only and MOSTWAS models, with the strength of association

marginally larger with the MOSTWAS model in each case (**Supplemental Figure S29**). QQ-plots for TWAS $Z$-statistics and $P$-values are provided in **Supplemental Figure S30A** and **Supplemental Figure S31** for both local-only and MOSTWAS models, showing earlier departure for the local-only models. Overall, using all heritable genes with cross-validation $R^2$ with the best MOSTWAS model in TCGA-BRCA, we identified 21 survival-associated loci at Benjamini-Hochberg[1] FDR adjusted $P < 0.05$. Of these 21 loci, 11 persisted when subjected to permutation testing at a significance threshold of FDR-adjusted $P < 0.05$ (**Figure 4.4C**). Our results in TCGA-BRCA showed improved transcriptomic prediction using MOSTWAS over local-only modeling and the strength of MOSTWAS to detect gene-trait associations that are influenced by distal variation.

#### 4.3.1.1 Functional hypothesis generation with MOSTWAS

An advantage of MOSTWAS is its ability to aid in functional hypothesis generation for mechanistic follow-up studies. The added-last test allows users to identify genes where trait association from distal variation is significant given the strength of the local association. For 8 of the TWAS-associated 11 loci, at FDR-adjusted $P < 0.05$ we found significant distal variation added-last associations (see Section 4.1.4), suggesting that distal variation may contribute to the gene-trait association. All 8 of these loci showed distal association with the gene of interest mediated through a set of four transcription factors (*NAA50*, *ATP6V1A*, *ROCK2*, *USF3*), all highly interconnected with the critical MAPK pathway[194–199]. These regulatory sites serve as an example of how distal genomic regions can be prioritized for functional follow-up studies to elucidate the mechanisms underlying the SNP-gene-trait associations.

### 4.3.2 Brain gene expression and psychiatric disorders

We also applied MOSTWAS to transcriptomic data on samples of prefrontal cortex, a tissue that has been used previously in studying neuropsychiatric traits and disorders with TWAS[200,201]. There has been ample evidence in brain tissue, especially the prefrontal cortex, that non-coding variants (up to 80%) regulate distal genes, providing a prime example to assess MOSTWAS[200,202]. Using ROS/MAP data on germline SNPs, tumor mRNA expression, DNA methylation, and miRNA expression, we trained MeTWAS, DePMA, and traditional local-only predictive models for the mRNA expression of all genes with germline heritability $h^2 > 0$ at $P < 0.05$. Consistent with results in TCGA-BRCA, estimates of heritability for genes were considerably larger when we considered distal variation using MOSTWAS methods (Supplemental Table S3). We also find that MeTWAS and DePMA perform better in cross-validation $R^2$ than local-only models (**Figures 4.4D-F**). Mean

predictive $R^2$ for local-only models was 0.029 (25% to 75% inter-quartile interval (0.0,0.015)), for MeTWAS models was 0.079 (0.019, 0.082), and for DePMA models was 0.045 (0.013, 0.037).

In addition to cross-validation, we used 87 samples in ROS/MAP with genotype and mRNA expression data that were not used in model training to test the portability of MOSTWAS models in independent external cohorts. As shown in **Figure 4.4A**, DePMA models obtain the highest predictive adjusted $R^2$ in the external cohort (0.042 (25% quantile 0.009, 75% quantile 0.057)), with MeTWAS models (0.040 (0.010, 0.054)) outperforming local-only models (0.031 (0.007, 0.039)), considering only genes that attained cross-validation adjusted $R^2 \geq 0.01$ using a given method. Overall, among genes with cross-validation adjusted $R^2 \geq 0.01$, 187 out of 267 genes achieved external predictive $R^2 \geq 0.01$ using local-only models, 683 out of 911 using MeTWAS, and 2,135 out of 2,934 using DePMA (**Figure 4.3D-F**).

We next conducted association tests for known Alzheimer's disease risk loci using local-only and the best MOSTWAS model (comparing MeTWAS and DePMA cross-validation $R^2$) trained in ROS/MAP and summary-level GWAS data from IGAP. From literature, we identified 14 known common and rare loci of late-onset Alzheimer's disease[2,203–205], 11 of which had MOSTWAS models with cross-validation $R^2 \geq 0.01$. Five of these 11 loci (*APOE*, *CLU*, *PLCG2*, *SORL1*, *ZCWPW1*) showed significant association at Benjamini-Hochberg FDR-adjusted $P \leq 0.05$ (Supplemental Table S4). We also compared these all 11 associations to those identified by local-only models and by latent Dirichlet process regression (DPR) as implemented in TIGAR[54], with raw $P$-values of association shown in **Figure 4.5B**. MOSTWAS showed stronger associations at 8 of these loci than both local-only and DPR models. We followed up on the 5 significantly associated loci using the permutation and added-last tests. The added-last test assesses whether the association from distal loci, given the strength of the association in the local locus, is significant. Three of these loci (*APOE*, *SORL1*, *ZCWPW1*) persisted permutation testing at FDR-adjusted $P < 0.05$ and showed significant associations with distal variants, given the association with local variants, at FDR-adjusted $P < 0.05$ (Supplemental Table S4).

We then conducted a transcriptome-wide association study for risk of major depressive disorder (MDD) using summary statistics from the Psychiatric Genomics Consortium (PGC) genome-wide meta-analysis that excludes data from the UK Biobank and 23andMe[190]. QQ-plots for TWAS $Z$-statistics and $P$-values are provided in **Supplemental Figure S30B** and **Supplemental Figure S31**. for both local-only and MOSTWAS models. Overall, using all heritable genes with cross-validation $R^2$ with the best MOSTWAS model in ROS/MAP, we identified 102 MDD risk-associated loci with FDR-adjusted $P < 0.05$ that persisted when subjected to permutation testing at an FDR-adjusted significance threshold of $P < 0.05$ (colored red in **Figure 4.4D**). We downloaded genome-wide

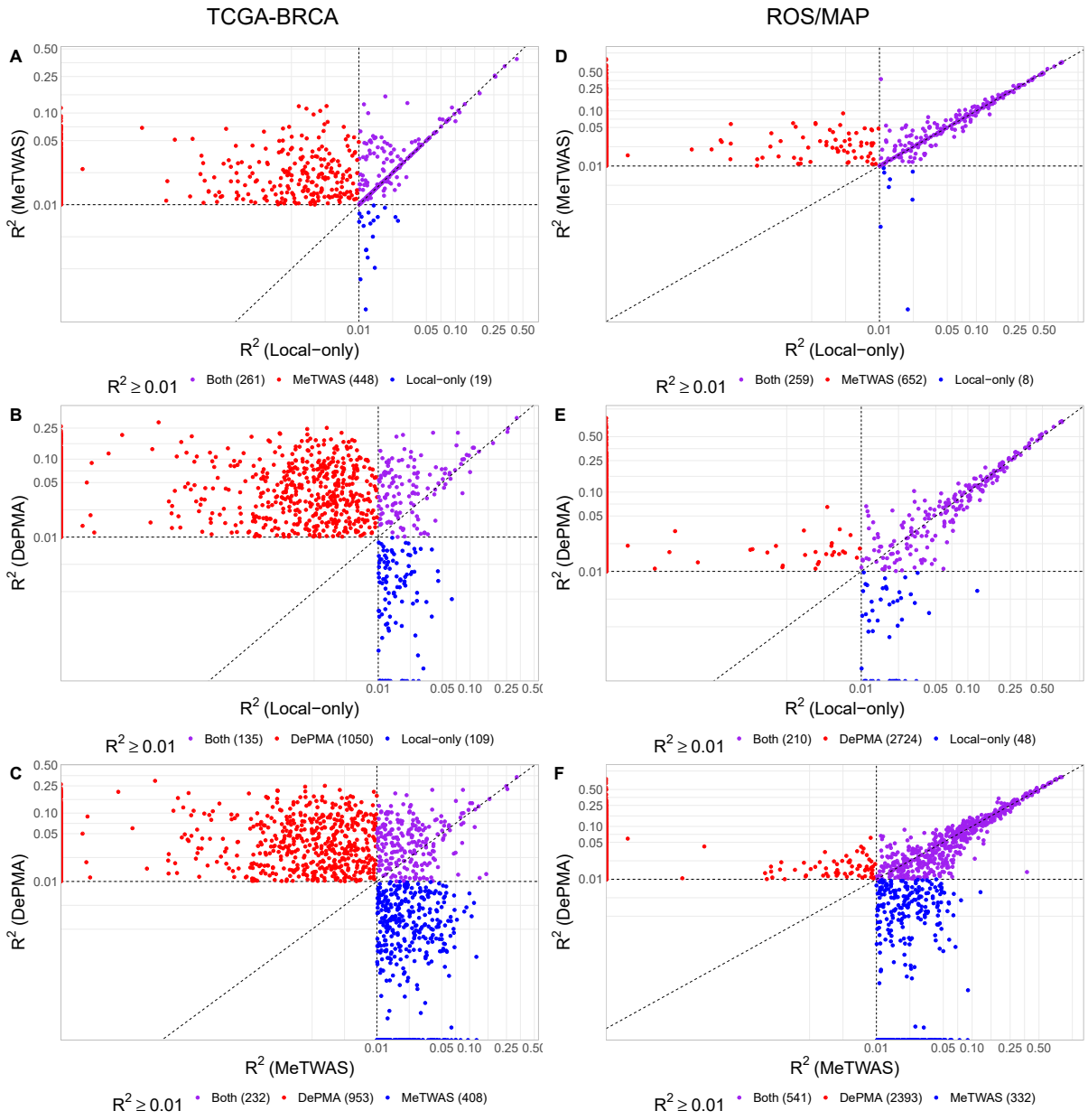


**Figure 4.4:** *Comparison of predictive adjusted $R^2$ in cross-validation using local-only, MeTWAS, and DePMA models.* If a given gene does not have $h^2 > 0$ with $P < 0.05$, we set the predictive adjusted $R^2$ to 0 here for comparison. We compare local-only and MeTWAS in TCGA-BRCA (A) and ROS/MAP (D), local-only and DePMA in TCGA-BRCA (B) and ROS/MAP (E), and MeTWAS and DePMA in TCGA-BRCA (C) and ROS/MAP (F).

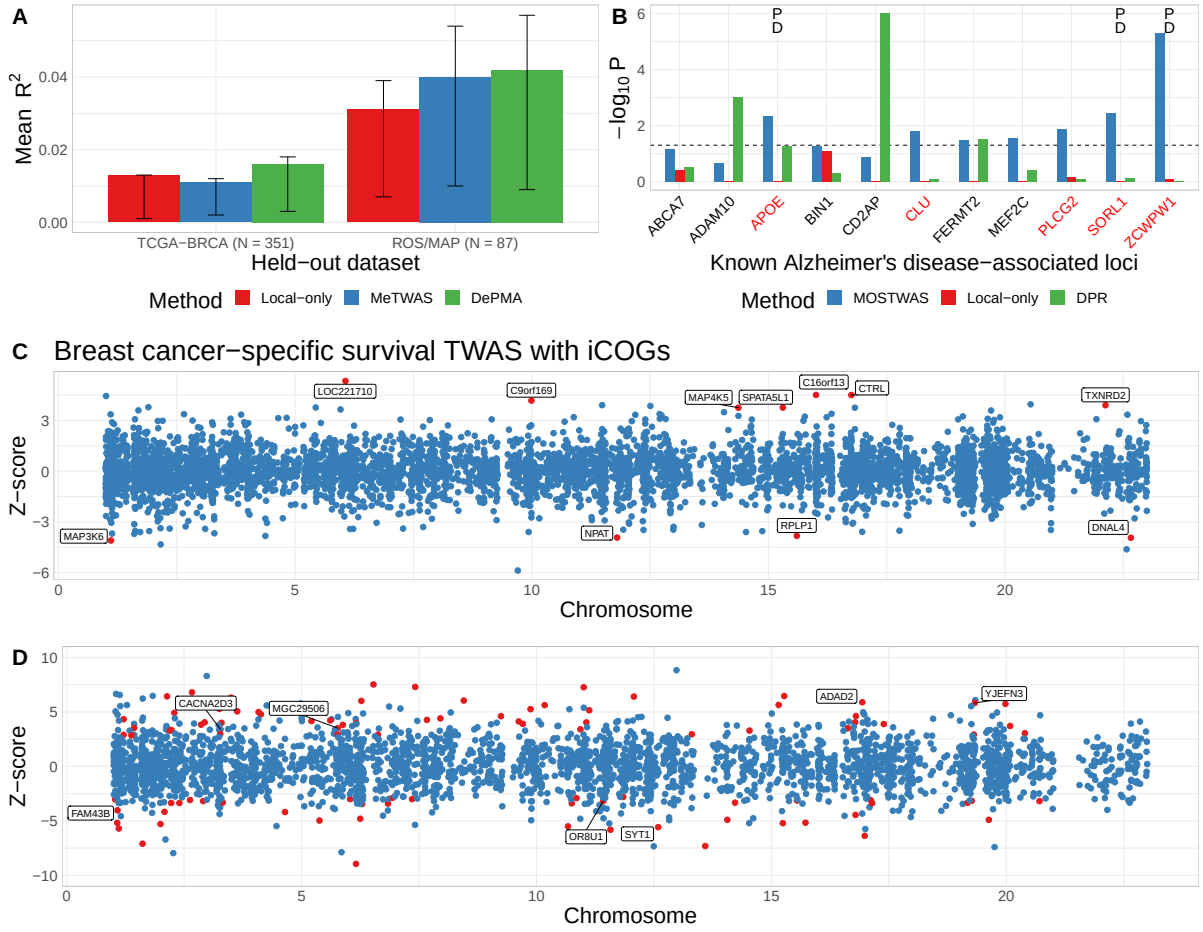


**Figure 4.5:** *External validation of MOSTWAS and gene-trait associations using MOSTWAS models.* (A) Predictive adjusted $R^2$ in held-out cohorts from TCGA-BRCA and ROS/MAP in local-only, MeTWAS, and DePMA models that have in-sample significant heritability and cross-validation $R^2 \geq 0.01$. The interval shows the 25% and 75% quantiles for external cohort predictive $R^2$. (B) Associations with 12 known Alzheimer's risk lock, as identified in literature, using MOSTWAS, local-only, and TIGAR Dirichlet process regression (DPR). (C) TWAS associations for breast cancer-specific survival using GWAS summary statistics from iCOGs. Loci are colored and labelled if the overall association achieves FDR-adjusted $P < 0.05$ and the permutation test also achieves FDR-adjusted $P < 0.05$. (D) TWAS associations for major depressive disorder risk using GWAS summary statistics from PGC. Loci are colored red if the overall association achieves FDR-adjusted $P < 0.05$ and the permutation test also achieves FDR-adjusted $P < 0.05$. We label the 12 loci that were independently validated with UK Biobank GWAX summary statistics at FDR-adjusted $P < 0.05$ for both the overall association test and permutation test.

association study by proxy (GWAX) summary statistics from the UK Biobank[191] for replication analysis of loci identified using PGC summary statistics. We found that 7 of these 102 loci (labelled in **Figure 4.4D**) also show an association in UK Biobank GWAX that is in the same direction as in PGC. Summary statistics for TWAS associations in PGC and UK Biobank are provided in Supplemental Table S5. It is important to note the UK Biobank dataset is not a GWAS dataset as it defines a case of MDD as any subject who has the disorder or a first-degree relative with MDD, leading to lower power to detect associations in this dataset.

We observed that MOSTWAS models generally had higher predictive $R^2$ that local-only models both in training and independent cohorts. We also found that MOSTWAS has recapitulated 5 known Alzheimer's risk loci that were not detected by local-only modeling (both PrediXcan[42] and TIGAR[54]), 3 of which had significant distal associations using our added-last test. We also illustrated that the MOSTWAS detected MDD-risk loci that were replicable across independent GWAS and GWAX cohorts[190,191].

### 4.3.3 Comparison of computational time

To assess the difference in computational burden between local-only, MeTWAS, and DePMA modelling, we randomly selected a set of 50 genes that are heritable across all three models from TCGA-BRCA and computed per-gene time for fitting using a 24-core, 3.0 GHz processor. We found that MeTWAS (mean of 225 seconds per gene) and DePMA (mean 312 seconds per gene) takes approximately 6-10 times longer to fit than a traditional local-only model (mean 36 seconds) (**Supplemental Figure S27**). Model-fitting here includes heritability estimation, estimating the SNP-expression weights, and cross-validation. We have implemented parallel options within a given gene and recommend fitting an entire set of genes on an RNA-seq panel via a batch computing approach[206]. Using parallel implementation with 5 cores and batch computing, we analyzed 15,568 genes from TCGA-BRCA in approximately 28 hours.

## 4.4 Discussion

Here, through a variety of simulations and real applications in two settings, we have shown that multi-omic methods that prioritize distal variation in TWAS gave added predictive performance and power to detect gene-trait associations, especially when distal variation contributed to trait heritability. We proposed two methods (MeTWAS and DePMA) for identifying and including distal genetic variants

in gene expression prediction models. We have provided implementations of these methods in the MOSTWAS (Multi-omic Strategies for Transcriptome-Wide Association Studies) R package, available freely on Github. MOSTWAS contains functions to train expression models with both MeTWAS and DePMA and outputs models with 5-fold cross-validation $R^2 \geq 0.01$ and significant germline heritability. The package also contains functions and documentation for simulation analyses[55], the weighted burden and follow-up permutation and distal-SNPs added last tests for TWAS[50,3] using GWAS summary statistics, and file-formatting. We also provide guidelines for parallelization to speed up computational time.

Not only does MOSTWAS improve transcriptomic imputation both in- and out-of-sample, it also provides a test for the identification of heritable mediators that may affect the eventual transcription of the gene of interest. These identified mediators can give some insight into the underlying mechanisms for SNP-gene-trait associations to improve detection of gene-trait associations and prioritize functional follow-up studies. Using MOSTWAS and iCOGs summary-level GWAS statistics for breast cancer-specific survival[162], we identified 11 survival-associated loci that are enriched for p53 binding and oxidoreductase activity pathways[207,208]. These loci include two genes (*MAP3K6* and *MAP4K5*) encoding mitogen-activated protein kinases, which are signalling transduction molecules involved in the progression of aggressive breast cancer hormone subtypes[209]. TWAS using MOSTWAS models was able to recapitulate 5 out of 14 known Alzheimer's disease risk loci in IGAP GWAS summary statistics statistics[2], which were not recoverable with local-only models. We showed the utility of the distal-SNPs added last test to prioritize significant distal SNP-gene-trait associations from follow-up. In PGC GWAS summary-level data for major depressive disorder[190], we found 102 risk loci, 7 of which were replicated in independent GWAX summary statistics from the UK Biobank[191]. Three of these seven loci (*SYT1*, *CACNA2D3*, *ADAD2*) encode important proteins involved in synaptic transmission in the brain and RNA editing. Studies have shown that variation at these loci may lead to loss of function at synapses and RNA editing that lead to psychiatric disorders[210–214]. All survival- or risk-associated loci identified by MOSTWAS were not detected using local-only models.

An admitted and considerable limitation of MOSTWAS is the increased computational burden over local-only modelling, especially in DePMA's permutation-based mediation analysis for multiple genome-wide mediators. We believe a Monte-Carlo resampling method will aid in scalability by making some standard distributional assumptions on the effect sizes of SNPs and mediators in the DePMA mediation model[215]. Nevertheless, we believe that MOSTWAS's gain in predictive performance and power to detect gene-trait associations may outweigh this computational time. Another limitation of MOSTWAS is the general lack of rich multi-omic panels, like TCGA-BRCA and

ROS/MAP, that provide a large set of mediating biomarkers that may be mechanistically involved in gene regulation. However, we believe that mRNA expression data could be re-used as mediator data to identify distal-eQTLs local to genes that code for transcription factors[67,68,72], which is an area of future development in MOSTWAS.

In conclusion, MOSTWAS provides a user-friendly and intuitive tool that extends transcriptomic imputation and association studies to include distal genetic variants. MOSTWAS enables users to utilize rich reference multi-omic datasets for enhanced gene mapping to better understand the genetic etiology of polygenic traits and diseases with more direct insight into functional follow-up studies.

# CHAPTER 5: CELL-TYPE DECONVOLUTION IN TARGETED RNA EXPRESSION PANELS

In this chapter, we outline a semi-reference-free cell-type deconvolution method using mRNA expression data from targeted panels. As mentioned in Chapter 1, targeted panels are particularly attractive for clinical settings and for longitudinal studies that use archival specimens[18]. However, a major limiting factor for targeted panels, especially for cell-type deconvolution, is the limited feature space; reference-free deconvolution methods rely on identifying genes that can indicate different cell-types, but targeted panels do not afford a large enough feature space to search for these cell-type specific genes[13,15,17,14]. We introduce DeCompress, a semi-reference-free method that uses compressed sensing to expand the targeted expression panel to a larger feature space using a reference RNA-seq or microarray dataset as a reference. We benchmark DeCompress against reference-free methods in simulated and published datasets and show that DeCompress generally estimates cell-type proportions with less error than competing reference-free methods. We then show some advantages of including these estimated cell-type proportions in clinical and academic settings (e.g. eQTL mapping, subtyping and outcome prediction) using data from the Carolina Breast Cancer Study (CBCS)[122,24]. DeCompress is available as an R software package at `https://github.com/bhattacharya-a-bt/DeCompress`.

## 5.1 Overview of DeCompress

DeCompress takes in two expression matrices from similar bulk tissue as inputs: the *target* matrix **T**, an $n \times k$ matrix from a targeted panel of gene expression, and the *reference* matrix **R**, an $N \times K$ matrix from an RNA-seq or microarray panel, such that $K > k$. For a user-defined $c$ cell-types, DeCompress outputs $\hat{\mathbf{S}}$, a $c \times K'$ matrix of cell-type specific expression profiles and $\hat{\mathbf{P}}$, a $c \times n$ matrix of cell-type proportions. The method follows three general steps, as detailed in **Figure 5.1**: (1) selection of approximate cell-type specific genes, (2) compressed sensing to expand the feature space of **T**, and (3) ensemble reference-free deconvolution on expanded expression dataset. DeCompress is freely available as an R package on Github (`https://github.com/bhattacharya-a-bt/DeCompress`).

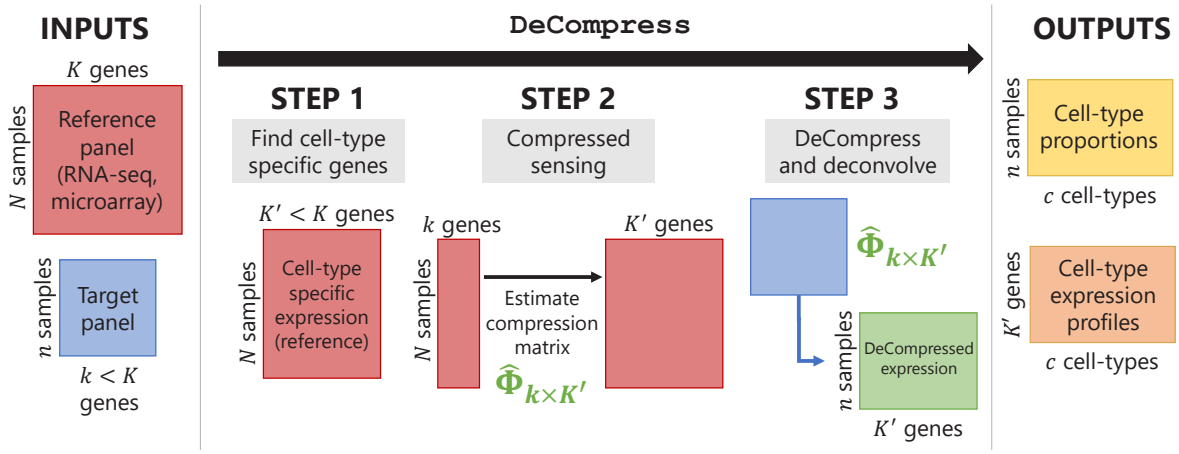


**Figure 5.1:** *Schematic for the DeCompress algorithm.* DeCompress takes in a reference RNA-seq or microarray matrix with $N$ samples and $K$ genes, and the target expression with $n$ samples and $k < K$ genes. The algorithm has three general steps: (1) finding the $K' < K$ genes in the reference that are cell-type specific, (2) training the compressed sensing model that projects the feature space in the target from $k$ genes to the $K'$ cell-type specific genes, and (3) decompressing the target to an expanded dataset and deconvolving this expanded dataset. DeCompress outputs cell-type proportions and cell-type specific profiles for the $K'$ genes.

### 5.1.1 Selection of cell-type specific genes

The first step of DeCompress is to find a set of $K' < K$ genes that are representative of the different cell types that comprise the bulk tissue. These $K'$ genes, called the cell-type specific (CTS) genes, can be supplied by the user if prior gene signatures can be applied. If any such gene signatures are not available, DeCompress borrows methods from two previous reference-free deconvolution methods to select a parsimonious gene set.

We include methods from Zaitsev et al's LINear Subspace identification for gene Expression Deconvolution (LINSEED) method[13] that assumes mutual linearity (i.e. $y_1 = ky_2$, where $y_1$ and $y_2$ are the expressions of gene 1 and gene 2, respectively) between cell-type specific genes to generate gene signatures. Briefly, LINSEED transforms the gene expression space to form a $c$-vertex simplex, where each vertex represents a distinct cluster of mutually linear genes corresponding to a cell type. The algorithm then picks the closest genes to each vertex to represent a cell-type specific gene signature[13]. We also include Li and Wu's feature selection method, TOols for the Analysis of heterogeneouS Tissues (TOAST)[14], which iteratively searches for cell type-specific genes and performs reference-free estimation at each step. TOAST uses a novel hypothesis testing framework to conduct cross-cell type differential analysis and identify gene signatures[14].

### 5.1.2 Compressed sensing framework

After a suitable set of $K'$ CTS genes are determined, we take the $K'$ corresponding columns of **R** to form $\mathbf{R'}_{N\times K'}$ and the $k$ genes corresponding to columns in **T** to form $\mathbf{R}^{(\mathbf{k})}_{\mathbf{N}\times\mathbf{k}}$. Consider the following matrix equation, where □ is a $k \times K'$ compression matrix that projects $\mathbf{R}^{(\mathbf{k})}$ to $\mathbf{R'}$:

$$\mathbf{R'}_{N\times K'} = \mathbf{R}^{(\mathbf{k})}_{\mathbf{N}\times\mathbf{k}}\square_{\mathbf{k}\times\mathbf{K'}} \tag{5.1}$$

We can break down Equation 5.1 into a system of equations. For the $i$th column of $\mathbf{R'}$, denoted $r'_i$, we wish to find a $k$-length sparse vector $\phi_i,\ 1 \leq i \leq K'$ such that

$$r'_i = \mathbf{R}^{(\mathbf{k})}_{\mathbf{N}\times\mathbf{k}}\phi_i. \tag{5.2}$$

We estimate $\hat{\phi}_i$ with the following optimization methods: least angle regression (using R package lars)[216], elastic net with elastic net mixture penalty $\alpha \in \{0, 0.5, 1\}$ (using the R package glmnet)[48], and $l_1$, $l_2$, and total variation $l_1$ (TV-L1) non-linear optimization (using R package R1magic)[217–220].. Functions in DeCompress allow the user to select any to all of these optimization methods and picks the best method through 5-fold cross-validation.

Especially when $N$ is sufficiently large, non-linear optimization is computationally expensive (see comparison of run times in **Supplemental Figure S32**). We implement parallelization across columns of $\mathbf{R'}$ using the future package in R[221] and recommend linear optimization methods as they are faster and give generally similar prediction (**Supplemental Figure S33**).

#### 5.1.2.1 Optimization methods for compressed sensing

Compressed sensing in DeCompress aims to estimate the $k \times K'$ compression matrix □ in the equation:

$$\mathbf{R'}_{N\times K'} = \mathbf{R}^{(\mathbf{k})}_{\mathbf{N}\times\mathbf{k}}\square_{\mathbf{k}\times\mathbf{K'}}. \tag{5.3}$$

We convert this into a system of equations. For the $i$th column of $\mathbf{R'}$, denoted $r'_i$, we wish to find a $k$-length sparse vector $\phi_i,\ 1 \leq i \leq K'$ such that

$$r'_i = \mathbf{R}^{(\mathbf{k})}_{\mathbf{N}\times\mathbf{k}}\phi_i. \tag{5.4}$$

DeCompress implements several regularized regression or optimization methods to estimate $\hat{\phi}_i$:

- *Elastic net*[48] finds

$$\hat{\phi}_i = \underset{\phi_i}{\arg\min} \left\{ \|r'_i - \mathbf{R^{(k)}}\phi_i\|_2^2 + \lambda \left[ \frac{(1-\alpha)}{2} \|\phi_i\|_2^2 + \alpha \|\beta\|_1 \right] \right\}. \tag{5.5}$$

  We have implemented $\alpha \in \{0, 0.5, 1\}$, where $\alpha = 0$ represents ridge regression with no sparsification of $\phi_i$ and $\alpha = 1$ represents traditional LASSO[47]. This optimization is carried out in DeCompress with the glmnet package[48].

- *Least angle regression* (LARS) minimizes the LASSO objective function in Expression 5.5 that speeds ups stage-wise forward selection. The algorithm starts with all elements of $\phi_i$ equal to zero and finds the predict most correlated with the response. The largest step possible is take in the direction of these predictor until some other predictor has as much correlation with the residual. LARS then proceeds in a direction equiangular between these two predictors until a third variable shares an equal correlation with the residual. The full mathematical justification and details are provided by Efron et al[216].

- $l_1$ *non-linear optimization* solves the following optimization using the `nlm` function in R, as implemented in the R1magic package[217]:

$$\hat{\phi}_i = \underset{\phi_i}{\arg\min} \left\{ \sum_{i=1}^{N} |\mathbf{R^{(k)}}\mathbf{T}\phi_i - r'_i|^2 + \lambda |\phi_i| \right\}, \tag{5.6}$$

  where **T** is a $K' \times K'$ matrix of sparsity bases and $\lambda$ is a tuned penalty parameter.

- $l_2$ *non-linear optimization* solves the following optimization using the `nlm` function in R, as implemented in the R1magic package[217]:

$$\hat{\phi}_i = \underset{\phi_i}{\arg\min} \left\{ \sum_{i=1}^{N} |\mathbf{R^{(k)}}\mathbf{T}\phi_i - r'_i|^2 + \lambda \sqrt{|\phi_i|} \right\}, \tag{5.7}$$

  where **T** is a $K' \times K'$ matrix of sparsity bases and $\lambda$ is a tuned penalty parameter.

- *total-variation* $l_1$ *non-linear optimization* solves the following optimization using the `nlm` function in R, as implemented in the R1magic package[217]:

$$\hat{\phi}_i = \underset{\phi_i}{\arg\min} \left\{ \sum_{i=1}^{N} \|\mathbf{R^{(k)}}\mathbf{T}\phi_i - r'_i\|_F^2 + \lambda TV(\phi_i) \right\}, \tag{5.8}$$

where $\mathbf{T}$ is a $K' \times K'$ matrix of sparsity bases, $\lambda$ is a penalty parameter, and $TV(\cdot)$ is the total-variation function, such that for a generic $n$-length vector $\nu$ with $j$th element $\nu_j$

$$TV(\nu) = \sum_{i=1}^{n-1} |\nu_i - \nu_{i+1}|.$$

#### 5.1.3 Ensemble deconvolution on expanded dataset

After the estimated compression matrix $\hat{\Phi}$ is obtained, we then expand the expression matrix from the targetted panel $\mathbf{T}_{n\times k}$ into a larger features space by multiplying $\mathbf{T}$ with $\hat{\Phi}$:

$$\tilde{\mathbf{T}}_{n\times K'} = \mathbf{T}_{n\times k} \square_{k\times K'}.$$

This expanded expression matrix $\tilde{\mathbf{T}}$, called the *decompressed* expression matrix, is then used for ensemble deconvolution. DeCompress includes multiple options for deconvolution, summarized in **Supplemental Table S6**: (1) reference-free methods, such as deconf[12], CellDistinguisher[15], TOAST with non-negative matrix factorization[14], Linseed[13], and DeconICA[17], and (2) reference-based methods using cell-type specific expression profiles from factorization of $\mathbf{R}'_{N\times K'}$, unmix from the DESeq2 package[16]. These methods are summarized in detail in the **Supplemental Methods**. The optimal estimated cell-type proportion matrix $\hat{\mathbf{P}}$ and cell-type specific expression profiles matrix $\hat{\mathbf{S}}$ are selected from the method that best recreates $\tilde{\mathbf{T}}$ (i.e. minimizes $\|\tilde{\mathbf{T}} - \hat{\mathbf{S}}^T\hat{\mathbf{P}}\|$).

### 5.2 Methods for benchmarking and real data analysis

#### 5.2.1 In-silico GTEx mixing experiments

We downloaded median tissue-specific expression profiles from the Genotype-Tissue Expression (GTEx) Project[222,223] for mammary tissue, lymphocytes, fibroblasts, and adipose tissue. Call these median expression profiles $\mathbf{E}_{\text{profile}}$. We randomly generated a matrix of mixing proportions $\mathbf{P}$ for $n$ samples and $c \in \{2, 3, 4\}$ of the tissue types. We then generated mixed expression profiles with the following model:

$$\mathbf{E}_{\text{mixed}} = \mathbf{E}_{\text{profile}}\mathbf{P}^T.$$

We then multiplied each element of $\mathbf{E}_{\text{mixed}}$ with a randomly generated error term drawn from a Normal distribution with 0 mean and standard deviation of either 4 or 8 (low and high noise). This simulates natural perturbation to mixed expression profiles. We then randomly generated 25 simulated pseudo-targeted panels each of $K \in \{200, 500, 800, 1000\}$ genes that have means and variances above the median mean and variance of all genes in the simulated genes. These simulated datasets have sample size 200. For benchmarking, in each of these simulated datasets, we selected 100 of the 200 samples as a test set for deconvolution. The other 100 samples are considered only in DeCompress deconvolution and simulated expression for all genes are kept as the reference. We added more multiplicative noise to the reference drawn from a Normal distribution with zero mean and standard deviation of 10 to simulate batch differences between the reference and target.

### 5.2.2 Benchmarking in published datasets

We downloaded four datasets, summarized in **Supplemental Table S7**: (1) microarray expression for mixed rat brain, liver, and lung biospecimens (GEO Accession Number: GSE19830), commonly used as a benchmarking dataset in deconvolution studies ($N = 42$)[6], (2) RNA-seq expression (GEO Accession Number: GSE123604) for a mixture of breast cancer cells, fibroblasts, normal mammary cells, and Burkitt's lymphoma cells ($N = 40$)[5], (3) microarray expression for laser capture micro-dissected prostate tumors ($N = 30$)[7], and (4) RNA-seq expression for a mixture of two lung adenocarcinoma cell lines ($N = 40$)[8]. Here, we detail the process of generating pseudo-targeted panels from these RNA-seq or microarray datasets. Assume the downloaded datasets are coded in the matrix **E** with $K$ rows corresponding to genes and $n$ columns corresponding to samples. We take the $K'$ genes such that the means and variances of each of these $K'$ genes are in the top 50% of means and variances of all $K$ genes. This restriction is placed on the $K'$ genes so as to not include lowly expressed genes with no variation across cell-types or other conditions. We then generated 25 pseudo-targeted panels with randomly selected 200, 500, 800, and 100 of the $K'$ genes.

For the rat mixture dataset, we used 30 of the 42 samples as a reference microarray matrix (with multiplicative noise, as in GTEx, to simulate a batch effect) and deconvolved on the remaining 12 samples in the target matrix. In the remaining three datasets, we obtained normalized RNA-seq reference matrices from The Cancer Genome Atlas[180]: TCGA-BRCA breast tumor expression for the breast cancer cell line mixture, TCGA-PRAD prostate tumor expression for the prostate tumor microarray study, and TCGA-LUAD for the lung adenocarcinoma mixing study.

### 5.2.3 Benchmarking in Carolina Breast Cancer Study

We lastly used expression data from the CBCS for validation and analysis[122,24]. Paraffin-embedded tumor blocks were requested from participating pathology laboratories for each samples, reviewed, and assayed for gene expression using the NanoString nCounter system, as discussed previously[24]. As described before[102,224], the expression data was pre-processed and normalized using quality control steps from the NanoStringQCPro package, upper quartile normalization using DESeq2[99,16], and estimation and removal of unwanted technical variation using the RUVSeq and limma packages[29,130]. The resulting normalized dataset comprised of samples from 1,199 patients (628 women of African descent and 571 women of European descent). A study pathologist analyzed tumor microarrays (TMAs) from 148 of the 1,199 patients to estimate area of dissections originating from epithelial tumor, intratumoral stroma, immune infiltrate, and adipose tissue[102]. These cell-type proportions of the 148 samples were used for benchmarking of DeCompress against other reference-free methods.

## 5.3 Results

### 5.3.1 Benchmarking DeCompress with reference-free deconvolution methods

We benchmarked DeCompress performance across 6 datasets (see **Supplemental Table S7** ): (1) *in-silico* mixing experiments using tissue-specific expression profiles from the Genotype-Tissue Expression (GTEx) Project[222,223], (2) expression from 4 published datasets with known cell-type proportions[6,5,7,8], and (3) and tumor expression from the Carolina Breast Cancer Study[122,24]. We compared the performance of DeCompress against 5 other reference-free deconvolution methods (**Supplemental Table S6**): deconf[12], Linseed[13], and DeconICA[17], iterative non-negative matrix factorization with feature selection using TOAST (TOAST + NMF)[14], and CellDistinguisher[15]. Estimated cell-type proportions are compared to simulated or reported true cell-type proportions by calculate the mean square error (MSE) between the two matrices. In total, we observed that DeCompress best recapitulates cell-type proportions compared to other reference-free deconvolution methods.

#### 5.3.1.1 In-silico GTEx mixing

We generated artificial targeted panels by mixing median tissue specific expression profiles from GTEx *in-silico* with randomly simulated cell-type proportions for mammary tissue, EBV-transformed lymphocytes, transformed fibroblasts, and subcutaneous adipose. We added multiplicative noise to the mixed expression to simulate measurement error and contributions to the bulk expression signal from other sources at two levels. **Figure 5.2A** shows the performance of DeCompress compared to other reference-free methods across 25 simulated targeted panels of increasing sample sizes and increasing number of genes from GTEx *in-silico* mixing experiments. In general, we find that DeCompress gives more accurate estimates of cell-type proportions than the other 5 methods at both settings for multiplicative noise. As the number of genes in the targeted panel increased, we largely see the difference in MSE between DeCompress and the other methods increase. Linseed and DeconICA, methods that search of mutually independent axes of variation that correspond to cell-types, consistently perform poorly on these simulated datasets. deconf, TOAST + NMF (matrix factorization-based methods) and CellDistinguisher (topic modeling) perform similarly to one another and only moderately worse in comparison to DeCompress.

We also investigated how the number of component cell-types affects the performance of all six reference-free methods. We generated another set of *in-silico* mixed targeted panels (500 genes) using 2 (mammary tissue and lymphocytes), 3 (mammary, lymphocytes, fibroblasts), and 4 (mammary, fibroblasts, lymphocytes, and adipose) and applied all six methods to estimate the cell-type proportions. **Figure 5.2B** provides boxplots of the MSE across 25 simulated targeted panels using DeCompress and the other 5 benchmarked methods. For all 6 methods, the median MSE for these datasets remained similar as the number of cell-types increased, though the variance in the MSE decreases considerably. In particular, the performance of DeconICA increases considerably as more cell-types were used for mixing, as highlighted by their documentation[17]. Here again, we found that DeCompress gave the smallest median MSE between the true and estimated cell proportions. In total, results from these *in-silico* mixing experiments show both the accuracy and precision of DeCompress in estimated cell-type proportions.

#### 5.3.1.2 Publicly available datasets

Although *in-silico* mixing experiments with GTEx data showed strong performance of DeCompress, we sought to benchmark DeCompress against reference-free methods in previously published datasets with known cell-type mixture proportions. We downloaded expression data from a

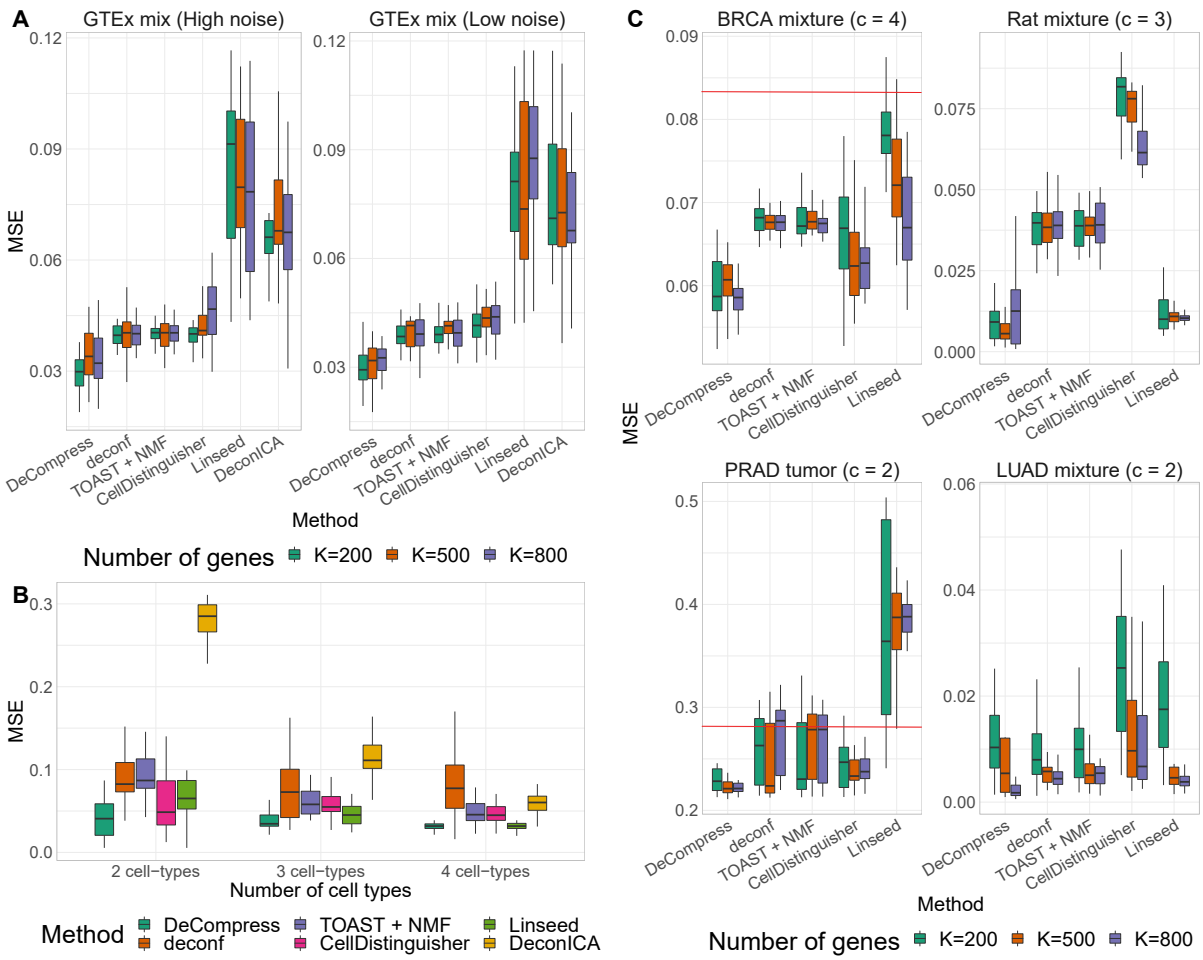


**Figure 5.2:** *Benchmarking results for in-silico GTEx mixing experiments and real data examples.* (A) Boxplots of mean square error ($Y$-axis) between true and estimated cell-type proportions in *in-silico* GTEx mixing experiments across simulated targeted panels of 200, 500, 800, and 1,000 genes ($X$-axis), with 25 simulated datasets per number of genes. GTEx mixing was done at two levels of multiplicative noise, such that errors were drawn from a Normal distribution with zero mean and standard deviation 8 (left) and 4 (right). Boxplots are colored by the benchmarked method (legend at bottom). (B) Boxplots of MSE ($Y$-axis) between true and estimated cell-type proportions over 25 simulated GTEx mixed expression datasets with 500 genes, multiplicative noise drawn from a Normal distribution with zero mean and standard deviation 10, and 2 (left), 3 (middle), and 4 (right) different cell-types. Boxplots are collected by the benchmarked method. (C) Boxplots of mean square error ($Y$-axis) between true and estimated cell-type proportions in 25 simulated targeted panels of 200, 500, 800, and 1,000 genes ($X$-axis), using four different datasets: breast cancer cell-line mixture (top-left)[5], rat brain, lung, and liver cell-line mixture (top-right)[6], prostate tumor samples (bottom-left)[7], and lung adenocarcinoma cell-line mixture (bottom-right)[8]. Boxplots are colored by the benchmarked method. The red line indicates the median null MSE when generating cell-type proportions randomly. If a red line is not provided, then the median null MSE is above the scale provided on the $Y$-axis.

breast cancer cell-line mixture (RNA-seq)[5], rat brain, lung, and liver cell-line mixture (microarray)[6], prostate tumor with cell-type proportions estimated with laser-capture microdissection (microarray)[7], and lung adenocarcinoma cell-line mixture (RNA-seq)[8] and generated pseudo-targeted panels with 200, 500, 800, and 1000 genes. For the rat mixture dataset, we trained the compression sensing model on a randomly selected training split; for the other three cancer-related datasets, reference RNA-seq data was downloaded from The Cancer Genome Atlas (TCGA)[180]. We then performed reference-free deconvolution in these datasets using DeCompress and the other reference-free methods.

Overall, DeCompress showed the lowest MSE across all three datasets, in comparison to the other reference-free methods (**Figure 5.2C**). The patterns observed in the GTEx results are evident in these real datasets, as well. As the number of genes in the targeted panel increases, the variance in the distribution of cell-type proportions decreases. Deconvolution using Linseed gave variable performance across datasets, with very precise estimates of MSE in the rat microarray and lung adenocarcinoma datasets while highly variable estimates in the breast cancer and prostate cancer datasets. We do not present DeconICA in these comparisons due to its large errors across all datasets (see **Supplemental Figure S34** for comparisons to DeconICA). Specific to DeCompress, we assessed the performance of different deconvolution methods (4 reference-free methods and unmix from the DESeq2 package[16]) on the decompressed expression matrix for the breast, prostate, and lung cancer datasets (**Supplemental Figure S35**). We found that unmix gives accurate estimates of cell-type proportions in the breast cancer and prostate tumor datasets, where the component cell-types are like those in bulk tumors. However, in the case of the lung adenocarcinoma mixing dataset (mixture of two lung cancer cell lines), unmix performs poorly, perhaps owing to a dissimilarity to the TCGA-LUAD reference. We lastly investigated the scenario when the reference and target assays measure different bulk tissue. Using the breast cancer cell-line mixtures pseudo-targets and a TCGA-LUAD reference, DeCompress estimated cell-type proportions with larger errors, such that the distribution of MSEs intersect with a null distribution of MSEs from randomly generated cell-type proportion matrices (**Supplemental Figure S36**).

Carolina Breast Cancer Study (CBCS) expression We finally benchmarked DeCompress against the other 5 reference-free deconvolution methods in breast tumor expression data from the Carolina Breast Cancer Study (CBCS)[122,24] on 406 breast cancer-related genes on 1,199 samples. We used RNA-seq breast tumor expression from TCGA to train the compression matrix for deconvolution in CBCS using DeCompress; 393 of the 406 genes on the CBCS panel were measured in TCGA-BRCA. For validation, a study pathologist analyzed 148 tumor microarrays (TMAs) to estimate cell-type

proportions for epithelial tumor, adipose, stroma, and immune infiltrate, which we treat here as a "gold standard."

To determine whether the decompressed expression matrix accurately predicts expression for samples in the target, we split the 393 genes into 5 groups and trained TCGA-based predictive models of genes in each group using those in the other four. Overall, in-sample cross-validation prediction per-sample in TCGA is strong (median adjusted $R^2 = 0.53$), with a drop-off in out-sample performance in CBCS (median adjusted $R^2 = 0.38$), shown in **Figure 5.3A**. We also trained models stratified by estrogen-receptor (ER) status, a major, biologically-relevant classification in breast tumors[107,106]. These ER-specific models showed slightly better out-sample performance (median adjusted $R^2 = 0.34$), though in-sample performance was similar to overall models with the same median $R^2$ (**Figure 5.3B**). Next, as in the GTEx mixing simulations and the 4 published datasets, DeCompress recapitulated true cell-type proportions with the minimum error (**Figure 5.3B**), approximately 33% less error than TOAST + NMF, the second-most accurate method. To provide some context to the magnitude of these errors, we randomly generated 10,000 cell-type proportion matrices for 148 samples and 4 cell-types. The mean MSE is provided in **Figure 5.3A**, showing that 2 of the 5 benchmarked methods (CellDistinguisher and DeconICA) exceeded this randomly generated null MSE value. We also observed that correlations between true and DeCompress-estimated cell-type proportions are positive and significantly non-zero for three of four cell-type components (**Figure 5.3C**). Unlike those from TOAST + NMF, DeCompress estimates of compartment-specific cell-type proportions were positively correlated with the truth ( **Figure 5.3C** and Supplemental Figure S37).

### 5.3.2 Comparison of computational speed

The computational cost of DeCompress is high, owing primarily to training the compressed sensing models. Non-linear estimation of the columns of the compression matrix is particularly slow (**Supplemental Figure S38**). In practice, we recommend running an elastic net method (LASSO, elastic net, or ridge regression) which are both faster (**Supplemental Figure S32**) and give larger cross-validation $R^2$ (**Supplemental Figure S33**). The median cross-validation $R^2$ for elastic net and ridge regression is approximately 16% larger than least angle regression and LASSO, and nearly 25% larger than the non-linear optimization methods. Using CBCS data with 1,199 samples and 406 genes, we ran all benchmarked deconvolution methods 25 times and recorded the total runtimes (**Supplemental Figure S38**). For DeCompress, we used TCGA-BRCA data with 1,212 samples as the reference. As shown in **Supplemental Figure S38**, running DeCompress in serial (approximately 62

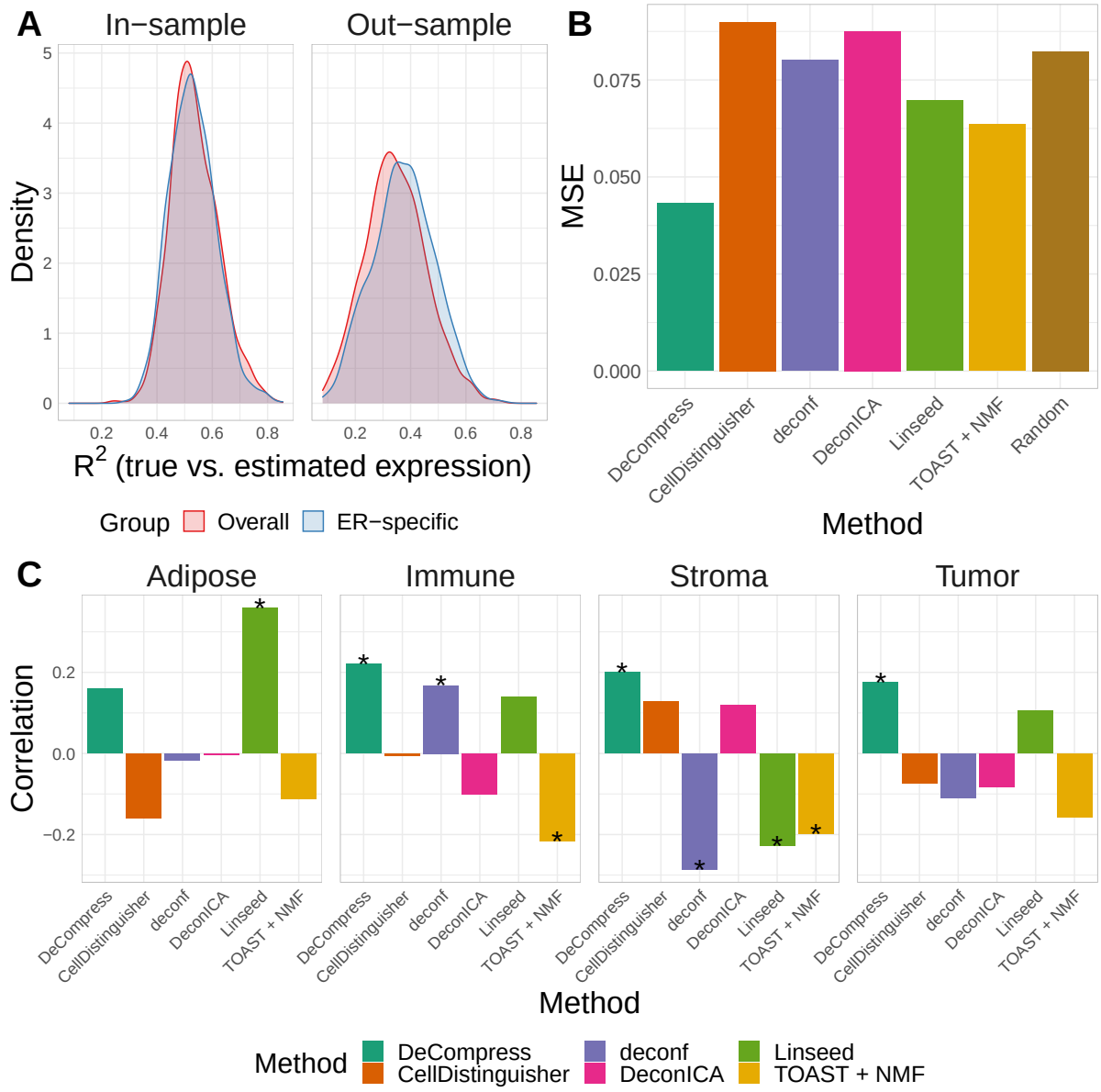


**Figure 5.3:** *Benchmarking results with Carolina Breast Cancer Study expression data.* (A) Kernel density plots of predicted adjusted $R^2$ per-sample in in-sample TCGA prediction (left) through cross-validation and out-sample prediction in CBCS (right), colored by overall and ER-specific models. (B) MSE ($Y$-axis) between true and estimated cell-type proportions in CBCS across all methods ($X$-axis). Random indicates the mean MSE over 10,000 randomly generated cell-type proportion matrices. (C) Spearman correlations ($Y$-axis) between compartment-wise true and estimated proportions across all benchmarked methods ($X$-axis). Correlations marked with a star are significantly different from 0 at $P < 0.05$.

minutes) takes around 40 times longer than the slowest reference-free deconvolution method (TOAST + NMF, approximately 1.5 minutes), though DeCompress can comparable in runtime to TOAST + NMF if run in parallel with enough workers (approximately 2.6 minutes). These computations were conducted on a high-performance cluster (RedHat Linux operating system) with 25 GB of RAM.

### 5.3.3 Applications of DeCompress in the Carolina Breast Cancer Study

Given the strong performance of DeCompress in benchmarking experiments, we estimated cell-type proportions for 1,199 subjects in CBCS with transcriptomic data assayed with NanoString nCounter. Using TCGA-BRCA expression as a training set, we iteratively searched for cell type-specific features[14] (Step 1 in Figure 5.1) and included canonical cell-type markers for guidance using a priori knowledge[225–227]. After expanding the targeted CBCS expression to these genes, we estimated proportions for 5 compartments. As reference-free methods output proportions for agnostic compartments, identifying approximate cell-types for compartments is often difficult. Here, we first outline a framework for assigning modular identifiers for compartments identified by DeCompress, guided by compartment-specific gene signatures. Then, we present some advantages of using compartment-specific proportions in downstream analyses of breast cancer outcomes and gene regulation.

Date of death and cause of death were identified by linkage to the National Death Index. All diagnosed with breast cancer have been followed for vital status from diagnosis until date of death or date of last contact. Breast cancer-related deaths were classified as those that listed breast cancer (International Statistical Classification of Disease codes 174.9 and C-50.9) as the underlying cause of death on the death certificate. Of the 1,199 samples deconvolved, 1,153 had associated survival data with 330 total deaths, 201 attributed to breast cancer.

#### 5.3.3.1 Identifying approximate cell-types for compartments

We leveraged compartment-specific gene signatures to annotate each compartment with modular identifiers. First, we computed Spearman correlations between the compartment-specific gene expression profiles and median tissue-specific expression profiles from GTEx[222,223] and single cell RNA-seq profiles of MCF7 breast cancer cells[228] (**Figure 5.4A**). Here, we find that Compartment 4 (C4) shows strong positive correlations with fibroblasts, lymphocytes, multiple collagenous organs (such as blood vessels, skin and the colon[229]), and MCF7 cells. The C3 gene signature was significantly correlated with expression profiles of secretory organs (salivary glands, pancreas, liver)

and contained a strong marker of HER2-enriched breast cancer (*ERBB2*)[230]. In fact, we see significant Spearman correlations between C3 and C4 proportions and ER and HER2 scores[109], scores that represent over-expression of genes up-regulated in ER-positive and HER2-enriched breast tumors; namely, the strong positive correlation ($\rho = 0.53$) between C3 proportion and HER2 scores provided more evidence that C3 may be a HER2-enriched tumor compartment.

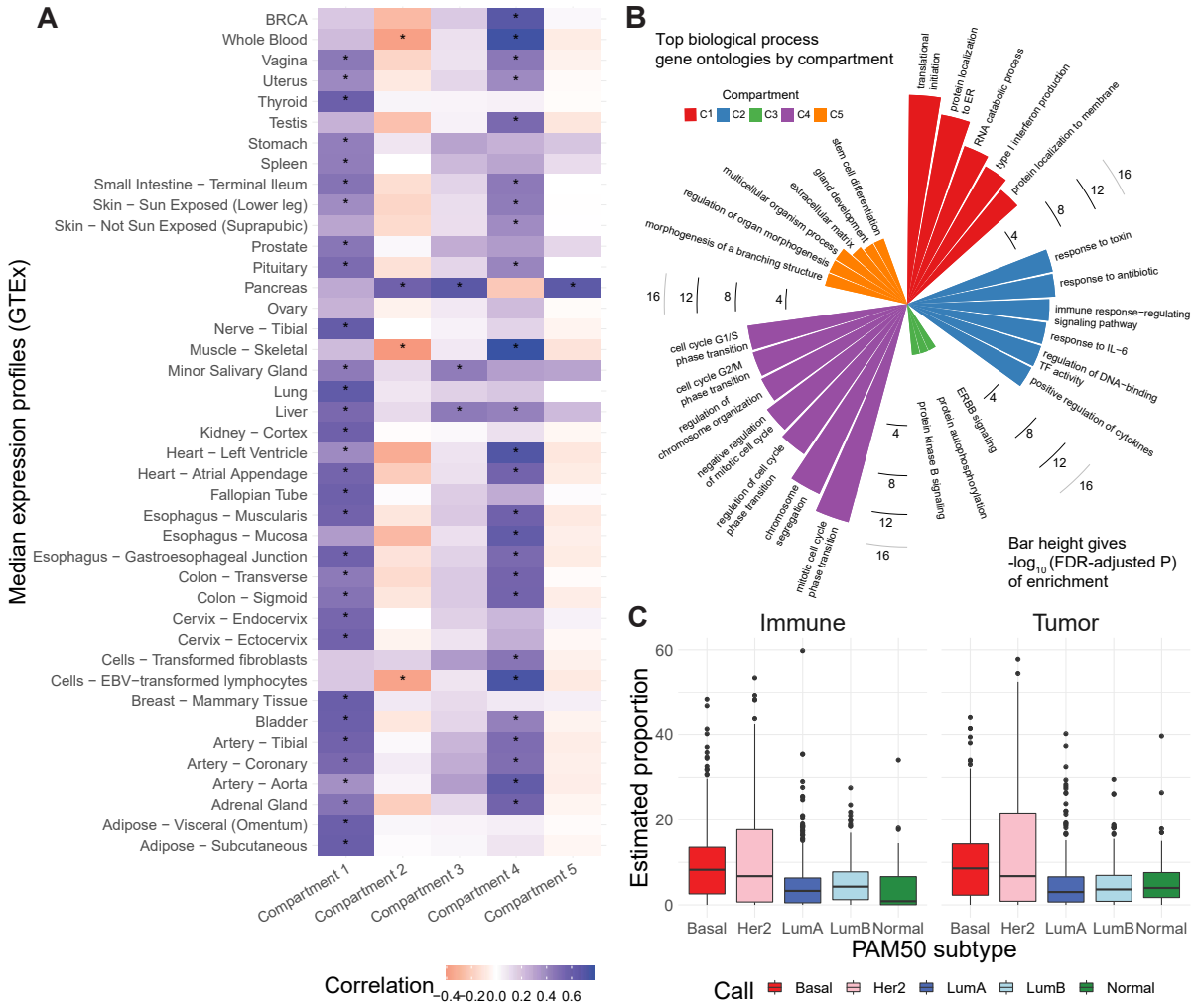


**Figure 5.4:** *Identification of Decompress-estimated compartments.* (A) Heatmap of Pearson correlations between compartment-specific gene signatures ($X$-axis) and GTEx median expression profiles and MCF7 single-cell profiles ($Y$-axis). Significant correlations at nominal $P < 0.01$ are indicated with an asterisk. (B) Barplot of $-\log_{10}$ FDR-adjusted $P$-values for top gene ontologies ($Y$-axis) enriched in compartment-specific gene signatures. (C) Boxplots of estimated immune (left) and tumor (C3 and C4 compartments, right) proportions ($Y$-axis) across PAM50 molecular subtypes ($X$-axis)

We conducted over-representation analysis (ORA)[231] of gene signatures for all five compartments, revealing cell cylce regulation ontologies for C3 that are consistent with the hypothesis generated from GTEx profiles at FDR-adjusted $P < 0.05$ (**Figure 5.4B**) We conducted gene set enrichment analysis (GSEA) for the C4 gene signature[232], revealing significant enrichments for cell differentiation and development process ontologies (**Supplemental Figure S9**). ORA analysis also assigned immune-related ontologies to the C2 gene signatures at FDR-adjusted $P < 0.05$ and ERBB signaling to C4, though these enrichments did not achieve statistical significance. C1 and C5 gene signatures were not enriched for ontologies that allowed for conclusive cell-type assignment, showing catabolic, morphogenic and extracellular process ontologies (**Figure 5.4B**). From these results, we hypothesized that C3 and C4 resembled epithelial tumor cells, C2 resembled an immune compartment, and C1 and C5 resembled stromal and mammary tissue.

Distributions of hypothesized immune (C2) and tumor (C3 and C4 proportions) revealed significant differences across PAM50 molecular subtypes (**Figure 5.4C**; Kruskal-Wallis test of differences with $P < 2.2 \times 10^{-16}$)[109]. These trends across subtypes were consistent with *a priori* knowledge, as well: Basal and HER2-enriched subtypes, the most aggressive subtypes, had the largest proportions of the estimated tumor and immune compartments, while Luminal A, Luminal B, and Normal-like subtypes showed lower proportions[233,109,132]. Furthermore, we found strong differences in C4 and total tumor compartment estimates across race (**Supplemental Figure S40A**) C3 and C4 also have strong correlations with ER- (estrogen receptor) and HER2-scores, gene-expression based continuous variables that indicate clinical subtypes based on *ESR1* and *ERBB2* gene modules (**Supplemental Figure S40B**); however, none of the C3, C4, immune, or tumor compartment estimates showed significant differences across clinical ER stautus determined by immunohistochemistry (**Supplemental Figure S40C**).

### 5.3.4 Incorporating estimated compartment improves outcome prediction

Next, we considered the impact of including the tumor (C3, C4, and combining C3/C4) and immune (C2) compartments in survival models. We constructed Cox models for breast-cancer specific mortality[154] with the following covariates: race, age, PAM50 molecular subtype, compartment proportion, and an interaction between subtype and compartment proportion. **Supplemental Table S8** shows hazard ratio estimates and 90% FDR-adjusted confidence intervals[234] from Cox models with the C3, C4, tumor, and immune compartments, along with comparisons to a reduced baseline model that excludes the compartment estimates and interaction terms. General relationships stay similar

across the baseline and interaction models (e.g. protective hazard ratios of Luminal A subtypes in comparison to the reference Basal subtypes). We also estimated, in the C4-compartment interaction model, that increased C4 proportion was associated with shorter survival (hazard ratio 1.69, FDR-adjusted $P = 0.026$). We also compared these compartment-specific interaction models with the nested baseline model that did not contain the compartment proportions using a partial likelihood ratio test. We found that only the interaction model with the C4 proportions gave a significantly better model fit ($\chi^2 = 11.52$ on 4 degrees of freedom, $P = 0.02$). Estimated survival Kaplan-Meier curves stratified by molecular subtype and median-stratified C3 and C4 proportions showed significant differences between low and high proportion groups within molecular subtypes (**Supplemental Table S8**). Namely, we observed that the C3 high and low proportion groups only split the HER2-enriched molecular subtype based on survival outcomes, reinforcing the ERBB signaling annotations assigned to C3 in ORA analysis. However, the HER2-enriched subtype was enriched for C3-high samples (127 out of 147 samples in the C3-high group). We also found that the C4 groups split the Basal and Luminal B subtype groups, though the Basal subtype was disproportionately enriched for C4-high subjects (315 out of 339 subjects). In sum, these results illustrate that incorporating computationally-derived estimates of compartments may aid in outcome prediction.

#### 5.3.4.1 Incorporating compartment proportions into eQTL models detects more tissue-specific gene regulators

We investigated how incorporating estimated compartment proportions affect cis-eQTL mapping in breast tumors, a common application of deconvolution methods in assessing sources of variation in gene regulation[85,235]. In previous eQTLs studies using CBCS expression, several bulk breast tumor cis-eGenes were found in healthy mammary, subcutaenous adipose, or lymphocytes from GTEx[102]. We included DeCompress proportion estimates for the tumor (C3 and C4 estimates) and immune (C2) compartments in a race-stratified, genetic ancestry-adjusted cis-eQTL interaction model, as proposed by Geeleher et al and Westra et al[85,84]. We found that sets of compartment-specific cis-eGenes generally had few intersections with bulk cis-eGenes (**Figure 5.5A**), but we detected more eQTLs in tumor- and immune-specific compartments (**Supplemental Figure S41**). At FDR-adjusted $P < 0.05$, of 209 immune-specific cis-eGenes identified in women of European ancestry (EA), 7 were also mapped in the bulk models (with no compartment proportions covariates), and no tumor-specific cis-eGenes were identified with the bulk models. Similarly, at FDR-adjusted $P < 0.05$, in women of African ancestry (AA), 27 of 331 and 9 of 124 cis-eGenes identified with the immune- and tumor-compartment interaction models were also mapped with the bulk models, respectively.

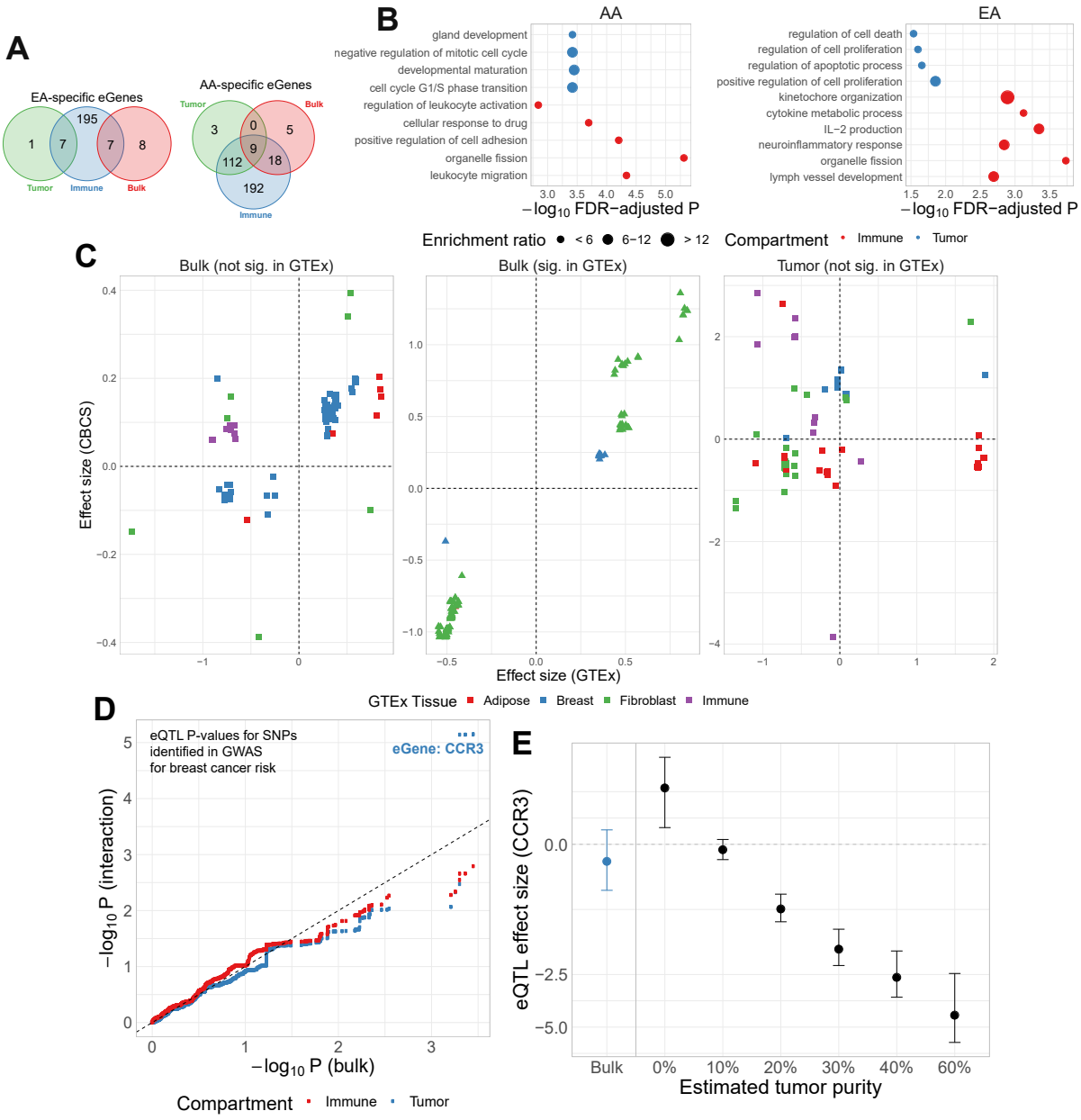


**Figure 5.5:** *Compartment-specific cis-eQTL mapping in the Carolina Breast Cancer Study.* (A) Venn diagram of bulk, tumor-, and immune-specific cis-eGenes identified European-ancestry (left) and African-ancestry samples (right) in CBCS. (B) Enrichment analysis of immune- (red) and tumor-specific (blue) cis-eGenes in CBCS plotting the $-log_{10}$ $P$-value of enrichment ($X$-axis) and description of gene ontologies ($Y$-axis). The size of the point represents the relative enrichment ratio for the given ontology. (C) Scatterplots of GTEx ($X$-axis) and CBCS effect size ($Y$-axis) for significant CBCS cis-eQTLs that were mapped in GTEx. Each point is colored by the GTEx tissue in which the cis-eQTL has the lowest $P$-value. Reference dotted lines for the $X$- and $Y$-axes are provided. (D) For risk variants from GWAS for breast cancer from iCOGs[9–11], scatterplot of $-log_{10}$ P-values of bulk ($X$-axis) and compartment-specific cis-eQTLs ($Y$-axis), colored blue for tumor- and red for immune-specific models. A 45-degree reference line is provided. In the top right corner, 3 tumor-specific cis-eQTLs are labeled with the eGene CCR3 as they are significant at FDR-adjusted $P < 0.05$. (E) Tumor-specific eQTL effect sizes and 95% confidence intervals ($Y$-axis) for rs56387622 on CCR3 expression across various estimates of tumor purity. The eQTL effect size from the bulk model is given in blue.

Manhattan plots for cis-eQTLs across the whole genome across bulk, tumor, and immune show the differences in eQTL architecture in these compartment-specific eQTL mappings in EA and AA samples (**Supplemental Figures S42** and **S43**, respectively). Furthermore, we generally detected more cis-eQTLs at FDR-adjusted $P < 0.05$ with the immune-specific interactions than the bulk and tumor-specific interactions (EA: 565 bulk cis-eQTLs, 65 tumor cis-eQTLs, 8927 immune cis-eQTLs; AA: 237 bulk cis-eQTLs, 449 tumor cis-eQTLs, 7676 immune cis-eQTLs; **Supplemental Figure S41**).

We analyzed the sets of EA and AA tumor- and immune-specific eGenes in CBCS with ORA analysis for biological processes (**Figure 5.5B**). We found that, in general, these sets of eGenes were concordant with the compartment in which they were mapped. All at FDR-adjusted $P < 0.05$, AA tumor-specific eGenes showed enrichment for cell cycle and developmental ontologies, while immune-specific eGenes were enriched for leukocyte activiation and migration and response to drug pathways. Similarly, EA tumor-specific eGenes showed enrichments for cell death and proliferation ontologies, and immune-specific eGenes showed cytokine and lymph vessel-associated processes. These results from cis-eQTL analysis provide an example of the advantage of including DeCompress-estimated compartment proportions in downstream genomic analyses.

We then cross-referenced bulk and tumor-specific cis-eGenes found in the CBCS EA sample with cis-eGenes detected in healthy tissues from GTEx: mammary tissue, fibroblasts, lymphocytes, and adipose, similar to previous pan-cancer germline eQTL analyses[102,137]. We attributed several of the bulk cis-eGenes to healthy GTEx tissue (all but 2), but tumor specific cis-eGenes were less enriched in these healthy tissues (**Supplemental Figure S44**). We compared the cis-eQTL effect sizes for significant CBCS cis-eSNPs found in GTEx. As shown in **Figure 5.5C**, 98 of 220 bulk cis-eQTLs detected in CBCS that were also found in GTEx were mapped in healthy tissue, with strong positive correlation between effect sizes (Spearman $\rho = 0.93$). The remaining 122 eQTLs that could not be detected in healthy GTEx tissue contained some discordance in the direction of effects, though correlations between these effect sizes were also strongly positive ($\rho = 0.71$). In contrast, we were unable to detect any of the CBCS tumor-specific cis-eQTLs in GTEx healthy tissue, and the correlation of these effect sizes across CBCS and GTEx was poor ($\rho = -0.07$).

We next extracted 932 breast cancer risk-associated SNPs in women of European ancestry[9–11] at FDR-adjusted $P < 0.05$ that were available on the CBCS OncoArray panel[124]. **Figure 5.5D** shows the raw $-\log_{10} P$-values of the association of these SNPs with their top cis-eGenes in the bulk and tumor- and immune-specific interaction models. In large part, none of these eQTLs reached FDR-adjusted $P < 0.05$, except for 3 cis-eQTLs, with their strengths of association favoring the bulk eQTLs. However, we detected 3 tumor-specific EA cis-eQTLs in near-perfect linkage disequilibrium of

$r^2 > 0.99$ (strongest association with rs56387622) with the chemokine receptor *CCR3*, previously found to be associated with breast cancer outcomes in luminal-like subtypes[236,237]. As estimated tumor purity increases, the cancer risk allele C at rs56387622 has a consistently stronger negative effect on CCR3 expression (**Figure 5.5E**).

## 5.4 Discussion

Here, we presented DeCompress, a semi-reference-free deconvolution method catered towards targeted expression panels that are commonly used for archived tissue in clinical and academic settings[18,25]. Unlike traditional reference-based methods that require cell-type specific expression profiles, DeCompress requires only a reference RNA-seq or microarray data on similar bulk tissue to train a compressed sensing model that projects the targeted panel into a larger feature space for deconvolution. Such reference datasets are much more widely available than cell-type specific expression on the same targeted panel. We benchmarked DeCompress against reference-free methods[12,13,17,14,15] using *in-silico* GTEx mixing experiments[222,223], 4 published datasets with known cell-type proportions[6,5,7,8], and a large, heterogeneous NanoString nCounter dataset from the CBCS[122,233]. In these analyses, we showed that DeCompress recapitulated true cell-type proportions with the minimum error and the strongest compartment-specific positive correlations, especially when the reference dataset is properly aligned with the tissue assayed in the target. Lastly, we outlined the advantages of incorporating these computationally derived estimates in downstream analyses of survival outcomes and eQTL mapping in breast cancer.

A disadvantage of DeCompress is its computational cost, owing mainly to its lengthy compressing sensing training step. We recommend running mainly linear optimization methods in this step and have implemented parallelization options to bring computation time on par with the iterative framework proposed in TOAST[14]. However, DeCompress estimates cell-type proportions both accurately and precisely, compared to other reference-free methods, and provides a strong computational alternative that is much faster than costly lab-based measurement of composition. Another disadvantage, which also affects reference-based methods, is the proper selection of a reference dataset. As seen in the lung adenocarcinoma example, where TCGA-LUAD data was not an accurate reflection of a mixture of adenocarcinoma cell-lines, DeCompress performance is slightly worse than with datasets with properly matched references. Yet, DeCompress performance is on par with that of the other reference-free methods. The compression model may also be sensitive to phenotypic variation in the

reference, as evidenced by the increase in out-sample prediction $R^2$ in ER-specific models compared to overall models in CBCS. This specificity may be leveraged to train more accurate models by using more than one reference dataset to reflect clinical or biological heterogeneity in the targeted panel.

A universal challenge of reference-free deconvolution methods, like DeCompress, is selecting an appropriate number of compartments. Previous groups have detailed how important *a priori* knowledge is for deconvolving well-studied tissues, such as blood and brain[238,239]. However, diseased tissues, like bulk cancerous tumors, especially in understudied subtypes or populations, are more difficult to deconvolve due to the similarity between compartments, many of which are rare, when comparing across individuals of different subtypes or phenotypes (e.g. activated and inactivated stroma in breast tumors)[106,226,240,108]. For this reason, though DeCompress includes several data-driven approaches in estimating the number of compartments from variation in the gene expression, we recommend applying prior domain knowledge about the tissue of interest. Another challenge for all reference-free methods is assigning gene module-based annotations to the unidentified estimated compartments. Several previous reference-free methods have leveraged *in vitro* mixtures of highly distinct cell lines in training and testing previous reference-free deconvolution methods[6,13], namely the rat cell line mixture (GSE19830). Though this dataset is easy to deconvolve and thus useful in testing methodology, the extreme differences in gene expression between these three tissue types renders this dataset sub-optimal for methods benchmarking. Furthermore, assigning estimated compartments to known tissues in this dataset is straightforward and does not capture how difficult this task in typical deconvolution applications. Instead, our applications in breast cancer expression with CBCS provided such a difficult statistical challenge. Our outlined approach of first comparing compartment-specific gene signatures to known tissue profiles from GTEx or single-cell profiles, then analyzing these signatures with ORA or GSEA, and lastly searching for known biological trends provides a structured framework for addressing the compartment identification problem.

Our downstream eQTL analysis in CBCS breast tumor expression also provided some insight into gene regulation, similar to recent work into deconvolving immune subpopulation eQTL signals from bulk blood eQTLs[235]. In breast cancer, Geeleher et al previously showed that a similarly implemented interaction eQTL model gave better mapping of compartment-specific eQTLs[84,85]. Our results are consistent with this finding, especially since tumor- and immune-specific eGenes were enriched for commonly associated ontologies. However, unlike Geeleher et al, we generally detected a larger number of immune- and tumor-specific eQTLs and eGenes than in the bulk, unadjusted models. We believe that this larger number of compartment-specific eGenes may be due to the specificity of the genes assayed by the CBCS nCounter panel. As the panel included 406 genes, all previously

implicated in breast cancer pathogenesis, proliferation, or response[102,233,241], the interaction model will detect for SNPs that have large effects on cell-type specific genes. The interaction term is interpreted as the difference in eQTL effect sizes between a samples of 0% and 100% of the given compartment; accordingly, for genes implicated in specific breast cancer pathways, we expect to see large differences in cell-type specific eQTL effects[242–244]. Though this interaction model is straight-forward in its interpretation for the tumor compartment (i.e. a sample of 100% tumor cells versus 100% tumor-associated normal cells), this interpretation may be tenuous for less well-defined compartments, like an immune compartment that includes several different immune cells. In addition, we did not consider trans-acting eQTLs that are often attributed to cell-type heterogeneity, though we believe that methods employing mediation or cross-condition analysis can be integrated with compartment estimates to map cell-type specific trans-eQTLs relevant in breast cancer[245,72,67].

Relevant to risk and proliferation of breast cancer, we detected a locus of cis-eSNPs associated with expression of *CCR3* (C-C chemokine receptor type 3) that were GWAS-identified risk SNPs[9–11] but were not significantly associated with *CCR3* expression using the bulk models and were not detected in GTEx. If one or more causal SNPs in this genomic region affects *CCR3* expression only in cancer cells and the effect on *CCR3* expression is the main mechanism by which the locus predisposes individuals to breast cancer, we can hypothesize that an earlier perturbation in the development of cancer (e.g. transcription factor or microRNA activation) may cause this SNP's tumorigenic effect. Given this perturbation in precancerous mammary cells, individuals with the risk allele would convey the tumorigenic effects of decreased CCR3 expression. It has been previously shown that increased peritumoral *CCR3* expression is associated with improved survival times in luminal-like breast cancers[236,237]. The CCR3 receptor has been shown to be the primary binding site of CCL11 (eotaxin-1), an eosinophil-selective chemoattractant cytokine[246,247], and accordingly CCR3 antagonism prohibited chemotaxis of basophils and eosinophils, a phenomenon observed in breast cancer activation and proliferation[248,249]. Without DeCompress and the incorporation of compartment estimated in the eQTL model, this association between eSNP and *CCR3* expression would not have been detected[250].

DeCompress, our semi-reference-free deconvolution method, provides a powerful method to estimate cell-type specific proportions for targeted expression panels that have a limited number of genes that only requires RNA-seq or microarray expression from a similar bulk tissue. Our method's estimates recapitulate known compartments with less error than reference-free methods, and provides compartments that are biologically relevant, even in complex tissues like bulk breast tumors. We provided examples of using these estimated compartment proportions in downstream studies of

outcomes and eQTL analysis. Given the wide applications of reference-free deconvolution, the popularity of targeted panels in both academic and clinical settings, and increasing need for analyzing heterogeneous tissues, we anticipate creative implementations of DeCompress to provide further insight into expression variation in complex diseases.

# CHAPTER 6: CONCLUSION

Here, we proposed several approaches for the analysis of biological data with various sources of variation. In Chapter 2, we propose a framework for the normalization of NanoString nCounter RNA expression data, especially in long-term longitudinal, multi-phase or multi-site cohorts. We compare our iterative framework with the commercially available nSolver software, showing that the nSolver software insufficiently removes technical variation, leading to potentially inflated biological associations due to confounding. In Chapter 3, we conduct a transcriptome-wide association study for breast cancer-specific survival that leverages race-specific, ancestry-adjusted breast tumor eQTLs. We show that the cis-germline genetically regulated expression of many important breast cancer-related genes are different across both race and clinical or molecular subtype. We then identify two novel genetic regions that are associated for breast cancer mortality. This work informs future research into disentangling genetic ancestry differences from subtype heterogeneity in breast cancer.

In Chapter 4, we propose an extension to transcriptomic prediction and association studies by considering distal germline varation. We prioritize distal eQTLs in prediction leading to gains in both expression predictive accuracy and power to detect gene-trait associations. We showed the advantage of this TWAS extension in identifying relevant pathways in both breast cancer proliferation and neuropsychiatric disorder. Our novel extension to test distal associations above and beyond the local genetic-trait association aids in generating hypotheses for potential gene regulatory mechanisms. Future work here is necessary to improve computational efficiency, but this approach is encouraging in tissues or diseases that are governed by complex networks of gene regulation. Lastly, in Chapter 5, we outline a semi-reference-free cell-type deconvolution method for targeted mRNA expression panels. We show the utility of this method over reference-free methods in a variety of settings and use deconvolved expression to better explain survival outcomes and compartment-specific eQTLs in bulk breast cancer tissue. This method can be integrated with other computational approaches to adjust complex genomic analyses for cell-type heterogeneity.

# APPENDIX: SUPPLEMENTAL FIGURES AND TABLES

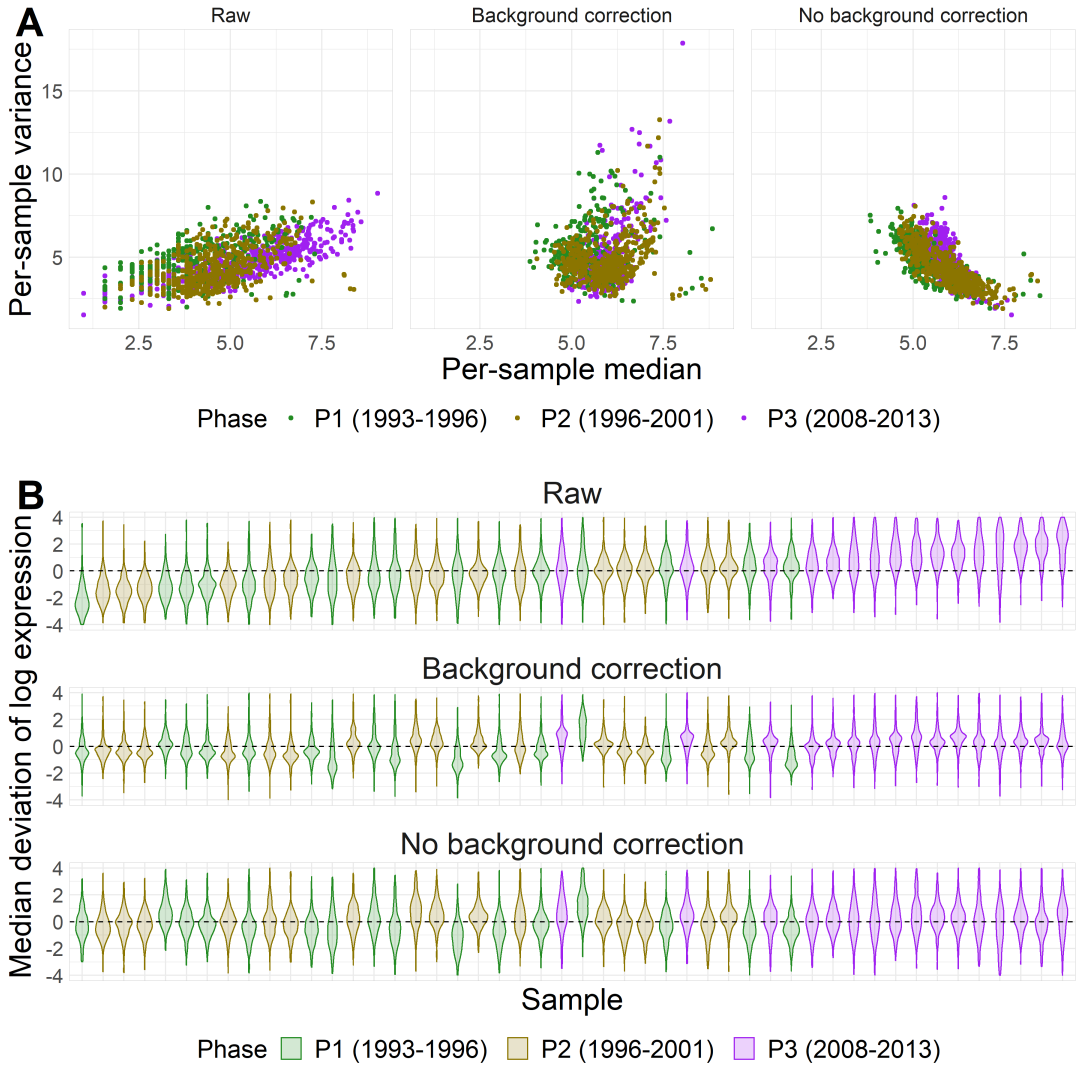


**Figure S1:** *Comparison of per-sample expression with and without background threshold.* (A) Scatter plot of per-sample median and per-sample variance of CBCS expression across raw expression (left), nSolver-normalized data with background correction (middle), and without background correction (right), with samples colored by study phase. (B) Relative log-expression (RLE) plots of raw expression (top), nSolver-normalized expression with background correction (middle), and nSolver-normalized expression without background correction (bottom) for 90 randomly selected CBCS breast cancer samples, ordered from left to right by increasing per-sample median in the raw expression. The dotted line gives a reference for a deviation of 0.

| Software | Implementation | Method | Memory Used (CBCS Used) | Notes |
|---|---|---|---|---|
| nSolver | GUI with R backend | Positive- and housekeeping-control scaling | 0.41 GB (using NanoStringNorm) | Can be implemented entirely in R using the NanoStringNorm package |
| RUVSeq | R pacakge (Bioconductor) | Factor analysis on housekeeping genes or technical replicates | 3.92 GB | RUV-III has been created by the same group for NanoString data using technical replicates |
| NanoStringDiff | R package (Bioconductor) | Generalized linear model | 0.37 GB | Normalized data is recommended to be used only for downstream differential expression analysis using an empirical Bayes shrinkage approach |
| RCRnorm | R package (CRAN) | Bayesian random-coefficient hierarchical regression | 20.93 GB | High computational cost, even in fast mode |

**Table S1:** Summary of normalization software compared in benchmarking. We provide the implementation of the software, a brief summary of the methods used by the software, total memory used on a submitted job on a high performance cluster with 25 GB allocated RAM, and any miscellaneous notes about the methods (i.e. alternative implementations and disadvantages of each method). The memory used is calculated from a submitted job that processed the CBCS expression data (417 genes, 1264 samples).

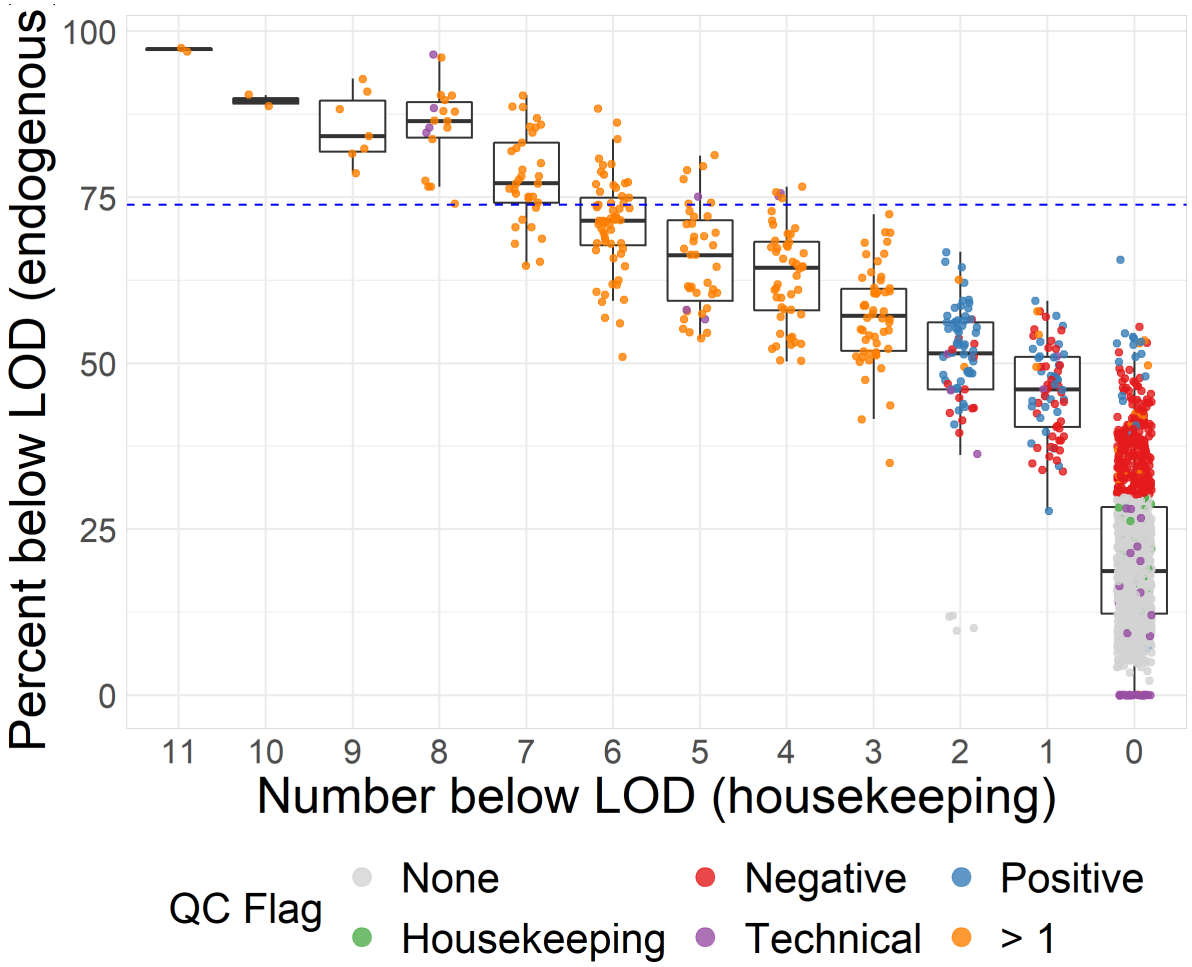


**Figure S2:** *Comparison of quality control flags and sample quality in CBCS.* Boxplot of percent of zero-counts in endogenous genes (Y-axis) over varying numbers of zero-counts in the 11 housekeeping genes (X-axis), colored by various QC flags.

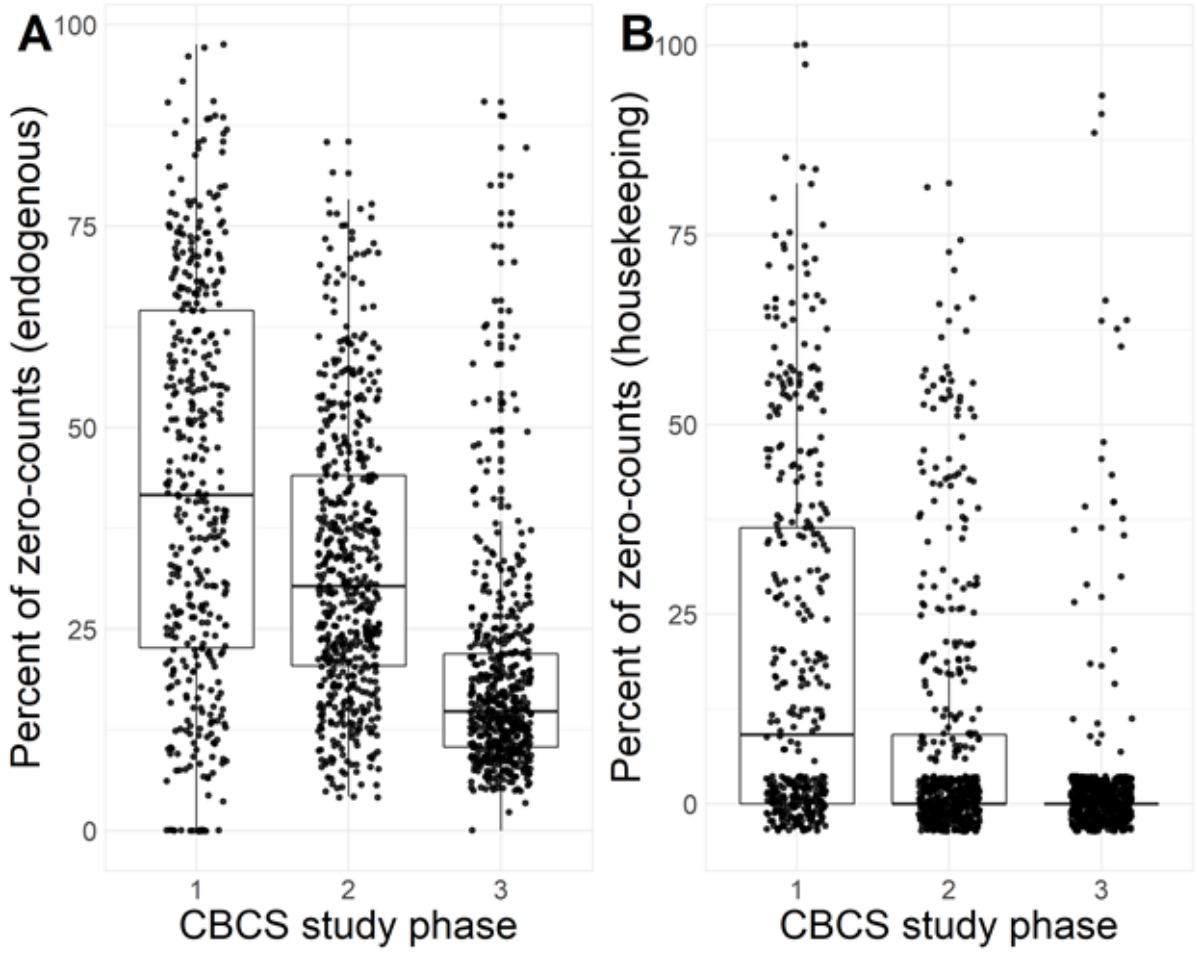


**Figure S3:** *Comparison of sample quality with sample age in CBCS.* Boxplots of percent of zero-counts per sample by CBCS study phase with percent of zero-counts of 406 endogenous genes (A) and percent of zero-counts of 11 housekeeping genes (B).

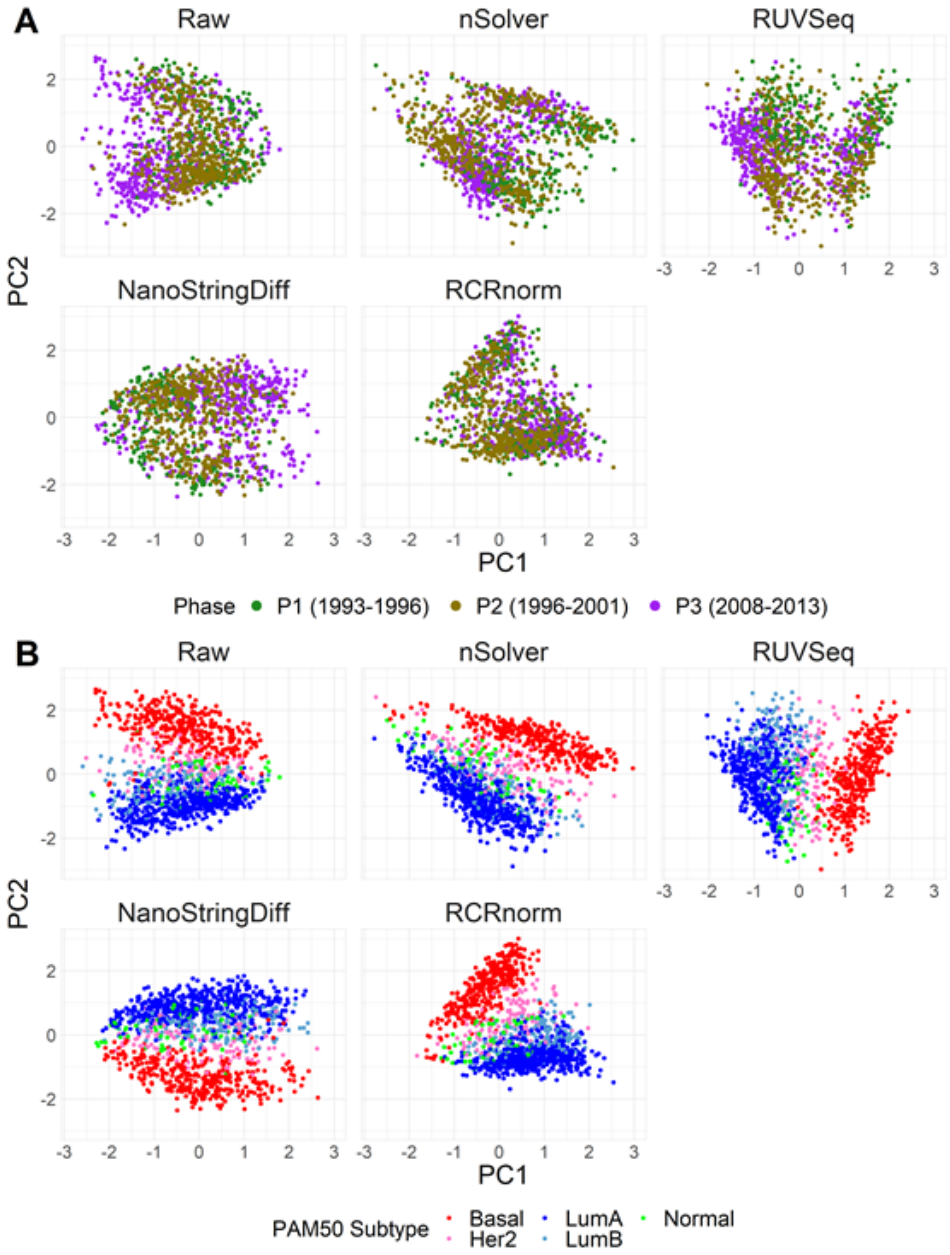


**Figure S4:** *Comparison of normalization methods on reflecting technical and biological variables.* Scatter plots of first two principal components of raw, nSolver-, RUVSeq-, NanoStringDiff-, and RCRnorm-normalized CBCS expression data colored by study phase (A) and PAM50 subtype call (B). PC1 ($X$-axis) captures the maximum variation in expression (approximately 9-12% across all datasets), and PC2 ($Y$-axis) captures the second most (approximately 3-4%).

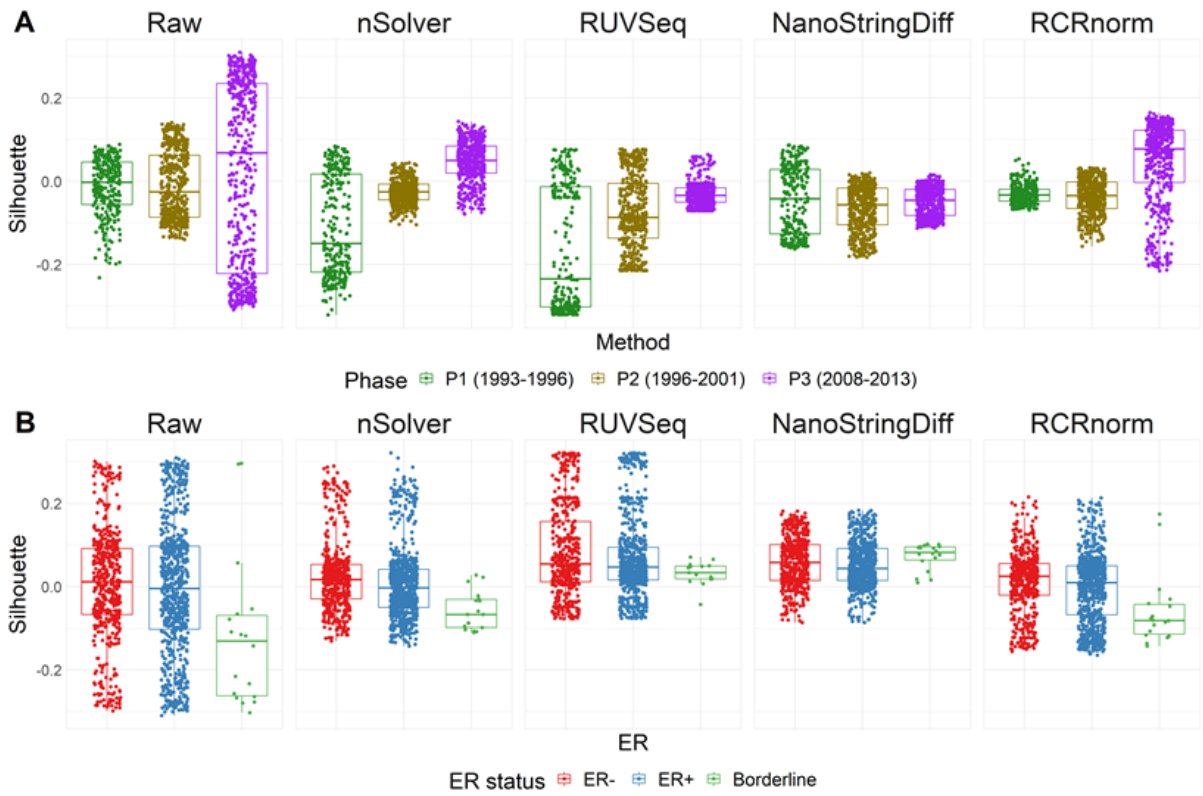


**Figure S5:** *Silhouette analysis of normalized data acros study phase and ER stuatus* Boxplots of silhouette widths of raw, nSolver-, RUVSeq-, NanoStringDiff-, and RCRnorm-normalized CBCS expression data colored by ER status (A) and study phase (B).

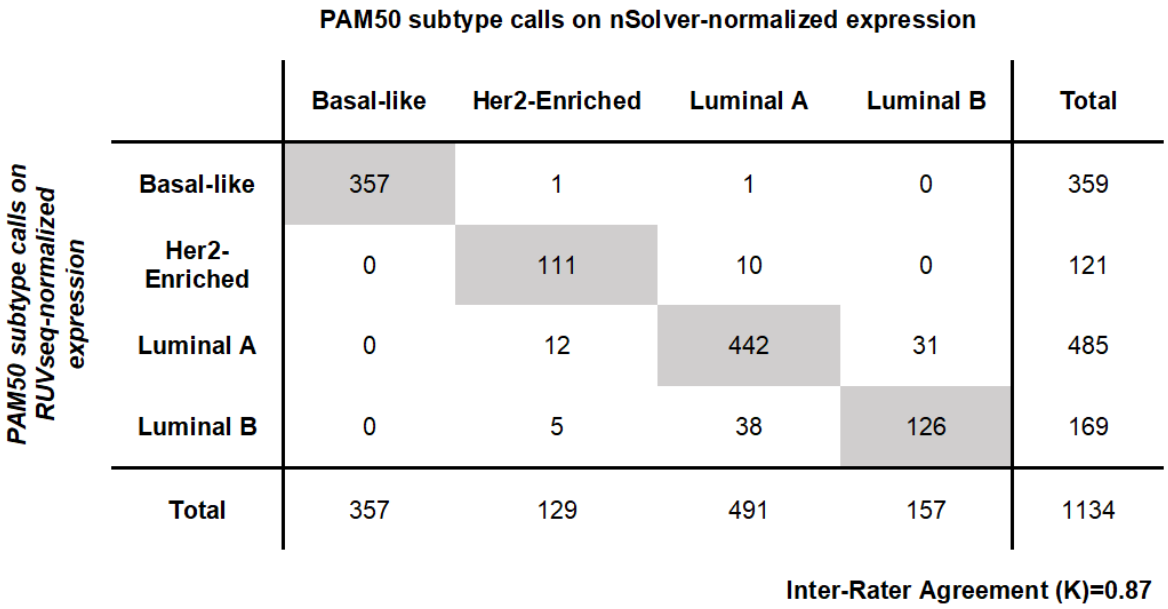


**Figure S6:** *Confusion matrix of PAM50 calls using nSolver-normalized and RUVSeq-normalized expression*

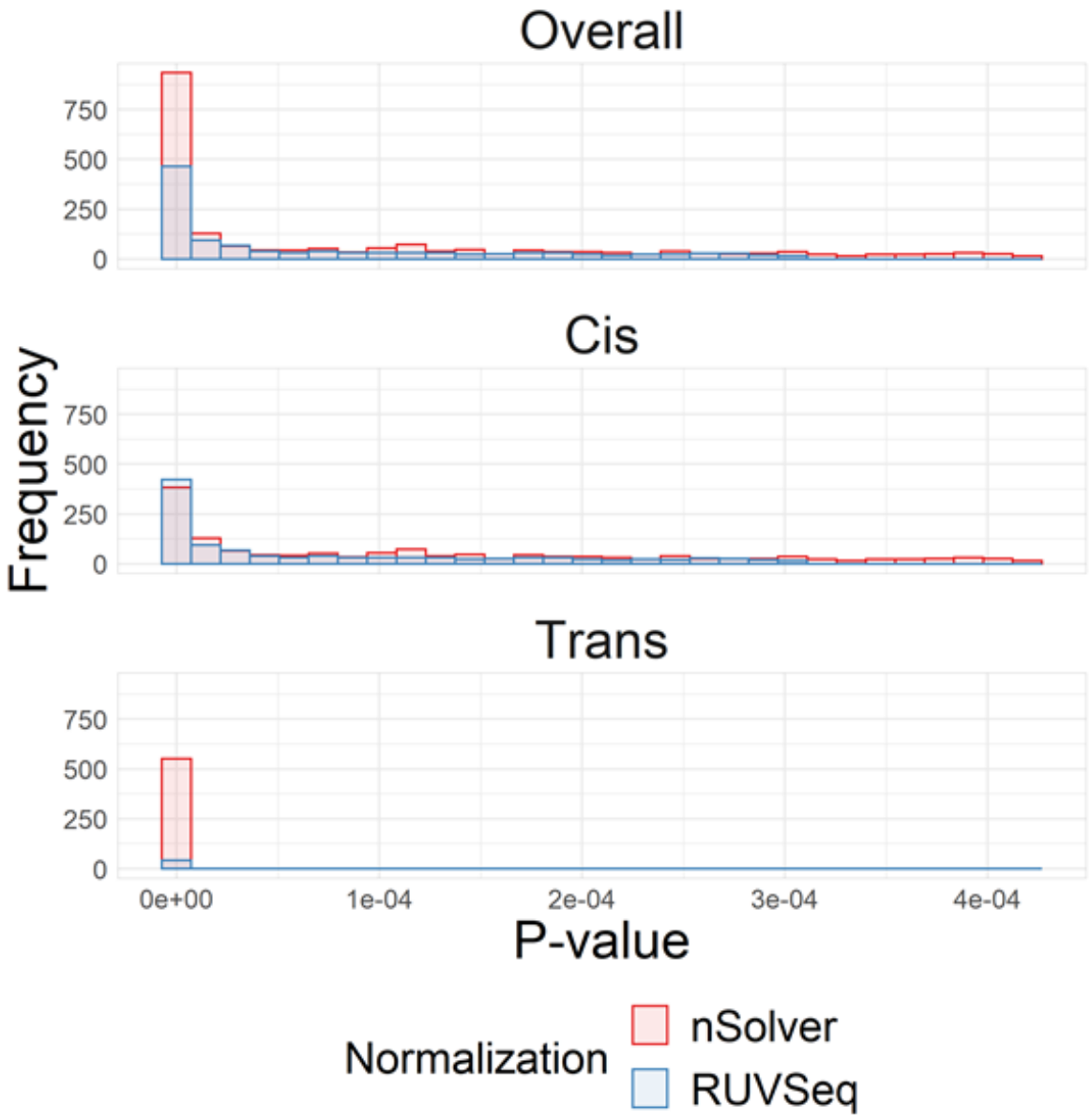


**Figure S7:** *Gene expression patterns across normalization methods in CBCS.* Histograms of raw $P$-values of eQTL associations using nSolver-normalized (red) and RUVSeq-normalized (blue) data across overall (top), cis-eQTLs only (middle), and trans-eQTLs only (bottom) for eQTL associations with FDR-adjusted $P < 0.05$.

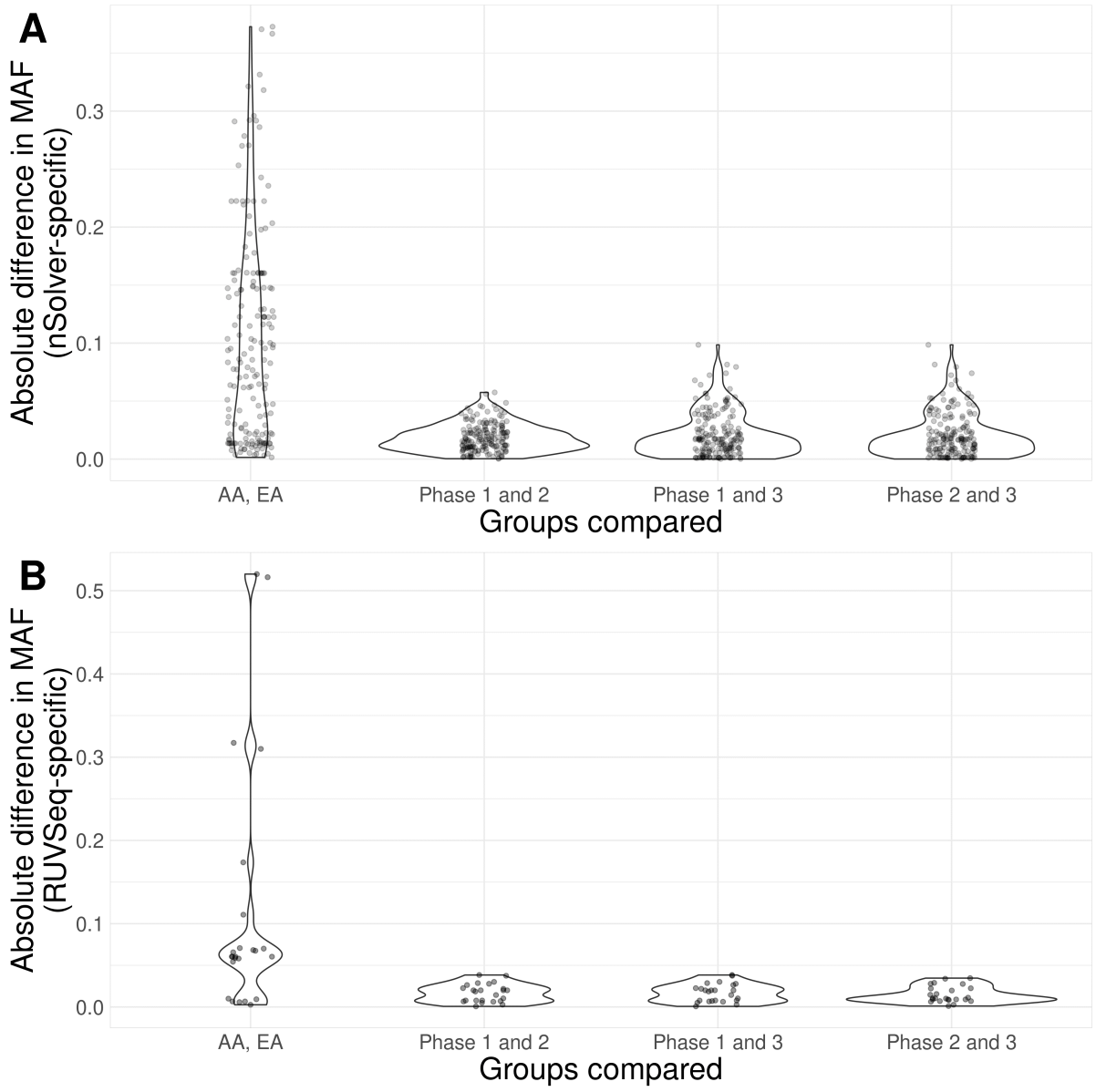


**Figure S8:** *Comparison of minor allele frequencies of trans-eSNPs in nSolver- and RUVSeq-normalized CBCS data.* Violin plots of absolute differences in minor allele frequencies of trans-eSNPs specific to nSolver-normalized data (A) and RUVSeq-normalized data (B) between groups of African ancestry women (AA) and European ancestry women (EA) and between the three study phases.

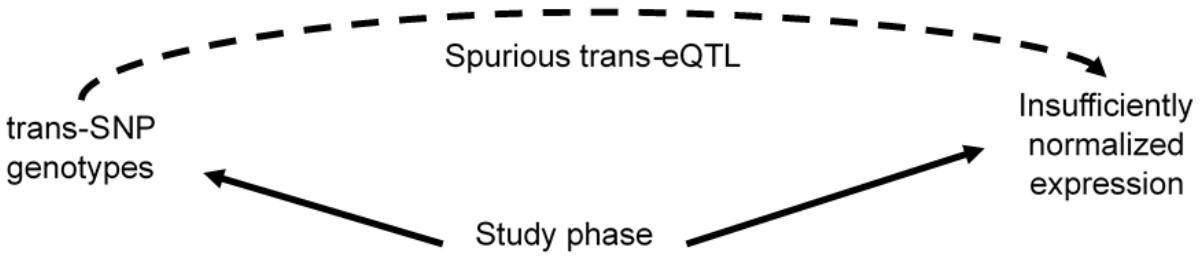


**Figure S9:** *Proposed causal relationships leading to perceived bias in detected trans-eQTLs.* Violin plots of absolute differences in minor allele frequencies of trans-eSNPs specific to nSolver-normalized data (A) and RUVSeq-normalized data (B) between groups of African ancestry women (AA) and European ancestry women (EA) and between the three study phases.

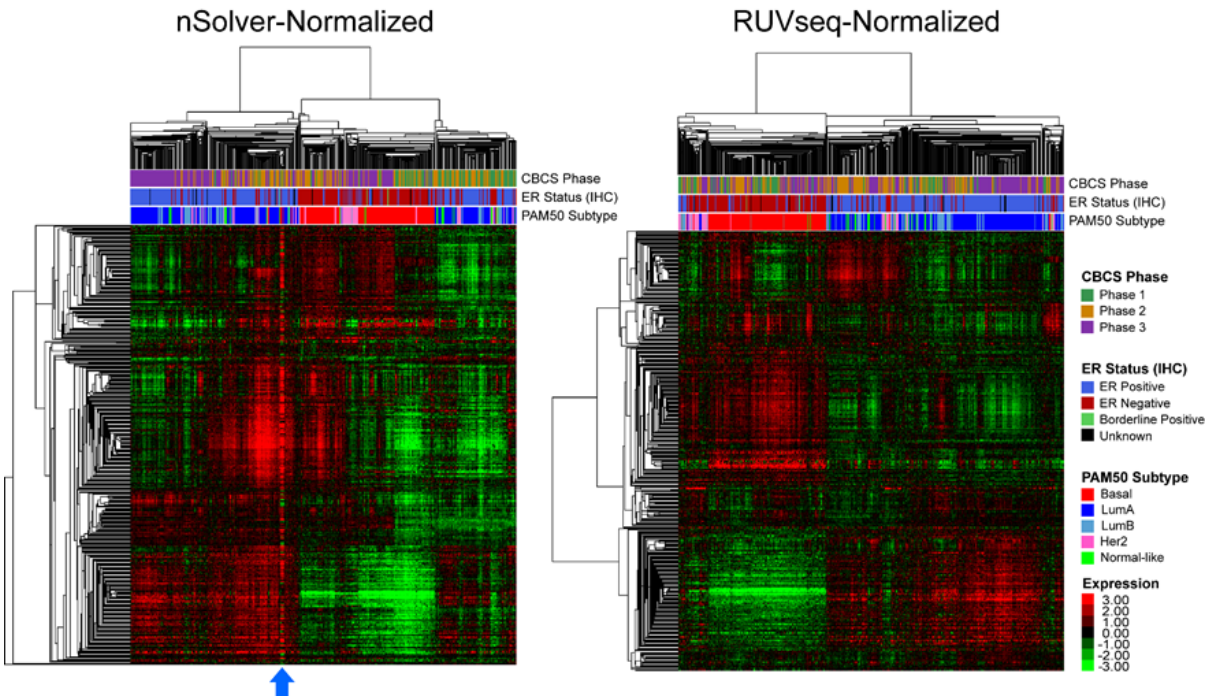


**Figure S10:** *Expression patterns in nSolver- and RUVSeq-normalized CBCS data.* Heatmap of nSolver-normalized (left) and RUVSeq-normalized (right) expression of 417 breast cancer-related genes with hierarchical clustering of samples (horizontal) and genes (vertical). Samples are classified as Basal-like (red), HER2-enriched (pink), luminal A (dark blue), luminal B (light blue), and normal-like (green). The left heatmap uses nSolver-normalized normalized data without quality control based on post-normalization visual inspection. The blue arrow indicates 14 samples without any pre- or post-normalization quality control flags, but show deviations from expression patterns.

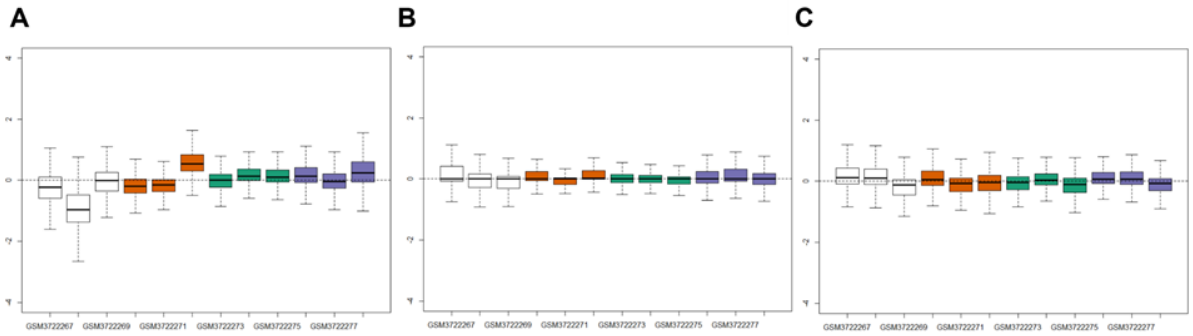


**Figure S11:** *Technical variation across study groups in Sabry et al data.* Relative log-expression (RLE) plots of raw expression (A), nSolver-normalized expression (B), and RUVSeq-normalized expression (C) for Sabry et al's natural killer Nanostring expression profile. Boxplots are colored by various treatment groups.

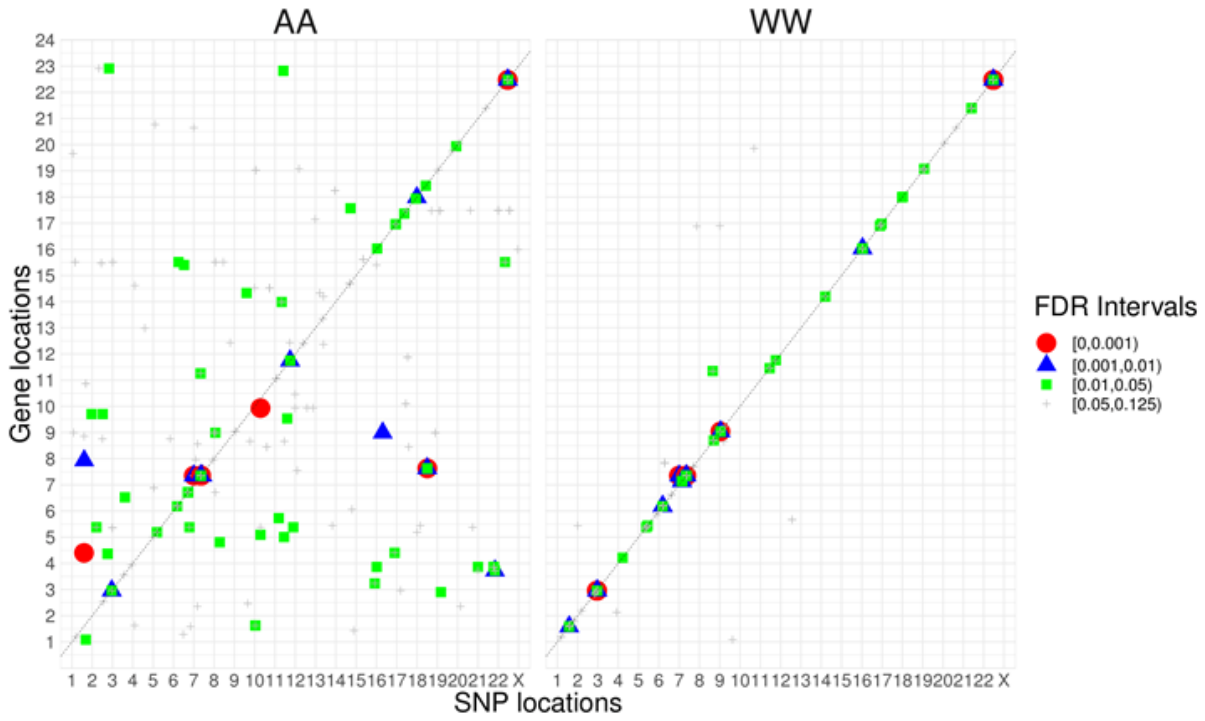


**Figure S12:** *Cis-trans plot of race-stratified eQTL analyses* AA eQTLs are shown on the left and WW on the right. Each point represents an eQTL with $BBFDR < 0.125$ with the location of the 5' end of the corresponding eGenes on the Y-axis and the genomic location of the corresponding eSNP on the X-axis. A 45-degree line is provided as a reference for cis-eQTLs.

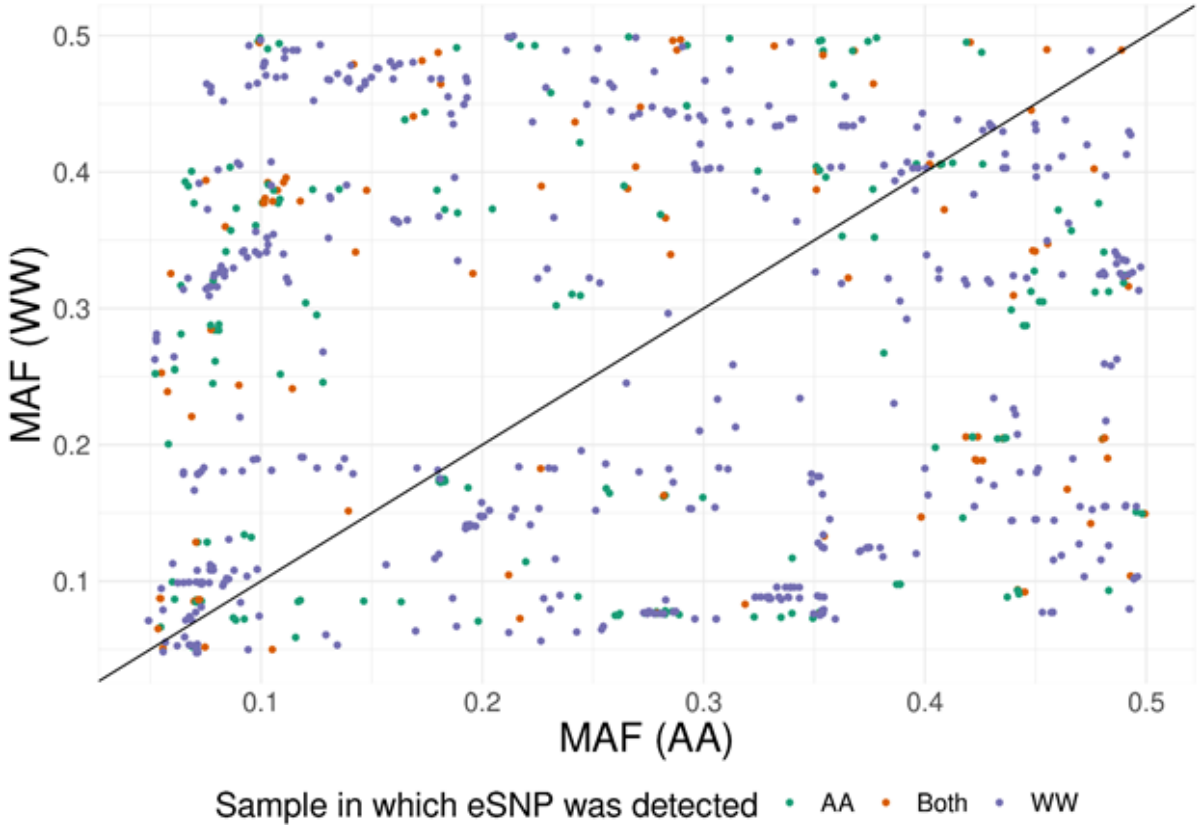


**Figure S13:** *Minor allele frequency differences of eSNPs across race.* Scatter plot of minor allele frequencies (MAF) of all significant eSNPs ($BBFDR < 0.05$) in either the AA or WW sample, with the MAF in the AA sample on the X-axis and in the WW sample on the Y-axis. Points are colored by the sample in which the eSNP was detected. The 45-degree line is provided for reference.

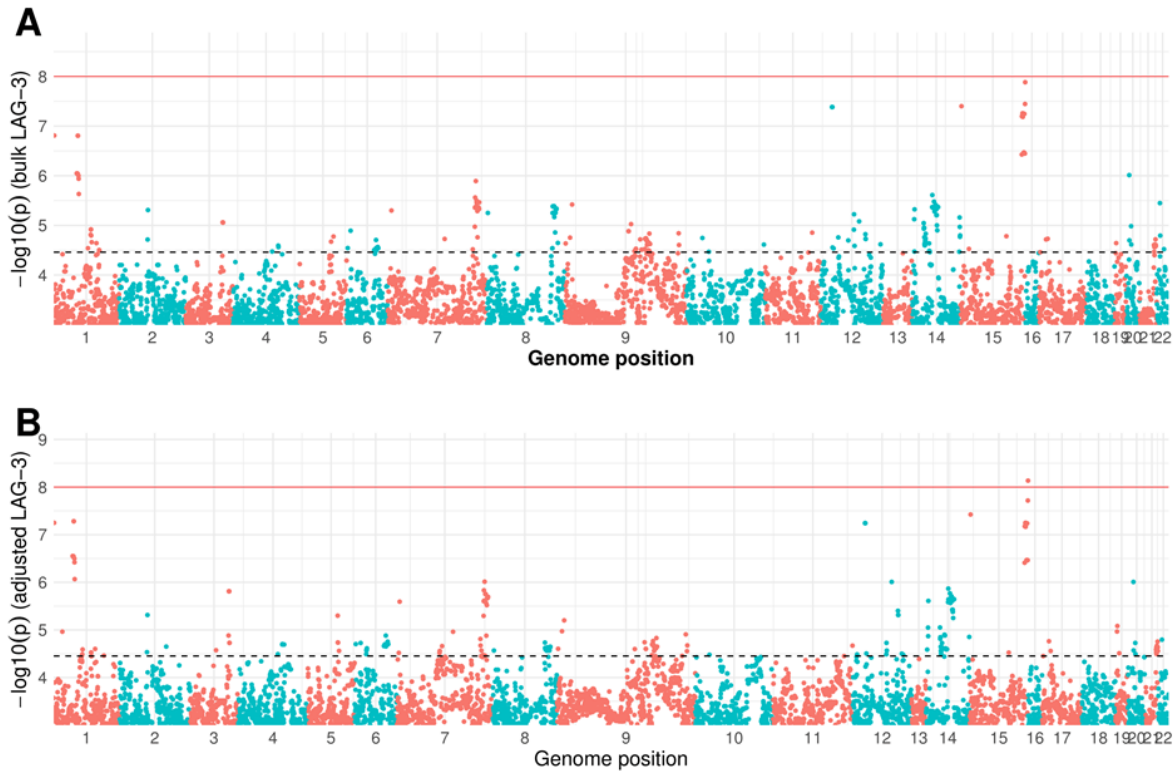


**Figure S14:** *Impact of tumor purity adjustment on eQTLs.* Example Manhattan plots for eQTL analysis in bulk tumor *LAG-3* expression (A) and tumor purity-adjusted *LAG-3* expression (B) in WW women. Red line represents a genome-wide significance threshold of $P = 1 \times 10^{-8}$ and the dotted black line corresponds to $BBFDR < 0.05$.

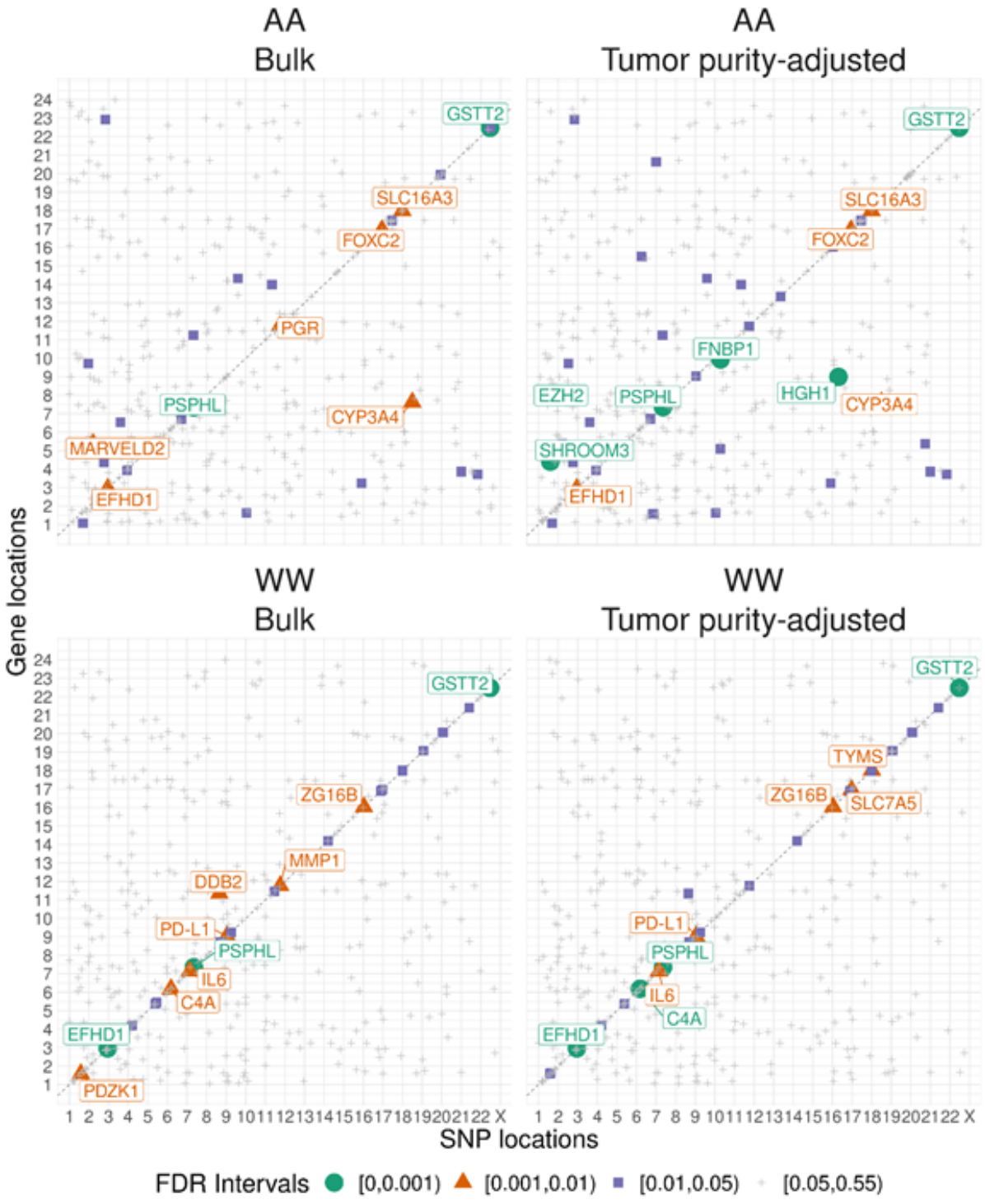


**Figure S15:** *Impact of tumor purity adjustment on eQTLs across race.* Cis-trans plots, as in Supplementary Figure S12, across self-identified race (top to bottom) and across adjustment for tumor purity (eQTLs in bulk tumor expression on left and eQTLs in tumor purity-adjusted expression on left)

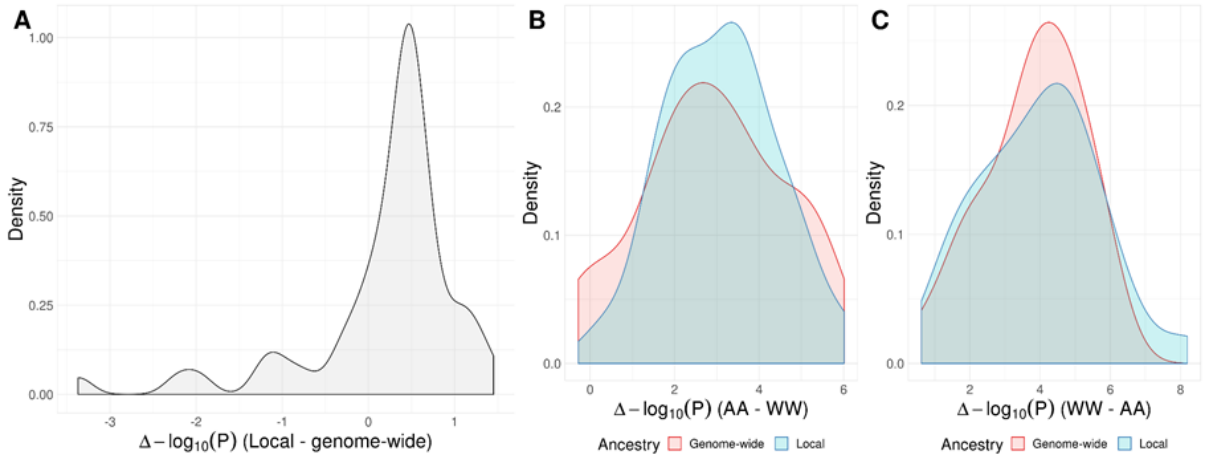


**Figure S16:** *Impact of local ancestry adjustment on cis-eQTLs.* (A) Kernel density plot of difference in $-\log_{10} P$-values for lead cis-eQTLs identified with local ancestry adjustments and genome-wide ancestry adjustments. (B) Kernel density plot of difference in $-\log_{10} P$-values of association of eQTLs between AA and WW women with genome-wide ancestry adjustment (red) and local ancestry adjusted (blue) for lead eQTLs identified for AA-specific cis-eGenes. (C) Kernel density plot of difference in $-\log_{10} P$-values of association of eQTLs between WW and AA women with genome-wide ancestry adjustment (red) and local ancestry adjusted (blue) for lead eQTLs identified for WW-specific cis-eGenes.

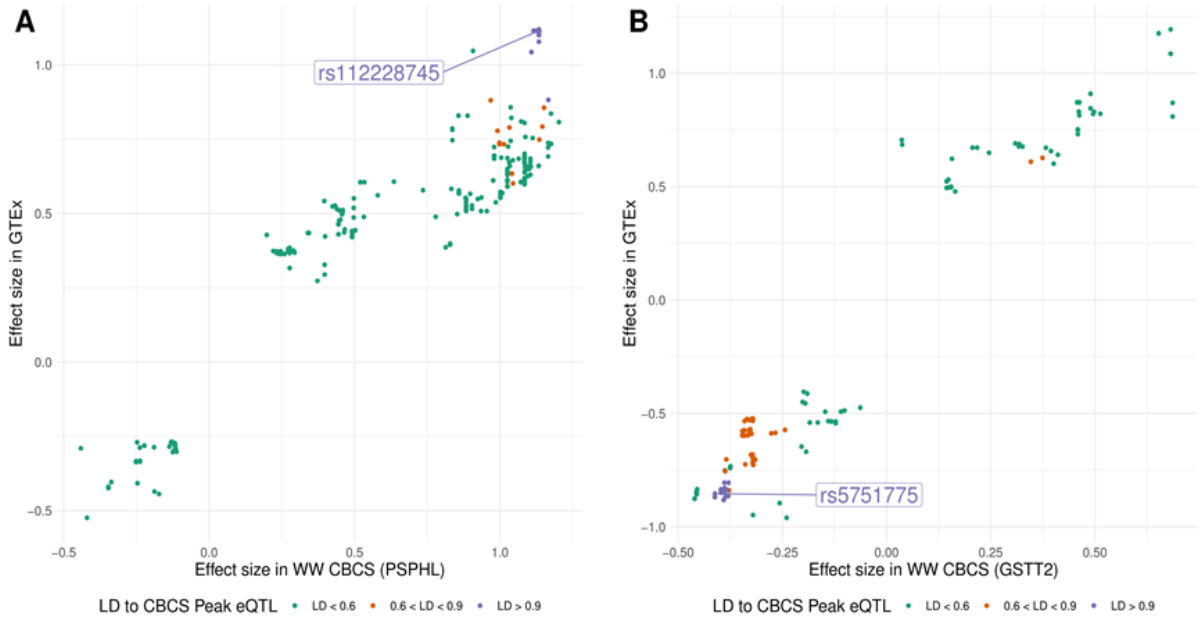


**Figure S17:** *Comparison of eQTL effect sizes across CBCS and GTEx.* Each point represents a significant eQTL for *PSPHL* (A) and *GSTT2* (B) found in both GTEx and the CBCS WW sample, colored by the strength of linkage disequilibrium to the top eSNP in CBCS. Absolute effect size of significant eQTLs in WW CBCS is plotted on the X-axis and absolute effect size of significant eQTLs in GTEx multiplied by the sign of the effect size in CBCS is plotted on the Y-axis.

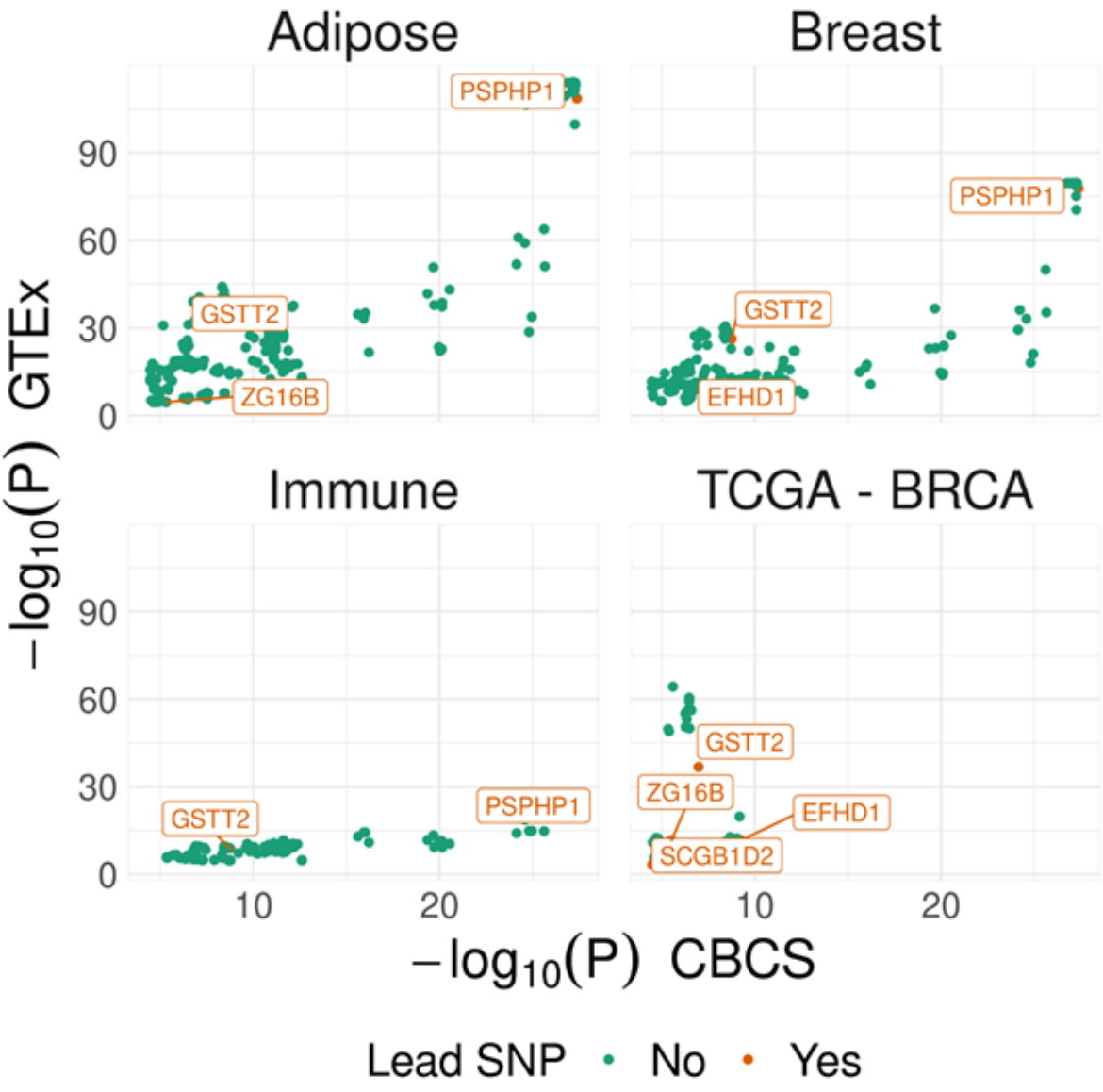


**Figure S18:** *Overlap of WW CBCS cis-eQTLs in GTEx and TCGA-BRCA.* Each point represents a given cis-eSNP-eGene pair (cis-eQTL), with the $-\log_{10} P$-value of the association in CBCS on the X-axis and the $-\log_{10} P$-value of the association in the external dataset on the Y-axis. Each cis-eQTL that is colored orange and labelled is the lead cis-eSNP in CBCS (i.e. the lowest P-value for that eGene in CBCS).

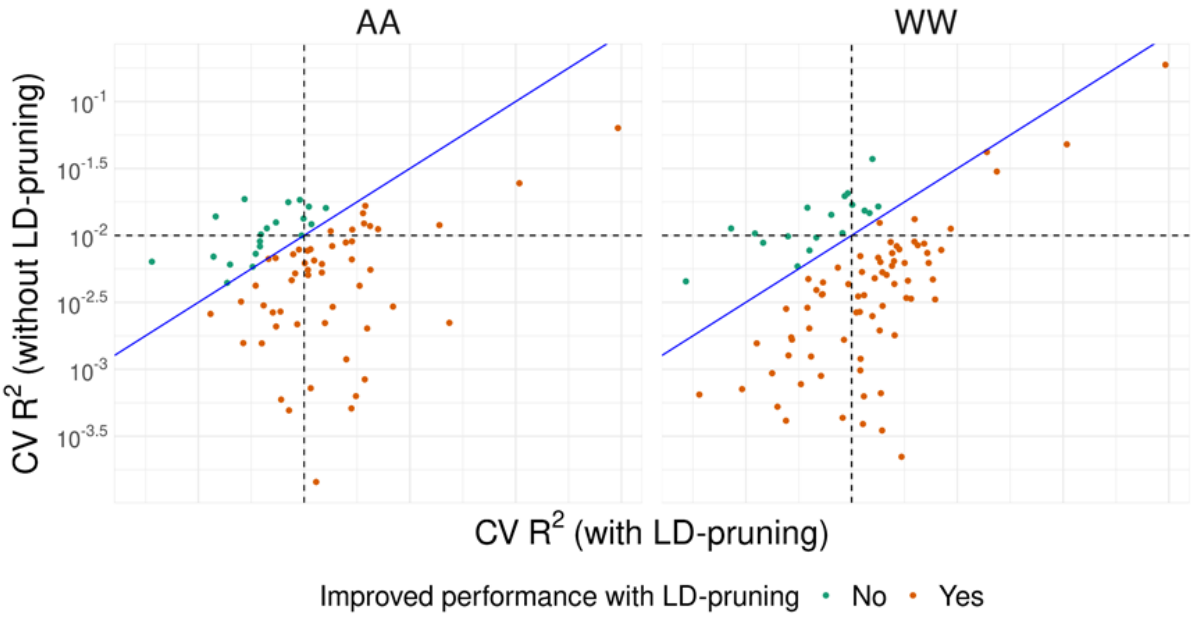


**Figure S19:** *Comparison of use of LD-pruning on model performance.* For genes with cis-$h^2$ with $P < 0.10$, cross-validation $R^2$ with (X-axis) and without (Y-axis) LD-pruning of genotype design matrix. Points are colored orange if there is increased CV $R^2$ with LD-pruning. The blue line gives the 45-degree line and the dotted black lines show thresholds for $R^2 = 0.01$, for reference.

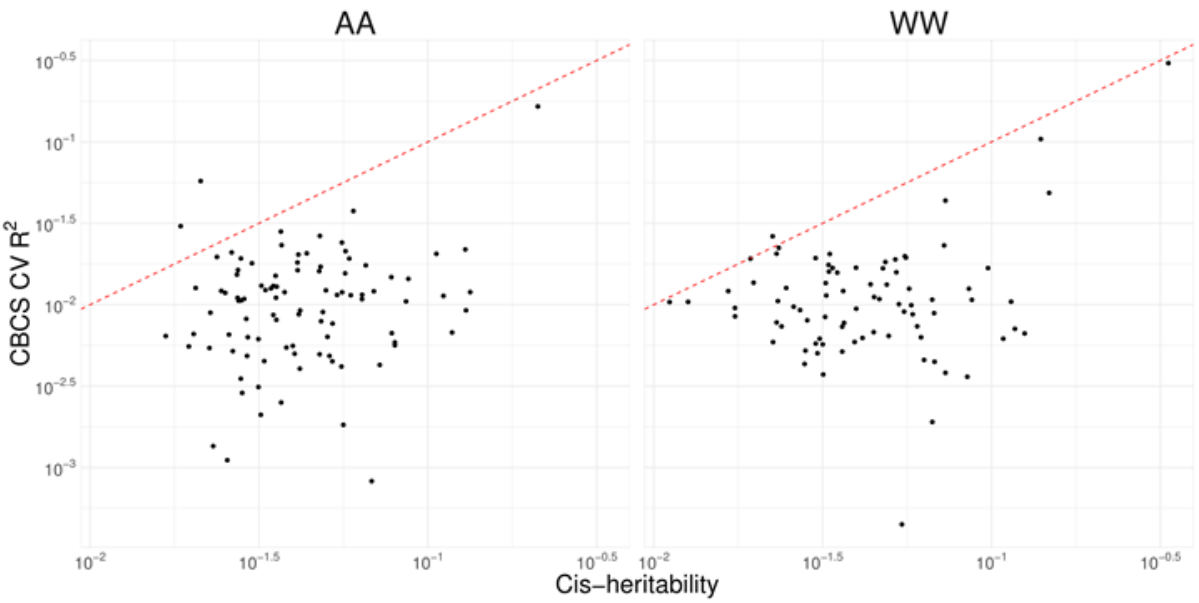


**Figure S20:** *Comparison of heritability and cross-validation predictive performance.* Comparison of cis-$h^2$ estimates (X-axis) and cross-validation $R^2$ (Y-axis) for each gene with likelihood ratio test $P < 0.10$ for cis-$h^2 = 0$ across AA and WW women in CBCS training set. The 45-degree line (i.e. $Y = X$) is provided for reference in red.

| Training/test set | AA | | WW | |
|---|---|---|---|---|
| | **All genes** **Mean $R^2$ (25%, 75%)** | **Prioritized genes** **Mean $R^2$ (25%, 75%)** | **All genes** **Mean $R^2$ (25%, 75%)** | **Prioritized genes** **Mean $R^2$ (25%, 75%)** |
| CBCS training set | 0.012 (0.007, 0.014) | 0.016 (0.006, 0.016) | 0.012 (0.007, 0.014) | 0.016 (0.006, 0.016) |
| Sample size: 628 AA, 571 WW | 417 genes | 81 genes | 417 genes | 100 genes |
| Held-out CBCS test set | 0.007 (0.002, 0.008) | 0.013 ($8.1 \times 10^{-4}$, 0.013) | 0.008 (0.002, 0.010) | 0.014 (0.002, 0.010) |
| Sample size: 1121 AA, 1070 WW | 166 genes | 50 genes | 166 genes | 50 genes |
| TCGA-BRCA test set | 0.006 ($5.1 \times 10^{-4}$, 0.006) | 0.009 ($6.0 \times 10^{-4}$, 0.007) | 0.002 ($1.0 \times 10^{-4}$, 0.002) | 0.005 ($1.5 \times 10^{-4}$, 0.002) |
| Sample size: 179 AA, 735 WW | 412 genes | 149 genes | 412 genes | 149 genes |

**Table S2:** *Mean cross-validation or external validation $R^2$ across CBCS training set, held-out CBCS test set, and TCGA-BRCA test set.* The 25% and 75% quantiles are provided in parentheses with the number of genes with the number of genes considered for these sample statistics. Note that here we define a prioritized gene as one with cis-$h^2 \geq 0$ with $P < 0.1$ in the training set.

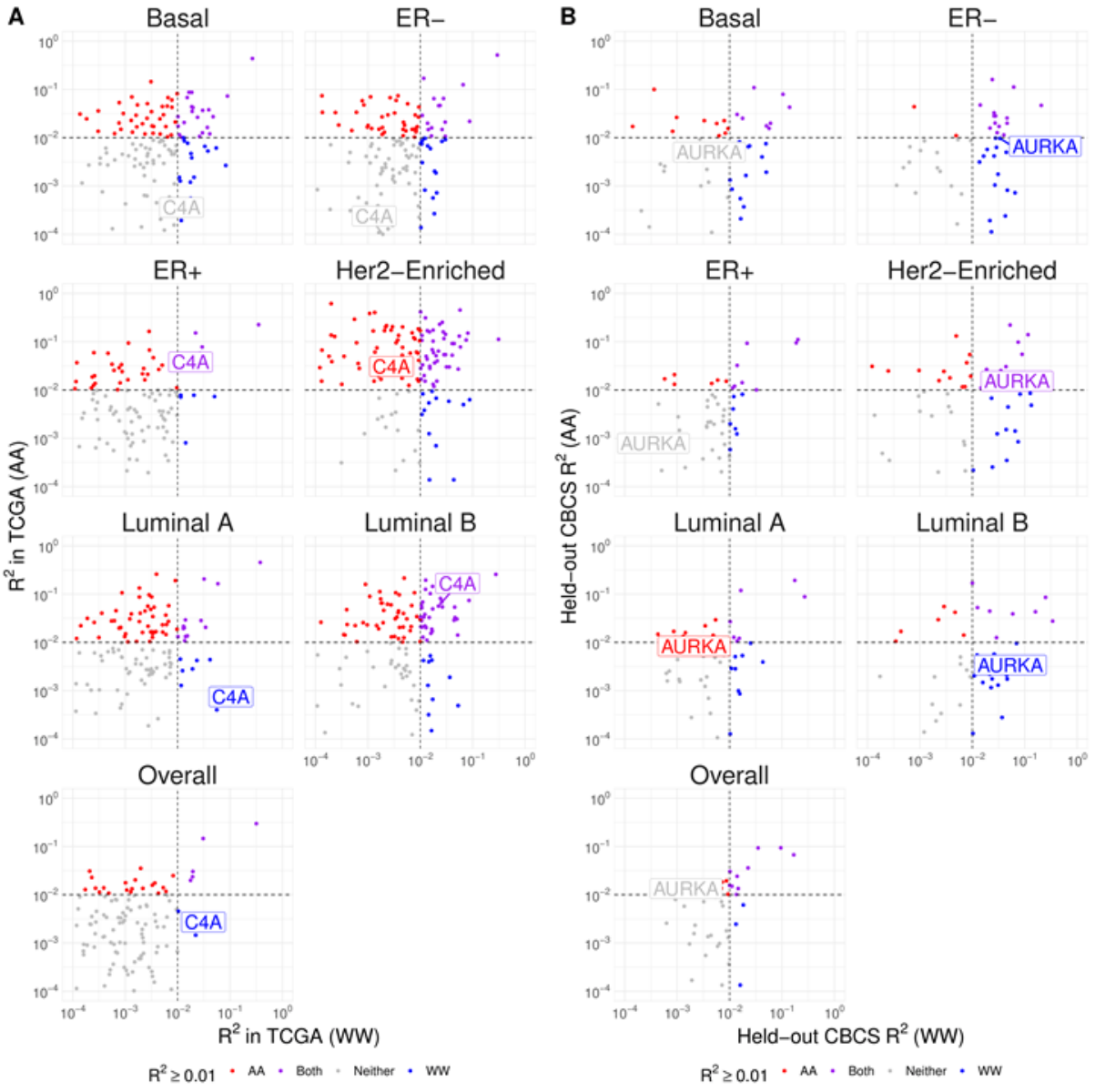


**Figure S21:** *Performance of CBCS expression models in independent external cohorts.* Comparison of EV $R^2$ across race, stratified by PAM50 molecular subtype and estrogen receptor status in TCGA (A) and CBCS (B). Squared Spearman correlation in WW (X-axis) and AA (Y-axis) for each of the available genes are plotted. Note that both scales are logarithmic. Dotted lines represent $R^2 = 0.01$. Colors represent the model with which a given gene can be predicted at cross-validation $R^2 > 0.01$. A representative gene with variable $R^2$ across subtypes is labelled.

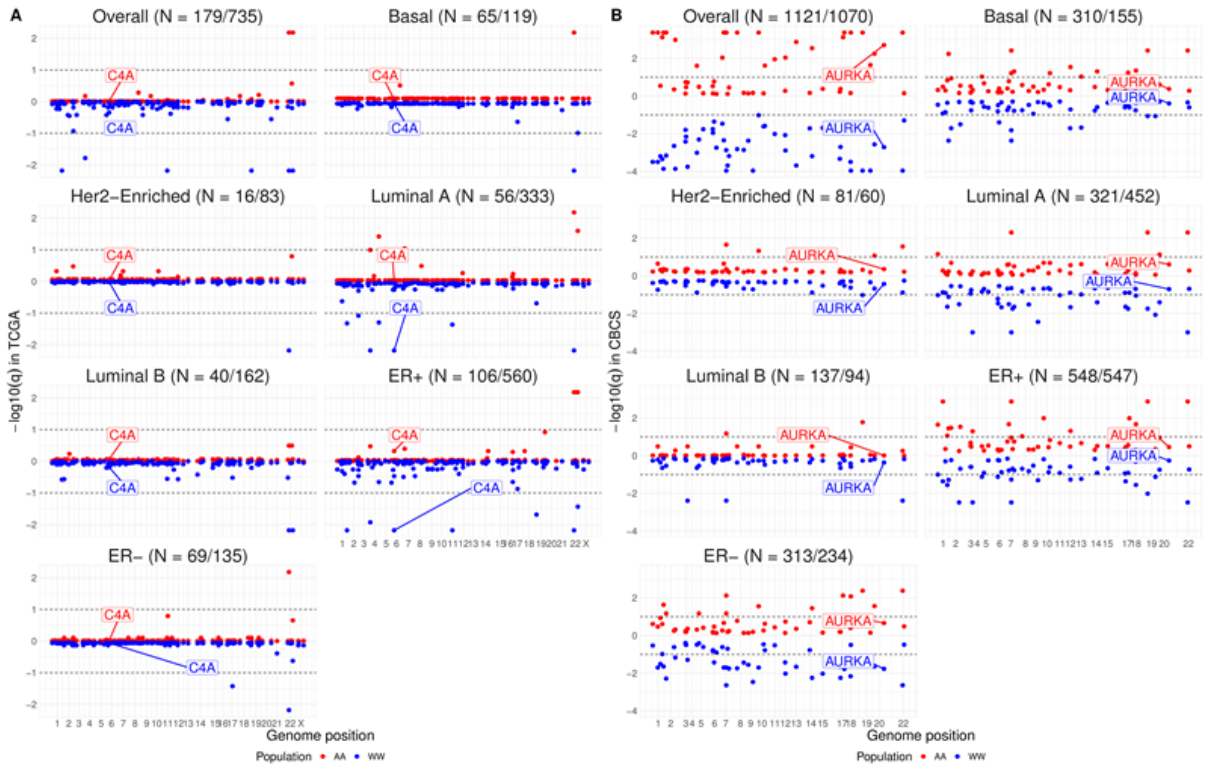


**Figure S22:** *Assessment of sampling variability on external predictive* $R^2$*.* Storey's $-\log_{10} q$-values from $P$-values of permutation tests over 10,000 permutations to assess significance of external validation $R^2$ in TCGA (A) and held-out CBCS (B). Dotted lines represent $q = 0.10$. Sample sizes are provided in the form (AA/WW). A representative gene with variable permutation $q$-value across subtype is labelled.

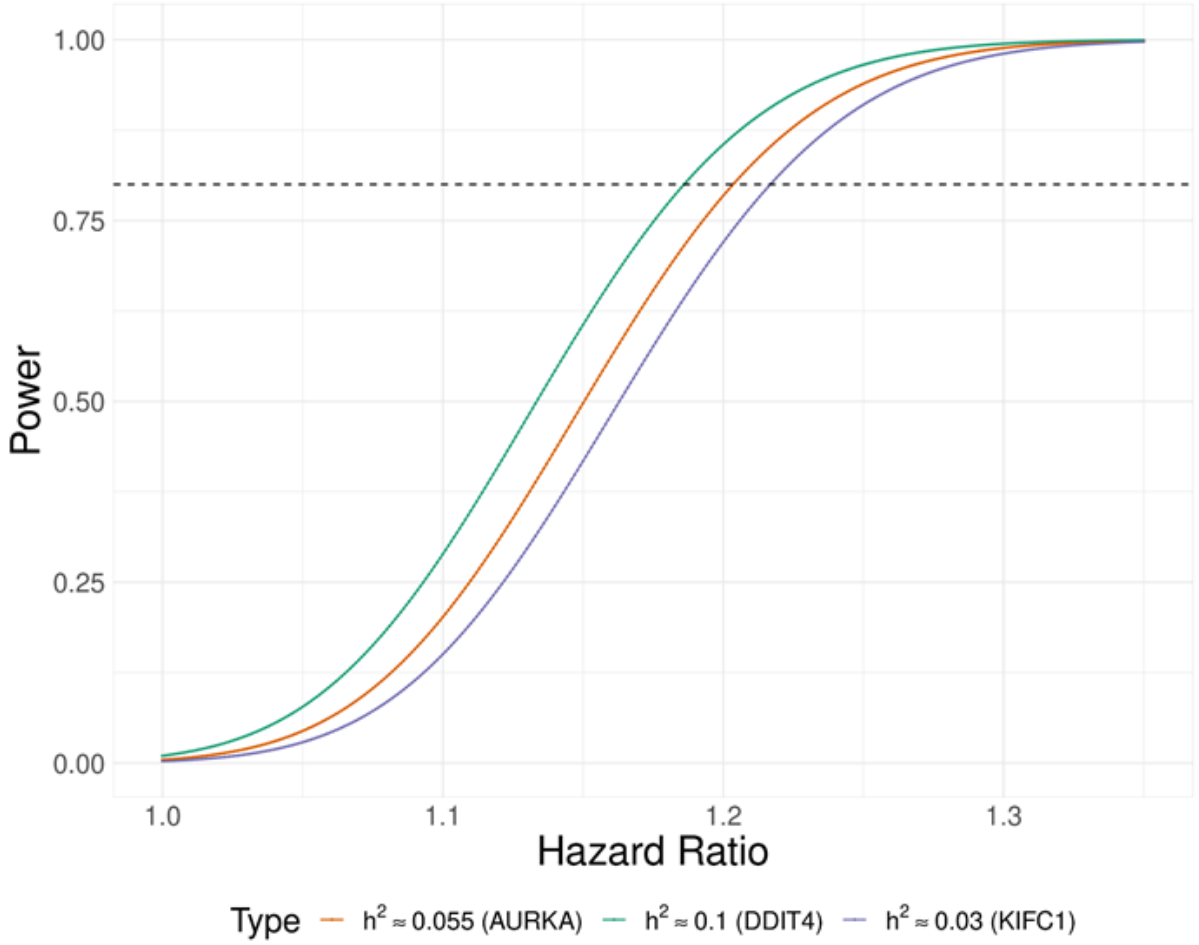


**Figure S23:** *Power analysis of TWAS for survival in CBCS.* Comparison of power of TWAS in CBCS sample of $N = 3,828$ and 348 breast cancer-specific deaths. Power (Y-axis) to detect a given hazard ratio (X-axis) is plotted. Curves correspond to genes of varying cis-$h^2$: *DDIT4* (green) has high $h^2$ across AA and WW, *AURKA* (orange) has average $h^2$ across AA and WW, and *KIFC1* (purple) has the lowest $h^2$ across AA and WW. Power calculations are derived from 1,000 re-samplings of the empirical distribution function of the GReX of a given gene. Dotted line represents 80% power.

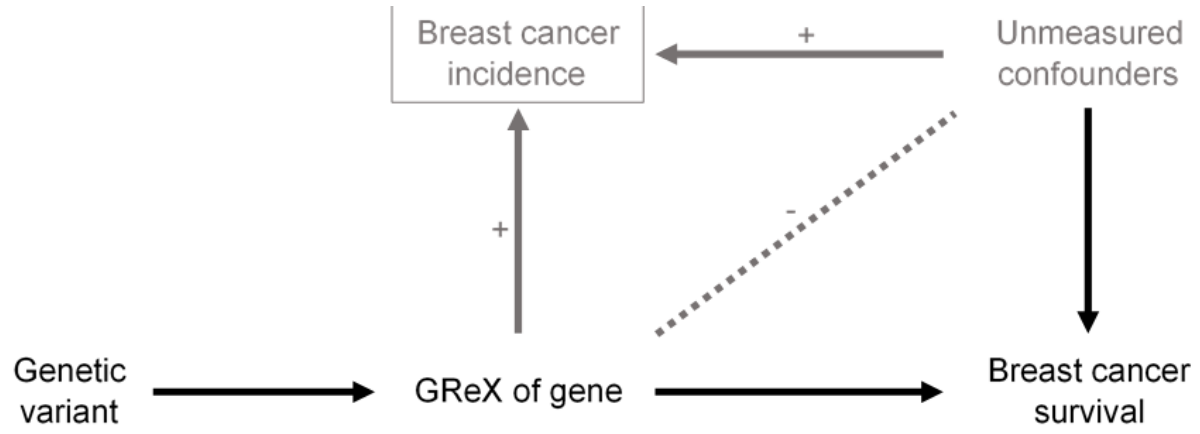


**Figure S24:** *Directed acyclic graph showing potential backdoor confounding in a case-only study.* Modified from Paternoster et al. Directed acyclic graph that shows how collider bias is introduced (grey path) in case-only studies. Here, in this case-only study, we condition on breast cancer incidence, which may open up a potential collider bias with unmeasured confounders in the measure of association between the GReX of a gene and breast cancer survival.

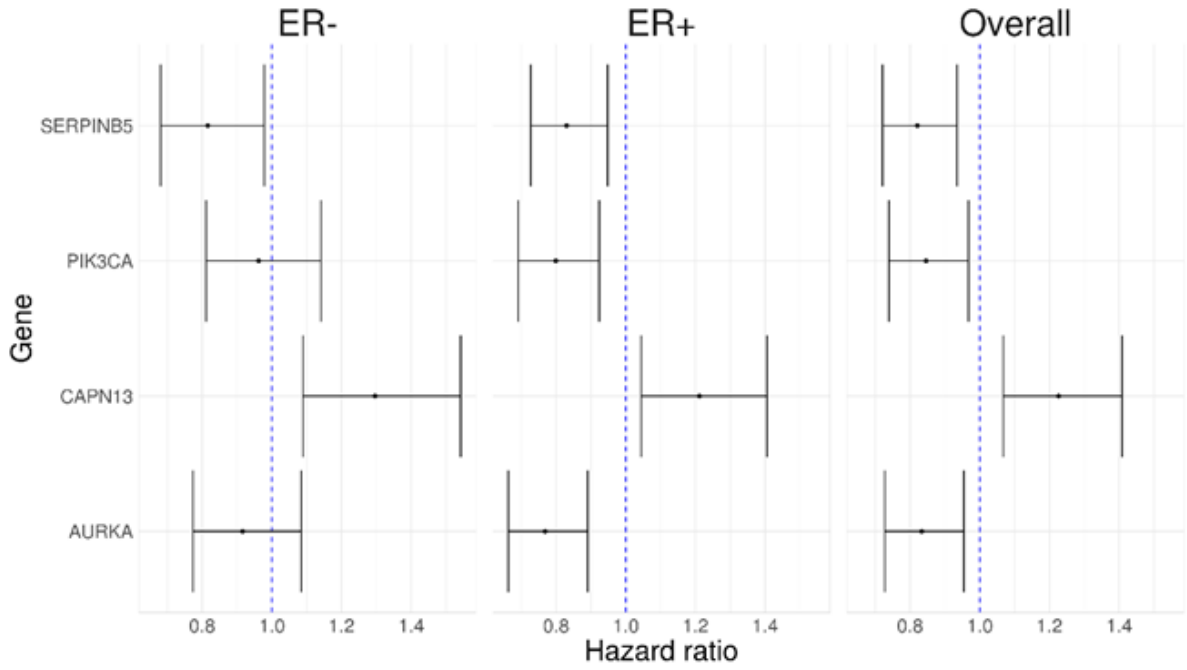


**Figure S25:** *Subtype-specific follow-up on TWAS associations.* Caterpillar plots for hazard ratio of breast cancer-specific survival in AA women for an increase of one standard deviation of GReX across models unadjusted for estrogen receptor subtype and stratifying for estrogen receptor subtype.

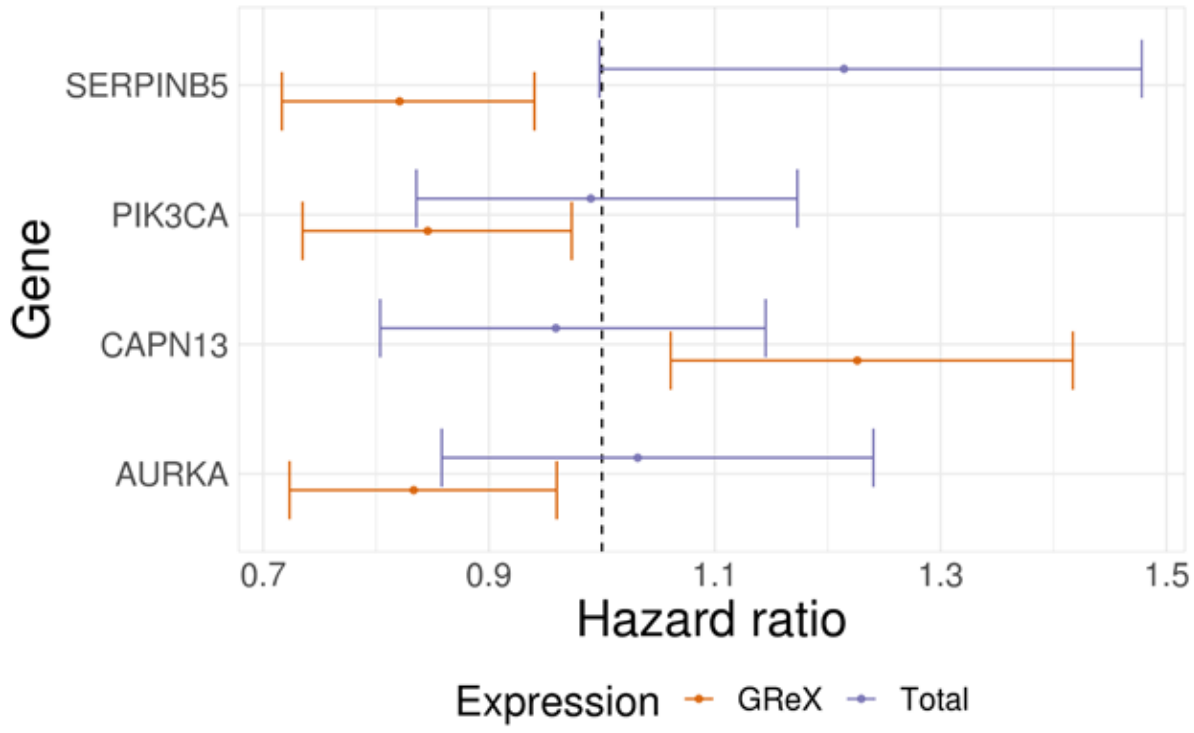


**Figure S26:** *Associations of gene expression and GReX for four TWAS-detected loci in CBCS.* Hazard ratios and 95% confidence intervals, adjusted for false discovery via Benjamini-Hochberg, as estimated from breast cancer-specific Cox models in AA women. Association with total expression (purple) and GReX (orange) of 4 TWAS-detected genes are compared.

| | TCGA-BRCA | ROS/MAP |
|---|---|---|
| **Local-only** | 0.037 (0.053) | 0.079 (0.119) |
| **MeTWAS** | 0.040 (0.066) | 0.135 (0.099) |
| **DePMA** | 0.383 (0.194) | 0.405 (0.118) |

**Table S3:** *Comparison of $h^2$ across local-only, MeTWAS, and DePMA predictive models.* The mean and standard deviation of $h^2$ across all genes that are significantly heritable with the genetic loci considered in the design matrix of each predictive model.

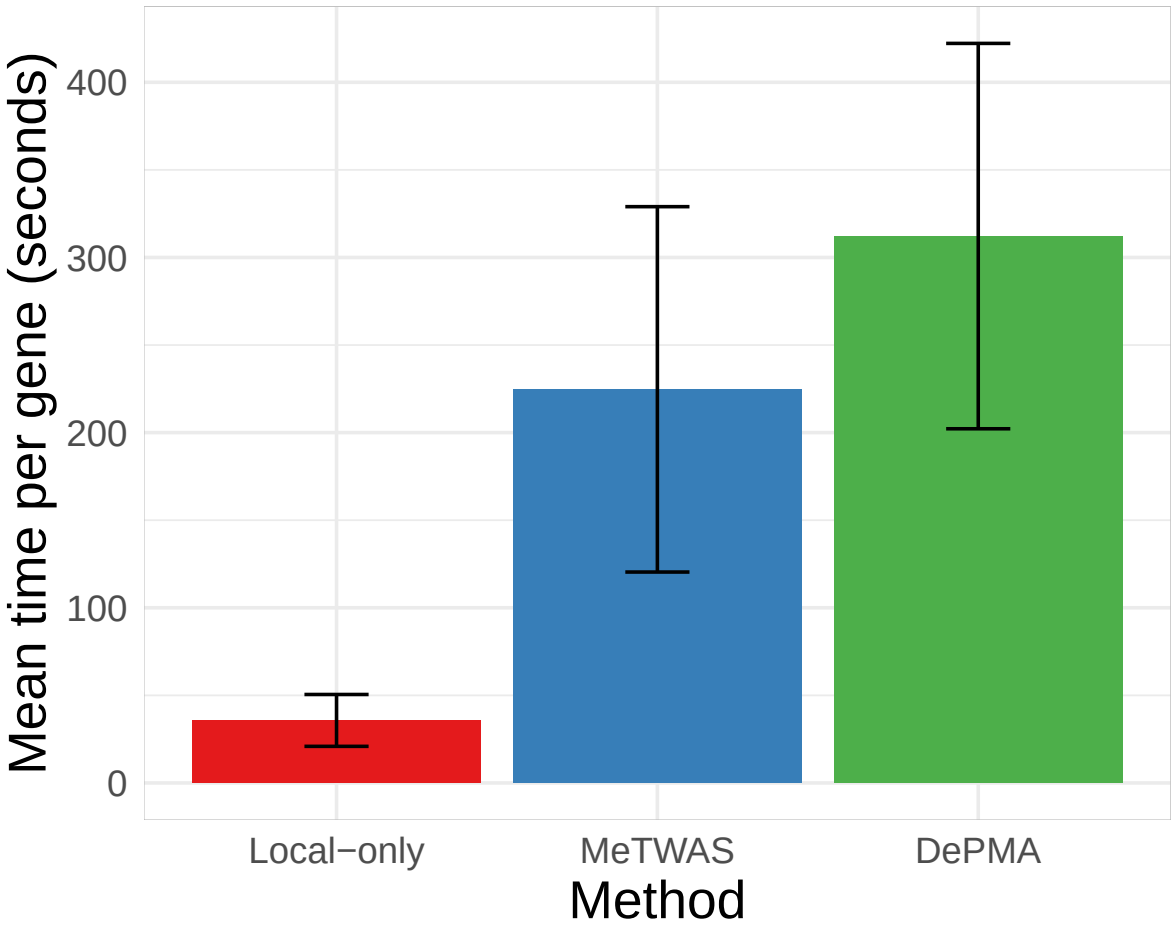


**Figure S27:** *Comparison of computation times between local-only and MOSTWAS modelling.* Mean and standard deviation of per-gene computation time across 50 randomly selected genes in TCGA-BRCA. Computations here were done with a 24-core, 3.0 GHz processor.

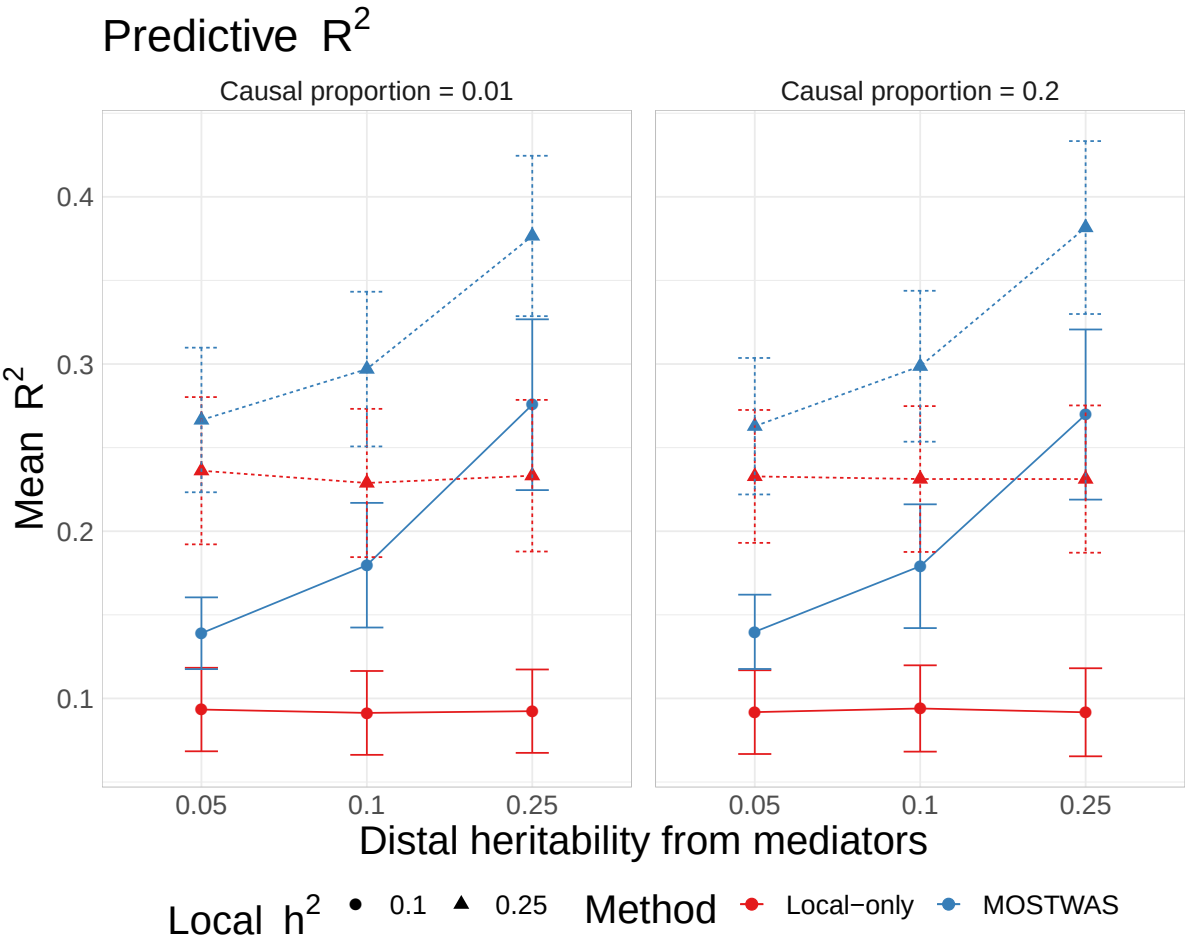


**Figure S28:** *Comparison of predictive $R^2$ in simulations.* Mean adjusted $R^2$ across various local and distal expression heritabilities, trait heritabilities, and causal proportions using local-only (red) and the best MOSTWAS (blue) models. The error bars reflect a width of 1 standard deviation of the 1,000 simulated adjusted $R^2$ values.

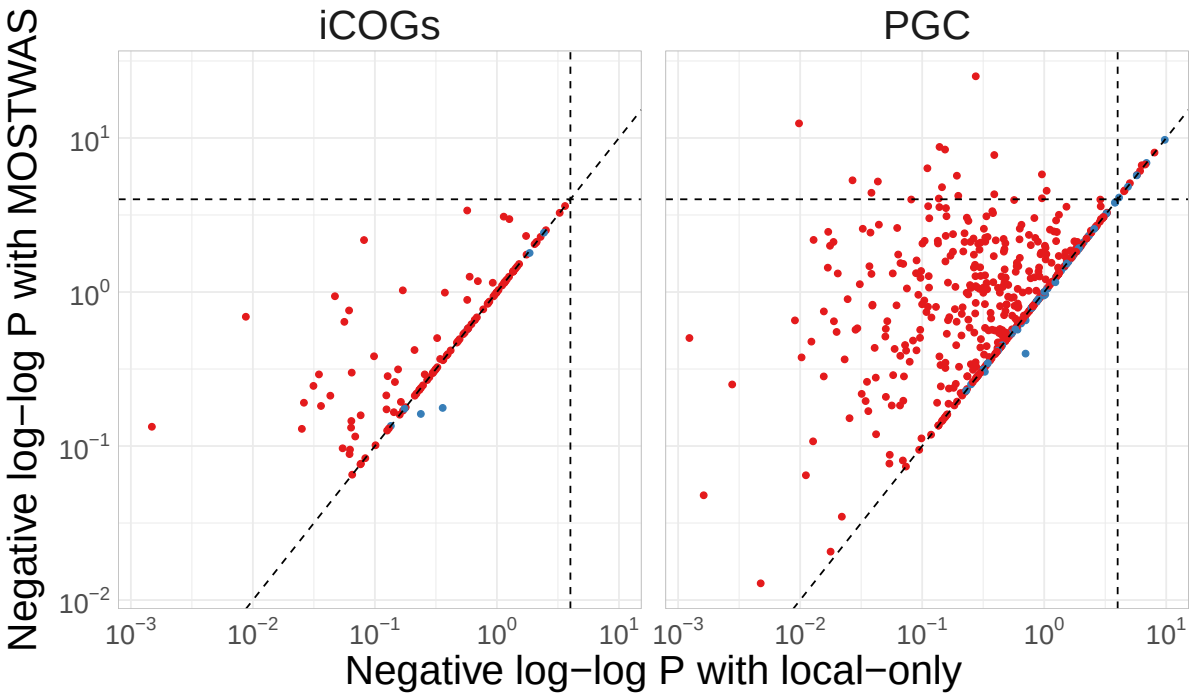


**Figure S29:** *Gene-trait associations in iCOGs and PGC using local-only and MOSTWAS models.* $-\log_{10} P$-values of weighted burden gene-trait associations using iCOGs survival GWAS in European-ancestry women (left) and PGC MDD risk GWAS in predominantly European-ancestry patients (right) among genes that were predicted at cross-validation $R^2 \geq 0.01$ using both local-only and MOSTWAS models. The $X$- and $Y$-axes display the $-\log_{10} P$-values for local-only and the best MOSTWAS model, respectively. Note that the scales of both axes are on a doubly logarithmic scale. Points are colored red if $P$-value of association is less than or equal using the MOSTWAS model. The horizontal and vertical reference lines indicate overall Bonferroni-corrected significance thresholds.

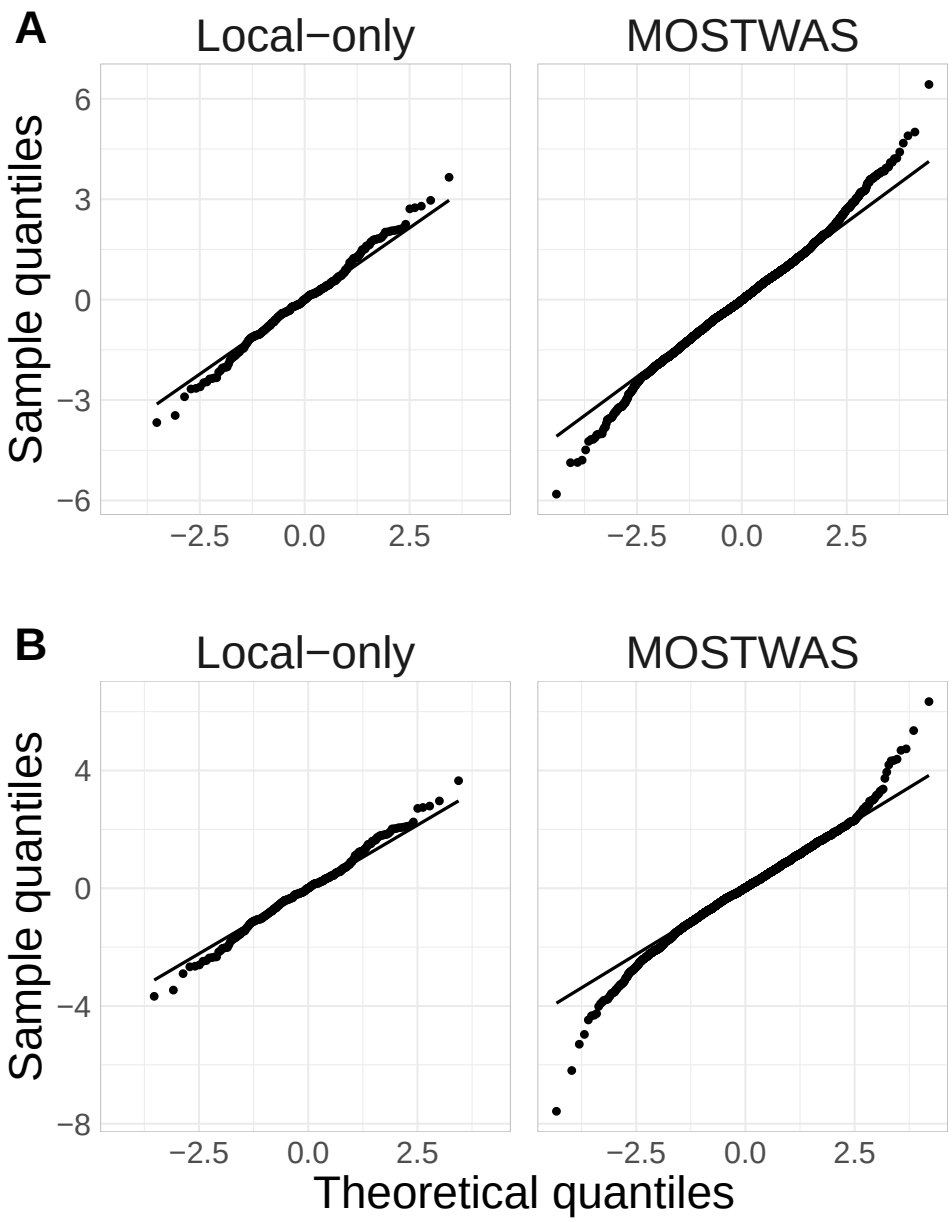


**Figure S30:** *Comparison of QQ-plots from TWAS associations.* QQ-plots from TWAS for breast cancer-specific survival in iCOGs (A) and MDD in PGC (B) with local-only models (left) and MOSTWAS (right)

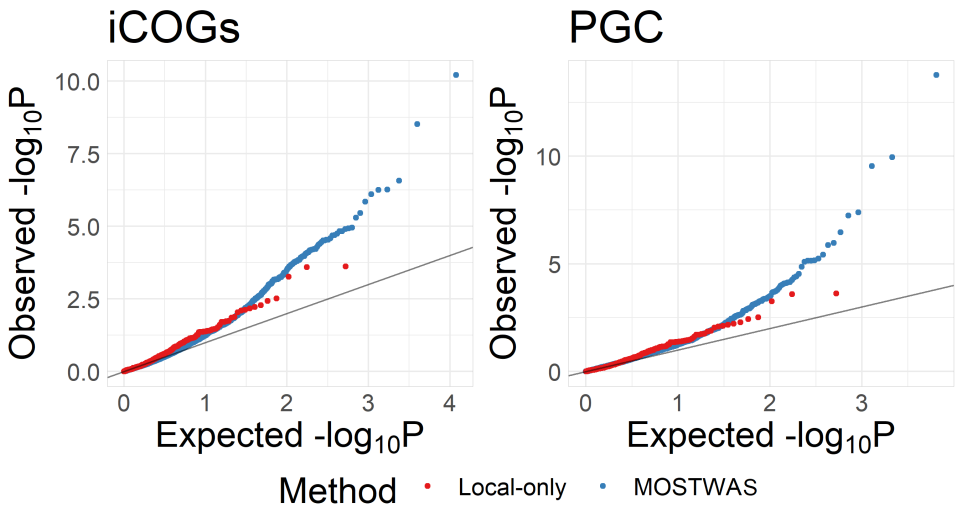


**Figure S31:** *Comparison of $P$-value QQ-plots from TWAS associations.* QQ-plots of $-\log_{10} P$-values from TWAS for breast cancer-specific survival in iCOGs (A) and MDD in PGC (B) with local-only models (left) and MOSTWAS (right)

| Gene | Z-statistic (FDR-adjusted $P$) | Cross-validation $R^2$ | TOP GWAS SNP location ($P$) | Permutation FDR-adjusted $P$ | Added last FDR-adjusted $P$ |
|---|---|---|---|---|---|
| **ABCA7** | -1.82 (0.09) | 0.011 | Chr19:553,066 (0.135) | NA | NA |
| **ADAM10** | -1.25 (0.23) | 0.014 | Chr15:59,052,072 ($1.68 \times 10^{-4}$) | NA | NA |
| **APOE** | 2.82 (0.02) | 0.119 | Chr19:45,545,562 ($3.0 \times 10^{-5}$) | $5.0 \times 10^{-3}$ | 0.03 |
| **BIN1** | 1.91 (0.08) | 0.010 | Chr22:24,199,787 ($8.53 \times 10^{-4}$) | NA | NA |
| **CD2AP** | 1.52 (0.15) | 0.011 | Chr6:47,432,637 ($1.23 \times 10^{-4}$) | NA | NA |
| **CLU** | -2.41 (0.04) | 0.012 | Chr8:27,465,312 ($1.33 \times 10^{-4}$) | 0.83 | 0.44 |
| **FERMT2** | 2.13 (0.06) | 0.017 | Chr14:53,305,626 ($1.38 \times 10^{-4}$) | NA | NA |
| **MEF2C** | 2.20 (0.06) | 0.016 | Chr5:88,359,039 (0.020) | NA | NA |
| **PLCG2** | -2.48 (0.04) | 0.010 | Chr16:81,879,218 (0.037) | 0.66 | 0.07 |
| **SORL1** | 2.91 (0.02) | 0.043 | Chr11:121,446,813 (0.032) | 0.04 | $4.5 \times 10^{-3}$ |
| **ZCWPW1** | -4.56 ($6.1 \times 10^{-5}$) | 0.018 | Chr7:100,435,157 (0.074) | 0.03 | $1.3 \times 10^{-5}$ |

**Table S4:** *Summary statistics for known Alzheimer's risk-associated loci identified by MOSTWAS models.* TWAS associations (weighted $Z$-score and FDR-adjusted[1] $P$-value) with late-onset Alzheimer's risk from GWAS statistics from IGAP[2]. The top IGAP GWAS SNP in the identified loci with its location and $P$-value are provided. For the 6 loci with significant TWAS associations, the FDR-adjusted $P$-value for the follow-up distal SNP added last test is provided.

| Gene | Cross-validation $R^2$ | PGC $Z$-statistic (UKBB $Z$) | Top GWAS SNP location ($P$) | Permutation FDR-adjusted $P$ |
|---|---|---|---|---|
| **ADAD2** | 0.050 | 5.89 (4.16) | Chr5:35,639,107 ($4.05 \times 10^{-3}$) | $3.5 \times 10^{-5}$ |
| **CACNA2D3** | 0.033 | 3.41 (2.88) | Chr7:12,268,243 ($1.27 \times 10^{-2}$) | 0.046 |
| **FAM43B** | 0.035 | -4.03 (-2.85) | Chr2:73,148,399 ($2.09 \times 10^{-2}$) | 0.028 |
| **MGC29506** | 0.022 | 3.51 (5.54) | Chr5:139,536,922 ($1.48 \times 10^{-3}$) | $3.5 \times 10^{-5}$ |
| **OR8U1** | 0.022 | -3.19 (-4.21) | Chr11:56,676,947 ($4.90 \times 10^{-5}$) | 0.049 |
| **SYT1** | 0.015 | -5.58 (-3.16) | Chr7:12,269,417 ($1.29 \times 10^{-2}$) | 0.040 |
| **YJEFN3** | 0.010 | 5.82 (7.22) | Chr7:12,276,011 ($1.35 \times 10^{-2}$) | 0.038 |

**Table S5:** *Summary statistics for 7 MDD risk-associated loci identified by MOSTWAS models.* TWAS associations with major depressive disorder from GWAS statistics from Psychiatric Genomics Consortium that were replicated with GWAX summary statistics in UK Biobank. The top PGC GWAS SNP in the identified loci with its location and $P$-value are provided.

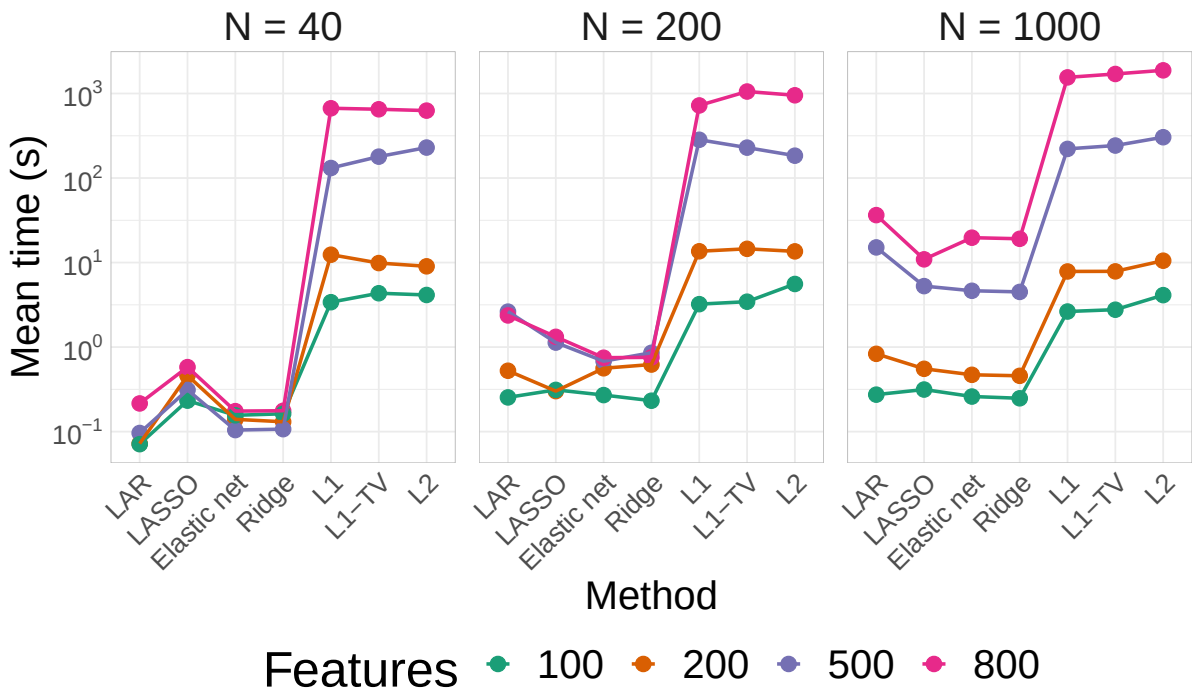


**Figure S32:** *Comparison of run-times for various methods implemented for compressed sensing in DeCompress.* Over sample sizes of $N = 40$, $N = 200$, and $N = 1000$ and feature sizes of 200, 500, 800, and 100, we plot the mean time of estimation compression model over the 7 methods implemented in DeCompress: least angle regression (LAR), LASSO, elastic net with $\alpha = 0.5$, ridge regression, non-linear optimization with $l_1$ norm, non-linear optimization with total variation-adjusted $l_1$ norm, and non-linear optimization with $l_2$ norm.

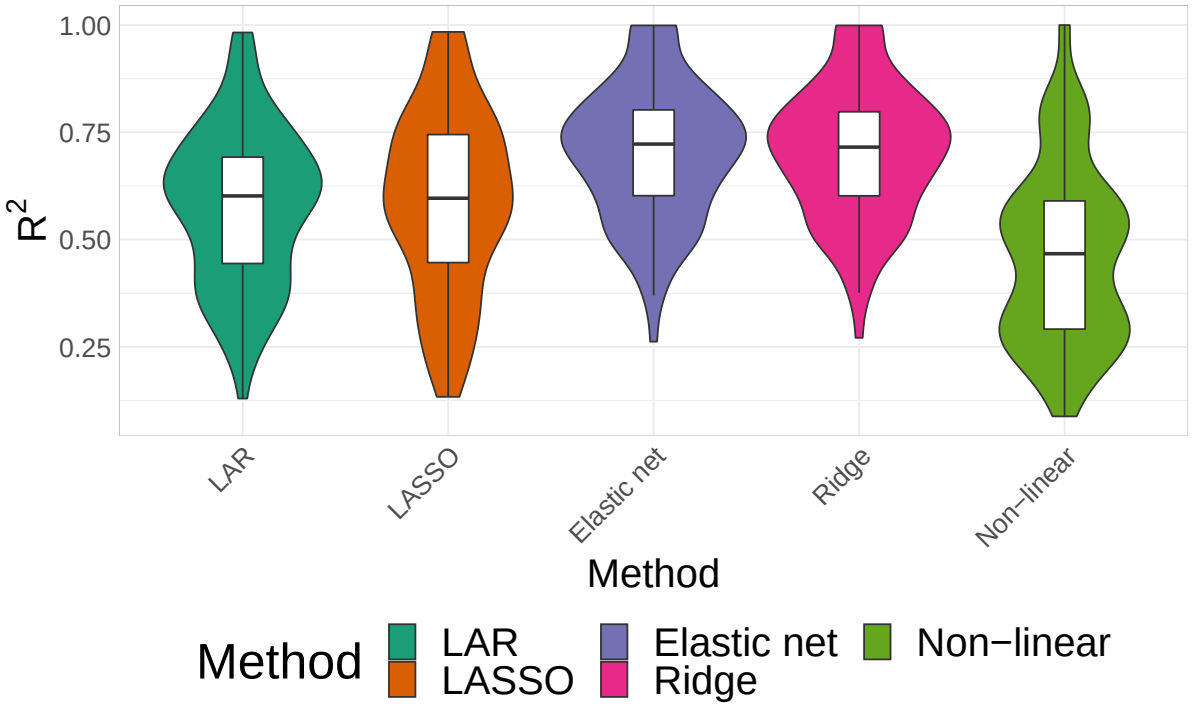


**Figure S33:** *Comparison of predictive performance of optimization methods used in DeCompress's compressing sensing step.* Violin plots for distributions of cross-validation $R^2$ ($Y$-axis) of the various optimization methods ($X$-axis) employed by DeCompress for compression sensing for 100 randomly selected genes from CBCS. From left to right, least angle regression, LASSO, elastic with $alpha = 0.5$, ridge regression, and non-linear optimization with $l_1$ norm. Non-linear optimization with either the total variation-adjusted $l_1$ norm or the $l_2$ norm gives similar results as with the $l_1$ norm, and hence is omitted.

| Method | Summary | Implementation |
|---|---|---|
| deconf [12] | Non-negative least squares on normalized expression matrix in $\log_2$-space, seeded by initial non-negative matrix factorization. | R package CellMix [251] |
| TOAST [14] | Feature selection used in combination with iterative reference-free deconvolution. Feature selection is done using a method for cross-cell type differential analysis for data from a mixed sample [252]. | R package TOAST [14] |
| CellDistinguisher [15] | Topic modeling based on a set of input cell-type distinguishing genes. CellDistinguisher includes a method to infer distinguishing genes using the gene-gene conditional expression vectors in a space where the number of vectors and number of dimensions are both equal to the number of genes. This step relies on a large input number of genes to properly function. | R package CellDistinguisher [15] |
| Linseed [13] | Solving a convex hull problem by projecting the gene expression data and find corners using an assumption that cell-type specific genes are mutually linear. The cell-type specific expression genes are then inputted into the Digital Sorting Algorithm, a gene-signature based deconvolution method [87]. | R package linseed |
| DeconICA [17] | Deconvolution using Independent Component Analysis (ICA), a matrix factorization method for dimension reduction by projecting the expression into a space such that distributions of the data point projections on the new axes are as mutually independent as possible. | R package DeconICA [17] |
| unmix [16,117] | Non-negative least squares on the non-$\log_2$ scale with loss calculated in a variance stabilized space. This is a reference-based method, and is seeded in DeCompress using the estimated cell-type specific expression profiles estimated from the reference. | R package DESeq2 [16] |

**Table S6:** *Summary of deconvolution methods benchmarked against or employed in DeCompress.*

| Dataset | Accession Number | Description |
|---|---|---|
| In-silico GTEx mixing [222,223] | dbGAP: phs000424.v7.p2 | Median tissue-specific expression profiles were mixed at randomly generated mixing proportions to simulate targeted panels. |
| Rat tissue cell-line mixture [6] | GEO: GSE19830 | Rat brain, liver, and lung biospecimens from one animal were mixed at the cRNA homogenate level in different proportions. Expression was measured using microarray. |
| Human breast cancer cell-line mixture [5] | GEO: GSE123604 | Total mRNA was prepared from Namalwa (Burkitt's lymphoma), Hs343T (fibroblasts from mammary gland adenocarcinoma), hTERT-HME1 (normal mammary epithelial cells), and MCF7 (estrogen receptor positive breast cancer cells). Cell lines were mixed in different proportions and expression was measured using RNA-seq. |
| Human prostate tumor laser capture microdissection [7] | GEO: GSE97284 | Gene expression profiling of laser capture microdissected epithelial and stromal specimens from prostate tumors using microarray. |
| Human lung cancer cell-line mixture [8] | GEO: GSE64098 | Two lung adenocarcinoma cell lines (NCI-H1975 and HCC827) were mixed at different proportions and expression was measure using RNA-Seq. |
| Bulk breast tumors from the Carolina Breast Cancer Study [122,24] | Request GEO download token from authors | Expression from bulk breast tumors were measured using NanoString nCounter. A pathologist estimated cell-type proportions for 148 samples from tumor microarrays. |

**Table S7:** *Summary of datasets used in benchmarking*

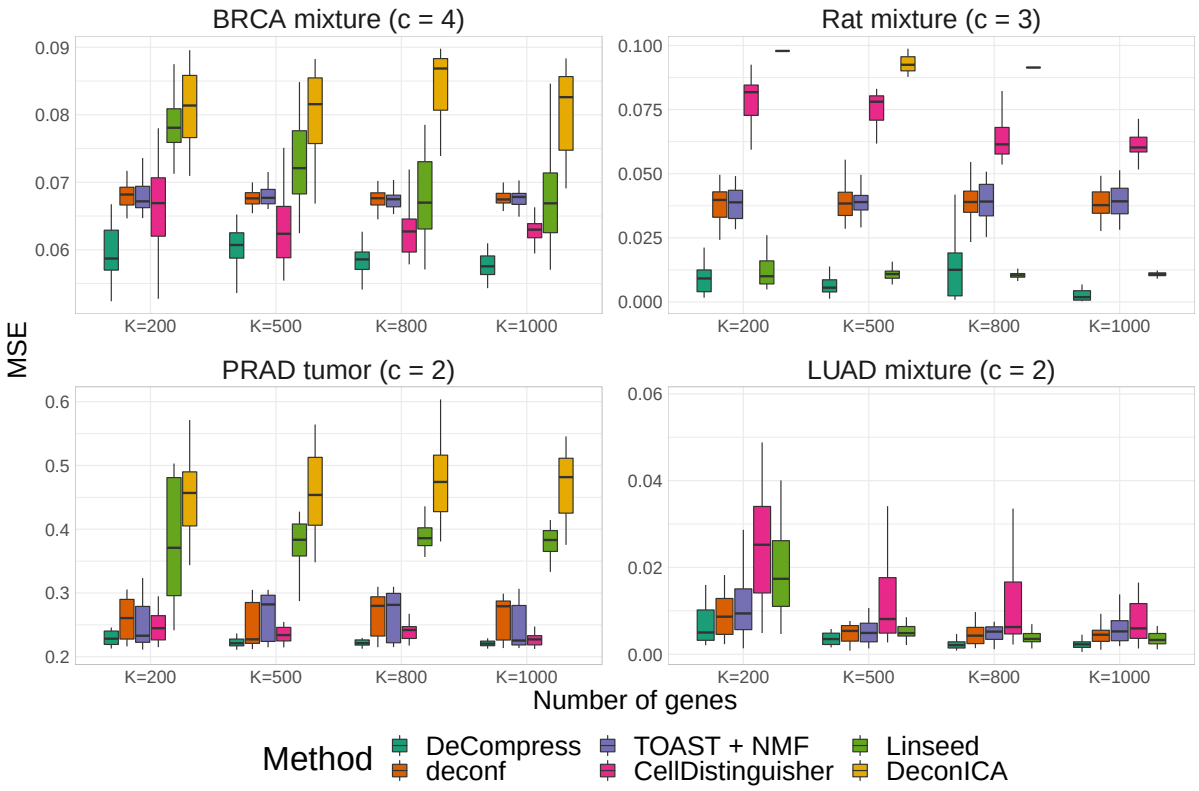


**Figure S34:** *Benchmarking of deconvolution performance using DeCompress and 5 other reference-free deconvolution in published data examples.* Boxplots of MSE ($Y$-axis) over 25 pseudo-targeted panels using four published datasets over 200, 500, 800, and 1000 genes ($X$-axis). This plot shows the same results as Figure 5.2C with the addition of DeconICA.

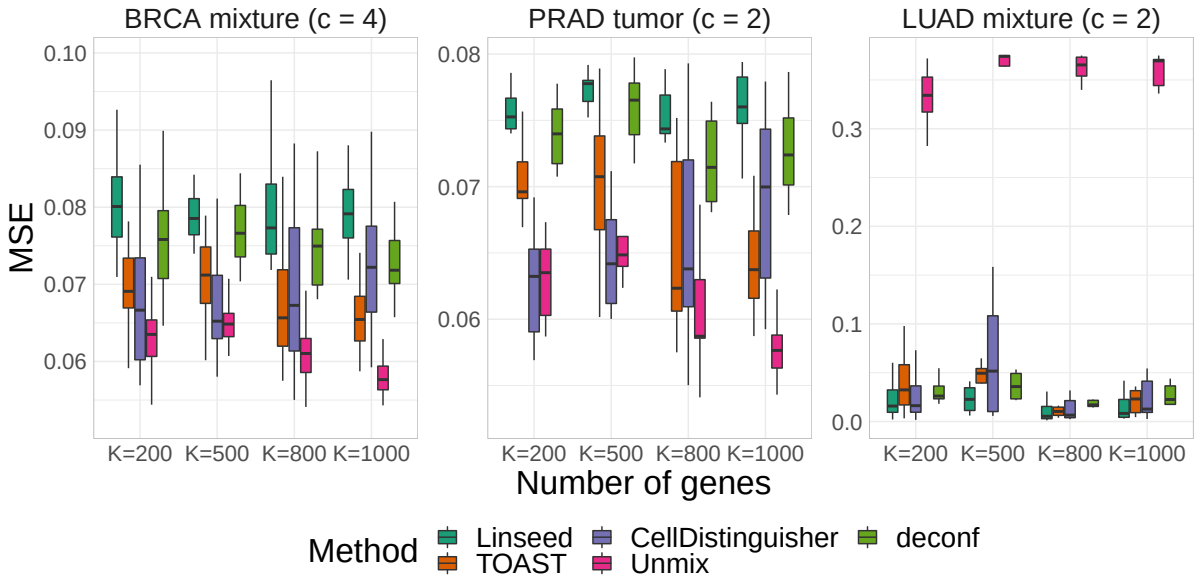


**Figure S35:** *Comparison of deconvolution performance using decompressed matrix in DeCompress across various methods.* Boxplots of MSE ($Y$-axis) between true and estimated cell-type proportions across pseudo-targeted panels of differing numbers of genes. We compare four reference-free methods (deconf[12], Linseed[13], iterative non-negative matrix factorization with feature selection using TOAST[14], CellDistinguisher[15]) and a reference-based method (unmix[16]) that uses cell-type specific expressions estimated from the reference. Here, we present results from the breast cancer cell line mixtures[5], prostate tumor[7], and lung adenocarcinoma cell line mixtures[8]. We do not include DeconICA[17] in this benchmarking due to large errors across all three datasets.

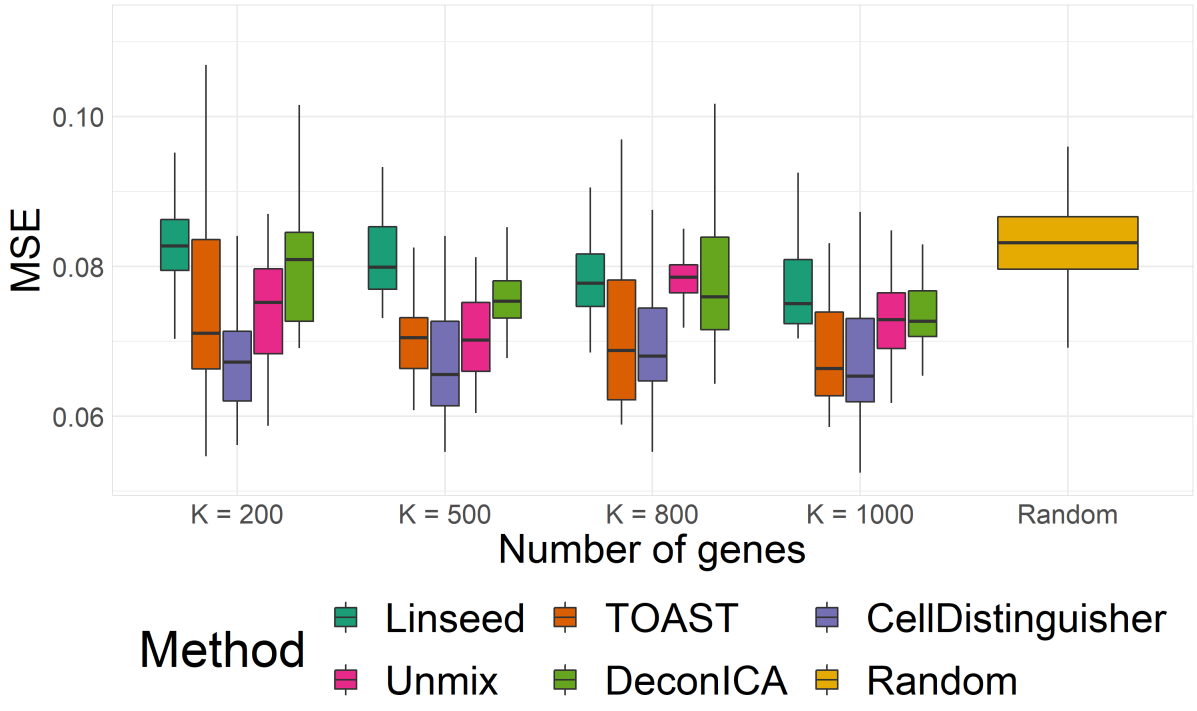


**Figure S36:** *Deconvolution of breast cancer cell mixture using TCGA-LUAD reference.* MSE ($Y$-axis) across 25 psuedo-targeted panels with different numbers of genes ($X$-axis) of using various reference-free deconvolution methods on decompresed breast cancer cell line data using TCGA-LUAD reference data. The yellow box-plot gives a distribution of the MSE for 1,000 randomly generated cell-type proportions

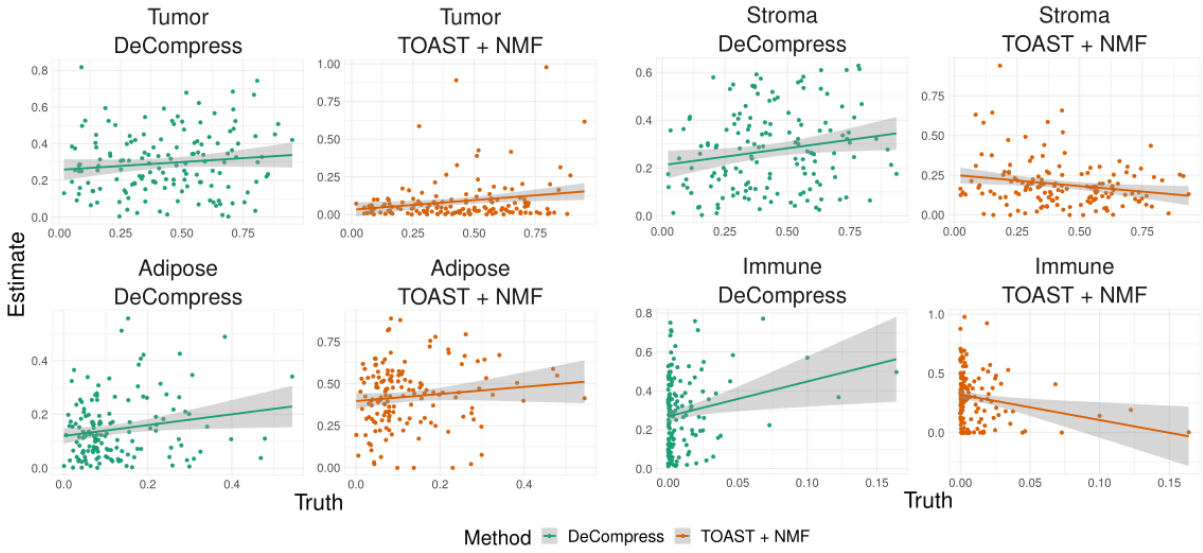


**Figure S37:** *Scatter-plot of known and estimated cell-type proportions in CBCS using DeCompress and TOAST + NMF.* Plots of true ($X$-axis) and estimated ($Y$-axis) cell-type proportions in CBCS using DeCompress and TOAST + NMF (most accurate benchmarked reference-free method). True cell-type proportions are taken as measurement by a study pathologist for 148 samples. A reference smoothed linear trend line is provided for reference.

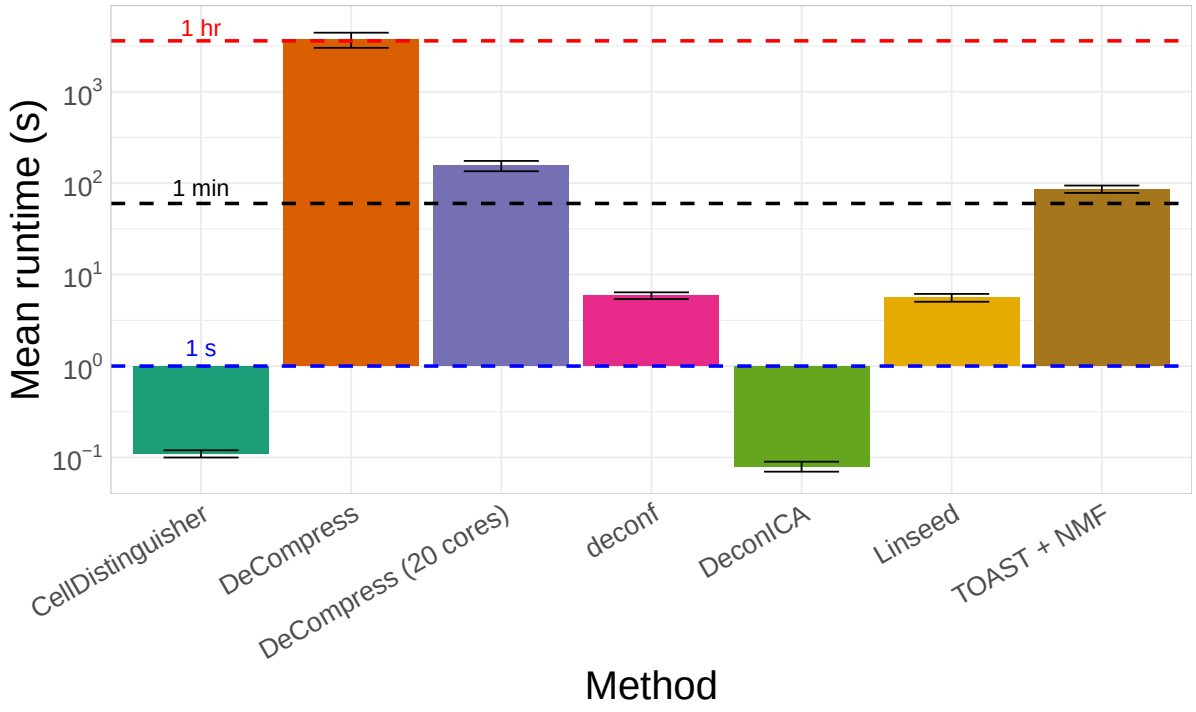


**Figure S38:** *Comparison of run-times for DeCompress and benchmarked reference-free deconvolution methods.* Mean runtimes in seconds ($X$-axis on logarithmic scale) for methods benchmarked ($Y$-axis): CellDistinguisher, DeCompress (in serial), DeCompress (in parallel with 20 cores), deconf, DeconICA, Linseed, iterative non-negative matrix factorization with feature selection using TOAST. These runtimes were generated by running all methods on CBCS data (1,199 samples with 407 genes). DeCompress was run using TCGA-BRCA (1,212 samples) as a reference. The error bar gives an interval of one standard deviation around the mean runtime. The blue, black, and red dotted lines provide references for 1 second, 1 minute, and 1 hour.

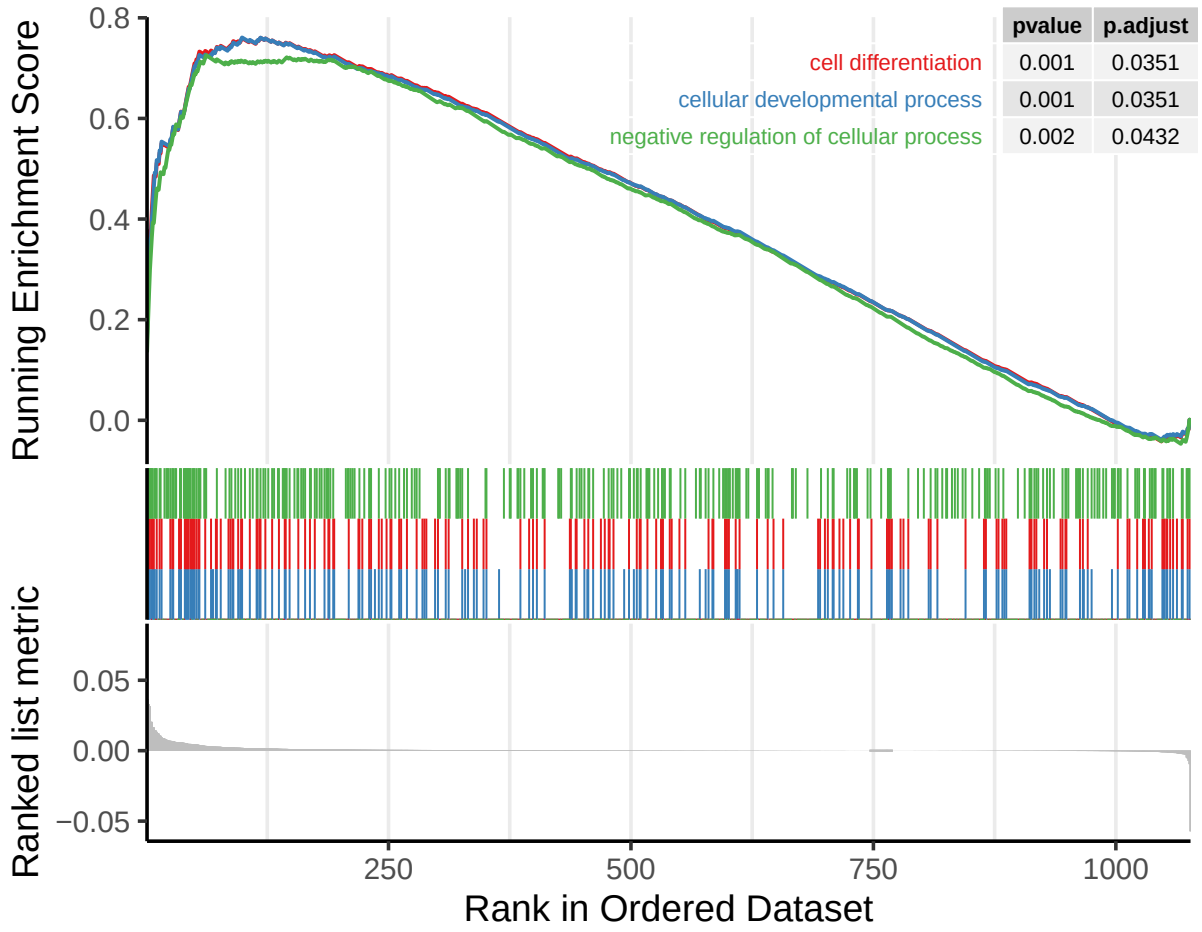


**Figure S39:** *Gene set enrichment plot for combined C3 and C4 gene signature.* The green, blue, and red lines in the top panel of the plot represents the running enrichment score (ES) for the corresponding gene ontology as the analysis goes down the ranked list. The peak gives the final ES. The green, blue, and red lines in the middle of the plot shows where the members of ontological groups in the dataset first appear in the ranked list. The bottom panel shows the value of the ranking metric as it moves down the list of the ranked genes.

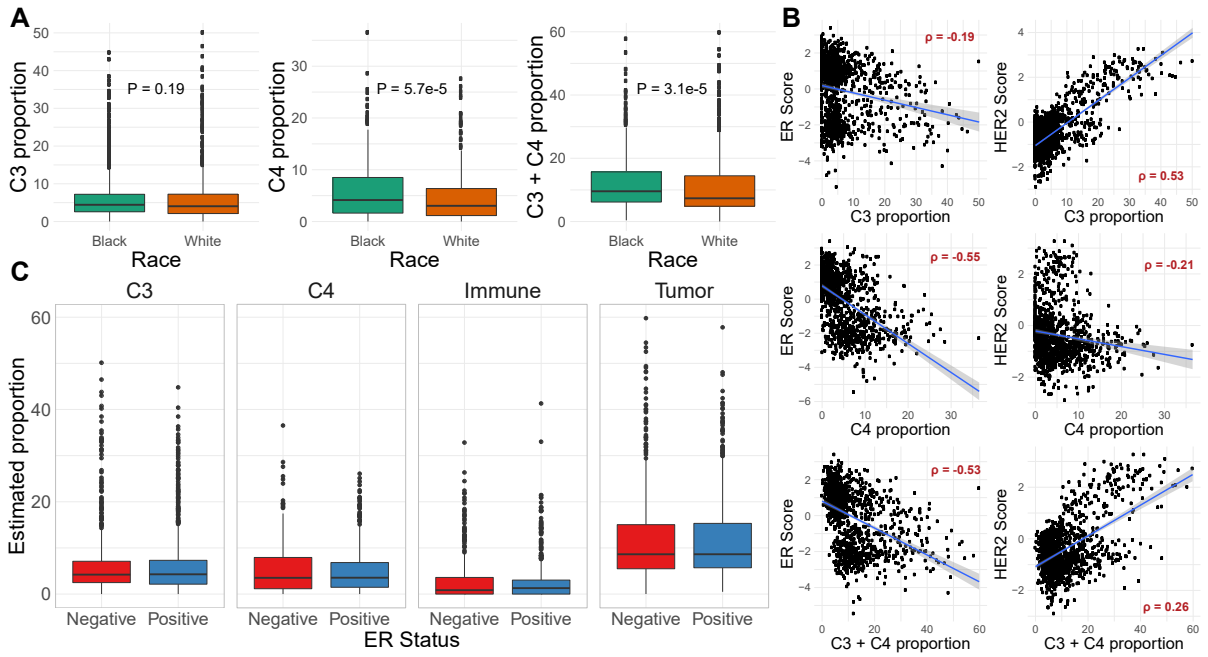


**Figure S40:** *Comparison of compartment proportion estimates with race and different clinical subtype metrics.* (A) Boxplot of C3, C4, and C3 + C4 proportions across race with $P$-value of Wilcoxon rank-sum test provided. (B) Scatterplot of compartment proportions ($X$-axis) and ER or HER2 score from PAM50 classification algorithm. A regression line is provided with a Spearman correlation $\rho$ for reference. (C) Boxplot of C3, C4, immune, and tumor compartment estimates acros clinical ER status.

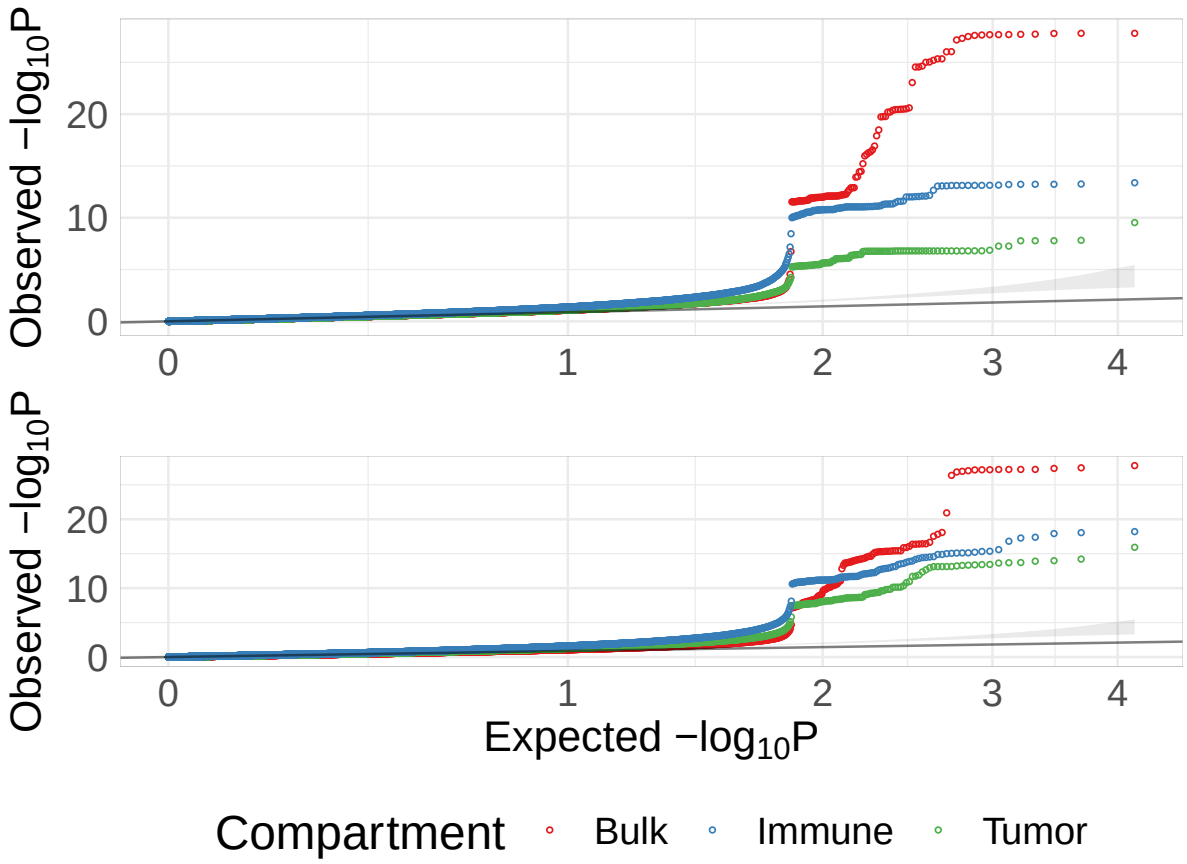


**Figure S41:** *QQ-plots for bulk, tumor-, and immune-specific eQTL models.* QQ-plots from cis-eQTL analysis with expected $-\log_{10} P$-values ($X$-axis) and observed $-\log_{10} P$-values ($Y$-axis) colored by bulk (red), immune- (blue), and tumor-specific (green) models. A 45-degree line is provided for reference.

| Baseline | | |
|---|---|---|
| ***Covariate*** | ***Hazard Ratio (90% adjusted CI)*** | ***FDR-adjusted P*** |
| *PAM50: HER2* | 1.37 (0.97,1.95) | 0.310 |
| *PAM50: LumA* | 0.55 (0.39, 0.79) | 0.041 |
| *PAM50: LumB* | 1.22 (0.86, 1.72) | 0.220 |
| *Race: White* | 0.76 (0.59, 1.00) | 0.110 |
| *Age (in 10 yrs)* | 0.84 (0.75, 0.95) | 0.072 |
| *Compartment (in 10%)* | | |
| *HER2/Compartment* | | |
| *LumA/Compartment* | | |
| *LumB/Compartment* | | |
| **C3** | | |
| *PAM50: HER2* | 1.65 (0.87, 3.11) | 0.260 |
| *PAM50: LumA* | 0.52 (0.30, 0.89) | 0.064 |
| *PAM50: LumB* | 1.15 (0.64, 2.05) | 0.782 |
| *Race: White* | 0.76 (0.57, 1.03) | 0.214 |
| *Age (in 10 yrs)* | 0.84 ( 0.74, 0.96) | 0.064 |
| *Compartment (in 10%)* | 1.07 (0.68, 1.68) | 0.782 |
| *HER2/Compartment* | 0.86 (0.50, 2.41) | 0.782 |
| *LumA/Compartment* | 1.12 (0.59, 2.11) | 0.782 |
| *LumB/Compartment* | 1.11 (0.51, 2.41) | 0.782 |
| **C4** | | |
| *PAM50: HER2* | 2.57 (1.42, 4.67) | 0.026 |
| *PAM50: LumA* | 0.90 (0.51, 1.60) | 0.761 |
| *PAM50: LumB* | 2.30 (1.34, 3.94) | 0.026 |
| *Race: White* | 0.76 (0.59, 0.98) | 0.125 |
| *Age (in 10 yrs)* | 0.86 (0.77, 0.96) | 0.045 |
| *Compartment (in 10%)* | 1.69 (1.21, 2.37) | 0.026 |
| *HER2/Compartment* | 0.47 (0.19, 1.16) | 0.200 |
| *LumA/Compartment* | 0.66 (0.27, 1.59) | 0.475 |
| *LumB/Compartment* | 0.40 (0.15, 1.02) | 0.146 |
| **Tumor** | | |
| *PAM50: HER2* | 2.51 (1.17, 5.41) | 0.070 |
| *PAM50: LumA* | 0.75 (0.37, 1.50) | 0.450 |
| *PAM50: LumB* | 1.88 (0.90, 3.92) | 0.124 |
| *Race: White* | 0.77 (0.57, 1.04) | 0.124 |
| *Age (in 10 yrs)* | 0.84 (0.74, 0.96) | 0.070 |
| *Compartment (in 10%)* | 1.32 (1.01, 1.74) | 0.101 |
| *HER2/Compartment* | 0.69 (0.48, 1.00) | 0.101 |
| *LumA/Compartment* | 0.87 (0.55, 1.39) | 0.562 |
| *LumB/Compartment* | 0.75 (0.41, 1.38) | 0.446 |
| **Immune** | | |
| *PAM50: HER2* | 1.64 (1.08, 2.50) | 0.096 |
| *PAM50: LumA* | 0.51 (0.33, 0.79) | 0.042 |
| *PAM50: LumB* | 1.30 (0.86, 1.98) | 0.369 |
| *Race: White* | 0.77 (0.59, 1.01) | 0.183 |
| *Age (in 10 yrs)* | 0.84 (0.75, 0.95) | 0.042 |
| *Compartment (in 10%)* | 1.03 (0.70, 1.53) | 0.878 |
| *HER2/Compartment* | 0.48 (0.19, 1.19) | 0.250 |
| *LumA/Compartment* | 1.47 (0.65, 3.33) | 0.494 |
| *LumB/Compartment* | 0.74 (0.29, 1.84) | 0.606 |

**Table S8:** *Results from bulk and compartment-specific survival models with PAM50 molecular subtype.* Hazard ratio estimates, 90% FDR-adjusted confidence intervals, and FDR-adjusted P-values for baseline and compartment-specific interaction Cox models for breast cancer-specific survival.

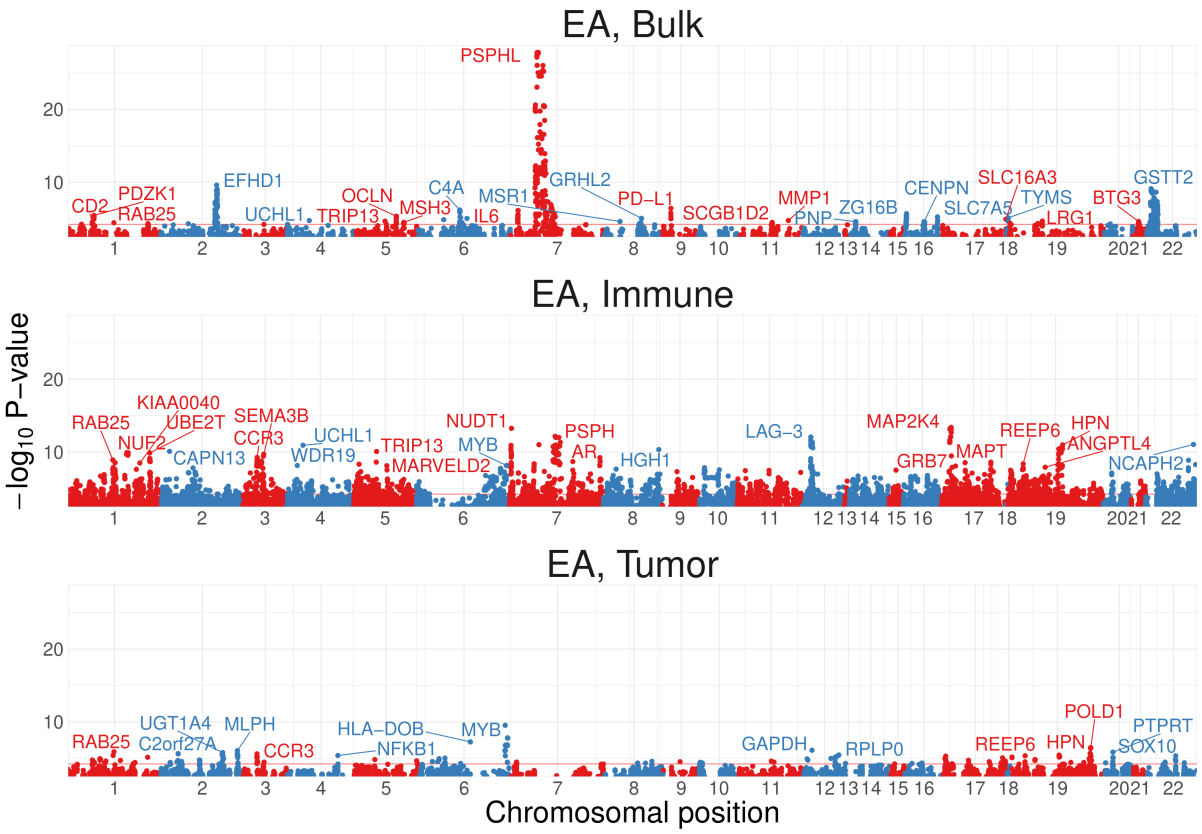


**Figure S42:** *Manhattan plot of cis-eQTLs across the genome in EA CBCS samples.* $-\log_{10} P$-values of eQTL association ($Y$-axis) across chromosomal position of *cis*-eQTLs across bulk (top), immune (middle), and tumor (bottom) models. Top *cis*-eGenes are labelled.

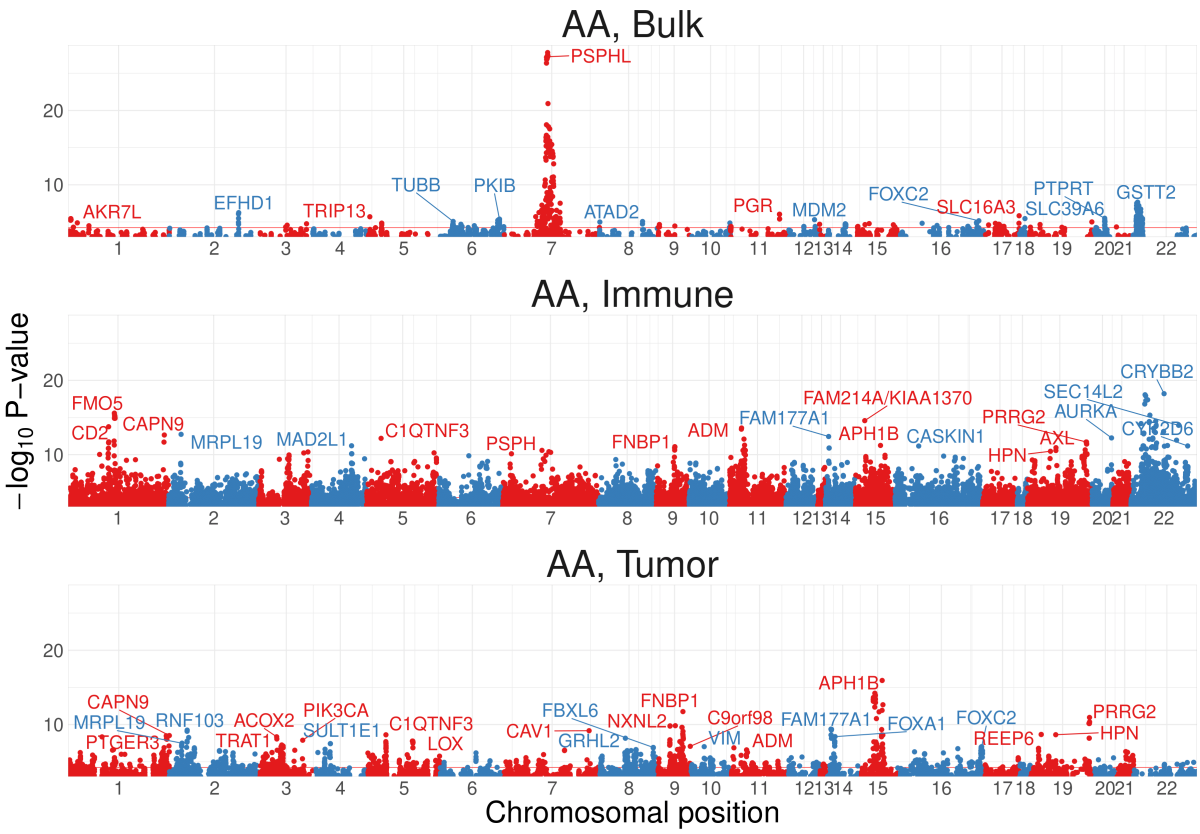


**Figure S43:** *Manhattan plot of cis-eQTLs across the genome in AA CBCS samples.* $-\log_{10} P$-values of eQTL association ($Y$-axis) across chromosomal position of *cis*-eQTLs across bulk (top), immune (middle), and tumor (bottom) models. Top *cis*-eGenes are labelled.

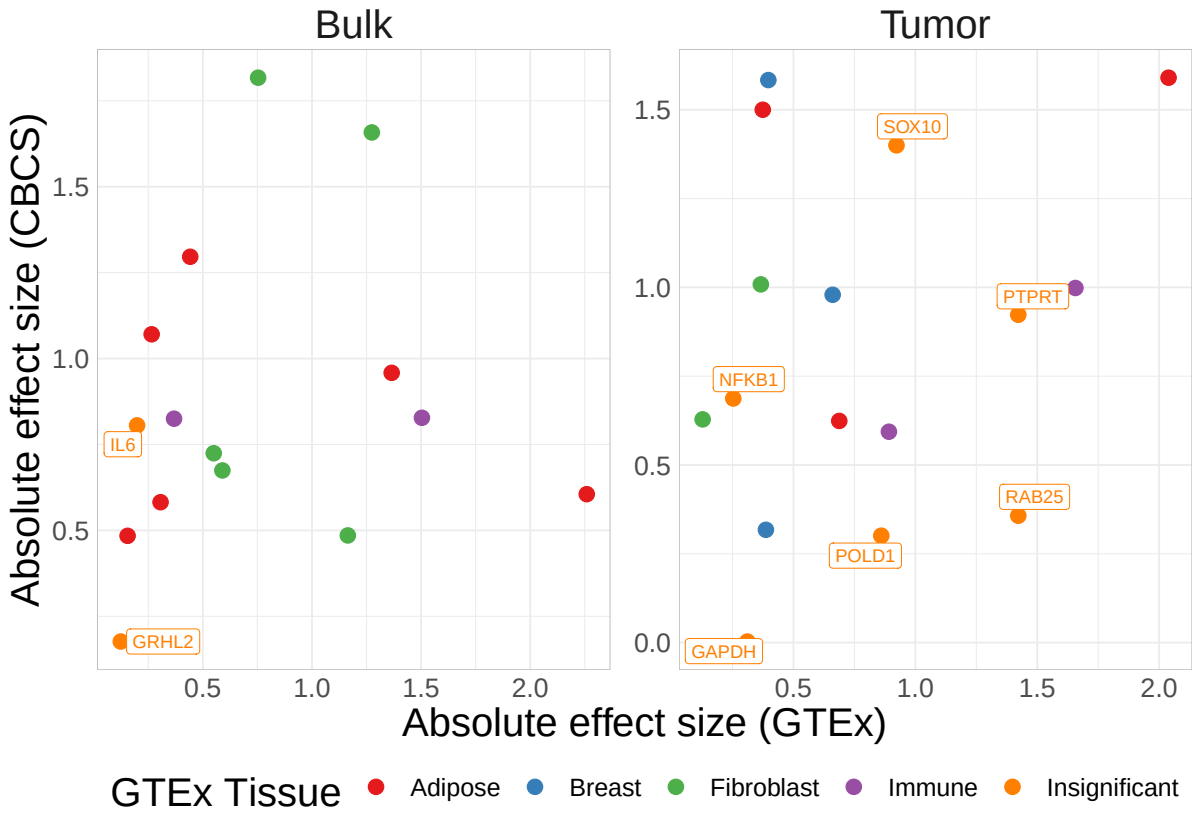


**Figure S44:** *Cross-referencing of bulk and tumor-specific CBCS EA cis-eGenes with GTEx.* Comparison of absolute effect sizes of eGenes with significant *cis*-eQTLs in EA CBCS ($Y$-axis) and GTEx ($X$-axis) over tissue type, stratified by bulk and tumor-specific eQTLs. eGenes are colored by the GTEx tissue that shows the eQTL with smallest $P$-value.

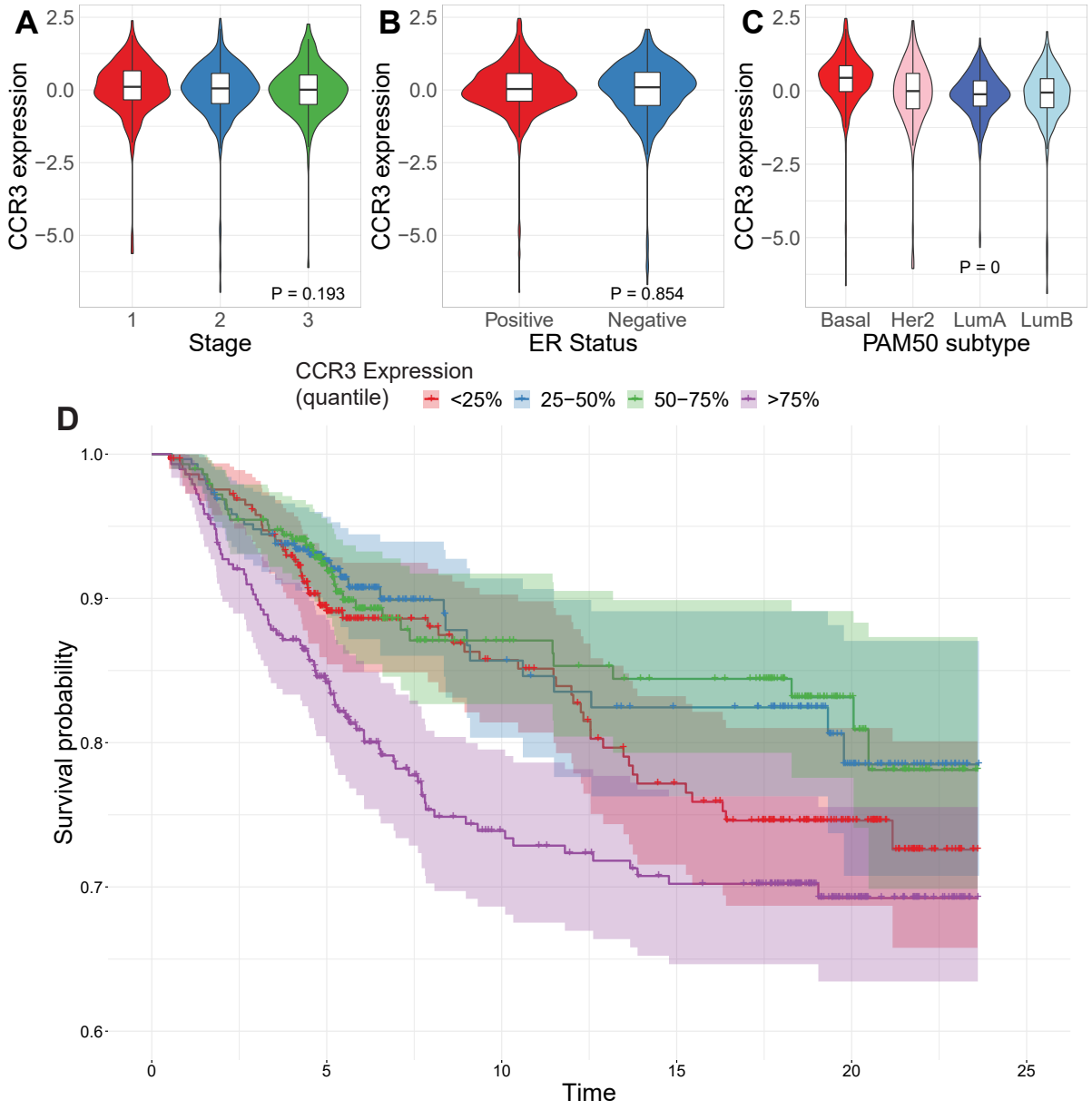


**Figure S45:** *Associations of CCR3 expression across clinical variables, subtypes, and mortality.* Violin plots of *CCR3* expression across breast tumor stage (A), estrogen status (B), and PAM50 molecular subtype (C). (D) Kaplan-Meier curves for breast cancer-specific survival across four quantiles of *CCR3* expression.

# REFERENCES


[1] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple. Technical Report 1, 1995.

[2] Jean Charles Lambert, Carla A. Ibrahim-Verbaas, Denise Harold, Adam C. Naj, Rebecca Sims, Céline Bellenguez, Gyungah Jun, Anita L. DeStefano, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nature Genetics*, 45(12):1452–1458, 12 2013.

[3] Alexander Gusev, Arthur Ko, Huwenbo Shi, Gaurav Bhatia, Wonil Chung, Brenda W J H Penninx, Rick Jansen, Eco J C de Geus, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, 3 2016.

[4] May Sabry, Agnieszka Zubiak, Simon P Hood, Poppy Simmonds, Helena Arellano-Ballestero, Eily Cournoyer, Meghavi Mashar, A Graham Pockley, et al. Tumor- and cytokine-primed human natural killer cells exhibit distinct phenotypic and transcriptional signatures. *PloS one*, 14(6):e0218674, 2019.

[5] Kai Kang, Qian Meng, Igor Shats, David M. Umbach, Melissa Li, Yuanyuan Li, Xiaoling Li, and Leping Li. CDSeq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLOS Computational Biology*, 15(12):e1007510, 12 2019.

[6] Shai S. Shen-Orr, Robert Tibshirani, Purvesh Khatri, Dale L. Bodian, Frank Staedtler, Nicholas M. Perry, Trevor Hastie, Minnie M. Sarwal, et al. Cell type-specific gene expression differences in complex tissues. *Nature Methods*, 7(4):287–289, 4 2010.

[7] Svitlana Tyekucheva, Michaela Bowden, Clyde Bango, Francesca Giunchi, Ying Huang, Chensheng Zhou, Arrigo Bondi, Rosina Lis, et al. Stromal and epithelial transcriptional map of initiation progression and metastatic potential of human prostate cancer. *Nature Communications*, 8(1), 12 2017.

[8] Aliaksei Z Holik, Charity W Law, Ruijie Liu, Zeya Wang, Wenyi Wang, Jaeil Ahn, Marie-Liesse Asselin-Labat, Gordon K Smyth, et al. RNA-seq mixology: designing realistic control experiments to compare protocols and analysis methods. *Nucleic Acids Research*, 45(5), 2016.

[9] Kyriaki Michailidou, Per Hall, Anna Gonzalez-Neira, Maya Ghoussaini, Joe Dennis, Roger L Milne, Marjanka K Schmidt, Jenny Chang-Claude, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature Genetics*, 45(4):353–361, 4 2013.

[10] Kyriaki Michailidou, Jonathan Beesley, Sara Lindstrom, Sander Canisius, Joe Dennis, Michael J Lush, Mel J Maranian, Manjeet K Bolla, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature Genetics*, 47(4):373–380, 4 2015.

[11] Kyriaki Michailidou, Sara Lindström, Joe Dennis, Jonathan Beesley, Shirley Hui, Siddhartha Kar, Audrey Lemaçon, Penny Soucy, et al. Association analysis identifies 65 new breast cancer risk loci. *Nature*, 551(7678):92–94, 2017.

[12] Dirk Repsilber, Sabine Kern, Anna Telaar, Gerhard Walzl, Gillian F Black, Joachim Selbig, Shreemanta K Parida, Stefan HE Kaufmann, et al. Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics*, 11(1):27, 12 2010.

[13] Konstantin Zaitsev, Monika Bambouskova, Amanda Swain, and Maxim N. Artyomov. Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nature Communications*, 10(1), 12 2019.

[14] Ziyi Li and Hao Wu. TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biology*, 20(1):190, 12 2019.

[15] Lee A. Newberg, Xiaowei Chen, Chinnappa D. Kodira, and Maria I. Zavodszky. Computational de novo discovery of distinguishing genes for biological processes and cell types in complex tissues. *PLOS ONE*, 13(3):e0193067, 3 2018.

[16] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 12 2014.

[17] Urszula Czerwinska. DeconICA, 5 2018.

[18] Gary K Geiss, Roger E Bumgarner, Brian Birditt, Timothy Dahl, Naeem Dowidar, Dwayne L Dunaway, H Perry Fell, Sean Ferree, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature Biotechnology*, 26(3):317–325, 3 2008.

[19] Margaret H. Veldman-Jones, Roz Brant, Claire Rooney, Catherine Geh, Hollie Emery, Chris G. Harbron, Mark Wappett, Alan Sharpe, et al. Evaluating Robustness and Sensitivity of the NanoString Technologies nCounter Platform to Enable Multiplexed Gene Expression Analysis of Clinical Samples. *Cancer Research*, 75(13):2587–2593, 7 2015.

[20] Fara Brasó-Maristany, Simone Filosto, Steven Catchpole, Rebecca Marlow, Jelmar Quist, Erika Francesch-Domenech, Darren A Plumb, Leila Zakka, et al. PIM1 kinase regulates cell death, tumor growth and chemotherapy response in triple-negative breast cancer. *Nature Medicine*, 22(11):1303–1313, 11 2016.

[21] Alejandra Urrutia, Darragh Duffy, Vincent Rouilly, Céline Posseme, Raouf Djebali, Gabriel Illanes, Valentina Libri, Benoit Albaud, et al. Standardized Whole-Blood Transcriptional Profiling Enables the Deconvolution of Complex Induced Immune Responses. *Cell Reports*, 16(10):2777–2791, 9 2016.

[22] D. W. Scott, G. W. Wright, P. M. Williams, C.-J. Lih, W. Walsh, E. S. Jaffe, A. Rosenwald, E. Campo, et al. Determining cell-of-origin subtypes of diffuse large B-cell lymphoma using gene expression in formalin-fixed paraffin-embedded tissue. *Blood*, 123(8):1214–1217, 2 2014.

[23] Stanley W. K. Ng, Amanda Mitchell, James A. Kennedy, Weihsu C. Chen, Jessica McLeod, Narmin Ibrahimova, Andrea Arruda, Andreea Popescu, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*, 540(7633):433–437, 12 2016.

[24] Melissa A. Troester, Xuezheng Sun, Emma H. Allott, Joseph Geradts, Stephanie M. Cohen, Chiu-Kit Tse, Erin L. Kirk, Leigh B. Thorne, et al. Racial Differences in PAM50 Subtypes in the Carolina Breast Cancer Study. *JNCI: Journal of the National Cancer Institute*, 110(2):176–182, 2 2018.

[25] Brett Wallden, James Storhoff, Torsten Nielsen, Naeem Dowidar, Carl Schaper, Sean Ferree, Shuzhen Liu, Samuel Leung, et al. Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Medical Genomics*, 8(1):54, 12 2015.

[26] André Filipe Vieira and Fernando Schmitt. An Update on Breast Cancer Multigene Prognostic Tests-Emergent Clinical Biomarkers. *Frontiers in medicine*, 5:248, 2018.

[27] Johann A. Gagnon-Bartsch and Terence P. Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics (Oxford, England)*, 13(3):539–552, 7 2012.

[28] Ramyar Molania, Johann A Gagnon-Bartsch, Alexander Dobrovic, and Terence P Speed. A new normalization for Nanostring nCounter gene expression data. *Nucleic Acids Research*, 47(12):6073–6083, 7 2019.

[29] Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology*, 32(9):896–902, 9 2014.

[30] nSolver™ 4.0 Analysis Software User Manual, 2018.

[31] Daryl Waggott, Kenneth Chu, Shaoming Yin, Bradly G Wouters, Fei-Fei Liu, and Paul C Boutros. Gene expression NanoStringNorm: an extensible R package for the pre-processing of NanoString mRNA and miRNA data. *BIOINFORMATICS APPLICATIONS NOTE*, 28(11):1546–1548, 2012.

[32] James R Perkins, John M Dawes, Steve B McMahon, David LH H Bennett, Christine Orengo, and Matthias Kohl. ReadqPCR and NormqPCR: R packages for the reading, quality checking and normalisation of RT-qPCR quantification cycle (Cq) data. *BMC Genomics*, 13(1):296, 7 2012.

[33] Hong Wang, Craig Horbinski, Hao Wu, Yinxing Liu, Shaoyi Sheng, Jinpeng Liu, Heidi Weiss, Arnold J. Stromberg, et al. NanoStringDiff: a novel statistical method for differential expression analysis based on NanoString nCounter data. *Nucleic Acids Research*, 44(20):gkw677, 7 2016.

[34] Hao Wu, Chi Wang, and Zhijin Wu. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*, 14(2):232–243, 4 2013.

[35] Jenna Lilyquist, Kathryn J Ruddy, Celine M Vachon, and Fergus J Couch. Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 27(4):380–394, 4 2018.

[36] Maria Escala-Garcia, Qi Guo, Thilo Dörk, Sander Canisius, Renske Keeman, Joe Dennis, Jonathan Beesley, Julie Lecarpentier, et al. Genome-wide association study of germline variants and breast cancer-specific mortality. *British Journal of Cancer*, 120(6):647–657, 3 2019.

[37] Ailith Pirie, Qi Guo, Peter Kraft, Sander Canisius, Diana M Eccles, Nazneen Rahman, Heli Nevanlinna, Constance Chen, et al. Common germline polymorphisms associated with breast cancer-specific survival. *Breast Cancer Research*, 17(1):58, 12 2015.

[38] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences of the United States of America*, 106(23):9362–9367, 6 2009.

[39] Alexander Gusev, S. Hong Hong Lee, Gosia Trynka, Hilary Finucane, Bjarni J. J Vilhjálmsson, Han Xu, Chongzhi Zang, Stephan Ripke, et al. Partitioning Heritability of Regulatory and Cell-Type-Specific Variants across 11 Common Diseases. *American Journal of Human Genetics*, 95(5):535, 11 2014.

[40] Dongwon Lee, David U Gorkin, Maggie Baker, Benjamin J Strober, Alessandro L Asoni, Andrew S McCallion, and Michael A Beer. A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics*, 47(8):955–961, 8 2015.

[41] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*, 169(7):1177–1186, 6 2017.

[42] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, 9 2015.

[43] Lang Wu, Wei Shi, Jirong Long, Xingyi Guo, Kyriaki Michailidou, Jonathan Beesley, Manjeet K Bolla, Xiao-Ou Shu, et al. A transcriptome-wide association study of 229,000 women identifies new candidate susceptibility genes for breast cancer. *Nature genetics*, 50(7):968–978, 2018.

[44] François Aguet, Alvaro N Barbeira, Rodrigo Bonazzola, Andrew Brown, Stephane E Castel, Brian Jo, Silva Kasela, Sarah Kim-Hellmuth, et al. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv*, pp. 787903, 2019.

[45] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedländer, Peter A.C. 'T Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, et al. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

[46] Alexis Battle, Sara Mostafavi, Xiaowei Zhu, James B. Potash, Myrna M. Weissman, Courtney McCormick, Christian D. Haudenschild, Kenneth B. Beckman, et al. Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, 24(1):14–24, 1 2014.

[47] Robert Tibshirani and Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, 58:267–288, 1994.

[48] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22, 2 2010.

[49] Jian Yang, Andrew Bakshi, Zhihong Zhu, Gibran Hemani, Anna A E Vinkhuyzen, Sang Hong Lee, Matthew R Robinson, John R B Perry, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature genetics*, 47(10):1114–1120, 10 2015.

[50] Bogdan Pasaniuc, Noah Zaitlen, Huwenbo Shi, Gaurav Bhatia, Alexander Gusev, Joseph Pickrell, Joel Hirschhorn, David P. Strachan, et al. Fast and accurate imputation of summary statistics enhances evidence of functional enrichment. *Bioinformatics*, 30(20):2906–2914, 10 2014.

[51] G. K. Robinson. That BLUP is a Good Thing: The Estimation of Random Effects.

[52] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic Modeling with Bayesian Sparse Linear Mixed Models. *PLoS Genetics*, 9(2), 2 2013.

[53] Yiming Hu, Mo Li, Qiongshi Lu, Haoyi Weng, Jiawei Wang, Seyedeh M. Zekavat, Zhaolong Yu, Boyang Li, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature Genetics*, 51(3):568–576, 3 2019.

[54] Sini Nagpal, Xiaoran Meng, Michael P. Epstein, L. C. Tsoi, Matthew Patrick, Greg Gibson, Philip L. De Jager, David A. Bennett, et al. TIGAR: An Improved Bayesian Tool for Transcriptomic Data Imputation Enhances Gene Mapping of Complex Traits. *American Journal of Human Genetics*, 105(2):258–266, 8 2019.

[55] Nicholas Mancuso, Malika K. Freund, Ruth Johnson, Huwenbo Shi, Gleb Kichaev, Alexander Gusev, and Bogdan Pasaniuc. Probabilistic fine-mapping of transcriptome-wide association studies. *Nature Genetics*, 51(4):675–682, 4 2019.

[56] Ryan Sun and Xihong Lin. Set-Based Tests for Genetic Association Using the Generalized Berk-Jones Statistic. Technical report, 2017.

[57] Ping Zeng and Xiang Zhou. Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nature Communications*, 8(1), 12 2017.

[58] Peter Müller and Riten Mitra. Bayesian Nonparametric Inference-Why and How. *Bayesian Analysis*, 8(2):269–302, 2013.

[59] David M Blei and Michael I Jordan. Variational Inference for Dirichlet Process Mixtures. Technical Report 1, 2006.

[60] Peter Carbonetto and Matthew Stephens. Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. *Bayesian Analysis*, 7(1):73–108, 2012.

[61] Farhad Hormozdiari, Gleb Kichaev, Wen-Yun Yang, Bogdan Pasaniuc, and Eleazar Eskin. Identification of causal genes for complex traits. *Bioinformatics*, 31, 2015.

[62] Gleb Kichaev, Wen Yun Yang, Sara Lindstrom, Farhad Hormozdiari, Eleazar Eskin, Alkes L. Price, Peter Kraft, and Bogdan Pasaniuc. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genetics*, 10(10), 10 2014.

[63] N. H. Barton, A. M. Etheridge, and A. Véber. The infinitesimal model: Definition, derivation, and implications. *Theoretical Population Biology*, 118:50–73, 12 2017.

[64] Frank W. Stearns. One hundred years of pleiotropy: A retrospective, 11 2010.

[65] Alexandra C. Nica and Emmanouil T. Dermitzakis. Expression quantitative trait loci: Present and future, 6 2013.

[66] Chang Liu, Yi Xiao, Zhonghua Tao, and Xi-Chun Hu. Identification of immune microenvironment subtypes of breast cancer in TCGA set: Implications for immunotherapy. *Journal of Clinical Oncology*, 37(15_suppl):e14205–e14205, 5 2019.

[67] Brandon L. Pierce, Lin Tong, Lin S. Chen, Ronald Rahaman, Maria Argos, Farzana Jasmine, Shantanu Roy, Rachelle Paul-Brutus, et al. Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 South Asians. *PLoS Genetics*, 10(12), 12 2014.

[68] Boel Brynedal, Jin Myung Choi, Towfique Raj, Robert Bjornson, Barbara E. Stranger, Benjamin M. Neale, Benjamin F. Voight, and Chris Cotsapas. Large-Scale trans-eQTLs Affect Hundreds of Transcripts and Mediate Patterns of Transcriptional Co-regulation. *American Journal of Human Genetics*, 100(4):581–591, 4 2017.

[69] Fan Yang, Jiebiao Wang, Brandon L. Pierce, and Lin S. Chen. Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome Research*, 27(11):1859–1871, 11 2017.

[70] François Aguet, Andrew A. Brown, Stephane E. Castel, Joe R. Davis, Yuan He, Brian Jo, Pejman Mohammadi, Yo Son Park, et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 10 2017.

[71] David P. Mackinnon, Chondra M. Lockwood, and Jason Williams. Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivariate Behavioral Research*, 39(1):99–128, 2004.

[72] Nayang Shan, Zuoheng Wang, and Lin Hou. Identification of trans-eQTLs using mediation analysis with multiple mediators. *BMC Bioinformatics*, 20, 3 2019.

[73] Kevin J Gleason, Fan Yang, Brandon L Pierce, Xin He, and Lin S Chen. Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits. *bioRxiv*, pp. 579581, 2019.

[74] Kyle Kai How Farh, Alexander Marson, Jiang Zhu, Markus Kleinewietfeld, William J. Housley, Samantha Beik, Noam Shoresh, Holly Whitton, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–343, 2 2015.

[75] Panos Roussos, Amanda C. Mitchell, Georgios Voloudakis, John F. Fullard, Venu M. Pothula, Jonathan Tsang, Eli A. Stahl, Anastasios Georgakopoulos, et al. A Role for Noncoding Variation in Schizophrenia. *Cell Reports*, 9(4):1417–1429, 11 2014.

[76] Mads Engel Hauberg, Wen Zhang, Claudia Giambartolomei, Oscar Franzén, David L. Morris, Timothy J. Vyse, Arno Ruusalepp, Menachem Fromer, et al. Large-Scale Identification of Common Trait and Disease Variants Affecting Gene Expression. *American Journal of Human Genetics*, 100(6):885–894, 6 2017.

[77] Wen Zhang, Georgios Voloudakis, Veera M. Rajagopal, Ben Readhead, Joel T. Dudley, Eric E. Schadt, Johan L. M. Björkegren, Yungil Kim, et al. Integrative transcriptome imputation reveals tissue-specific and shared biological mechanisms mediating susceptibility to complex traits. *Nature Communications*, 10(1):3834, 12 2019.

[78] Roadmap Epigenomics Consortium, Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–329, 2 2015.

[79] Yang I. Li, Bryce Van De Geijn, Anil Raj, David A. Knowles, Allegra A. Petti, David Golan, Yoav Gilad, and Jonathan K. Pritchard. RNA splicing is a primary link between genetic variation and disease. *Science*, 352(6285):600–604, 4 2016.

[80] Marta Tellez-Gabriel, Benjamin Ory, Francois Lamoureux, Marie Francoise Heymann, and Dominique Heymann. Tumour heterogeneity: The key advantages of single-cell analysis, 12 2016.

[81] Kevin McGregor, Sasha Bernatsky, Ines Colmegna, Marie Hudson, Tomi Pastinen, Aurélie Labbe, and Celia M.T. Greenwood. An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies. *Genome Biology*, 17(1):84, 12 2016.

[82] Alexandre Kuhn, Doris Thu, Henry J. Waldvogel, Richard L.M. M Faull, and Ruth Luthi-Carter. Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nature Methods*, 8(11):945–947, 11 2011.

[83] Jerry Guintivano, Martin J. Aryee, and Zachary A. Kaminsky. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics*, 8(3):290–302, 3 2013.

[84] G ; André, H-J Westra, D Arends, T Esko, M J Peters, C Schurmann, and K Schramm. Cell Specific eQTL Analysis without Sorting Cells. *Cell Specific eQTL Analysis without Sorting Cells. PLoS Genet*, 24(5):1005223.

[85] Paul Geeleher, Aritro Nath, Fan Wang, Zhenyu Zhang, Alvaro N. Barbeira, Jessica Fessler, Robert L. Grossman, Cathal Seoighe, et al. Cancer expression quantitative trait loci (eQTLs) can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity. *Genome Biology*, 19(1):130, 12 2018.

[86] Dimitri P Bertsekas. *Convex Optimization Algorithms*. Athena Scientific, Belmot, Massachusetts, 1999.

[87] Yi Zhong, Ying-Wooi Wan, Kaifang Pang, Lionel ML L Chow, and Zhandong Liu. Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, 14(1):89, 2013.

[88] Gerald Quon, Syed Haider, Amit G Deshwar, Ang Cui, Paul C Boutros, and Quaid Morris. Computational purification of individual tumor gene expression profiles leads to significant improvements in prognostic prediction. *Genome Medicine*, 5(3):29, 2013.

[89] Zeya Wang, Shaolong Cao, Jeffrey S. Morris, Jaeil Ahn, Rongjie Liu, Svitlana Tyekucheva, Fan Gao, Bo Li, et al. Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *iScience*, 9:451–460, 11 2018.

[90] Julian Besag. On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):259–279, 7 1986.

[91] Binbin Chen, Michael S. Khodadoust, Chih Long Liu, Aaron M. Newman, and Ash A. Alizadeh. Profiling tumor infiltrating immune cells with CIBERSORT. In *Methods in Molecular Biology*, volume 1711, pp. 243–259. Humana Press Inc., 2018.

[92] Meichen Dong, Aatish Thennavan, Eugene Urrutia, Yun Li, Charles M Perou, Fei Zou, and Yuchao Jiang. SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Briefings in Bioinformatics*, 1 2020.

[93] Jiebiao Wang, Bernie Devlin, and Kathryn Roeder. Using multiple measurements of tissue to estimate subject- and cell-type-specific gene expression. *Bioinformatics*, 8 2019.

[94] Xuefeng Wang, Eric P Xing, and Daniel J Schaid. Kernel methods for large-scale genomic data analysis. *Briefings in bioinformatics*, 16(2):183–192, 3 2015.

[95] Jose Bioucas-Dias. A Variable Splitting Augmented Lagrangian Approach to Linear Spectral Unmixing. 4 2009.

[96] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, 4 edition, 2002.

[97] Saskia Freytag, Johann Gagnon-Bartsch, Terence P. Speed, and Melanie Bahlo. Systematic noise degrades gene co-expression signals but can be corrected. *BMC Bioinformatics*, 16(1):309, 12 2015.

[98] R. A. Irizarry, Bridget Hobbs, Francois Collin, Yasmin D. Beazer□Barclay, Kristen J. Antonellis, Uwe Scherf, and Terence P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 4 2003.

[99] James H Bullard, Elizabeth Purdom, Kasper D Hansen, and Sandrine Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11(1):94, 12 2010.

[100] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, 10 2010.

[101] Gaoxiang Jia, Xinlei Wang, Qiwei Li, W. E.I. Lu, Ximing Tang, Ignacio Wistuba, and Yang Xie. Rcrnorm: An integrated system of random-coefficient hierarchical regression models for normalizing nanostring ncounter data. *Annals of Applied Statistics*, 13(3):1617–1647, 9 2019.

[102] Arjun Bhattacharya, Montserrat García-Closas, Andrew F. Olshan, Charles M. Perou, Melissa A. Troester, and Michael I. Love. A framework for transcriptome-wide association studies in breast cancer in diverse study populations. *Genome Biology*, 21(1):42, 12 2020.

[103] Xiaofeng Dai, Liangjian Xiang, Ting Li, and Zhonghu Bai. Cancer hallmarks, biomarkers and breast cancer molecular subtypes, 2016.

[104] M Elizabeth, H Hammond, Daniel F Hayes, Mitch Dowsett, D Craig Allred, Karen L Hagerty, Sunil Badve, Patrick L Fitzgibbons, et al. American Society of Clinical Oncology/College of American Pathologists Guideline Recommendations for Immunohistochemical Testing of Estrogen and Progesterone Receptors in Breast Cancer. *Journal of Clinical Oncology*, 28(16):2784 – 2795, 2010.

[105] Christina Curtis, Sohrab P. Shah, Suet Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 6 2012.

[106] Charles M. Perou, Therese Sørile, Michael B. Eisen, Matt Van De Rijn, Stefanie S. Jeffrey, Christian A. Ress, Jonathan R. Pollack, Douglas T. Ross, et al. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, 8 2000.

[107] Therese Sørlie, Robert Tibshirani, Joel Parker, Trevor Hastie, J. S. Marron, Andrew Nobel, Shibing Deng, Hilde Johnsen, et al. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America*, 100(14):8418–8423, 7 2003.

[108] Katherine A. Hoadley, Christina Yau, Toshinori Hinoue, Denise M. Wolf, Alexander J. Lazar, Esther Drill, Ronglai Shen, Alison M. Taylor, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, 173(2):291–304, 4 2018.

[109] Joel S Parker, Michael Mullins, Maggie C U Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 27(8):1160–1167, 3 2009.

[110] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(C):53–65, 1987.

[111] A. C. Picornell, I. Echavarria, E. Alvarez, S. López-Tarruella, Y. Jerez, K. Hoadley, J. S. Parker, M. Del Monte-Millán, et al. Breast cancer PAM50 signature: Correlation and concordance between RNA-Seq and digital multiplexed gene expression technologies in a triple negative breast cancer series. *BMC Genomics*, 20(1):452, 6 2019.

[112] Gábor J. Székely and Maria L. Rizzo. The Energy of Data. *Annual Review of Statistics and Its Application*, 4(1):447–479, 3 2017.

[113] N Mantel. The detection of disease clustering and a generalized regression approach. *Cancer research*, 27(2):209–220, 2 1967.

[114] Patrick Breheny, Arnold Stromberg, and Joshua Lambert. P-Value histograms: Inference and diagnostics. *High-Throughput*, 7(3), 9 2018.

[115] Kavleen Sikand, Jagjit Singh, Jey Sabith Ebron, and Girish C Shukla. Housekeeping gene selection advisory: glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and $\beta$-actin are targets of miR-644a. *PloS one*, 7(10):e47510, 2012.

[116] Robert D. Barber, Dan W. Harmer, Robert A. Coleman, and Brian J. Clark. GAPDH as a housekeeping gene: analysis of GAPDH mRNA expression in a panel of 72 human tissues. *Physiological Genomics*, 21(3):389–395, 5 2005.

[117] Chelsea K Raulerson, Arthur Ko, John C Kidd, Kevin W Currin, Sarah M Brotman, Maren E Cannon, Ying Wu, Cassandra N Spracklen, et al. Adipose Tissue Gene Expression Associations Reveal Hundreds of Candidate Genes for Cardiometabolic Traits. 2019.

[118] Chaeyoung Lee. Genome-wide expression quantitative trait loci analysis using mixed models, 8 2018.

[119] Ning Jiang, Minghui Wang, Tianye Jia, Lin Wang, Lindsey Leach, Christine Hackett, David Marshall, and Zewei Luo. A robust statistical method for association-based eQTL analysis. *PLoS ONE*, 6(8), 2011.

[120] Min Kang Hyun, Chun Ye, and Eleazar Eskin. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–1925, 12 2008.

[121] Weiguang Mao, Ryan Hausler, and Maria Chikina. DataRemix: a universal data transformation for optimal inference from gene expression datasets.

[122] Beth Newman, Patricia G. Moorman, Robert Millikan, Bahjat F. Qaqish, Joseph Geradts, Tim E. Aldrich, and Edison T. Liu. The Carolina Breast Cancer Study: integrating population-based epidemiology and molecular biology. *Breast Cancer Research and Treatment*, 35(1):51–60, 1995.

[123] Brionna Y Hair, Sandi Hayes, Chiu-Kit Tse, Mary Beth Bell, and Andrew F Olshan. Racial differences in physical activity among breast cancer survivors: implications for breast cancer care. *Cancer*, 120(14):2174–2182, 7 2014.

[124] Christopher I Amos, Joe Dennis, Zhaoming Wang, Jinyoung Byun, Fredrick R Schumacher, Simon A Gayther, Graham Casey, David J Hunter, et al. The OncoArray Consortium: A Network

for Understanding the Genetic Architecture of Common Cancers. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 26(1):126–135, 2017.

[125] Adam Auton, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, et al. A global reference for human genetic variation, 9 2015.

[126] Jared O'Connell, Deepti Gurdasani, Olivier Delaneau, Nicola Pirastu, Sheila Ulivi, Massimiliano Cocca, Michela Traglia, Jie Huang, et al. A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genetics*, 10(4):e1004234, 4 2014.

[127] Olivier Delaneau, Jonathan Marchini, and Jean-François Zagury. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9(2):179–181, 2 2012.

[128] Bryan N. Howie, Peter Donnelly, and Jonathan Marchini. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genetics*, 5(6):e1000529, 6 2009.

[129] Andrey A Shabalin. Gene expression Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics (Oxford, England)*, 28(10):1353–1358, 5 2012.

[130] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 4 2015.

[131] Seth Rummel, Cayla E Penatzer, Craig D Shriver, and Rachel E Ellsworth. PSPHL and breast cancer in African American women: causative gene or population stratification? *BMC genetics*, 15:38, 3 2014.

[132] Monica D'Arcy, Jodie Fleming, Whitney R. Robinson, Erin L. Kirk, Charles M. Perou, and Melissa A. Troester. Race-associated biological differences among Luminal A breast tumors. *Breast Cancer Research and Treatment*, 152(2):437–448, 7 2015.

[133] Lu Lu, Ashutosh K. Pandey, M. Trevor Houseal, and Megan K. Mulligan. The Genetic Architecture of Murine Glutathione Transferases. *PLOS ONE*, 11(2):e0148230, 2 2016.

[134] Mingfeng Zhang, Soren Lykke-Andersen, Bin Zhu, Wenming Xiao, Jason W Hoskins, Xijun Zhang, Lauren M Rost, Irene Collins, et al. Characterising cis-regulatory variation in the transcriptome of histologically normal and tumour-derived pancreatic tissues. *Gut*, 67(3):521–533, 2018.

[135] Corinna Cortes. Support-Vector Networks. Technical report, 1995.

[136] David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch, Chih-Chung Chang, and Chih-Chen Lin. *Misc Functions of the Department of Statistics (Formerly e1071)*. Comprehensive R Archive Network (CRAN), 2019.

[137] C Calabrese, K Lehmann, L Urban, F Liu, S Erkek, NA A Fonseca, A Kahles, H Kilpinen, et al. Assessing the Gene Regulatory Landscape in 1,188 Human Tumors. *bioRxiv*, pp. 225441, 11 2017.

[138] Jing Gong, Shufang Mei, Chunjie Liu, Yu Xiang, Youqiong Ye, Zhao Zhang, Jing Feng, Renyan Liu, et al. PancanQTL: Systematic identification of cis -eQTLs and trans -eQTLs in 33 cancer

types. *Nucleic Acids Research*, 46(D1):D971–D976, 1 2018.

[139] Brian K. Maples, Simon Gravel, Eimear E. Kenny, and Carlos D. Bustamante. RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *American Journal of Human Genetics*, 93(2):278–288, 8 2013.

[140] Alicia R. Martin, Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, et al. Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *American Journal of Human Genetics*, 100(4):635–649, 4 2017.

[141] Yizhen Zhong, Minoli A. Perera, and Eric R. Gamazon. On Using Local Ancestry to Characterize the Genetic Architecture of Human Traits: Genetic Regulation of Gene Expression in Multiethnic or Admixed Populations. *American Journal of Human Genetics*, 104(6):1097–1115, 6 2019.

[142] Huaying Fang, Qin Hui, Julie Lynch, Jacqueline Honerlaw, Themistocles L. Assimes, Jie Huang, Marijana Vujkovic, Scott M. Damrauer, et al. Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-wide Association Studies. *The American Journal of Human Genetics*, 105(4):763–772, 10 2019.

[143] Lauren S. Mogil, Angela Andaleon, Alexa Badalamenti, Scott P. Dickinson, Xiuqing Guo, Jerome I. Rotter, W. Craig Johnson, Hae Kyung Im, et al. Genetic architecture of gene expression traits across diverse populations. *PLOS Genetics*, 14(8):e1007586, 8 2018.

[144] Soheil Baharian, Maxime Barakatt, Christopher R. Gignoux, Suyash Shringarpure, Jacob Errington, William J. Blot, Carlos D. Bustamante, Eimear E. Kenny, et al. The Great Migration and African-American Genomic Diversity. *PLoS Genetics*, 12(5), 5 2016.

[145] Katarzyna Bryc, Eric Y. Durand, J. Michael Macpherson, David Reich, and Joanna L. Mountain. The genetic ancestry of african americans, latinos, and european Americans across the United States. *American Journal of Human Genetics*, 96(1):37–53, 1 2015.

[146] Jian Yang, S Hong Lee, Michael E Goddard, and Peter M Visscher. GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics*, 88(1):76–82, 1 2011.

[147] Jeffrey B Endelman. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome*, 4(3):250–255, 2011.

[148] Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4):407–409, 4 2014.

[149] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, Pamela Sklar, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet*, 81:559–575, 2007.

[150] Nilanjan Chatterjee, Jianxin Shi, and Montserrat García-Closas. Developing and evaluating polygenic risk prediction models for stratified disease prevention, 7 2016.

[151] John D. Storey, Andrew J. Bass, Alan Dabney, David Robinson, and Gregory Warnes. qvalue: Q-value estimation for false discovery rate control, 2019.

[152] David Tritchler. On inverting permutation tests. *Journal of the American Statistical Association*, 79(385):200–207, 1984.

[153] Kouros Owzar, Zhiguo Li, Nancy Cox, and Sin-Ho Jung. Power and Sample Size Calculations for SNP Association Studies With Censored Time-to-Event Outcomes. *Genetic Epidemiology*, 36(6):538–548, 9 2012.

[154] Peter C Austin and Jason P Fine. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Statistics in medicine*, 36(27):4391–4400, 11 2017.

[155] Maarten van Iterson, Erik W. van Zwet, Bastiaan T. Heijmans, and Bastiaan T. Heijmans. Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology*, 18(1):19, 12 2017.

[156] S. M. Gogarten, T. Bhangale, M. P. Conomos, C. A. C. Laurie, C. P. McHugh, I. Painter, X. Zheng, D. R. Crosslin, et al. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, 28(24):3329–3331, 12 2012.

[157] Yuqian Yulu Liao, Yuqian Yulu Liao, Junyu Li, Junyu Li, Ying Fan, and Binghe Xu. Polymorphisms in AURKA and AURKB are associated with the survival of triple-negative breast cancer patients treated with taxane-based adjuvant chemotherapy. *Cancer management and research*, 10:3801–3808, 2018.

[158] Tatsunori Shimoi, Akinobu Hamada, Marifu Yamagishi, Mitsuharu Hirai, Masayuki Yoshida, Tadaaki Nishikawa, Kazuki Sudo, Akihiko Shimomura, et al. PIK3CAmutation profiling in patients with breast cancer, using a highly sensitive detection system. *Cancer Science*, 109(8):2558–2566, 8 2018.

[159] Magdalena Cizkova, Aurélie Susini, Sophie Vacher, Géraldine Cizeron-Clairac, Catherine Andrieu, Keltouma Driouch, Emmanuelle Fourme, Rosette Lidereau, et al. PIK3CA mutation impact on survival in breast cancer patients and in ER$\alpha$, PR and ERBB2-based subgroups. *Breast cancer research : BCR*, 14(1):R28, 2 2012.

[160] Sajjad Rafiq, Sofia Khan, William Tapper, Andrew Collins, Rosanna Upstill-Goddard, Susan Gerty, Carl Blomqvist, Kristiina Aittomäki, et al. A Genome Wide Meta-Analysis Study for Identification of Common Variation Associated with Breast Cancer Prognosis. *PLoS ONE*, 9(12):e101488, 12 2014.

[161] Sofia Khan, Rainer Fagerholm, Latha Kadalayil, William Tapper, Kristiina Aittomäki, Jianjun Liu, Carl Blomqvist, Diana Eccles, et al. Meta-analysis of three genome-wide association studies identifies two loci that predict survival and treatment outcome in breast cancer. *Oncotarget*, 9(3):4249–4257, 1 2018.

[162] Qi Guo, Marjanka K. Schmidt, Peter Kraft, Sander Canisius, Constance Chen, Sofia Khan, Jonathan Tyrer, Manjeet K. Bolla, et al. Identification of Novel Genetic Markers of Breast Cancer Survival. *JNCI: Journal of the National Cancer Institute*, 107(5), 5 2015.

[163] Kevin L. Keys, Angel C.Y. Y Mak, Marquitta J. White, Walter L. Eckalbar, Andrew W. Dahl, Joel Mefford, Anna V. Mikhaylova, María G. Contreras, et al. On the cross-population portability of gene expression prediction models. *bioRxiv*, pp. 552042, 2019.

[164] Forike K. Martens and A. Cecile J.W. W Janssens. How the Intended Use of Polygenic Risk Scores Guides the Design and Evaluation of Prediction Studies. *Current Epidemiology Reports*, pp. 1–7, 4 2019.

[165] Joshua D. Hoffman, Rebecca E. Graff, Nima C. Emami, Caroline G. Tai, Michael N. Passarelli, Donglei Hu, Scott Huntsman, Dexter Hadley, et al. Cis-eQTL-based trans-ethnic meta-analysis

reveals novel genes associated with breast cancer risk. *PLoS Genetics*, 13(3), 2017.

[166] Nasim Mavaddat, Kyriaki Michailidou, Joe Dennis, Michael Lush, Laura Fachal, Andrew Lee, Jonathan P. Tyrer, Ting-Huei Chen, et al. Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. *The American Journal of Human Genetics*, 104(1):21–34, 1 2019.

[167] Thomas U Ahearn, Haoyu Zhang, Kyriaki Michailidou, Roger L Milne, Manjeet K Bolla, Joe Dennis, Alison M Dunning, Michael Lush, et al. Common breast cancer risk loci predispose to distinct tumor subtypes. Technical report.

[168] Alexander Gusev, Kate Lawrenson, Xianzhi Lin, Paulo C. Lyra, Siddhartha Kar, Kevin C. Vavra, Felipe Segato, Marcos A. S. Fonseca, et al. A transcriptome-wide association study of high-grade serous epithelial ovarian cancer identifies new susceptibility genes and splice variants. *Nature Genetics*, 51(5):815–823, 5 2019.

[169] M Vecchi, S Confalonieri, P Nuciforo, M A Viganò, M Capra, M Bianchi, D Nicosia, F Bianchi, et al. Breast cancer metastases are molecularly distinct from their primary tumors. *Oncogene*, 27(15):2148–2158, 4 2008.

[170] Shun-Fa Yang, Chao-Bin Yeh, Ying-Erh Chou, Hsiang-Lin Lee, and Yu-Fan Liu. Serpin peptidase inhibitor (SERPINB5) haplotypes are associated with susceptibility to hepatocellular carcinoma. *Scientific Reports*, 6(1):26605, 7 2016.

[171] Sarah J. Storr, Nicola Thompson, Xuan Pu, Yimin Zhang, and Stewart G. Martin. No Title. *Pathobiology*, 82(3-4):133–141, 8 2015.

[172] Sarah J. Storr, Siwei Zhang, Tim Perren, Mark Lansdown, Hiba Fatayer, Nisha Sharma, Renu Gahlaut, Abeer Shaaban, et al. The calpain system is associated with survival of breast cancer patients with large but operable inflammatory and non-inflammatory tumours treated with neoadjuvant chemotherapy. *Oncotarget*, 7(30):47927–47937, 7 2016.

[173] Ludovic Leloup and Alan Wells. Calpains as potential anti-cancer targets. *Expert opinion on therapeutic targets*, 15(3):309–323, 3 2011.

[174] C. B. Begg and E. C. Zabor. Detecting and Exploiting Etiologic Heterogeneity in Epidemiologic Studies. *American Journal of Epidemiology*, 176(6):512–518, 9 2012.

[175] María Elena Martínez, Giovanna I Cruz, Abenaa M Brewster, Melissa L Bondy, and Patricia A Thompson. What can we learn about disease etiology from case-case analyses? Lessons from breast cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 19(11):2710–2714, 11 2010.

[176] Lavinia Paternoster, Kate Tilling, and George Davey Smith. Genetic epidemiology and Mendelian randomization for informing disease therapeutics: Conceptual and methodological challenges. *PLOS Genetics*, 13(10):e1006944, 10 2017.

[177] Ruth E. Mitchell, Lavinia Paternoster, and George Davey Smith. Mendelian Randomization in Case Only Studies: A Promising Approach to be Applied With Caution. *The American Journal of Cardiology*, 122(12):2169–2171, 12 2018.

[178] Frank Dudbridge, Richard J. Allen, Nuala A. Sheehan, A. Floriaan Schmidt, James C. Lee, R. Gisli Jenkins, Louise V. Wain, Aroon D. Hingorani, et al. Adjustment for index event bias in genome-wide association studies of subsequent events. *Nature Communications*, 10(1):1561,

12 2019.

[179] Kathleen Conway, Eloise Parrish, Sharon N Edmiston, Dawn Tolbert, Chiu-Kit Tse, Patricia Moorman, Beth Newman, and Robert C Millikan. Risk factors for breast cancer characterized by the estrogen receptor alpha A908G (K303R) mutation. *Breast cancer research : BCR*, 9(3):R36, 2007.

[180] Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G. Van Meir, Daniel J. Brat, Gena M. Mastrogianakis, Jeffrey J. Olson, Tom Mikkelsen, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, 10 2008.

[181] Philip L. De Jager, Yiyi Ma, Cristin McCabe, Jishu Xu, Badri N. Vardarajan, Daniel Felsky, Hans Ulrich Klein, Charles C. White, et al. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Scientific Data*, 5, 8 2018.

[182] Michael E. Sobel. Asymptotic Confidence Intervals for Indirect Effects in Structural Equation Models. *Sociological Methodology*, 13:290, 1982.

[183] Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete Multivariate Analysis: Theory and Practice*. Springer, New York, 1975.

[184] Michael E. Sobel. Direct and Indirect Effects in Linear Structural Equation Models. *Sociological Methods & Research*, 16(1):155–176, 8 1987.

[185] Evan Koch, Mickey Ristroph, and Mark Kirkpatrick. Long Range Linkage Disequilibrium across the Human Genome. *PLoS ONE*, 8(12):e80754, 12 2013.

[186] Leeyoung Park. Population-specific long-range linkage disequilibrium in the human genome and its influence on identifying common disease variants. *Scientific Reports*, 9(1):1–13, 12 2019.

[187] Janis E Wigginton, David J Cutler, and Gonçalo R Abecasis. A Note on Exact Tests of Hardy-Weinberg Equilibrium. Technical report, 2005.

[188] David A. Bennett, Julie A. Schneider, Aron S. Buchman, Lisa L. Barnes, Patricia A. Boyle, and Robert S. Wilson. Overview and Findings from the Rush Memory and Aging Project. *Current Alzheimer Research*, 9(6):646–663, 7 2013.

[189] Philip L. De Jager, Joshua M. Shulman, Lori B. Chibnik, Brendan T. Keenan, Towfique Raj, Robert S. Wilson, Lei Yu, Sue E. Leurgans, et al. A genome-wide scan for common variants affecting the rate of age-related cognitive decline. *Neurobiology of Aging*, 33(5):1–1017, 5 2012.

[190] Naomi R. Wray, Stephan Ripke, Manuel Mattheisen, Maclej Trzaskowski, Enda M. Byrne, Abdel Abdellaoui, Mark J. Adams, Esben Agerbo, et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, 50(5):668–681, 5 2018.

[191] Jimmy Z. Liu, Yaniv Erlich, and Joseph K. Pickrell. Case-control association mapping by proxy using family history of disease. *Nature Genetics*, 49(3):325–331, 3 2017.

[192] Xingyi Guo, Weiqiang Lin, Jiandong Bao, Qiuyin Cai, Xiao Pan, Mengqiu Bai, Yuan Yuan, Jiajun Shi, et al. A Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci Identified in Genome-wide Association Studies. *American Journal of Human Genetics*, 102(5):890–903, 5 2018.

[193] Alejandro Quiroz-Zárate, Benjamin J. Harshfield, Rong Hu, Nick Knoblauch, Andrew H. Beck, Susan E. Hankinson, Vincent Carey, Rulla M. Tamimi, et al. Expression Quantitative Trait loci (QTL) in tumor adjacent normal breast tissue and breast tumor tissue. *PLOS ONE*, 12(2):e0170181, 2 2017.

[194] Max J. Dörfel and Gholson J. Lyon. The biological functions of Naa10 - From amino-terminal acetylation to human disease, 8 2015.

[195] Irina Lambertz, Candy Kumps, Shana Claeys, Sven Lindner, Anneleen Beckers, Els Janssens, Daniel R Carter, Alex Cazes, et al. Biology of Human Tumors Upregulation of MAPK Negative Feedback Regulators and RET in Mutant ALK Neuroblastoma: Implications for Targeted Treatment. *Clinical Cancer Research*, 2015.

[196] Bradleigh Whitton, Haruko Okamoto, Graham Packham, and Simon J. Crabb. Vacuolar ATPase as a potential therapeutic target and mediator of treatment resistance in cancer, 8 2018.

[197] Masahiro Matsubara and Mina J. Bissell. Inhibitors of Rho kinase (ROCK) signaling revert the malignant phenotype of breast cancer cells in 3D context. *Oncotarget*, 7(22):31602–31622, 5 2016.

[198] Fei Chang, Yunpeng Zhang, Jun Mi, Qian Zhou, Fuxiang Bai, Xin Xu, David E. Fisher, Qinfeng Sun, et al. ROCK inhibitor enhances the growth and migration of BRAF-mutant skin melanoma cells. *Cancer Science*, 109(11):3428–3437, 11 2018.

[199] Ying Ni, Spencer Seballos, Benjamin Fletcher, Todd Romigh, Lamis Yehia, Jessica Mester, Leigha Senter, Farshad Niazi, et al. Germline compound heterozygous poly-glutamine deletion in USF3 may be involved in predisposition to heritable and sporadic epithelial thyroid carcinoma. *Human Molecular Genetics*, 26(2):243–257, 2017.

[200] Alexander Gusev, Nicholas Mancuso, Hyejung Won, Maria Kousi, Hilary K. Finucane, Yakir Reshef, Lingyun Song, Alexias Safi, et al. Transcriptome-wide association study of schizophrenia and chromatin activity yields mechanistic disease insights. *Nature Genetics*, 50(4):538–548, 4 2018.

[201] Towfique Raj, Yang I. Li, Garrett Wong, Jack Humphrey, Minghui Wang, Satesh Ramdhani, Ying Chih Wang, Bernard Ng, et al. Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nature Genetics*, 50(11):1584–1592, 11 2018.

[202] Nancy Y. A. Sey, Benxia Hu, Won Mah, Harper Fauni, Jessica Caitlin McAfee, Prashanth Rajarajan, Kristen J. Brennand, Schahram Akbarian, et al. A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles. *Nature Neuroscience*, pp. 1–11, 3 2020.

[203] Christiane Reitz. Genetic loci associated with Alzheimer's disease. *Future Neurology*, 9(2):119–122, 3 2014.

[204] Rebecca Sims, Sven J. Van Der Lee, Adam C. Naj, Céline Bellenguez, Nandini Badarinarayan, Johanna Jakobsdottir, Brian W. Kunkle, Anne Boland, et al. Rare coding variants in PLCG2, ABI3, and TREM2 implicate microglial-mediated innate immunity in Alzheimer's disease. *Nature Genetics*, 49(9):1373–1384, 9 2017.

[205] Xiang Zhen Yuan, Sen Sun, Chen Chen Tan, Jin Tai Yu, and Lan Tan. The Role of ADAM10 in Alzheimer's Disease, 2017.

[206] Bernd Bischl, Michel Lang, Olaf Mersmann, Jorg Rahnenfuhrer, and Claus Weihs. BatchJobs and BatchExperiments: Abstraction Mechanism for Using R in Batch Environments. *The Journal of Statistical Software*, 64(11):1–25, 2015.

[207] Feifei Bao, Peter R. LoVerso, Jeffrey N. Fisk, Victor B. Zhurkin, and Feng Cui. p53 binding sites in normal and cancer cells are characterized by distinct chromatin context. *Cell Cycle*, 16(21):2073–2085, 11 2017.

[208] Joseph X. Zhou, Zerrin Isik, Caide Xiao, Irit Rubin, Stuart A. Kauffman, Michael Schroeder, and Sui Huang. Systematic drug perturbations on cancer cells reveal diverse exit paths from proliferative state. *Oncotarget*, 7(7):7415–7425, 2016.

[209] Dena A. J. Ahmad, Ola H. Negm, M. Layth Alabdullah, Sameer Mirza, Mohamed R. Hamed, Vimla Band, Andrew R. Green, Ian O. Ellis, et al. Clinicopathological and prognostic significance of mitogen-activated protein kinases (MAPK) in breast cancers. *Breast Cancer Research and Treatment*, 159(3):457–467, 10 2016.

[210] M. M. Ryan, H. E. Lockstone, S. J. Huffaker, M. T. Wayland, M. J. Webster, and S. Bahn. Gene expression analysis of bipolar disorder reveals downregulation of the ubiquitin cycle and alterations in synaptic genes. *Molecular Psychiatry*, 11(10):965–978, 10 2006.

[211] Kate Baker, Sarah L. Gordon, Detelina Grozeva, Margriet Van Kogelenberg, Nicola Y. Roberts, Michael Pike, Edward Blair, Matthew E. Hurles, et al. Identification of a human synaptotagmin-1 mutation that perturbs synaptic vesicle cycling. *Journal of Clinical Investigation*, 125(4):1670–1678, 4 2015.

[212] Samuel Heyes, Wendy S. Pratt, Elliott Rees, Shehrazade Dahimene, Laurent Ferron, Michael J. Owen, and Annette C. Dolphin. Genetic disruption of voltage-gated calcium channels in psychiatric and neurological disorders, 11 2015.

[213] Yiannis A. Savva, Leila E. Rieder, and Robert A. Reenan. The ADAR protein family. *Genome Biology*, 13(12):259, 12 2012.

[214] William Slotkin and Kazuko Nishikura. Adenosine-to-inosine RNA editing and human disease, 11 2013.

[215] Kristopher J Preacher and James P Selig. Advantages of Monte Carlo Confidence Intervals for Indirect Effects. *Communication Methods and Measures*, 6:77–98, 2012.

[216] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. LEAST ANGLE REGRESSION. Technical Report 2, 2004.

[217] Mehmet Suzen. Compressive Sampling: Sparse Signal Recovery Utilities [R package R1magic version 0.3.2], 4 2015.

[218] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2 2006.

[219] Emmanuel J. Candès and Justin Romberg. Quantitative robust uncertainty principles and optimally sparse decompositions. *Foundations of Computational Mathematics*, 6(2):227–254, 6 2006.

[220] David L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 4 2006.

[221] Henrik Bengtsson. R package: future: Unified Parallel and Distributed Processing in R for Everyone, 2020.

[222] John Lonsdale, Jeffrey Thomas, Mike Salvatore, Rebecca Phillips, Edmund Lo, Saboor Shad, Richard Hasz, Gary Walters, et al. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6):580–585, 6 2013.

[223] Kristin G. Ardlie, David S. DeLuca, Ayellet V. Segrè, Timothy J. Sullivan, Taylor R. Young, Ellen T. Gelfand, Casandra A. Trowbridge, Julian B. Maller, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 5 2015.

[224] Arjun Bhattacharya, Alina M Hamilton, Helena Furberg, Eugene Pietzak, Mark P Purdue, Melissa A Troester, Katherine A Hoadley, and Michael I Love. An approach for normalization and quality control for NanoString RNA expression data. *bioRxiv*, pp. 2020.04.08.032490, 4 2020.

[225] Elham Azizi, Ambrose J Carr, George Plitas, Linas Mazutis, Alexander Y Rudensky, Dana Pe'er, Andrew E Cornish, Catherine Konopacki, et al. Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment Resource Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*, 174:1293–1308, 2018.

[226] Sunny Z Wu, Daniel L Roden, Chenfei Wang, Holly Holliday, Kate Harvey, Aurélie S Cazet, Kendelle J Murphy, Brooke Pereira, et al. Single-cell analysis reveals diverse stromal subsets associated with immune evasion 1 in triple-negative breast cancer 2 3 Authors 4. *bioRxiv*, 18(4).

[227] Quy H Nguyen, Nicholas Pervolarakis, Kerrigan Blake, Dennis Ma, Ryan Tevia Davis, Nathan James, Anh T Phung, Elizabeth Willey, et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nature Communications*, 8(2028):1–12, 2018.

[228] Dominic G. Rothwell, Yaoyong Li, Mahmood Ayub, Catriona Tate, Gillian Newton, Yvonne Hey, Louise Carter, Suzanne Faulkner, et al. Evaluation and validation of a robust single cell RNA-amplification protocol through transcriptional profiling of enriched lung cancer initiating cells. *BMC Genomics*, 15(1), 12 2014.

[229] Bryan A. Smith, Nikolas G. Balanis, Avinash Nanjundiah, Katherine M. Sheu, Brandon L. Tsai, Qingfu Zhang, Jung Wook Park, Michael Thompson, et al. A Human Adult Stem Cell Signature Marks Aggressive Variants across Epithelial Cancers. *Cell Reports*, 24(12):3353–3366, 9 2018.

[230] Aleix Prat, Tom As Pascual, Carmine De Angelis, Carolina Gutierrez, Antonio Llombart-Cussac, Tao Wang, Javier Cort, Brent Rexer, et al. HER2-Enriched Subtype and ERBB2 Expression in HER2-Positive Breast Cancer Treated with Dual HER2 Blockade.

[231] Yuxing Liao, Jing Wang, Eric J Jaehnig, Zhiao Shi, and Bing Zhang. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Research*, 47:199–205, 2019.

[232] Guangchuang Yu, Li Gen Wang, Yanyan Han, and Qing Yu He. ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS A Journal of Integrative Biology*, 16(5):284–287, 5 2012.

[233] Melissa A Troester, Xuezheng Sun, Emma H Allott, Joseph Geradts, Stephanie M Cohen, Chiu-Kit Tse, Erin L Kirk, Leigh B Thorne, et al. No Title. 110(2), 2 2018.

[234] Yoav Benjamini, Daniel Yekutieli, Don Edwards, Juliet Popper Shaffer, Ajit C. Tamhane, Peter H. Westfall, Burt Holland, Yoav Benjamini, et al. False discovery rate-adjusted multiple confidence intervals for selected parameters, 2005.

[235] Raúl Aguirre-Gamboa, Niek de Klein, Jennifer di Tommaso, Annique Claringbould, Monique GP van der Wijst, Dylan de Vries, Harm Brugge, Roy Oelen, et al. Deconvolution of bulk blood eQTL effects into immune cell subpopulations. *BMC Bioinformatics*, 21(1):243, 12 2020.

[236] Di He Gong, Lei Fan, Hai Yan Chen, Ke Feng Ding, and Ke Da Yu. Intratumoral expression of CCR3 in breast cancer is associated with improved relapse-free survival in luminal-like disease. *Oncotarget*, 7(19):28570–28578, 5 2016.

[237] Jeronay K. Thomas, Hina Mir, Neeraj Kapur, Sejong Bae, and Shailesh Singh. CC chemokines are differentially expressed in Breast Cancer and are associated with disparity in overall survival. *Scientific Reports*, 9(1):1–12, 12 2019.

[238] Lovisa E. Reinius, Nathalie Acevedo, Maaike Joerink, Göran Pershagen, Sven-Erik Dahlén, Dario Greco, Cilla Söderhäll, Annika Scheynius, et al. Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS ONE*, 7(7):e41361, 7 2012.

[239] Carolina M. Montaño, Rafael A. Irizarry, Walter E. Kaufmann, Konrad Talbot, Raquel E. Gur, Andrew P. Feinberg, and Margaret A. Taub. Measuring cell-type specific differential methylation in human brain tissue. *Genome Biology*, 14(8):R94, 8 2013.

[240] Y. P. Chen, Y. Q. Wang, J. W. Lv, Y. Q. Li, M. L.K. K Chua, Q. T. Le, N. Lee, A. Dimitrios Colevas, et al. Identification and validation of novel microenvironment-based immune molecular subgroups of head and neck squamous cell carcinoma: Implications for immunotherapy. *Annals of Oncology*, 30(1):68–75, 1 2019.

[241] Monica D'Arcy, Jodie Fleming, Whitney R. Robinson, Erin L. Kirk, Charles M. Perou, Melissa A. Troester, Monica D'Arcy, Jodie Fleming, et al. Race-associated biological differences among Luminal A breast tumors. *Breast Cancer Research and Treatment*, 152(2):437–448, 7 2015.

[242] Fan Wang, Zachariah Dohogne, Jin Yang, Yu Liu, and Benjamin Soibam. Predictors of breast cancer cell types and their prognostic power in breast cancer patients. *BMC Genomics*, 19(1):137, 12 2018.

[243] Melissa A. Troester, Katherine A. Hoadley, Therese Sørlie, Brittney Shea Herbert, Anne Lise Børresen-Dale, Per Eystein Lønning, Jerry W. Shay, William K. Kaufmann, et al. Cell-type-specific responses to chemotherapeutics in breast cancer. *Cancer Research*, 64(12):4218–4226, 6 2004.

[244] Martin H. Schaefer and Luis Serrano. Cell type-specific properties and environment shape tissue specificity of cancer genes. *Scientific Reports*, 6(1):1–14, 2 2016.

[245] Fan Yang, Kevin J. Gleason, Jiebiao Wang, The GTEx consortium, Jubao Duan, Xin He, Brandon L Pierce, and Lin S Chen. CCmed: cross-condition mediation analysis for identifying robust trans-eQTLs and assessing their effects on human traits. *bioRxiv*, pp. 803106, 2019.

[246] Karin Jöhrer, Claudia Zelle-Rieser, Alexander Perathoner, Patrizia Moser, Martina Hager, Reinhold Ramoner, Hubert Gander, Lorenz Höltl, et al. Up-regulation of functional chemokine receptor CCR3 in human renal cell carcinoma. *Clinical Cancer Research*, 11(7):2459–2465, 4 2005.

[247] Tomomitsu Miyagaki, Makoto Sugaya, Takashi Murakami, Yoshihide Asano, Yayoi Tada, Takafumi Kadono, Hitoshi Okochi, Kunihiko Tamaki, et al. CCL11-CCR3 interactions promote survival of anaplastic large cell lymphoma cells via ERK1/2 activation. *Cancer Research*, 71(6):2056–2065, 3 2011.

[248] Shannon A. Bryan, Peter J. Jose, Joanna R. Topping, Robert Wilhelm, Carol Soderberg, Denis Kertesz, Peter J. Barnes, Timothy J. Williams, et al. Responses of leukocytes to chemokines in whole blood and their antagonism by novel CC-Chemokine Receptor 3 antagonists. *American Journal of Respiratory and Critical Care Medicine*, 165(12):1602–1609, 6 2002.

[249] Michael K Samoszuk, Vince Nguyen, Iris Gluzman, and Justin H Pham. Occult Deposition of Eosinophil Peroxidase in a Subset of Human Breast Carcinomas. Technical Report 3, 1996.

[250] Kaur Alasoo, Julia Rodrigues, Subhankar Mukhopadhyay, Andrew J. Knights, Alice L. Mann, Kousik Kundu, Christine Hale, Gordon Dougan, et al. Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nature Genetics*, 50(3):424–431, 3 2018.

[251] Renaud Gaujoux and Cathal Seoighe. Gene expression CellMix: a comprehensive toolbox for gene expression deconvolution. 29(17):2211–2212, 2013.

[252] Ziyi Li, Zhijin Wu, Peng Jin, and Hao Wu. Dissecting differential signals in high-throughput data from complex tissues.