DeCompress: tissue compartment deconvolution of targeted mRNA expression panels using compressed sensing

Arjun Bhattacharya[1], Alina M. Hamilton[2], Melissa A. Troester[2,3], and Michael I. Love[1,4*]

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 27516

[2]Department of Pathology and Laboratory Medicine, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA, 27516

[3]Department of Epidemiology, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA, 27516

[4]Department of Genetics, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA, 27516

*To whom correspondence should be addressed. Email: milove@email.unc.edu.

Present Address: Michael I. Love, Department of Biostatistics, Department of Genetics, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA, 27516

ABSTRACT

Targeted mRNA expression panels, measuring up to 800 genes, are used in academic and clinical settings due to low cost and high sensitivity for archived samples. Most samples assayed on targeted panels originate from bulk tissue comprised of many cell types, and cell-type heterogeneity confounds biological signals. Reference-free methods are used when cell-type-specific expression references are unavailable, but limited feature spaces render implementation challenging in targeted panels. Here, we present *DeCompress*, a semi-reference-free deconvolution method for targeted panels. *DeCompress* leverages a reference RNA-seq or microarray dataset from similar tissue to expand the feature space of targeted panels using compressed sensing. Ensemble reference-free deconvolution is performed on this artificially expanded dataset to estimate cell-type proportions and gene signatures. In simulated mixtures, four public cell line mixtures, and a targeted panel (1199 samples; 406 genes) from the Carolina Breast Cancer Study, *DeCompress* recapitulates cell-type proportions with less error than reference-free methods and finds biologically relevant compartments. We integrate compartment estimates into *cis*-eQTL mapping in breast cancer, identifying a tumor-specific *cis*-eQTL for *CCR3* (C-C Motif Chemokine

Receptor 3) at a risk locus. *DeCompress* improves upon reference-free methods without requiring expression profiles from pure cell populations, with applications in genomic analyses and clinical settings.

INTRODUCTION

Academic and clinical settings have prioritized the collection of tissue samples of mixed cell types for molecular profiling and biomarker studies (1–3). Bulk tissue, especially from cancerous tumors, is comprised of different cell types, many rare, and each contributing varied biological signal to an assay (e.g. mRNA expression) (4, 5). This cell-type heterogeneity makes it difficult to distinguish variability that reflects shifts in cell populations from variability that reflects changes in cell-type-specific expression (6). Since RNA-seq technology was developed, cell-type deconvolution from mRNA expression has become important in genetic and genomic association studies: either using compositions in regression models as covariates to adjust for the association between cell type and phenotype (7–10), or using them as inputs to solve for cell-type specific quantities (11, 12). Cell-type deconvolution methods can be reference-based (supervised) (13–19) or reference-free (unsupervised) (20–26), depending on whether cell-type-specific expression profiles are available for the component cell-types. When reference panels are unavailable, as in understudied tissues or populations (27), reference-free deconvolution is the only viable option. Even in cases where reference expression profiles are available, reference-based methods may provide inaccurate proportion estimates if the mixed tissue and references represent different clinical settings or phenotypes (28).

Given the advent of single-cell technologies and studies into cell trajectories, the concept of cell types in bulk tissue has been debated (29). Especially in perturbed or diseased tissues, like cancer, individual cells may present in different states, or various cells of possibly different identities may contribute, in aggregate, to the same biological process and have similar molecular profiles (30–32). While previous reference-free methods rely on searching the feature space for compartment-specific molecular features from the entire transcriptome and thus require a large feature space (22, 24–26), reference-free deconvolution methods can, with fewer assumptions, identify tissue compartments, or isolated units of a tissue that represent either a biological process or a cell type (33). Thus, reference-free methods have important advantages over reference-based methods but may require a large number of features for optimal performance (25, 34).

Many important datasets may have fewer expression targets than those required for existing reference-free deconvolution methods. Targeted mRNA expression assays are optimized for gene expression quantification in samples stored clinically and use a panel of up to 800 genes without requiring cDNA synthesis or amplification steps (35–37). These technologies offer key advantages in sensitivity, technical reproducibility, and strong robustness for profiling formalin-fixed, paraffin-embedded (FFPE) samples (35, 38). Given these advantages, targeted expression profiling is increasingly being used for molecular studies (36, 37, 39–42), especially prospective studies involving FFPE samples stored over several years (43) and diagnostic assays in clinical settings (3, 44). Due to its viability in diagnostics, it is important to identify reference-free deconvolution methods that overcome the need for searching for compartment-specific genes from the assay's feature space (22, 24–26), given the limited feature space in targeted panels.

Previous groups have proposed methods for efficiently reconstructing full gene expression profiles from sparse measurements of the transcriptome, borrowing techniques from image reconstruction using compressed sensing (45, 46) and machine learning (47–50). For example, Cleary *et al* developed a blind compressed sensing method that recovers gene expression from multiple composite measurements of the transcriptome (up to 100 times fewer measurements than genes) by using modules of interrelated genes in an unsupervised manner. Another imputation method by Viñas *et al* (51) used recent machine learning methodology (52) to provide efficient and accurate transcriptomic reconstruction in healthy, unperturbed tissue from the Genotype-Tissue Expression (GTEx) Project (53, 54). The performance of these methods provides a promising avenue to expand the feature space of targeted panels, rendering them more applicable for reference-free deconvolution methods.

Here, we present *DeCompress*, a semi-reference-free deconvolution method for targeted panels. *DeCompress* requires a reference RNA-seq or microarray dataset from the same bulk tissue assayed by the targeted expression panel to train a compressed sensing model to expand the feature space in a targeted panel. We show the advantages of using *DeCompress* over other reference-free methods with simulation analyses and real data applications. Lastly, we examine the impact of tissue compartment deconvolution on downstream analyses, such as *cis*-eQTL analysis using expression data from the Carolina Breast Cancer Study (CBCS) (55). *DeCompress* is available freely as an R package on GitHub at https://github.com/bhattacharya-a-bt/DeCompress.

MATERIAL AND METHODS

The Decompress algorithm

DeCompress takes in two expression matrices from similar bulk tissue as inputs: the *target* expression matrix from a targeted panel of gene expression with n samples and k genes, and a *reference* expression matrix from an RNA-seq and microarray panel with N samples and $K > k$ genes. Ideally, both the target and reference expression matrices should be on the raw expression scale (not log-transformed), as we presume the total RNA abundance for a given gene in bulk tissue is a linear combination of that gene's compartment-specific RNA abundance. We refer to DeCompress as a semi-reference-free method, as it requires a reference expression matrix but not compartment-specific expression profiles (as in reference-based methods). For a user-defined number of compartments, *DeCompress* outputs compartment proportions for all samples in the target and the compartment-specific expression profiles for the genes used in deconvolution. The method follows three general steps, as detailed in **Figure 1**: (1) selection of the compartment-specific genes from the reference, (2) compressed sensing to expand the targeted panel to a *DeCompressed* expression matrix with these compartment-specific genes, and (3) ensemble deconvolution on the DeCompressed dataset. Full mathematical and algorithmic details for *DeCompress* are provided in **Supplemental Methods**. *DeCompress* is available as an R package on GitHub (https://github.com/bhattacharya-a-bt/DeCompress).

The first step of *DeCompress* is to use the reference dataset to find a set of $K' < K$ genes that are representative of different compartments that comprise the bulk tissue. These K' genes, called the compartment-specific genes, can be supplied by the user if prior gene signatures can be applied. If any such gene signatures are not available, *DeCompress* borrows from previous reference-free methods to determine this set of genes (*Linseed* (22) or *TOAST* (25)). If the user cannot determine the total number of compartments, using the reference, the number of compartments can be estimated by assessing the cumulative total variance explained by successive singular value decomposition modes.

After a set of compartment-specific genes are determined, *DeCompress* uses the reference to infer a model that predicts the expression of each of these compartment-specific genes from the genes in the target. Predictive modeling procedures borrow ideas from compressed sensing (45, 46, 56), a technique that was developed to reconstruct a full image from sparse measurements of it: the estimation procedure can be broken down into solving a system of equations using either linear or non-linear regularized optimization, with options for parallelization when the sample size of the reference dataset is large. These optimization methods are detailed in **Supplemental Methods**. The predictive models are curated into a *compression* matrix, which is then used to expand the original target (with $k < K' < K$ genes) into the artificially *DeCompressed* expression matrix (with the K' compartment-specific genes). In practice, we observed that regularized linear regression (lasso, ridge, or elastic net regression (48)) provides the best prediction of gene expression (**Supplemental Figure S1**), and the user may either model the gene expression using the traditional Gaussian family or assume that the errors follow a Poisson distribution to account for the scale of the original data (not log-transformed).

Lastly, ensemble deconvolution is performed on the DeCompressed expression matrix to estimate (1) compartment proportions on the samples in the target, and (2) the compartment-specific expression profiles for the K' genes used in deconvolution. Several options for reference-free deconvolution are provided in *DeCompress*. We also provide options that uses a reference-based method, *unmix* from the DESeq2 package (57), based on compartment expression profiles estimated from the reference RNA-seq or microarray dataset (i.e. an approximate compartment expression profile is estimated from a non-negative matrix factorization of the reference dataset). Estimates from the method that best recovers the DeCompressed expression matrix is chosen. **Supplemental Table S1** provides summaries of the methods employed in *DeCompress*.

Benchmarking analysis

Using simulations and published datasets, we benchmarked *DeCompress* against five other reference-free methods: *deconf* (20), *CellDistinguisher* (26), *Linseed* (22), *DeconICA* (24), and iterative non-negative matrix factorization with feature selection using TOAST (25) (see **Supplemental Table S1**). All these datasets provide a matrix of known compartment proportions. To measure the performance of each method, we calculate the error between the estimated and true compartment proportions as the mean square error (MSE) (i.e. the mean row-wise MSE between the two matrices). We also permute the columns the estimated matrix (corresponding to compartments) to align compartments accordingly between the known and estimated proportions to minimize the MSE for each method.

In-silico mixing with GTEx

We performed *in-silico* mixing experiments using expression data from the Genotype-Tissue Expression (GTEx) Project (dbGAP accession number phs000424.v7.p2) (53, 54). Here, we obtained median transcripts per kilobase million (TPM) data for four tissue types: mammary tissue, EBV-transformed lymphocytes, transformed fibroblasts, and subcutaneous adipose. We randomly generated compartment proportions for each of these tissue types and simulated mixed RNA-seq expression data for 200 samples. We then scaled these mixed expression profiles with multiplicative noise randomly generated from a Normal distribution with 0 mean and standard deviations of 4 and 8. We then generated 25 pseudo-targeted expression panels by randomly selecting 200, 500, and 800 of the genes with mean and standard deviations above the median mean and standard deviations of all genes. For benchmarking, we randomly select 100 samples for the target matrix. For *DeCompress*, the simulated RNA-seq data on the other 100 samples are used as the reference matrix. We added more normally-distributed multiplicative noise with zero mean and unit variance to simulate a batch difference between the reference and target matrix. For comparison to compartments with dissimilar expression profiles, we repeated these simulations for four other tissues: mammary tissue, pancreas, pituitary, and whole blood. Full details for this simulation framework are provided in **Supplemental Methods**.

Existing mixing experiments

We also benchmarked *DeCompress* in four published mixing experiments: (1) microarray expression for mixed rat brain, liver, and lung biospecimens (GEO Accession Number: GSE19830), commonly used as a benchmarking dataset in deconvolution studies ($N = 42$) (11), (2) RNA-seq expression (GSE123604) for a mixture of breast cancer cells, fibroblasts, normal mammary cells, and Burkitt's lymphoma cells ($N = 40$) (23), (3) microarray expression (GSE97284) for laser capture micro-dissected prostate tumors ($N =$

30) (58), and (4) RNA-seq expression (GSE64098) for a mixture of two lung adenocarcinoma cell lines ($N = 40$) (59, 60). As in the in-silico mixing using GTEx data, we generated pseudo-targeted panels by randomly selecting 200, 500, and 800 of the genes with mean and standard deviations above the median mean and standard deviations of all genes. For the rat mixture dataset, we used 30 of the 42 samples as a reference microarray matrix (with multiplicative noise, as in GTEx) 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: 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. These datasets are summarized in **Supplemental Table S2**.

Applications in Carolina Breast Cancer Study (CBCS) data

We lastly used expression data from the Carolina Breast Cancer Study for validation and analysis (55). 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 (43). As described before (10, 61), the expression data (406 genes and 11 housekeeping genes) was pre-processed and normalized using quality control steps from the *NanoStringQCPro* package, upper quartile normalization using *DESeq2* (57, 62), and estimation and removal of unwanted technical variation using the *RUVSeq* and *limma* packages (63, 64). The resulting normalized dataset comprised of samples from 1,199 patients, comprising of 628 women of African descent (AA) and 571 women of European descent (EA). 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 (10). These compartment proportions of the 148 samples were used for benchmarking of *DeCompress* against other reference-free methods.

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.

Over-representation and gene set enrichment analysis

We conducted over-representation (ORA) and gene set enrichment analysis (GSEA) to identify significantly enriched gene ontologies using *WebGestaltR* (65). Specifically, we considered biological process ontologies categorized by The Gene Ontology Consortium (66, 67) at FDR-adjusted $P < 0.05$.

Survival analysis

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 built cause-specific Cox models (68) by modeling the hazard function of breast cancer-specific mortality with the following covariates: race, PAM50 molecular subtype (69), age, compartment-specific proportions, and an interaction term between molecular subtype and compartment proportion. We compared these compartment-specific survival models with the nested baseline model that did not include compartment proportions using partial likelihood ratio tests. We tested for the statistical significance of parameter estimates using Wald-type tests, adjusting for multiple testing burden using the Benjamini-Hochberg procedure at a 10% false discovery rate (70).

eQTL analysis

CBCS genotype data is measured on the OncoArray. Approximately 50% of the SNPs for the OncoArray were selected as a “GWAS backbone” (Illumina HumanCore), which aimed to provide high coverage for many common variants through imputation. The remaining SNPs were selected from lists supplied by six disease-based consortia, together with a seventh list of SNPs of interest to multiple disease-focused groups. Approximately 72,000 SNPs were selected specifically for their relevance to breast cancer. The sources for the SNPs included in this backbone, as well as backbone manufacturing, calling, and quality control, are discussed in depth by the OncoArray Consortium (71, 72). All samples were imputed using the October 2014 (v.3) release of the 1000 Genomes Project (73) as a reference panel in the standard two-stage imputation approach, using *SHAPEIT2* for phasing and *IMPUTEv2* for imputation (74–76). All genotyping, genotype calling, quality control, and imputation was done at the DCEG Cancer Genomics Research Laboratory (71, 72).

From the provided genotype data, we excluded variants (1) with a minor frequency less than 1% based on genotype dosage and (2) that deviated significantly from Hardy-Weinberg equilibrium at $P < 10^{-8}$ using the appropriate functions in *PLINK v1.90b3* (77). Finally, we intersected genotyping panels for the AA and EA samples, resulting in 5,989,134 autosomal variants. We excluded 334,391 variants on the X chromosome. CBCS genotype data was coded as dosages, with reference and alternative allele coding as in the National Center for Biotechnology Information's Single Nucleotide Polymorphism Database (dbSNP) (78).

As previously described (10), using the 1,199 samples (621 AA, 578 EA) with expression data, we assessed the additive relationship between the gene expression values and genotypes with linear regression analysis using *MatrixeQTL* (79). We consider a baseline linear model with log-transformed gene expression of a gene of interest as the dependent variable, SNP dosage as the primary predictor of interest, and the following covariates: age, BMI, post-menopausal status, and the first 5 principal components of the joint AA and EA genotype matrix. We also considered a compartment-specific interaction model that adds compartment proportion from *DeCompress* and an interaction term between the SNP dosage and compartment proportion (8, 9). This interaction model subtly changes the interpretation of the main SNP dosage effect, representing an estimate of the eQTL effect size at 0% compartment-specific cells. Thus, we recover compartment-specific eQTLs by testing the interaction effect, which measures how the magnitude of an eQTL differs between the two cell types. The interaction model was fit using *MatrixeQTL*'s linear-cross implementation. It is important to note that we model the log-transformed expression here, as existing methods for modeling expression on genotype do not support interaction terms (80–82).

We compared eQTLs mapped in CBCS here with eQTLs in GTEx. We downloaded healthy tissue eQTLs from the Genotype-Tissue Expression (GTEx) Project and cross-referenced eGenes and corresponding eSNPs between CBCS and GTEx in healthy breast mammary tissue, EBV-transformed lymphocytes, transformed fibroblasts, and subcutaneous adipose tissue. We considered these tissues mainly due to their high relative composition in bulk breast tumor samples, as shown previously in many studies (23, 83–85). The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 05/14/20. We also downloaded iCOGs GWAS summary statistics for breast cancer risk (86–88) to assess any overlap between CBCS eQTLs and GWAS-detected risk variants.

RESULTS

Overview of the DeCompress algorithm

DeCompress takes in two expression matrices from similar bulk tissue as inputs: an expression matrix from a targeted panel of gene expression with n samples and k genes, and an expression matrix from an RNA-seq and microarray panel with N samples and $K > k$ genes. For shorthand, we will refer to RNA-seq or microarray panel as the *reference* and the targeted expression panel as the *target*. *DeCompress* outputs tissue compartment proportions for a user-defined number of all samples in the target and the compartment-specific expression profiles for the genes used in deconvolution. The method follows three general steps, as detailed in **Figure 1**: (1) feature selection of the compartment-specific genes from the reference, (2) compressed sensing to expand the targeted panel to a *DeCompressed* expression matrix with these compartment-specific genes, and (3) ensemble deconvolution on the DeCompressed dataset using existing reference-free methods. We provide further details about *DeCompress* in **Methods** and full mathematical and algorithmic details in **Supplemental Methods**.

Benchmarking DeCompress against other reference-free deconvolution methods

We benchmarked *DeCompress* performance across 6 datasets (see **Supplemental Table S2**): (1) *in-silico* mixing experiments using tissue-specific expression profiles from the Genotype-Tissue Expression (GTEx) Project (53, 54), (2) expression from 4 published datasets with known compartment proportions (11, 23, 58, 59), and (3) and tumor expression from the Carolina Breast Cancer Study (43, 55). We compared the performance of *DeCompress* against 5 other reference-free deconvolution methods (summarized in **Supplemental Table S1**): *deconf* (20), *Linseed* (22), *DeconICA* (24), iterative non-negative matrix factorization with feature selection using *TOAST* (*TOAST* + *NMF*) (25), and *CellDistinguisher* (26). Estimated compartment proportions are compared to simulated or reported true compartment proportions with the mean square error (MSE) between the two matrices (see **Methods**). In total, we observed that *DeCompress* recapitulates compartment proportions with the least error compared to reference-free deconvolution methods.

In-silico GTEx mixing

We generated artificial targeted panels by mixing median tissue specific expression profiles from GTEx *in-silico* with randomly simulated compartment 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 (see **Methods**). **Figure 2A** shows the performance of *DeCompress* compared to other reference-free methods across 25 simulated targeted panels of increasing numbers of genes on the simulated targeted panels. In general, we find that *DeCompress* gives more accurate estimates of compartment proportions than the other 5 methods at both settings for multiplicative noise. As the number of genes in the targeted panel increased, the difference in MSE between *DeCompress* and the other methods remains largely constant. *Linseed* and *DeconICA*, methods that search for mutually independent axes of variation that correspond to compartments, consistently perform poorly on these simulated datasets, possibly due to the relative similarity between the expression profiles for these compartments and the small number of genes on the targeted panels. *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 compartments 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 compartment proportions. **Figure 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 compartments increased, though the range in the MSE decreases considerably. In particular, the performance of *DeconICA* increases considerably as more compartments were used for mixing, as mentioned in its documentation (24). 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 compartment proportions.

The four cell types we used for the above analyses simulated bulk mammary tissue but contained compartments with highly correlated gene expression profiles (**Supplemental Figure 2A**). We recreated the *in-silico* mixing experiments with four compartments with minimal correlations: mammary tissue, pancreas, pituitary gland, and whole blood (**Supplemental Figure 2A**). In mixtures with these tissues, we found that *DeCompress* also outperformed the reference-free methods, with a clear decrease in median MSE as the number of genes on the simulated targeted panels are increased (**Supplemental Figure 2B**). This trend between MSE and number of genes in this setting provides some evidence that dissimilar compartments may be easier to deconvolve with more genes on the targeted panel.

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 compartment mixture proportions. We downloaded expression data from a breast cancer cell-line mixture (RNA-seq) (23), rat brain, lung, and liver cell-line mixture (microarray) (11), prostate tumor with compartment proportions estimated with laser-capture microdissection (microarray) (58), and lung adenocarcinoma cell-line mixture (RNA-seq) (59) and generated pseudo-targeted panels with 200, 500, and 800 genes (see **Methods**). For the rat mixture dataset, we trained the compression sensing model on a randomly selected training split with added noise to simulate a batch effect between the training and targeted panel; for the other three cancer-related datasets, reference RNA-seq data was downloaded from The Cancer Genome Atlas (TCGA) (2). We then performed semi-reference-free deconvolution in these datasets using *DeCompress* and the reference-free methods.

Overall, *DeCompress* showed the lowest MSE across all three datasets, in comparison to the other reference-free methods (**Figure 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 range in the distribution of MSEs decreases. Deconvolution using *Linseed* gave variable performance across datasets (high variability in model performance), with very small ranges in MSEs in the rat microarray and lung adenocarcinoma datasets while highly variable MSEs 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 S3** 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 (57)) on the DeCompressed expression matrix for the breast, prostate, and lung cancer datasets (**Supplemental Figure S4**). We found that *unmix* gives accurate estimates of compartment proportions in the breast cancer and prostate tumor datasets, where the component compartments are like those in bulk tumors. However, in the case of the lung adenocarcinoma mixing dataset (mixture of two lung cancer cell lines), *unmix* does not consistently outperform the reference-free methods, perhaps owing to a dissimilarity between the lung adenocarcinoma mixture dataset and TCGA-LUAD reference dataset. We lastly investigated a scenario where 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 compartment proportions with larger errors, such that the distribution of MSEs intersect with a null distribution of MSEs from randomly generated compartment proportion matrices (**Supplemental Figure S5**).

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) (43, 55) 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 trained a computational algorithm to estimate compartment proportions using 148 tumor microarrays (TMAs) (89). We treat these estimated compartment proportions for epithelial tumor, adipose, stroma, and immune infiltrate 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 3A**. We also trained models stratified by estrogen-receptor (ER) status, a major, biologically-relevant classification in breast tumors (90, 91). 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 3B**). Next, as in the GTEx mixing simulations and the 4 published datasets, *DeCompress* recapitulated true compartment proportions with the minimum error (**Figure 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 compartment proportion matrices for 148 samples and 4 compartments. The mean MSE is provided in **Figure 3B**, 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 compartment proportions are positive and significantly non-zero for three of four compartment components (**Figure 3C**). Unlike those from *TOAST* + *NMF*, *DeCompress* estimates of compartment-specific compartment proportions were positively correlated with the truth (**Supplemental Figure S6**).

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 S7**). In practice, we recommend running an elastic net method (LASSO, elastic net, or ridge regression) which are both faster (**Supplemental Figure S7**) and give larger cross-validation R^2 (**Supplemental Figure S1**). 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 S8**). For *DeCompress*, we used TCGA-BRCA data with 1,212 samples as the reference. As shown in **Supplemental Figure S8**, running *DeCompress* in serial (approximately 62 minutes) takes around 40 times longer than the slowest reference-free deconvolution method (*TOAST* + *NMF*, approximately 1.5 minutes), though *DeCompress* is 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.

Applications of DeCompress in the Carolina Breast Cancer Study

Given the strong performance of DeCompress in benchmarking experiments, we estimated compartment proportions for 1,199 subjects in CBCS with transcriptomic data assayed with NanoString nCounter. Using TCGA breast cancer (TCGA-BRCA) expression as a training set, we iteratively searched for cell type-specific features (25) (Step 1 in **Figure 1**) and included canonical compartment markers for guidance using *a priori* knowledge (30, 92, 93) (see **Methods**). 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 descriptors 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 assess performance of using compartment-specific proportions in downstream analyses of breast cancer outcomes and gene regulation.

Identifying approximate modules for DeCompress-estimated 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 (53, 54) and single cell RNA-seq profiles of MCF7 breast cancer cells (94) (**Figure 4A**). Here, we find that Compartment 4 (C4) shows strong positive correlations with fibroblasts, lymphocytes, multiple collagenous organs (such as blood vessels, skin, bladder, vagina, and uterus (95–97)), and MCF7 cells. We hypothesize that strong correlation with lymphocytes reflects tumor-infiltrating lymphocytes. 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*) (98).

We conducted over-representation analysis (ORA) (65) of gene signatures for all five compartments, revealing cell cycle regulation ontologies for C4 that are consistent with the hypothesis generated from GTEx profiles at FDR-adjusted $P < 0.05$ (**Figure 4B**). We conducted gene set enrichment analysis (GSEA) for the C4 gene signature (99), 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 this enrichment did not achieve statistical significance. C1 and C5 gene signatures were not enriched for ontologies that allowed for conclusive compartment assignment, showing catabolic, morphogenic, and extracellular process ontologies (**Figure 4B**). From these results, we hypothesized that C3 and C4 resembled epithelial tumor cells, C2 an immune compartment (possibly excluding lymphocytes that may infiltrate tumors), and C1 and C5 presumptively stromal and/or mammary tissue.

Distributions of the hypothesized immune (C2) and tumor (C3 + C4 proportions) revealed significant differences across PAM50 molecular subtypes (**Figure 4C**; Kruskal-Wallis test of differences with $P < 2.2 \times 10^{-16}$) (69). These trends across subtypes were consistent with evidence that Basal-like and HER2-enriched subtypes had the largest proportions of estimated tumor and immune compartments, while Luminal A, Luminal B, and Normal-like subtypes showed lower proportions (43, 69, 100). Furthermore, we found strong differences in C4 and total tumor compartment estimates across race (**Supplemental Figure S10A**). 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 S10B**); however, none of the C3, C4, immune, or tumor compartment estimates showed significant differences across clinical ER status determined by immunohistochemistry (**Supplemental Figure S10C**). We considered the incorporation of estimates of compartment proportions in building models of breast cancer survival (**Supplemental Results** and **Supplemental Table S3**).

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

We investigated how incorporating estimated compartment proportions affect *cis*-expression quantitative trait loci (*cis*-eQTL) mapping in breast tumors, a common application of deconvolution methods in assessing sources of variation in gene regulation (9, 101). In previous eQTL studies using CBCS expression, several bulk breast tumor *cis*-eGenes (i.e. the gene of interest in an eQTL association between SNP and gene expression) were found in healthy mammary, subcutaneous adipose, or lymphocytes from GTEx (10). We included *DeCompress* proportion estimates for the tumor (C3 + C4 estimates) and immune (C2) compartments in a race-stratified, genetic ancestry-adjusted *cis*-eQTL interaction model (see **Methods**), as proposed by Geeleher *et al* and Westra *et al* (8, 9). We found that sets of compartment-specific *cis*-eGenes generally had few intersections with bulk *cis*-eGenes (**Figure 5A**), though we detected more *cis*-eQTLs with the immune- and tumor-specific interaction models (**Supplemental Figure S11**). 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 proportion 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. 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 S12** and **S13**, 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 S11**). All eQTLs with FDR-adjusted $P < 0.05$ are provided in **Supplemental Data** (https://github.com/bhattacharya-a-bt/DeCompress_supplement) (102).

We analyzed the sets of EA and AA tumor- and immune-specific eGenes in CBCS with ORA analysis for biological processes (**Figure 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 activation 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. 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 (see **Methods**), similar to previous pan-cancer germline eQTL analyses (10, 103). We attributed several of the bulk *cis*-eGenes to healthy GTEx tissue (all but 2), but tumor-specific *cis*-eGenes were less enriched in healthy tissues (**Supplemental Figure S14**). We compared the *cis*-eQTL effect sizes for significant CBCS *cis*-eSNPs found in GTEx. As shown in **Figure 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 high ($\rho = 0.71$). In contrast, we were unable to detect any of the CBCS tumor-specific *cis*-eQTLs in as significant eQTLs in GTEx healthy tissue, and the correlation of these effect sizes across CBCS and GTEx was poor (Spearman $\rho = -0.07$). These results suggest that this compartment-specific eQTL mapping, especially those that are tumor-specific, identified eQTLs that are not enriched for eQTLs from healthy tissue.

To evaluate any overlap of compartment-specific eQTLs with SNPs implicated with breast cancer risk, we extracted 932 risk-associated SNPs in women of European ancestry from iCOGS (86–88) at FDR-adjusted $P < 0.05$ that were available on the CBCS OncoArray panel (71). **Figure 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 three *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 $r^2 \geq 0.99$ (strongest association with rs56387622) with chemokine receptor *CCR3*, a gene whose expression was previously found to be associated with breast cancer outcomes in luminal-like subtypes (104, 105). As estimated tumor purity increases, the cancer risk allele C at rs56387622 has a consistently strong negative effect on *CCR3* expression (**Figure 5E**). We find that *CCR3* expression is insignificantly different across tumor stage and ER status but is significantly different across PAM50 molecular subtype (**Supplemental Figure S15**). In sum, results from our *cis*-eQTL analysis show the advantage of including *DeCompress*-estimated compartment proportions in downstream genomic analyses to identify compartment-specific associations that may be relevant in disease pathways.

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 (3, 35). Unlike traditional reference-based methods that require compartment-specific expression profiles, *DeCompress* requires only a reference RNA-seq or microarray dataset 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 compartment-specific expression on the same targeted panel. We benchmarked *DeCompress* against reference-free methods (20, 22, 24–26) using *in-silico* GTEx mixing experiments (53, 54), 4 published datasets with known compartment proportions (11, 23, 58, 59), and a large, heterogeneous NanoString nCounter dataset from the CBCS (43, 55). In these analyses, we showed that *DeCompress* recapitulated true compartment 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. We tested the performance of *DeCompress* by incorporating compartment estimates in eQTL mapping to reveal immune- and tumor-compartment-specific breast cancer eQTLs.

While *DeCompress* has several important strengths, it has some limitations. First, *DeCompress* has a high computational cost, owing mainly to its lengthy compressed 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 (25). However, *DeCompress* estimates compartment 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. Second, *DeCompress*, as a semi-reference-free method, shares the limitations of reference-based methods – namely concerns with 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 has slightly lower performance than datasets properly matched to their references. Yet, in this setting, *DeCompress* performance was on par with that of the other reference-free methods that do not use a misaligned reference. Lastly, also in common with 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. Researchers may employ more systematic methods of assessing the similarity of the reference and target datasets, like measuring the distance between the two matrices (i.e. norms based on the singular values of matrices) or comparing the correlation structure of overlapping genes in the feature spaces of the reference and target. These evaluations will help with selecting a proper reference for a targeted panel to be deconvolved using *DeCompress*.

DeCompress also shares some challenges with reference-free deconvolution methods, such as the selection of an appropriate number of compartments. Previous groups have emphasized reliance on *a priori* knowledge for deconvolving well-studied tissues, such as blood and brain (106, 107). 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 may be rare or reflect transient cell states (30, 91, 108, 109). For this reason, we included several data-driven approaches in estimating the number of compartments from variation in the gene expression and recommended applying prior domain knowledge about the tissue of interest. It is also important to carefully consider the gene module-based annotations for the unidentified estimated compartments, especially in bulk tissue where traditional ideas of compartments are inapplicable (29). Several previous reference-free methods have leveraged *in vitro* mixtures of highly distinct cell lines in training and testing previous reference-free deconvolution methods (11, 22), namely the rat cell line mixture (GSE19830) (11). 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 the difficulty of 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 checking hypotheses against 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 (101). In breast cancer, Geeleher *et al* previously showed that a similarly implemented interaction eQTL model gave better mapping of compartment-specific eQTLs (8, 9). 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 targeted panel. As the panel included 406 genes, all previously implicated in breast cancer pathogenesis, proliferation, or response (10, 43, 110), the interaction model will detect SNPs that have large effects on compartment-specific genes. The interaction term is interpreted as the difference in eQTL effect sizes between 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 compartment-specific eQTL effects (111–113). 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. This interaction term's effect size may also be inflated for compartment estimates that have low mean and high variance across the samples. In addition, we did not consider *trans*-acting eQTLs that are often attributed to compartment heterogeneity, though we believe that methods employing mediation or cross-condition analysis can be integrated with compartment estimates to map compartment-specific *trans*-eQTLs relevant in breast cancer (114–116).

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 (86–88) 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 (104, 105). The CCR3 receptor has been shown to be the primary binding site of CCL11 (eotaxin-1), an eosinophil-selective chemoattractant cytokine (117, 118), and accordingly CCR3 antagonism prohibited chemotaxis of basophils and eosinophils, a phenomenon observed in breast cancer activation and proliferation (119, 120). Without *DeCompress* and the incorporation of estimated compartment proportions in the eQTL model, this association between eSNP and *CCR3* expression would not have been detected in this dataset (121).

DeCompress, our semi-reference-free deconvolution method, provides a powerful method to estimate compartment-specific proportions for targeted expression panels that have a limited number of genes and 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 provide 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 and dynamic tissues, we anticipate creative implementations of *DeCompress* to give further insight into expression variation in complex diseases.

DATA AVAILABILITY

The *DeCompress* package is available as R software on GitHub: https://github.com/bhattacharya-a-bt/DeCompress. Sample code for replication and results from the eQTL analysis are provided: https://github.com/bhattacharya-a-bt/DeCompress_supplement (102). CBCS expression data is publicly available at GSE148426. CBCS genotype datasets analyzed in this study are not publicly available as many CBCS patients are still being followed and accordingly is considered sensitive; the data is available from M.A.T upon reasonable request. GTEx median expression profiles are available from dbGAP accession number phs000424.v7.p2. Data from the published mixture experiments are available from GEO: GSE19830, GSE123604, GSE97284, and GSE64098. Single-cell expression profiles of MCF7 cells were obtained from GSE52716. Expression data from The Cancer Genome Atlas is available from the Broad GDAC Firehose repository (https://gdac.broadinstitute.org/) with accession number phs000178.v11.p8.

SUPPLEMENTARY DATA

Additional File 1: Supplemental Methods, Results, Tables, and Figures

ACKNOWLEDGEMENTS

We think the Carolina Breast Cancer Study participants and volunteers. We also thank Katie Hoadley, Linnea Olsson, Chuck Perou, Luca Pinello, and Naim Rashid for valuable discussion during the research process. We thank Erin Kirk and Jessica Tse for their invaluable support during the research process. We thank the DCEG Cancer Genomics Research Laboratory and acknowledge the support from Stephen Chanock, Rose Yang, Mereditch Yeager, Belynda Hicks, and Bin Zhu.

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from: the GTEx Portal (https://storage.googleapis.com/gtex_analysis_v7/rna_seq_data/GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8_gene_median_tpm.gct.gz) with dbGaP accession number phs000424.v7.p2 on 05/14/20.

FUNDING

Susan G. Komen® provided financial support for CBCS study infrastructure. Funding was provided by the National Institutes of Health, National Cancer Institute P01-CA151135, P50-CA05822, and U01-CA179715 to M.A.T. M.I.L. is supported by P01-CA142538, and A.B. and M.I.L are supported by P30-ES010126. A.M.H. is supported by 1T32GM12274 (National Institute of General Medical Sciences). The Translational Genomics Laboratory is supported in part by grants from the National Cancer Institute (3P30CA016086) and the University of North Carolina at Chapel Hill University Cancer Research Fund. The funders had no role in the design of the study, the collection, analysis, or interpretation of the data, the writing of the manuscript, or the decision to submit the manuscript for publication.

CONFLICT OF INTEREST

The author have no conflicts of interest to disclose. This study was approved by the Office of Human Research Ethics at the University of North Carolina at Chapel Hill, and written informed consent was obtained from each participant. All experimental methods abided by the Helsinki Declaration.

REFERENCES

1. A. Bennett,D., A. Schneider,J., Arvanitakis,Z. and S. Wilson,R. (2013) Overview and Findings from the Religious Orders Study. *Curr. Alzheimer Res.*, **9**, 628–645.

2. Weinstein,J.N., Collisson,E.A., Mills,G.B., Shaw,K.R.M., Ozenberger,B.A., Ellrott,K., Sander,C., Stuart,J.M., Chang,K., Creighton,C.J., *et al.* (2013) The cancer genome atlas pan-cancer analysis project. *Nat. Genet.*, **45**, 1113–1120.

3. Wallden,B., Storhoff,J., Nielsen,T., Dowidar,N., Schaper,C., Ferree,S., Liu,S., Leung,S., Geiss,G., Snider,J., *et al.* (2015) Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. *BMC Med. Genomics*, **8**, 54.

4. Tellez-Gabriel,M., Ory,B., Lamoureux,F., Heymann,M.F. and Heymann,D. (2016) Tumour heterogeneity: The key advantages of single-cell analysis. *Int. J. Mol. Sci.*, **17**.

5. McGregor,K., Bernatsky,S., Colmegna,I., Hudson,M., Pastinen,T., Labbe,A. and Greenwood,C.M.T. (2016) An evaluation of methods correcting for cell-type heterogeneity in DNA methylation studies.

Genome Biol., **17**, 84.

6. Kuhn,A., Thu,D., Waldvogel,H.J., Faull,R.L.M.M. and Luthi-Carter,R. (2011) Population-specific expression analysis (PSEA) reveals molecular changes in diseased brain. *Nat. Methods*, **8**, 945–947.

7. Guintivano,J., Aryee,M.J. and Kaminsky,Z.A. (2013) 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**, 290–302.

8. André,G.;, Westra,H.-J., Arends,D., Esko,T., Peters,M.J., Schurmann,C. and Schramm,K. Cell Specific eQTL Analysis without Sorting Cells. *Cell Specif. eQTL Anal. without Sorting Cells. PLoS Genet*, **24**, 1005223.

9. Geeleher,P., Nath,A., Wang,F., Zhang,Z., Barbeira,A.N., Fessler,J., Grossman,R.L., Seoighe,C. and Stephanie Huang,R. (2018) Cancer expression quantitative trait loci (eQTLs) can be determined from heterogeneous tumor gene expression data by modeling variation in tumor purity. *Genome Biol.*, **19**, 130.

10. Bhattacharya,A., García-Closas,M., Olshan,A.F., Perou,C.M., Troester,M.A. and Love,M.I. (2020) A framework for transcriptome-wide association studies in breast cancer in diverse study populations. *Genome Biol.*, **21**, 42.

11. Shen-Orr,S.S., Tibshirani,R., Khatri,P., Bodian,D.L., Staedtler,F., Perry,N.M., Hastie,T., Sarwal,M.M., Davis,M.M. and Butte,A.J. (2010) Cell type-specific gene expression differences in complex tissues. *Nat. Methods*, **7**, 287–289.

12. Kim-Hellmuth,S., Aguet,F., Oliva,M., Muñoz-Aguirre,M., Wucher,V., Kasela,S., Castel,S.E., Hamel,A.R., Viñuela,A., Roberts,A.L., *et al.* (2019) Cell type specific genetic regulation of gene expression across human tissues. *bioRxiv*, 10.1101/806117.

13. Bertsekas,D.P. (1999) Convex Optimization Algorithms Athena Scientific, Belmot, Massachusetts.

14. Zhong,Y., Wan,Y.-W., Pang,K., Chow,L.M.L. and Liu,Z. (2013) Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC Bioinformatics*, **14**, 89.

15. Quon,G., Haider,S., Deshwar,A.G., Cui,A., Boutros,P.C. and Morris,Q. (2013) Computational purification of individual tumor gene expression profiles leads to significant improvements in

prognostic prediction. *Genome Med.*, **5**, 29.

16. Wang,Z., Cao,S., Morris,J.S., Ahn,J., Liu,R., Tyekucheva,S., Gao,F., Li,B., Lu,W., Tang,X., *et al.* (2018) Transcriptome Deconvolution of Heterogeneous Tumor Samples with Immune Infiltration. *iScience*, **9**, 451–460.
17. Chen,B., Khodadoust,M.S., Liu,C.L., Newman,A.M. and Alizadeh,A.A. (2018) Profiling tumor infiltrating immune cells with CIBERSORT. In *Methods in Molecular Biology*. Humana Press Inc., Vol. 1711, pp. 243–259.
18. Wang,J., Devlin,B. and Roeder,K. (2019) Using multiple measurements of tissue to estimate subject- and cell-type-specific gene expression. *Bioinformatics*, 10.1093/bioinformatics/btz619.
19. Dong,M., Thennavan,A., Urrutia,E., Li,Y., Perou,C.M., Zou,F. and Jiang,Y. (2020) SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.*, 10.1093/bib/bbz166.
20. Repsilber,D., Kern,S., Telaar,A., Walzl,G., Black,G.F., Selbig,J., Parida,S.K., Kaufmann,S.H. and Jacobsen,M. (2010) Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC Bioinformatics*, **11**, 27.
21. Wang,X., Park,J., Susztak,K., Zhang,N.R. and Li,M. (2019) Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.*, **10**.
22. Zaitsev,K., Bambouskova,M., Swain,A. and Artyomov,M.N. (2019) Complete deconvolution of cellular mixtures based on linearity of transcriptional signatures. *Nat. Commun.*, **10**.
23. Kang,K., Meng,Q., Shats,I., Umbach,D.M., Li,M., Li,Y., Li,X. and Li,L. (2019) CDSeq: A novel complete deconvolution method for dissecting heterogeneous samples using gene expression data. *PLOS Comput. Biol.*, **15**, e1007510.
24. Czerwinska,U. (2018) DeconICA. 10.5281/ZENODO.1250070.
25. Li,Z. and Wu,H. (2019) TOAST: improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biol.*, **20**, 190.
26. Newberg,L.A., Chen,X., Kodira,C.D. and Zavodszky,M.I. (2018) Computational de novo discovery of distinguishing genes for biological processes and cell types in complex tissues. *PLoS One*, **13**, e0193067.

27. Schelker,M., Feau,S., Du,J., Ranu,N., Klipp,E., MacBeath,G., Schoeberl,B. and Raue,A. (2017) Estimation of immune cell content in tumour tissue using single-cell RNA-seq data. *Nat. Commun.*, **8**, 2032.

28. Yousefi,P., Huen,K., Quach,H., Motwani,G., Hubbard,A., Eskenazi,B. and Holland,N. (2015) Estimation of blood cellular heterogeneity in newborns and children for epigenome-wide association studies. *Environ. Mol. Mutagen.*, **56**, 751–758.

29. Clevers,H. (2017) What Is Your Conceptual Definition of “Cell Type” in the Context of a Mature Organism? *Cell Syst.*, **4**, 255–259.

30. Wu,S.Z., Roden,D.L., Wang,C., Holliday,H., Harvey,K., Cazet,A.S., Murphy,K.J., Pereira,B., Al-Eryani,G., Hou,R., *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**.

31. Barkley,D. and Yanai,I. (2019) Plasticity and Clonality of Cancer Cell States. *Trends in Cancer*, **5**, 655–656.

32. van der Leun,A.M., Thommen,D.S. and Schumacher,T.N. (2020) CD8+ T cell states in human cancer: insights from single-cell analysis. *Nat. Rev. Cancer*, **20**, 218–232.

33. Peng,X.L., Moffitt,R.A., Torphy,R.J., Volmar,K.E. and Yeh,J.J. (2019) De novo compartment deconvolution and weight estimation of tumor samples using DECODER. *Nat. Commun.*, **10**, 1–11.

34. Li,Z., Wu,Z., Jin,P. and Wu,H. Dissecting differential signals in high-throughput data from complex tissues. 10.1093/bioinformatics/btz196.

35. Geiss,G.K., Bumgarner,R.E., Birditt,B., Dahl,T., Dowidar,N., Dunaway,D.L., Fell,H.P., Ferree,S., George,R.D., Grogan,T., *et al.* (2008) Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nat. Biotechnol.*, **26**, 317–325.

36. Marczyk,M., Fu,C., Lau,R., Du,L., Trevarton,A.J., Sinn,B. V., Gould,R.E., Pusztai,L., Hatzis,C. and Symmans,W.F. (2019) The impact of RNA extraction method on accurate RNA sequencing from formalin-fixed paraffin-embedded tissues. *BMC Cancer*, **19**, 1189.

37. Mercer,T.R., Gerhardt,D.J., Dinger,M.E., Crawford,J., Trapnell,C., Jeddeloh,J.A., Mattick,J.S. and Rinn,J.L. (2012) Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat. Biotechnol.*, **30**, 99–104.

38. Veldman-Jones,M.H., Brant,R., Rooney,C., Geh,C., Emery,H., Harbron,C.G., Wappett,M., Sharpe,A., Dymond,M., Barrett,J.C., *et al.* (2015) Evaluating Robustness and Sensitivity of the NanoString Technologies nCounter Platform to Enable Multiplexed Gene Expression Analysis of Clinical Samples. *Cancer Res.*, **75**, 2587–2593.

39. Brasó-Maristany,F., Filosto,S., Catchpole,S., Marlow,R., Quist,J., Francesch-Domenech,E., Plumb,D.A., Zakka,L., Gazinska,P., Liccardi,G., *et al.* (2016) PIM1 kinase regulates cell death, tumor growth and chemotherapy response in triple-negative breast cancer. *Nat. Med.*, **22**, 1303–1313.

40. Urrutia,A., Duffy,D., Rouilly,V., Posseme,C., Djebali,R., Illanes,G., Libri,V., Albaud,B., Gentien,D., Piasecka,B., *et al.* (2016) Standardized Whole-Blood Transcriptional Profiling Enables the Deconvolution of Complex Induced Immune Responses. *Cell Rep.*, **16**, 2777–2791.

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

42. Ng,S.W.K., Mitchell,A., Kennedy,J.A., Chen,W.C., McLeod,J., Ibrahimova,N., Arruda,A., Popescu,A., Gupta,V., Schimmer,A.D., *et al.* (2016) A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*, **540**, 433–437.

43. Troester,M.A., Sun,X., Allott,E.H., Geradts,J., Cohen,S.M., Tse,C.-K., Kirk,E.L., Thorne,L.B., Mathews,M., Li,Y., *et al.* (2018) Racial Differences in PAM50 Subtypes in the Carolina Breast Cancer Study. *JNCI J. Natl. Cancer Inst.*, **110**, 176–182.

44. Vieira,A.F. and Schmitt,F. (2018) An Update on Breast Cancer Multigene Prognostic Tests-Emergent Clinical Biomarkers. *Front. Med.*, **5**, 248.

45. Candès,E.J. and Romberg,J. (2006) Quantitative robust uncertainty principles and optimally sparse decompositions. *Found. Comput. Math.*, **6**, 227–254.

46. Candès,E.J., Romberg,J. and Tao,T. (2006) Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, **52**, 489–509.

47. Efron,B., Hastie,T., Johnstone,I. and Tibshirani,R. (2004) LEAST ANGLE REGRESSION.

48. Friedman,J., Hastie,T. and Tibshirani,R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.*, **33**, 1–22.

49. Liao,S.J. (1999) An explicit, totally analytic approximate solution for Blasius' viscous flow problems. *Int. J. Non. Linear. Mech.*, **34**, 759–778.

50. Goodfellow,I.J., Pouget-Abadie,J., Mirza,M., Xu,B., Warde-Farley,D., Ozair,S., Courville,A. and Bengio,Y. (2014) Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*.pp. 2672–2680.

51. Viñas,R., Azevedo,T., Gamazon,E.R. and Liò,P. Gene Expression Imputation with Generative Adversarial Imputation Nets. 10.1101/2020.06.09.141689.

52. Yoon,J., Jordon,J. and Van Der Schaar,M. (2018) GAIN: Missing Data Imputation using Generative Adversarial Nets.

53. Lonsdale,J., Thomas,J., Salvatore,M., Phillips,R., Lo,E., Shad,S., Hasz,R., Walters,G., Garcia,F., Young,N., *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

54. Ardlie,K.G., DeLuca,D.S., Segrè,A. V., Sullivan,T.J., Young,T.R., Gelfand,E.T., Trowbridge,C.A., Maller,J.B., Tukiainen,T., Lek,M., *et al.* (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).*, **348**, 648–660.

55. Newman,B., Moorman,P.G., Millikan,R., Qaqish,B.F., Geradts,J., Aldrich,T.E. and Liu,E.T. (1995) The Carolina Breast Cancer Study: integrating population-based epidemiology and molecular biology. *Breast Cancer Res. Treat.*, **35**, 51–60.

56. Donoho,D.L. (2006) Compressed sensing. *IEEE Trans. Inf. Theory*, **52**, 1289–1306.

57. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.

58. Tyekucheva,S., Bowden,M., Bango,C., Giunchi,F., Huang,Y., Zhou,C., Bondi,A., Lis,R., Van Hemelrijck,M., Andrén,O., *et al.* (2017) Stromal and epithelial transcriptional map of initiation progression and metastatic potential of human prostate cancer. *Nat. Commun.*, **8**.

59. Holik,A.Z., Law,C.W., Liu,R., Wang,Z., Wang,W., Ahn,J., Asselin-Labat,M.-L., Smyth,G.K. and Ritchie,M.E. (2016) RNA-seq mixology: designing realistic control experiments to compare protocols and analysis methods. *Nucleic Acids Res.*, **45**.

60. Liu,R., Holik,A.Z., Su,S., Jansz,N., Chen,K., Leong,H.S., Blewitt,M.E., Asselin-Labat,M.-L., Smyth,G.K. and Ritchie,M.E. (2015) Why weight? Modelling sample and observational level variability improves power in RNA-seq analyses. *Nucleic Acids Res.*, **43**, 97.

61. Bhattacharya,A., Hamilton,A.M., Furberg,H., Pietzak,E., Purdue,M.P., Troester,M.A., Hoadley,K.A. and Love,M.I. (2020) An approach for normalization and quality control for NanoString RNA expression data. *Brief. Bioinform.*, 10.1093/bib/bbaa163.

62. Bullard,J.H., Purdom,E., Hansen,K.D. and Dudoit,S. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**, 94.

63. Risso,D., Ngai,J., Speed,T.P. and Dudoit,S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, **32**, 896–902.

64. Ritchie,M.E., Phipson,B., Wu,D., Hu,Y., Law,C.W., Shi,W. and Smyth,G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47–e47.

65. Liao,Y., Wang,J., Jaehnig,E.J., Shi,Z. and Zhang,B. (2019) WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.*, **47**, 199–205.

66. Consortium,T.G.O. (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**, D330–D338.

67. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., *et al.* (2000) Gene ontology: Tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

68. Austin,P.C. and Fine,J.P. (2017) Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Stat. Med.*, **36**, 4391–4400.

69. Parker,J.S., Mullins,M., Cheang,M.C.U., Leung,S., Voduc,D., Vickery,T., Davies,S., Fauron,C., He,X., Hu,Z., *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.*, **27**, 1160–1167.

70. Benjamini,Y. and Hochberg,Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple.

71. Amos,C.I., Dennis,J., Wang,Z., Byun,J., Schumacher,F.R., Gayther,S.A., Casey,G., Hunter,D.J.,

Sellers,T.A., Gruber,S.B., *et al.* (2017) The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol. Biomarkers Prev.*, **26**, 126–135.

72. Lilyquist,J., Ruddy,K.J., Vachon,C.M. and Couch,F.J. (2018) Common Genetic Variation and Breast Cancer Risk-Past, Present, and Future. *Cancer Epidemiol. Biomarkers Prev.*, **27**, 380–394.

73. Auton,A., Abecasis,G.R., Altshuler,D.M., Durbin,R.M., Bentley,D.R., Chakravarti,A., Clark,A.G., Donnelly,P., Eichler,E.E., Flicek,P., *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

74. O'Connell,J., Gurdasani,D., Delaneau,O., Pirastu,N., Ulivi,S., Cocca,M., Traglia,M., Huang,J., Huffman,J.E., Rudan,I., *et al.* (2014) A General Approach for Haplotype Phasing across the Full Spectrum of Relatedness. *PLoS Genet.*, **10**, e1004234.

75. Delaneau,O., Marchini,J. and Zagury,J.-F. (2012) A linear complexity phasing method for thousands of genomes. *Nat. Methods*, **9**, 179–181.

76. Howie,B.N., Donnelly,P. and Marchini,J. (2009) A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet.*, **5**, e1000529.

77. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A.R., Bender,D., Maller,J., Sklar,P., De Bakker,P.I.W., Daly,M.J., *et al.* (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet*, **81**, 559–575.

78. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) DbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

79. Shabalin,A.A. (2012) Gene expression Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, **28**, 1353–1358.

80. Palowitch,J., Shabalin,A., Zhou,Y.H., Nobel,A.B. and Wright,F.A. (2018) Estimation of cis-eQTL effect sizes using a log of linear model. *Biometrics*, **74**, 616–625.

81. Sun,W. (2012) A Statistical Framework for eQTL Mapping Using RNA-seq Data. *Biometrics*, **68**, 1–11.

82. Mohammadi,P., Castel,S.E., Brown,A.A. and Lappalainen,T. (2017) Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res.*, **27**, 1872–1884.

83. Ellsworth,R.E., Blackburn,H.L., Shriver,C.D., Soon-Shiong,P. and Ellsworth,D.L. (2017) Molecular heterogeneity in breast cancer: State of the science and implications for patient care. *Semin. Cell*

Dev. Biol., **64**, 65–72.

84. Turashvili,G. and Brogi,E. (2017) Tumor Heterogeneity in Breast Cancer. *Front. Med.*, **4**.

85. Wen,Y., Wei,Y., Zhang,S., Li,S., Liu,H., Wang,F., Zhao,Y., Zhang,D. and Zhang,Y. (2016) Cell subpopulation deconvolution reveals breast cancer heterogeneity based on DNA methylation signature. *Brief. Bioinform.*, 10.1093/bib/bbw028.

86. Michailidou,K., Hall,P., Gonzalez-Neira,A., Ghoussaini,M., Dennis,J., Milne,R.L., Schmidt,M.K., Chang-Claude,J., Bojesen,S.E., Bolla,M.K., *et al.* (2013) Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.*, **45**, 353–361.

87. Michailidou,K., Beesley,J., Lindstrom,S., Canisius,S., Dennis,J., Lush,M.J., Maranian,M.J., Bolla,M.K., Wang,Q., Shah,M., *et al.* (2015) Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.*, **47**, 373–380.

88. Michailidou,K., Lindström,S., Dennis,J., Beesley,J., Hui,S., Kar,S., Lemaçon,A., Soucy,P., Glubb,D., Rostamianfar,A., *et al.* (2017) Association analysis identifies 65 new breast cancer risk loci. *Nature*, **551**, 92–94.

89. Sandhu,R., Chollet-Hinton,L., Kirk,E.L., Midkiff,B. and Troester,M.A. (2016) Digital histologic analysis reveals morphometric patterns of age-related involution in breast epithelium and stroma. *Hum. Pathol.*, **48**, 60–68.

90. Sørlie,T., Tibshirani,R., Parker,J., Hastie,T., Marron,J.S., Nobel,A., Deng,S., Johnsen,H., Pesich,R., Geisler,S., *et al.* (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 8418–8423.

91. Perou,C.M., Sørile,T., Eisen,M.B., Van De Rijn,M., Jeffrey,S.S., Ress,C.A., Pollack,J.R., Ross,D.T., Johnsen,H., Akslen,L.A., *et al.* (2000) Molecular portraits of human breast tumours. *Nature*, **406**, 747–752.

92. Azizi,E., Carr,A.J., Plitas,G., Mazutis,L., Rudensky,A.Y., Pe'er,D., Cornish,A.E., Konopacki,C., Prabhakaran,S., Nainys,J., *et al.* (2018) 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.e36.

93. Nguyen,Q.H., Pervolarakis,N., Blake,K., Ma,D., Davis,R.T., James,N., Phung,A.T., Willey,E.,

Kumar,R., Jabart,E., *et al.* (2018) Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.*, **8**, 1–12.

94. Rothwell,D.G., Li,Y., Ayub,M., Tate,C., Newton,G., Hey,Y., Carter,L., Faulkner,S., Moro,M., Pepper,S., *et al.* (2014) Evaluation and validation of a robust single cell RNA-amplification protocol through transcriptional profiling of enriched lung cancer initiating cells. *BMC Genomics*, **15**.

95. Smith,B.A., Balanis,N.G., Nanjundiah,A., Sheu,K.M., Tsai,B.L., Zhang,Q., Park,J.W., Thompson,M., Huang,J., Witte,O.N., *et al.* (2018) A Human Adult Stem Cell Signature Marks Aggressive Variants across Epithelial Cancers. *Cell Rep.*, **24**, 3353-3366.e5.

96. Uhlen,M., Zhang,C., Lee,S., Sjöstedt,E., Fagerberg,L., Bidkhori,G., Benfeitas,R., Arif,M., Liu,Z., Edfors,F., *et al.* (2017) A pathology atlas of the human cancer transcriptome. *Science (80-.).*, **357**.

97. Uhlen,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A., *et al.* (2015) Tissue-based map of the human proteome. *Science (80-.).*, **347**, 1260419–1260419.

98. Prat,A., As Pascual,T., De Angelis,C., Gutierrez,C., Llombart-Cussac,A., Wang,T., Cort,J., Rexer,B., Par,L., Forero,A., *et al.* HER2-Enriched Subtype and ERBB2 Expression in HER2-Positive Breast Cancer Treated with Dual HER2 Blockade. 10.1093/jnci/djz042.

99. Yu,G., Wang,L.G., Han,Y. and He,Q.Y. (2012) ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.*, **16**, 284–287.

100. D'Arcy,M., Fleming,J., Robinson,W.R., Kirk,E.L., Perou,C.M., Troester,M.A., D'Arcy,M., Fleming,J., Robinson,W.R., Kirk,E.L., *et al.* (2015) Race-associated biological differences among Luminal A breast tumors. *Breast Cancer Res. Treat.*, **152**, 437–448.

101. Aguirre-Gamboa,R., de Klein,N., di Tommaso,J., Claringbould,A., van der Wijst,M.G., de Vries,D., Brugge,H., Oelen,R., Võsa,U., Zorro,M.M., *et al.* (2020) Deconvolution of bulk blood eQTL effects into immune cell subpopulations. *BMC Bioinformatics*, **21**, 243.

102. Bhattacharya,A., Hamilton,A.M., Troester,M.A. and Love,M.I. (2020) Code and summary results for DeCompress. 10.5281/zenodo.3979913.

103. Calabrese,C., Lehmann,K., Urban,L., Liu,F., Erkek,S., Fonseca,N.A., Kahles,A., Kilpinen,H., Markowski,J., 3,P.G., *et al.* (2017) Assessing the Gene Regulatory Landscape in 1,188 Human

Tumors. *bioRxiv*, 10.1101/225441.

104. Gong,D.H., Fan,L., Chen,H.Y., Ding,K.F. and Yu,K. Da (2016) Intratumoral expression of CCR3 in breast cancer is associated with improved relapse-free survival in luminal-like disease. *Oncotarget*, **7**, 28570–28578.

105. Thomas,J.K., Mir,H., Kapur,N., Bae,S. and Singh,S. (2019) CC chemokines are differentially expressed in Breast Cancer and are associated with disparity in overall survival. *Sci. Rep.*, **9**, 1–12.

106. Reinius,L.E., Acevedo,N., Joerink,M., Pershagen,G., Dahlén,S.-E., Greco,D., Söderhäll,C., Scheynius,A. and Kere,J. (2012) Differential DNA Methylation in Purified Human Blood Cells: Implications for Cell Lineage and Studies on Disease Susceptibility. *PLoS One*, **7**, e41361.

107. Montaño,C.M., Irizarry,R.A., Kaufmann,W.E., Talbot,K., Gur,R.E., Feinberg,A.P. and Taub,M.A. (2013) Measuring cell-type specific differential methylation in human brain tissue. *Genome Biol.*, **14**, R94.

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

109. Hoadley,K.A., Yau,C., Hinoue,T., Wolf,D.M., Lazar,A.J., Drill,E., Shen,R., Taylor,A.M., Cherniack,A.D., Thorsson,V., *et al.* (2018) Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. *Cell*, **173**, 291-304.e6.

110. D'Arcy,M., Fleming,J., Robinson,W.R., Kirk,E.L., Perou,C.M. and Troester,M.A. (2015) Race-associated biological differences among Luminal A breast tumors. *Breast Cancer Res. Treat.*, **152**, 437–448.

111. Wang,F., Dohogne,Z., Yang,J., Liu,Y. and Soibam,B. (2018) Predictors of breast cancer cell types and their prognostic power in breast cancer patients. *BMC Genomics*, **19**, 137.

112. Troester,M.A., Hoadley,K.A., Sørlie,T., Herbert,B.S., Børresen-Dale,A.L., Lønning,P.E., Shay,J.W., Kaufmann,W.K. and Perou,C.M. (2004) Cell-type-specific responses to chemotherapeutics in breast cancer. *Cancer Res.*, **64**, 4218–4226.

113. Schaefer,M.H. and Serrano,L. (2016) Cell type-specific properties and environment shape tissue

specificity of cancer genes. *Sci. Rep.*, **6**, 1–14.

114. Yang,F., Gleason,K.J., Wang,J., consortium,T.Gte., Duan,J., He,X., Pierce,B.L. and Chen,L.S. (2019) CCmed: cross-condition mediation analysis for identifying robust trans-eQTLs and assessing their effects on human traits. *bioRxiv*, 10.1101/803106.

115. Shan,N., Wang,Z. and Hou,L. (2019) Identification of trans-eQTLs using mediation analysis with multiple mediators. *BMC Bioinformatics*, **20**.

116. Pierce,B.L., Tong,L., Chen,L.S., Rahaman,R., Argos,M., Jasmine,F., Roy,S., Paul-Brutus,R., Westra,H.J., Franke,L., *et al.* (2014) Mediation Analysis Demonstrates That Trans-eQTLs Are Often Explained by Cis-Mediation: A Genome-Wide Analysis among 1,800 South Asians. *PLoS Genet.*, **10**.

117. Jöhrer,K., Zelle-Rieser,C., Perathoner,A., Moser,P., Hager,M., Ramoner,R., Gander,H., Höltl,L., Bartsch,G., Greil,R., *et al.* (2005) Up-regulation of functional chemokine receptor CCR3 in human renal cell carcinoma. *Clin. Cancer Res.*, **11**, 2459–2465.

118. Miyagaki,T., Sugaya,M., Murakami,T., Asano,Y., Tada,Y., Kadono,T., Okochi,H., Tamaki,K. and Sato,S. (2011) CCL11-CCR3 interactions promote survival of anaplastic large cell lymphoma cells via ERK1/2 activation. *Cancer Res.*, **71**, 2056–2065.

119. Bryan,S.A., Jose,P.J., Topping,J.R., Wilhelm,R., Soderberg,C., Kertesz,D., Barnes,P.J., Williams,T.J., Hansel,T.T. and Sabroe,I. (2002) Responses of leukocytes to chemokines in whole blood and their antagonism by novel CC-Chemokine Receptor 3 antagonists. *Am. J. Respir. Crit. Care Med.*, **165**, 1602–1609.

120. Samoszuk,M.K., Nguyen,V., Gluzman,I. and Pham,J.H. (1996) Occult Deposition of Eosinophil Peroxidase in a Subset of Human Breast Carcinomas.

121. Alasoo,K., Rodrigues,J., Mukhopadhyay,S., Knights,A.J., Mann,A.L., Kundu,K., Hale,C., Dougan,G. and Gaffney,D.J. (2018) Shared genetic effects on chromatin and gene expression indicate a role for enhancer priming in immune response. *Nat. Genet.*, **50**, 424–431.

FIGURE LEGENDS

Figure 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.

Figure 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 various methods (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 number of genes in each simulated dataset. **(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 reference-free method tested. **(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) (23), rat brain, lung, and liver cell-line mixture (top-right) (11), prostate tumor samples (bottom-left) (58), and lung adenocarcinoma cell-line mixture (bottom-right) (59). 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.

Figure 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 $P < 0.05$.

Figure 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)** Bar plot 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 + C4 compartments, right) proportions (Y-axis) across PAM50 molecular subtypes (X-axis)

Figure 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 (86–88), 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.

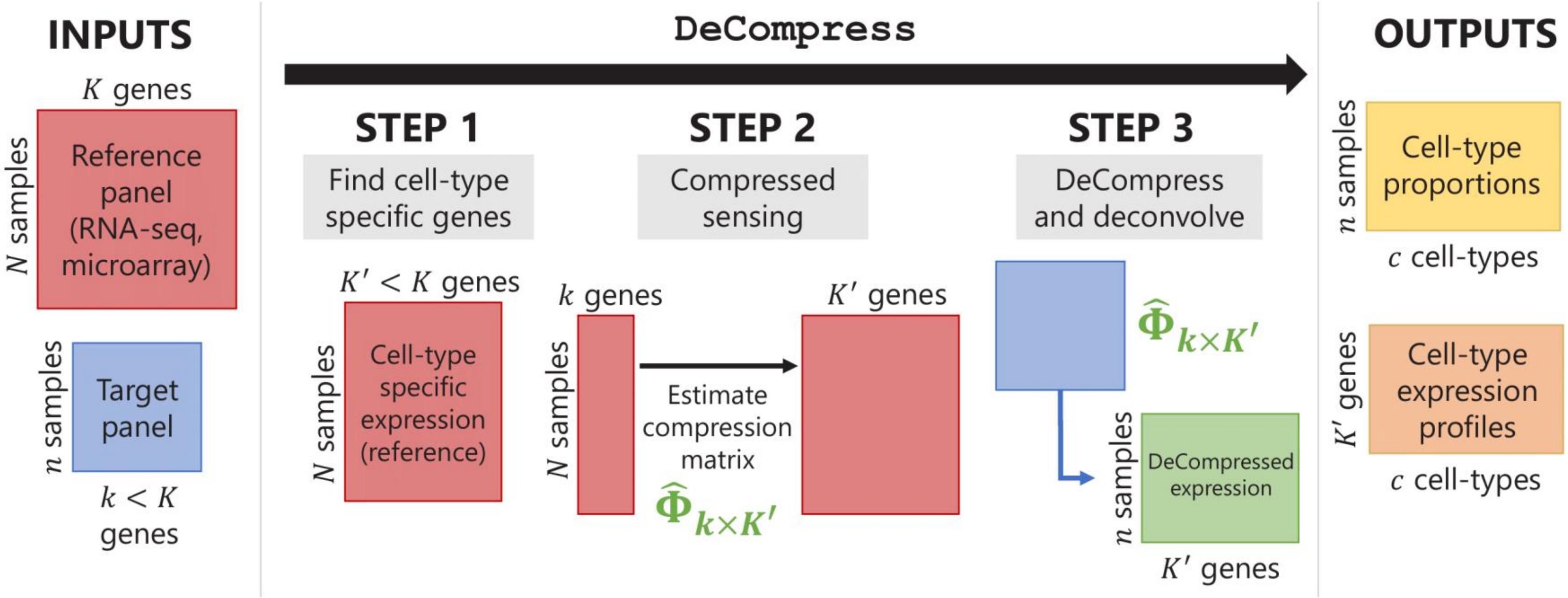

Figure 1: *Schematic for the DeCompress algorithm*. *DeCompress* takes in a *reference* RNA-seq or microarray matrix with N samples and K genes, and the *target* expression with n samples and $k < K$ genes. The algorithm has three general steps: (1) finding the $K' < K$ genes in the reference that are 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.

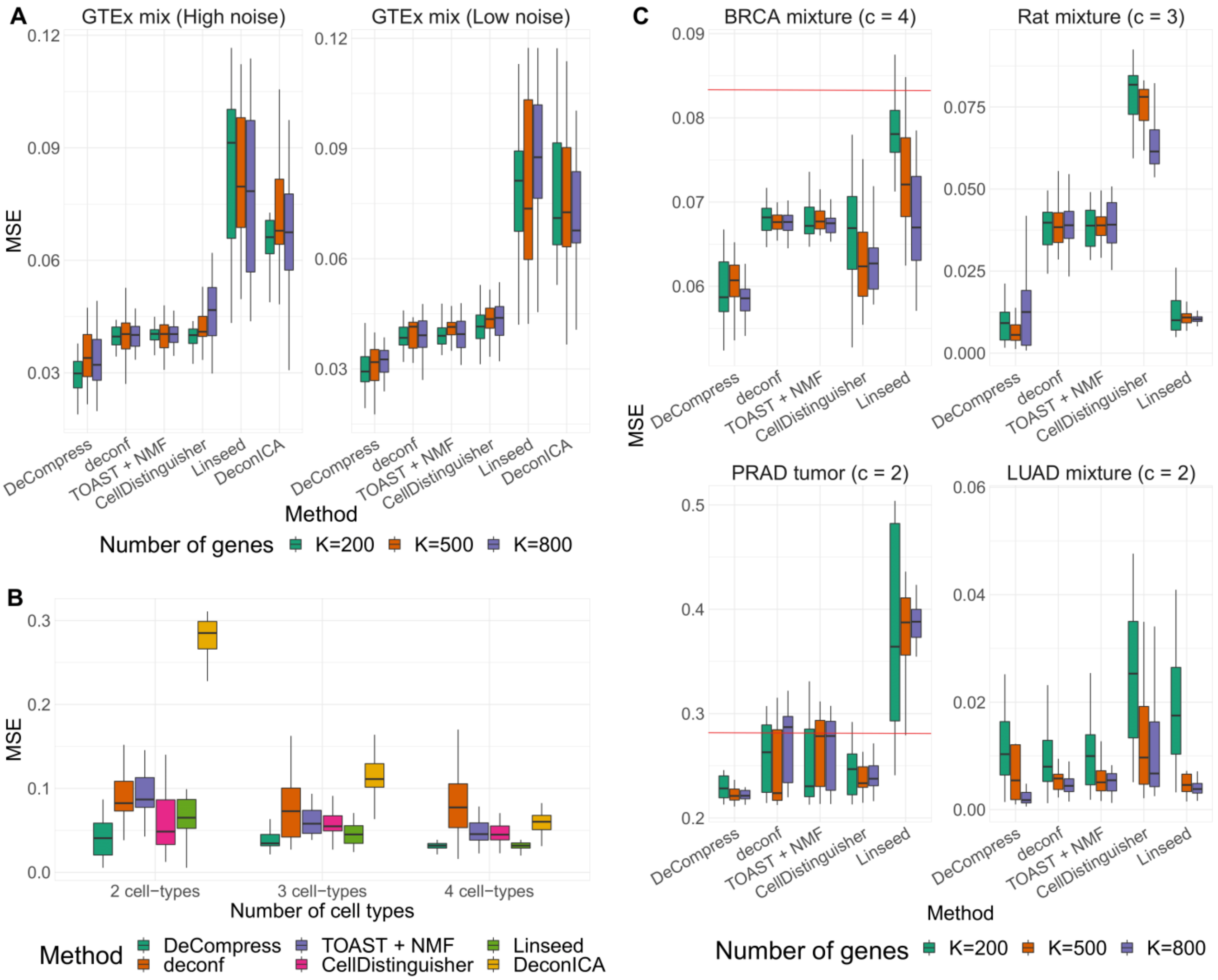

Figure 2: *Benchmarking results for in-silico GTEx mixing experiments and real data examples*. **(A)** Boxplots of mean square error (Y-axis) between true and estimated cell-type proportions in *in-silico* GTEx mixing experiments across various methods (X-axis), with 25 simulated datasets per number 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 number of genes in each simulated dataset. **(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 reference-free method tested. **(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) (23), rat brain, lung, and liver cell-line mixture (top-right) (11), prostate tumor samples (bottom-left) (58), and lung adenocarcinoma cell-line mixture (bottom-right) (59). 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.

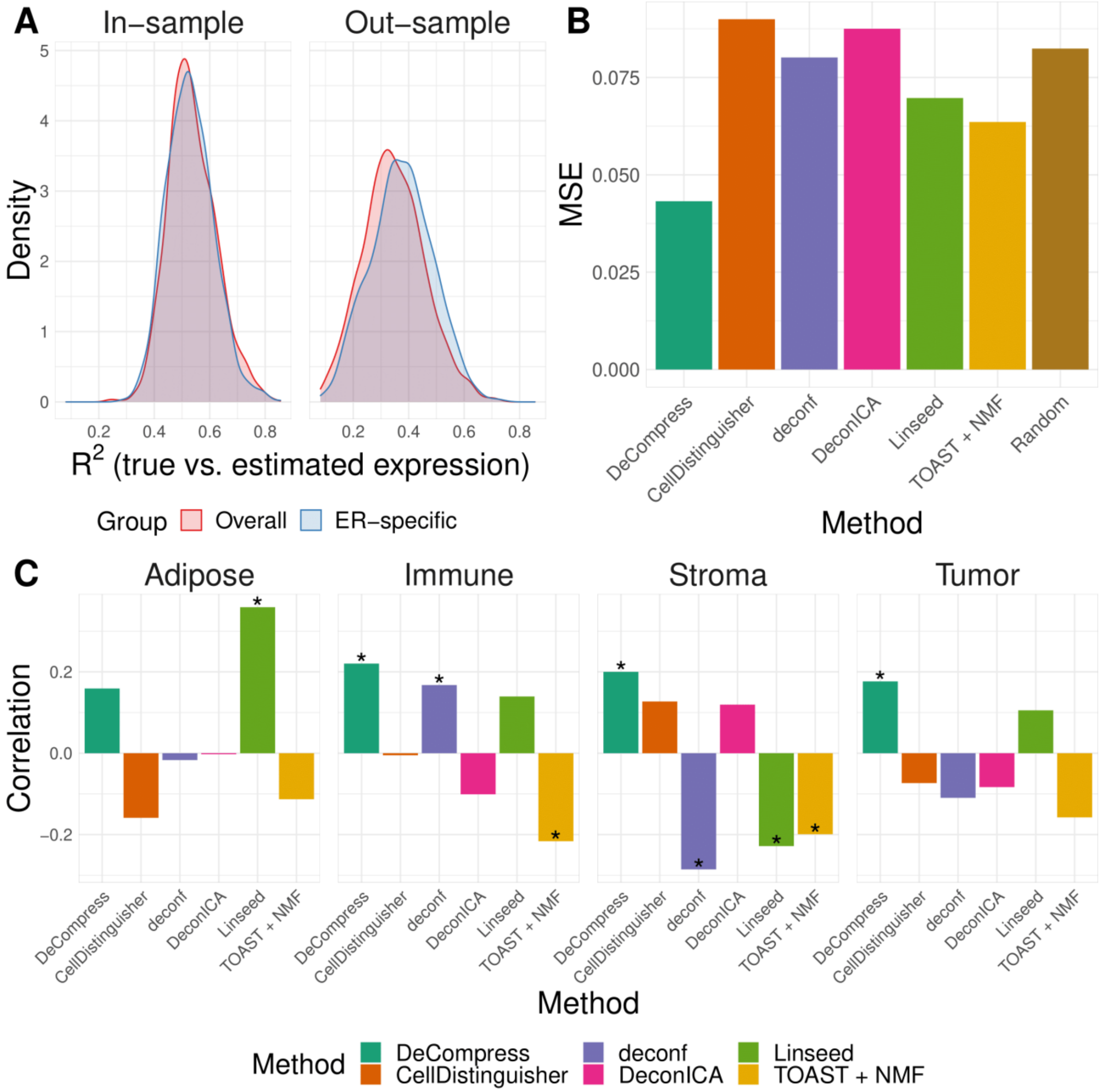

Figure 3: *Benchmarking results with Carolina Breast Cancer Study expression data*. **(A)** Kernel density plots of predicted adjusted R^2 per-sample in in-sample TCGA prediction (left) through cross-validation and out-sample prediction in CBCS (right), colored by overall and ER-specific models. **(B)** 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 $P < 0.05$.

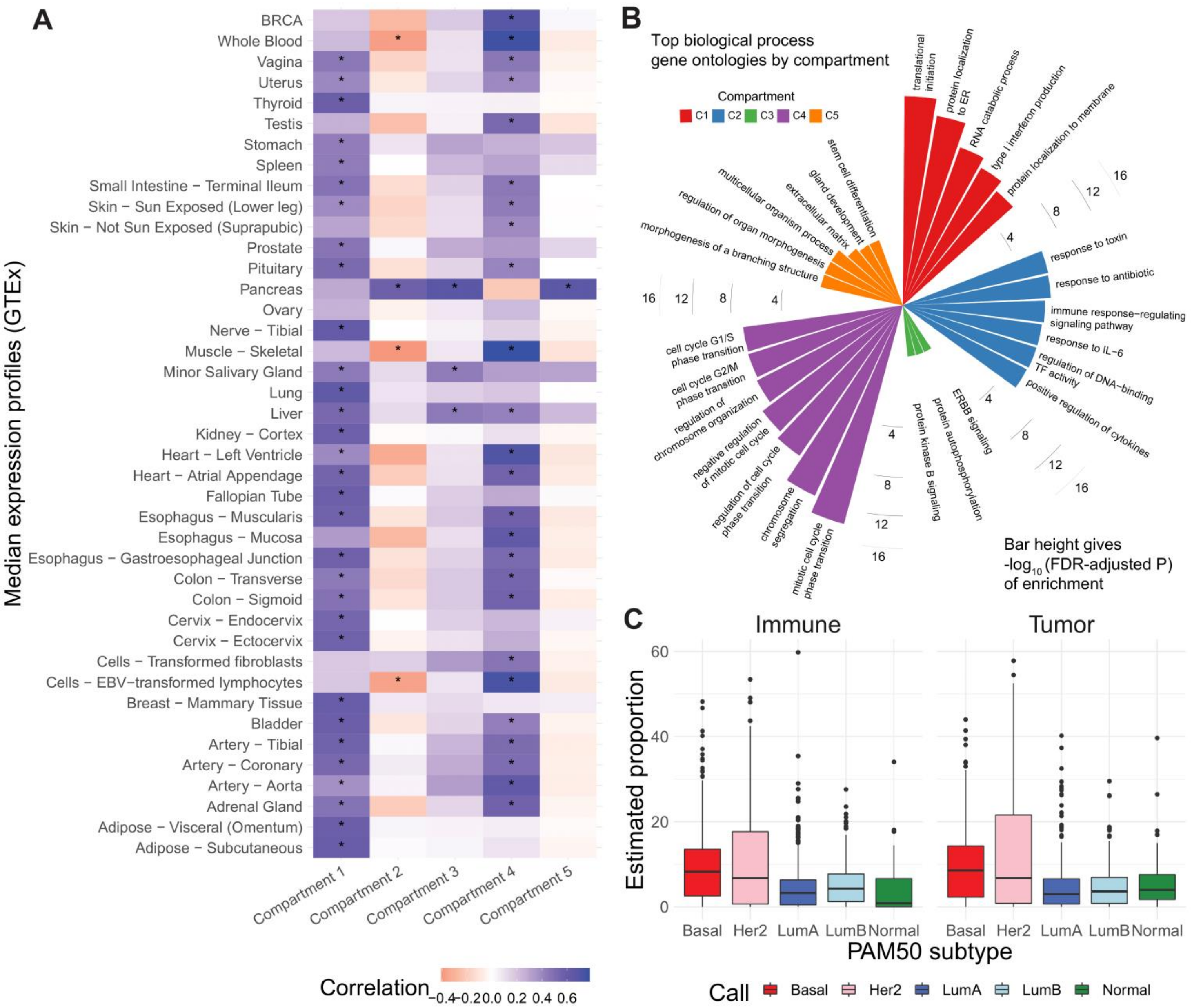

Figure 4: *Identification of Decompress-estimated compartments.* **(A)** Heatmap of Pearson correlations between compartment-specific gene signatures (X-axis) and GTEx median expression profiles and MCF7 single-cell profiles (Y-axis). Significant correlations at nominal $P < 0.01$ are indicated with an asterisk. **(B)** Bar plot 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 + C4 compartments, right) proportions (Y-axis) across PAM50 molecular subtypes (X-axis)

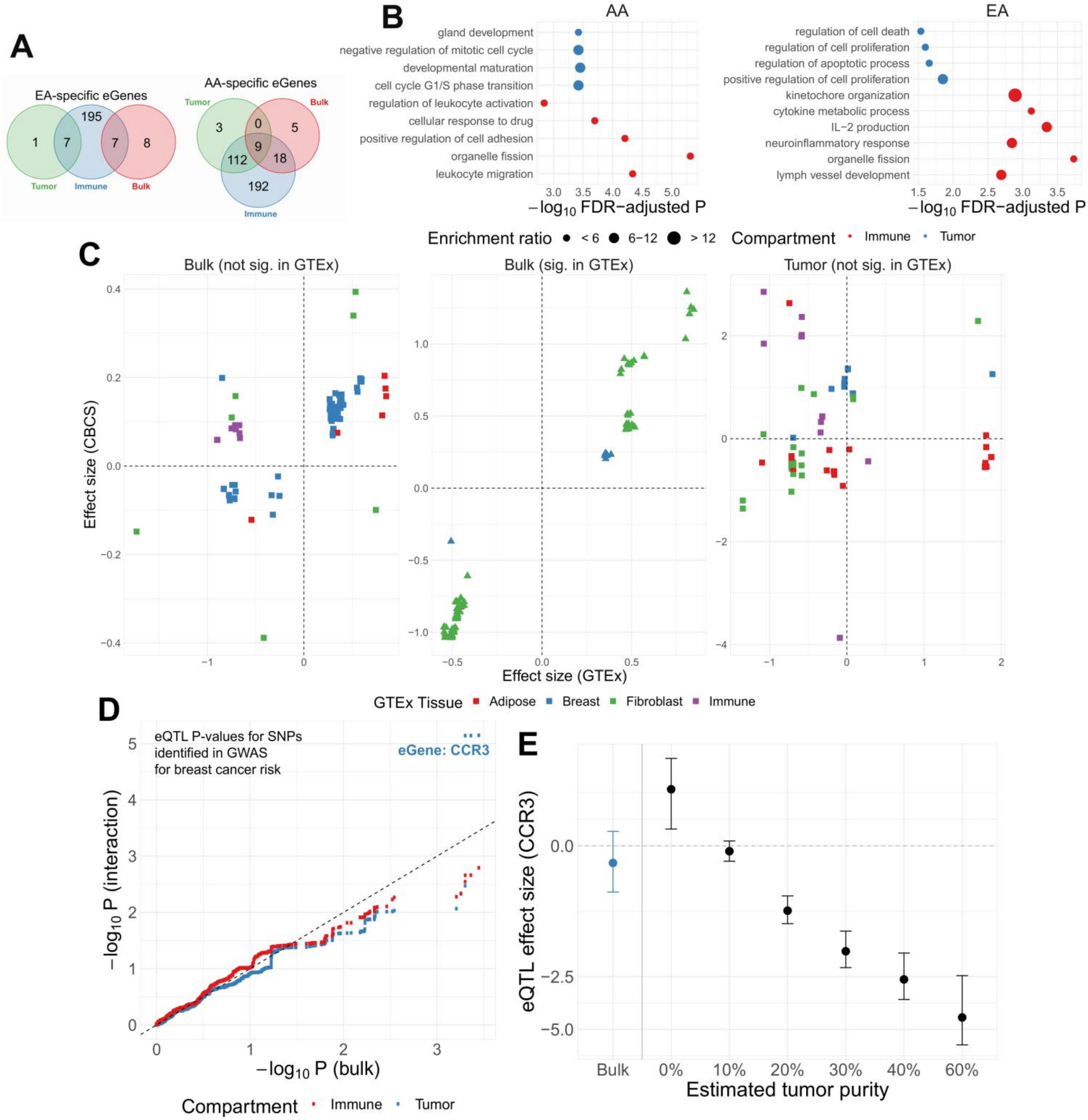

Figure 5: *Compartment-specific cis-eQTL mapping in the Carolina Breast Cancer Study.* **(A)** Venn diagram of bulk, tumor-, and immune-specific *cis*-eGenes identified European-ancestry (left) and African-ancestry samples (right) in CBCS. **(B)** Enrichment analysis of immune- (red) and tumor-specific (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 (86–88), 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.