Evaluating cancer cell line and patient-derived xenograft recapitulation of tumor and non-diseased tissue gene expression profiles

Avery S. Williams[1*], Elizabeth J. Wilk[1*], Jennifer L. Fisher[1], Brittany N. Lasseigne, Ph.D.[1#]

1. The Department of Cell, Developmental and Integrative Biology, Heersink School of Medicine, The University of Alabama at Birmingham, Birmingham, Alabama, USA

*Co-first author
#Corresponding author

Avery S. Williams: averywil@uab.edu
Elizabeth J. Wilk: lizzyr@uab.edu
Jennifer L. Fisher: jfisher7@uab.edu
Brittany N. Lasseigne: bnp0001@uab.edu

CRediT Statement:
ASW: Formal analysis, Data curation, Writing - original draft, Writing - review & editing, Visualization. **EJW:** Methodology, Data curation, Validation, Formal analysis, Writing - review & editing, Project administration **JLF:** Methodology, Validation, Writing - review & editing. **BNL:** Conceptualization, Methodology, Resources, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments: We would like to thank members of the Lasseigne Lab for their thoughtful feedback.

Data Availability Statement: This study uses publicly available data. All accession information is given in the Methods. Code (https://doi.org/10.5281/zenodo.7813872) and Docker image (https://doi.org/10.5281/zenodo.7813907) are available on Zenodo. The Docker image is also available on DockerHub (https://hub.docker.com/r/lizzyr/rstudio_modelselection).

Funding Statement: This work was supported in part by OD R03OD030604 (to BNL; supporting BNL, ASW, and JLF), R00HG009678 (to BNL; supporting BNL, EJW, AWS, JLF), an ACS-IRG (to BNL; supporting JLF), the UAB Lasseigne Lab Start-Up funds (to BNL; supporting BNL, ASW, EJW, and JLF), and the UAB Pilot Center for Precision Animal Modeling (C-PAM) 1U54OD030167 (supporting BNL and EJW). The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Conflict of Interest Disclosure: The authors have nothing to disclose.

Abstract

Preclinical models like cancer cell lines and patient-derived xenografts (PDXs) are vital for studying disease mechanisms and evaluating treatment options. It is essential that they accurately recapitulate the disease state of interest to generate results that will translate in the clinic. Prior studies have demonstrated that preclinical models do not recapitulate all biological aspects of human tissues, particularly with respect to the tissue of origin gene expression signatures. Therefore, it is critical to assess how well preclinical model gene expression profiles correlate with human cancer tissues to inform preclinical model selection and data analysis decisions. Here we evaluated how well preclinical models recapitulate human cancer and non-diseased tissue gene expression patterns with respect to the most variable genes, tumor purity, and tissue specificity by using publicly available gene expression profiles across multiple sources. We found that using the full gene set improves correlations between preclinical model and tissue global gene expression profiles, confirmed that GBM PDX global gene expression correlation to GBM tumor global gene expression outperforms GBM cell line to GBM tumor global gene expression correlations, and demonstrated that preclinical models in our study often failed to reproduce tissue-specific expression. While including additional genes for global gene expression comparison between cell lines and tissues decreases the overall correlation, it improves the relative rank between a cell line and its tissue of origin compared to other tissues. Our findings underscore the importance of using the full gene expression set measured when comparing preclinical models and tissues and confirm that tissue-specific patterns are better preserved in GBM PDX models than in GBM cell lines. Future studies can build on these findings to determine the specific pathways and gene sets recapitulated by particular preclinical models to facilitate model selection for a given study design or goal.

1. Introduction

In vitro preclinical models like cell lines and patient-derived xenografts (PDXs) are critical for advancing oncology studies but have limitations in how well they recapitulate human disease gene expression patterns and perform in the lab.[1] For example, cancer cell lines are generally less expensive to acquire and maintain, are more readily available from repositories, are more easily genetically manipulated, and have higher proliferation rates than PDXs.[2] However, because cancer cell lines are homogeneous, they fail to recapitulate disease and tissue heterogeneity and the microenvironment.[2] Additionally, over time, cancer cell lines are prone to mislabeling, cross-contamination, and contamination by chemicals (e.g., endotoxins, impure media) or biological entities (e.g., mycoplasma, bacteria, yeast).[3] In contrast, PDX models contain an *in vivo* microenvironment, are heterogenous, and have been shown generally to better match the patient's tumor profile than cell lines.[5] In regards to both model types, the process and method of culturing can influence cell gene expression patterns.[4] As resources like the CelloSaurus and groups like the International Cell Line Authentication Committee continue to identify preclinical models that, through authentication testing, have been determined to be likely misidentified or contaminated, studies must continue to re-evaluate which preclinical models should be used in research.[6]

However, despite demonstrating efficacy in preclinical model systems like cancer cell lines and PDXs, most novel cancer drugs fail in clinical trials. Of all oncology drugs entering Phase I clinical trials, only 5% are approved. The main bottleneck of drug failure is in Phase II of clinical trials, suggesting that preclinical model systems are effective in predicting human drug toxicity (Phase I), but fail to accurately recapitulate many diseases for the sake of drug testing (Phase II).[7,8] Further, previous studies suggest that cancer cell lines often do not accurately represent the transcriptomic profile of their tissue or disease of origin, or that they may be better surrogates for particular cases, such as primary versus secondary tumors, or for a particular molecular pathway.[9,10] While disease and genetic background heterogeneity in humans and system or microenvironmental effects contribute to this, a better understanding of how cancer cell line or PDX model gene expression profiles recapitulate the human tumor or non-diseased tissue being modeled is critical.

Here we evaluated how restricting to the most variable genes impacts how well cancer cell lines or glioblastoma (GBM) PDX global gene expression profiles correlate with human tissue gene expression profiles. We used gene expression profiles from the Cancer Cell Line Encyclopedia (CCLE, n = 943),[11] the Human Protein Atlas (HPA; n = 194),[12] a publicly available osteosarcoma and bone sample set (SRP090849, n = 212),[13] GBM PDXs from the Mayo Clinic Brain Tumor PDX National Resource (n = 65),[14] The Cancer Genome Atlas (TCGA, n = 10,098)[15], and Genotype-Tissue Expression project (GTEx, n = 17,510)[16]. We further investigated whether the inclusion or exclusion of tumor purity-correlated genes impacts those correlations and the ability of the preclinical models to capture tissue-specific gene expression signatures. Our findings illustrate that while GBM PDX samples outperformed GBM cancer cell lines concerning both tissue and disease context of origin, cell lines better model tissues of different origins, further underscoring that PDX models more faithfully capture tissue and disease-specific gene expression profiles. The degree to which either model captured the gene expression profile of the same tissue of origin, however, varied as expected depending on the gene set included when gene subsets were used. We find that it is generally best to assess global profiles, but if considering based on a smaller subset, like a specific pathway, it may be best to focus more consideration on this gene set of interest for informed decision making.

2. Methods

2.1 Data Collection

All analyses were conducted using the R programming language (version 4.2.2) within the RStudio interface (version 2022.02.01+461). We used the Recount3 R package to access publicly available RNA-seq gene counts (version 1.4.0)[17] GENCODE version 26 annotations (n=63,856 genes). All included data were originally generated using either an Illumina Genome Analyzer or HiSeq instrument. Data accessed through Recount3 was previously preprocessed with the Monorail pipeline, which facilitates comparison between cohorts and minimizes batch differences.[17] For this study, we used data from human tumor and non-diseased tissues, cell lines, and patient-derived xenograft (PDX) mouse models. Human tissue profiles were from The Cancer Genome Atlas (TCGA, n = 10,098 across 33 cancer types)[15], a publicly available osteosarcoma tissue data set (SRP090849; osteosarcoma n = 188, non-diseased bone n = 9)[13] to provide a bone tissue comparison for bone tissue-derived cell lines in subsequent analyses, the Genotype-Tissue Expression (GTEx) Project (n = 17,510 across 30 tissue types),[16] and the Human Protein Atlas (HPA, n = 171).[18] We sourced cell line gene expression data from the Cancer Cell Line Encyclopedia (CCLE, n = 1,004),[11] the HPA (n = 23),[18] and SRP090849

(bone-derived cell lines, n = 15).[13] In total, this data included 64 different cancers derived from 41 tissues. Finally, we sourced GBM PDX gene expression data from the Mayo Clinic's Glioblastoma PDX National Resource (n = 65).[14]

2.2 Data Curation

We excluded some samples from subsequent analyses based on mislabeling or possible contamination (Supplemental Table 1). To avoid possibly confounding our analyses, we excluded the GTEx samples marked as bone marrow as they are derived from a leukemia cell line. Five HPA non-diseased tissue samples (ERR315357, ERR315383, ERR315386, ERR315403, and ERR315438) were relabeled as well for metadata consistency. We selected primary tumor samples within the TCGA dataset. We also confirmed that each CCLE cell line had not been flagged as having evidence of contamination or misidentification in The Cellosaurus, a Swiss Institute of Bioinformatics resource for biomedical cell lines.[6] From this assessment, we removed 66 of the CCLE cell lines from our analysis,[6] as well as one additional sample (SRR8615727) which was missing significant portions of its metadata, including which cell line it was derived from. We also relabeled 4 HPA cell lines (SRR629581, SRR629582, SRR4098609, and SRR4098610) for metadata consistency in subsequent analyses. We removed one PDX sample (SRR9294079) as it is IDH-mutant and no longer classified as GBM by the World Health Organization as of 2021.[19]

2.3 Data Categorization

In order to match preclinical models to corresponding tissues for cross-disease and cross-tissue comparisons we assigned preclinical models by their putative tissue of origin (Supplemental Tables 2 and 3). For example, all leukemia-derived cell lines regardless of subtype were grouped as "Blood Cancers". With these adjustments, all cell lines were identified as part of a parent tissue group. However, some supplementary tissues (i.e., those without corresponding preclinical models in our analyses) were also included to further evaluate model performance in contexts separate from their origins (e.g., GBM cell lines and PDXs evaluated for their correlation to other non-brain tumor tissue types).

2.4 Data Normalization

From raw counts, we calculated transcripts per million (TPM) using the gene lengths available in the Recount3-downloaded ranged-summarized experiment (RSE) object with merged Ensembl/Havana GENCODE version 26 annotations[20] via the getTPM() function of the recount R package (version 1.20.0) to facilitate a comparison of the proportion of total reads mapped to a gene across samples.[21]

2.5 Gene Subsets

Previous studies evaluating gene expression differences between cell lines and tissues by restricting to the 5,000 most variable genes[9,10,22] suggest this subset represents the most biologically informative signal.[10] This restriction of the global gene set may be useful for reducing computational memory requirements and run times. We evaluated the difference in variability across samples by gene set size with principal component analysis (PCA) with the mixOmics R package (version 6.18.1) pca() function to determine principal components,[23] and then used the PLSDAbatch package (version 0.2.1) Scatter_Density() function to generate PCA and density rugplots to investigate the variance between sample origin types (i.e., cell line, PDX, non-disease tissue, tumor tissue; between brain-derived tissues and GBM-derived models in Figure 2B; between all samples of all sources in Supplemental Figure 1A), as well as with respect to other variables that may impact variance (i.e., read length, tissue type, sample source, sex; Supplemental Figure 1B-E).[24]
We identified tumor purity-correlated genes based on TCGA tumor gene expression significantly correlated with ABSOLUTE-generated tumor purity. We also defined tissue-specific gene expression from GTEx using the Harminozome (accessed November 2022).[16,25,26] We chose this resource due to its breadth of tissue-specific gene sets available across various tissues.

2.6 Correlation & Pathway Analysis

We used Spearman's rank correlation with the cor() function of the stats base R package[27] to evaluate model performance in recapitulating tissue gene expression patterns (Supplemental Table 4). We visualized the correlation distributions with the hist() function in the graphics base R package (GBM cell lines correlated to brain tumor tissue: Figure 5A; all other groups: Supplemental Figure 2).[27] We computed the significance between correlation performance using the Wilcoxon rank sum test using the rstatix (version 0.7.0) and coin

(version 1.4-2) R packages.[28–30] We studied their performance within many sub-groups. To investigate pathway enrichment of genes highly and lowly correlated and anticorrelated between brain-derived models and brain tumor and non-diseased tissue used the gprofiler2 R package (GOSt results for these gene groups are in Supplemental Table 5).[31] We generated visualization of correlations, pathways, and all other figures using the ggplot2 (version 3.3.5), ggpubr (version 0.4.0), ComplexHeatmap (version 2.10.0), mixOmics (version 6.18.1), and PLSDAbatch (version 0.2.1) R packages and compiled with BioRender.[23,32–36]

3. Results

3.1 The correlation between cancer cell line and tissue of origin gene expression when restricting to the most variable genes is reduced and less specific compared to all genes

Previous studies evaluating gene expression differences between cell lines and tissues by restricting to the 5,000 most variable genes[9,10,22] suggest this subset represents the most biologically informative signal.[10] We evaluated how the gene subset size impacts the correlation between cancer cell line and tumor tissue gene expression profiles originating from the same tissue type (i.e., with matched origin tissue). From CCLE,[11] HPA,[12] and SRP090849[13] cancer cell line gene expression profiles, we calculated their global Spearman correlation to the corresponding non-diseased tissue (GTEx,[16] HPA, SRP090849) or tumor tissue profiles (TCGA,[15] SRP090849). For this analysis, we determined the 100, 1,000, 5,000, and 10,000 most variable genes across all included TCGA tumor tissue samples or GTEx non-diseased tissue samples. We then evaluated how restricting each of those subsets of genes, impacted the correlation between cancer cell lines and matched tissue types compared to the global profile (n = 63,856 genes) (Figure 1A). We also evaluated including or excluding previously identified high tumor purity-correlated genes (i.e., TCGA tumor gene expression significantly correlated with ABSOLUTE-generated tumor purity).[10,37,38]

Compared to the median correlation between the cancer cell lines and their matched tumor tissue in the global gene profile (rho = 0.82) (Supplemental Table 4), the 100 most variable gene subset based on TCGA variance, including tumor purity-associated genes, had the highest median correlation value (rho = 0.73). However, the 5,000 most variable gene subset based on TCGA variation, excluding tumor purity-associated genes, had the lowest correlation value (rho = 0.47). The median correlation rho between cancer cell lines and matched tumor tissue in the 5,000 most variable genes excluding tumor purity-correlated genes, was 0.05 greater in GTEx-derived variation than in TCGA. Therefore, we found that restricting to the most variable genes reduces the cell line-to-tissue of origin correlation and that cell lines generally correlated more to the same tissue they were derived from when using the most variable genes based on GTEx samples. Further, while excluding purity-associated genes improved the correlation for gene sets based on the most variable GTEx genes, it did not for the most variable genes across TCGA samples (Figure 1A).

We also evaluated the percentage of cancer cell lines that correlated most highly with tumor tissues from the same origin compared to all tissues included in this study (i.e., both cancer- and non-disease-derived tissues). Intriguingly, even though cancer cell line model gene expression profiles were most strongly correlated with their tissue of origin when only the top 100 variable genes were included (Figure 1A), very few correlated most highly with their tissue of origin when compared to any tissue (i.e., 0% and 0.1% when the variable genes were selected based on TCGA and tumor purity-correlated genes were included and excluded, respectively; 1% when the variable genes were selected based on GTEx, regardless of the inclusion of tumor purity-correlated genes). Further, while 20% of cell lines correlated most highly to tumor tissues of the same origin with the global gene expression profile, that percentage was higher in all gene subsets greater than 100 except for in the 1,000 genes based on TCGA variation with tumor purity-correlated genes included (Figure 1B).

Next, we investigated if there was a difference between tissues in terms of how well the gene expression of cancer cell lines derived from a particular tumor tissue correlated with gene expression from the same tumor tissue type depending on the gene subset used (Figure 1C). We observed that some tissue types had less variation in their global gene expression correlation to cancer cell lines from that same tissue type regardless of gene subset size, GTEx or TCGA tissue gene expression most variable gene subsetting, or purity gene inclusion. For example, the difference between the median correlation of all of the most variable gene set correlations (i.e., all subset sizes, tumor purity gene inclusion/exclusion) by GTEx or TCGA variable gene subsetting was ~0.01 for soft tissue cancer-derived tissue (i.e., soft tissue/connective tissue cancer, sarcomas; rho values ranged between 0.53 and 0.67) and ~0.02 for bone and brain cancer-derived tissue (rho values ranged between 0.64 and 0.79 and 0.53 and 0.71, respectively). However, the difference in rho was at least 0.1 in 14 of the 22 tumor tissue types we analyzed. This suggests that for many tissues, how the most

variable genes are defined for the purpose of comparing cancer cell lines to tumor tissues from the same tissue of origin impacts the results. Based on these findings, for subsequent analyses, we included all genes.

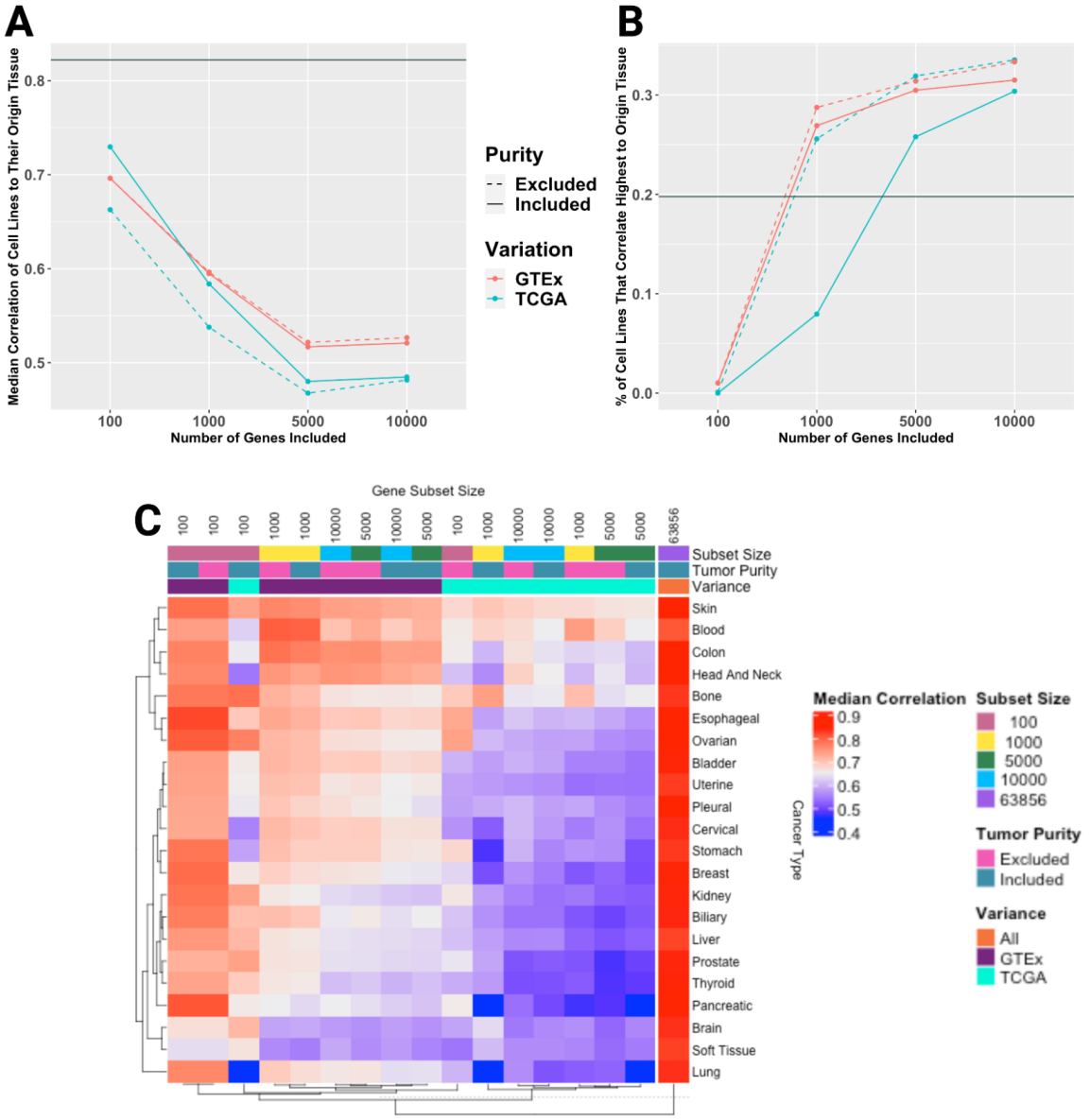


Figure 1. A. Line graph of the difference in median correlation of cell lines to their matched origin tumor tissue based on the set size of most variable genes, whether that variation is based on GTEx (red) or TCGA (blue) most variable gene subsetting, and whether tumor purity-correlated genes are included (solid line) or excluded (dashed line). Global correlation is represented as a gray solid line. **B**. Percent of cell lines that correlate highest with their matched origin tumor tissue out of all tumor and non-diseased tissues included based on the set size of the most variable genes, whether that variation is based on GTEx (red) or TCGA (blue) tissue variation, and whether tumor purity-correlated genes are included (solid line) or excluded (dashed line). Global correlation is represented as a gray solid line. **C**. Heatmap of cell line median correlation to their matched origin tumor tissue based on the set size of most variable genes, whether that variation is based on GTEx or TCGA tissue variation, and whether tumor purity-correlated genes are included.

3.2 Global gene correlations between GBM models and brain tissues underscore GBM PDX gene expression specificity and GBM cell line gene expression ambiguity

We next correlated GBM cell line and PDX gene expression profiles to brain cancer and non-diseased brain tissue profiles. Principal component analysis (PCA) demonstrated that principal component 1 (39% of the variance) and principal component 2 (15% of the variance) separated non-diseased brain tissue from GBM PDXs, GBM cell lines, and brain tumor tissue (Figure 2A). To evaluate how well the GBM PDXs and cell lines recapitulated GBM-specific, brain tumor-specific (i.e., GBM and lower-grade glioma, LGG), and brain non-diseased tissue gene expression profiles, we evaluated how well each correlated (Figure 2B). We found that GBM PDX models best correlated to brain tumor tissues, and both GBM cell lines and PDX models correlated better to tumor tissue than non-diseased tissue. We found that, though within a modest correlation range (all samples of both models' median correlations to either tissue type were between Spearman rho = 0.77-0.89) PDXs significantly correlated more to brain tissue, both tumor and non-diseased, than cell lines ($p < 0.0001$ and $W = 410$ when correlated to brain tumor tissue, $p < 0.0001$ and $W = 455$ when correlated to brain non-diseased tissue, Figure 2C). Finally, we asked how well GBM PDX and cell line global gene expression profiles correlated to other tumor and non-diseased tissues from TCGA and GTEx because other studies have suggested cell line gene expression profiles may better resemble other tumor types than the one from which they were originally derived.[39] Here, we found that the GBM cell lines correlated significantly better to non-brain tissues than the GBM PDX gene expression profiles did (Bonferroni-corrected Wilcox p-value < 0.05, Figure 2D) in every comparison with the exception of two tumor tissues (i.e., adrenal and eye), and five non-diseased tissues (i.e., pituitary, colon, nerve, testis, and ovary). This suggests that the GBM PDXs better recapitulate their tissue of origin global gene expression patterns and that the GBM cell line global gene expression profiles are less tissue-specific and therefore correlate better with both cancer or non-diseased tissues different from their own origin than the PDX models.

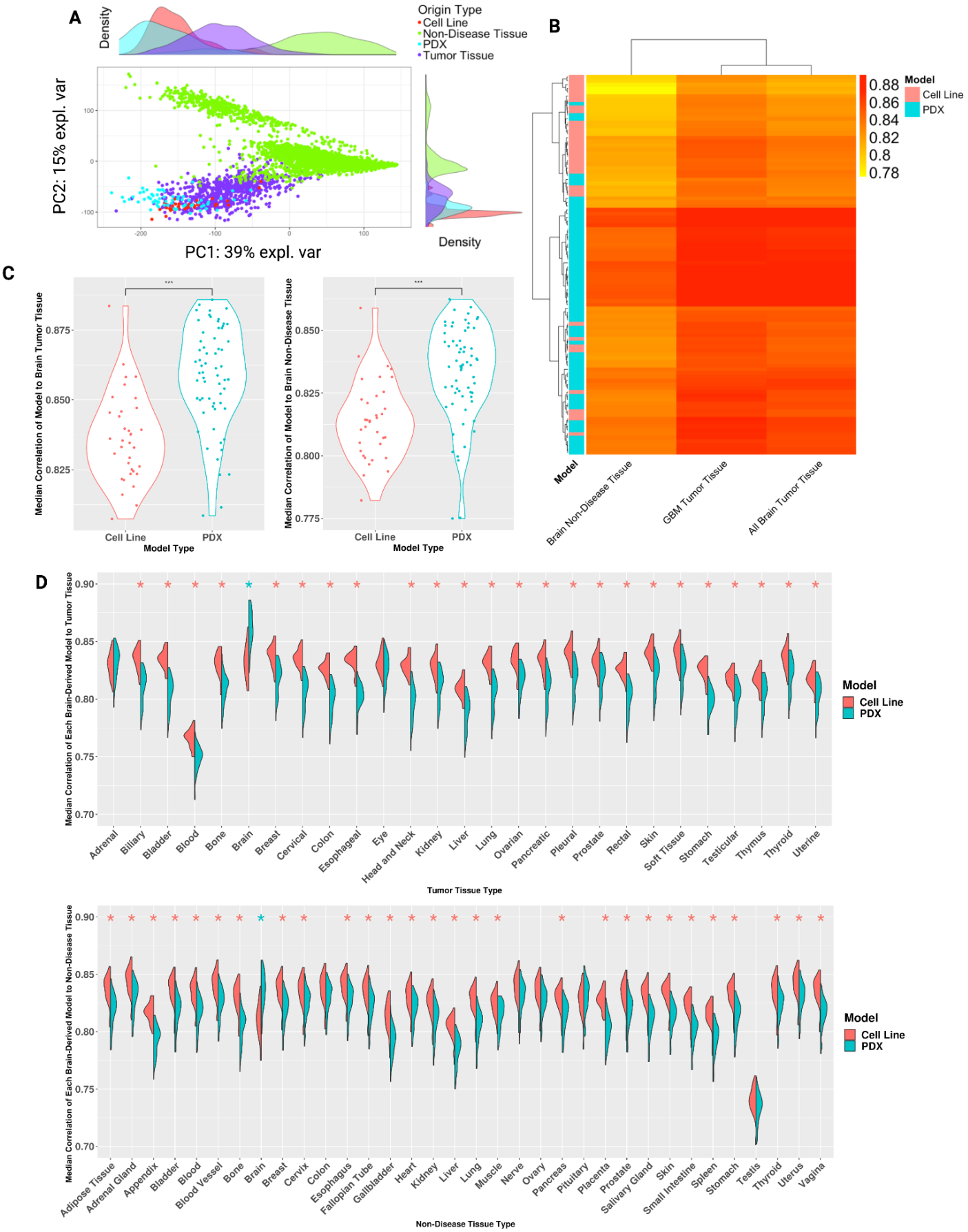

A
Density
Origin Type
Cell Line
Non-Disease Tissue
PDX
Tumor Tissue
PC2: 15% expl. var
PC1: 39% expl. var
Density
B
Model
Cell Line
PDX
0.88
0.86
0.84
0.82
0.8
0.78
Model
Brain Non-Disease Tissue
GBM Tumor Tissue
All Brain Tumor Tissue
C
Median Correlation of Model to Brain Tumor Tissue
Cell Line
PDX
Model Type
Median Correlation of Model to Brain Non-Disease Tissue
Cell Line
PDX
Model Type
D
Median Correlation of Each Brain-Derived Model to Tumor Tissue
Adrenal
Biliary
Bladder
Blood
Bone
Brain
Breast
Cervical
Colon
Esophageal
Eye
Head and Neck
Kidney
Liver
Lung
Ovarian
Pancreatic
Pleural
Prostate
Rectal
Skin
Soft Tissue
Stomach
Testicular
Thymus
Thyroid
Uterine
Tumor Tissue Type
Model
Cell Line
PDX
Median Correlation of Each Brain-Derived Model to Non-Disease Tissue
Adipose Tissue
Adrenal Gland
Appendix
Bladder
Blood
Blood Vessel
Bone
Brain
Breast
Cervix
Colon
Esophagus
Fallopian Tube
Gallbladder
Heart
Kidney
Liver
Lung
Muscle
Nerve
Ovary
Pancreas
Pituitary
Placenta
Prostate
Salivary Gland
Skin
Small Intestine
Spleen
Stomach
Testis
Thyroid
Uterus
Vagina
Non-Disease Tissue Type
Model
Cell Line
PDX

Figure 2. A. Principal component analysis scatterplot with density rug plot of gene expression profiles from GBM cell lines (red), GBM PDX models (teal), non-diseased brain tissue (green), and GBM and LGG tissue samples (purple). **B**. Heatmap of median performance of each GBM cell line and PDX sample to brain non-disease, GBM tumor, and all available brain tumor (GBM, LGG) tissue. **C**. Violin plot showing median correlation of each GBM cell line and PDX sample to brain tumor (left) and non-diseased (right) tissue split by model type. Significant Wilcox p-value ($p < 0.05$) indicated by asterisks. **D**. Split violin plot of correlation of each GBM cell line and PDX sample to the tumor (top) and non-diseased (bottom) tissues. Wilcoxon rank sum test with Bonferroni procedure for multiple hypothesis correction, p-adjusted values < 0.05 indicated by asteriks and colored by model type with significantly greater median correlation.

3.3 Models often fail to recapitulate origin tissue tissue-specific gene expression

Next, we evaluated how well cancer cell lines recapitulated non-diseased and tumor tissue gene expression patterns in those tissues' tissue-specific genes. Here we defined genes as having tissue-specific gene expression based on prior work identifying genes with either high or low gene expression in each tissue relative to all other tissues in the GTEx dataset.[16,26] For each tissue, we calculated the correlation between the gene expression of that tissue's tissue-specific genes in cancer cell lines derived from that tissue to the non-diseased gene expression profiles of that tissue and cancer gene expression profiles of that tissue (Figure 3A). These analyses show that the tissue-specific gene expression correlation is higher between cancer cell lines and tumor tissues than between cancer cell lines and non-diseased tissue for that tissue (Bonferroni-corrected Wilcoxon p-value between the correlation of cancer cell line to matched tumor tissue tissue-specific genes and cancer cell line to matched non-diseased tissue tissue-specific genes $p < 0.05$, for 14 out of 16 tissues).

We next studied the performance of the 100 cancer cell lines with the highest median global gene expression correlation to all studied tissue and disease contexts with respect to a given tissue's tissue-specific genes. When visualized as a heatmap of the cancer cell lines' correlation to each tissue and disease context for that tissue's tissue-specific genes (Figure 3B), we found there were subclades of some cancer cell lines from the same tissue of origin clustering together (e.g., some brain-derived and liver-derived cell lines), but this was not widespread. Some tissues clustered by tumor or non-diseased state, though again not all (e.g., within the first split of the dendrogram, eight of the ten tissues were tumor-derived, but the other two were non-diseased prostate and bladder). Though cell lines included here were derived from 25 different tissues, all of the top 100 cancer cell lines, regardless of origin, correlated highly (rho > 0.72) with kidney, breast, skin, thyroid, ovary, pancreas, uterus, and bladder tumor tissue as well as prostate and bladder non-diseased tissue. Further, all included cancer cell lines correlated less with brain, lung, blood, and stomach non-diseased tissue (rho < 0.66). As cancer cell lines were more correlated on average to tumor gene expression profiles, this further underscores that tissue status (i.e., from a tumor or non-diseased tissue) is a critical driver of how well a cancer cell line's global expression profile correlates with that tissue. These tissues had low variation in the strength of correlation across cell lines (e.g., bladder non-diseased tissue had a difference of 0.04 between the lowest and highest correlation rhos); however, a few tissues, such as colon tumor and non-diseased esophagus, had higher variation in how well they correlated with the 100 tested cancer cell lines (i.e., the difference between lowest and highest correlation rho between cell line and tissue for colon tumor and non-diseased esophagus were 0.25 and 0.22, respectively). Further, 18 of 35 tissues had a difference between their lowest and highest rho correlation value of at least 0.1. Colon tumor samples predominantly correlated highly with colon-derived cell lines; all 6 of the colon-derived cell lines correlated highest to colon tumor tissue (rho > 0.7). However, other tissues, including pancreatic tumor tissue, had less specificity and correlated highly with cell lines across origins. Only 3 of the 6 pancreatic tumor-derived cell lines included were in the top 15 rho values of pancreas tumor tissue's correlation to cell lines, while the lowest correlation between pancreas tumor tissue and pancreas tumor-derived cell line was the 36th highest rho value at rho = ~0.81, and no correlation to pancreas tumor tissue out of any of the 100 cell lines was lower than rho = ~0.75. Ovary tumor tissue correlated highly with all of the top 100 cell lines (the lowest rho was = ~0.82), and with minimal specificity to cell lines of the same origin; while 10 of the 19 ovary cell lines were in the top 15 rho values of ovarian tumor tissue's correlation to cell lines, one of those 19 was ranked 95th for its strength of correlation (rho = ~0.825). This supports that some tissues are more broadly recapitulated by cancer cell lines regardless of the cell lines' origin, while others have greater specificity, whether that be to cell lines derived from the same matched tissue or a specific set of tissues.

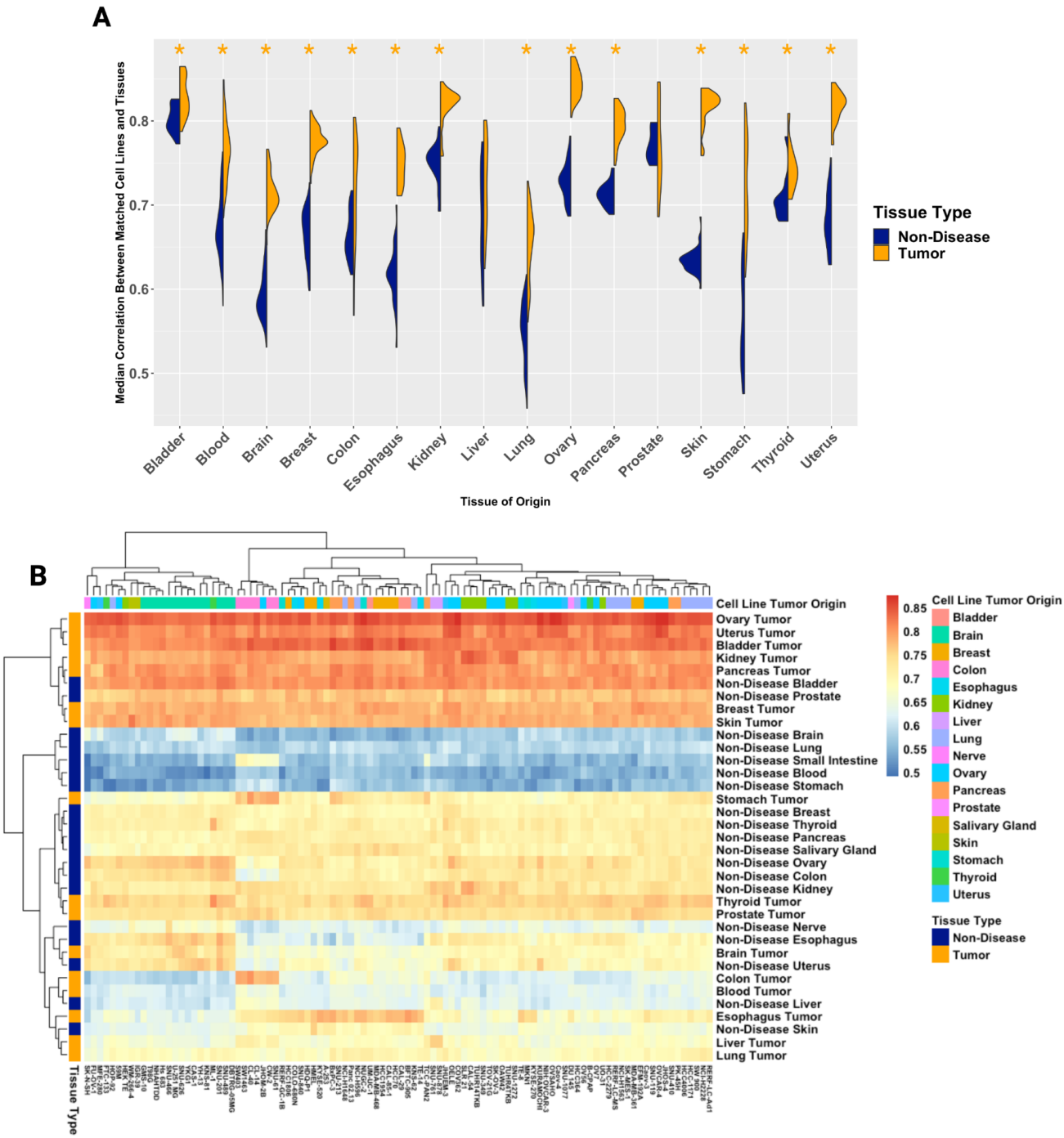


Figure 3. A. Violin plots of the median correlation of each cell line sample to its matched tumor (orange) and non-diseased (dark blue) tissue in each matched tissue's specific genes. Wilcoxon rank sum test with Bonferroni procedure for multiple hypothesis correction, p-adjusted values < 0.05 indicated by asterisks and colored by tissue type with significantly greater median correlation. **B**. Heatmap of correlation of the top 100 most generally highly correlated cell lines to each tumor and non-diseased tissue.

We again focused on GBM cell lines and PDXs and found that while GBM PDXs were more correlated to both brain-derived tumor and non-diseased brain tissue than GBM cell lines in the global profile (Figure 2C) when restricted to tissue-specific genes, this difference was even greater (median Spearman correlation rho of PDX gene expression to tumor gene expression and to non-diseased brain tissue gene expression was 0.126

and 0.155 greater than to the GBM cell lines, respectively) (Figure 4A). We then assessed how well the GBM PDXs and GBM cell lines recapitulated the tissue-specific gene expression of other tissues (correlation to tumor tissue: Figure 4B; correlation to control tissue: Supplemental Figure 2). We again found that brain-derived cancer cell lines better recapitulated gene expression of other tissues compared to GBM PDX models, but there was more variability here than compared to the full global gene expression profile. The range of median correlation to each tumor tissue for GBM cancer cell lines was 0.54-0.85 and for GBM PDXs was 0.46-0.88 using tissue-specific genes (Figure 4B), while in the full gene set median correlation for cancer cell lines was 0.75-0.86 and PDXs was 0.71-0.89 (Figure 2D).

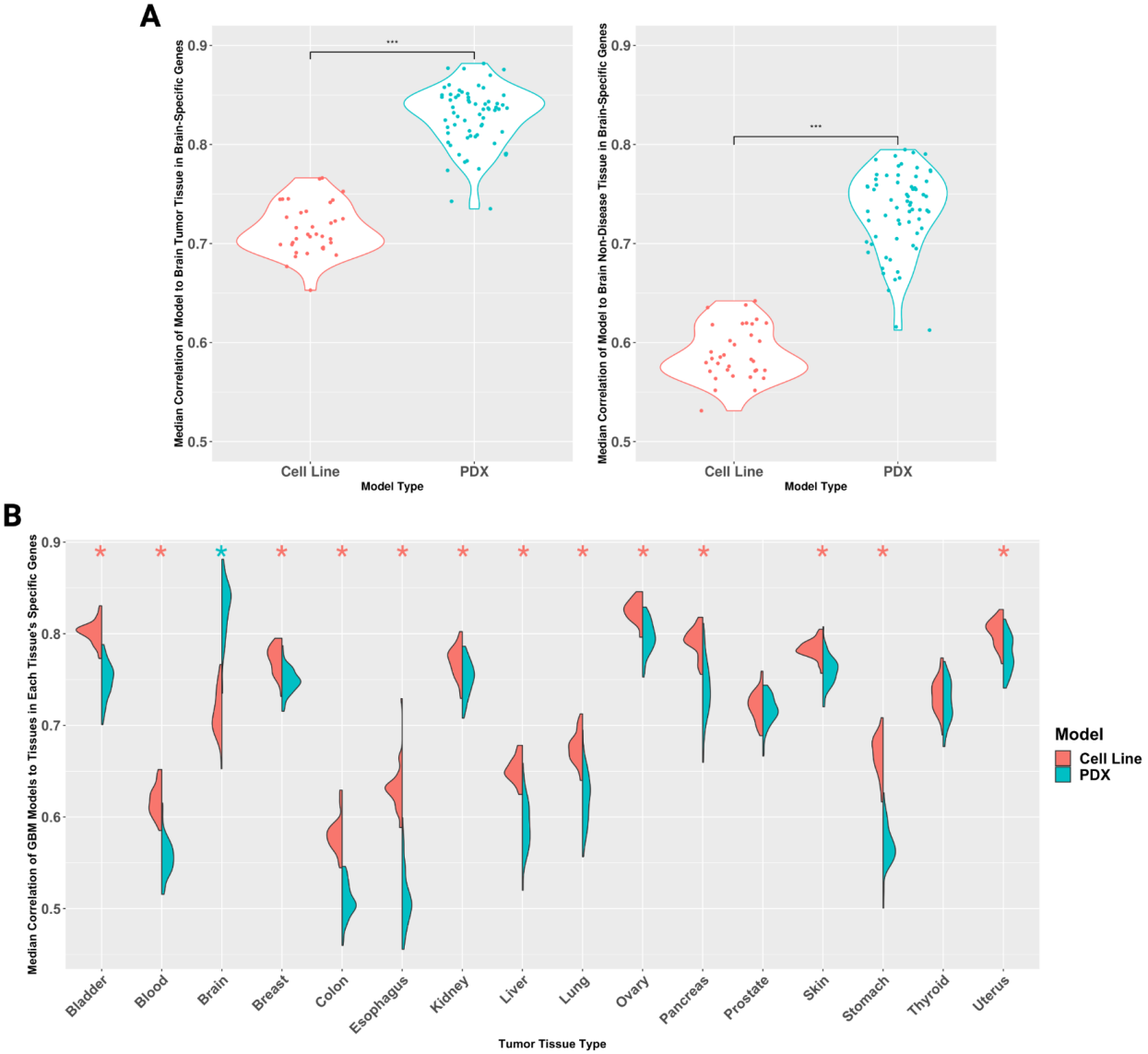


Figure 4. A. Violin plots with Wilcoxon test showing the median correlation of each GBM cancer cell line and PDX sample to brain tumor and non-diseased brain tissue with brain-specific genes. Significant Wilcox p-value ($p < 0.05$) indicated by asterisks. **B**. Split violin plots of the correlation of each GBM cancer cell line and PDX sample to each tumor tissue type by that tissue's specific genes. Wilcoxon rank sum test with Bonferroni procedure for multiple hypothesis correction, p-adjusted values < 0.05 indicated by asteriks and colored by model type with significantly greater median correlation.

3.4 Gene ontology analysis of genes significantly correlated and anti-correlated between brain-derived models and brain tissue gene expression

After determining which genes were highly or lowly correlated or anticorrelated between GBM models and brain tumor or non-disease tissue gene expression profiles (Figure 5A; Supplemental Figure 3), we evaluated if pathways or biological terms were overrepresented within these sets using g:Profiler (Supplemental Figure 4; Supplemental Table 5).[40] We found that most genes significantly correlated between GBM cell lines and brain tumor or non-diseased tissues did not overlap with genes significantly correlated between GBM PDXs and brain tumor or non-diseased tissues (Figure 5B; Supplemental Table 5). The top terms significantly enriched for each group were unique to a given comparison (Figure 5C; Supplemental Table 5). When evaluating between GBM cell lines and brain tumor tissue, we found that cellular response to heat shock terms were overrepresented in correlated genes, while terms related to immune response (e.g., NF-kappaB complex, Innate Immune System, Neutrophil degranulation) were overrepresented in anticorrelated genes (Figure 5C; Supplemental Table 5). Between GBM cell lines and non-diseased brain tissue gene expression, highly correlated genes were, surprisingly, overrepresented for oxidative phosphorylation and related terms (e.g., respiratory chain complex), and anti-correlated to cellular and anatomical development and cellular motility (Figure 5C; Supplemental Table 5). The terms significantly enriched in the correlated gene set between GBM PDXs and tumor brain tissue varied widely, from Bardet-Biedl syndrome, microtubule cytoskeleton, “skin 1; Langerhans”, to exercise-induced circadian regulation). The anticorrelated gene set had top enriched terms for oxidoreductase activity and metabolic pathways (daunorubicin metabolic pathway, polyketide metabolic process, etc.) (Figure 5C; Supplemental Table 5). Gene sets significantly correlated between non-diseased brain tissue and the GBM PDXs included cell-cell signaling and synaptic signaling. The terms from the significantly anticorrelated gene set included mostly metabolic pathways (e.g., cholesterol biosynthesis, alcohol biosynthetic process, and steroid biosynthesis). Our results here suggest that cell lines and PDX models recapitulate disease contexts in unique biological terms and pathways, and, further, uniquely fail to capture other specific biological terms and pathways.

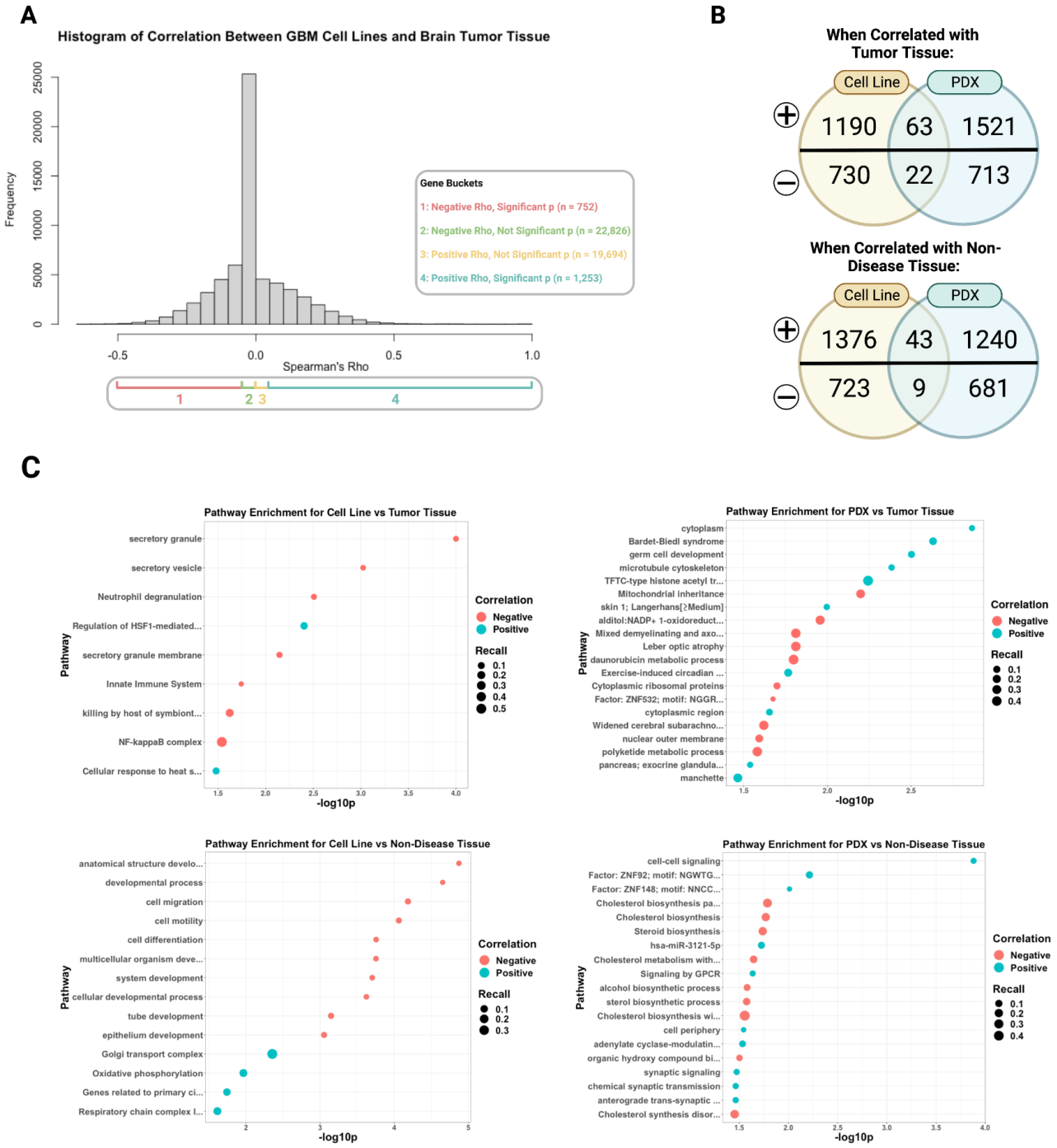


Figure 5. A. Histogram of the distribution of Spearman's correlation between GBM cell lines and brain tumor tissue and the separation into groups by positive rho (i.e., correlated) or negative rho (i.e., anticorrelated) as well as the p-value (i.e., $p \geq 0.05$ is insignificant, $p < 0.05$ is significant). **B**. Venn diagrams with the number of genes significantly correlated (top of each diagram, denoted by "+") or anti-correlated (bottom of each diagram, denoted by "-") between GBM models (cell lines on the left, PDXs on the right) and brain tumor (top diagram) and non-diseased (bottom diagram) tissue. **C**. Dot plots of the most significant (top 10 if at least 10 were significant) pathways represented in significant genes with either positive (blue) or negative (red) correlation between, clockwise from top left, GBM cell lines and brain tumor tissue, GBM PDXs and brain tumor tissue, GBM PDXs and brain non-diseased tissue, and GBM cell lines and brain non-diseased tissue, respectively.

Dots are sized by recall or the proportion of genes from a specific term represented in the data divided by the total genes possible for that term.[41]

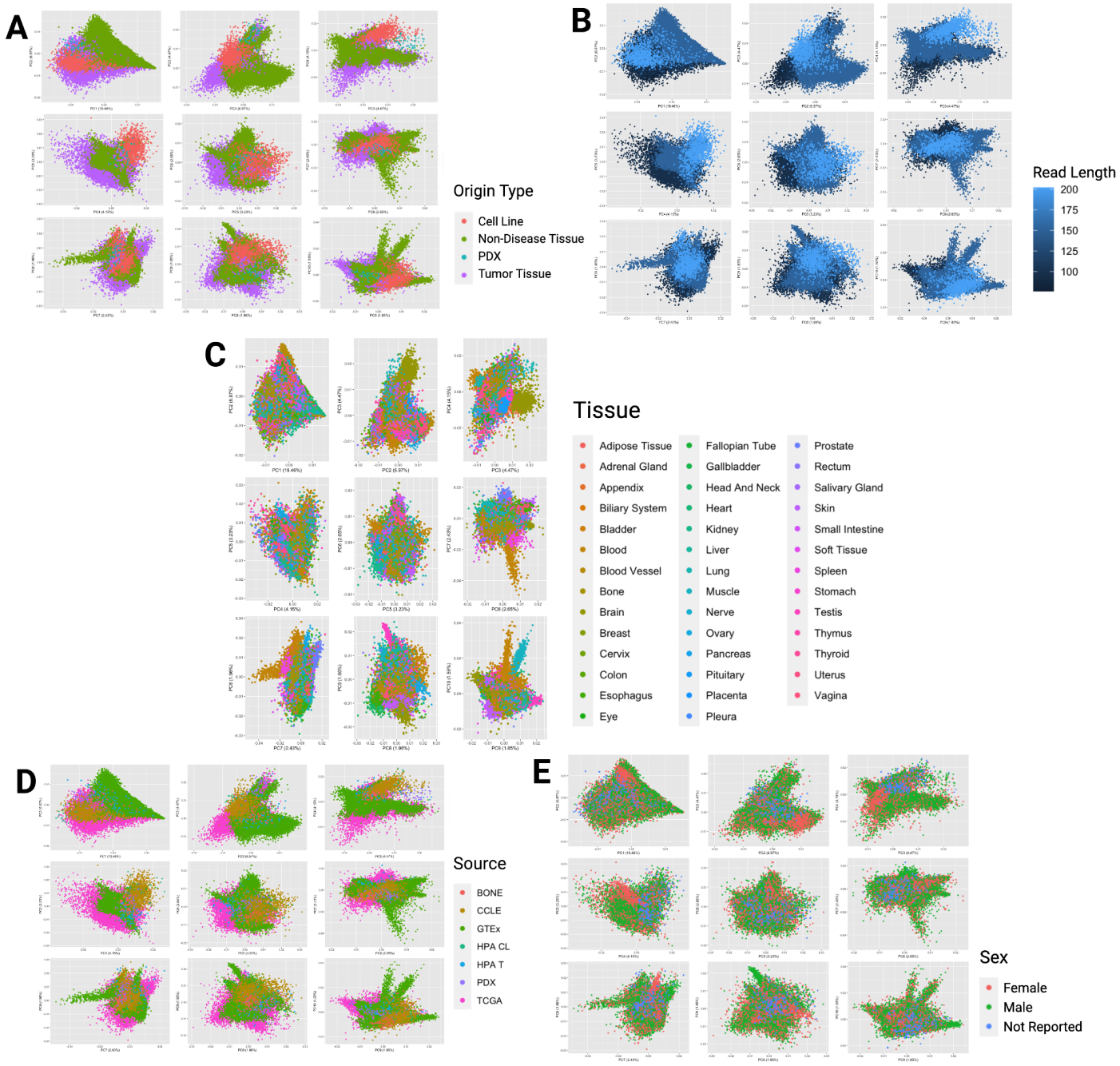
4. Discussion

In this study, we evaluated how well cancer cell lines derived from many different tissues recapitulate human tissue expression patterns across different contexts (i.e., tissue, disease) and concerning gene inclusion (i.e., most variable genes, purity associated genes, tissue-specific genes). We also performed additional analyses in GBM models, given the availability of GBM PDX gene expression profiles. We found that using the full gene set improves correlations between preclinical model and tissue global gene expression profiles, confirmed that GBM PDX global gene expression correlation to GBM tumor global gene expression outperforms GBM cell line to GBM tumor global gene expression correlations, and demonstrated that preclinical models in our study often failed to reproduce tissue-specific expression. While including additional genes for global gene expression comparison between cell lines and tissues decreases the overall correlation, it improves the relative rank between a cell line and its tissue of origin compared to other tissues.

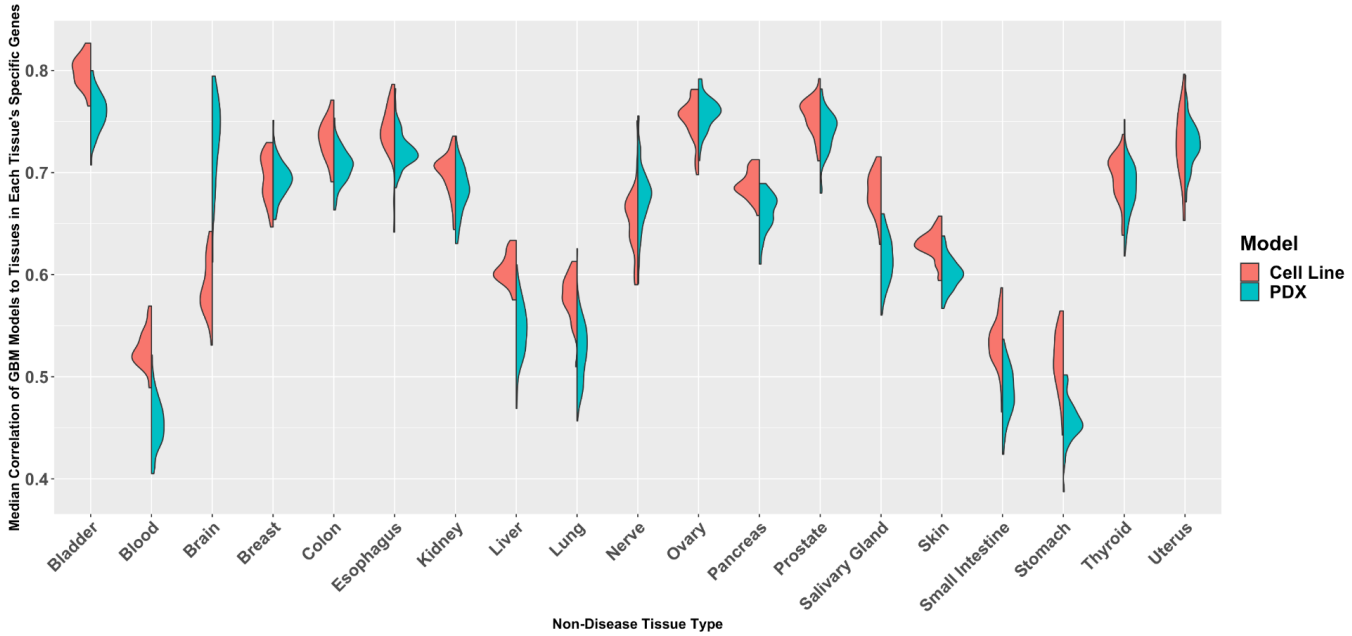
There are several limitations to the current study. While we removed 66 cancer cell lines from our study that were known to be problematic in the Cellosaurus (i.e., evidence of contamination or misidentification), other cancer cell lines included may be misidentified or contaminated but not reported. Another limitation of our study is that while we assessed how genes associated with tumor purity and tissue-specific expression impact the global gene expression correlation between cancer cell lines and human tissues, future studies will need to assess if particular pathways or other gene sets are conserved or not and the impact of other tumor-specific characteristics like subtype or tumor grade. Finally, given the robust number of GBM cell lines and PDX gene expression profiles, our comparative analyses between cancer cell lines and PDXs were focused on these. While we found that GBM PDX models better recapitulated tissue-specific gene expression than GBM cell lines when correlated with cancer and non-diseased brain tissues, this may or may not be true for cancer cell lines and PDX models derived from other tissues.

Our findings 1) underscore the importance of using the full gene expression set measured when comparing preclinical models and tissues and 2) confirm that tissue-specific patterns are better preserved in GBM PDX models than in GBM cell lines. The latter is not surprising, as these cancer cell lines and PDX models are from GBM, so restricting tissue-specific gene expression of non-brain tissues should reduce that correlation. However, given the important role of preclinical models in oncology research, our study suggests that, at least for brain tumor research, careful consideration of what aspects of tumor gene expression are reproduced in cancer cell lines compared to PDXs may be critical. Future studies can build on these findings to determine the specific pathways and gene sets recapitulated by particular preclinical models to facilitate model selection for a given study design or goal.

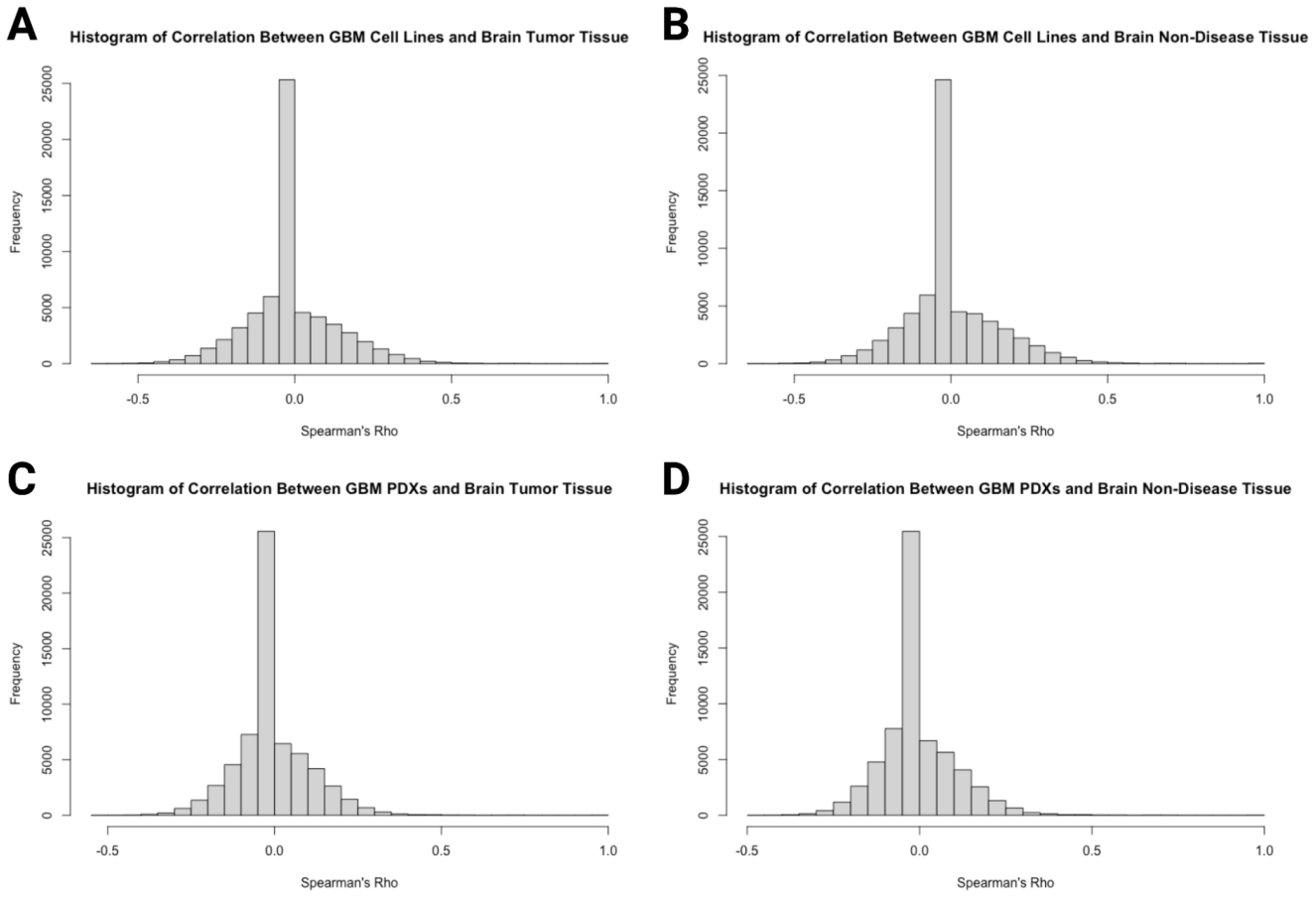
Supplementary Figures:



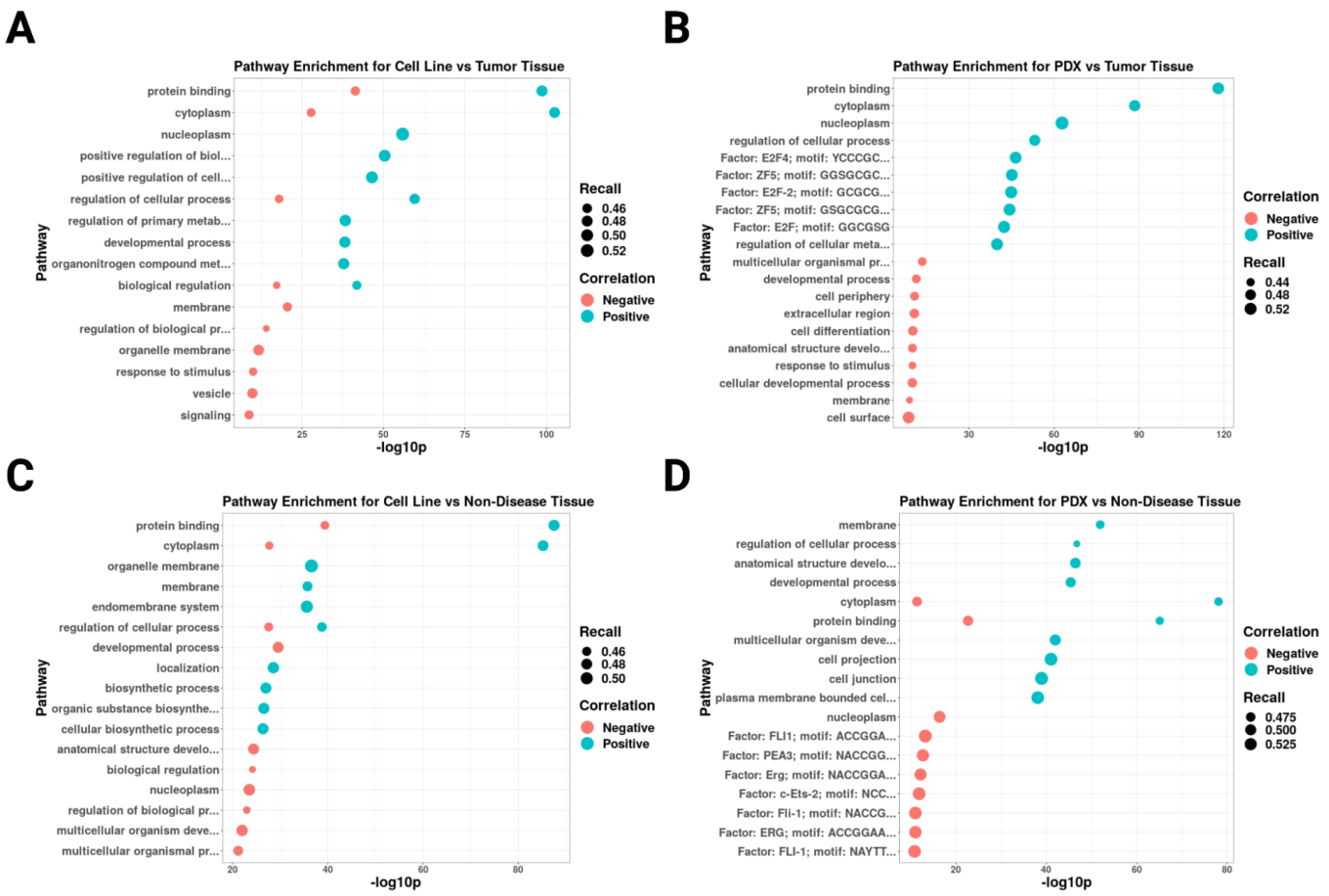
Supplementary Figure 1. A. Principal component analysis scatterplot with density rug plot of gene expression profiles from cancer cell lines (red), GBM PDX models (teal), non-diseased tissue (green), and tumor tissue samples (purple) in principal components 1-10. **B.** Principal component analysis scatterplot with density rug plot of gene expression profiles from the same groups colored by read length. **C.** Principal component analysis scatterplot with density rug plot of gene expression profiles from the same groups colored by tissue type, regardless of disease state. **D.** Principal component analysis scatterplot with density rug plot of gene expression profiles from the same groups colored by the data resource samples originated from. **E.** Principal component analysis scatterplot with density rug plot of gene expression profiles of the same groups colored by sex of the individual that samples were originally derived from.



Supplementary Figure 2. Split violin plots of the correlation of each GBM cancer cell line and PDX sample to each non-diseased tissue type by that tissue's tissue-specific genes.



Supplementary Figure 3. Histogram of the distribution of Spearman's correlation rho between **A.** GBM cell lines and brain tumor tissue, **B.** GBM cell lines and brain non-diseased tissue, **C.** GBM PDXs and brain non-diseased tissue, and **D.** GBM PDXs and brain tumor tissue, respectively.



Supplementary Figure 4. Dot plots of the most significant (top 10 if at least 10 were significant) pathways represented in insignificant genes with either positive (blue) or negative (red) correlation between **A.** GBM cell lines and brain tumor tissue, **B.** GBM PDXs and brain tumor tissue, **C.** GBM PDXs and brain non-diseased tissue, and **D.** GBM cell lines and brain non-diseased tissue, respectively. Dots are sized by recall or the proportion of genes from a specific term represented in the data divided by the total genes possible for that term.

Supplementary Tables:

[SupplementaryTable1_ProblematicSamples.xlsx]:

Supplementary Table 1: Problematic Samples Based on Metadata Mislabeling or Possible Contamination.

[SupplementaryTable2_CommonTissueTypes.xlsx]:

Supplementary Table 2: Tissue Types Common Between Data Sources.

[SupplementaryTable3_CancersRepresented.xlsx]:

Supplementary Table 3: Cancers Represented Between Data Sources.

[SupplementaryTable4_MedianCorrelations.xlsx]:

Supplementary Table 4: Median Correlation of Each Individual Model Sample to Each Tissue Grouping.

[SupplementaryTable5_PathwayTerms.xlsx]:

Supplementary Table 5: Pathway Terms Represented in GOSt Results for Each Gene Set of Interest.

References

1. Seyhan AA. Lost in translation: the valley of death across preclinical and clinical divide – identification of problems and overcoming obstacles. *Translational Medicine Communications*. 2019;4(1):1-19.

2. Sajjad H, Imtiaz S, Noor T, Siddiqui YH, Sajjad A, Zia M. Cancer models in preclinical research: A chronicle review of advancement in effective cancer research. *Animal Model Exp Med*. 2021;4(2):87-103.

3. Stacey GN. Cell culture contamination. *Methods Mol Biol*. 2011;731:79-91.

4. Januszyk M, Rennert RC, Sorkin M, et al. Evaluating the Effect of Cell Culture on Gene Expression in Primary Tissue Samples Using Microfluidic-Based Single Cell Transcriptional Analysis. *Microarrays (Basel)*. 2015;4(4):540-550.

5. Huo KG, D'Arcangelo E, Tsao MS. Patient-derived cell line, xenograft and organoid models in lung cancer therapy. *Transl Lung Cancer Res*. 2020;9(5):2214-2232.

6. Bairoch A. The Cellosaurus, a Cell-Line Knowledge Resource. *J Biomol Tech*. 2018;29(2):25-38.

7. Zhang Z, Zhou L, Xie N, et al. Overcoming cancer therapeutic bottleneck by drug repurposing. *Signal Transduct Target Ther*. 2020;5(1):113.

8. Petsko GA. When failure should be the option. *BMC Biol*. 2010;8:61.

9. Batchu S, Kellish AS, Hakim AA. Assessing alveolar rhabdomyosarcoma cell lines as tumor models by comparison of mRNA expression profiles. *Gene*. 2020;760:145025.

10. Yu K, Chen B, Aran D, et al. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nat Commun*. 2019;10(1):3574.

11. Ghandi M, Huang FW, Jané-Valbuena J, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*. 2019;569(7757):503-508.

12. Thul PJ, Lindskog C. The human protein atlas: A spatial map of the human proteome. *Protein Sci*. 2018;27(1):233.

13. Scott MC, Temiz NA, Sarver AE, et al. Comparative Transcriptome Analysis Quantifies Immune Cell Transcript Levels, Metastatic Progression, and Survival in Osteosarcoma. *Cancer Res*. 2018;78(2):326-337.

14. Vaubel RA, Tian S, Remonde D, et al. Genomic and Phenotypic Characterization of a Broad Panel of Patient-Derived Xenografts Reflects the Diversity of Glioblastoma. *Clin Cancer Res*. 2020;26(5):1094-1104.

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

16. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013;45(6):580-585.

17. Wilks C, Zheng SC, Chen FY, et al. recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol*. 2021;22(1):323.

18. Uhlén M, Fagerberg L, Hallström BM, et al. Proteomics. Tissue-based map of the human proteome. *Science*. 2015;347(6220):1260419.

19. Louis DN, Perry A, Wesseling P, et al. The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol*. 2021;23(8):1231-1251.

20. Aken BL, Ayling S, Barrell D, et al. The Ensembl gene annotation system. *Database* . 2016;2016. doi:10.1093/database/baw093

21. Leonardo Collado-Torres [aut, cre], Abhinav Nellore [ctb], Andrew E. Jaffe [ctb], Margaret A. Taub [ctb], Kai Kammers[ctb], Shannon E. Ellis [ctb], Kasper Daniel Hansen [ctb], BenLangmead [ctb], Jeffrey T. Leek [aut, ths]. *Recount*. Bioconductor; 2017. doi:10.18129/B9.BIOC.RECOUNT

22. Chen B, Sirota M, Fan-Minogue H, Hadley D, Butte AJ. Relating hepatocellular carcinoma tumor samples and cell lines using gene expression data in translational research. *BMC Med Genomics*. 2015;8 Suppl 2(Suppl 2):S5.

23. Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for ’omics feature selection and multiple data integration. *PLoS Comput Biol*. 2017;13(11):e1005752.

24. Wang Y, Lê Cao KA. PLSDA-batch: a multivariate framework to correct for batch effects in microbiome data. *Brief Bioinform*. 2023;24(2). doi:10.1093/bib/bbac622

25. Rouillard AD, Gundersen GW, Fernandez NF, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* . 2016;2016. doi:10.1093/database/baw100

26. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science*. 2015;348(6235):648-660.

27. Core Team R. R: A language and environment for statistical computing. Version 3.6. 0. Vienna, Austria. */ra-language-and-environment-forstatistical-computing*.

28. Kassambara A. Pipe-Friendly Framework for Basic Statistical Tests [R package rstatix version 0.7.2]. Published online February 1, 2023. Accessed February 2, 2023. https://CRAN.R-project.org/package=rstatix

29. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. A Lego System for Conditional Inference. *Am Stat*. 2006;60(3):257-263.

30. Hothorn T, Hornik K, van de Wiel MA, Zeileis A. Implementing a Class of Permutation Tests: The coin Package. *J Stat Softw*. 2008;28:1-23.

31. Kolberg L, Raudvere U, Kuzmin I, Vilo J, Peterson H. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Res*. 2020;9. doi:10.12688/f1000research.24956.2

32. Wickham H. *Ggplot2 Elegant Graphics for Data Analysis*. Springer London, Limited; 2016.

33. Kassambara A. “ggplot2” Based Publication Ready Plots [R package ggpubr version 0.5.0]. Published online November 16, 2022. Accessed February 2, 2023. https://CRAN.R-project.org/package=ggpubr

34. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*. 2016;32(18):2847-2849.

35. Wang Y, Cao KAL. A multivariate method to correct for batch effects in microbiome data. *bioRxiv*. Published online October 28, 2020:2020.10.27.358283. doi:10.1101/2020.10.27.358283

36. BioRender. Accessed March 14, 2023. https://www.biorender.com/

37. Carter SL, Cibulskis K, Helman E, et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat Biotechnol*. 2012;30(5):413.

38. Aran D, Sirota M, Butte AJ. Systematic pan-cancer analysis of tumour purity. *Nat Commun*. 2015;6:8971.

39. Warren A, Chen Y, Jones A, et al. Global computational alignment of tumor and cell line transcriptional profiles. *Nat Commun*. 2021;12(1):22.

40. Raudvere U, Kolberg L, Kuzmin I, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. 2019;47(W1):W191-W198.

41. Gene list functional enrichment analysis and namespace conversion with gprofiler2. Published August 23, 2021. Accessed March 16, 2023. https://cran.r-project.org/web/packages/gprofiler2/vignettes/gprofiler2.html