medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Variation in bulk RNA-seq and estimated cell type proportion using deconvolution when comparing pancreatic cancer samples within the same individual

Rick J. Jansen, PhD[1]; Sarah A. Munro, PhD[2]; Samuel O. Antwi, PhD[3] ; Kari G. Rabe, MS[4] ; Hugues Sicotte, PhD[5]

[1]Masonic Cancer Center, University of Minnesota, Minneapolis, MN

[2]Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN

[3] Department of Quantitative Health Sciences, Division of Epidemiology, Mayo Clinic, Jacksonville, FL

[4]Department of Quantitative Health Sciences, Division of Clinical Trials and Biostatistics, Mayo Clinic, Rochester, MN

[5] Department of Quantitative Health Sciences, Division of Computational Biology, Mayo Clinic, Rochester, MN

Abstract

Introduction: There is great promise in using genomic data to inform individual cancer treatment plans. Assessing intratumor genetic heterogeneity, studies have shown it may be possible to target biopsies to tumor subclones driving disease progression or treatment resistance. Here, we explore if the interpretation of tumor gene expression analysis varies across two specimens from the same patient. **Methods:** We performed bulk RNA-seq using FFPE samples from 16 patients who also had a previous separate bulk RNA-seq performed and deposited in TCGA. We used three different deconvolution methods to compare cell type proportions for these paired data. We normalized study-specific gene expression values per gene by calculating transcripts per million and adjusted for batch effect across study to compare median expression values. We also compared the reliability of gene expression measurements. We selected *KRAS, TP53, SMAD4*, and *CDKN2A*, as the most mutated genes in pancreatic cancer, and *CTNNB1, JUN, SMAD3, SMAD7*, and *TCF7,* as these tend to be enriched in pancreatic cancer compared with adjacent normal tissue. **Results:** We found that average cell type proportion varied the most between studies (i.e., samples for each patient) for NK and macrophages (using adjusted p-value 0.05/21=0.002). For the differential expression analysis, we did not observe significant differences in average expression of any of the selected genes. We observed substantial (kappa=0.75) for only *JUN* with low to moderate concordance (i.e., Kappa value 0.25-0.5) when using a median cut point for the remaining 8 genes across the two studies. **Discussion:** Together, the findings suggest that more than one tumor sample may be

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

needed for effective treatment planning. Any potential difference in observed expression values across the paired samples could be related to the different cell type proportions across the samples. The sample size was small, and each study used different sequencing technologies, so any interpretation should be confirmed with additional studies.

Introduction

Pancreatic Ductal Adenocarcinoma (PDAC) is a highly aggressive malignancy with heterogeneous tumor microenvironment.(Oketch et al., 2024) As a result, patients are often diagnosed at a late stage contributing to its high mortality rate and poor treatment response.(Grünwald et al., 2021; Hwang et al., 2022) In 2020, approximately 495,773 new cases were diagnosed worldwide, ranking it as the 12th most common malignancy.(Cai et al., 2021; J. X. Hu et al., 2021; Li et al., 2024; Rawla et al., 2019) In the United States, the American Cancer Society reported approximately 60,430 new cases and 48,220 deaths in 2021, ranking pancreatic cancer as the third leading cause of cancer death.(Cai et al., 2021; J. X. Hu et al., 2021) Similarly, in the European Union, it is projected that approximately 111,500 people will die from pancreatic cancer by 2025. These statistics underscore the urgent need for improved prevention, early detection, and treatment strategies to mitigate the escalating impact of pancreatic cancer on global health.(Ying et al., 2025)

The promise and utility of using genomic data to inform individual cancer treatment plans has been longstanding. However, minimal attention has been focused on using genomic data to inform biopsy protocols or guide biopsy sampling strategies to enhance the diagnostic and prognostic yield of tissue sampling. By assessing intratumor genetic heterogeneity, studies in other cancers have shown it may be possible to target biopsies to tumor subclones driving disease progression or treatment resistance.(Blanco-Heredia et al., 2024) The tumor microenvironment (TME) in PDAC is a complex biological barrier with multiple components, such as desmoplasia, hypoxia, presence of various cell types, and complex signaling pathways, making treatment challenging.(Binkowski et al., 2024, 2025; Giannoukakos et al., 2024; Padwal et al., 2024)

The TME's complexity necessitates innovative therapeutic strategies, and nanomedicine offers potential solutions for targeted drug delivery and modulation of the microenvironment.(Nair et al., 2024) CNV studies analyze the expression levels of genes with amplifications or deletions in malignant tissues compared to normal pancreatic tissues to understand the CNV landscape and identify correlations with survival outcomes.(Giannoukakos et al., 2024) These variations can influence gene expression, leading to the dysregulation of critical cellular processes and contributing to cancer development and progression.(Mazur et al., 2010) Liquid biopsy, particularly RNA

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

sequencing (RNA-Seq) of blood samples, is emerging as a non-invasive method for cancer diagnostics and monitoring treatment response.(Giannoukakos et al., 2024) RNA-seq allows for the analysis of gene expression and identification of potential biomarkers in body fluids.(Padwal et al., 2024) Studies employing RNA-Seq have identified potential biomarkers for early detection of pancreatic cancer, including with proteins.(Butera et al., 2024; Karmakar et al., 2020; R. Liu et al., 2024) Furthermore, some research indicates that a NETseq ensemble classifier based on peripheral blood RNA-seq data may serve as a tool for non-invasive detection of neuroendocrine tumors and assessment of treatment response.(Padwal et al., 2024)

Here, we explored the importance of targeted biopsy sampling by using bulk RNA-seq data to analyze variation gene expression across paired tumor samples from the same group of individuals.

Materials and Methods

We performed bulk RNA-seq extracted from FFPE samples of 16 patients who also had bulk RNA-seq performed on a different tumor section and deposited in TCGA. The pipeline and workflow details for the TCGA can be found on their website. (https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/) The second sample for these16 patients was processed using the NovaSeq S4 PE100 with Illumina's TruSeq Total Stranded RNA prep reagents (https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-stranded-total-rna.html) and used a second DNase treatment to minimize any potential DNA contamination. Table 1 shows select patient characteristics for our study. (**Table 1**) The Institutional Review Board of Mayo Clinic gave approval for this work.

Table 1. Select patient characteristics for 16 patients with paired samples.

Characteristic	N = 16[1]
Age at PDAC Diagnosis	
<65	8 (50%)
65+	8 (50%)
Median (Range)	63 (43,75)
Sex	
Female	5 (31%)

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Male	11 (69%)
Race/Ethnicity	
Multiracial	1 (6.3%)
White - Non-Hispanic	15 (94%)
Ever Smoker	9 (60%)
Unknown	1
Body Mass lindex (BMI) (kg/m²)	31.2 (21.8,36.8)
Obese (BMI > 30)	6 (55%)
Unknown	5
Diabetes	
<3 years	1 (13%)
No	7 (88%)
Unknown	8
Pancreatitis	4 (36%)
Unknown	5
Family History of PDAC	1 (10%)
Unknown	6
Family History of Pancreatitis	1 (10%)
Unknown	6
Histological Classification	
Undifferentiated (anaplastic) carcinoma	1 (6.3%)
Ductal adenocarcinoma	13 (81%)

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Ductal adenocarcinoma in association with IPMN (tubular)	1 (6.3%)
Mucinous adenocarcinoma in association with IPMN (muconuodular)	1 (6.3%)
Tumor Stage	
IB	1 (6.3%)
IIA	5 (31%)
IIB	10 (63%)
Surgery Type	
Distal pancreatectomy	5 (31%)
Whipple	11 (69%)
PDAC site	
Body	3 (19%)
Body/Tail	1 (6.3%)
Head	9 (56%)
Head/Body	1 (6.3%)
Tail	1 (6.3%)
Uncinate	1 (6.3%)
Tumor Grade	

For alignment and quantification of the Mayo RNA-Seq data, we used a pipeline based on HiSat2 and subread (CHURP(Baller et al., 2019), HiSat2(Kim et al., 2019), and featureCounts(Liao et al., 2014)). The TCGA and Mayo data sets were analyzed independently for deconvolution. For each data set, raw counts were normalized to transcripts per million (tpm). Cell type reference signatures were created from two different published single cell RNA-seq studies: sig1 from GSE229413 and sig2 from GSE205049. Annotations from the original publications were used to define cell types and to subset cells to those from tumor samples (not normal tissue). We filtered each signature set to

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

remove genes with all zeros in all samples and then filtered each data set to remove any genes with mean expression across all samples that was less than the overall median gene expression value. This was done to ensure that the most robust genes were kept for the signatures. We kept only the genes that were found in both scRNA-seq signatures to have the same set of genes. We then used granulator(Pfister S et al., 2024) to run deconvolution for each bulk RNA-seq data set with each scRNA-seq reference signature. The granulator R package is designed to run multiple deconvolution algorithms, and we report the non-negative algorithm results for the methods defined by the granulator package as dtangle, nnls, and qprogwc. We used these three different deconvolution methods to compare cell type proportion estimates for both runs of RNA-seq separately.

We evaluated the impact of batch correction and normalization on median expression. We present in the main manuscript the log2 normalized study-specific gene expression values per gene by calculating tpm across all shared genes and then compare median expression values of key pancreatic cancer genes between studies. We use the ComBat-seq function in the R package sva to adjust for batch effect using an empirical Bayes framework. EdgeR to was used to normalize the batch corrected counts within gene and transformed the values by the log2 function. We selected *KRAS, TP53, SMAD4*, and *CDKN2A*, as these are the genes which are commonly mutated in PDAC.(H. feng Hu et al., 2021) We also selected *CTNNB1, JUN, SMAD3, SMAD7*, and *TCF7*, as these transcription factor enrichment identified genes have been observed to have altered gene expression in PDAC compared with adjacent normal tissue.(Atay, 2020)

Using a reference scRNA-seq study (GSE205049), we evaluated median differences in cell type proportion between the PDAC sample and the adjacent normal paired tissue samples from 9 patients using a Wilcoxon Rank Sum Test for 23 cell types. After deconvolution, we created scatter plots to show the correlation across the two RNA-seq tumor samples for each patient and evaluated significant differences in median cell type proportion using a paired Wilcoxon rank sum test for 21 cell types across the two paired samples. We used the ESTIMATE algorithm implemented in the tidyestimate package in R to estimate tumor purity (stromal and immune cell signatures) using the normalized RNA-seq data from each study separately. Additionally, we evaluated median differences using a paired Wilcoxon rank sum test across the selected nine genes previously determined to be important in PDAC. We visualized differences in log2 normalized gene expression using violin plot and statistically evaluated the significance of the median difference using a Wilcoxon Rank Sum Test with Holm adjusted p-value with values <0.05 indicating significance. We also explored the effects of batch correction, normalization, and transformation on median differences. A plot was also generated to show correlations between each of the genes based on Spearman correlation coefficients.

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Results

We observed that average cell type proportion varied the most between studies (i.e., between samples for each patient) for NK and macrophages (**Figure 1A and B**). We visually observed across all plots that Macrophages, NK cells, T cells, cycling, and dendritic cells are among the cell types that vary the most. Based on the two single cell reference samples, we observed that NK cells are significantly (adj p-value <0.05) enriched in PDAC tissue compared to adjacent normal tissue while Dendric cells (DC1 & 2) and CD16+ monocyte populations are significantly reduced (**Figure 2A**). Using the Wilcoxon test in our paired samples, we observe the most significant difference in the deconvolution-based cell types are for CD4 (adjusted p-value = 0.00613), endothelial (adjusted p-value = 0.00000172), fibroblasts (adjusted p-value = 0.000161), granulocytes (adjusted p-value = 0.000895), neural (adjusted p-value = 0.000378), NK (adjusted p-value = 0.00224), and monocytes (adjusted p-value = 0.00816; **Figure 2B**).

Figure 1. **A)** Sig 1 **& B)** Sig 2 deconvolution cell type proportion estimates for 16 paired samples. Each panel represents a different deconvolution method. Each color represents an individual cell type. Each point represents a patient's proportion for The Cancer Genome Atlas Program (TCGA) sample (y-axis) and the second study sample (x-axis).

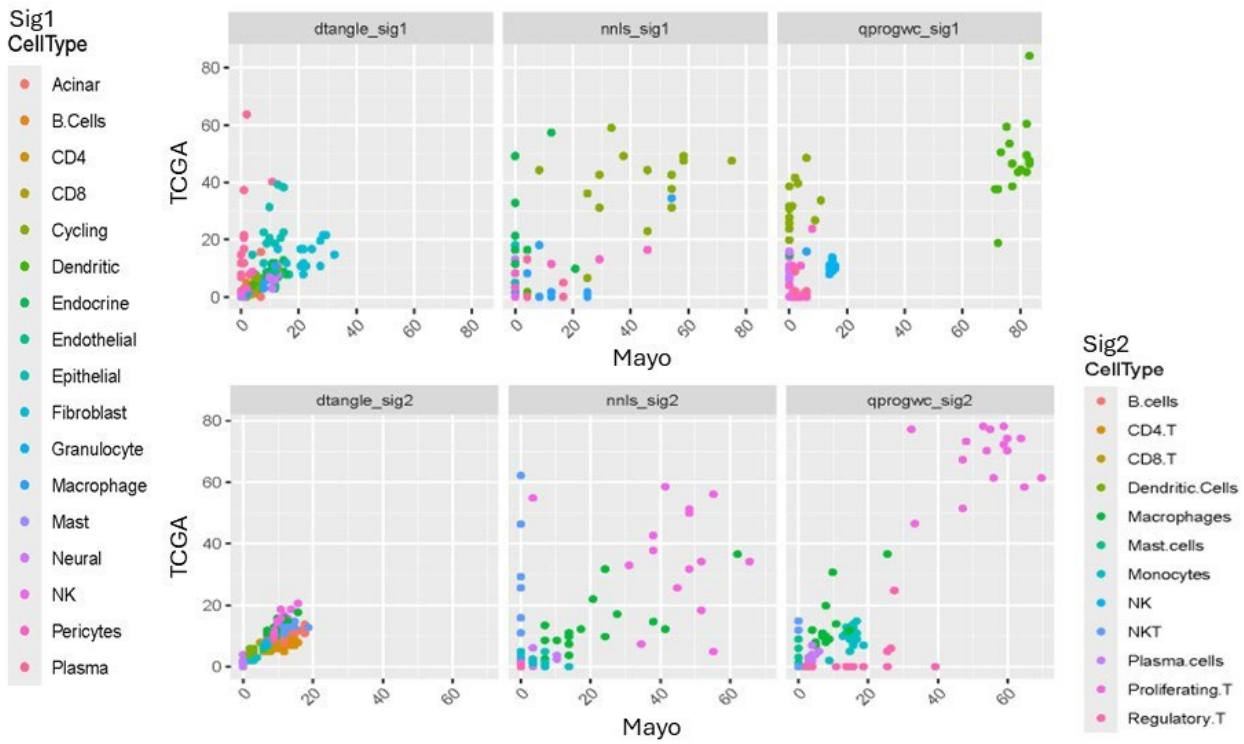
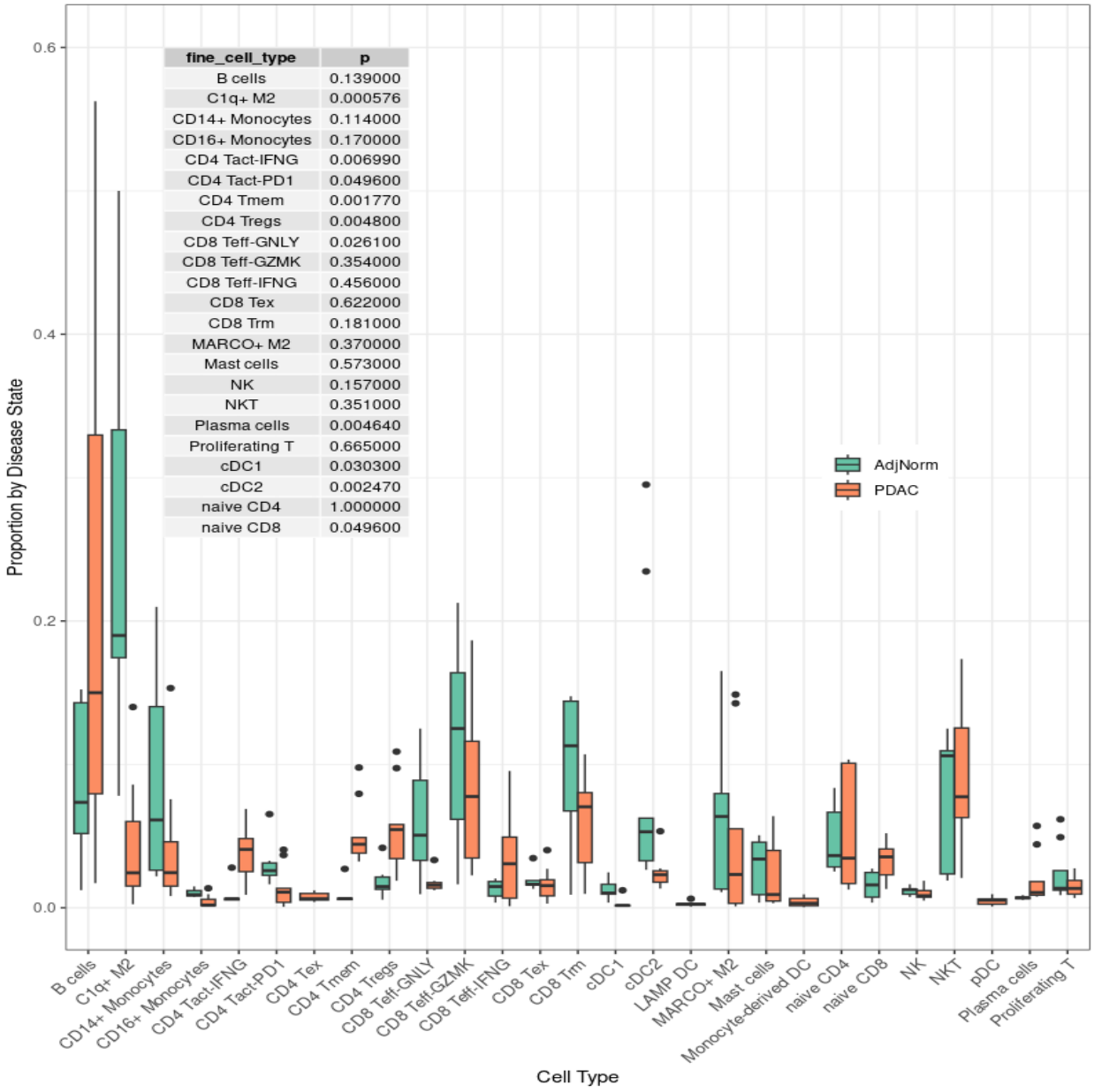


Figure 2.A) Comparison of PDAC (orange) and Adjacent normal (green) cell type proportions. **B)** Comparison of Mayo (salmon) vs TCGA (blue -green) cell type proportions

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

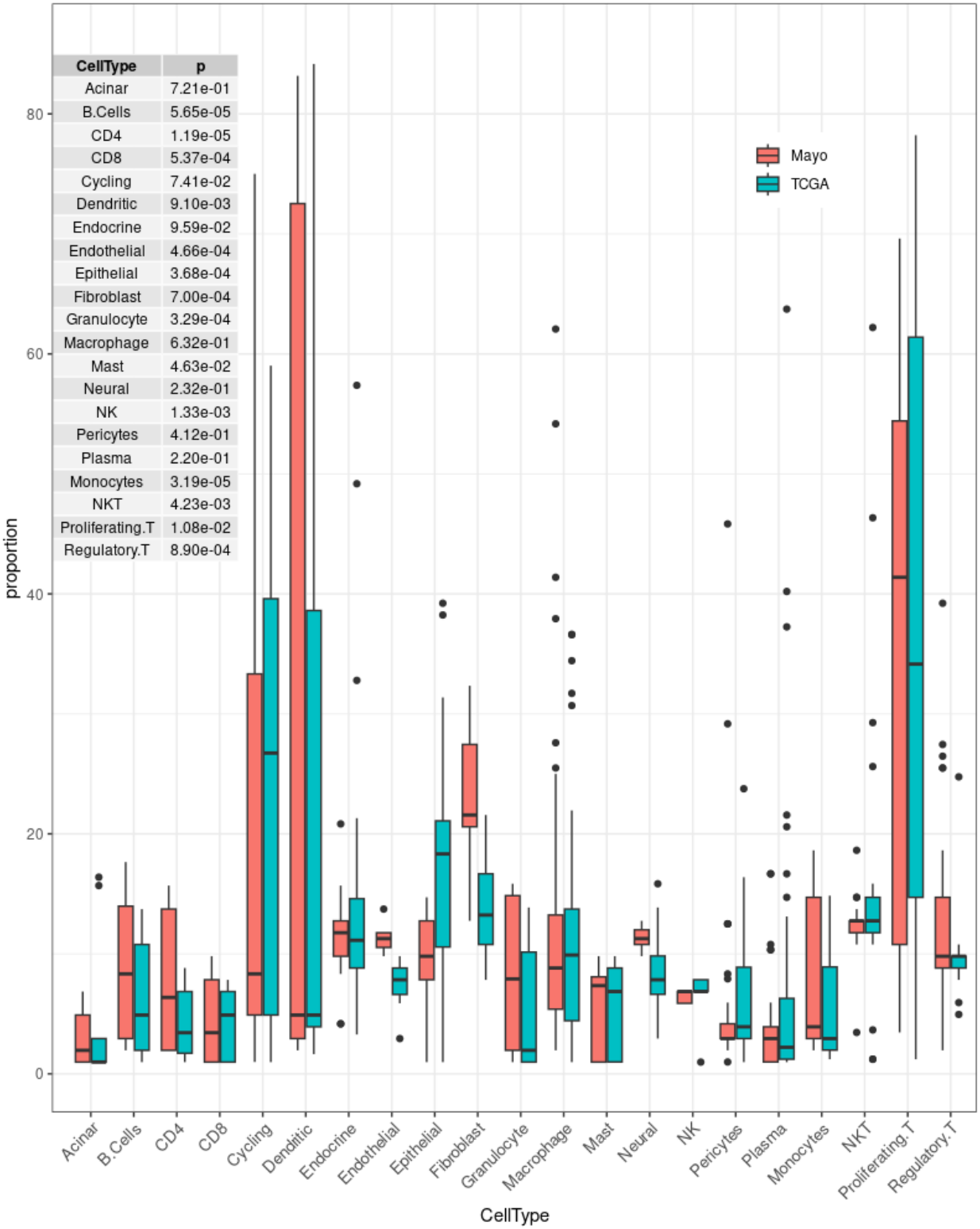
across 21 different cell types. P-values are from the Wilcoxon rank sum test. The scatterplot shows the correlation between TCGA RNA-seq-based deconvolution results and Mayo RNA-seq-based deconvolution. The top row represents more cell types and uses a different single cell reference sample than the bottom row. Each of the three columns represent a different deconvolution method: dtangle, nnis, and qprogwc. Each point is colored according to the cell type, as indicated by the legend on the left and right sides. The primary focus should be on identifying patterns for specific cell types that deviate from the diagonal.

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .



fine_cell_type	p
B cells	0.139000
C1q+ M2	0.000576
CD14+ Monocytes	0.114000
CD16+ Monocytes	0.170000
CD4 Tact-IFNG	0.006990
CD4 Tact-PD1	0.049600
CD4 Tmem	0.001770
CD4 Tregs	0.004800
CD8 Teff-GNLY	0.026100
CD8 Teff-GZMK	0.354000
CD8 Teff-IFNG	0.456000
CD8 Tex	0.622000
CD8 Trm	0.181000
MARCO+ M2	0.370000
Mast cells	0.573000
NK	0.157000
NKT	0.351000
Plasma cells	0.004640
Proliferating T	0.665000
cDC1	0.030300
cDC2	0.002470
naive CD4	1.000000
naive CD8	0.049600

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .



CellType	p
Acinar	7.21e-01
B.Cells	5.65e-05
CD4	1.19e-05
CD8	5.37e-04
Cycling	7.41e-02
Dendritic	9.10e-03
Endocrine	9.59e-02
Endothelial	4.66e-04
Epithelial	3.68e-04
Fibroblast	7.00e-04
Granulocyte	3.29e-04
Macrophage	6.32e-01
Mast	4.63e-02
Neural	2.32e-01
NK	1.33e-03
Pericytes	4.12e-01
Plasma	2.20e-01
Monocytes	3.19e-05
NKT	4.23e-03
Proliferating.T	1.08e-02
Regulatory.T	8.90e-04

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

We estimated tumor infiltration scores in each study separately using normalized RNA-seq data in the ESTIMATE function in R. In summary, this plot (**Figure 3**) demonstrates the positive relationship between the estimated stromal and immune cell infiltration and the ESTIMATE score, suggesting that tumors with high stromal and immune cells also likely have high infiltration of the tumor. This is expected because a higher presence of non-tumor cells will naturally lower the proportion of tumor cells in the sample. The blue-green (TCGA) and salmon (Mayo) circles highlight specific tumor samples with distinct characteristics in terms of their microenvironment. It is important to note that ESTIMATE scores can only be interpreted relatively, and it cannot be inferred across studies.

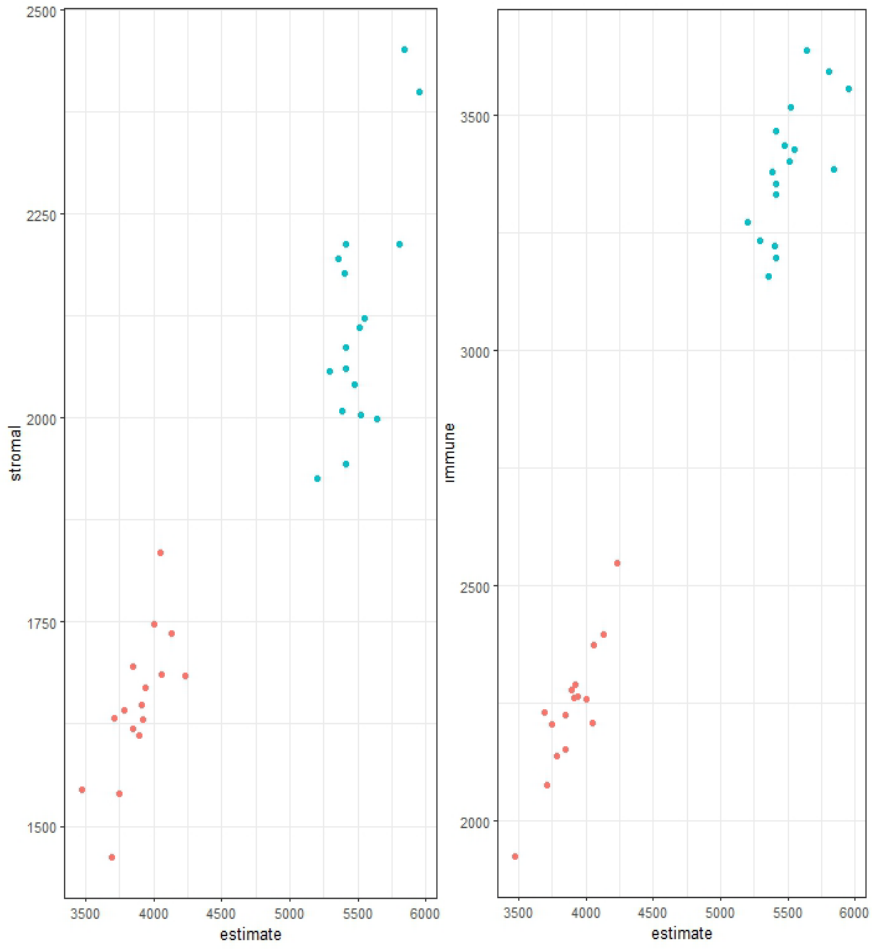


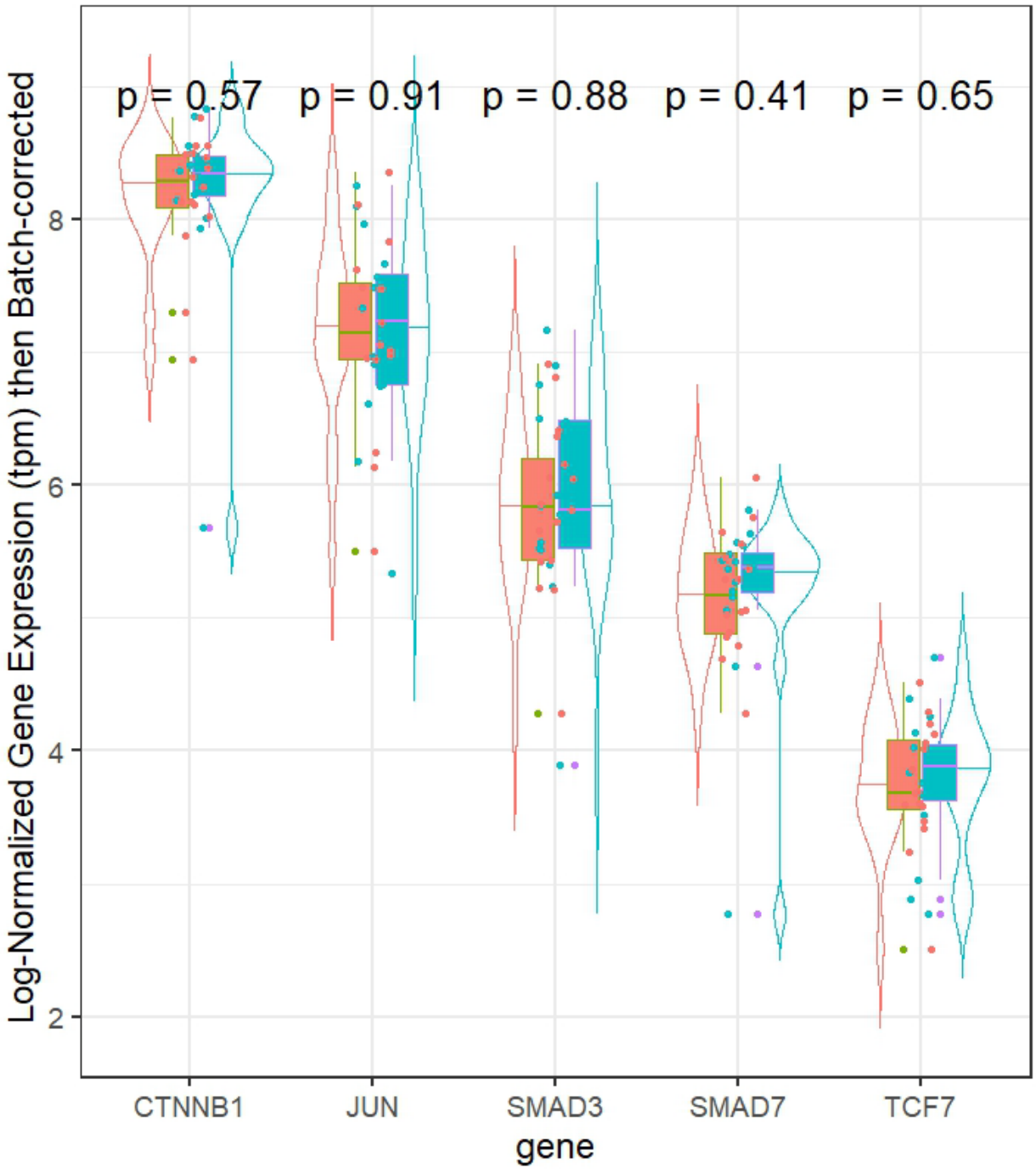
Figure 3 Tumor infiltration estimates for Mayo (salmon) and TCGA (blue-green) sample data. The ESTIMATE algorithm from R was used to estimate A) stromal and B) immune infiltration in the sample using the normalized gene expression data for each study separately. The tumor infiltration estimate is relative and cannot be compared across studies.

Violin plots of gene expression in the paired Mayo and TCGA sample data for the four PDAC specific genes comparing the effect of batch correction, normalization, and transformations are included (**Supplemental Figure S1a and b**). The plots highlight the importance of adjusting both within sample (normalization) and across sample (batch correction) to avoid bias. When looking at the first four PDAC specific genes, we did not observe significant median expression differences across paired samples. However, there may be hints of differences for KRAS (adj. p-value = 0.18) and SMAD4 (adj. p-value = 0.59, **Figure 4A**). When evaluating the five transcription factor enriched genes, there were no

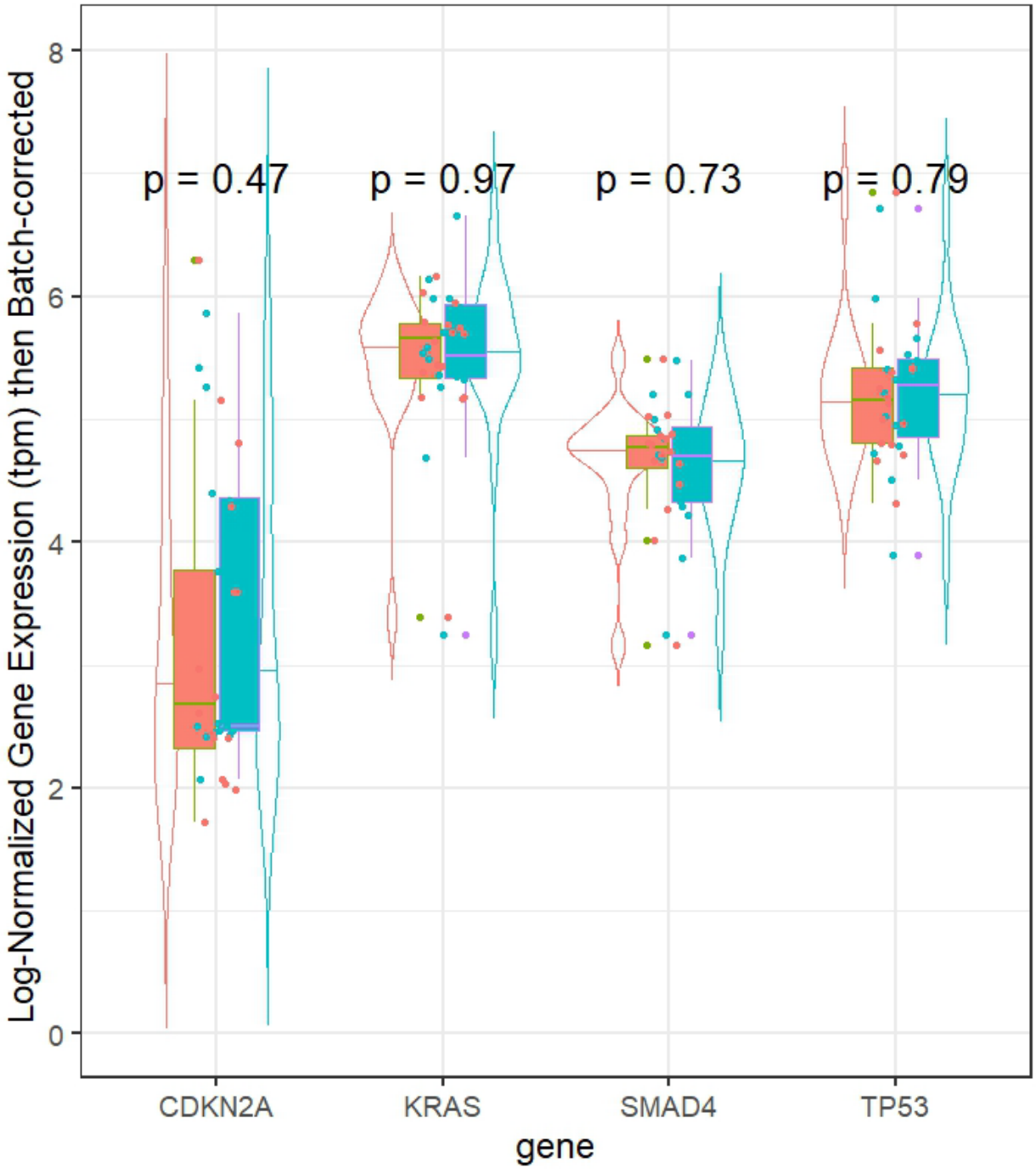
medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

statistically significant results. Visually, CTNNB1(adj. p-value = 0.30) and TCF7 (adj. p-value = 0.56) varied the most between the paired TCGA and Mayo sample data (**Figure 4B**).

Figure 4. Violin plot of normalized gene expression in transcripts per million (tpm) for Mayo (salmon) and TCGA (blue-green) samples in **A)** four selected highly mutated pancreatic cancer genes and **B)** five validated highly variably expressed genes. Each gene also has an adjusted p-value for the Wilcoxon comparison between paired samples.



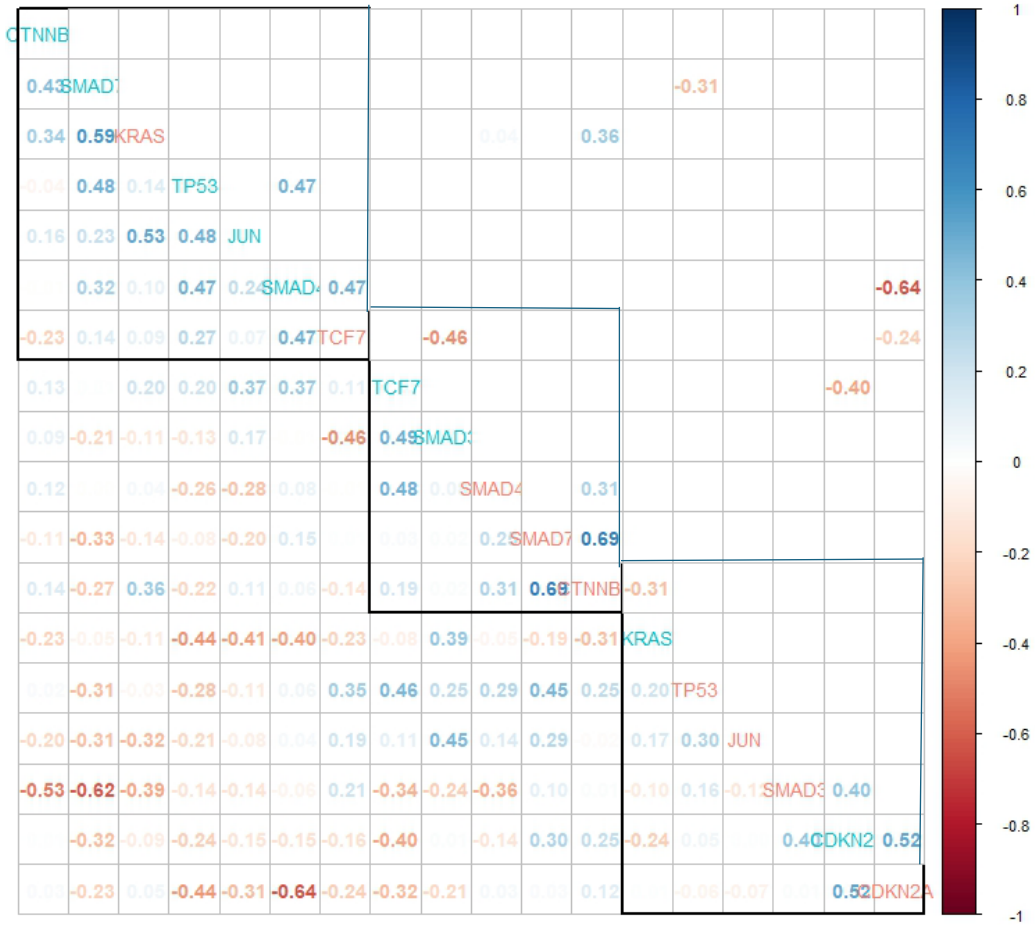
medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .



We used the median as a cut point for each gene and calculated one score for each patient separately to determine concordance. (**Table S1**) We observed a substantial concordance (kappa=0.75) for *JUN*, moderate concordance for *KRAS, TP53, SMAD3* and *CDKN2A* (Kappa =0.5) and low concordance value for *TCF7*, *SMAD7, SMAD4, CTNNB1*(Kappa = 0.25) across the two studies suggesting that sufficiently representative genomic data cannot be

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

collected with one sample. Likewise, the correlation of the gene expression values for the nine selected genes in general vary significantly across the paired samples (**Figure 5)**. In this correlation matrix, each row and column represent a gene and the number in the cell represents the Spearman correlation coefficient with colors ranging from highly positively correlated values (blue) to highly negatively correlated values (red). The coefficients which reached statistical significance ranged from <-0.24 to >0.24 and are shown in the upper triangle. The hierarchical clustering method was used to order the genes in the matrix. We observed moderate positive correlations between TP53/SMAD4 (0.47) among only TCGA samples; SMAD7/CTNNB1 (0.69), KRAS/CTNNB1 (0.36), and SMAD4/CTNNB1 (0.31) using only Mayo samples. We observed a moderate negative correlation between TCF7/CDKN2A (-0.24) within the TCGA samples only. We observed moderate positive correlation across study genes for CDKN2A/CDKN2A (0.52), SMAD4/TCF7 (0.47), and SMAD3/CDKN2A (0.40). Moderate negative correlations were observed for the following genes across studies: SMAD4/CDKN2A (-0.64), TCF7/SMAD3 (-0.46), SMAD7/TP53 (-0.31), and TCF7/CDKN2A (-0.24).



medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Figure 5. Correlation plot showing the Spearman correlation in select gene expression values for TCGA data (blue-green) and Mayo data (salmon). The genes are arranged based on hierarchical clustering. The lower triangle shows all correlation coefficients with higher correlations having higher transparency. The higher triangle shows only the correlation values which have significant p-values. Correlation coefficients with values close to 1 have deeper blue and coefficients with values close to -1 have deeper red color.

Discussion

In this study, we observed minimal nonsignificant differences in expression values for selected genes established to have a role in PDAC. This slight difference could be related to the different cell type proportions across the samples that we estimated using deconvolution methods. Studies show that pancreatic tumors are heterogeneous and different regions of the tumor may have distinct genetic and molecular profiles, (Lodestijn et al., 2021) so observing a difference across samples is not surprising. We did not observe significant differences in median log normalized, batch-corrected expression values for our selected genes. However, this may be due to a small sample size. Collectively, the results suggest there should be additional research investigating how closely a single sample fully captures the heterogeneity of a patient's tumor and determine how important this might be for treatment planning as genomic results are increasingly used in a clinical setting. Focusing research on how many samples to obtain and strategically deciding where to biopsy will be important as single cell and spatial technologies advance.

When evaluating median gene expression across studies, we did not observe any statistically significant differences; however, there may be hints of a difference. Performing follow-up scRNA-seq studies with a larger sample size could help clarify if there is an actual difference by looking at cell type specific expression. As we see using the deconvolution and tumor infiltration estimates, there is large variation in the estimated cell type proportions and tumor cells in the sample.

The correlation matrix heatmap provides a comprehensive overview of the pairwise relationships between the expression levels of the given genes. It highlights strong positive and negative correlations that can provide valuable insights into gene co-regulation, functional relationships, and potential regulatory mechanisms. We used hierarchical clustering to order the genes, and that method grouped the genes with sample type intermingled (Mayo vs. TCGA). This could indicate that because cell type proportions vary across samples, there are also potential co-regulation pathways that differ by cell type. Additionally, the difference in correlation structure by sample type could suggest a difference in functional relationships or regulatory mechanisms utilized across the two samples. Cross-sample correlations could potentially suggest a regional relationship

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

between the genes or could just represent random association. Mechanistic studies in the laboratory would need to confirm any of these suggested relationships.

Some of the key limitations in this research include the small sample size and that each of the paired samples were processed using different sequencing technologies. Any findings should be confirmed with additional studies.

The consistency of RNA-seq results across multiple samples from the same person is a complex issue. Factors that can influence consistency include: 1) biological variation as gene expression fluctuates due to circadian rhythms, physiological state (e.g., stress, illness), and cellular heterogeneity and 2) technical variation as there are many processing steps in the RNA-seq workflow and bioinformatic analysis.(Auer & Doerge, 2010) Careful experimental design and quality control are essential to minimize technical variation. While some variability is inherent in biological systems, RNA-seq can provide highly consistent results when proper experimental design and analysis are employed. (Su et al., 2014)

Tumor heterogeneity between different sites of neoplasia in a single patient, and among tumors from different patients, is an emerging theme in cancer research.(Alizadeh et al., 2015) Heterogeneity confounds researchers' understanding of tumor evolution and their ability to design effective treatments.(Javed et al., 2024; Y. Liu et al., 2024, 2025) Due to the limitations of methods that characterize tumors in 'bulk', such as averaging across all tumor clones, even a single RNA-Seq sample, may not fully represent the tumor heterogeneity.(Y. Liu et al., 2024, 2025) Ideally, researchers would obtain multiple regions from each tumor to capture spatial heterogeneity.(Giannoukakos et al., 2024; Sergeyev et al., 2024; Yaqoob et al., 2024)

While computational deconvolution of different expression components in a sample can distinguish between cells from different lineages, it has limited applicability in samples with low transcriptional diversity.(Y. Liu et al., 2025) Even at the single cell level, several factors affect the ability of a single RNA-seq sample to represent tumor heterogeneity.(Hwang et al., 2022) Single-cell RNA-Seq is a robust technology that can analyze tens of thousands of cells simultaneously in a cost-effective and efficient manner.(Hwang et al., 2022) However, sensitivity for low expressed genes still presents a challenge, and metrics that quantify similarity and difference between samples from the same clonal origin are needed.(Hwang et al., 2022; Y. Liu et al., 2025)

The long-term goal of this area of research is to harness the power of genomic analysis to optimize biopsy decisions and ultimately improve patient outcomes. Challenges remain in fully validating genetic and genomic biopsy strategies and supporting their widespread clinical implementation and will be the focus of future work. Attention should focus on

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

standardizing tissue collection/processing protocols, establishing evidence-based thresholds for acting on genomic findings, and integrating molecular and image data into existing clinicopathologic risk assessment frameworks. Research in the future should focus building off the results presented here to get a clearer picture if multiple samples taken at the time of biopsy would help develop more effective treatment plans.

Funding

Partial support from a UMN Data Science Initiative (DSI) and Research Computing SEED grant and NIH grant P30 CA77598 utilizing the Biostatistics Core shared resource of the Masonic Cancer Center, University of Minnesota and by the National Center for Advancing Translational Sciences of the National Institutes of Health Award Number UL1TR002494. Additional partial funding support for this work was provided by National Institutes of Health (NIH) COBRE grant P20GM109024, NIH P50 CA102701, and NIH R25 CA92049. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Alizadeh, A. A., Aranda, V., Bardelli, A., Blanpain, C., Bock, C., Borowski, C., Caldas, C., Califano, A., Doherty, M., Elsner, M., Esteller, M., Fitzgerald, R., Korbel, J. O., Lichter, P., Mason, C. E., Navin, N., Pe'Er, D., Polyak, K., Roberts, C. W. M., ... Zucman-Rossi, J. (2015). Toward understanding and exploiting tumor heterogeneity. *Nature Medicine*, *21*(8), 846–853. https://doi.org/10.1038/nm.3915

Atay, S. (2020). Integrated transcriptome meta-analysis of pancreatic ductal adenocarcinoma and matched adjacent pancreatic tissues. *PeerJ*, *8*. https://doi.org/10.7717/PEERJ.10141

Auer, P. L., & Doerge, R. W. (2010). Statistical design and analysis of RNA sequencing data. *Genetics*, *185*(2), 405–416. https://doi.org/10.1534/genetics.110.114983

Baller, J., Kono, T., Herman, A., & Zhang, Y. (2019). ChURP: A lightweight CLI framework to enable novice users to analyze sequencing datasets in parallel. *ACM International Conference Proceeding Series*, 3333156. https://doi.org/10.1145/3332186.3333156

Binkowski, B., Klamer, Z., Gao, C., Staal, B., Repesh, A., Tran, H.-L., Brass, D. M., Bartlett, P., Gallinger, S., Blomqvist, M., Morrow, J. B., Allen, P., Shi, C., Singhi, A., Brand, R., Huang, Y., Hostetter, G., & Haab, B. B. (2024). Multiplexed Glycan Immunofluorescence Identification of Pancreatic Cancer Cell Subpopulations in Both Tumor and Blood Samples. *BioRxiv : The Preprint Server for Biology*. https://doi.org/10.1101/2024.08.22.609143

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Binkowski, B., Klamer, Z., Gao, C., Staal, B., Repesh, A., Tran, H.-L., Brass, D. M., Bartlett, P., Gallinger, S., Blomqvist, M., Morrow, J. B., Allen, P., Shi, C., Singhi, A., Brand, R., Huang, Y., Hostetter, G., & Haab, B. B. (2025). Multiplexed glycan immunofluorescence identification of pancreatic cancer cell subpopulations in both tumor and blood samples. *Science Advances*, *11*(10). https://doi.org/10.1126/SCIADV.ADT0029.

Blanco-Heredia, J., Souza, C. A., Trincado, J. L., Gonzalez-Cao, M., Gonçalves-Ribeiro, S., Gil, S. R., Pravdyvets, D., Cedeño, S., Callari, M., Marra, A., Gazzo, A. M., Weigelt, B., Pareja, F., Vougiouklakis, T., Jungbluth, A. A., Rosell, R., Brander, C., Tresserra, F., Reis-Filho, J. S., ... De Mattos-Arruda, L. (2024). Converging and evolving immuno-genomic routes toward immune escape in breast cancer. *Nature Communications*, *15*(1). https://doi.org/10.1038/S41467-024-45292-1

Butera, F., Sero, J. E., Dent, L. G., & Bakal, C. (2024). Actin networks modulate heterogeneous NF-κB dynamics in response to TNFα. *ELife*, *13*. https://doi.org/10.7554/ELIFE.86042

Cai, J., Chen, H., Lu, M., Zhang, Y., Lu, B., You, L., Zhang, T., Dai, M., & Zhao, Y. (2021). Advances in the epidemiology of pancreatic cancer: Trends, risk factors, screening, and prognosis. *Cancer Letters*, *520*, 1–11. https://doi.org/10.1016/j.canlet.2021.06.027

Giannoukakos, S., D'Ambrosi, S., Koppers-Lalic, D., Gómez-Martín, C., Fernandez, A., & Hackenberg, M. (2024). Assessing the complementary information from an increased number of biologically relevant features in liquid biopsy-derived RNA-Seq data. *Heliyon*, *10*(6), e27360. https://doi.org/10.1016/j.heliyon.2024.e27360

Grünwald, B. T., Devisme, A., Andrieux, G., Vyas, F., Aliar, K., McCloskey, C. W., Macklin, A., Jang, G. H., Denroche, R., Romero, J. M., Bavi, P., Bronsert, P., Notta, F., O'Kane, G., Wilson, J., Knox, J., Tamblyn, L., Udaskin, M., Radulovich, N., ... Khokha, R. (2021). Spatially confined sub-tumor microenvironments in pancreatic cancer. *Cell*, *184*(22), 5577-5592.e18. https://doi.org/10.1016/J.CELL.2021.09.022

Hu, H. feng, Ye, Z., Qin, Y., Xu, X. wu, Yu, X. jun, Zhuo, Q. feng, & Ji, S. rong. (2021). Mutations in key driver genes of pancreatic cancer: molecularly targeted therapies and other clinical implications. *Acta Pharmacologica Sinica*, *42*(11), 1725–1741. https://doi.org/10.1038/S41401-020-00584-2

Hu, J. X., Lin, Y. Y., Zhao, C. F., Chen, W. B., Liu, Q. C., Li, Q. W., & Gao, F. (2021). Pancreatic cancer: A review of epidemiology, trend, and risk factors. *World Journal of Gastroenterology*, *27*(27), 4298–4321. https://doi.org/10.3748/wjg.v27.i27.4298

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Hwang, W. L., Jagadeesh, K. A., Guo, J. A., Hoffman, H. I., Yadollahpour, P., Reeves, J. W., Mohan, R., Drokhlyansky, E., Van Wittenberghe, N., Ashenberg, O., Farhi, S. L., Schapiro, D., Divakar, P., Miller, E., Zollinger, D. R., Eng, G., Schenkel, J. M., Su, J., Shiau, C., ... Regev, A. (2022). Single-nucleus and spatial transcriptome profiling of pancreatic cancer identifies multicellular dynamics associated with neoadjuvant treatment. *Nature Genetics*, *54*(8), 1178–1191. https://doi.org/10.1038/S41588-022-01134-8

Javed, S., Qureshi, T. A., Wang, L., Azab, L., Gaddam, S., Pandol, S. J., & Li, D. (2024). An insight to PDAC tumor heterogeneity across pancreatic subregions using computed tomography images. *Frontiers in Oncology*, *14*, 1378691. https://doi.org/10.3389/fonc.2024.1378691

Karmakar, S., Rauth, S., Nallasamy, P., Perumal, N., Nimmakayala, R. K., Leon, F., Gupta, R., Barkeer, S., Venkata, R. C., Raman, V., Rachagani, S., Ponnusamy, M. P., & Batra, S. K. (2020). RNA Polymerase II-Associated Factor 1 Regulates Stem Cell Features of Pancreatic Cancer Cells, Independently of the PAF1 Complex, via Interactions With PHF5A and DDX3. *Gastroenterology*, *159*(5), 1898-1915.e6. https://doi.org/10.1053/J.GASTRO.2020.07.053

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology 2019 37:8*, *37*(8), 907–915. https://doi.org/10.1038/s41587-019-0201-4

Li, Z., Zhang, X., Sun, C., Li, Z., Fei, H., & Zhao, D. (2024). Global, regional, and national burdens of early onset pancreatic cancer in adolescents and adults aged 15-49 years from 1990 to 2019 based on the Global Burden of Disease Study 2019: a cross-sectional study. *International Journal of Surgery (London, England)*, *110*(4), 1929–1940. https://doi.org/10.1097/JS9.0000000000001054

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923–930. https://doi.org/10.1093/BIOINFORMATICS/BTT656

Liu, R., Qiu, M., Deng, X., Zhang, M., Gao, Z., Wang, Y., Mei, H., Zhai, M., Zhang, Q., Hao, J., Yang, Z., & Wang, H. (2024). Erianin inhibits the progression of pancreatic cancer by directly targeting AKT and ASK1. *Cancer Cell International*, *24*(1). https://doi.org/10.1186/S12935-024-03533-9

Liu, Y., Carbonetto, P., Willwerscheid, J., Oakes, S. A., Macleod, K. F., & Stephens, M. (2024). Dissecting tumor transcriptional heterogeneity from single-cell RNA-seq data

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

by generalized binary covariance decomposition. *BioRxiv : The Preprint Server for Biology*. https://doi.org/10.1101/2023.08.15.553436

Liu, Y., Carbonetto, P., Willwerscheid, J., Oakes, S. A., Macleod, K. F., & Stephens, M. (2025). Dissecting tumor transcriptional heterogeneity from single-cell RNA-seq data by generalized binary covariance decomposition. *Nature Genetics*, *57*(1), 263–273. https://doi.org/10.1038/S41588-024-01997-Z.

Lodestijn, S. C., Miedema, D. M., Lenos, K. J., Nijman, L. E., Belt, S. C., El Makrini, K., Lecca, M. C., Waasdorp, C., van den Bosch, T., Bijlsma, M. F., & Vermeulen, L. (2021). Marker-free lineage tracing reveals an environment-instructed clonogenic hierarchy in pancreatic cancer. *Cell Reports*, *37*(3). https://doi.org/10.1016/J.CELREP.2021.109852

Mazur, P. K., Grüner, B. M., Nakhai, H., Sipos, B., Zimber-Strobl, U., Strobl, L. J., Radtke, F., Schmid, R. M., & Siveke, J. T. (2010). Identification of epidermal Pdx1 expression discloses different roles of Notch1 and Notch2 in murine Kras(G12D)-induced skin carcinogenesis in vivo. *PloS One*, *5*(10). https://doi.org/10.1371/JOURNAL.PONE.0013578

Nair, S. T., Abhi, C., Kamalasanan, K., Pavithran, K., Unni, A. R., Sithara, M. S., Sarma, M., & Mangalanandan, T. S. (2024). Pathophysiology-Driven Approaches for Overcoming Nanomedicine Resistance in Pancreatic Cancer. *Molecular Pharmaceutics*, *21*(12), 5960–5988. https://doi.org/10.1021/acs.molpharmaceut.4c00801

Oketch, D. J. A., Giulietti, M., & Piva, F. (2024). Copy Number Variations in Pancreatic Cancer: From Biological Significance to Clinical Utility. *International Journal of Molecular Sciences*, *25*(1). https://doi.org/10.3390/ijms25010391

Padwal, M. K., Parghane, R. V., Chakraborty, A., Ujaoney, A. K., Anaganti, N., Basu, S., & Basu, B. (2024). Developing a peripheral blood RNA-seq based NETseq ensemble classifier: A potential novel tool for non-invasive detection and treatment response assessment in neuroendocrine tumor patients receiving 177Lu-DOTATATE PRRT. *Journal of Neuroendocrinology*. https://doi.org/10.1111/JNE.13462

Pfister S, Kuettel V, & Ferrero E. (2024). *granulator: Rapid benchmarking of methods for *in silico* deconvolution of bulk RNA-seq data*. R Package Version 1.14.0.

Rawla, P., Sunkara, T., & Gaduputi, V. (2019). Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. *World Journal of Oncology*, *10*(1), 10–27. https://doi.org/10.14740/wjon1166

Sergeyev, O., Bezuglov, V., Soloveva, N., Smigulina, L., Denisova, T., Dikov, Y., Shtratnikova, V., Vavilov, N., Williams, P. L., Korrick, S., Lee, M. M., Zgoda, V., Hauser, R., & Suvorov,

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

A. (2024). Intraindividual variability of semen quality, proteome, and sncRNA profiles in a healthy cohort of young adults. *Andrology*. https://doi.org/10.1111/andr.13739

Su, Z., Łabaj, P. P., Li, S., Thierry-Mieg, J., Thierry-Mieg, D., Shi, W., Wang, C., Schroth, G. P., Setterquist, R. A., Thompson, J. F., Jones, W. D., Xiao, W., Xu, W., Jensen, R. V., Kelly, R., Xu, J., Conesa, A., Furlanello, C., Gao, H., ... Shi, L. (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature Biotechnology*, *32*(9), 903–914. https://doi.org/10.1038/nbt.2957

Yaqoob, A., Verma, N. K., Aziz, R. M., & Shah, M. A. (2024). RNA-Seq analysis for breast cancer detection: a study on paired tissue samples using hybrid optimization and deep learning techniques. *Journal of Cancer Research and Clinical Oncology*, *150*(10), 455. https://doi.org/10.1007/s00432-024-05968-z

Ying, H., Kimmelman, A. C., Bardeesy, N., Kalluri, R., Maitra, A., & DePinho, R. A. (2025). Genetics and biology of pancreatic ductal adenocarcinoma. *Genes & Development*, *39*(1–2), 36–63. https://doi.org/10.1101/gad.351863.124

medRxiv preprint doi: https://doi.org/10.1101/2025.05.05.25326976; this version posted May 6, 2025. The copyright holder for this preprint **(which was not certified by peer review)** is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .