Journal of proteome research

Open Access

pubs.acs.org/jpr

Article

Evaluating and Optimizing Mass Spectrometry Proteomics Data to Deconvolve Cell-Type-Specific Protein Expression in Tumors

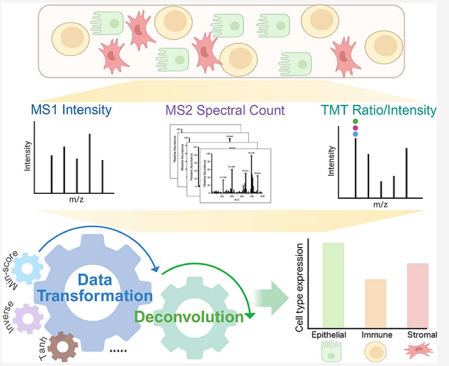
Yingnan Song, Qi Zhou, and Chen Huang*

Cite This: *J. Proteome Res.* 2026, 25, 471−483

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Understanding intratumoral heterogeneity is essential for elucidating tumor biology. Compared to RNA expression, omics-level characterization of cell-type-specific protein expression remains a technical challenge. Bulk mass spectrometry (MS) provides abundant proteomics resources to infer cell-type specificity via data deconvolution; however, it is unclear which proteomic quantification formats are optimal, as they differ from the data types for which most deconvolution methods were designed. Here, leveraging recently generated large-cohort proteogenomics data, we systematically evaluated different MS proteomics quantification formats and preprocessing strategies to resolve cell-type-specific protein expression. Our results indicate that while label-free spectral counts can be used directly, TMT MS1 intensities and MS2 ratios are less suitable and require appropriate data transformation. We demonstrate that a 'min-score' transformation significantly improves MS1 intensity-based deconvolution, providing useful insights for subtyping pancreatic cancer. Moreover, we identified the coefficient of variation (CV) as a robust statistical indicator of deconvolution suitability. Finally, we developed "ProTransDeconv", an R package integrating data transformation, deconvolution, and quality checks for major MS proteomics data formats. This work provides practical guidance for deconvolving bulk proteomics to study cell-type-specific protein-level dysregulation.



KEYWORDS: *deconvolution, proteomics, cancer protein markers, intratumor heterogeneity*

INTRODUCTION

Intratumoral heterogeneity refers to the diverse cellular composition of bulk tumors. In addition to malignant tumor cells, a significant portion of the tumor mass consists of infiltrating immune cells, stromal cells, endothelial cells, and other cell types.[1] Understanding the interactions among these tumor components is essential for improving or innovating tumor treatments. For example, checkpoint inhibitor-based immunotherapy aims to enhance the cytotoxic effects of infiltrating T cells against malignant tumor cells.[2] However, factors that may compromise treatment efficacy can originate from various components of the tumor microenvironment, which remains incompletely characterized.[3]

To better dissect intratumoral heterogeneity, emerging techniques, such as single-cell RNA sequencing (scRNA-seq), have been developed to profile cell-type-specific gene expression at single-cell resolution. Although these techniques provide tumor portraits at an unprecedented level of detail, their results lack critical insights into protein-level regulation. Proteins constitute the majority of diagnostic biomarkers and druggable targets, and thus knowledge of their cell-type-specific mechanisms has significant clinical value. For instance, it has been shown that TGFB1 expressed in tumor epithelial cells has antitumor roles but exerts protumor effects when expressed in stromal cells.[4] Accumulating studies have shown that protein-level dysregulation is common in tumors and that mRNA abundances do not fully reflect protein expression.[5] These rationales underscore the importance of directly detecting and quantifying cell-type-specific proteins.

Mass spectrometry (MS)-based proteomics provides the most comprehensive protein identification and quantification. However, unlike RNA, protein signals cannot be amplified. The substantial MS sample input requirement thus makes single-cell proteomic profiling difficult in achieving sufficient coverage and depth to systematically study protein-level biology.[6] Other experimental techniques, such as laser-capture microdissection (LCM), are capable of enriching specific tissues or cell groups for MS analysis, but they are labor-intensive and have limited applications in large-cohort cancer studies.[7] For instance, LCM-MS has been commonly used to study pancreatic ductal adenocarcinoma (PDAC), a cancer type characterized by low cellularity and complex intratumoral heterogeneity.[8,9] However, the most comprehensive of these

Received: October 9, 2025
Revised: November 17, 2025
Accepted: November 26, 2025
Published: December 2, 2025

studies only utilized a medium-sized cohort (32 patients),[9] which is substantially smaller than those used in bulk-level profiling, such as the study performed by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) involving 140 patients.[10]

An alternative approach to circumvent these technical challenges is data deconvolution, which infers tumor composition and/or cell-type-specific gene expression from bulk gene expression data.[11] The underlying mathematical principle is that bulk gene expression (or other molecular signals) represents a linear combination of the expression from constituent cell types, weighted by their proportions. Specifically, for a given gene:

$$E_{\text{bulk}} = \sum_i (E_i \times P_i)$$

where E_{bulk} represents the bulk expression, and E_i and P_i denote the cell-type-specific expression and population proportion for each constituent cell type i, respectively.

While the inference of cell populations can be readily conducted using RNA-seq[12] or DNA methylation data,[13] the unique role of proteomics data in the deconvolution lies in obtaining the cell-type-specific protein expression. This approach is particularly valuable because large cohort studies typically have detailed clinical records (e.g., tumor stage, patient prognosis) and multiomics profiles.[14,15] Consequently, the cell-type-specific protein expression derived from data deconvolution can provide refined biological and clinical insights through multiomics integration and clinically guided prioritization.

Unlike the count-based quantification in RNA-seq, MS proteomics data are quantified using various methods, such as MS2 spectral counts and MS1 intensity for label-free MS, and MS1 intensity and MS2 ratios for iTRAQ- or TMT-based labeled MS.[16] It remains unclear which of these methods, if any, are suitable for data deconvolution to study cell-type-specific protein expression or modification. Notably, previous studies on RNA-seq-based deconvolution have suggested that data format is a critical factor affecting deconvolution performance.[17] Given the greater diversity of data formats in MS proteomics, driven by variability in MS techniques and data processing workflows, a systematic comparison of deconvolution performance is essential to determine which data type is optimal for this application. Furthermore, most deconvolution algorithms for obtaining cell-type-specific expression were developed for RNA-seq.[18] Therefore, it is imprudent to assume that arbitrary proteomics data formats can be directly used in these algorithms.

In this study, we conducted a comprehensive evaluation of deconvolution suitability for protein quantification formats widely used in recent large-scale proteogenomic studies, such as those from the CPTAC[14] and International Cancer Proteome Consortium (ICPC).[19] We focused on these studies because they have generated the most abundant data resources for protein expression and quantification in cancer research. Additionally, they provide matched genomic and transcriptomic data, enabling orthogonal validation of some of our inferences. For proteomic quantification formats less suitable for deconvolution, we propose several data transformation methods to enhance data linearity. We envision that our study will provide general guidance for leveraging these proteogenomic data to detect cell-type-specific protein biomarkers and druggable targets, as well as to understand protein-level pathway activities and intercellular crosstalk.

MATERIALS AND METHODS

Datasets

The harmonized ccRCC[20] and PDAC[10] bulk TMT proteomics and transcriptomics datasets, along with and patient clinical data were downloaded from the CPTAC Pan-Cancer Data Portal[14] (https://proteomic.datacommons.cancer.gov/pdc/cptac-pancancer). The CRC label-free proteomic dataset was downloaded from its original publication,[21] and the corresponding transcriptomics data were downloaded from the GDC data portal (https://portal.gdc.cancer.gov/). The RNA expression data from 81 cell types were downloaded from the Human Protein Atlas (HPA)[22] (https://www.proteinatlas.org/download/tsv/rna_single_cell_type_cell_types.tsv.zip). The immunohistochemistry (IHC) protein expression data from 149 cell types were also downloaded from HPA (https://www.proteinatlas.org/download/tsv/normal_ihc_cell_types.tsv.zip). The accession IDs for the reference methylation datasets used in first-phase deconvolution are listed in Supplementary Table 1.

Inferring Proportions for Constituent Cell Types

Methylation-based deconvolution to infer cell-type proportions was performed using EDec.[23] The DNA methylation references used for each cell type were generated from purified cell groups or cell lines (listed in the Supplementary Table 1). Bulk tumors were deconvolved into three cell types for colorectal cancer (CRC) and clear cell renal cell carcinoma (ccRCC): immune cells, stromal cells, and epithelial cells. For pancreatic ductal adenocarcinoma (PDAC), bulk tumors were deconvolved into an additional cell type: exocrine and endocrine cells. RNA-expression-based abundances of immune cells and stromal cells were inferred using ESTIMATE.[24] For CRC, whole-exome sequencing (WES)-based tumor purity was inferred using ABSOLUTE and obtained from the GDC Data Portal via TCGAbiolinks.[25] For ccRCC and PDAC, whole-genome sequencing (WGS)-based tumor purity was inferred using CNVEX (https://github.com/mctp/cnvex) and obtained from the CPTAC Pan-Cancer Data Portal.

To further evaluate the reliability of cell-type proportion estimates from deconvolution, we analyzed microarray-based RNA expression data from flow-cytometry-sorted cell subpopulations in colon cancer (GSE39396).[26] Cell-type-specific marker genes for immune cells (CD45+), stromal cells (FAP+), and epithelial cells (EpCAM+) were identified using limma,[27] with a fold-change cutoff >4 and an adjusted p-value <0.01 (Benjamini-Hochberg procedure).

Selection of General Cell-Type-Specific Reference Proteins

General cell-type-specific protein markers were selected based on three levels of evidence: (1) RNA-level specificity identified from cross-tissues comparisons; (2) strong protein-RNA correlation, ensuring that the RNA-level specificity can be translated to the protein level. (3) Direct protein-level specificity identified by immunofluorescence or immunochemistry, as reported by HPA.

First, RNA expression data for 81 purified cell types, downloaded from the Human Protein Atlas (HPA), were used to identify cell-type-specific marker genes at the RNA level. The original cell types described by HPA were reclassified into three meta-cell types based on their annotations: "Blood &

https://doi.org/10.1021/acs.jproteome.5c00980
J. Proteome Res. 2026, 25, 471−483

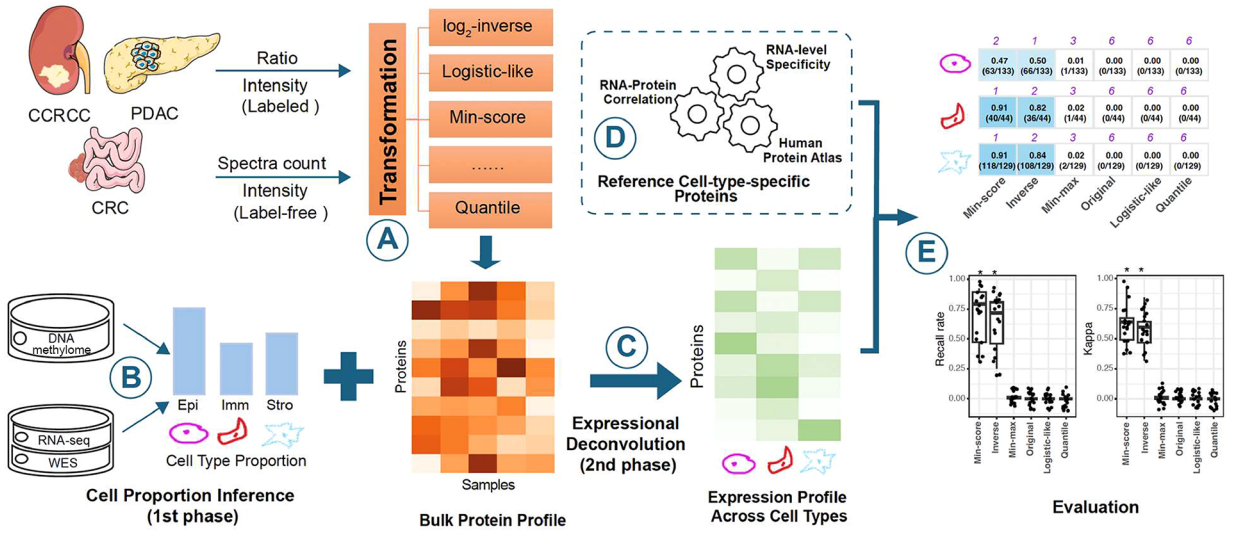


Figure 1. Schematic of the study workflow. Proteomics data (TMT ratio and intensity; label-free spectral count and intensity) from three cancer types (ccRCC, PDAC, and CRC) undergo various data transformations (A). In parallel, tissue compositions—proportions of epithelial cells, immune cells, and stromal cells (Epi, Imm, and Stro)—are inferred from DNA methylation data and cross-validated using other omics platforms (B). Multiple deconvolution methods (csSAM, EDec, Rodeo, and bMIND) are then used to estimate cell-type-specific protein expression (C). Finally, the deconvolution results are evaluated against a set of reference protein markers (derived from multiple lines of evidence, D) using specific evaluation metrics (E).

Immune Cells" was relabeled as "immune cells"; "Endothelial Cells" and "Mesenchymal Cells" were merged into "stromal Cells"; and "Glandular Epithelial Cells," "Squamous Epithelial Cells," and "Specialized Epithelial Cells" were merged into "epithelial cells". For each meta-cell type, marker genes were selected using a fold-change cutoff of >4 and an adjusted p-value <0.01 (t test on $\log_2$-transformed TPM).

Second, these RNA-level marker genes were filtered to include only those exhibiting high protein-RNA correlation in each proteogenomic dataset (Spearman's correlation coefficient >0.5 and adjusted $p < 0.001$). This list was further refined using HPA data to select proteins with known IHC-based cell-type specificity. For the IHC annotations, "Lymphoid Tissue," "Lymphocytes," and "Hematopoietic Cells" were classified as immune cells; "Stromal Fibroblasts" and "Stroma" were classified as stromal cells; and all nine epithelial cell types were classified as epithelial cells. Based on the semi-quantitative IHC results, "High" or "Medium" staining signals were considered positive, while "Low" or "Not Detected" was considered negative. Final protein markers were required to be positive in their corresponding cell type and negative in all others.

Cancer-Type-Specific Reference Proteins

If a protein is specific to a particular cell type, its expression in a bulk tumor sample is primarily driven by the proportion of that cell type. Therefore, a positive correlation is expected between a cell type's proportion and its specific proteins' bulk expression levels. Based on this assumption, we identified cancer-type-specific protein markers. For each protein, we calculated Spearman's correlation between its bulk-level expression and the proportions of immune, stromal, or epithelial cells derived from the first-phase deconvolution. For each of the three cell types, we selected the top 25% of proteins that were most highly correlated and also statistically significant (Benjamini–Hochberg adjusted $p < 0.05$). Finally, we removed the general cell-type-specific marker proteins from these lists, leaving a refined set of cancer-type-specific reference proteins.

Second-Phase Deconvolution and Evaluation

EDec (the second-phase functionality), Rodeo,[28] csSAM,[29] and bMIND[30] were used to perform deconvolution to obtain mean cell-type-specific expression for each protein. The inputs for these second-phase deconvolution tools are bulk protein expression matrices (with different formats) and the proportions of constituent cell types derived from the first-phase deconvolution. The bulk protein expression matrices were filtered to keep the proteins with fewer than 30% missing values. From the deconvolution results, a protein was designated as cell-type-specific if its expression in that cell type was greater than the combined expression from all the other cell types. In other words, the specificity defined by the following formula is greater than 0.5.

$$\text{Specificity} = \frac{x_i}{\sum_{i=1}^{N} xi} \tag{1}$$

where x_i represents the mean expression of protein i in a given cell type, and N is the total number of cell types.

The recall rate and Cohen's kappa value as described in eqs 2 and 3 were used to evaluate the suitability of various MS proteomics quantification formats.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

where TP denotes true positive and FN denotes false negative prediction based on reference cell-type-specific proteins.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{3}$$

where P_o represents observed agreement between inferred and reference cell-type-specific proteins, and P_e represents the expected agreement by chance.

https://doi.org/10.1021/acs.jproteome.5c00980
J. Proteome Res. 2026, 25, 471−483

Identification of Subtype-Specific Clinical Marker Proteins in PDAC

The TMT intensity CPTAC PDAC proteomics data were transformed using the "min-score" method and deconvolved using the second-phase functionality of EDec. Subsequently, cell-type-specific marker proteins were selected based on eq 1. The prognostic value of a protein was then calculated using the *C*-index (or concordance index)[31] approach implemented in the R package "survcomp". The *C*-index provides a single, easily interpretable value to assess the association between a continuous variable (e.g., protein expression) and right-censored data (e.g., patent survival). A *C*-index <0.5 indicates a trend toward a favorable prognosis, while a *C*-index >0.5 indicates a trend toward an adverse prognosis. For selected proteins, their power to stratify patients into good and poor survival groups was further determined by a log-rank test. Gene ontology (GO) biological process enrichment was performed utilizing the specific group of proteins (e.g., stromal proteins with adverse prognostic values) against the background of all detected proteins using WEBGESTALT.[32]

Data Availability

All the datasets used for this study were publicly available via the provided downloaded links and GEO IDs. All the data analysis was performance in R (version 4.2.0). The R package proTransDeconv is released under an MIT license and is freely available at https://github.com/HuangLabAtUAB/ProTransDeconv.

RESULTS

Overall Study and Workflow

Starting from bulk gene expression (or other molecular profiles), deconvolution can be applied first to infer tumor composition (the first phase) and subsequently to estimate cell-type-specific gene expression (the second phase). Our study focuses on the second phase, which aims to identify cell-type-specific proteins to enhance understanding of protein-level dysregulation and inform the development of protein biomarkers and druggable targets. We used omics data types other than proteomics (e.g., DNA methylation, whole-exome sequencing [WES], and transcriptomic data) for the first phase because these data provide more features, have well-established deconvolution methods, and are commonly available in proteogenomic studies. In essence, the primary value of proteomics-based deconvolution lies in the second phase, which we systematically evaluated here (Figure 1).

We primarily focused on the CPTAC datasets,[14] which represent the most comprehensive large-cohort cancer proteomics data and include both label-free and labeled MS proteomics. Label-free MS proteomics was quantified using MS2 spectral count and MS1 intensity. Labeled MS proteomics, such as Tandem Mass Tag (TMT) MS, was quantified by MS2 ion reporter ratio or by recently developed MS1 intensity approaches.[14] In addition to the publicly available data formats, we applied several methods commonly used for data transformation or normalization (Table 1). Notably, since most of the quantitative proteomics analyses use nonparametric statistical methods, such as the Wilcoxon test for identifying differential protein expression and Spearman's correlation for determining protein-RNA correlation, our data transformations will not affect any existing data-driven results and can be broadly applied to proteomics datasets if necessary.

Table 1. Data Transformation Methods Applied to Proteomics Datasets in This Study

method	equation
inverse[a]	$x_{ij}^{\text{trans}} = 2^{x_{ij}}$
logistic-like[a]	$x_{ij}^{\text{trans}} = 1 - e^{-ax_{ij}};\ a = \frac{1}{\max(x_i)}$
min-max[a]	$x_{ij}^{\text{trans}} = \frac{x_{ij} - \min(\boldsymbol{x}_i)}{\max(\boldsymbol{x}_i) - \min(\boldsymbol{x}_i)}$
min-score[a]	$x_{ij}^{\text{trans}} = \frac{x_{ij} - \min(\boldsymbol{x}_i)}{sd(\boldsymbol{x}_i)}$
tanh[a]	$x_{ij}^{\text{trans}} = \frac{\tanh(x_{ij}) + 1}{2}$
ratio-shift[a]	$x_{ij}^{\text{trans}} = x_{ij} - min\ (\boldsymbol{x}_i)$
quantile[a,b]	$x_{ij}^{\text{trans}} = \mu_{r_{ij}}$

[a] x_{ij} and x_{ij}^{trans} denote the original and transformed expression value of protein i in sample j, respectively. $\boldsymbol{x}_i$ denotes the original expression vector of protein i across all the samples. [b] In quantile normalization, the gene expression values in each sample are first sorted in ascending order. r_{ij} denotes the rank of protein i in the sorted values of sample j, and $\mu_{r_{ij}}$ denotes the mean of the r_{ij}-th smallest values across all samples.

To balance interpretability and model complexity, our deconvolution focuses on three high-level cell types in solid tumors: immune cells, stromal cells, and epithelial cells. Our main objective was to determine which MS proteomics quantification formats and data transformation methods are suitable for robust inference of cell-type-specific protein expression, given the proportions of these constituent cell types.

We used the CPTAC omics data generated from three cancer types with pronounced intratumoral heterogeneity: colorectal cancer (CRC, n = 95),[21] clear cell renal cell carcinoma (ccRCC, n = 103),[20] and pancreatic ductal adenocarcinoma (PDAC, n = 140).[10] These cohorts yielded 8763 (ccRCC), 8663 (PDAC), and 4054 (CRC) quantified proteins, respectively, and importantly included diverse MS proteomics quantification formats.

To date, most of the deconvolution tools for the second phase have been based on least-squares regression (LSR). We included three tools with variable LSR implementations: (1) csSAM, which is based on a standard LSR;[29] (2) EDec (the second-phase functionality), which is based on a non-negative constrained LSR;[23] and (3) Rodeo, which is based on robust linear regression.[28] In addition, we included bMIND, which is on Bayesian estimation rather than LSR.[30] This complex design ensures that our conclusions are not biased toward a specific cancer type or a deconvolution algorithm but instead reflect the general suitability of the proteomics data formats (Figure 1).

Inference of Tumor Composition and Collection of Reference Cell-Type-Specific Marker Proteins

For the first phase of deconvolution, we used EDec (the first-phase functionality),[23] a DNA methylation-based deconvolution method to infer the proportions of constituent cell types. Additionally, we used ESTIMATE (RNA-based),[12] ABSOLUTE (WES-based),[33] and mutation allele frequency (MAF, WES-based) for cross-omics validation. The inferred proportions and abundances for different cell types were consistent among these orthogonal omics data platforms (Figure 2). For instance, ESTIMATE scores for tumor-infiltrated immune and stromal cells were highly correlated with corresponding cell proportions inferred by EDec (Pearson's correlation 0.78 and

https://doi.org/10.1021/acs.jproteome.5c00980
J. Proteome Res. 2026, 25, 471−483

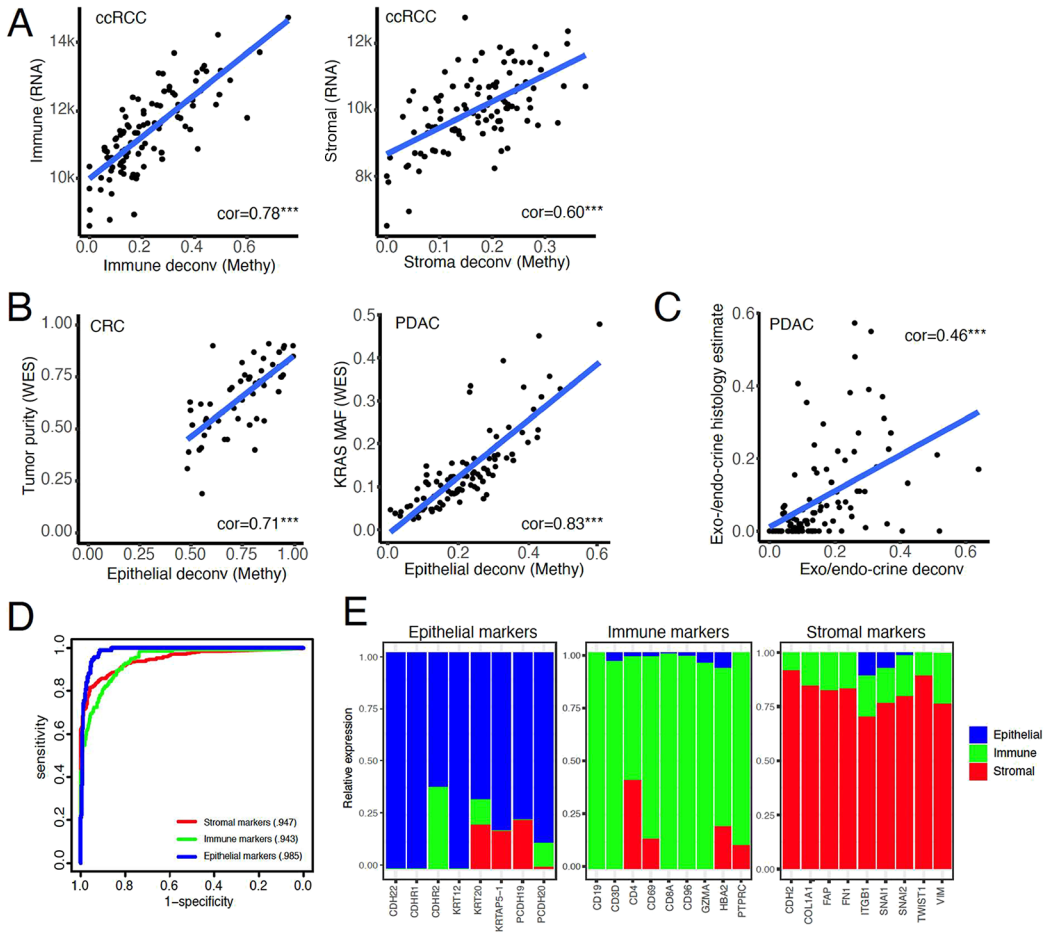


Figure 2. Cross-validation of inferred constituent cell proportions. (A) Correlation between DNA methylation-based cell proportions and RNA-seq-based cell abundance for immune cells (left) and stromal cells (right) in ccRCC. (B) Correlation between the DNA methylation-based tumor epithelial cell proportion and WES-based tumor purity (left) in CRC or *KRAS* mutated allele frequency (MAF) in PDAC (right). (C) Correlation between DNA methylation-based and histologically estimated exocrine/endocrine cell proportions in PDAC. (D) AUROC curves for CRC cell-type-specific markers resulting from deconvolution. (E) Bar plots showing cell-type-specific expression for representative epithelial (left), immune (middle), and stromal (right) genes.

0.60, respectively, $p < 0.001$, Figure 2A). Similarly, the somatic mutation allele frequency (MAF) of *KRAS*, a known clonal mutation in pancreatic ductal adenocarcinoma (PDAC),[10] correlated with the tumor epithelial cell fraction inferred by EDec (Pearson's correlation 0.83, $p < 0.001$) (Figure 2B). Furthermore, the EDec deconvolution results aligned with histology-based tumor composition assessments, even if they were generated from different sample aliquots (Pearson's correlation 0.46, $p < 0.001$) (Figure 2C).

Moreover, we tested the feasibility of inferred cell proportions in performing the second phase of deconvolution on transcriptomics data, which yields cell-type-specific RNA expression. For the colorectal cancer, where RNA-seq from physically purified cell groups is available to serve as ground truth,[26] we found that the cell-type specificity from deconvolution is highly accurate (AUROC > 0.95, Figure 2D). The expression patterns of well-known epithelial, immune, and stromal genes were readily reproduced by deconvolution (Figure 2E). Together, these results suggest that cell proportions inferred from the methylation data, which are commonly available in the CPTAC and ICPC datasets, are well-suited for performing the second-phase deconvolution of whole bulk proteomics data.

A systematic evaluation of cell-type-specific protein expression requires known protein markers for constituent cell types as ground truth. However, unlike RNA profiling, genome-wide cell-type-specific protein expression is currently unavailable. To overcome this challenge, we defined protein markers using evidence orthogonal to the deconvolution process itself. Specifically, a protein was classified as a cell-type-specific marker if it satisfied the following criteria: (1) its corresponding RNA expression level is highly and selectively expressed in one cell type relative to others in purified cell populations, and (2) its protein abundance strongly correlates with mRNA levels in large-cohort proteogenomic datasets. For the first criterion, we obtained cell-type-specific transcriptome profiles from flow cytometry-sorted cell groups.[22] Notably, these purified cell groups were derived from various bulk tissue types, enhancing the generalizability of these markers across different cancer types. The second criterion ensures that protein abundance is primarily determined by RNA expression rather than variability in protein translation or stability. Finally, we confirmed the cell-type specificity of the protein markers used in our evaluation by checking them against antibody-detected expression patterns from the Human Protein Atlas (HPA)[22] (Figure 1 and Supplementary Figure 1, see also Materials and Methods and Supplementary Table 2).

https://doi.org/10.1021/acs.jproteome.5c00980
J. Proteome Res. 2026, 25, 471−483

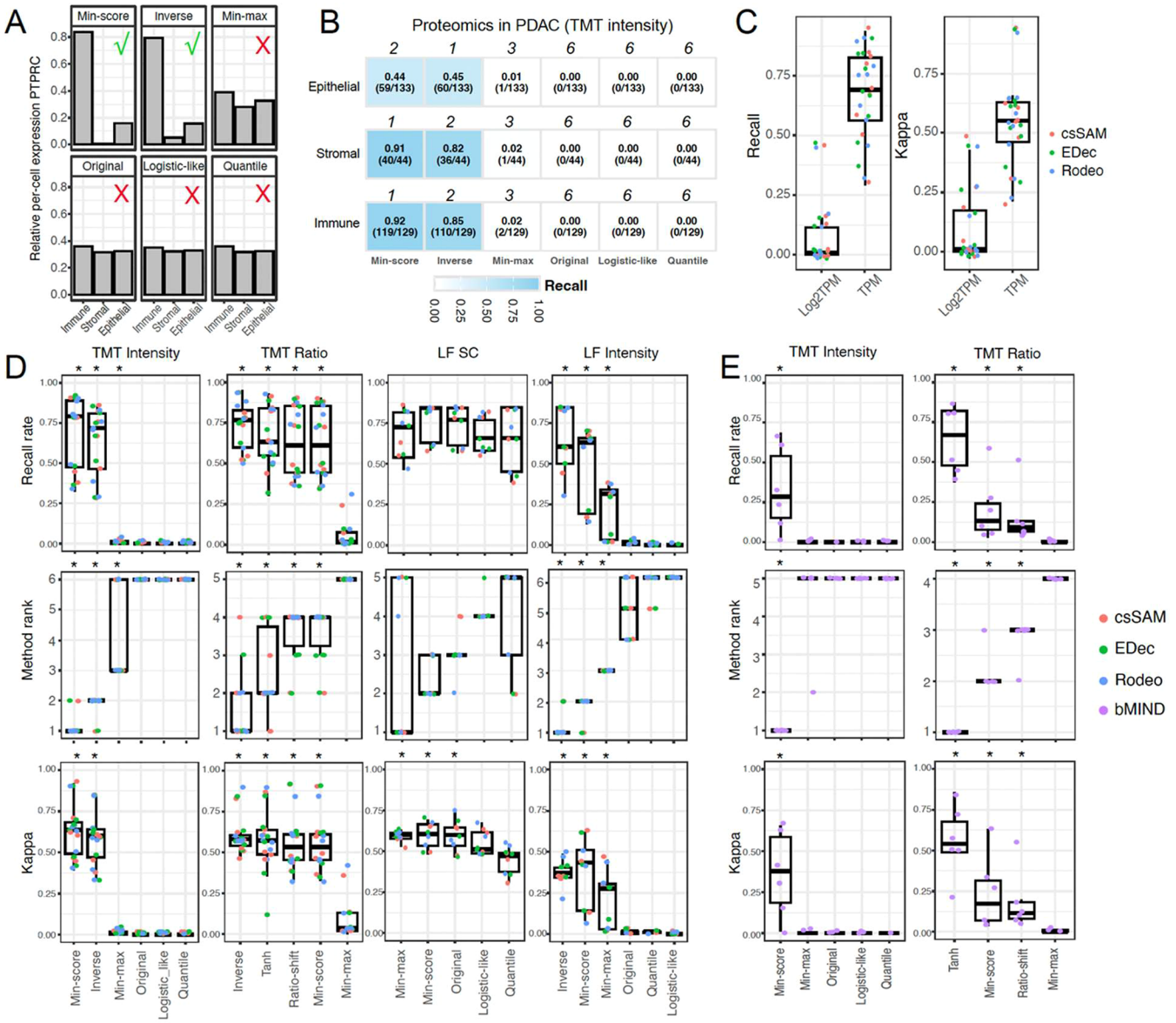


Figure 3. Performance evaluation based on general cell-type-specific protein markers. (A) Cell-type-specific protein expression of the pan-immune cell marker *PTPRC* resulting from csSAM deconvolution of different proteomics data formats. The green tick or red cross indicates success or failure in identifying this protein as immune-cell-specific. (B) Heatmap showing recall rates for csSAM-based deconvolution using PDAC TMT intensity proteomic datasets. Numbers in parentheses indicate the recalled and total protein counts, respectively. The ranks of recall rates across different data formats are listed above the heatmap. (C) Recall rates and kappa values for deconvolution results using transcript-per-million (TPM) or $\log_2$ TPM of RNA-seq data. (D) Boxplots showing recall rates (top), ranks of recall rates (middle), and Cohen's kappa values (bottom) resulting from LSR-based deconvolution using different proteomics data formats. For each proteomics data format, metrics were collected across various deconvolution methods, cancer types, and cell types. (E) Same as (D), but results were generated from bMIND using compatible data formats. For (D) and (E), asterisks denote significantly superior performance (*p < 0.001, Wilcoxon rank-sum test and median recall rate difference >0.2) for the indicated group relative to the lowest-performing group. LF, label free; SC, spectral count.

Performance Evaluation Based on General Reference Cell-Type-Specific Proteins

We next performed the second-phase deconvolution using the algorithms described above. These methods take the proportions of constituent cell types and bulk protein expression as input to solve for cell-type-specific protein expression. With the orthogonally derived reference marker proteins as a benchmark, we evaluated the suitability of the bulk proteomics data from various MS quantification and data transformation methods for deconvolution. Since reference proteins were identified using purified cells from multiple tissues, their cell-type specificity was expected to be conserved across different cancer types and to be recalled in the deconvolution results. Accordingly, the primary performance metric was the recall rate of ground-truth cell-type-specific proteins. For instance, the protein expression of *PTPRC*, an immune cell marker gene, can be successfully recalled as immune cell-specific by both the deconvolution based on min-score- and inverse-transformed proteomics data (Figure 3A). To account for multiple cell types and cancer types, we also recorded the ranking of recall rates across different data formats (Figure 3B). Additionally, we adopted the Cohen's kappa value as a third metric,[34] which measures the ability of deconvolution results to distinguish proteins with different cell-

https://doi.org/10.1021/acs.jproteome.5c00980
J. Proteome Res. 2026, 25, 471−483

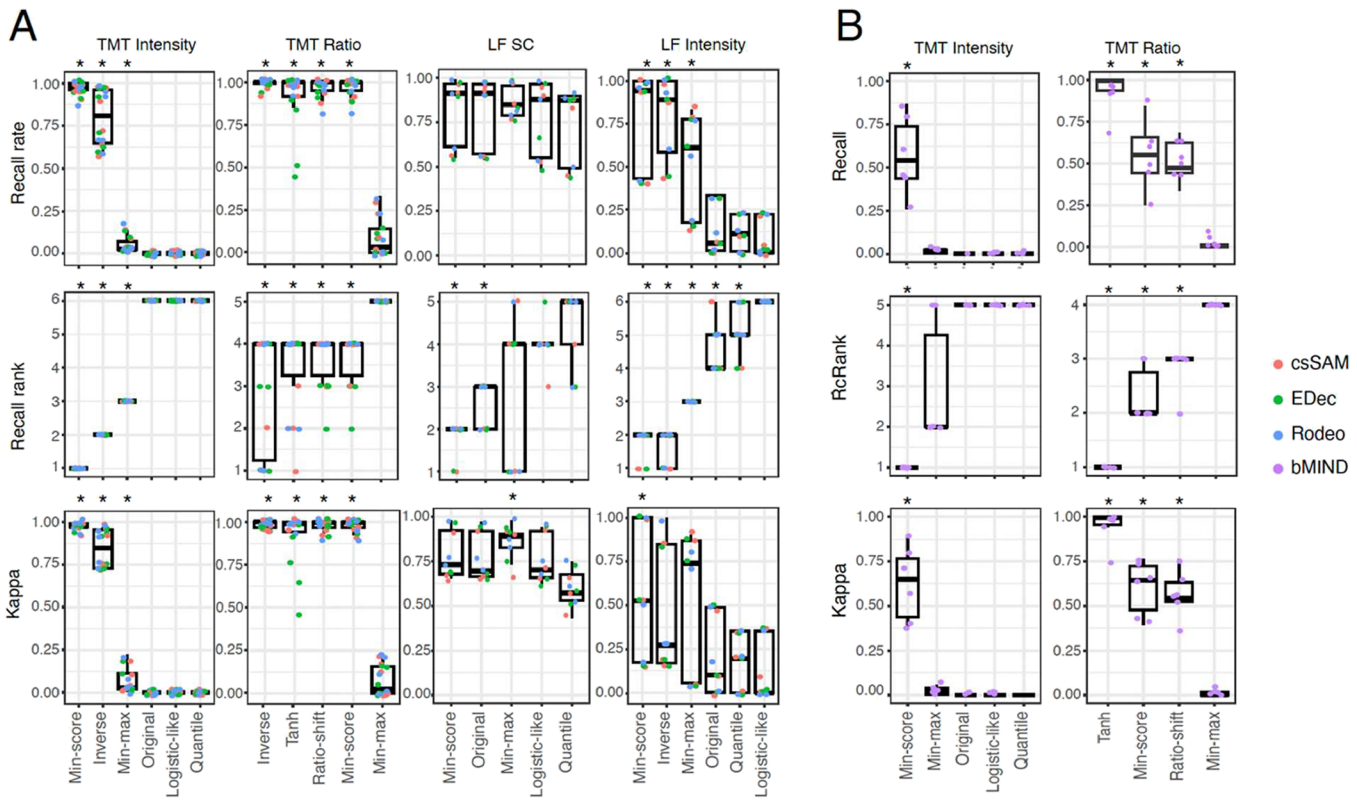


Figure 4. Performance evaluation using cancer-specific cell-type-specific protein markers. (A) Boxplots showing recall rates (top) of cancer-specific marker proteins, ranks of recall rates (middle), and Cohen's kappa values (bottom) resulting from LSR-based deconvolution using different proteomics data formats. For each proteomic data format, metrics were collected across various deconvolution methods, cancer types, and cell types. (B) Same as (A), but results were generated from bMIND using compatible data formats. Asterisks denote significantly superior performance (*p < 0.01, Wilcoxon rank-sum test and median recall rate difference >0.2) for the indicated group relative to the lowest-performing group. LF, label free; SC, spectral count.

type specificity (see Materials and Methods). Because recall primarily focuses on true positives, Cohen's Kappa provide a complementary metric by reflecting overall agreement across all classes while accounts for probability attributable to random chance.[35] As a pilot test, we performed deconvolution on RNA-seq data. Across all methods and cancer types, TPM at the linear scale consistently outperformed log2 scale (median recall rate, MRR: 69.2% vs 0.009%, p < 0.001, Wilcoxon rank-sum test), aligning with previous findings[18] and confirming the utility of both metrics for deconvolution evaluation (Figure 3C).

Overall, the LSR-based deconvolution methods showed better data compatibility and similar performance across different cancer and cell type combinations (≥81% pairwise overlap of recall rank-based results across all three tools; p < 0.001, Fisher's Exact test; Supplementary Figure 2A), thus providing robust results for evaluating different data formats. First, TMT log2 intensity, a format widely used for recent CPTAC proteomics quantification,[14] exhibited poor performance and could identify few or no reference cell-type-specific proteins across all deconvolution tools (MRR < 5%). The log2 inverse by the "inverse"-transformation could effectively address this issue (MRR 71.9%) and is the second-best data transformation method for this data type, although the wide range of the values was not compatible with bMIND. The best deconvolution suitability was achieved by the 'min-score' transformation for both LSF-based methods (MRR 79.1%) and bMIND (MRR 28.6%) (Figure 3D,E).

The TMT log2 ratio, also widely used, was incompatible with deconvolution due to the presence of negative values. Among transformation methods, the "inverse" transformation yielded the best performance with LSR-based tools (MRR 76.6%), while the "tanh" transformation performed best with bMIND (MRR 66.8%). For label-free MS proteomics, original spectral counts showed inherent suitability for deconvolution (MRR 76.9%), and the "min-score" transformations further improving performance (MRR 81.6%). Similar to TMT log2 intensity, label-free log2 intensity was unsuitable for deconvolution but could be improved by "inverse" or "min-score" transformations (MRR 60.5% and 63.2%, respectively) (Figure 3D,E).

Evaluation using Cohen's kappa yielded conclusions consistent with those from recall rates, with a Pearson's correlation of 0.96 between the two metrics ($p < 2.2 \times 10^{-16}$). For example, the median kappa value for deconvolution using TMT log2 intensity was 0.03, compared to 0.68 after "min-score" transformation (p < 0.01, Wilcoxon rank-sum test). Recall rate comparisons for different transformation methods were consistent across individual cell types (≥75% pairwise overlap of recall rank-based results; p < 0.001, Fisher's Exact test, Supplementary Figure 2B). Collectively, these results indicate that specific transformations are required for optimal deconvolution performance on different proteomics data formats, and that some transformations, such as "min-score", are broadly applicable across deconvolution algorithms.

https://doi.org/10.1021/acs.jproteome.5c00980
J. Proteome Res. 2026, 25, 471−483

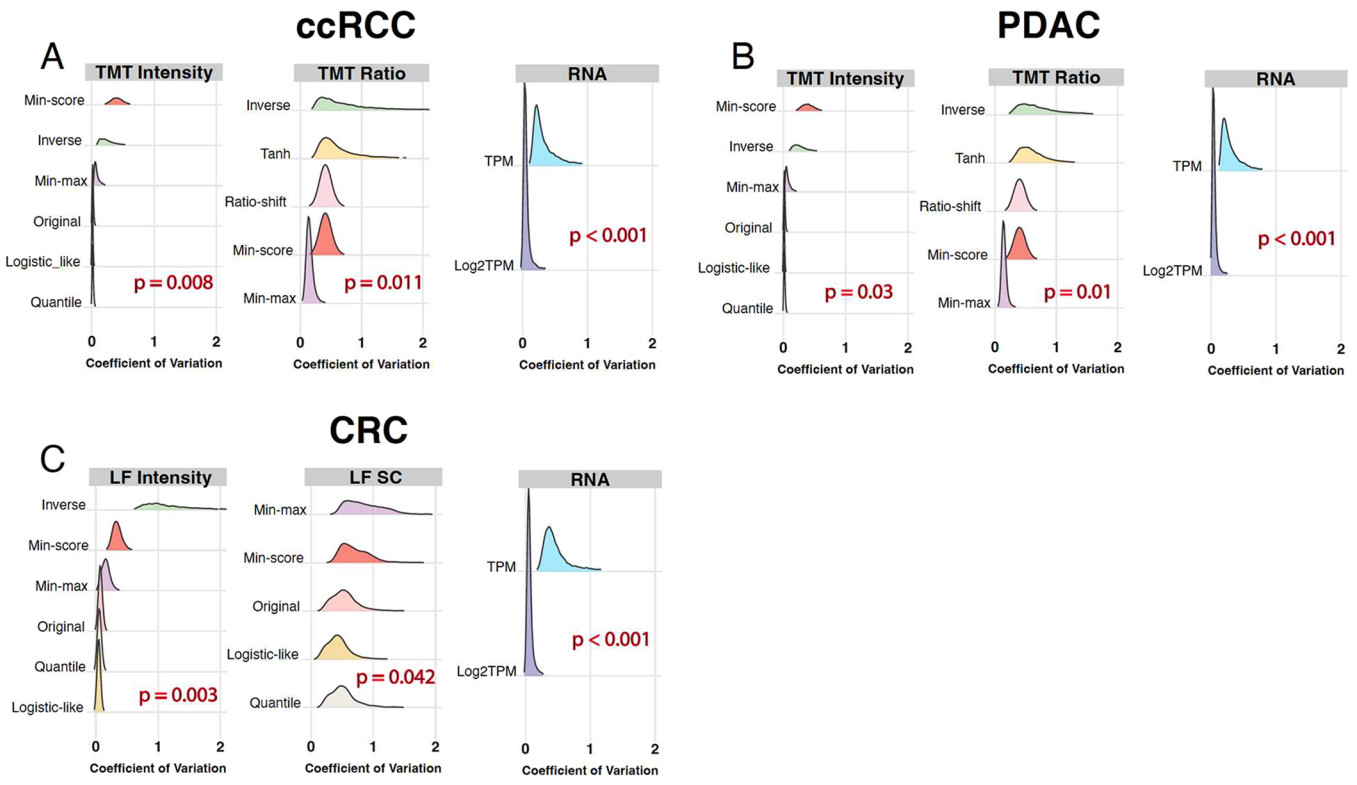


Figure 5. Association between coefficient of variation and deconvolution performance. Coefficient of variation (CV) distributions for different proteomics data formats in ccRCC using TMT intensity and TMT ratios (A), PDAC using TMT intensity and TMT ratios (B), and CRC using label-free intensity and spectral counts (C). Within each panel, data formats are ordered from top to bottom according to their performance in LSR-based methods. *P*-values were calculated using Jonckheere's trend test to assess the monotonic association between higher CV and better performance ranking. For comparison, CV distributions of transcript-per-million (TPM) and $\log_2$-transformed TPM RNA-seq data are included as references in each cancer cohort. LF, label free; SC, spectral count.

Performance Evaluation Using Cancer-Specific Reference Proteins

To strengthen the evidence derived from general cell-type-specific proteins, we investigated whether the same conclusions would hold if reference proteins were identified in a cancer-specific manner. We hypothesized that a protein specific to a particular cell type in a given cancer sample would have its bulk expression primarily driven by the proportion of that cell type. Accordingly, for each cancer type, we selected cell-type-specific reference proteins whose expression strongly correlated with the corresponding cell type proportion (see Materials and Methods). Interestingly, a substantial number of cancer-specific reference proteins remain after we remove the overlap with the general reference (median number: epithelial 43, immune 80, stromal 106; Supplementary Table 2), suggesting their involvement in cancer-specific biological processes. We then investigated whether the deconvolution results could recapitulate these marker proteins.

Most quantification formats and transformation methods exhibited performance consistent with our previous findings using general cell-type-specific proteins (Figure 4A,B). For TMT log2 intensity, the deconvolution using "min-score"-transformed or "inverse"-transformed data formats achieved the best and second-best performance across all deconvolution methods (MRR 98.4% and 81.0%, respectively). For the TMT log2 ratio, the "inverse" transformation performed best with LSR-based methods, while "tanh" was best for bMIND (MRR 100% and 99.5%, respectively). For label-free spectral counts, the min-score-transformed data and original count data achieved the top two performances (MRR 91.2% and 90.1%, respectively). Finally, for label-free log2 intensity, transformation by "min-score" or "inverse log2" was required for optimal deconvolution (MRR 94.4% and 88.9%, respectively).

Coefficient of Variation as an Indicator of Deconvolution Suitability

Given the diverse formats for proteomic data quantification, we investigated if there is a straightforward statistical metric that could indicate a dataset's suitability for deconvolution. We evaluated several descriptive metrics derived from large-scale omics data to assess their ability to distinguish high-performing quantification formats (or transformation methods) from less effective ones.

Among all the statistical descriptors examined, including variance, coefficient of variation (CV), distribution skewness, and median absolute deviation (MAD), only the CV consistently exhibited a clear monotonic relationship with deconvolution performance ($p < 0.05$, Jonckheere's trend test). Higher CV ranges were strongly associated with better suitability for deconvolving cell-type-specific proteins (Figure 5A–C and Supplementary Figure 3). For instance, in the TMT intensity-based proteomics, the distribution range was greatest for the "min-score"-transformed data and was reduced in data formats with lower deconvolution suitability (Figure 5A,B). In comparison, the MAD values displayed similar ranges for "min-score"-, "quantile"-transformed data and the original data despite substantial differences in their deconvolution performance (Supplementary Figure 3B). Moreover, CV was higher in the linear RNA format (TPM) than in the log2 scale,

https://doi.org/10.1021/acs.jproteome.5c00980
J. Proteome Res. 2026, 25, 471−483

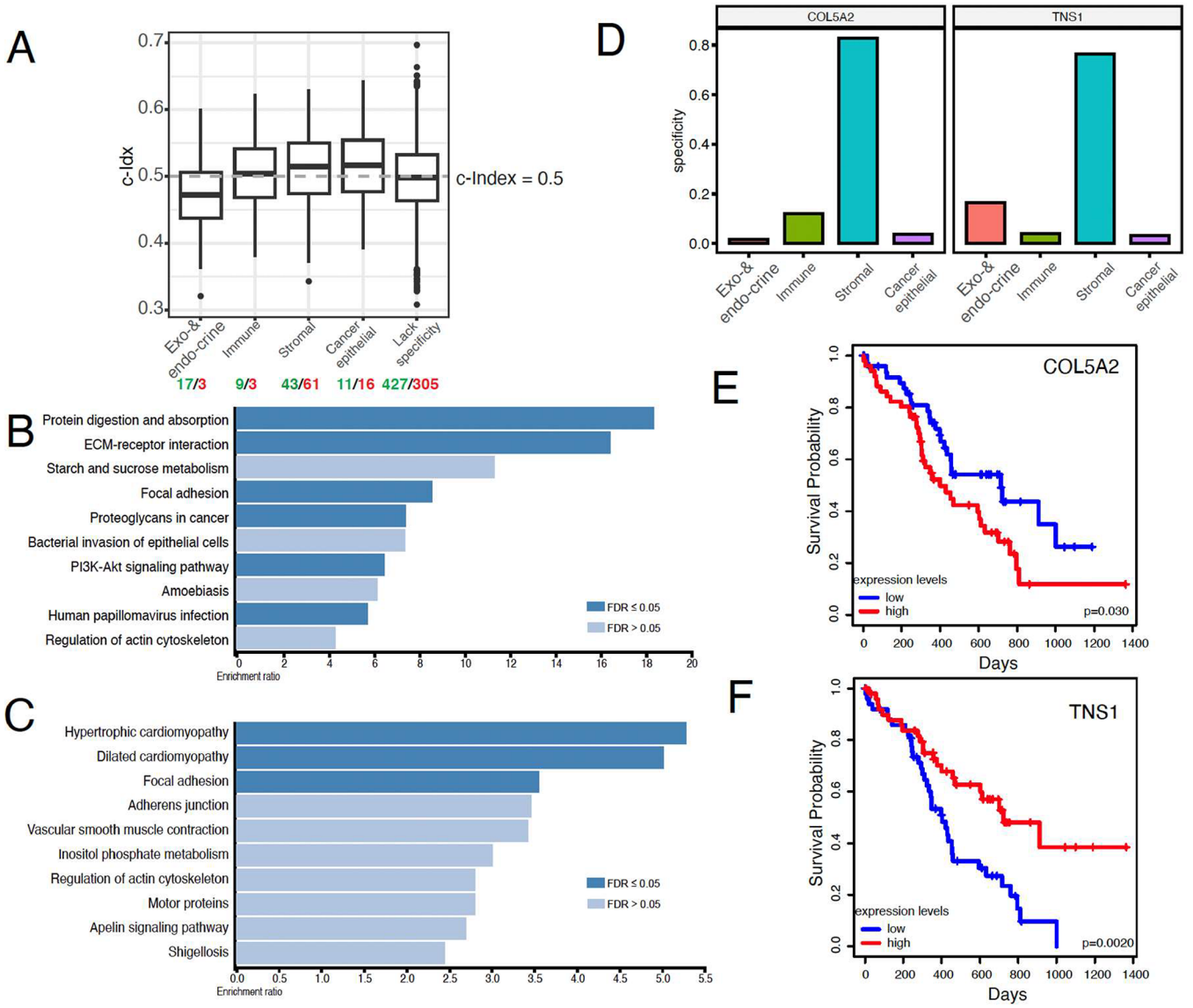


Figure 6. Protein-level deconvolution provides insights into PDAC subtyping. (A) Distribution of *C*-index values for cell-type-specific marker proteins identified by deconvolution. Numbers at the bottom indicate the counts of favorable (green) and adverse (red) prognostic protein markers for each cell type. (B and C) Top ten KEGG pathways enriched for stromal marker proteins associated with favorable (B) and adverse (C) prognoses. (D) Cell-type-specific protein expression profiles for the stromal marker proteins COL5A2 (left) and TNS1 (right). (E and F) Kaplan−Meier survival curves for patients stratified by median protein expression levels of *COL5A2* (E) and *TNS1* (F).

supporting its role in identifying the linear RNA format as more suitable for deconvolution (Figure 5A−C). The only partial exception was with the label-free spectral count data, where the "min-max"-transformed format had the largest CV distribution yet slightly lower median recall rate than 'min-score'-transformed format (Figure 5C). Nonetheless, the "min-max" still ranked among the top-performing methods for label-free proteomics, and the "min-score"-transformed format also exhibited a higher CV distribution compared to all other formats.

Taken together, our results suggest that normalized data variation across samples, which reflects biological heterogeneity in underlying tumor composition and captured by CV, is a critical property enabling effective second-phase deconvolution.

An R Package to Evaluate and Optimize Proteomic Data Deconvolution

Our results showed that while methods like "min-score" achieved top performance across multiple datasets, no single data transformation method can be universally applied to all proteomics data formats. To enable data deconvolution with customized data transformations, we developed an R package, "proTransDeconv", which implements the aforementioned data transformation, deconvolution, and evaluation functionalities for proteomic data to identify cell-type-specific proteins. The input for proTransDeconv consists of large-cohort proteomic data and a corresponding cell proportion matrix. The package performs optimal data transformations to enhance deconvolution performance and generates evaluation results based on general and cancer-specific recall rates (Supplementary Figure 4). proTransDeconv is released

https://doi.org/10.1021/acs.jproteome.5c00980
J. Proteome Res. 2026, 25, 471−483

under an MIT license and is freely available at https://github.com/HuangLabAtUAB/ProTransDeconv.

Cell-Type-Specific Protein Expression Sheds Novel Light on PDAC Subtyping

As proteins serve as major biomarkers and drug targets, understanding cell-type-specific protein expression has significant clinical value. Here, we present a case study using our optimized proteomic deconvolution methods to clarify the molecular subtyping of PDAC, focusing on two controversial issues. First, early studies identified an exocrine/endocrine-like subtype,[36] which some later studies dismissed as resulting from normal tissue contamination during cancer dissection.[37] Consequently, it remains unclear whether the abundance of exocrine/endocrine tissue and the definition of this subtype are clinically meaningful. Second, while stromal cells have been generally considered as a pro-tumor and antitreatment component,[38] a recent study identified a group of PDAC tumors characterized by high expression of specific endothelial cell (a stromal component cell group) marker genes also showed high immune cell infiltration and better survival, highlighting the need for a deeper understanding of stromal cell subsets.[10]

Using the large-scale data generated from the CPTAC PDAC cohort,[10] we first inferred proportions for four cell types (i.e., cancer epithelial, immune, exocrine and endocrine, and stromal) from methylation data with EDec (see Materials and Methods). Next, we performed proteomics deconvolution to identify cell-type-specific protein markers using the "min-score"-transformed TMT log2 intensity, which we determined to be optimal for this cancer type. Compared to the other categories of cell-type specificity, the expression of endocrine/exocrine cell markers was associated with a significantly lower C-index ($p < 0.001$, Wilcoxon rank-sum test), indicating a trend toward a favorable prognosis (Figure 6A). Moreover, among proteins with significant prognostic value (log-rank test $p < 0.05$), most endocrine/exocrine cell markers (17/20) were associated with better survival outcomes ($p < 0.05$, Fisher's exact test). These results suggest that the presence of endocrine/exocrine tissues indicates a less malignant tumor state, rather than merely reflecting sample contamination during cancer dissection.

In parallel, we identified both adverse (61/104) and favorable (43/104) prognostic markers among stromal proteins, indicating that heterogeneous stromal components contribute to the tumor microenvironment (Figure 6A and Supplementary Table 3). Pathway enrichment analysis further revealed two groups of stromal markers associated with distinct functional categories. The adverse ("bad") stromal proteins were linked to extracellular matrix (ECM) and proteoglycan activities, which are hallmarks of malignant cancer-associated fibroblasts (CAFs) and known to promote tumor growth and metastasis[39] (Figure 6B). Conversely, the favorable ("good") stromal markers were associated with smooth muscle activities, reflecting normal pancreatic tissue composition[40] (Figure 6C). This finding reinforces our previous hypothesis that a diminished normal stromal component is associated with tumor progression.[10] Two representative proteins from these distinct stromal functional categories are COL5A2, a collagen involved in ECM remodeling,[41] and TNS1, an actin-binding protein involved in muscle contractility.[42] Both are highly and specifically expressed in stromal cells, with levels more than fourfold greater than in any other cell type (Figure 6D). Despite this shared expression pattern, they have opposite associations with patient prognosis: higher COL5A2 expression is associated with poor survival (Figure 6E), whereas higher TNS1 expression corresponds to favorable survival (Figure 6F).

Thus, cell-type-specific protein expression provides a more detailed understanding of the functional roles of endocrine/exocrine tissue and distinct stromal functions, collectively enhancing PDAC subtyping to improve precision oncology and guide the development of targeted therapies.

DISCUSSION

Although omics data deconvolution approaches have been developed for over a decade, few have been tested or applied in the proteomics field. This gap represents an urgent, unmet need for three reasons. First, because proteins are major diagnostic markers and druggable targets, a cell-type-specific understanding has great translational value. Second, from a technical perspective, achieving a cell-type-specific understanding of proteins relies on deconvolution even more than it does for RNA. While single-cell proteomics has emerged as a powerful technique, its coverage is limited to hundreds of proteins—far fewer than scRNA-seq—and thus falls short of comprehensive profiling.[6] Moreover, the high demands of sample processing and experimental costs make single-cell proteomics currently impractical for large-cohort patient studies. Finally, protein-level deconvolution is essential to maximize the value of existing large-scale proteogenomics data through secondary analysis. Leveraging multiomics integration and clinical data association, data mining of TCGA genomics and transcriptomics data has proven invaluable for prioritizing druggable targets and generating mechanistic hypotheses, even though scRNA-seq analysis has become routine for small-cohort studies.[15] Likewise, similar analytical approaches are now being applied to proteogenomics datasets, which will make lasting contributions to cancer research.[14,43] The cell-type-specific protein expression profiles derived from deconvolution would enhance the resolution of protein-level analyses and increase the scientific value of large public proteomics datasets, as demonstrated in our case study uncovering stromal heterogeneity in PDAC.

The challenge of MS proteomics data normalization and harmonization lies in the diverse data quantification formats resulting from different MS experimental techniques and data processing workflows. We reported that the selection and optimization of proteomics data formats are key factors for optimal deconvolution. Similar to previous studies on RNA-seq, the inverse log2 transformation of the TMT MS1 log2 intensity can improve deconvolution performance. However, due to its wide range, the inverse intensity is incompatible with bMIND. On the other hand, the MS2 log2 ratio required only a value shift to overcome negative values for the deconvolution, although its log-inverse can add additional improvement. Moreover, of all the methods tested, our novel "min-score" was the most effective approach for optimizing MS1 log2 intensity data for deconvolution, and it also showed improvement for MS2 log2 ratio data.

For label-free MS proteomics, spectral count achieved optimal performance in its original format. Notably, previous studies suggest that MS1 intensity is superior for quantifying protein abundance and is more robust against random ion exclusion in label-free MS.[44] Our results suggest that besides quantification accuracy, the data format (e.g., data range and

https://doi.org/10.1021/acs.jproteome.5c00980
J. Proteome Res. 2026, 25, 471−483

linearity) is a more critical factor for deconvolution performance. Compared to intensity data, MS2 spectral count is more similar to RNA-seq, as both use count as the basic quantification unit. Thus, spectral count preserves a linearity that is beneficial for deconvolution algorithms.

We provided evidence that the deconvolution suitability for each proteomics data format is consistent among different deconvolution methods, reinforcing the importance role of the data's inherent properties. By examining possible metrics reflecting data properties, we further identified that the coefficient of variation (CV) can serve as a valuable indicator for evaluating deconvolution suitability, which was also validated in RNA-seq data.

To maximize the translational impact of our study, we developed proTransDeconv, an R package that provides a streamlined pipeline for proteomics data deconvolution with functions for transformation, deconvolution, and evaluation functionalities. Applying this tool to the CPTAC PDAC cohort,[10] we demonstrated that cell-type-specific protein markers offer a new layer of insights into tumor subtypes previously defined only by bulk gene expression. Furthermore, our tool can be used to perform a pan-cancer characterization of cell-type-specific protein expression, which would reveal how a tumor's anatomic location impacts its intrinsic biology and microenvironment at the protein level.

Our study has several limitations. First, we did not include all proteomics data formats, such as those used in Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC) or specialized data-independent acquisition (DIA) proteomics. SILAC is not suitable for studying patient tissues, and DIA proteomics has not yet been widely applied in large-cohort cancer studies. Second, our cell-type-specific evaluation was limited to the cohort level rather than the sample level. This limitation stems from the scarcity of deconvolution approaches capable of achieving sample-level resolution, possibility due to their limited accuracy. We anticipate that similar evaluation designs could be applied to sample-level cell-type-specific protein analysis once these algorithms become widely adopted. Third, the deconvolution pipeline used in this study requires additional data modalities (e.g., DNA methylation, WES, or RNA-seq) to infer the tumor composition, which may restrict its application in studies lacking these data types. Nonetheless, it is readily implementable in large-scale consortia such as CPTAC and ICPC, where multiomics data are routinely generated. Moreover, with the increasing prevalence of multiomics profiling in future cohorts, this requirement is expected to become less restrictive.

In summary, our study evaluated the suitability of deconvolution for diverse MS proteomics data formats and provided insights for selecting or optimizing MS data to study cell-type-specific protein abundance.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jproteome.5c00980.

Figure S1: representative immunohistochemistry images from the Human Protein Atlas showing cell-type specificity; Figure S2: consistency of evaluation results across different deconvolution pipelines and cell-type specificities; Figure S3: comparison of variance, mean absolute deviation, and data skewness across different proteomics data formats; and Figure S4: workflow and functionality of the R package 'ProTransDeconv'. (PDF)

Details of the reference methylation datasets used for 1st-phase deconvolution with EDec (XLSX)

Reference marker proteins used for evaluation (XLSX)

Cell-type-specific proteins with significant prognostic values (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Chen Huang − *Department of Genetics and O'Neal Comprehensive Cancer Center, University of Alabama at Birmingham, Birmingham, Alabama 35233, United States;* orcid.org/0000-0003-2777-7365; Email: huangc@uab.edu

Authors

Yingnan Song − *Department of Genetics, University of Alabama at Birmingham, Birmingham, Alabama 35233, United States*

Qi Zhou − *Department of Genetics, University of Alabama at Birmingham, Birmingham, Alabama 35233, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jproteome.5c00980

Author Contributions

Conceptualization, C.H.; Methodology, Y.S. and C.H.; Formal Analysis, Y.S. and C.H.; visualization: Y.S., Q.Z., and C.H.; Tool implementation, Y.S.; Writing − original draft, Y.S. and C.H., final manuscript − Y.S., Q.Z., and C.H.; Supervision and Funding Acquirement: C.H.

Funding

This work was supported by NIH/NIGMS grants 1R35GM154953 (to C.H.).

Notes

This research did not involve human or animal participants. The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Araf Mahmud for helpful and constructive discussions throughout the project.

■ REFERENCES

(1) de Visser, K. E.; Joyce, J. A. The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. *Cancer Cell* **2023**, *41* (3), 374−403.

(2) Bagchi, S.; Yuan, R.; Engleman, E. G. Immune Checkpoint Inhibitors for the Treatment of Cancer: Clinical Impact and Mechanisms of Response and Resistance. *Annu. Rev. Pathol* **2021**, *16*, 223−249.

(3) Binnewies, M.; Roberts, E. W.; Kersten, K.; Chan, V.; Fearon, D. F.; Merad, M.; Coussens, L. M.; Gabrilovich, D. I.; Ostrand-Rosenberg, S.; Hedrick, C. C.; et al. Understanding the tumor immune microenvironment (TIME) for effective therapy. *Nat. Med.* **2018**, *24* (5), 541−550.

(4) Calon, A.; Lonardo, E.; Berenguer-Llergo, A.; Espinet, E.; Hernando-Momblona, X.; Iglesias, M.; Sevillano, M.; Palomo-Ponce, S.; Tauriello, D. V.; Byrom, D.; et al. Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat. Genet.* **2015**, *47* (4), 320−329.

(5) Mani, D. R.; Krug, K.; Zhang, B.; Satpathy, S.; Clauser, K. R.; Ding, L.; Ellis, M.; Gillette, M. A.; Carr, S. A. Cancer proteogenomics: current impact and future prospects. *Nat. Rev. Cancer* **2022**, *22* (5), 298−313.
(6) Beck, L.; Geiger, T. MS-based technologies for untargeted single-cell proteomics. *Curr. Opin Biotechnol* **2022**, *76*, No. 102736.
(7) Datta, S.; Malhotra, L.; Dickerson, R.; Chaffee, S.; Sen, C. K.; Roy, S. Laser capture microdissection: Big data from small samples. *Histol. Histopathol.* **2015**, *30* (11), 1255−1269.
(8) (a) Padden, J; Ahrens, M; Kälsch, J; Bertram, S; Megger, DA; Bracht, T; Eisenacher, M; Kocabayoglu, P; Meyer, HE; Sipos, B; et al. Immunohistochemical Markers Distinguishing Cholangiocellular Carcinoma (CCC) from Pancreatic Ductal Adenocarcinoma (PDAC) Discovered by Proteomic Analysis of Microdissected Cells. *Mol. Cell. Proteomics* **2016**, *15* (3), 1072. (b) Li, Q. K.; Hu, Y.; Chen, L.; Schnaubelt, M.; Zhou, D. C.; Li, Y.; Lu, R.J.-H.; Thiagarajan, M.; Hostetter, G.; Newton, C. J.; et al. Neoplastic cell enrichment of tumor tissues using coring and laser microdissection for proteomic and genomic analyses of pancreatic ductal adenocarcinoma. *Clin. Proteomics* **2022**, *19* (1), 36. (c) Wang, X.; Niu, J.; Qian, S.; Shen, S.; Straubinger, R. M.; Qu, J. Species-Deconvolved Proteomics for In Situ Investigation of Tumor-Stroma Interactions after Treatment of Pancreatic Cancer Patient-Derived Xenografts with Combined Gemcitabine and Paclitaxel. *J. Proteome Res.* **2023**, *22* (7), 2436−2449.
(9) Grunwald, B. T.; Devisme, A.; Andrieux, G.; Vyas, F.; Aliar, K.; McCloskey, C. W.; Macklin, A.; Jang, G. H.; Denroche, R.; Romero, J. M.; et al. Spatially confined sub-tumor microenvironments in pancreatic cancer. *Cell* **2021**, *184* (22), 5577−5592.e5518.
(10) Cao, L.; Huang, C.; Cui Zhou, D.; Hu, Y.; Lih, T. M.; Savage, S. R.; Krug, K.; Clark, D. J.; Schnaubelt, M.; Chen, L.; et al. Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell* **2021**, *184* (19), 5031−5052.e5026.
(11) Nguyen, H.; Nguyen, H.; Tran, D.; Draghici, S.; Nguyen, T. Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges. *Nucleic Acids Res.* **2024**, *52* (9), 4761−4783.
(12) Yoshihara, K.; Shahmoradgoli, M.; Martinez, E.; Vegesna, R.; Kim, H.; Torres-Garcia, W.; Trevino, V.; Shen, H.; Laird, P. W.; Levine, D. A.; et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **2013**, *4*, 2612.
(13) Ferro Dos Santos, M. R.; Giuili, E.; De Koker, A.; Everaert, C.; De Preter, K. Computational deconvolution of DNA methylation data from mixed DNA samples. *Brief Bioinform* **2024**, *25* (3), No. bbae234.
(14) Li, Y.; Dou, Y.; Da Veiga Leprevost, F.; Geffen, Y.; Calinawan, A. P.; Aguet, F.; Akiyama, Y.; Anand, S.; Birger, C.; Cao, S.; et al. Proteogenomic data and resources for pan-cancer analysis. *Cancer Cell* **2023**, *41* (8), 1397−1406.
(15) Tomczak, K.; Czerwinska, P.; Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol. (Pozn)* **2015**, *19* (1A), A68−A77.
(16) Rozanova, S.; Barkovits, K.; Nikolov, M.; Schmidt, C.; Urlaub, H.; Marcus, K. Quantitative Mass Spectrometry-Based Proteomics: An Overview. *Methods Mol. Biol.* **2021**, *2228*, 85−116.
(17) Jin, H.; Wan, Y. W.; Liu, Z. Comprehensive evaluation of RNA-seq quantification methods for linearity. *BMC Bioinf.* **2017**, *18* (Suppl 4), 117.
(18) Avila Cobos, F.; Alquicira-Hernandez, J.; Powell, J. E.; Mestdagh, P.; De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* **2020**, *11* (1), 5650.
(19) Rodriguez, H.; Pennington, S. R. Revolutionizing Precision Oncology through Collaborative Proteogenomics and Data Sharing. *Cell* **2018**, *173* (3), 535−539.
(20) Clark, D. J.; Dhanasekaran, S. M.; Petralia, F.; Pan, J.; Song, X.; Hu, Y.; da Veiga Leprevost, F.; Reva, B.; Lih, T. M.; Chang, H. Y.; et al. Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell* **2019**, *179* (4), 964−983.e931.
(21) Zhang, B.; Wang, J.; Wang, X.; Zhu, J.; Liu, Q.; Shi, Z.; Chambers, M. C.; Zimmerman, L. J.; Shaddox, K. F.; Kim, S.; et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* **2014**, *513* (7518), 382−387.
(22) Uhlen, M.; Oksvold, P.; Fagerberg, L.; Lundberg, E.; Jonasson, K.; Forsberg, M.; Zwahlen, M.; Kampf, C.; Wester, K.; Hober, S.; et al. Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.* **2010**, *28* (12), 1248−1250.
(23) Onuchic, V.; Hartmaier, R. J.; Boone, D. N.; Samuels, M. L.; Patel, R. Y.; White, W. M.; Garovic, V. D.; Oesterreich, S.; Roth, M. E.; Lee, A. V.; et al. Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. *Cell Rep* **2016**, *17* (8), 2075−2086.
(24) Yoshihara, K.; Shahmoradgoli, M.; Martínez, E.; Vegesna, R.; Kim, H.; Torres-Garcia, W.; Treviño, V.; Shen, H.; Laird, P. W.; Levine, D. A.; Carter, S. L.; Getz, G.; Stemke-Hale, K.; Mills, G. B.; Verhaak, R. G. W. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **2013**, *4*, 2612.
(25) Colaprico, A.; Silva, T. C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T. S.; Malta, T. M.; Pagnotta, S. M.; Castiglioni, I.; et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **2016**, *44* (8), No. e71.
(26) Calon, A.; Espinet, E.; Palomo-Ponce, S.; Tauriello, D. V.; Iglesias, M.; Cespedes, M. V.; Sevillano, M.; Nadal, C.; Jung, P.; Zhang, X. H.; et al. Dependency of colorectal cancer on a TGF-beta-driven program in stromal cells for metastasis initiation. *Cancer Cell* **2012**, *22* (5), 571−584.
(27) Ritchie, M. E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C. W.; Shi, W.; Smyth, G. K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **2015**, *43* (7), No. e47.
(28) Jaakkola, M. K.; Elo, L. L. Computational deconvolution to estimate cell type-specific gene expression from bulk data. *NAR Genom. Bioinform.* **2021**, *3* (1), No. lqaa110.
(29) Shen-Orr, S. S.; Tibshirani, R.; Khatri, P.; Bodian, D. L.; Staedtler, F.; Perry, N. M.; Hastie, T.; Sarwal, M. M.; Davis, M. M.; Butte, A. J. Cell type-specific gene expression differences in complex tissues. *Nat. Methods* **2010**, *7* (4), 287−289.
(30) Wang, J.; Roeder, K.; Devlin, B. Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data. *Genome Res.* **2021**, *31* (10), 1807−1818.
(31) Harrell, F. E., Jr.; Lee, K. L.; Mark, D. B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* **1996**, *15* (4), 361−387.
(32) Elizarraras, J. M.; Liao, Y.; Shi, Z.; Zhu, Q.; Pico, A. R.; Zhang, B. WebGestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics. *Nucleic Acids Res.* **2024**, *52* (W1), W415−W421.
(33) Carter, S. L.; Cibulskis, K.; Helman, E.; McKenna, A.; Shen, H.; Zack, T.; Laird, P. W.; Onofrio, R. C.; Winckler, W.; Weir, B. A.; et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **2012**, *30* (5), 413−421.
(34) Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* **1960**, *20* (1), 37−46.
(35) McHugh, M. L. Interrater reliability: the kappa statistic. *Biochem Med. (Zagreb)* **2012**, *22* (3), 276−282.
(36) (a) Collisson, E. A.; Sadanandam, A.; Olson, P.; Gibb, W. J.; Truitt, M.; Gu, S.; Cooc, J.; Weinkle, J.; Kim, G. E.; Jakkula, L.; et al. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nat. Med.* **2011**, *17* (4), 500−503. (b) Bailey, P.; Chang, D. K.; Nones, K.; Johns, A. L.; Patch, A. M.; Gingras, M. C.; Miller, D. K.; Christ, A. N.; Bruxner, T. J.; Quinn, M. C.; et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **2016**, *531* (7592), 47−52.
(37) (a) Puleo, F.; Nicolle, R.; Blum, Y.; Cros, J.; Marisa, L.; Demetter, P.; Quertinmont, E.; Svrcek, M.; Elarouci, N.; Iovanna, J.; et al. Stratification of Pancreatic Ductal Adenocarcinomas Based on Tumor and Microenvironment Features. *Gastroenterology* **2018**, *155* (6), 1999−2013.e1993. (b) Maurer, C.; Holmstrom, S. R.; He, J.; Laise, P.; Su, T.; Ahmed, A.; Hibshoosh, H.; Chabot, J. A.; Oberstein,

https://doi.org/10.1021/acs.jproteome.5c00980
J. Proteome Res. 2026, 25, 471−483

P. E.; Sepulveda, A. R.; et al. Experimental microdissection enables functional harmonisation of pancreatic cancer subtypes. *Gut* **2019**, *68* (6), 1034−1043.
(38) Hosein, A. N.; Brekken, R. A.; Maitra, A. Pancreatic cancer stroma: an update on therapeutic targeting strategies. *Nat. Rev. Gastroenterol Hepatol* **2020**, *17* (8), 487−505.
(39) Moffitt, R. A.; Marayati, R.; Flate, E. L.; Volmar, K. E.; Loeza, S. G.; Hoadley, K. A.; Rashid, N. U.; Williams, L. A.; Eaton, S. C.; Chung, A. H.; et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nat. Genet.* **2015**, *47* (10), 1168−1178.
(40) Golosow, N.; Grobstein, C. Epitheliomesenchymal interaction in pancreatic morphogenesis. *Developmental biology* **1962**, *4* (2), 242−255.
(41) Li, X.; Zhang, Q.; Yu, S. M.; Li, Y. The Chemistry and Biology of Collagen Hybridization. *J. Am. Chem. Soc.* **2023**, *145* (20), 10901−10916.
(42) Shih, Y. P.; Sun, P.; Wang, A.; Lo, S. H. Tensin1 positively regulates RhoA activity through its interaction with DLC1. *Biochim. Biophys. Acta* **2015**, *1853* (12), 3258−3265.
(43) (a) Li, Y.; Porta-Pardo, E.; Tokheim, C.; Bailey, M. H.; Yaron, T. M.; Stathias, V.; Geffen, Y.; Imbach, K. J.; Cao, S.; Anand, S.; et al. Pan-cancer proteogenomics connects oncogenic drivers to functional states. *Cell* **2023**, *186* (18), 3921−3944.e3925. (b) Savage, S. R.; Yi, X.; Lei, J. T.; Wen, B.; Zhao, H.; Liao, Y.; Jaehnig, E. J.; Somes, L. K.; Shafer, P. W.; Lee, T. D.; et al. Pan-cancer proteogenomics expands the landscape of therapeutic targets. *Cell* **2024**, *187* (16), 4389−4407.e4315.
(44) Choi, H.; Glatter, T.; Gstaiger, M.; Nesvizhskii, A. I. SAINT-MS1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *J. Proteome Res.* **2012**, *11* (4), 2619−2624.

https://doi.org/10.1021/acs.jproteome.5c00980
J. Proteome Res. 2026, 25, 471−483