

Survival driven deconvolution (deSurv) reveals prognostic and interpretable cancer subtypes

Amber M. Young^{a,1,2}, Alisa Yurovsky^b, Didong Li^a, and Naim U. Rashid^{a,c}

^aUniversity of North Carolina at Chapel Hill, Biostatistics, Street, City, State, Zip; ^bStony Brook University, Street, City, State, Zip

This manuscript was compiled on November 5, 2025

Molecular subtyping in cancer is an ongoing problem that relies on the identification of robust and replicable gene signatures. While transcriptomic profiling has revealed recurrent gene expression patterns in various types of cancer, the prognostic value of these signatures is typically evaluated in retrospect. This is due to the reliance on unsupervised learning methods for identifying cell-type-specific signals and clustering patients into molecular subtypes. Here we present a Survival-driven Deconvolution tool (deSurv) that integrates bulk RNA-sequencing data with patient survival information to identify cell-type-enriched gene signatures associated with prognosis. Applying deSurv to various cohorts in pancreatic cancer, we uncover prognostic and biologically interpretable subtypes that reflect the complex interactions between stroma, tumor, and immune cells in the tumor microenvironment. Our approach highlights the value of using patient outcomes during gene signature discovery.

one | two | optional | optional | optional

Molecular subtyping has become a cornerstone of precision oncology, enabling the classification of patients into biologically distinct groups that inform prognosis and guide therapeutic decisions (1–5). Defining these subtypes depends on identifying robust genomic or transcriptomic signatures that capture the molecular programs driving disease heterogeneity (6–9).

However, tumors are composed of diverse cell types—including malignant, stromal, immune, and endothelial populations—whose gene-expression signals are deeply interwoven. This complexity makes it difficult to isolate the signatures that truly characterize each subtype or to distinguish tumor-intrinsic programs from those shaped by the surrounding microenvironment. Furthermore, some molecular programs may delineate distinct aspects of tumor biology but not demonstrate clinical relevance with respect to patient outcomes (10–12). This disconnect highlights a key challenge in translational genomics: distinguishing the molecular features that merely describe tumor heterogeneity from those that are predictive of disease course or therapeutic response.

Single-cell transcriptomic technologies have transformed our understanding of tumor ecosystems by resolving the diversity of cell types and states within the tumor microenvironment (13). Yet their high cost and limited cohort sizes often preclude systematic evaluation of clinical outcomes, so the molecular programs they uncover are typically assessed downstream in bulk datasets for prognostic or therapeutic relevance. Bulk transcriptomic studies—while lacking cellular resolution—have been conducted across many independent patient cohorts, which collectively encompass hundreds to thousands of clinically annotated samples (14, 15). These datasets provide the statistical power to evaluate molecular programs in relation to patient outcomes, highlighting a trade-off between biological resolution and clinical scalability.

To address the limited resolution of bulk data, numerous

computational deconvolution approaches have been developed to infer the relative abundance of distinct cell types and their gene-expression programs, summarized in this review (16). These methods often rely on reference signatures derived from sorted or single-cell data to estimate cellular composition and attribute gene-expression patterns to known cell types. While such approaches have been valuable for characterizing the tumor microenvironment, they depend on predefined references and therefore cannot discover new or context-specific molecular programs.

Unsupervised matrix factorization methods, particularly Nonnegative Matrix Factorization (NMF) (17), have been instrumental in revealing latent molecular programs that define tumor heterogeneity, especially in pancreatic cancer (18–21). NMF excels at identifying biologically meaningful patterns in gene-expression data (22) but is limited in prognostic ability due to its unsupervised nature (23). Additionally, the non-uniqueness of NMF solutions can make replicability across studies challenging. Consequently, the uncovered programs frequently capture broad biological variation, such as cell type composition, rather than the specific molecular features that influence prognosis or treatment response.

To bridge this gap, we introduce DeSurv, a semi-supervised deconvolution framework that integrates NMF with the Cox proportional hazards model (24). By integrating survival modeling directly into the matrix factorization framework, DeSurv learns *de novo* molecular programs that are both biologically interpretable and prognostic of patient outcomes. Importantly,

Significance Statement

Understanding which molecular programs influence patient survival remains a central challenge in cancer genomics. We developed DeSurv, a survival-driven deconvolution framework that integrates nonnegative matrix factorization with Cox regression to identify latent gene-expression programs directly linked to clinical outcomes. Applied to pancreatic ductal adenocarcinoma, DeSurv uncovered interpretable tumor and stromal factors that generalize across independent cohorts and align with established subtypes. By coupling survival modeling with matrix factorization, DeSurv bridges the gap between biologically descriptive and prognostically relevant signatures, revealing that patient outcomes are shaped by the combined activity of basal and immune-stromal programs. This approach provides a broadly applicable strategy for discovering outcome-associated transcriptional programs in complex cancers.

Please provide details of author contributions here.

Please declare any conflict of interest here.

² To whom correspondence should be addressed. E-mail: ayoung31@live.unc.edu

DeSurv incorporates an automated cross-validation framework to determine key parameters, including the optimal number of latent factors—addressing a major limitation of fully unsupervised approaches that rely on heuristic or subjective choices. Applied to pancreatic adenocarcinoma (PDAC) data, DeSurv uncovered molecular programs representing blended tumor, stromal, and immune features, suggesting that integrating survival information can reveal novel prognostic interactions between tumor and microenvironmental processes.

Results

Model Overview. We have developed an integrated framework, DeSurv, that couples Nonnegative Matrix Factorization (NMF) with Cox proportional hazards regression to identify latent gene-expression programs associated with patient survival (Fig. 1). The model takes as input a bulk expression matrix of p genes by n patients (X_{Train}) together with corresponding survival times (y_{Train}) and censoring indicators (δ_{Train}) (Fig. 1A).

DeSurv optimizes a joint objective combining the NMF reconstruction loss and the Cox model's log-partial likelihood, weighted by a supervision parameter (α) that determines the relative contribution of each term (Fig. 1B):

$$(1 - \alpha) \mathcal{L}_{NMF}(X_{Train} \approx WH) \\ -\alpha \mathcal{L}_{Cox}(X_{Train}^T W \beta, y_{Train}, \delta_{Train}) \quad [1]$$

When $\alpha = 0$, the method reduces to standard unsupervised NMF; when $\alpha > 0$, survival information directly guides the learned factors toward prognostic structure.

Within this framework, the product ($X_{Train}^T W$) represents patient-level factor scores - the inferred burden of each latent program across subjects. These factor scores serve as covariates in the Cox model, and their regression coefficients (β) indicate whether higher activity of a given program corresponds to improved or reduced survival.

Model training yields gene weights (\hat{W}), factor loadings (\hat{H}), and Cox coefficients ($\hat{\beta}$) (Fig. 1C), where the inner dimension (k) specifies the number of latent factors. Genes with high gene weights in one factor and low gene weights in all others define the factor-specific signature genes (Fig. 1D). By integrating survival supervision into the factorization, DeSurv not only reconstructs the underlying expression structure, preserving biological interpretability, but also guides latent factors to be prognostically informative. Subsequent analyses can therefore focus on the survival-associated gene programs (Fig. 1E).

The DeSurv pipeline provides an automated method for choosing the NMF rank (k) and supervision parameter (α). We implemented a five-fold cross-validation procedure to identify the optimal hyperparameters for the DeSurv model—namely, the number of latent factors (k) and the supervision strength (α). In each fold, the gene-expression matrix and corresponding survival outcomes were partitioned into training and test sets at the patient level (Fig. 2A). The DeSurv model was applied to the training set to estimate the gene weights (\hat{W}) and regression coefficients ($\hat{\beta}$) (Fig. 2B). The expression matrix for the test set (X_{Test}) was projected onto the learned basis (\hat{W}) to obtain patient-level factor scores. These scores were multiplied by regression coefficients ($\hat{\beta}$) to generate the linear predictor for each patient (Fig. 2C), which was then compared against

the true survival outcomes to assess predictive performance (Fig. 2D).

We evaluated the cross-validated concordance index (C-index) across a grid of model ranks ($k = \{2, \dots, 12\}$) and supervision parameters ($\alpha = \{0, 0.05, 0.10, \dots, 1\}$) (Fig. 2E). To reduce overfitting, we selected the smallest combination of α and k within one standard error of the maximum cross validated c-index. The cross-validated C-index varied modestly across the parameter space, but models with $\alpha > 0$ consistently outperformed the unsupervised baseline ($\alpha = 0$). For each value of k , there existed at least one $\alpha > 0$ yielding higher predictive accuracy than the unsupervised equivalent, indicating that incorporating survival information improves predictive performance. Interestingly, the supervised models achieved their best performance at smaller values of k compared with unsupervised NMF, suggesting that survival-guided factorization produces a more parsimonious yet more prognostic representation.

For comparison, we also evaluated standard NMF using the NMF R package, which lacks an objective means for cross-validation. Conventional selection of the factor rank k relies on heuristic metrics such as cophenetic correlation, dispersion, residual error, explained variance, silhouette score, and matrix sparseness (Fig. 2F). In our analysis, cophenetic correlation and dispersion peaked at $k = 3$ and declined thereafter, indicating reduced clustering stability at higher ranks. In contrast, explained variance and reconstruction residuals plateaued around $k = 7$, suggesting that at least this many factors are needed to adequately reconstruct the expression matrix. The silhouette score was highest at $k = 2-3$, implying better cluster separation at smaller ranks, whereas the sparsity of W and H was most balanced near $k = 7$, favoring interpretability at moderate rank. These criteria therefore yield conflicting recommendations: stability metrics favor smaller k , while reconstruction and interpretability metrics favor moderate k . In the cross validation at $\alpha = 0$, survival prediction under unsupervised NMF did not improve appreciably until $k \geq 5$ and peaked at $k = 6$.

In contrast, DeSurv's joint optimization identifies a simpler model ($k = 3$ and $\alpha = 0.7$) that achieves superior prognostic performance compared to standard NMF with a more parsimonious model, underscoring the advantage of integrating survival supervision into the matrix factorization process.

DeSurv provides biologically interpretable gene signatures. For the selected DeSurv model ($k = 3, \alpha = 0.7$), we identified the top 50 factor-specific genes from each latent factor to characterize their underlying biological functions. Figures 3A–C summarize the results of an over-representation analysis (ORA) performed on these genes. The enrichment profiles reveal three distinct biological programs. Factor 1 is strongly enriched for immune-related pathways, including cytokine signaling, T-cell activation, and interferon response. Factor 2 shows enrichment for genes associated with normal pancreatic or exocrine function, suggesting that this component captures residual normal-tissue signal. Factor 3 is enriched for pathways linked to epithelial–mesenchymal transition, cell cycle progression, and KRAS signaling—features typically associated with the Basal-like subtype of pancreatic ductal adenocarcinoma (PDAC).

Survival analysis of the factor signature scores on the training data demonstrates clear prognostic stratification. High

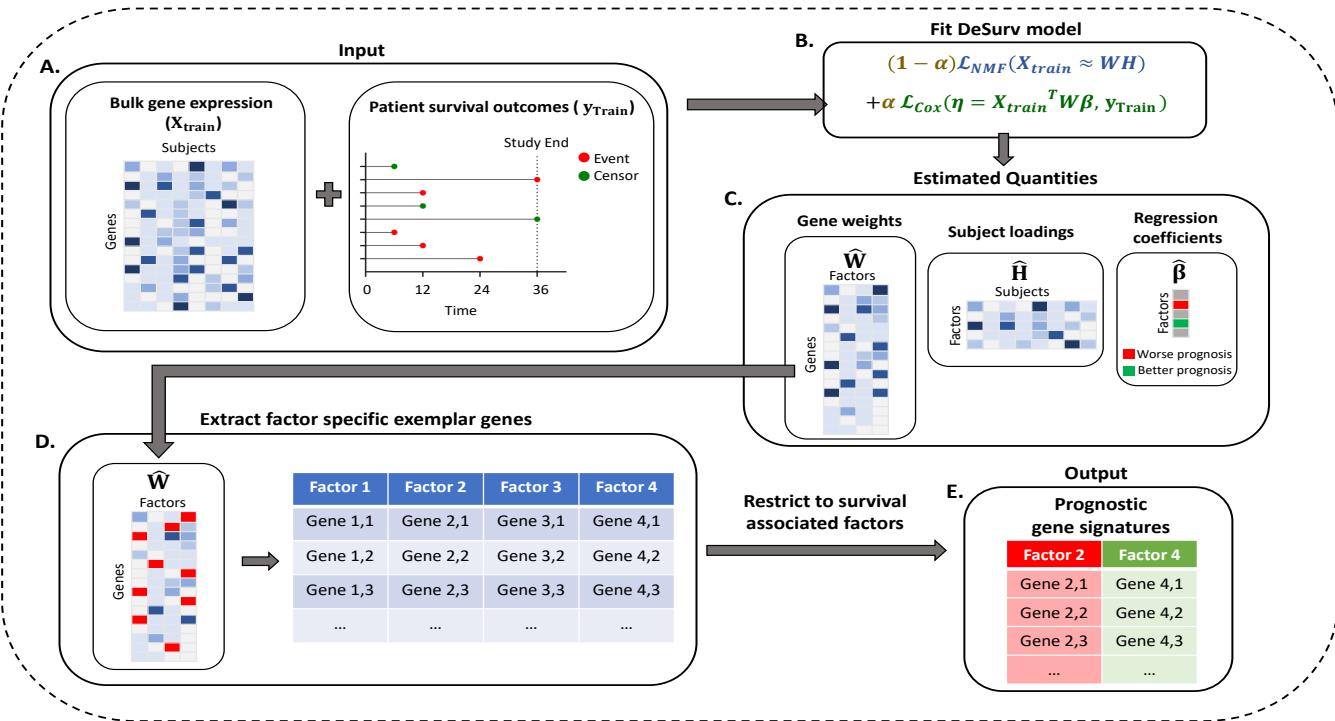


Fig. 1. DeSurv overview. Overview of the DeSurv training pipeline. **A.** The input is a preprocessed bulk gene expression matrix and patient survival outcomes. **B.** The DeSurv model optimizes a joint NMF + Cox loss function. **C.** The DeSurv model estimates three quantities: Gene weights matrix (\hat{W}), Subject loadings matrix (\hat{H}), and regression coefficients from the Cox model ($\hat{\beta}$). **D.** We extract the exemplar genes from each signature in the gene weights matrix. **E.** The exemplar genes from survival associated factors provide prognostic gene signatures.

expression of Factor 1 is associated with longer overall survival (HR = 0.37, $p < 2e-16$), whereas Factor 3 is linked to significantly shorter survival (HR = 1.43, $p = 2e-4$). These trends are consistent with known biology: immune-infiltrated tumors generally exhibit improved prognosis, while Basal-like tumors are aggressive and clinically refractory. Factor 2, corresponding to the normal/exocrine signal, shows no significant association with survival (HR = 0.99, $p = 0.91$), indicating that while this component contributes to reconstructing the expression matrix, it does not carry prognostic information. This behavior illustrates a key property of DeSurv—its joint optimization ensures that the latent factors balance both survival relevance and expression reconstruction, preserving biologically necessary structure even when not directly linked to outcome.

Figure 3D further examines the correlation between DeSurv factors and previously published PDAC gene signatures. Factor 1 aligns with signatures from classical tumor, iCAF and restCAF stroma, and immune infiltration, all of which are associated with favorable prognosis. In contrast, Factor 3 correlates with Basal-like tumor, proCAF, and activated stromal programs, which are generally linked to poor outcomes. These overlaps suggest that each factor represents a composite axis of tumor–stroma–immune interaction, capturing biologically coherent programs that align with known PDAC ecosystems but arise de novo from the DeSurv model.

Clustering patients from external cohorts on DeSurv gene signatures produces prognostic subtypes.

To evaluate the

generalizability of DeSurv-derived prognostic signatures, we applied the exemplar genes from the survival-associated factors in the trained model (Factors 1 and 3) to multiple independent PDAC cohorts—including Dijk, PACA-AU (RNA-seq and microarray), Moffitt, and Puleo datasets—using their bulk gene-expression profiles. Unsupervised consensus clustering was performed on each cohort based on the expression of these prognostic genes. Three clusters were selected as the optimal solution, supported by inspection of the cumulative distribution function (CDF) plots, relative change in area under the CDF curve, consensus heatmaps, and cluster consensus scores (Fig. S4). Importantly, the number of clusters was determined without incorporating any survival information, ensuring an unbiased evaluation of downstream prognostic performance.

Expression heatmaps of the DeSurv prognostic signatures revealed three distinct transcriptional patterns across the external cohorts (Fig. 4A). Cluster 1 displayed high expression of the basal-like (Factor 3) gene set; Cluster 2 showed elevated expression of the immune/iCAF (Factor 1) signature and reduced basal activity; and Cluster 3 exhibited intermediate expression of both programs. These patterns suggest that survival prediction is primarily driven by the presence or absence of basal and immune/iCAF programs, rather than by a reciprocal contrast between basal versus classical tumor states or iCAF versus myCAF stromal signatures.

Additionally, DeSurv-derived clusters showed strong correspondence with established molecular classifiers. Column annotations in Fig. 4A and the contingency table in Fig. 4B demonstrate substantial overlap with PurIST and DeCAF

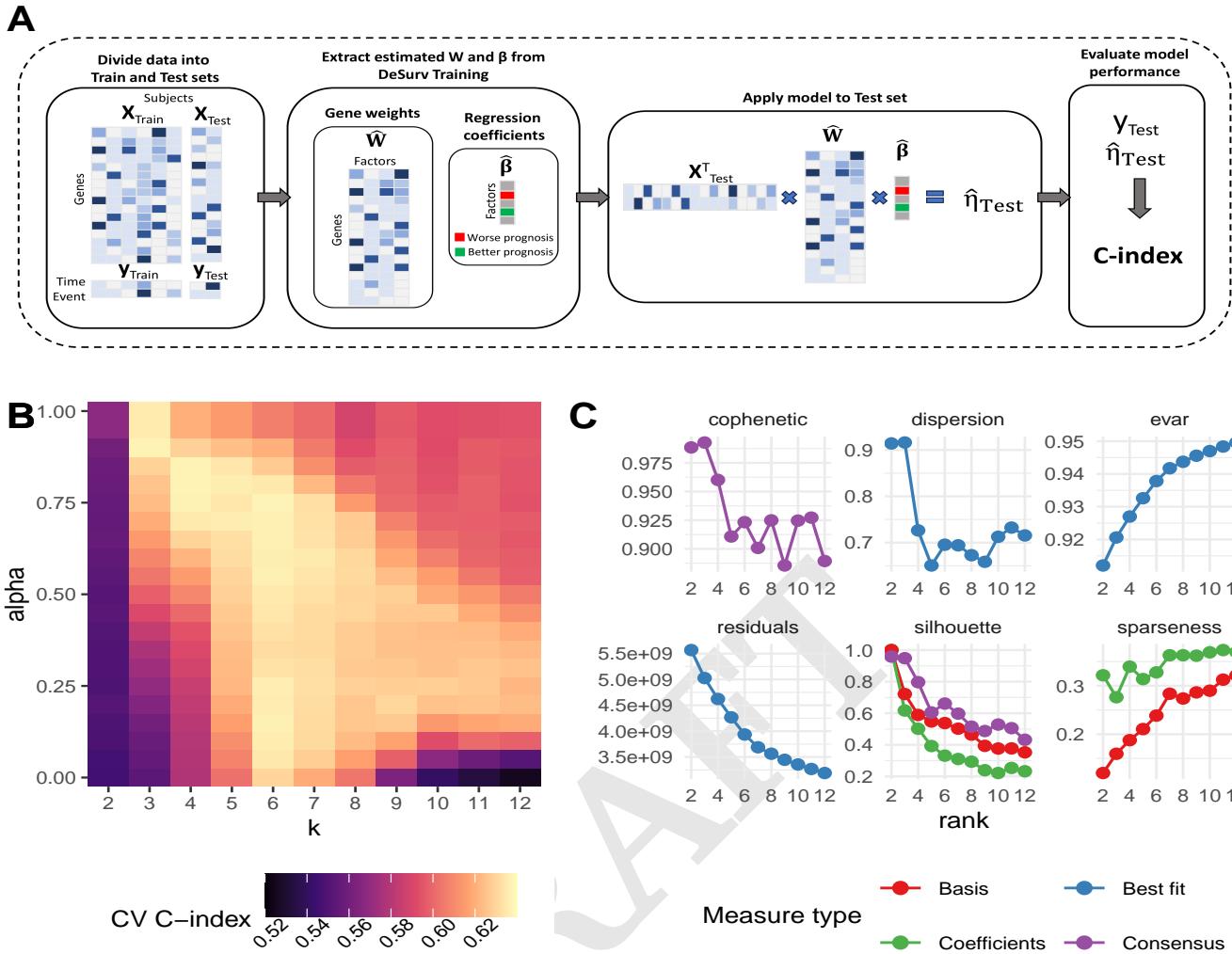


Fig. 2. A. Overview of DeSurv cross-validation pipeline. B. Heatmap of cross-validated C-index with columns k and rows α . C. Plots of various metrics used to choose k in the standard NMF setting.

subtype labels: for example, X% of patients in Cluster 1 were classified as PurIST Basal-like and DeCAF proCAF, X% of Cluster 2 as Classical + restCAF, and X% of Cluster 3 as Classical + proCAF. The rarity of the Basal + restCAF subtype likely contributes to the discovery of 3 rather than 4 clusters. A chi-squared test confirmed a significant association between the DeSurv-defined clusters and existing classifier labels ($\chi^2 = 210.16$ $p < 2.2e-16$).

We next assessed the prognostic relevance of these clusters using overall survival data from the external cohorts (Fig. XC–D). Because DeSurv was trained exclusively on the TCGA-PAAD and CPTAC datasets—without access to the expression or outcome data of the external cohorts—this analysis represents an independent validation of the model’s predictive generalizability. Kaplan–Meier analysis demonstrated that the DeSurv-derived clusters were strongly associated with survival ($p = 4.8e-6$). As expected, the Basal-like cluster (Cluster 1) was associated with the poorest survival, the Classical + restCAF cluster (Cluster 2) exhibited the most favorable outcomes, and the Classical + proCAF cluster (Cluster 3) showed intermediate survival. This pattern mirrors previous findings

describing the interplay between tumor-intrinsic and stromal subtypes in PDAC (e.g., (25)), supporting the robustness and biological validity of the DeSurv-derived signatures.

Analysis of scRNA-seq. Next, we verify our findings at the cellular level using the Elyada scRNA-seq data (cite data). We found that our DeSurv factor 1 signature was expressed primarily in iCAF and B cells, factor 2 was expressed in Acinar and Classical PDAC 2, and factor 3 was expressed primarily in Basal-like PDAC cells (Figures 4B–E). This is consistent with our ORA analysis which found that factor 1 was mostly immune and factor 3 was basal-like, and with the overlap we saw with published gene lists that identified factor 1 as iCAF/classical/immune and factor 3 as basal-like. Interestingly, we do not see a large overlap of factor 1 with the classical PDAC cell types despite some overlap with classical gene lists. Need more here...

Discussion

We present DeSurv, a survival-driven deconvolution framework that integrates nonnegative matrix factorization with Cox pro-

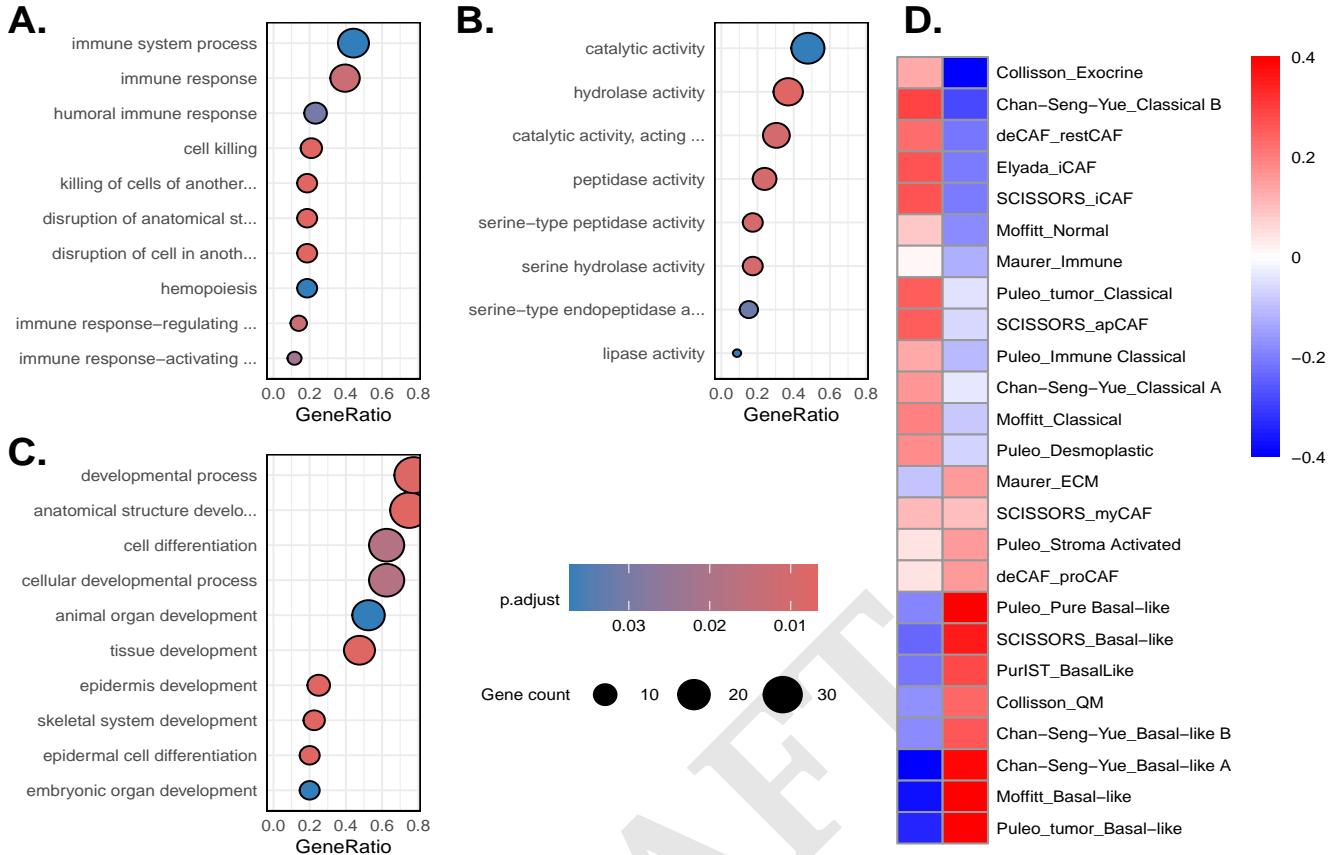


Fig. 3. A-C. Dotplots of ORA analysis of top 50 genes from each DeSurv derived factor. D. Correlation between the DeSurv gene weights matrix and previously published PDAC gene signatures.

portional hazards modeling to uncover latent gene-expression programs associated with patient outcomes. By coupling unsupervised matrix decomposition with direct survival supervision, DeSurv bridges the gap between descriptive molecular subtyping and prognostic modeling. Unlike conventional NMF, which identifies factors that explain transcriptional variance without regard to clinical relevance, DeSurv simultaneously optimizes reconstruction fidelity and survival discrimination, yielding interpretable gene programs that are both biologically coherent and clinically informative.

In pancreatic ductal adenocarcinoma (PDAC), DeSurv identified three major transcriptional programs corresponding to immune/iCAF, normal/exocrine, and basal-like factors. The immune/iCAF factor was associated with prolonged survival and enriched for inflammatory and immune-response pathways, consistent with prior evidence linking tumor-infiltrating immune and inflammatory fibroblast populations to improved outcomes. In contrast, the basal-like factor was strongly associated with poor survival and enriched for epithelial-mesenchymal transition and cell-cycle programs characteristic of aggressive tumor phenotypes. The normal/exocrine factor, while not prognostic, captured background pancreatic tissue signal and contributed to accurate reconstruction, demonstrating that DeSurv can disentangle technical and biological sources of variation while preserving those necessary for modeling.

Importantly, DeSurv outperformed unsupervised NMF in cross-validated concordance index, achieving higher predictive

accuracy with fewer factors. This suggests that explicitly incorporating survival information regularizes the factorization, producing a more parsimonious representation that focuses on biologically and clinically relevant variation. Moreover, when applied to independent PDAC cohorts—including Dijk, PACA-AU, Moffitt, and Puleo—DeSurv-derived gene signatures stratified patients into three reproducible clusters with clear prognostic differences. These clusters aligned strongly with existing PurIST and DeCAF classifiers, recapitulating the interplay between tumor-intrinsic (Basal versus Classical) and stromal (proCAF versus restCAF/iCAF) subtypes. Notably, survival prediction appeared to depend on the presence or absence of basal and immune/iCAF programs rather than binary subtype identities, indicating that the co-activation or suppression of these programs better captures the biological continuum of PDAC progression.

Beyond PDAC, DeSurv provides a generalizable framework for discovering prognostic transcriptional programs across complex cancers where tumor and stromal signals are interwoven. By integrating survival modeling into matrix factorization, it unifies molecular deconvolution and outcome prediction within a single, interpretable model. This approach complements both single-cell-based tumor-stroma characterization and bulk expression subtyping, offering a scalable route to translate transcriptomic heterogeneity into clinically actionable prognostic insight. Future work will extend DeSurv to multi-omic data and time-varying outcomes, further enhancing its ability to resolve dynamic tumor-microenvironment inter-

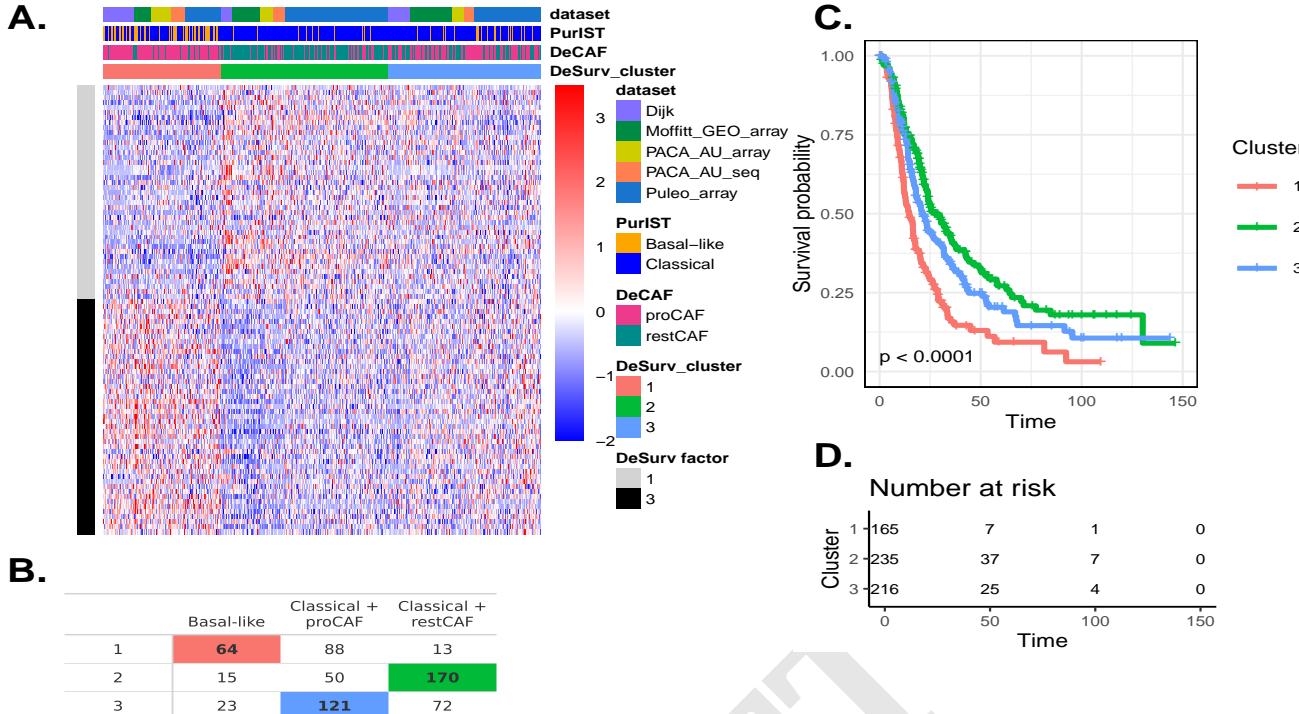


Fig. 4. External cohort validation. A. Heatmap of expression values for genes in the DeSurf factor signatures in external cohorts. B. Table of DeSurf derived clusters by PurlST + DeCAF classifier subtypes in external cohorts. C. Kaplan Meier curves describing survival outcomes for patients from external cohorts in each of the DeSurf derived clusters.

actions that shape disease progression and therapy response.

Materials and methods

NMF. Let $X \in R_{\geq 0}^{p \times n}$ denote a nonnegative gene expression matrix of p features (genes) across n subjects. The goal of NMF is to approximate X as the product of two low-rank, nonnegative matrices

$$X \approx WH, \quad [2]$$

where $W \in R_{\geq 0}^{p \times k}$ contains the gene weights, and $H \in R_{\geq 0}^{k \times n}$ contains the factor scores for each subject. The number of latent factors k determines the dimensionality of the shared low-rank representation. The NMF loss is defined as the residual sum of squares.

$$\mathcal{L}(W, H)_{NMF} = \|X - WH\|_F^2. \quad [3]$$

Cox partial log-likelihood. To incorporate survival outcomes, let T_i denote the event time and C_i the censoring time for subject i . The observed time is $y_i = \min(T_i, C_i)$, and the event indicator is δ_i . Given that W is shared across datasets, we define a lower dimensional transformation of the data:

$$Z = X^T W \in R^{n \times k}, \quad [4]$$

where each row Z_i^T represents the factor signature scores for subject i . These scores serve as covariates in a Cox proportional hazards model:

$$h_i(t) = h_0(t) \exp(Z_i^T \beta), \quad [5]$$

where $h_0(t)$ is the baseline hazard and $\beta \in R^k$ are the factor specific coefficients. The Cox log partial likelihood is then

$$\ell(W, \beta) = \sum_{i=1}^n \delta_i \left[Z_i^T \beta - \log \left(\sum_{j: y_j \geq y_i} \exp(Z_j^T \beta) \right) \right]. \quad [6]$$

DeSurf. Building on the definitions above, DeSurf combines the unsupervised NMF reconstruction loss and the supervised Cox partial likelihood into a single joint objective. The combined loss function is

$$\mathcal{L}(W, H, \beta) = (1 - \alpha) \mathcal{L}_{NMF}(X \approx WH) - \alpha \mathcal{L}_{cox}(W, \beta), \quad [7]$$

where $\mathcal{L}_{cox}(W, \beta)$ is the elastic net penalized log partial likelihood:

$$\mathcal{L}_{cox}(W, \beta) = \ell(W, \beta) + \lambda(\xi \|\beta\|_1 + \frac{(1 - \xi)}{2} \|\beta\|_2^2), \quad [8]$$

where λ represents the penalty weight and ξ is the balance parameter between the L1 and L2 penalty terms.

The hyperparameter $\alpha \in [0, 1]$ controls the relative contribution of each component:

- $\alpha = 0$ recovers standard NMF, focusing purely on reconstruction;
- $\alpha = 1$ corresponds to a fully supervised Cox model in the low-dimensional space $Z = X^T W$

Intermediate values of α encourage discovery of latent molecular programs that are both biologically coherent and prognostically informative.

Optimization framework. DeSurv is optimized using an alternating minimization scheme (Algorithm 1) that iteratively updates W , H , and β until convergence. The sub-problems for H and β are convex in the corresponding parameter conditional on the others.

Algorithm 1 DeSurv algorithm

Input: $X \in \mathbb{R}_{\geq 0}^{p \times n}$, $y \in \mathbb{R}_{\geq 0}^n$, $\delta \in \mathbb{R}_{0,1}^n$, $W^{(0)}$, $H^{(0)}$, $\beta^{(0)}$, tol , $maxit$

- 1: $eps = \infty$
- 2: $t = 0$
- 3: $loss = 0$
- 4: **while** $eps < tol$ **and** $t < maxit$ **do**
- 5: $W^{(t)} = \operatorname{argmin}_{W \geq 0} \mathcal{L}(W, H^{(t-1)}, \beta^{(t-1)})$
- 6: $H^{(t)} = \operatorname{argmin}_{H \geq 0} \mathcal{L}(W^{(t)}, H, \beta^{(t-1)})$
- 7: $\beta^{(t)} = \operatorname{argmin}_{\beta} \mathcal{L}(W^{(t)}, H^{(t)}, \beta)$
- 8: $lossNew = \mathcal{L}(W^{(t)}, H^{(t)}, \beta^{(t)})$
- 9: $eps = |lossNew - loss|/loss$
- 10: $loss = lossNew$
- 11: $t = t + 1$
- return W, H, β

Update for H . The nonnegative factor matrix H is updated using standard multiplicative updates that guarantee nonnegativity and monotonic decrease in reconstruction error as derived in (26):

$$H_{ij} = H_{ij} \frac{(W^T X)_{ij}}{(W^T W H)_{ij}}. \quad [9]$$

Update for β . Given W , the coefficients β are updated by coordinate descent using elastic-net regularization:

$$\hat{\beta}_r = \frac{S(\frac{1}{n} \sum_{i=1}^n w(\tilde{\eta})_i v_{i,r} \left[z(\tilde{\eta})_i - \sum_{j \neq r} v_{ij} \beta_j \right], \lambda \xi)}{\frac{1}{n} \sum_{i=1}^n w(\tilde{\eta})_i v_{i,r}^2 + \lambda(1 - \xi)}, \quad [10]$$

where $S(., \lambda \xi)$ is the soft-thresholding operator. The term $w(\tilde{\eta})_i$ is the i th diagonal element of the hessian of the partial log-likelihood and $z(\tilde{\eta})_i$ is an approximation of the Newton-Raphson update for $\tilde{\eta}$. Details can be found in (27). The parameters (λ, ξ) control the strength and type of regularization.

Coupled W update. The shared basis W is updated through a hybrid multiplicative rule that incorporates both NMF reconstruction gradients and Cox partial likelihood gradients.

$$W^{(t+1)} = W^{(t)} \odot \max \left(\frac{\frac{(1-\alpha)}{np} X H^T + \frac{2\alpha}{N_{event}} \nabla_W \ell(W^{(t)}, \beta)}{\frac{(1-\alpha)}{np} W^{(t)} H H^T}, 0 \right) \quad [11]$$

The quantity $\nabla_W \ell(W^{(t)}, \beta)$ denotes the gradient of the Cox partial likelihood with respect to W evaluated at $W^{(t)}$. This update allows the survival signal to propagate into the latent factors while preserving nonnegativity. Backtracking and gradient balancing are used in this update to ensure decrease in the overall loss and avoid one component dominating the update.

Publicly Available Datasets Preprocessing. We trained and validated DeSurv using seven publicly available pancreatic ductal adenocarcinoma (PDAC) transcriptomic datasets spanning both RNA-seq and microarray platforms (Table @ref(tab:datasets)). The TCGA and CPTAC datasets were used for training the DeSurv model, and the remaining 5 were used as external validation. Samples from all datasets were filtered to include only non-metastatic primary PDAC samples. All datasets were $\log_2 + 1$ transformed and filtered to the top highly expressed and variable genes. The training data were rank transformed by subject to mitigate cross dataset batch effects and combined, restricting to genes that were kept in each dataset.

Model Training. The TCGA and CPTAC datasets were used for model training. Each dataset was filtered to the top 2000 highly expressed and variable genes. These gene lists were then intersected, resulting in XXX genes incorporated in model training.

Cross Validation. To identify optimal hyperparameters, we employed stratified five-fold cross-validation based on event status. An exhaustive search was performed across a grid of hyperparameters $k = 2, \dots, 12$, $\alpha \in \{0.1, 0.2, \dots, 0.9\}$, $\lambda \in 10^{\{-3, \dots, 3\}}$, and $\xi \in \{0, 0.1, 0.2, \dots, 1\}$. Within each fold, models were fit on 80% of the data and evaluated on the held-out 20%. Because NMF solutions are non-unique and sensitive to initialization, each fold was repeated with 20 random seeds for W , H , and β , resulting in a total of 100 trained models per parameter configuration of k , $lambda$, and eta . Warm-start initializations were used across $alpha$ to accelerate convergence.

Final hyperparameters were selected as the combination that produced average C-index (across initializations and folds) within one standard error of the maximum (1 s.e. rule).

Standard NMF. As an unsupervised baseline, we trained conventional NMF models that minimized only the reconstruction error, corresponding to setting $alpha = 0$ in the DeSurv formulation. For each rank $k \in 2, \dots, 12$, 100 random initializations were performed to address non-uniqueness. The initialization with the smallest reconstruction error for each k was selected. To select the rank, k , in the unsupervised setting, we used standard metrics including cophenetic coefficient, dispersion, explained variance, residuals, silhouette score, and sparseness. The resulting gene-factor matrix W and validation data were subsequently used as input to a Cox proportional hazards model to evaluate the prognostic value of the unsupervised factors.

Factor-specific gene signatures. For each factor f and gene g , we calculated the difference in weight of gene g in factor f and the maximum weight of gene g across all other factors. We define this difference as s_{gf} .

$$s_{gf} = W_{gf} - \max_{r \neq f} (W_{gr}) \quad [12]$$

We then rank s_{gf} across all genes $g = 1, \dots, p$ from largest to smallest. Genes that appear at the top of this list for each factor are considered to be factor-specific.

Clustering. Clustering was performed using the Consensus-ClusterPlus package in R (28). Each validation dataset was

Table 1. Publicly available pancreatic ductal adenocarcinoma (PDAC) datasets used for model training and validation. Expression data were rank-transformed across genes within samples to mitigate platform- and scale-related effects.

Dataset	Platform	Samples (n)	Data type	Reference
TCGA-PAAD	RNA-seq (Illumina HiSeq)	144	Discovery / Training	(?)
CPTAC-PDAC	RNA-seq (Proteogenomic)	129	Validation	(?)
Dijk <i>et al.</i>	Microarray (Affymetrix)	90	Validation	@Dijk2020unsupervised
Moffitt <i>et al.</i>	Microarray (Affymetrix)	123	Validation	(?)
PACA-AU (array)	Microarray (Agilent)	63	Validation	(?)
PACA-AU (RNA-seq)	RNA-seq (Illumina HiSeq)	52	Validation	(?)
Puleo <i>et al.</i>	Microarray (Affymetrix)	288	Validation	(?)

restricted to the top 50 factor-specific genes from all prognostic factors prior to clustering. The number of clusters was selected via inspection of cumulative distribution function (CDF) plots, relative change in area under the CDF curve, consensus heatmaps, and cluster consensus scores (Fig Sx). After clustering, clusters were manually aligned across datasets based on expression of the factor-specific signatures in each cluster.

Survival analysis. For pooled survival analysis, patients with available OS time and censoring indicator were included. Overall survival estimates were calculated using the Kaplan-Meier method. Association between overall survival and cluster membership were evaluated via the cox proportional hazards models using the coxph function from the ‘survival’ R package (version 3.2-13), where cluster was considered as a multi-level categorical predictor. The log-rank test was used to evaluate the association between cluster membership and overall survival and derive the p-values. In the pooled analyses, a stratified cox proportional hazards model was utilized, where dataset of origin was used as a stratification factor to account for variation in baseline hazard across studies.

ACKNOWLEDGMENTS. Please include your acknowledgments here, set in a single paragraph. Please do not include any acknowledgments in the Supporting Information, or anywhere else in the manuscript.

- Pareja F, et al. (2016) Triple-negative breast cancer: The importance of molecular and histologic subtyping, and recognition of low-grade variants. *NPJ breast cancer* 2(1):1–11.
- Dienstmann R, Salazar R, Tabernero J (2018) Molecular subtypes and the evolution of treatment decisions in metastatic colorectal cancer. *Am Soc Clin Oncol Educ Book* 38(38):231–8.
- Zhou X, et al. (2021) Clinical impact of molecular subtyping of pancreatic cancer. *Frontiers in cell and developmental biology* 9:743908.
- Seiler R, et al. (2017) Impact of molecular subtypes in muscle-invasive bladder cancer on predicting response and survival after neoadjuvant chemotherapy. *European urology* 72(4):544–554.
- Prat A, et al. (2015) Clinical implications of the intrinsic molecular subtypes of breast cancer. *The Breast* 24:S26–S35.
- Bertucci F, et al. (2005) Gene expression profiling identifies molecular subtypes of inflammatory breast cancer. *Cancer research* 65(6):2170–2178.
- Jézéquel P, et al. (2015) Gene-expression molecular subtyping of triple-negative breast cancer tumours: Importance of immune response. *Breast Cancer Research* 17(1):43.
- Marisa L, et al. (2013) Gene expression classification of colon cancer into molecular subtypes: Characterization, validation, and prognostic value. *PLoS medicine* 10(5):e1001453.
- Collisson EA, Bailey P, Chang DK, Biankin AV (2019) Molecular subtypes of pancreatic cancer. *Nature reviews Gastroenterology & hepatology* 16(4):207–220.
- Hoadley KA, et al. (2014) Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. *Cell* 158(4):929–944.
- Tirosh I, et al. (2016) Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* 352(6282):189–196.
- Rashid NU, et al. (2020) Purity independent subtyping of tumors (PurIST), a clinically robust, single-sample classifier for tumor subtyping in pancreatic cancer. *Clinical Cancer Research* 26(1):82–92.
- Ren X, et al. (2021) Insights gained from single-cell analysis of immune cells in the tumor microenvironment. *Annual review of immunology* 39(1):583–609.
- Tomczak K, Czerwińska P, Wiznerowicz M (2015) Review the cancer genome atlas (TCGA): An immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia* 2015(1):68–77.
- Zhang J, et al. (2019) The international cancer genome consortium data portal. *Nature biotechnology* 37(4):367–369.
- Nguyen H, Nguyen H, Tran D, Draghici S, Nguyen T (2024) Fourteen years of cellular deconvolution: Methodology, applications, technical evaluation and outstanding challenges. *Nucleic Acids Research* 52(9):4761–4783.
- Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *nature* 401(6755):788–791.
- Bailey P, Chang DK, et al. (2016) Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* 531(7592):47–52.
- Collisson EA, et al. (2011) Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. *Nature medicine* 17(4):500–503.

20. Moffitt RA, et al. (2015) Virtual microdissection identifies distinct tumor-and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature genetics* 47(10):1168–1178.
21. Peng XL, Moffitt RA, Torphy RJ, Volmar KE, Yeh JJ (2019) De novo compartment deconvolution and weight estimation of tumor samples using DECODER. *Nature communications* 10(1):4729.
22. Brunet J-P, Tamayo P, Golub TR, Mesirov JP (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences* 101(12):4164–4169.
23. Bair E, Tibshirani R (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS biology* 2(4):e108.
24. Cox DR (1972) Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* 34(2):187–202.
25. Peng XL, et al. (2024) Determination of permissive and restraining cancer-associated fibroblast (DeCAF) subtypes. *bioRxiv*.
26. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791.
27. Simon N, Friedman JH, Hastie T, Tibshirani R (2011) Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of statistical software* 39:1–13.
28. Wilkerson MD, Hayes DN (2010) ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking. *Bioinformatics* 26(12):1572–1573.

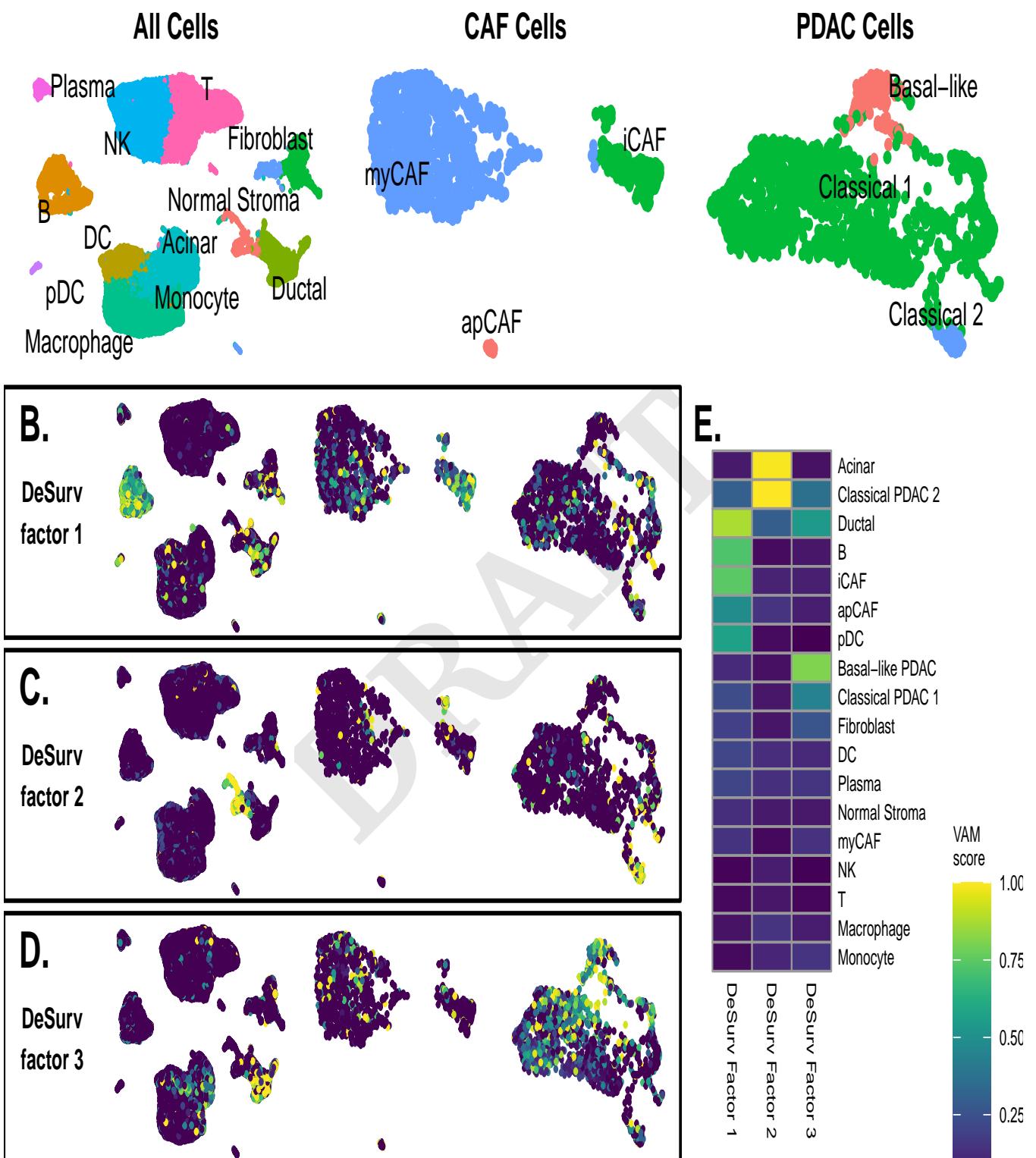


Fig. 5. A. Cell type clusters from the Elyada scRNA-seq data. B-D. VAM scores for the DeSurv factor signatures in each cell shown on the UMAP of cells in the Elyada-sc data. E. Heatmap of average VAM scores by cell type and factor signature in the Elyada-sc data.