Survival driven deconvolution (deSurv) reveals clinically relevant tumor and stromal gene signatures

Amber Young[a,1,2], **Alisa**[b], **Didong**[a], **and Naim**[a,c]

[a]University of North Carolina at Chapel Hill, Department, Street, City, State, Zip; [b]Another University Department, Street, City, State, Zip

This manuscript was compiled on August 20, 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, bladder, and colorectal cancer, we uncover previously unrecognized gene signatures linked to tumor, stromal, and immune compartments, including â€¦ Several identified signatures exhibit consistent prognostic value across cohorts and cancer types and demonstrate potential as therapeutic targets or biomarkers. 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 stratification of cancer patients based on distinct gene expression patterns. This stratification informs prognosis, guides therapeutic decisions, and enhances our understanding of tumor biology. However, despite considerable progress, current approaches often depend on unsupervised learning techniques, which may not reliably capture the prognostic relevance of specific cell-type contributions. As a result, many proposed gene expression signatures are evaluated retrospectively for clinical relevance and may lack consistent replication across independent cohorts and cancer types.

A key limitation lies in the disconnect between molecular subtyping and clinical outcomes. Most methods do not explicitly incorporate survival information during signature discovery, potentially overlooking gene programs with true prognostic value. Moreover, the complex interplay between malignant, stromal, and immune compartments in the tumor microenvironment presents an additional challenge to disentangling biologically meaningful and clinically actionable signals.

To address these issues, we developed deSurv, a Survival-driven Deconvolution framework that integrates bulk RNA-sequencing data with patient survival outcomes to uncover cell-type-enriched gene signatures with prognostic relevance. By aligning molecular signals with clinical endpoints, deSurv provides a more targeted and interpretable approach to gene signature discovery. Applying this method to diverse cancer cohorts, we identify novel prognostic markers within tumor, stromal, and immune compartments, offering new insights into the cellular basis of patient outcomes and revealing candidates for biomarker development or therapeutic intervention.

Results

DeSurv incorporates patient survival information directly into deconvolution. We developed deSurv, a survival-driven deconvolution framework that integrates bulk RNA-sequencing data with patient outcome information to identify cell-type-enriched gene signatures with prognostic value. Unlike standard unsupervised methods that cluster expression patterns without regard to patient survival, deSurv incorporates time-to-event data directly into the signature discovery process. We applied deSurv to bulk RNA-seq profiles from pancreatic, bladder, and colorectal cancer cohorts, each with matched clinical follow-up data, encompassing a total of n = XXXX patients across discovery and validation sets (Table 1). An overview of DeSurv can be found in Figure @ref(fig:fig-schema)

```
knitr::include_graphics("desurv_schematic_v5.pdf")
```

DeSurv captures distinct cell-type specific gene signatures. In PDAC, deSurv identified cell-type-specific signatures spanning tumor, stromal, and immune compartments. Many signatures were distinct from those obtained using unsupervised methods (Figure 2a and 2b). Representative heatmaps illustrate the differential expression of these signatures across patients (Figure 2c).

DeSurv extracts prognostic tumor signatures. Tumor signatures derived using deSurv stratified patients into groups with significantly different survival outcomes (log-rank $P < 0.001$ in all datasets), often outperforming unsupervised methods (Figures 3a-c). For example, a deSurv tumor signature achieved a

Significance Statement

Authors must submit a 120-word maximum statement about the significance of their research paper written at a level understandable to an undergraduate educated scientist outside their field of speciality. The primary goal of the Significance Statement is to explain the relevance of the work in broad context to a broad readership. The Significance Statement appears in the paper itself and is required for all research papers.

Please provide details of author contributions here.

Please declare any conflict of interest here.

[2] To whom correspondence should be addressed. E-mail: bob@email.com

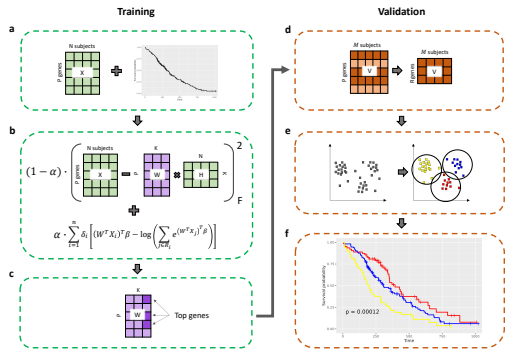

Fig. 1. DeSurv overview

concordance index (C-index) of 0.72 compared to 0.61 for the nearest unsupervised equivalent (Table 1). Performance gains were consistent in independent validation cohorts (Figures 3d-e) , indicating that survival integration during signature discovery enhances prognostic robustness.

DeSurv extracts prognostic stromal factors. Figure: Panel A , Panel B, Panel C,

At k=9, DeSurv finds an iCAF factor that is not found in standard NMF. When we cluster on this factor in the validation datasets the resulting clusters are prognostic for patient survival. Note that almost all other stromal factors are not associated with survival.

Cross-cancer robustness of prognostic signatures. Several deSurv-derived signatures retained prognostic value when applied to other cancer types. A tumor signature discovered in PDAC was also prognostic in colorectal cancer (log-rank P = xx), and a stromal signature from pancreatic cancer predicted improved survival in bladder cancer (Figure 4a). Heatmaps of hazard ratios across cross-cancer applications revealed that X% of signatures demonstrated statistically significant associations in at least two cancer types (Figure 4b), suggesting a degree of pan-cancer prognostic relevance.

Discussion

Materials and methods

Standard NMF. Let X be a bulk gene expression matrix of p genes by n subjects. Standard NMF seeks to reconstruct X using two nonnegative matrices W and H such that $X = WH$, where W is a $p \times k$ matrix of gene weights and H is a $k \times n$ matrix of sample weights. This is done by minimizing the loss function

$$||X - WH||_F^2 \qquad [1]$$

where F represents the Frobenius norm.

Multiplicative updates were proposed by Lee and Seung with the following update rules (1):

$$W_{ij} = \frac{W_{ij}(XH^T)_{ij}}{(WHH^T)_{ij}} \qquad [2]$$

$$H_{ij} = H_{ij}\frac{(W^TX)_{ij}}{(W^TWH)_{ij}} \qquad [3]$$

These updates are alternated until convergence to a stationary point.

Proportional Hazards. To determine how the lower dimensional representation of X is associated with patient survival outcomes, we take $Z = W^TX$ to be the covariates passed to the proportional hazards model. The matrix Z can be interpreted as the transformation of the data matrix X into the lower dimensional space, such that Z_{ri} represents a score for the contribution of factor r to subject i.

Let $y_i = \min(T_i, C_i)$ where T_i is the event time and C_i is the censoring time for the ith subject; let δ_i represent the indicator that the event time for the ith subject is observed. The the log partial likelihood is

$$\ell(W,\beta) = \sum_{i=1}^{n}\delta_i\left[Z_i^T\beta - \log\left(\sum_{j=1}^{n}\exp\left(Z_j^T\beta\right)\mathbb{1}(y_j \geq y_i)\right)\right] \qquad [4]$$

DeSurv. DeSurv is a semi-supervised extension of NMF that incorporates the cox proportional hazards directly into the NMF model to encourage the discovered factors to be associated with patient survival. We propose the following loss function

$$\mathcal{L}(W,H,\beta) = \frac{(1-\alpha)}{2np}||X-WH||_F^2 - \alpha(\frac{2}{n}\ell(W,\beta) - \lambda p_\xi(\beta)) + \frac{\gamma}{2pk}||W||_F^2 + \frac{\nu}{2n} \qquad [5]$$

where α is the hyperparameter that balances the contribution of the NMF and proportional hazards model to the overall loss. Penalty terms $||W||_F^2$, and $||H||_F^2$ provide additional stability for the model. We take

$$p_\xi(\beta) = \xi||\beta||_1 + \frac{(1-\xi)}{2}||\beta||_2^2 \qquad [6]$$

to be an elastic net penalty on the regression coefficients β. The L1 component allows factors that are not associated with survival to be shrunk out of the proportional hazards model, while still contributing to the reconstruction of X.

Update Rules for DeSurv. To solve this loss function, we propose an update scheme that alternates between updating W, H, and β until convergence. The algorithm is summarized in Algorithm 1.

Update for W. To get the update rule for W, we must find

$$\underset{W\geq 0}{\operatorname{argmin}}\,\mathcal{L}(W,H,\beta) \qquad [7]$$

The derivative of the loss with respect to W is

$$\frac{\partial\mathcal{L}}{\partial W} = \frac{(1-\alpha)}{s}(WH - X)H^T - \frac{2\alpha}{n}\frac{\partial\ell}{\partial W} + \frac{\gamma}{pk}W \qquad [8]$$

where $\frac{\partial\ell}{\partial W}$ is the derivative of the partial likelihood with respect to W

Algorithm 1 DeSurv algorithm

Input: $X \in \mathbb{R}_{\geq 0}^{p \times n}$, $y \in \mathbb{R}_{\geq 0}^{n}$, $\delta \in \mathbb{R}_{0,1}^{n}$

$eps \leftarrow \infty$
$iter \leftarrow 0$
$W_{jr} \sim Unif(0, \max(X))$ for $j = 1, \ldots, p$ and $r = 1, \ldots, k$
$H_{ri} \sim Unif(0, \max(X))$ for $r = 1, \ldots, k$ and $i = 1, \ldots, n$
while $eps < tol$ **and** $iter < maxit$ **do**
    $W \leftarrow \text{argmin}_{W \geq 0} \mathcal{L}(W, H, \beta)$
    $H \leftarrow \text{argmin}_{H \geq 0} \mathcal{L}(W, H, \beta)$
    $\beta \leftarrow \text{argmin}_{\beta} \mathcal{L}(W, H, \beta)$
    $errNew \leftarrow \mathcal{L}(W, H, \beta)$
    $relErr \leftarrow |errNew - err|/err$
    $err \leftarrow errNew$
    $iter \leftarrow iter + 1$
return W, H, β

$$\frac{\partial \ell}{\partial W} = \sum_{i=1}^{n} \delta_i \left[X_i - \frac{\sum_{j=1}^{n} e^{Z_j^T \beta} \mathbb{1}(y_j \geq y_i) X_j}{\sum_{j=1}^{n} e^{Z_j^T \beta} \mathbb{1}(y_j \geq y_i)} \right] \beta^T \quad [9]$$

A multiplicative update for W can then be formed as

$$W = W \odot \max\left(\frac{XH^T + \frac{2\alpha s}{n(1-\alpha)} \frac{\partial \ell}{\partial W}}{WHH^T + \frac{\gamma s}{pk(1-\alpha)} W}, 0 \right) \quad [10]$$

note that the $\max(*, 0)$ is necessary because the derivative of the partial likelihood is not guaranteed to be nonnegative.

Update for H. The update for H is a standard NMF multiplicative update.

$$H = H \odot \frac{W^T X}{W^T W H + \frac{\nu s}{nk(1-\alpha)} H} \quad [11]$$

Update for β. To get the update for β we take

$$\beta = \underset{\beta}{\text{argmin}}\, \mathcal{L}(W, H, \beta) \quad [12]$$

which is equivalent to

$$\beta = \underset{\beta}{\text{argmax}}\, \frac{2}{n} \ell(W, \beta) - \lambda p_{\xi}(\beta) \quad [13]$$

Note that this is the same form as a standard cox partial likelihood update. So the β update as derived in (**?**) is

$$\hat{\beta}_r = \frac{S(\frac{1}{n} \sum_{i=1}^{n} w(\tilde{\eta})_i v_{i,r} \left[z(\tilde{\eta})_i - \sum_{j \neq r} v_{ij} \beta_j, \right], \lambda \xi)}{\frac{1}{n} \sum_{i=1}^{n} w(\tilde{\eta})_i v_{i,r}^2 + \lambda(1 - \xi)} \quad [14]$$

where

$$S(x, \lambda) = sgn(x)(|x| - \lambda)_+. \quad [15]$$

$$w(\tilde{\eta})_r = \ell''(\tilde{\eta})_{r,r} = \sum_{i \in C_r} \left[\frac{e^{\tilde{\eta}_r} \sum_{j \in R_i} e^{\tilde{\eta}_j} - \left(e^{\tilde{\eta}_r}\right)^2}{\left(\sum_{j \in R_i} e^{\tilde{\eta}_j}\right)^2} \right] \quad [16]$$

$$z(\tilde{\eta})_r = \tilde{\eta}_r - \frac{\ell'(\tilde{\eta})_r}{\ell''(\tilde{\eta})_{r,r}} = \tilde{\eta}_r + \frac{1}{w(\tilde{\eta})_r} \left[\delta_r - \sum_{i \in C_r} \left(\frac{e^{\tilde{\eta}_r}}{\sum_{j \in R_i} e^{\tilde{\eta}_j}} \right) \right] \quad [17]$$

A. Publicly Available Datasets.

Text

B. Model Training. DeSurv was applied to the TCGA dataset. The data were log-transformed for variance stabilization and then quantile normalized to ensure comparability of expression values across samples. Next, the data was filtered to the top 5000 highly expressed and variable genes. Models were trained across a grid of hyperparameters $\alpha \in \{0, .95\}$, $\lambda \in$, $\xi \in$, $\gamma \in$, $\nu \in$, and $k = 2, \ldots, 15$

B.1. Hyperparameter selection. The hyperparameters α, λ, ξ, γ, and ν were selected to adequately balance the supervised and unsupervised portions of the model using a metric we defined as the c-index of the proportional hazards model divided by the reconstruction error. The parameters were chosen to maximize this metric. Since the reconstruction error exclusively decreases as the dimension k increases, this metric was not adequate to choose k.

B.2. Top genes. The top genes were extracted from each factor of W in the selected model at each value of k. A top gene was defined as ...

C. Model Validation. The remaining 7 publicly available datasets, CPTAC, Dijk, Linehan, Moffitt, PACA microarray, PACA RNAseq, and Puleo, were used to validate our models. The datasets were log transformed and merged. To mitigate between study and between platform heterogeneities, the samples were rank transformed. For each selected model, the merged data was restricted to the top genes in each factor and clustered to determine patient subtypes.

C.1. Clustering. Consensus clustering was performed with the ConsensusClusterPlus package in R, using the Kmeans algorithm and euclidean distance. Each repetition samples 80% of subjects and 80% of the top genes. To account for the difference in sample size across studies, subjects were sampled with weight \$1/Dn_d \$ where D is the number of studies in the validation set, and n_d is the number of subjects in dataset d. The number of clusters ranged from 2-3.

C.2. Survival Analysis. After clustering, stratified cox models were fit to the clusters in the merged validation dataset. From these models, the following metrics were obtained: the hazard ratio, c-index, BIC, and p-values for the likelihood ratio test. These metrics were used to compare our approach to the unsupervised NMF equivalent.

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.

1. Lee D, Seung HS (2000) Algorithms for non-negative matrix factorization. *Advances in neural information processing systems* 13.