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

Abstract

Please provide an abstract of no more than 250 words.

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). The pipeline …

## 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 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.

# Materials and methods

## Standard NMF

Let be a bulk gene expression matrix of genes by subjects. Standard NMF seeks to reconstruct X using two nonnegative matrices and such that , where is a matrix of gene weights and is a matrix of sample weights. This is done by minimizing the loss function

where represents the Frobenius norm.

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

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 to be the covariates passed to the proportional hazards model. The matrix can be interpreted as the transformation of the data matrix into the lower dimensional space, such that represents a score for the contribution of factor to subject .

Let where is the event time and is the censoring time for the th subject; let represent the indicator that the event time for the th subject is observed. The the log partial likelihood is

## 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

where is the hyperparameter that balances the contribution of the NMF and proportional hazards model to the overall loss. Penalty terms , and provide additional stability for the model. We take

to be an elastic net penalty on the regression coefficients . 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 .

## Update Rules for DeSurv

To solve this loss function, we propose an update scheme that alternates between updating , , and until convergence. The algorithm is summarized in Algorithm .

## Update for

To get the update rule for , we must find

The derivative of the loss with respect to is

where is the derivative of the partial likelihood with respect to

A multiplicative update for W can then be formed as

note that the is necessary because the derivative of the partial likelihood is not guaranteed to be nonnegative.

## Update for

The update for is a standard NMF multiplicative update.

## Update for

To get the update for we take

which is equivalent to

Note that this is the same form as a standard cox partial likelihood update. So the update as derived in is

where

Text

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 , , , , , and

The hyperparameters , , , , and 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 increases, this metric was not adequate to choose .

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

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.

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 is the number of studies in the validation set, and is the number of subjects in dataset . The number of clusters ranged from 2-3.

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.

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