NCI R21: Cancer-specific gene set testing

Rob Frost

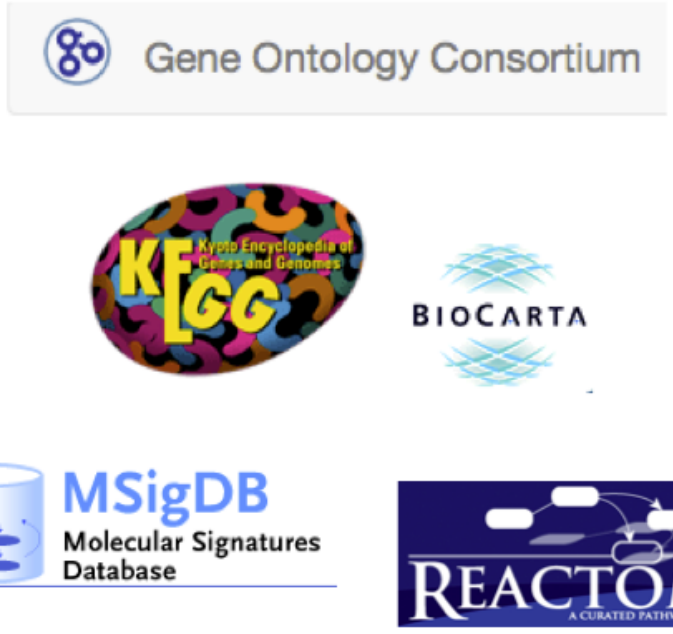
Department of Biomedical Data Science
Geisel School of Medicine at Dartmouth

Outline

- Motivation
- Aims
- **Tissue-adjusted pathway analysis of cancer (TPAC)**

Gene set testing, or pathway analysis

Test hypotheses about statistics computed for functionally related groups of genes rather than just single genes.



Improves **interpretability, replication and statistical power**.

Gene set testing challenges for cancer genomics

1. Mismatch between gene set annotations and gene activity in neoplastic tissue.
2. Failure to account for gene activity in associated normal tissue during gene set testing of cancer data.

Grant Aims

1. Mismatch between gene set annotations and gene activity in neoplastic tissue.

   $\rightarrow$ **Aim 1: Customize existing gene set collections for common human solid cancers.**

2. Failure to account for gene activity in associated normal tissue during gene set testing of cancer data.

Grant Aims

1. Mismatch between gene set annotations and gene activity in neoplastic tissue.
2. Failure to account for gene activity in associated normal tissue during gene set testing of cancer data.

$\rightarrow$ **Aim 2: Develop cancer gene set testing methods that adjust for gene activity in the associated normal tissue.**

New Results

Tissue-adjusted pathway analysis of cancer (TPAC)

H. Robert Frost

doi: https://doi.org/10.1101/2022.03.17.484779

Tissue-adjusted pathway analysis of cancer (TPAC)

Computes a tumor-specific gene set scores to convert a tumor-by-gene expression matrix:

$$\mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{bmatrix}$$

into a tumor-by-set matrix:

$$\mathbf{S} = \begin{bmatrix} s_{1,1} & \cdots & s_{1,m} \\ \vdots & \ddots & \vdots \\ s_{n,1} & \cdots & s_{n,m} \end{bmatrix}$$

TPAC features

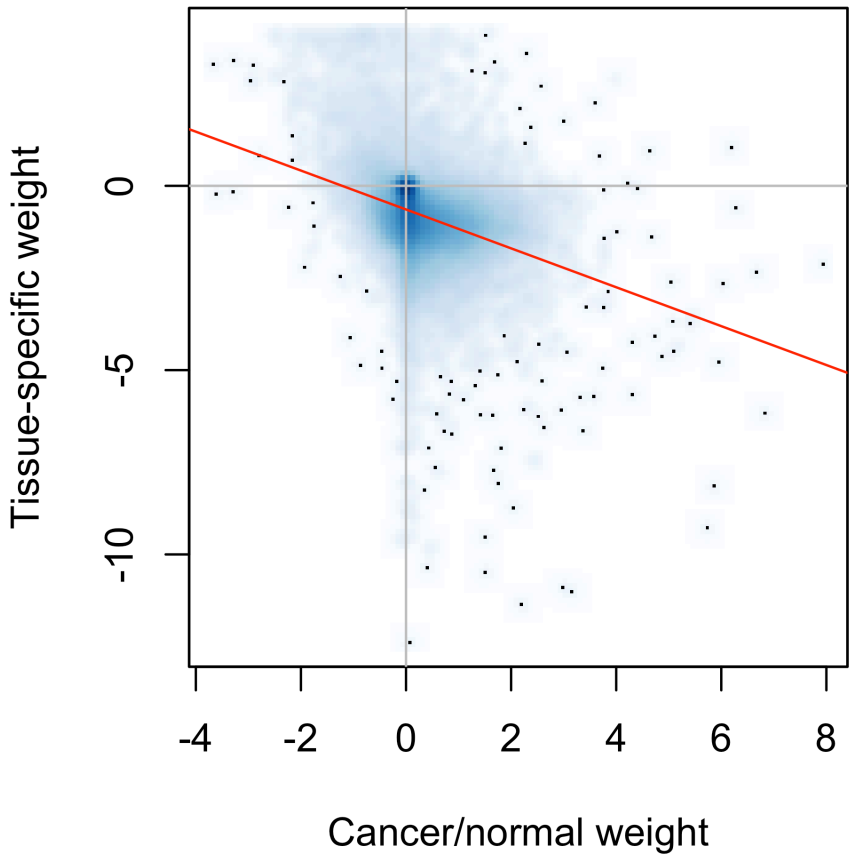
- Sample-level scores enable downstream analyzes (visualization, clustering, DE testing, etc.) on the level of gene sets.
- Scores have a gamma null distribution which enables sample-level inference.
- **Score generation leverages information regarding the specificity of genes in the associated normal tissue.**

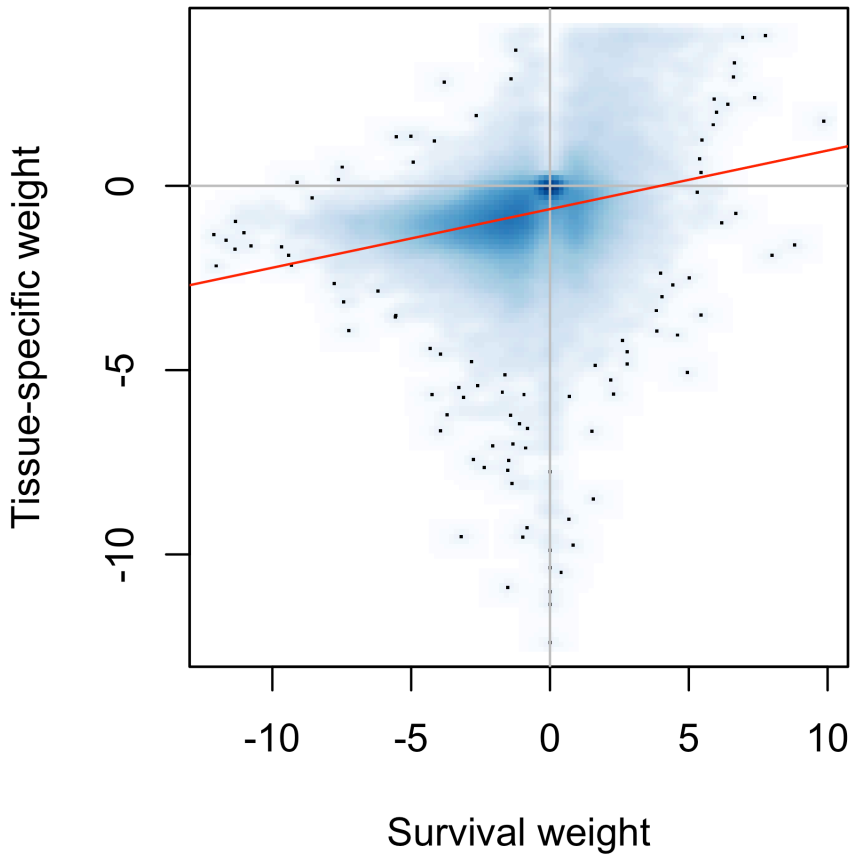
  Initial support for 21 TCGA cancer types and 18 corresponding normal tissue types.

Association between tissue-specificity and cancer

Genes with elevated expression in a given tissue compared to other tissues:

- Tend to be down-regulated in the corresponding cancer.
- Are favorably prognostic.





Frost, HR. Analyzing cancer gene expression data through the lens of normal tissue specificity. PLOS Comp Bio, 2021.

TPAC algorithm

- TPAC generates three scores matrices: $\mathbf{S}$, $\mathbf{S}^+$, and $\mathbf{S}^-$
- Elements of $\mathbf{S}$, $\mathbf{S}^+$, and $\mathbf{S}^-$ are computed as gamma CDF values for modified Mahalanobis distances (i.e., variance-adjusted multivariate Euclidean distances).

TPAC algorithm, continued

- Distances are measured from the mean in the associated normal tissue rather than mean across tumors in $\mathbf{X}$.
- $\mathbf{S}^+$ captures the up-regulated portion, $\mathbf{S}^-$ the down-regulated portion, and $\mathbf{S}$ both up and down-regulation.
- Normal tissue-specificity is used to adjust the sample covariance matrix used to compute the Mahalanobis distances.

Choice of $\mathbf{S}$, $\mathbf{S}^+$ or $\mathbf{S}^-$

$\mathbf{S}^+$:

- Large values in $\mathbf{S}^+$ correspond to tumors where expression of pathway genes is elevated relative to the associated normal tissue.
- Use of normal tissue-specificity to adjust sample variances prioritizes expression differences for genes that are normally suppressed in the associated normal tissue.

Choice of $\mathbf{S}$, $\mathbf{S}^+$ or $\mathbf{S}^-$, continued

$\mathbf{S}^-$:

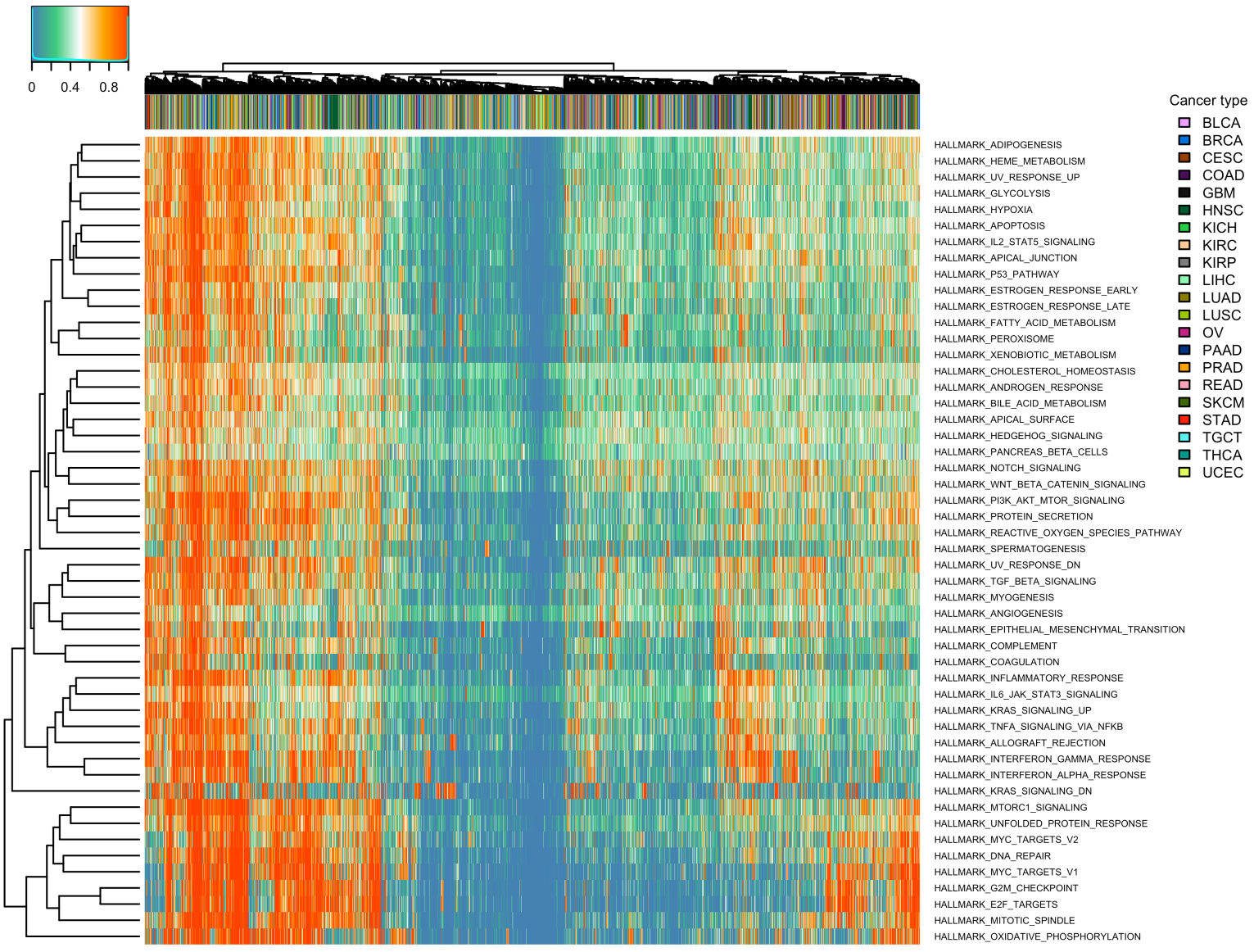
- Large values in $\mathbf{S}^-$ correspond to tumors where expression of pathway genes is down-regulated relative to the associated normal tissue.
- Use of normal tissue-specificity to adjust sample variances leads to larger $\mathbf{S}^-$ values when tissue-specific genes are down-regulated in the tumor.

Choice of $\mathbf{S}$, $\mathbf{S}^+$ or $\mathbf{S}^-$, continued

$\mathbf{S}$:

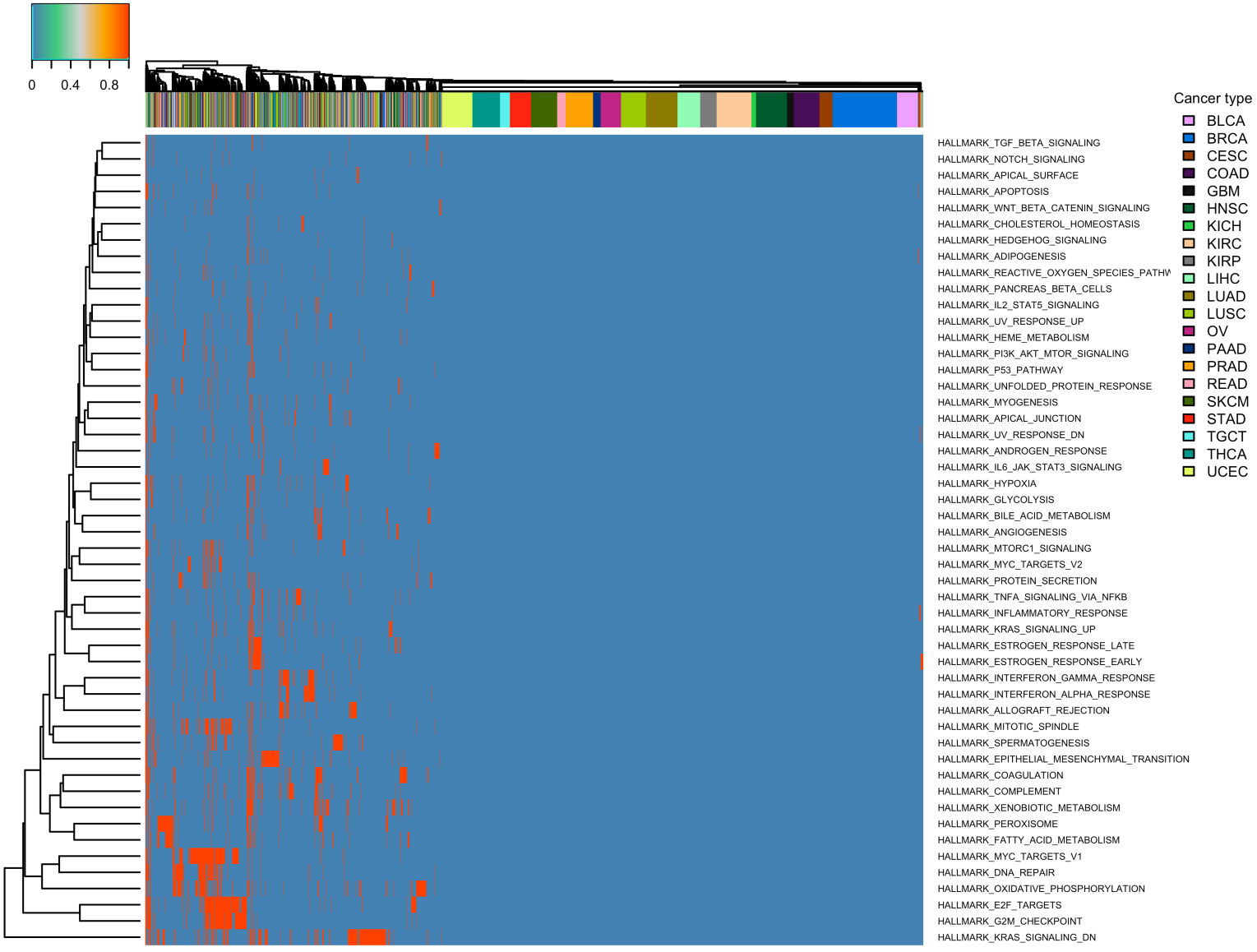
- Large values in $\mathbf{S}$ correspond to tumors where expression of pathway genes exhibit a combination of up and down-regulation relative to the associated normal tissue.

Landscape of pan-cancer pathway dysregulation



S matrix for MSigDB Hallmark pathways and 21 TCGA cohorts.

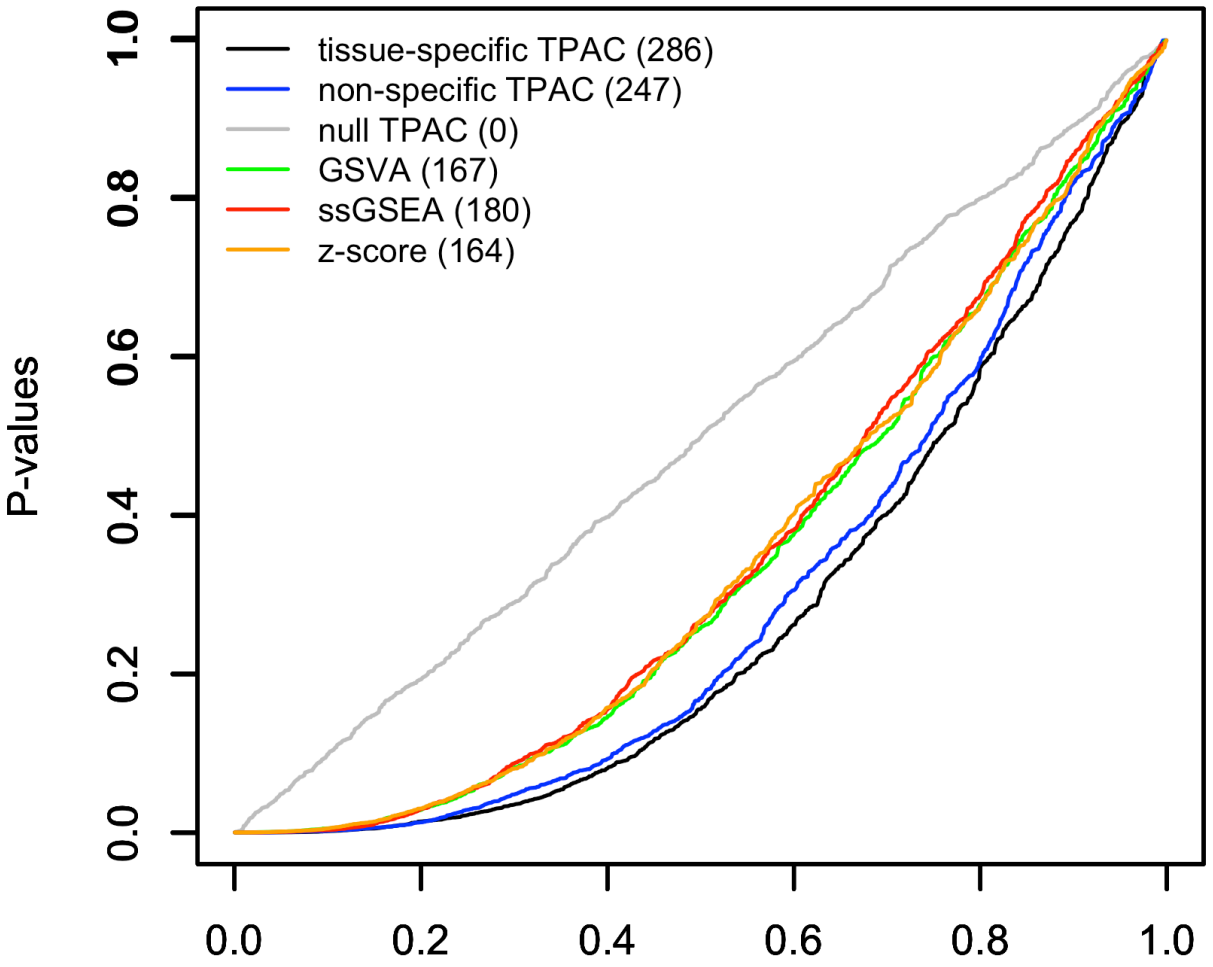
Single tumor inference



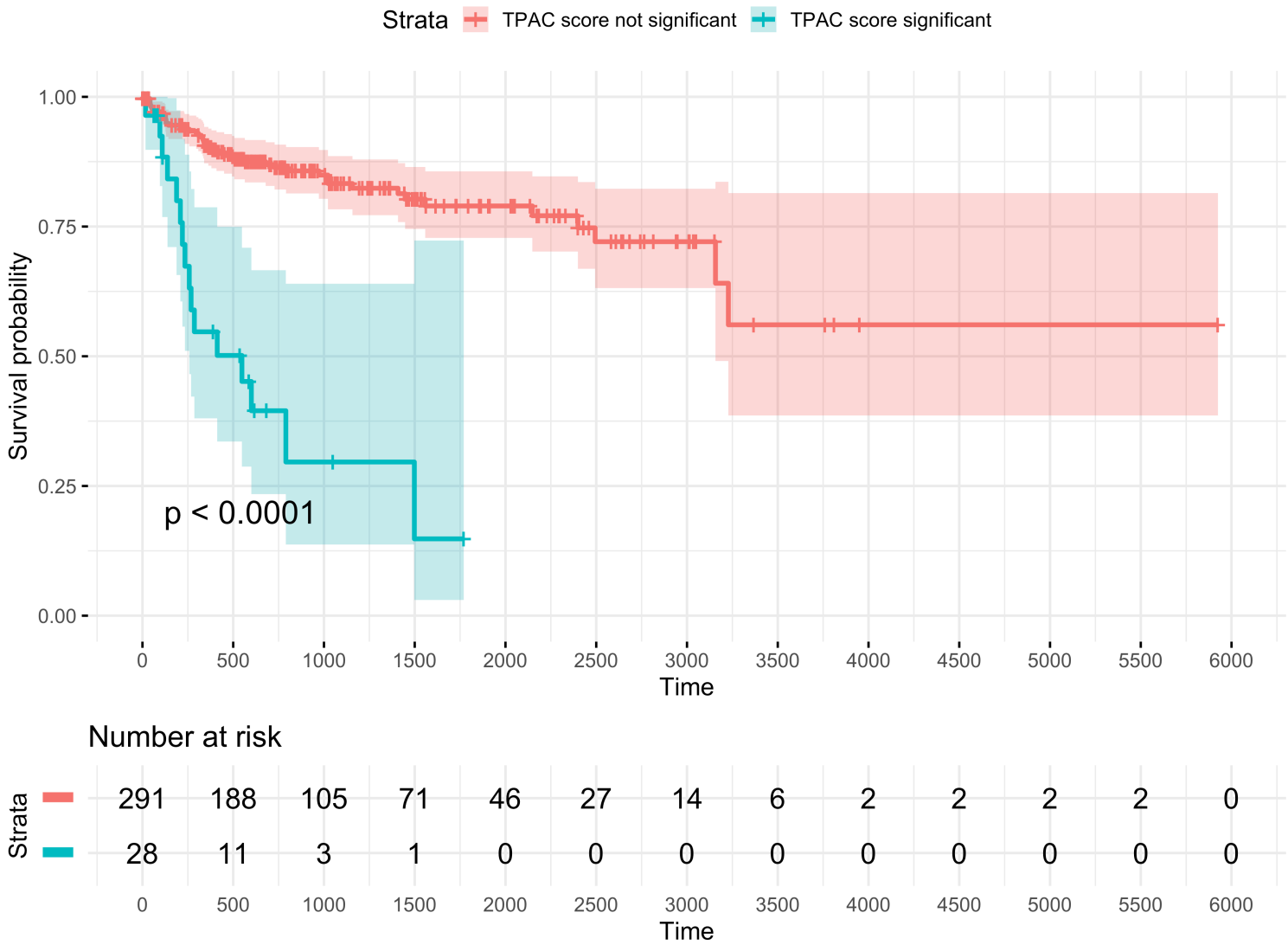
S discretized according to FDR ≤ 0.3

Survival analysis

QQ plot of p-values from Cox model relative to PFI for each S column predictor (#s are FDR ≤ 0.1).



Kaplan-Meir based on TPAC significance for Hallmark MYC Targets



KM for PFI using FDR cutoff of 0.25 to stratify samples.

Conclusions

- TPAC is a novel single sample gene set testing method for cancer transcriptomics.
- Leverages normal tissue-specificity to improve performance.
- Generated scores can be used with or without a probabilistic interpretation.

Acknowledgments

- NIH grants R21CA253408, P20GM130454, P30CA023108
- Department of Biomedical Data Science at Dartmouth
- Dartmouth Cancer Center

Questions?