

Isoform-level Expression Improves Survival Signal Selection and Prediction

Yung-Han Chang[1], S. Taylor Head[2], Sean Bresnahan[2], Ryan Sun[1], Arjun Bhattacharya[2,3]

[1]Department of Biostatistics, The University of Texas MD Anderson Cancer Center UTHealth Houston Graduate School of Biomedical Sciences, Houston, TX, USA [2]Department of Epidemiology, University of Texas MD Anderson Cancer Center, Houston, TX, USA, [3]Institute for Data Science in Oncology, University of Texas MD Anderson Cancer Center, Houston, TX, USA

Cancer UT MD Anderson

Motivation

Gene expression can predict breast cancer survival: PAM50

- **PAM50** is a widely used gene-expression signature in breast cancer that summarizes the expression of 50 genes to characterize molecular subtypes and derive risk of recurrence.

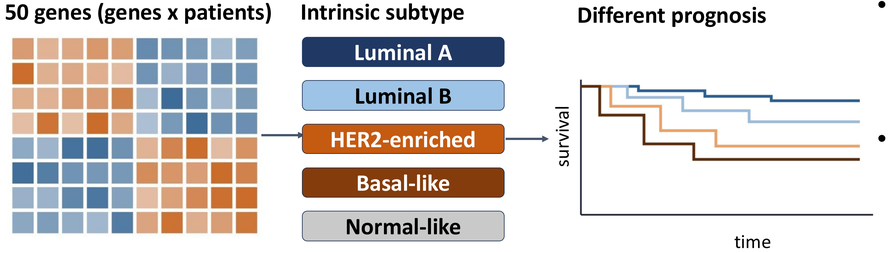


Figure 1: PAM50 illustration. PAM50 uses the expression of 50 genes to classify breast cancer patients into intrinsic molecular subtypes. These expression-defined subtypes are associated with distinct clinical outcomes and can also be used to derive recurrence risk scores for prognosis prediction.

Isoform expression can identify signals missed at gene-level

- A gene can produce multiple isoforms through alternative splicing, and different isoforms may have distinct effects on disease. By focusing only on gene-level expression, we may miss important isoform-specific signals.

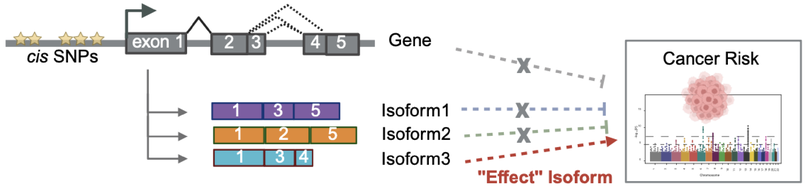


Figure 2: Importance of isoform-level expression. A gene may show no significant association with cancer risk when using total gene expression only. Isoform-level analysis can uncover signals that are masked at the gene level. From Chang et al., *British Journal of Cancer* (2025).

Methods and Results

Can isoforms improve survival prediction beyond genes?

- We want to evaluate whether isoform-level expression provides additional information for survival prediction beyond gene-level expression.
- We propose using a penalized regression framework to select isoforms and predict restricted mean survival time (RMST), which can be interpreted as the expected survival time up to a prespecified time point τ.
- The objective function:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n} \sum_{i=1}^{n} w_i \left[Y_i - \frac{\tau}{1 + e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}} \right]^2 + P_\lambda(\boldsymbol{\beta})$$

$$\text{where } Y_i = \min(t_i, \tau), \quad \text{and } \mu_i = \frac{\tau}{1+e^{-\mathbf{x}_i^\top \beta}} \in (0, \tau).$$

- We fit LASSO, elastic net, sparse group lasso with within-gene grouping, and sparse group lasso with exon-similarity grouping.

Simulation scenarios and results

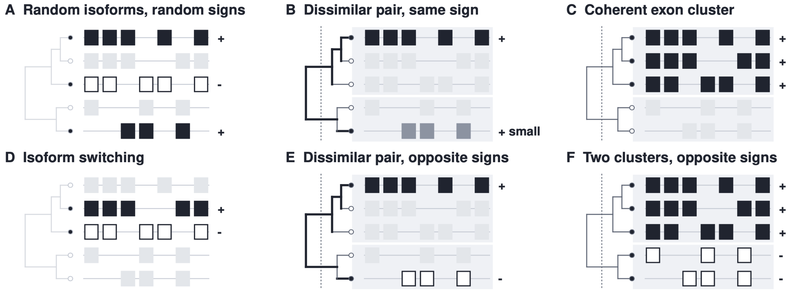


Figure 3: Simulation scenarios. Survival outcomes were simulated using real TCGA gene and isoform expression data (1,000 genes; 30 causal genes) under a Weibull model with 50% censoring. Scenarios varied effect direction, effect magnitude, isoform similarity (exon sharing), and within-gene grouping structure.

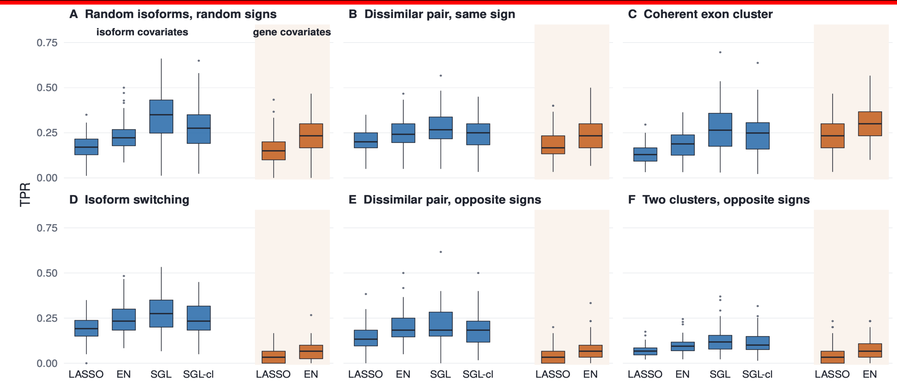


Figure 4: True positive rate (TPR) for gene- and isoform-level penalized models. Opposing isoform effects (D–F) substantially reduce gene-level selection accuracy, while isoform-level models retain signal. SGL with within-gene grouping generally achieves the highest TPR. EN: elastic net; SGL: sparse group lasso with within-gene grouping; SGL-cl: sparse group lasso with exon-similarity-based grouping.

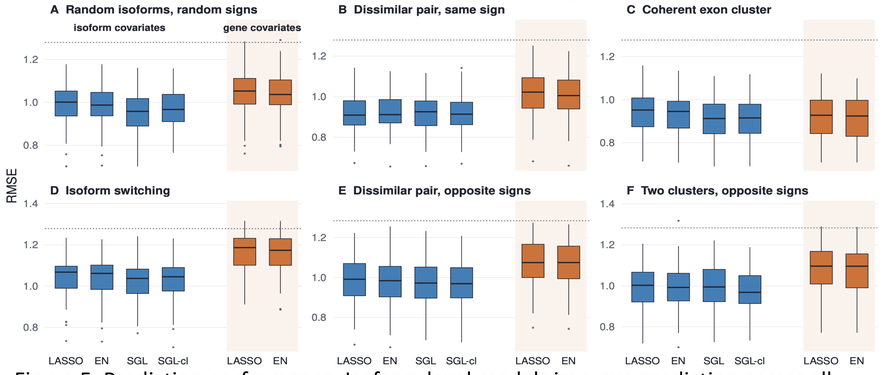


Figure 5: Prediction performance. Isoform-level models improve prediction across all scenarios, with the largest gains when isoform effects have opposite signs (D–F). The four isoform-level penalties perform similarly.

Future directions

- Incorporate isoform relationships as prior information through similarity-informed penalties or Bayesian priors.
- Apply the framework to TCGA and validate it using Carolina Breast Cancer Study data.