NCI R21: Cancer-specific gene set testing

Rob Frost

Department of Biomedical Data Science
Geisel School of Medicine at Dartmouth

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

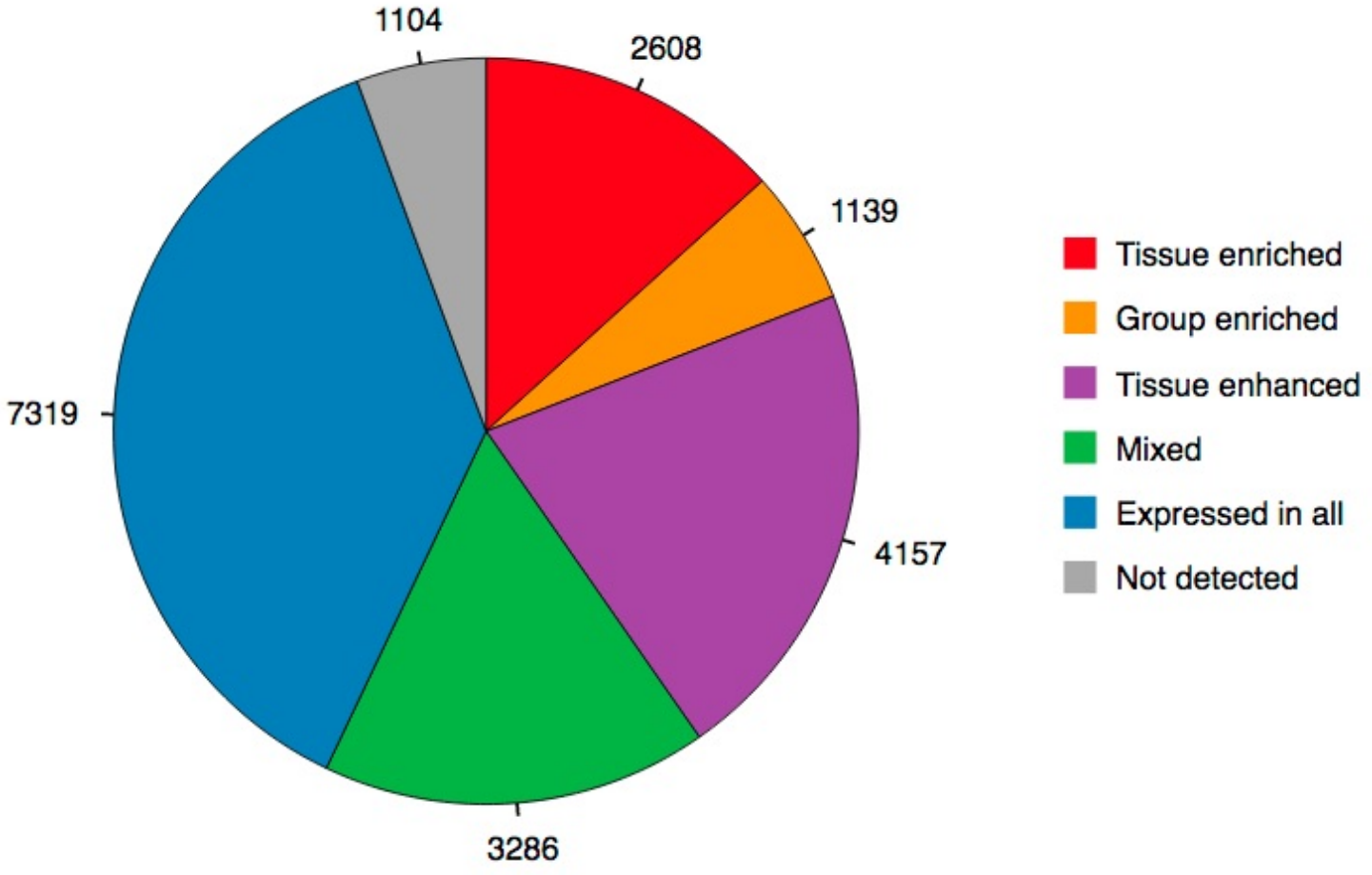
1. Mismatch between gene set annotations and gene activity in neoplastic tissue.

   $\rightarrow$ **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.

Aim 1 Scientific Premise

Majority of human genes have tissue-specific expression.



From `https://www.proteinatlas.org/humanproteome`

Aim 1 Scientific Premise, continued

Gene set testing is typically performed in a tissue-agnostic fashion ignoring:

- Source tissue of the experimental data under analysis.
- Tissue specificity of gene set members.
- Tissue associated with the experimental support for gene set annotations.

Aim 1 Scientific Premise, continued

Impact of ignoring tissue-specificity:

- Testing gene sets representing biological functions not present in experimental tissue $\rightarrow$ **low power**
- Pathway is active in experimental tissue but annotations based on a different tissue (or cell line) $\rightarrow$ **biased results**

Aim 1 Scientific Premise, continued

Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Computation and application of tissue-specific gene set weights

H. Robert Frost [1,*]

[1] Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, NH 03755, USA.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

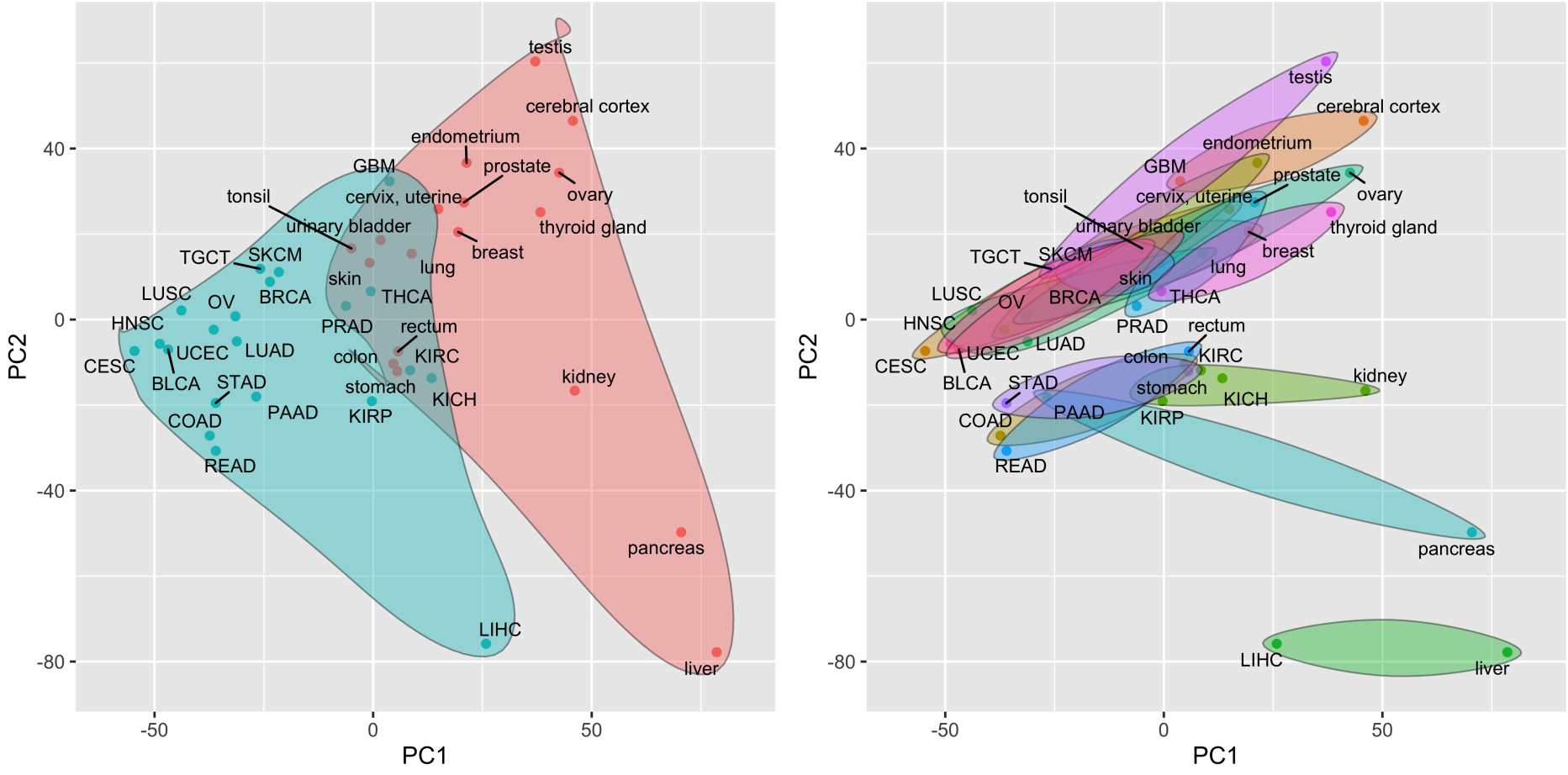
Abstract

Motivation: Gene set testing, or pathway analysis, has become a critical tool for the analysis of high-dimensional genomic data. Although the function and activity of many genes and higher-level processes is tissue-specific, gene set testing is typically performed in a tissue agnostic fashion, which impacts statistical power and the interpretation and replication of results.

Results: To address this challenge, we have developed a bioinformatics approach to compute tissue-specific weights for individual gene sets using information on tissue-specific gene activity from the Human Protein Atlas (HPA). We used this approach to create a public repository of tissue-specific gene set weights for 37 different human tissue types from the HPA and all collections in the Molecular Signatures Database (MSigDB). To demonstrate the validity and utility of these weights, we explored three different

Aim 1 Scientific Premise, continued

Normal tissue-specificity $\neq$ cancer-specificity



Aim 1 Approach

Create optimized versions of each MSigDB collection for each TCGA cohort.

Inputs:

- $\mathbf{T}$: $t \times p$ matrix of tumor gene expression data.
- $\mathbf{A}$: $m \times p$ gene set annotation matrix.

Output:

- $\mathbf{A}^*$: $m \times p$ matrix of gene set annotation proportions that capture annotation tumor-specificity.

Aim 1 Approach, continued

Given $\mathbf{T}$ and $\mathbf{A}$, compute $\mathbf{A}^*$ by:

- Cluster p genes into k disjoint clusters according to tumor expression data in $\mathbf{T}$.
- Compute a measure of entropy between each gene set and clusters and filter set to minimize measure.
- Repeat on many bootstrap resampled data sets for different k values.
- Elements in $\mathbf{A}^*$ are set to the proportion of iterations in which an annotation was retained.

Can use $\mathbf{A}^*$ directly or discretize.

Aim 1 Evaluation and Deliverables

Evaluation:

- Compare regression models (e.g., Cox) using predictors based on optimized and unoptimized pathways in terms of power, predictive performance, parsimony, and replication.

Deliverables:

- Public repository with optimized versions of each MSigDB collection for each TCGA cohort (both discretized collections and proportions).
- R package implementing the optimization logic.

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.

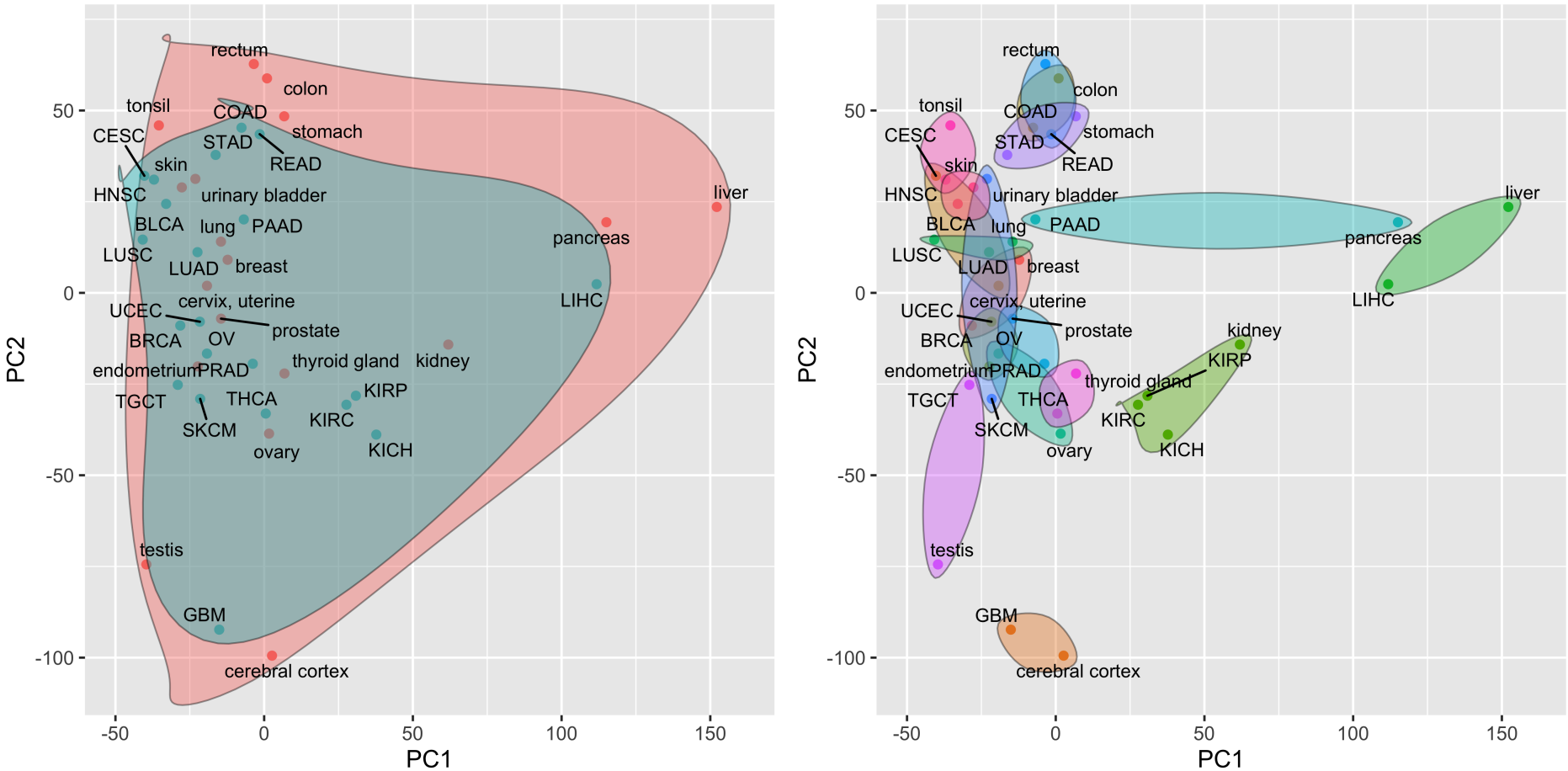
$\rightarrow$ **Aim 2: Develop cancer gene set testing methods that adjust for gene activity in the associated normal tissue.**

Aim 2 Scientific Premise

TCGA abbrev.	Cancer type	HPA tissue	Cancer/normal correlation	Most correlated
BLCA	Bladder Urothelial Carcinoma	urinary bladder	0.901	*
BRCA	Breast Invasive Carcinoma	breast	0.924	*
CESC	Cervical Squamous Cell Carcinoma and ...	cervix, uterine	0.852	urinary bladder
COAD	Colon Adenocarcinoma	colon	0.93	*
GBM	Glioblastoma Multiforme	cerebral cortex	0.895	*
HNSC	Head and Neck Squamous Cell Carcinoma	tonsil	0.862	skin
KICH	Kidney Chromophobe	kidney	0.932	*
KIRC	Kidney Renal Clear Cell Carcinoma	kidney	0.931	*
KIRP	Kidney Renal Papillary Cell Carcinoma	kidney	0.925	*
LIHC	Liver Hepatocellular Carcinoma	liver	0.929	*
LUAD	Lung Adenocarcinoma	lung	0.925	*
LUSC	Lung Squamous Cell Carcinoma	lung	0.886	urinary bladder
OV	Ovarian Serous Cystadenocarcinoma	ovary	0.817	cervix, uterine
PAAD	Pancreatic Adenocarcinoma	pancreas	0.886	stomach
PRAD	Prostate Adenocarcinoma	prostate	0.951	*
READ	Rectum Adenocarcinoma	rectum	0.912	colon
SKCM	Skin Cutaneous Melanoma	skin	0.838	urinary bladder
STAD	Stomach Adenocarcinoma	stomach	0.921	*
TGCT	Testicular Germ Cell Tumors	testis	0.698	urinary bladder
THCA	Thyroid Carcinoma	thyroid gland	0.934	*
UCEC	Uterine Corpus Endometrial Carcinoma	endometrium	0.868	cervix, uterine

Aim 2 Scientific Premise, continued

Relative expression between cancers largely explained by relative expression in associated normal tissues.



Aim 2 Scientific Premise, continued

Cancer Research

Research Article

Transcriptomic differences between primary colorectal adenocarcinomas and distant metastases reveal metastatic colorectal cancer subtypes

Yasmin Kamal, Stephanie L Schmit, Hannah J Hoehn, Christopher I. Amos, and H Robert Frost

DOI: 10.1158/0008-5472.CAN-18-3945 Check for updates

Published OnlineFirst June 25, 2019
doi: 10.1158/0008-5472.CAN-18-3945

Abstract

Approximately 20% of colorectal cancer patients present with metastases at the time of diagnosis, and therapies that specially target these metastases are lacking. We present a novel approach for investigating transcriptomic differences between primary colorectal cancers (CRC) and distant metastases, which may help to identify primary tumors with high risk for future dissemination and to

Aim 2 Approach

Single sample gene set test that adjusts for gene activity in associated normal tissue.

Inputs:

- **T**: $t \times p$ matrix of tumor gene expression data.
- **N**: $n \times p$ matrix of expression data from associated normal.
- **A**: $m \times p$ gene set annotation matrix.

Output:

- **S**: $t \times m$ matrix of tumor-specific gene set scores.

Aim 2 approach, continued

Given **T**, **N**, and **A**, compute **S** by:

- Compute sample covariance matrix for normal expression data: $\hat{\mathbf{\Sigma}}_N$
- Compute modified Mahalanobis distance for all m gene sets, save in $n \times m$ matrix **M**. For set k:

$$\mathbf{M}[, k] = diag(\mathbf{T}_k(\Sigma_{k,N})^{-1}\mathbf{T}_k^T)$$

- Compute modified Mahalanobis distances on permuted **T**. Store in $\mathbf{M}_p$.
- Fit gamma distribution to each column in $\mathbf{M}_p$.
- Compute elements of **S** using estimated gamma CDF:

$$\mathbf{S}[, k] = F_{\gamma(\hat{\alpha}_k, \hat{\beta}_k)}(\mathbf{M_p}[, k]) \quad (1)$$

Aim 2 Evaluation and Deliverables

Evaluation:

- Test single sample pathway scores as predictors in regression models. Proposed method should improve performance vs. existing single sample gene set tests.

Deliverables:

- R package implementing the gene set testing method. Will require pre-computing $\hat{\mathbf{\Sigma}}_N$.

Current Status

- Grant started in September.
- Basic computing infrastructure implemented.
- Evaluating Aim 1 approach on MSigDB collections and TCGA data.

Acknowledgments

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

Questions?