Addressing Scale Uncertainty in Differential Set Analysis

Kyle McGovern

The Pennsylvania State University
kvm6065@psu.edu

GLBIO 2024
May 8, 2024

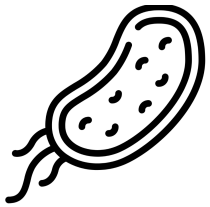
What scientific question are we asking of our sequence count data?

What is Our Scientific Question?

Differential Expression/Abundance (DE/DA) Analyses

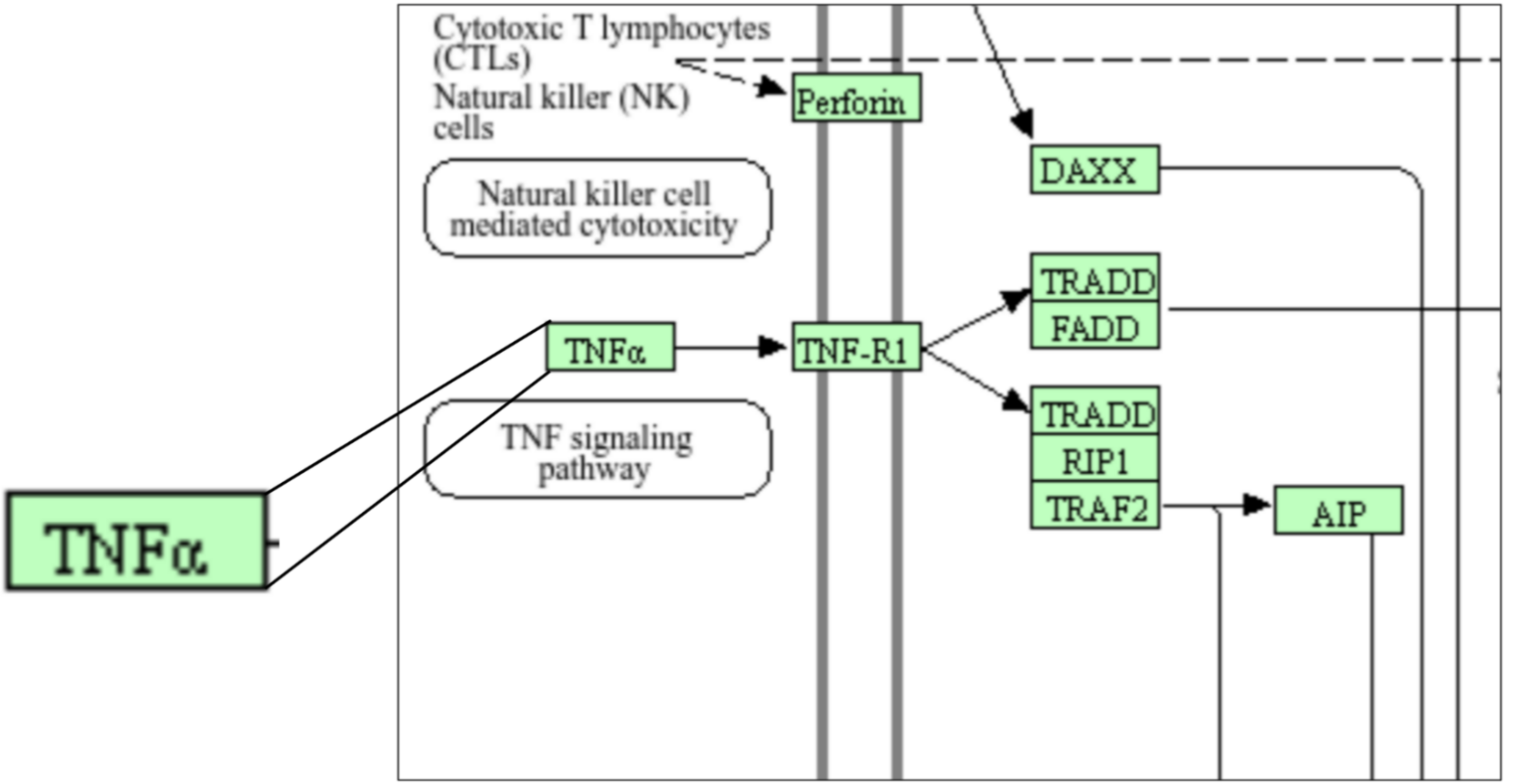
- Is the TNFα gene upregulated in tumor tissue?

- Is the *S. pyogenes* bacetrium killed by an antibiotic?



Methods for DE/DA: ALDEx2, DESeq2, etc.

But what if we have questions of higher level biological processes (e.g., gene pathways)?



What is Our Scientific Question?

Differential Set Analysis (DSA)

- Is the **Apoptosis pathway** upregulated in tumor tissue?
- Are **anaerobic bacetria** killed by an antibiotic?

Methods for DSA: GSEA, CAMERA, etc.

Review of the Problem of Scale

Example Experiment

Conisder 16S rRNA sequencing of fecal samples to measure D taxa in the colons of N individuals with and without IBS.

Review of the Problem of Scale

Example Experiment

Conisder 16S rRNA sequencing of fecal samples to measure D taxa in the colons of N individuals with and without IBS.

Problem of Scale

- We want to analyze absolute abundances (W)
- Absolute abundances are only known if the composition ($W^{\parallel}$) and scale ($W^{\perp}$) are known
- Sequence count data only measure the composition

$$\underbrace{W_{dn}}_{\substack{\text{Absolute Abundance} \\ \text{Taxa d, Patient n} \\ \text{(Unmeasured)}}} = \underbrace{W^{\parallel}_{dn}}_{\substack{\text{Composition} \\ \text{Taxa d, Patient n} \\ \text{(Measured)}}} \times \underbrace{W^{\perp}_{n}}_{\substack{\text{Scale} \\ \text{(e.g., total \# of microbes in} \\ \text{patient n's colon)} \\ \text{(Unmeasured)}}}$$

DE/DA Estimation of LFCs

Example Experiment

Conisder 16S rRNA sequencing of fecal samples to measure D taxa in the colons of N individuals with and without IBS.

The Log Fold Change (LFC) in absolute abundance of taxa d is

$$\theta_d = \underset{n \in \text{IBS}}{\text{mean}}(\log W_{dn}) - \underset{n \in \text{Healthy}}{\text{mean}}(\log W_{dn})$$

DE/DA Estimation of LFCs

The Log Fold Change (LFC) in absolute abundance of taxa d is

$$\theta_d = \underset{n \in \text{IBS}}{\text{mean}}(\log W_{dn}) - \underset{n \in \text{Healthy}}{\text{mean}}(\log W_{dn})$$

Using the relationship $W_{dn} = W_{dn}^{\parallel} \times W_n^{\perp}$:

$$\begin{aligned}
\theta_d &= \underset{n \in \text{IBS}}{\text{mean}}(\log W_{dn}) - \underset{n \in \text{Healthy}}{\text{mean}}(\log W_{dn}) \\
&= \underbrace{\left[\underset{n \in \text{IBS}}{\text{mean}}(\log W_{dn}^{\parallel}) - \underset{n \in \text{Healthy}}{\text{mean}}(\log W_{dn}^{\parallel})\right]}_{\theta_d^{\parallel}} + \underbrace{\left[\underset{n \in \text{IBS}}{\text{mean}}(\log W_n^{\perp}) - \underset{n \in \text{Healthy}}{\text{mean}}(\log W_n^{\perp})\right]}_{\theta^{\perp}} \\
&= \theta_d^{\parallel} + \theta^{\perp}.
\end{aligned}$$

DE/DA Estimation of LFCs

The LFC can be decomposed into compositional and scale terms as

$$\underbrace{\theta_d}_{\text{LFC (Microbe d)}} = \underbrace{\theta_d^{\parallel}}_{\substack{\text{LFC in} \\ \text{composition (Microbe d)}}} + \underbrace{\theta^{\perp}}_{\substack{\text{LFC in} \\ \text{scale}}}$$

DE/DA Estimation of LFCs

The LFC can be decomposed into compositional and scale terms as

$$\underbrace{\theta_d}_{\text{LFC (Microbe d)}} = \underbrace{\theta_d^{\parallel}}_{\substack{\text{LFC in} \\ \text{composition (Microbe d)}}} + \underbrace{\theta^{\perp}}_{\substack{\text{LFC in} \\ \text{scale}}}$$

Using vector notation for all genes/microbes:

$$\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_D \end{bmatrix} = \begin{bmatrix} \theta_1^{\parallel} \\ \vdots \\ \theta_D^{\parallel} \end{bmatrix} + \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \theta^{\perp}$$

$$\theta = \theta^{\parallel} + \mathbf{1}\theta^{\perp}$$

DE/DA Estimation of LFCs

DE/DA Estimation of LFCs

Methods like DESeq2, ALDEx2, etc. are functions of the sequence count data Y:

$$
\begin{aligned}
f(Y) &= \underbrace{\hat{\theta}}_{\text{Estimated LFC}} \\
&= \underbrace{\hat{\theta}^{\parallel}}_{\substack{\text{Estimated LFC} \\ \text{in Composition}}} + \underbrace{\mathbf{1}\hat{\theta}^{\perp}}_{\text{Scale Assumption}}
\end{aligned}
$$

DE/DA Estimation of LFCs

Methods like DESeq2, ALDEx2, etc. are functions of the sequence count data Y:

$$f(Y) = \underbrace{\hat{\theta}}_{\text{Estimated LFC}}$$

$$= \underbrace{\hat{\theta}^{\parallel}}_{\substack{\text{Estimated LFC} \\ \text{in Composition}}} + \underbrace{\mathbf{1}\hat{\theta}^{\perp}}_{\text{Scale Assumption}}$$

Scale Assumption $\hat{\theta}^{\perp}$

- Remember $\theta^{\perp}$ is unmeasured
- Methods like ALDEx2, etc. estimate $\theta^{\perp}$ through **normalization**
- Normalizations imply assumptions about the unmeasured $\theta^{\perp}$

Scale Assumptions and Normalization

Scale assumptions $\hat{\theta}^{\perp}$ are made through noramlization:

- Total Sum Scaling (TSS) Normalization assumes $W_n^{\perp} = 1$, implying

$$\hat{\theta}^{\perp} = 0$$

- Centered Log Ratio assumes $W_n^{\perp} = 1/\text{gm}(W_n^{\parallel})$, implying

$$\hat{\theta}^{\perp} = -\text{mean}(\theta^{\parallel})$$

Errors in Scale Assumptions and LFCs

- Scale Assumption Error:

$$\underbrace{\theta^{\perp}}_{\text{True LFC in Scale}} = \underbrace{\hat{\theta}^{\perp}}_{\text{Scale Assumption}} + \underbrace{\epsilon^{\perp}}_{\text{Scale Assumption Error}}$$

Errors in Scale Assumptions and LFCs

- Scale Assumption Error:

$$\underbrace{\theta^{\perp}}_{\text{True LFC in Scale}} = \underbrace{\hat{\theta}^{\perp}}_{\text{Scale Assumption}} + \underbrace{\epsilon^{\perp}}_{\text{Scale Assumption Error}}$$

- Ignoring compositional error, i.e., assuming:

$$\theta^{\parallel} = \hat{\theta}^{\parallel}$$

Errors in Scale Assumptions and LFCs

- Scale Assumption Error:

$$\underbrace{\theta^{\perp}}_{\text{True LFC in Scale}} = \underbrace{\hat{\theta}^{\perp}}_{\text{Scale Assumption}} + \underbrace{\epsilon^{\perp}}_{\text{Scale Assumption Error}}$$

- Ignoring compositional error, i.e., assuming:

$$\theta^{\parallel} = \hat{\theta}^{\parallel}$$

- The relationship between the LFC and scale error is then

$$\underbrace{\theta}_{\text{True LFC}} = \hat{\theta}^{\parallel} + \mathbf{1}\hat{\theta}^{\perp} + \mathbf{1}\epsilon^{\perp}$$

$$= \underbrace{\hat{\theta}}_{\text{Estimated LFC}} + \underbrace{\mathbf{1}\epsilon^{\perp}}_{\text{Scale Assumption Error}}$$

LFC Sensitivity Analysis

$$\underbrace{\theta}_{\text{True LFC}} = \underbrace{\hat{\theta}}_{\text{Estimated LFC}} + \underbrace{\epsilon^{\perp}}_{\text{Scale Assumption Error}}$$

LFC Sensitivity Analysis

How do the results of a method as a function of error $\epsilon^{\perp}$?

For DE/DA LFC Sensitivity Analysis is Not Interesting

[Insert image from Justin's slides]

But is LFC Sensitivity Analysis for DSA interesting?

DSA Target of Inference

Let $S = \{s_1, \ldots, s_K\}$ where $s_k \in \{1, \ldots, D\}$ be a set of genes or microbes.

The goal of DSA is to infer ϕ_S where

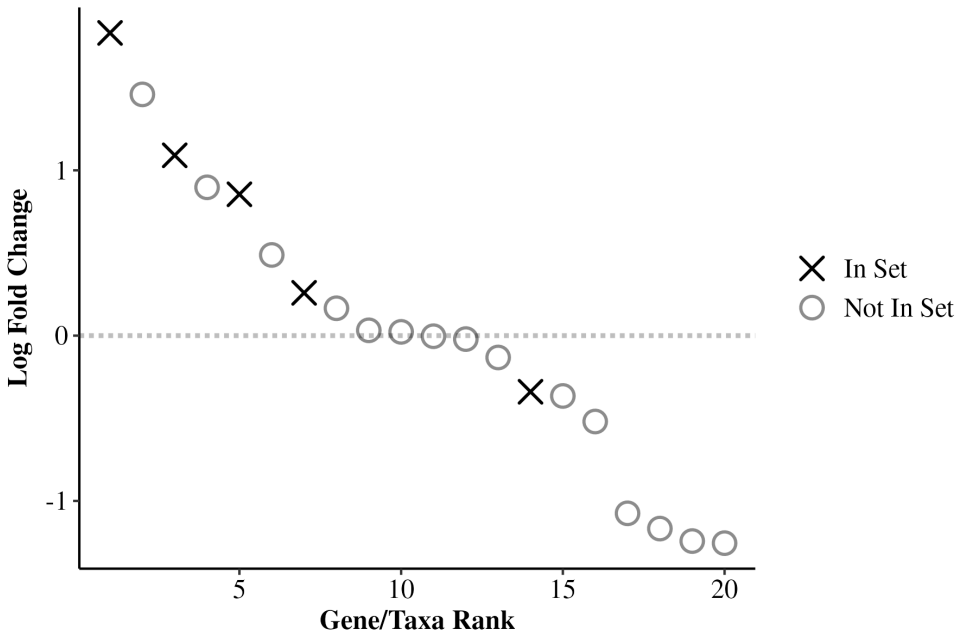
$$\phi_S = \begin{cases} 1 & \text{If } S \text{ is enriched} \\ -1 & \text{If } S \text{ is depleted} \\ 0 & \text{If } S \text{ is neither enriched/depleted} \end{cases}$$

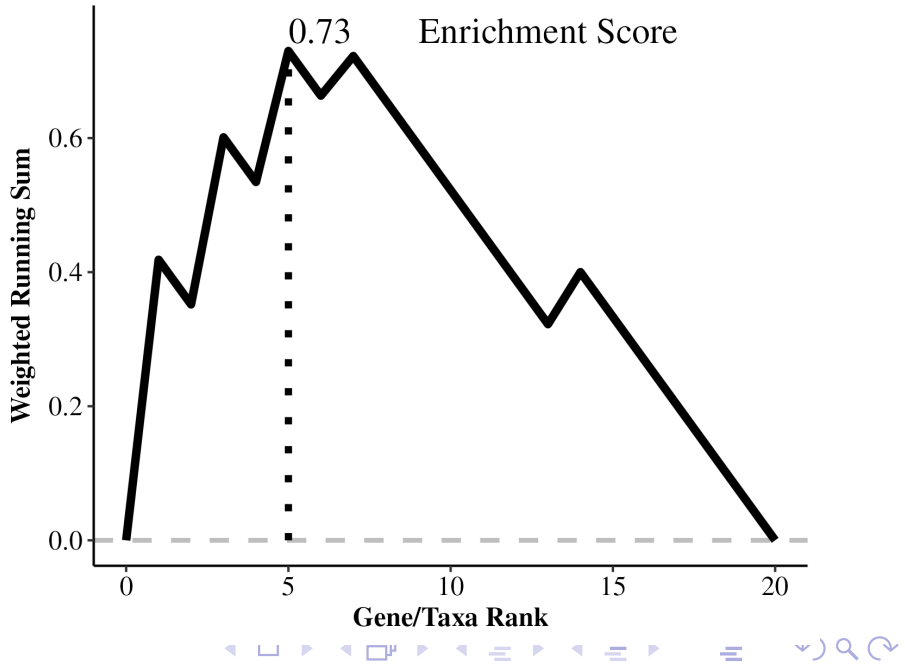
The Gene Set Enrichment Analysis (GSEA) method is a popular tool for estimating ϕ_S

GSEA with Gene Label Permutations

GSEA test statistic calculation:

1. Estimate LFCs $\hat{\theta} = f(Y)$ and rank in descending order
2. Calculate a **weighted** running sum iterating down ranked LFCs
   - Genes in S increase running sum by an amount proportional to the LFC
   - Genes not in S decrease running sum by constant amount
3. Enrichment Score is supremum of running sum





GSEA with Gene Label Permutations

Permutation Statistical Test:

1. Calculate Enrichment Score (ES) for Gene Set S
2. Permute gene set labels Q times, compute null ESs
3. p-value: $\#(\text{ES} > \text{Permuted ES})/Q$

Gene Name	RB1	IL2	APP	AR	HTT	IL6	B2M	RGN	CAT
Set S	X		X	X	X				
Perm 1		X	X				X	X	
Perm 2	X					X	X	X	
...									
Perm Q	X			X			X		X

GSEA and LFC Sensitivity Analysis

The GSEA method can be represented mathematically as

$$\phi_S = u(\theta)$$

GSEA and LFC Sensitivity Analysis

The GSEA method can be represented mathematically as

$$\phi_S = u(\theta)$$

Plugging in

$$\underbrace{\theta}_{\text{True LFC}} = \underbrace{\hat{\theta}}_{\text{Estimated LFC}} + \underbrace{\mathbf{1}\epsilon^{\perp}}_{\text{Scale Assumption Error}}$$

GSEA and LFC Sensitivity Analysis

The GSEA method can be represented mathematically as

$$\phi_S = u(\theta)$$

Plugging in

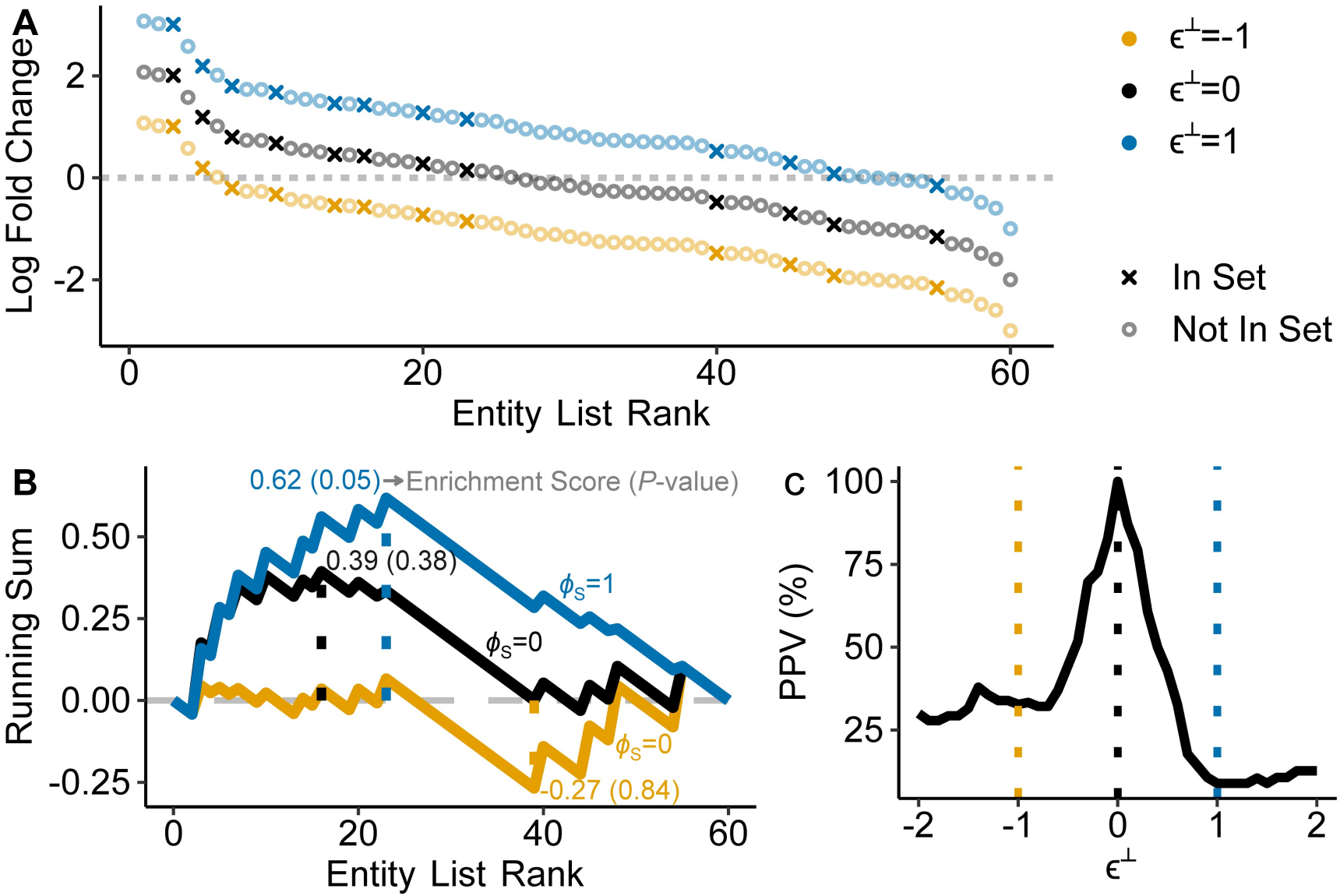
$$\underbrace{\theta}_{\text{True LFC}} = \underbrace{\hat{\theta}}_{\text{Estimated LFC}} + \underbrace{\mathbf{1}\epsilon^{\perp}}_{\text{Scale Assumption Error}}$$

We get

$$\phi_S = u(\hat{\theta} + \mathbf{1}\epsilon^{\perp})$$

How does ϕ_S change with scale assumption error $\epsilon^{\perp}$?

GSEA LFC Sensitivity Analysis Simulation Results



GSEA LFC Sensitivity Analysis Real Data Results

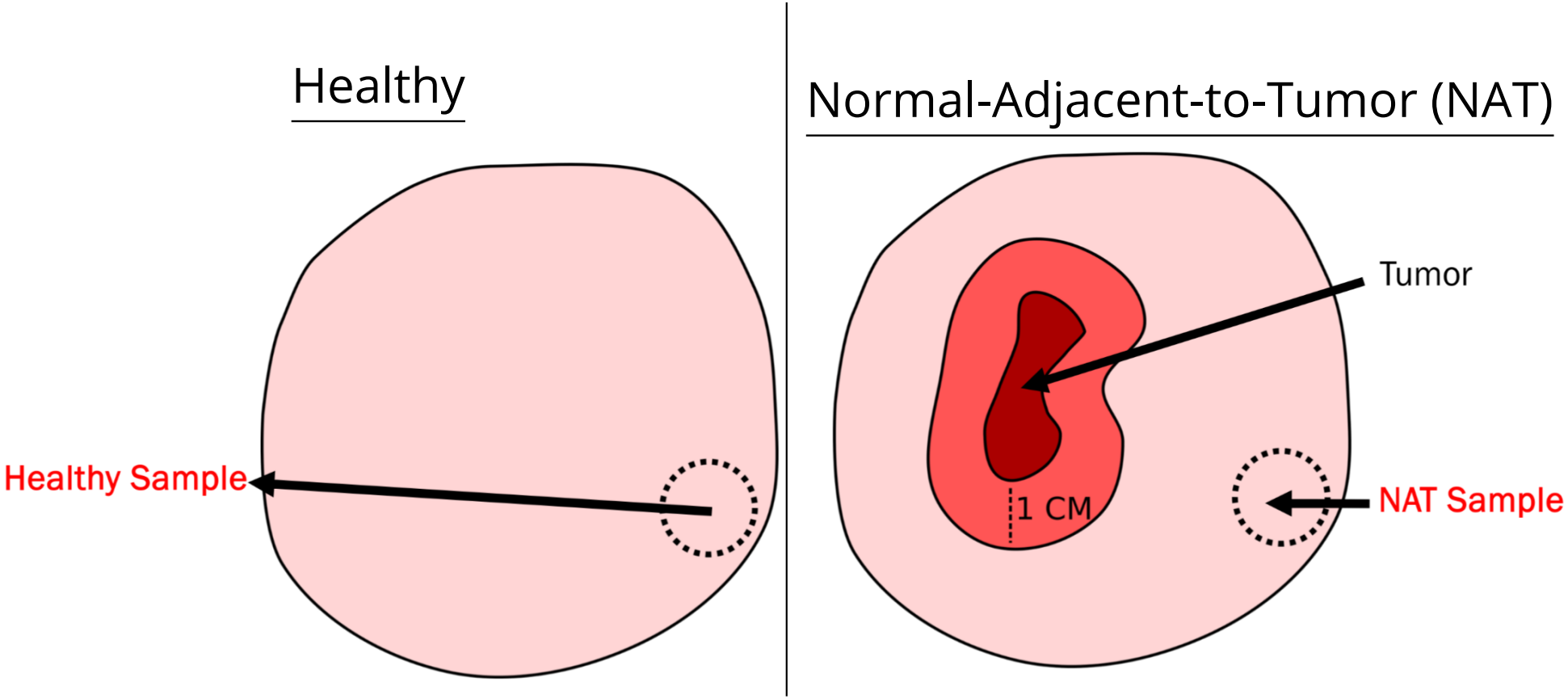
Article | Open access | Published: 20 October 2017

Comprehensive analysis of normal adjacent to tumor transcriptomes

Dvir Aran, Roman Camarda, Justin Odegaard, Hyojung Paik, Boris Oskotsky, Gregor Krings, Andrei Goga, Marina Sirota & Atul J. Butte

Nature Communications **8**, Article number: 1077 (2017) | Cite this article

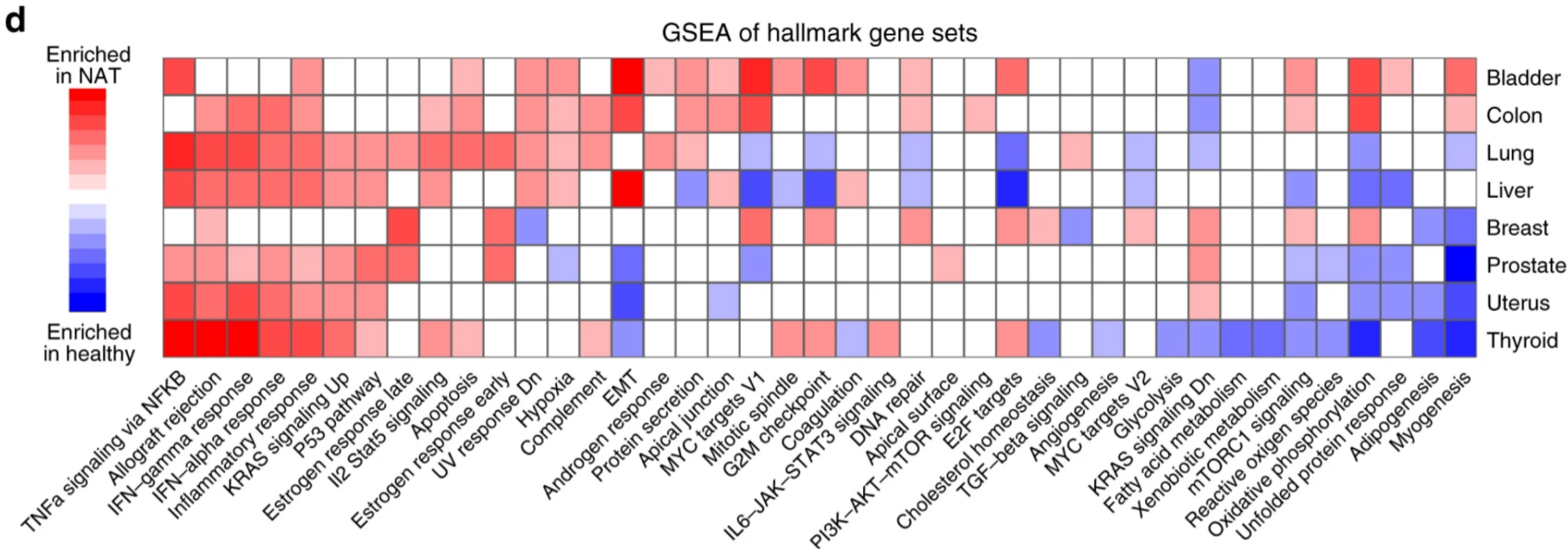
Healthy vs. NAT Tissue Research Question



Research Question

Is NAT tissue an appropriate proxy for healthy tissue in cancer research?

Aran et al. GSEA Results



NAT vs. Healthy Thyroid Tissue

Results using the fast GSEA (fgsea) package:

```
## Run vanilla fgsea
simple_fgsea_res <- fgsea(stats=lfcs, pathways=gmt.file)
```

Myogenesis

adj. p-val	**9e-9**
NES	-2.1

Inflamatory Response

adj. p-val	**3e-3**
NES	1.5

NAT vs. Healthy Thyroid Tissue

Results using the fast GSEA (fgsea) package:

```
## Run vanilla fgsea
simple_fgsea_res <- fgsea(stats=lfcs, pathways=gmt.file)
```

Myogenesis

adj. p-val	**9e-9**
NES	-2.1

Inflamatory Response

adj. p-val	**3e-3**
NES	1.5

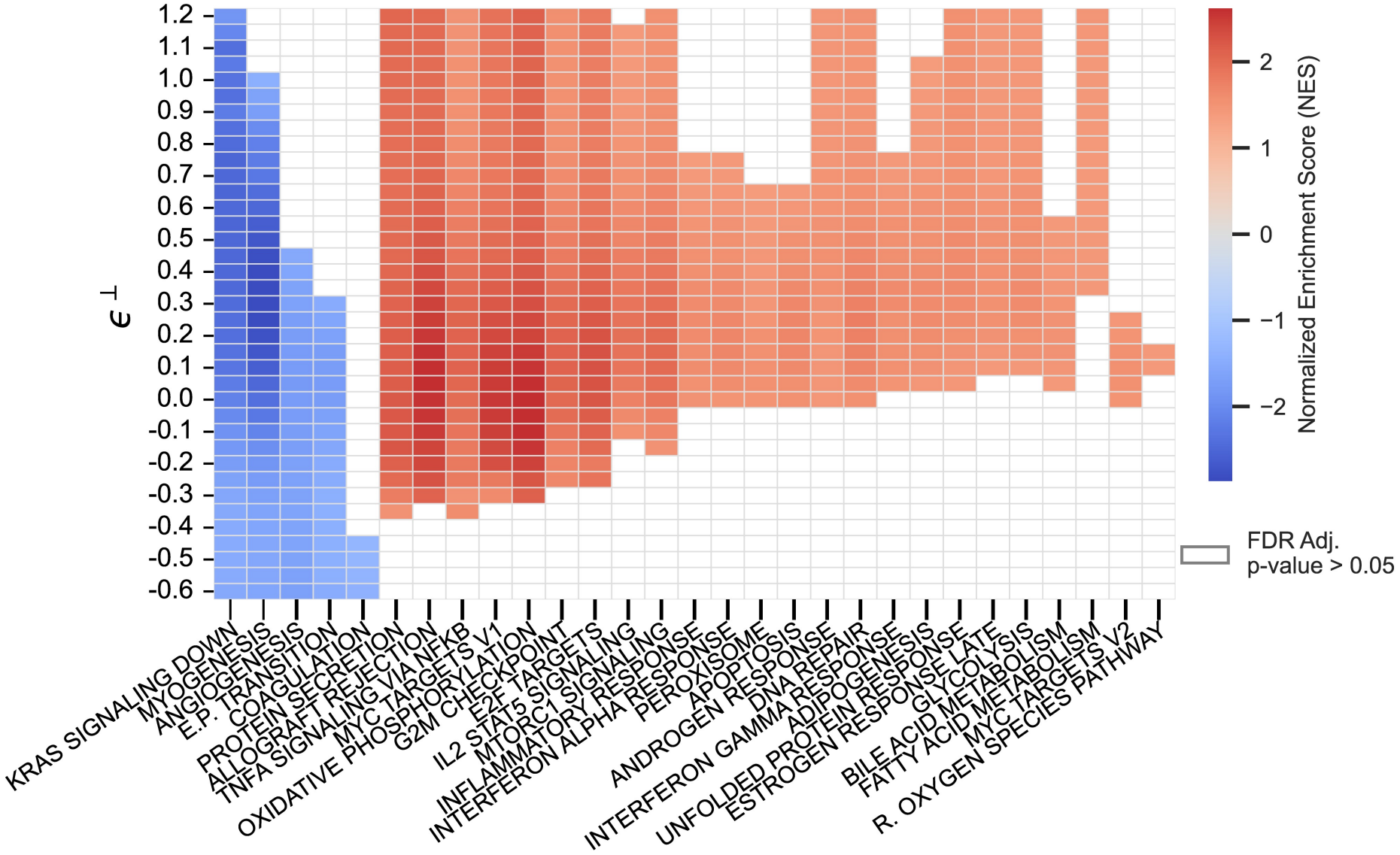
Results using fgsea **LFC Sensitivity Analysis** wrapper:

```
## Run LFC Sensitivity Analysis FGSEA
lfc_fgsea_res <- fgsea.error(lfcs, gmt.file,
                             epsilon=c(-0.4, 0, 0.4))
```

$\epsilon^{\perp}$	-0.4	0	0.4
adj. p-val	**2e-6**	**9e-9**	**2e-9**
NES	-1.4	-2.1	-2.5

$\epsilon^{\perp}$	-0.4	0	0.4
adj. p-val	1	**3e-3**	**3e-3**
NES	-0.8	1.5	1.6

LFC Sensitivity Analysis: Thyroid NAT vs Healthy



Addressing the Problem of Inter-gene/microbe Correlations

The Problem with Inter-gene/microbe correlations

Are Random Gene Sets a Reasonable Method for Generating a Null Distribution?

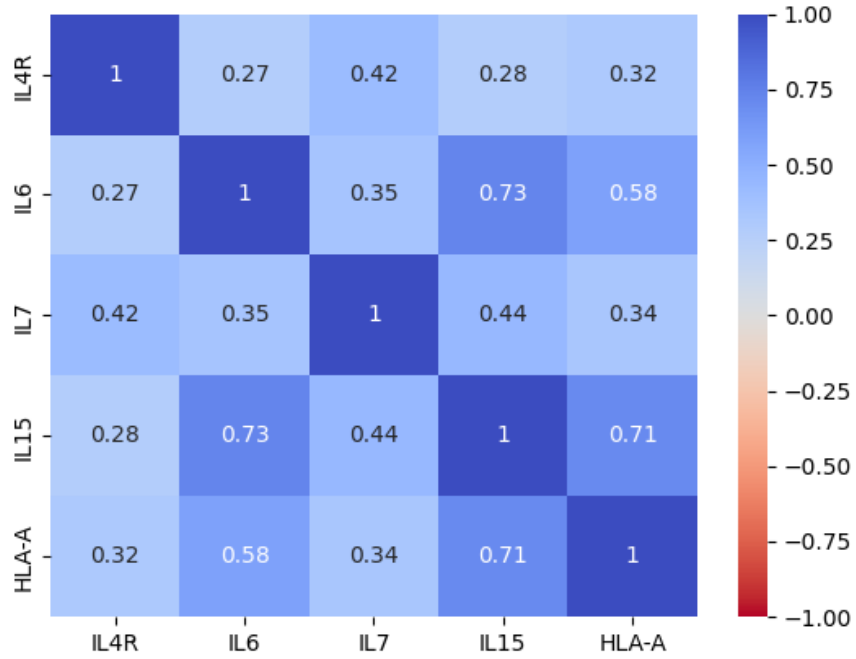
Gene Name	RB1	IL2	APP	AR	HTT	IL6	B2M	RGN	CAT
Set S	X		X	X	X				
Perm 1		X	X				X	X	
Perm 2	X					X	X	X	
...									
Perm Q	X			X			X		X

The Problem with Inter-gene/microbe Correlations

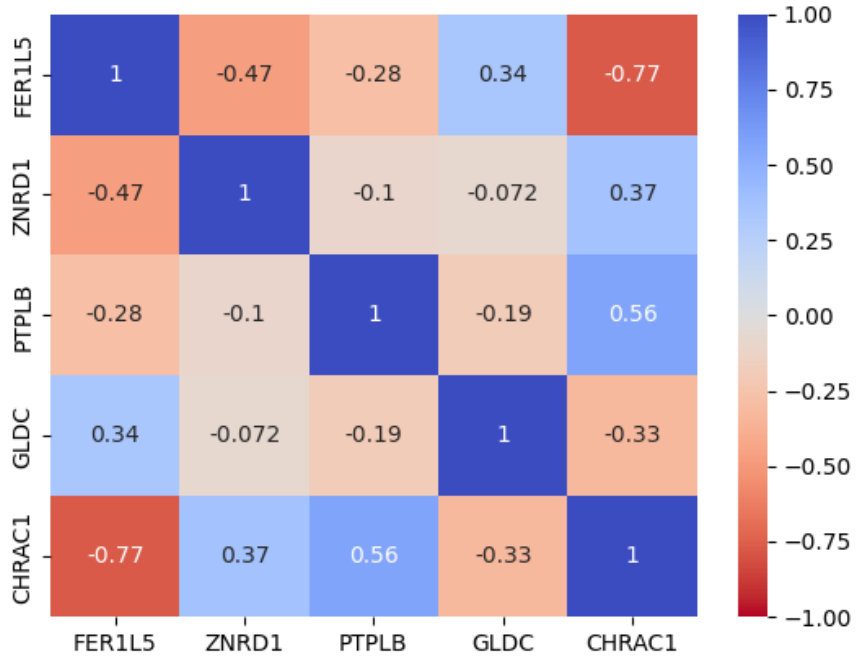
Core Issue

Biologically related genes tend to be (positively) correlated, random genes tend to not be correlated.

Correlation Matrix: 5 Biologically Related Genes



Correlation Matrix: 5 Random Genes



Correlations Inflate Flase Positives

Wu et al. considered a simulation with

- 10000 genes, **all with an LFC of 0**
- One gene set of 100 genes **with an average correlation of 0.05**

Correlations Inflate Flase Positives

Wu et al. considered a simulation with

- 10000 genes, **all with an LFC of 0**
- One gene set of 100 genes **with an average correlation of 0.05**

Expected Results:

- No pathways/genes are enriched meaning the false positive rate should be around 5%

Correlations Inflate Flase Positives

Wu et al. considered a simulation with

- 10000 genes, **all with an LFC of 0**
- One gene set of 100 genes **with an average correlation of 0.05**

Expected Results:

- No pathways/genes are enriched meaning the false positive rate should be around 5%

Actual Results:

- GSEA w/ gene label permutations had a false positive rate of up to 40%!

Two Methods for Addressing Inter-gene/microbe Correlations

Two Methods for Addressing Inter-gene/microbe Correlations

1. GSEA with **sample label permutations**

GSEA with gene label permutations

$$\hat{\theta} \implies \text{Permute Set } S \implies \text{Enrichment Score}$$

GSEA with sample label permutations

$$\log \hat{W} \implies \text{Permute Sample Labels} \implies \hat{\theta}^* \implies \text{Enrichment Score}$$

Two Methods for Addressing Inter-gene/microbe Correlations

2. CAMERA: Two-sample T-test **adjusted for correlation in** S

Let:

- $\theta_{\in S}, \theta_{\notin S}$ be LFCs of genes in/not-in set S
- ρ_S be the average correlation in gene expression for set S

$$T = \frac{\hat{\theta}_{\in S} - \hat{\theta}_{\notin S}}{s_p \sqrt{\frac{1+(n_1-1)\hat{\rho}_S}{n_1} + \frac{1}{n_2}}}$$

where

$$\hat{\rho}_S = f(\log \hat{W})$$

Scale Sensitivity Analysis

- Review: GSEA with Gene Label Permutations

  A function of LFCs

$$\phi_S = u(\theta)$$

  LFC Sensitivity Analysis:

$$\phi_S = u(\hat{\theta} + \mathbf{1}\epsilon^{\perp})$$

Scale Sensitivity Analysis

- GSEA with Sample Label Permutations / CAMERA
  A function of log absolute abundances:

$$\phi_S = u(\log W)$$

Scale Sensitivity Analysis

- GSEA with Sample Label Permutations / CAMERA
  A function of log absolute abundances:

$$\phi_S = u(\log W)$$

$$\underbrace{\log W}_{\substack{D \times N \text{ Matrix} \\ \text{Abs Abundances}}} = \underbrace{\log W^{\|}}_{\substack{D \times N \text{ Matrix} \\ \text{Compositions}}} + \underbrace{\log W^{\perp}}_{\substack{N \text{ length vector} \\ \text{of scale}}}$$

Scale Sensitivity Analysis

- GSEA with Sample Label Permutations / CAMERA
  A function of log absolute abundances:

$$\phi_S = u(\log W)$$

$$\underbrace{\log W}_{\substack{D \times N \text{ Matrix} \\ \text{Abs Abundances}}} = \underbrace{\log W^{\parallel}}_{\substack{D \times N \text{ Matrix} \\ \text{Compositions}}} + \underbrace{\log W^{\perp}}_{\substack{N \text{ length vector} \\ \text{of scale}}}$$

$$\underbrace{\begin{bmatrix} W_1^{\perp} \\ \vdots \\ W_N^{\perp} \end{bmatrix}}_{\text{True Scale}} = \underbrace{\begin{bmatrix} \hat{W}_1^{\perp} \\ \vdots \\ \hat{W}_N^{\perp} \end{bmatrix}}_{\text{Scale Assumption}} + \underbrace{\begin{bmatrix} \epsilon_1^{\perp} \\ \vdots \\ \epsilon_N^{\perp} \end{bmatrix}}_{\substack{\text{Scale Assumption} \\ \text{Error}}}$$

Scale Sensitivity Analysis

- LFC Sensitivity Analysis

$$\phi_S = u(\theta)$$
$$\phi_S = u(\hat{\theta} + \epsilon^{\perp})$$

- Scale Sensitivity Analysis

$$\phi_S = u(\log W)$$
$$\phi_S = u(\log \hat{W} + [\epsilon_1^{\perp}, \ldots, \epsilon_N^{\perp}]^T)$$

Two Scale Sensitivity Analyses

Scale Sensitivity Analysis (GSEA w/ sample label permutations and CAMERA)

$$\phi_S = u(\log W)$$
$$\phi_S = u(\log \hat{W} + \left[\epsilon_1^{\perp}, \ldots, \epsilon_n^{\perp}\right])$$

Here we consider 4500 gene sets in Healthy vs. Tumor breast tissue:

Constant Error

$$\epsilon_i^{\perp} = \begin{cases} 0 & \text{if } i \in \text{Healthy} \\ \delta^{\perp} & \text{if } i \in \text{Tumor} \end{cases}$$

Sample-specific Error

$$\epsilon_i^{\perp} \in [-\gamma^{\perp}, \gamma^{\perp}]$$

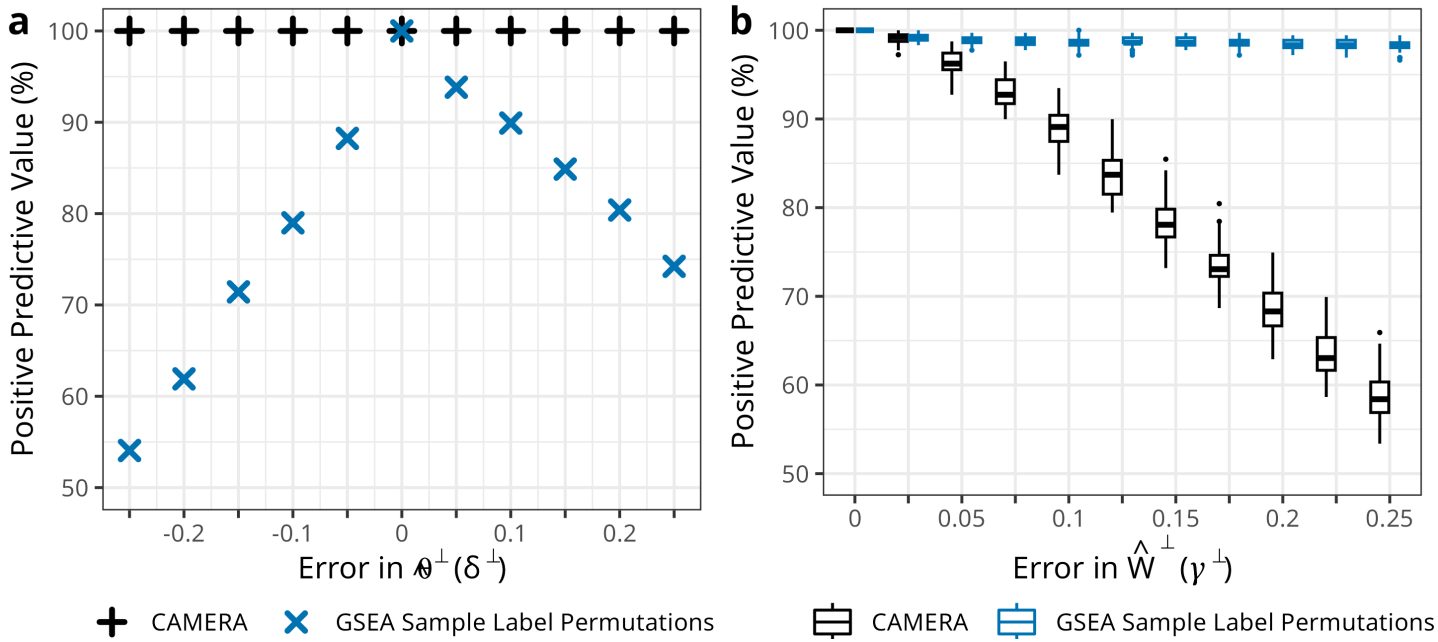
Two Scale Sensitivity Analyses

Constant Error

$$\epsilon_i^{\perp} = \begin{cases} 0 & \text{if } i \in \text{Healthy} \\ \delta^{\perp} & \text{if } i \in \text{Tumor} \end{cases}$$

Sample-specific Error

$$\epsilon_i^{\perp} \in [-\gamma^{\perp}, \gamma^{\perp}]$$



References

1. Gatti, et al. Heading down the wrong pathway: on the influence of correlation within gene sets. BMC Genomics. 2010 Oct 18;11:574. doi: 10.1186/1471-2164-11-574.
2. McGovern, et al. Addressing erroneous scale assumptions in microbe and gene set enrichment analysis. PLoS Comput Biol. 2023 Nov 20;19(11):e1011659. doi: 10.1371/journal.pcbi.1011659.
3. Subramanian, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005 Oct 25;102(43):15545-50. doi: 10.1073/pnas.0506580102. Epub 2005 Sep 30.
4. Wu, et al. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012 Sep 1;40(17):e133. doi: 10.1093/nar/gks461. Epub 2012 May 25.
5. Aran, et al. Comprehensive analysis of normal adjacent to tumor transcriptomes. Nat Commun. 2017 Oct 20;8(1):1077. doi: 10.1038/s41467-017-01027-z. PMID: 29057876; PMCID: PMC5651823.