Scale Sensitivity as a Tool for Differential Set Analyses

Kyle C. McGovern

Silverman Lab
The Pennsylvania State University
Program in Bioinformatics and Genomics

Differential Set Analysis

Differential (Expression/Abundance) Analysis

Which Genes (or Taxa) are changing in amount between conditions?

Differential Set Analysis

aka Gene (or Microbe) Set Enrichment Analysis

Which pathways (sets of genes) or sets of microbes are changing in amount between conditions?

Review and Problem Statement

Scale Reliant Inference

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

Scale Reliant Inference

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

We measure sequence count data Y which provides information about $W^{\parallel}$ but not $W^{\perp}$.

Scale Reliant Inference

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

We measure sequence count data Y which provides information about $W^{\parallel}$ but not $W^{\perp}$.

$$\theta = f(W^{\parallel}, W^{\perp})$$

Scale Reliant Inference

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

We measure sequence count data Y which provides information about $W^{\parallel}$ but not $W^{\perp}$.

$$\theta = f(W^{\parallel}, W^{\perp})$$

Example (Differential Abundance / Expression Analysis)

Using 16S rRNA-seq data, we want to identify which taxa change in abundance between cases and controls (LFC estimation).

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

Scale-Composition of LFC Estimand

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

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

$$\theta_d = \underbrace{\left[\underset{n \in \text{case}}{\text{mean}}(\log W_{dn}^{\parallel}) - \underset{n \in \text{control}}{\text{mean}}(\log W_{dn}^{\parallel})\right]}_{\theta_d^{\parallel}} + \underbrace{\left[\underset{n \in \text{case}}{\text{mean}}(\log W_n^{\perp}) - \underset{n \in \text{control}}{\text{mean}}(\log W_n^{\perp})\right]}_{\theta^{\perp}}$$

$$= \theta_d^{\parallel} + \theta^{\perp}.$$

Scale-Composition of LFC Estimand

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

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

$$\theta_d = \underbrace{\left[\underset{n \in \text{case}}{\text{mean}}(\log W_{dn}^{\parallel}) - \underset{n \in \text{control}}{\text{mean}}(\log W_{dn}^{\parallel})\right]}_{\theta_d^{\parallel}} + \underbrace{\left[\underset{n \in \text{case}}{\text{mean}}(\log W_n^{\perp}) - \underset{n \in \text{control}}{\text{mean}}(\log W_n^{\perp})\right]}_{\theta^{\perp}}$$

$$= \theta_d^{\parallel} + \theta^{\perp}.$$

- $\theta_d^{\parallel}$ is the LFC in composition of the d-th taxon (How do the proportions change between conditions?)
- $\theta^{\perp}$ is the LFC in scales (How do the totals change between conditions?)

LFC Vector Notation

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

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

LFC Sensitivity Analysis

Scale Assumptions

Methods like DESeq2 or ALDEx2 estimate LFCs ($\hat{\theta}$).

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

Scale Assumptions

Methods like DESeq2 or ALDEx2 estimate LFCs ($\hat{\theta}$).

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

Because the data provides information about $W^{\parallel}$, ignore error in $\hat{\theta}^{\parallel}$ (assuming $\hat{\theta}^{\parallel} = \theta^{\parallel}$).

Scale Assumptions

Methods like DESeq2 or ALDEx2 estimate LFCs ($\hat{\theta}$).

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

Because the data provides information about $W^{\parallel}$, ignore error in $\hat{\theta}^{\parallel}$ (assuming $\hat{\theta}^{\parallel} = \theta^{\parallel}$).

Since the data lacks information about $\theta^{\perp}$ we call $\hat{\theta}^{\perp}$ a **scale assumption**.

Scale Assumptions

Methods like DESeq2 or ALDEx2 estimate LFCs ($\hat{\theta}$).

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

Because the data provides information about $W^{\parallel}$, ignore error in $\hat{\theta}^{\parallel}$ (assuming $\hat{\theta}^{\parallel} = \theta^{\parallel}$).

Since the data lacks information about $\theta^{\perp}$ we call $\hat{\theta}^{\perp}$ a **scale assumption**.

Assumptions are subject to error $\theta^{\perp} = \hat{\theta}^{\perp} + \epsilon^{\perp}$.

Scale Assumptions

Methods like DESeq2 or ALDEx2 estimate LFCs ($\hat{\theta}$).

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

Because the data provides information about $W^{\parallel}$, ignore error in $\hat{\theta}^{\parallel}$ (assuming $\hat{\theta}^{\parallel} = \theta^{\parallel}$).

Since the data lacks information about $\theta^{\perp}$ we call $\hat{\theta}^{\perp}$ a **scale assumption**.

Assumptions are subject to error $\theta^{\perp} = \hat{\theta}^{\perp} + \epsilon^{\perp}$.

LFC Sensitivity Analysis

$$\theta = \hat{\theta} + \mathbf{1}\epsilon^{\perp}$$

Review Scale Assumptions

- CLR Normalization ($W_n^{\perp} = 1/GM(W^{\parallel} \cdot n)$) leads to

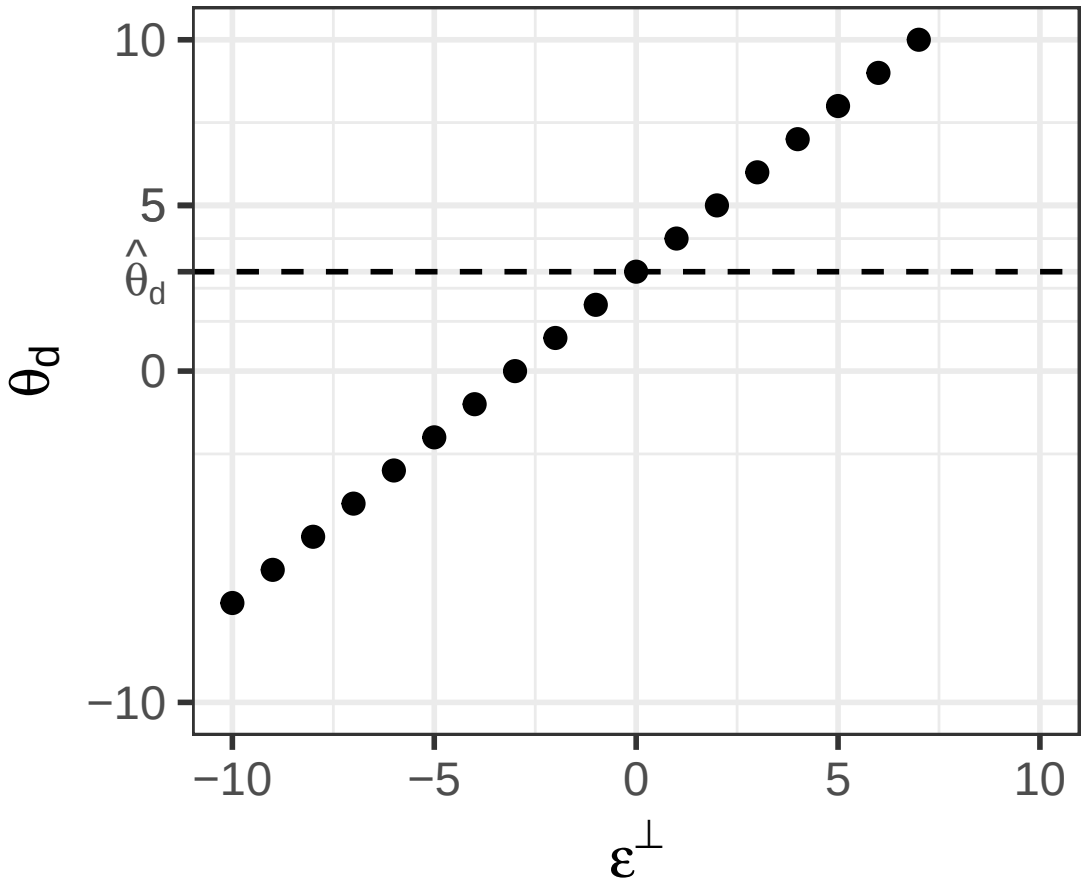
$$\hat{\theta}^{\perp}_{\mathsf{CLR}} = \underset{case}{\text{mean}} \log(1/GM(W^{\parallel}_{\cdot n})) - \underset{control}{\text{mean}} \log(1/GM(W^{\parallel}_{\cdot n}))$$

- TSS Normalization (dividing by sequencing depth) implies

$$\hat{\theta}^{\perp}_{\mathsf{TSS}} = 0$$

LFC Sensitivity Analysis is Boring on Its Own

$$\theta_d = \hat{\theta}_d + \epsilon^{\perp}$$



Aside: Interpretation of $\epsilon^{\perp}$

At a given level of error $\epsilon^{\perp}$, the true ratio of the average scale in case vs. control is $e^{\epsilon^{\perp}}$ higher than assumed.

Example

If $\epsilon^{\perp} = 1$ then the true average scale (ratio) is $e^1 \approx 2.7$ times higher than assumed.

Differential Set Analysis (DSA)

Target Estimand for DSA

Let S be a set of genes or microbes e.g., $S = \{\text{Taxa 1}, \text{Taxa 3}, \text{Taxa 9}\}$

The goal of DSA is to infer ϕ_S where

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

Target Estimand for DSA

Let S be a set of genes or microbes e.g., $S = \{\text{Taxa 1}, \text{Taxa 3}, \text{Taxa 9}\}$

The goal of DSA is to infer ϕ_S where

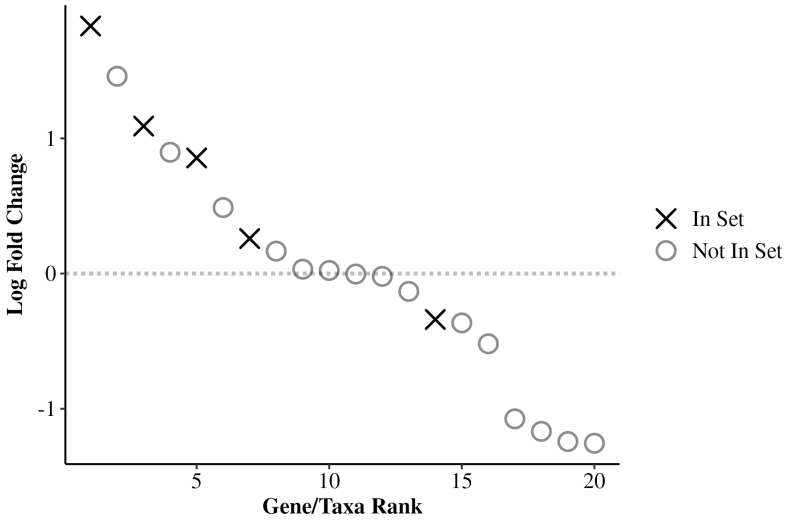
$$\phi_S = \begin{cases} 1 & \text{If } S \text{ is enriched} \\ -1 & \text{If } S \text{ is depleted} \\ 0 & \text{If } S \text{ is neither enriched/depleted} \end{cases}$$

- Many researchers define ϕ_S as a function of θ (LFCs) which in turn is a function of W:

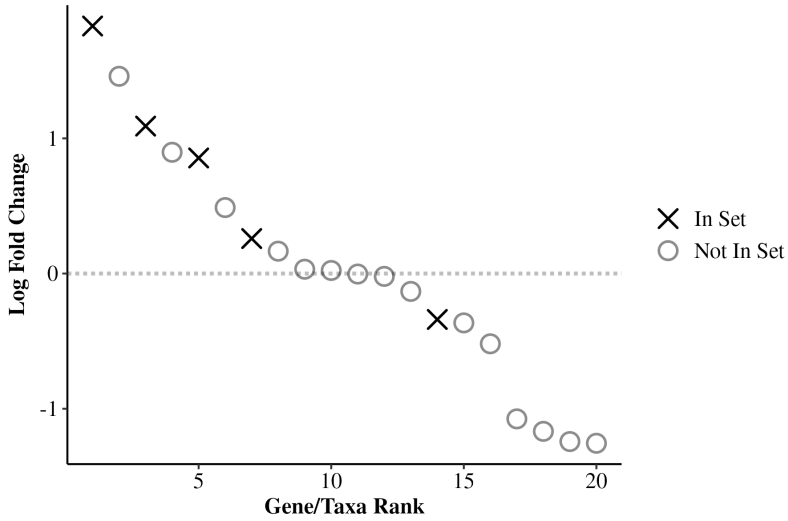
$$\phi_S = u(\theta).$$

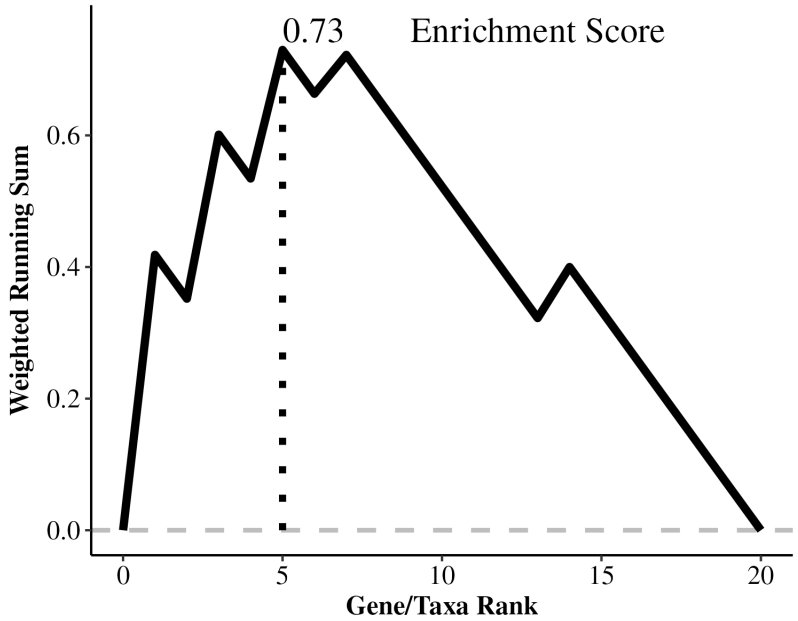
- Particularly common is defining u based on the GSEA algorithm.

Gene (and Microbe) Set Enrichment Analysis (GSEA)

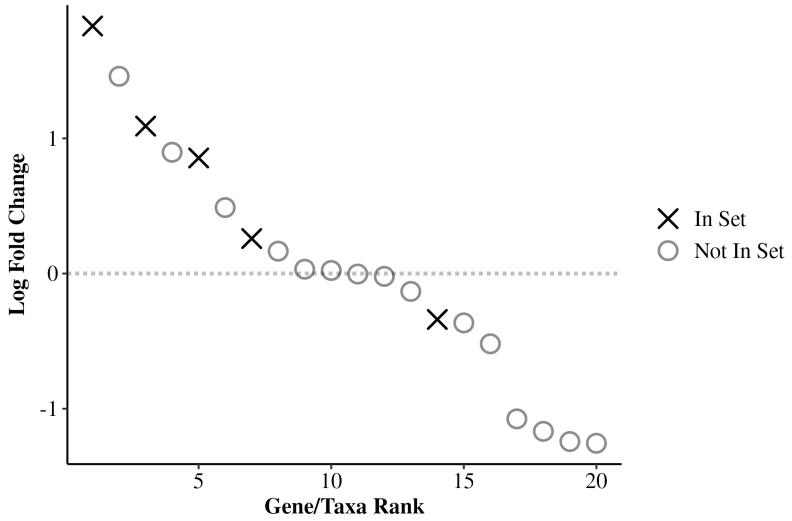


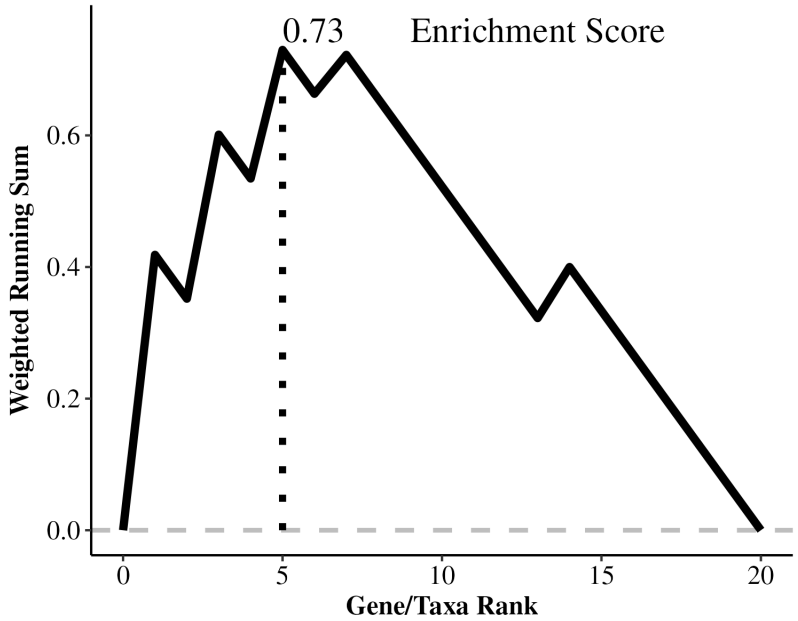
Gene (and Microbe) Set Enrichment Analysis (GSEA)





Gene (and Microbe) Set Enrichment Analysis (GSEA)





Statistical significance (p-values) estimated by permutation test:

Gene Label permutations

- Our focus to start
- Computationally simpler
- Requires less data
- Ignores for inter-gene (or inter-taxa) correlations

Sample Label permutations

- Discussed later

LFC Sensitivity Analysis for GSEA

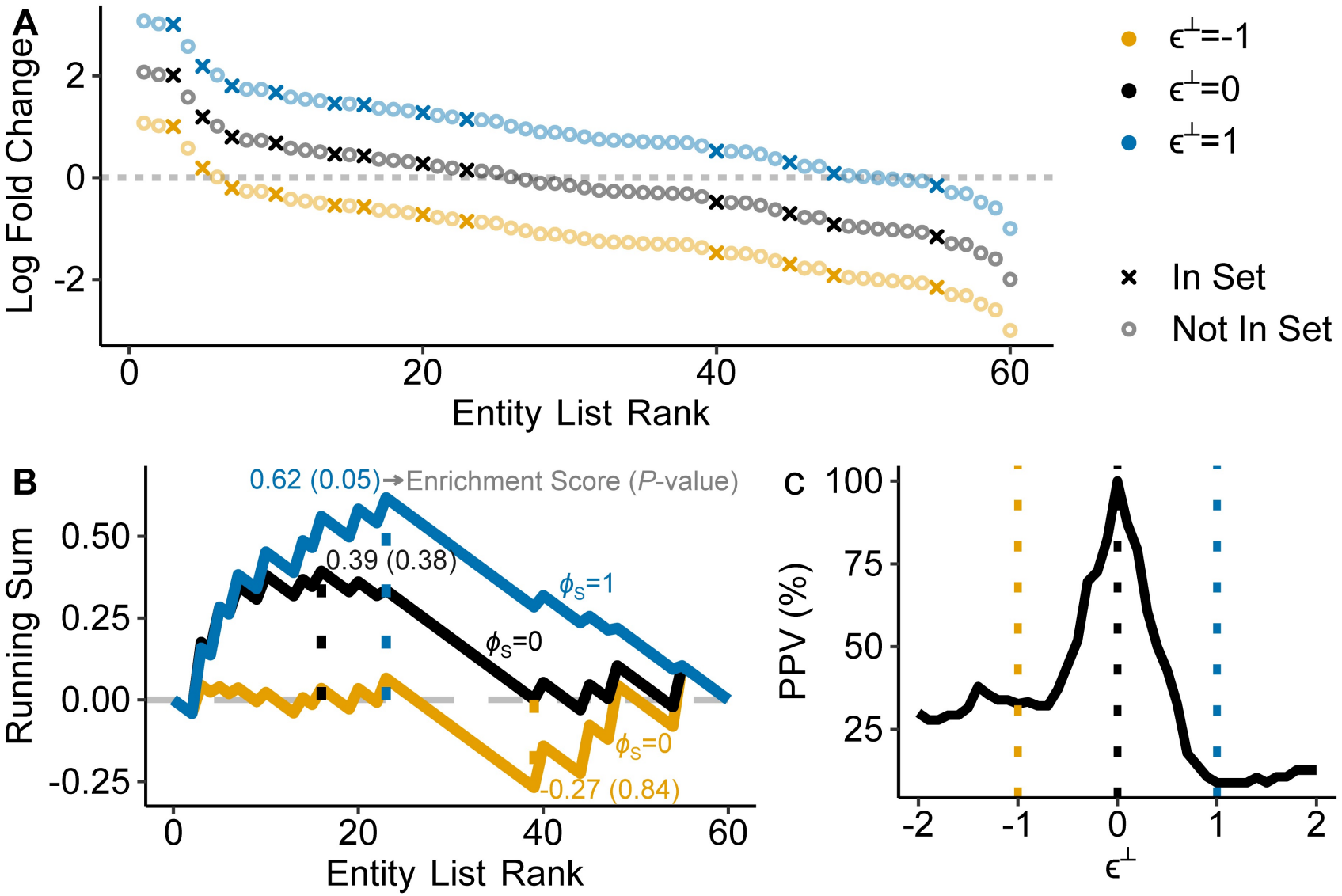
$$\phi_S = u(\theta)$$

LFC Sensitivity Analysis for GSEA

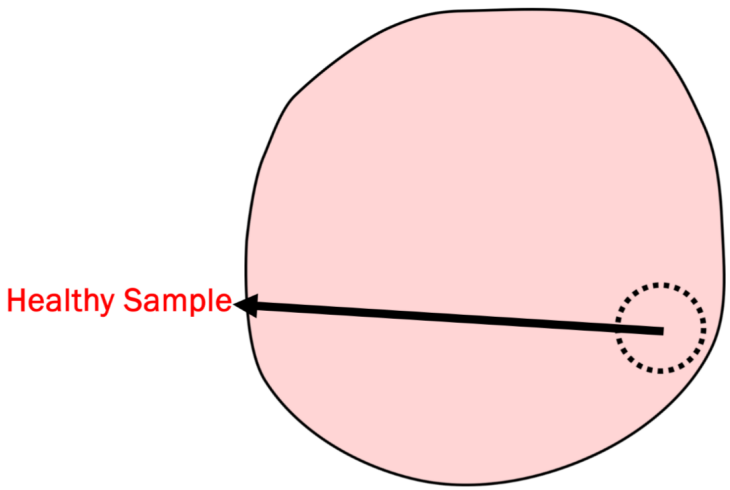
$$\phi_S = u(\theta)$$

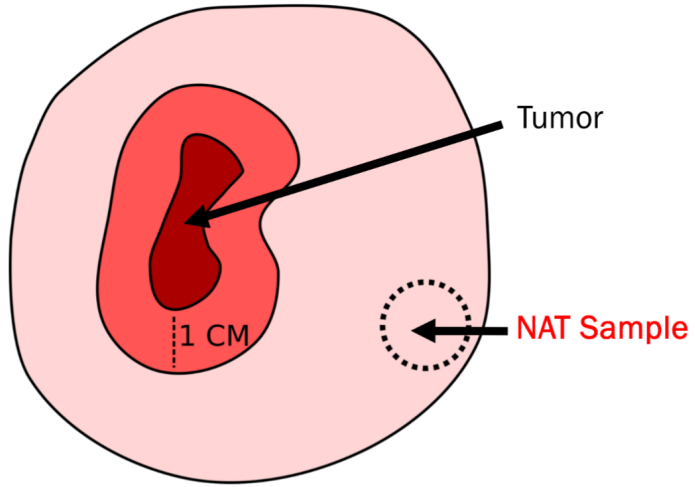
$$\phi_S = u(\hat{\theta} + \mathbf{1}\epsilon^{\perp})$$

Simulated Data Example



Real Data Example

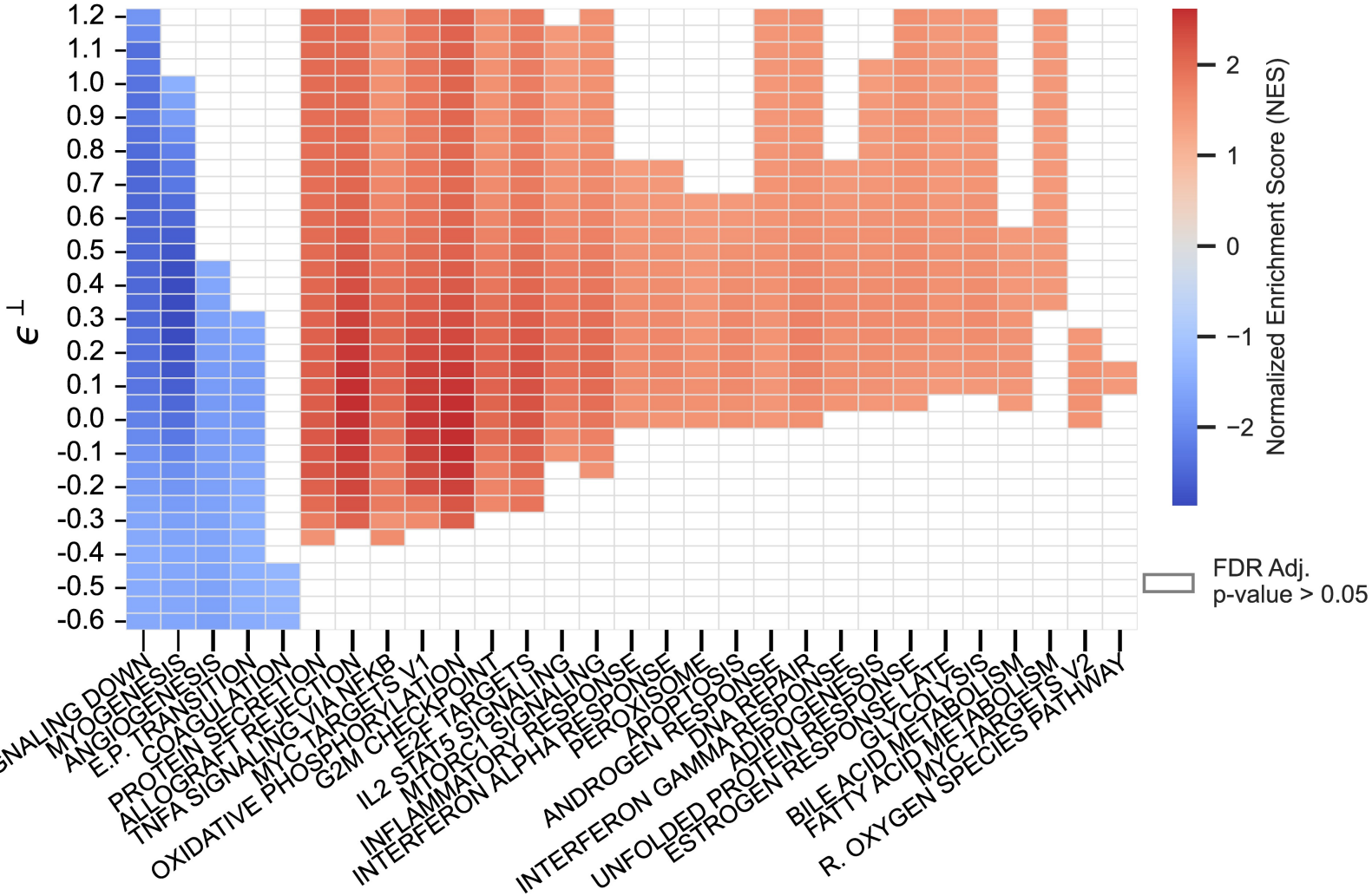




Motivation

- Normal Adjacent to Tumor (NAT) is often used as surrogate for healthy control.
- But is it really the same as healthy tissue?
- We reanalyze RNA-seq data from healthy verus NAT thyroid tissue.

Real Data Example



McGovern, Nixon & and Silverman (2023) Addressing Erroneous Scale Assumptions in Microbe and Gene Set Enrichment Analysis, *PLoS Computational Biology*

Software: Just 1 Exrtra Argument

```
head(lfcs, 3)
# 1/2-SBSRNA4        A1BG    A1BG-AS1
#   1.4861493  -0.2447069  -0.6440828
```

```
head(pathways, 1)
# $HALLMARK_TNFA_SIGNALING_VIA_NFKB
#  [1] "JUNB"     "CXCL2"    "ATF3"
```

```
# typical gsea using fgsea
simple_fgsea_res <-       fgsea(lfcs, pathways)
# LFC Sensitivity Analysis fgsea wrapper
lfc_fgsea_res    <- fgsea.error(lfcs, pathways, epsilon=c(-0.4, 0, 0.4))
```

Software: Just 1 Exrtra Argument

```
head(lfcs, 3)
# 1/2-SBSRNA4        A1BG    A1BG-AS1
#   1.4861493  -0.2447069  -0.6440828
```

```
head(pathways, 1)
# $HALLMARK_TNFA_SIGNALING_VIA_NFKB
#  [1] "JUNB"     "CXCL2"    "ATF3"
```

```
# typical gsea using fgsea
simple_fgsea_res <-       fgsea(lfcs, pathways)
# LFC Sensitivity Analysis fgsea wrapper
lfc_fgsea_res    <- fgsea.error(lfcs, pathways, epsilon=c(-0.4, 0, 0.4))
```

epsilon <dbl>	pathway <chr>	pval <dbl>	padj <dbl>
-0.4	HALLMARK_INFLAMMATORY_RESPONSE	0.9828034393	1.000000000
0.0	HALLMARK_INFLAMMATORY_RESPONSE	0.0011709727	0.004503741
0.4	HALLMARK_INFLAMMATORY_RESPONSE	0.0006580979	0.002193660

Inter-Entity Correlations

Problem with Inter-Entity Correlations

- Genes (and microbes) tend to be correlated.
- Entity Label permutation test ignores these correlations.
- This leads to elevated rates of false positives (Wu et al. 2012, *Nucleic Acids Res*)

GSEA with Sample Label Permutations (GSEA-S)

- GSEA but we permute which samples are in case versus control.
- Permutation based null models retains inter-entity correlations (addressing false positives)

GSEA with Sample Label Permutations (GSEA-S)

- GSEA but we permute which samples are in case versus control.
- Permutation based null models retains inter-entity correlations (addressing false positives)

Can't run LFC Sensitivity Analysis for GSEA-S as described before because $\hat{\theta}$ also changes when sample labels are permuted.

GSEA with Sample Label Permutations (GSEA-S)

- GSEA but we permute which samples are in case versus control.
- Permutation based null models retains inter-entity correlations (addressing false positives)

Can't run LFC Sensitivity Analysis for GSEA-S as described before because $\hat{\theta}$ also changes when sample labels are permuted.

Updated LFC Sensitivity Analysis for GSEA-S

Same as before but when sampling from null (permutation distribution) need to use

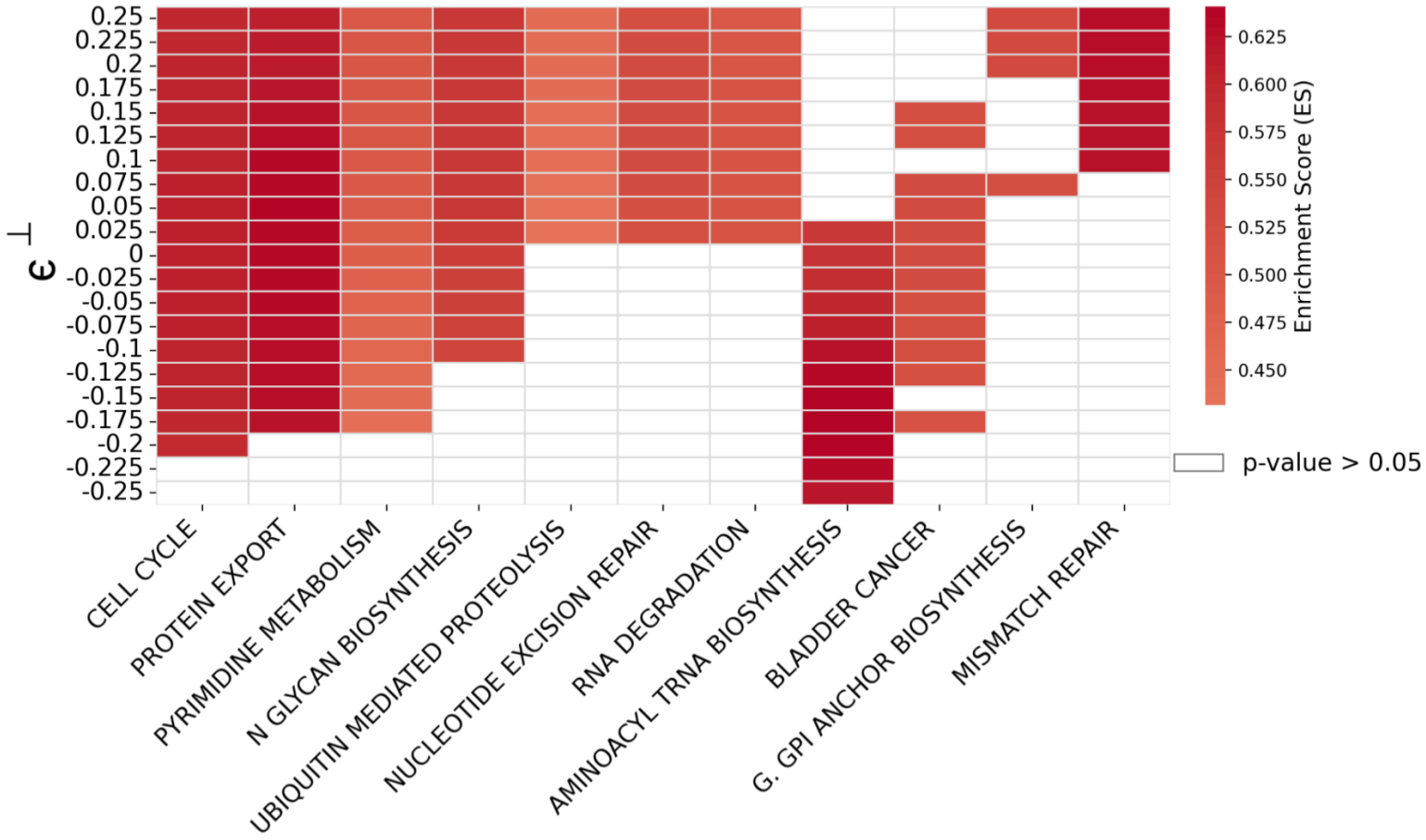
$$\phi_{S \text{ perm}} = u(\hat{\theta}_{\text{perm}} + \mathbf{1}\epsilon^{\perp})$$

Real Data Example

- GSEA-S on NAT vs. Healthy Thyroid dataset revealed no positives (!)

Real Data Example

- GSEA-S on NAT vs. Healthy Thyroid dataset revealed no positives (!)
- Instead we performed an Updated LFC Sensitivity Analysis with GSEA-S using breast Tumor vs. Healthy tissue:



Software: Just 1 more argument

```
gsea_res <-                 gsea_s(W, X, path_inds, iterations=500)
gsea_lfcs_sens_res <- gsea_s.error(W, X, path_inds, iterations=500, epsilon=c(0, -0.25))
```

	epsilon_perp	enrichment_score	p_value
KEGG_PROTEIN_EXPORT	0.00	0.6329117	0.02254098
KEGG_PROTEIN_EXPORT	-0.25	0.6114013	0.07566462

Summary Recommendations

Few Samples ($N < 20$) Use GSEA (entity label permutations) and LFC Sensitivity Analysis

Many Samples ($N \geq 20$) Use GSEA-S (sample label permutations) and Updated LFC Sensitivity Analysis

Beyond GSEA and GSEA-S

- Not all methods for DSA can be represented as

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

- Some (e.g., CAMERA) can only be written as:

$$\phi_S = u(W)$$

- See McGovern et al. (2023) *PLoS Comp. Bio* for sensitivity analysis algorithm using scale models.

Acknowledgements

- Dr. Justin Silverman (PI)
- Dr. Michelle Nixon
- Dr. Greg Gloor
- Maxwell Konaris
- Tinghua Chen
- Won Gu
- Andrew Sugarman
- Manan Saxena

Future Directions and Upcoming Works

Sparse Scale Simulation Random Variables

- Scale Models discussed by Dr. Nixon involved quantifying uncertainty in $\theta^{\perp}$ directly.
- However if we know that only a few taxa are changing (a sparsity assumption), then more powerful scale models can be developed.
- See Dr. Justin Silverman's talk Tuesday at 2:45pm in Clap Hall Auditorium for more information. *Sparse Approaches to Differential Abundance and Expression Analyses: Potential and Pitfalls*

Covariance / Network Inference

- We have focused on LFC and DSA estimands.
- Many research want to estimate networks and interactions between genes or taxa.
- Core to these methods is covariance estimation

$$\Sigma_{d_1,d_2} = \mathsf{Cov}(\log W_{d_1 \cdot}, \log W_{d_2 \cdot}).$$

- Current methods (including proportionality) suffer from substantial unacknowledged bias.
- We are developing Bayesian and Frequentist methods that address this problem.

Powerful Frequentist Hypothesis Tests for DA/DE

- Bayesian scale models as discussed by Dr. Nixon require defining a distribution of uncertainty over the scale:

$$\theta^{\perp} \sim \mathcal{N}(0, \sigma^2).$$

- We are developing a Frequentist framework where scale uncertainty is incorporated as bounds:

$$\theta^{\perp} \in [\theta_l^{\perp}, \theta_u^{\perp}].$$

- This framework allows for the development of novel, Frequentist hypothesis tests in differential expression and differential abundance:

$$H_0 : \theta_d = 0.$$