Addressing Scale Uncertainty in Gene and Microbe Set Enrichment Analysis

Kyle McGovern

The Pennsylvania State University
kvm6065@psu.edu

GLBIO 2024
April 28, 2024

Review of Key Concepts

Consider as an example an 16S rRNA-seq experiment measuring D taxa in the colons of N patients:

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

Review of Key Concepts

Consider as an example an 16S rRNA-seq experiment measuring D taxa in the colons of N patients:

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

Further consider as an example estimation of the LFC (Log Fold Change) of taxa d in patients with and without Ulcerative Colitis:

$$\underbrace{\theta_d}_{\text{LFC in Absolute Abundance}} = \underbrace{\theta_d^{\parallel}}_{\text{LFC in Composition}} + \underbrace{\theta^{\perp}}_{\text{LFC in Scale}} .$$

Review of Key Concepts

Methods like ALDEx2, DESeq2, Limma, etc. estimate LFCs using sequence count data Y:

$$f(Y) = \hat{\theta}_d$$

$$= \underbrace{\hat{\theta}_d^{\parallel}}_{\substack{\text{Estimated LFC in the} \\ \text{measured composition}}} + \underbrace{\hat{\theta}^{\perp}}_{\substack{\text{Estimated LFC in the} \\ \text{unmeasured scale}}} .$$

Review of Key Concepts

Methods like ALDEx2, DESeq2, Limma, etc. estimate LFCs using sequence count data Y:

$$f(Y) = \hat{\theta}_d$$
$$= \underbrace{\hat{\theta}_d^{\parallel}}_{\substack{\text{Estimated LFC in the} \\ \text{measured composition}}} + \underbrace{\hat{\theta}^{\perp}}_{\substack{\text{Estimated LFC in the} \\ \text{unmeasured scale}}} .$$

Estimates $\hat{\theta}^{\perp}$ come from normalization, for example:

- Total Sum Scaling (TSS): $\hat{\theta}^{\perp} = 0$

Review of Key Concepts

Methods like ALDEx2, DESeq2, Limma, etc. estimate LFCs using sequence count data Y:

$$f(Y) = \hat{\theta}_d$$
$$= \underbrace{\hat{\theta}_d^{\parallel}}_{\substack{\text{Estimated LFC in the} \\ \text{measured composition}}} + \underbrace{\hat{\theta}^{\perp}}_{\substack{\text{Estimated LFC in the} \\ \text{unmeasured scale}}} .$$

Estimates $\hat{\theta}^{\perp}$ come from normalization, for example:

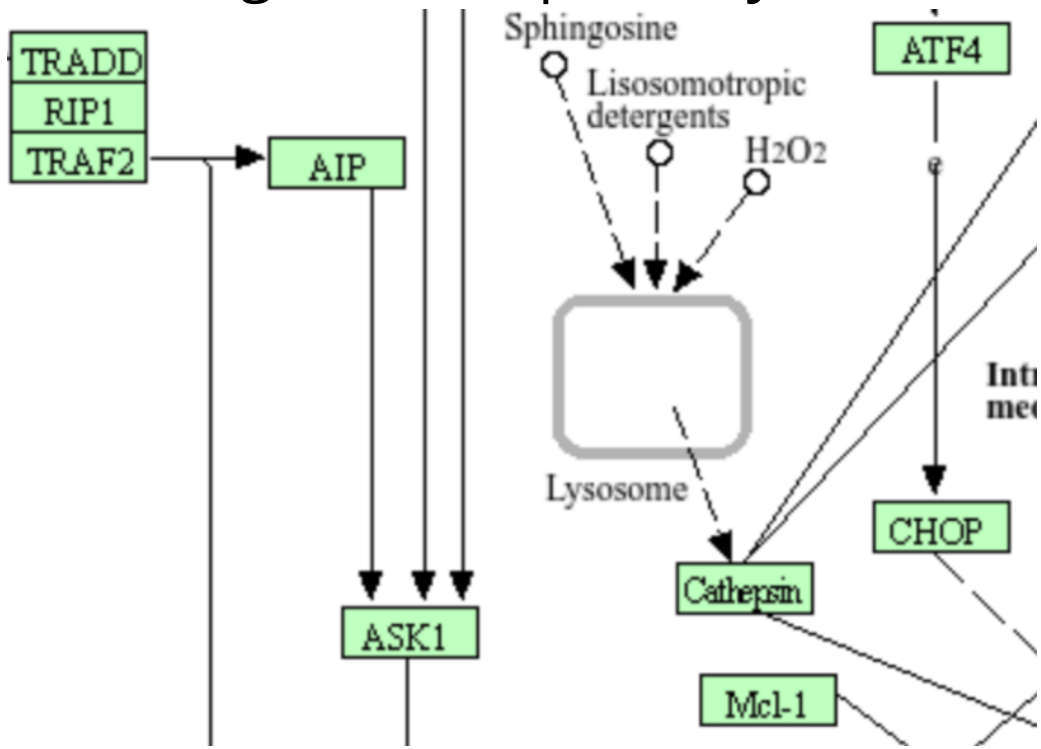
- Total Sum Scaling (TSS): $\hat{\theta}^{\perp} = 0$
- Centered Log Ratio (CLR): $\hat{\theta}^{\perp} = -\text{mean}(\hat{\theta}^{\parallel})$

Differential Set Analysis (DSA)

Rather than estimating **LFCs** of **single** genes/taxa

ASK1

What if we are interested in a **set** of genes in a pathway?



Differential Set Analysis (DSA) is used to estimate enrichment or depletion of a gene/taxa set

Key Points of this Talk

1. Errors in scale assumptions (i.e., estimates $\hat{\theta}^{\perp}$, $\hat{W}^{\perp}$) inflate false positive rates in DSA

Key Points of this Talk

1. Errors in scale assumptions (i.e., estimates $\hat{\theta}^{\perp}$, $\hat{W}^{\perp}$) inflate false positive rates in DSA
2. Errors in DSA estimates are a **non-linear** function of scale errors

Key Points of this Talk

1. Errors in scale assumptions (i.e., estimates $\hat{\theta}^{\perp}$, $\hat{W}^{\perp}$) inflate false positive rates in DSA
2. Errors in DSA estimates are a **non-linear** function of scale errors
3. We have developed three solutions to these errors:
    1. LFC Sensitivity Analysis
    2. LFC Sensitivity Testing
    3. Compositional Weighting Methods

Three Methods for DSA

In this presentation 3 common DSA methods will be considered

1. **Gene Set Enrichment Analysis (GSEA) with Gene Label permutations**
2. Gene Set Enrichment Analysis (GSEA) with Sample Label permutations
3. CAMERA

GSEA with Gene Label Permutations

The GSEA Algorithm Step-by-Step

1. Pick a set of genes S (e.g., the apoptosis signaling pathway):

$$S = \{\text{ASK1, CHOP, TRAF2}\}$$

GSEA with Gene Label Permutations

The GSEA Algorithm Step-by-Step

1. Pick a set of genes S (e.g., the apoptosis signaling pathway):

$$S = \{\text{ASK1, CHOP, TRAF2}\}$$

2. Estimate LFCs $\hat{\theta} = f(Y)$ (i.e., with DESeq2, ALDEx2, limma)

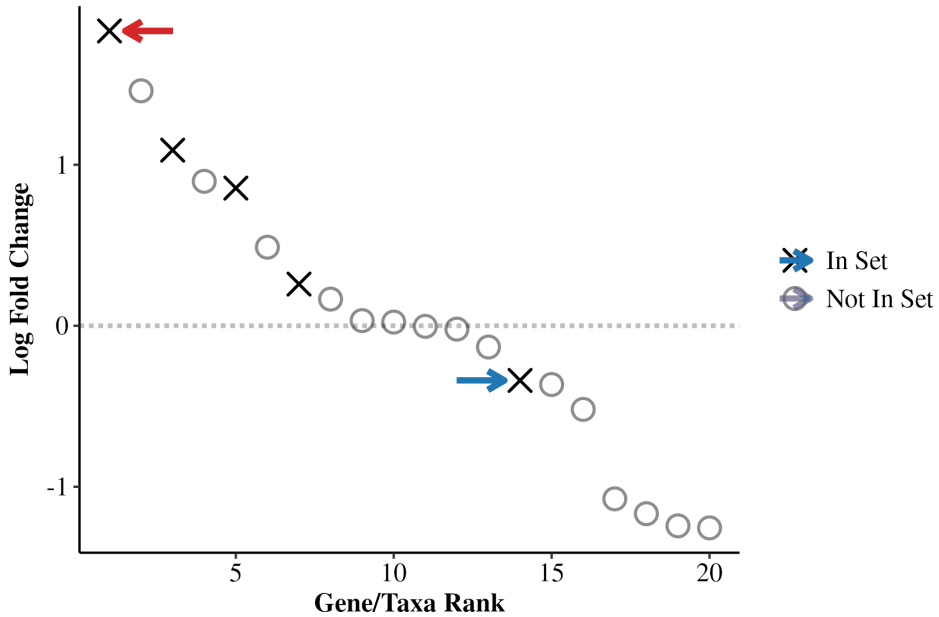
GSEA with Gene Label Permutations

The GSEA Algorithm Step-by-Step

1. Pick a set of genes S (e.g., the apoptosis signaling pathway):

$$S = \{\text{ASK1, CHOP, TRAF2}\}$$

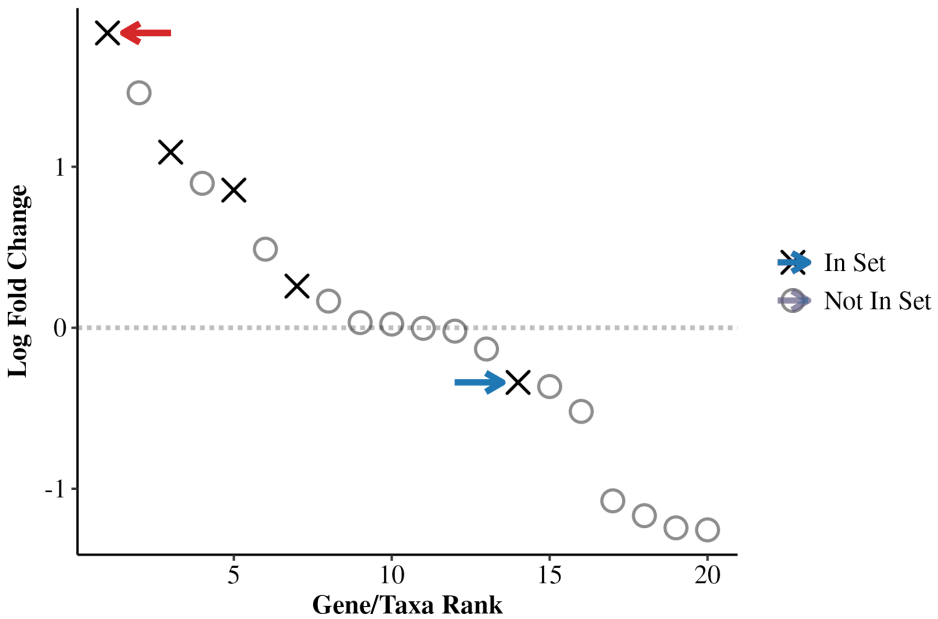
2. Estimate LFCs $\hat{\theta} = f(Y)$ (i.e., with DESeq2, ALDEx2, limma)
3. Order the LFCs from largest to smallest

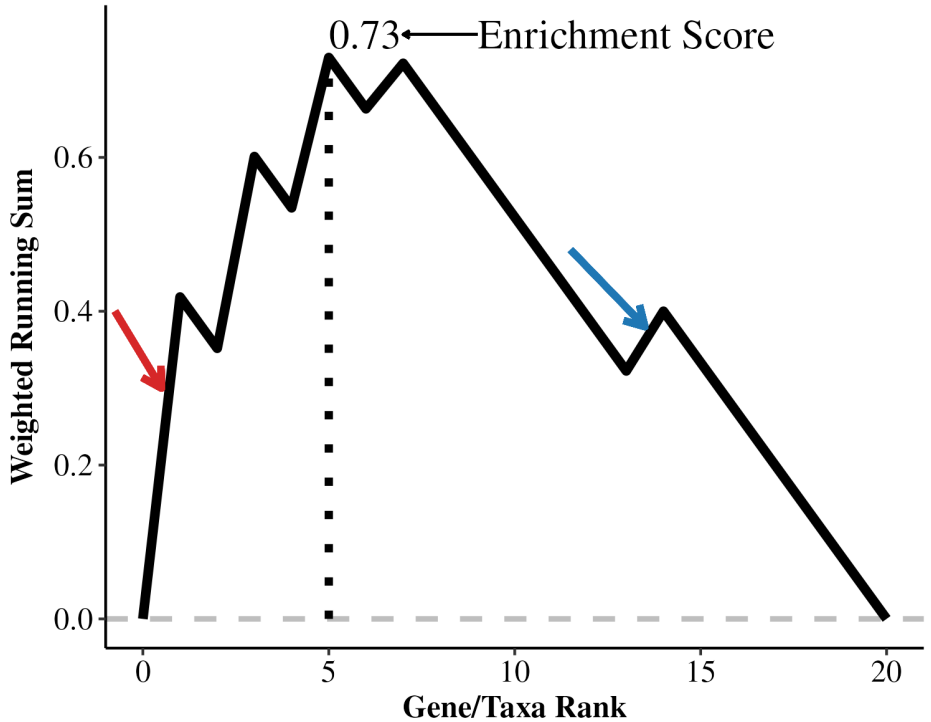


GSEA with Gene Label Permutations

The GSEA Algorithm Step-by-Step

3. Calculate a running sum **weighted** by the LFC
4. Calculate an enrichment score (max distance from 0 of **weighted** running sum)





GSEA with Gene Label Permutations

The GSEA Algorithm Step-by-Step

5. Calculate a null distribution of Enrichment Scores (ESs)

$$
\begin{aligned}
S &= \{\textbf{ASK1}, \textbf{CHOP}, \textbf{TRAF2}\} &&\Longrightarrow \text{ES} \\
S_1^* &= \{\textbf{ASK1}, \textbf{CHOP}, \text{B2M}\} &&\Longrightarrow \text{ES}_1^* \\
S_2^* &= \{\text{BRCA1}, \text{EGFR}, \text{XRCC4}\} &&\Longrightarrow \text{ES}_2^*
\end{aligned}
$$

GSEA with Gene Label Permutations

The GSEA Algorithm Step-by-Step

5. Calculate a null distribution of Enrichment Scores (ESs)

$$
\begin{aligned}
S &= \{\textbf{ASK1}, \textbf{CHOP}, \textbf{TRAF2}\} &&\Longrightarrow \text{ES} \\
S_1^* &= \{\textbf{ASK1}, \textbf{CHOP}, \text{B2M}\} &&\Longrightarrow \text{ES}_1^* \\
S_2^* &= \{\text{BRCA1}, \text{EGFR}, \text{XRCC4}\} &&\Longrightarrow \text{ES}_2^*
\end{aligned}
$$

6. Use null distribution to calculate a p-value

DSA Target Estimand

The goal of DSA is to estimate a **target estimand** ϕ_S:

$$\phi_S = \begin{cases} 1 & \text{Gene Set } S \text{ is significantly enriched} \\ -1 & \text{Gene Set } S \text{ is significantly depleted} \\ 0 & \text{Gene Set } S \text{ is not significantly changing.} \end{cases}$$

In GSEA the target estimand is a function of the **true** LFCs:

$$\phi_S = g(\theta)$$

LFC Sensitivity Analysis

In GSEA the target estimand is a function of the **true** LFCs:

$$\phi_S = g(\theta)$$

But we don't know the true LFCs, we only have **estimates**:

$$\begin{aligned}\hat{\phi}_S &= g(\hat{\theta}) \\ &= g(\hat{\theta}^{\parallel} + \underbrace{\hat{\theta}^{\perp}}_{\substack{\text{Estimated LFC in Scale} \\ \text{(Normalization Assumption)}}}).\end{aligned}$$

LFC Sensitivity Analysis

In GSEA the target estimand is a function of the **true** LFCs:

$$\phi_S = g(\theta)$$

But we don't know the true LFCs, we only have **estimates**:

$$\begin{aligned}\hat{\phi}_S &= g(\hat{\theta}) \\ &= g(\hat{\theta}^{\|} + \underbrace{\hat{\theta}^{\perp}}_{\substack{\text{Estimated LFC in Scale} \\ \text{(Normalization Assumption)}}}).\end{aligned}$$

Our DSA estimate $\hat{\phi}_S$ **depends on our scale estimate** $\hat{\theta}^{\perp}$!

LFC Sensitivity Analysis

In GSEA the target estimand is a function of the **true** LFCs:

$$\phi_S = g(\theta)$$

But we don't know the true LFCs, we only have **estimates**:

$$\begin{aligned}\hat{\phi}_S &= g(\hat{\theta}) \\ &= g(\hat{\theta}^{\|} + \underbrace{\hat{\theta}^{\perp}}_{\substack{\text{Estimated LFC in Scale} \\ \text{(Normalization Assumption)}}}).\end{aligned}$$

Our DSA estimate $\hat{\phi}_S$ **depends on our scale estimate** $\hat{\theta}^{\perp}$!

LFC Sensitivity Analysis

A sensitivity analysis of how error in $\hat{\theta}^{\perp}$ affects ϕ_S

LFC Sensitivity Analysis

Error $\epsilon^{\perp}$ in our estimate of the unmeasured scale $\theta^{\perp}$:

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

LFC Sensitivity Analysis

Error $\epsilon^{\perp}$ in our estimate of the unmeasured scale $\theta^{\perp}$:

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

How does the **true** ϕ_S change with error $\epsilon^{\perp}$?

$$\begin{aligned} \phi_s &= g(\hat{\theta}^{\parallel} + \hat{\theta}^{\perp} + \epsilon^{\perp}) \\ &= g(\hat{\theta} + \epsilon^{\perp}) \end{aligned}$$

LFC Sensitivity Analysis

Error $\epsilon^{\perp}$ in our estimate of the unmeasured scale $\theta^{\perp}$:

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

How does the **true** ϕ_S change with error $\epsilon^{\perp}$?

$$\begin{aligned} \phi_s &= g(\hat{\theta}^{\parallel} + \hat{\theta}^{\perp} + \epsilon^{\perp}) \\ &= g(\hat{\theta} + \epsilon^{\perp}) \end{aligned}$$

LFC Sensitivity Analysis Algorithm:

1. Get estimated LFCs $\hat{\theta}$ (e.g., from ALDEx2, limma, DESeq2, etc.)

LFC Sensitivity Analysis

Error $\epsilon^{\perp}$ in our estimate of the unmeasured scale $\theta^{\perp}$:

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

How does the **true** ϕ_S change with error $\epsilon^{\perp}$?

$$\begin{aligned}\phi_s &= g(\hat{\theta}^{\parallel} + \hat{\theta}^{\perp} + \epsilon^{\perp}) \\ &= g(\hat{\theta} + \epsilon^{\perp})\end{aligned}$$

LFC Sensitivity Analysis Algorithm:

1. Get estimated LFCs $\hat{\theta}$ (e.g., from ALDEx2, limma, DESeq2, etc.)
2. Run GSEA with $\epsilon^{\perp} = 0$ (i.e., $\hat{\phi}_S = g(\hat{\theta})$)

LFC Sensitivity Analysis

Error $\epsilon^{\perp}$ in our estimate of the unmeasured scale $\theta^{\perp}$:

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

How does the **true** ϕ_S change with error $\epsilon^{\perp}$?

$$\begin{aligned} \phi_s &= g(\hat{\theta}^{\parallel} + \hat{\theta}^{\perp} + \epsilon^{\perp}) \\ &= g(\hat{\theta} + \epsilon^{\perp}) \end{aligned}$$

LFC Sensitivity Analysis Algorithm:

1. Get estimated LFCs $\hat{\theta}$ (e.g., from ALDEx2, limma, DESeq2, etc.)
2. Run GSEA with $\epsilon^{\perp} = 0$ (i.e., $\hat{\phi}_S = g(\hat{\theta})$)
3. Rerun GSEA with $\epsilon^{\perp} \neq 0$ and compare to $\epsilon^{\perp} = 0$ (i.e., $\phi_S = g(\hat{\theta} + \epsilon^{\perp})$

Interpreting Error $\epsilon^{\perp}$ and LFC Sensitivity Analysis Results

Consider error $\epsilon^{\perp} = \pm 0.5$:

1. This error corresponds to the true $\theta^{\perp}$ being $e^{0.5} = 1.65$ times lower/higher than $\hat{\theta}^{\perp}$

Interpreting Error $\epsilon^{\perp}$ and LFC Sensitivity Analysis Results

Consider error $\epsilon^{\perp} = \pm 0.5$:

1. This error corresponds to the true $\theta^{\perp}$ being $e^{0.5} = 1.65$ times lower/higher than $\hat{\theta}^{\perp}$
2. Example results if a Gene set S is sensitive to error:

$\epsilon^{\perp} = -0.5$	$\epsilon^{\perp} = 0$	$\epsilon^{\perp} = 0.5$
$\phi_S = 0$	$\phi_S = 1$	$\phi_S = 0$

Interpreting Error $\epsilon^{\perp}$ and LFC Sensitivity Analysis Results

Consider error $\epsilon^{\perp} = \pm 0.5$:

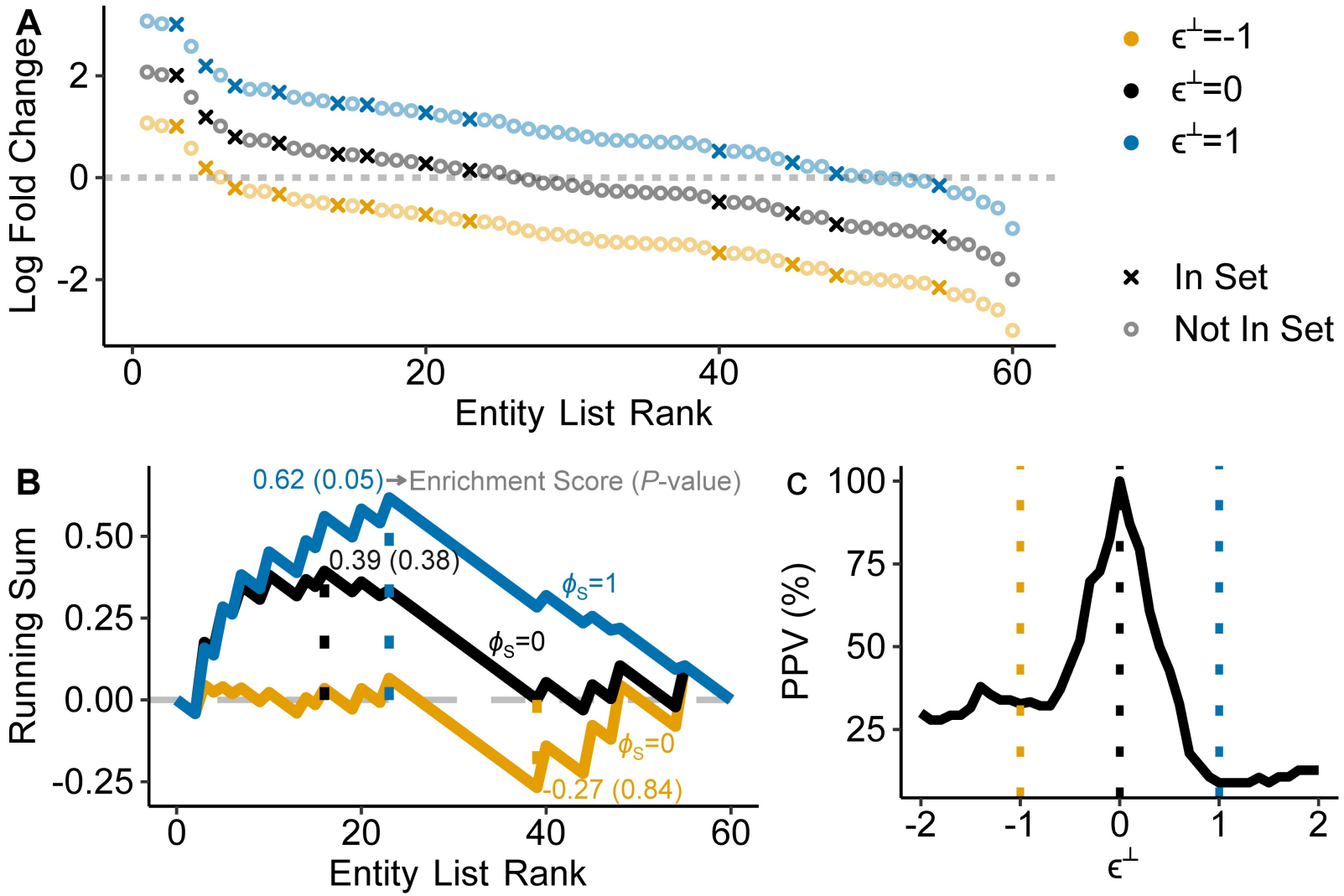
1. This error corresponds to the true $\theta^{\perp}$ being $e^{0.5} = 1.65$ times lower/higher than $\hat{\theta}^{\perp}$
2. Example results if a Gene set S is sensitive to error:

$\epsilon^{\perp} = -0.5$	$\epsilon^{\perp} = 0$	$\epsilon^{\perp} = 0.5$
$\phi_S = 0$	$\phi_S = 1$	$\phi_S = 0$

3. Example results if a Gene set S is not sensitive to error:

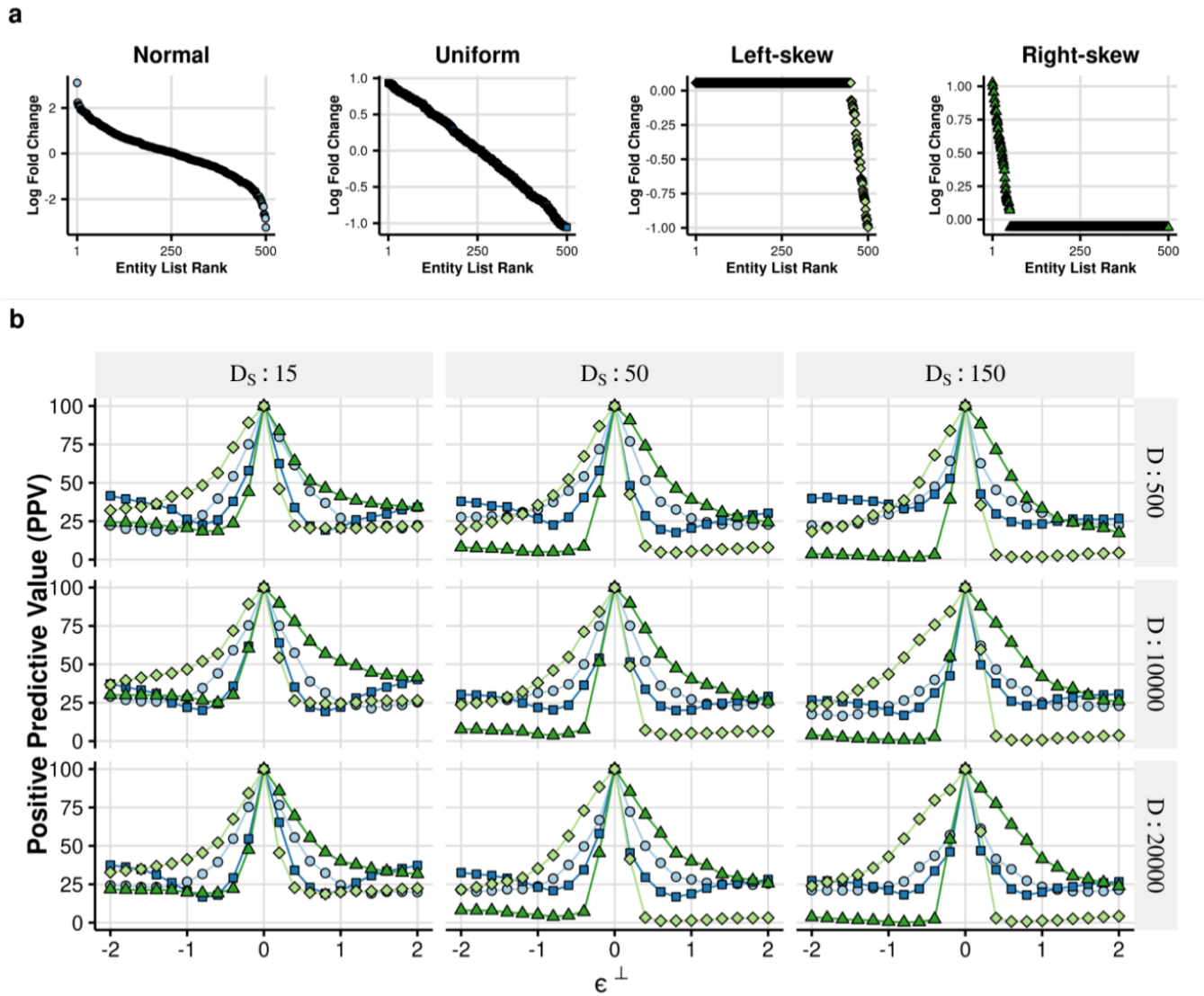
$\epsilon^{\perp} = -0.5$	$\epsilon^{\perp} = 0$	$\epsilon^{\perp} = 0.5$
$\phi_S = 1$	$\phi_S = 1$	$\phi_S = 1$

LFC Sensitivity Analysis Simulation



PPV (Positive Predictive Value) is the % of positives that are true positives

Multiple Simulations of Different LFC Distributions, Gene Set Sizes, and Total #'s of Genes



Real Data Analysis

RNA-seq: normal-adjacent-to-tumor vs. healthy thyroid tissue*

1. LFCs estimated with Songbird (Morton et al., 2019)
2. GSEA performed with fgsea (Korotkevich et al., 2021)

Real Data Analysis

RNA-seq: normal-adjacent-to-tumor vs. healthy thyroid tissue*

1. LFCs estimated with Songbird (Morton et al., 2019)
2. GSEA performed with fgsea (Korotkevich et al., 2021)
3. fgsea results for the Inflammatory Response Pathway:

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

(Normalized) Enrichment Score	Adjusted p-value
1.53	0.003

Real Data Analysis

RNA-seq: normal-adjacent-to-tumor vs. healthy thyroid tissue*

1. LFCs estimated with Songbird (Morton et al., 2019)
2. GSEA performed with fgsea (Korotkevich et al., 2021)
3. fgsea results for the Inflammatory Response Pathway:

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

(Normalized) Enrichment Score	Adjusted p-value
1.53	0.003

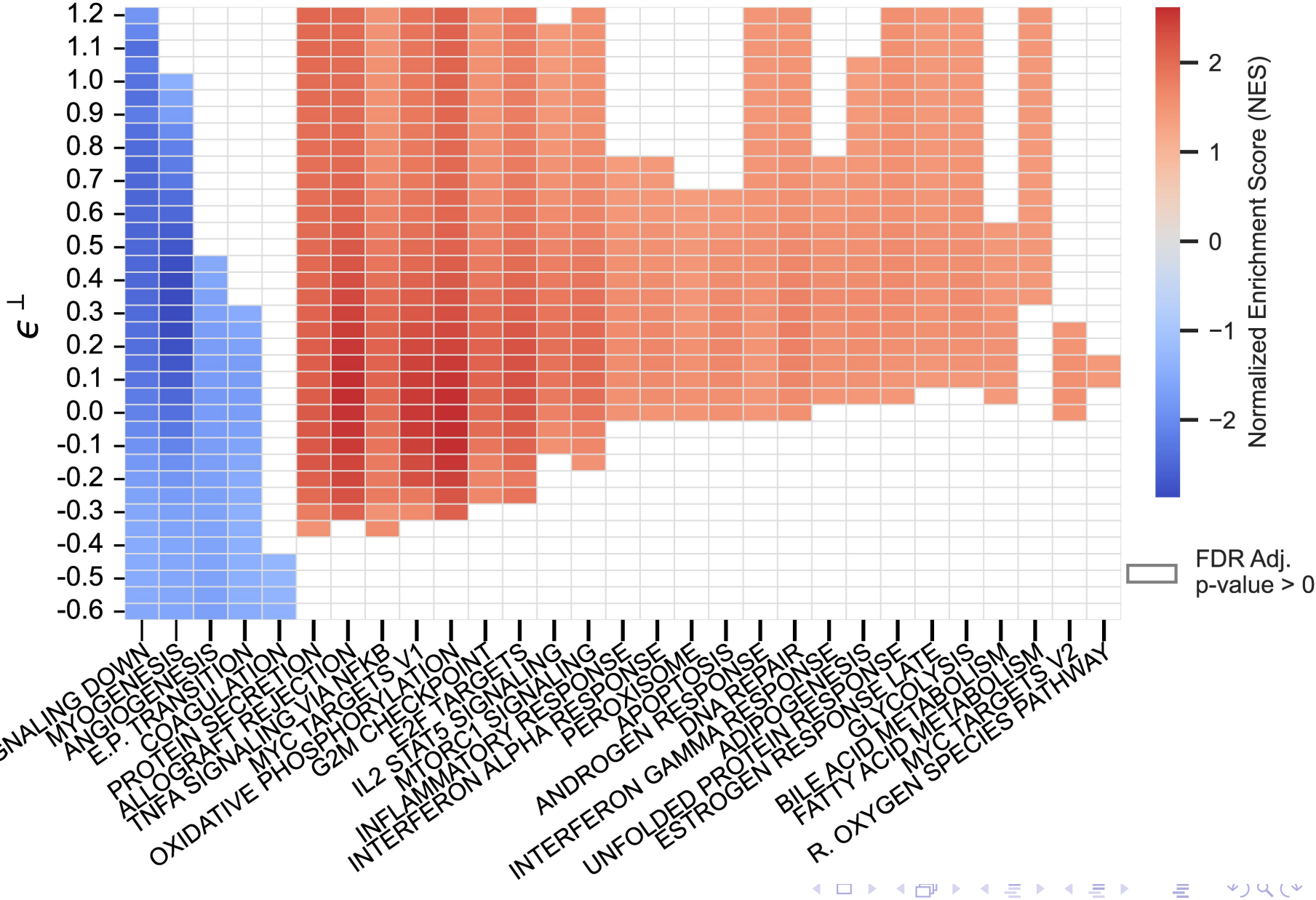
4. LFC Sensitvity Analysis results:

```
## Run LFC Sensitivity Analysis FGSEA
lfc_fgsea_res <- fgsea.error(lfcs, gmt.file,
                             epsilon=c(-0.5, 0, 0.5))
```

$\epsilon^{\perp}$	(Normalized) Enrichment Score	Adjusted p-value
-0.5	-0.89	1
0	1.53	0.003
0.5	1.5	0

*(Aran et al., 2017)

Complete Results for Real Data Analysis



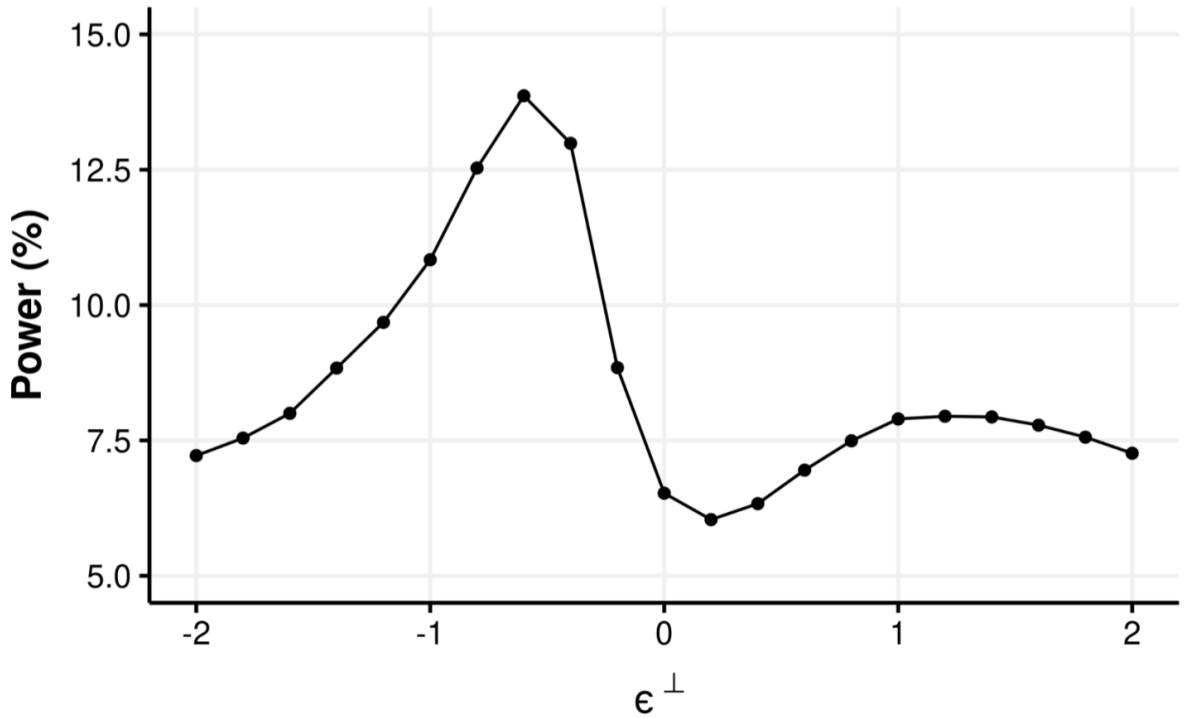
Two Additional Methods to Address False Positives

Two Additional Methods to Address False Positives

1 The LFC Sensitivity Test

Let $p_{\epsilon^\perp}$ be the GSEA p-value at $\epsilon^\perp$. Take the maximum GSEA p-value across all possible scale errors $\epsilon^\perp \in (-\infty, \infty)$:

$$p = \sup_{\epsilon^\perp \in (-\infty, \infty)} p_{\epsilon^\perp}$$



Two Additional Methods to Address False Positives

2. GSEA with Compositional Weighting (GSEA-CW)

The GSEA target estimand (i.e., goal of inference) is

$$\phi_S = g(\theta)$$

Two Additional Methods to Address False Positives

2. GSEA with Compositional Weighting (GSEA-CW)

The GSEA target estimand (i.e., goal of inference) is

$$\phi_S = g(\theta)$$

The GSEA-CW target estimand is different, and implies a different scientific question

$$\psi_S = g(\theta^{\|})$$

Two Additional Methods to Address False Positives

2. GSEA with Compositional Weighting (GSEA-CW)

The GSEA target estimand (i.e., goal of inference) is

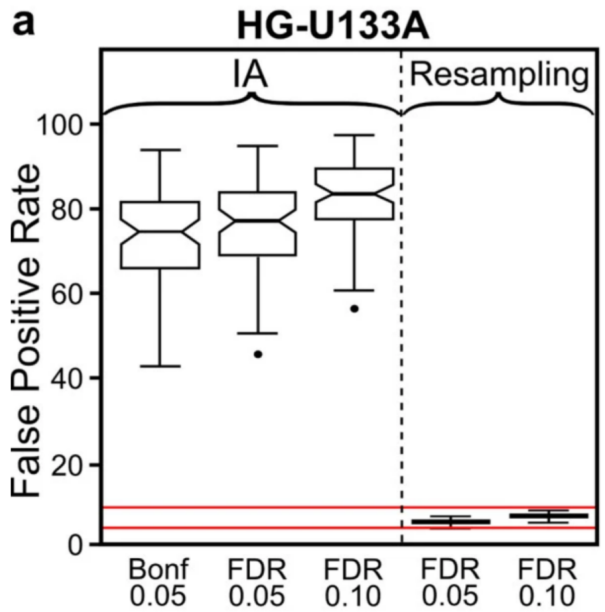
$$\phi_S = g(\theta)$$

The GSEA-CW target estimand is different, and implies a different scientific question

$$\psi_S = g(\theta^{\|})$$

DSA Methods that Account for Inter-gene/inter-taxa Correlations

Inter-gene/taxa correlations can massively inflate the false positive rate of GSEA with gene label permutations (Gatti et al., 2010)



Two methods that handle inter-gene/taxa correlations:

1. GSEA with **sample label** permutations
2. limma's CAMERA method

The CAMERA Method

Two methods that handle inter-gene/taxa correlations:

1. GSEA with **sample label** permutations

   Permute the sample labels (e.g., whether sample *n* came from healthy or tumor tissue), then re-estimate LFCs $\hat{\theta}$

The CAMERA Method

Two methods that handle inter-gene/taxa correlations:

1. GSEA with **sample label** permutations

   Permute the sample labels (e.g., whether sample n came from healthy or tumor tissue), then re-estimate LFCs $\hat{\theta}$

2. limma's CAMERA method

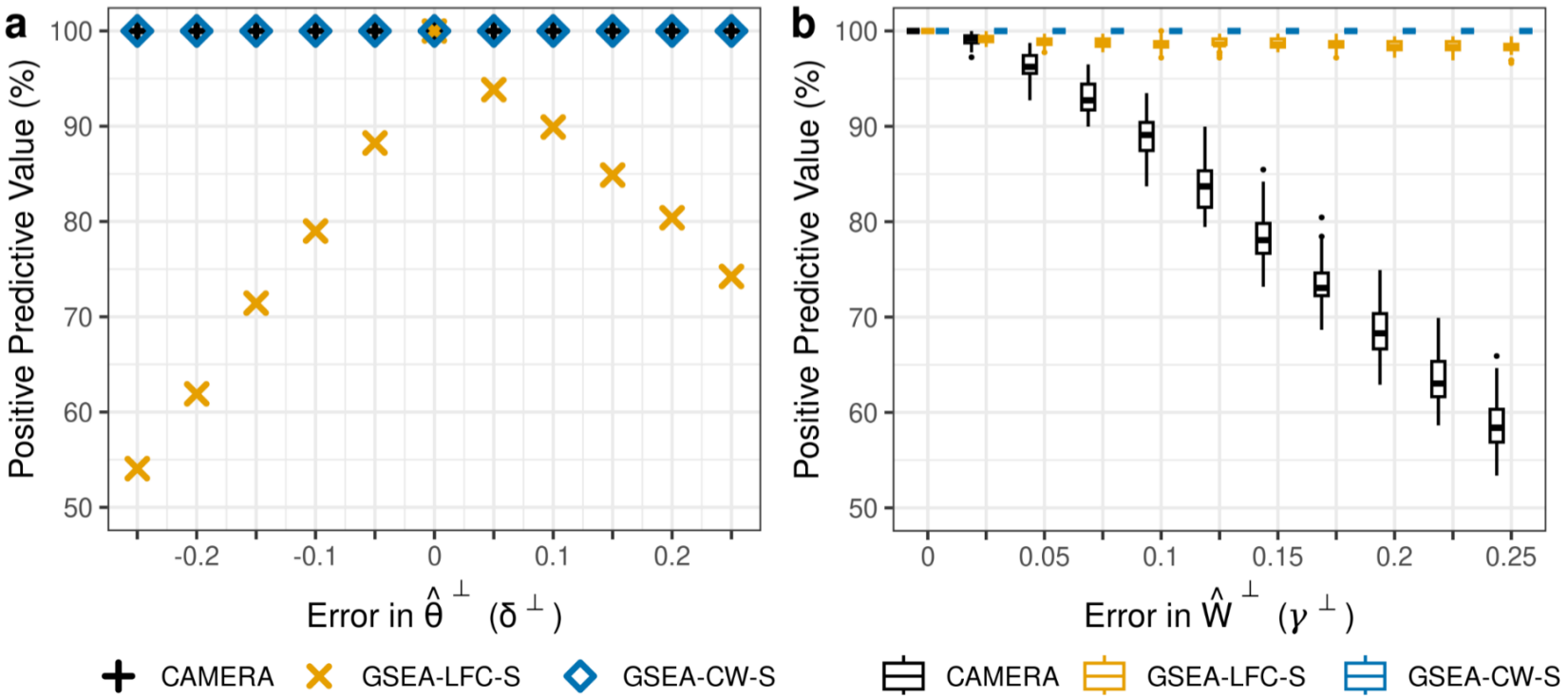
   A two-sample t-test for the set S that uses a Variance Inflation Factor (VIF) to account for correlation:

$$\frac{\hat{\theta}_{\in S} - \hat{\theta}_{\notin S}}{s_p \sqrt{\text{VIF}/m_1 + 1/m_0}}$$

Scale Sensitivity Analyses

Here I consider two real data studies:

1. Constant Error $\delta^{\perp}$ added to just samples from normal-adjacent-to-tumor
2. Variable Error simulated from Uniform$[-\gamma^{\perp}, \gamma^{\perp}]$ added to all samples



* GSEA-LFC-S is GSEA with sample label permutations, GSEA-CW-S is GSEA with compositional weighting and sample label permutations

Future Directions

The LFC Sensitivity Test considers all possible errors $\epsilon^{\perp} \in (-\infty, \infty)$:

$$p = \sup_{\epsilon^{\perp} \in (-\infty, \infty)} p_{\epsilon^{\perp}}$$

Future Directions

The LFC Sensitivity Test considers all possible errors $\epsilon^{\perp} \in (-\infty, \infty)$:

$$p = \sup_{\epsilon^{\perp} \in (-\infty, \infty)} p_{\epsilon^{\perp}}$$

But why consider *all* possible errors? For instance $\theta^{\perp} = 10$ might imply total transcription is 22,000 times higher in tumor than healthy tissue!

Future Directions

The LFC Sensitivity Test considers all possible errors $\epsilon^{\perp} \in (-\infty, \infty)$:

$$p = \sup_{\epsilon^{\perp} \in (-\infty, \infty)} p_{\epsilon^{\perp}}$$

But why consider *all* possible errors? For instance $\theta^{\perp} = 10$ might imply total transcription is 22,000 times higher in tumor than healthy tissue!

Rather assume $\epsilon^{\perp} \in [-1, 1]$:

$$p = \sup_{\epsilon^{\perp} \in [-1, 1]} p_{\epsilon^{\perp}}$$

Future Directions

The LFC Sensitivity Test considers all possible errors $\epsilon^{\perp} \in (-\infty, \infty)$:

$$p = \sup_{\epsilon^{\perp} \in (-\infty, \infty)} p_{\epsilon^{\perp}}$$

But why consider *all* possible errors? For instance $\theta^{\perp} = 10$ might imply total transcription is 22,000 times higher in tumor than healthy tissue!

Rather assume $\epsilon^{\perp} \in [-1, 1]$:

$$p = \sup_{\epsilon^{\perp} \in [-1, 1]} p_{\epsilon^{\perp}}$$

Future Directions

In the previous presentation, SSRVs were a Bayesian approach to LFC estimation

$$\theta^{\perp} = \hat{\theta}^{\perp} + \epsilon^{\perp}$$
$$\epsilon^{\perp} \sim P$$

Where P is a user-defined distribution of scale uncertainty.

Future Directions

In the previous presentation, SSRVs were a Bayesian approach to LFC estimation

$$\theta^{\perp} = \hat{\theta}^{\perp} + \epsilon^{\perp}$$
$$\epsilon^{\perp} \sim P$$

Where P is a user-defined distribution of scale uncertainty.

What if we take a *purely Frequentist* approach? Rather than define P we just assume:

$$\theta^{\perp} \in [\hat{\theta}^{\perp}_{L}, \hat{\theta}^{\perp}_{U}]$$

Let $p_{\hat{\theta}^{\perp}}$ be the p-value under a single normalization assumption $\hat{\theta}^{\perp}$, we can define a new p-value:

$$p = \sup_{\hat{\theta}^{\perp} \in [\hat{\theta}^{\perp}_{L}, \hat{\theta}^{\perp}_{U}]} p_{\hat{\theta}^{\perp}}$$

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.