Addressing Scale Uncertainty in Gene and Microbe Set Enrichment Analysis

Kyle McGovern

The Pennsylvania State University
kvm6065@psu.edu

GLBIO 2024
April 24, 2024

Review of Key Concepts

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

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

Consider estimation of the (Log Fold Change) LFC 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:

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

Estimate $\hat{\theta}^{\perp}$ comes 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:

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

Estimate $\hat{\theta}^{\perp}$ comes 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)

This presentation will focus on Differential Set Analysis (DSA) rather than Differential Expression/Abundance (DE/DA)

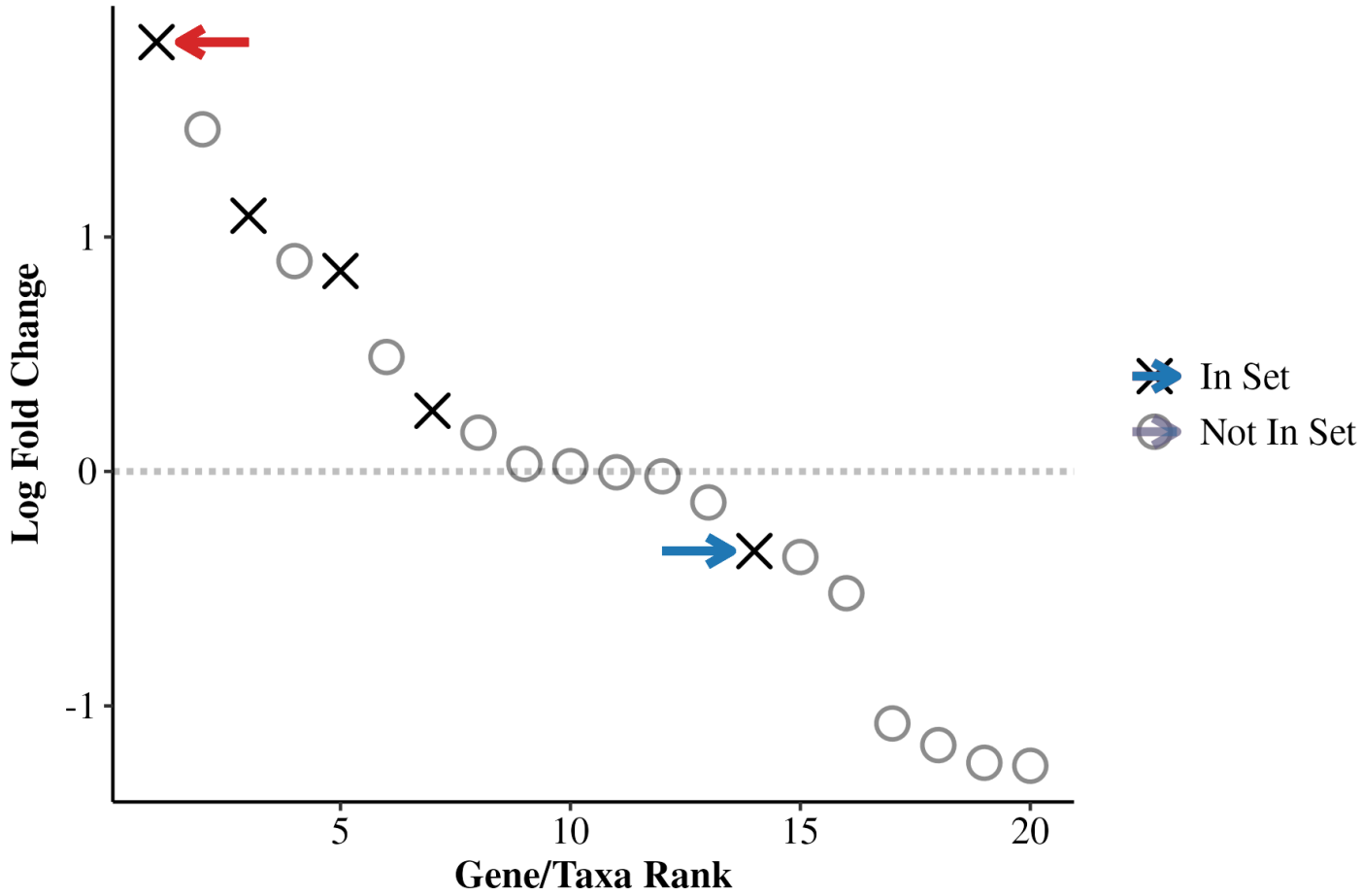
Goals of This Presentation

1. Show that **incorrect scale estimates** (i.e., $\hat{\theta}^{\perp}$ or $\hat{W}^{\perp}$) inflate false positives in DSA
2. Propose methods to address these false positives

GSEA with Gene Label Permutations

The GSEA Algorithm Step-by-Step

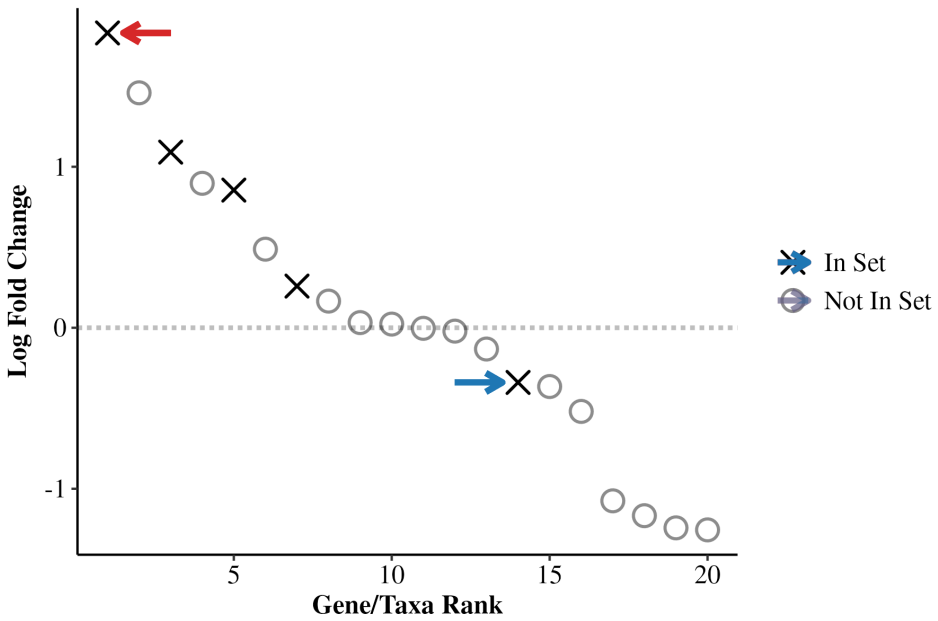
1. Pick a set of genes S (e.g., the apoptosis signaling pathway)
2. Estimate LFCs $\hat{\theta} = f(Y)$ (i.e., with DESeq2, ALDEx2, limma)
3. Order the LFCs from largest to smallest

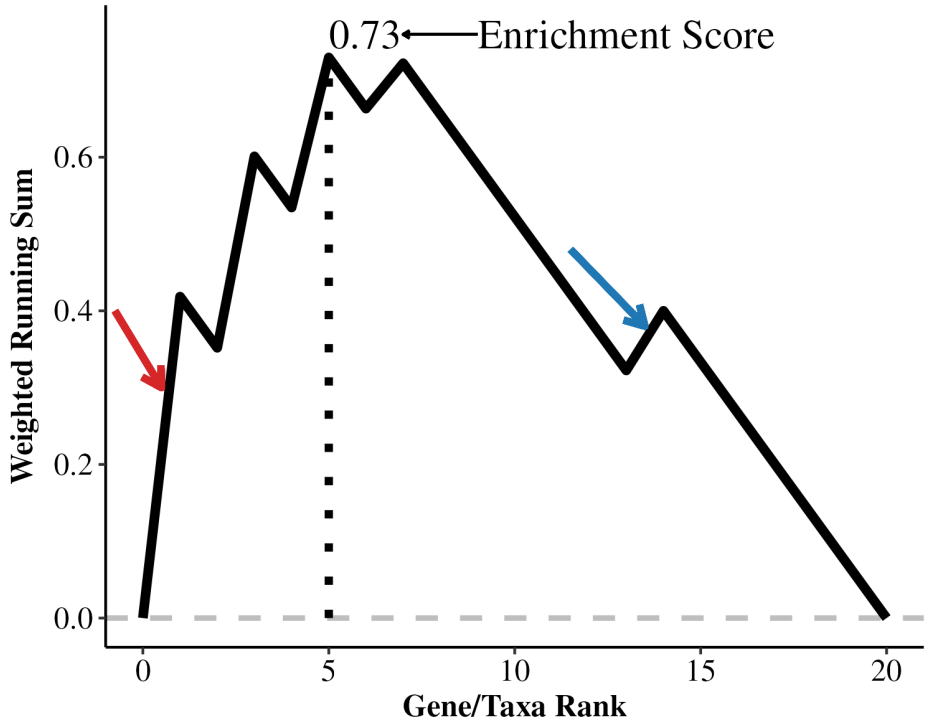


GSEA with Gene Label Permutations

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 Target of Inference

We can think of the GSEA algorithm as a function of the LFCs:

$$\underbrace{\phi_S}_{\substack{\text{DSA Target} \\ \text{Estimand}}} = u\Big(\underbrace{\theta}_{\substack{\text{True, Unkown} \\ \text{LFCs}}} \Big)$$

DSA Target Estimand:

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

We want to estimate enrichment with the true LFC θ:

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

We want to estimate enrichment with the true LFC θ:

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

But we don't know the true LFC θ! So instead we estimate:

$$\begin{aligned}\hat{\phi}_S &= u(\hat{\theta}) \\ &= u(\hat{\theta}^{\parallel} + \hat{\theta}^{\perp}).\end{aligned}$$

We want to estimate enrichment with the true LFC θ:

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

But we don't know the true LFC θ! So instead we estimate:

$$\begin{aligned}\hat{\phi}_S &= u(\hat{\theta}) \\ &= u(\hat{\theta}^{\parallel} + \hat{\theta}^{\perp}).\end{aligned}$$

The scale $\theta^{\perp}$ is unmeasured, so what if our estimate $\hat{\theta}^{\perp}$ is wrong?

LFC Sensitivity Analysis

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

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

LFC Sensitivity Analysis

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

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

We can use $\epsilon^{\perp}$ to learn how sensitive the **truth** ϕ_S is to error:

$$\begin{aligned}\phi_s &= u(\hat{\theta}^{\parallel} + \hat{\theta}^{\perp} + \mathbf{1}\epsilon^{\perp}) \\ &= u(\hat{\theta} + \mathbf{1}\epsilon^{\perp})\end{aligned}$$

LFC Sensitivity Analysis

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

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

We can use $\epsilon^{\perp}$ to learn how sensitive the **truth** ϕ_S is to error:

$$\begin{aligned}\phi_S &= u(\hat{\theta}^{\parallel} + \hat{\theta}^{\perp} + \mathbf{1}\epsilon^{\perp}) \\ &= u(\hat{\theta} + \mathbf{1}\epsilon^{\perp})\end{aligned}$$

LFC Sensitivity Analysis Algorithm:

1. Get estimated LFCs $\hat{\theta}$ (e.g., from ALDEx2, limma, DESeq2, etc.)
2. Estimate set enrichment/depletion for set S: $\hat{\phi}_S = u(\hat{\theta})$
3. Compare estimate ($\hat{\phi}_S$) to truth (ϕ_S) under different amounts of error $\epsilon^{\perp}$

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

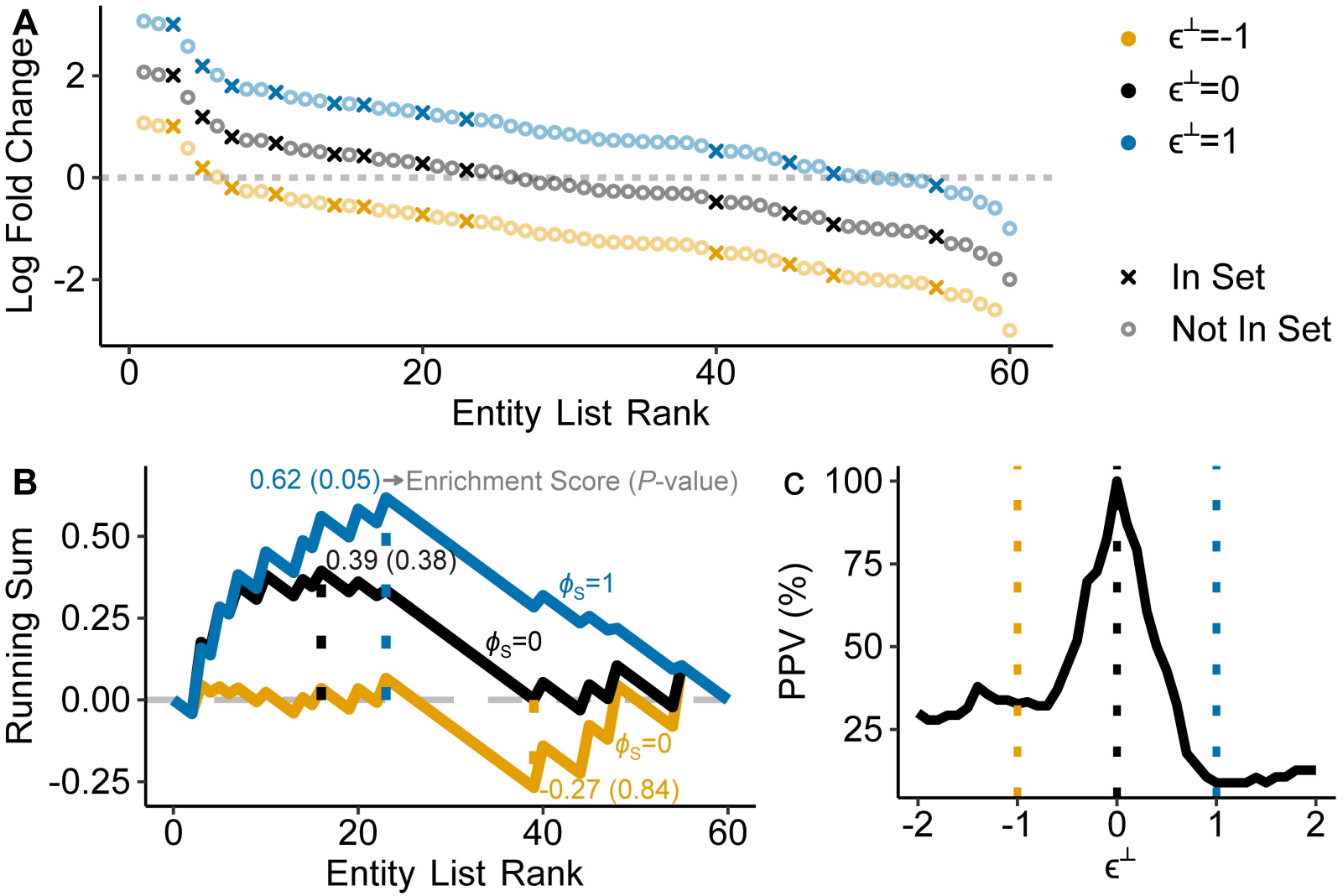
1. This error corresponds to the true $\theta^{\perp}$ being $e^{0.5} = 1.65$ times lower/higher than $\hat{\theta}^{\perp}$
2. Example results if a Gene set S is sensitive to error:

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

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

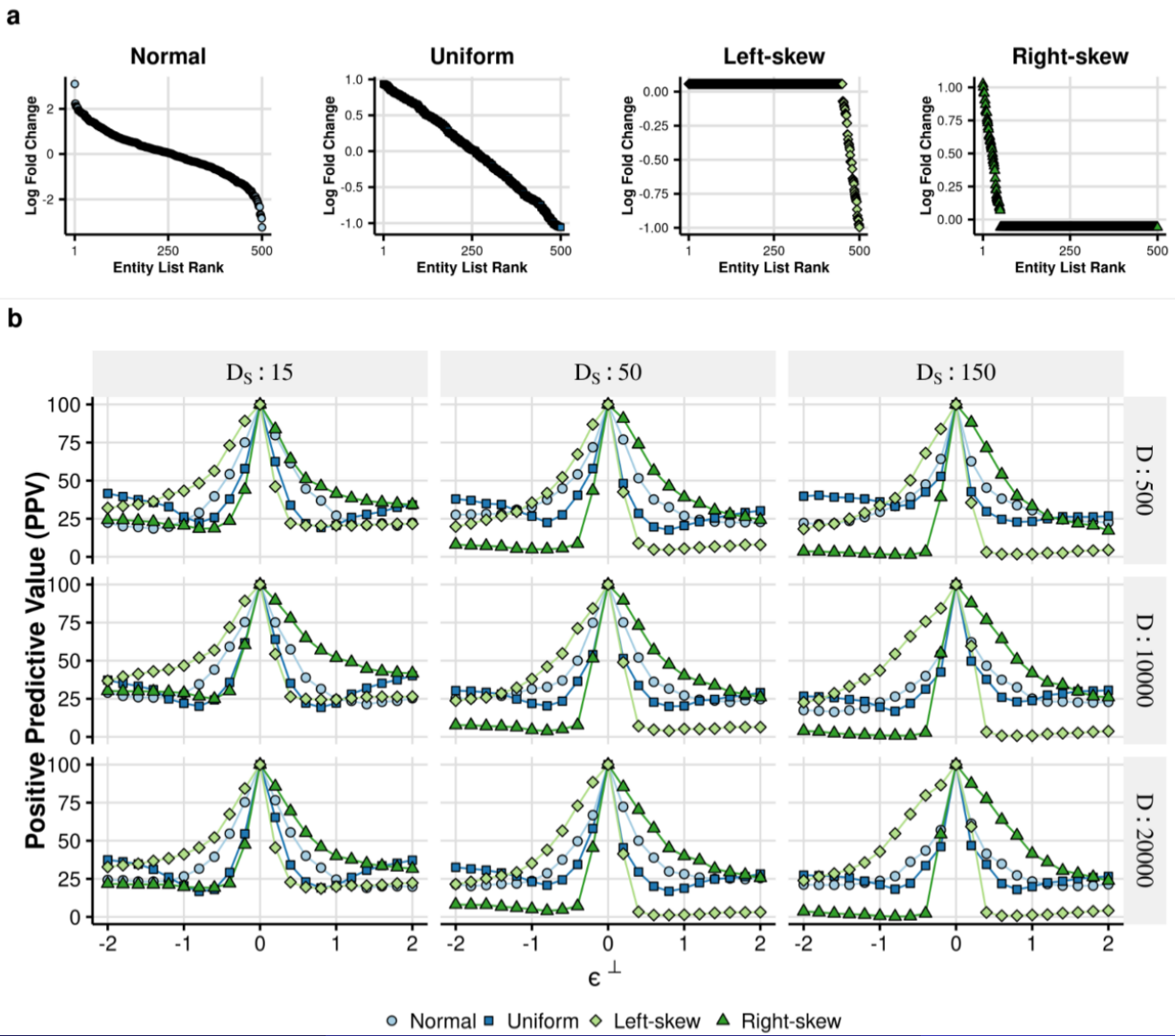
$\epsilon^{\perp} = -0.5$	$\epsilon^{\perp} = 0$	$\epsilon^{\perp} = 0.5$
$\phi_S = 1$	$\phi_S = 1$	$\phi_S = 1$

LFC Sensitivity Analysis Simulation



Note: 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

Real Data Analysis:

1. RNA-seq: normal-adjacent-to-tumor vs healthy thtroid tissue
2. LFCs estimated with Songbird (Morton et al., 2019), GSEA performed with fgsea (Korotkevich et al., 2021)

Real Data Analysis

Real Data Analysis:

1. RNA-seq: normal-adjacent-to-tumor vs healthy thtroid tissue
2. LFCs estimated with Songbird (Morton et al., 2019), GSEA performed with fgsea (Korotkevich et al., 2021)

fgsea results for the Inflammatory Response Pathway:

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

Enrichment Score	Adjusted p-value
1.53	0.003

Real Data Analysis

Real Data Analysis:

1. RNA-seq: normal-adjacent-to-tumor vs healthy thtroid tissue
2. LFCs estimated with Songbird (Morton et al., 2019), GSEA performed with fgsea (Korotkevich et al., 2021)

fgsea results for the Inflammatory Response Pathway:

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

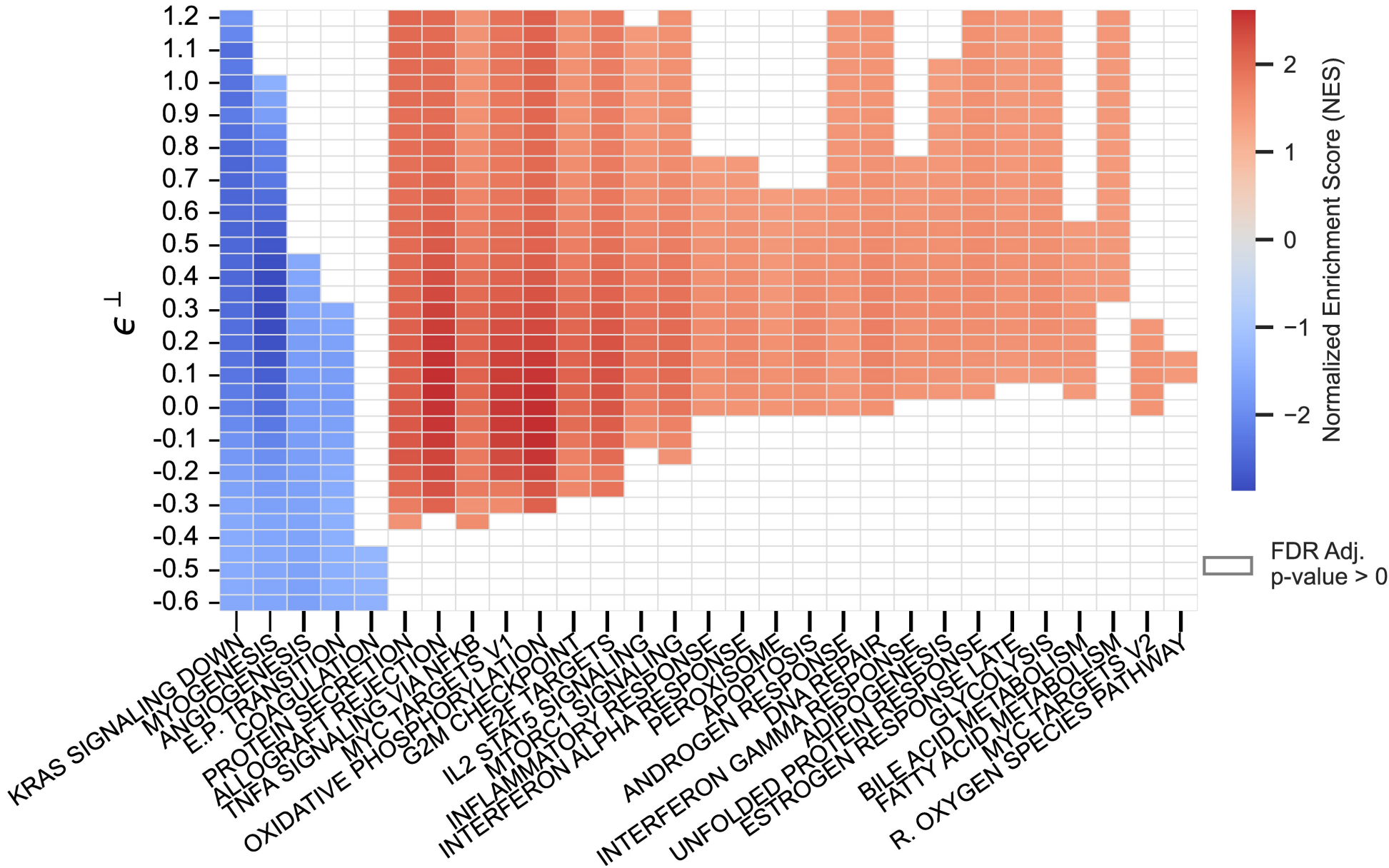
Enrichment Score	Adjusted p-value
1.53	0.003

LFC Sensitvity Results:

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

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

Complete Results for Real Data Analysis



Two Alternatives to LFC Senstivity Analysis

To address false positives in GSEA I present two additional methods:

Two Alternatives to LFC Senstivity Analysis

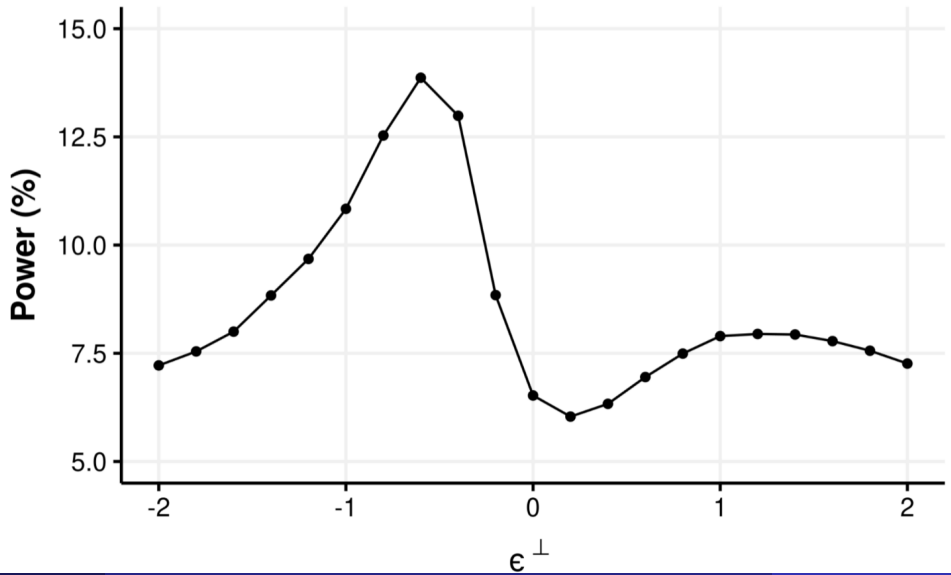
To address false positives in GSEA I present two additional methods:

1. The LFC Sensitivity Test

   Let $p_{\epsilon^\perp}$ be the GSEA p-value at $\epsilon^\perp$. The LFC Sensitivity Test's new p-value:

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

   This test has non-zero power in real data analysis:



Two Alternatives to LFC Senstivity Analysis

To address false positives in GSEA I present two additional methods:

2. GSEA with Compositional Weighting (GSEA-CW)

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

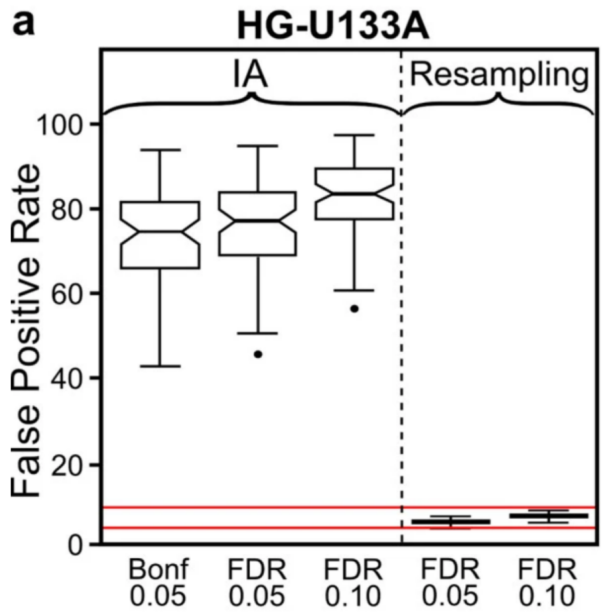
$$\phi_S = u(\theta)$$

   A different target estimand (using the CLR) implies a different inferential goal:

$$\begin{aligned}\psi_S &= u(\theta - \text{mean}(\theta)) \\ &= u(\theta^{\|} + \theta^{\perp} - \text{mean}(\theta^{\|}) - \theta^{\perp}) \\ &= u(\theta^{\|} - \text{mean}(\theta^{\|}))\end{aligned}$$

DSA Methods that Account for Inter-gene/inter-entity 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

   Rather than permute gene set S; estimate absolute abundances $\hat{W}$, permute sample labels, and re-estimate the LFC $\hat{\theta}$.

The CAMERA Method

Two methods that handle inter-gene/taxa correlations:

1. GSEA with **sample label** permutations

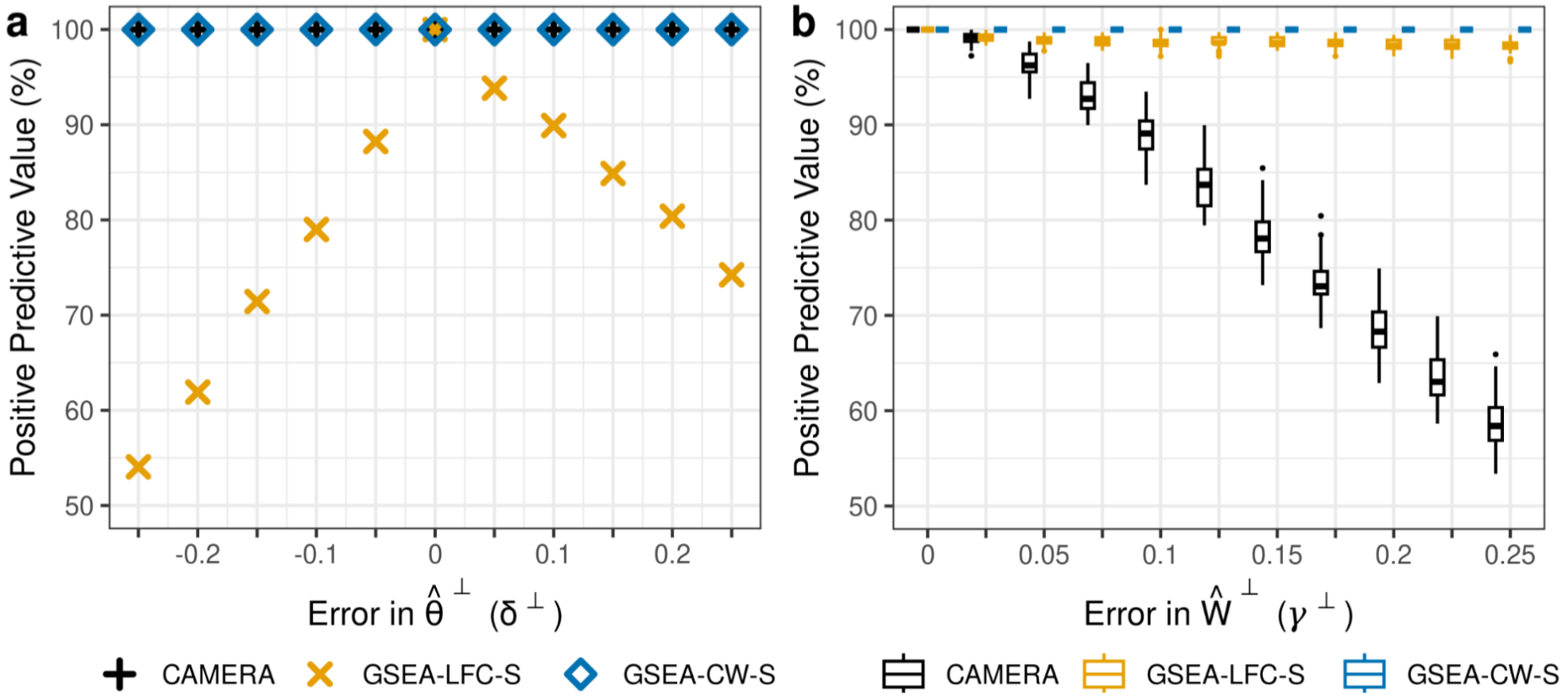
   Rather than permute gene set S; estimate absolute abundances $\hat{W}$, permute sample labels, and re-estimate the LFC $\hat{\theta}$.

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



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

Future Directions

In the LFC Sensitivity Test we calculate a p-value:

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

But why consider all possible scale assumptions?

Future Directions

In the LFC Sensitivity Test we calculate a p-value:

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

But why consider all possible scale assumptions?

For instance $\theta^{\perp} = 10$ implies a $e^{10} = 22026$ fold increase in scale!

Future Directions

In the LFC Sensitivity Test we calculate a p-value:

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

But why consider all possible scale assumptions?

For instance $\theta^{\perp} = 10$ implies a $e^{10} = 22026$ fold increase in scale!

Consider a reseacher willing to make an *interval assumption* $\epsilon^{\perp} \in [-1, 1]$:

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

This could be useful beyond GSEA, but **differential expression/abundance as well(!)**