Addressing Scale Uncertainty in Gene and Microbe Set Enrichment Analysis

Kyle McGovern

The Pennsylvania State University
kvm6065@psu.edu

GLBIO 2024
May 2, 2024

Review of Key Concepts

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

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

The Log Fold Change (LFC) in abundance between patients before/after taking an antibiotic:

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

Review of Key Concepts

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

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

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

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

Review: The Bayesian Approach

We have seen a Bayesian approach to Scale Uncertainty

$$\theta^{\perp} = \underbrace{\hat{\theta}^{\perp}}_{\substack{\text{(Normalization) Estimate} \\ \text{of LFC in Scale}}} + \underbrace{\epsilon^{\perp}}_{\substack{\text{Error in} \\ \text{Normalization}}}$$

$$\epsilon^{\perp} \sim \underbrace{\mathcal{N}\left(0, \gamma^2\right)}_{\text{Scale Model (Prior)}} .$$

Equivalently we can express this prior as

$$\theta^{\perp} \sim \mathcal{N}\left(\hat{\theta}^{\perp}, \gamma^2\right).$$

Review: The Bayesian Approach

We have seen a Bayesian approach to Scale Uncertainty

$$\theta^{\perp} = \underbrace{\hat{\theta}^{\perp}}_{\substack{\text{(Normalization) Estimate} \\ \text{of LFC in Scale}}} + \underbrace{\epsilon^{\perp}}_{\substack{\text{Error in} \\ \text{Normalization}}}$$

$$\epsilon^{\perp} \sim \underbrace{\mathcal{N}\left(0, \gamma^2\right)}_{\text{Scale Model (Prior)}} .$$

Equivalently we can express this prior as

$$\theta^{\perp} \sim \mathcal{N}\left(\hat{\theta}^{\perp}, \gamma^2\right).$$

Question: What **Frequentist** alternatives can we use to handle scale error $\epsilon^{\perp}$ made through normalization?

Frequentist vs. Bayesian Statistical Methods

Bayesian Statistics

1. Data are fixed
2. Parameters are random variables
3. Priors on Paramters

$$\epsilon^{\perp} \sim \mathcal{N}(0, \gamma^2)$$

4. Bayesian Constructs:
   - Credible Intervals
   - Posterior Predictive p-values

Frequentist Statistics

1. Data are random variables
2. Parameters are fixed
3. NO Priors on Paramters

$$\cancel{\epsilon^{\perp} \sim \mathcal{N}(0, \gamma^2)}$$

4. Frequentist Constructs:
   - Confidence Intervals
   - p-values

Question: What Frequentist alternatives can we use to handle scale error $\epsilon^{\perp}$ made through normalization?

One Possible Answer: Treat $\epsilon^{\perp}$ as a **nuisance parameter**

Frequentism and Nuisance Parameters

Consider the general case of

1. Some data X
2. A parameter of interest μ
3. An **unmeasured** nuisance parameter $\lambda \in \Lambda$
4. A function f that returns an estimate ($\hat{\mu}$) and p-value (p):

$$f(X, \lambda) = (\hat{\mu}, p)$$

Key Point

Changing λ changes $\hat{\mu}$ and p.

Two Methods for Handling Nuisance Parameters

1. Frequentist Senstivity Analyses

   Simply visualize how the nuisance parameter λ affects $\hat{\mu}$, p:

Two Methods for Handling Nuisance Parameters

1. Frequentist Senstivity Analyses

   Simply visualize how the nuisance parameter λ affects $\hat{\mu}$, p:

2. Frequentist Sensitivity Tests

   Take the "worst-case" p-value over all possible λ:

$$p = \sup_{\lambda \in \Lambda} p_\lambda$$

Gene and Microbe Set Enrichment Analysis: An Example of Frequentist Sensitivity Analysis and Sensitivity Testing

Gene and Microbe Set Enrichment Analysis: An Example of Frequentist Sensitivity Analysis and Sensitivity Testing

[Placeholder visulize diff between single/pathways]

Example Experiment

Article | Open access | Published: 20 October 2017

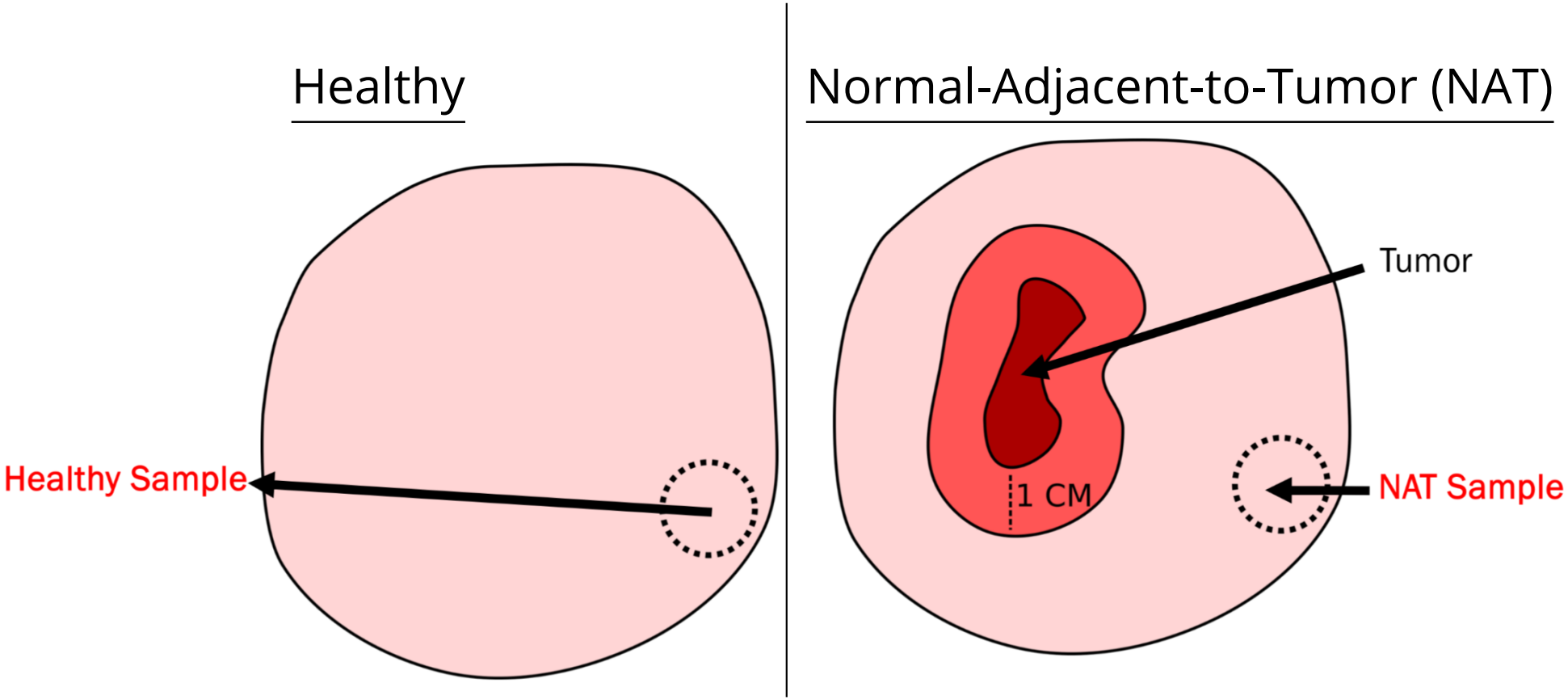
Comprehensive analysis of normal adjacent to tumor transcriptomes

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

40k Accesses | **320** Citations | **137** Altmetric | Metrics

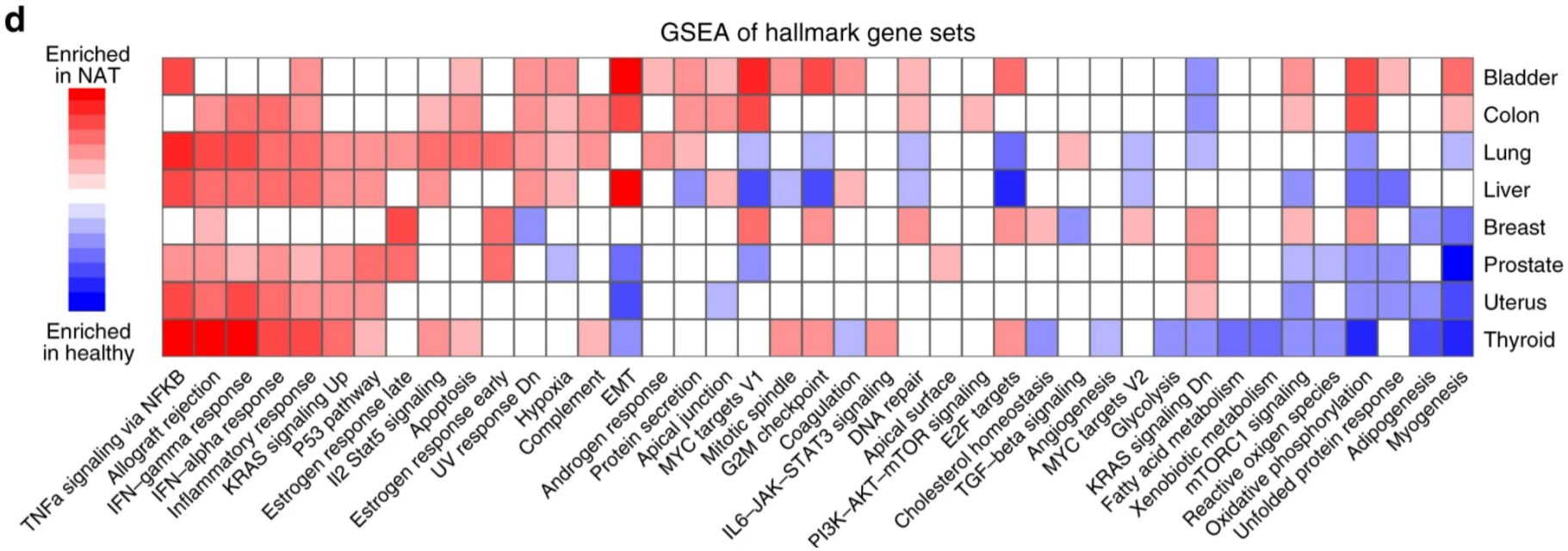
Example Experiment



Research Question

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

Example Experiment



Key Points about Pathway Enrichment Results

1. Inflammatory-related pathways enriched in NAT (e.g., TNF-α signaling and interferon response)
2. Metabolic/Differentiation pathways enriched in healthy (e.g., Myogenesis)

Exploring how error in scale (normalization) assumptions affect these results

The GSEA Algorithm

RESEARCH ARTICLE | BIOLOGICAL SCIENCES |

Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, +7, and Jill P. Mesirov Authors Info & Affiliations

September 30, 2005 | 102 (43) 15545-15550 | https://doi.org/10.1073/pnas.0506580102

Key Points about the GSEA Algorithm

1. GSEA's input is estimated LFCs $\hat{\theta}$ and a gene/microbe set S
   - e.g., $S = \{\text{EGFR}, \text{B2M}, \ldots, \text{RAD1}\}$
2. GSEA returns an Enrichment Score (ES) (i.e., effect size) and p-value
3. GSEA uses a weighting schema that **depends on** $\hat{\theta}$:
   - Changes to $\hat{\theta}$ affect the ES / p-value
4. GSEA is a permutation test using gene label permutations:

Original Gene Set

$$S = \{\text{EGFR}, \text{B2M}, \ldots, \text{RAD1}\}$$

Permuted Gene Sets

$$S_1^* = \{\text{EGFR}, \text{B2M}, \ldots, \text{RAD1}\}$$
$$S_2^* = \{\text{EGFR}, \text{B2M}, \ldots, \text{RAD1}\}$$

GSEA with Gene Label Permutations

The GSEA algorithm proposed by Subramanian et al. involves X key steps:

1. Pick Gene Set(s)
   - e.g., $S = \{\text{EGFR}, \text{B2M}, \ldots, \text{MTOR}\}$
2. Estimate LFCs $\hat{\theta} = f(Y)$
   - e.g., with DESeq2, limma, ALDEx2, Songbird, etc.
3. Rank the genes in descending order by LFC
4. **Weight the genes by their LFC**
5. Calculate a Running Sum, Enrichment Score, and p-value

GSEA with Gene Label Permutations

Mathematically the GSEA Algorithm can be written as a function g:

$$g(\hat{\theta}, S) = (\text{ES}, p)$$

What about scale error $\epsilon^{\perp}$?

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

Frequentist LFC Sensitivity Analysis

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

If we assume little compositional error ($\theta^{\parallel} \approx \hat{\theta}^{\parallel}$):

$$\underbrace{\theta_d}_{\text{True LFC}} \approx \underbrace{\hat{\theta}_d^{\parallel}}_{\text{Estimated LFC in Composition}} + \underbrace{\hat{\theta}^{\perp}}_{\text{Estimated LFC in Scale}} + \underbrace{\epsilon^{\perp}}_{\text{Scale Estimation Error}}$$

$$\approx \underbrace{\theta_d}_{\text{Estimated LFC}} + \underbrace{\epsilon^{\perp}}_{\text{Scale Estimation Error}}$$

Frequentist LFC Senstivity Analysis

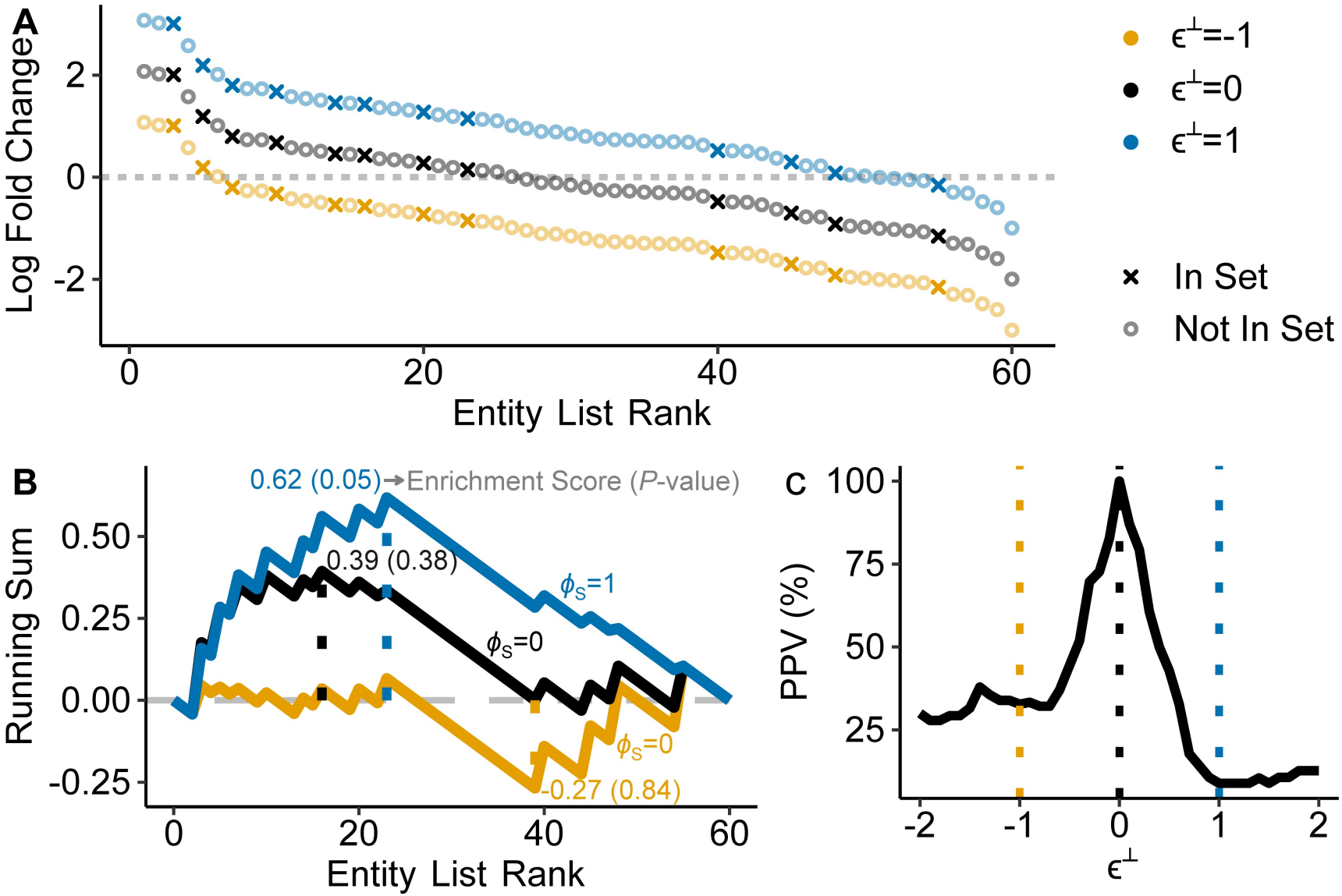
$$\underbrace{\theta_d}_{\text{True LFC}} \approx \underbrace{\theta_d}_{\text{Estimated LFC}} + \underbrace{\epsilon^{\perp}}_{\text{Scale Estimation Error}}$$

Frequentist LFC Sensitivity Analysis

$$g(\hat{\theta} + \epsilon^{\perp}, S) = (\text{ES}, p)$$

All we need to do is rerun GSEA for different $\epsilon^{\perp}$ and visualize how Enrichment Score (ES) and p-value changes

Frequentist LFC Senstivity Analysis: A Simulation



Returning to NAT vs. Healthy Tissue

Aran et al. found:

1. Myogenesis is enriched in healthy thyroid tissue
2. INF-γ is enriched in NAT thyroid tissue

Results using the fast GSEA (fgsea) package:

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

Myogenesis

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

Inflamatory Response

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

Returning to NAT vs. Healthy Tissue

Results using the fast GSEA (fgsea) package:

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:

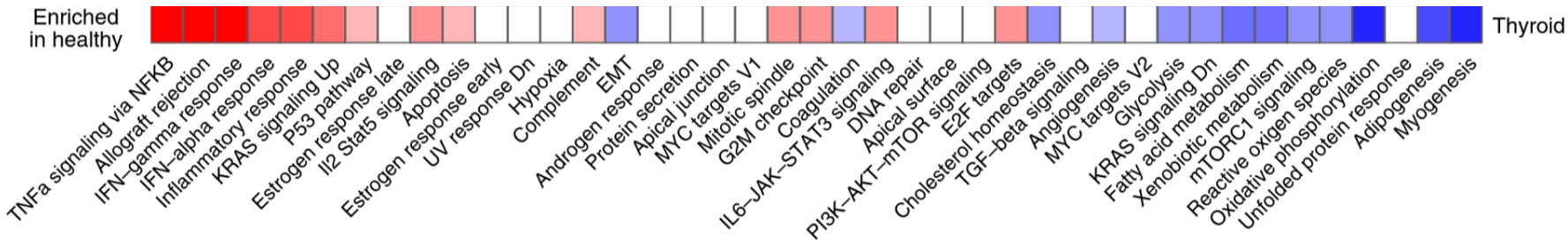
```
## Run LFC Sensitivity Analysis FGSEA
lfc_fgsea_res <- fgsea.error(lfcs, gmt.file,
                             epsilon=c(-0.4, 0, 0.4))
```

$\epsilon^{\perp}$	-0.4	0	0.5
adj. p-val	**2e-6**	**9e-9**	**2e-9**
NES	-1.4	-2.1	-2.5

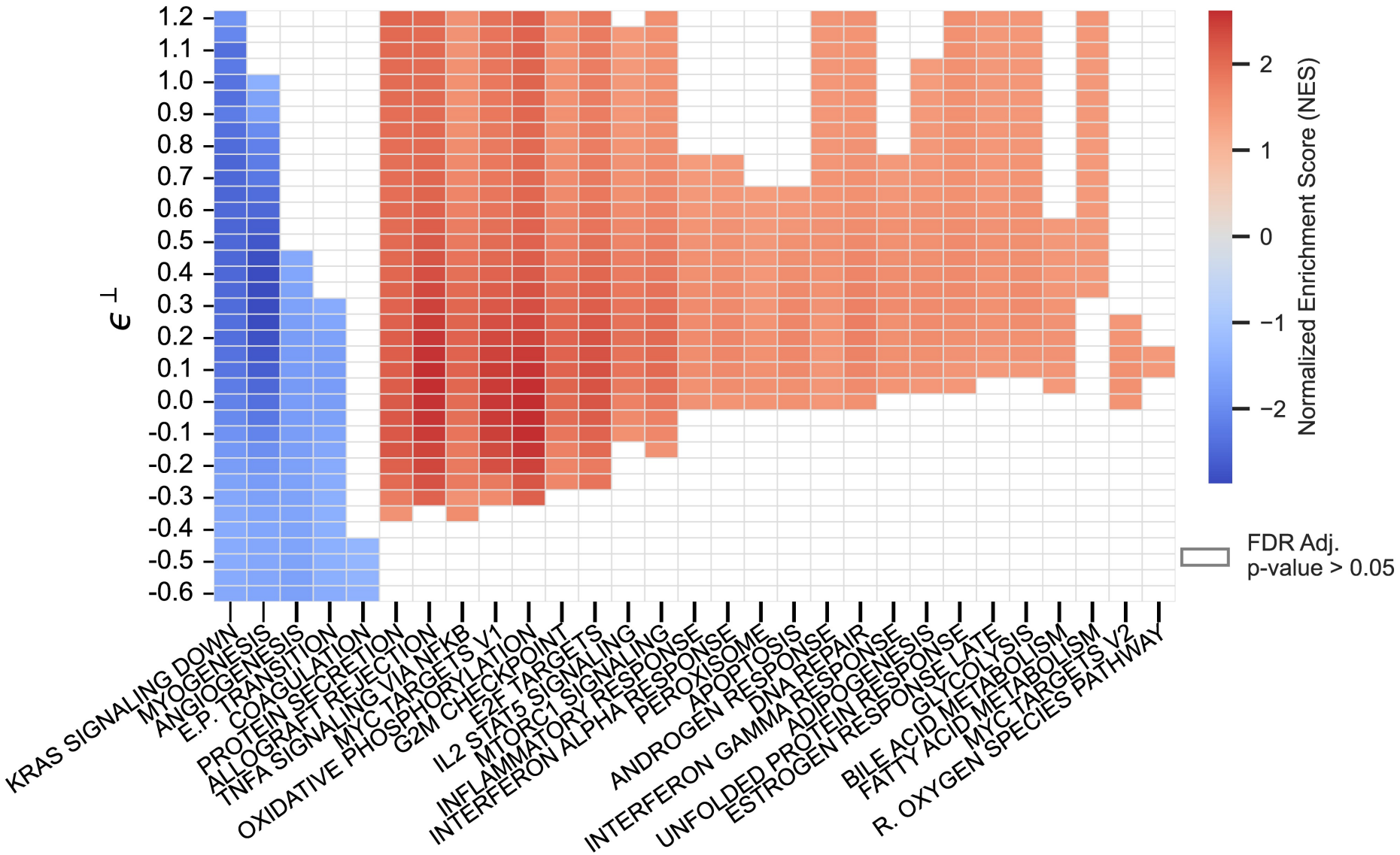
$\epsilon^{\perp}$	-0.4	0	0.5
adj. p-val	1	**3e-3**	**3e-3**
NES	-0.8	1.5	1.6

Returning to NAT vs. Healthy Tissue

Aran et al.'s original results for Thyroid tissue



Thyroid NAT vs Healthy LFC Sensitivity Analysis



Frequentist LFC Sensitivity Testing

Let $p_{\epsilon^\perp}$ be the GSEA p-value at $\epsilon^\perp$, the **LFC Sensitivity Test**:

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

Remarkably this test has non-zero power:

1. Hallmark Gene Sets: 0 /50 Significant
2. C2 Gene Sets: X/Y Significant

Improving Power of Test

Why consider all possible $\epsilon^\perp \in (-\infty, \infty)$? For instance $\epsilon^\perp = 10$ implies expression is 22,000 times higher in NAT than healthy tissue!

Bayesian Approacj