False Discovery Rate Methods and Me

Josh Habiger

March 28, 2026

Wheat Microbiome Data Sets

- Anderson and Habiger (2012).Characterization and Identification of Productivity-Associated Rhizobacteria in Wheat.
  - From the abstract: "The rhizosphere is populated by a numerous and diverse array of rhizobacteria, and many impact productivity in largely unknown ways... **Operational taxonomic units (OTUs) that were significantly associated with biomass productivity were identified using an exact test adjusted for the false discovery rate** ..."

⋮

- Ramos-Lopez, Espindola...Habiger,...(2026+). Rhizosphere-Mediated Effects on Winter Wheat Yield in Oklahoma and the Development of Diagnostic Probes Targeting Beneficial Microorganisms.
  - From the abstract: "The rhizosphere comprises a complex system of interactions among plant roots, soil conditions, and microbial communities that influence plant phenotypic traits such as growth and yield. ... This study aims to **identify and select taxa that are differentially abundant** in high-yielding winter wheat (Triticum aestivum L.) varieties across different Oklahoma soils at jointing growth stages Feekes 6 and 9 ...

The Anderson and Habiger (2012) Data and Research Questions

- Collaborator:
  - I measured the abundance of 778 rhizobacteria, or taxa, in 5 wheat rhizosphere samples.
  - **Which of the 778 bacteria are associated with productivity?**

```
url = "https://raw.githubusercontent.com/Jhabige/Jhabige.github.io/refs/heads/master/assets/bacteria.txt"
bacteria=read.table(url, row.names=1)
biomass=c(0.86, 1.34,1.81,2.37,3.0)
colnames(bacteria)= biomass
bacteria[1:5,]
```

```
##      0.86 1.34 1.81 2.37  3
## 1847    0    0    4   11 12
## 588     0    0    0    0 14
## 1697    0    0    4    1 13
## 1573    0    0    2    3  7
## 385     2    0    2    2 12
```

```
dim(bacteria)
```

```
## [1] 778   5
```

*Remark: More details and R code at https://jhabige.github.io/research/

An Answer

- Suppose we
  1. get a p-value P_i for each bacteria using a log-linear model
  2. discover bacteria as associated with productivity if $P_i < .05$
- We'd have this awkward conversation
  - Me: "Check it out! I found 135 statistically significant bacteria!"

An Answer

- Suppose we
  1. get a p-value P_i for each bacteria using a log-linear model
  2. discover bacteria as associated with productivity if $P_i < .05$
- We'd have this awkward conversation
  - Me: "Check it out! I found 135 statistically significant bacteria!"
  - Collaborator: "Thanks! I'll look into these!"

An Answer

- Suppose we
  1. get a p-value P_i for each bacteria using a log-linear model
  2. discover bacteria as associated with productivity if $P_i < .05$
- We'd have this awkward conversation
  - Me: "Check it out! I found 135 statistically significant bacteria!"
  - Collaborator: "Thanks! I'll look into these!"
  - Collaborator: "Hey a lot of these don't make sense. Is there some mistake?"

An Answer

- Suppose we
  1. get a p-value P_i for each bacteria using a log-linear model
  2. discover bacteria as associated with productivity if $P_i < .05$
- We'd have this awkward conversation
  - Me: "Check it out! I found 135 statistically significant bacteria!"
  - Collaborator: "Thanks! I'll look into these!"
  - Collaborator: "Hey a lot of these don't make sense. Is there some mistake?"
  - Me: "Oh. Yeah a lot of them could be false discoveries."

An Answer

- Suppose we
  1. get a p-value P_i for each bacteria using a log-linear model
  2. discover bacteria as associated with productivity if $P_i < .05$
- We'd have this awkward conversation
  - Me: "Check it out! I found 135 statistically significant bacteria!"
  - Collaborator: "Thanks! I'll look into these!"
  - Collaborator: "Hey a lot of these don't make sense. Is there some mistake?"
  - Me: "Oh. Yeah a lot of them could be false discoveries."
  - Collaborator: "What? How many?"

An Answer

- Suppose we
  1. get a p-value P_i for each bacteria using a log-linear model
  2. discover bacteria as associated with productivity if $P_i < .05$
- We'd have this awkward conversation
  - Me: "Check it out! I found 135 statistically significant bacteria!"
  - Collaborator: "Thanks! I'll look into these!"
  - Collaborator: "Hey a lot of these don't make sense. Is there some mistake?"
  - Me: "Oh. Yeah a lot of them could be false discoveries."
  - Collaborator: "What? How many?"
  - Me: "Well, if p-values are valid, we could expect about $778 \times 0.05 = 39$ false discoveries or a **false discovery rate** of

$$\widehat{FDR}(.05) = \frac{778 \times .05}{135} = 0.29$$

An Answer

- Suppose we
  1. get a p-value P_i for each bacteria using a log-linear model
  2. discover bacteria as associated with productivity if $P_i < .05$
- We'd have this awkward conversation
  - Me: "Check it out! I found 135 statistically significant bacteria!"
  - Collaborator: "Thanks! I'll look into these!"
  - Collaborator: "Hey a lot of these don't make sense. Is there some mistake?"
  - Me: "Oh. Yeah a lot of them could be false discoveries."
  - Collaborator: "What? How many?"
  - Me: "Well, if p-values are valid, we could expect about $778 \times 0.05 = 39$ false discoveries or a **false discovery rate** of

$$\widehat{FDR}(.05) = \frac{778 \times .05}{135} = 0.29$$

  - Collaborator: "If p-values are valid? 29 percent?"

An Answer

- Suppose we
  1. get a p-value P_i for each bacteria using a log-linear model
  2. discover bacteria as associated with productivity if $P_i < .05$
- We'd have this awkward conversation
  - Me: "Check it out! I found 135 statistically significant bacteria!"
  - Collaborator: "Thanks! I'll look into these!"
  - Collaborator: "Hey a lot of these don't make sense. Is there some mistake?"
  - Me: "Oh. Yeah a lot of them could be false discoveries."
  - Collaborator: "What? How many?"
  - Me: "Well, if p-values are valid, we could expect about $778 \times 0.05 = 39$ false discoveries or a **false discovery rate** of

$$\widehat{FDR}(.05) = \frac{778 \times .05}{135} = 0.29$$

  - Collaborator: "If p-values are valid? 29 percent?"
  - Me: "Yeah ... well ... you know we used P<.05 so that's why 29 percent"

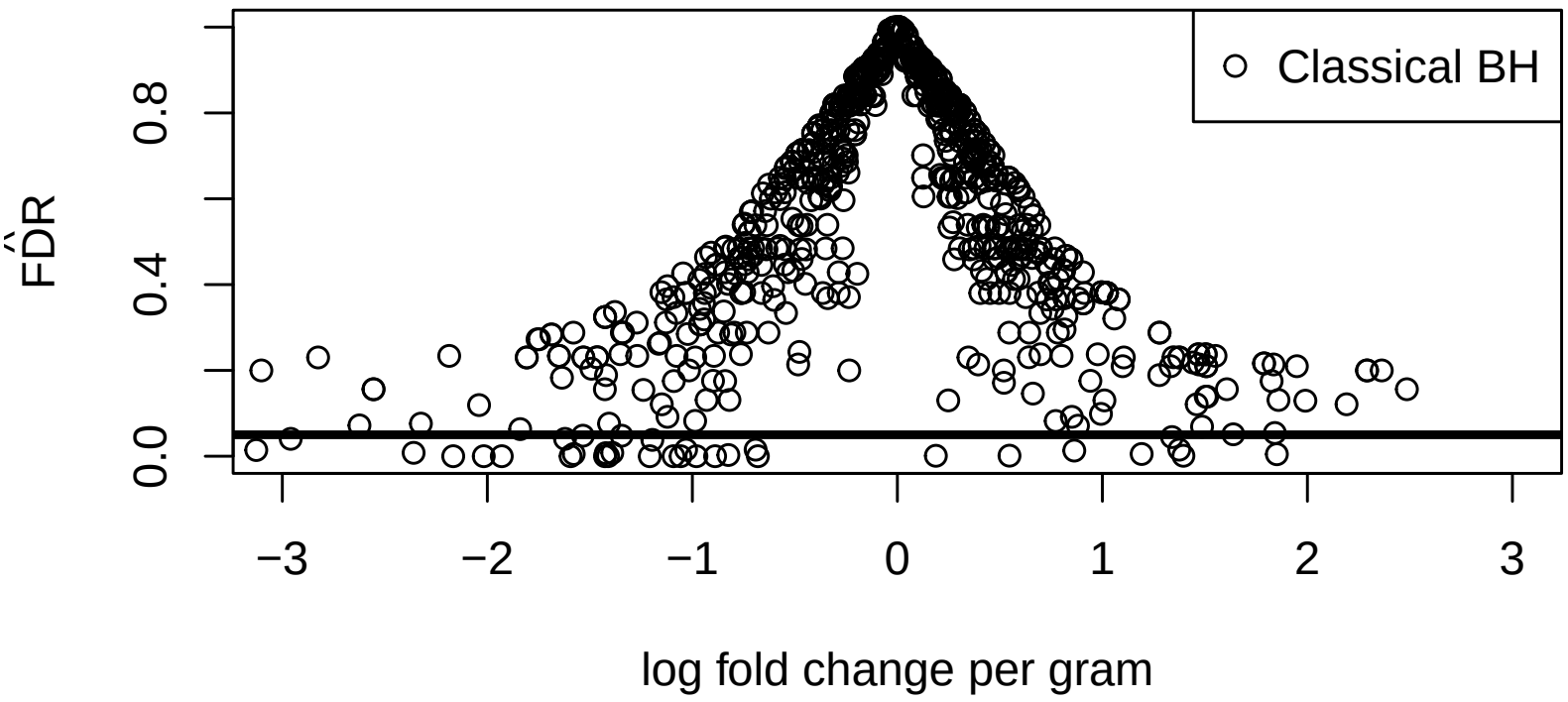
Overview of Classical and (Empirical) Bayes FDR Methods

Consider data $X_m = (X_{mn})$ and null hypothesis H_m for $m = 1, 2, ..., M$.

1. Specify statistical models and null hypothesis
2. Compute statistical significance summary
   - **p-values**: $P_m = \Pr(T(X_m) \geq T(x_m)|H_m)$
   - **posterior null probabilities**: $LFDR_m = \Pr(H_m|x_m)$
3. Estimate FDR for rejection threshold t and choose t s.t. $\widehat{FDR}(t) \leq \alpha$
   - **Classical BH:** $\widehat{FDR}(t) = \frac{Mt}{\sum_m I(P_m \leq t)}$
   - **(Empirical) Bayes:** $\widehat{FDR}(t) = \frac{\sum_m LFDR_m I(LFDR_m \leq t)}{\sum_m I(LFDR_m \leq t)}$

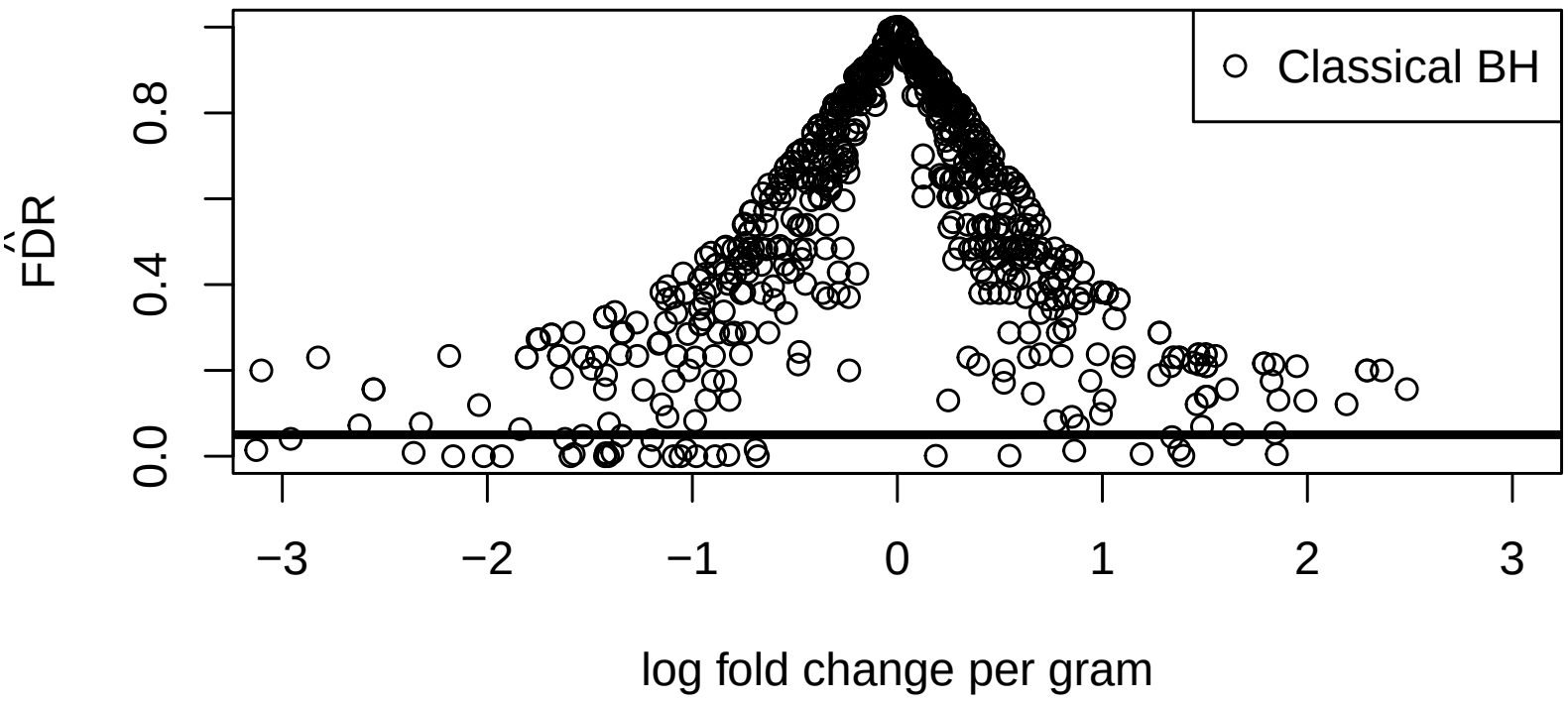
BH Procedure

- Me: "Here's 34 discoveries using $\widehat{FDR}(.008) \leq .05$ with estimated log fold changes for assessing **practical significance**"



BH Procedure

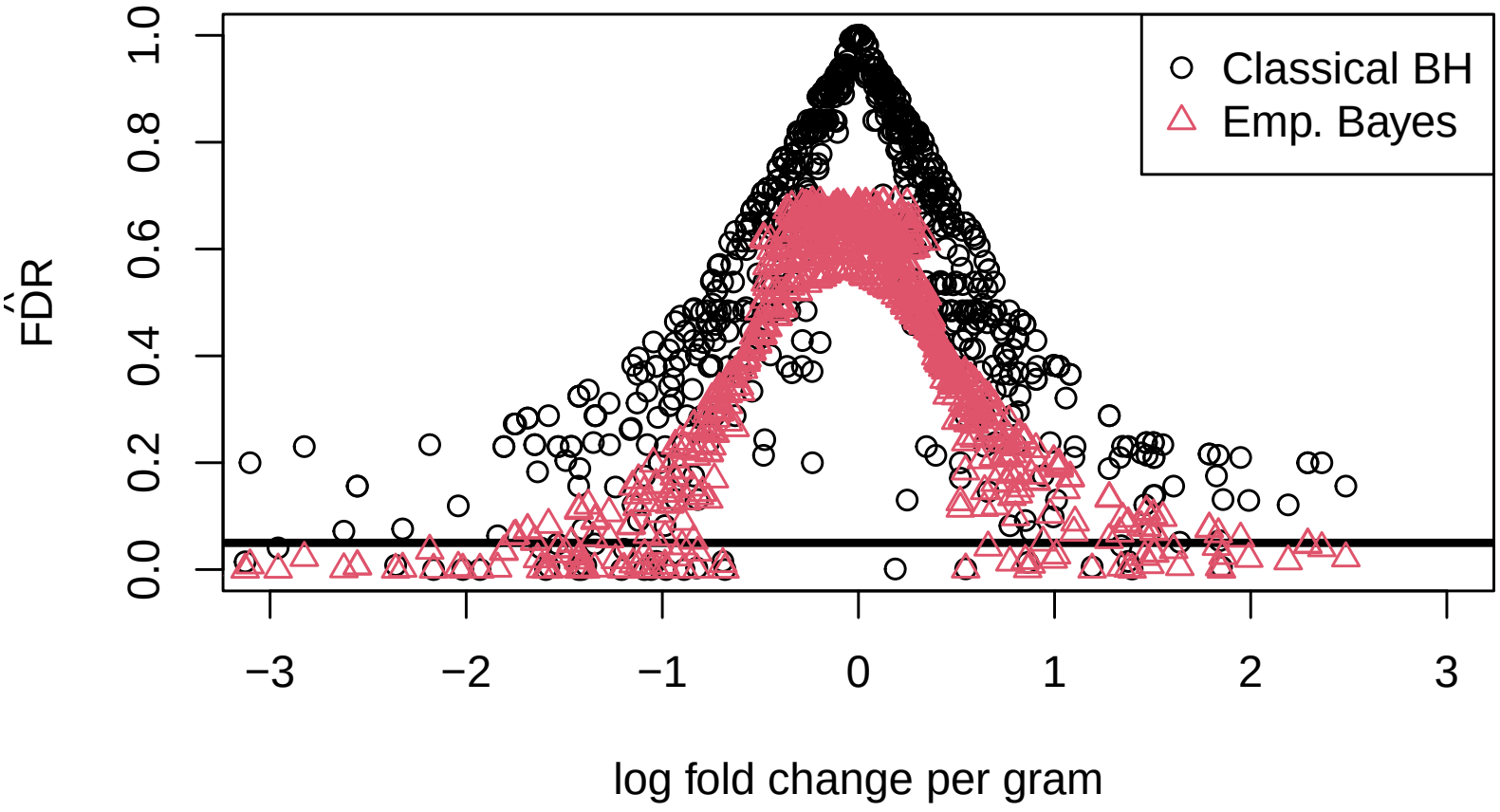
- Me: "Here's 34 discoveries using $\widehat{FDR}(.008) \leq .05$ with estimated log fold changes for assessing **practical significance**"



- Collaborator: "Great! But I'm not interested in all of these 'negligible associations'. Why can't I discover the stronger ones instead?!>!"

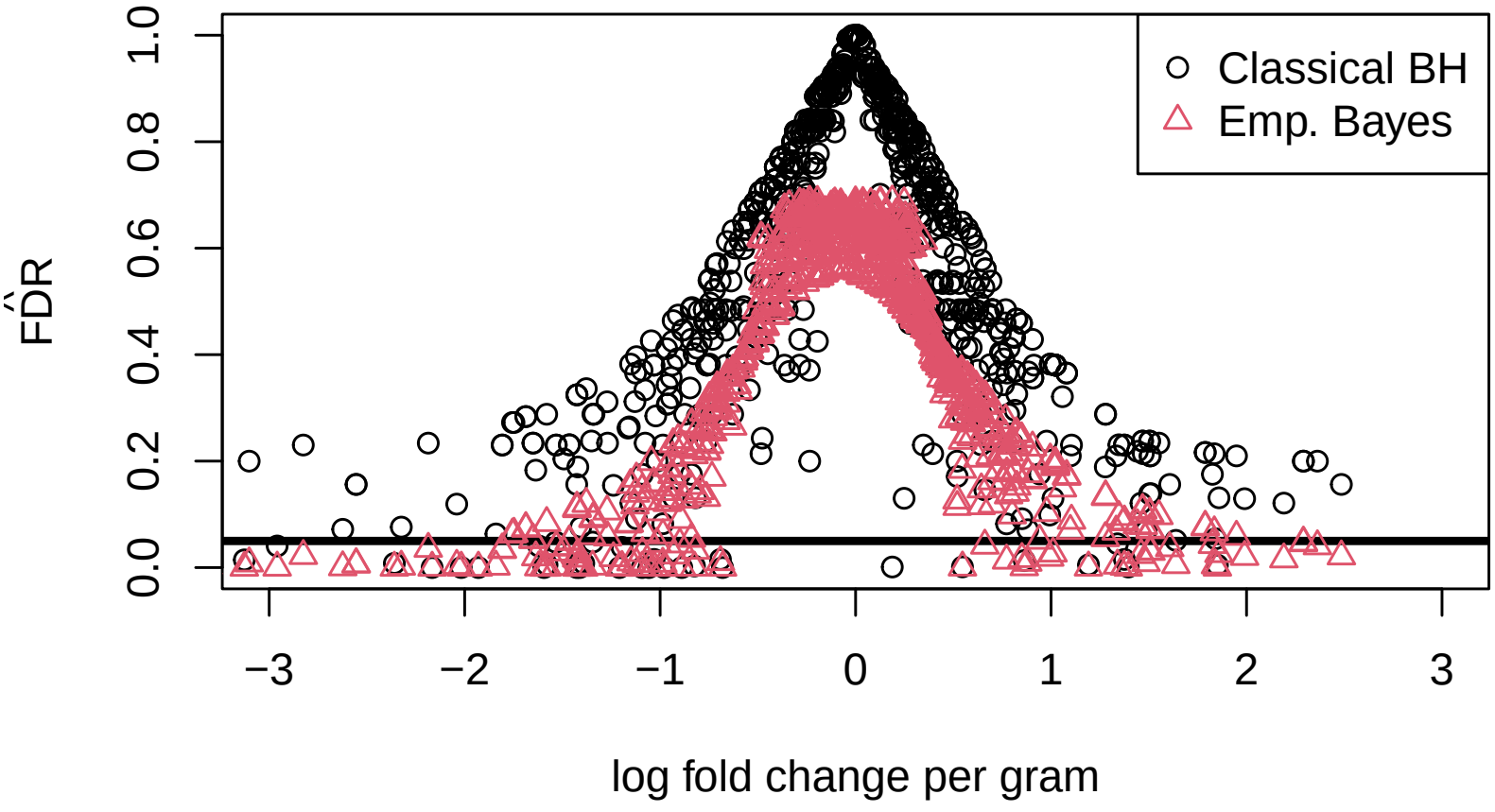
Empirical Bayes Procedure: Habiger, Watts, Anderson (**2017**)

- Me: "Ok. Now we have 97 (instead of 34) discoveries and they're all non-negligible associations!"



Empirical Bayes Procedure: Habiger, Watts, Anderson (**2017**)

- Me: "Ok. Now we have 97 (instead of 34) discoveries and they're all non-negligible associations!"



- Collaborator: "I retired three years ago."

Motivating Questions for FDR Methods Research

Exact P-values Could Be Valid

Classical BH procedure has $FDR \leq \pi_0 \alpha \leq \alpha$ (for finite M) if P-values are uniformly distributed under nulls and positively correlated.

- **Question: Suppose $X_{mn} \sim$ Pois. Are p-values from glm(..., family="Poisson") valid?**

Exact P-values Could Be Valid

Classical BH procedure has $FDR \leq \pi_0 \alpha \leq \alpha$ (for finite M) if P-values are uniformly distributed under nulls and positively correlated.

- **Question: Suppose $X_{mn} \sim$ Pois. Are p-values from glm(..., family="Poisson") valid?**
- Solution: Use **conditional exact test p-value** via
  1. Find **sufficient statistic** under log-linear model: $T_m = \sum_n X_{mn} \times \text{biomass}_n$
  2. Condition on **ancillary statistic** $S_m = \sum_n X_{mn}$
  3. Use **MC integration** to get estimate exact p-value

$$P(t_m) = Pr_0(T_m \geq t_m | S_m = s_m)$$

Weighted Adaptive BH procedures: Hetergoeneity in Multiple Testing

- Observe FDR analysis is conditional upon $S_1, S_2, .., S_M$.
- **Did the BH procedure exploit observable heterogeneity?**

Weighted Adaptive BH procedures: Hetergoeneity in Multiple Testing

- Observe FDR analysis is conditional upon $S_1, S_2, .., S_M$.
- **Did the BH procedure exploit observable heterogeneity?**
- Solution: Weighted FDR and weighted adaptive FDR procedures
  1. Compute weighted p-values: $Q_m = P_m / w_m$
  2. Apply BH procedure (or adaptive BH procedure) to Q_ms

Weighted Adaptive BH procedures: Hetergoeneity in Multiple Testing

- Observe FDR analysis is conditional upon $S_1, S_2, .., S_M$.
- **Did the BH procedure exploit observable heterogeneity?**
- Solution: Weighted FDR and weighted adaptive FDR procedures
  1. Compute weighted p-values: $Q_m = P_m / w_m$
  2. Apply BH procedure (or adaptive BH procedure) to Q_ms
- **Question: How to compute optimal weights**?

Weighted Adaptive BH procedures: Hetergoeneity in Multiple Testing

- Observe FDR analysis is conditional upon $S_1, S_2, .., S_M$.
- **Did the BH procedure exploit observable heterogeneity?**
- Solution: Weighted FDR and weighted adaptive FDR procedures
  1. Compute weighted p-values: $Q_m = P_m / w_m$
  2. Apply BH procedure (or adaptive BH procedure) to Q_ms
- **Question: How to compute optimal weights**?
- They are "complex" but procedures are robust. So weights can be based on
  - S_m or any other covariate
  - standard errors in imaging
  - training data in multi-omics (incomplete data is used as training)
- Key tools:
  - **Optional Stopping** for finite sample proofs
  - **Glivenko Cantelli** for asymptotics
  - Clever collaborators!

Discreteness Effect

- **Is $P(t_m)$ uniformly distributed under H_m?**

Discreteness Effect

- **Is $P(t_m)$ uniformly distributed under H_m?**
- General P-value: $P(t_m, u_m) = \Pr_0(T_m > t_m) + u_m \Pr_0(T_m = t_m)$
  - $P(t_m, 1)$ is the usual natural (valid) p-value
  - $P(t_m, 0.5)$ is the mid-p-value
  - $P(t_m, U_m)$ is randomized (valid and uniform)

Discreteness Effect

- **Is $P(t_m)$ uniformly distributed under H_m?**
- General P-value: $P(t_m, u_m) = \Pr_0(T_m > t_m) + u_m \Pr_0(T_m = t_m)$
  - $P(t_m, 1)$ is the usual natural (valid) p-value
  - $P(t_m, 0.5)$ is the mid-p-value
  - $P(t_m, U_m)$ is randomized (valid and uniform)
- **Question: Generate or specify u?**
  - Randomized p-value immediately fixes most issues for most procedures
  - Its a bias vs. variance question that cannot be ignored in multiple testing
  - Less bias/variance is possible!
- Key tools:
  - **Decision theory** framework
  - **Law of iterated expectation**
  - **Sufficiency principle and NP lemma**

Empirical Bayes Details

Sun and Cai (2007): Procedures of the form $\widehat{LFDR}_m < t$ can be asymptotically optimal and control FDR if $\widehat{LFDR}_m \xrightarrow{p} LFDR_m$

- Optimality intuition:

$$LDR_m = \Pr(H_m|X_m = x_m) = \frac{\Pr(H_m)f_0(x_m)}{\Pr(H_m)f_0(x_m) + [1 - \Pr(H_m)]f_1(x_m)} = \frac{1}{1 + \frac{1-\Pr(H_m)}{\Pr(H_m)} \frac{f_1(x_m)}{f_0(x_m)}}$$

Empirical Bayes Details

Sun and Cai (2007): Procedures of the form $\widehat{LFDR}_m < t$ can be asymptotically optimal and control FDR if $\widehat{LFDR}_m \xrightarrow{p} LFDR_m$

- Optimality intuition:

$$LDR_m = \Pr(H_m|X_m = x_m) = \frac{\Pr(H_m)f_0(x_m)}{\Pr(H_m)f_0(x_m) + [1 - \Pr(H_m)]f_1(x_m)} = \frac{1}{1 + \frac{1-\Pr(H_m)}{\Pr(H_m)} \frac{f_1(x_m)}{f_0(x_m)}}$$

- **Question: How can we get $\widehat{LFDR}_m \xrightarrow{p} LFDR_m$?**

Empirical Bayes Details

Sun and Cai (2007): Procedures of the form $\widehat{LFDR}_m < t$ can be asymptotically optimal and control FDR if $\widehat{LFDR}_m \xrightarrow{p} LFDR_m$

- Optimality intuition:

$$LDR_m = \Pr(H_m|X_m = x_m) = \frac{\Pr(H_m)f_0(x_m)}{\Pr(H_m)f_0(x_m) + [1 - \Pr(H_m)]f_1(x_m)} = \frac{1}{1 + \frac{1-\Pr(H_m)}{\Pr(H_m)} \frac{f_1(x_m)}{f_0(x_m)}}$$

- **Question: How can we get $\widehat{LFDR}_m \xrightarrow{p} LFDR_m$?**
- Wheat microbiome example
  - Model: Finite mixture of k multinomial distributions.
  - Estimation: **Maximum likelihood estimates** via **EM algorithm**

Advice and Questions fo Aspiring FDR Methods Researchers

- Question: "Let $T_1, T_2, ..., T_M$ be test statistics or p-values based on data $X_1, X_2, ..., X_M$". **How do you already know you can improve?**

Advice and Questions fo Aspiring FDR Methods Researchers

- Question: "Let $T_1, T_2, ..., T_M$ be test statistics or p-values based on data $X_1, X_2, ..., X_M$". **How do you already know you can improve?**
- Hint: Recall $LDR_m = \Pr(H_m|X_m = x_m) = ... = \frac{1}{1+\frac{1-\Pr(H_m)}{\Pr(H_m)}\frac{f_1(x_m)}{f_0(x_m)}}$

Advice and Questions fo Aspiring FDR Methods Researchers

- Question: "Let $T_1, T_2, ..., T_M$ be test statistics or p-values based on data $X_1, X_2, ..., X_M$". **How do you already know you can improve?**
- Hint: Recall $LDR_m = \Pr(H_m|X_m = x_m) = ... = \frac{1}{1+\frac{1-\Pr(H_m)}{\Pr(H_m)}\frac{f_1(x_m)}{f_0(x_m)}}$
- **Question: Is** $\frac{f_1(T(x_m))}{f_0(T(x_m))} = \frac{f_1(x_m)}{f_0(x_m)}$**?**

Advice and Questions fo Aspiring FDR Methods Researchers

- Question: "Let $T_1, T_2, ..., T_M$ be test statistics or p-values based on data $X_1, X_2, ..., X_M$". **How do you already know you can improve?**
- Hint: Recall $LDR_m = \Pr(H_m | X_m = x_m) = ... = \frac{1}{1 + \frac{1-\Pr(H_m)}{\Pr(H_m)} \frac{f_1(x_m)}{f_0(x_m)}}$
- **Question: Is** $\frac{f_1(T(x_m))}{f_0(T(x_m))} = \frac{f_1(x_m)}{f_0(x_m)}$**?**
- Next steps:
  1. Find a data set that has not yet been reduced to p-values and do the simplest possible FDR analysis to get started
  2. Verify that the BH procedure is an empirical Bayes procedure.
  3. Read one earlier LFDR paper with asymptotic analysis and one classical FDR paper with martingale proof carefully

Example: Future Empirical Bayes Methods for Wheat Microbiome

- Data: "Rhizosophere-Mediated Effects" has
  - Multivariate continuous/discrete outcomes for 36 wheat plants: number of kernels, shoot weight, etc.
  - Exposure - Feek stage 6 vs Feek stage 9?
  - Compositional (count data) Mediators - taxa in the rhizosphere like before
  - repeated measures (several sampling depths)
  - data for several soil types and several varieties
- Basic question: Which mediators are found in multiple soil types / varieties
- First step: Get some mediation analysis for one soil sample, one variety, one continuous outcome and one meditator

Overtime

Other research interests

- post selective FDR in replicability analysis: Is $P < .05$ or $P < .005$ really a reasonable selection model for published data?

- HD classification / discriminant analysis with discrete data: Can a "simple" linear discriminant function
  - allow for zero inflation and overdispersion,
  - address overfitting and perform variable selection, AND
  - avoid iterative procedures

Pictures