Detecting Replicating Signals using Adaptive Filtering Procedures with the Application in High-throughput Experiments

Jingshu Wang

University of Pennsylvania

Student Seminar, 2018

Joint work with Weijie Su (Upenn), Chiara Sabatti (Stanford) and Art B. Owen (Stanford)
The cornerstone of science

One of the core principles of the scientific process is that other scientists are able to repeat your experiment and either confirm or refute your results.

This is referred to as reproducibility or replication.

(Figure from https://thenib.com/repeat-after-me)
The replication crisis

(Cover of Economist, Oct. 2013)
The replication crisis

- Publication bias

(Cover of Economist, Oct. 2013)
The replication crisis

- Publication bias
- p-hacking

(Cover of Economist, Oct. 2013)
The replication crisis

- Publication bias
- p-hacking
- Idiosyncratic aspects of single large-scale experiments

(Cover of Economist, Oct. 2013)
## Replicability analysis

<table>
<thead>
<tr>
<th>Case</th>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
<th>Study 4</th>
<th>Study 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$10^{-200}$</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>B</td>
<td>$10^{-10}$</td>
<td>$10^{-9}$</td>
<td>$10^{-8}$</td>
<td>$10^{-7}$</td>
<td>$10^{-6}$</td>
</tr>
</tbody>
</table>

**Table:** Individual p-values in each study
Replicability analysis

<table>
<thead>
<tr>
<th>Case</th>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
<th>Study 4</th>
<th>Study 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$10^{-200}$</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>B</td>
<td>$10^{-10}$</td>
<td>$10^{-9}$</td>
<td>$10^{-8}$</td>
<td>$10^{-7}$</td>
<td>$10^{-6}$</td>
</tr>
</tbody>
</table>

**Table:** Individual p-values in each study

- Meta-analysis can not guarantee finding replicating signals
### Replicability analysis

<table>
<thead>
<tr>
<th>Case</th>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
<th>Study 4</th>
<th>Study 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$10^{-200}$</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>B</td>
<td>$10^{-10}$</td>
<td>$10^{-9}$</td>
<td>$10^{-8}$</td>
<td>$10^{-7}$</td>
<td>$10^{-6}$</td>
</tr>
</tbody>
</table>

**Table:** Individual p-values in each study

- Meta-analysis can not guarantee finding replicating signals
- Need a framework to evaluate the consistency across studies
Replicability analysis

<table>
<thead>
<tr>
<th>Case</th>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
<th>Study 4</th>
<th>Study 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$10^{-200}$</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>B</td>
<td>$10^{-10}$</td>
<td>$10^{-9}$</td>
<td>$10^{-8}$</td>
<td>$10^{-7}$</td>
<td>$10^{-6}$</td>
</tr>
</tbody>
</table>

Table: Individual p-values in each study

- Meta-analysis can not guarantee finding replicating signals
- Need a framework to evaluate the consistency across studies
- Also need to properly accounts for experimental and sampling heterogeneity
### Replicability analysis

<table>
<thead>
<tr>
<th>Case</th>
<th>Study 1</th>
<th>Study 2</th>
<th>Study 3</th>
<th>Study 4</th>
<th>Study 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>$10^{-200}$</td>
<td>0.4</td>
<td>0.5</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>B</td>
<td>$10^{-10}$</td>
<td>$10^{-9}$</td>
<td>$10^{-8}$</td>
<td>$10^{-7}$</td>
<td>$10^{-6}$</td>
</tr>
</tbody>
</table>

**Table:** Individual p-values in each study

- Meta-analysis can not guarantee finding replicating signals
- Need a framework to evaluate the consistency across studies
- Also need to properly accounts for experimental and sampling heterogeneity
- Can extend to finding common genetic signals across similar disease, tissues, cell types ...
A framework: partial conjunction (PC) hypothesis testing

A partial conjunction hypothesis $H^{r/n}_r (r \geq 2)$ is testing

$H^{r/n}_0$: less than $r$ out of $n$ hypotheses are nonnull

$H^{r/n}_1$: at least $r$ out of $n$ hypotheses are nonnull
Test for a PC hypothesis

How to test for $H_0^{2/3}$?

$H_{01}$  $H_{02}$  $H_{03}$

$H_0^{2/3}$ is false $\iff$

Combined p-value for $H_0^{2/3}$

$f(p_1, p_2, p_3) = \max \{ g_{12}(p_1, p_2), g_{13}(p_1, p_3), g_{23}(p_2, p_3) \}$
Test for a PC hypothesis

How to test for $H_0^{2/3}$?

$H_0^{2/3}$ is false $\iff H_{01} \cap H_{02}$ is false &

$H_{01} \cap H_{02}$ is false &
Test for a PC hypothesis

How to test for $H_0^{2/3}$?

$H_0^{2/3}$ is false $\iff$ $H_{01} \cap H_{02}$ is false & $H_{01} \cap H_{03}$ is false &
Test for a PC hypothesis

How to test for $H_0^{2/3}$?

$H_0^{2/3}$ is false $\iff$ $H_0^{2/3}$ is false & $H_0^{2/3}$ is false & $H_0^{2/3}$ is false
Test for a PC hypothesis

How to test for $H_0^{2/3}$?

$H_0^{2/3}$ is false $\iff H_{01} \cap H_{02}$ is false & $H_{01} \cap H_{03}$ is false & $H_{02} \cap H_{03}$ is false

Combined p-value for $H_0^{2/3}$

$$f(p_1, p_2, p_3) = \max\{g_{12}(p_1, p_2), g_{13}(p_1, p_3), g_{23}(p_2, p_3)\}$$
**Test for a PC hypothesis**

A combined p-value for the PC null has the form:

\[
f(p) = f(p_1, p_2, \cdots, p_n) = \max_{u \subseteq 1:n} g_u(p_u)
\]

where \( g_u(p_u) \) is a combined p-value for the global null hypothesis on \( u \).
A combined p-value for the PC null has the form:

\[ f(p) = f(p_1, p_2, \cdots, p_n) = \max_{u \subseteq \{1:n\}, |u| = n-r+1} g_u(p_u) \]

where \( g_u(p_u) \) is a combined p-value for the global null hypothesis on \( u \).

The symmetric form (Benjamini and Heller, 2008)

If \( g_u \equiv g \) for some symmetric and non-decreasing function \( g \), then

\[ f(p) = g(p(r), p(r+1), \cdots, p(n)) \]

For example, \( g \) can be the Simes/Bonferroni combined p-value, or from Fisher’s method.
Multiple testing of PC Hypotheses in high-throughput experiments

Each row a gene / SNP
Multiple testing of PC Hypotheses in high-throughput experiments

- Each row a gene / SNP
- For Each row testing for PC null hypothesis $H_{0j}^{r/n}$
Multiple testing of PC Hypotheses in high-throughput experiments

- Each row a gene / SNP
- For Each row testing for PC null hypothesis $H_{0j}^{r/n}$
- $M$ can be very large
Multiple testing of PC Hypotheses in high-throughput experiments

- Each row a gene / SNP
- For Each row testing for PC null hypothesis $H_{0j}^{r/n}$
- $M$ can be very large
- Simultaneous error control:
Multiple testing of PC Hypotheses in high-throughput experiments

- Each row a gene / SNP
- For Each row testing for PC null hypothesis $H_{0j}^{r/n}$
- $M$ can be very large
- Simultaneous error control:
  - FWER: $\mathbb{P}[V \geq 1]$
Multiple testing of PC Hypotheses in high-throughput experiments

<table>
<thead>
<tr>
<th>Study 1</th>
<th>Study 2</th>
<th>⋯</th>
<th>Study n</th>
</tr>
</thead>
<tbody>
<tr>
<td>P_{11}</td>
<td>P_{12}</td>
<td>⋯</td>
<td>P_{1n}</td>
</tr>
<tr>
<td>P_{21}</td>
<td>P_{22}</td>
<td>⋯</td>
<td>P_{2n}</td>
</tr>
<tr>
<td>P_{31}</td>
<td>P_{32}</td>
<td>⋯</td>
<td>P_{3n}</td>
</tr>
<tr>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
<td>⋮</td>
</tr>
<tr>
<td>P_{M1}</td>
<td>P_{M2}</td>
<td>⋯</td>
<td>P_{Mn}</td>
</tr>
</tbody>
</table>

- Each row a gene / SNP
- For Each row testing for PC null hypothesis $H_{0j}^{r/n}$
- $M$ can be very large
- Simultaneous error control:
  - **FWER**: $\mathbb{P}[V \geq 1]$
  - **FDR**: $\mathbb{E}\left[\frac{V}{\max(R, 1)}\right]$
Multiple testing of PC Hypotheses in high-throughput experiments

Each row a gene / SNP

For Each row testing for PC null hypothesis $H_{0j}^{r/n}$

$M$ can be very large

Simultaneous error control:

- **FWER**: $\mathbb{P}[V \geq 1]$
- **FDR**: $\mathbb{E}\left[\frac{V}{\max(R, 1)}\right]$
- **PFER**: $\mathbb{E}[V]$
Direct multiple testing correction

- Compute a p-value $p_{j/n}^r$ for each PC null $H_{0j}^{r/n}$ ($j = 1, \cdots, M$)
Direct multiple testing correction

- Compute a p-value $p_{j}^{r/n}$ for each PC null $H_{0j}^{r/n}$ ($j = 1, \cdots, M$)

- Apply standard multiple testing procedure on $(p_{1}^{r/n}, \cdots, p_{M}^{r/n})$
Direct multiple testing correction is too conservative

FDR control

M = 10000

BH−Pr r/n
BH−P^B
BH−P^F
BH−P^S
repfdr
Adafilter BH

Jingshu Wang (Statistics)
Why is it conservative?

A simple example for $n = r = 2$

- The PC p-values: $p_{j}^{2/2} = p(2)_j = \max(p_{1j}, p_{2j})$
Why is it conservative?

A simple example for \( n = r = 2 \)

- The PC p-values: \( p_{2j}^{2/2} = p(2)_j = \max(p_{1j}, p_{2j}) \)
- use Bonferroni to control FWER: reject \( j \) if \( p(2)_j \leq \alpha/M \)
Why is it conservative?

A simple example for \( n = r = 2 \)

- The PC p-values: \( p_j^{2/2} = p(2)_j = \max(p_{1j}, p_{2j}) \)
- use Bonferroni to control FWER: reject \( j \) if \( p(2)_j \leq \alpha/M \)
- Assume that the fractions of rows having null and nonnull hypotheses as \((0, 0), (0, 1), (1, 0)\) and \((1, 1)\) are \( \pi_{00}, \pi_{01}, \pi_{10} \) and \( \pi_{11} \)
Why is it conservative?

A simple example for $n = r = 2$

- The PC p-values: $p_{j}^{2/2} = p_{(2)j} = \max(p_{1j}, p_{2j})$
- use Bonferroni to control FWER: reject $j$ if $p_{(2)j} \leq \alpha/M$
- Assume that the fractions of rows having null and nonnull hypotheses as $(0, 0), (0, 1), (1, 0)$ and $(1, 1)$ are $\pi_{00}, \pi_{01}, \pi_{10}$ and $\pi_{11}$

- The actual FWER:

$$\mathbb{P} [V \geq 1] \leq \mathbb{E} [V] = \sum_{j=1}^{M} \mathbb{P} [p_{(2)j} \leq \alpha/M]$$

$$\leq \pi_{00} \frac{\alpha^2}{M} + (\pi_{01} + \pi_{10}) \alpha \ll \alpha$$

if $\pi_{00} \approx 1$ as in most problems.
Empirical Bayesian solutions

- Estimate the fractions $\pi_{ij}$
- Calculate and control Bayesian FDR

(Flutre et al. 2013, Heller and Yekutieli 2014, Zhao 2017, Urbut et al. 2017 ···)
Empirical Bayesian solutions

- Estimate the fractions $\pi_{ij}$
- Calculate and control Bayesian FDR

(Flutre et. al. 2013, Heller and Yekutieli 2014, Zhao 2017, Urbut et. al. 2017 · · · )

Difficulty

- When $n$ is large, very hard to estimate all $\pi_{ij}$
- Can be time consuming
Empirical Bayesian solutions

- Estimate the fractions $\pi_{ij}$
- Calculate and control Bayesian FDR

(Flutre et. al. 2013, Heller and Yekutieli 2014, Zhao 2017, Urbut et. al. 2017 · · ·)

Difficulty

- When $n$ is large, very hard to estimate all $\pi_{ij}$
- Can be time consuming

Our solution: a frequentist approach avoid estimating $\pi_{ij}$
Idea for our adaptive-filtering procedure

**Why conservative?:**
PC null hypothesis is composite $\Rightarrow$

$$P \left[ p_j^{r/n} \leq \gamma \right] \leq \gamma$$

too loose under some null scenario
Idea for our adaptive-filtering procedure

Why conservative?:
PC null hypothesis is composite ⇒
\[ \Pr \left[ p_j^{r/n} \leq \gamma \right] \leq \gamma \]
too loose under some null scenario

For \( n = r = 2 \), \( H_{0/2}^2 \) contains three scenarios
Idea for our adaptive-filtering procedure

- Why conservative?:
  - PC null hypothesis is composite ⇒
    \[ P \left( p_j^{r/n} \leq \gamma \right) \leq \gamma \]
    is too loose under some null scenario
  - For \( n = r = 2 \), \( H_0^{2/2} \) contains three scenarios
    - \((0, 0)\): Type-I error \( \gamma^2 \)
Why conservative?:
PC null hypothesis is composite ⇒
\[ P \left[ p_{r/n} \leq \gamma \right] \leq \gamma \]
too loose under some null scenario
For \( n = r = 2 \), \( H_0^{2/2} \) contains three scenarios
- \((0, 0)\): Type-I error \( \gamma^2 \)
- \((0, 1)\): Type-I error \( \leq \gamma \)
Idea for our adaptive-filtering procedure

- Why conservative?:
  PC null hypothesis is composite ⇒
  \[ \mathbb{P}\left[p_j^{r/n} \leq \gamma\right] \leq \gamma \]
  too loose under some null scenario

- For \( n = r = 2 \), \( H_0^{2/2} \) contains three scenarios
  - \((0, 0)\): Type-I error \( \gamma^2 \)
  - \((0, 1)\): Type-I error \( \leq \gamma \)
  - \((1, 0)\): Type-I error \( \leq \gamma \)
Intuitive idea of AdaFilter

- Idea: find filtering regions
  \[ A_\gamma \in \mathbb{R}^n: \]

\[
\mathbb{P} \left[ p_{j/n}^r \leq \gamma \mid p_j \in A_\gamma \right] \leq \gamma
\]

still valid under \( H_{0j}^{r/n} \)
Intuitive idea of AdaFilter

• Idea: find filtering regions
  \[ A_{\gamma} \in \mathbb{R}^n : \]
  \[ \mathbb{P} \left[ p_{j}^{r/n} \leq \gamma \mid p_{j} \in A_{\gamma} \right] \leq \gamma \]

  still valid under \( H_{0j}^{r/n} \)

• For \( n = r = 2 \), under \( H_{0j}^{2/2} \):
  \[ \mathbb{P} \left[ p(2)_j \leq \gamma \mid p(1)_j \leq \gamma \right] \leq \gamma \]
Illustration for \( r = n = 2 \)

How to estimate \( \mathbb{E}[V] \)?

- Bonferroni: \( \gamma \cdot M \)

A tighter estimate:

\[
\gamma \cdot M \sum_{j=1}^{M} \frac{1}{p(j)} \leq \gamma \leq \sum_{j=1}^{M} \frac{1}{p(j)}
\]
Illustration for $r = n = 2$

How to estimate $E[V]$?

- Bonferroni: $\gamma \cdot M$
- A tighter estimate:

$$\gamma \cdot \left( \sum_{j=1}^{M} 1_{p(1)j \leq \gamma} \right)$$
Illustration for $r = n = 2$

How to estimate $E[V]$?

- Bonferroni: $\gamma \cdot M$
- A tighter estimate:

$$
\gamma \cdot \left( \sum_{j=1}^{M} 1_{p(1)_j \leq \gamma} \right)
$$
How to estimate $\mathbb{E}[V]$?

- Bonferroni: $\gamma \cdot M$
- A tighter estimate:

$$\gamma \cdot \left( \sum_{j=1}^{M} 1_{p(1)j \leq \gamma} \right)$$

How to estimate $\mathbb{E}\left[ \frac{V}{\max(R, 1)} \right]$?

- BH procedure:

$$\frac{\gamma \cdot M}{\sum_{j=1}^{M} 1_{p(2)j \leq \gamma}}$$
Illustration for $r = n = 2$

How to estimate $\mathbb{E}[V]$?
- Bonferroni: $\gamma \cdot M$
- A tighter estimate:

$$\gamma \cdot \left( \sum_{j=1}^{M} 1_{p(1)_j \leq \gamma} \right)$$

How to estimate $\mathbb{E} \left[ \frac{V}{\max(R, 1)} \right]$?
- BH procedure:

$$\frac{\gamma \cdot M}{\sum_{j=1}^{M} 1_{p(2)_j \leq \gamma}}$$

- A tighter estimate:

$$\frac{\gamma \cdot \left( \sum_{j=1}^{M} 1_{p(1)_j \leq \gamma} \right)}{\sum_{j=1}^{M} 1_{p(2)_j \leq \gamma}}$$
Illustration for $n = r = 2$: adaptively choosing $\gamma$

- Control FWER at level $\alpha$: find maximum $\gamma$ such that

$$\gamma \cdot \left( \sum_{j=1}^{M} 1_{p_{(1),j} \leq \gamma} \right) \leq \alpha$$

- Control FDR at level $q$: find maximum $\gamma$ such that

$$\frac{\gamma \cdot \left( \sum_{j=1}^{M} 1_{p_{(1),j} \leq \gamma} \right)}{\sum_{j=1}^{M} 1_{p_{(2),j} \leq \gamma}} \leq q$$
Illustration for \( n = r = 2 \): adaptively choosing \( \gamma \)

- Control FWER at level \( \alpha \): find maximum \( \gamma \) such that

\[
\gamma \cdot \left( \sum_{j=1}^{M} 1_{p(1)_j \leq \gamma} \right) \leq \alpha
\]

- Control FDR at level \( q \): find maximum \( \gamma \) such that

\[
\frac{\gamma \cdot \left( \sum_{j=1}^{M} 1_{p(1)_j \leq \gamma} \right)}{\sum_{j=1}^{M} 1_{p(2)_j \leq \gamma}} \leq q
\]
Illustration for $n = r = 2$: adaptively choosing $\gamma$

- Control FWER at level $\alpha$: find maximum $\gamma$ such that

$$\gamma \cdot \left( \sum_{j=1}^{M} 1_{p(1)_j \leq \gamma} \right) \leq \alpha$$

- Control FDR at level $q$: find maximum $\gamma$ such that

$$\gamma \cdot \frac{\left( \sum_{j=1}^{M} 1_{p(1)_j \leq \gamma} \right)}{\sum_{j=1}^{M} 1_{p(2)_j \leq \gamma}} \leq q$$

- Reject $H_{0j}^{r/n}$ if $p(2)_j \leq \gamma$
Formal definition for general $n$ and $r$

Define

Selection p-value: $S_j = (n - r + 1)p(r)_j$

Filtering p-value: $F_j = (n - r + 1)p(r-1)_j$
Formal definition for general $n$ and $r$

- Define

  Selection p-value: $S_j = (n - r + 1)p_{(r)j}$
  Filtering p-value: $F_j = (n - r + 1)p_{(r-1)j}$

- Filtering property:

  $\mathbb{P} [S_j \leq \gamma | F_j \leq \gamma] \leq \gamma$

  satisfied under $H_{0j}^{r/n}$ and independence
Formal definition for general $n$ and $r$

- Control for FWER at level $\alpha$ (AdaFilter Bonferroni):

$$
\gamma_0 = \max \left\{ \gamma \in \{ \alpha, \frac{\alpha}{2}, \ldots, \frac{\alpha}{M} \} : \gamma \cdot \frac{1}{M} \sum_{j=1}^{M} 1_{f_j} \leq \gamma \leq \alpha \right\}.
$$
Formal definition for general \( n \) and \( r \)

- Control for FWER at level \( \alpha \) (AdaFilter Bonferroni):

\[
\gamma_0 = \max \left\{ \gamma \in \{ \alpha, \frac{\alpha}{2}, \cdots, \frac{\alpha}{M} \} : \gamma \cdot \sum_{j=1}^{M} 1_{f_j \leq \gamma} \leq \alpha \right\}.
\]

- Control for FDR at level \( q \) (AdaFilter BH):

\[
\gamma_0 = \max \left\{ \gamma \in \mathcal{I}_{q,M} : \gamma \cdot \sum_{j=1}^{M} 1_{f_j \leq \gamma} \leq q \cdot \sum_{j=1}^{M} 1_{s_j \leq \gamma} \right\}
\]

where \( \mathcal{I}_{q,M} = \left\{ \frac{k}{m} \cdot q : k \in 1:M, m \in 1:M, k \leq m \right\} \)
Validity

AdaFilter Bonferroni for the null hypotheses \( \{ H_{0j}^{r/n} : j = 1, 2, \ldots, M \} \) controls FWER(PFER) at the nominal level \( \alpha \).

- Need independence across studies
Theoretical Properties

Validity

AdaFilter Bonferroni for the null hypotheses \( \{H_{0j/n}^r : j = 1, 2, \ldots, M\} \)
controls FWER(PFER) at the nominal level \( \alpha \).

- Need independence across studies
- FDR control of AdaFilter BH is still a conjecture
Validity

AdaFilter Bonferroni for the null hypotheses \( \{ H_0^{r/n} : j = 1, 2, \ldots, M \} \) controls FWER(PFER) at the nominal level \( \alpha \).

- Need independence across studies
- FDR control of AdaFilter BH is still a conjecture
- Both FWER and FDR control is very robust to dependence within each study in simulation
Theoretical Properties

- Researcher A finds significant genes for each study separately and takes the union of discovered genes.

Table: Toy example of p-values

<table>
<thead>
<tr>
<th>j</th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.9</td>
</tr>
</tbody>
</table>

None can be rejected by researcher A at $\alpha = 0.05$. $j=1$ can be rejected by researcher B using AdaFilter.
Theoretical Properties

- Researcher A finds significant genes for each study separately and takes the union of discovered genes.
- Researcher B claims a significant gene only when it is rejected by AdaFilter with $r = 2$.

Table: Toy example of $p$-values

<table>
<thead>
<tr>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_j$</td>
<td>$S_j$</td>
</tr>
<tr>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>0.5</td>
<td>0.9</td>
</tr>
<tr>
<td>0.5</td>
<td>0.9</td>
</tr>
</tbody>
</table>

None can be rejected by researcher A at $\alpha = 0.05$. Study 1 can be rejected by researcher B using AdaFilter with $r = 2$. 

Jingshu Wang (Statistics)
Theoretical Properties

- Researcher A finds significant genes for each study separately and takes the union of discovered genes.
- Researcher B claims a significant gene only when it is rejected by AdaFilter with $r = 2$. 

Table: Toy example of p-values

<table>
<thead>
<tr>
<th></th>
<th>Study 1</th>
<th>Study 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$j_1$</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>$j_2$</td>
<td>0.5</td>
<td>0.9</td>
</tr>
</tbody>
</table>

None can be rejected by researcher A at $\alpha = 0.05$. $j_1$ can be rejected by researcher B using AdaFilter.
Theoretical Properties

- Researcher A finds significant genes for each study separately and takes the union of discovered genes.
- Researcher B claims a significant gene only when it is rejected by AdaFilter with $r = 2$.

Who will get more discoveries?

<table>
<thead>
<tr>
<th>Study</th>
<th>$p$-value</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.04</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.9</td>
</tr>
</tbody>
</table>

None can be rejected by researcher A at $\alpha = 0.05$. $j = 1$ can be rejected by researcher B using AdaFilter.
Theoretical Properties

- Researcher A finds significant genes for each study separately and takes the union of discovered genes.
- Researcher B claims a significant gene only when it is rejected by AdaFilter with $r = 2$.

Who will get more discoveries?

<table>
<thead>
<tr>
<th>$j$</th>
<th>Study 1</th>
<th>Study 2</th>
<th>$F_j$</th>
<th>$S_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.9</td>
<td>0.5</td>
<td>0.9</td>
</tr>
</tbody>
</table>

None can be rejected by researcher A at $\alpha = 0.05$. 
Theoretical Properties

- Researcher A finds significant genes for each study separately and takes the union of discovered genes.
- Researcher B claims a significant gene only when it is rejected by AdaFilter with $r = 2$.

Who will get more discoveries?

<table>
<thead>
<tr>
<th>$j$</th>
<th>Study 1</th>
<th>Study 2</th>
<th>$F_j$</th>
<th>$S_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.9</td>
<td>0.5</td>
<td>0.9</td>
</tr>
</tbody>
</table>

- None can be rejected by researcher A at $\alpha = 0.05$.
- $j = 1$ can be rejected by researcher B using AdaFilter.
Theoretical Properties

lack of ‘Complete Monotonicity’

For a given $r$, the rejection of $H_{0_j}^{r/n}$ is not monotone in $(p_{ji})_{M \times n}$
Theoretical Properties

lack of ‘Complete Monotonicity’

For a given $r$, the rejection of $H_{0j}^{r/n}$ is not monotone in $(p_{ji})_{M \times n}$

<table>
<thead>
<tr>
<th>j</th>
<th>Study 1</th>
<th>Study 2</th>
<th>$F_j$</th>
<th>$S_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.9</td>
<td>0.5</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Theoretical Properties

lack of ‘Complete Monotonicity’

For a given $r$, the rejection of $H_{0j}^{r/n}$ is not monotone in $(p_{ji})_{M \times n}$

<table>
<thead>
<tr>
<th>$j$</th>
<th>Study 1</th>
<th>Study 2</th>
<th>$F_j$</th>
<th>$S_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.9</td>
<td>0.5</td>
<td>0.9</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$j$</th>
<th>Study 1</th>
<th>Study 2</th>
<th>$F_j$</th>
<th>$S_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.01</td>
<td>0.9</td>
<td>0.01</td>
<td>0.9</td>
</tr>
</tbody>
</table>
Theoretical Properties

lack of ‘Complete Monotonicity’

For a given $r$, the rejection of $H_{0j}^{r/n}$ is not monotone in $(p_{ji})_{M \times n}$

<table>
<thead>
<tr>
<th></th>
<th>Study 1</th>
<th>Study 2</th>
<th>$F_j$</th>
<th>$S_j$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.04</td>
<td>0.04</td>
<td>0.05</td>
</tr>
<tr>
<td>2</td>
<td>0.5</td>
<td>0.9</td>
<td>0.5</td>
<td>0.9</td>
</tr>
</tbody>
</table>

None can now be rejected by AdaFilter
Simulations: Type-I error

\[ \pi_0 = 0.8, \pi_0 = 0.98 \]

\[ b = 100, b = 1000 \]

\[ \frac{2}{2}, \frac{2}{4}, \frac{4}{4}, \frac{2}{8}, \frac{4}{8}, \frac{8}{8} \]

\[ 0.0, 0.2, 0.4, 0.6 \]

**Case**

**False Discovery Proportion**

**BH−P \text{r/n}**, **BH−P \text{r/n}**, **BH−P \text{r/n}**, **repfdr**, **AdaFilter BH**

\[ M = 10000, 6 \text{ r/n combinations} \]
Simulations: Type-I error

- \( M = 10000 \), 6 r/n combinations
- \( \pi_0 \): fraction of complete null genes

\( \pi_0 = 0.8 \)

\( \pi_0 = 0.98 \)
Simulations: Type-I error

- \( M = 10000 \), 6 r/n combinations
- \( \pi_0 \): fraction of complete null genes
- test statistics for each study \( i \):
  \[ Z_i \sim \Sigma 0.5 \otimes I_{b \times b} \]
Simulations: Type-I error

- $M = 10000, \, 6 \, r/n$ combinations
- $\pi_0$: fraction of complete null genes
- test statistics for each study $i$: $Z_i \sim \Sigma_{0.5} \otimes I_{b \times b}$
- two-sided p-values
Simulations: Type-I error

- $M = 10000$, 6 r/n combinations
- $\pi_0$: fraction of complete null genes
- test statistics for each study $i$:
  - $Z_i \sim \Sigma_{0.5} \otimes I_{b \times b}$
- two-sided p-values
- Nominal FDR controlled at $\alpha = 0.05$
Simulations: Power

- $M = 10000$, 6 r/n combinations
- $\pi_0$: fraction of complete null genes
- Test statistics for each study $i$:
  \[ Z_i \sim \sum_{0.5} \otimes I_{b \times b} \]
- Two-sided p-values
Example 1: microarray experiments on Duchenne Muscular Dystrophy (DMD)

<table>
<thead>
<tr>
<th>GEO ID</th>
<th>Platform</th>
<th>Description</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>GDS 214</td>
<td>custom Affymetrix</td>
<td>4 healthy, 26 DMD</td>
<td>Muscle</td>
</tr>
<tr>
<td>GDS 563</td>
<td>Affymetrix U95A</td>
<td>11 healthy, 12 DMD</td>
<td>Quadriceps</td>
</tr>
<tr>
<td>GDS 1956</td>
<td>Affymetrix U133A</td>
<td>18 healthy, 10 DMD</td>
<td>Muscle</td>
</tr>
<tr>
<td>GDS 3027</td>
<td>Affymetrix U133A</td>
<td>14 healthy, 23 DMD</td>
<td>Quadriceps</td>
</tr>
</tbody>
</table>

- **DMD**: X-linked disorder for progressive muscle degeneration
- Four datasets share no common sample
- Goal: select bio-markers for diagnosis and treatment of DMD
Replication analysis of DMD experiments

- Individual p-values calculated using Limma and rescaled
- Multiple probes for the same gene are combined using Bonferroni
- *Study GDS 1956* is the least powerful study
## AdaFilter results

FDR is controlled at level $\alpha = 0.05$

<table>
<thead>
<tr>
<th></th>
<th>$r = 2$</th>
<th>$r = 3$</th>
<th>$r = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M$</td>
<td>13912</td>
<td>9848</td>
<td>1871</td>
</tr>
<tr>
<td># of filtered in</td>
<td>1458</td>
<td>529</td>
<td>108</td>
</tr>
<tr>
<td># of rejected</td>
<td>494</td>
<td>142</td>
<td>32</td>
</tr>
</tbody>
</table>
# AdaFilter results

FDR is controlled at level $\alpha = 0.05$

<table>
<thead>
<tr>
<th>M</th>
<th>$r = 2$</th>
<th>$r = 3$</th>
<th>$r = 4$</th>
</tr>
</thead>
<tbody>
<tr>
<td># of filtered in</td>
<td>1458</td>
<td>529</td>
<td>108</td>
</tr>
<tr>
<td># of rejected</td>
<td>494</td>
<td>142</td>
<td>32</td>
</tr>
</tbody>
</table>

Known marker genes of the rejected genes for $r = 4$

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>GDS 214</th>
<th>GDS 563</th>
<th>GDS 1956</th>
<th>GDS 3027</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYH3</td>
<td>5.47e-14</td>
<td>2.18e-69</td>
<td>3.31e-07</td>
<td>2.49e-20</td>
</tr>
<tr>
<td>MYH8</td>
<td>5.74e-06</td>
<td>9.09e-11</td>
<td>2.58e-03</td>
<td>5.16e-33</td>
</tr>
<tr>
<td>MYL5</td>
<td>8.97e-04</td>
<td>3.06e-06</td>
<td>1.87e-03</td>
<td>6.63e-08</td>
</tr>
<tr>
<td>MYL4</td>
<td>1.48e-06</td>
<td>7.94e-08</td>
<td>1.21e-02</td>
<td>2.66e-08</td>
</tr>
</tbody>
</table>
Metabolites data (Shin et. al. 2015)

- $M = 2,182,555$ markers
- 275 annotated metabolites measures
- 8 super-pathways, 71 pathways
- Test statistics are given
Correlation among the metabolites measures

Calculated Correlation Matrix

Real data example
Goal: Find genes that affect multiple super pathways
Calculate individual p-values

- Goal: Find genes that affect multiple super pathways
- \( n = 8 \): Each super-pathway \( i \) is a “study”
Calculate individual p-values

- Goal: Find genes that affect multiple super pathways
- \( n = 8 \): Each super-pathway \( i \) is a “study”
- Under the null, \( Z_{ji} \sim \mathcal{N}(0, \Sigma_i) \)
Real data example

Calculate individual p-values

- Goal: Find genes that affect multiple super pathways
- $n = 8$: Each super-pathway $i$ is a “study”
- Under the null, $Z_{ji} \sim \mathcal{N}(0, \Sigma_i)$
- The individual p-values $p_{ji}$ is calculated using a chi-square test
Comparison of power (FDR controlled at 0.05)

Metabolics GWAS

# of significant SNPs

BH\(_{-P}^{B}_{r/n}\)

BH\(_{-P}^{F}_{r/n}\)

BH\(_{-P}^{S}_{r/n}\)

repfdr

AdaFilter BH

Jingshu Wang (Statistics)
Significant SNPs mapped genes

- GCKR
- CCNT2-AS1
- SLC17A3
- SLCO1B1
- SLC2A9
- ZNF19
- SLCL2A9
- CPS1
- RGS14
- ABCC1
- CCNT2-AS1
- GCKR

Amino acid
Carbohydrate
Cofactors and vitamins
Energy
Lipid
Nucleotide
Peptide
Xenobiotics
The End