Conditionalized testing: Improving multiple testing methods for inflated p-values

Jakub Pečánka

Department of Medical Statistics and Bioinformatics
Leiden University Medical Center (LUMC)

Joint work with Jules Ellis (Radboudumc, Nijmegen) and Jelle Goeman (LUMC)

Channel Network Conference 2017, Hasselt, Belgium

April 2017
Real data example

Differential expression of mice data from Uslu (2014) shows:
Starting point

Framework

- **Goal**: simultaneously test $m$ null hypotheses $H_{01}, \ldots, H_{0m}$
  - **Multiplicity of testing** (a need to correct for it)

- **Idealized setting**: p-values of the true nulls are **standard uniform**

- Often **not the case**
  - **Inflated p-values**: values near 1 are "over-represented"

- The popular solution of ignoring the large p-values is **anti-conservative**
  - **Super-uniformity** of the p-values: both a problem and an opportunity

Terminology

Inconsistent terminology in literature: Inflated/deflated in the sense of value (our usage) vs. in the sense of distribution (e.g. in Listgartena (2010) inflated means too many small p-values)
### Motivation

#### Starting point

#### Framework

- **Goal:** simultaneously test $m$ null hypotheses $H_{01}, \ldots, H_{0m}$
  - \textit{multiplicity of testing} (a need to correct for it)

- **Idealized setting:** p-values of the true nulls are **standard uniform**

- **Often not the case**
  - \textit{Inflated p-values:} values near 1 are "over-represented"

- The popular solution of ignoring the large p-values is **anti-conservative**
  - \textit{super-uniformity} of the p-values: both a problem and an opportunity

#### Terminology

- **Inconsistent terminology in literature:** Inflated/deflated in the sense of value (**our usage**) vs. in the sense of distribution (e.g. in Listgarten (2010) *inflated* means *too many small p-values*)
Why are p-values inflated?

Example of inflated p-values

- **Interval null hypotheses:**
  - \( H_0 : \mu \leq 0 \) for location in normal model
  - \( H_0 : \mu \in (-\delta, \delta), \delta > 0 \) for location in normal model
  - \( H_0 : \text{OR} \in (0.9, 1.1) \) for risk in logistic regression model

- **Modelling issues:**
  - model misspecification (e.g. normality)
  - nuisance parameters (the way a p-value is calculated plays a crucial role)
  - overestimation of the test statistic variance (e.g. gene expression)
  - theoretical vs effective degrees of freedom (e.g. low sample size: a lack of asymptotic approximation by chi-square distribution)
Why are p-values inflated?

Example of inflated p-values

- **Interval null hypotheses:**
  - $H_0 : \mu \leq 0$ for location in normal model
  - $H_0 : \mu \in (-\delta, \delta)$, $\delta > 0$ for location in normal model
  - $H_0 : \text{OR} \in (0.9, 1.1)$ for risk in logistic regression model

- **Modelling issues:**
  - model misspecification (e.g. normality)
  - nuisance parameters (the way a p-value is calculated plays a crucial role)
  - overestimation of the test statistic variance (e.g. gene expression)
  - theoretical vs effective degrees of freedom (e.g. low sample size: a lack of asymptotic approximation by chi-square distribution)

Our goal

- Address the scenario with interval (composite) null hypotheses
Simulated data: normal data

normally distributed data and normal model with $H_0: \mu \leq 0$ and $H_1: \text{not } H_0$

truth: $\mu = 0$
(100000 p-values, known unit variance)

truth: $\mu \sim \text{Unif}(-1, 0)$
(100000 p-values, known unit variance)

truth: $\mu \sim \text{Unif}(-2, 0)$
(100000 p-values, known unit variance)
Other consequences of inflation

Data-adaptive methods

- Inflation of p-values near 1 is particularly harmful to certain **data-adaptive methods** that attempt to *estimate the proportion of true nulls* using "large" p-values (e.g. Storey’s method)

Example: Bonferroni plug-in procedure

- **Fact**: the classical Bonferroni method only requires correction by \( m_0 \) (the number of true nulls) in order to strongly control FWER \( \Rightarrow \) estimate \( m_0 \) using the large p-values (above \( \lambda \)) which likely correspond to true nulls

- The **Bonferroni plug-in procedure** as defined by Finner and Gontscharuk (2009) based on the work of Storey (2002): reject \( H_i \) if \( p_i < \alpha/\hat{m}_0 \), where

\[
\hat{m}_0 = (1 - \lambda)^{-1} (1 + \sum_j I\{p_j > \lambda\}),
\]

where \( \lambda \in (0, 1) \).

- With inflated p-values this **is still valid** but it becomes **inefficient**
OUR METHOD
Conditionalized testing

1. Before looking at the data, set $\lambda \in (0, 1)$ (e.g. $\lambda = 0.5$)
2. Calculate $p_1, \ldots, p_m$
3. Remove all p-values above $\lambda$
4. Scale the remaining p-values by dividing them by $\lambda$
5. Perform the selected multiple testing procedure as if only the remaining scaled p-values were originally observed

Comments

1. Conditionalization is applicable to virtually any multiple testing procedure
2. Especially useful if many of the nulls are "deeply" true
3. Even if that is not the case, the associated power loss is typically small (especially when $\lambda$ is not extremely small)
Proposed method

Conditionalized testing

1. Before looking at the data, set $\lambda \in (0, 1)$ (e.g. $\lambda = 0.5$)
2. Calculate $p_1, \ldots, p_m$
3. Remove all p-values above $\lambda$
4. Scale the remaining p-values by dividing them by $\lambda$
5. Perform the selected multiple testing procedure as if only the remaining scaled p-values were originally observed

Comments

1. *Conditionalization* is applicable to **virtually any multiple testing procedure**
2. Especially useful if many of the nulls are "deeply" true
3. Even if that is not the case, the associated **power loss is typically small** (especially when $\lambda$ is not extremely small)
THEORY
General theoretical result

Definition: **Supra-uniformity**

The distribution of $p$ is **supra-uniform** if for all $\lambda, \gamma \in [0, 1]$ with $\gamma \leq \lambda$ it holds $P(p < \gamma | p \leq \lambda) \leq \gamma / \lambda$. (This is equivalent to $p$ having **convex CDF**)

**Theorem**

Let $M$ be a multiple testing procedure which controls FWER (FDR) if the p-values of the **true hypotheses are independent and supra-uniform**. Let the p-values be independent and let the p-values of the true null hypotheses be supra-uniform. Then the conditionalized version of $M$ (with any $\lambda$) controls FWER (FDR).

**Essentially**

With independent and inflated (with convex CDF) p-values conditionalization does not break error control of most standard procedures (e.g. Bonferroni, Holm, Hommel, BH).
SIMULATION
Simulation setup

Setup

- Exchangeable normal model with $\rho = 0$ (all p-values independent)
- Number of tests: $N = 100$
- Fixed $\lambda = 0.5$
- Distribution of effects:
  - **False hypotheses**: $n_1$ positive effects of size $\mu_1 = 0.1$
  - **True hypotheses**: $n_0$ non-positive effects equi-distantly covering $[\mu_0, 0]$, where $\mu_0$ ranges between 0 and −0.2

Considered methods

- Bonferroni, Holm, Hochberg, Hommel, Benjamini-Hochberg (BH), Benjamini-Yekutieli (BY), SFG
RESULTS

ERROR CONTROL
Simulation setup 1 ($N = 100$): FWER (without FDR controlling methods)

Notice: 1. Price to pay "on the border" ($\mu = 0$) vs. gains under the "deep null" ($\mu < 0$)
2. With many false nulls conditionalizing doesn't always help especially when $\mu$ near 0 (except SFG)
3. Non-monotonicity of FWER as a function of $\mu \Rightarrow$ complex behavior
RESULTS

POWER
Simulation results

Simulation setup 1 ($N = 100$): POWER (ratios)

Power ratio $k \geq 1$
exchangeable normal model, $\lambda = 0.5$, $\rho = 0$, $N=100$, $n_1 = 1$, $M=100K$

Power ratio $k \geq 5$
exchangeable normal model, $\lambda = 0.5$, $\rho = 0$, $N=100$, $n_1 = 10$, $M=100K$

Power ratio $k \geq 25$
exchangeable normal model, $\lambda = 0.5$, $\rho = 0$, $N=100$, $n_1 = 50$, $M=100K$

Power ratio $k \geq 40$
exchangeable normal model, $\lambda = 0.5$, $\rho = 0$, $N=100$, $n_1 = 80$, $M=100K$
Simulation results

Simulation setup 1 \((N = 100)\): POWER

Power \(k\geq1\)

exchangeable normal model, \(\lambda = 0.5, \rho = 0, N=100, n_1 = 1, M=100K\)

Power \(k\geq5\)

exchangeable normal model, \(\lambda = 0.5, \rho = 0, N=100, n_1 = 10, M=100K\)

Power \(k\geq25\)

exchangeable normal model, \(\lambda = 0.5, \rho = 0, N=100, n_1 = 50, M=100K\)

Power \(k\geq40\)

exchangeable normal model, \(\lambda = 0.5, \rho = 0, N=100, n_1 = 80, M=100K\)


Uslu, V.V. et. al. (2014), Longrange enhancers regulating Myc expression are required for normal facial morphogenesis. Nature Genetics.


The end.
A broken promise: Differential abundance methods do not control the false discovery rate

Stijn Hawinkel\textsuperscript{1}, Federico Mattiello\textsuperscript{1}, Luc Bijnens\textsuperscript{2,3} and Olivier Thas\textsuperscript{1,4}

\textsuperscript{1}Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, \textsuperscript{2}Quantitative Sciences, Janssen Pharmaceutical companies of Johnson and Johnson, \textsuperscript{3}Center for Statistics, Hasselt University, \textsuperscript{4}National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Australia

April 24, 2017
The human microbiome

- The community of microbes living in and on the human body
- Crucial functions in digestion and immunity
- Our “forgotten organ”
The human microbiome

- Changes in the microbiome correlate with disease development
- Opportunities for the microbiome as
  - Diagnostic or prognostic marker
  - Drug target
  - Probiotic
Bacteria are counted by sequencing a marker gene (usually 16S rRNA)

Data structure:

<table>
<thead>
<tr>
<th>Samples</th>
<th>taxon 1</th>
<th>...</th>
<th>taxon p</th>
<th>Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample 1</td>
<td>$x_{11}$</td>
<td>...</td>
<td>$x_{1p}$</td>
<td>$Z_1$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>sample n</td>
<td>$x_{n1}$</td>
<td>...</td>
<td>$x_{np}$</td>
<td>$Z_n$</td>
</tr>
</tbody>
</table>

Around 75% zero counts
Differential abundance

Which taxa have a **different abundance** between two groups (e.g. treatment-control)?

- Association of **mean abundance** of a single bacterial strain (**taxon**) with a patient covariate
  - Identify signature diagnostic taxa
  - Discover taxa affected by a treatment
Test for differential abundance

Taxon per taxon

- t-test or Wilcoxon rank sum test (**SAMseq**)
- Parametric model
  - log-normal: **limma-voom**
  - negative binomial: **edgeR, DESeq2**
  - zero-inflated normal: **metagenomeSeq**

1000’s of taxa $\Rightarrow$ Multiplicity correction needed!

- False discovery rate
  - Benjamini-Hochberg
  - Benjamini-Yekutieli (any **dependence** structure)
Parametric simulation

1. Propose a parametric model
2. Estimate parameter values from real datasets
3. Choose:
   - Sample size
   - Effect size (fold change)
4. Generate Monte-Carlo samples (taxon per taxon)
Taxon correlation networks

- Generating datasets taxon per taxon implicitly assumes independence between taxa
- Bacteria in the same niche do not grow independently!
- Our novelty:
  - Estimate taxa correlation networks from real data
  - Incorporate the estimated correlation in the simulations
Real data resampling

1. **Mock variable**: Split a dataset without biological signal randomly in two groups and test for differential abundance
   - Evaluate type I error

2. **Evaluation-verification** method
   - Choose a covariate known to be associated with microbiome composition
   - Split dataset in **small** evaluation and **large** verification set
   - Test for differential abundance in both sets
   - Evaluate results in evaluation set based on those of the verification set
Sensitivity: Synthetic data

![Graph showing sensitivity for various statistical tests and sample sizes]

- Distribution: Negative_binomial, Beta–binomial(Cor)
- Sample size: 5, 25, 50, 100
- Sensitivity: 0.00, 0.25, 0.50, 0.75, 1.00

Tests:
- t-test
- Wilcoxon
- DESeq2
- edgeR
- limma-voom
- metagenomeSeq
- SAMseq

Adjustment methods:
- Benjamini-Hochberg
- Benjamini-Yekutieli
- Plug-in
Sensitivity: Evaluation-verification

Verification Method
- Wilcoxon_TSS
- DESeq2
- metagenomeSeq
- edgeR
- limma-voom
- SAMseq

IBD
Penicillin use
Gender

Sample Size
Sensitivity

25 5 25 5 25 5 25 5 25 5 25 5 25 5 25 5 25 5

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30

0.00 0.15 0.30
False discovery rate: Evaluation-verification method

Verification Method
- Wilcoxon_TSS
- DESeq2
- metagenomeSeq
- edgeR
- limma-voom
- SAMseq

SampleSize
- 5
- 25

FDR
- 0.00
- 0.25
- 0.50
- 0.75
- 1.00

IBD
- Penicillin use
- Gender

SampleSize
Why is the FDR out of control?

- Multiple correction methods rely on correct P-values
- P-values should be uniformly distributed under H0
P-value distribution under H0: mock variable

Normalization:
- None
- Library sizes
- Relative log-expression
- Trimmed mean of M-values
- Cumulative sum scaling
- Rarefying

Sample sizes:
5, 25, 100

Liberal area:

Stool (AGP)

Mid. vagina

Stool (HMP)

Tongue. obs."
P-value distribution and zero frequency (mock variable)
P-value distribution and zero frequency

- Count distributions and associated GLMs require large sample sizes for *convergence to asymptotic distributions*
- High frequencies of *zeroes* and *skewed* count distributions yield overtly liberal methods
- Remedies:
  - Thorough trimming of rare taxa
  - Explicitly model zero counts
Questions?

Thank you for your attention!
A typical genomics experiment tests 100’s to 1000’s of null hypotheses.

Most of which are **true**!

Interest is mainly in taxa that are **differentially abundant**.

- $H_0$ taxa
- $H_1$ taxa

Specificity = 0.94

Sensitivity = 0.5

FDR = 0.5 !

$P_{\text{adjusted}} < 0.05$
A general and powerful stage-wise testing procedure for differential expression and differential transcript usage

Koen Van den Berge\textsuperscript{1,2}, Charlotte Soneson\textsuperscript{3,4}, Mark D. Robinson\textsuperscript{3,4} & Lieven Clement\textsuperscript{1,2}

\textsuperscript{1} Department of Applied Mathematics, Computer Science & Statistics, Ghent University
\textsuperscript{2} Bioinformatics Institute Ghent, Ghent University
\textsuperscript{3} Institute of Molecular Life Sciences, University of Zurich
\textsuperscript{4} SIB Swiss Institute of Bioinformatics, University of Zurich

April 21, 2017
Motivation: Modern RNA-seq experiments

Gene expression is measured

- for thousands of genes
- with few biological replicates, typically 3 to 5
- in often multiple conditions.

Complex designs result in multiple hypotheses of interest:
1. Is the gene differentially expressed in the different conditions?
2. Does the pattern of differential expression change over time?
Motivation: Modern RNA-seq experiments

Gene expression is measured
- for thousands of genes
- with few biological replicates, typically 3 to 5
- in often multiple conditions.

Complex designs result in multiple hypotheses of interest:
1. Is the gene is differentially expressed in the different conditions?
2. Does the pattern of differential expression change over time?
Motivation: Modern RNA-seq experiments

Recent developments in read alignment allow transcript-level analysis.

https://en.wikibooks.org/wiki/Proteomics/Protein_Primary_Structure/Alternative_Splicing
Multiple hypotheses per gene

Transcript-level analysis and complex designs result in multiple hypotheses of interest per gene. The conventional strategy would be to

1. assess each hypothesis separately
2. on some FDR level $\alpha$
3. provide the biologist with list of top-genes for every contrast

However,

- Interpretation occurs at the gene level while FDR is not controlled at this level
Multiple hypotheses per gene

Transcript-level analysis and complex designs result in multiple hypotheses of interest per gene. The conventional strategy would be to:

1. assess each hypothesis separately
2. on some FDR level $\alpha$
3. provide the biologist with list of top-genes for every contrast

However,

- Interpretation occurs at the gene level while FDR is not controlled at this level
Overall (gene-level) FDR

- Define a false positive gene as a gene for which one of the null hypotheses have been falsely rejected.
- The overall FDR (gene-level FDR) is defined as the expected proportion of false positive genes over all rejected genes.

How should we control FDR on this level?
Stage I: Control FDR on gene level

A simple strategy would be to

1. Aggregate p-values across hypotheses (i.e. omnibus test)
2. Control FDR on level $\alpha_I$ on the aggregated p-values

Additionally takes advantage of aggregated tests with higher sensitivity

However, we lose resolution on the biology
Stage I: Control FDR on gene level

A simple strategy would be to

1. Aggregate p-values across hypotheses (i.e. omnibus test)
2. Control FDR on level $\alpha_I$ on the aggregated p-values

Additionally takes advantage of aggregated tests with higher sensitivity

However, we lose resolution on the biology
Stage-wise testing procedure \(^1\)

Let \(H^S_g\) be the screening hypothesis for gene \(g\) testing the global \(H_0\)

1. **Screening Stage:**
   - Assess the screening hypothesis \(H^S_g\) for all genes in the set \(G\).
   - Let \(p^S_1, \ldots, p^S_G\) be the unadjusted p-values from the screening stage test.
   - Apply the Benjamini Hochberg (BH) FDR procedure to \(p^S_1, \ldots, p^S_G\) at FDR level \(\alpha_I\). Let \(R\) be the number of rejected screening hypotheses.

2. **Confirmation Stage:** For all \(R\) genes that pass the screening stage.
   - Let \(\alpha_{II} = R\alpha_I / G\) be the FDR-adjusted significance level from the first stage.
   - Let \(p_{1g}, \ldots, p_{ng}\) be the p-values from \(H_{1g}, \ldots, H_{ng}\) for gene \(g\).
   - Adopt a multiple testing procedure to assess all \(ng\) hypotheses while controlling the within gene family wise error rate (FWER) at the adjusted level \(\alpha_{II}\).

\(^1\)Heller et al. 2009, Bioinformatics.
Stage-wise testing procedure \(^1\)

Let \(H^S_g\) be the screening hypothesis for gene \(g\) testing the global \(H_0\)

1. **Screening Stage:**
   - Assess the screening hypothesis \(H^S_g\) for all genes in the set \(G\).
   - Let \(p^S_1, \ldots, p^S_G\) be the unadjusted p-values from the screening stage test.
   - Apply the Benjamini Hochberg (BH) FDR procedure to \(p^S_1, \ldots, p^S_G\) at FDR level \(\alpha_I\). Let \(R\) be the number of rejected screening hypotheses.

2. **Confirmation Stage:** For all \(R\) genes that pass the screening stage.
   - Let \(\alpha_{II} = R\alpha_I/G\) be the FDR-adjusted significance level from the first stage.
   - Let \(p_{1g}, \ldots, p_{ng}\) be the p-values from \(H_{1g}, \ldots, H_{ng}\) for gene \(g\).
   - Adopt a multiple testing procedure to assess all \(ng\) hypotheses while controlling the within gene family wise error rate (FWER) at the adjusted level \(\alpha_{II}\).

\(^1\)Heller et al. 2009, Bioinformatics.
Stage-wise testing procedure

Control FDR on aggregated tests across all genes

Aggregate evidence (Omnibus test)

Only retain significant genes

Control FWER within a gene on FDR adjusted significance level

Assess each hypothesis separately
Simulation: DGE experiments with complex design

- Based on Hammer et al. (2010), Genome Research
- Two conditions (control - SNL)
- Two timepoints (2 weeks - 2 months)

Interest in:
1. DE between conditions at 2 weeks
2. DE between conditions at 2 months
3. Different DE between timepoints (interaction)
Results: DGE experiments with complex designs

▶ We show that the procedure correctly controls the OFDR
▶ The omnibus test enriches for genes with interaction effects
▶ While maintaining equivalent power for main effects
Results: transcript-level analysis

The procedure naturally unites gene-level with transcript-level tests.

- Benefits from high gene-level power combined with transcript-level resolution of the results
- Performances are at least par or better compared to a conventional analysis
- Better FDR control
Conclusion

Differential expression experiments with complex designs:
1. Gene-level FDR control
2. Higher power for interactions
3. Equivalent power for main effects

Transcript-level analysis:
1. Gene-level power
2. Transcript-level results
3. Better FDR control

R package stageR: https://github.com/statOmics/stageR

Preprint on biorXiv:

A general and powerful stage-wise testing procedure for differential expression and differential transcript usage

Koen Van den Berge, Charlotte Soneson, Mark D. Robinson, Lieven Clement

doi: https://doi.org/10.1101/109082
Appendix: FDR is not controlled at the gene level

<table>
<thead>
<tr>
<th></th>
<th>H_1</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>1</td>
</tr>
<tr>
<td>gene2</td>
<td>1</td>
</tr>
<tr>
<td>gene3</td>
<td>1</td>
</tr>
<tr>
<td>gene4</td>
<td>1</td>
</tr>
<tr>
<td>gene5</td>
<td>1</td>
</tr>
<tr>
<td>gene6</td>
<td>0</td>
</tr>
<tr>
<td>gene7</td>
<td>1</td>
</tr>
<tr>
<td>gene8</td>
<td>1</td>
</tr>
<tr>
<td>gene9</td>
<td>1</td>
</tr>
<tr>
<td>gene10</td>
<td>1</td>
</tr>
<tr>
<td>gene11</td>
<td>1</td>
</tr>
<tr>
<td>gene12</td>
<td>1</td>
</tr>
<tr>
<td>gene13</td>
<td>1</td>
</tr>
<tr>
<td>gene14</td>
<td>1</td>
</tr>
<tr>
<td>gene15</td>
<td>1</td>
</tr>
<tr>
<td>gene16</td>
<td>1</td>
</tr>
<tr>
<td>gene17</td>
<td>0</td>
</tr>
<tr>
<td>gene18</td>
<td>0</td>
</tr>
<tr>
<td>gene19</td>
<td>0</td>
</tr>
<tr>
<td>gene20</td>
<td>1</td>
</tr>
<tr>
<td>gene21</td>
<td>0</td>
</tr>
<tr>
<td>gene22</td>
<td>1</td>
</tr>
<tr>
<td>gene23</td>
<td>0</td>
</tr>
<tr>
<td>gene24</td>
<td>0</td>
</tr>
<tr>
<td>gene25</td>
<td>1</td>
</tr>
<tr>
<td>gene26</td>
<td>1</td>
</tr>
<tr>
<td>gene27</td>
<td>1</td>
</tr>
<tr>
<td>gene28</td>
<td>1</td>
</tr>
<tr>
<td>gene29</td>
<td>1</td>
</tr>
<tr>
<td>gene30</td>
<td>0</td>
</tr>
</tbody>
</table>

- 0 represents not significant
- 1 represents significance
- 1 represents false positive

\[ FDR_{hyp} = \frac{1}{20} = 5\% \]

Define

- a false positive gene as a gene with at least one false null hypothesis
- Gene-level FDR (OFDR) controls the expected proportion of false positive genes over all rejected genes

\[ FDR_{gene} = \frac{1}{20} = 5\% \]
Appendix: FDR is not controlled at the gene level

<table>
<thead>
<tr>
<th>Gene</th>
<th>H1</th>
<th>H2</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene4</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene8</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene10</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene11</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene12</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene13</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene14</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene15</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene16</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene17</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gene18</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene19</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene20</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene21</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gene22</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene23</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene24</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene25</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene26</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene27</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene28</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene29</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene30</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- 0 represents not significant
- 1 represents significance
- 1 represents false positive

\[
FDR_{hyp} = \frac{2}{40} = 5\%
\]

Define

- a false positive gene as a gene with at least one false null hypothesis
- Gene-level FDR (OFDR) controls the expected proportion of false positive genes over all rejected genes

\[
FDR_{gene} = \frac{2}{28} = 7.1\%
\]
Appendix: FDR is not controlled at the gene level

<table>
<thead>
<tr>
<th></th>
<th>H₁</th>
<th>H₂</th>
<th>H₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene2</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene3</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene6</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gene7</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene8</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gene9</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene10</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene11</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene12</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene13</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene14</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene15</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene16</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene17</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene18</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene19</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene20</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gene21</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene22</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene23</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene24</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene25</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gene26</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene27</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene28</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene29</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gene30</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- 0 represents not significant
- 1 represents significance
- **1 represents false positive**

\[
FDR_{hyp} = \frac{3}{60} = 5\%
\]

Define

- a false positive gene as a gene with at least one false null hypothesis
- Gene-level FDR (OFDR) controls the expected proportion of false positive genes over all rejected genes

\[
FDR_{gene} = \frac{3}{30} = 10\%
\]
Appendix: FDR is not controlled at the gene level

<table>
<thead>
<tr>
<th>gene</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
<th>H4</th>
</tr>
</thead>
<tbody>
<tr>
<td>gene1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gene6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene7</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene8</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene9</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene11</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene12</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene13</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene14</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene15</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene16</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene17</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene18</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene19</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene20</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gene21</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene22</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene23</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene24</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene25</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene26</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>gene27</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>gene28</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>gene29</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>gene30</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

▶ 0 represents not significant
▶ 1 represents significance
▶ 1 represents false positive

\[ FDR_{hyp} = \frac{4}{80} = 5\% \]

Define

▶ a false positive gene as a gene with at least one false null hypothesis
▶ Gene-level FDR (OFDR) controls the expected proportion of false positive genes over all rejected genes

\[ FDR_{gene} = \frac{4}{30} = 13.3\% \]