Correcting false discovery rates for their bias toward false positives

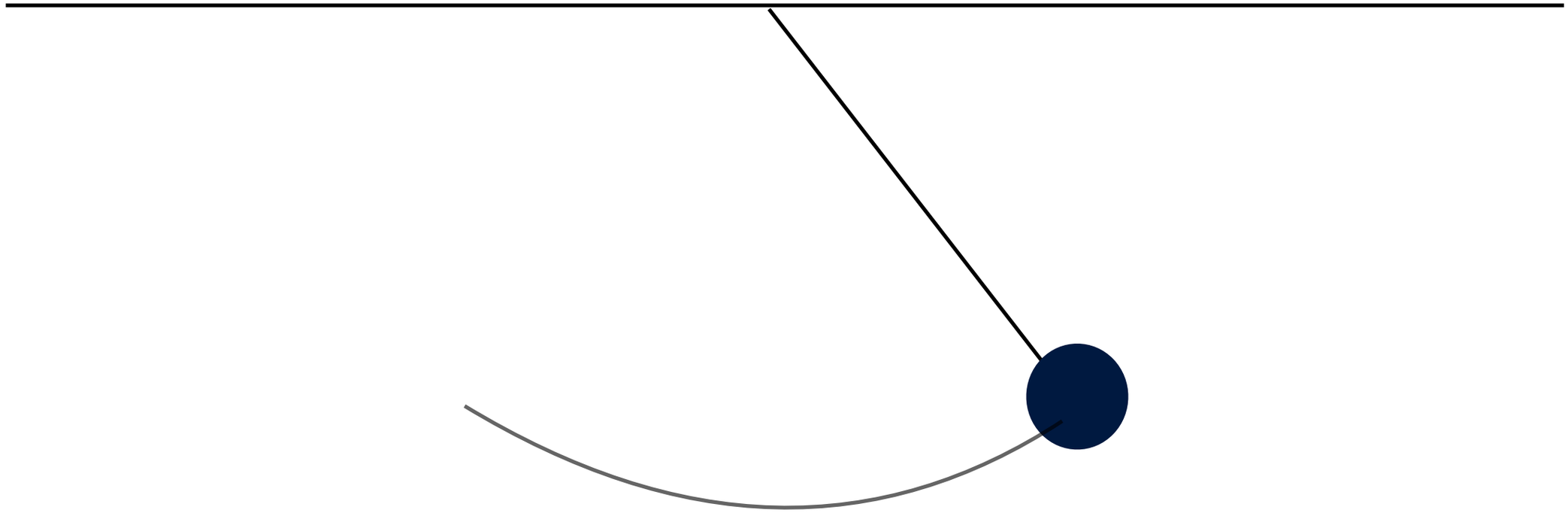
2017 Annual Meeting of the Statistical Society of Canada
University of Manitoba, Winnipeg

13 June 2017

David Bickel
University of Ottawa

uOttawa

Swung too far?



Family-wise error rates

Too many false negatives

False discovery rates

Too many false positives

Local false discovery rates

Large variance

The rise of false discovery rates

FDR software used in genomics

- Desktop software:
  - GSEA
  - Cyber-T
  - MACS
- Web software:
  - DAVID
  - Toppfunn
  - GREAT

Why the FDR became popular

Approach	FWER control	FDR control
Significance measure	Adjusted p-value	q-value
Interpretation	*(Achieved FWER)*	*(Achieved FDR)*
Decisions	*(Rejection only)*	*(Rejection only)*
Prior distribution	**None**	**None**
Many tests	*(Highly conservative)*	**Adequate**
Few tests	**Adequate**	**Adequate**

Multiple comparisons

Table 1 Summary of strengths and weakness of the four major approaches to multiple hypothesis testing.

	Error-rate control approaches		Posterior probability approaches	
Approach	**FWER control**	**FDR control**	**Classical Bayes**	**Empirical Bayes**
Significance measure	Adjusted p-value	q-value	LFDR	Estimated LFDR
Interpretation	*(Achieved FWER)*	*(Achieved FDR)*	**Level of belief**	**Estimated prob.**
Decisions	*(Rejection only)*	*(Rejection only)*	**Optimal, flexible**	Flexible
Prior distribution	**None**	**None**	*(Specified)*	Estimated
Many tests	*(Highly conservative)*	**Adequate**	**Adequate**	**Adequate**
Few tests	**Adequate**	**Adequate**	**Adequate**	*(Estimation error)*

Each of the last five rows has practical advantages and disadvantages of each approach according to the consideration given in the first column. A **bold entry** means the approach of a column is among the best for the consideration, an underlined entry means it is advantageous but notably less so, and an (*italicized entry in parentheses*) means it is relatively disadvantageous.

D. R. Bickel, *Statistical Applications in Genetics and Molecular Biology* **12**, 529–543

FDRs & local FDRs

$$\begin{aligned}
\mathrm{FDR}(0.01) &= \frac{\text{average number of false discoveries at 0.01 significance}}{\text{average number of discoveries at 0.01 significance}} \\
&= \frac{\text{average number of p-values} < 0.01 \text{ for equivalently expressed genes}}{\text{average number of p-values} < 0.01}
\end{aligned}$$

$$\begin{aligned}
\widehat{\mathrm{FDR}}(\alpha) &= \frac{\text{estimated average number of false discoveries}}{\text{estimated average number of discoveries}} \\
&= \frac{\text{estimated average number of false discoveries}}{\text{number of discoveries}} \\
&= \begin{cases} \frac{\alpha d}{\#(p(x_i) \le \alpha)} & \text{if } \frac{\alpha d}{\#(p(x_i) \le \alpha)} < 1 \\ 1 & \text{if } \frac{\alpha d}{\#(p(x_i) \le \alpha)} > 1. \end{cases}
\end{aligned}$$

$$\mathrm{FDR}(\alpha) \approx \frac{\mathrm{LFDR}(p_1) + \mathrm{LFDR}(p_2) + \cdots + \mathrm{LFDR}\left(p_{\#(p(x_i) \le \alpha)}\right)}{\#(p(x_i) \le \alpha)}$$

Interpreting the false discovery rate

- If a discovery of differential expression is made whenever the p-value is less than 0.05, then the false discovery rate is the average of all LFDRs corresponding to discoveries

$$\text{FDR}(0.05) = \text{mean}(\text{LFDR}(p(x_i)) | p(x_i) < 0.05)$$

- false discovery rate = probability that randomly selected discovery is false

$$\text{FDR}(0.05) = P(A_i = 0 | p(x_i) < 0.05)$$

Local false discovery rates

- local false discovery rate (LFDR) = posterior probability of equivalent expression: $\mathrm{LFDR}\left(0.00832\right) = P\left(A_i = 0 | p\left(X_i\right) = 0.00832\right)$

- evidence of differential expression = likelihood ratio:

$$\frac{L_i\left(1\right)}{L_i\left(0\right)} \approx \frac{P\left(p\left(X_i\right) \approx 0.00832 | A_i = 1\right)}{P\left(p\left(X_i\right) \approx 0.00832 | A_i = 0\right)}$$

- posterior odds that gene i of p-value 0.00832 is differentially expressed:

$$\frac{1 - \mathrm{LFDR}\left(0.00832\right)}{\mathrm{LFDR}\left(0.00832\right)} = \frac{P\left(A_i = 1 | p\left(X_i\right) = 0.00832\right)}{P\left(A_i = 0 | p\left(X_i\right) = 0.00832\right)} = \frac{P\left(A_i = 1\right)}{P\left(A_i = 0\right)} \times \frac{L_i\left(1\right)}{L_i\left(0\right)}$$

Achieved FDR

$$\mathrm{FDR}(0.01) = \frac{\text{average number of false discoveries at 0.01 significance}}{\text{average number of discoveries at 0.01 significance}}$$

$$= \frac{\text{average number of p-values} < 0.01 \text{ for equivalently expressed genes}}{\text{average number of p-values} < 0.01}$$

$$\widehat{\mathrm{FDR}}(\alpha) = \begin{cases} \frac{\alpha d}{\#(p(x_i) \le \alpha)} & \text{if } \frac{\alpha d}{\#(p(x_i) \le \alpha)} < 1 \\ 1 & \text{if } \frac{\alpha d}{\#(p(x_i) \le \alpha)} > 1 \end{cases}$$

$$\widehat{\mathrm{FDR}}(p(x_j)) = \begin{cases} \frac{p(x_j) d}{\#(p(x_i) \le p(x_j))} & \text{if } \frac{p(x_j) d}{\#(p(x_i) \le p(x_j))} < 1 \\ 1 & \text{if } \frac{p(x_j) d}{\#(p(x_i) \le p(x_j))} > 1 \end{cases}$$

Bias in false discovery rates

$$\begin{aligned}\mathrm{FDR}\left(p\left(x_j\right)\right) &\approx \frac{\mathrm{LFDR}\left(p_1\right)+\mathrm{LFDR}\left(p_2\right)+\cdots+\mathrm{LFDR}\left(p_{\#\left(p\left(x_i\right)\le p\left(x_j\right)\right)}\right)}{\#\left(p\left(x_i\right)\le p\left(x_j\right)\right)} \\ &= \frac{\mathrm{LFDR}\left(p_1\right)+\mathrm{LFDR}\left(p_2\right)+\cdots+\mathrm{LFDR}\left(p_j\right)}{\#\left(p\left(x_i\right)\le p\left(x_j\right)\right)}\end{aligned}$$

$$p_1 < p_2 < \cdots < p_j$$

$$\mathrm{LFDR}\left(p_1\right) < \mathrm{LFDR}\left(p_2\right) < \cdots < \mathrm{LFDR}\left(p_j\right)$$

$$\mathrm{LFDR}\left(p_j\right) > \frac{\mathrm{LFDR}\left(p_1\right)+\mathrm{LFDR}\left(p_2\right)+\cdots+\mathrm{LFDR}\left(p_j\right)}{\#\left(p\left(x_i\right)\le p\left(x_j\right)\right)} = \mathrm{FDR}\left(p\left(x_j\right)\right)$$

$$\mathrm{FDR}\left(p\left(x_j\right)\right) < \mathrm{LFDR}\left(p_j\right)$$

Correcting the bias

$$p_1 < p_2 < \cdots < p_j$$

$$\mathrm{LFDR}(p_1) < \mathrm{LFDR}(p_2) < \cdots < \mathrm{LFDR}(p_j)$$

$$\widehat{\mathrm{FDR}}(p(x_j)) = \begin{cases} \dfrac{p(x_j)d}{\#\left(p(x_i) \le p(x_j)\right)} & \text{if } \dfrac{p(x_j)d}{\#\left(p(x_i) \le p(x_j)\right)} < 1 \\ 1 & \text{if } \dfrac{p(x_j)d}{\#\left(p(x_i) \le p(x_j)\right)} > 1 \end{cases}$$

$$\widehat{\mathrm{LFDR}}(p_j) = \left(\frac{1}{j-1+1} + \frac{1}{j-2+1} + \cdots + \frac{1}{j-j+1}\right)\widehat{\mathrm{FDR}}(p(x_j))$$

D. R. Bickel, deposited in uO Research at https://goo.gl/GcUjJe

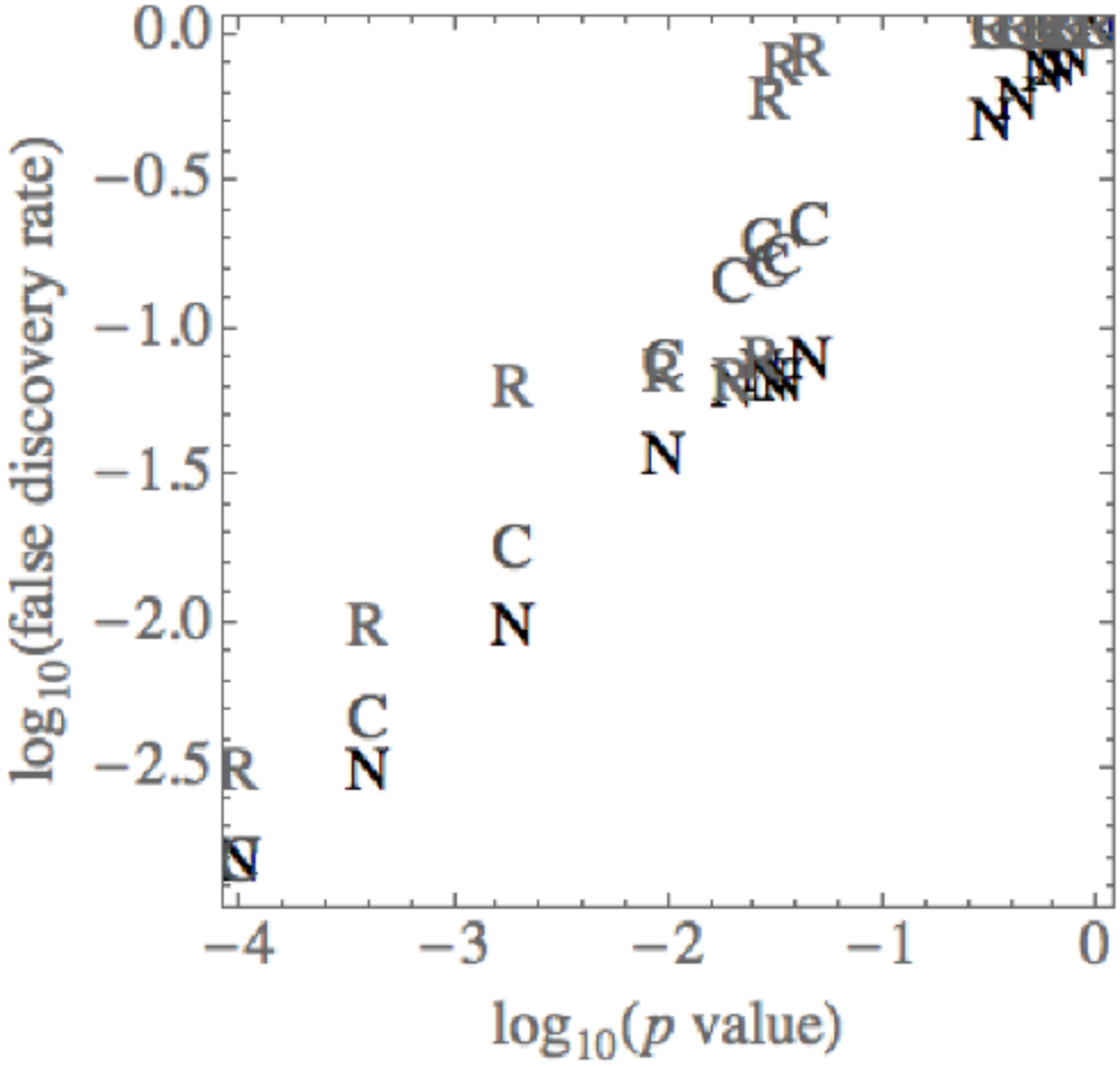
Re-ranked FDRs

Let $q(\alpha)$ denote the smallest value of q such that all hypothesis with p values in $[0, \alpha]$ are rejected according to some procedure that guarantees that the FDR, NFDR, or an estimate of either is no higher than q.

$$\text{RFDR}\left(x_{(i)}\right) = q\left(p\left(x_{([i/F^{\star}(\text{NFDR})])}\right)\right) \text{ if } [i/F^{\star}(\text{NFDR})] \leq d \text{ else } \text{RFDR}\left(x_{(i)}\right) = 1$$

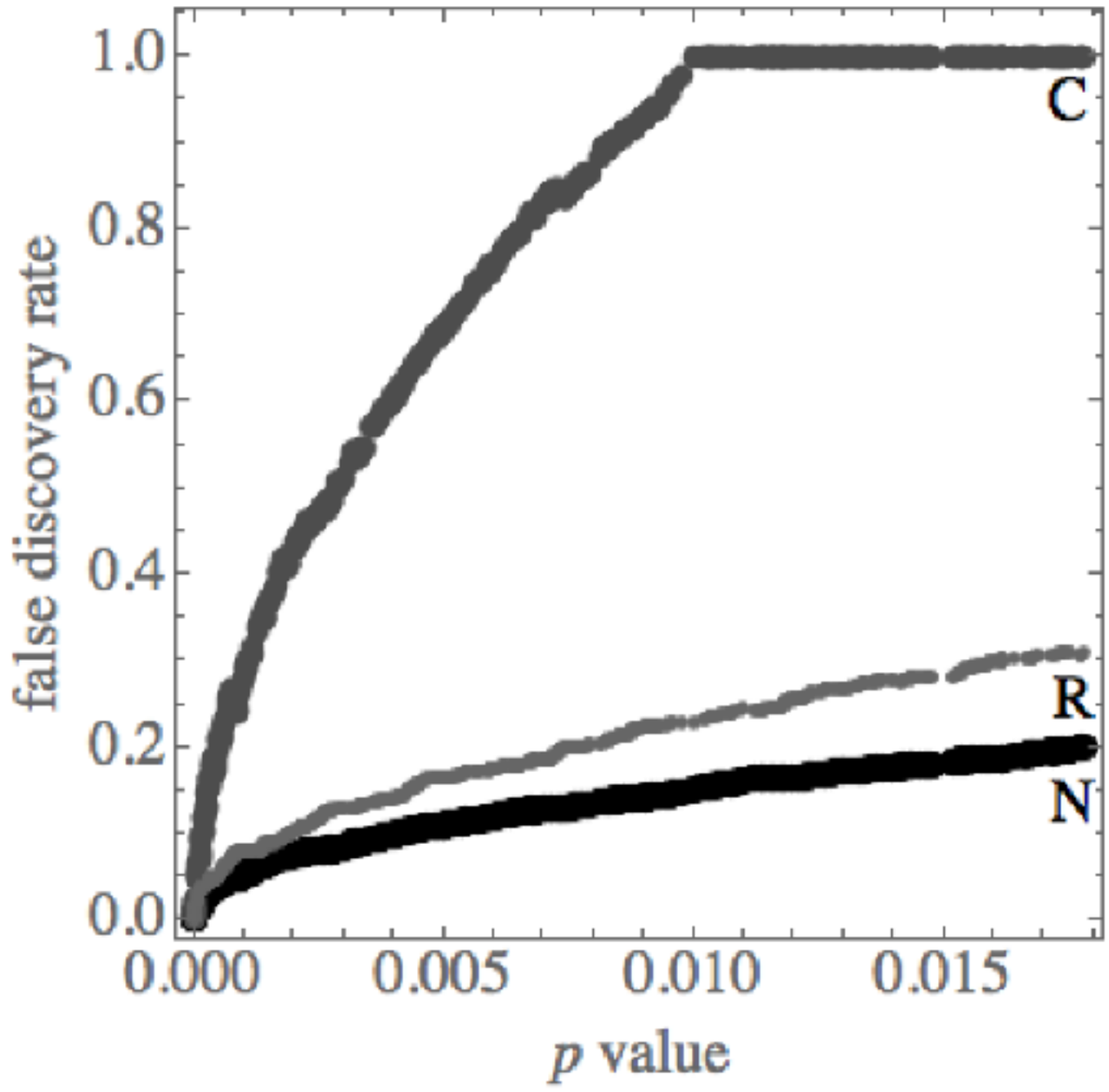
$$F^{\star}(\text{NFDR}) = 1 - e^{-1}$$

Biomedical data



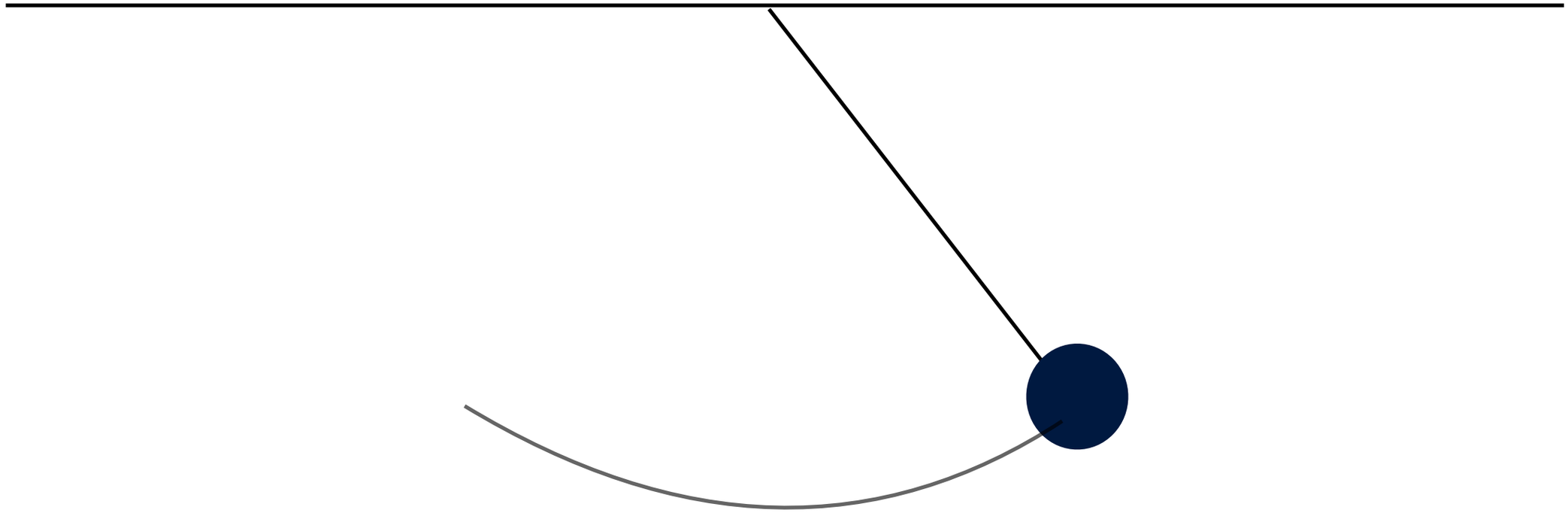
D. R. Bickel, deposited in uO Research at https://goo.gl/GcUjJe

Gene expression data



D. R. Bickel, deposited in uO Research at https://goo.gl/GcUjJe

The right balance



Family-wise error rates

Too many false negatives

CFDR RFDR

Other local FDR estimators?

False discovery rates

Too many false positives

Slides and preprint: www.davidbickel.com

Acknowledgements

- Collaborators:
  - Alexandre Blais
  - Abbas Rahal

- Funding:
  - Agriculture and Agri-Food Canada
  - Faculty of Medicine of the University of Ottawa