Simultaneous Hypothesis Testing using Internal Negative Controls

Qingyuan Zhao

Statistical Laboratory, University of Cambridge

March 29, 2023, Statistics Seminar, University of Pennsylvania

(Joint work with Zijun Gao)

Motivating application

- Use proteomic profiling to identify cell membrane proteins in certain brain regions.

CellPress
OPEN ACCESS

Neuron

Neuroresource

In situ cell-type-specific cell-surface proteomic profiling in mice

S. Andrew Shuster,[1,2,5] Jiefu Li,[1,5] URee Chon,[1,2] Miley C. Sinantha-Hu,[1] David J. Luginbuhl,[1] Namrata D. Udeshi,[3] Dominique Kiki Carey,[3] Yukari H. Takeo,[1] Qijing Xie,[1,2] Chuanyun Xu,[1] D.R. Mani,[3] Shuo Han,[4] Alice Y. Ting,[4] Steven A. Carr,[3] and Liqun Luo[1,6,*]

[1]Department of Biology and Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305, USA
[2]Neurosciences Program, Stanford University, CA 94305, USA
[3]The Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[4]Departments of Genetics, Biology, and Chemistry, Chan Zuckerberg Biohub, Stanford University, Stanford, CA 94305, USA
[5]These authors contributed equally
[6]Lead contact
*Correspondence: lluo@stanford.edu
https://doi.org/10.1016/j.neuron.2022.09.025

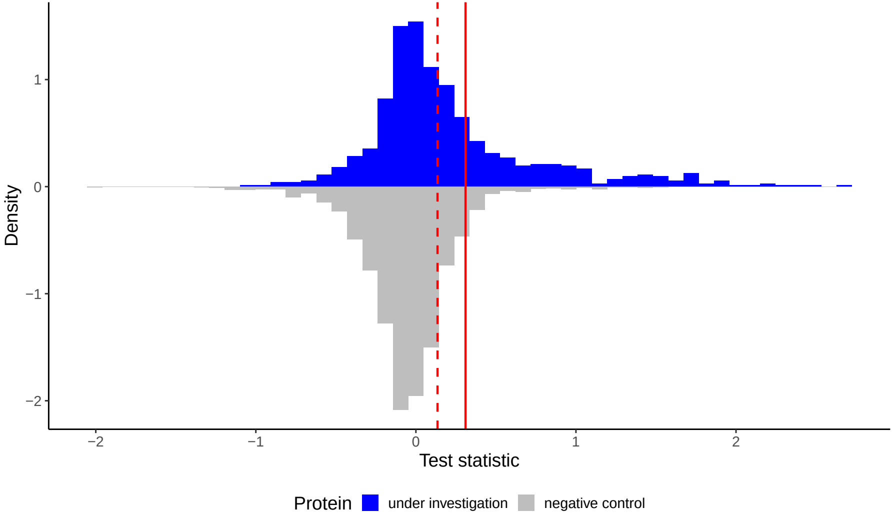
Motivating application: More detail

- Shuster *et al.* (2022) extracted developing Purkinje cells from mice and prepared them for mass spectrometry under two conditions: HRP+H_2O_2 and HRP only (control).
- A common challenge: lack of biological repeats.
- Instead, they used a heuristic in Hung *et al.* (2014)[1] to select a "cut-off".
- This is based on using the UniPort database to classify proteins as
  - Under investigation: annotated with plasma membrane; ($n = 740$)
  - Negative controls:[2] nuclear, mitochondrial, or cytoplasmic but not plasma membrane. ($m = 2,067$)

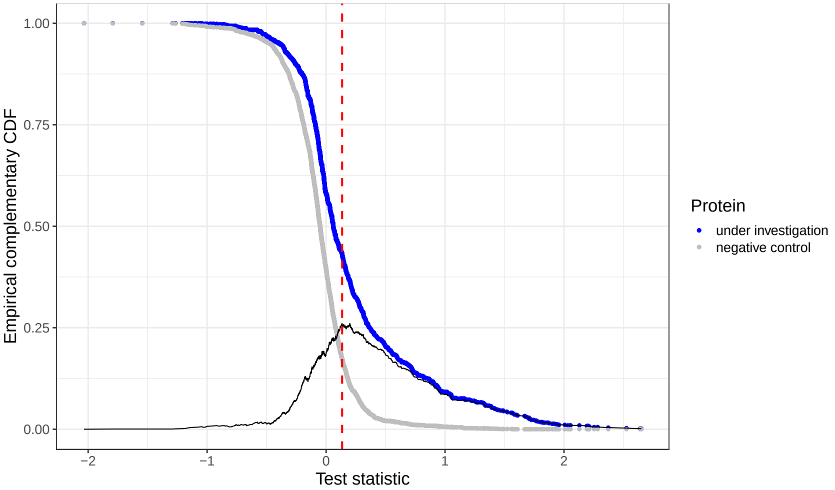
[1] V. Hung *et al.*, *Molecular Cell* **55**, 332–341 (2014).

[2] Shuster *et al.* (2022) referred to internal negative control proteins as "false positives" and proteins under investigation as "true positives".

Illustration of the dataset



Dashed threshold: Hung *et al.* (2014)



- Seems quite ad hoc (but actually not).

Solid threshold: A “new” method

There are two ways to obtain this threshold:

1. Use negative controls to form an empirical null distribution, and then apply the Benjamini-Hochberg procedure;
2. Use negative controls to directly estimate the false discovery rate of a rejection set.

Literature

- Empirical null: Efron (2004); population stratification (Price *et al.* 2006); batch effect (Leek *et al.* 2010).
- Negative control: **internal** (e.g. non-membrane proteins) vs. **external** (e.g. HRP only); essential concept in scientific methods; applications in genomics (Gagnon-Bartsch and Speed 2012), epidemiology (Lipsitch, Tchetgen Tchetgen, and Cohen 2010), causal inference (Miao, Geng, and Tchetgen Tchetgen 2018).
- Use negative controls in multiple testing: several informal proposals (Nix, Courdy, and Boucher 2008; Listgarten *et al.* 2013; Slattery *et al.* 2011; Parks, Raphael, and Lawrence 2018; Zhang *et al.* 2008; Song *et al.* 2007).
- Closely related to conformal inference/prediction and semi-supervised learning.

Outline

1. Setup
2. Method 1: Empirical null
3. Method 2: Empirical process
4. Method 3: Local FDR control/Decision-theoretic perspective
5. Discussion

Setup

- $n + m$ hypotheses: $\mathcal{I} = \{1, \ldots, n\}$ are under investigation; $\mathcal{I}_0 \subseteq \mathcal{I}$ is the unknown set of true null hypotheses; $\mathcal{I}_{\mathrm{nc}} = \{n + 1, \ldots, n + m\}$ are known to be true (negative controls).
- Each hypothesis H_i is associated with a test statistic T_i with CDF F_i. Small T_i indicates evidence against H_i.
- Common error rates in multiple testing: familywise error rate (FWER), false discover rate (FDR), tail probability of false discovery proportion (FDP), local false discovery rate (local-FDR).

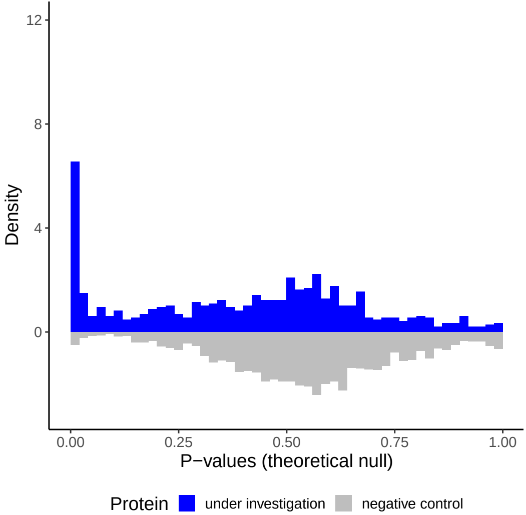
RANC p-values: Two definitions

- We define $p_i = \hat{F}(T_i)$ for $i = 1, \dots, n$, where $\hat{F}$ is the empirical CDF of $(-\infty, T_{n+1}, \dots, T_{n+m})$:
$$\hat{F}(t) = \frac{1 + \sum_{j \in \mathcal{I}_{\mathrm{nc}}} 1_{\{T_j \le t\}}}{1 + m}.$$

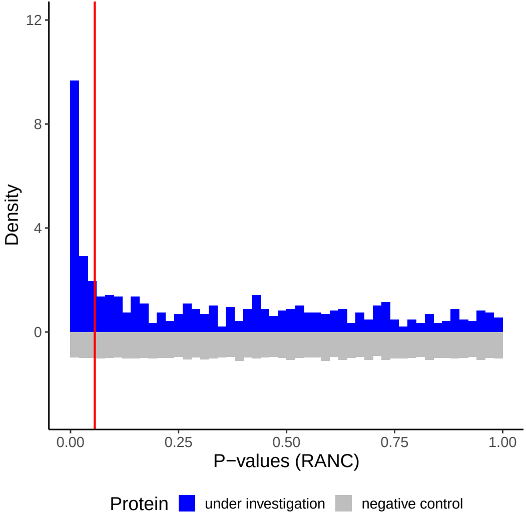
- Equivalently, p_i is simply the normalized rank of T_i among $(T_j)_{j \in \{i\} \cup \mathcal{I}_{\mathrm{nc}}}$:
$$p_i = \frac{1 + (\text{number of negative control statistics } \le T_i)}{1 + (\text{number of negative control statistics})}.$$

- This is why we call p_i the **Rank Among Negative Control (RANC)** p-value.
- If we assume $(T_i)_{i \in \mathcal{I}_0 \cup \mathcal{I}_{\mathrm{nc}}}$ is exchangeable, p_i is exactly the permutation test p-value of exchangeability for T_i.

An illustration using the proteomic dataset



(a) P-values from a two-sample "*t*-test".



(b) RANC p-values. Red: BH with $\mathrm{FDR} = 0.2$.

Useful definitions

Next: properties of RANC p-values and implications for multiple testing.

Partial/conditional exchangeability

$(X_i)_{i \in \mathcal{I}}$ is **exchangeable on a subset** $(X_i)_{i \in \mathcal{J}}$ for some $\mathcal{J} \subseteq \mathcal{I}$, if for any permutation $g : \mathcal{I} \to \mathcal{I}$ such that $g(i) = i$ for all $i \notin \mathcal{J}$, we have $(X_{g(i)})_{i \in \mathcal{I}} \overset{\mathrm{d}}{=} (X_i)_{i \in \mathcal{I}}$.
When this holds for $\mathcal{J} = \mathcal{I}$, we simply say $(X_i)_{i \in \mathcal{I}}$ is **exchangeable**.

A set $\mathcal{D} \subseteq \mathbb{R}^n$ is increasing if $\mathcal{D}$ contains all $y \succeq x \in \mathcal{D}$.

PRDS

$(X_i)_{i \in \mathcal{I}}$ exhibits **positive regression dependence on a subset (PRDS)** $(X_j)_{j \in \mathcal{J}}$ for some $\mathcal{J} \subseteq \mathcal{I}$, if $\mathbb{P}\big((X_j)_{j \in \mathcal{I}} \in \mathcal{D} \mid X_j = x\big)$ is increasing in x for any increasing set $\mathcal{D} \subseteq \mathbb{R}^{|\mathcal{I}|}$ and $j \in \mathcal{J}$.
When this holds for $\mathcal{J} = \mathcal{I}$, we simply say $(X_i)_{i \in \mathcal{I}}$ is **PRD**.

Validity

Proposition

Fix some $i \in \mathcal{I}_0$ and suppose the following assumptions are satisfied:

1. $F_i(t) \leq F_j(t)$ for all $j \in \mathcal{I}_{\mathrm{nc}}$ and t;
2. $(F_j(T_j))_{j \in \{i\} \cup \mathcal{I}_{\mathrm{nc}}}$ is exchangeable.

Then the RANC p-value p_i is valid in the sense that $\mathbb{P}(p_i \leq \alpha) \leq \alpha$ for all $0 < \alpha < 1$.

- Allows null statistics to be conservative (or equivalent, negative controls to be anti-conservative). Useful for one-sided testing and misclassification of negative controls.

PRDS

Theorem

Suppose one of the two sets of conditions below holds:

1.
   1. $T_i \overset{d}{=} T_j$ for any $i \in \mathcal{I}_0$ and $j \in \mathcal{I}_{\text{nc}}$;
   2. $(T_i)_{i \in \mathcal{I}} \perp\!\!\!\perp (T_j)_{j \in \mathcal{I}_{\text{nc}}}$;
   3. $(T_j)_{j \in \mathcal{I}_{\text{nc}}}$ is mutually independent;
   4. $(T_i)_{i \in \mathcal{I}}$ is PRDS on $(T_i)_{i \in \mathcal{I}_0}$;
2. $(T_i)_{i \in \mathcal{I} \cup \mathcal{I}_{\text{nc}}}$ is exchangeable on $(T_i)_{i \in \mathcal{I}_0 \cup \mathcal{I}_{\text{nc}}}$.

Then the RANC p-values are valid and $(p_i)_{i \in \mathcal{I}}$ is PRDS on $(p_i)_{i \in \mathcal{I}_0}$.

- Allows some dependence between test statistics.
- Proof is based on the following heuristic: if we **swap any $T_i, i \in \mathcal{I}_0$ with the next smallest NC statistic**, the probability of $(p_i)_{i \in \mathcal{I}}$ is in an increasing set can only increase.

Multiple testing procedures

Enabled by validity

- FWER: Bonferroni's correction, Holm's procedure, graph-based procedures (fixed sequence, fall-back, etc.).

Enabled by validity + PRDS

- Global/intersection null: Simes' test;
- FWER: Hochberg-Hommel procedure (closure of Simes' test);
- FDP control: Lehmann-Romano step-down procedure;
- FDR control: Benjamini-Hochberg procedure.

- Remark: By using the monotonicity of Simes' test, the sufficient condition for it can be relaxed to stochastic dominance + exchangeability of $(F_i(T_i))_{i \in \mathcal{I}_0 \cup \mathcal{I}_{\mathrm{nc}}}$.

Empirical estimation of FDR

- Dates back to at least Storey, Taylor, and Siegmund (2004) and Genovese and Wasserman (2004). For simplicity, assume $T_i \in [0,1]$ for all i.
- The empirical processes for false rejections, all rejections, and the FDP are defined as

$$V(t) := \sum_{i \in \mathcal{I}_0} 1_{\{T_i \le t\}}, \quad R(t) := \sum_{i \in \mathcal{I}} 1_{\{T_i \le t\}}, \quad \text{and} \quad \mathrm{FDP}(t) := \frac{V(t)}{R(t) \vee 1}, \quad 0 \le t \le 1.$$

- Further define

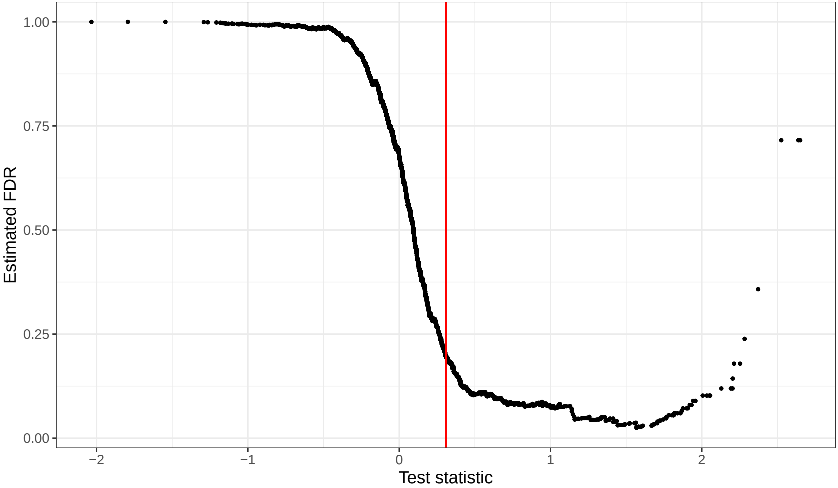
$$V_{\mathrm{nc}}(t) := \sum_{j \in \mathcal{I}_{\mathrm{nc}}} 1_{\{T_j \le t\}}, \quad \bar{V}_{\mathrm{nc}}(t) := \frac{n \cdot (V_{\mathrm{nc}}(t) + 2)}{m+1}, \quad 0 \le t \le 1.$$

- An estimator of $\mathrm{FDR}(t) = \mathbb{E}\left[\mathrm{FDP}(t)\right]$ based on negative controls is

$$\widehat{\mathrm{FDR}}_\lambda(t) := \frac{\hat{\pi}(\lambda) \cdot \bar{V}_{\mathrm{nc}}(t)}{R(t) \vee 1}, \quad \text{for some } 0 < \lambda \le 1,$$

where $\hat{\pi}(\lambda)$ is an estimator of the null proportion $\pi = |\mathcal{I}_0|/|\mathcal{I}|$ ($\hat{\pi}(\lambda) = 1$ when $\lambda = 1$).

An illustration using the proteomic dataset



Solid red: rejection threshold (FDR $= 0.2$).

Relation to method 1

In method 2, H_i is rejected if

$$T_i \leq \tau_q := \sup\left\{0 \leq t \leq \lambda : \widehat{\mathrm{FDR}}_\lambda(t) \leq q\right\}.$$

Proposition

A hypothesis H_i, $i \in \mathcal{I}$ is rejected by the above step-up procedure with $\lambda = 1$ if and only if it is rejected by the BH procedure with the following modified RANC p-values:

$$\tilde{p}_i = \frac{2 + \sum_{j \in \mathcal{I}_{\mathrm{nc}}} 1_{\{T_j \leq T_i\}}}{1 + m} \wedge 1.$$

- Extends the well known empirical process interpretation of the Benjamini-Hochberg procedure.

FDR control

Useful definition (Zhao, Small, and Su 2019)

For two random variables X, Y supported on $[0,1]$, we say X is **uniformly stochastically larger** than Y if $\mathbb{P}(X \le t) > 0$, $\mathbb{P}(Y \le t) > 0$, and $\mathbb{P}(X \le s \mid X \le t) \le \mathbb{P}(Y \le s \mid Y \le t)$ for all $0 < s \le t \le 1$.

Theorem

The above step-up procedure controls the FDR at level q under the following conditions:

1. T_i is uniformly stochastically larger than T_j for all $i \in \mathcal{I}_0$ and $j \in \mathcal{I}_{\text{nc}}$;
2. $(T_i)_{i \in \mathcal{I} \cup \mathcal{I}_{\text{nc}}}$ is mutually independent.

- Our proof extends Storey, Taylor, and Siegmund (2004) by showing the following is a backward super-martingale:

$$M(t) = \frac{V(t)}{(1 + V_{\text{nc}}(t))/(1+m)}.$$

- From there, the condition about uniformly stochastic dominance naturally arises.

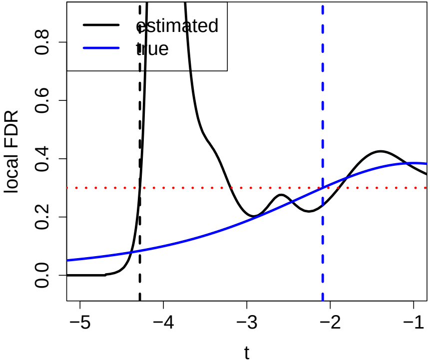
Local FDR

- Consider the two-mixture model: $(H_i, T_i), i = 1, \ldots, n$ are i.i.d., $H_i \sim \text{Bernoulli}(1 - \pi)$, $T_i \mid H_i \sim F_{H_i}$. So the marginal CDF is $F(t) = \pi F_0(t) + (1 - \pi) F_1(t)$.
- Let the corresponding density functions be f_0 and f_1.
- The local FDR at t is defined as (Efron *et al.* 2001)

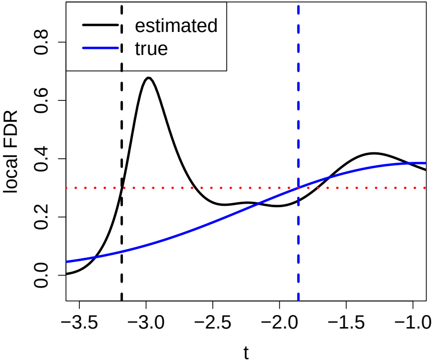
$$\text{local-FDR}(t) = \mathbb{P}(H_i = 0 \mid T_i = t) = \frac{\pi f_0(t)}{f(t)} = \frac{\pi f_0(t)}{\pi f_0(t) + (1 - \pi) f_1(t)}.$$

Estimation of local FDR: PDF-based methods

- Common to plug in estimator of the density function(s), but often doesn't work well.
- Example: $\pi = 0.5$, $F_0 = t_{10}$, $F_1 = Exp(1)$, $n = 400$, $m = 1000$, $q = 0.3$.



(a) Kernel density estimator.



(b) Kernel density estimator on z-scores.

A CDF-based method

- Solve the next optimization problem for some given $\lambda = q/\pi$ and $0 < q < 1$:

$$\hat{\tau}_{\lambda,n,m} = \arg\min_t F_{0,m}(t) - \lambda F_n(t),$$

  where $F_{0,m}$ and F_n are the empirical CDFs of $(T_i)_{i \in \mathcal{I}_{\text{nc}}}$ and $(T_i)_{i \in \mathcal{I}}$, respectively.
- Intuitively, $\hat{\tau}_{\lambda,n,m}$ should converge to

$$\tau^*_\lambda = \arg\min_t F_0(t) - \lambda F(t).$$

- By taking the derivative, we obtain

$$f_0(\tau^*_\lambda) = (q/\pi) f(\tau^*_\lambda) \Leftrightarrow \text{local-FDR}(\tau^*_\lambda) = q.$$

- The heuristic in Hung *et al.* (2014) corresponds to using $\lambda = 1$ or $q = \pi$, which is quite sensible!

Remarks

- By assuming f_0 and f are differentiable at τ_λ^* and $(f_0/f)'(\tau_\lambda^*) > 0$, we show in the paper that $\hat{\tau}_{\lambda,n,m} - \tau_\lambda^* = O_p((n \wedge m)^{-1/3})$. Such convergence is uniform over λ if (f_0/f) is monotone.
- The optimization can be rewritten as a decision-theoretic problem, which dates back to at least Sun and Cai (2007). Let $\hat{H}_i(t) = 1_{\{T_i \le t\}}$, then

$$\tau_\lambda^* = \arg\min_t \mathbb{E}_{F_0}\left[\hat{H}_i(t)\right] - \lambda \mathbb{E}_F\left[\hat{H}_i(t)\right] = \arg\min_t \mathbb{E}\left[(1-q)1_{\{H_i < \hat{H}_i(t)\}} + q 1_{\{H_i > \hat{H}_i(t)\}}\right].$$

- By using order stats. $p_{(1)} < \cdots < p_{(n)}$ of RANC p-values, the optimization can be rewritten as

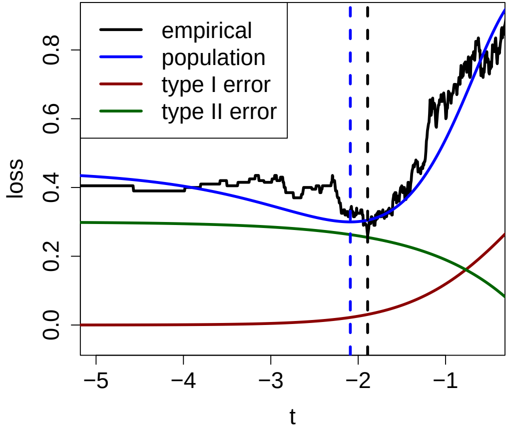
$$\hat{\tau}_{\lambda,n,m} = T_{(i^*)}, \text{ where } i^* = \arg\min_i \frac{m+1}{m} p_{(i)} - \frac{\lambda i}{n}.$$

  This can then be inverted to estimate the entire local-FDR curve:

$$\hat{q}(t) = \inf_q \{q : \hat{\tau}(q) \ge t\}.$$

  This is basically Grenander (1956)'s estimator of a monotone density function.

Illustration of the CDF-based method



- Even if the density functions doesn't exist, a simple argument shows that the regret is $O_p(n^{-1/2})$.

And the twist...

Shortly after completing the first draft of the paper, we discovered that all three methods have been independently proposed in the last 2 years:

- Method 1: Bates, Candès, Lei, Romano, and Sesia (2021) called this conformal p-value and considered outlier detection in machine learning outputs.
- Method 2: Mary and Roquain (2022) called this semi-supervised multiple testing and considered applications to astrostatistics.
- Method 3: Soloff, Xiang, and Fithian (2022) developed basically the same estimator.

Some distinctions

- Neither Bates *et al.* (2021) or Mary and Roquain (2022) paid attention to the possibility that the hypotheses could be one-sided/the negative controls could be misclassified.
- Soloff, Xiang, and Fithian (2022) worked with the conventional setup with known F_0 and showed that the method controls expected maximum local-FDR. It is unclear if this still holds when F_0 must be estimated.
- Some of our theoretical treatments that I skipped appear novel.
- Surprisingly, the connection between empirical null and conformal inference seems has never been pointed out.
- In particular, **negative control** is arguably a better name:
  - Highlights the nature of the method;
  - Can be be immediately understood by practitioners and be falsified (example in paper).

Lesson

Good methodological ideas may hide in

- the old literature;
- the most recent literature;
- the the literature that calls things different or doesn't make the expected citations;
- the practice.

Link to paper & slides: `http://www.statslab.cam.ac.uk/~qz280/publication/ranc`.

References

1. S. Bates *et al.*, *Annals of Statistics*, to appear, arXiv: 2104.08279v3 [stat.ME] (2021).
2. B. Efron, *Journal of the American Statistical Association* **99**, 96–104 (2004).
3. B. Efron *et al.*, *Journal of the American statistical association* **96**, 1151–1160 (2001).
4. J. A. Gagnon-Bartsch, T. P. Speed, *Biostatistics* **13**, 539–552 (2012).
5. C. Genovese, L. Wasserman, *The annals of statistics* **32**, 1035–1061 (2004).
6. U. Grenander, *Scandinavian Actuarial Journal* **1956**, 125–153 (1956).
7. V. Hung *et al.*, *Molecular Cell* **55**, 332–341 (2014).
8. J. T. Leek *et al.*, *Nature Reviews Genetics* **11**, 733–739 (2010).
9. M. Lipsitch, E. J. Tchetgen Tchetgen, T. Cohen, *Epidemiology* **21**, 383–388 (2010).
10. J. Listgarten *et al.*, *Bioinformatics* **29**, 1526–1533 (2013).
11. D. Mary, E. Roquain, *Electronic Journal of Statistics* **16** (2022).
12. W. Miao, Z. Geng, E. J. Tchetgen Tchetgen, *Biometrika* **105**, 987–993 (2018).
13. D. A. Nix, S. J. Courdy, K. M. Boucher, *BMC bioinformatics* **9**, 1–9 (2008).
14. M. M. Parks, B. J. Raphael, C. E. Lawrence, *BMC bioinformatics* **19**, 1–8 (2018).
15. A. L. Price *et al.*, *Nature Genetics* **38**, 904–909 (2006).
16. S. A. Shuster *et al.*, *Neuron* **110**, 1–14 (2022).
17. M. Slattery *et al.*, *Cell* **147**, 1270–1282 (2011).
18. J. A. Soloff, D. Xiang, W. Fithian, arXiv: 2207.07299 [stat.ME] (2022).
19. J. S. Song *et al.*, *Genome biology* **8**, 1–13 (2007).
20. J. D. Storey, J. E. Taylor, D. Siegmund, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **66**, 187–205 (2004).
21. W. Sun, T. T. Cai, *Journal of the American Statistical Association* **102**, 901–912 (2007).
22. Y. Zhang *et al.*, *Genome biology* **9**, 1–9 (2008).
23. Q. Zhao, D. S. Small, W. Su, *Journal of the American Statistical Association* **114**, 1291–1304 (2019).