Variance Moderation of Locally Efficient Estimators and Supervised Clustering with Applications in High-Dimensional Biology

Nima S. Hejazi, Mark J. van der Laan, & Alan E. Hubbard   *Graduate Group in Biostatistics*

School of Public Health
University of California, Berkeley

Overview

1. We introduce and implement a general approach for applying variance moderation techniques to locally efficient estimators in semiparametric statistical models.
2. The approach allows for such estimators to be utilized for differential expression analysis by stabilizing their small-sample properties.
3. Focusing on targeted maximum likelihood estimation (TMLE), we illustrate how the approach generalizes to influence function-based estimators.
4. We estimate the average treatment effect (ATE) in a study of occupational exposure to benzene, identifying **3280** significant genes after controlling the FDR at 5%.
5. We further illustrate that our focus on influence function-based estimators allows for supervised clustering.

Introduction & Data

- With the growing number of methods for measuring biomarkers there arises a need for methodologies able to simultaneously analyze multiple kinds of exposome data.
- Data was generated by the **Illumina Human Ref-8 BeadChips** platform.
- There were 125 subjects, for which background characteristics and expression measures for $\sim 22,000$ genes were obtained.
- Covariates in W were age, sex, and smoking status; all were discretized.
- The treatment (A) is degree of Benzene exposure: none, <1ppm, and >5ppm.
- The outcome (Y) is a vector of gene expression measures, normalized by median.

Methodology I: Variance Moderation & Asymptotic Linear

- Let observed data $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where W represents potential baseline confounders, A the exposure of interest, and $Y = (Y_b, b = 1, \dots, B)$ a vector of potential biomarkers.
- We consider, as an example, the *average treatment effect* (ATE), as the causal parameter of interest, which is identified by the observed data parameter:

$$\Psi_b(P_0) = \mathbb{E}_W[Q_0^b(A = 1, W) - Q_0^b(A = 0, W)], \tag{1}$$

where $Q_0^b(A, W) \equiv \mathbb{E}_{P_0}(Y_b \mid A, W)$ and may be estimated via *ensemble machine learning* [6, 1, 7].

- Similarly to the estimator $\hat{\beta}$ in a linear model, the $\Psi_b(P_n)$ is **asymptotically linear** (for Ψ_b) [4, 5]:

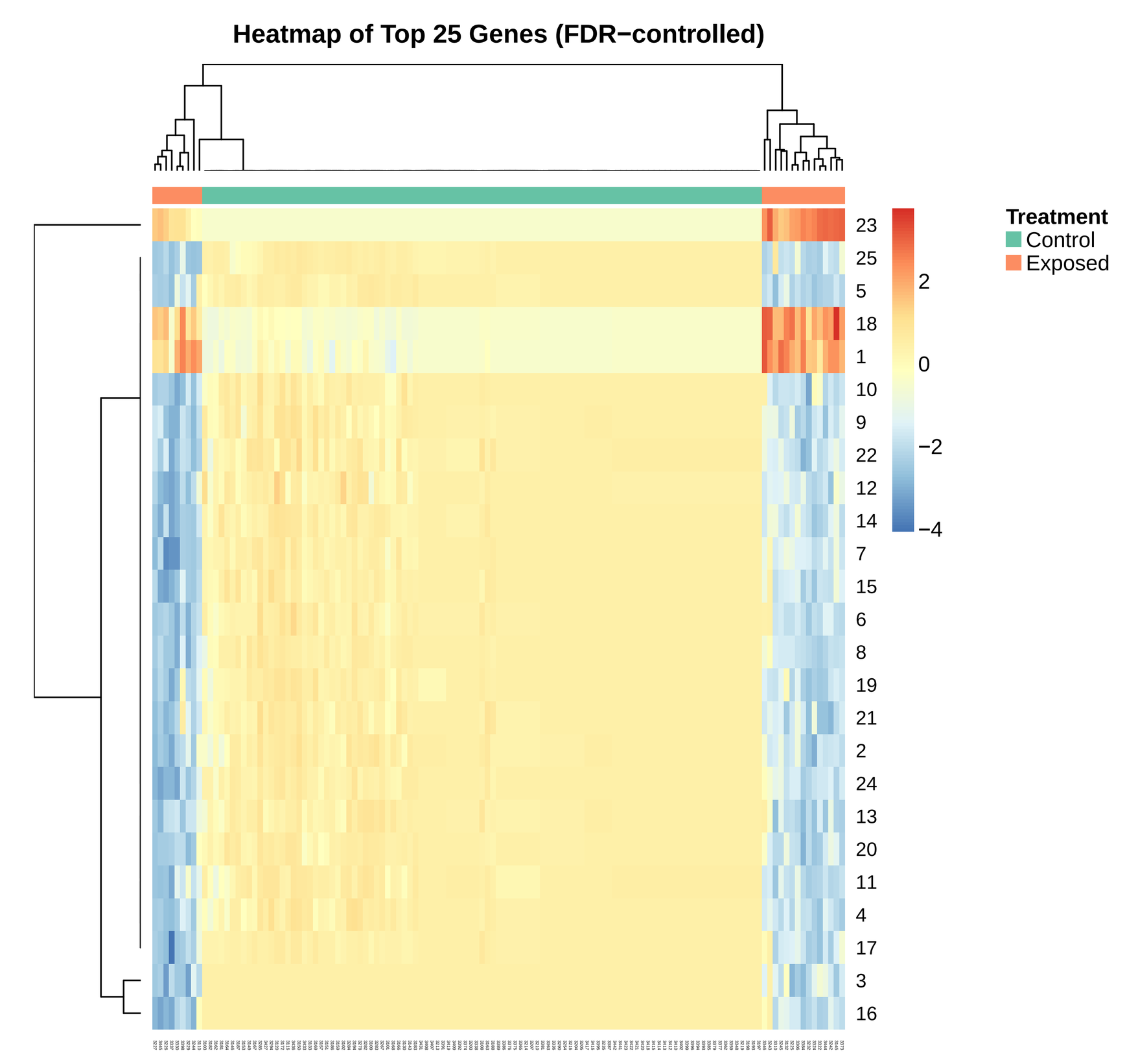
$$\sqrt{n}(\Psi_b(P_n) - \Psi_b(P_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_b(O_i) + o_p(1). \tag{2}$$

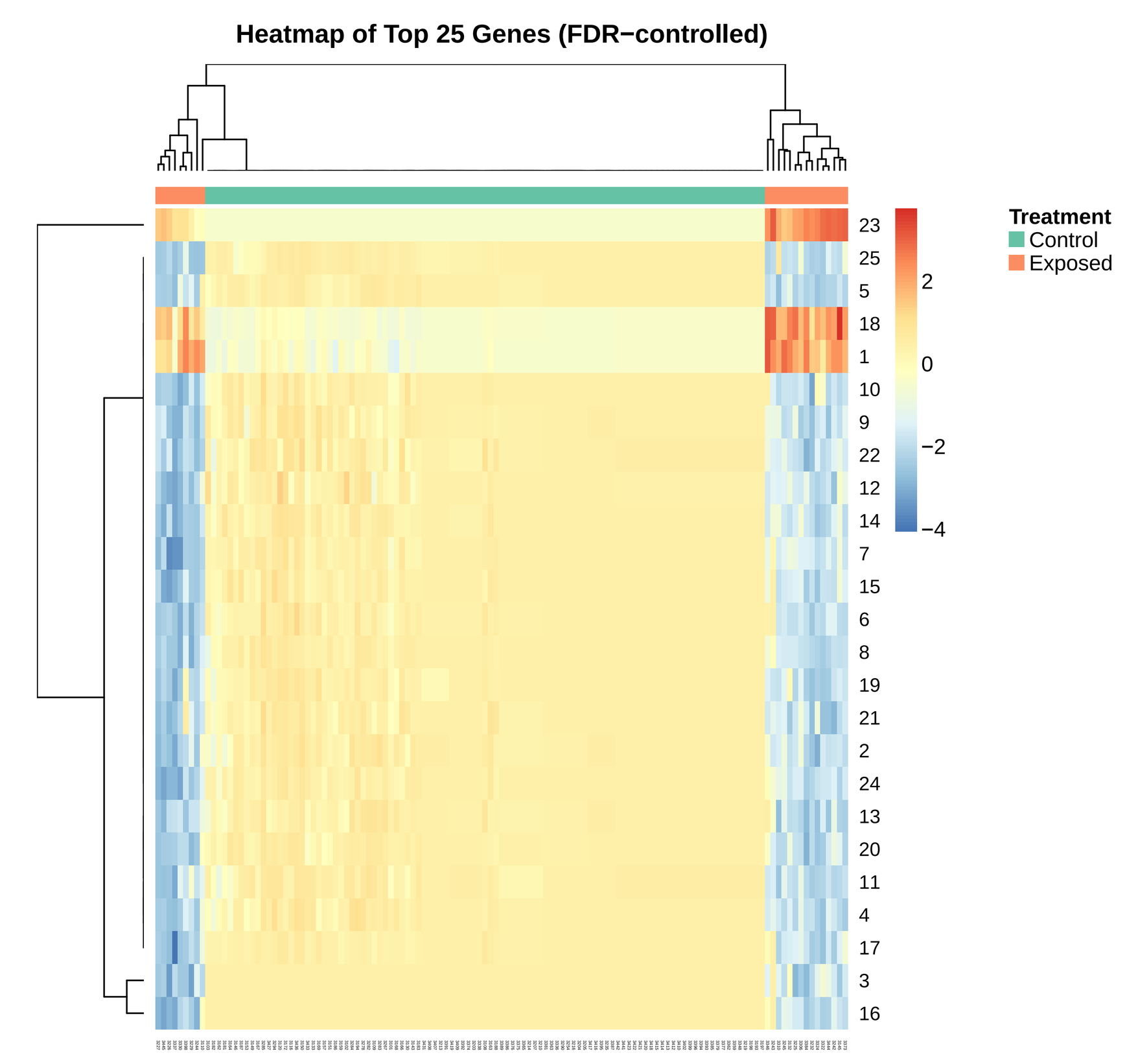
- Ψ_b has (efficient) influence function, relative to the nonparametric model $\mathcal{M}$:

$$D_b(P_0)(o) = \left(\frac{I(a = 1)}{g(1 \mid w)} - \frac{I(a = 0)}{g(0 \mid W)} \right) \cdot [y_b - Q_0^b(a, w)] + Q_0^b(1, w) - Q_0^b(0, w) - \Psi_b(P_0)(o). \tag{3}$$

- The moderated t-statistic [2, 3] may be applied readily to asymptotically linear estimators: $\tilde{t}_b = \frac{\sqrt{n}(\Psi_b(P_n) - \psi_0)}{S_b(D_{b,n})}$, where $\tilde{S}_{b,n}^2 = \frac{d_0 S_0^2 + d_b S_b^2(D_{b,n})}{d_0 + d_b}$ where d_b is the degrees of freedom for the b^{th} biomarker, d_0 is the degrees of freedom for the remaining $(B - 1)$ biomarkers, S_b is the standard deviation for the b^{th} biomarker and S_0 is the common standard deviation across all biomarkers.

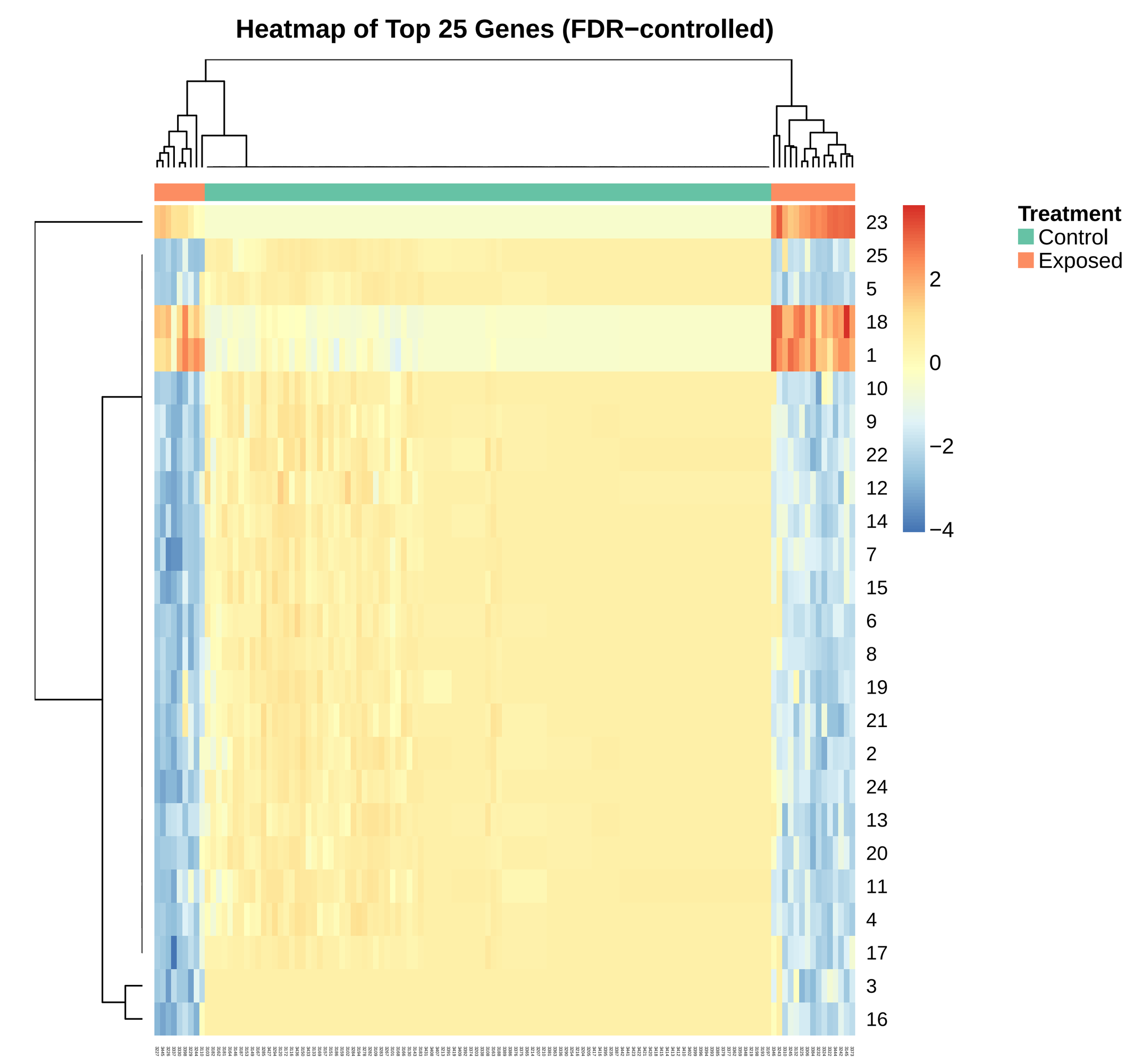
Methodology II: Supervised Distance Matrices

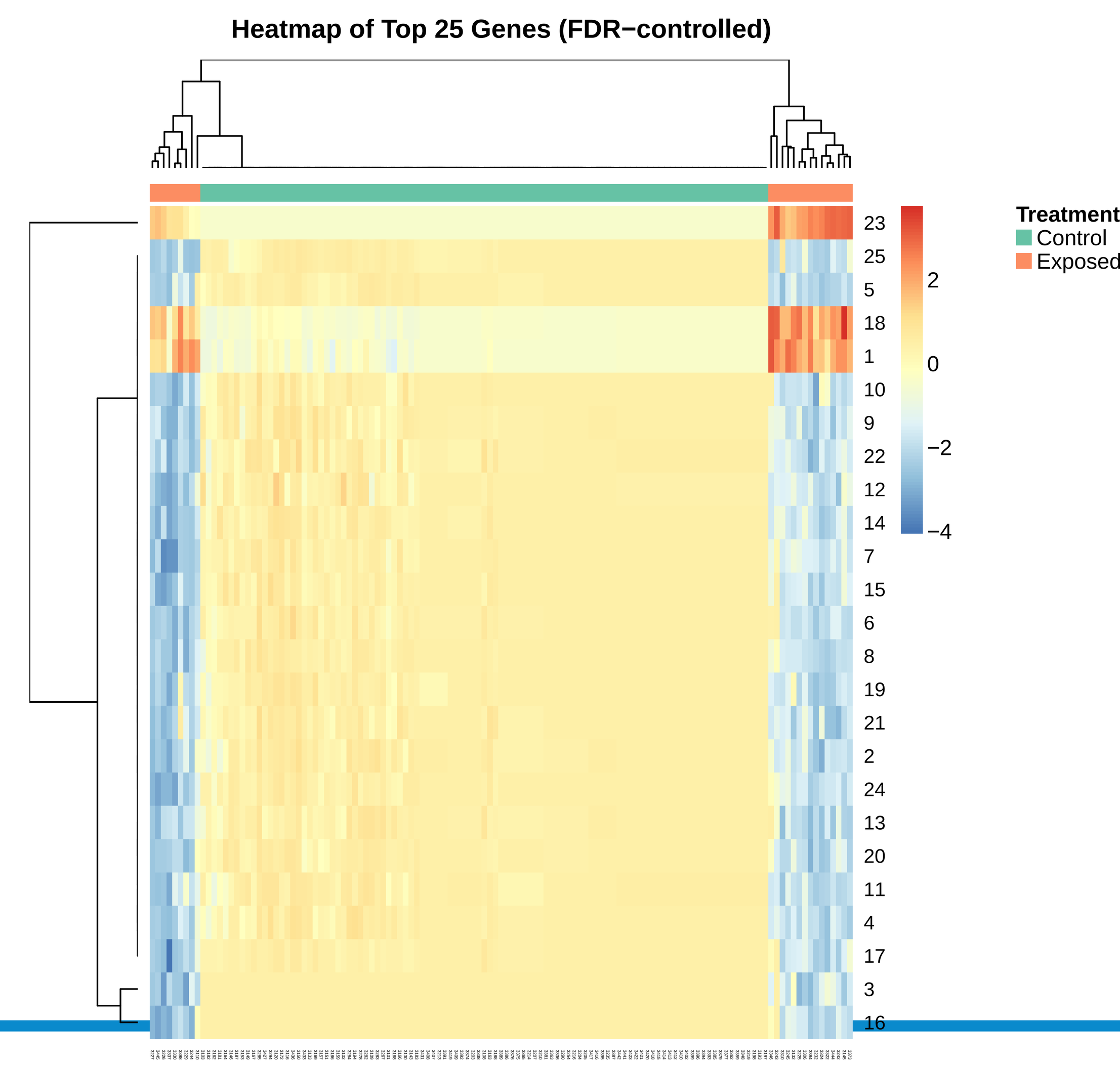




- The raw p-values are bimodally distributed, with a uniform distribution outside of the peaks, and clusters near 0 and 1.
- These raw p-values must be adjusted on account of the $\sim 22,000$ simultaneous tests.
- Using the Benjamini-Hochberg procedure to adjust for multiple comparisons yields an expected distribution of p-values.
- **3280** genes have Benjamini-Hochberg adjusted p-values falling below the 5% FDR.

Results & Discussion





- The heatmap visualizes the ATE difference induced by benzene exposure.
- The x-axis shows the 125 subjects, while the y-axis shows the top 25 genes showing highest differential ATE (based on BH adjusted p-values).
- Blue indicates a depression in the ATE, while red indicates an increase in the ATE, based

References

[1] Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.
[2] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3 (1):1–25, 2004.
[3] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.
[4] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.
[5] Mark J van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.
[6] Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super Learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.
[7] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

Contact Information

- **N.S. Hejazi**, Ph.D. student, Group in Biostatistics, NHEJAZI@BERKELEY.EDU
- **M.J. van der Laan**, Professor of Biostatistics & Statistics, LAAN@BERKELEY.EDU
- **A.E. Hubbard**: Professor of Biostatistics, HUBBARD@BERKELEY.EDU