# Targeted Learning with the Moderated T-Statistic

NIMA HEJAZI, ANDRE KUREPA WASCHKA, & MARY COMBS  Division of Biostatistics, UC Berkeley

School of Public Health
UNIVERSITY OF CALIFORNIA, BERKELEY

## OVERVIEW

1. In this project we introduce and implement a method to identify genes differentially expressed (based on the ATE) across subjects with varying levels of benzene exposure.
2. We use targeted maximum likelihood estimation (TMLE), relying on the influence curve of the proposed estimator, with the moderated t-statistic for gene expression.
3. The parameter of interest is the expected difference in gene expression if all subjects had received maximal benzene exposure as opposed to not.
4. We identify **3280** genes with (BH) adjusted p-values below the 5% FDR.

## 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

The procedure for **Targeted Learning with the Moderated T-Statistic** works as follows:

- Let $O = (W, A, Y) \sim P_0$, where W represents confounders, A the exposure of interest, and $Y = (Y_b, b = 1, \ldots, B)$ a vector of potential biomarkers. The proposed target parameter is $\Psi_b(P_0) = E_W[E_0(Y_b|A = 1, W) - E_0(Y_b|A = 0, W)]$.
- To estimate $\Psi$, define ${Q_0}^b(A, W) \equiv E_0(Y_b|A, W) \implies \Psi(P_n)_b = \frac{1}{n}\sum_{i=1}^{n} {Q_n}^b(1, W_i) - {Q_n}^b(0, W_i)$, where ${Q_n}^b$ represents an initial estimate of ${Q_0}^b$ (later referred to as ${Q_0}^{(b,0)}$). Super Learner is applied to derive an initial estimate of ${Q_0}^b$.
- The TMLE estimate of $\Psi_b$ is: $\hat{\Psi}_b(P_n) = \frac{1}{n}\sum_{i=1}^{n}[Q_n^{(b,1)}(1, W_i) - Q_n^{(b,1)}(0, W_i)]$, where $Q_n^{(b,a)}$ is a main terms logistic regression. Thus, $Q_n^{(b,1)}(A, W)] = Q_n^{(b,0)}(A, W)] + \epsilon h_{\hat{g}}(A, W)$. The initial Super Learner fit $Q_n^{(b,0)}(A, W)]$ is treated as an offset and $h_{\hat{g}}(A, W) = (\frac{I(A=1}{\hat{g}(1|W)} - \frac{I(A=0)}{\hat{g}(0|W)})$.
- $\hat{\Psi}_b(P_n)$ is an asymptotically linear estimator of $\Psi_b$ [1] with influence curve $IC(O_i)$ if it satisfies: $\sqrt{n}(\Psi_b(P_n) - \Psi_b(P_0)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC(O_i) + o_p(1)$. Based on this, the plug-in IC for the ATE is:

  $IC_{b,n}(O_i) = (\frac{I(A_i=1)}{g_n(1|W_i)} - \frac{I(A_i=0)}{g_n(0|W_i)})(Y_{b,i} - Q_n^{(b,1)}(A_i, W_i)) + Q_n^{(b,1)}(1, W_i) - Q_n^{(b,1)}(0, W_i) - \Psi_b(P_n)$
- The moderated t-statistic [2] for an asymptotically linear parameter estimate: $\tilde{t}_j = \frac{\sqrt{n}(\Psi_j(P_n) - \psi_0)}{S_j(IC_{j,n})}$. Our goal is to define $\Psi$ as the difference (per gene) in outcome between receiving the maximum and minimum levels of treatment. Let: $\Psi_j$*=$E[E[Y_j \mid A = max(A), W] - E[Y_j \mid A = min(A), W]]$.
- The moderated t-statistic based on this parameter is $\tilde{t}_j = \frac{\sqrt{n}(\hat{\Psi}_{j,n}^{max} - \hat{\Psi}_{j,n}^{\neq max})}{\tilde{S}_{j,n}^2}$ where $\tilde{S}_{j,n}^2 = \frac{d_0 S_0^2 + d_j S_j^2(IC_{j,n})}{d_0 + d_j}$ where $d_j$ is the degrees of freedom for the $j^{th}$ gene, $d_0$ is the degrees of freedom for the remaining genes, $S_j$ is the standard deviation for the $j^{th}$ gene and $S_0$ is the common standard deviation across all genes towards which empirical Bayes performs shrinkage.

## RESULTS

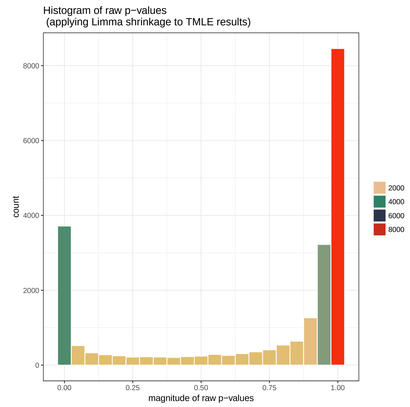


**Figure 1:** raw p-values from applying Limma

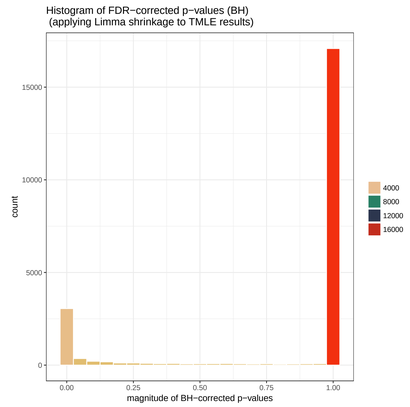


**Figure 2:** BH-corrected p-values from applying Limma

- 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.

## DISCUSSION & CONCLUSIONS

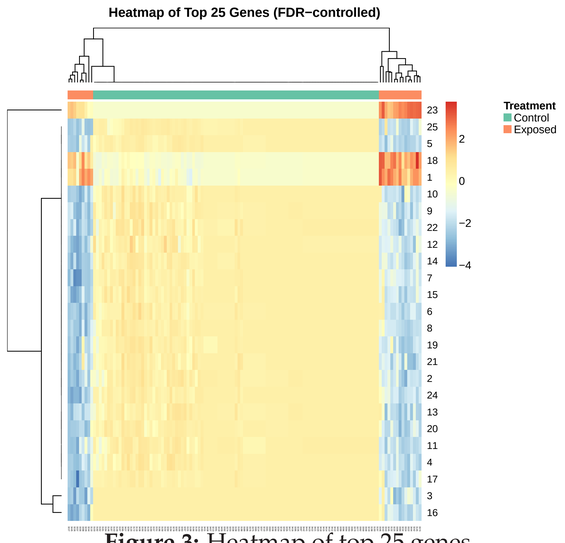


**Figure 3:** Heatmap of top 25 genes

- 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 on exposure to the maximal level of benzene as opposed to not.
- The results of our analysis indicate that the moderated t-statistic applied to the ATE constitutes a powerful approach for assessing variable importance (based on exposure) in the context of high-dimensional investigations of biomarkers

## REFERENCES

[1] Mark J van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.

[2] Gordon K Smyth. Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer, 2005.

## ACKNOWLEDGEMENTS

We thank Prof. Alan E. Hubbard for his generous guidance and support throughout this project.