Data-Adaptive Estimation and Inference in the Analysis of Differential Methylation

for the annual retreat of the *Center for Computational Biology*,
given 18 November 2017

Nima Hejazi

Division of Biostatistics
University of California, Berkeley
stat.berkeley.edu/~nhejazi

nimahejazi.org
twitter/@nshejazi
github/nhejazi

slides: goo.gl/xabp3Q

This slide deck is for a brief (about 15-minute) talk on a new statistical algorithm for using nonparametric and data-adaptive estimates of variable importance measures for differential methylation analysis. This talk was most recently given at the annual retreat of the Center for Computational Biology at the University of California, Berkeley.

Source: `https://github.com/nhejazi/talk_methyvim`
Slides: `https://goo.gl/JDhSEg`
With notes: `https://goo.gl/xabp3Q`

Preview: Summary

- DNA methylation data is *extremely* high-dimensional — we can collect data on 850K genomic sites with modern arrays!
- Normalization and QC are critical components of properly analyzing modern DNA methylation data. There are many choices of technique.
- A relative scarcity of techniques for estimation and inference exists — analyses are often limited to the general linear model.
- Statistical causal inference provides an avenue for answering richer scientific questions, especially when combined with modern advances in machine learning.

We'll go over this summary again at the end of the talk. Hopefully, it will all make more sense then.

Motivation: Let's meet the data

- Observational study of the impact of disease state on DNA methylation.
- Phenotype-level quantities: 216 subjects, binary disease status (FASD) of each subject, background info on subjects (e.g., sex, age).
- Genomic-level quantities: $\sim 850,000$ CpG sites interrogated using the *Infinium MethylationEPIC BeadChip* by Illumina.
- **Questions**: How do disease status and differential methylation relate? Is a coherent biomarker-type signature detectable?

- FASD is an abbreviation for Fetal Alcohol Spectrum Disorders.
- We're mostly interested in the interplay between disease and DNA methylation.
- In particular, we'd like to construct some kind of importance score for CpG sites impacted by the exposure/disease of interest.
- Re: dimensionality, c.f., RNA-seq analyses are $\sim 30,000$ in dimension at the gene level.

DNA Methylation

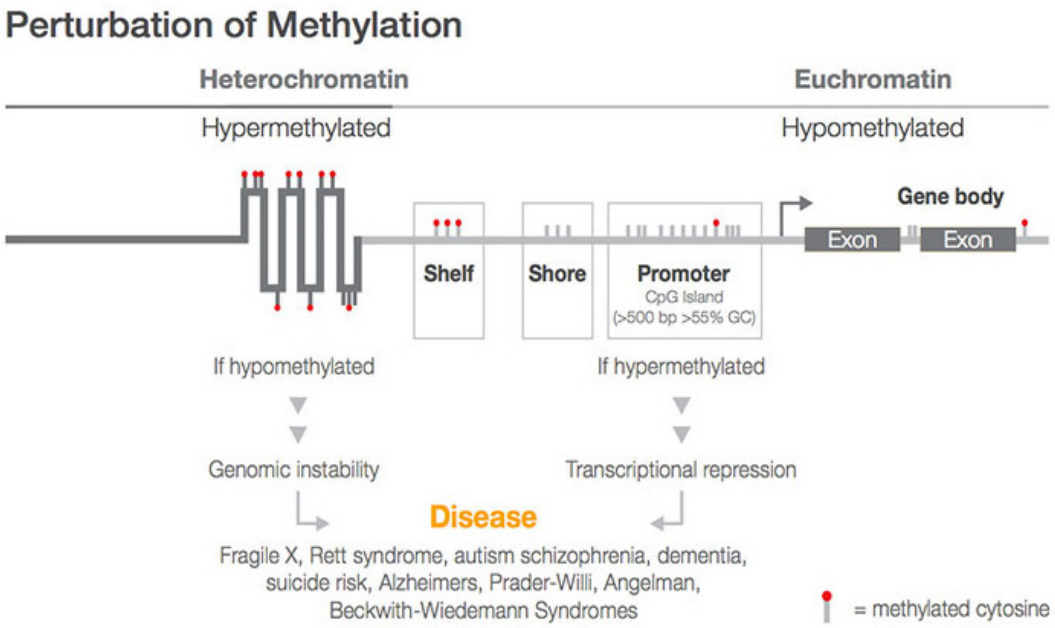


Figure: `https://www.illumina.com/techniques/sequencing/methylation-sequencing.html` (source)

- The biology of these structures is quite complicated: many different structures — CpG islands, shores, etc.
- Both hypermethylation and hypomethylation are linked to disease states.
- It's especially important to examine these processes, keeping in mind that the process may be complex and non-monotonic.

Data analysis? Linear Models!

- Standard operating procedure: For each CpG site ($g = 1, \ldots, G$), fit a linear model:

$$\mathbb{E}[y_g] = X\beta_g$$

- Test the coefficent of interest using a standard t-test:

$$t_g = \frac{\hat{\beta}_g - \beta_{g,H_0}}{s_g}$$

- Such models are a matter of convenience: does $\hat{\beta}_g$ answer our scientific questions? Perhaps not.
- Is consideration being given to whether the data could have been generated by a linear model? Perhaps not.

- CpG sites are thought to function in networks. Treating them as acting independently is not faithful to the underlying biology.
- The linear model is a great starting point for analyses whne the data is generated using complex technology — no need to make the analysis more complicated.
- That being said, the data is difficult and expensive to collect, so why restrict the scope of the questions we'd like to ask.

Motivation: Science Before Statistics

```
What is the effect of disease status
on DNA methylation at a specific CpG
site, controlling for the observed
methylation status of the neighbors
of the given CpG site?
```


- Again, CpG sites are thought to function in networks. Treating them as acting independently is not faithful to the underlying biology.
- This means that we should take into account the methylation status of neighboring CpG sites when assessing differential methylation at a single site.
- This is a coherent scientific question that we can set out to answer statistically. It's motivated by the established science and possible to do with modern statistical methodology.

Data analysis? A Data-Adaptive Approach

1. Isolate a subset of CpG sites for which there is cursory evidence of differential methylation.
2. Assign CpG sites into neighborhoods (e.g., bp distance). If there are many neighbors, apply clustering (e.g., *PAM*) to select a subset.
3. Estimate *variable importance measure* (VIM) at each screened CpG site, with disease as intervention (A) and controlling for neighboring CpG sites (W).
4. Apply a variant of the Benjamini & Hochberg method for FDR control, accounting for initial screening.

- Pre-screening is a critical step since we cannot perform computationally intensive estimation on all the sites. This is flexible — just use your favorite method (as long as allows a ranking to be made).
- The variable importance step merely comes down to the creation of a score. We use TMLE to statistically estimate parameters from causal models. The procedure is general enough to accomodate any inference technique.

Pre-Screening — Pick Your Favorite Method

- The estimation procedure is computationally intensive — apply it only to sites that appear promising.
- Consider estimating univariate (linear) regressions of intervention on CpG methylation status. Fast, easy.
- Select CpG sites with a marginal p-value below, say, 0.01. Apply data-adaptive procedure to this subset.
- The modeling assumptions do not matter since the we won't be pursuing inference under such a model.
- Software implementation is extensible. Users are encouraged to add their own. (It's easy!)

- We'll be adding to the available routines for pre-screening too! For now, we have limma, and more are on the way.

Too Many Neighbors? Clustering

- There are many options: *k*-means, *k*-medoids, etc., as well as many algorithmic solutions.
- For convenience, we use Partitioning Around Medoids (PAM), a well-established algorithm.
- With limited sample sizes, the number of neighboring sites that may be controlled for is limited.
- To faithfully answer the question of interest, choose the neighboring sites that are the most representative.
- This is an *optional* step — it need only be applied when there is a large number of CpG sites in the neighborhood of the target CpG site.

- The number of sites that we can control for is roughly a function of sample size. This impacts the definition of the parameter that we estimate, and allows enough flexibility to obtain either very local or more regional estimates.

Nonparametric Variable Importance

- Let's consider a simple target parameter: the average treatment effect (ATE):

$$\Psi_g(P_0) = \mathbb{E}_{W,0}[\mathbb{E}_0[Y_g \mid A = 1, W_{-g}] - \mathbb{E}_0[Y_g \mid A = 0, W_{-g}]]$$

- Under certain (untestable) assumptions, interpretable as difference in methylation at site g with intervention and, possibly contrary to fact, the same under no intervention, controlling for neighboring sites.
- Provides a *nonparametric* (model-free) measure for those CpG sites impacted by a discrete intervention.
- Let the choice of parameter be determined by our scientific question of interest.

By allowing scientific questions to inform the parameters that we choose to estimate, we can do a better job of actually answering the questions of interest to our collaborators. Further, we abandon the need to specify the functional relationship between our outcome and covariates; moreover, we can now make use of advances in machine learning.

Target Minimum Loss-Based Estimation

- We use *targeted minimum loss-based estimation* (TMLE), a method for inference in semiparametric infinite-dimensional statistical models.
- No need to specify a functional form or assume that we know the true data-generating distribution.
- **Asymptotic linearity:**

$$\Psi_g(P_n^*) - \Psi_g(P_0) = \frac{1}{n}\sum_{i=1}^{n} IC(O_i) + o_P\left(\frac{1}{\sqrt{n}}\right)$$

- **Limiting distribution:**

$$\sqrt{n}(\Psi_n - \Psi) \to N(0, Var(D(P_0)))$$

- **Statistical inference:**

$$\Psi_n \pm z_\alpha \cdot \frac{\sigma_n}{\sqrt{n}}$$

Under the additional condition that the remainder term $R(\hat{P}^*, P_0)$ decays as $o_P\left(\frac{1}{\sqrt{n}}\right)$, we have that $\Psi_n - \Psi_0 = (P_n - P_0) \cdot D(P_0) + o_P\left(\frac{1}{\sqrt{n}}\right)$, which, by a central limit theorem, establishes a Gaussian limiting distribution for the estimator, with variance $V(D(P_0))$, the variance of the efficient influence curve (canonical gradient) when Ψ admits an asymptotically linear representation.

The above implies that Ψ_n is a $\sqrt{n}$-consistent estimator of Ψ, that it is asymptotically normal (as given above), and that it is locally efficient. This allows us to build Wald-type confidence intervals, where σ_n^2 is an estimator of $V(D(P_0))$. The estimator σ_n^2 may be obtained using the bootstrap or computed directly via $\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} D^2(\bar{Q}_n^*, g_n)(O_i)$

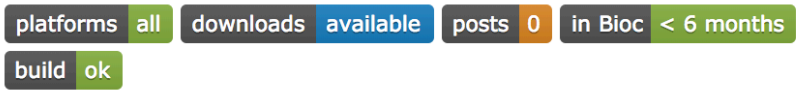
Corrections for Multiple Testing

- Multiple testing corrections are critical. Without these, we systematically obtain misleading results.
- The Benjamini & Hochberg procedure for controlling the False Discovery Rate (FDR) is a well-established technique for addressing the multiple testing issue.
- We use a modified BH-FDR procedure to account for the pre-screening step of the proposed algorithm.
- This modified BH-FDR procedure for multi-stage analyses (FDR-MSA) works by adding a p-value of 1.0 for each site that did not pass pre-screening then performs BH-FDR as normal.

- Note that $\text{FDR} = \mathbb{E}\left[\frac{V}{R}\right] = \mathbb{E}\left[\frac{V}{R} \mid R > 0\right] P(R > 0)$.
- BH-FDR procedure: Find $\hat{k} = max\{k : p_{(k)} \leq \frac{k}{M} \cdot \alpha\}$
- FDR-MSA will only incur a loss of power if the initial screening excludes variables that would have been rejected by the BH procedure when applied to the subset on which estimation was performed.
- BH-FDR control is a rank-based procedure, so we must assume that the pre-screening does not disrupt the ranking with respect to the estimation subset, which is provably true for screening procedures of a given type.
- MSA controls type I error with any procedure that is a function of only the type I error itself — e.g., FWER. This does not hold for the FDR in complete generality.

Software package: R/methyvim

methyvim



DOI: 10.18129/B9.bioc.methyvim

Differential Methylation Analysis with Targeted Minimum Loss-Based Estimates of Variable Importance Measures

Bioconductor version: Release (3.6)

Figure: `https://bioconductor.org/packages/methyvim`

- Variable importance for discrete interventions.
- Future releases will support continuous interventions.
- Take it for a test drive!

- Contribute on GitHub.
- Reach out to us with questions and feature requests.

Data analysis the methyvim way

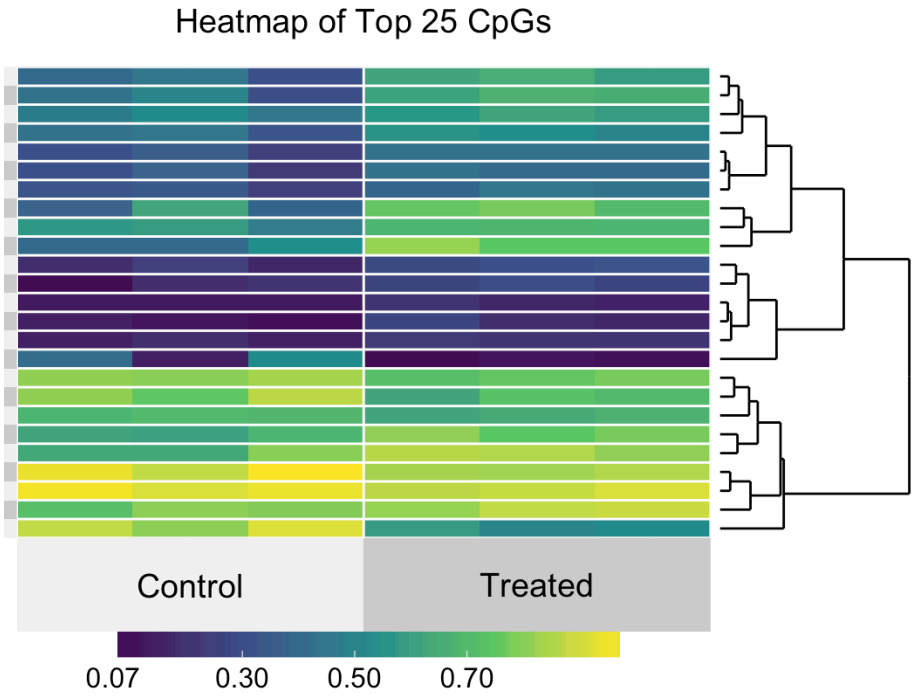


Figure: `http://code.nimahejazi.org/methyvim`

Review: Summary

- DNA methylation data is *extremely* high-dimensional — we can collect data on 850K genomic sites with modern arrays!
- Normalization and QC are critical components of properly analyzing modern DNA methylation data. There are many choices of technique.
- A relative scarcity of techniques for estimation and inference exists — analyses are often limited to the general linear model.
- Statistical causal inference provides an avenue for answering richer scientific questions, especially when combined with modern advances in machine learning.

It's always good to include a summary.

References I

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25.

Smyth, G. K. (2005). LIMMA: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer Science & Business Media.

References II

Tuglus, C. and van der Laan, M. J. (2009). Modified FDR controlling procedure for multi-stage analyses. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–15.

van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media.

van der Laan, M. J. and Rose, S. (2017). *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Science & Business Media.

Acknowledgments

Collaborators:
Mark van der Laan — University of California, Berkeley
Alan Hubbard
Lab of Martyn Smith
Lab of Nina Holland
Rachael Phillips

Funding source:
National Library of Medicine (of NIH): T32LM012417

Thank you.

Slides: goo.gl/JDhSEg

Notes: goo.gl/xabp3Q

Source (repo): goo.gl/m5As73

`stat.berkeley.edu/~nhejazi`

`nimahejazi.org`

`twitter/@nshejazi`

`github/nhejazi`