October 2017

# Multiple Testing Correction with Repeated Correlated Outcomes: Applications to Epigenetics

Katie Leap

*University of Massachusetts Amherst*

# Multiple Testing Correction with Repeated Correlated Outcomes:

# Applications to Epigenetics

A Thesis Presented

by

KATRINA N. LEAP

Submitted to the Graduate School of the

University of Massachusetts Amherst in partial fulfillment

of the requirements for the degree of

MASTER OF SCIENCE

September 2017

Public Health

Biostatistics

Multiple Testing Correction with Repeated Correlated Outcomes:

Applications to Epigenetics

A Thesis Presented

by

KATRINA N. LEAP

Approved as to style and content by:

Kenneth P. Kleinman, Chair

Brian W. Whitcomb, Member

Matthias Steinrücken, Member

Susan E. Hankinson, Department Head

Department of Biostatistics & Epidemiology

# ACKNOWLEDGEMENTS

I would first like to thank my advisor and thesis chair, Ken Kleinman, ScD, for bringing this project to my attention and always having my research interests in mind, even if they bore him to death. I am so grateful for his consistent guidance and lenient deadlines; the combination of the two has enabled this work to be completed.

I would like to thank all of my collaborators: Marie-France Hivert and Andres Cardenas for their insight into underlying epigenetic mechanisms and the proper framework to think about this problem; Sam Lent for coming up with the research question that prompted this study; Cat Briggs for her tireless data analysis and for coming into work on the weekend to get this done in time; Tim Adams for his help in optimizing the data generation code and saving me months of computational time.

I would also like to thank my committee members Brian Whitcomb and Matthias Steinrücken for their timely help shaping my approach at the beginning and refining the details at the end, and for their patience through the lack of updates in the middle.

Lastly I would like to thank my friends and family. Thank you for listening to me rant about the cluster, for persevering through math explanations of any length that did not interest you, for overlooking my constantly divided attention, and for asking the questions I forget to ask. A special shout out to Brook Frye, Justin Baldwin, and Mahim Naveed for help with my code and sanity; to Jess Buske for making sure my housework was attended to at least every other month; and to Sarah Murcek, Meghan Weakley, and Rebekah Whittier for sticking with me through the ups and downs and reminding me of all the jokes that matter.

# ABSTRACT

MULTIPLE TESTING CORRECTION WITH REPEATED CORRELATED OUTCOMES:

APPLICATIONS TO EPIGENETICS

SEPTEMBER 2017

KATRINA N. LEAP, B.I.A., CHATHAM UNIVERSITY

M.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Kenneth P. Kleinman

Epigenetic changes (specifically DNA methylation) have been associated with adverse health outcomes; however, unlike genetic markers that are fixed over the lifetime of an individual, methylation can change. Given that there are a large number of methylation sites, measuring them repeatedly introduces multiple testing problems beyond those that exist in a static genetic context. Using simulations of epigenetic data, we considered different methods of controlling the false discovery rate. We considered several underlying associations between an exposure and methylation over time.

We found that testing each site with a linear mixed effects model and then controlling the false discovery rate (FDR) had the highest positive predictive value (PPV), a low number of false positives, and was able to differentiate between differential methylation that was present at only one time point vs. a persistent relationship. In contrast, methods that controlled FDR at a single time point and ad hoc methods tended to have lower PPV, more false positives, and/or were unable to differentiate these conditions.

Validation in data obtained from Project Viva found a difference between fitting longitudinal models only to sites significant at one time point and fitting all sites longitudinally.

# TABLE OF CONTENTS

Page

CHAPTER

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

Epigenetic processes regulate how genes are expressed without changing the genetic sequence; DNA methylation is a mechanism that can turn a gene on or off and has been implicated in many diseases (Jirtle & Skinner, 2007). Although the genetic sequence of an individual is identical in all cells and stable over their lifetime, epigenetic markers like methylation are vital to cell differentiation and thus vary from cell to cell (Franchini, Schmitz, & Petersen-Mahrt, 2012). Monozygotic twin studies have shown cross-sectionally that while younger twin pairs have identical methylation patterns, the patterns of older twins diverge, indicating that methylation changes over time (Fraga, et al., 2005). Previous studies have examined the relationship between early exposures and site-specific differential methylation in a cross-sectional manner and some have studied global methylation levels within-person over time, but the effect of early exposures on site-specific methylation later in life is poorly understood.

Although epigenetic methods draw heavily on those used for genome-wide association studies (GWAS), the changeable nature of epigenetic markers necessitates methods that draw from longitudinal analysis. Despite using similar methods, genetic and epigenetic data differ in a few key ways. Genetic loci can consist of one of four nucleotides, but methylation is a binary indicator. With GWAS data, the many genetic loci are considered to be predictors of an outcome, but with epigenetic data, methylation status is considered the outcome and the disease status or exposure is the predictor. While multiple testing correction for genetic applications can be addressed by controlling the Benjamini-Hochberg False Discovery Rate (FDR) for multiple tests of association, epigenetic data introduces a longitudinal aspect previously unexplored. Furthermore, it is highly unlikely that methylation status at any point is independent of its previous status and we would expect that it would be correlated over time within-person.

Our purpose in this article is to explore multiple approaches to this problem and evaluate them using simulation studies and then apply them to a particular research question motivated by Project Viva, a prospective cohort study established to study the effect of prenatal diet on maternal and child health well-documented elsewhere (Oken, et al., 2015). Because of the correlation structure within-person over time, we use a linear mixed effects model to detect associations between an exposure and a large number of outcomes: methylation at each site. However, this approach is computationally expensive, so we will explore methods to decrease the number of such models that need to be fit. Methods will be evaluated based on whether they can detect whether associations persist over time and whether they can control the FDR adequately.

# CHAPTER 2

# BACKGROUND

For the purposes of this analysis, we are considering the longitudinal relationship between a single point exposure and multiple outcomes. Each individual has one exposure measure regardless of how many outcomes are measured. The outcomes are site-specific methylation at $T$ different time points. At each observation time, there are $M$ measured sites. Thus we can consider each site at each time point individually, allowing for $M \times T$ possible associations, or we can construct a longitudinal model of the $M$ sites, generating $M$ associations.

## Methylation

Methylation refers to the addition of a methyl group ($CH_3$) to cytosine, one of the amino acids in DNA; methylation status is therefore binary: either the cytosine has a methyl group or it does not. However, methylation is not typically assessed on single cells, but rather on a mixture of cells. Because of this, methylation outcomes represent a proportion of cells in the sample that are methylated at any given site; these methylation levels are Beta-distributed and termed Beta-values (Du, et al., 2010). As the Beta-values will be heteroscedastic, or having a non-constant variance, we can transform them to M-values, defined as $M_i = \log_2\left(\frac{Beta_i}{1-Beta_i}\right)$ (Du, et al., 2010). In addition to addressing the issue of constant variance, this transformation means that the values are no longer bounded between 0 and 1, which makes them more suited for analysis with linear models.

## Methylation over time

Previous studies have been done on global methylation within-person over time. One study compared cord blood to blood collected at 3 years in 165 children and found that while on average, methylation was higher in the second time point, within-person differences were lower. Furthermore, 62% of individuals' methylation changed very little, 12% decreased and 27% increased (Herbstman, et al., 2013). Another study of 77 adults comparing blood draws an average of 8 years apart found that about a third had large changes in global methylation (≥ 10%); 19.5% were increases and 13.0% were decreases (Wu, Wang, Delgado-Cruzata, Santella, & Terry, 2012). A third study examined longitudinal global methylation changes within-individuals clustered by families; in one of their cohorts, 30% had a ≥10% change in methylation over time and the other cohort had 18%. Additionally, they found a familial clustering effect on methylation changes (Bjornsson, et al., 2008).

A study similar to the one motivating this one examined the effect of both birth weight and gestational age on longitudinal site-specific DNA methylation. They used cord blood as their first time point and followed up at age 7 and 17; their statistical analysis consisted of identifying sites associated with the predictor in the first time point and conducting longitudinal analysis of only those sites (Simpkin, et al., 2015).

## Linear mixed effects models

In order to account for correlation within individuals and to evaluate the effect of time on the relationship between a predictor and a repeated outcome, we can use a linear mixed effects (LME) model. While linear regression forces a common intercept and slope for all individuals, mixed effects models allow individuals to have different random intercepts and/or slopes. This means that while the effect of the predictor may be similar across individuals, some individuals start with a higher or lower baseline and can have an individual deviance from the effect of the predictor. We use the `lme4` package in R to fit LME models for the simulation studies (Bates, Maechler, Bolker, & Walker, 2015) and `robustlmm` for the applied analysis (Koller, 2016).

## Controlling the false discovery rate

The False Discovery Rate (FDR) was defined by Benjamini and Hochberg in 1995 in terms of the following variables:

**Table 1: Relationship between $m_0$ and R**

| | Declared Non-Significant | Declared Significant | Total |
|---|---|---|---|
| **True Null Hypotheses** | **U** | **V** | $m_0$ |
| **Non-True Null Hypotheses** | **T** | **S** | $m - m_0$ |
| | $m$ - **R** | **R** | $m$ |

$m$ represents the null hypotheses being tested simultaneously; in our setting, $m$ is 470,870 or the number of sites for which we have methylation outcomes. **R** is observable and represents all of the sites that our method returns as significantly associated with the predictor, but all of the other variables are not observable. The rate at which we make false discoveries is $\boldsymbol{Q} = \frac{\boldsymbol{V}}{\boldsymbol{R}}$, where **V** is our Type I error.

The method for controlling the FDR is an algorithm that orders the p-values found by the test of the null hypothesis and then compares them iteratively to a function of the rank of the p-value and the desired controlled FDR value. More details can be found in the paper, but important to note is that this procedure rejects null hypotheses based on both the number of hypotheses there are total and their ordered values; it does not adjust the p-values as a Bonferroni correction would (Benjamini & Hochberg, 1995). However, we use the `p.adjust` function in R, which changes the p-value to the q-value and allows us to filter the sites as we might if the p-values had been adjusted (R Core Team, 2017).

## Motivating example: Project Viva

Our study was motivated by an analysis of the effect of maternal glucose levels on a child’s DNA methylation, as a hypothesized mechanism of the observed relationship between a mother’s gestational diabetes mellitus and the risk of obesity, cardiovascular disease, and type 2 diabetes later in the child’s life. To this end, exclusion criteria included mothers with diagnosed diabetes, as well as those who were underweight, premature births, and non-white children. Glucose was measured as non-fasting 50-gram glucose challenge test in trimester 2.

Preliminary analysis showed a relationship between the maternal glucose and methylation status at the first time point at different locations, but it is unknown whether these relationships continue over time. Methylation was assessed using the Illumina 450k Assay: there are about 470,870 locations measured. There are 660 observations on 430 children over three time points: in the cord blood, in early childhood (~ 3 years old) and in mid-childhood (~ 7 years old).

## Plausible underlying associations

The longitudinal relationship between exposure and outcome could take on several different shapes. In order to assess the methods' abilities to detect underlying associations, we considered four different plausible patterns of association over time: null association, constant association, attenuated association, and acute association. These will be defined more formally in the methods section.

## Behavior of persistent associations

As mentioned briefly already, a key component of this analysis is whether differential methylation persists over time. This has implications for interventions: if differential methylation at sites associated with disease later in life corrects itself within a finite time frame after exposure, it is unlikely that an intervention targeting the exposure would be worthwhile. Therefore, we do not want our method to identify sites that are correlated at the first time point, but do not have a persistent correlation.

Additionally, we want to distinguish this behavior from an attenuated relationship, especially given that we might expect the association to attenuate if the cells are able to restore the original methylation after exposure. We want to detect a difference between complete restoration and a weakened association over time.

# CHAPTER 3

# METHODS

Simulations were used to investigate differences between our three proposed methods of correcting for multiple testing. All three of the methods use the Benjamini-Hochberg False Discovery Rate, but vary when in the process rejections are made. For all methods, data was first generated according to one of four underlying assumptions. Then sites thought to be significantly associated with the exposure were identified.

The first method uses a pairwise correlation test to identify sites, corrects for multiple testing, and then tests whether the correlation persists over time in the selected sites with a linear mixed effects model. We call this method an FDR method because FDR is used to select a subset of sites to fit to an LME. This method is similar to the one used in the study of gestational age and birth weight (Simpkin, et al., 2015).

The second method uses a pairwise correlation test to identify sites, tests whether the correlation persists over time in a subset of these sites with a linear mixed effects model, and then corrects for multiple testing. We call this method a rank method because ranked p-values are used to select a subset of sites to fit to an LME. We arbitrarily chose 1000 as the size of the subset and select the 1000 lowest p-values.

The third method uses a linear mixed effects model to both identify sites and test whether the correlation persists over time and corrects for multiple testing. We call this method All LME because all sites are selected to fit to an LME.

Because the data is simulated, we then calculated the percentage of the true associations found, as well as a measure of the false associations. Correlation is calculated using the `cor` function in R and is Pearson's correlation coefficient (R Core Team, 2017), but any test of correlation could have been used.

Our general method is as follows:

1. Generate a data set according to a specified data generating method
2. Run pairwise correlation tests between the predictor and the outcomes
3. Select sites that pass selection criteria
4. Fit an LME to each of the selected sites
5. Return the sites that achieve a significance threshold

For each of our four data generating assumptions, we repeated this process 1000 times. The All LME method omits steps 2 and 3.

## Data generating assumptions

For all of our data generation, we used a continuous predictor simulated from a standard normal distribution. We then simulated methylation status outcomes as continuous variables for 300 individuals at 3 time points according to the linear relationships described below. These values are not bounded by 0 and 1 because their analogue is an M-value rather than a Beta-value. We have a total of 470,560 sites in the simulation study: 470,000 locations have no association with the predictor; 500 have a weak association; 50 have a moderate association; and 10 have a strong association. These numbers were chosen based on estimates provided by our collaborators at the Harvard Medical School Department of Population Medicine.

We calibrated the actual values for these strengths of association by varying the three different effect sizes while generating simulated data sets and measuring the percentage achieving a significance of $p < 0.05$ after FDR correction on a Pearson's correlation test on just the first time point, averaged over 1000 replicates. We chose the effect sizes that resulted in:

- Weak association: about 10% of 500 locations achieved significance
- Moderate association: about 50% of 50 locations achieved significance
- Strong association: just under 100% of 10 locations achieved significance

For this simulation, the effect sizes were 0.1753, 0.2483, and 0.4089 respectively, but these have no real world analogue.

Because of the computational expense, raw data was not simulated for the 470,000 uncorrelated sites. Raw data was simulated for the 560 truly associated locations, from which p-values were calculated, while simulated p-values (from a uniform distribution) were used for the 470,000 uncorrelated sites. If one of the uncorrelated locations was selected spuriously and therefore needed to be fit to an LME, a bootstrap method was employed to select an effect size to use to generate correlated data.

## Null

The first possible relationship between the exposure and outcomes is that of no relationship. This means that the exposure and outcomes are independent of each other. Each site's methylation is constant, except for measurement error, and is not associated with the exposure. Measurement error refers both to our inability to measure methylation perfectly and our lack of knowledge about the behavior of methylation over time.

Under the assumption of no relationship, the outcome is simulated independently of the simulated predictor for the 560 sites. In the following equation, $y$ is the continuous methylation outcome over three indices: the location $l = 1 \dots 470{,}560$; the individual $i = 1 \dots 300$; and the time point $t = 1, 2, 3$.

$$y_{l,it} = b_{l,i} + e_{l,it} \qquad \textbf{(1)}$$

For each location, the equation describes a linear mixed effects model with fixed intercept of 0 and a random intercept. There are two error terms: $b_{l,i} \sim N(0,0.05)$ and $e_{l,it} \sim N(0,0.95)$, where $N(\mu, \sigma^2)$ denotes a normal distribution with mean $\mu$ and variance $\sigma^2$. The first error term is shared between the three time points within an individual and the second is simulated independently for every time point. Therefore, even though there is no relationship between the predictor and the outcome, the three time points are correlated within-individual.

## Constant

The next possible relationship is that the exposure influences the outcome in a constant manner. For example, a person with a higher level of exposure would have a higher methylation outcome and this relationship would be constant over time. Because the exposure is fixed, this means that the methylation status would be fixed over time, except for measurement error. This is a fairly simplistic model, but if the data followed this pattern, it would have large implications for possible interventions.

Under the assumption of a constant relationship, the outcome is simulated using the simulated exposure, denoted here as $g_i \sim N(0,1)$, as mentioned in the Data generating assumptions section. The effect of the predictor on the outcome is denoted as $\beta_l$ because it is dependent on the location. It is simulated as three distinct values, representing strong, moderate or weak associations, but we would expect each $\beta_l$ to be unique. We simulate 10 strongly associated sites, 50 moderately associated, and 500 weakly associated.

$$\beta_l = \begin{cases} 0 & l \leq 470{,}000 \\ 0.1753 & 470{,}000 < l \leq 470{,}500 \\ 0.2483 & 470{,}500 < l \leq 470{,}550 \\ 0.4089 & l > 470{,}550 \end{cases} \quad \textbf{(2)}$$

$$y_{l,it} = \beta_l g_i + b_{l,i} + e_{l,it} \quad \textbf{(3)}$$

Note that for $l < 470{,}000$, there is no association between $g$ and $y_l$. The error terms in (3) are simulated as in (1). Here we can index on time, but it is not featured in the equation.

## Attenuated

It is possible that the exposure influences the outcome in a linear manner as described above, but that the strength of the relationship decreases over time. This is a more likely model than stable levels of methylation over time, given that methylation changes over time.

Under the assumption of a linear relationship that attenuates, the outcome is simulated using the simulated exposure and a time-dependent factor. At the first time point, the predictor is the same as in the constant relationship. At the second time point, the $\beta$ term is multiplied by 0.8. At the third time point, the $\beta$ term is multiplied by 0.6. We represent the different relationships with an indicator function, denoted $I(x)$, where $I(x) = 1$ if $x$ is true and 0 otherwise.

$\beta_l$ is defined as in ($\beta_l = \begin{cases} 0 & l \leq 470{,}000 \\ 0.1753 & 470{,}000 < l \leq 470{,}500 \\ 0.2483 & 470{,}500 < l \leq 470{,}550 \\ 0.4089 & l > 470{,}550 \end{cases}$ **(2)**.

$$y_{l,it} = \beta_l g_i \, I(t = 1) + \beta_l g_i \, I(t = 2) \times 0.8 + \beta_l g_i \, I(t = 3) \times 0.6 + b_{l,i} + e_{l,it} \quad \textbf{(4)}$$

The error terms in (4) are simulated as in (1).

## Acute

The final relationship we considered was that the exposure influences the outcome in the time point closest to it, but it is not associated at later time points. This is another biologically plausible relationship, in which a sudden event affects the methylation for only a short period.

Under the assumption of an acute relationship, the outcome at the first time point is simulated as in the constant relationship, but the outcomes at the second and third time points are simulated as in the null relationship. The indicator function is denoted again $I(x)$.

$\beta_l$ is defined as in (2).

$$y_{l,it} = \beta_l g_i \, I(t = 1) + b_{l,i} + e_{l,it} \qquad \textbf{(5)}$$

The error terms in (5) are simulated as in (1).

## Multiple testing correction

Although the previous study used just the cord blood correlations to select sites to fit an LME model to, we consider different selection methods because there might be differences between the time points in terms of their relationship with the predictor. For example, if the relationship is constant, we lose power by only considering the first time point. We consider three different ways of applying our two-step methods, FDR or rank, each based on a pairwise correlation test.

The LME model that we fit was the constant model with a parameter for continuous time, $y_{l,it} = \beta_l g_i + \beta_2 t + b_{l,i} + e_{l,it}$, which allows a random intercept for each individual and a fixed effect of both the exposure and time.

## Using the first time point

The first approach to question of whether there is a relationship between exposure and outcome is to calculate the correlation between the exposure and site-specific methylation at the first time point. Considering only the first time point would reduce the number of tests required and thereby the multiple testing problem would be reduced somewhat. Additionally, we might think that the methylation from the time point closest to the exposure has a unique relationship with it and would want to prioritize effects detectable in the first time point.

We denote this approach as **T1** and use it with our two methods: using the False Discovery Rate (FDR) method before fitting a linear mixed effects (LME) model or using FDR after fitting the LME model. These are denoted T1-FDR and T1-1000 respectively.

### T1-FDR:

Given the resultant p-values from testing the pairwise correlation between predictor and outcome at just the first time point, we use the FDR method to select sites that pass the significance threshold ($q < 0.05$) and fit an LME model to each of these sites. Finally, we return the sites with (unadjusted) $p < 0.05$ from the p-values returned by the LME.

### T1-1000:

Given the same p-values as in T1-FDR, we do not make any multiple testing correction before fitting an LME. Instead, we choose the 1000 lowest p-values to fit to an LME. Once we have fit an LME model to the sites, we correct the resultant 1,000 p-values using FDR with the number of tests ($m$ in the FDR algorithm) equal to 470,560 and return those that achieve the significance threshold.

## All time points

Instead of testing a subset of our outcomes, we could use all of our time points, looking at the relationship between exposure and outcome for each observation as if they had been measured at the same time point. For $\rho_{lt}$, where $l = 1 \dots 470{,}560$ and $t = 1, 2, 3$, we use all values of $\rho_{lt}$.

This would prioritize the strong associations because strongly associated loci would appear three times. However, because controlling the rate of false discoveries means more tests achieve significance when there are a larger number of true associations, this approach would allow more power.

We denote this approach as **AT** and use it with our two methods as before.

### AT-FDR:

While the previous approach tested only the first time point, this approach tests pairwise correlation on all of the data, tripling the number of tests that are done ($m = 470{,}560 \times 3$). Ignoring the time index, we adjust the resultant p-values using FDR and select those that pass the significance threshold (q < 0.05). The perceived benefit of this method is that some sites might not achieve significance at the first time point, but would at the second or third. We then fit an LME model for each of the selected sites and return the sites with (unadjusted) p < 0.05 from the p-values returned by the LME.

### AT-1000:

This procedure calculates $M \times T$ pairwise correlations as in the AT-FDR, but does not adjust the p-values as in the T1-1000. We choose the 1000 lowest p-values from the $M \times T$ tests, ignoring the time index. After fitting the selected sites to LME models, we correct the resultant p-values using FDR ($m = 470{,}560$) and return those that achieve the significance threshold.

**Table 2: Overview of Selection Approaches**

| NAME | PROCESS | | | |
|---|---|---|---|---|
| **T1-FDR** | ρ T1 | → | FDR → | LME |
| **AT-FDR** | ρ all T | → | FDR → | LME |
| **ET-FDR** | ρ each T | → | FDR → | LME |
| **T1-1000** | ρ T1 | → 1000 | LME → | FDR |
| **AT-1000** | ρ all T | → 1000 | LME → | FDR |
| **ET-1000** | ρ each T | → 1000* | LME → | FDR |
| **ALL-LME** | fit LME | → | FDR | |

*ET-1000 could select up to 3000 values

## Each time point

Another approach we might take would be to treat the outcomes at each time point as its own data set. If we assume that our outcomes are correlated over time, we might be able to find some of the weaker associations in one time point but not the others. For $\rho_{lt}$, where $l = 1 \dots 470{,}560$ and $t = 1, 2, 3$, we use the p-values obtained from $\rho_{l1}$ and $\rho_{l2}$ and $\rho_{l3}$.

Thus we assess the correlations between exposure and outcomes and correct for multiple testing within each time point. This will increase our Type I error, but we will find more correlations. If the outcome is not stable over time, we would notice that the number of associations found is different between time points.

We denote this approach as **ET** and use it with our two methods as before.

### ET-FDR:

This approach tests pairwise correlation on all of the data, as does the previous one, but adjusts for multiple testing differently. Here, we test for correlation on all of the data, but use the FDR method within each time point ($m = 470{,}560$). We then select the distinct sites that pass the significance threshold (q < 0.05), regardless of whether they passed in just one time point or all three, and fit these sites to an LME for each site. Finally, we return the sites with (unadjusted) p < 0.05 from the p-values returned by the LME.

### ET-1000:

This procedure calculates $M \times T$ pairwise correlations as in the ET-FDR method, then chooses the 1000 lowest p-values from each of the time points, as in the T1-1000 method, resulting in 3000 selections. However, we select only the unique sites, and thus will fit between 1000 and 3000 sites to LME models. We correct the resultant p-values using FDR and return those that achieve the significance threshold.

### All-LME

All of the sites are fit with an LME and FDR is used to correct for multiple testing ($m = 470{,}560$).

## Applied methods

Instead of using pairwise correlation tests in the applied analysis, we used robust linear regression to assess association at the first time point (Venables & Ripley, 2002). The covariates included in the model were maternal glucose, maternal BMI, gestational age, maternal college education, parity, smoking during pregnancy, child's sex, child's age at time of sample collection, and imputed cell counts. To implement this, we used the `rlm` function from the `MASS` package in R. For the longitudinal analysis, we used robust linear mixed effects models from the `robustlmm` package using a method developed by the author of the package (Koller, 2016) with the same covariates as the robust linear regression and the addition of a random intercept for each individual. We did not fit random slopes in this model, although previous research would suggest this would be beneficial, because three time points is relatively few, we had few subjects with all three times observed, and because of the computational expense.

Batch effects were corrected using ComBat, which uses empirical Bayes to adjust experimental variation that has non-biological origins (Johnson, Rabinovic, & Li, 2007), from the Bioconductor `sva` package (Leek, et al., 2017).

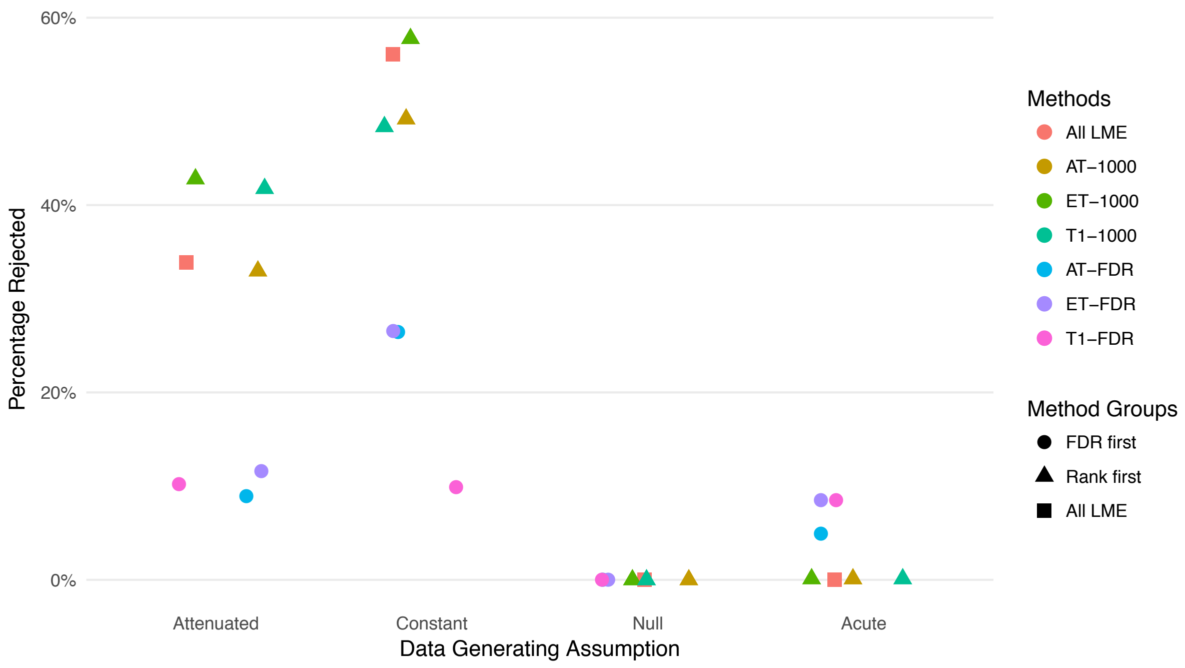
Fitting all 470,870 sites with a robust linear mixed effects model took about 2.25 minutes per site, or nearly 2 years if run serially. We did initially try ordinary least squares regression and linear mixed effects models, which are far less computationally expensive, but they found no significant results and were not used in the preliminary analysis that motivated this study.

The preliminary analysis stratified by sex, but we have chosen to use sex as a covariate in the model in the interest of both power considerations and computational time. Additionally, they used Beta-values in their analysis, but we opted for M-values because bounded values are less appropriate for fitting to linear models.

# CHAPTER 4

# RESULTS

**Figure 1: Percentage Rejected in the Weak Association Group**



## Simulation Results

To evaluate our different methods of adjusting for multiple testing, we considered the percentage of true associations rejected in each of the weak, moderate, and strong association groups; the number of false rejections made by each method; and the positive predictive value of each method.

We found that all methods were similar in their ability to detect associations in the strong and moderate association groups. Therefore, we focus on the weak association group where there were a total of 500 sites with a true simulated correlation.

The percentage of true associations found out of the number of true associations is an approximation of the simulated power of the test, which is the probability that the test rejects the null hypothesis when the alternative hypothesis is true. However, because calculations of power depend on the effect sizes expected and because we did not use biologically plausible effect sizes and instead calibrated an effect size based on expected power, we cannot consider these measures as power. That is to say, the stipulation of our experiment was that we would not expect to be able to detect many of the weakly associated sites and we aimed to increase the percentage that we could detect. Therefore, these percentages of rejections in the weak association group are approximations of the power relative to each other, with the understanding the the T1-FDR method was fixed at ≤ 0.10.

As seen in Figure 1, the methods that rank the initial p-values and utilize FDR after fitting an LME reject the largest percentages of the weakly associated sites. This is true in both an attenuated relationship and a constant relationship between predictor and outcome over time. The methods that correct with FDR before fitting an LME reject a far smaller percentage of sites, while the method of no pre-selection and only fitting an LME is much closer to the rank methods.

Important to note is that under the acute association, the methods that use FDR first reject a similar percentage to the attenuated and constant association data models, while the rank methods make few rejections and closely resemble the null data model.

**Table 3: Number of False Rejections**

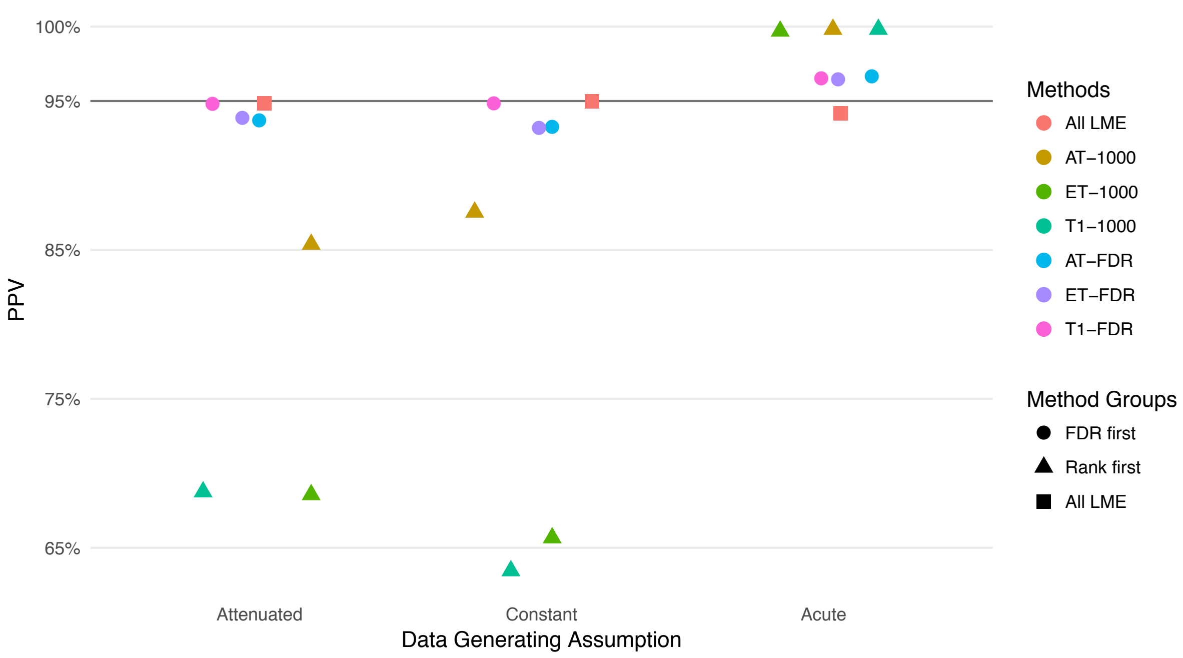
| METHOD | CONSTANT | ATTENUATED | ACUTE | NULL |
|---|---|---|---|---|
| **ALL LME** | 18 | 12 | 0 | 0 |
| **AT-1000** | 43 | 37 | 0 | 0 |
| **ET-1000** | 182 | 123 | 0 | 0 |
| **T1-1000** | 171 | 119 | 0 | 0 |
| **AT-FDR** | 13 | 5 | 2 | 0 |
| **ET-FDR** | 14 | 6 | 3 | 0 |
| **T1-FDR** | 5 | 5 | 3 | 0 |

Averaged over 1000 replicates, rounded to the nearest integer.

When considering 470,560 possible outcomes, setting an acceptable Type I error of α = 0.05 would result in 23,528 false rejections. Representing the false rejections we found as a percentage would be approximately α = 0.0005, which is difficult to understand practically. Therefore, we consider Type I error in terms of the number of false rejections we make by method and data model, as seen in Table 3. We found fewer than 50 false rejections for all methods except the rank methods that used either the first time point or each time point separately. Because we are controlling the false discovery rate, more false rejections are allowed when more true rejections are found.

The positive predictive value (PPV) is the complement of the false discovery rate (FDR) and therefore, we would expect to see a consistent value of 95% for all methods if we held the FDR at 5%. However, as seen in Figure 2, the rank methods do not control the FDR as well. The inflated value for the acute relationship is likely due to the low number of rejections made.

**Figure 2: Positive Predictive Value by Method**



Pre-selecting with FDR limits the power of the test and performs identically for attenuated, constant and acute relationships, but controls Type I error well. Pre-selecting with ranked p-values gives more power and differentiates between attenuated/constant and acute, but does not control Type I error as well. Testing each site with a linear mixed effects model and controlling the FDR has the consistently highest PPV, a low number of false positives, and is able to differentiate between differential methylation that was present at only one time point vs. a persistent relationship.

## Viva Data Results

Fitting the robust linear regression with the cord blood and the covariates previously mentioned found three sites that had a significant effect of maternal glucose on the methylation outcomes after correction with FDR. If we instead fit these three sites to a robust linear mixed effects model, the fixed effect of maternal glucose on the outcome was significant with unadjusted $p < 0.05$.

Using pairwise tests at the second and third time points yielded more significant sites than limiting to the cord blood: our all time points selection method identified 195 sites and the method using each time point identified 306. The rank method identified close to the maximum number of sites it could. The results are summarized in Table 4 below.

**Table 4: Number of sites selected and significant in the Project Viva cohort**

| METHOD | SITES SELECTED | SITES SIGNIFICANT |
|---|---|---|
| **T1-FDR** | 3 | 3 |
| **AT-FDR** | 195 | 76 |
| **ET-FDR** | 306 | 85 |
| **T1-1000** | 1000 | 0 |
| **AT-1000** | 999 | 0 |
| **ET-1000** | 2985 | 0 |
| **ALL-LME** | | 0 |

All of the rank methods identified no sites significant using the longitudinal analysis and analyzing all of the p-values from the different LME models had the same result. The FDR methods selected fewer sites, but returned some of them as significant.

# CHAPTER 5

# DISCUSSION

A key component of testing longitudinal differential epigenetic outcomes is determining whether changes persist over time. This is important because assuming that a differential outcome is constant over the lifetime of an individual is a very strong assumption and it is unknown what functional form an association would take over time. From a pragmatic standpoint, it is important to consider that the relationship might not be constant because differences that correct themselves would not require interventions or treatments. We also consider an attenuated relationship to determine if fitting a constant model to attenuated data is an acceptable choice; fitting an attenuated model would require more parameters to be fit and decrease the degrees of freedom in the model, which is undesirable.

Testing for correlation only does not detect whether changes persist over time, thus a longitudinal model is necessary. We found that pre-selecting with a correlation test diminishes the ability of even a linear mixed effects model to return differential results when the data is generated such that an association exists at the first time point but is not present in the later time points. This is seen in the percentage of weak association sites found; the all LME and rank methods do not find any significant sites in the acute relationship. Because the longitudinal model answers the question of whether there is a persistent relationship, it should *not* find significant sites under the acute data generating assumption, given that the relationship does not persist. However, the FDR methods reject a similar number of sites between constant, attenuated, and acute, which implies that if one uses these methods, the results returned could be coming from an acute relationship and we could not distinguish this.

With regard to our different selection methods, T1, AT and ET, we would assume that under the null, ET would select three times as many sites as AT would. This is because if there is no relationship, there is not an expectation that sites would be returned as significant more than once. Thus, the spuriously significant sites at each time point would have little to no overlap with the other time points. In our simulation results, we see that our ET-1000 method returns roughly three times more false rejections than AT-1000 under both the constant and attenuated assumptions.

Our applied results in Project Viva suggest that the underlying relationship between maternal glucose and a child's methylation outcomes over time may not be linear. The pattern of the number of sites found by each method is similar to that of the simulated acute relationship, which suggests that the relationship does not persist over time. Notably, we do not select any significant sites with the all LME or rank methods, but we do with the FDR methods. However, we cannot rule out a true relationship of some form because we have not explored non-linear relationships or other possible functional forms of the relationship. Additionally, we did not consider random slopes because we lack sufficient repeated measures, but this would be a logical next step.

This does not imply that maternal glucose levels are not mechanistically involved in increasing a child's risk of obesity or diabetes later in life, nor does it suggest that differential methylation is not involved. Rather, it implies that differential methylation in cord blood may not be reflected in peripheral blood cells later in life. Differential methylation patterns might still exist in other somatic cells in the mechanistic pathways of diabetes.

While simulation studies put us in a position of omniscience, data analysis comes from a position of ignorance.  The question in data analysis is not which procedure to perform, but what kind of relationship we are interested in. If we always fit a longitudinal model, we will miss an acute relationship. We might take this to mean that there is no relationship at all, but a relationship present in cord blood might persist in other cells, or the relationship might take a different form. Thus fitting a longitudinal model is appropriate when the question is whether the association persists, but inappropriate if the question is whether there is an association at any of the times. However, often we are asking both questions and we would recommend that both analyses be performed.

Therefore, if the research question hinges on whether the association with the exposure persists over time, pre-selecting with a correlation test will likely not answer this question and a full longitudinal analysis is necessary. We recommend that pairwise tests with the time point of interest be conducted separately from the longitudinal analysis, with both pairwise and longitudinal fitting all sites. If there are significant results in the first time point and not in the longitudinal analysis, it can be concluded that the relationship might not persist. If computational resources are limited, fitting a large number of sites selected from all of the time points and adjusting the $m$ in an FDR procedure might be an acceptable approximation, although the FDR will not be as well-controlled.

# BIBLIOGRAPHY

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software , 67*, 1-48.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B , 57*, 289-300.

Bjornsson, H. T., Sigurdsson, M. I., Fallin, M. D., Irizarry, R. A., Aspelund, T., Cui, H., et al. (2008). Intra-individual Change Over Time in DNA Methylation With Familial Clustering. *JAMA , 299* (24), 2877-2883.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., et al. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics , 11*, 587.

Fraga, M. F., Ballestar, E., Paz, M. F., Ropero, S., Setien, F., Ballestar, M. L., et al. (2005). Epigenetic differences arise during the lifetime of monozygotic twins. *PNAS , 102*, 10604–10609.

Franchini, D.-M., Schmitz, K.-M., & Petersen-Mahrt, S. K. (2012). 5-Methylcytosine DNA Demethylation: More Than Losing a Methyl Group. *Annual Review of Genetics , 46*, 419-41.

Herbstman, J. B., Wang, S., Perera, F. P., Lederman, S. A., Vishnevetsky, J., Rundle, A. G., et al. (2013). Predictors and consequences of global DNA methylation in cord blood and at three years. *PLoS ONE , 8* (9).

Jirtle, R. L., & Skinner, M. K. (2007). Environmental epigenomics and disease susceptibility. *Nature reviews Genetics , 8*, 253–262.

Johnson, W. E., Rabinovic, A., & Li, C. (2007). Adjusting batch effects in microarray expression data using Empirical Bayes methods. *Biostatistics , 8* (1), 118-127.

Koller, M. (2016). robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models. *Journal of Statistical Software , 75*, 1-24.

Leek, J. T., Johnson, W. E., Park, H. S., Fertig, E. J., Jaffe, A. E., Storey, J. D., et al. (2017). sva: Surrogate Variable Analysis.

Oken, E., Baccarelli, A. A., Gold, D. R., Kleinman, K. P., Litonjua, A. A., De Meo, D., et al. (2015). Cohort Profile: Project Viva. *International Journal of Epidemiology , 44* (1), 37-48.

R Core Team. (2017). *R: A language and environment for statistical computing*. Retrieved from https://www.R-project.org/

Simpkin, A. J., Suderman, M., Gaunt, T. R., Lyttleton, O., McArdle, W. L., Ring, S. M., et al. (2015). Longitudinal analysis of DNA methylation associated with birth weight and gestational age. *Human Molecular Genetics , 24*, 3752-3763.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth Edition ed.). Springer.

Wu, H.-C., Wang, Q., Delgado-Cruzata, L., Santella, R. M., & Terry, M. (2012). Genomic Methylation Changes Over Time in Peripheral Blood Mononuclear Cell DNA: Differences by Assay Type and Baseline Values. *Cancer Epidemiol Biomarkers Prev , 21* (8), 1314-1318.