# Outline

1. Write out and talk about model for the data (maybe discuss M vs. Beta value here)
2. Discuss methods without cell type data (i.e. Houseman and CATE).
3. When do these perform well?
   (a) Standardized effect size for $L$, taking into account variability $\Xi$.
   (b) Discuss $\frac{1}{p}\Gamma^T\Sigma^{-1}\Gamma$ result in Bai and Li, i.e. how informative methylation data are for cell type.
   (c) Figures: simulation results. We will ALWAYS underestimate $\Omega$ when data are not informative. Effect on results is small when $\Omega$ is small (can we prove this?)
   (d) Give Amish/Hutterite data example. Talk about size of $\frac{1}{p}\Gamma^T\Sigma^{-1}\Gamma$, MAJOR differences in the estimates for $\pi_0$, the fact that the data suggest that there is confounding (p-value for $\alpha$). Large difference in $\hat{\pi}_0$ along with the fact that $\frac{1}{p}\Gamma^T\Sigma^{-1}\Gamma = O\left(\frac{1}{n}\right)$ seem to indicate the data are not informative enough to estimate the correlation between $C$ and $X$.
4. If data are not informative for cell type, we need another method to estimate the correlation between $C$ and $X$. Obvious alternative: collect cell type or use training data to estimate the correlation between $C$ and $X$.
5. In some circumstances, we can do well with only partially observed data (in the absence of other confounders). Plot simulation figures. When does this happen? What happens when we have additional confounders?
6. Recommendations: if you have strong prior assumptions that there are additional confounders that you can measure, it is always a good idea to measure them.
   (a) Talk about how large correlation needs to be to start making serious errors. If standardized correlation is small, maybe you don't need to worry about it and can estimate it from the data.
   (b) With strong prior knowledge, it is always a good idea to include covariates that may have an effect on response and are correlated with the variable of interest.

# 1 Notation

- All matrices (including vectors) are bold and all vectors are column vectors, unless otherwise stated.
- Throughout the paper, the matrix $\boldsymbol{X} = \boldsymbol{X}_{d\times n} \in \mathbb{R}^{d\times n}$ contains $d$ covariates of interest (including an intercept) measured on $n$ individuals. For example, in the real data example $d = 2$ where one covariate is the intercept and the other is the Amish vs. Hutterite contrast.
- Cell type proportions for $K$ cells measured on $n$ individuals is given as the matrix $\boldsymbol{C} = \boldsymbol{C}_{K\times n} \in [0-1]^{K\times n}$.
- $P_{\boldsymbol{A}}$ is the orthogonal projection onto the linear subspace generated by the columns of $\boldsymbol{A}$ and $P^{\perp}_{\boldsymbol{A}}$ projects onto the subspace orthogonal to Image $(\boldsymbol{A})$.
- vec $(\boldsymbol{A})$ is the vectorized version of a matrix, obtain by column stacking.
- When the matrix $\boldsymbol{U} \in \mathbb{R}^{p\times n}$ is random and the covariance of the elements of $\boldsymbol{U}$ are separable across rows and columns, we will write
$$\boldsymbol{U} \sim \left(\boldsymbol{\mu}_{p\times n}, \left[\boldsymbol{S}_{p\times p}, \boldsymbol{R}_{n\times n}\right]\right)$$
to indicate that $\boldsymbol{\mu}$ is the mean of $\boldsymbol{U}$, $\boldsymbol{S}$ is the covariance across rows and $\boldsymbol{R}$ is the covariance across columns. That is, Var (vec $(\boldsymbol{U})$) $= \boldsymbol{R}\otimes\boldsymbol{S}$. For example, if we measured the methylation on $p$ CpG sites across $n$ independent individuals, $\boldsymbol{R} = I_n$ and $\boldsymbol{S}$ would be the covariance across the $p$ CpG sites in a single sample. We characterize random variables by their first two moments, as it is often times more important to accurately estimate the mean and variance than it is to use the correct probability model (cite contiguity paper).

## 2 The Model

Let $\boldsymbol{Y}_{p\times n}$ be the methylation response measured on $p$ CpG sites across $n$ independent individuals, $\boldsymbol{X}_{d\times n} = \left[\vec{x}_1 \quad \cdots \quad \vec{x}_n\right]$ the matrix of $d$ covariates of interest (i.e. disease status, sex, etc.) measured on the same $n$ individuals and $\boldsymbol{C}_{K\times n} = \left[\vec{c}_1 \quad \cdots \quad \vec{c}_n\right]$ be the cell composition matrix for $K$ cell types across the $n$ individuals Assuming that methylation status is ONLY affect by the covariate(s) of interest $\boldsymbol{X}$ and cell type composition $\boldsymbol{C}$, full data model for $\boldsymbol{Y}$ is then:

$$\boldsymbol{Y}_{p\times n} = \boldsymbol{B}_{p\times d}\boldsymbol{X}_{d\times n} + \boldsymbol{L}_{p\times K}\boldsymbol{C}_{K\times n} + \boldsymbol{E}_{p\times n} \tag{1}$$

Where $\boldsymbol{E}_{p\times n} \sim \left(\boldsymbol{0}_{p\times n}, \left[\boldsymbol{\Sigma}_{p\times p} = \text{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right), I_n\right]\right)$, $\boldsymbol{L} = \begin{bmatrix}\boldsymbol{\ell}_1^T \\ \vdots \\ \boldsymbol{\ell}_p^T\end{bmatrix}$ and $\boldsymbol{B} = \begin{bmatrix}\boldsymbol{\beta}_1^T \\ \vdots \\ \boldsymbol{\beta}_p^T\end{bmatrix}$. That is,

$$\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{C}) \sim (\boldsymbol{B}\boldsymbol{X} + \boldsymbol{L}\boldsymbol{C}, [\boldsymbol{\Sigma}, I_n]) \tag{2}$$

We are interested in estimating the effect $\boldsymbol{\beta}_i$ for each CpG site $i = 1, \ldots, p$.

However, sometimes we do not have access to the cell type composition data. If cell type is uncorrelated with the covariate(s) of interest, we can account for it using factor analysis correction techniques (e.g. SVA). When cell type and the covariate(s) of interest are correlated, we assume that

$$\boldsymbol{C}_{K\times n} = \boldsymbol{\Omega}_{K\times d}\boldsymbol{X}_{d\times n} + \boldsymbol{\Xi}_{K\times n}, , \boldsymbol{\Xi} = \left[\vec{\xi}_1 \quad \cdots \quad \vec{\xi}_n\right] \tag{3}$$

$$\boldsymbol{\Xi} \sim (\boldsymbol{0}, [\Lambda_{K\times K}, I_n]) \tag{4}$$

Note that since cell type composition is proportion data, the variance $E\left(\vec{\xi}_i\vec{\xi}_i^T \mid \vec{x}_i\right)$ is a function of the mean $E\left(\vec{c}_i \mid \vec{x}_i\right)$, which is not captured by our assumptions on $\boldsymbol{\Xi}$. We show by simulation that we can ignore the mean-variance relationship and still perform accurate inference on $\hat{\boldsymbol{B}}$.

When $\boldsymbol{C}$ is unobserved, we can write the full data model as

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{L}\boldsymbol{\Omega}\boldsymbol{X} + \boldsymbol{L}\boldsymbol{\Xi} + \boldsymbol{E} = \boldsymbol{B}\boldsymbol{X} + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Omega}}\boldsymbol{X} + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Xi}} + \boldsymbol{E} \tag{5}$$

where $\tilde{\boldsymbol{L}} = \begin{bmatrix}\tilde{\boldsymbol{\ell}}_1^T \\ \vdots \\ \tilde{\boldsymbol{\ell}}_p^T\end{bmatrix} = L\Lambda^{1/2}$, $\tilde{\Omega} = \Lambda^{-1/2}\Omega$ and $\tilde{\boldsymbol{\Xi}} \sim (\boldsymbol{0}, [I_K, I_n])$. The first and second moments conditional on only observing the covariates $\boldsymbol{X}$ are

$$\boldsymbol{Y} \mid \boldsymbol{X} \sim \left(\boldsymbol{B}\boldsymbol{X} + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Omega}}\boldsymbol{X}, \left[\tilde{\boldsymbol{L}}\tilde{\boldsymbol{L}}^T + \boldsymbol{\Sigma}, I_n\right]\right). \tag{6}$$

## 3 Estimation Methods With Unobserved Cell Type

We start by assuming $\boldsymbol{X}$ and cell type $\boldsymbol{C}$ are both observed. We see from equation **3** that if we were to estimate $\tilde{\boldsymbol{\Omega}}$ using OLS (assuming $\boldsymbol{\Lambda}$ was known), we would get

$$\tilde{\boldsymbol{\Omega}}^{(\text{OLS})} = \boldsymbol{\Lambda}^{-1/2}\boldsymbol{C}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1} = \tilde{\boldsymbol{\Omega}} + \tilde{\boldsymbol{\Xi}}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}. \tag{7}$$

Note that there are two sources of empirical correlation. The first is due to the true correlation $\boldsymbol{\Omega}$ and the second is due to the residual $\tilde{\boldsymbol{\Xi}}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}$, which we will refer to as spurious correlation. When $\boldsymbol{C}$ is observed (i.e. $\boldsymbol{\Omega}\boldsymbol{X} + \boldsymbol{\Xi}$ is observed), the unbiased OLS estimate for $\boldsymbol{B}$ is

$$\hat{\boldsymbol{B}}^{(\text{OLS})} = \boldsymbol{Y}\boldsymbol{X}^T\left[\boldsymbol{X}\boldsymbol{X}^T\right]^{-1} - \boldsymbol{Y}P^{\perp}_{\boldsymbol{X}^T}\tilde{\boldsymbol{\Xi}}^T\left[\tilde{\boldsymbol{\Xi}}P^{\perp}_{\boldsymbol{X}^T}\tilde{\boldsymbol{\Xi}}^T\right]^{-1}\tilde{\boldsymbol{\Omega}}^{(\text{OLS})} = \boldsymbol{Y}\boldsymbol{X}^T\left[\boldsymbol{X}\boldsymbol{X}^T\right]^{-1} - \hat{\tilde{\boldsymbol{L}}}^{(\text{OLS})}\tilde{\boldsymbol{\Omega}}^{(\text{OLS})} \tag{8}$$

with variance

$$\text{Var}\left(\hat{\boldsymbol{\beta}}_i^{(\text{OLS})}\right) = \sigma_i^2\left(\left[\boldsymbol{X}\boldsymbol{X}^T\right]^{-1} + \left(\tilde{\boldsymbol{\Omega}}^{(\text{OLS})}\right)^T\left[\tilde{\boldsymbol{\Xi}}P^{\perp}_{\boldsymbol{X}^T}\tilde{\boldsymbol{\Xi}}^T\right]^{-1}\tilde{\boldsymbol{\Omega}}^{(\text{OLS})}\right) \tag{9}$$

which by Gauss-Markov is the smallest variance possible among all unbiased estimators for $\boldsymbol{\beta}_i$. If cell type $\boldsymbol{C}$ is NOT observed, to correct for cell type heterogeneity it stands to reason that our goal should be to estimate the part of $\boldsymbol{C}$

uncorrelated with $\boldsymbol{X}$ and the part that is correlated with $\boldsymbol{X}$. That is, we wish to estimate $P^{\perp}_{\boldsymbol{X}^T}\tilde{\boldsymbol{\Xi}}^T$ and $\tilde{\boldsymbol{\Omega}}^{(\text{OLS})}$ = true correlation + spurious correlation from the methylation data $\boldsymbol{Y}$.

Luckily, considerable attention has been given to developing methods to estimate $\boldsymbol{B}$ in the presence of confounding that is correlated with the design matrix $\boldsymbol{X}$ (LEAPP, Houseman et al., CATE). Using **Equ. 5** as a reference, all three methods assume that $\tilde{\boldsymbol{\Xi}}$ is independent of $\boldsymbol{X}$ and that $\boldsymbol{B}$ is sparse. That is, the covariate of interest affects only a small fraction of the CpG sites. Both of these assumptions tend to hold in real data. To get estimates $\hat{\boldsymbol{B}} = \begin{bmatrix} \hat{\boldsymbol{\beta}}_1^T \\ \vdots \\ \hat{\boldsymbol{\beta}}_p^T \end{bmatrix}$ and $\hat{\text{Var}}\left(\hat{\boldsymbol{\beta}}_i\right)$ for CpG sites $i = 1, \ldots, p$, all three methods use the following general procedure:

1. Compute $\hat{\tilde{\boldsymbol{L}}}$ and $\hat{\boldsymbol{\Sigma}}$, estimates for $\boldsymbol{L}$ and $\boldsymbol{\Sigma}$. This is done by regressing out $\boldsymbol{X}$ from **Equ. 5** and then performing factor analysis on the residual matrix $\boldsymbol{Y}P^{\perp}_{X^T}$.

2. Compute $\hat{\tilde{\boldsymbol{\Xi}}}^{\perp}$, an estimate for $\tilde{\boldsymbol{\Xi}}P^{\perp}_{X^T}$. Two possibilities for $\hat{\tilde{\boldsymbol{\Xi}}}^{\perp}$ are the BLUP for $\tilde{\boldsymbol{\Xi}}P^{\perp}_{X^T}$ and the GLS estimate $\left(\hat{\tilde{\boldsymbol{L}}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\tilde{\boldsymbol{L}}}\right)^{-1}\hat{\tilde{\boldsymbol{L}}}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y}P^{\perp}_{X^T}$.

3. Compute $\hat{\tilde{\boldsymbol{\Omega}}}$, an estimate for $\tilde{\boldsymbol{\Omega}}^{(\text{OLS})}$ by regressing $\boldsymbol{Y}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}$ onto $\hat{\tilde{\boldsymbol{L}}}$. The methods differ in the way they find $\hat{\tilde{\boldsymbol{\Omega}}}$.

   (a) Houseman uses the the GLS estimate for $\tilde{\boldsymbol{\Omega}}^{(\text{OLS})}$, $\hat{\tilde{\boldsymbol{\Omega}}} = \left(\hat{\tilde{\boldsymbol{L}}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\tilde{\boldsymbol{L}}}\right)^{-1}\hat{\tilde{\boldsymbol{L}}}^T\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{Y}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}$. If $\boldsymbol{B} \neq \boldsymbol{0}$, this is sub-optimal since $\mathbb{E}\left[\boldsymbol{Y}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1} \mid \boldsymbol{X}, \tilde{\boldsymbol{\Xi}}\right] = \boldsymbol{B} + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Omega}}^{(\text{OLS})}$, meaning the GLS estimate will be contaminated by the non-zero entries of $\boldsymbol{B}$.

   (b) CATE, LEAPP use a robust regression to find $\hat{\tilde{\boldsymbol{\Omega}}}$. This is a more appropriate estimator when $\boldsymbol{B} \neq \boldsymbol{0}$.

4. Estimate $\boldsymbol{B}$ as

$$\hat{\boldsymbol{B}} = \boldsymbol{Y}\boldsymbol{X}^T\left[\boldsymbol{X}\boldsymbol{X}^T\right]^{-1} - \boldsymbol{Y}\left(\hat{\tilde{\boldsymbol{\Xi}}}^{\perp}\right)^T\left[\hat{\tilde{\boldsymbol{\Xi}}}^{\perp}\left(\hat{\tilde{\boldsymbol{\Xi}}}^{\perp}\right)^T\right]^{-1}\hat{\tilde{\boldsymbol{\Omega}}} \approx \boldsymbol{Y}\boldsymbol{X}^T\left[\boldsymbol{X}\boldsymbol{X}^T\right]^{-1} - \hat{\tilde{\boldsymbol{L}}}\hat{\tilde{\boldsymbol{\Omega}}} \tag{10}$$

   We then set

$$\hat{\text{Var}}\left(\hat{\boldsymbol{\beta}}_i\right) = \hat{\sigma}_i^2\left(\left[\boldsymbol{X}\boldsymbol{X}^T\right]^{-1} + \hat{\tilde{\boldsymbol{\Omega}}}^T\left[\hat{\tilde{\boldsymbol{\Xi}}}^{\perp}\left(\hat{\tilde{\boldsymbol{\Xi}}}^{\perp}\right)^T\right]^{-1}\hat{\tilde{\boldsymbol{\Omega}}}\right) \tag{11}$$

These methods (CATE and LEAPP in particular) are attractive because one can use all of the CpG sites to estimate the confounding in the data without actually having to measure the confounding (if that is at all possible). They tend to perform well when the observed methylation data are informative enough to estimate the cell type effect, $\tilde{\boldsymbol{L}}$ (simulation plots when data are informative enough to estimate cell type?). However, since we must use the noisy estimate $\hat{\tilde{\boldsymbol{L}}}$ to estimate $\tilde{\boldsymbol{\Omega}}^{\text{OLS}}$, a term that controls the accuracy of any inference we do on $\boldsymbol{B}$ (see **equ. 10** and **equ. 11**), we should try to understand methylation data are not informative enough to infer differentially expressed CpGs in the presence of unmeasured cell type heterogeneity. We are particularly concerned with underestimating $\tilde{\boldsymbol{\Omega}}^{\text{OLS}}$, as this will tend to lead to non-conservative inference on $\boldsymbol{B}$.

# 4 When can we use these methods to correct for cell type and other confounding?

The fidelity of any (is this word too strong) cell type correction method depends on how informative the available methylation data are for estimating cell type heterogeneity. If $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T \to \boldsymbol{\Sigma}_X \succ 0$ and $\tilde{\boldsymbol{L}}$ is fixed, Hastie (cite paper) proves that under suitable conditions on $\boldsymbol{B}, \tilde{\boldsymbol{L}}$ and $\boldsymbol{\Sigma}$,

$$\hat{\boldsymbol{\beta}}_i \xrightarrow{P} \boldsymbol{\beta}_i \tag{12}$$

and

$$n\hat{\text{Var}}\left(\hat{\boldsymbol{\beta}}_i\right) = \hat{\sigma}_i^2\left(\left[\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T\right]^{-1} + \hat{\tilde{\boldsymbol{\Omega}}}^T\left[\frac{1}{n}\hat{\tilde{\boldsymbol{\Xi}}}^{\perp}\left(\hat{\tilde{\boldsymbol{\Xi}}}^{\perp}\right)^T\right]^{-1}\hat{\tilde{\boldsymbol{\Omega}}}\right) \xrightarrow{P} \sigma_i^2\left(\boldsymbol{\Sigma}_X^{-1} + \boldsymbol{\Omega}^T\boldsymbol{\Lambda}^{-1}\boldsymbol{\Omega}\right) \tag{13}$$

which is exactly what we would expect had we measured cell type. We also show through simulation that for $n$ large enough and $\tilde{\boldsymbol{L}}$ fixed, the above procedure with robust regression tends to control false discovery rate at a nominal level (show plots).

However, the methylation data may not be informative enough to estimate and correct for the effects of cell type heterogeneity for small to moderate sample sizes. Recall that when we correct for cell type we must first compute $\hat{\tilde{\boldsymbol{L}}}$, and use that noisy estimate of $\tilde{\boldsymbol{L}}$ to regress $\boldsymbol{Y}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}$ onto $\hat{\tilde{\boldsymbol{L}}}$. If the residual $\hat{\tilde{\boldsymbol{L}}} - \tilde{\boldsymbol{L}}$ is large, then our estimate $\hat{\tilde{\boldsymbol{\Omega}}}$ will be shrunk closer to $\mathbf{0}$. The intuition here is that if $\hat{\tilde{\boldsymbol{L}}}$ was just a random matrix, the regression $\boldsymbol{Y}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1} \sim \hat{\tilde{\boldsymbol{L}}}$ will give regression coefficients that are very close to 0. If our estimate for $\hat{\tilde{\boldsymbol{\Omega}}}$ is too small, we will underestimate the variance of $\hat{\boldsymbol{\beta}}_i$ and any inference we do will be anti-conservative.

To study this phenomenon, we will assume $\boldsymbol{B} = \mathbf{0}$ and that there is only one covariate of interest (we assume that the covariate and response have been mean centered, so we may ignore the intercept and let $d = 1$). That is, variability in the data is caused only by cell type heterogeneity which may be correlated with the variable of interest:

$$\boldsymbol{Y} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Omega}}\boldsymbol{X} + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Xi}} + \boldsymbol{E}. \tag{14}$$

If we let $\boldsymbol{E}^* = \left(\tilde{\boldsymbol{L}} - \hat{\tilde{\boldsymbol{L}}}\right)\left(\tilde{\boldsymbol{\Omega}}\boldsymbol{X} + \tilde{\boldsymbol{\Xi}}\right) + \boldsymbol{E}$, we can rewrite **11** as

$$\boldsymbol{Y} = \hat{\tilde{\boldsymbol{L}}}\left(\tilde{\boldsymbol{\Omega}}\boldsymbol{X} + \tilde{\boldsymbol{\Xi}}\right) + \boldsymbol{E}^*. \tag{15}$$

If the residual $\hat{\tilde{\boldsymbol{L}}} - \tilde{\boldsymbol{L}}$ is large in comparison to $\hat{\tilde{\boldsymbol{L}}}$, we have no way of separating the effect $\hat{\tilde{\boldsymbol{L}}}\left(\tilde{\boldsymbol{\Omega}}\boldsymbol{X} + \tilde{\boldsymbol{\Xi}}\right)$ from the error $\boldsymbol{E}^*$, which causes our estimate $\hat{\tilde{\boldsymbol{\Omega}}}$ to shrink toward $\mathbf{0}$.

If cell type were known, (i.e. if $\tilde{\boldsymbol{\Omega}}\boldsymbol{X} + \tilde{\boldsymbol{\Xi}}$ were known), we can estimate $\tilde{\boldsymbol{L}}$ using OLS where the estimates $\hat{\tilde{\boldsymbol{\ell}}}_i^{(\text{OLS})} \approx \tilde{\boldsymbol{\ell}}_i \pm 1.96\frac{\sigma_i}{\sqrt{n}}$. Surprisingly, even if we do not know cell type, we can use factor analysis to to get an estimate $\hat{\tilde{\boldsymbol{\ell}}}_i^{(\text{FA})} \approx \tilde{\boldsymbol{\ell}}_i \pm 1.96\frac{\sigma_i}{\sqrt{n}}$! Therefore, the effect $\hat{\tilde{\boldsymbol{L}}}\left(\tilde{\boldsymbol{\Omega}}\boldsymbol{X} + \tilde{\boldsymbol{\Xi}}\right)$ is roughly equal to or smaller than the noise $\boldsymbol{E}^*$ if, on average, $\tilde{\boldsymbol{\ell}}_i \leq O\left(\frac{\sigma_i}{\sqrt{n}}\right)$ for CpG sites $i = 1, \ldots, p$. Another way of saying this is $\frac{1}{p}\tilde{\boldsymbol{L}}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{L}} \leq O\left(\frac{1}{n}\right)$. From here on out, we will define

$$\mathcal{I}_{\text{confounding}} = \frac{1}{p}\tilde{\boldsymbol{L}}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{L}} \tag{16}$$

as the confounding information matrix. This is not to be confused with the Fisher Information matrix, although it has a similar role. $\mathcal{I}_{\text{confounding}}$ determines how much information about the confounding (in our case cell type heterogeneity) can be gleaned from the observed methylation data. When this information is on the order of the statistical error (i.e. when $\mathcal{I}_{\text{confounding}} \leq O\left(\frac{1}{n}\right)$), we have difficulty estimating and correcting for cell type heterogeneity.

Note that since we are assuming $\tilde{\boldsymbol{\Xi}} \sim (\mathbf{0}, [I_K, I_n])$ in **equ. 14**, we can rotate $\tilde{\boldsymbol{\Xi}}$ without changing the first and second moments. Therefore, the first and second moments of $\tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Omega}}\boldsymbol{X} + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Xi}} + \boldsymbol{E}$ and $\left(\tilde{\boldsymbol{L}}\boldsymbol{U}\right)\left(\boldsymbol{U}^T\tilde{\boldsymbol{\Omega}}\right)\boldsymbol{X} + \left(\tilde{\boldsymbol{L}}\boldsymbol{U}\right)\left(\boldsymbol{U}^T\tilde{\boldsymbol{\Xi}}\right) + \boldsymbol{E}$ are the same for any rotation matrix $\boldsymbol{U} \in \mathbb{R}^{K\times K}$. This means that it suffices to assume that

$$\mathcal{I}_{\text{confounding}} = \text{diag}\left(\delta_1, \ldots, \delta_K\right) \text{ with } \delta_1 \geq \delta_2 \geq \cdots \geq \delta_K > 0. \tag{17}$$

Lastly, since we have assumed $d = 1$, we can re-write **equ. 14** as

$$\boldsymbol{Y} = \tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Omega}}^{\text{OLS}}\boldsymbol{X} + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Xi}}P^{\perp}_{\boldsymbol{X}^T} + \boldsymbol{E} = \left\|\tilde{\boldsymbol{\Omega}}^{\text{OLS}}\right\|_2\left(\tilde{\boldsymbol{L}}\boldsymbol{Q}\boldsymbol{e}_1\right)\boldsymbol{X} + \tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Xi}}P^{\perp}_{\boldsymbol{X}^T} + \boldsymbol{E} \tag{18}$$

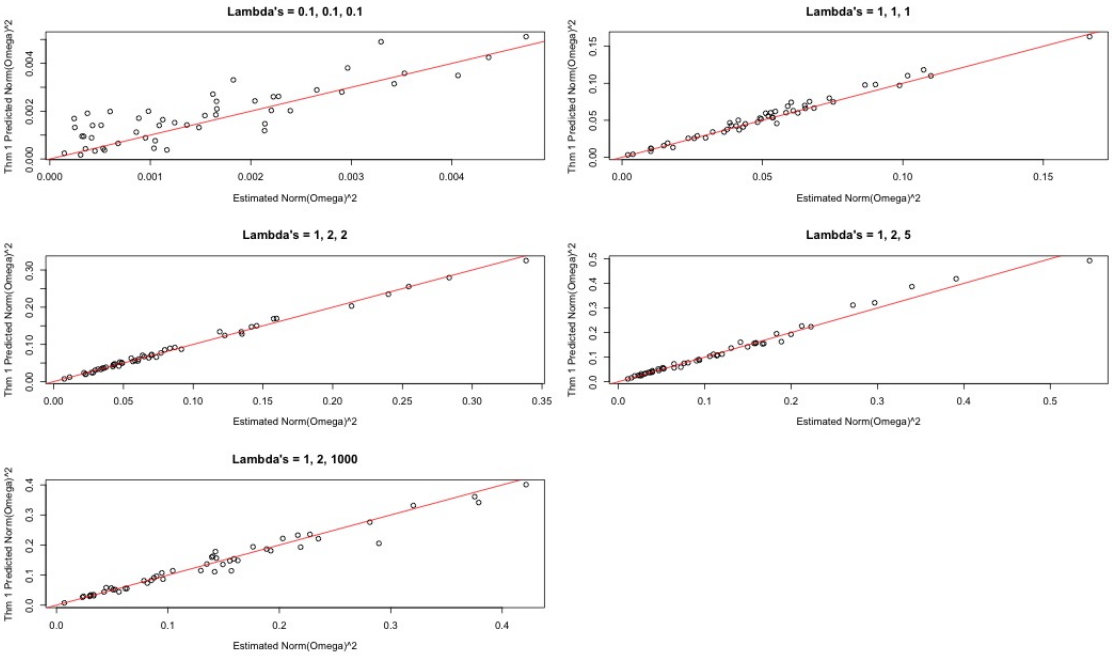
where $\tilde{\boldsymbol{\Omega}}^{\text{OLS}} = \left\|\tilde{\boldsymbol{\Omega}}^{\text{OLS}}\right\|_2\boldsymbol{Q}\boldsymbol{e}_1$ is the QR decomposition of $\tilde{\boldsymbol{\Omega}}^{\text{OLS}}$. From **equ. 11** and **equ. 18**, the accuracy of our inference is dependent on how well we estimate $\left\|\tilde{\boldsymbol{\Omega}}^{\text{OLS}}\right\|_2$ (it also depend on $\boldsymbol{Q}\boldsymbol{e}_1$, but we will ignore this for now). When we underestimate $\left\|\tilde{\boldsymbol{\Omega}}^{\text{OLS}}\right\|_2$, we tend to perform anti-conservative inference on $\boldsymbol{B}$.

We can consolidate these ideas into a single statement about the behavior of $\left\|\hat{\tilde{\boldsymbol{\Omega}}}\right\|_2$ when the methylation data are not informative enough to accurately estimate the cell type effect $\tilde{\boldsymbol{L}}$:

**Theorem 1** *Suppose that* $\tilde{\boldsymbol{\Omega}} = (\omega_1, \ldots, \omega_K)^T \neq 0$ *and* $\mathcal{I}_{confounding} = diag\,(\delta_1, \ldots, \delta_K)$ *such that* $\delta_i n \to \lambda_i$ *for* $i \geq k$ *and* $\delta_i n \to \infty$ *for* $i < k$. *Then as* $n, p \to \infty$,

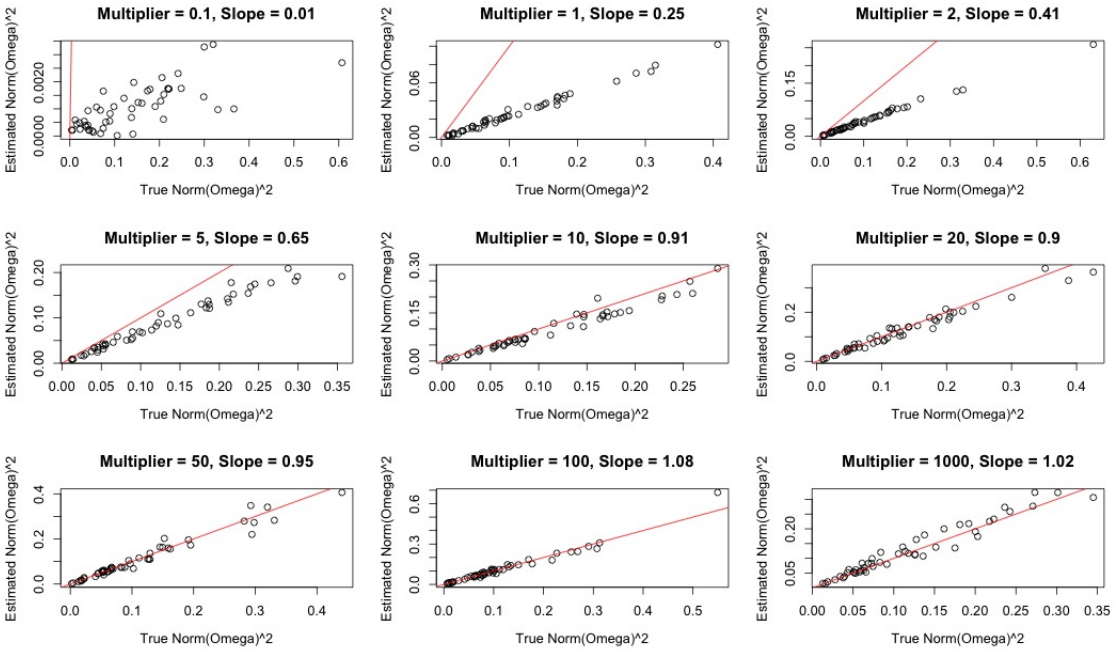
$$\left\|\hat{\tilde{\boldsymbol{\Omega}}}\right\|_2^2 \to \sum_{j=1}^{k-1} \omega_j^2 + \sum_{i=k}^{K} \left(\frac{\lambda_i}{\lambda_i + 1}\right)^2 \omega_i^2 < \left\|\tilde{\boldsymbol{\Omega}}\right\|_2^2 \tag{19}$$

I have a back-of-the-envelope proof of this. Rigorously proving this statement may take some work. Below is a proof of principle that leads me to believe the statement is correct:




(The above data were simulated assuming $\tilde{\boldsymbol{\Xi}} \sim MN_{3\times 100}\,(\mathbf{0}, I_3, I_{100})$, $\boldsymbol{E} \sim MN_{p\times 100}\,(\mathbf{0}, \text{diag}\,(0.15, \ldots, 0.15)\,, I_{100})$. The elements of $\tilde{\boldsymbol{L}}$ are independent normals with variance chosen to get the correct $\mathcal{I}_{\text{confounding}} = \text{diag}\left(\frac{\lambda_1}{100}, \frac{\lambda_2}{100}, \frac{\lambda_3}{100}\right)$)

We can illustrate the effect of the size of $\mathcal{I}_{\text{confounding}}$ on the extent of the shrinkage in $\hat{\tilde{\boldsymbol{\Omega}}}$ with a simple simulation. We first fix the effect due to cell type, $\tilde{\boldsymbol{L}}$, and the residual variability, $\boldsymbol{\Sigma}$, s.t. $\frac{1}{p}\tilde{\boldsymbol{L}}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{L}} = \frac{\lambda}{n}I_K$. We also fix a $2 \times n$ design matrix $\boldsymbol{X} = \begin{bmatrix} \text{Intercept} \\ \text{Disease Status (0's and 1's)} \end{bmatrix}$. We then simulate data according to **Equ. 11** by varying the observed correlation between cell type and covariate of interest ($\tilde{\boldsymbol{\Omega}}^{(\text{OLS})}$) and see how significantly $\hat{\tilde{\boldsymbol{\Omega}}}$ underestimates the observed correlation. Below are our results:




It is clear that the less informative the methylation data are for cell type (i.e. the smaller $\frac{1}{p}\tilde{\boldsymbol{L}}^T\boldsymbol{\Sigma}^{-1}\tilde{\boldsymbol{L}}$ is), the more difficult it is to estimate the observed correlation between cell type and the covariate(s) interest.

Based on the above results, we can identify realistic scenarios when it will be difficult to estimate the correlation between cell type and the covariate(s) of interest. The first is when the non-standardized cell type effect $\boldsymbol{L}$ is small, causing only subtle differences in methylation between cell types. This can occur when cells are taken from the same tissue with little observable cellular heterogeneity. Another related possibility is that the majority of CpG sites have the same methylation status across cell types and only a few CpG's have different methylation patterns across cells. If the cell type effects for only a few of the CpG sites are large, then underestimating the variance given by**equ. 11** can have serious repercussions on the accuracy of any false discovery procedure (need to add an FDR plot). A third may occur when the covariate(s) of interest explains the majority of the variability in cellular heterogeneity. For example, if our covariate of interest is asthma status and everyone with asthma has one cell type composition and everyone without asthma has a different cell type composition, then there is no way to distinguish the cell type effect from the asthma effect and it is impossible to estimate $\tilde{\boldsymbol{L}}$, regardless of the sample size. Mathematically speaking, this occurs when the residual variance $\boldsymbol{\Lambda}$ from **3** and **4** is small, meaning $\tilde{\boldsymbol{L}} = \boldsymbol{L}\boldsymbol{\Lambda}^{1/2}$ is small.

# 5 Real Data Example

Our motivating example came from whole-blood methylation data collected from individuals in the Amish and Hutterite communities, two founder populations with little genetic diversity. The goal of this experiment was to study any potential differences in the two population's methylome. We quantified the methylation from $p = 327,273$ CpG sites from 30 Amish and 30 Hutterite adolescents (ages 7-14) using 450K Illumina chips and measured the whole-blood cell composition for 29 of the Amish and all 30 of the Hutterite adolescents using flow cytometry. The design matrix for this experiment was

$$\boldsymbol{X} = \begin{bmatrix} \text{Intercept} \\ \text{Amish (1) or Hutterite (0)} \end{bmatrix} \tag{20}$$

and the cell type composition matrix was

$$\boldsymbol{C} = \begin{bmatrix} \text{Tcells} & \text{Bcells} & \text{Eosinophils} & \text{Neutroophils} & \text{Monocytes} \end{bmatrix}. \tag{21}$$

Since we have access to whole-blood cell composition for 59 of the 60 study subjects, we have the luxury of comparing inference on $\boldsymbol{B}$ when cell type is and is not observed (we only use the $n = 59$ subjects with observed cell type to compare inference). When we assumed cell type was observed, we modeled the data as

$$\boldsymbol{Y}_{p\times n} = \boldsymbol{B}_{p\times d}\boldsymbol{X}_{d\times n} + \boldsymbol{L}_{p\times K}\boldsymbol{C}_{K\times n} + \boldsymbol{\Gamma}_{p\times r}\boldsymbol{V}_{r\times n} + \boldsymbol{E}_{p\times n} \tag{22}$$

$$\boldsymbol{B} = \begin{bmatrix} \mu_1 & b_{(A-H)_1} \\ \vdots & \vdots \\ \mu_p & b_{(A-H)_p} \end{bmatrix} \tag{23}$$

where the term $\boldsymbol{\Gamma}_{p\times r}\boldsymbol{V}_{r\times n}$ was included to take into account the effect of unmeasured covariates $\boldsymbol{V}$ and $\mu_i$ and $b_{(A-H)_i}$ are the global mean and Amish-Hutterite contrast at CpG $i$. We further assumed that $\boldsymbol{V} \sim MN_{r\times n}(\boldsymbol{\alpha}_{r\times d}\boldsymbol{X}_{d\times n}, I_r, I_n)$. For simplicity we set $\boldsymbol{\alpha} = \boldsymbol{0}$, although the inference we did on $\boldsymbol{B}$ would not change if we assumed $\boldsymbol{\alpha} \neq \boldsymbol{0}$ because we needed to account for any spurious correlation between the unmeasured covariates $\boldsymbol{V}$ and $\boldsymbol{X}$. In order to correct for cell type heterogeneity in this scenario, we simply regressed out $\boldsymbol{C}$, meaning the cell type-corrected model was

$$\boldsymbol{Y}^{(c)} = \boldsymbol{B}\boldsymbol{X}^{(c)} + \boldsymbol{\Gamma}\boldsymbol{V}^{(c)} + \boldsymbol{E}^{(c)} \tag{24}$$

$$\boldsymbol{Y}^{(c)} = \boldsymbol{Y}P^{\perp}_{\boldsymbol{C}^T}, \quad \boldsymbol{V}^{(c)} = \boldsymbol{V}P^{\perp}_{\boldsymbol{C}^T}, \quad \boldsymbol{E}^{(c)} = \boldsymbol{E}P^{\perp}_{\boldsymbol{C}^T} \tag{25}$$

We estimated $r = 4$ additional unobserved factors using BCV (cite Wang, Hastie paper) and used CATE to compute any spurious correlation between $\boldsymbol{V}^{(c)}$ and $\boldsymbol{X}^{(c)}$, as well as $\hat{\boldsymbol{B}} = \begin{bmatrix} \hat{\mu}_1^{(\text{cell})} & \hat{b}_{(A-H)_1}^{(\text{cell})} \\ \vdots & \vdots \\ \hat{\mu}_p^{(\text{cell})} & \hat{b}_{(A-H)_p}^{(\text{cell})} \end{bmatrix}$. Our test statistic used to infer the difference in methylation at CpG site $i$ between Amish and Hutterite adolescents was then

$$\hat{t}_i^{(\text{cell})} = \frac{\hat{b}_{(A-H)_i}^{(\text{cell})}}{\sqrt{\hat{\text{Var}}\left(\hat{b}_{(A-H)_i}^{(\text{cell})}\right)}} \tag{26}$$

where $\hat{\text{Var}}\left(\hat{b}_{(A-H)_1}^{(\text{cell})}\right)$ was computed according to **equ. 11**. We used a $t$-distribution with $n - K - d - r = 48$ degrees of freedom to compute p-values at each CpG site $i = 1, \ldots, p$ to test for differential methylation.

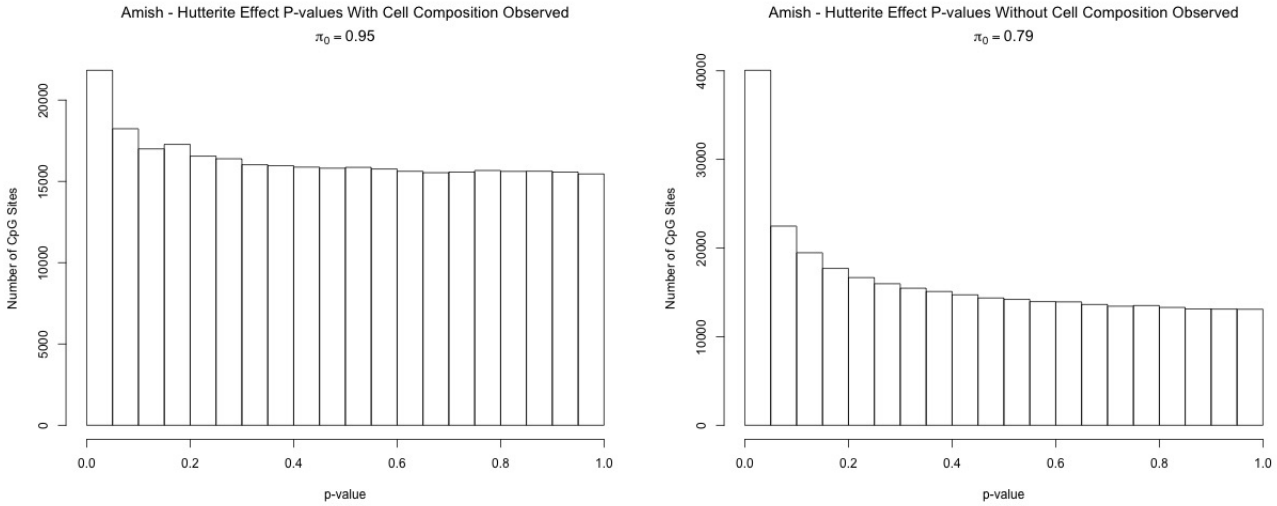
When cell type is UNOBSERVED, we combined unmeasured covariates $\boldsymbol{C}$ and $\boldsymbol{V}$ in **equ. 22** and modeled the data as

$$\boldsymbol{Y}_{p\times n} = \boldsymbol{B}_{p\times d}\boldsymbol{X}_{d\times n} + \tilde{\boldsymbol{L}}_{p\times K'}\tilde{\boldsymbol{\Omega}}_{K'\times d}\boldsymbol{X}_{d\times n} + \tilde{\boldsymbol{L}}_{p\times K'}\tilde{\boldsymbol{\Xi}}_{K'\times n} + \boldsymbol{E}_{p\times n} \tag{27}$$

where $\tilde{\boldsymbol{\Omega}} = \begin{bmatrix} \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Omega} \\ \boldsymbol{\alpha} \end{bmatrix}$, $\tilde{\boldsymbol{L}} = \begin{bmatrix} \boldsymbol{L}\Lambda^{1/2} & \boldsymbol{\Gamma} \end{bmatrix}$ and $\tilde{\boldsymbol{\Xi}} = \begin{bmatrix} \Lambda^{-1/2}\boldsymbol{\Xi} \\ \boldsymbol{V} \end{bmatrix}$. We used the exact same procedure as above to estimate $K' = 6$, $\hat{\tilde{\boldsymbol{\Omega}}}$ and $\hat{\boldsymbol{B}} = \begin{bmatrix} \hat{\mu}_1^{(\text{no cell})} & \hat{b}_{(A-H)_1}^{(\text{no cell})} \\ \vdots & \vdots \\ \hat{\mu}_p^{(\text{no cell})} & \hat{b}_{(A-H)_p}^{(\text{no cell})} \end{bmatrix}$. And just as we did when we observed each individual's whole-blood cell composition, we let

$$\hat{t}_i^{(\text{no cell})} = \frac{\hat{b}_{(A-H)_i}^{(\text{no cell})}}{\sqrt{\hat{\text{Var}}\left(\hat{b}_{(A-H)_i}^{(\text{no cell})}\right)}} \tag{28}$$

and computed p-values to test for differential methylation between both groups using a $t$-distribution with $n - d - K' = 51$ degrees of freedom. Below are p-value histrograms that compare our results when we assume cell composition are and are not observed:




The stark difference between these results is because of the correlation between whole-blood cell composition and group status. The table below gives the standardized regression coefficients after regressing cell type $\boldsymbol{C}$ onto Amish vs. Hutterite group status $\boldsymbol{X}$:

| Cell Type | Amish − Hutterite Standardized Effect |
|---|---|
| T cell | −0.62 |
| B cell | 0.65 |
| Eosinophil | −0.40 |
| Neutrophil | 1.41 |
| Monocyte | 0.65 |

Since cell heterogeneity and group status are so correlated, there is little variability left in $\boldsymbol{X}$ to estimate $\boldsymbol{B}$ after we correct for cell type (i.e. we are in the scenario when $\boldsymbol{\Lambda}^{1/2}$ is small), which in turn cause our test statistics to move closer to 0. We also estimated $\tilde{\boldsymbol{L}}$ and found that $\left(\frac{\left|\frac{1}{p}\hat{\tilde{\boldsymbol{L}}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\tilde{\boldsymbol{L}}}\right|}{\left|\frac{1}{59}I_5\right|}\right)^{1/5} \approx 1.3$. This implies that if $\tilde{\boldsymbol{L}}$ is non-zero and cellular heterogeneity were not observed, the observed CpG methylation data would not be informative enough to estimate cell type and we would perform considerable anti-conservative inference. However, just because $\left|\frac{1}{p}\hat{\tilde{\boldsymbol{L}}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\tilde{\boldsymbol{L}}}\right|$ is slightly larger than $\left|\frac{1}{59}I_5\right|$ does not mean the inference we do without measured cell type is anti-convervative. In fact, if $\tilde{\boldsymbol{L}} = \boldsymbol{0}$, we would expect $\frac{1}{p}\hat{\tilde{\boldsymbol{L}}}^T\hat{\boldsymbol{\Sigma}}^{-1}\hat{\tilde{\boldsymbol{L}}} \approx \frac{1}{59}I_5$. Therefore, we cannot say with 100% confidence that the results without cell type measurements are anti-conservative. We can only use our prior understanding of methylation patterns across different cell types and assume that $\tilde{\boldsymbol{L}} \neq \boldsymbol{0}$. If this were in fact the case, our sample size is not large enough to perform accurate inference without having observed cell composition.

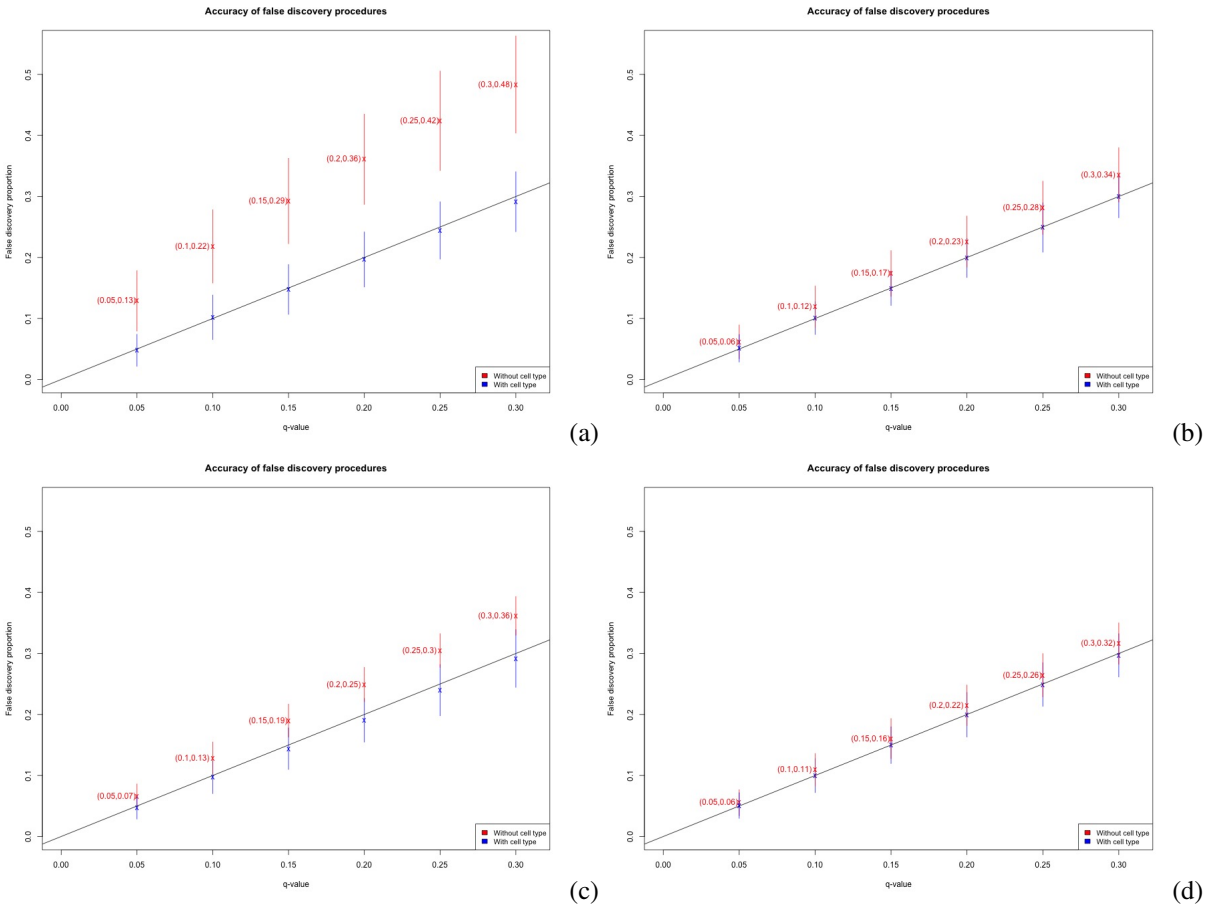
# 6 Simulation Study

Incorrectly estimating the correlation between cell type heterogeneity and the covariate(s) of interest can have deleterious effects on any procedure that attempts to control for false discovery. To explore this, we simulated data according to **equ. 5** with parameters based on estimates in the Amish-Hutterite study (see below table). We partitioned $n = 100$ individuals into two groups, 50 of them Amish and 50 of them Hutterites. In all of our simulations, methylation differed in only one cell type, i.e. $K = 1$.

| Simulation ID | Amish-Hutterite Stand. Effect, $B$ | Cell Type Stand. Effect, $L$ | $\mathcal{I}_{\text{condounding}} = \frac{1}{p} L^T \Sigma^{-1} L$ | Observed Correlation btwn. $\boldsymbol{X}$ and $\boldsymbol{C}$, $\tilde{\boldsymbol{\Omega}}^{\text{OLS}}$ |
|---|---|---|---|---|
| (a) | $0.95\delta_0 + 0.05N(0, 0.4)$ | $0.7\delta_0 + 0.3N(0, \tau_1)$ | $\frac{1.3}{n} I_1$ | 1.41 |
| (b) | $0.95\delta_0 + 0.05N(0, 0.4)$ | $0.7\delta_0 + 0.3N(0, \tau_3)$ | $\frac{1.3 \times 10}{n} I_1$ | 1.41 |
| (c) | $0.95\delta_0 + 0.05N(0, 0.4)$ | $0.7\delta_0 + 0.3N(0, \tau_2)$ | $\frac{1.3}{10n} I_1$ | 1.41 |
| (d) | $0.95\delta_0 + 0.05N(0, 0.4)$ | $0.7\delta_0 + 0.3N(0, \tau_4)$ | $\frac{1.3}{n} I_1$ | 1.41/3 |

The above table gives the parameters for the four different simulations, where the variances $\tau_i$ were chosen to get the correct $\mathcal{I}_{\text{confounding}}$. The below figures plot Storey's q-value against the true false discovery proportion

$$FDP(q) = \frac{\#\{\text{False positives with q-values} \leq q\}}{\#\{\text{CpG sites with q-values} \leq q\}} \tag{29}$$




It is not hard to see that when the correlation $\tilde{\boldsymbol{\Omega}}^{\text{OLS}}$ is large and we have little power to estimate it (plot (a)), we incur significantly more false positives than we would expect using a reasonable false discovery control procedure. The reason for this is two fold. First, when we underestimate $\tilde{\boldsymbol{\Omega}}^{\text{OLS}}$ our estimator for $\hat{B}$ is significantly biased and second, we will underestimate the variance of the components of $\hat{B}$ (see **equ. 9**). However, is we have enough statistical power to accurately estimate $\tilde{\boldsymbol{\Omega}}^{\text{OLS}}$ (plot (b)), our false discovery control procedure is accurate. We also note that if the cell type effect is very small in comparison to the effect of interest (simulation (c)), we significantly underestimate $\tilde{\boldsymbol{\Omega}}^{\text{OLS}}$ but still can accurately control false discovery. This is because the effect $\tilde{\boldsymbol{L}}\tilde{\boldsymbol{\Omega}}^{\text{OLS}}\boldsymbol{X}$ is much smaller in comparison to $\boldsymbol{BX}$, so we tend to find true non-zero Amish-Hutterite effects before Amish-Hutterite effects that are just due to differences in cell composition. Similar reasoning explains why we can control for false discovery when $\tilde{\boldsymbol{\Omega}}^{\text{OLS}}$ is small (as in simulation (d)). Even though we underestimate $\tilde{\boldsymbol{\Omega}}^{\text{OLS}}$, $\tilde{\boldsymbol{\Omega}}^{\text{OLS}} - \hat{\tilde{\boldsymbol{\Omega}}}$ is innocuously small.

# 7 Partially Observed Cell Type

# 8 What I need to add

- My model with partially observed cell type. I only have simulation results here. I should include a brief discussion about the caveats with real data, since additional batch effects may be difficult to estimate.
- A discussion when the residual $E$ is not normally distributed. The confidence we have in factor analysis will depend on the higher order cumulants (notably the fourth kumulant) when the residuals are not normally distributed. This is important when we attempt to diagnose datasets with confounding we cannot estimate (i.e. how much does $\frac{1}{p}\hat{\Gamma}^T\hat{\Sigma}^{-1}\hat{\Gamma}$ differ from $\frac{1}{p}\Gamma^T\Sigma^{-1}\Gamma$ as a function of the distribution of $E$).

# 9 Junk

$\boldsymbol{X}^T = \begin{bmatrix}\boldsymbol{Q}_1 & \boldsymbol{Q}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{R}_{d\times d} \\ \boldsymbol{0}_{n-d\times d}\end{bmatrix}$ be the QR decomposition of $\boldsymbol{X}^T$. $\boldsymbol{W}_{r\times n} \sim MN_{r\times n}(\boldsymbol{\alpha}_{r\times d}\boldsymbol{X}_{d\times n}, I_r, I_n)$ be a random effects matrix that represents additional confounding in the experiment that may be correlated with the covariates of interest. In most data sets, $\boldsymbol{\alpha}$ is small or 0, i.e. the correlation between the confounders and the covariates of interest is small. For simplicity, we assume $\boldsymbol{\alpha} = 0$, although the model and methods can be easily extended for $\boldsymbol{\alpha} \neq 0$. $\tilde{\boldsymbol{\Gamma}} = \begin{bmatrix}\boldsymbol{\Gamma} & \boldsymbol{L}\boldsymbol{\Lambda}^{1/2}\end{bmatrix}$, $\tilde{\boldsymbol{W}} = \begin{bmatrix}\boldsymbol{W} \\ \boldsymbol{\Lambda}^{-1/2}\boldsymbol{\Xi}\end{bmatrix}$ and $\tilde{\boldsymbol{\Omega}} = \begin{bmatrix}\boldsymbol{0}_{r\times d} \\ \boldsymbol{\Lambda}^{-1/2}\Omega\end{bmatrix}$. Note that $\tilde{\boldsymbol{W}} \sim MN_{(r+K)\times n}(\boldsymbol{0}, I_{r+K}, I_n)$.