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 Ξ.
   (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 Ω when data are not informative. Effect on results is small when Ω 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 π_0, the fact that the data suggest that there is confounding (p-value for α). 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.