Homework 8
BIOS-7659/CPBS-7659
Due 12/1 9AM on Canvas

1. Cell Type Composition

   Background on data set: The occurrence of gestational diabetes mellitus (GDM) during pregnancy is believed to alter obesity risk of offspring later in life. It was hypothesized that exposure to maternal GDM in utero will be associated with changes in DNA methylation patterns of key genes and pathways in the offspring, which will mediate the association between in utero exposure and childhood adiposity-related outcomes. Data was collected from the EPOCH (Exploring Perinatal Outcomes in CHildren) study, a historical prospective cohort that enrolled children aged 10.5 on average (T1) who were exposed or not exposed to maternal GDM during the intrauterine life. DNA was extracted from peripheral blood samples collected from children at the T1 EPOCH visit on 85 exposed to GDM and 85 unexposed to GDM, and methylation data was generated using the Illumina Infinium HumanMethylation450 BeadChip. We have provided data on a subset of 10 subjects (all Non-Hispanic Whites) from these two groups to identify methylation sites and nearby genes that show differential methylation between the two groups. The data for this problem is available through a link on Canvas to the "blood" folder on Dropbox.

   - Install the following packages from Bioconductor: `FlowSorted.Blood.450k`, `quadprog` and `IlluminaHumanMethylation450kmanifest`.
   - Use this code to read in the data:

```
baseDir1 <- "blood/plate1"
targets1 <- read.metharray.sheet(baseDir1)
baseDir2 <- "blood/plate2"
targets2 <- read.metharray.sheet(baseDir2)
targets <- rbind(targets1, targets2)

rgSet <- read.metharray.exp(targets=targets, extended=T)
#for part c), use "extended =F"

sampleNames(rgSet) = rgSet[[1]]
getManifest(rgSet)
clindat <- read.table("blood/demographic.txt", sep="\t", header=T)
pData(rgSet)$Sample_Group <- clindat$Exposure
pData(rgSet)$child_sex <- clindat$child_sex
```

   (a) Use SWAN normalization from Homework 7, then find differentially methylated positions based on exposure status using `dmpFinder()`. Are there any DMPs with q-value ≤ 0.10 (or p-value cutoff of 10^{-5})? Summarize the results in a table and include the direction (hyper or hypo methylated based on exposure status).

(b) Because blood is a heterogeneous collection of different cell types, it has been recognized that it may be important to adjust for cell-type composition in your analysis. Why is cell-type composition relevant for DNA methylation studies? What are the methods available to adjust for cell-type composition when cell-types are not directly measured (as in this example)?

- See Houseman *et al.* BMC Bioinformatics (2012) 13:86 and Jaffe & Irizarry, Genome Biology (2014) 15:R31 for more information.

(c) Estimate cell counts using `estimateCellCounts()`. Explain the graph that is displayed by this function. Then repeat part a), but include cell-type composition as covariates in your model. Summarize your results again in a table.

How do the results compare between the unadjusted and adjusted cell-type composition analysis? Provide possible explanations if you do not see differences. Based on the available clinical data file, discuss what other covariates that you may consider including in the model (no need to implement).

- Read the data again with `read.metharray.exp()` but using `extended=F`. Then, use `estimateCellCounts()`, with `meanPlot = T` to estimate cell counts for each of your samples using whole blood (`compositeCellType = "Blood"`) and these cell types `c("CD8T","CD4T","NK","Bcell","Mono","Gran")`.
- Note: `dmpFinder()` does not allow for covariates like cell types. Instead, you can use a simple linear model with `lm()`, but it will be slower than `dmpFinder()`. Be patient!

2. ChIP-Seq

- Download the data provided on Canvas (tup1_IP.txt, mock_IP.txt, input_IP.txt). These are three ChIP-Seq experiments in yeast from Park et al. 2013 PLoS One 8:12 e83506 (`http://www.ncbi.nlm.nih.gov/pubmed/24349523`). Tup1 is a transcriptional repressor and the mock and input are two different controls for comparison. The data (WIG files) were obtained from GEO
  - `http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE51251`
  - GSM1241096 (tup1_DMSO_illumina)
  - GSM1241087 (mock_DMSO_illumina)
  - GSM1241085 (input_DMSO_illumina)

  There was some additional processing of the WIG files so that they could be read into R.
- Install the `bPeaks` package from CRAN. This is a simple package that examines fold change between ChIP and control samples.
- Use the following code to read in the data

```
data(yeastCDS) #for gene location annotation
allData = dataReading("tup1_IP.txt", "mock_IP.txt",
     yeastSpecies = yeastCDS$Saccharomyces.cerevisiae) #read in data
```

(a) By examining the GEO links and reference, what methods were used for sequencing, basecalling, mapping reads and dealing with non-uniquely mapped reads? What is the difference between mock and input controls?

(b) Using `baseLineCalc()`, what is the average number of sequenced mapped at each position? Does the Tup1 ChIP or mock sample have more average reads? (Hint: This function and the function in part c) only need the last column of read counts from `allData$IPdata` and `allData$controlData`)

(c) Examining only chromosome V ("chrV"'), by subsetting the first column of `allData$IPdata` and `allData$controlData`, run `peakDetection()`, with the `baseLineIP` and `baseLineControl` values calculated in part b).

- How many peaks are detected? How does this function define peaks?
- What are the parameters `IPthreshold, controlThreshold, ratioThreshold` and `averageThreshold`?
- This function will create two pdfs "bPeaks_results_dataSummary.pdf" and "bPeaks_results_bPeaksDrawing.pdf." Include the "dataSummary" in your report and explain what is reported, then include the first page of "bPeaks-Drawing" and explain what is reported. You can upload these files separately on Canvas if needed.

(d) The `bPeaks` package uses simple fold change cutoffs. What alternative methods discussed in class would you apply for a more rigorous approach to detect peaks using a statistical testing framework? Describe the methods and statistical approach.

(e) Using `peakLocation()` with the resulting file "bPeaks_results.bed" and `yeastCDS$Saccharomyce` for the `cdsPositions`. How many of the peaks are in genes or promoters? What are those genes?

(f) Now repeat b), c) and e) using the input_IP sample instead ("input_IP.txt"). What differences do you find between the results using mock or input IP? How does that relate to the conclusions in the paper (see Discussion)?