**Epigenetic data in public health: 450k computer lab exercise**

**Johns Hopkins Bloomberg School of Public Health**

**Department of Mental Health Summer Session**

**June 3-4, 2020**

**Kelly Bakulski, Shan Andrews, David Sosnowski, & Brion Maher**

In this exercise, you will learn how to perform analyses with the Illumina 450k HumanMethylation BeadArray. First, you will learn to read in and process raw array image files into quantitative data. Second, you will assess data quality and remove markers and samples that fail quality control standards. You will examine options for background correction, normalization and batch correction. Third, you will analyze preprocessed methylation data at individual sites. Fourth, you will utilize the single site signals to test for enrichment in gene ontologies. Fifth, you will learn to perform regional DNA methylation analyses capitalizing on the correlation structures of nearby probes. Lastly, you will test for association between methylation and genotype at methylation quantitative trait loci. This exercise works with 450k data that is publicly available through the Genome Expression Omnibus (Accession #GSE42861; Liu et al. 2013 Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis *Nature Biotechnology*).

**1. Running RStudio Cloud**

Use a web browser to navigate to our course RStudio Cloud page: <https://rstudio.cloud/spaces/58018/projects>

First, you will need to create a free RStudio Cloud account. Once you create an account, you should have access to a workspace called “psychGenomics\_class” on the left-hand side of the screen. Click on that workspace and you should see a shared project titled, *JHSPH Summer Institute 2020*. Click on the project (it may take a moment to load) and you will be re-directed to an online version of RStudio. All files for the course should be under the “files” tab in the bottom right quadrant of RStudio. You also will notice a red “Temporary” sign at the top of the page. Next to this you can click and create your own copy of the project. This will allow you to add your own notes and pull the material to be run on your local computer. The original version of the project will remain available to you.

Note, to run this software locally on your computer, you will need to: 1) Install the latest version of R: <http://cran.us.r-project.org>; 2) Install R Studio: <http://www.rstudio.com/products/rstudio/download>/

After the install and download, open and run the file EpigeneticsLab2020\_01\_Install.Rmd in R to install the packages needed for the lab.

**2. Read in the image files.**

How are individual arrays (samples) labeled by Illumina in .idat files?

How many image files are there per array (sample)?

The rest of this script assumes that your working directory is the same as the folder where the raw files are located.

library(minfi)

setwd("/file/path/to/idats")

pheno<-read.csv("samplesheet.csv",header=TRUE)

RGset <- read.metharray.exp(file.path(getwd()), targets=pheno, verbose=TRUE)

How many samples are there in our dataset? **Fill in oval 1 [at the end of the document]**

How many RA case samples are in our dataset?

What are the rows of the phenotype file? Columns?

How many CpG sites are queried by this array? **Fill in oval 1**

What classes of data are in a Red-Green-Channel-Set (RGset)?

What human genome annotation version is our dataset linked to?

How many green fluorescent probes are on the array?

**3. Quality control measures**

1. *Generate raw methylation values. Visualize data*

rawMSet <- preprocessRaw(RGset)

#M signal per probe, per sample

Meth<-getMeth(rawMSet)

#U signal per probe, per sample

Unmeth<-getUnmeth(rawMSet)

#Overall intensity, summary value per sample

pd$MQC<-log2(colMedians(Meth))

pd$UQC<-log2(colMedians(Unmeth))

pdf("MvsUplot.pdf")

plot(pd$UQC, pd$MQC, main="M vs. U QC", pch=16, xlab="Log2 Median Unmethylated Intensity", ylab="Log2 Median Methylated Intensity", cex.lab=1.2, cex.main=2)

dev.off()

What is the minimum log2 median methylation signal in the dataset?

How many samples have log2 median methylation and unmethylated signals below 11? \_\_\_\_\_\_\_\_\_\_\_\_\_

Do you see differences in overall intensity by any array covariate? If so, which one(s)?

#Raw density plot

type<-getProbeType(rawMset)

beta.raw<-getBeta(rawMSet)

probe.type<-data.frame(Name=rownames(beta.raw), Type=type)

pdf("Density-plot-preproccessRaw.pdf")

densityPlot(beta.raw, sampGroups=pd$casestatus, main="Raw Beta by Tissue")

densityPlot(beta.raw, sampGroups=pd$Batch, main="Raw Beta by Batch")

plotBetasByType(beta.raw[,1], probeTypes=probe.type, main="Raw Beta by Probe Type, Sample 1")

dev.off()

Describe the shape of the beta density distribution.

Do you see differences in beta density by any array covariate? If so, which one(s)?

Which probe type (I or II) have a wider range of beta values?

#Raw PCA plots

How much variance does the first principal component explain (approximately)?

What principal component appears to capture sex?

What principal components describe batch?

#Detection P

detP <- detectionP(RGset)

failed <- detP>0.01

per.samp<-colMeans(failed)

probe.fail<-failed[rowMeans(failed)>0.1,]

sample.fail<-per.samp[per.samp>0.01]

How many samples failed (at P>0.01) at 1% or more of sites? **Put this in Box 1**

RGset.drop<-RGset[,!colnames(RGset) %in% names(sample.fail)]

How many samples and probes remain? **Put this in Oval 2**

How many probes failed (at P>0.01) at 10% or more of samples? **Put this in Box 2**

load("cross.probes.info.rda")

How many probes are cross-reactive (according to Chen et al 2013, with a cutoff of 47 and above)? **Put this in Box 3**

1. *Noob background correct the raw methylation values. Visualize data*

#Noob

noob<-preprocessNoob(RGset.drop, offset=15, dyeCorr=TRUE, verbose = TRUE)

noob.dropP<-noob[!rownames(noob) %in% rownames(probe.fail),]

noob.dropCross<-noob.dropP[!rownames(noob.dropP) %in% cross.probes.info$TargetID,]

How many probes remain after dropping detection P and cross-reactive probes? **Put this in Oval 3**

#Sex Check

sex<-getSex(GmRawSet)

table(pd$predictedSex, pd$gender)

How many samples have a discrepancy between annotated gender and predicted sex? **Put this in Box 4**

#Cell type

cell<-estimateCellCounts(RGset)

prin.cell<-princomp(cell)

What cell type explains most of the variability in total cell composition?

What is the range in estimated granulocyte percent?

How many samples have estimated monocytes >20%? **Put this in Box 5**

#Noob PCA plots

prin<-princomp(beta.noob)

Compare the raw PC plot painted by batch to the noob PC plot painted by batch.

Key covariates are gender, age, smoking, Gran, and casestatus.

How many samples have a missing value for a key covariate?  **Put this in Box 6**

How many probes/samples remain after dropping sex switches, cell type outliers, and missing covariate? **Put this in Oval 4**

**3. Single site analysis**

What is the lambda statistic for this single-site association screen? How do you interpret this value?

On what chromosome is the most significant CpG site (by p-value)?

Do any loci meet genome-wide significance?

What is the mean % methylation difference between cases and controls for the most significant cpg site (by p-value)?

**4. Gene ontology analysis**

How many entrez gene id’s are represented on the array background?

What is the most highly significant pathway, using the Fisher’s exact test?

**5. Region finding analysis**

Should you use the bootstrap or permutation option for the ‘nullMethod’ argument?

To what gene does the highest ranked DMR (by fwerArea value) map?

Are the cases hyper- or hypo- methylated relative to the controls in the region of the first DMR?

**6. meQTL analysis**

List two important considerations for meQTL studies.

How many SNP to CpG associations tests did you perform?

Analysis

**Box 6**

DNA samples dropped for missing key covariates

**Box 5**

DNA samples dropped for cell composition outliers

**Box 4**

DNA samples dropped for sex mismatch

Batch correction

Cell type estimation

**Box 3**

drop cross reactive sites

**Box 2**

sites failed because of detection P > 0.01 in >10% of sites

preprocessNoob

**Box 1**

DNA samples failed because of detection P >0.01 in >1% of sites