Homework 7
BIOS-7659/CPBS-7659
Due 11/13 in class

1. DNA Methylation QC and Normalization (Illumina 450K)

   - For problems #1 and #2 install the following packages from BioConductor: `shinyMethyl`, `minfi`, `bumphunter` and `IlluminaHumanMethylation450kanno.ilmn12.hg19`
   - The data for this problem is available through a link on Canvas to the "idats" folder on Dropbox. This is an Illumina 450K dataset from The Cancer Genome Atlas (TCGA)

     `http://cancergenome.nih.gov/`

     There are two files for the red and green channels from 6 subjects in the colon adenicarcinoma data set (COAD) from:

     `https://portal.gdc.cancer.gov`

     The clinical and demographic data has been abbreviated in the `targets.csv` file.
   - Here is the link to the `minfi` User's Guide:

     `http://www.bioconductor.org/packages/release/bioc/vignettes/minfi/inst/doc/minfi.html`

     Here is another reference:

     `https://www.bioconductor.org/help/course-materials/2014/BioC2014/minfi_BioC2014.pdf`

     For information on the control probes, see page 6-7 `http://www.filgen.jp/Product/Bioscience/Methyl/Methylation_report.pdf`
   - Run the following code to load the raw .idat files. But change your path name, and also the path in the Basename column in the "SampleSheet" file in the idats directory.

```
baseDir = c("/Users/Katerina/Desktop/7659/homeworks/hw7/idats")
targets = read.metharray.sheet(baseDir)
rgSet <- read.metharray.exp(targets = targets)
annotation(rgSet)
```

   (a) In clinical manuscripts, the first table often includes summaries of clinical and demographic data (e.g., disease status, race, age, etc.). Create a table with this information (means, standard deviations, etc), using `pData(rgSet)` to extract relevant information. How many unique subjects are there?

   (b) From the array annotation information given by `getManifest(rgSet)`, how many Type I and II probes are there?

(c) Use `densityPlot()`, `controlStripPlot()` and `densityBeanPlot()` to display QC plots. In the information from the "targets" file, use "id" for `sampNames` and repeat the QC plots on "sample_type" and "Sex" for `sampGroups` to see if there are differences in cancer versus normal subjects or by sex. Do you see any differences in the beta values between sample type or sex using the QC reports? Are there any samples that appear to be problematic?

(d) Describe their purpose of the different control probes on the array (see link above and `help(qcReport)`). Use `controlStripPlot()` to display the intensity values for the"BISULFITE CONVERSION I" and "NEGATIVE" probes. What do the ranges of these intensity values tell us about the quality of the data?

(e) Illumina also reports detection p-values, how are these calculated? Using the function `detectionP()`, which sample had the largest percentage of detection p-values ≥ 0.05? How many probes have average detection p-value ≥ 0.05 across the 6 samples?

- Save the methylation signals using the `preProcess` series of functions for the Raw data (without normalization) and the SWAN normalization method.

```
mset <- preprocessRaw(rgSet)
msetSWAN <- preprocessSWAN(rgSet)
```

(f) Use multidimensional scaling (MDS) plots to show how samples group by sex or cancer status with `mdsPlot()`. What do you conclude? Are conclusions different if you take more positions with the most methylation variability (1000 vs 10000 positions)? or by using the raw data `mset` compared to the SWAN normalized data `msetSWAN`?

(g) Plot the distribution of beta values before and after SWAN normalization using `plotBetasByType()`. What do you see in the density plots?

2. DNA Methylation Annotation and Differentially Methylated Positions (Illumina 450K)

- Continuing with the the data from problem #1, get genome annotation information using the following code:

```
gset <-mapToGenome(msetSWAN)
annotation <-getAnnotation(gset)
```

(a) What are CpG islands, shores, shelves and open seas? From `annotation()` how many CpG site probes are in each of these types?

(b) Using the SWAN normalized data from problem #,1 `msetSWAN`, find differentially methylated positions (DMP) for cancer status with `getM()`, followed by `dmpFinder()` (which currently does not handle paired samples, so you will need to run it assuming independence). Are there any DMPs with q-value ≤ 0.10? Using a p-value cutoff of 10^{-5}, how many DMPs show hyper or hypomethylation

due to cancer status? Use `plotCpg()` to plot the beta values and then M-values for the top four DMPs. What do trends and effect sizes do you see in the plots?

(c) Repeat part b) but for DMPs between male and females.

(d) Global methylation profiles vary by sex. There is a function `addSex()` to estimate whether each sample is male or female. Are the predicted and given labels correct for Sex? If not, revisit the MDS plot from part 1e)? Do the new predictions group in the plot? Also repeat the analysis in 2c). Now are there DMPs with q-value ≤ 0.10 (or p-value $\leq 10^{-5}$)? Some example code:

```
gset <-mapToGenome(msetSWAN)
gset = addSex(gset)
cbind(pData(gset)$predictedSex, pData(gset)$Sex)
```

(e) We learned about `bumphunter` in class (Jaffe et al., 2012, `https://www.ncbi.nlm.nih.gov/pubmed/22422453`). This sample data set is too small for `bumphunter` to identify significant regions by performing permutations or bootstrap. However, we can use the `getSegment()` function to find regions of extreme values for the differences found in part b).

See User's Guide `http://www.bioconductor.org/packages/release/bioc/vignettes/bumphunter/inst/doc/bumphunter.pdf`

and use the following code to plot one example region. Note, this is just an example. Report on, and provide a plot for a region that shows hypomethylation in more than one CpG site for the cancer subjects.

```
diffs <- dmp$intercept #NOTE: "dmp" is where you saved results from part b)
chr <- annotation$chr
pos <- annotation$pos
cl <- clusterMaker(chr, pos, maxGap = 300) #cluster probes

#Find regions with a stretch of differences
segs <- getSegments(diffs, f = cl, cutoff = 6)

#To plot one of the regions
ind = segs$dnIndex[[1]]
index <- which(cl==cl[ind[1]])
plot(pos[index],diffs[index],
   xlab=paste("position on", chr[ind[1]]), ylab="diff")
points(pos[ind], diffs[ind], pch=16, col=2)
abline(h = 0.05, col = "blue")
```