Homework 7 Solutions
BIOS-7659/CPBS-7659

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

a) We examined DNA methylation data from a subset of samples in the colon adenocarcinoma (COAD) data set obtained from the Genomics Data Commons Portal from NIH. The table below provides a summary of the demographic and clinical data. This is usually “Table 1” in clinical papers. The demographic data is identical between the two groups indicating that these are 3 unique subjects with paired samples of primary tumor and solid tissue normal for the same subject. NOTE: This table combines all samples regardless of pairing.

Table 1	
Sex	
Female	66.6% (n=4)
Male	33.3% (n=2)
Race	
Black/African	33.3% (n=2)
White	66.6% (n=4)
Status	
Solid Tissue Normal	50.0% (n=3)
Primary Tumor	50.0% (n=3)
Age	76.7 ± 6.5 (sd)
Height (cm)	166.3 ± 14.3 (sd)
Weight (kg)	61.9 ± 7.2 (sd)

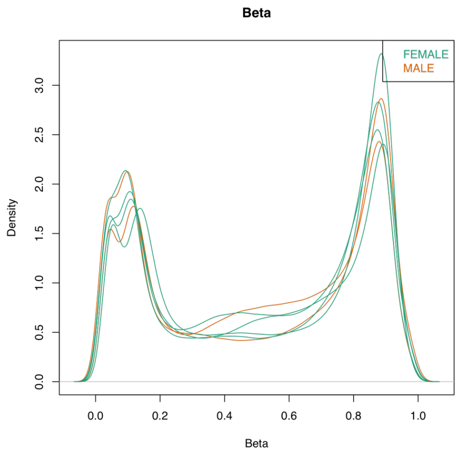
```
Code:
library(shinyMethyl)
library(minfi)
library(IlluminaHumanMethylation450kanno.ilmn12.hg19)
library(bumphunter)

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

table(pData(rgSet)$Sex)
table(pData(rgSet)$patient.race)
table(pData(rgSet)$sample_type)
c(mean(pData(rgSet)$patient.age), sd(pData(rgSet)$patient.age))
c(mean(pData(rgSet)$patient.height), sd(pData(rgSet)$patient.height))
c(mean(pData(rgSet)$patient.weight), sd(pData(rgSet)$patient.weight))
```

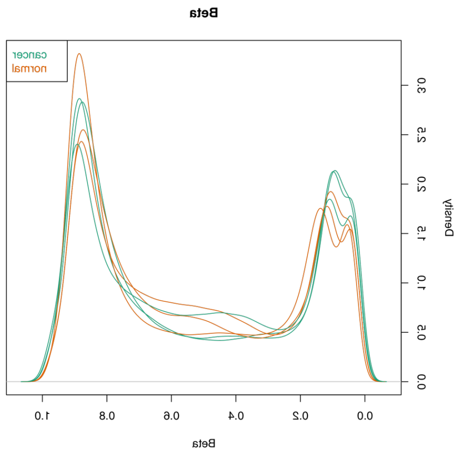
b) There are 135476 Type I probes and 350036 Type II probes. Type I probes are the earlier generation of probes used on the Illumina 27K platform, where there are two probes for each CpG site. Type II was introduced for the Illumina 450K where there is only one probe per CpG site, which allows for more coverage of the genome. The fluorescence resulting from extension of type I probes is not dependant on whether the target site is methylated or unmethylated [1]. For type II probes, green fluorescence denotes a methylated site while red fluorescence denotes an unmethylated site [1]. Type I probes are in CpG islands relatively more than Type II probes (57% vs 21%).

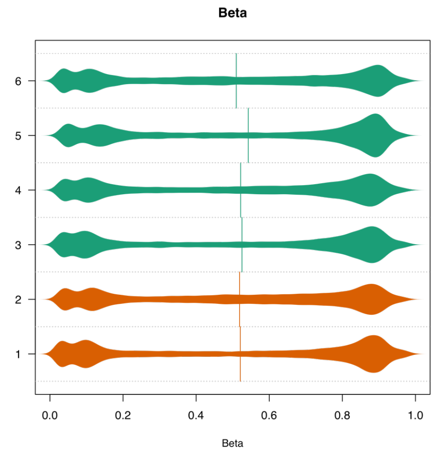
```
Code:
getManifest(rgSet) #annotation information
```

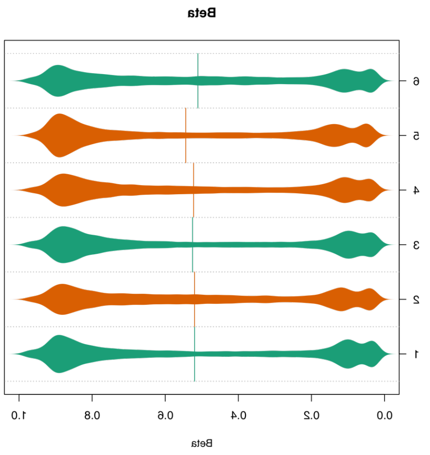
c) The beta value is the ratio of intensities of methylated values to the sum of intensities for methylated and unmethylated values. Although there seems to be more variation among the females, the density plots of the beta values do not show any major differences by sex or status. There are generally two peaks expected near 0 and 1, due to the bimodal nature of the beta value indicating either lack of methylation or methylation respectively.

However, there seem to be two small peaks around 0 as well. Bean plots are an alternative display for the density. For sex and status, the bean plot is showed below the respective histogram. Here the bimodal distribution near 0 for some samples is more obvious. Sample 5 appears to have a higher mean than the other samples and may be problematic. Overall, these plots indicate the need for normalization.





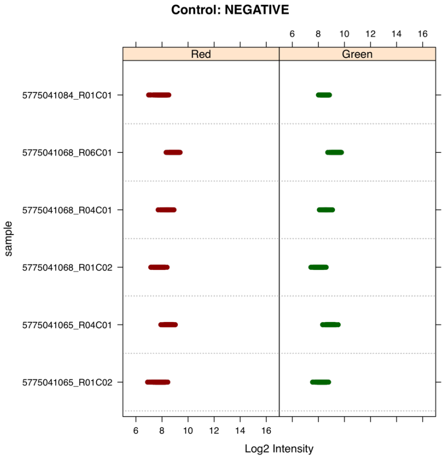


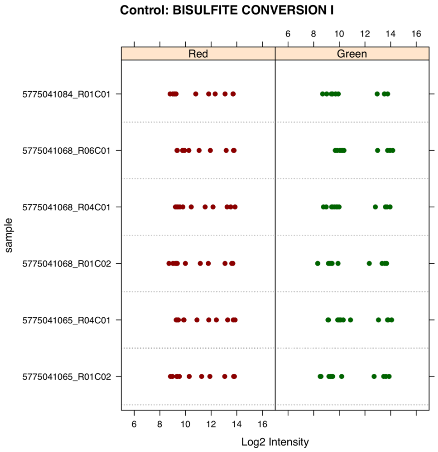


```
Code:
densityPlot(rgSet, sampGroups=targets$Sex, main="Beta", xlab="Beta")
densityPlot(rgSet, sampGroups=targets$Status, main="Beta", xlab="Beta")
densityBeanPlot(rgSet, sampGroups=targets$Sex)
densityBeanPlot(rgSet, sampGroups=targets$Status)
```

d) The Illumina 450K platform has 848 control probes. These contain negative control probes (613), probes for between array normalization (186) and others designed for quality control, including assessing the bisulfite conversion rate (49) [2]. The negative control probes should not hybridize to DNA. They are useful for determining background hybridization levels. We would be concerned if the negative control values were large, if any of the samples were much different than the others or if the ranges between the red and green channels varied. For our samples, we do not see any samples that are problematic and the ranges are similar between the two channels, and on the lower end compared to the bisulfite conversion probes.

Bisulfite Conversion probes are used to examine the efficiency of bisulphite conversion process for Type I and Type II probes separately. For Type I, there are 12 probes. Half of the probes are expected to have high signal in the green channel, indicating that the bisulfite conversion reaction was successful, and similarly, the other half are expected to have high signal in the red channel. In the plots we see a bimodal distribution in the green channel, with one peak similar to the background levels from the negative control probes and the other peak has higher levels. For the red, we see something similar but the probes with higher levels are more diffuse. However, we see the same pattern and distribution range for all samples, therefore there do not seem to be any problematic samples based on these types of control probes.





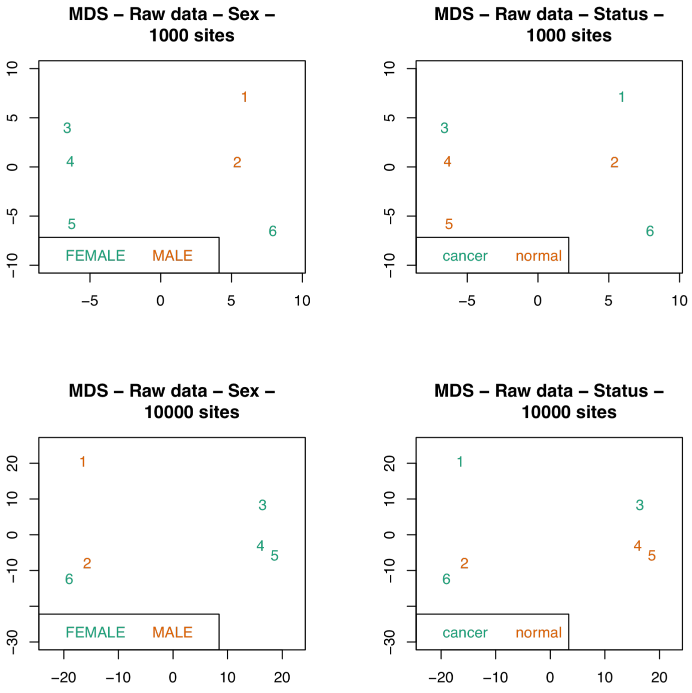
```
Code:
controlStripPlot(rgSet, controls = c("NEGATIVE"), sampNames =
       targets$bcr_sample_barcode)
controlStripPlot(rgSet, controls = c("BISULFITE CONVERSION I"),    sampNames =
       targets$bcr_sample_barcode)
```

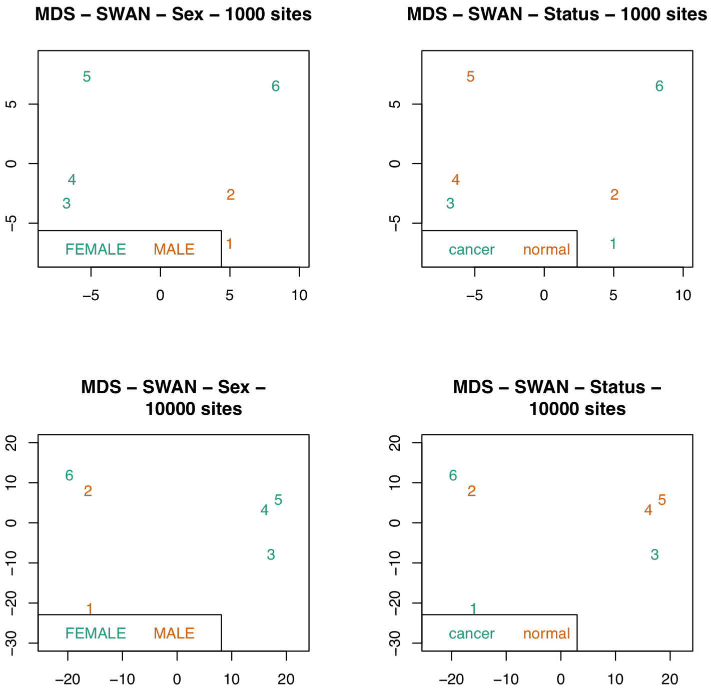
e) Detection p-values are a QC metric for each probe. Assuming a normal distribution and using the mean and standard deviation of the negative control probes, the detection p-value is the probability of observing a hybridization value more extreme at that probe. Large detection p-values indicate probes with values similar or smaller than the background probes and it is recommended to remove those probes. In this data set, sample 5, which also was flagged in the QC analysis, has the largest percentage (0.14%) of probes with detection p-values above 0.05, but this is a relatively small percentage so it is not of concern. Of the 6 samples, 853 probes have an average detection p-value greater than 0.05 and can be removed from the analysis.

```
Code:
detP = detectionP(rgSet)
failed = detP > 0.05
numfail.col = colMeans(failed)
which.max(numfail.col) #Sample 5
badProbes = rowMeans(detP) >= 0.05
sum(badProbes) #853
```

f) Multidimensional scaling is a method for visualizing high-dimensional data in two or three dimensions while retaining the level of similarity between observations (or sample). The built-in function `mdsPlot()` uses only a subset of the most variable positions. Below, we examined the effect of using the top 1000 compared to 10,000 positions and whether normalization by the Subset-quantile Within Array Normalization (SWAN) method affected the groupings. SWAN enforces the distributions to be similar between the two probe types using a quantile-based method.

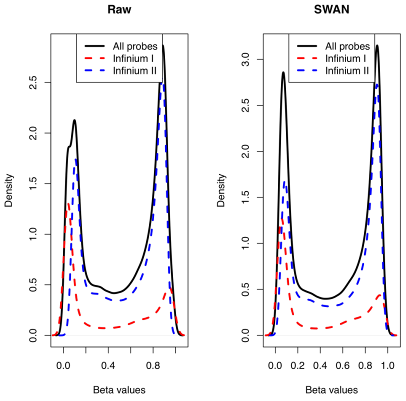
In all plots we see that there is one female (#6) that does not group with the other female samples. We also see that the cancer samples do not group. Increasing the number of positions changed the scales on the axes and brought some samples closer together (e.g., 3,4 and 5, also 2 and 6) but made others further apart (e.g., 1). Normalization had not discernible effect.





```
Code:
mset = preprocessRaw(rgSet)
set.seed(1234)
msetSWAN=preprocessSWAN(rgSet)

#Raw data
par(mfrow = c(2,2))
nump =1000
mdsPlot(mset, sampGroups = targets$Sex, sampNames = targets$id,
        numPositions=nump, ylim = c(-10,10), main = "MDS - Raw data - Sex -
        1000 sites")
mdsPlot(mset, sampGroups = targets$Status, sampNames = targets$id,
numPositions=nump, ylim = c(-10,10), main = "MDS - Raw data - Status -
        1000 sites")

nump =10000
mdsPlot(mset, sampGroups = targets$Sex, sampNames = targets$id,
numPositions=nump, ylim = c(-30,25), main = "MDS - Raw data - Sex -
        10000 sites")
mdsPlot(mset, sampGroups = targets$Status, sampNames = targets$id,
numPositions=nump, ylim = c(-30,25), main = "MDS - Raw data - Status -
        10000 sites")

#SWAN normalized data
par(mfrow = c(2,2))
nump =1000
mdsPlot(msetSWAN, sampGroups = targets$Sex, sampNames = targets$id,
        numPositions=nump, main = "MDS - SWAN - Sex - 1000 sites")
mdsPlot(msetSWAN, sampGroups = targets$Status, sampNames = targets$id,
        numPositions=nump, main = "MDS - SWAN - Status - 1000 sites")

nump =10000
mdsPlot(msetSWAN, sampGroups = targets$Sex, sampNames = targets$id,
        numPositions=nump, ylim = c(-30,20), main = "MDS - SWAN - Sex -
        10000 sites")
mdsPlot(msetSWAN, sampGroups = targets$Status, sampNames = targets$id,
        numPositions=nump, ylim = c(-30,20), main = "MDS - SWAN - Status -
        10000 sites")
```

g) From the density plot by type of probe, Type II probes have a larger peak close to 1, while Type I probes have a larger peak close to 0. After SWAN normalization, the peaks for the two probe types are aligned much better, which facilitates comparing values between the probe types.



```
Code:
par(mfrow = c(1,2))
plotBetasByType(mset[,1], main = "Raw")
plotBetasByType(msetSWAN[,1], main = "SWAN")
```

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

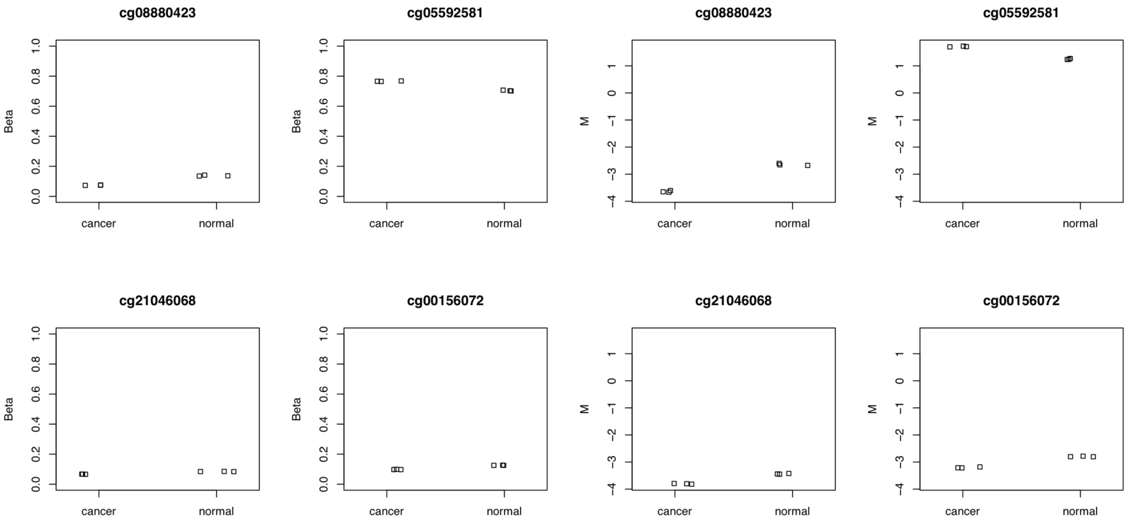
a) CpG islands are clusters of CpG and G+C rich regions within at least 200 bases, CG content >50% and Obs/Exp CpG ratio >0.60 [1]. Shelves and shores are neighboring regions to CpG islands. Everything else is considered the “Open Sea.” On this array there are 150254 probes in CpG islands and 176047 probes in the Open Sea

Island	N_Shelf	N_Shore	OpenSea	S_Shelf	S_Shore
150254	24844	62870	176047	22300	49196

```
Code:
gset <-mapToGenome(msetSWAN)
annotation <-getAnnotation(gset)
table(annotation$Relation_to_Island)
```

b) The `dmpFinder()` function uses linear regression or an F-test to identify sites associated with a continuous or categorical phenotypes respectively. There are no differentially methylated positions (DMP) at a q-value of 0.10, but there are 8 DMPs at a p-value of 10^-5. A negative intercept denotes hypomethylation at that site in cancer, while a positive intercept denotes hypermethylation. Only one of the 8 DMPs is hypermethylated, while the rest are hypomethylated. The beta and M-values for last four sites are displayed below. From the beta value plots we see that the changes in percent methylation are very small. However, the analysis is performed using M values. From the M value plots, the differences are still not extreme but compared to the previous plots, more visibly different between cancer and normal tissues. NOTE: This function does not handle paired tests, therefore we assume all samples are independent and ignore the inherent pairing.

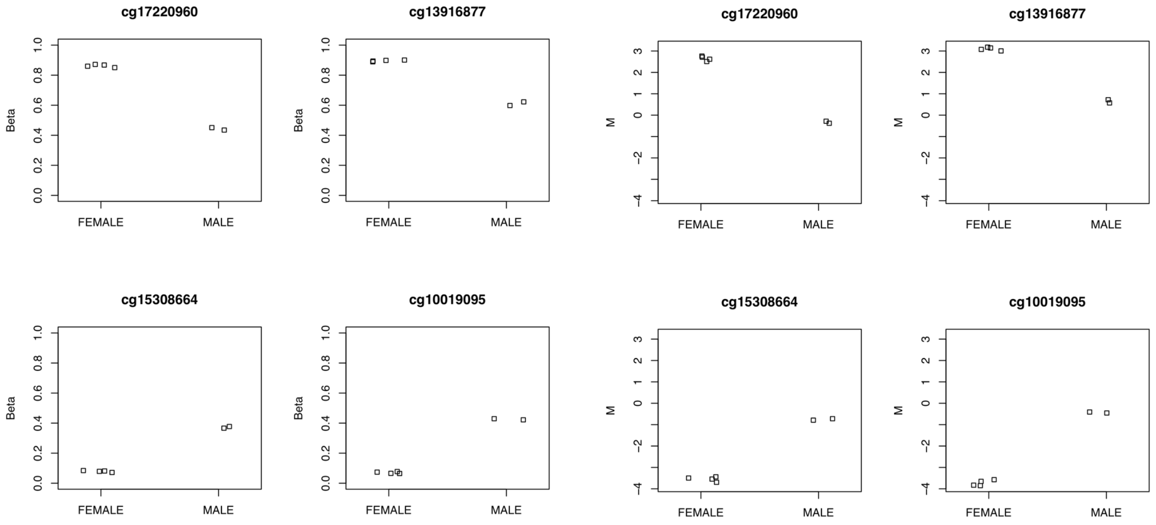
Site	**Intercept**	**F-stat**	**p-value**	**q-value**
cg14488592	-3.83	1541.67	2.51E-06	0.223346
cg13163765	-3.35	1526.90	2.56E-06	0.223346
cg10033612	-3.27	1371.66	3.17E-06	0.223346
cg12298140	-3.82	1198.88	4.15E-06	0.223346
cg08880423	-3.64	1198.46	4.15E-06	0.223346
cg05592581	1.71	1181.57	4.27E-06	0.223346
cg21046068	-3.80	1163.07	4.41E-06	0.223346
cg00156072	-3.20	980.89	6.19E-06	0.274470



```
Code:
set.seed(1234) #results may vary based on the seed
msetSWAN=preprocessSWAN(rgSet)
Mnorm = getM(msetSWAN, betaThreshold = 0.001, type = "beta")
dmp = dmpFinder(Mnorm, pheno  = targets$Status, type  = "categorical")
sum(dmp$q < .10)
sum(dmp$p < 10^-5)
dmp[dmp$p <= 10^-5,]
sum(dmp[dmp$p <= 10^-5,]$intercept<0)
cpgs = row.names(dmp[dmp$p < 10^-5,])[5:8]
grp <- pData(mset)$Status
par(mfrow=c(2,2))
plotCpg(msetSWAN, cpg=cpgs, pheno=grp, type="categorical")
par(mfrow=c(2,2))
plotCpg(msetSWAN, cpg=cpgs, pheno=grp, type="categorical",  measure = c("M"))
```

c) The DMP analysis was repeated using sex as the categorical phenotype. There are no DMPs at a q-value of 0.10, but there are 16 DMPs at a p-value of 10^-5. Eleven of the DMPs are hypomethylated in females and 5 DMPs are hypermethylated. Compared to the cancer status analysis, in the beta value plots we see much larger differences between males and females, with up to 40% differences in methylation percentages.

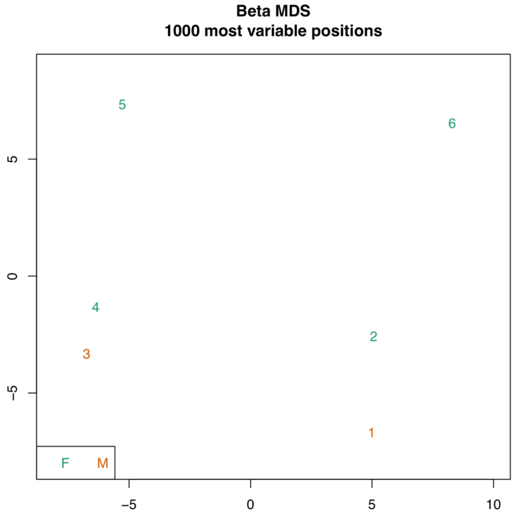
Site	Intercept	F-stat	p-value	q-value
cg10013343	-3.52	1203.37	4.12E-06	0.2274634
cg20648899	-1.65	1195.60	4.17E-06	0.2274634
cg11403874	-3.28	1188.06	4.23E-06	0.2274634
cg15479068	-2.77	1161.99	4.42E-06	0.2274634
cg17220960	2.65	1115.91	4.79E-06	0.2274634
cg13916877	3.10	1108.32	4.86E-06	0.2274634
cg15308664	-3.54	1074.63	5.16E-06	0.2274634
cg10019095	-3.73	1023.96	5.69E-06	0.2274634
cg13764106	-3.20	987.49	6.11E-06	0.2274634
cg10492999	1.92	986.69	6.12E-06	0.2274634
cg22137373	2.09	925.88	6.95E-06	0.2274634
cg02854536	-2.61	925.10	6.96E-06	0.2274634
cg00393263	2.38	887.75	7.56E-06	0.2274634
cg14215586	-3.62	842.10	8.39E-06	0.2274634
cg00488787	-3.71	813.48	8.99E-06	0.2274634
cg05949171	-2.44	789.26	9.55E-06	0.2274634



```
Code:
dmp2 = dmpFinder(Mnorm, pheno  = targets$Sex, type  = "categorical")
sum(dmp2$q < .10)
sum(dmp2$p < 10^-5)
dmp2[dmp2$p <= 10^-5,]
sum(dmp2[dmp2$p <= 10^-5,]$intercept<0)
```

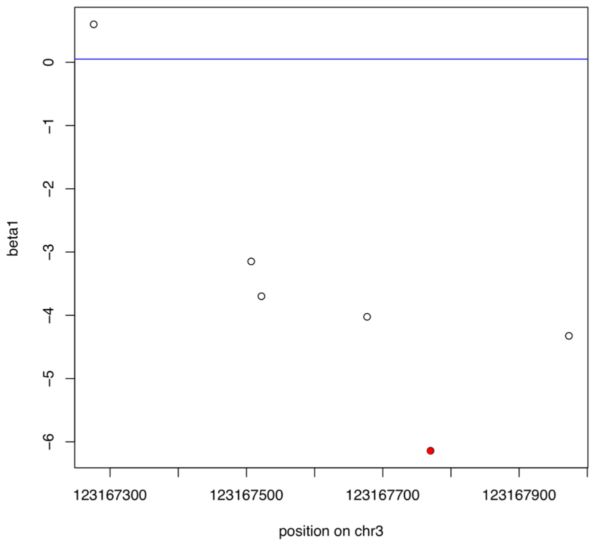
d) The function addSex() estimates the sex of the sample based on methylation status of the X and Y chromosomes. The results below indicate that two samples have different predicted sex than that given by the annotation file. The updated MDS plot now groups the males and females with respect to the y-axis, although the separation is not that strong between the two groups. Repeating the DMP analysis, there are more differences between males and females. We find 54 DMPs with a q-value of 0.10 and 42 with a p-value of 10^-5.

Predicted	Annotated
M	MALE
F	MALE
M	FEMALE
F	FEMALE
F	FEMALE
F	FEMALE



```
gset <-mapToGenome(msetSWAN)
gset = addSex(gset)
cbind(pData(gset)$predictedSex, pData(gset)$Sex)
mdsPlot(msetSWAN, sampGroups = pData(gset)$predictedSex, sampNames =targets$id)

dmp3 = dmpFinder(Mnorm, pheno= pData(gset)$predictedSex, type  = "categorical")
sum(dmp3$p < 10^-5)
sum(dmp3$q < 0.10)
```

e) After examining the top clusters that are hypomethylated in cancer, the fourth one was selected to display below because it has a series of negative intercepts. Note that the y-axis is the intercept from the model fit (not the methylation beta value). The blue line indicates no difference in methylation.



```
Code:
diffs = dmp$intercept #where you saved your dmp results
chr = annotation$chr
pos = annotation$pos
cl <- clusterMaker(chr, pos, maxGap = 300) #cluster probes
segs <- getSegments(diffs, f = cl, cutoff = 6) #find regions with

#Plot a few of the top regions
par(mfrow = c(2,2))
for(j in 1:4)
{
        ind = segs$dnIndex[[j]]
        index <- which(cl==cl[ind[1]])
        plot(pos[index],diffs[index],
             xlab=paste("position on", chr[ind[1]]), ylab="beta1")
        points(pos[ind], diffs[ind], pch=16, col=2)
        abline(h = 0.05, col = "blue")
}

j=4 #plot 4th one
ind = segs$dnIndex[[j]]
index <- which(cl==cl[ind[1]])
plot(pos[index],diffs[index], xlab=paste("position on", chr[ind[1]]),
       ylab="beta1")
points(pos[ind], diffs[ind], pch=16, col=2)
abline(h = 0.05, col = "blue")
```

References:

[1] Price et al., Epigenetics & Chromatin 2013 6:4
[2] Fortin et al., 2013, Genome Biology 15:503