Homework 7 - BIOS 7649

Dominic Adducci

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

```
## [read.metharray.sheet] Found the following CSV files:

## [1] "C:/Users/domin/Documents/Biostatistics Masters Program/Spring 2024/SMG-BIOS7659/Homework 7/Sampl

##                           array                      annotation
## "IlluminaHumanMethylation450k"                  "ilmn12.hg19"
```

Part 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?

Table 1: Demographic Data

Age(mean,sd)	76.67(6.47)
Height(mean,sd)	166.3(14.25)
Weight(mean,sd)	61.9(7.23)
Patient Status (n = 3)	
With Cancer	3
Without Cancer	3
Patient Sex (n = 3)	
Male	2
Female	4

Part B

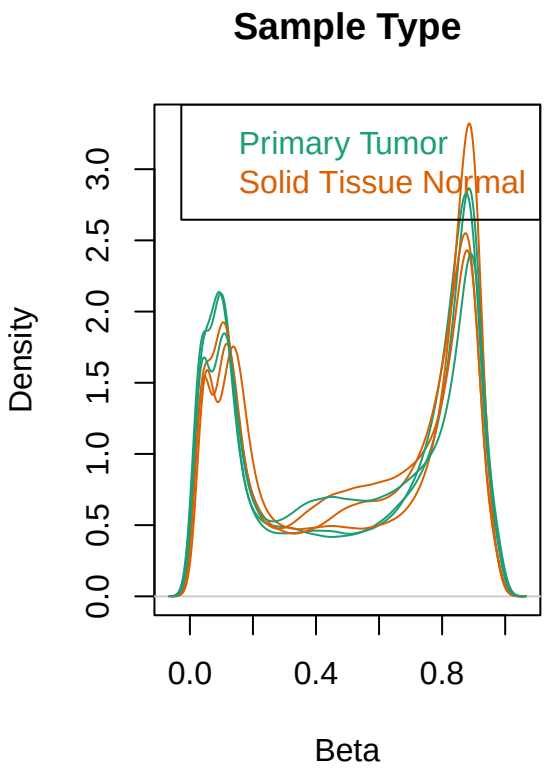
From the array annotation information given by **getManifest(rgSet)**, how many Type I and II probes are there? What is the difference between Type I vs II probes?

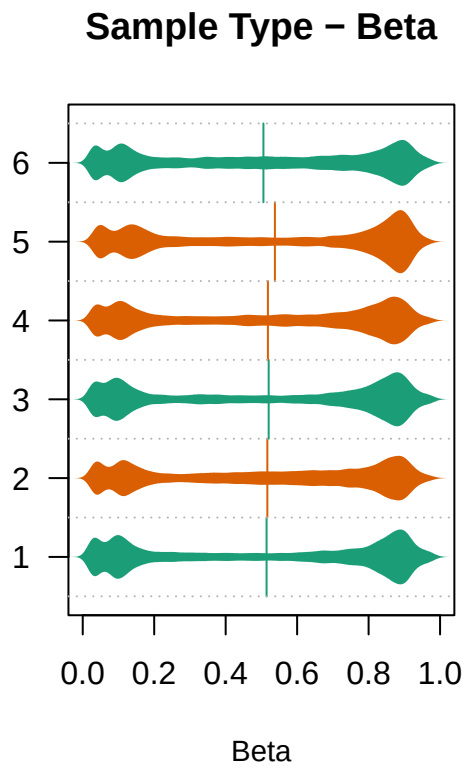
There are 135,476 Type I probes and 350,036 Type II probes.

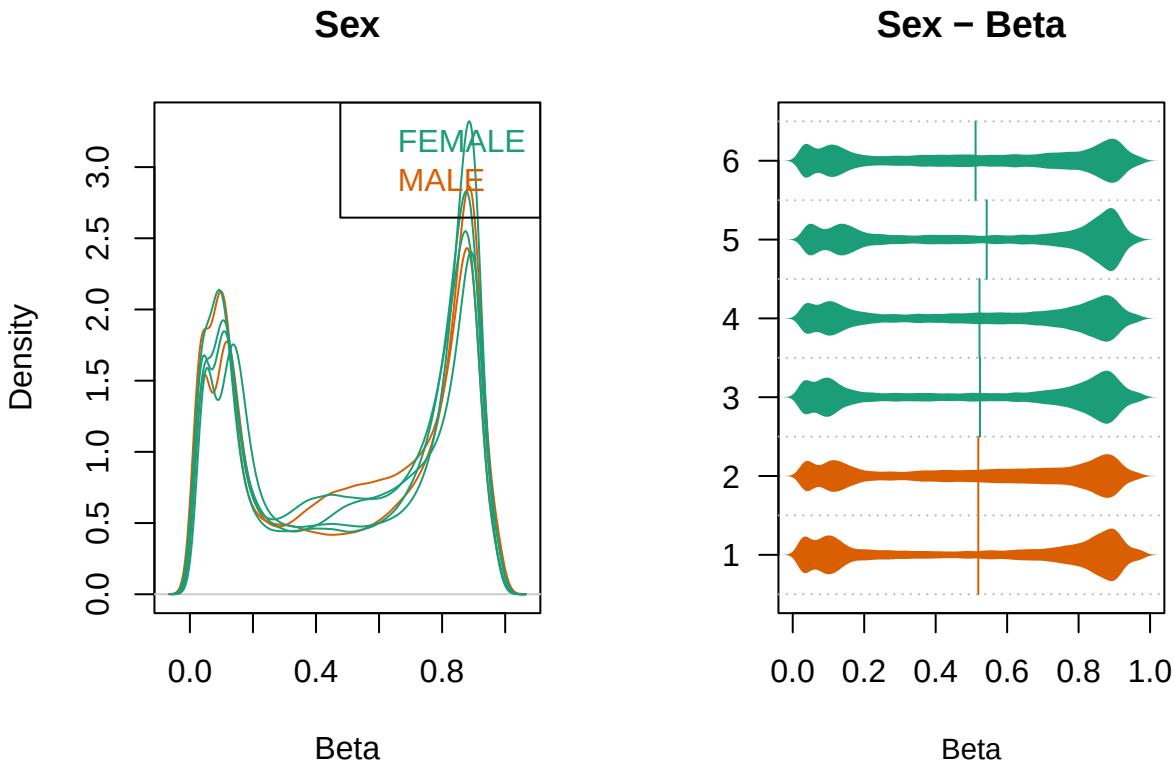
Part C

Use **densityPlot()** 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 and sex using the QC reports? Are there any samples that appear to be problematic?

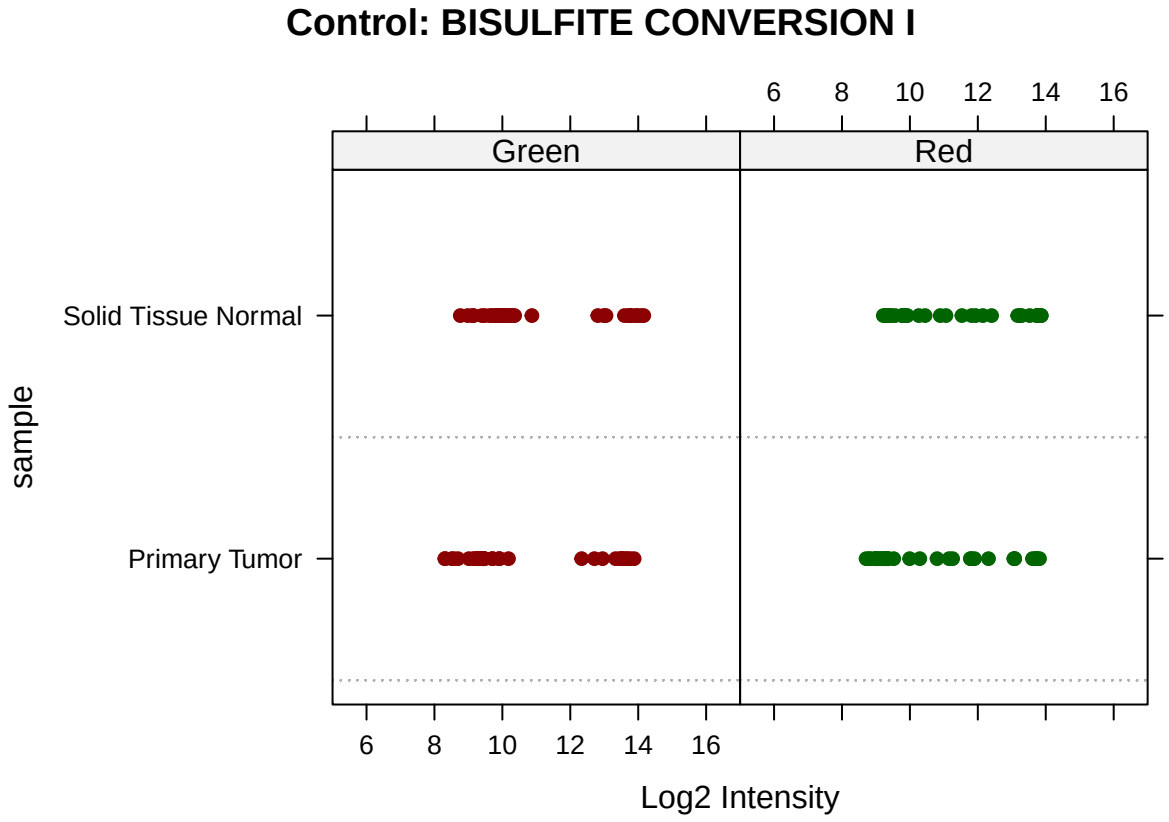


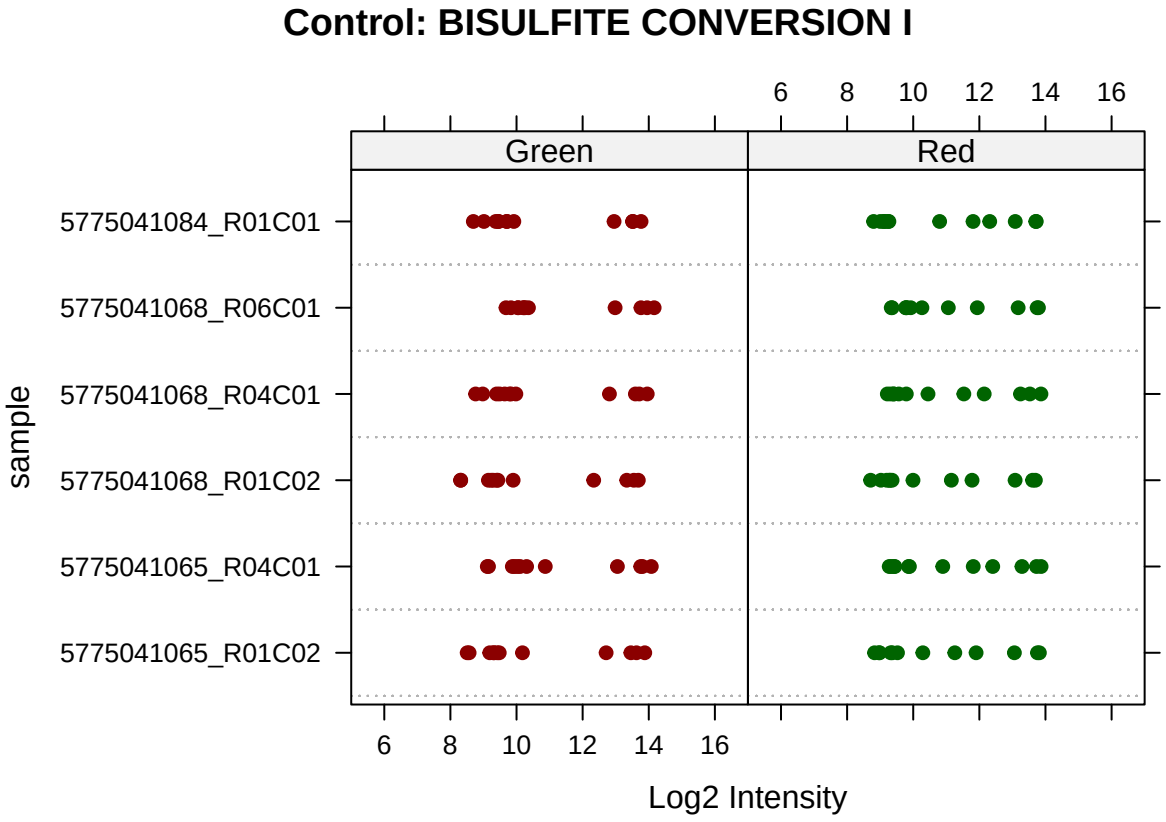


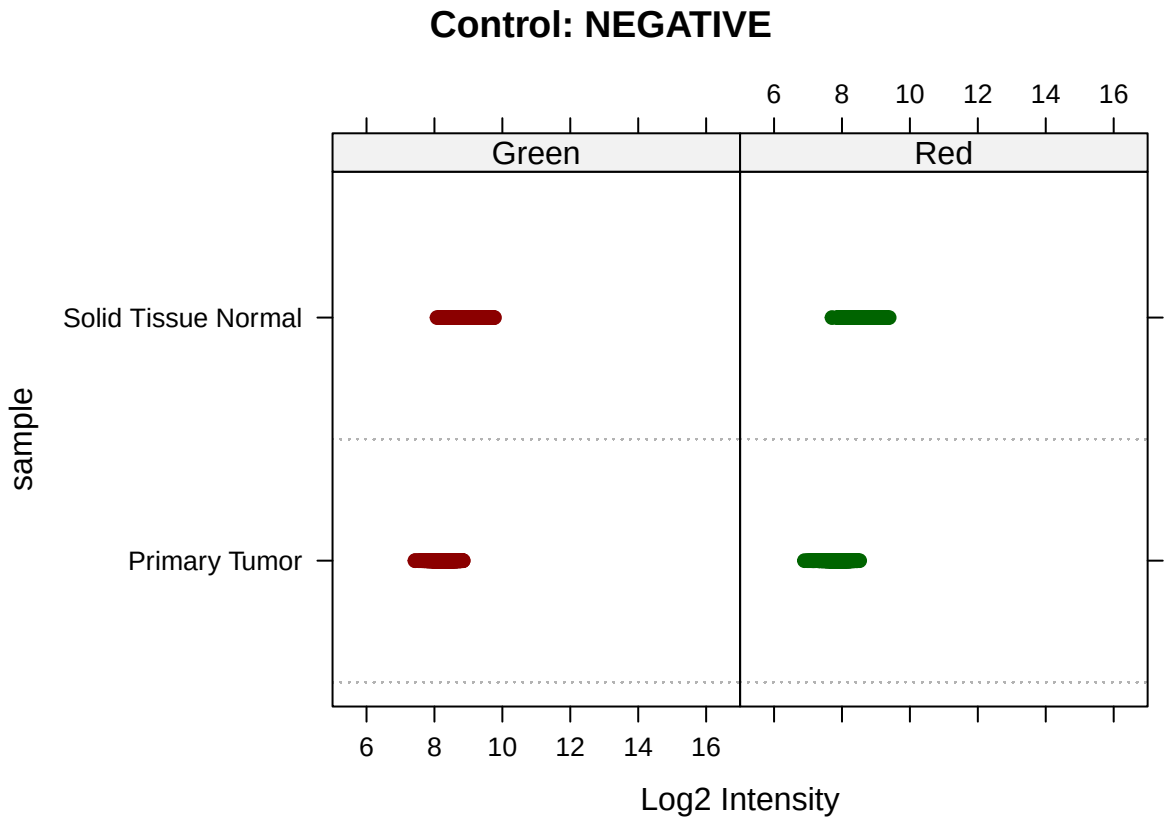


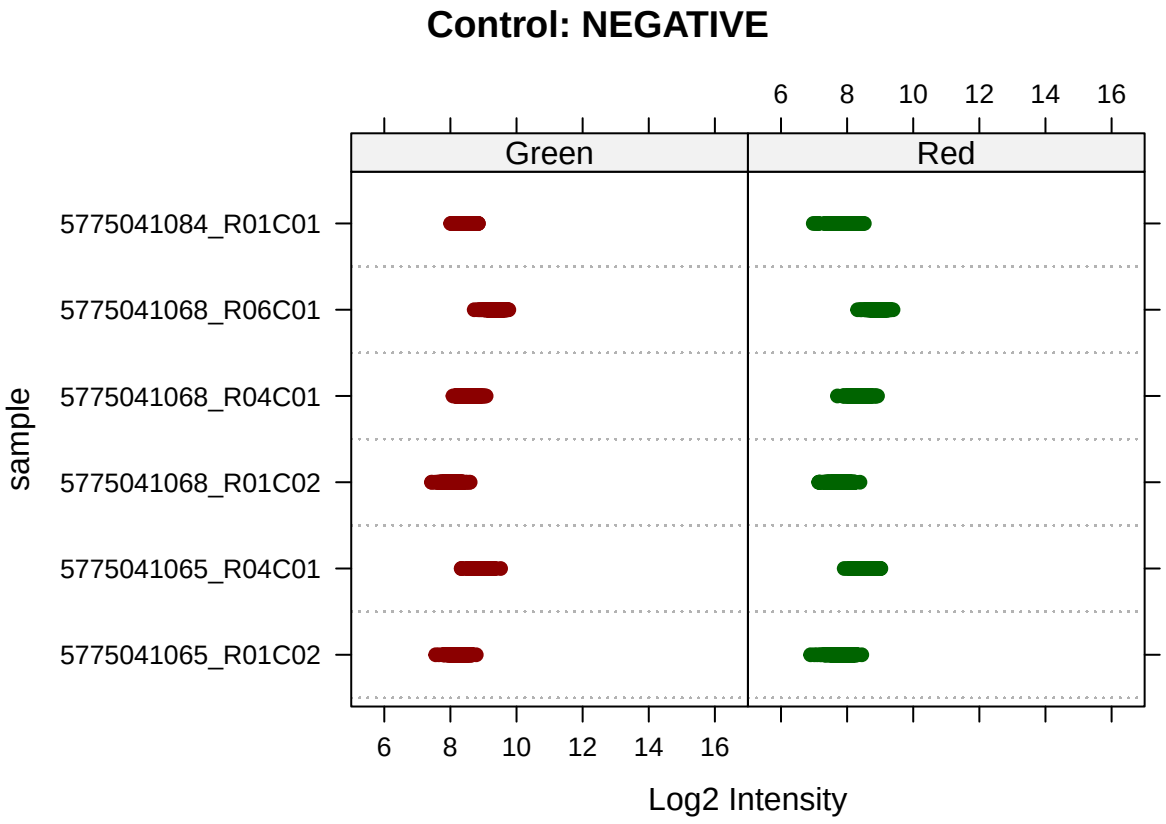
Part D

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?


Control: BISULFITE CONVERSION I
Green
Red
Solid Tissue Normal
Primary Tumor
sample
6
8
10
12
14
16
Log2 Intensity


Control: BISULFITE CONVERSION I
Green
Red
sample
5775041084_R01C01
5775041068_R06C01
5775041068_R04C01
5775041068_R01C02
5775041065_R04C01
5775041065_R01C02
6
8
10
12
14
16
Log2 Intensity


Control: NEGATIVE
Green
Red
Solid Tissue Normal
Primary Tumor
sample
6
8
10
12
14
16
Log2 Intensity



Part 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?

Table 2: P-Values >= to 0.05

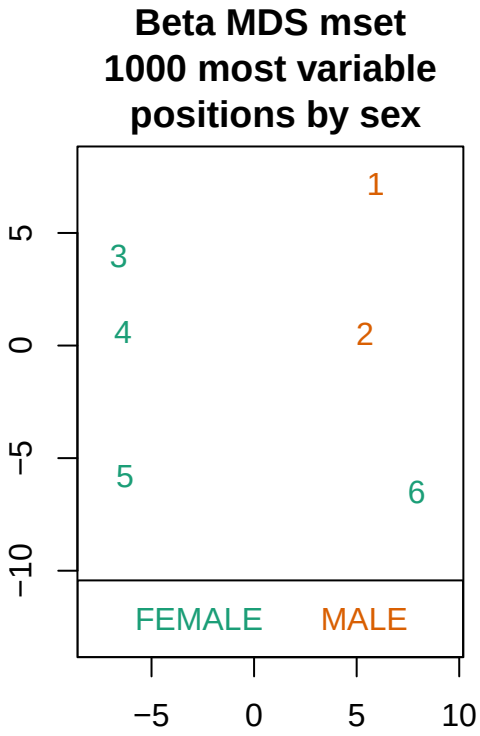
Sample	Sum of p-values >= 0.05
5775041065-R01C02	133
5775041065-R04C01	567
5775041068-R01C02	102
5775041068-R04C01	493
5775041068-R06C01	681
5775041084-R01C01	554

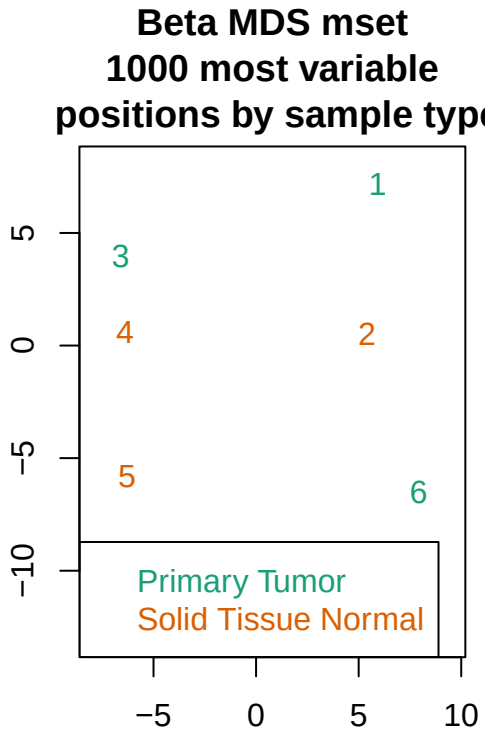
There are 853 probes that have an average detection p-value ≥ 0.05 across the 6 samples.

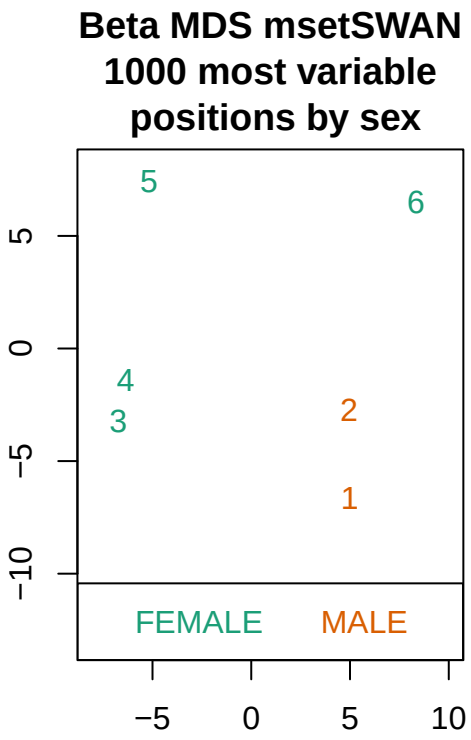
Part F

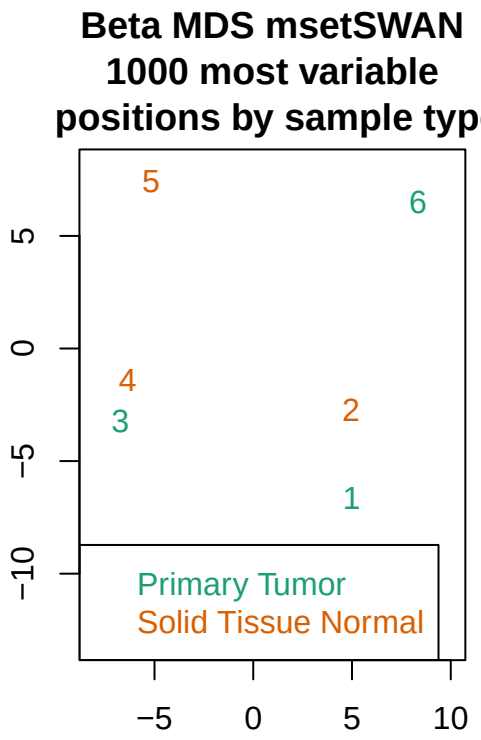
Use multidimensional scaling (MDS) plots to show how samples group by sex or cancer status with **mdsPlot()**. What do you conclude? Are conclusions different if you take more positions with the most

methylation variability (1000, vs 10000 positions)? Or by using the raw data **mset** compared to the SWAN normalized data **msetSWAN**.



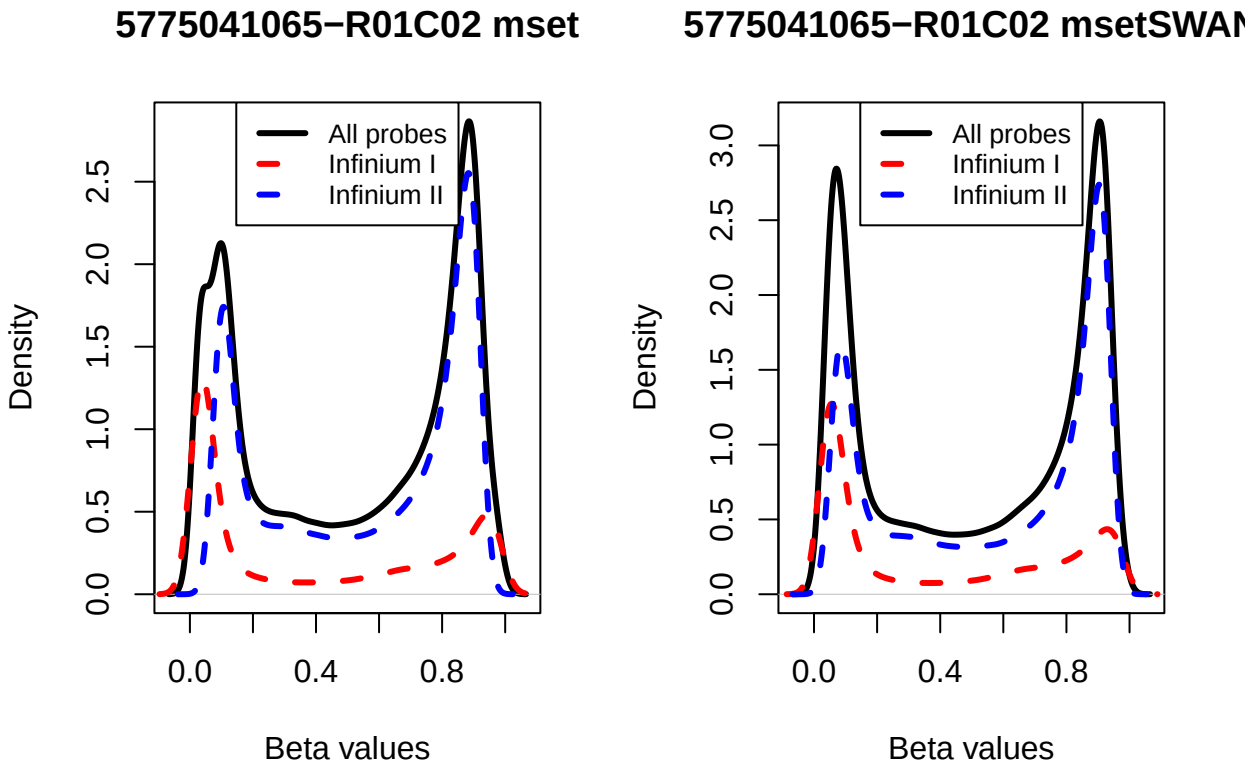


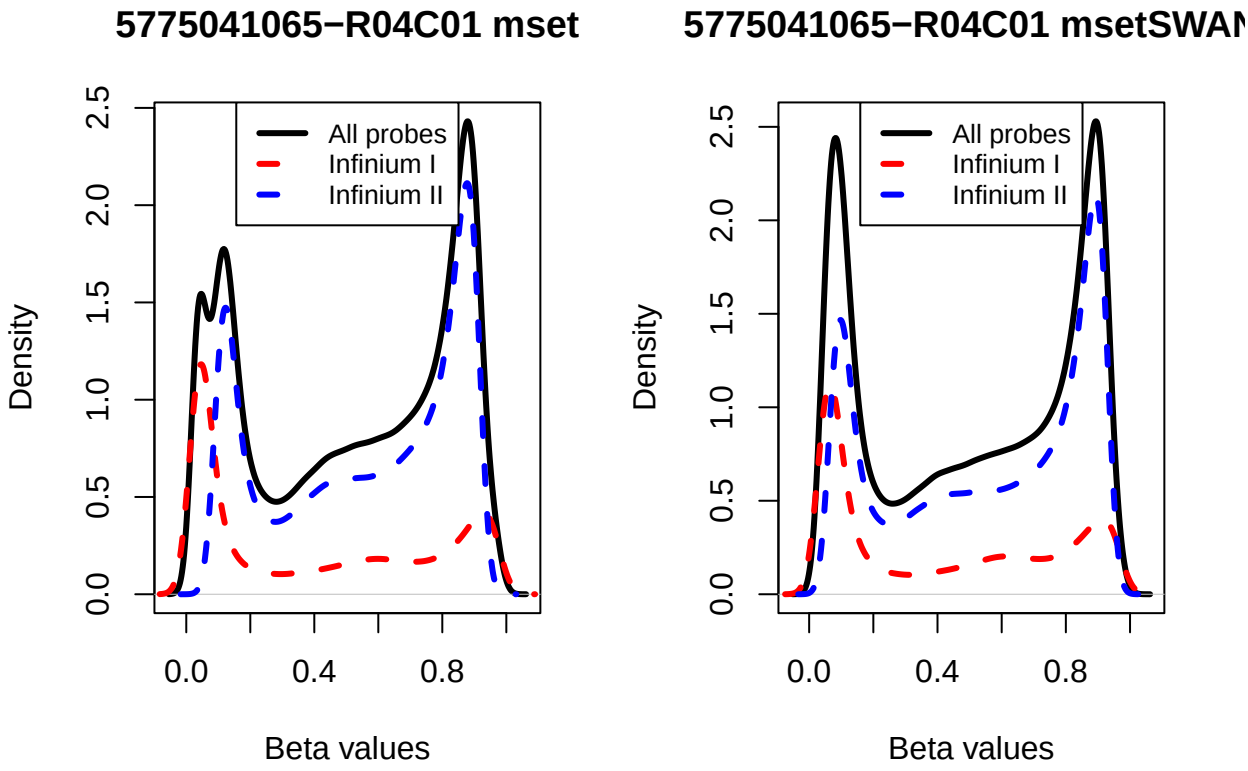


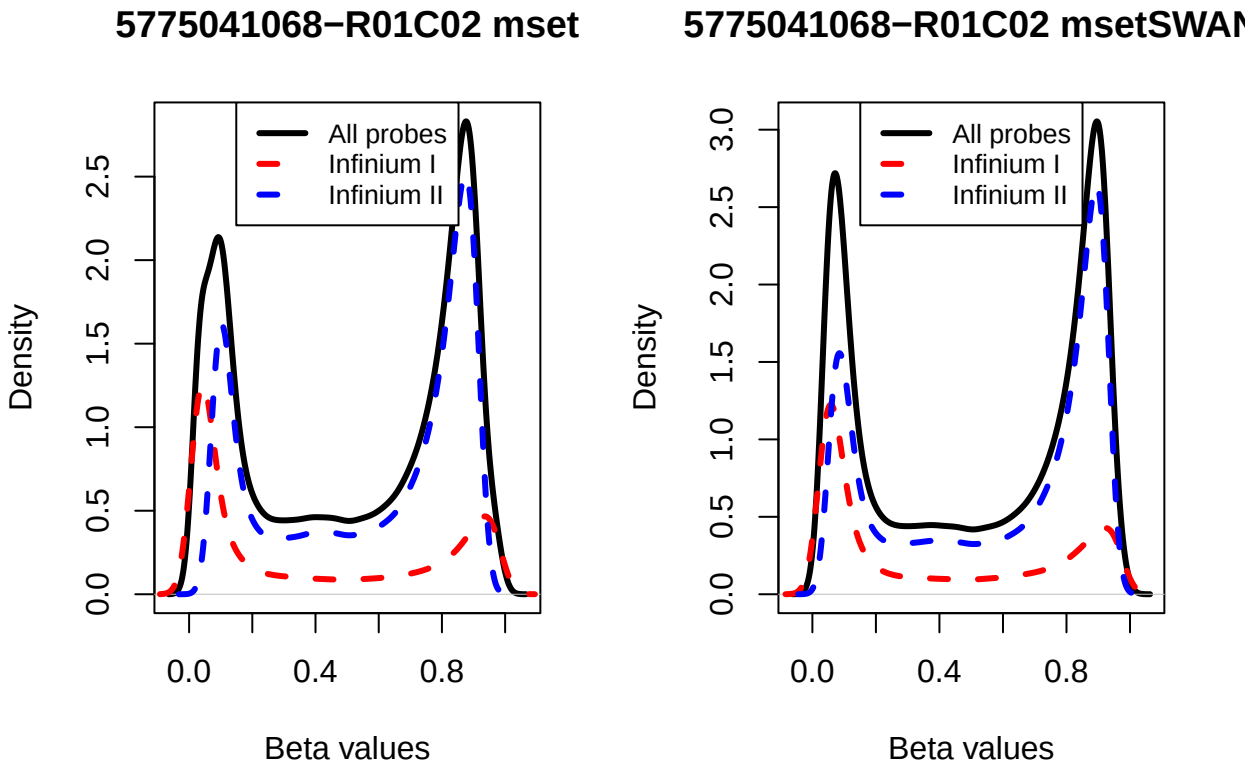


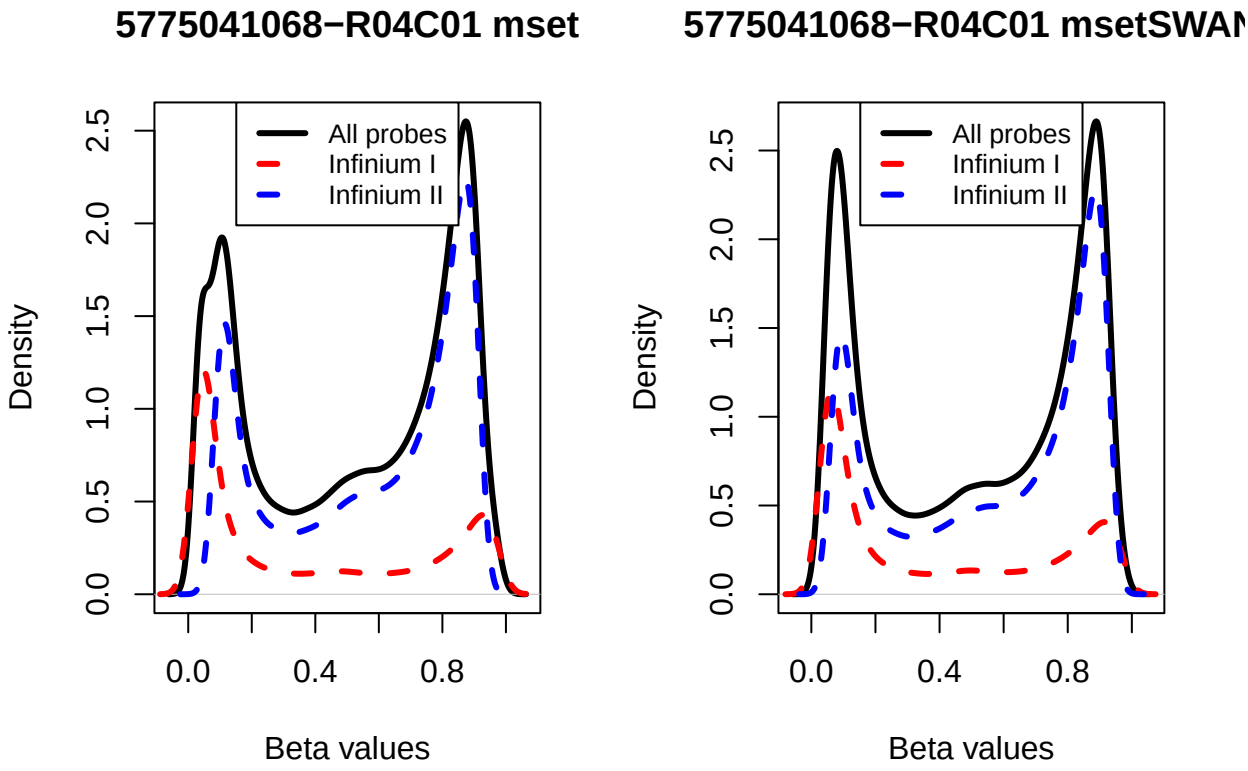
Part G

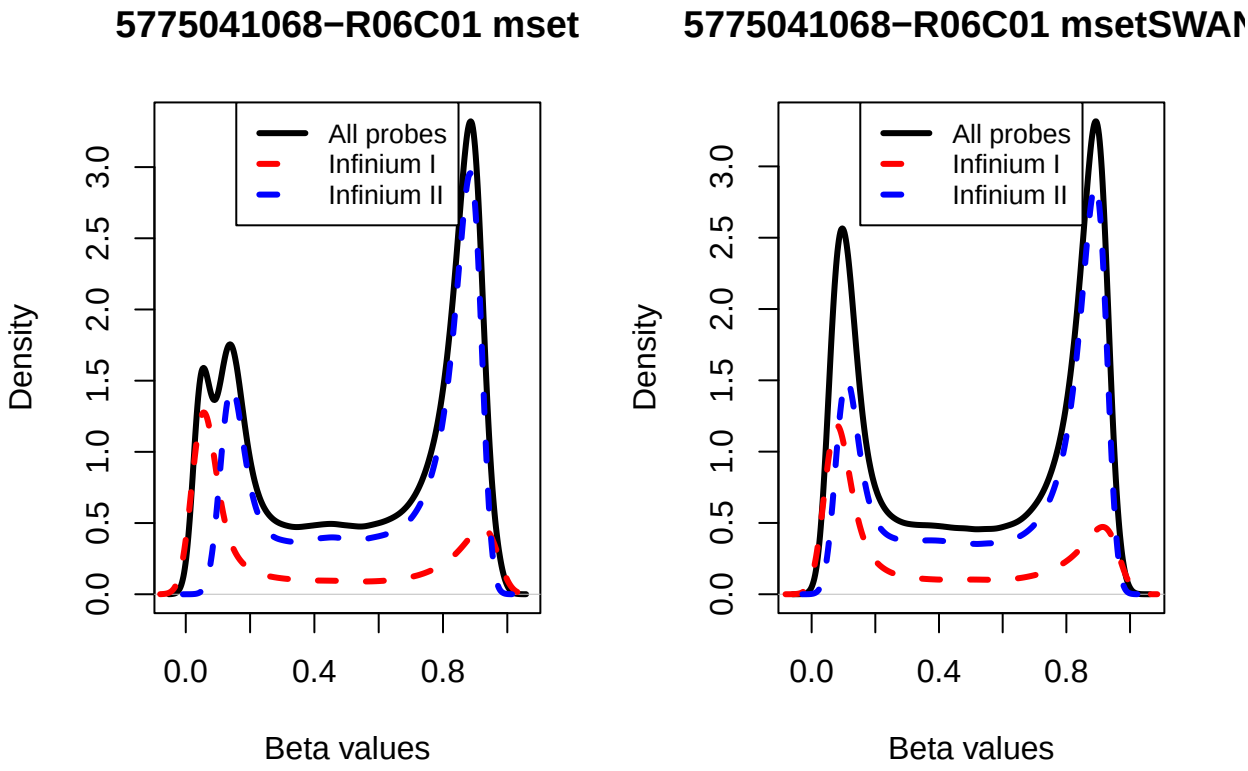
Plot the distribution of beta values before and after SWAN normalization using **plotBetasByType()**. What do you see in the density plots?

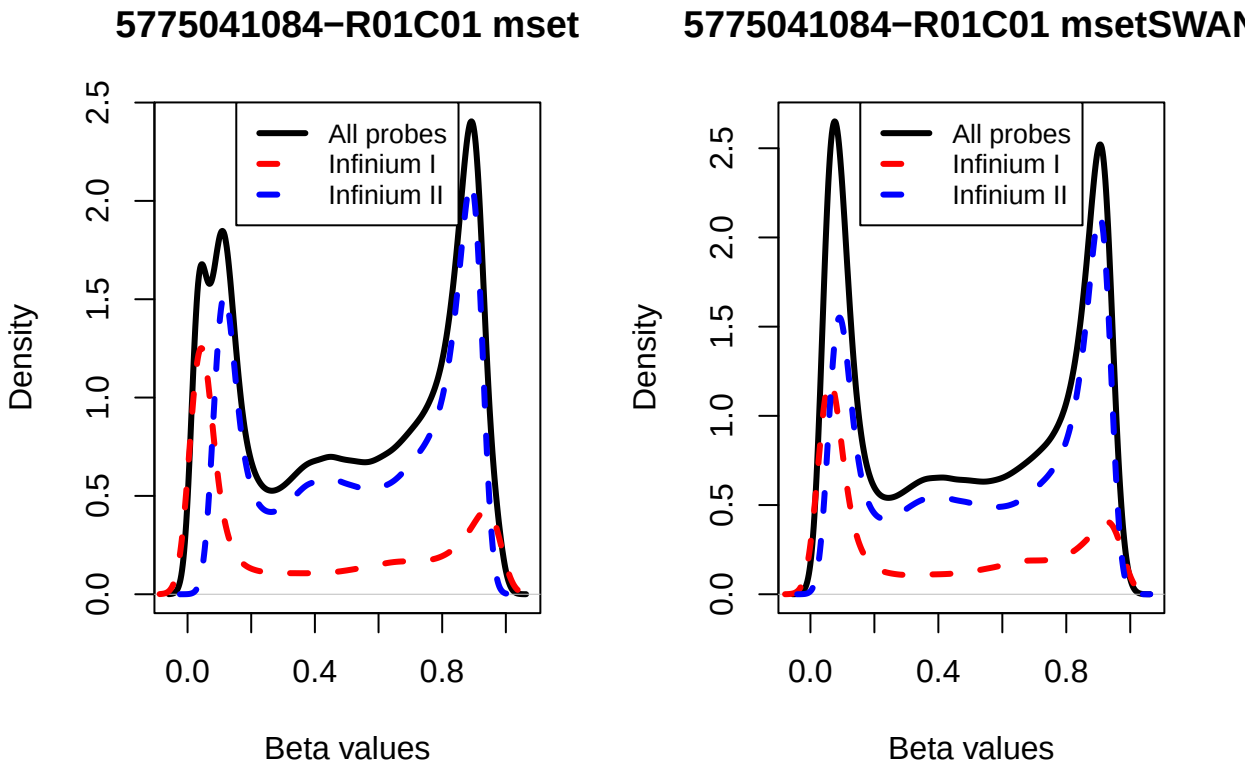

5775041065−R01C02 mset
All probes
Infinium I
Infinium II
Density
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.4
0.8
Beta values
5775041065−R01C02 msetSWAN
All probes
Infinium I
Infinium II
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.4
0.8
Beta values


5775041065−R04C01 mset
5775041065−R04C01 msetSWAN
All probes
Infinium I
Infinium II
Density
Beta values
0.0
0.5
1.0
1.5
2.0
2.5
0.4
0.8


5775041068−R01C02 mset
5775041068−R01C02 msetSWAN
All probes
Infinium I
Infinium II
Density
Beta values
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.4
0.8


5775041068−R04C01 mset
5775041068−R04C01 msetSWAN
All probes
Infinium I
Infinium II
Density
Beta values
0.0
0.5
1.0
1.5
2.0
2.5
0.0
0.4
0.8


5775041068−R06C01 mset
5775041068−R06C01 msetSWAN
All probes
Infinium I
Infinium II
Density
Beta values
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.4
0.8



Question 2: DNA Methylation Annotation and Differentially Methlyated Positions (Illumina 450K)

Part A

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

Table 3: Number of CpG Probes by Site

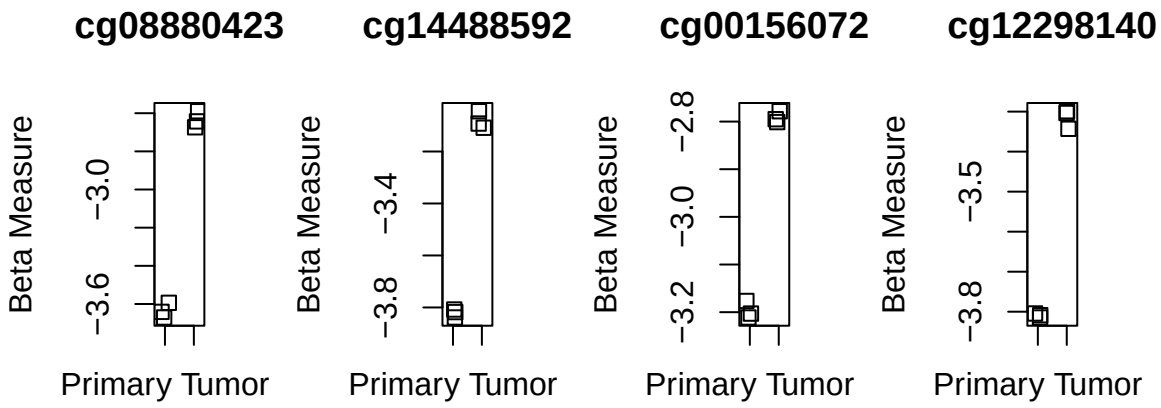
CpG Site	Number of Probes
CpG Islands	150254
CpG Upstream Shores	62870
CpG Downstream Shores	49197
CpG Upstream Shelves	24844
CpG Downstream Shelves	22300
CpG Open Sea	176047

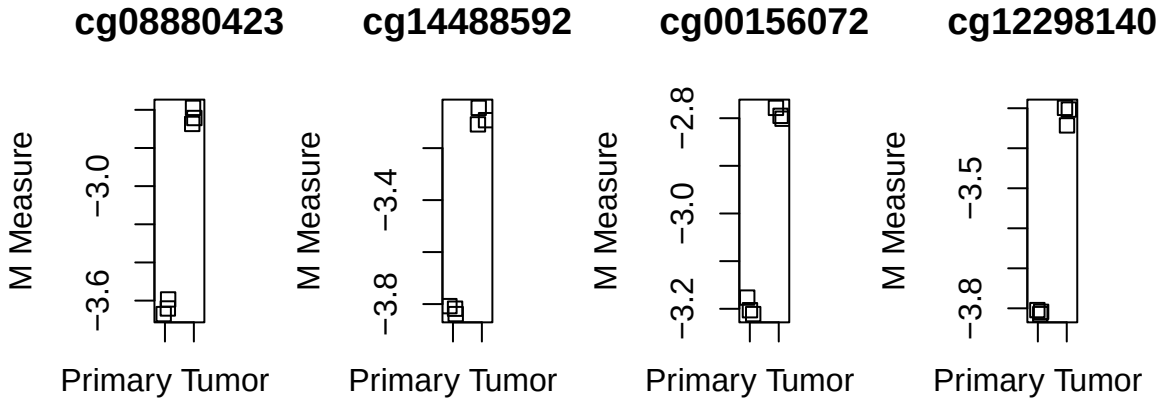
Part B

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?

There are 0 differentially methylated positions with q-values ≤ 0.10. There are 82 differentially methylated positions between cancer and non-cancer samples. Of these 82, 56 are hypomethylated and 26 are hypermethylated.

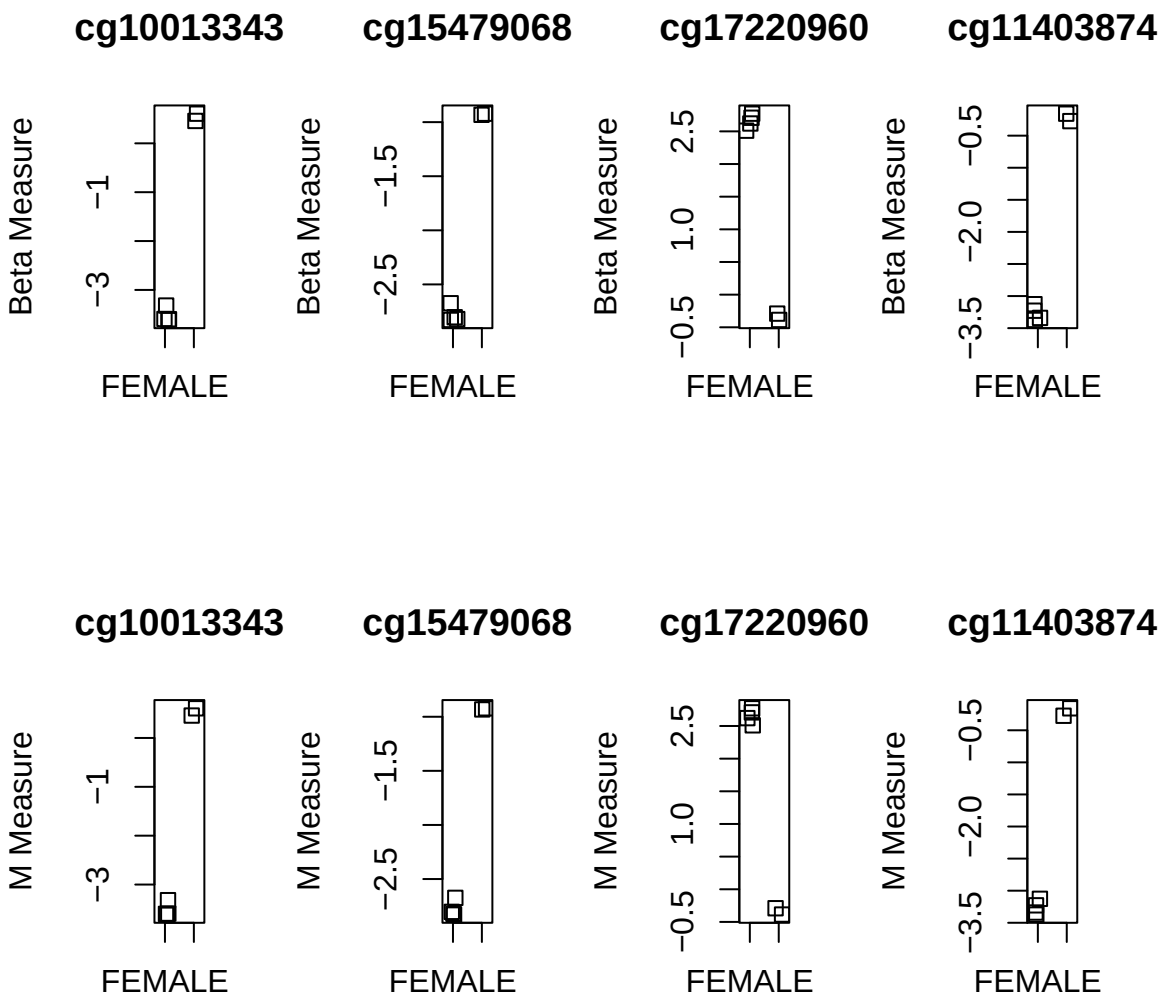




Part C

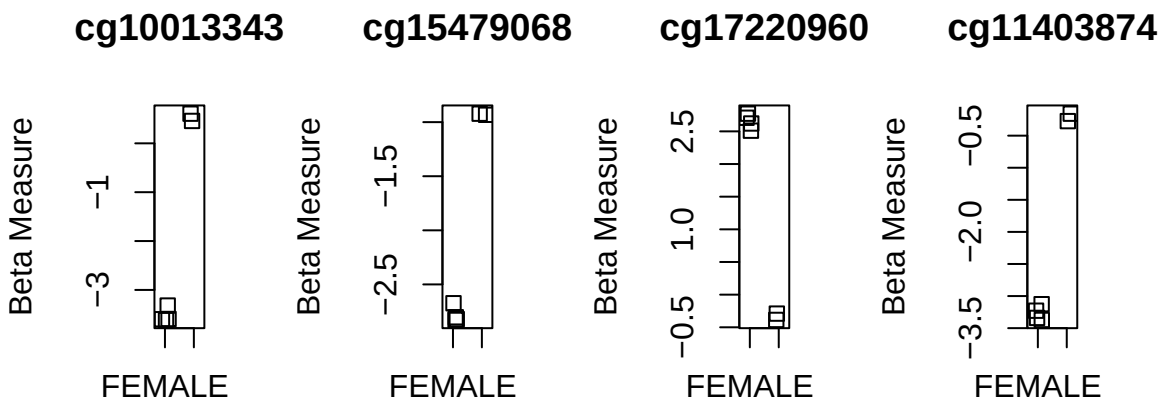
Repeat part b) but for DMPs between male and females.

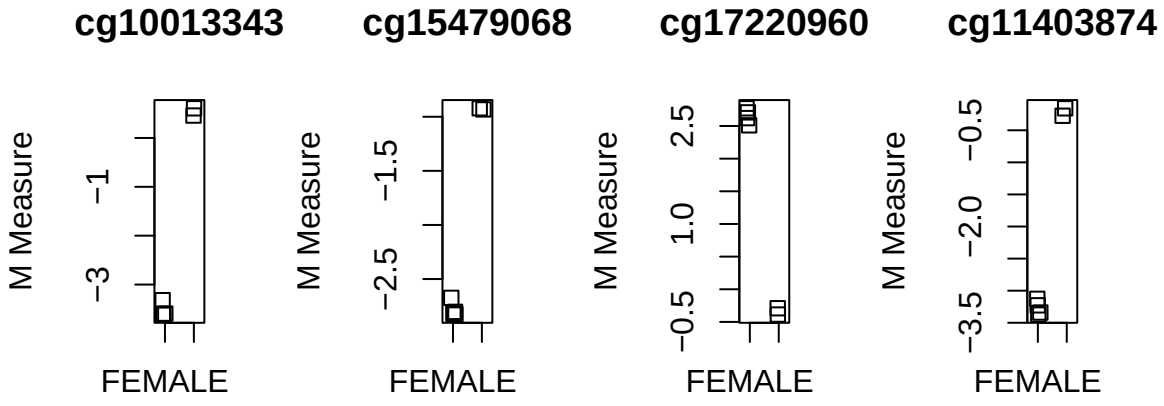
There are 0 differentially methylated positions with q-values ≤ 0.10. There are 147 differentially methylated positions between males and females. Of these 147, 85 are hypomethylated and 62 are hypermethylated.



Part D

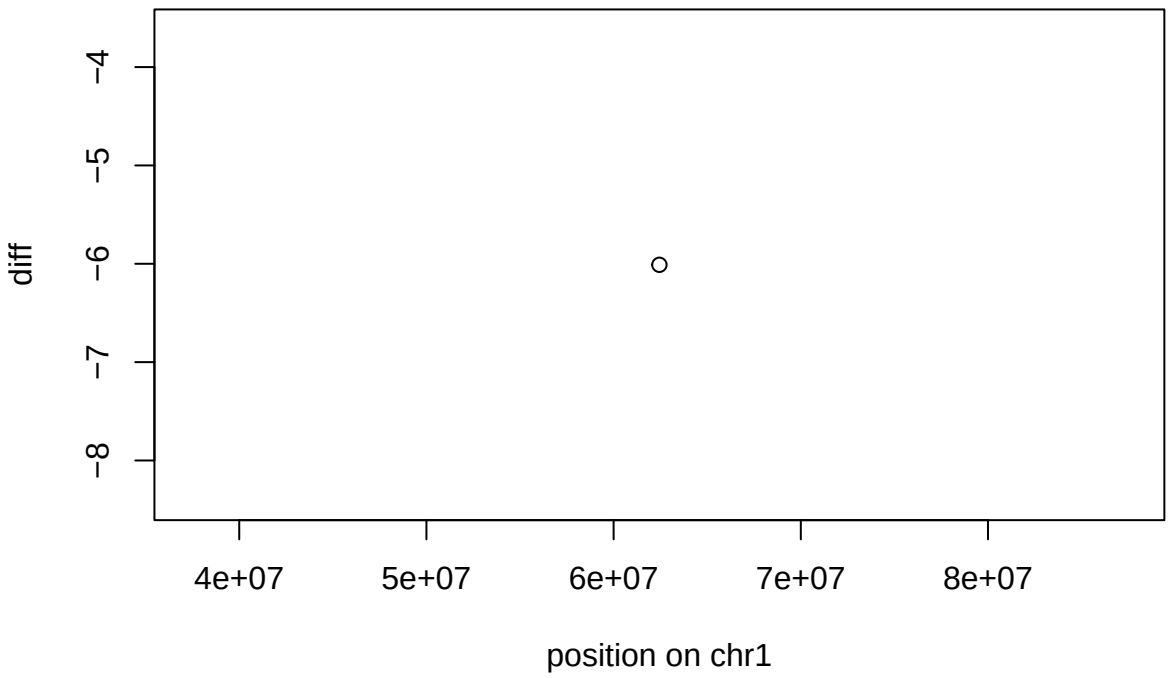
Global methylation profiles vary by sex. There is a function **addSex()** to estimate whether each samples is male or female. Are the predicted and given labels correct for Sex? If not, revisit the MDS plot from part 1e)? Do the new predictions group in the plot? Also repeat the analysis in 2c). Now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





Part E

This samples data is too small for **bumphunter** to identify significant regions by performing permutations or bootstrap. However, we can use the **getSegment()** function to find region of extreme values for the differences found in part b) and use the following code to plot one examples 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.


diff
−4
−5
−6
−7
−8
4e+07
5e+07
6e+07
7e+07
8e+07
position on chr1