BIOS 7659 Homework 7

Tim Vigers

15 November 2020

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

Load the data:

baseDir = "C:/Users/tim/Dropbox/Documents/School/Statistical Genomics/Homework Files/HW7/idats"  
targets = read.metharray.sheet(baseDir)

## [1] "C:/Users/tim/Dropbox/Documents/School/Statistical Genomics/Homework Files/HW7/idats/SampleSheet.csv"

rgSet = read.metharray.exp(targets = targets)  
annotation(rgSet)

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

## a) Table 1

df = as.data.frame(pData(rgSet))  
df = df[df$Status == "cancer",]  
t1 = tableby(~ patient.age\_at\_initial\_pathologic\_diagnosis +  
 patient.height + patient.weight + Sex + patient.race,  
 data = df)  
summary(t1,labelTranslations =   
 list(patient.age\_at\_initial\_pathologic\_diagnosis =   
 "Age at Diagnosis",  
 patient.height = "Height",patient.weight = "Weight",  
 patient.race = "Race"))

|  |  |
| --- | --- |
|  | Overall (N=3) |
| **Age at Diagnosis** |  |
| Mean (SD) | 76.667 (7.234) |
| Range | 72.000 - 85.000 |
| **Height** |  |
| Mean (SD) | 166.300 (15.934) |
| Range | 151.000 - 182.800 |
| **Weight** |  |
| Mean (SD) | 61.900 (8.080) |
| Range | 52.600 - 67.200 |
| **Sex** |  |
| FEMALE | 2 (66.7%) |
| MALE | 1 (33.3%) |
| **Race** |  |
| BLACK OR AFRICAN AMERICAN | 1 (33.3%) |
| WHITE | 2 (66.7%) |

There are three unique subjects in this dataset, each with two samples (one primary tumor sample and one from normal solid tissue).

## b) Type I and II probes

getManifest(rgSet)

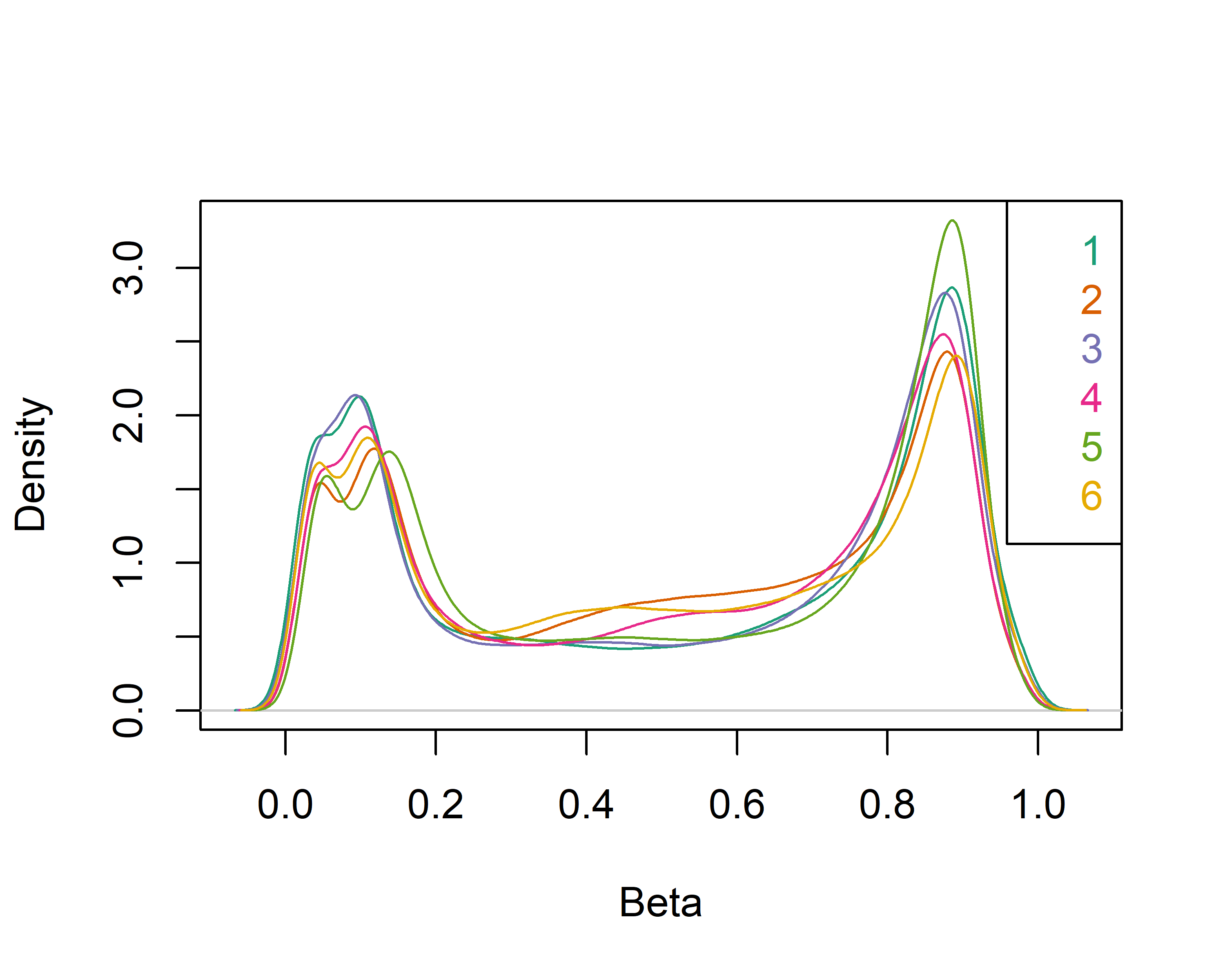
## IlluminaMethylationManifest object  
## Annotation  
## array: IlluminaHumanMethylation450k  
## Number of type I probes: 135476   
## Number of type II probes: 350036   
## Number of control probes: 850   
## Number of SNP type I probes: 25   
## Number of SNP type II probes: 40

There are 135,476 type I probes and 350,036 type II probes. Type I probes have two different sequences per CpG site, one for methylated and one for unmethylated CpGs. Type II probes use a two-color channel, which allows each probe to measure both methylated and unmethylated CpGs. As a result, type II probes take up half the physical space of type I probes. However, they have a lower dynamic range than type I probes, and are also more biased and less reproducible.

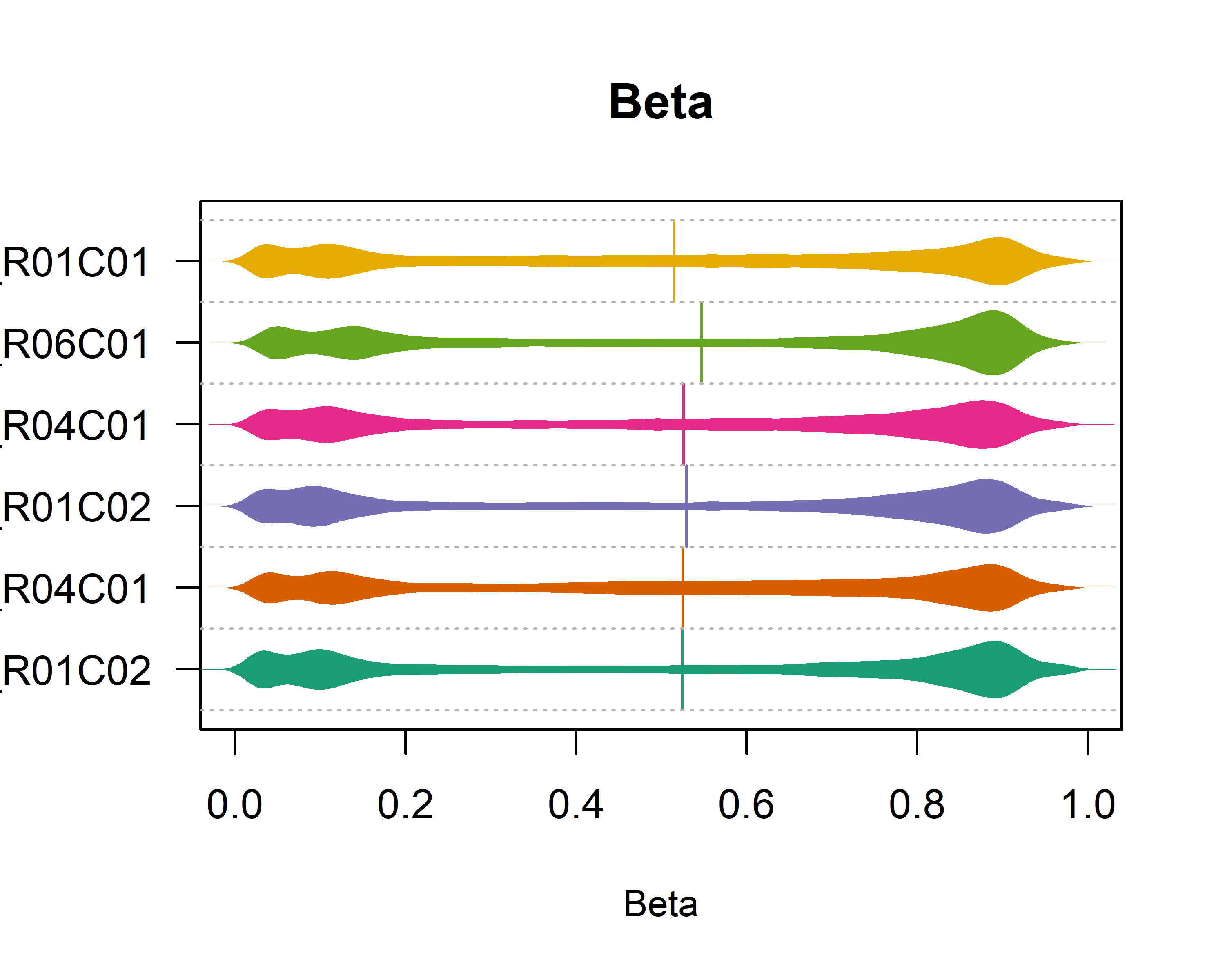
## c) QC Plots

### By ID

id = pData(rgSet)$id  
densityPlot(rgSet,sampGroups = id)

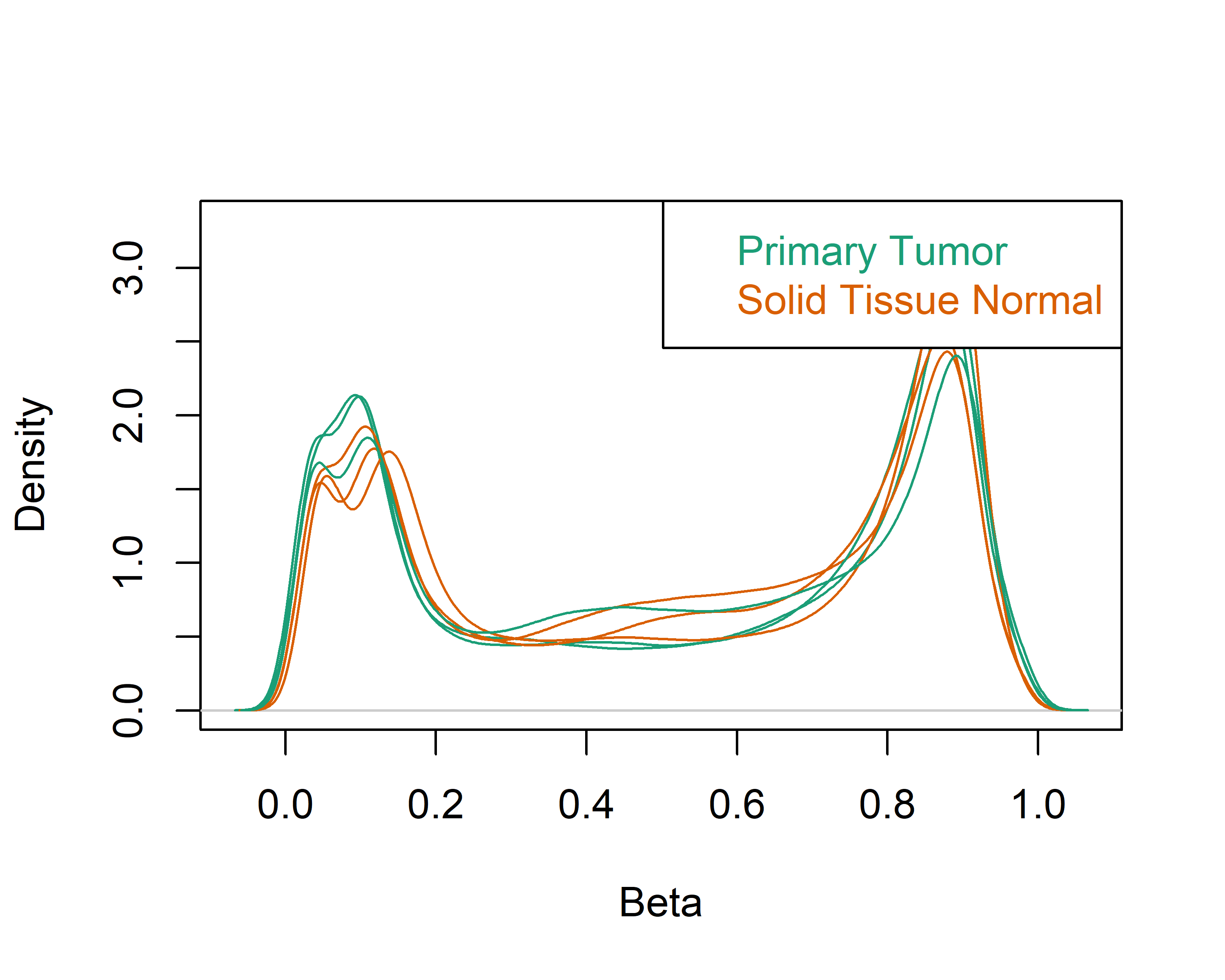


densityBeanPlot(rgSet,sampGroups = id)

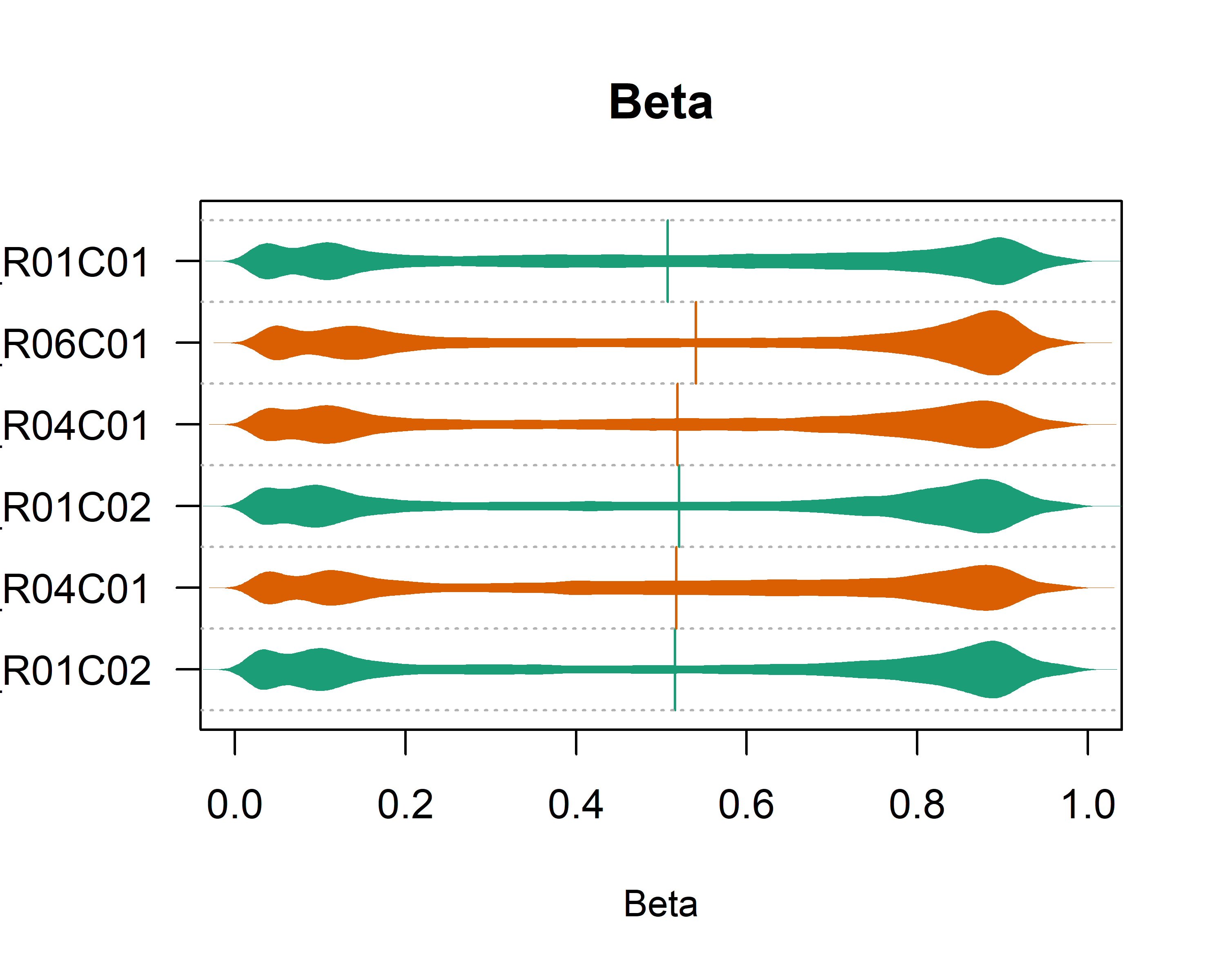


### By sample type

stype = pData(rgSet)$sample\_type  
densityPlot(rgSet,sampGroups = stype)

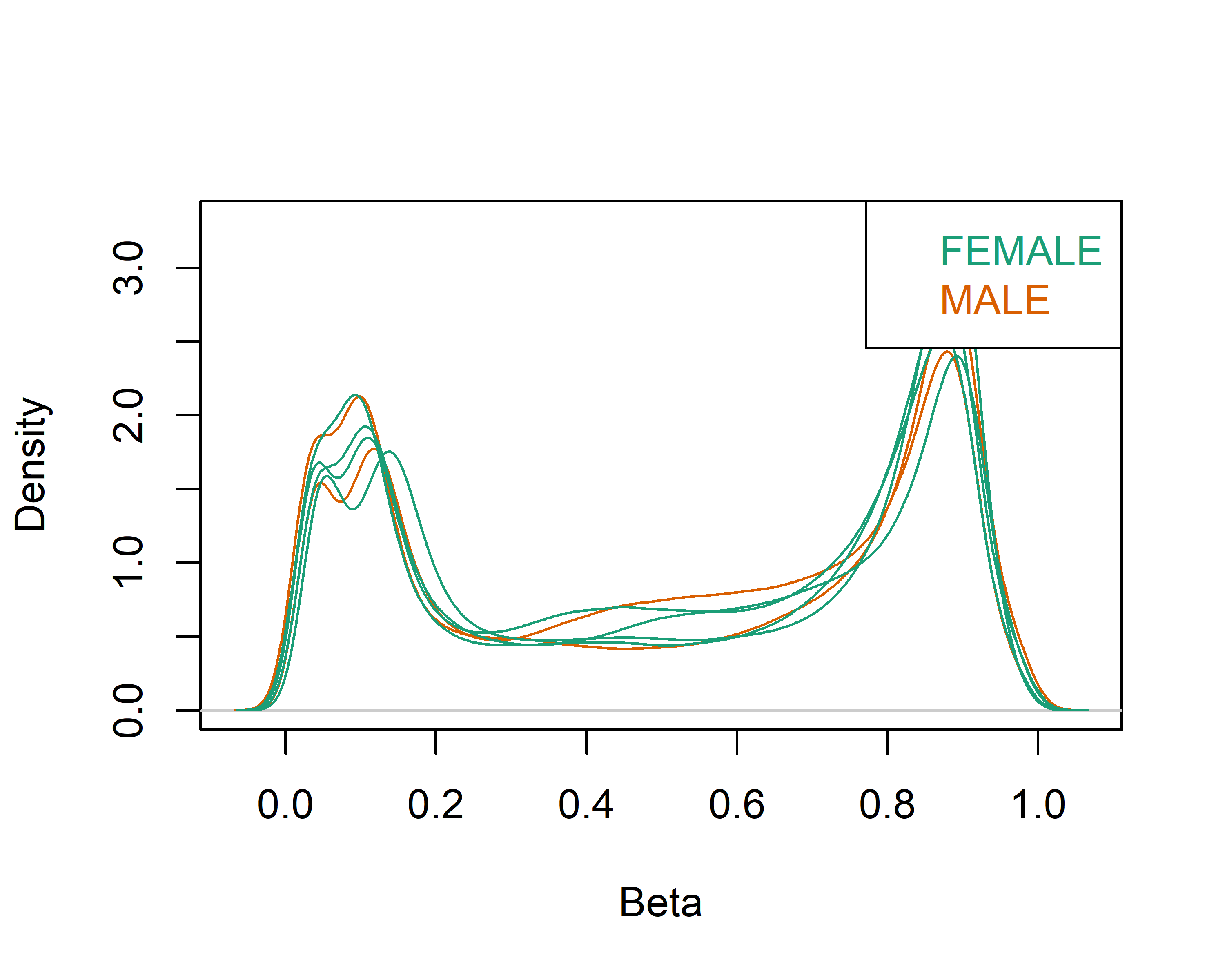


densityBeanPlot(rgSet,sampGroups = stype)

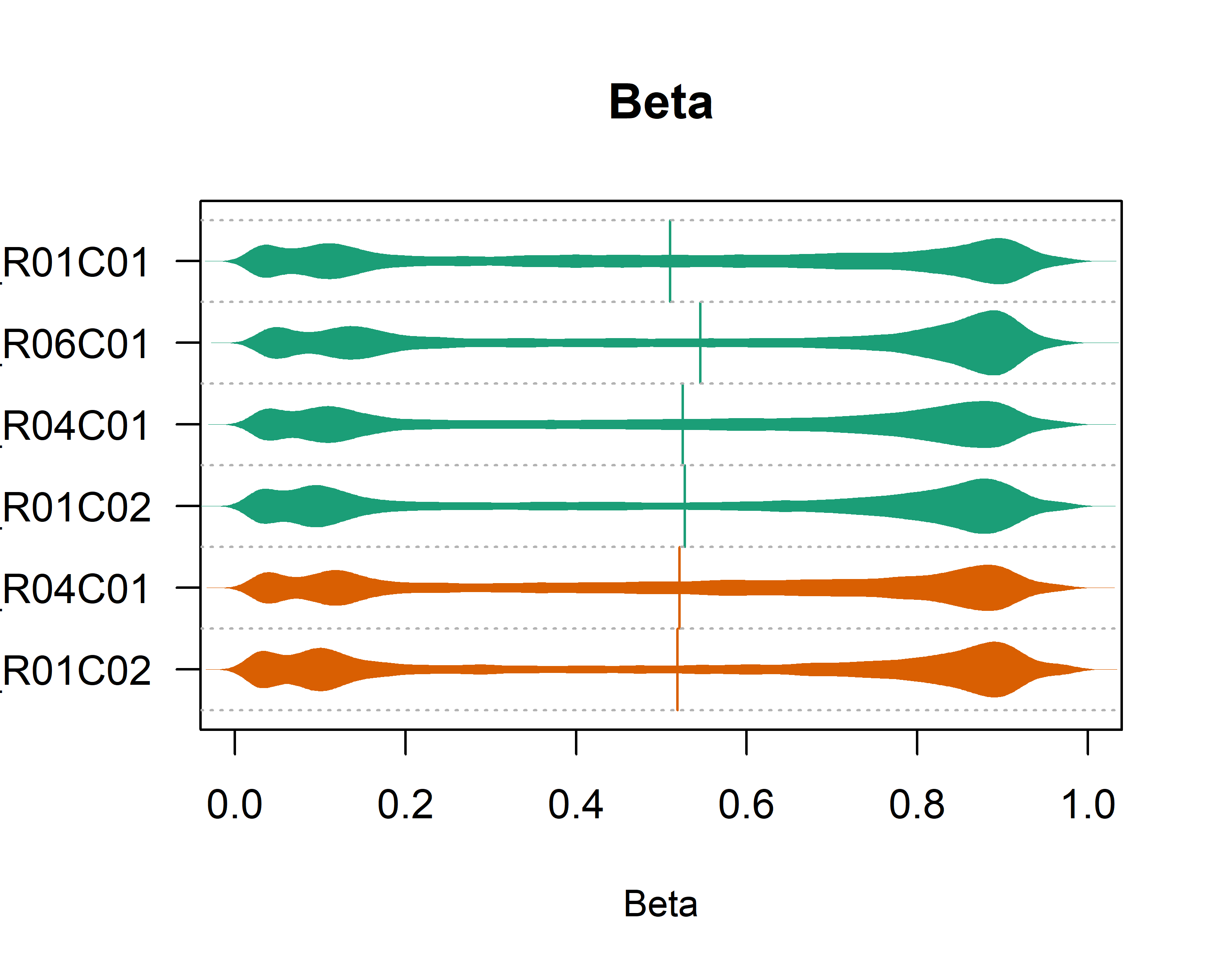


### By sex

sex = pData(rgSet)$Sex  
densityPlot(rgSet,sampGroups = sex)

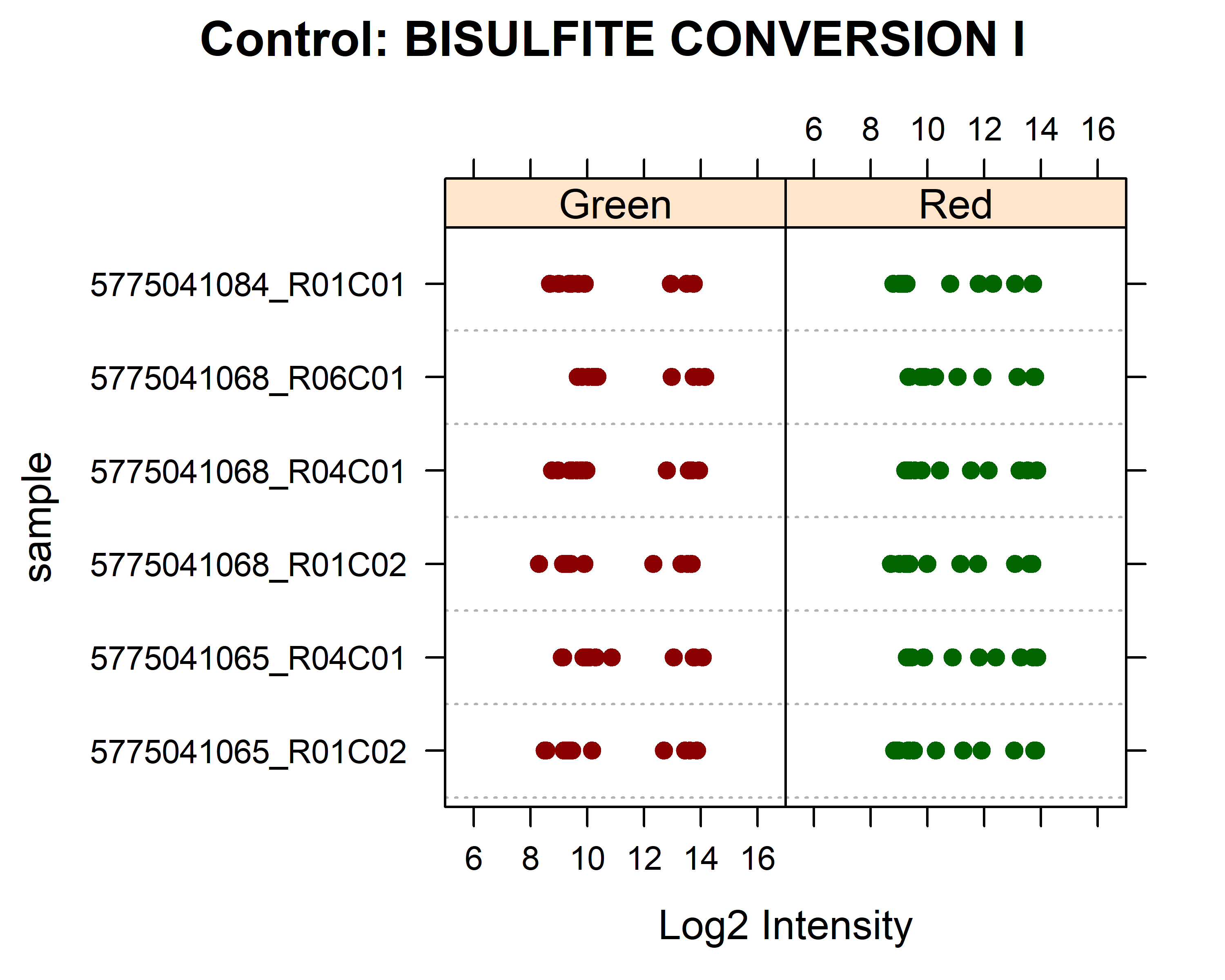


densityBeanPlot(rgSet,sampGroups = sex)

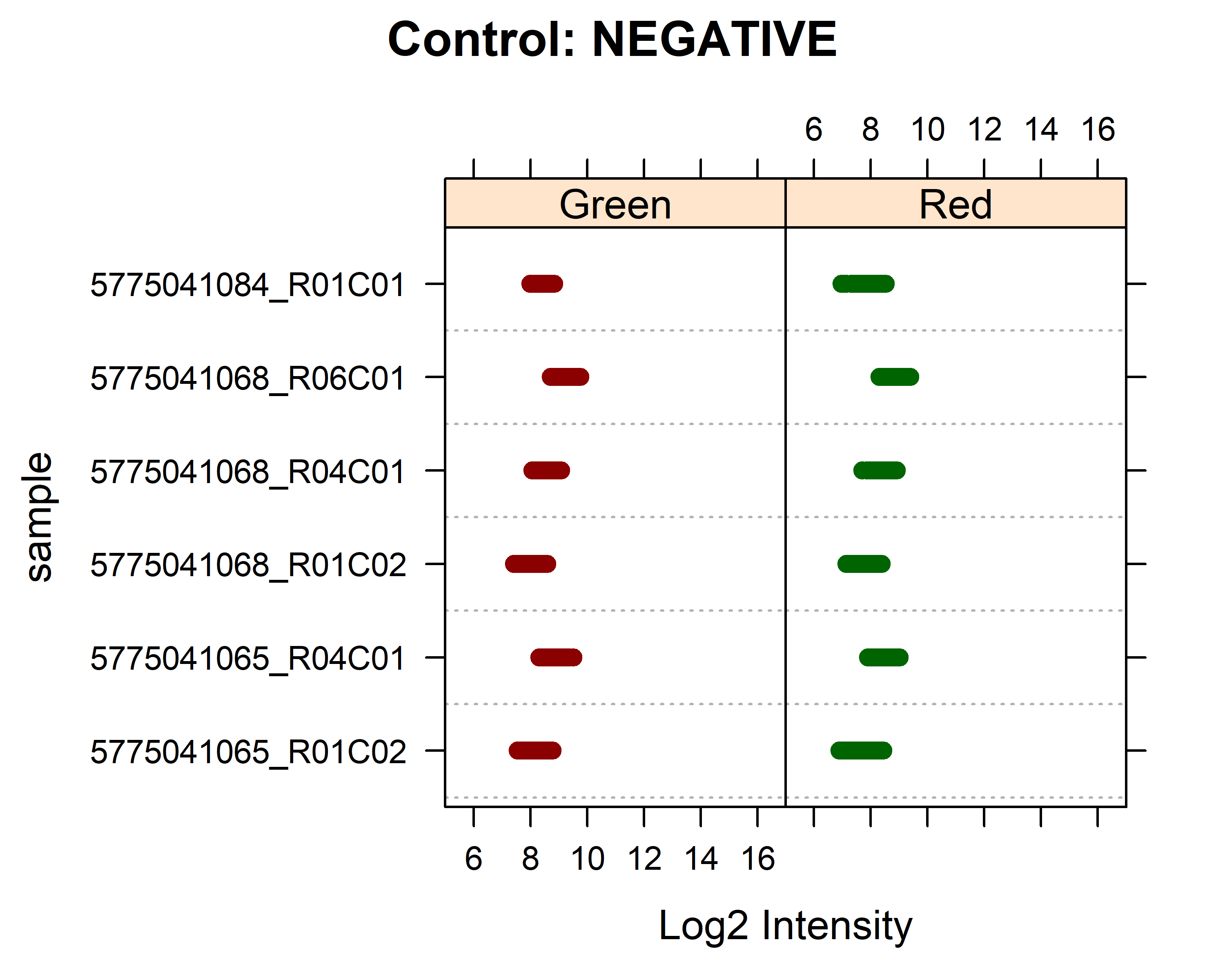


## d) Control probes

controlStripPlot(rgSet,controls = "BISULFITE CONVERSION I")



controlStripPlot(rgSet,controls = "NEGATIVE")



## e) Detection p values

# Count p values >= 0.05 per sample  
detect = detectionP(rgSet)  
colSums(detect >= 0.05)

## 5775041065\_R01C02 5775041065\_R04C01 5775041068\_R01C02 5775041068\_R04C01   
## 133 567 102 493   
## 5775041068\_R06C01 5775041084\_R01C01   
## 681 554

# Row means  
rmeans = rowMeans(detect)

Sample 5775041068\_R06C01 has the most detection p values with 681 (0.14%). Out of the 485512 probes, 853 have a mean p value .

### Save the methylation signals

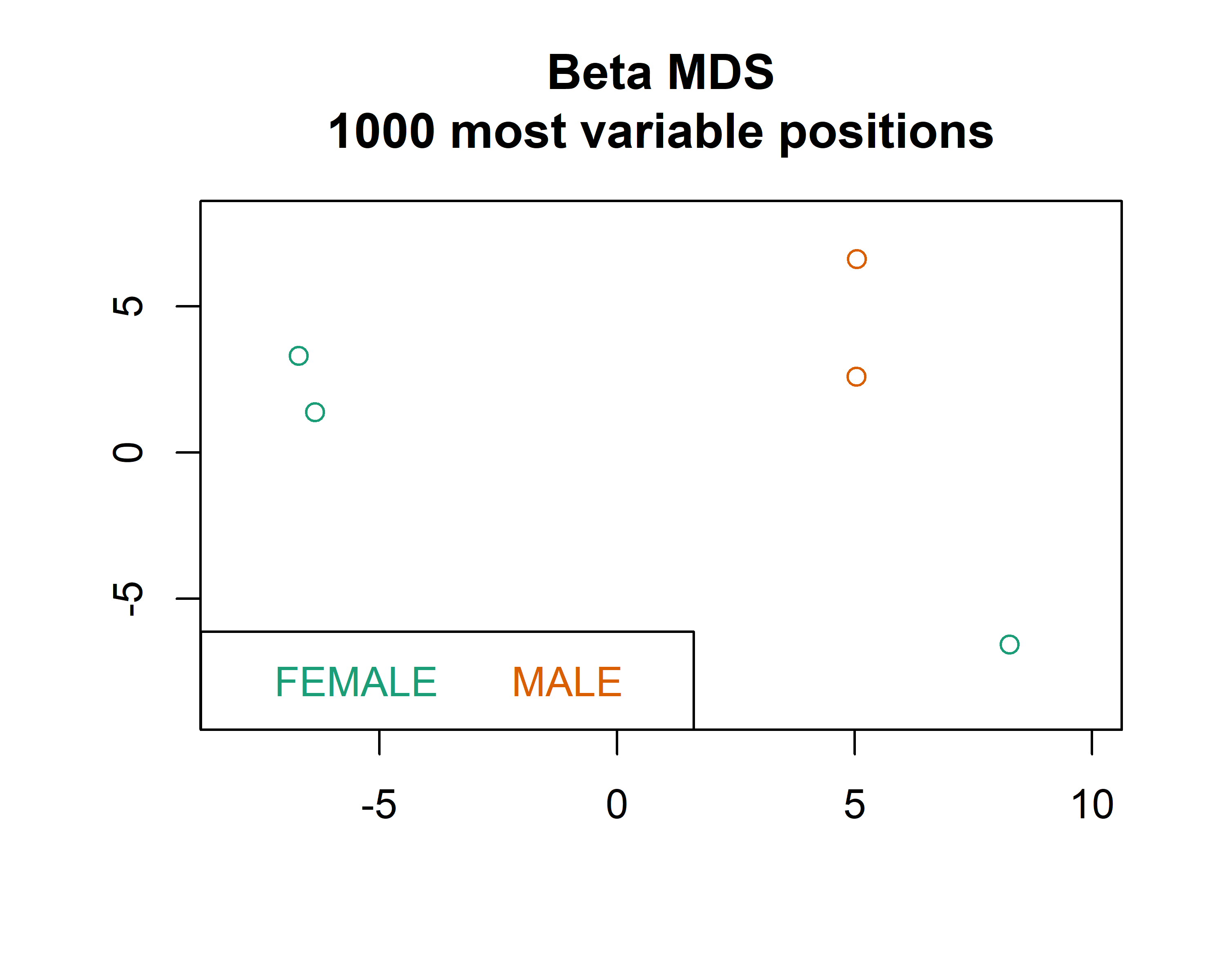
mset = preprocessRaw(rgSet)  
msetSWAN = preprocessSWAN(rgSet)

## f) Multidimensional scaling (MDS) plots

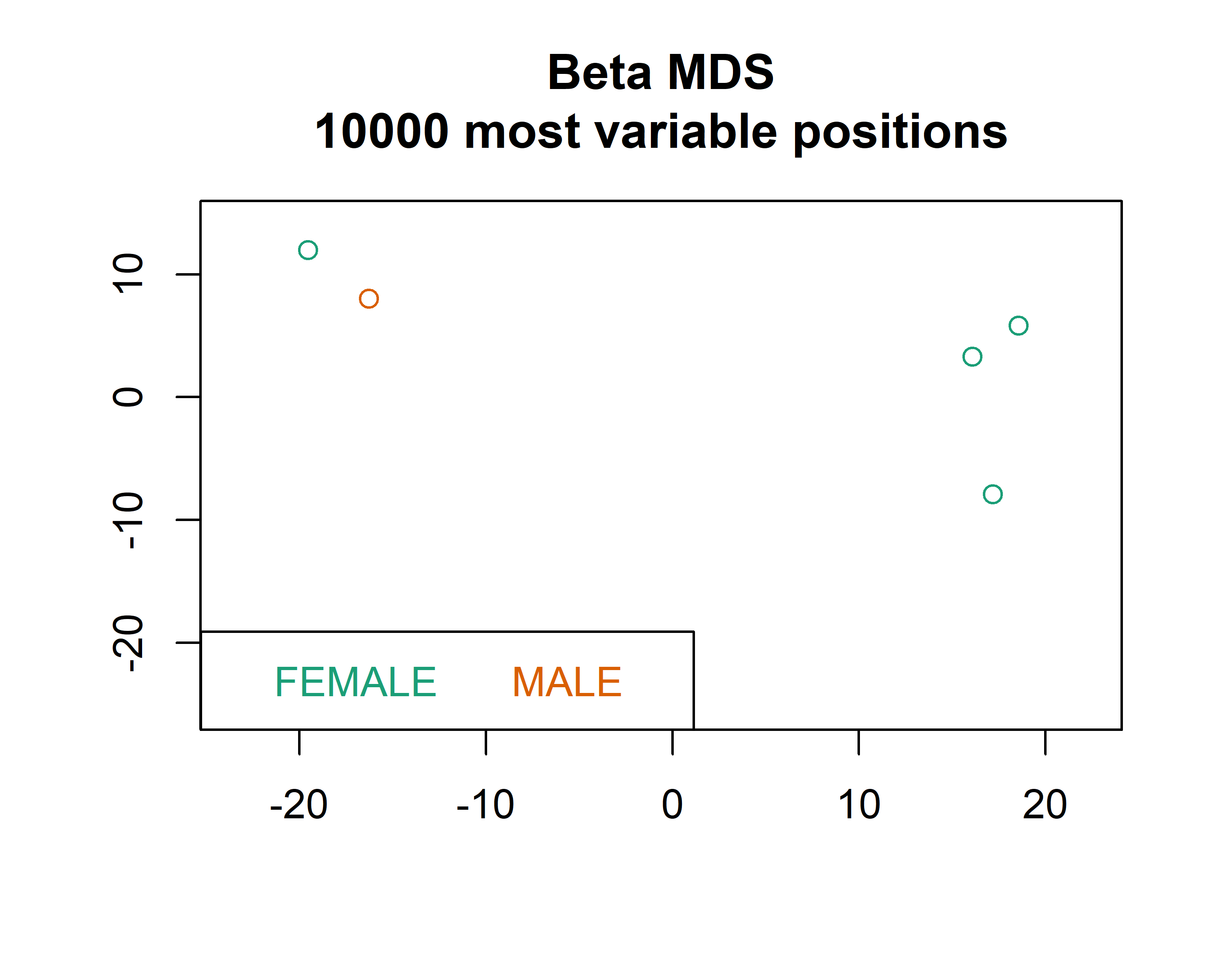
### By sex

#### SWAN-normalized

mdsPlot(msetSWAN,sampGroups = sex)

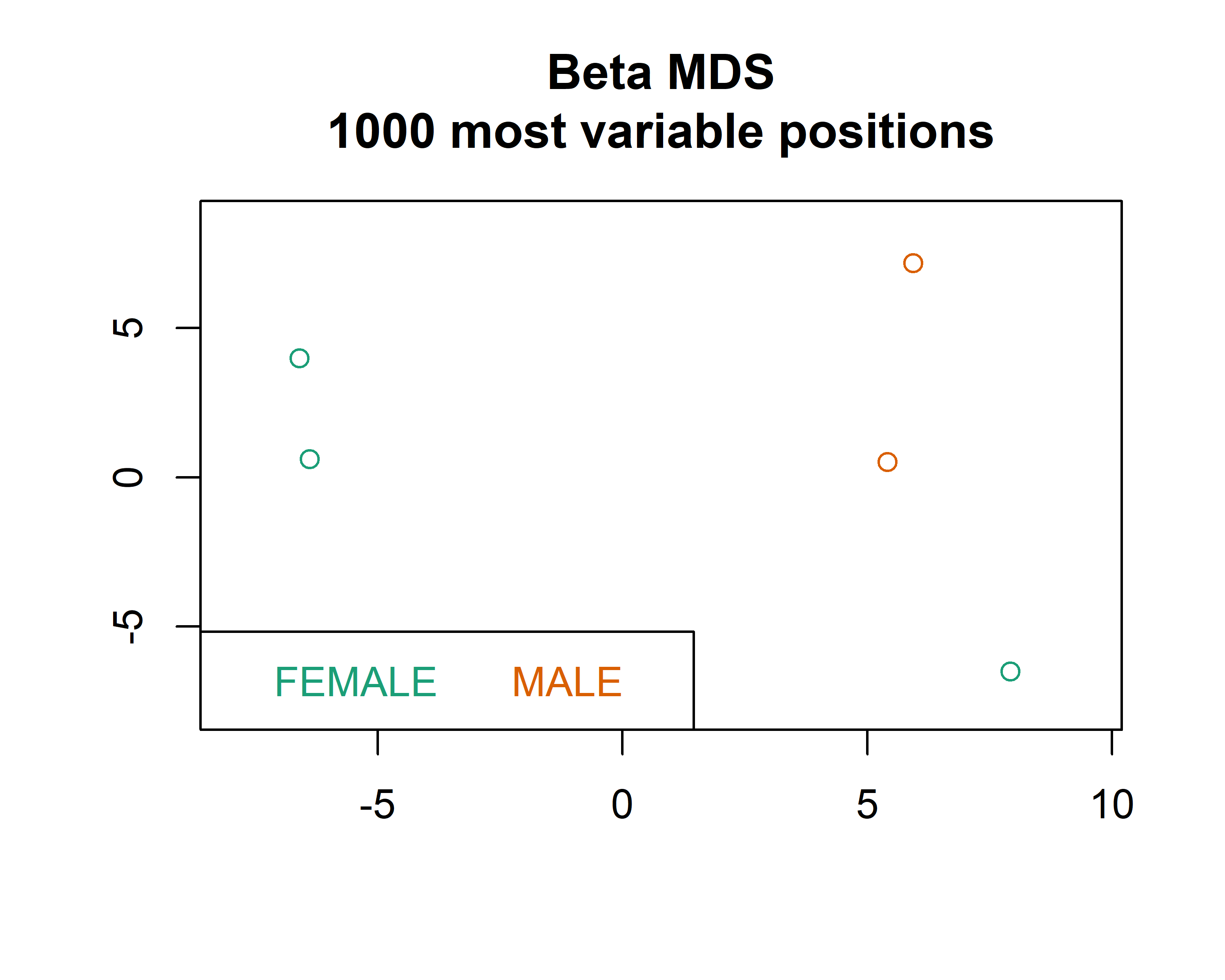


mdsPlot(msetSWAN,numPositions = 10000,sampGroups = sex)

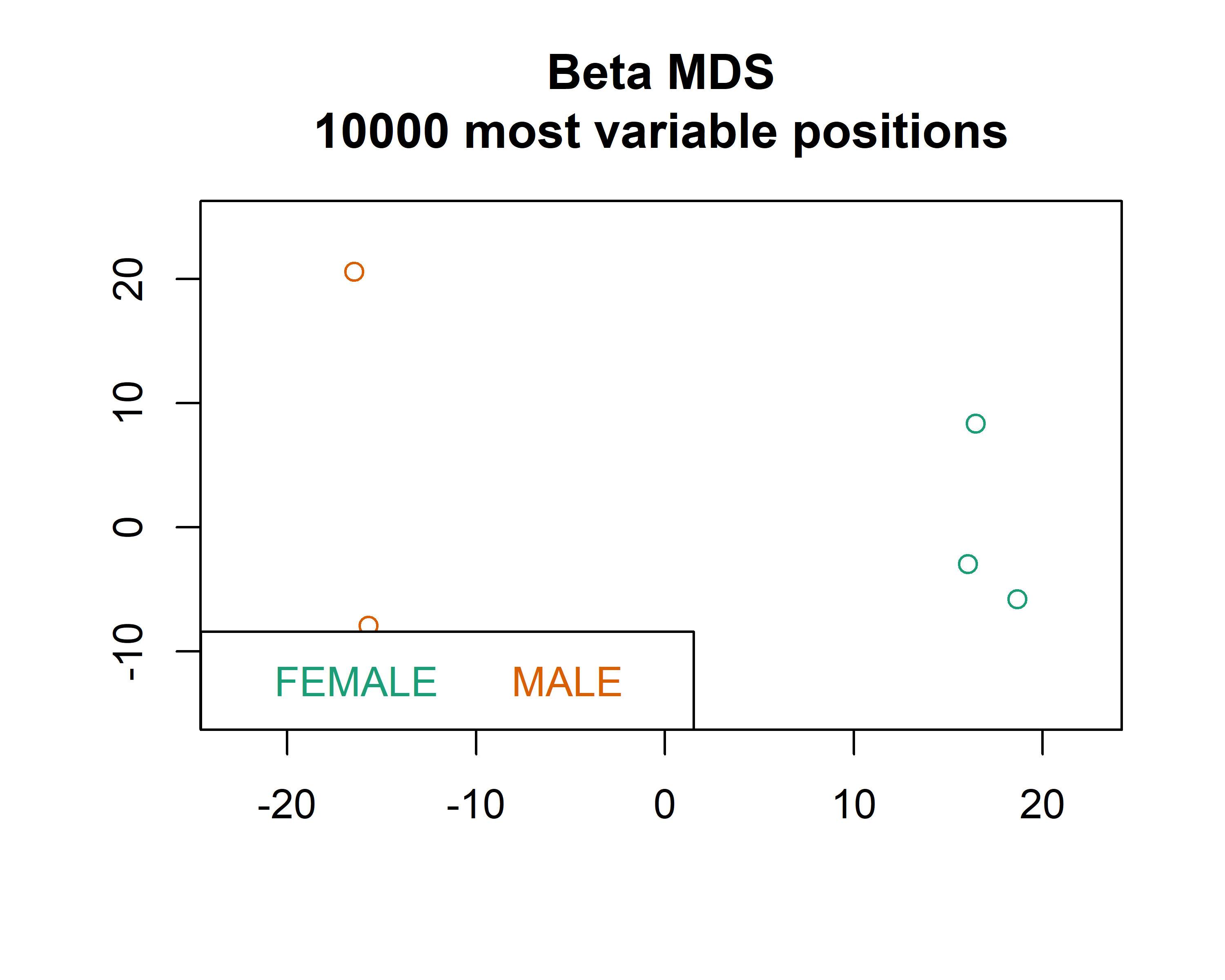


#### Raw data

mdsPlot(mset,sampGroups = sex)



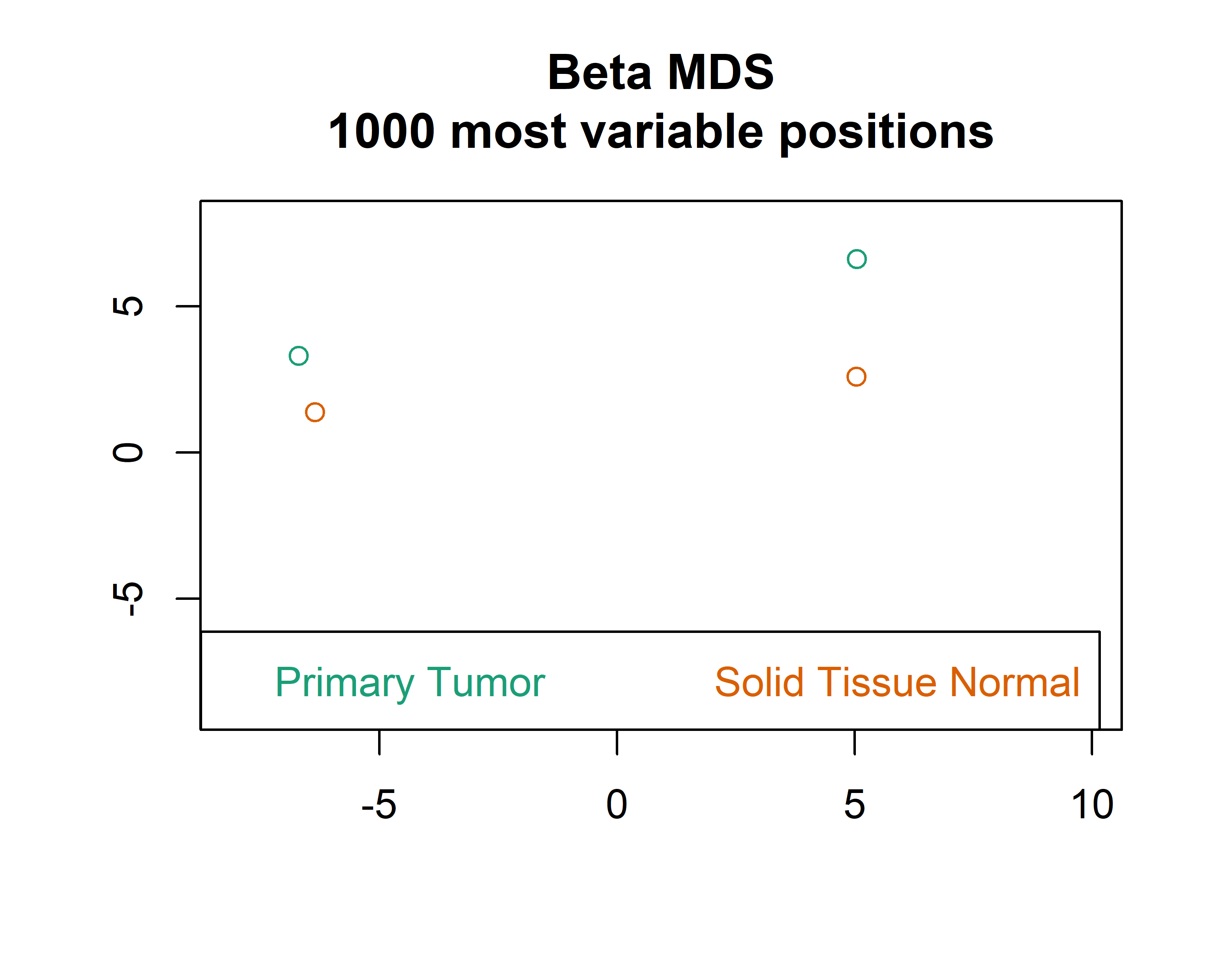
mdsPlot(mset,numPositions = 10000,sampGroups = sex)



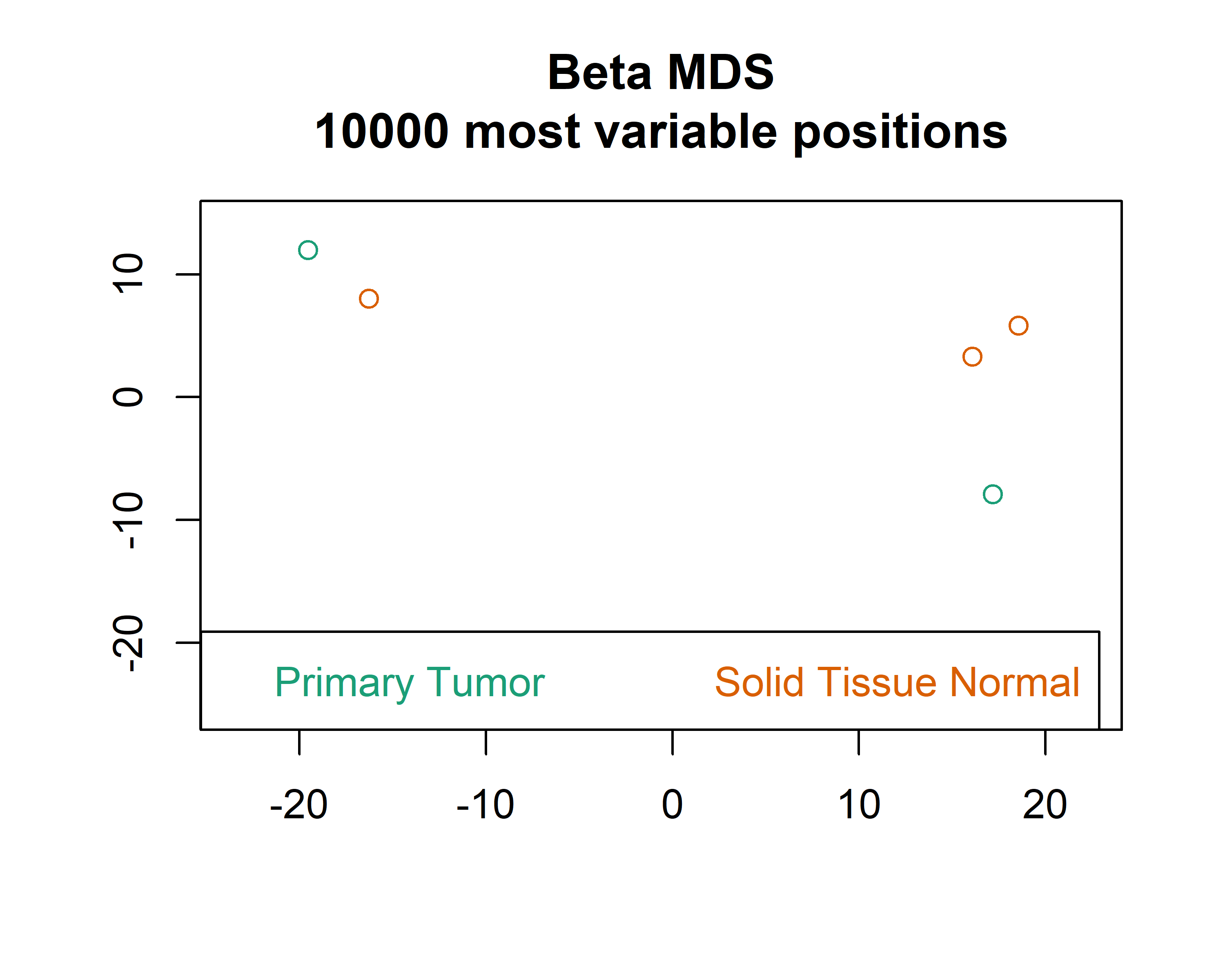
### By cancer status

#### SWAN-normalized

mdsPlot(msetSWAN,sampGroups = stype)

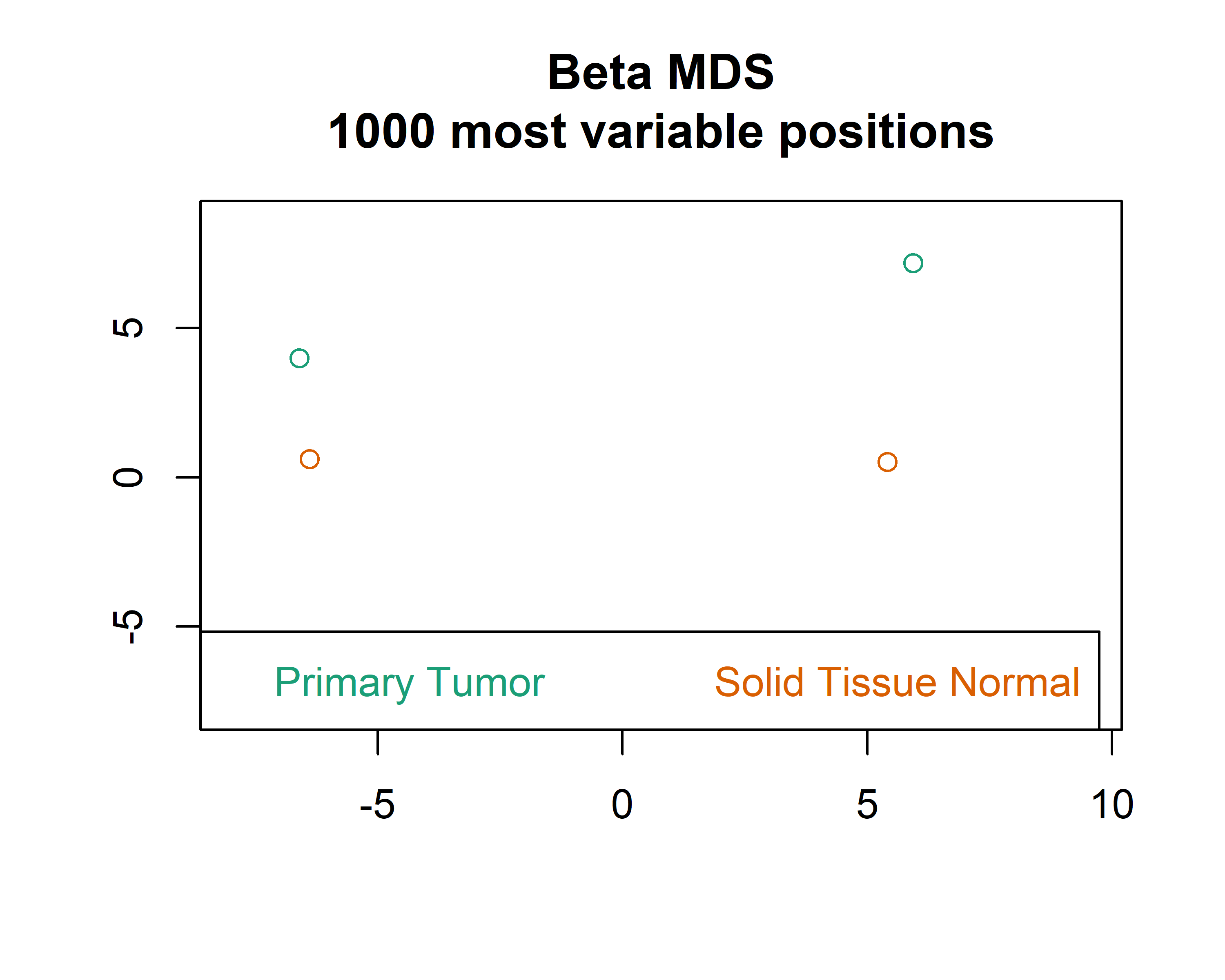


mdsPlot(msetSWAN,numPositions = 10000,sampGroups = stype)

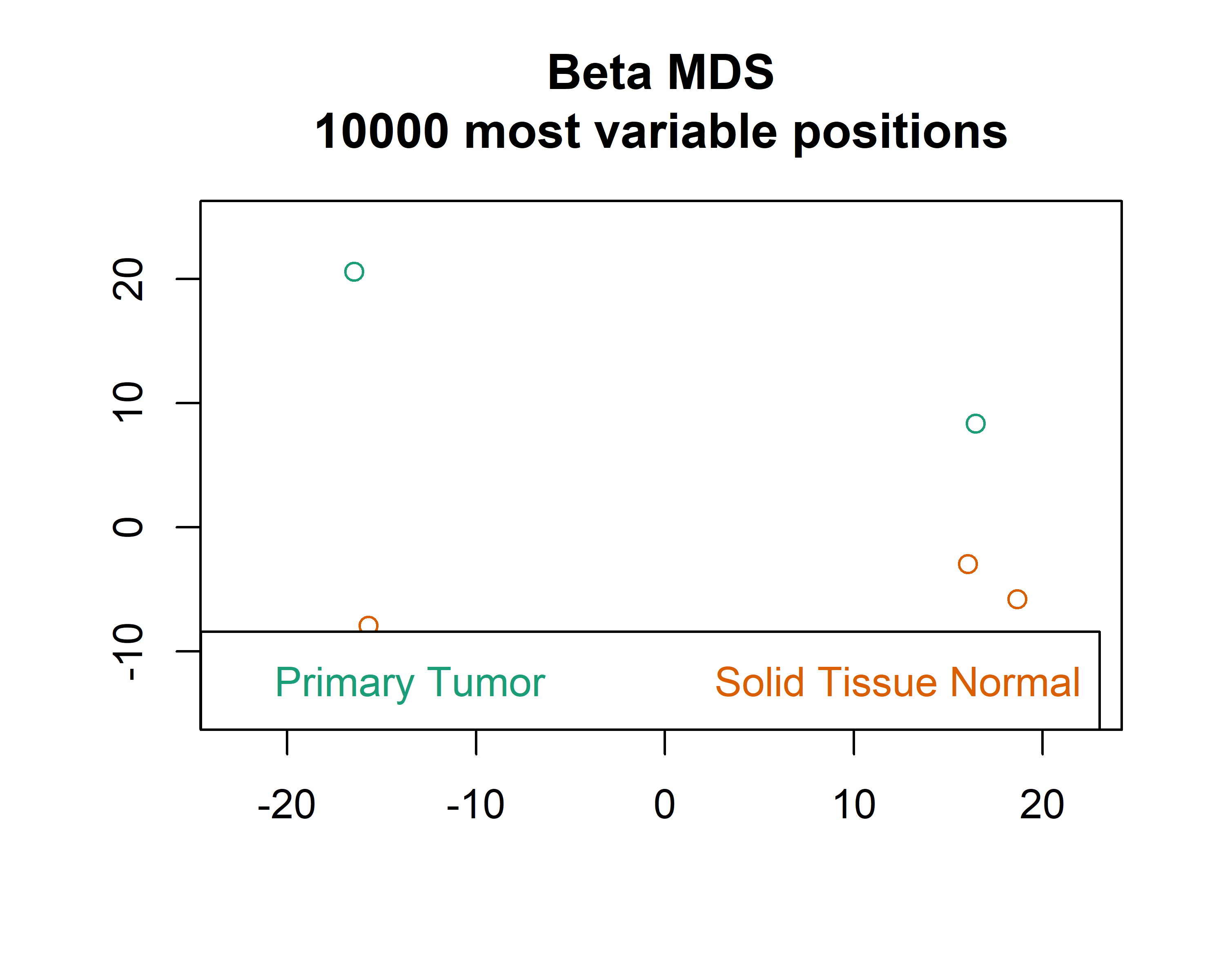


#### Raw data

mdsPlot(mset,sampGroups = stype)

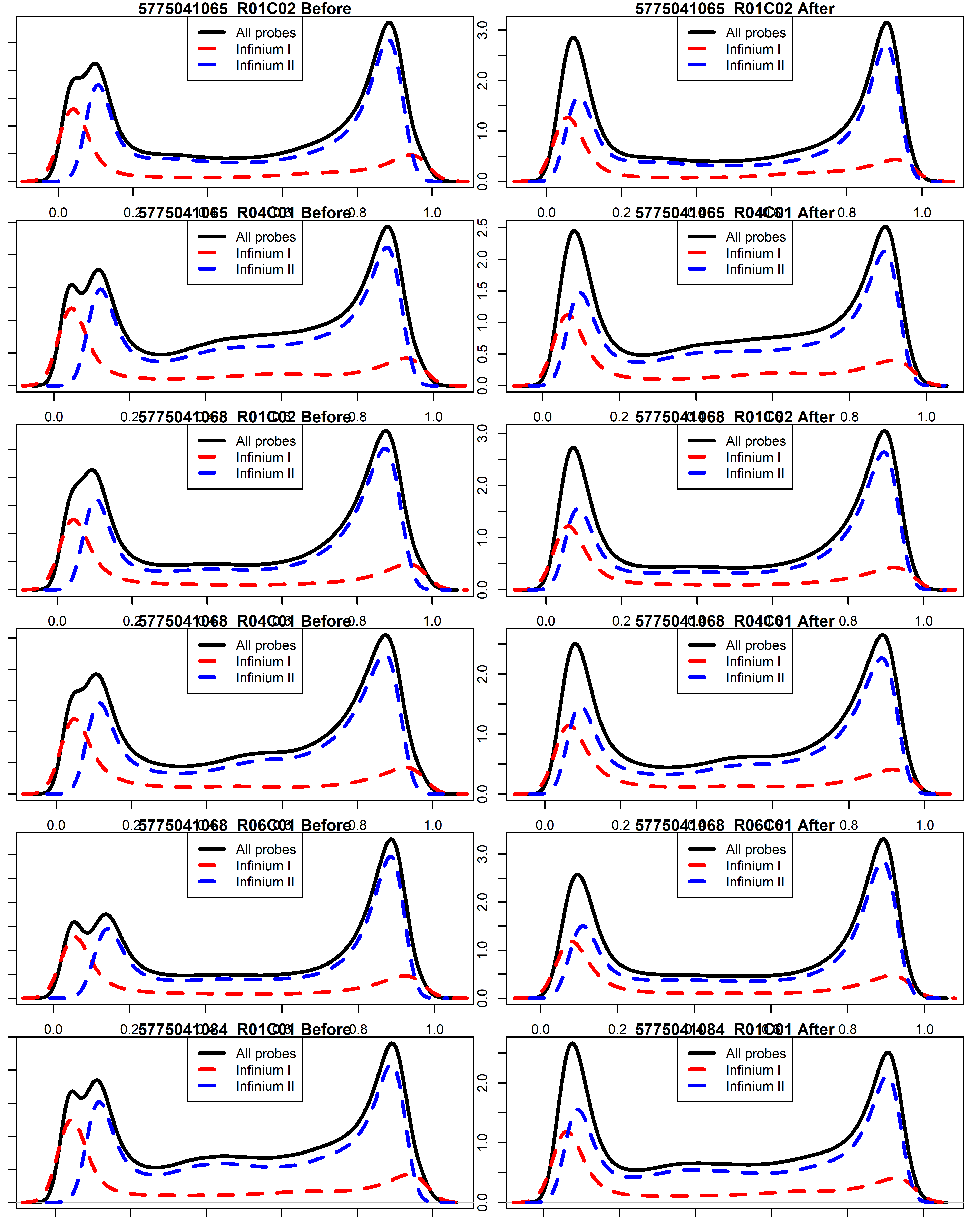


mdsPlot(mset,numPositions = 10000,sampGroups = stype)



## g) Distribution of beta values before and after SWAN normalization

par(mfrow=c(6,2),mar=c(1,1,1,1))  
for (i in 1:6) {  
 plotBetasByType(mset[,i],main = paste(colnames(mset[,i]),"Before"))  
 plotBetasByType(msetSWAN[,i],  
 main = paste(colnames(msetSWAN[,i]),"After"))  
}



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

Get genome annotation information:

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

## a) CpG islands, shores, shelves and open seas

table(annotation$Relation\_to\_Island) %>% as.data.frame(.) %>%  
 rename(.,Feature = Var1) %>% flextable(.)

| Feature | Freq |
| --- | --- |
| Island | 150254 |
| N\_Shelf | 24844 |
| N\_Shore | 62870 |
| OpenSea | 176047 |
| S\_Shelf | 22300 |
| S\_Shore | 49197 |

## b) Find DMP for cancer status

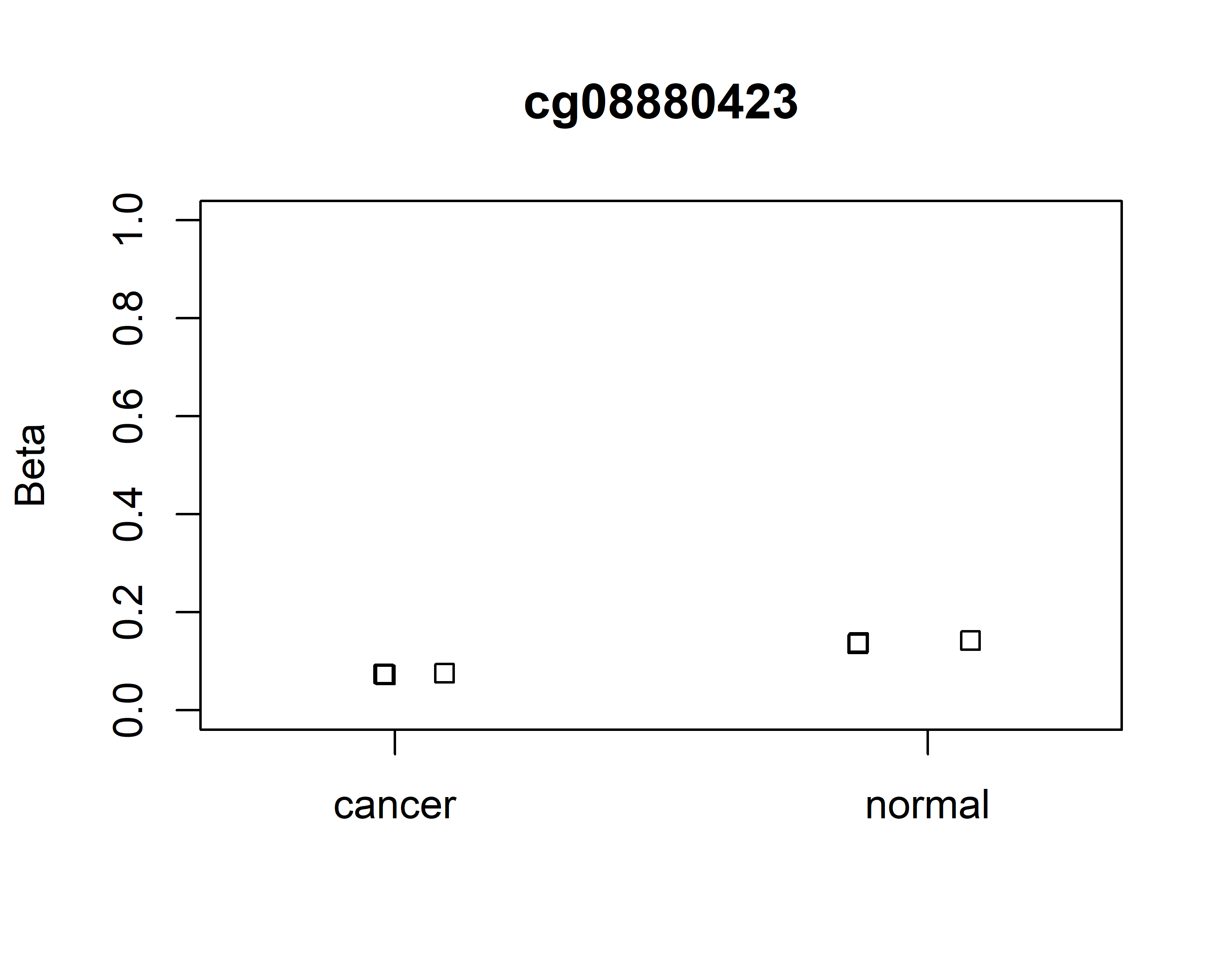
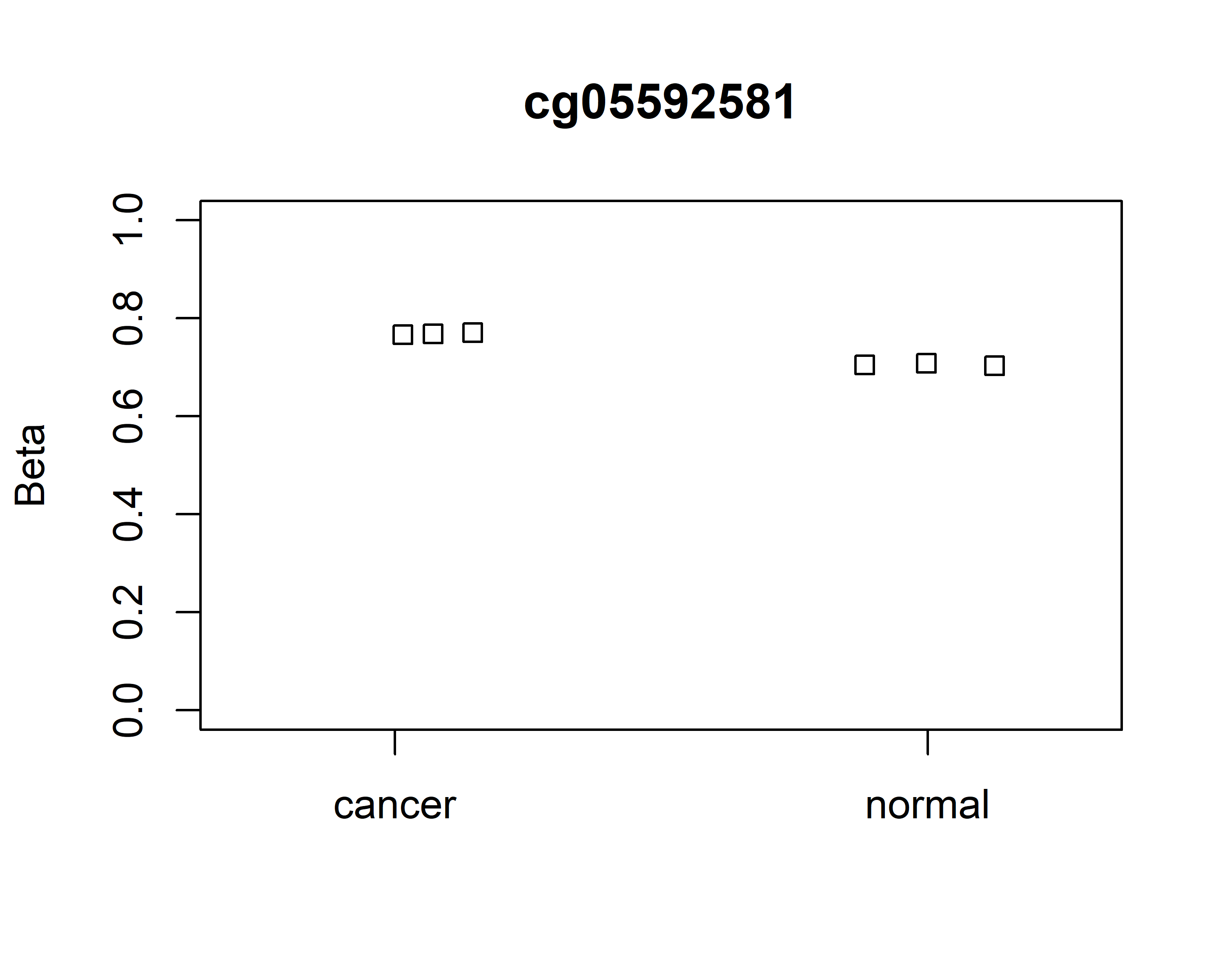
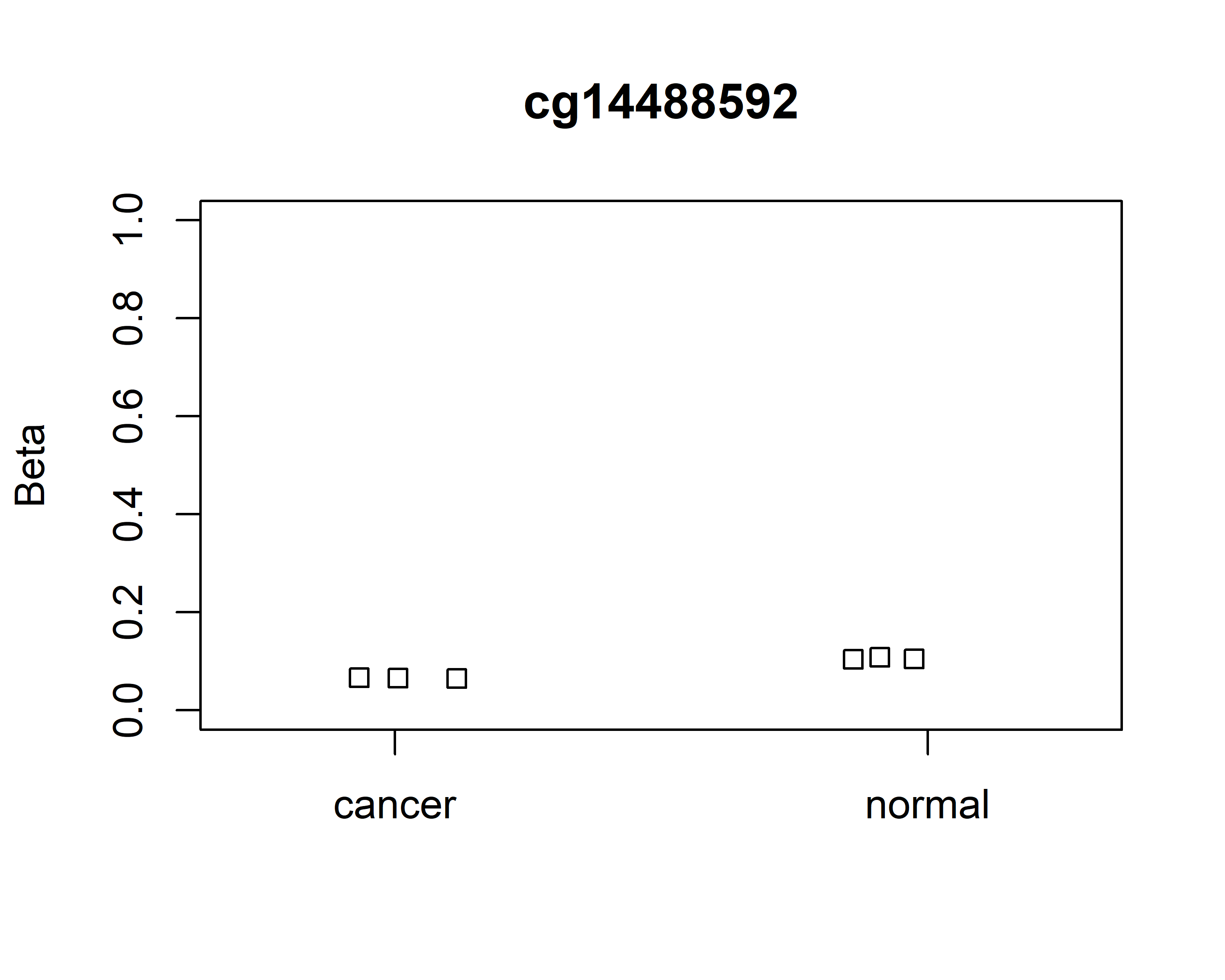
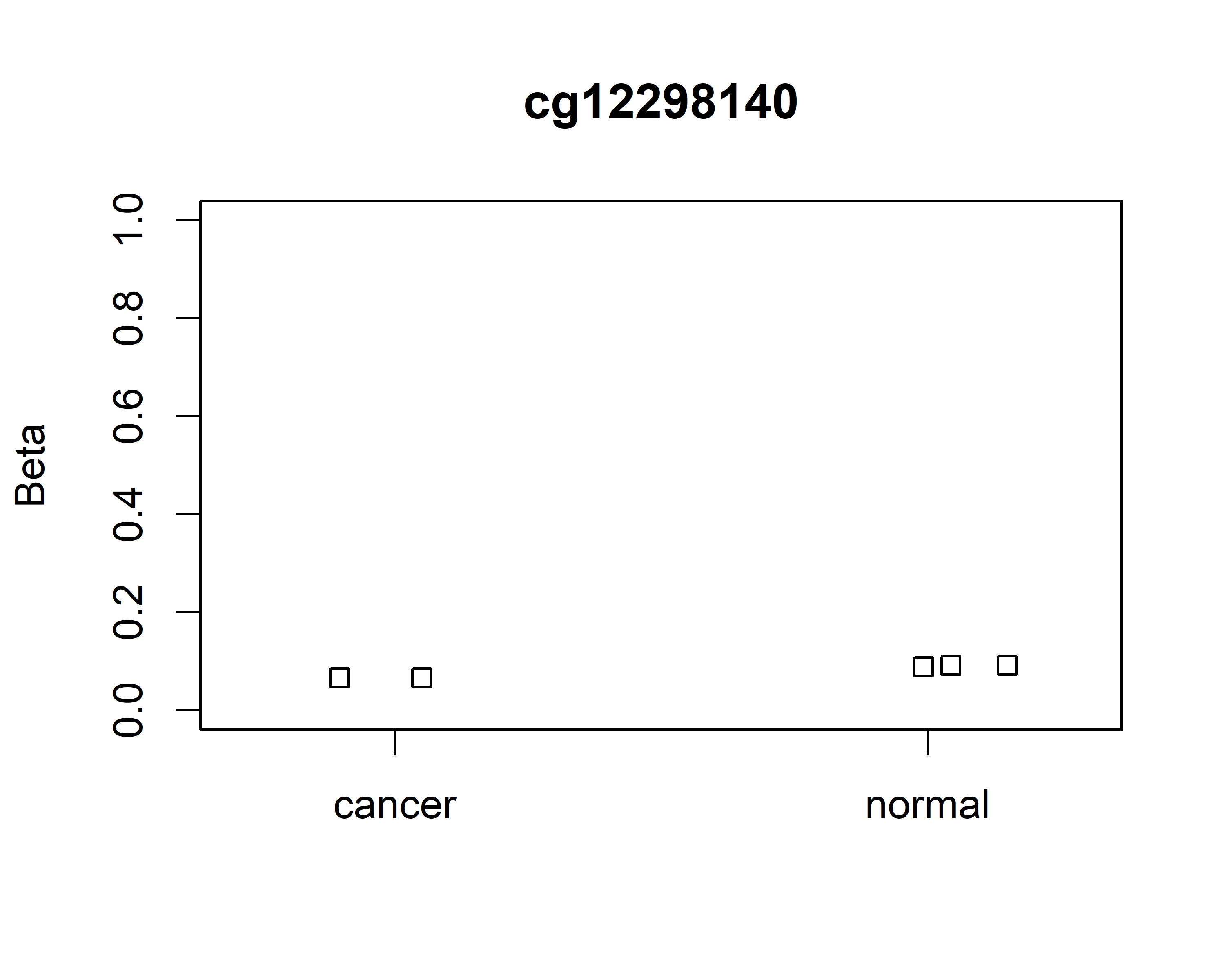
m = getM(msetSWAN)  
pheno = pData(rgSet)$Status  
dmp = dmpFinder(m,pheno = pheno,type = "categorical")  
dmp %>% round(.,3) %>% rownames\_to\_column(var = "CpG") %>%   
 head(.,10) %>% flextable(.)

| CpG | intercept | f | pval | qval |
| --- | --- | --- | --- | --- |
| cg12298140 | -3.815 | 1474.231 | 0 | 0.385 |
| cg14488592 | -3.825 | 1167.346 | 0 | 0.385 |
| cg05592581 | 1.735 | 1138.838 | 0 | 0.385 |
| cg08880423 | -3.639 | 1037.513 | 0 | 0.385 |
| cg00156072 | -3.196 | 834.794 | 0 | 0.385 |
| cg03150279 | -3.702 | 796.479 | 0 | 0.385 |
| cg00260655 | -4.009 | 793.936 | 0 | 0.385 |
| cg21046068 | -3.804 | 766.268 | 0 | 0.385 |
| cg13163765 | -3.344 | 740.721 | 0 | 0.385 |
| cg10033612 | -3.277 | 713.501 | 0 | 0.385 |

There are 0 DMPs with q-value . In the DMP table above, a positive intercept indicates that the methylation was higher in the cancer samples compared to the normal tissue samples. Of the 7 CpG sites significant at the p level, 1 was hypermethylated in cancer samples and 6 were hypomethylated. This makes some sense, because I would expect that in general cancer cells would have higher gene expression than normal cells.

### Plot CpGs

top4 = rownames(dmp[order(dmp$pval)[1:4],])  
plotCpg(msetSWAN,cpg = top4,pheno)



## c) Find DMP for sex

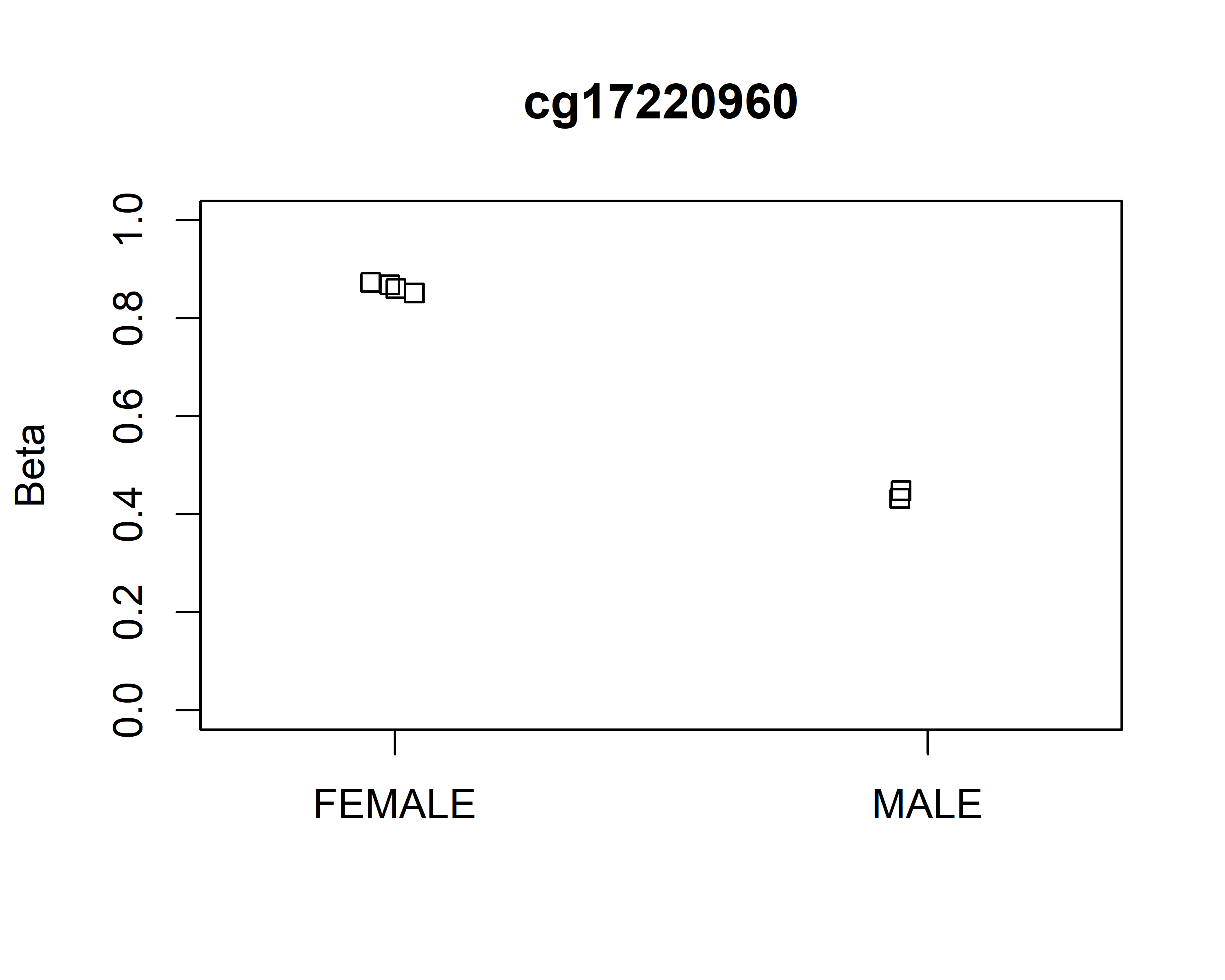
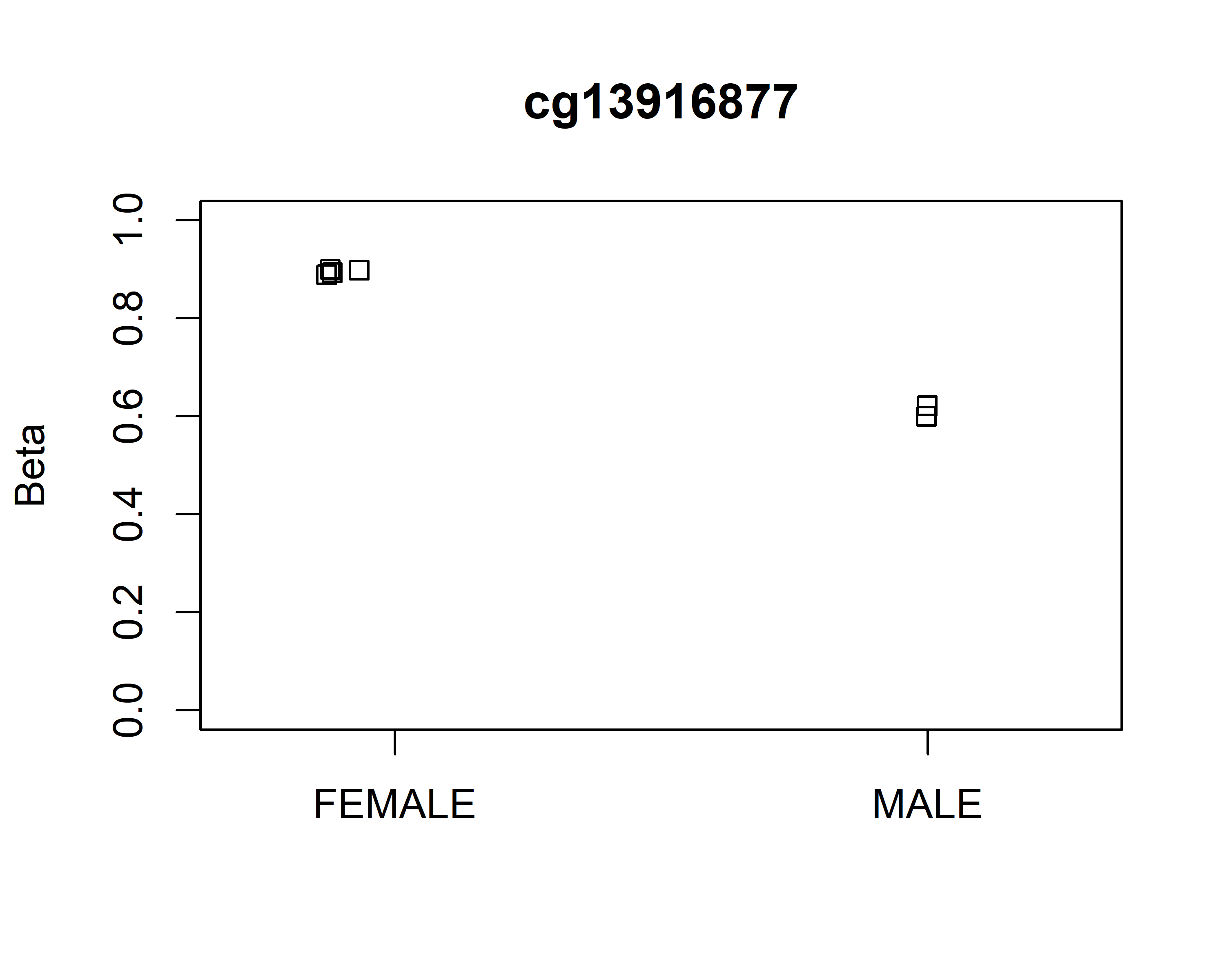
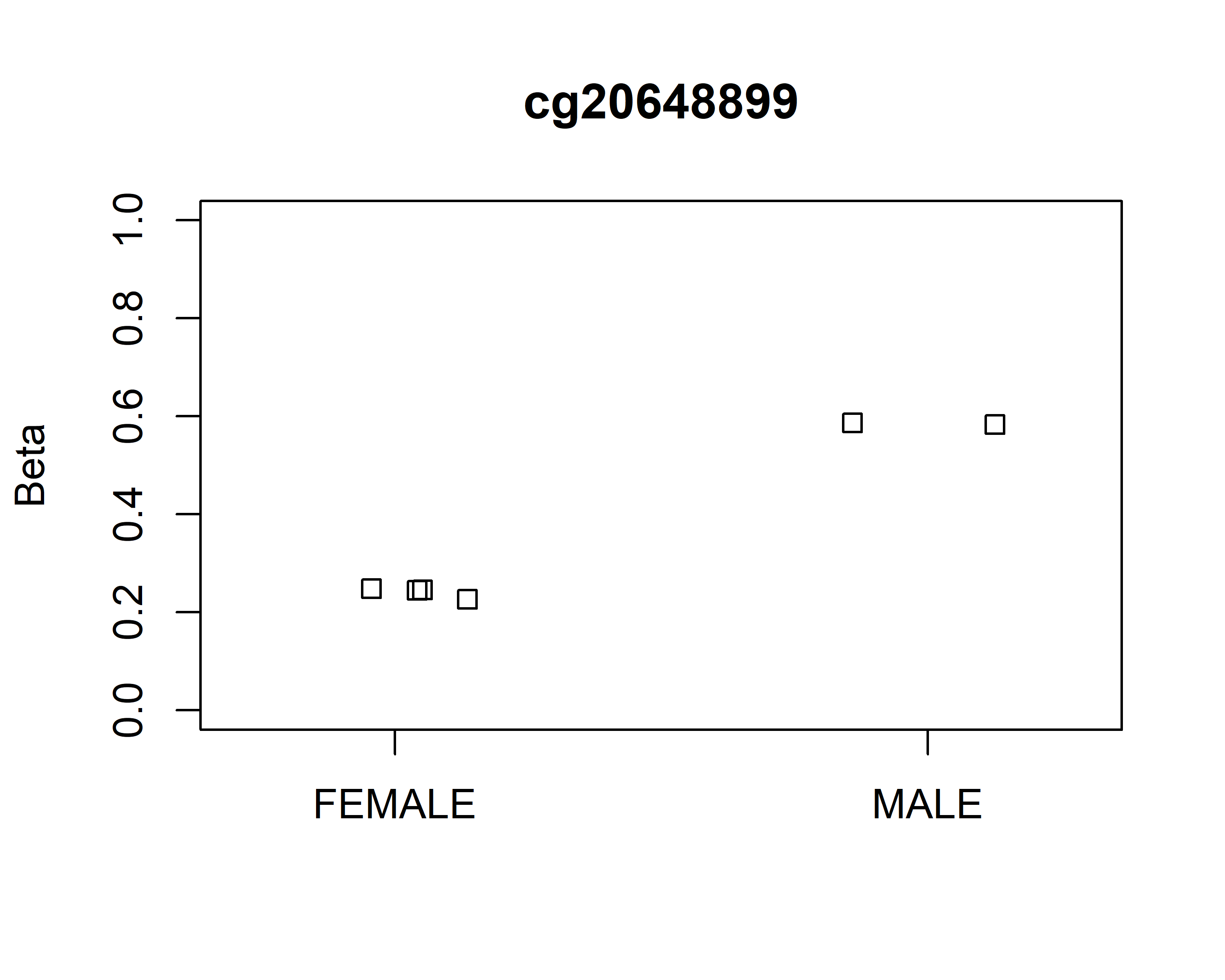
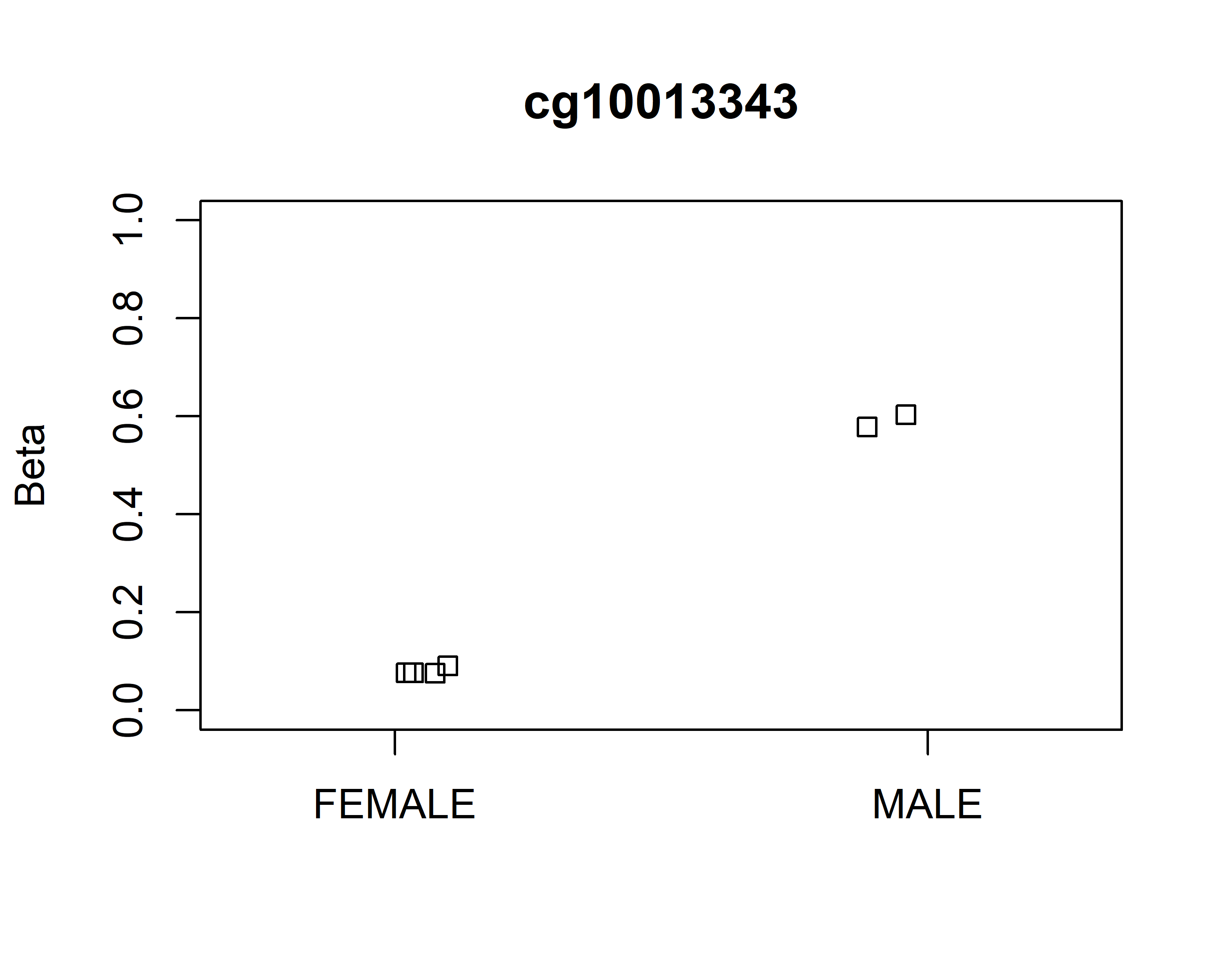
pheno = pData(rgSet)$Sex  
dmp = dmpFinder(m,pheno = pheno,type = "categorical")  
dmp %>% round(.,3) %>% rownames\_to\_column(var = "CpG") %>%   
 head(.,10) %>% flextable(.)

| CpG | intercept | f | pval | qval |
| --- | --- | --- | --- | --- |
| cg10013343 | -3.525 | 1259.157 | 0 | 0.224 |
| cg20648899 | -1.653 | 1234.053 | 0 | 0.224 |
| cg13916877 | 3.094 | 1233.287 | 0 | 0.224 |
| cg17220960 | 2.663 | 1159.953 | 0 | 0.224 |
| cg15479068 | -2.780 | 1100.499 | 0 | 0.224 |
| cg11403874 | -3.274 | 1069.127 | 0 | 0.224 |
| cg10019095 | -3.728 | 1062.031 | 0 | 0.224 |
| cg25742382 | 2.350 | 1058.829 | 0 | 0.224 |
| cg15308664 | -3.536 | 1046.970 | 0 | 0.224 |
| cg13764106 | -3.211 | 1028.167 | 0 | 0.224 |

There are 0 DMPs with q-value . Of the 17 CpG sites significant at the p level, 7 was hypermethylated in females and 10 were hypomethylated.

### Plot CpGs

top4 = rownames(dmp[order(dmp$pval)[1:4],])  
plotCpg(msetSWAN,cpg = top4,pheno)



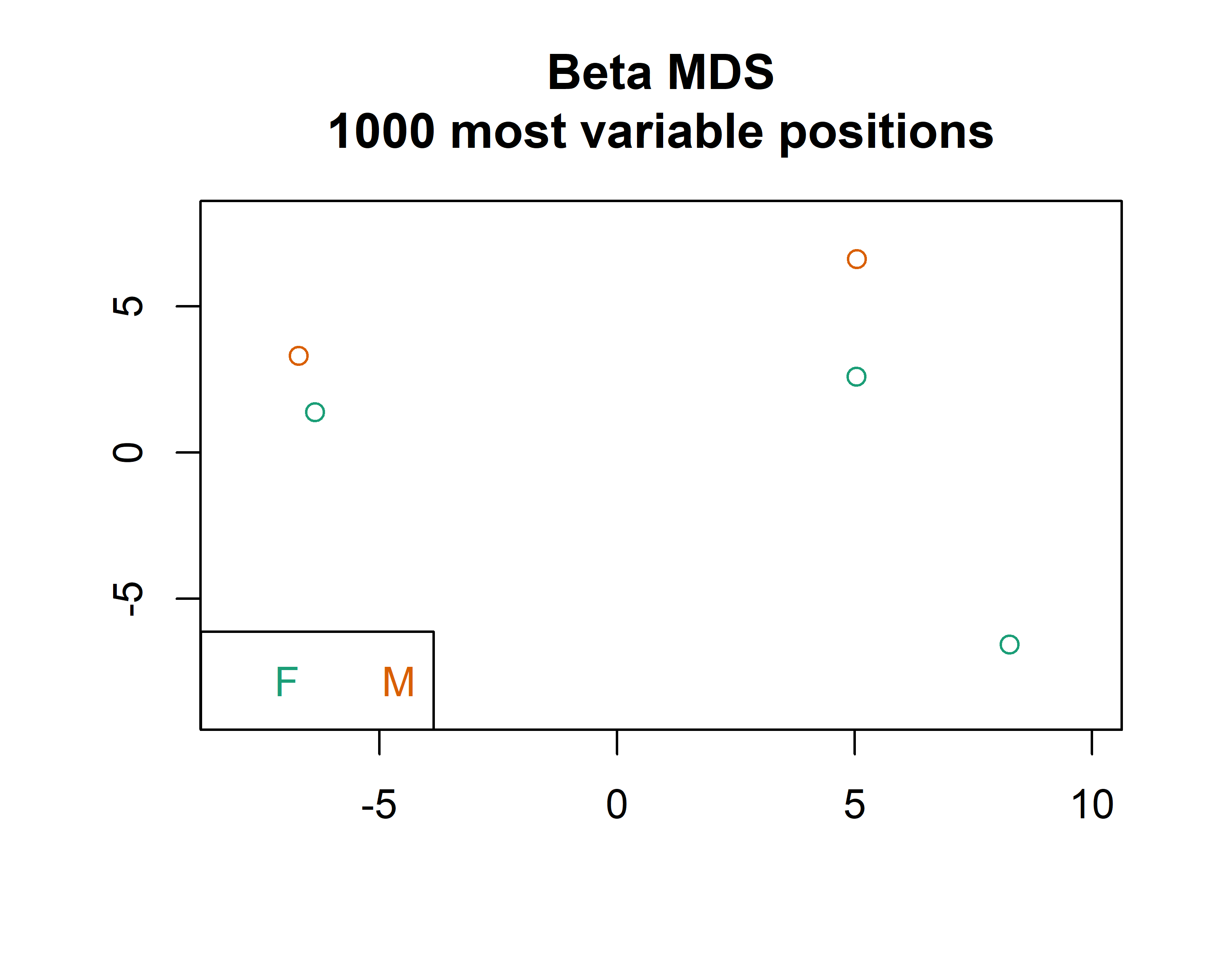
## d) Estimate whether each sample is male or female

gset <-mapToGenome(msetSWAN)  
gset = addSex(gset)  
pred\_sex =   
 as.data.frame(cbind(pData(gset)$predictedSex, pData(gset)$Sex))  
colnames(pred\_sex) = c("Predicted Sex","Label")  
flextable(pred\_sex)

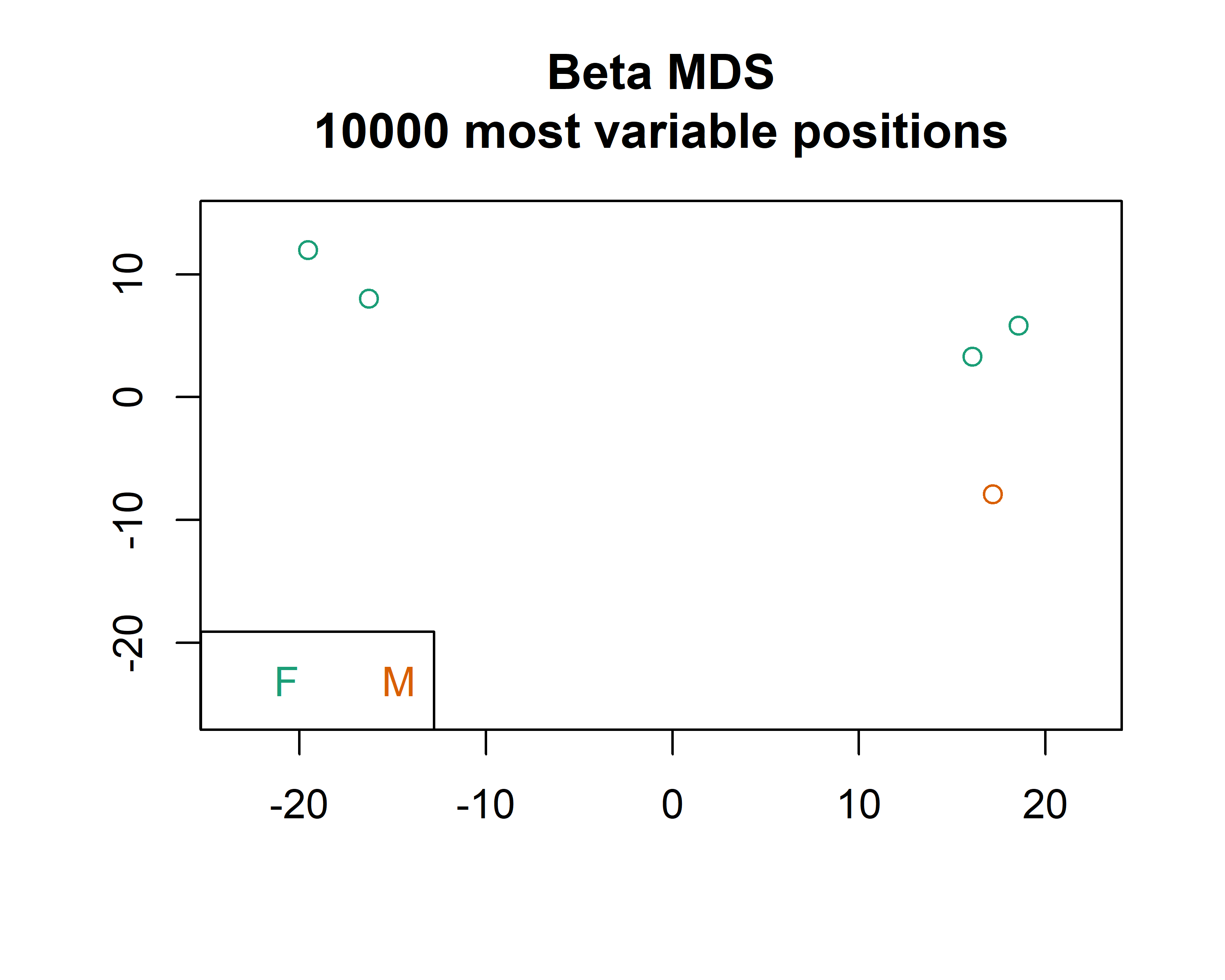
| Predicted Sex | Label |
| --- | --- |
| M | MALE |
| F | MALE |
| M | FEMALE |
| F | FEMALE |
| F | FEMALE |
| F | FEMALE |

### Re-do MDS plot

pred = pred\_sex$`Predicted Sex`  
mdsPlot(msetSWAN,sampGroups = pred)



mdsPlot(msetSWAN,numPositions = 10000,sampGroups = pred)



### Re-do 2c

dmp = dmpFinder(m,pheno = pred,type = "categorical")  
dmp %>% round(.,3) %>% rownames\_to\_column(var = "CpG") %>%   
 head(.,10) %>% flextable(.)

| CpG | intercept | f | pval | qval |
| --- | --- | --- | --- | --- |
| cg09307883 | 3.107 | 7219.882 | 0 | 0.024 |
| cg16788857 | -3.083 | 4699.921 | 0 | 0.024 |
| cg21130926 | 4.828 | 4408.130 | 0 | 0.024 |
| cg12077433 | 2.329 | 4149.080 | 0 | 0.024 |
| cg12186981 | 3.685 | 4039.051 | 0 | 0.024 |
| cg08250118 | 2.642 | 3610.564 | 0 | 0.024 |
| cg07889003 | 2.411 | 3354.124 | 0 | 0.024 |
| cg25791279 | -2.036 | 3162.311 | 0 | 0.024 |
| cg12165338 | 0.030 | 2879.515 | 0 | 0.024 |
| cg04003990 | 2.660 | 2874.105 | 0 | 0.024 |

After correcting the sex labels, there are 65 DMPs with q-value .

## e) bumphunter

# Re-run for cancer status  
pheno = pData(rgSet)$Status  
dmp = dmpFinder(m,pheno = pheno,type = "categorical")  
# Code from questions  
diffs <- dmp$intercept   
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 the first region identified  
j=1  
ind = segs$dnIndex[[j]]  
index <- which(cl==cl[ind])  
plot(pos[index],diffs[index],  
xlab=paste("position on", chr[ind]), ylab="diff")  
abline(h = 0, col = "blue")

