BIOS 7659 Homework 7

Tim Vigers

17 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)

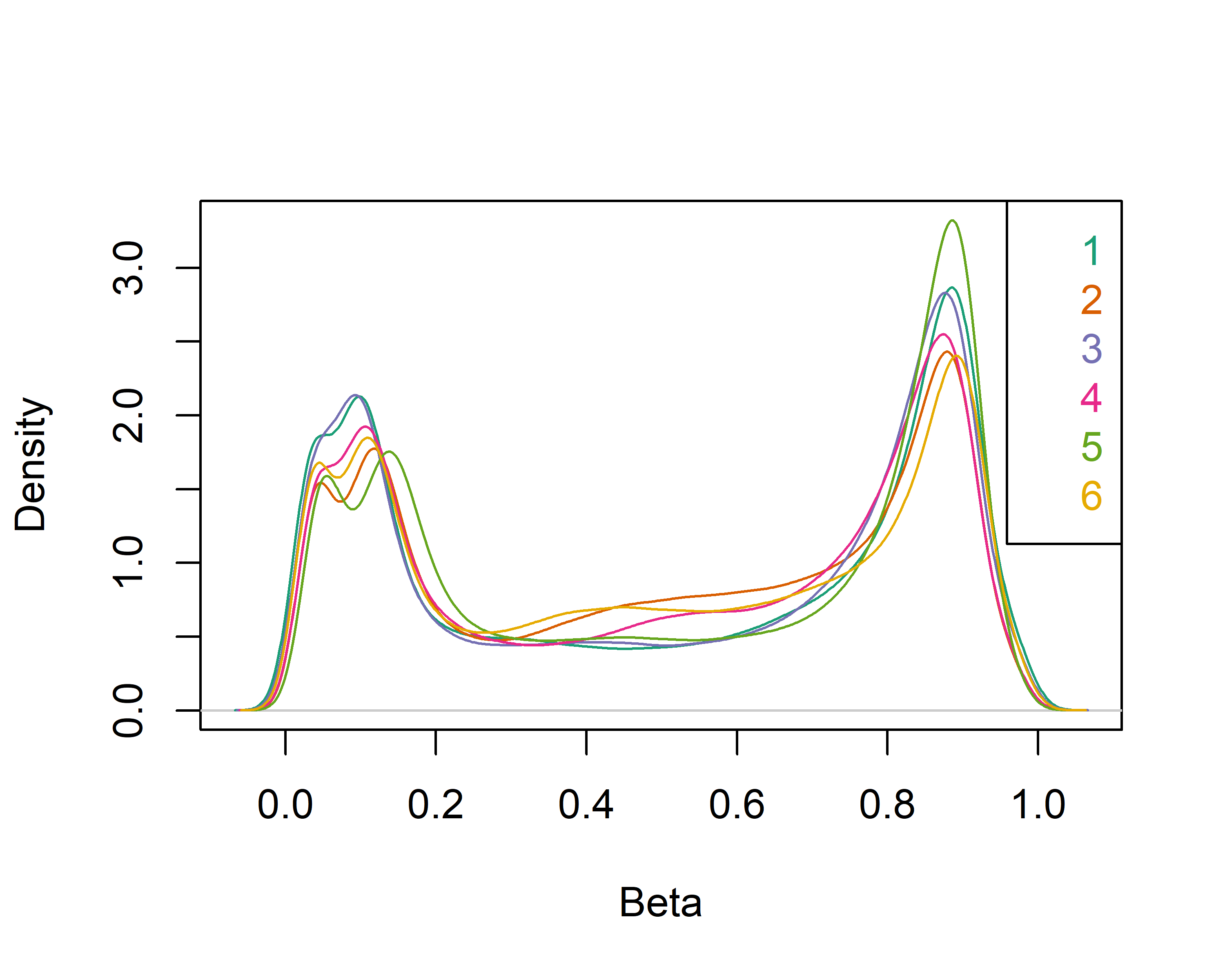
## IlluminaMethylationManifest object  
## Annotation  
## array: IlluminaHumanMethylation450k  
## Number of type I probes: 135476   
## Number of type II probes: 350036   
## Number of control probes: 850   
## Number of SNP type I probes: 25   
## Number of SNP type II probes: 40

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.

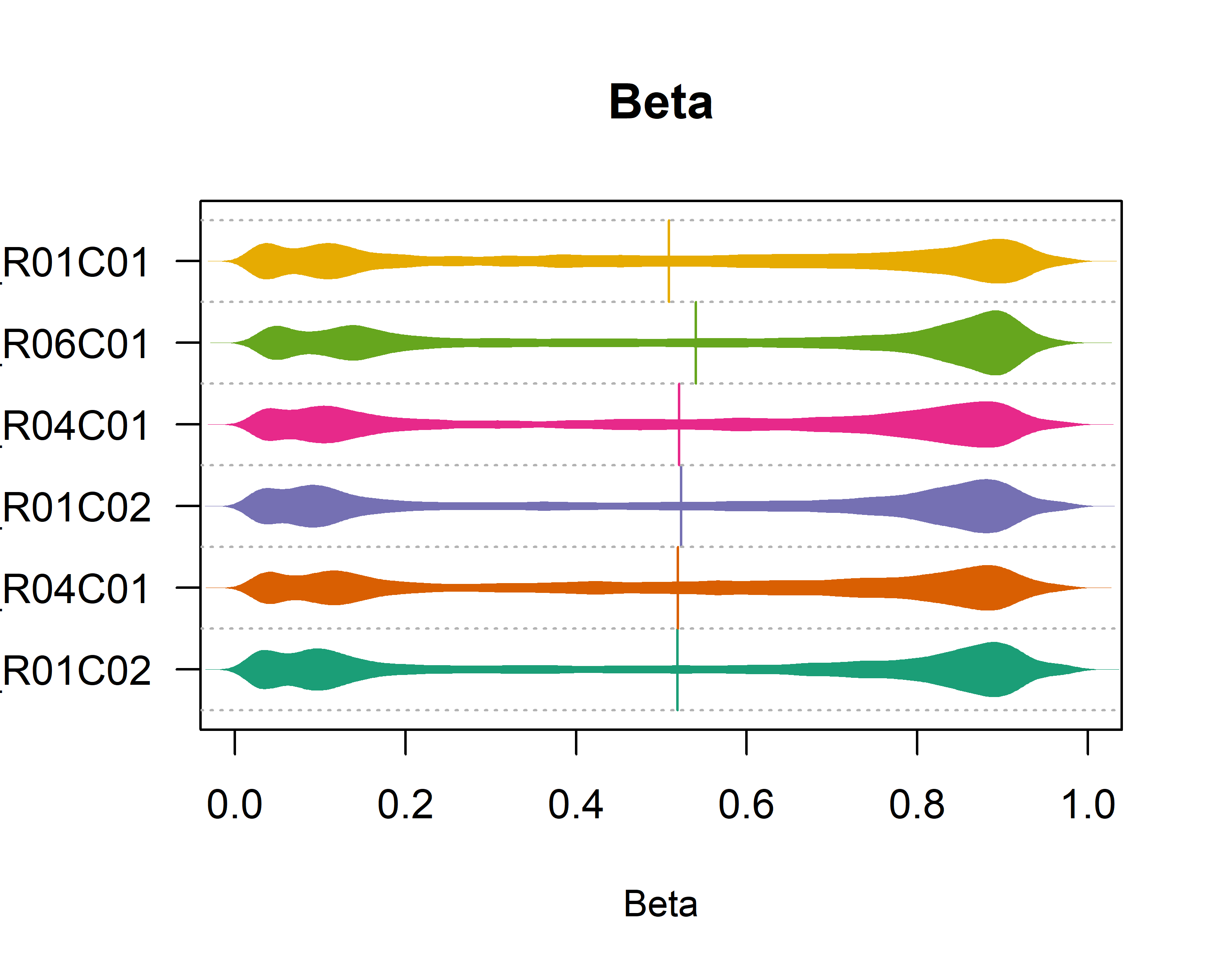
## c) QC Plots

### By ID

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



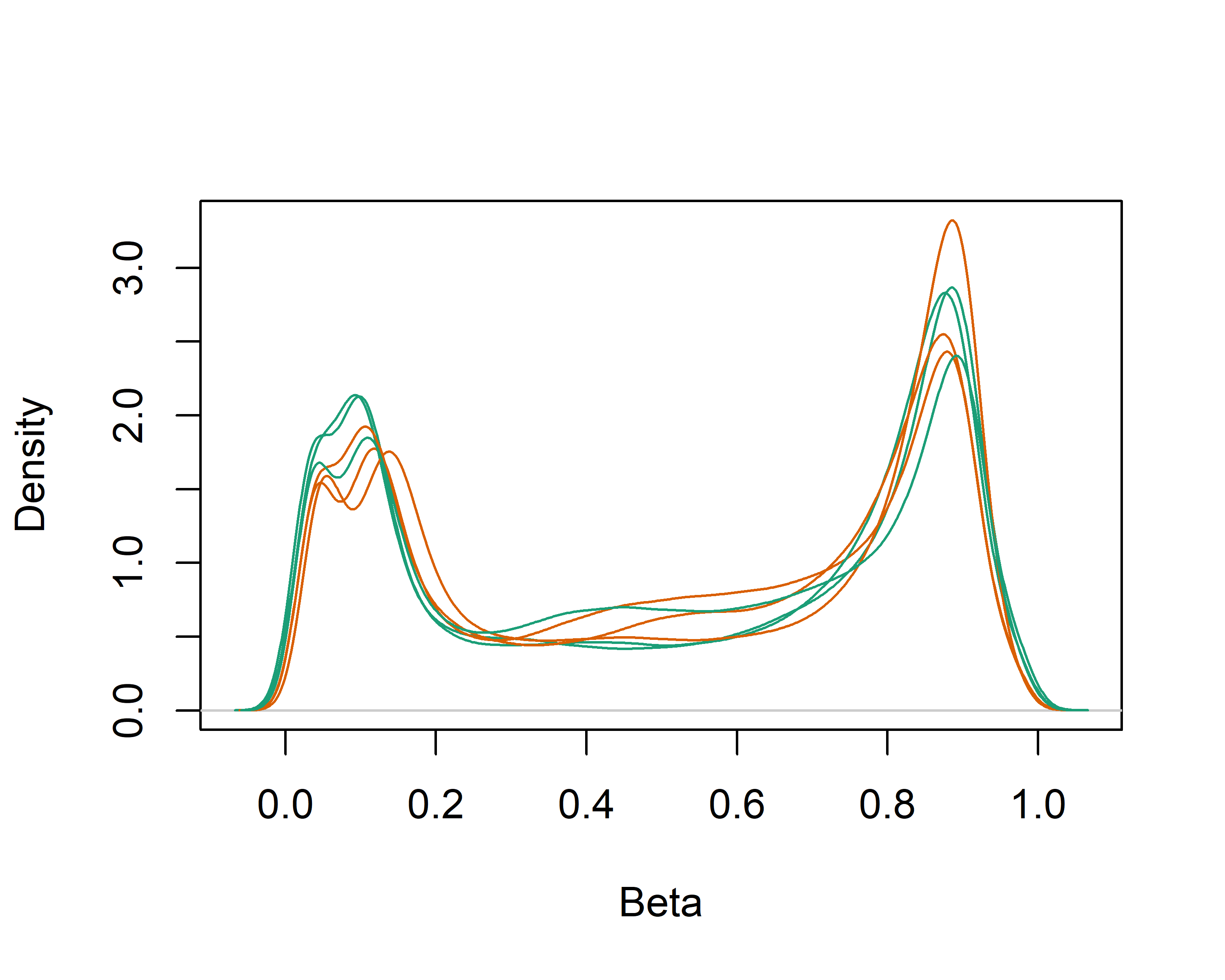
densityBeanPlot(rgSet,sampGroups = id)



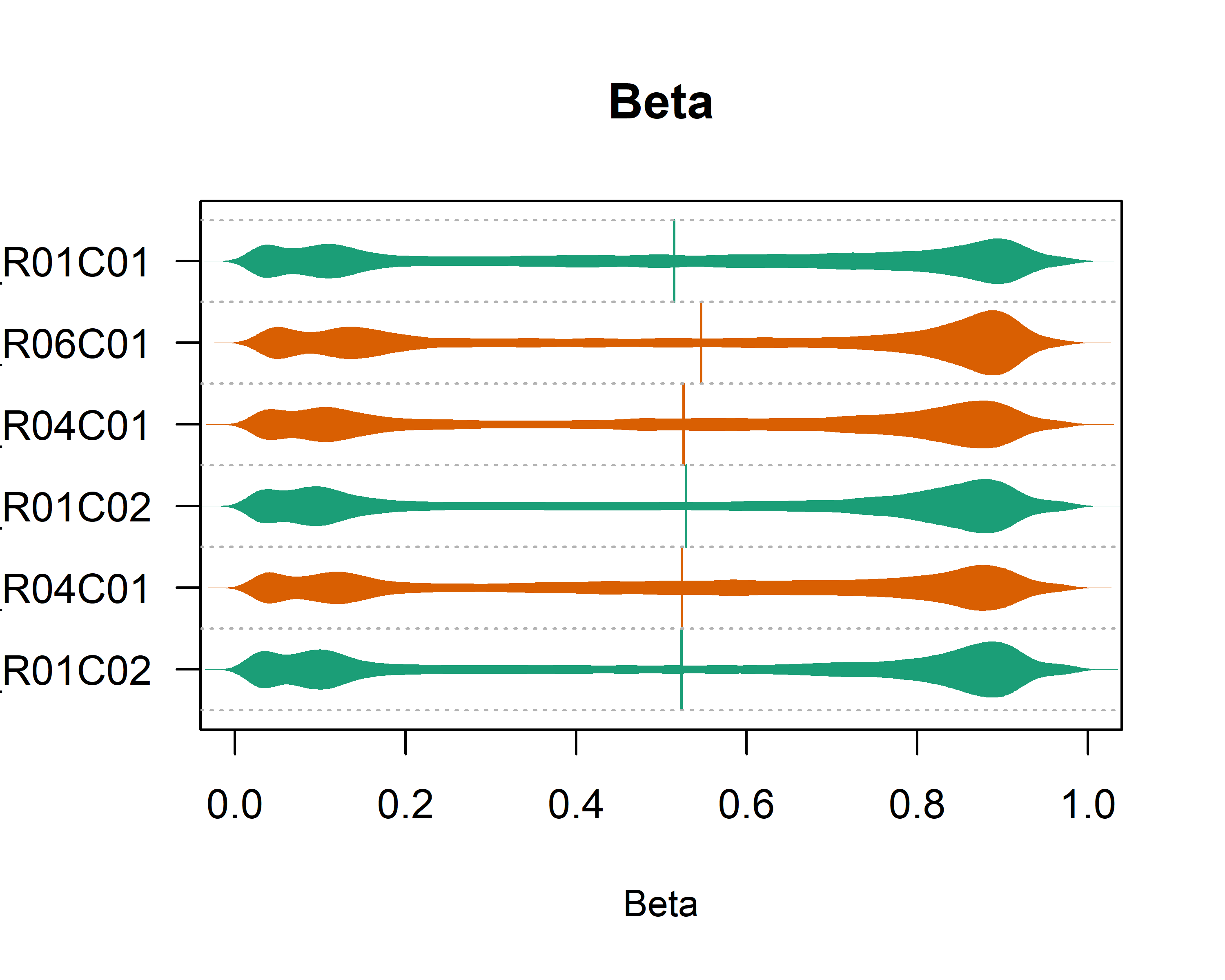
Unfortunately I get the error “‘sampNames’ is not a graphical parameter” when trying to use sampNames as requested in the prompt, so I used sample ID as the group ID. On first inspection, nothing in the density plot particularly jumped out at me, although the peak for sample 5 is a little higher than the other at the 1.0 end of the x axis. This sample also looks different from the others in the bean plot because its center is skewed further to the right than the others.

### By sample type

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



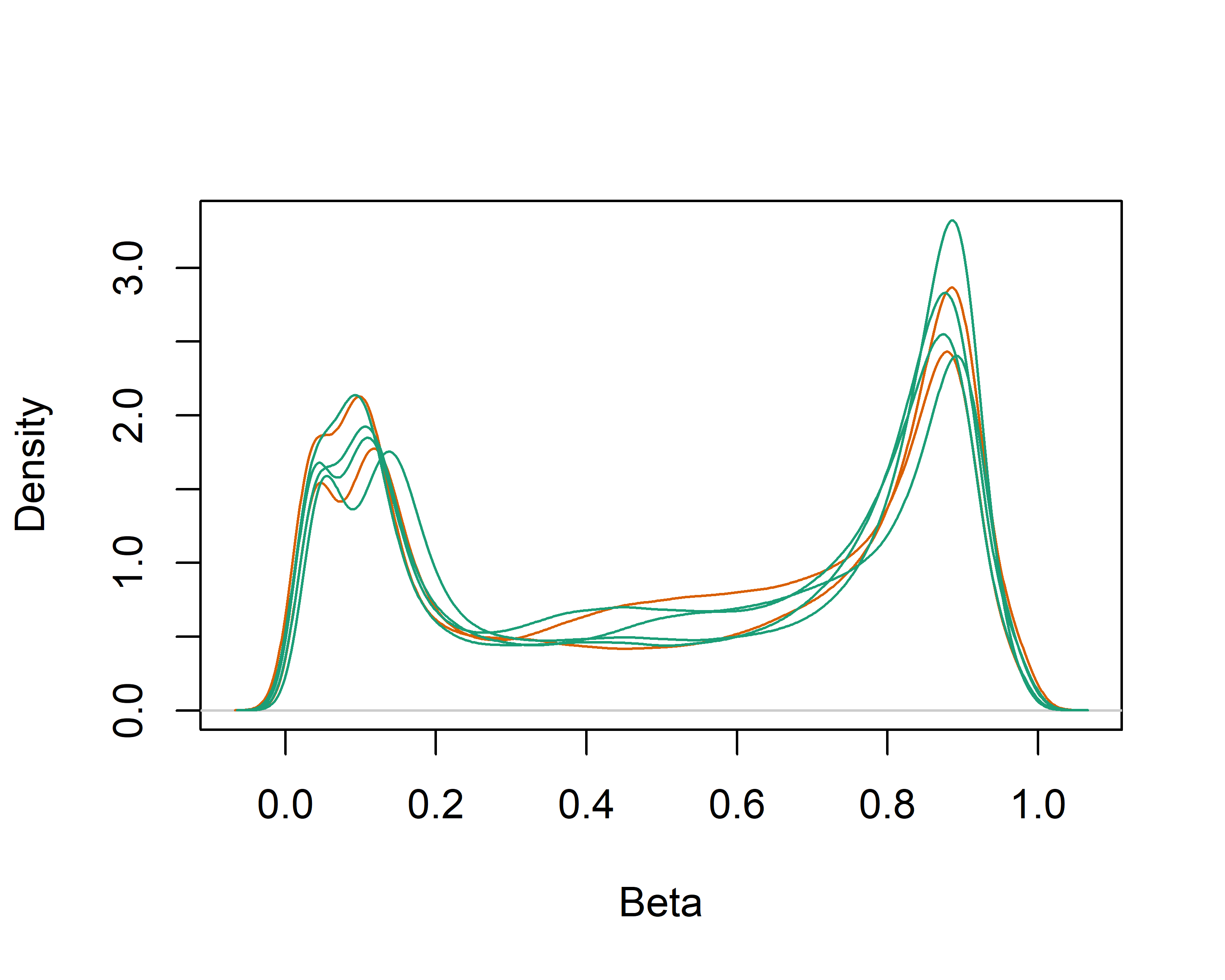
densityBeanPlot(rgSet,sampGroups = stype)



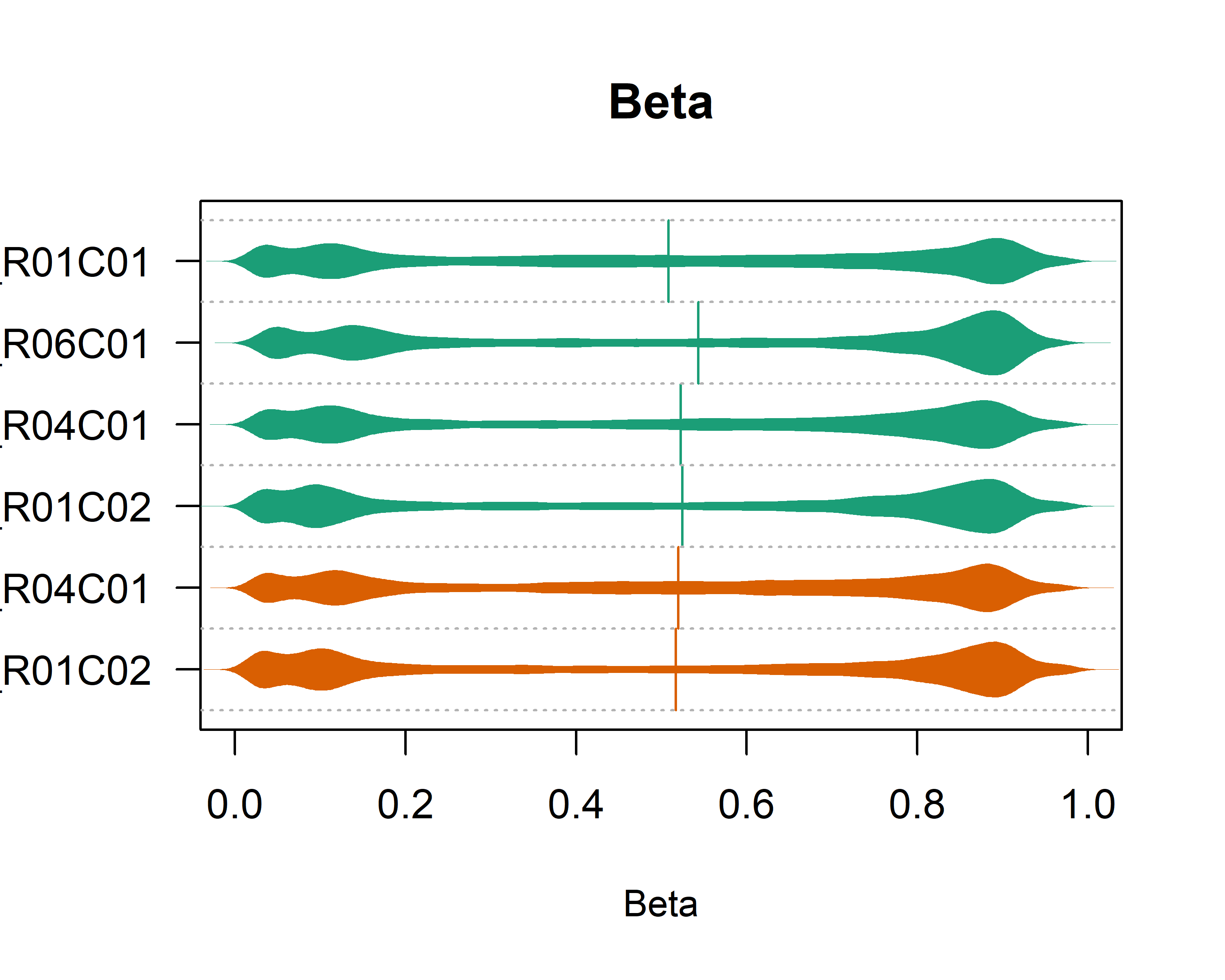
I removed the legend above because otherwise it covers most of the right-hand peak, but tumor samples are in green and normal tissue samples are in orange. No patterns particularly stand out to me here, except that sample 5 stands out even more when it’s colored orange. I imagine that people who look at these plots more often might notice a pattern by sample type, but all I can see in these plots is the fact that sample 5 (R06C01) looks different from the rest.

### By sex

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



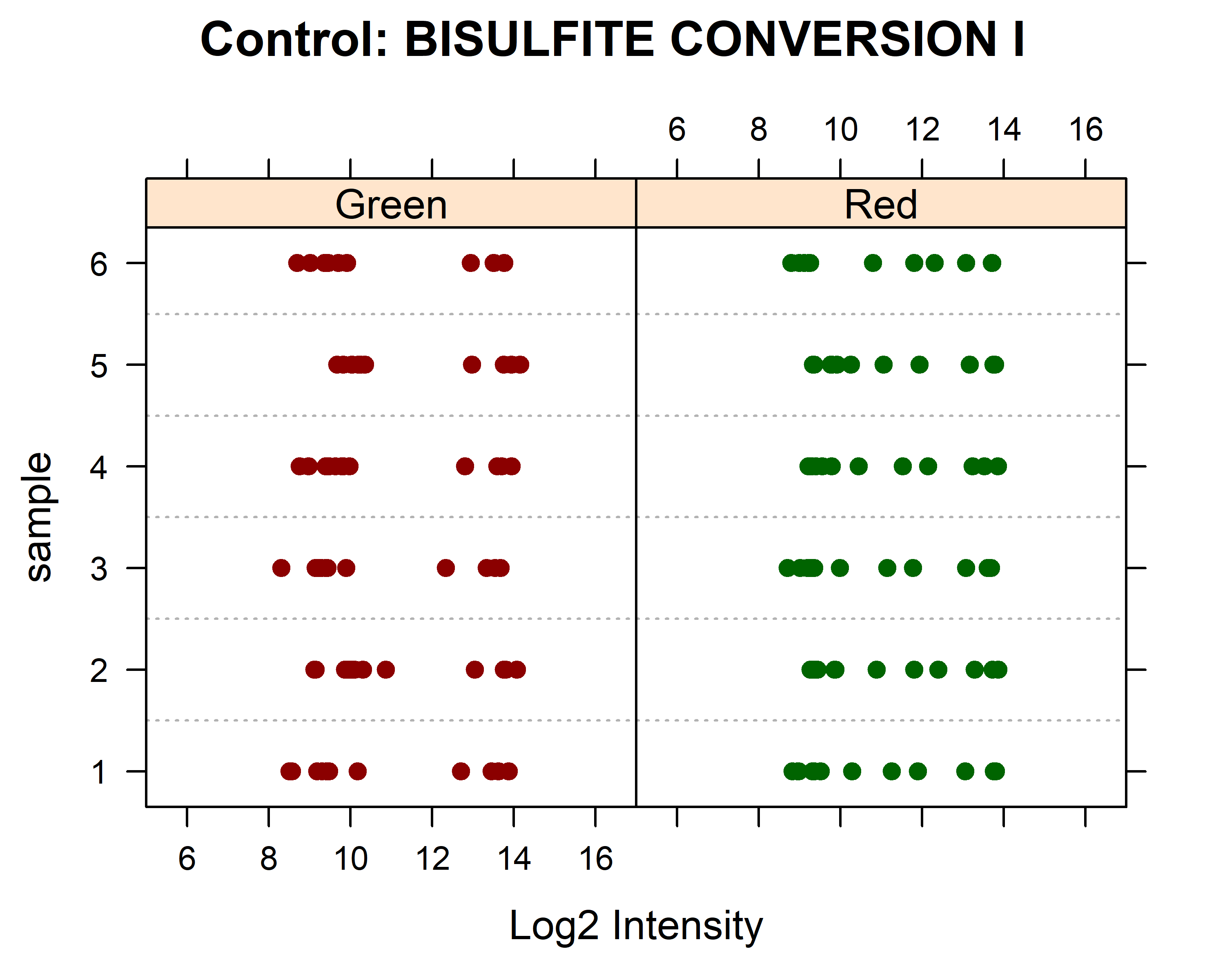
densityBeanPlot(rgSet,sampGroups = sex)



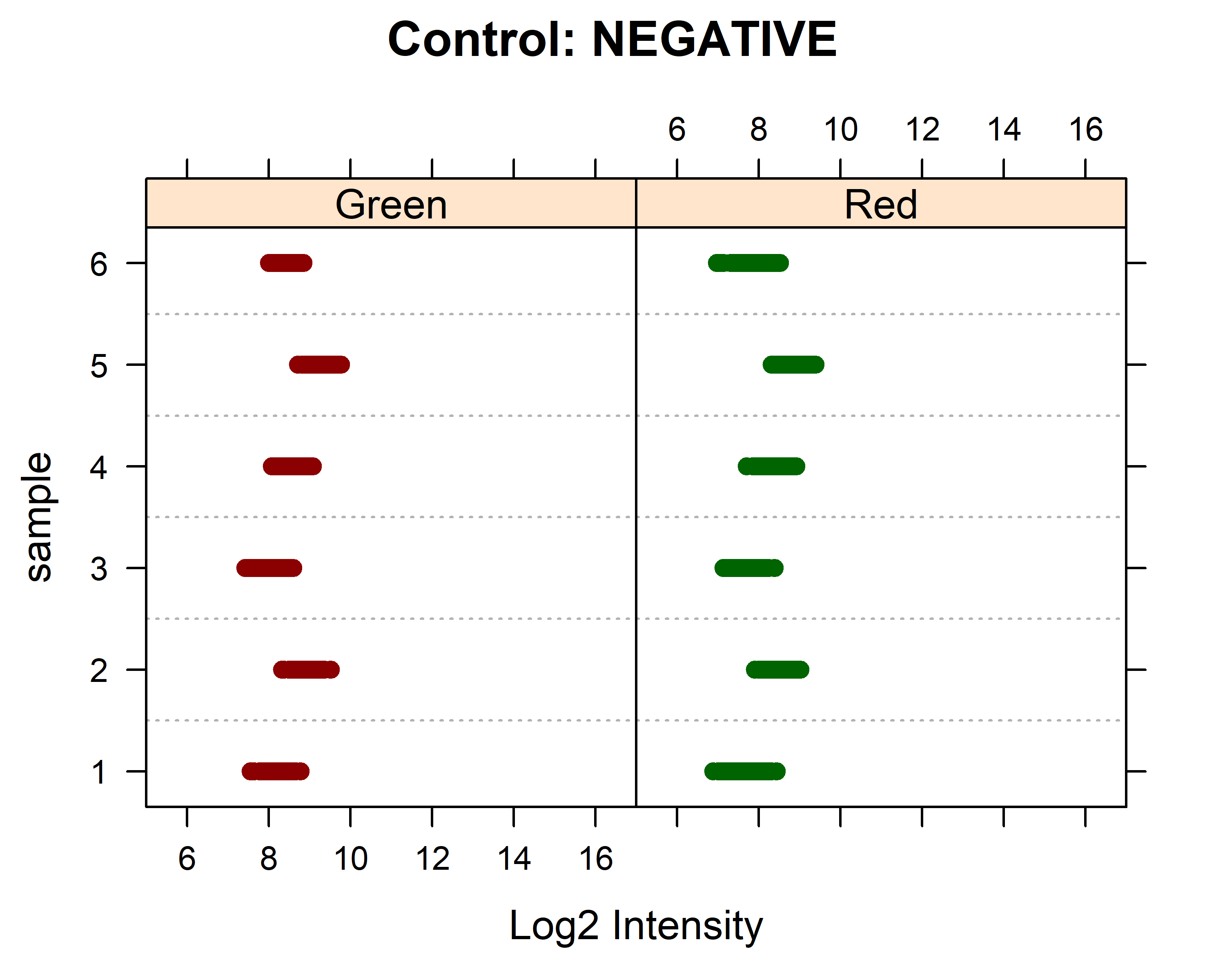
Again, I had to remove the legend in the density plot because it covers up the right-hand peak, but in these plots females are in green and males in orange. Again, sample 5 stands out to me here but I don’t see any patterns by sex. Sample 5 is not so different that it should be excluded, but I would try some between sample normalization and see if that helps it align more with the other samples.

## d) Control probes

controlStripPlot(rgSet,controls = "BISULFITE CONVERSION I",  
 sampNames = rgSet$id)



controlStripPlot(rgSet,controls = "NEGATIVE",  
 sampNames = rgSet$id)



Each 450K array includes internal control probes to help assess the quality of various pre-treatment steps. Bisulfite conversion probes are included to help assess the efficiency of bisulfite conversion, because “if the bisulfite conversion reaction was successful, the”C" (Converted) probes will match the converted sequence and get extended." Negative controls are included to help with background correction, by targeting bisulfite-converted sequences that don’t contain CpGs.

In the plots above, we want samples to be consistent within each of the red and green channels, so these plots look fairly reasonable. Samples 2 and 5 may look slightly different from the others, but it’s difficult to say for sure based on just these plots. Also, the range of intensities is about the same for both the red and green channel, which is a good sign for this data because it indicates that there isn’t significant red/green bias. Also, the intensity of the negative probes is generally lower than the bisulfite probes, which indicates that background noise should not be an issue with this dataset.

## 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)

Detection p values indicate whether a probe’s signal is greater than the average background signal. They are calculated as where x is the sum of two beads (type I probes) or two color intensities (type II probes) and is the normal CDF.

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)

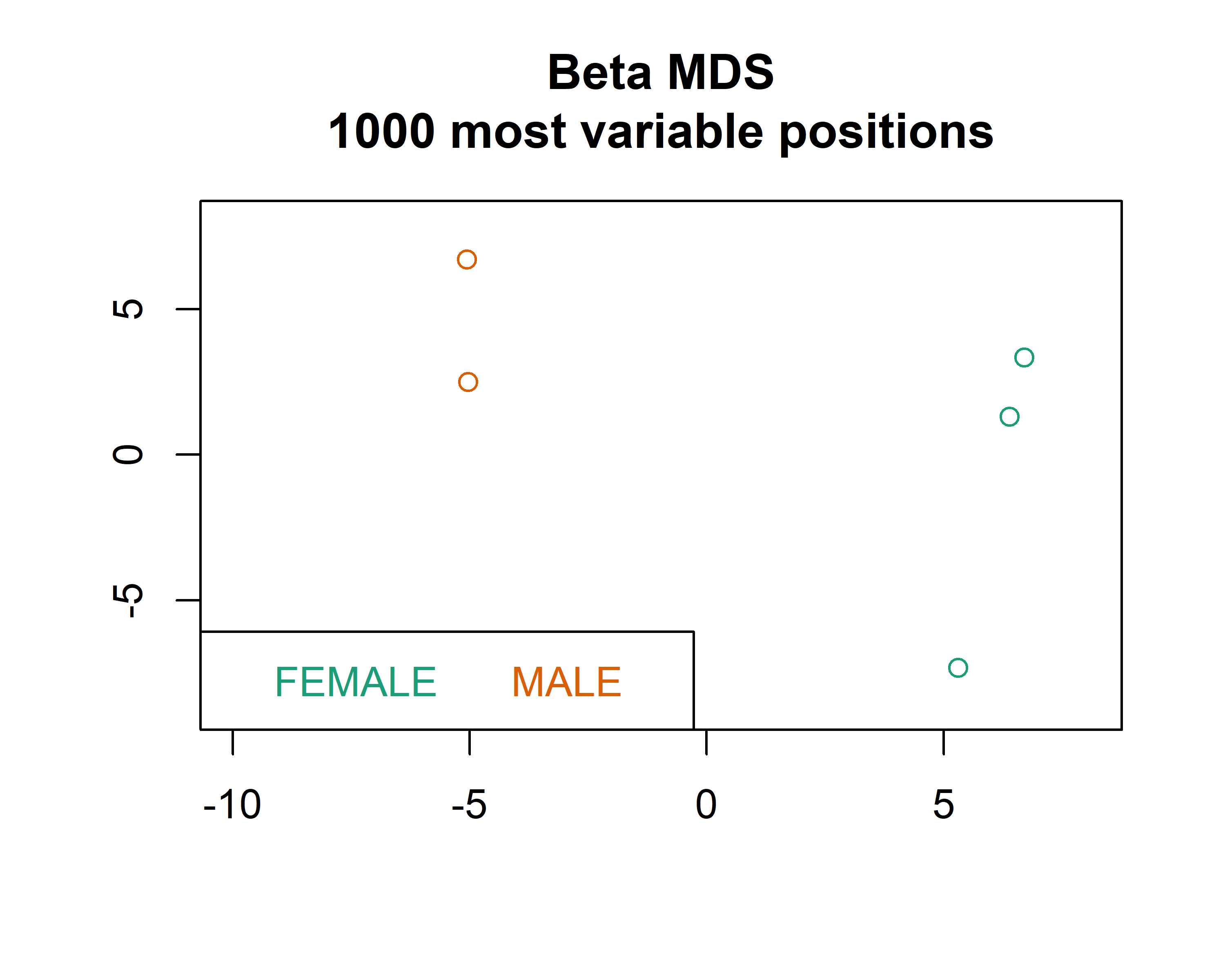
## f) Multidimensional scaling (MDS) plots

# ADD SAMPLE NAMES TO ALL PLOTS

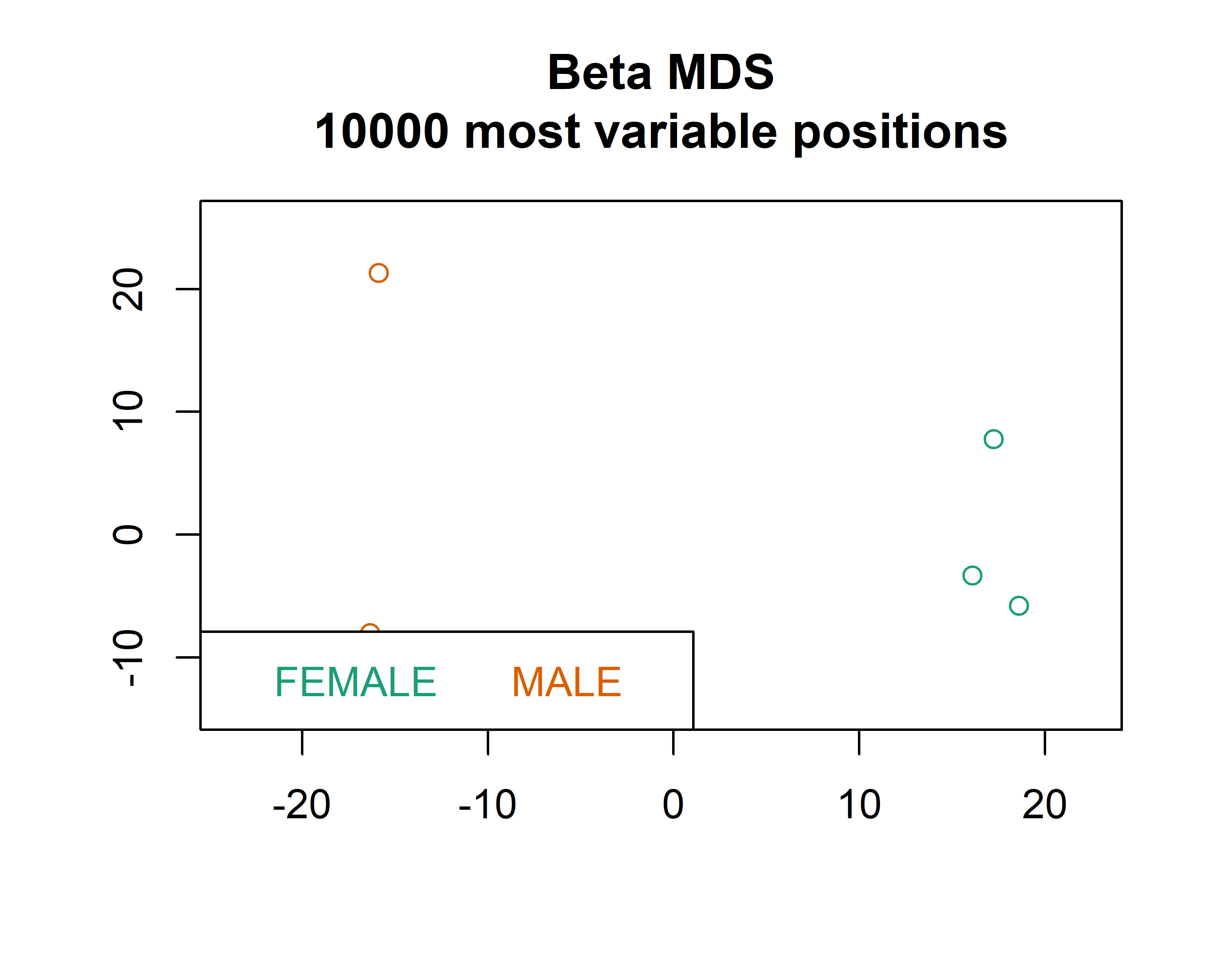
### By sex

#### SWAN-normalized

mdsPlot(msetSWAN,sampGroups = sex)

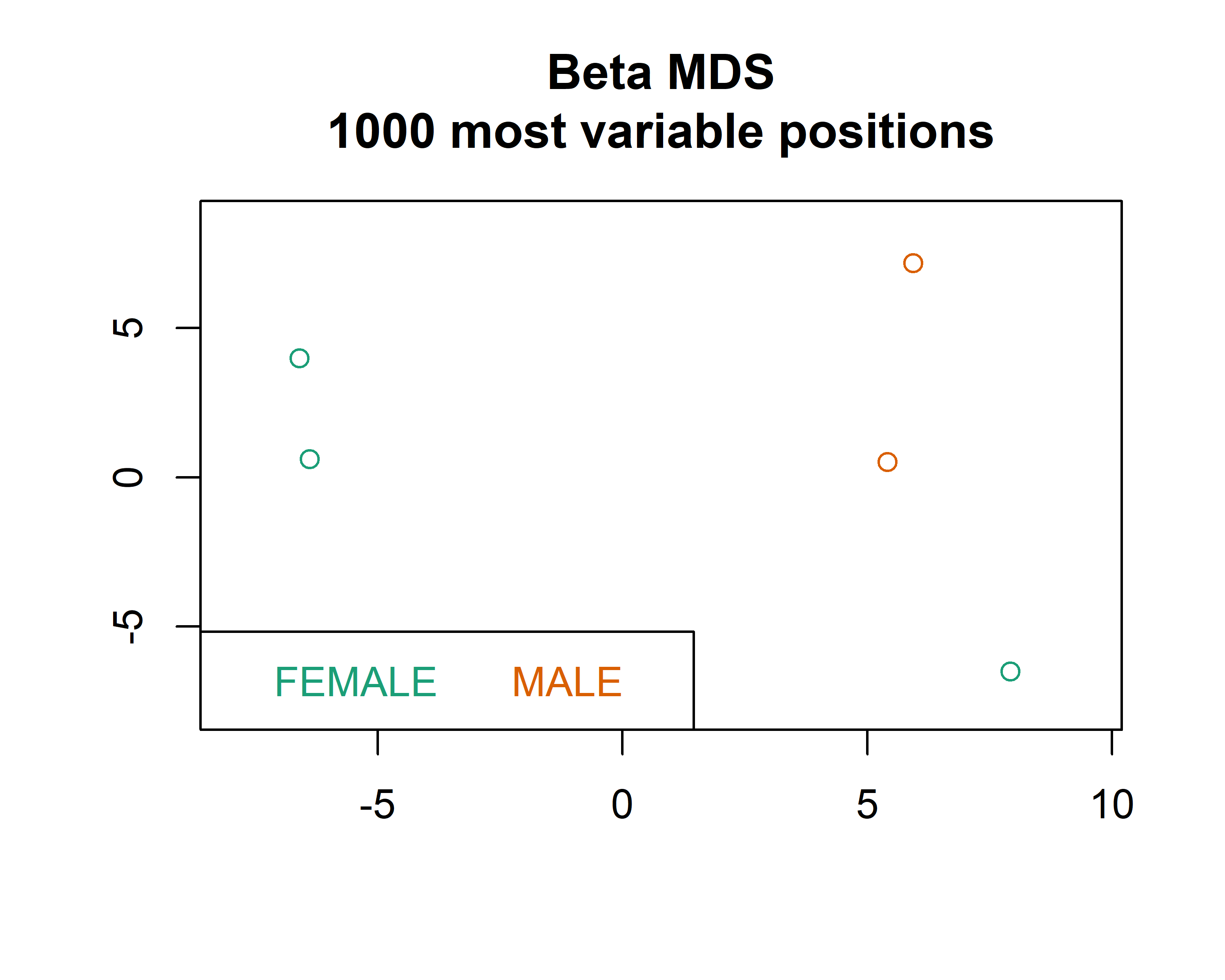


mdsPlot(msetSWAN,numPositions = 10000,sampGroups = sex)

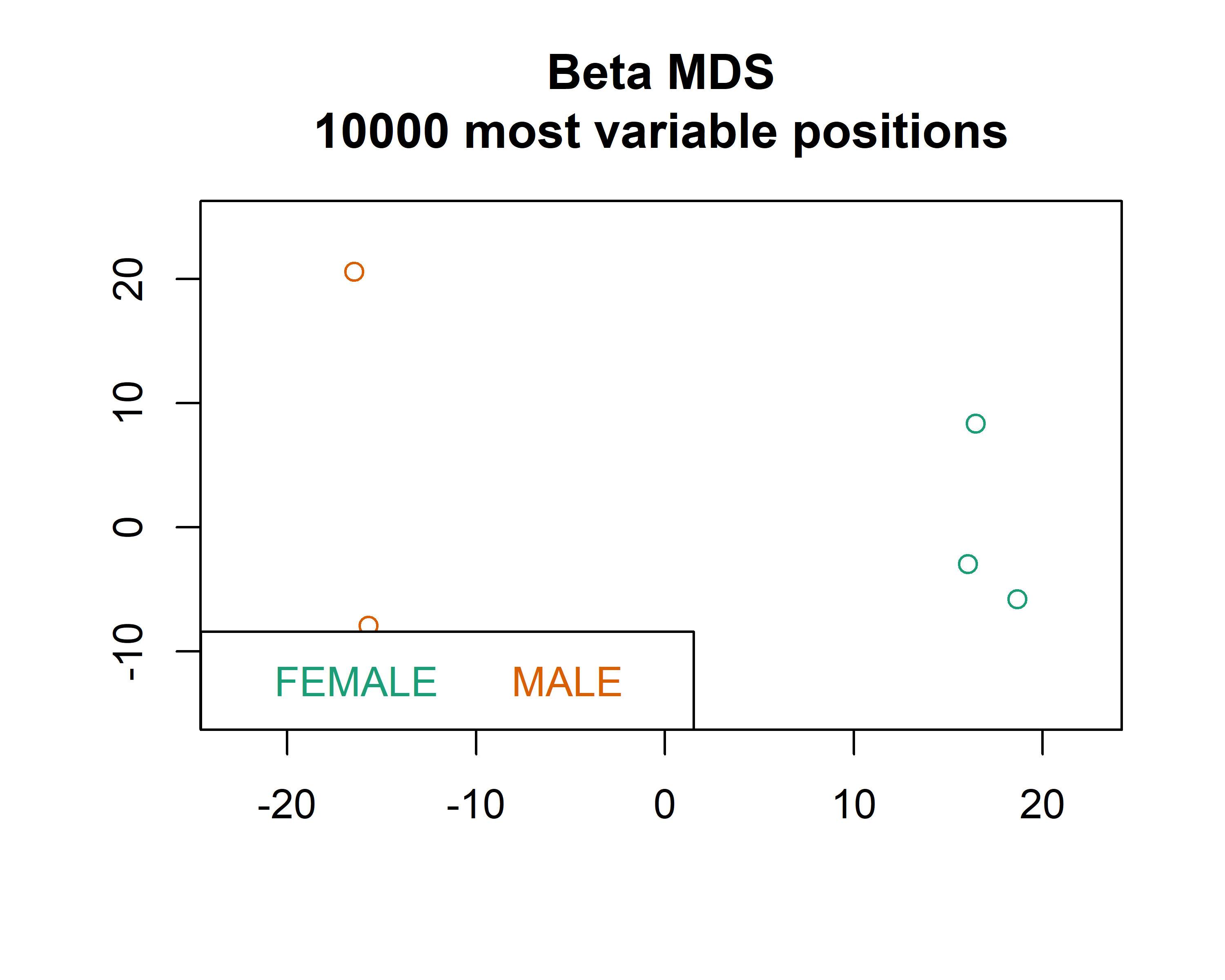


#### Raw data

mdsPlot(mset,sampGroups = sex)



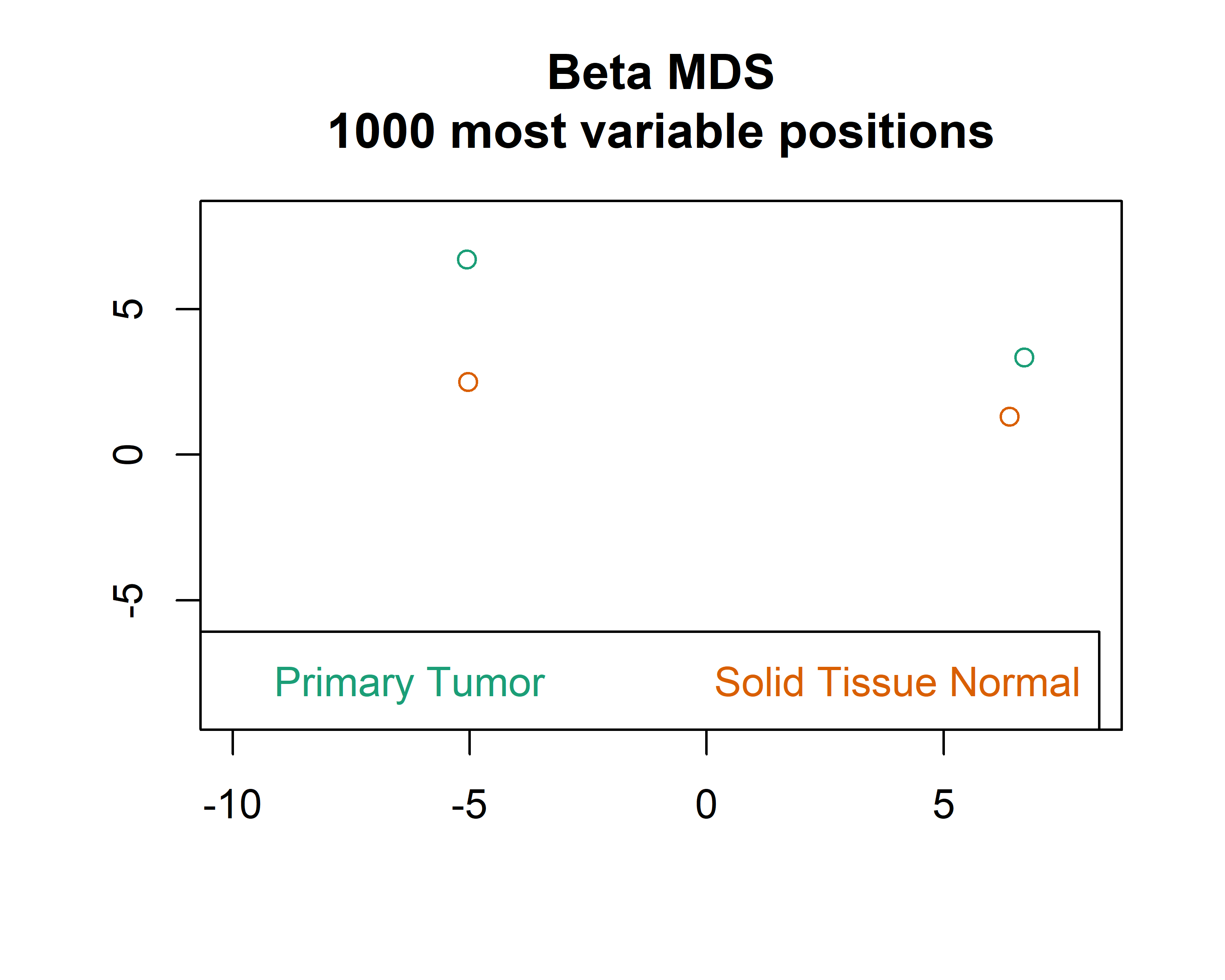
mdsPlot(mset,numPositions = 10000,sampGroups = sex)



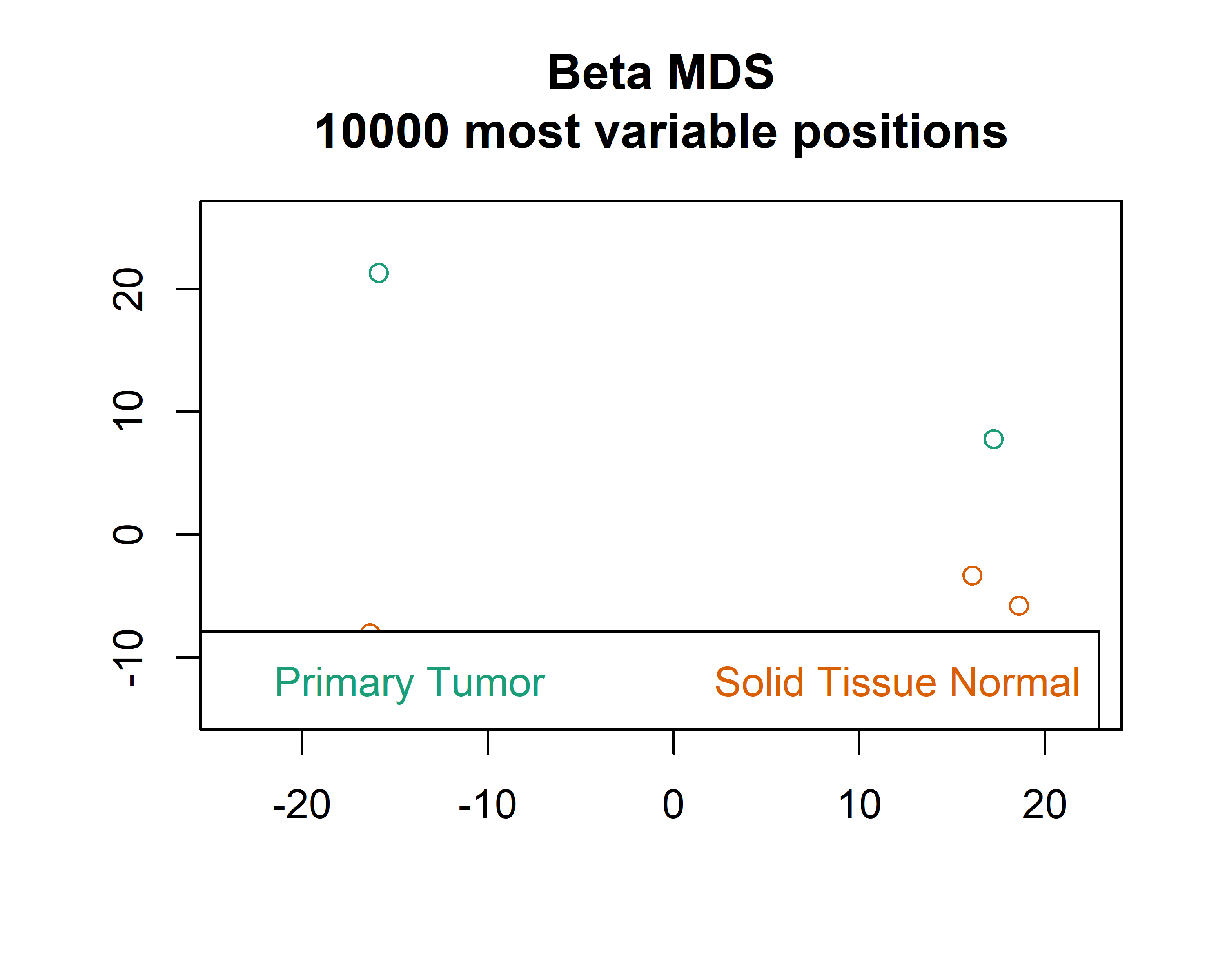
### By cancer status

#### SWAN-normalized

mdsPlot(msetSWAN,sampGroups = stype)

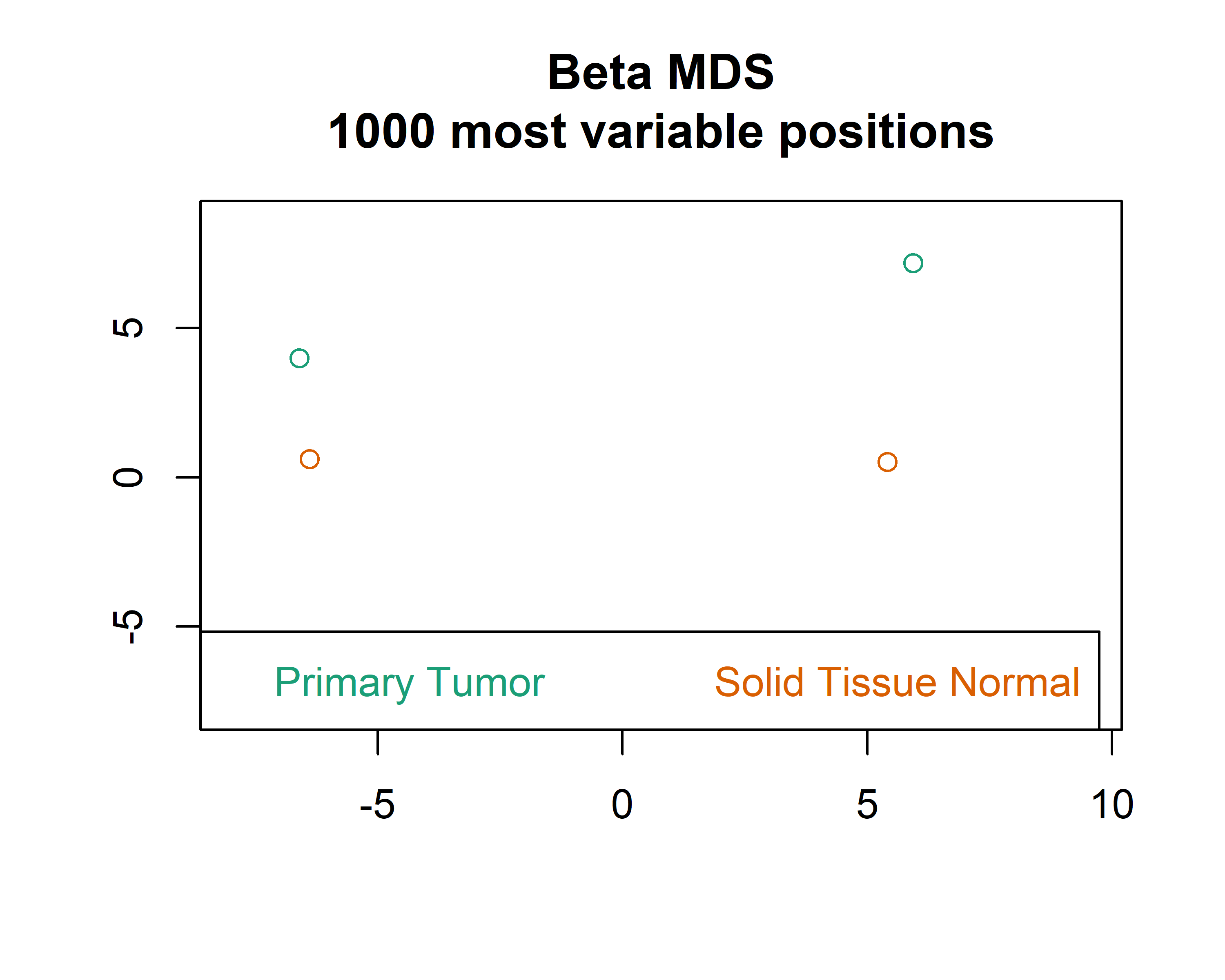


mdsPlot(msetSWAN,numPositions = 10000,sampGroups = stype)

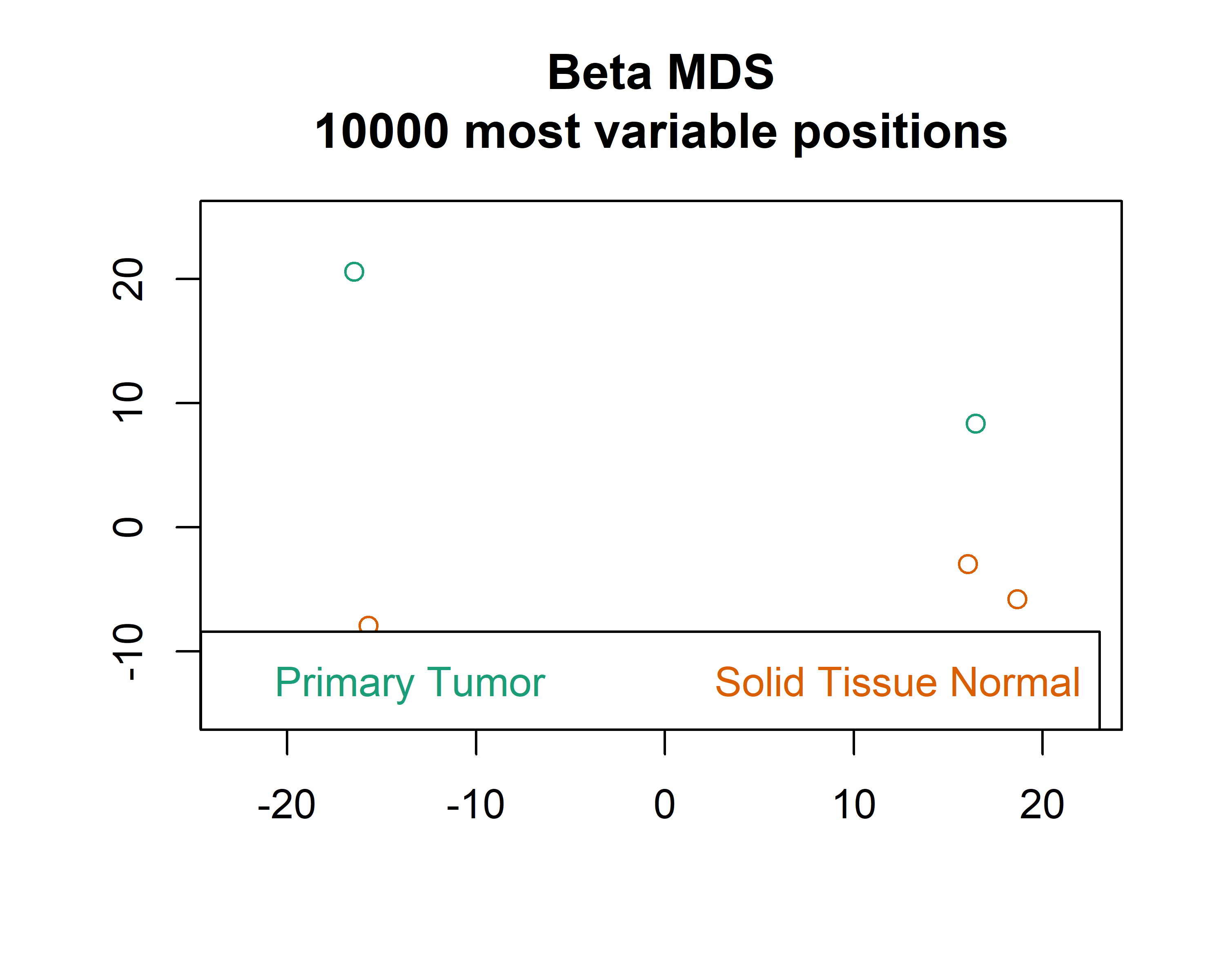


#### Raw data

mdsPlot(mset,sampGroups = stype)

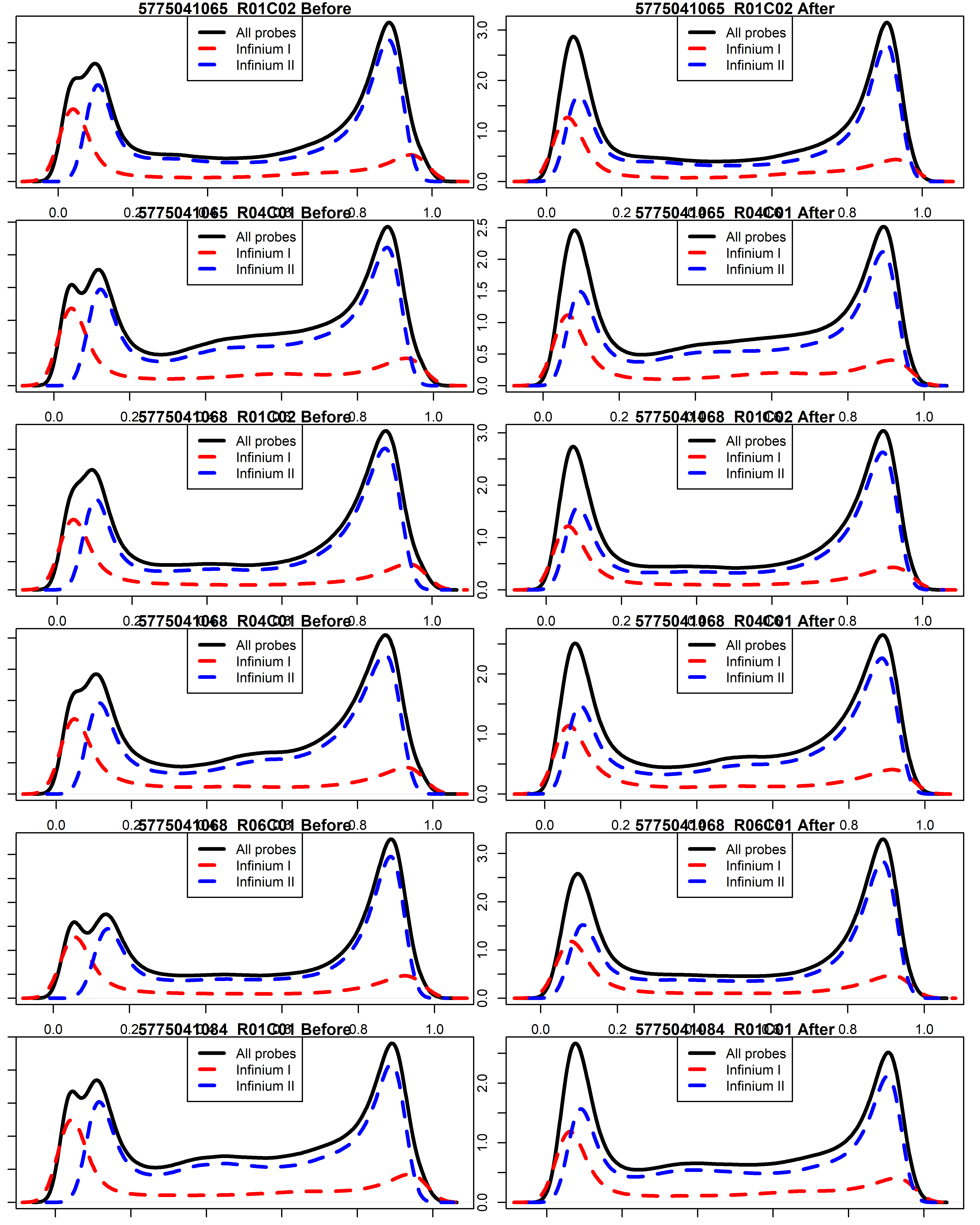


mdsPlot(mset,numPositions = 10000,sampGroups = stype)



## 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 Differentially 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

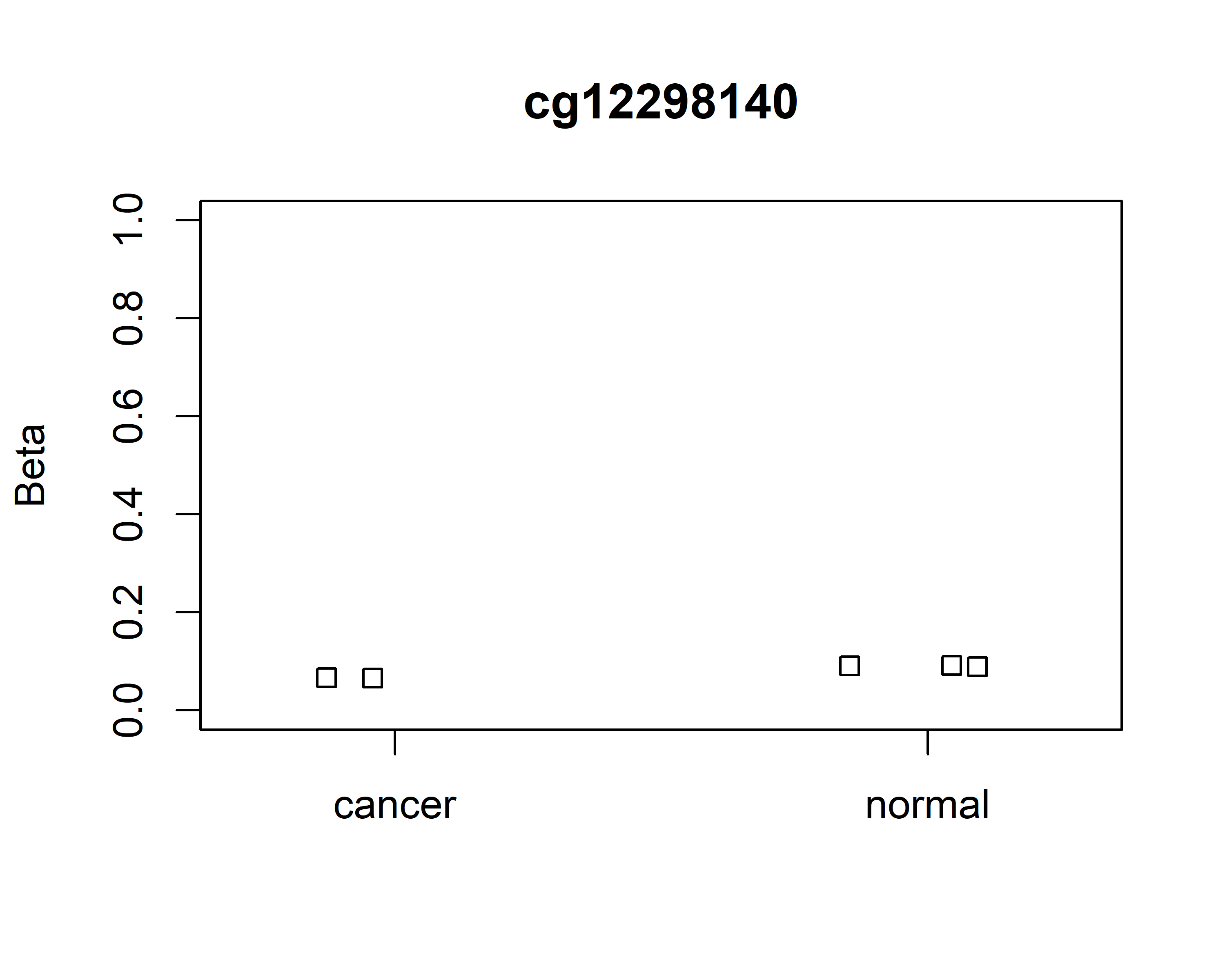
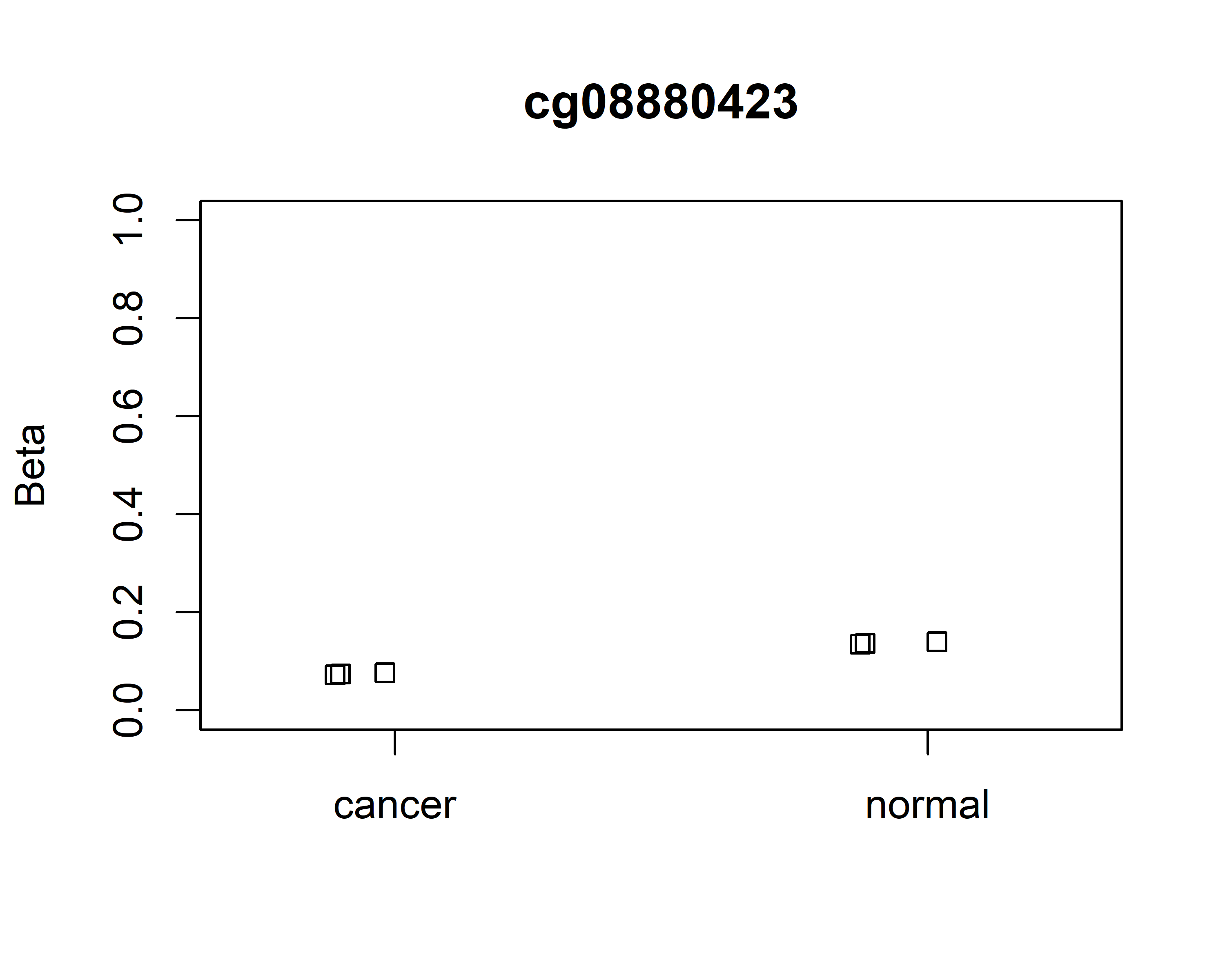
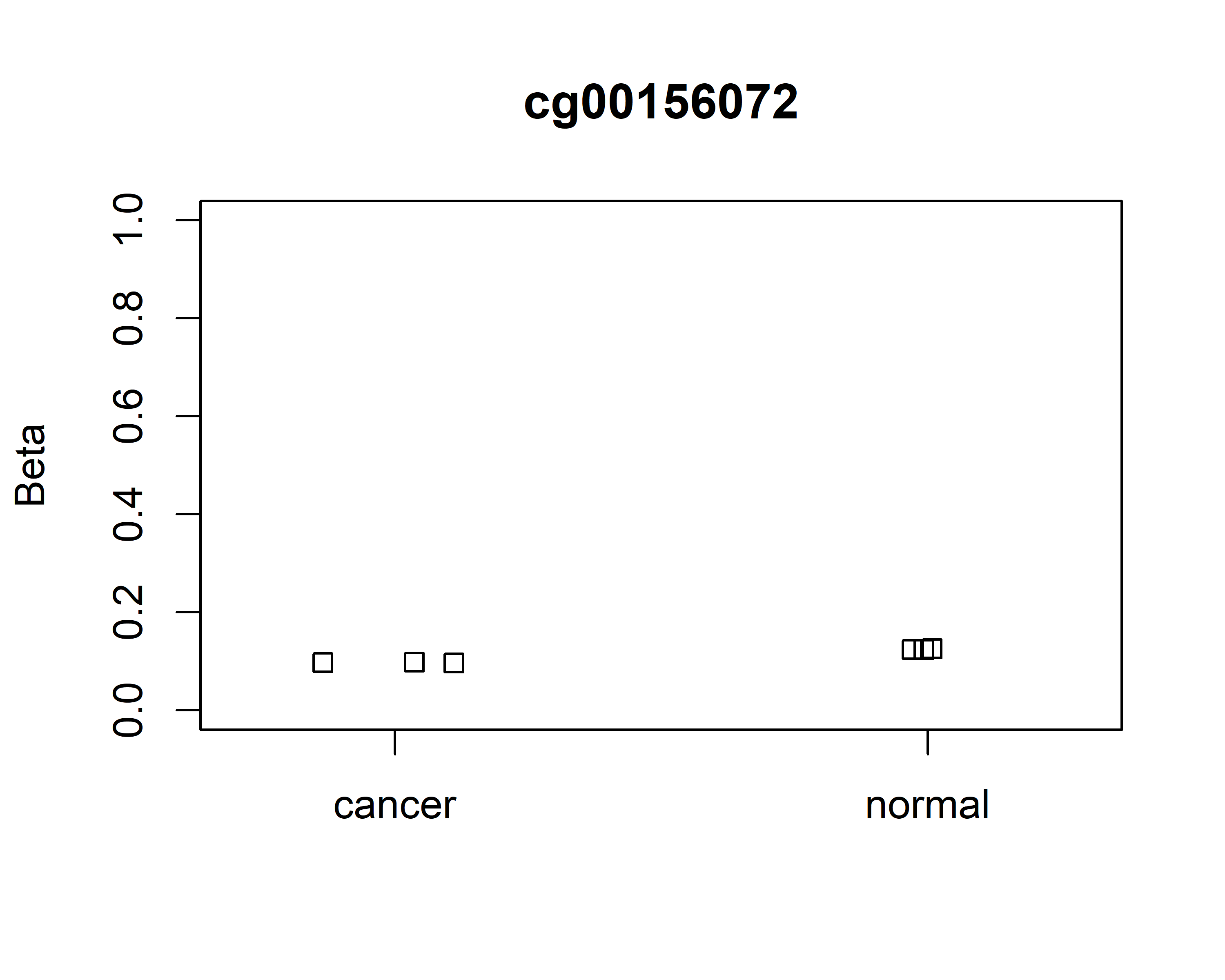
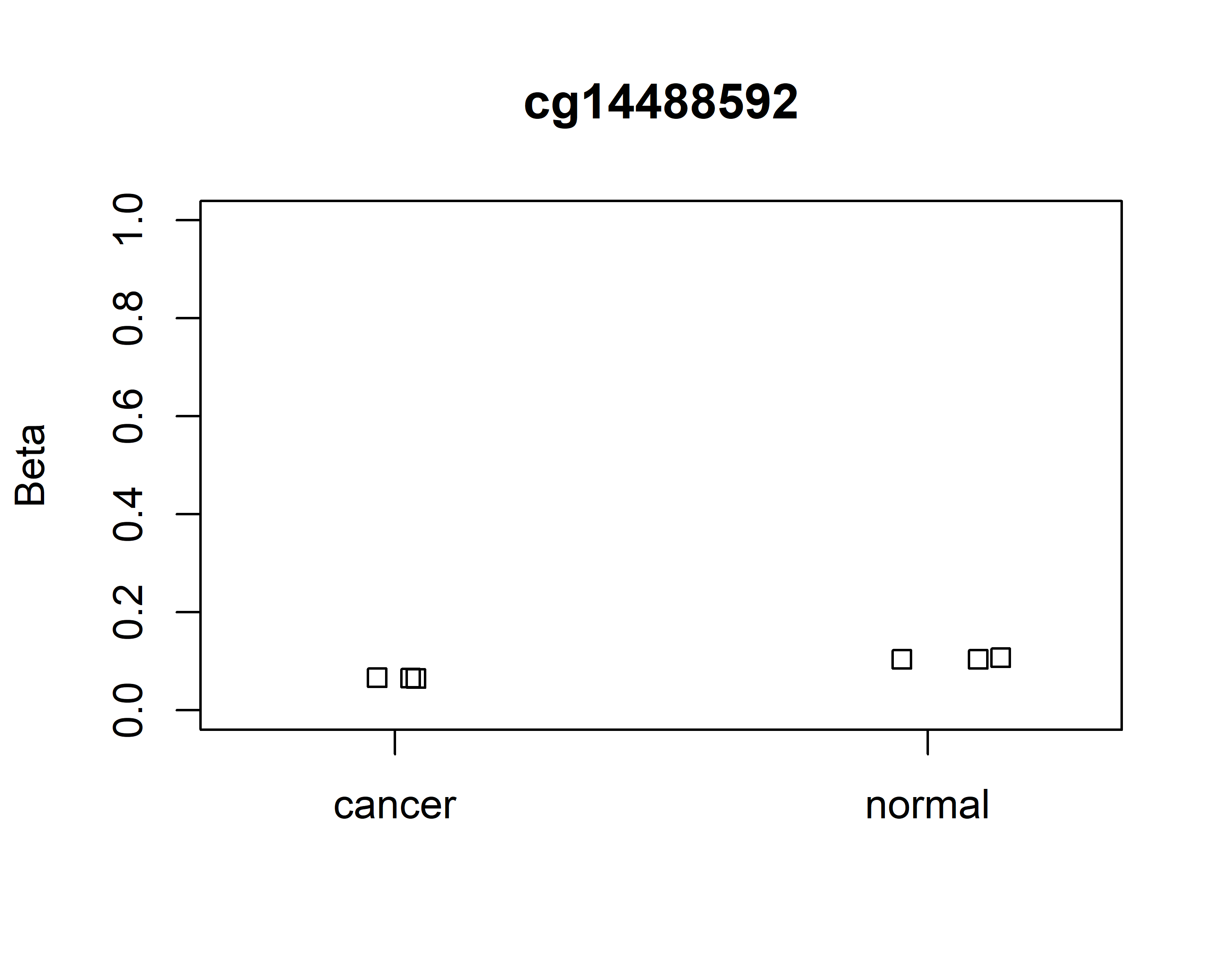
m = getM(msetSWAN)  
pheno = pData(rgSet)$Status  
dmp = dmpFinder(m,pheno = pheno,type = "categorical")  
dmp %>% round(.,3) %>% rownames\_to\_column(var = "CpG") %>%   
 head(.,10) %>% flextable(.)

| CpG | intercept | f | pval | qval |
| --- | --- | --- | --- | --- |
| cg14488592 | -3.824 | 1598.491 | 0 | 0.351 |
| cg00156072 | -3.213 | 1271.908 | 0 | 0.351 |
| cg08880423 | -3.637 | 1217.395 | 0 | 0.351 |
| cg12298140 | -3.824 | 1174.856 | 0 | 0.351 |
| cg05592581 | 1.731 | 1068.850 | 0 | 0.351 |
| cg13163765 | -3.356 | 997.655 | 0 | 0.351 |
| cg21046068 | -3.798 | 926.855 | 0 | 0.351 |
| cg03150279 | -3.712 | 806.109 | 0 | 0.406 |
| cg05239513 | -4.035 | 605.893 | 0 | 0.432 |
| cg17480467 | -4.221 | 527.857 | 0 | 0.432 |

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 8 CpG sites significant at the p level, 1 was hypermethylated in cancer samples and 7 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

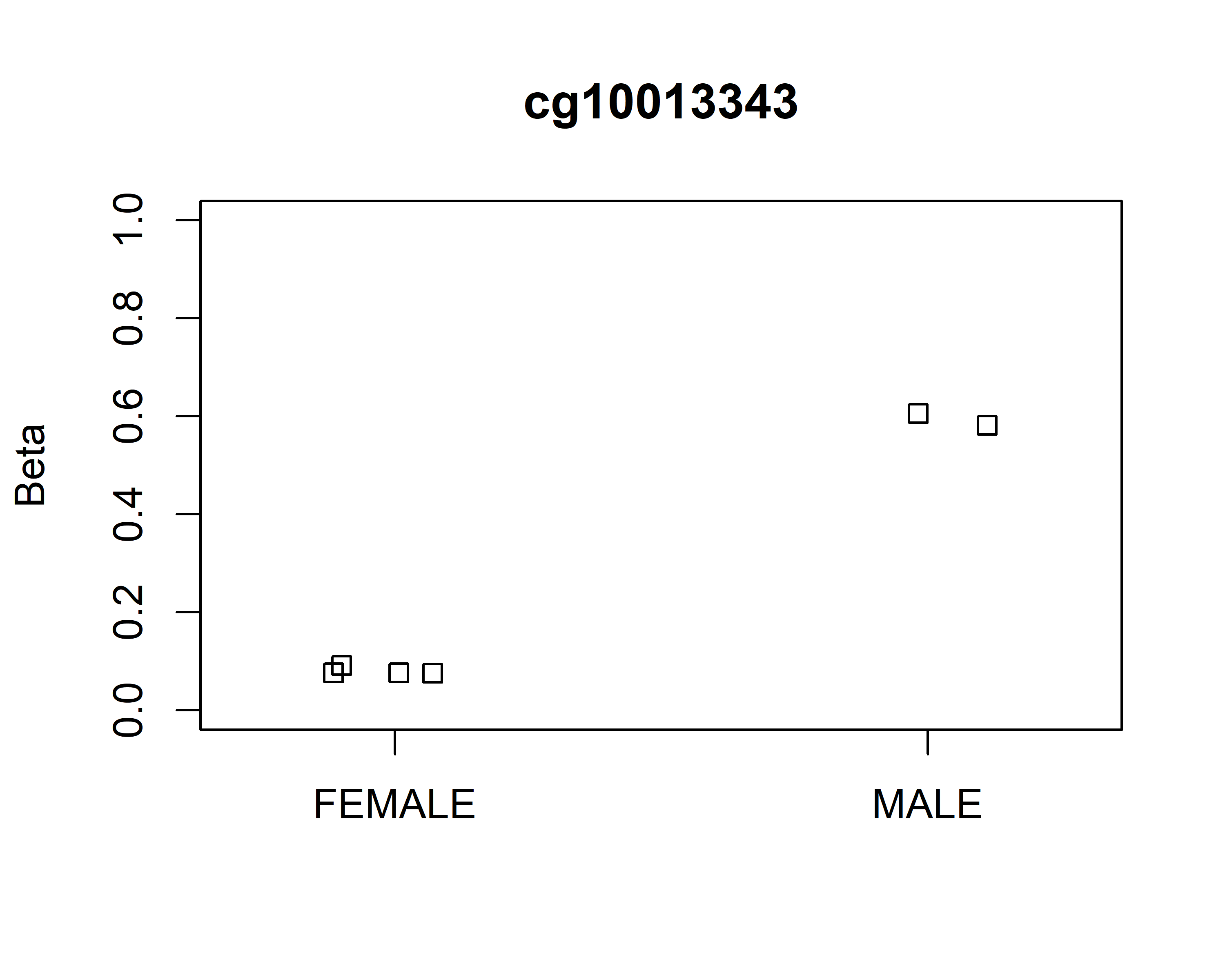
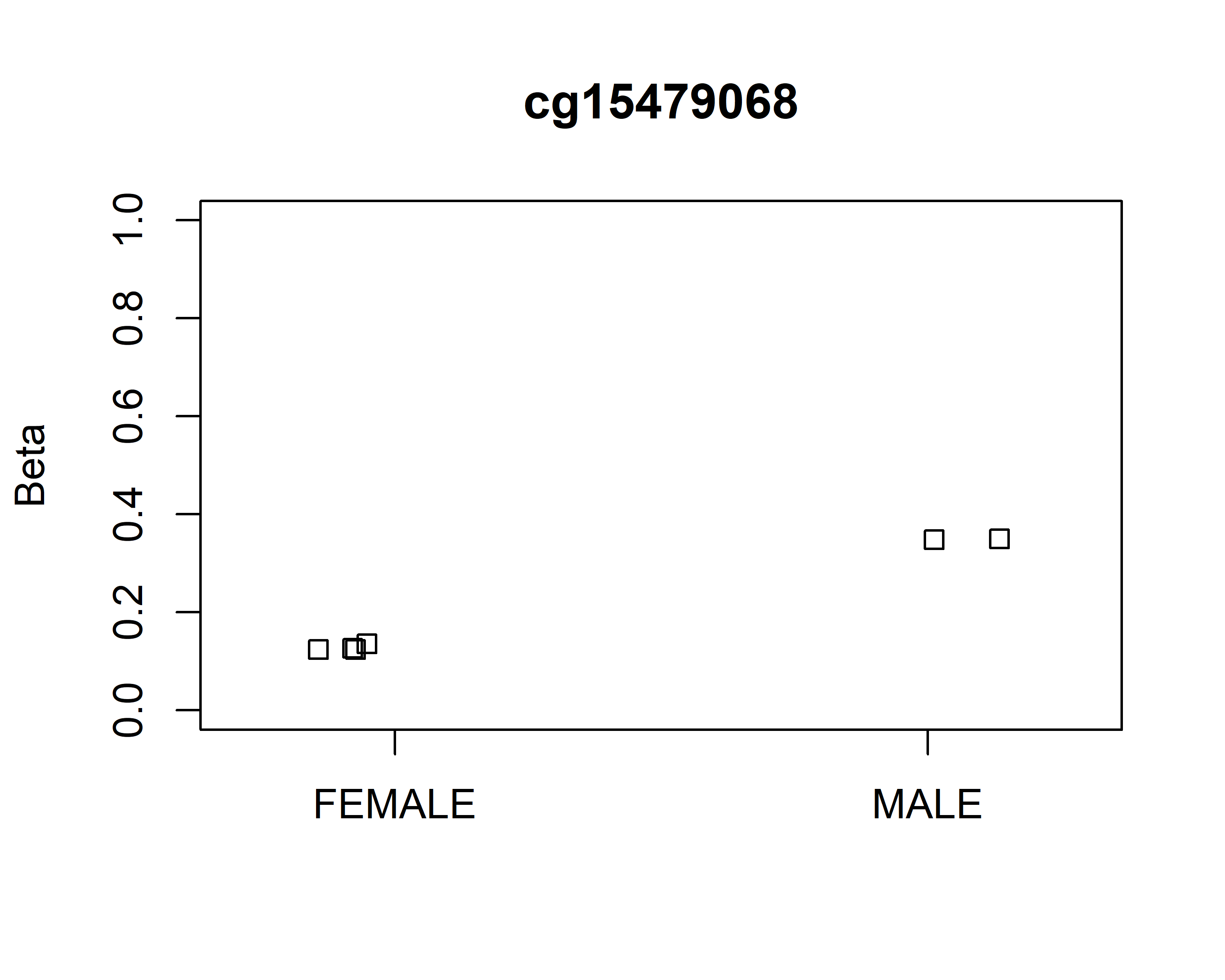
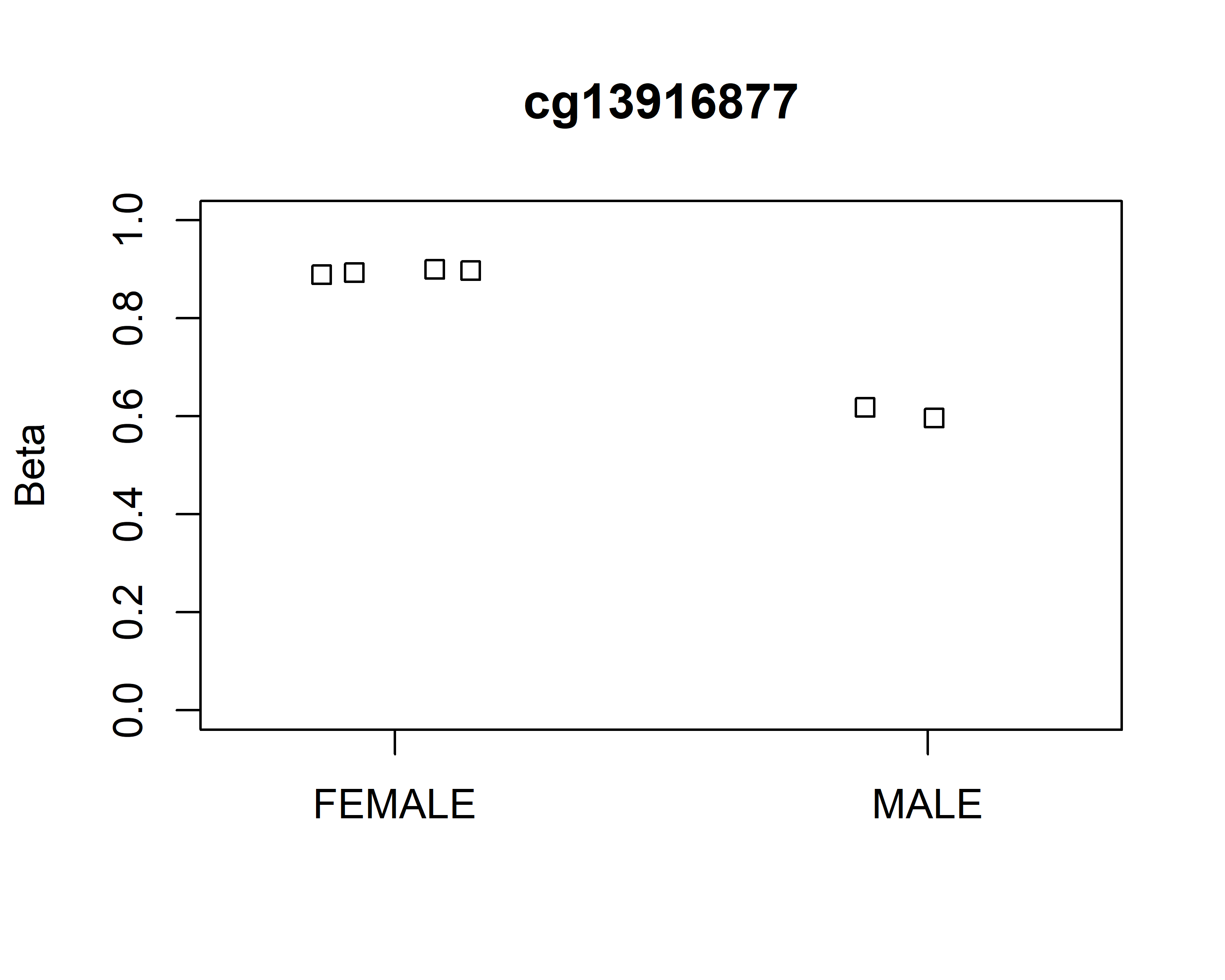
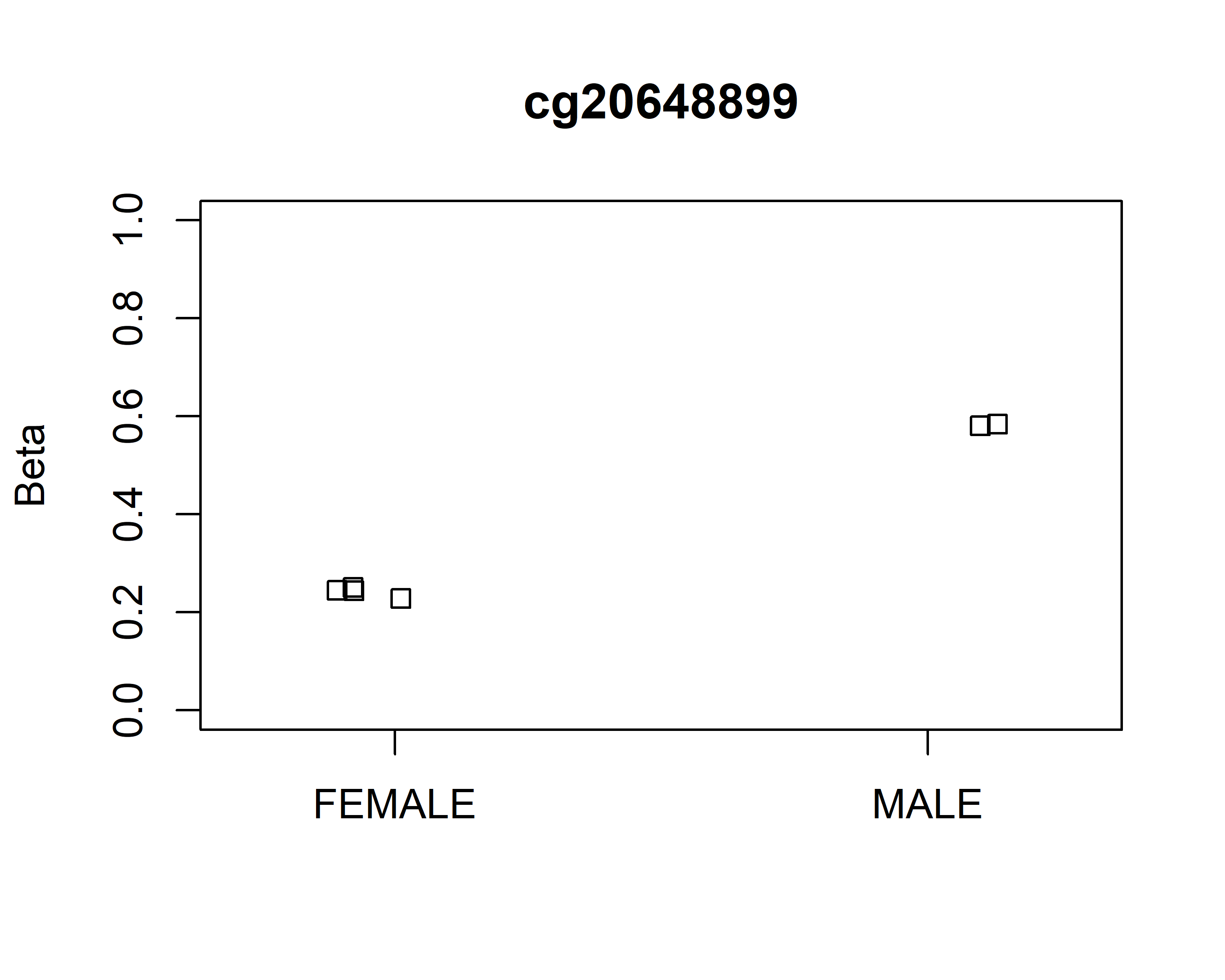
pheno = pData(rgSet)$Sex  
dmp = dmpFinder(m,pheno = pheno,type = "categorical")  
dmp %>% round(.,3) %>% rownames\_to\_column(var = "CpG") %>%   
 head(.,10) %>% flextable(.)

| CpG | intercept | f | pval | qval |
| --- | --- | --- | --- | --- |
| cg20648899 | -1.648 | 1297.239 | 0 | 0.234 |
| cg13916877 | 3.090 | 1290.182 | 0 | 0.234 |
| cg15479068 | -2.771 | 1223.905 | 0 | 0.234 |
| cg10013343 | -3.526 | 1223.488 | 0 | 0.234 |
| cg11403874 | -3.277 | 1143.907 | 0 | 0.234 |
| cg17220960 | 2.654 | 1124.061 | 0 | 0.234 |
| cg15308664 | -3.552 | 1080.010 | 0 | 0.234 |
| cg10019095 | -3.735 | 1032.094 | 0 | 0.234 |
| cg13764106 | -3.208 | 1023.031 | 0 | 0.234 |
| cg00393263 | 2.367 | 933.418 | 0 | 0.234 |

There are 0 DMPs with q-value . Of the 16 CpG sites significant at the p level, 6 was hypermethylated in females and 10 were hypomethylated.

### Plot CpGs

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



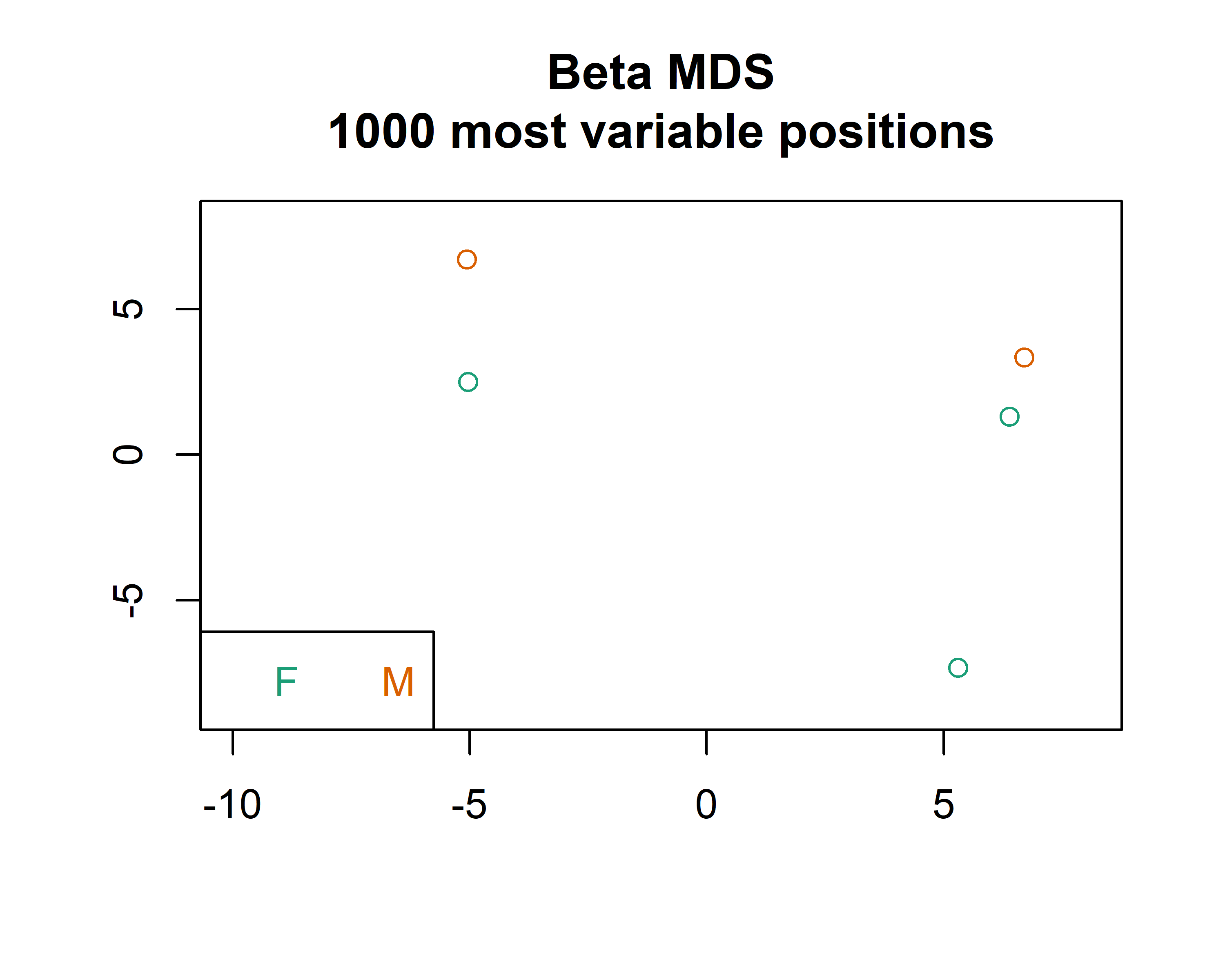
## d) Estimate whether each sample is male or female

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)

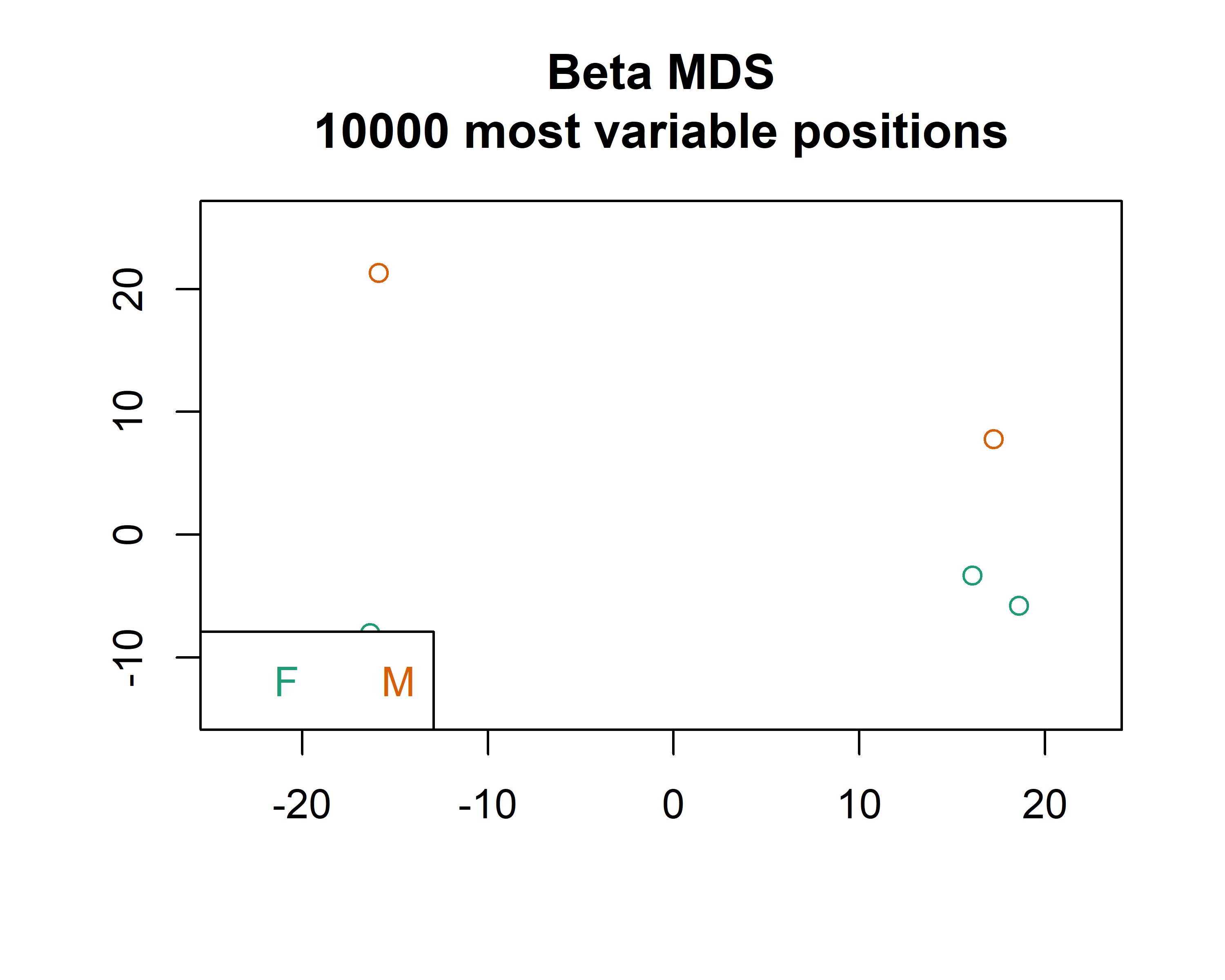
| Predicted Sex | Label |
| --- | --- |
| M | MALE |
| F | MALE |
| M | FEMALE |
| F | FEMALE |
| F | FEMALE |
| F | FEMALE |

### 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 | 3.117 | 6346.706 | 0 | 0.022 |
| cg16788857 | -3.091 | 4766.888 | 0 | 0.022 |
| cg07889003 | 2.409 | 4415.505 | 0 | 0.022 |
| cg21130926 | 4.838 | 4382.665 | 0 | 0.022 |
| cg04003990 | 2.661 | 3999.184 | 0 | 0.022 |
| cg12077433 | 2.332 | 3924.840 | 0 | 0.022 |
| cg12186981 | 3.697 | 3699.892 | 0 | 0.022 |
| cg08250118 | 2.650 | 3497.971 | 0 | 0.022 |
| cg16523141 | 3.003 | 2951.465 | 0 | 0.025 |
| cg12165338 | 0.028 | 2906.530 | 0 | 0.025 |

After correcting the sex labels, there are 54 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")

