Analysis of DNA methylation data

Moumita Karmakar

In the following, I provide a brief description of the analysis of the Illumina Infinium methylation microarray data.

Step 1: Reading data into R

`minfi` package is used to read the raw data files in `.idat` format. Sample numbers '879', '4667', '5505', '7725', '3F05_1_2', '3F05_10_2', 'Leuk_ZC', 'Leuk_FH', '3D03_10_2' are discarded to perform the baseline analysis. We have 32 samples for the baseline. The three important covariates of interest are i) history of Traumatic brain injury (TBI) ii) age and iii) lifetime breaching history.

Step 2: Preliminary analysis on the data

Before any detailed analysis, basic preliminary analysis should be performed to get a better idea about the data. Here, we are mainly interested to know how methylation pattern changes with respect to different categories of lifetime breaching history of 32 samples.

Breaches plot

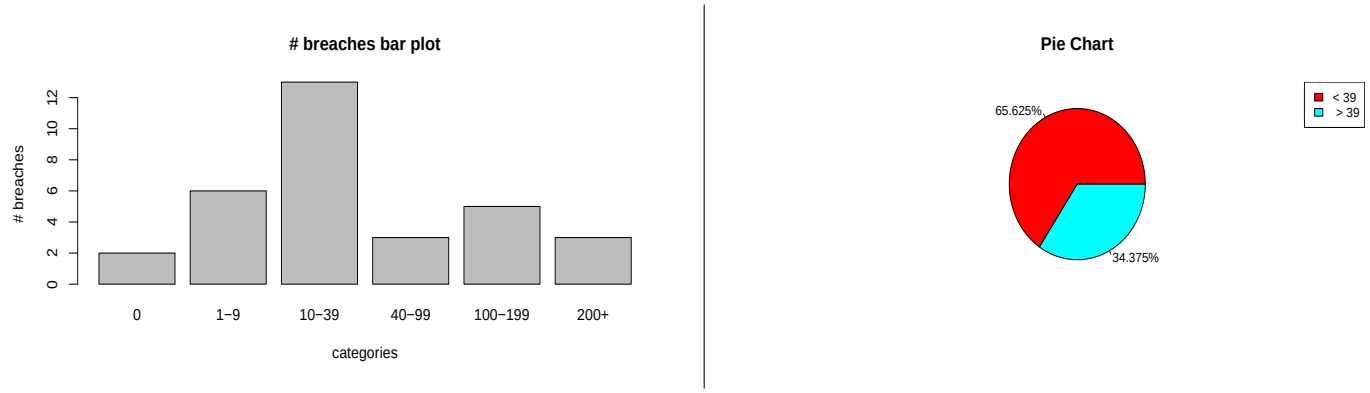


Figure 1: Bar plot for categories of lifetime breaching history

To have a knowledge about the distribution of lifetime breaching history we plotted the bar plot for categories of lifetime breaching history in Figure 1. From the above bar plot and pie chart, it is evident that more than 50% of the samples were exposed to less than 39 breaches. Based on that observation, I decided to divide the lifetime breaching history into two categories 'low' (less than 39) and 'high' (greater than 39).

Lifetime breaching history & lifetime history of mild TBI

In the following bar plot in Figure 2, we want to study the relation between lifetime breaching history and lifetime history of mild TBI. Lifetime history of mild TBI has two categories 'yes' and 'no'. The plot in

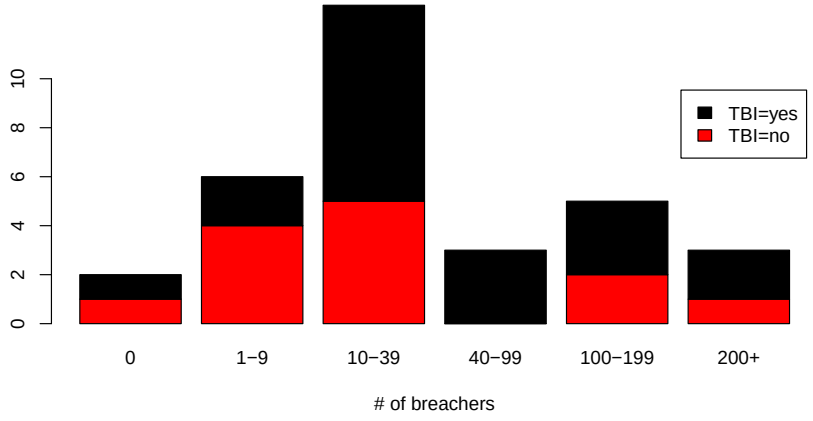


Figure 2: Divided bar plot showing proportion of samples having history of TBI in each breacher category

Figure 2 shows us the proportion of samples having history of TBI within each breacher category. It is important to note that for Breacher category '40-99', all samples have TBI.

Dendogram for sample mixup plot

We have pre and post blast data, denoted by '1' and '10' respectively. We plot the dendogram using the `hclust` function in `R` and the `manhattan` distance. Few duplicate samples (3F05_10_1', 3F05_10_2', '3D03_10_2') are identified from the above dendogram.

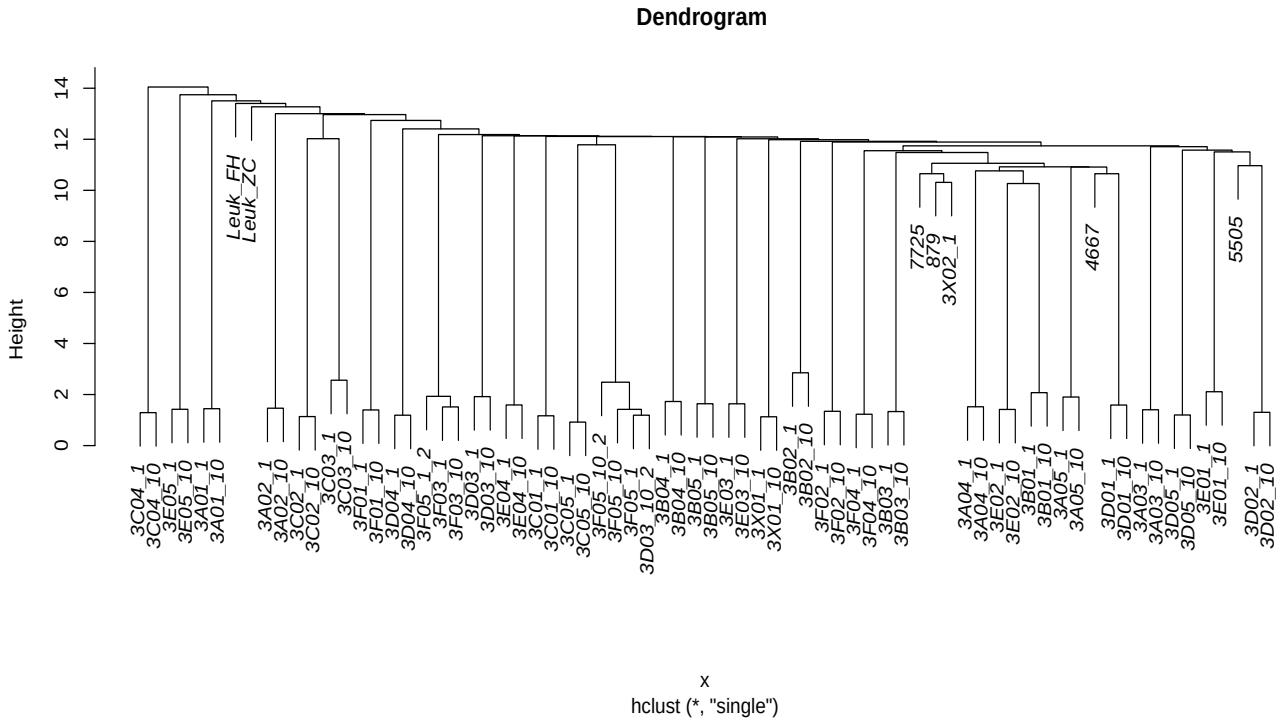


Figure 3: Dendogram for sample mixup plot

Sex plot

Our data have only one female sample. The following sex plot in Figure 4 based on median beta values of XY chromosomes conforms with data.

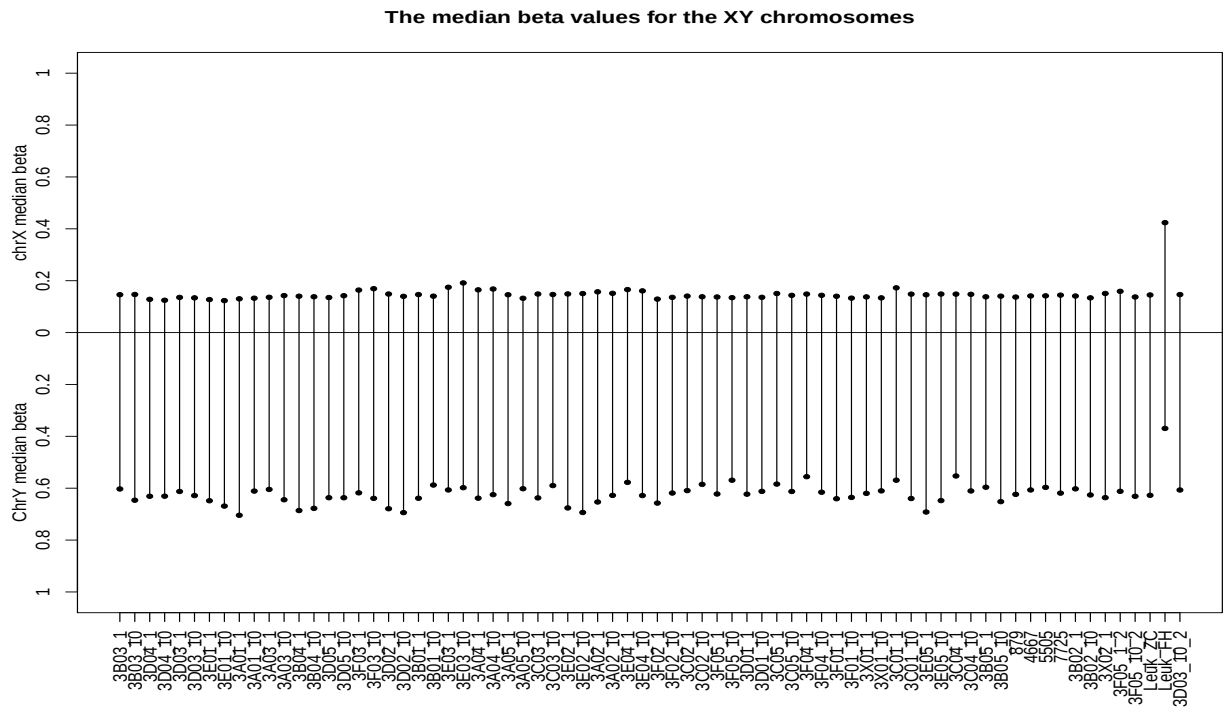


Figure 4: Sex plot

Data preprocessing or normalization

Preprocessing and normalization is also done using the `R` package `minfi`.

1. `RGchannelset` is converted to `methylset` using preprocessRaw function. No normalization is done in this step.

2. Background correction and control normalization is done using `preprocessIllumina` function.

3. We use `preprocessSWAN` to perform Subset-quantile within array normalization.

The swan plot in Figure 5 shows that the swan normalization reduces the within array technical differences between Infinium I and II type designs. Poor quality probes are then filtered out using detection 'p-value' less than 0.01. Probes that contain a SNP at the CpG interrogation are also discarded following (Naeem, Wong et al. 2014). I started the analysis with a total of 485512 CpG cites. After preprocessing and normalization, only 40135 CpG cites are retained.

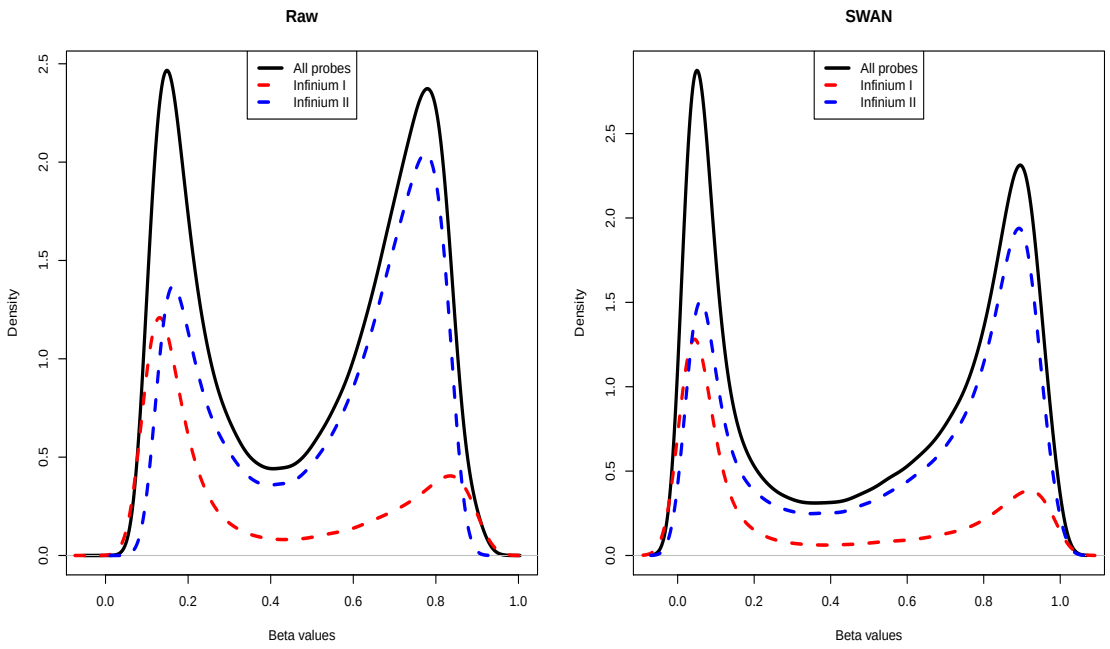


Figure 5: Swan plot

Data Analysis for baseline only

The rest of the data analysis is done using the `limma` package. A linear model is fitted using the three covariates: age, lifetime history of TBI & lifetime breaching history. We use `lmFit` and `eBayes` function in the `limma` package to identify differentially expressed methylated CpG positions with respect to breaching history and history of TBI.

In Figure 6, we plotted the methylation mean difference (based on M-value) on the horizontal axis and the number of differentially methylated CpG cites on the vertical axis. The red color represents a gain in methylation and blue represents a loss. From the plot it is clear that there is gain in methylation in high breaching compared to low breaching the for mean difference range $0.2 - 0.37$.

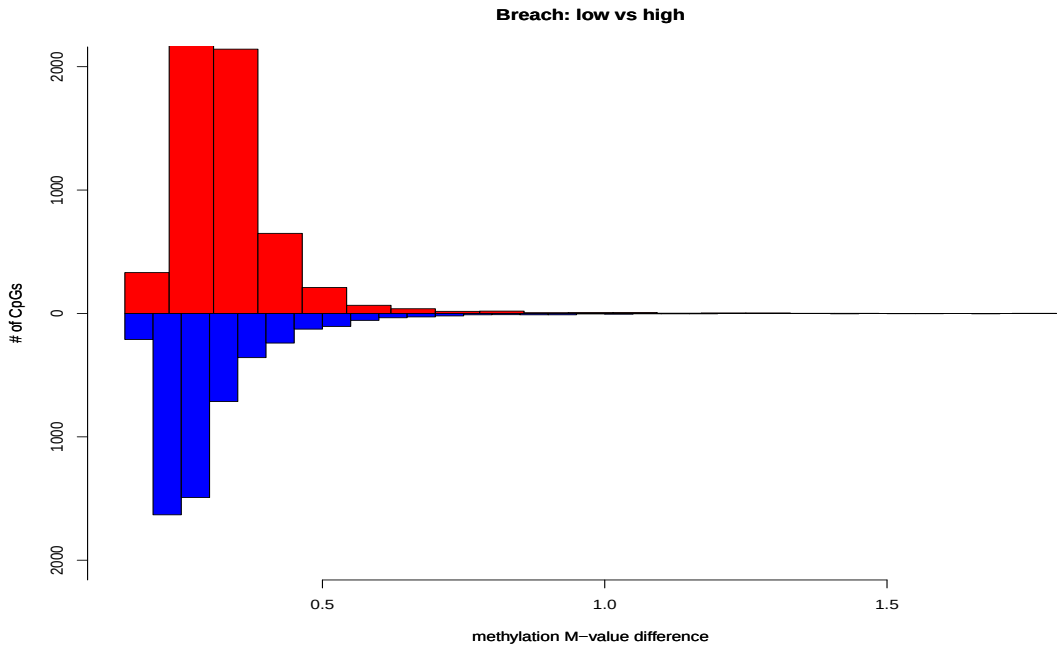


Figure 6: Mirrored histogram plot for breaching history

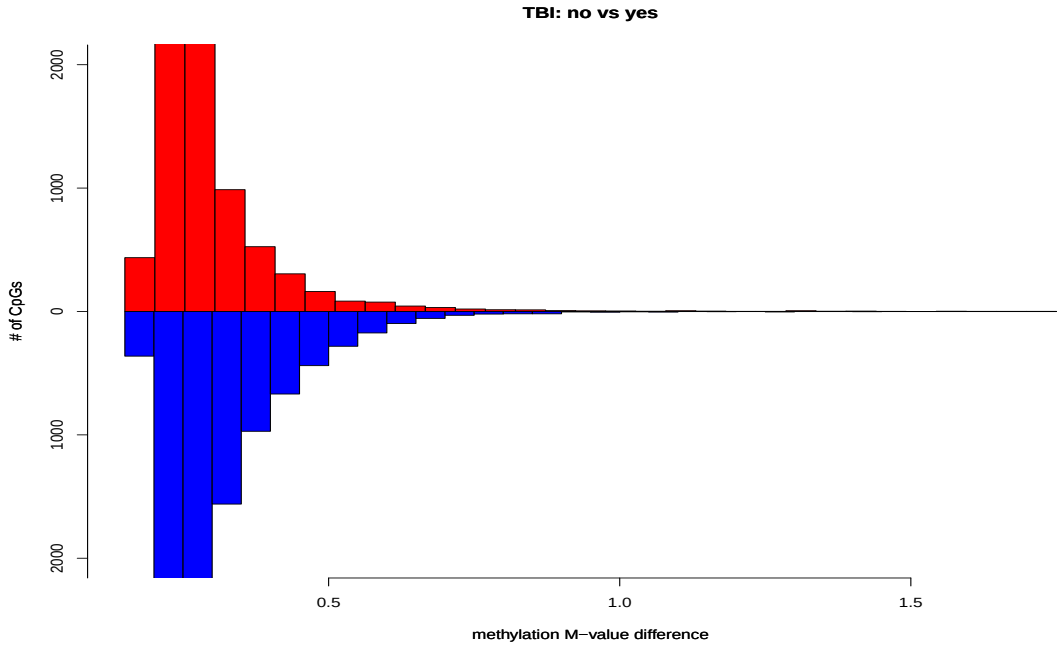


Figure 7: Mirrored histogram plot for history of mild TBI

Figure 7 is the mirrored histogram plot for history of mild TBI. The plot does not show any significant difference in the methylation level between two groups 'yes' or 'no'. Differential methylation test for each CpG position is performed both for subjective 'p-value = 0.05' and 'adj p-value (BH)'. Due to a huge number

of individual hypothesis tests and poor quality of predictor response signal, BH procedure fails to capture the significant CpG sites. Subjective 'p-value' is able to detect 14075 significant CpGs.

After the CpGs are selected, we use the `getAnnotation` function in `minfi` to acquire all annotation information related to these CpG cites. The bar plot in Figure 8 shows the proportion of CpGs belonging to different categories of Island. Most CpGs are in the 'Opensea' region. 'Island' is the second highest category of Island.

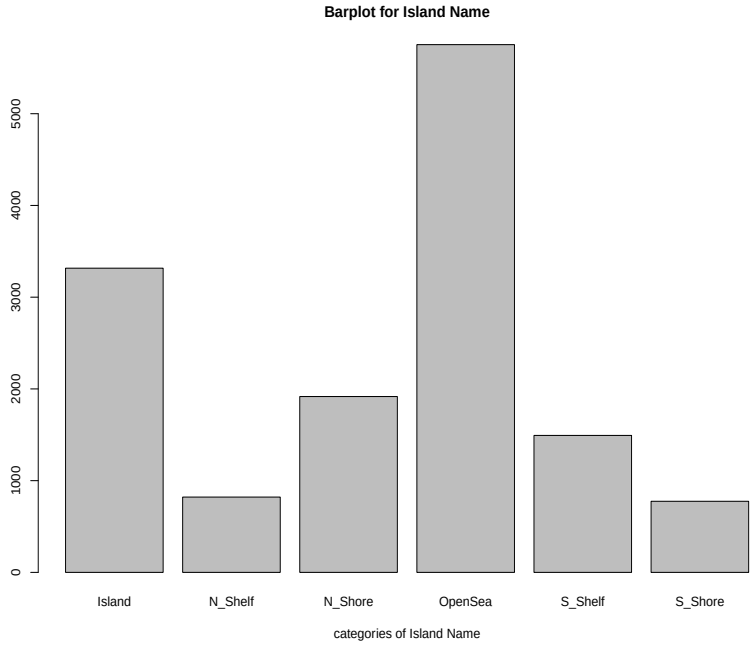


Figure 8: Bar plot showing proportion of CpGs belonging to different categories of Island

The following bar plot in Figure 9 shows the proportion of different regions 'promoter', 'Body', '1stExon', and '5'UTR' of gene. 'Body' has the maximum frequency and promoter or 'TSS' region has the second maximum count.

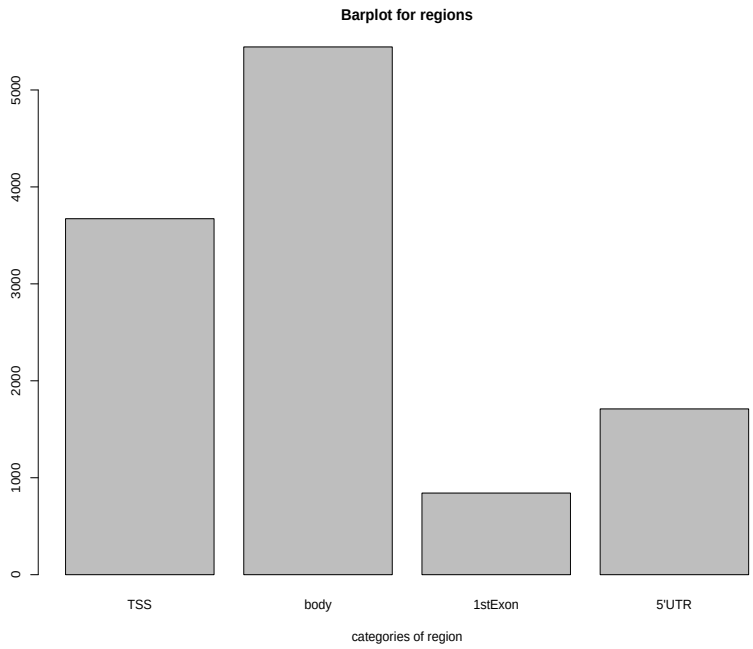


Figure 9: Proportion of different regions ‘promoter’, ‘Body’, ‘1stExon’, and ‘5’UTR’ of gene

To detect any methylation rate difference between ‘low’ and ‘high’ breaching group in terms of beta values, we created box plot for each category in Figure 10. The ‘low’ breaching group has higher within group variability compared to ‘high’ breaching group.

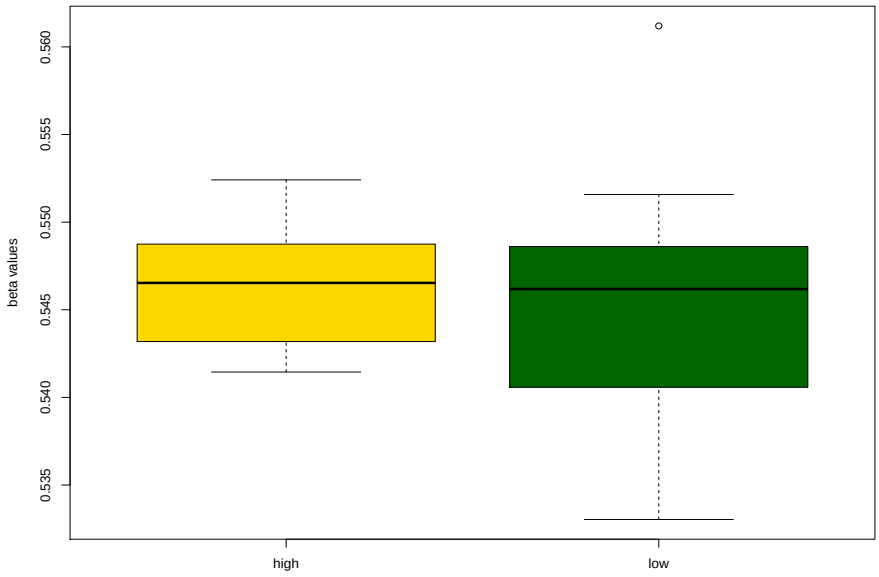
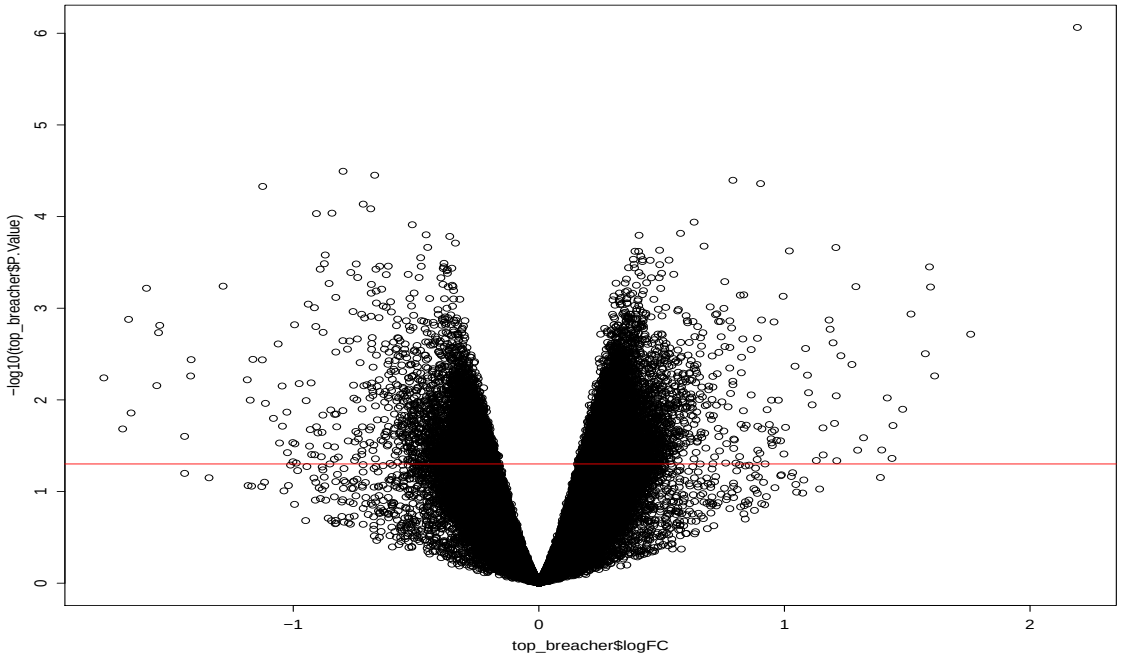


Figure 10: Boxplot of beta values for ‘low’ and ‘high’ breaching group


−log10(top_breacher$P.Value)
0
1
2
3
4
5
6
−1
0
1
2
top_breacher$logFC