**EPIGENETICS: MODULE 4**

**Taniya Pal (s373094)**

**Introduction**

Illumina Infinium Methylation Arrays provides the methylation data for CpG islands at a single base level, currently available for human and mouse only. It has two types of probes: Type I and Type II. Type I probe consists of two arrays, each giving red or green signal for methylation and unmethylation. Type II probe has only one array giving red signal for methylation and green signal for unmethylation During quality control usually for type I probe the red and green colors are of less significance than the bright or dark intensities of the signals. Along with the sample probes there are two types of control probes: 1) Negative control probes for detecting low signal intensities 2) SNP control probes for detecting sample duplication.

The initial steps of the preprocessing of data includes reading the red or green channel signal intensities stored in .idat files into numerical values, which are stored in a variable called RGset. The RGset is then converted to Mset which has two types of value: 1) β value 2) M value

The β value is calculated by the following formula:

M/M+U+α where, M=Methylation values,

U=Unmethylated values,

α=offset value which is included to prevent

the value of the denominator to become zero.

The β value ranges from 0 to 1.

Whereas the M value is given by the formula:

log(M/U) where, the M value ranges from -5<0<+5

Biologists usually prefer to use β value than M value as it is a indicator of how much methylated the sample is and is simple to interpret. Whereas statisticians prefer M value as it has a statistical background. M value may also be expressed as logit value of β.

**Quality Control**

Quality control is done for both the samples and the probes.

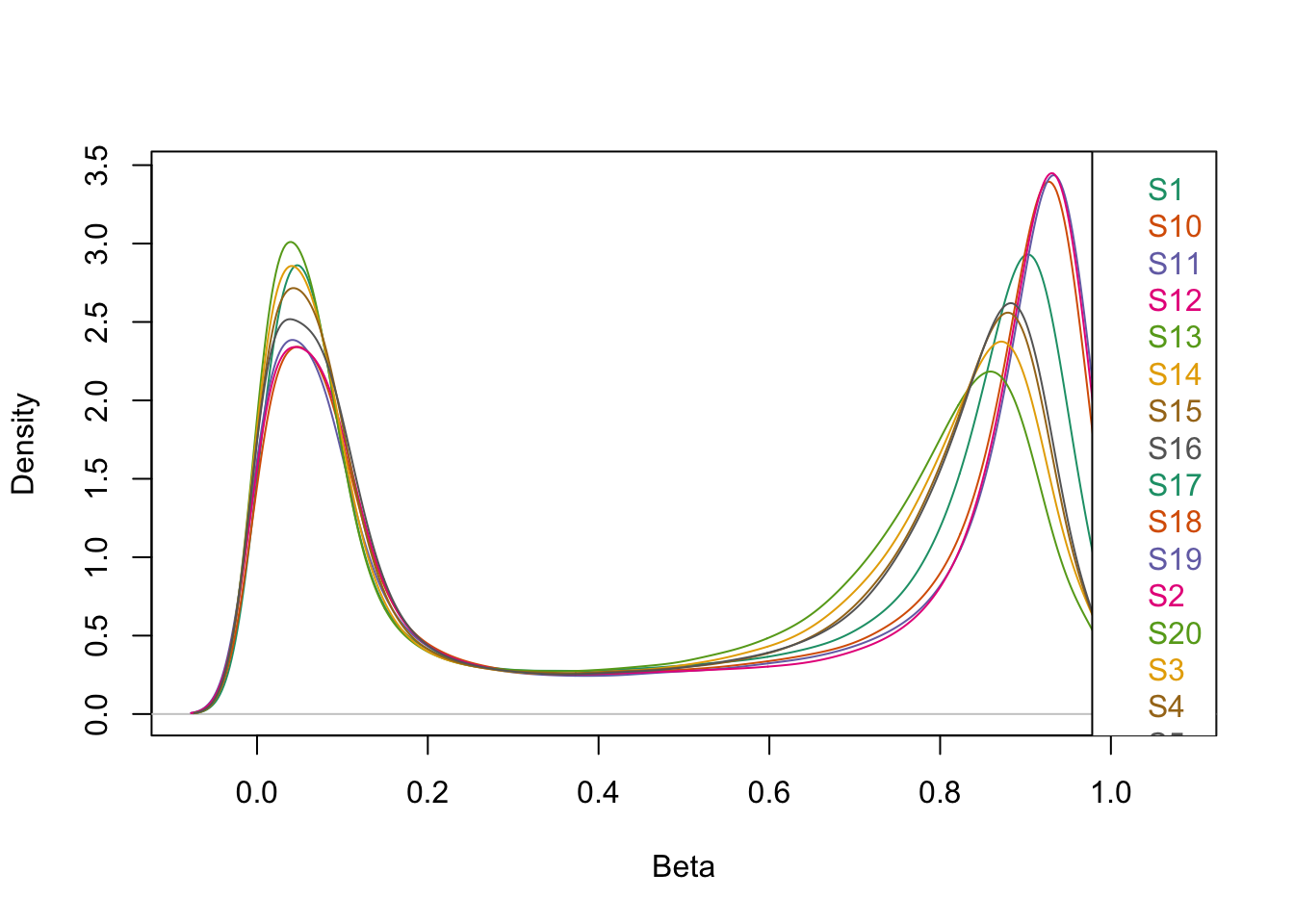
For quality control of probes,

* Depending on the background and negative control probes, the signal intensities may be high or low, hence a p value is assigned to the probes to declare it significant or non-significant.
* Measuring the significance, we drop the probes which fail in more than 10% of the samples.
* For common SNPs in the probes we need to decide whether to discard or keep them, this is decided based on the location of the SNP, if it is present exactly on the CpG site, they are kept as it has a significant effect on the results, otherwise they are discarded. But, at the same time, we have to be careful to not mix the SNPs in the sample probes to the SNPs in the control probes, as the latter is only to identify duplicate samples.
* The X and Y chromosome probes are excluded as they require special statistical treatment which the different R packages doesn’t take into consideration.
* The extreme methylation values are to be excluded both 0% and 100% as they are most likely due to technical artefacts.

For quality control of samples,

* The samples consisting of more than 50% failed probes are to be dropped.
* The gender of the samples should be checked, to eliminate any mixing up of the same.
* SNP control probes are included in the methylation analysis to prevent any sample duplication- if two samples have the same SNPs which is practically not possible, it indicates sample duplication.

On the given data, the quality control of the samples were done, and the following results were obtained:

Density plot between β values and sample names:

We can see here the density plot is bimodal, the β values lying in two extremes, representing the values are either methylated or unmethylated. The different alignments of the peaks of different samples for methylated and non-methylated β values can be rectified through normalization.

Density bean plot between β values and sample names:

Chart

Description automatically generatedThe density bean plot can be considered an alternative to box plot. The two ends of the bean plot indicate the distribution of the β values is bimodal, representing either methylation or unmethylation. Also, duplicate samples can also be spotted in the plot. The one dimensional scatter plot consists of one line for each observation in a batch. When there are multiple observation in a same batch, it increases the length of the line, hence making duplicate samples easy identifiable. In this plot, we can see such lines, thus we have duplicate samples in our sample batches.

**Multidimensional scaling plot (MDS plot) of β values**

Chart, scatter chart

Description automatically generated

The multidimensional scaling plot can be considered an alternative to PCA plot. In this plot, we can see the batch effect between the samples as few samples are grouped together at specific positions in the plot. This can be removed through batch correction.

**Quality Control plot of the samples**

Chart, scatter chart

Description automatically generated

We can see all the samples have passed the quality control, hence we can infer there are no sample with low quality.

**Inferring the sex of the samples**

There was no mismatch found between the predicted sex and the sex provided in the sample description file. Hence, we can infer that there is no discrepancy with respect to the sex of the samples.

Sample Duplication

A picture containing qr code

Description automatically generatedTo check the duplication of samples, the SNPs from the control probe was first extracted and a heatmap was plotted with the correlation values of the extracted SNPs.

From the heatmap, it was inferred that sample pairs (1,9) and (17,18) were duplicate samples. Hence further processing was done to remove one sample from each pair. Sample 1 from the first pair and sample 17 from the second pair were removed for this reason.

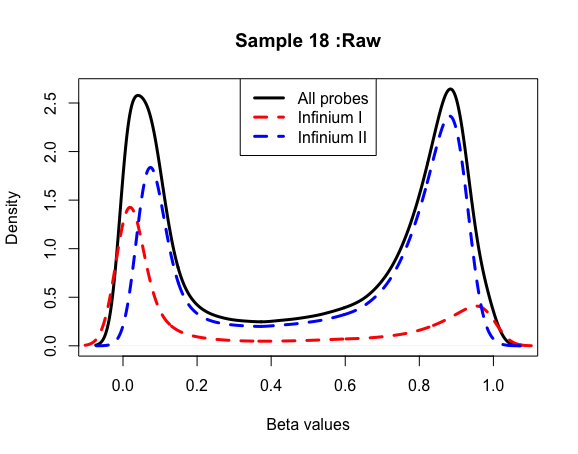
**Filtering of samples and probes**

The p value was updated after the removal of sample duplication as the number of samples decreased. The threshold for p value for the quality control of the samples was set to 10^(-16) in atleast 50% of the probes. And for the quality control of the probes, the threshold of p value was set to 10^(-16) for atleast 90% of the samples.

**Normalization**

Normalization involves removal of experimental errors and noises, which if not removed can mask significant differences between the samples. Type I probe has more CpG density than type II probe which has the potential to introduce bias during any statistical analysis. SWAN (Subset-quantile Within Array Normalization) uses a subset of probes which are biologically similar with respect to CpG content. Both samples and probes were normalized with the help of SWAN. It reduces the differences in β value distribution between Infinium Type I and II probes. According to Gordon et al, it also increases the correlation between technical duplicates, thus increasing the detection of number of differentially methylated probes.

**Before normalization the density plot of the β distribution between probes was as follows:**



**After normalization, the density plot appeared as follows:**

Chart, histogram

Description automatically generatedWe can see that the differences between the β distribution of type I and type II probes have been reduced.

**Between samples:**

Before normalization

Chart, histogram

Description automatically generated

After normalization

Chart, histogram

Description automatically generated

**Batch Correction:**

Varying conditions during wet lab experiments such as different laboratory conditions or experiment time can result in batch effect being incorporated in our results. These batch effects are nonlinear which makes it difficult to align the datasets while preserving the variations which are significant biologically. Normalization removes this to some extent but is not successful in completely removing it. “ComBat” package in R uses empirical Bayes method to remove the same which adjusts any outliers or any potential confounder other than batch. In order to visualize the batch effect, PCA plot before and after batch correction was plotted.

Before batch correction, the PCA plot was visualized as follows:

Chart, scatter chart

Description automatically generated

We can see here; the batch effect was observed between the different types of IIlumina arrays: 450k and EPIC. This was successfully removed by using the package “ComBat” in R.

After removal of batch effect:

Chart, scatter chart

Description automatically generated

We can see, after batch correction, the batch effect was removed.

**Estimating cell type proportion**

A picture containing background pattern

Description automatically generatedThe proportion of methylation between different blood cells was explored.

We can see CD8T and Neutrophil are unmethylated with respect to each other as well as Neutrophil and CD4T, B cell and Neutrophil.

Thereafter, regression analysis was done on the β values obtained from the genotype set. It was done to explore the relationship between the methylation values and BMI and other covariates such as sex and types of blood cells.

The significant CpGs were extracted based on a threshold of p value of <0.05. A volcano plot was made further highlighting the significant and non significant genes.

Enrichment analysis was also done on the significant CpGs based on the threshold p value in Gene Ontology and Kyoto Encyclopedia of Gnes and Genomes disease pathways. The results were stored in the files “Gene Ontology Analysis” and “KEGG analysis” respectively.

References:

1. Wang, Z., Wu, X. and Wang, Y., 2018. A framework for analyzing DNA methylation data from Illumina Infinium HumanMethylation450 BeadChip. *BMC bioinformatics*, *19*(5), pp.15-22.
2. Pidsley, R., Wong, C.C., Volta, M., Lunnon, K., Mill, J. and Schalkwyk, L.C., 2013. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC genomics*, *14*(1), pp.1-10.
3. Wilhelm-Benartzi, C.S., Koestler, D.C., Karagas, M.R., Flanagan, J.M., Christensen, B.C., Kelsey, K.T., Marsit, C.J., Houseman, E.A. and Brown, R., 2013. Review of processing and analysis methods for DNA methylation array data. *British journal of cancer*, *109*(6), pp.1394-1402.