Third task

Akvilė Višniauskaitė ir Ieva Pudžiuvelytė

5/6/2022

Introduction

This work will present data analysis from a research that studied the enchancer's at *IGF2* differential methylation association with abnormal dopamine synthesis in major psychosis (Pai et al., 2019).

Our samples were taken from the prefrontal cortex isolated neurons from patients with schizophrenia and bipolar disorder.

The study analysed data from individuals diagnosed with schizophrenia, bipolar disorder, and controls (29, 26 and 27 individuals respectively). In the analysis, study controlled for age, sex, post-mortem interval, genetic ancestry (determined by genotyping the same individuals).

Experiment design

The experiment design was multi-omics study with 55 cases (with schizophrenia or bipolar disorder) and 27 controls.

Objective of the research

According to authors, schizophrenia and bipolar disorder have got characteristic of periods of psychosis. The main objective of the research was to gather epigenomic profiling data to get a more accurate model of neuronal dysregulation in diseases with periods of psychosis.

Biological targets of the research

Researchers intended to look for specific patterns of DNA methylation in isolated neurons from the frontal cortex of individuals that had diseases.

- IGF2 - insulin growth factor 2 protein
- *IGF2* - IGF2 gene
- *Igf2* - enhancer of *IGF2*
- TH - tyrosine hydroxylase protein
- dopamine - a neuromodulatory molecule
- psychosis - an abnormal condition of the mind that results in difficulties determining what is real and what is not real

Results received

Authors found a strong association between methylation of *Igf2* and TH synthesis. TH is the bottleneck enzyme that is responsible for dopamine synthesis. If enhancer *Igf2* is hypomethylated, levels of TH are higher, which determines the higher production of dopamine. Apparently, dopamine is responsible for psychosis in the mental disorders of interest.

Additional information

Schizophrenia and bipolar disorder patients are consistently hypomethylated at *IGF2* locus when compared to controls. This locus remained significantly hypomethylated even after accounting lifestyle-related variables of smoking and anti-psychotic use.

The reaction chain of interest of the research (upward arrows show elevated expression or synthesis of the protein, product, or effect):

Hypomethylation of *Igf2* $\rightarrow \uparrow$ IGF2 $\rightarrow \uparrow$ TH $\rightarrow \uparrow$ dopamine $\rightarrow \uparrow$ psychosis

Data preparation

Sample keys heading is made of the following columns names:

- *id* - an identifier of the sample
- *sentrix_id* - Illumina's Sentrix BeadChip identifier (13 unique values) (National Institutes of Health, n.d.)
- *sentrix_row* - row number in the Sentrix array
- *sentrix_col* - column number in the Sentrix array
- *basename* - sample identifier in the research (joined values in a format: [*id*]_[*sentrix_id*]_R0[*sentrix_row*]C0[*sentrix_id*])
- *tissue_bank_id* - an identifying number of the tissue bank from which the sample was taken
- *tissue_bank* - the literal identifier of the tissue bank
- *tissue* - a tissue type from which the sample was taken
- *cell_type* - a cell type found in the sample
- *donor* - an integer number that identifies the donor of the sample (82 unique values)
- *pmi* - a post-mortem interval, unknown values were labeled as NA
- *race* - race of the donor (white, black, hispanic, or unknown (NA))
- *sex* - gender of the donor
- *diagnosis* - an experimental group of the donor (bipolar, schizophrenia, or control)
- *age* - age of the donor (years)

As it was noted in the article, there were 100 records in the sample keys dataset.

Calculating detection p-values

Getting detection p-value for each score of DNA modification. These p-values determine whether the measured intensity can be distinguished from the background.

All values that have got p-value higher than 0.01 are considered as bad and all samples that have more than 1% of bad detection p-values should be removed.

Although, in our data, none of the samples had more than 1% of bad values, therefore no sample was removed.

Predicting sample sex

This stage estimates sample sex based on methylation data.

Number of females and males after estimation matched original data (25 female and 75 male).

Converted 'M' and 'F' notation to 'male' and 'female'.

No mismatches between real and estimated sex were found.

Data normalisation

According to the documentation of *minfi* package (Fortin & Hansen, n.d.), *preprocessFunnnorm()* function is recommended for known large-scale differences (for example, cancer/normal) or between-tissue studies. Our chosen data spans only over one cell type of one tissue, therefore it was decided to opt for different normalisation methods.

Authors (Pai et al., 2019) noted that they used noob normalisation followed by the quantile one. Quantile normalisation performs processing of Type I and Type II array design differences. Whereas, *preprocessIllumina()* normalisation has only background subtraction and control normalisation implemented. Therefore, we decided to choose *preprocessSWAN()* normalisation, since this method performs within-array normalisation correction for technical differences between Type I and Type II array designs.

Filtering position data by detection p-values

There were 5835 positions found that had p-value higher than 0.01 in 1% of the samples. These positions were removed from the dataset. After this procedure, we have 861001 positions in each sample.

Removing methylation loci positions

2918 methylation loci that do not contain "CG" nucleotide pair (CH probes) or are close to DNR polymorphisms were removed. After the removal data contained 858083 positions in each sample.

Making three different data objects

The DNA modification score matrix was generated and was saved as well as the information about main matrix samples and information about main matrix positions into files for later manipulations with the data.

Interarray correlation outliers elimination

Identification and removal of samples with divergent modification scores.

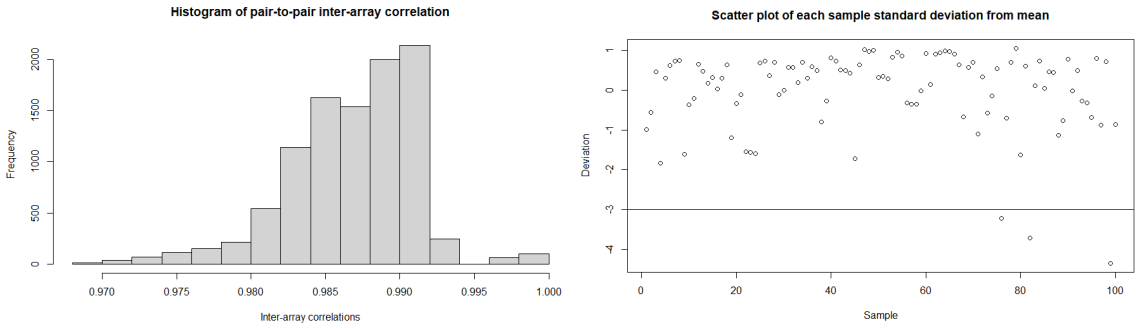

Figure 1: The correlation histogram (left) and scatter plot (right) to detect the outliers for elimination.

The histogram (Figure 1) identifies that our dataset contains values which distort the overall distribution. For further investigation, standard deviation from mean in each sample was calculated.

Scatter plot of each sample (column) (Figure 1) standard deviation from mean visually highlights the data outliers (under -3 limit of deviation).

Algorithm identified and removed 3 outliers:

- GSM3059462_200590490031_R08C01 - a control sample of 53-year-old male
- GSM3059520_200357150067_R08C01 - a sample of a 56-year-old male with bipolar disorder
- GSM3059454_200590490031_R01C01 - a sample of a 77-year-old female with schizophrenia

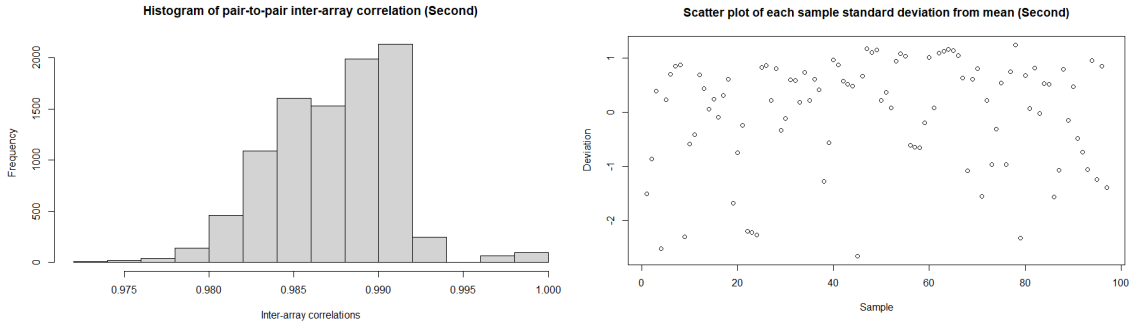

Figure 2: The correlation histogram (left) and scatter plot (right) of all dataset after the elimination of outliers.

There is a visible difference on the left side of the histogram (Figure 2) compared to the histogram before the removal of outliers (Figure 1). This change indicated that the distorting values were removed correctly.

No outliers were left in the recalculated scatter plot (Figure 2).

Quality control

After all data manipulations, our set has 97 samples with 858083 positions.

Data for quality control was separated into case (65 samples) and control (32 samples). Our main goal is to check if distortions in the methylation data exist.

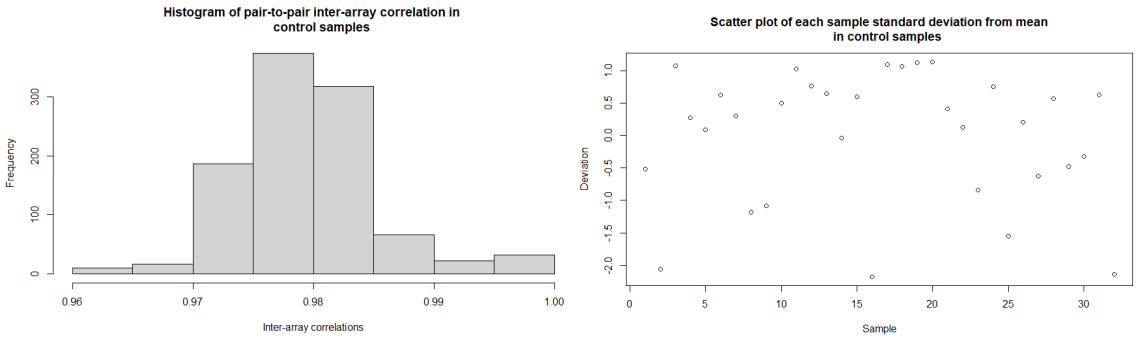

Figure 3: The pair-to-pair correlation histogram (left) and scatter plot (right) for control samples.

The histogram (Figure 3) represents a pair-to-pair correlation in control methylation data.

We can indicate both from histogram and scatter plot (Figure 3) that data is distributed normally and there is no need for data removal.

Sequentially, it was decided to check the distribution of case methylation data.

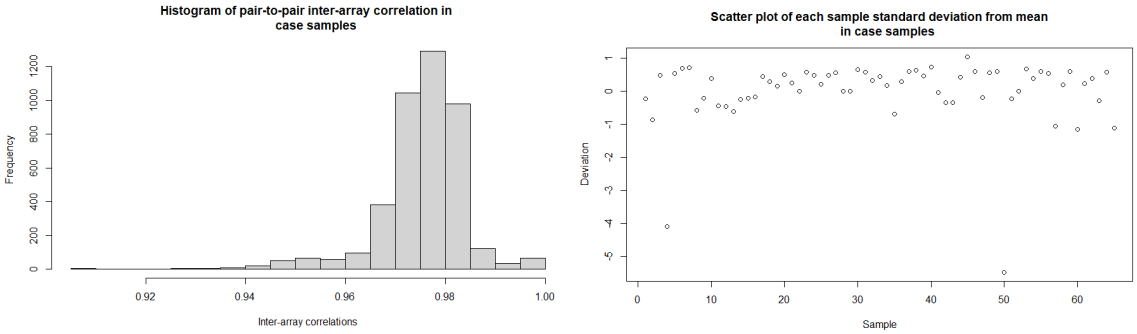

Figure 4: The pair-to-pair correlation histogram (left) and scatter plot (right) of all case samples.

The scatter plot (Figure 4) of all case samples indicates, that our data has distorted values in respect of all case samples mean. It demonstrates two outliers in standard deviation from methylation mean. For further analysis we separated case data into “bipolar” and “schizophrenia” cases.

The scatter plot (Figure 5) with bipolar cases does not show any big fluctuations from the mean methylation value of bipolar disorder case samples.

The scatter plot (Figure 6) of schizophrenia cases also does not show any wide variations from the mean methylation value.

These separated data cases indicated that there is no need to remove any samples.

Additionally, the sample-specific quality control for methylation data with getQC, addQC, and plotQC functions was estimated.

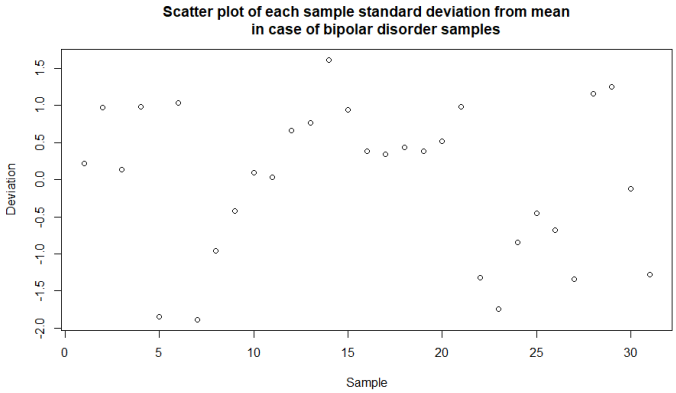

Figure 5: The scatter plot of the bipolar samples.

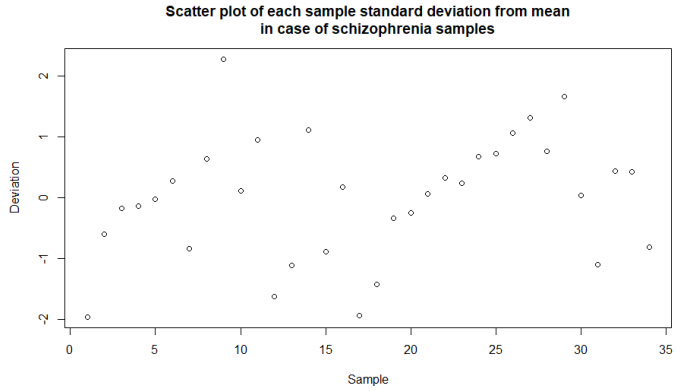

Figure 6: The scatterplot of the schizophrenia case samples.

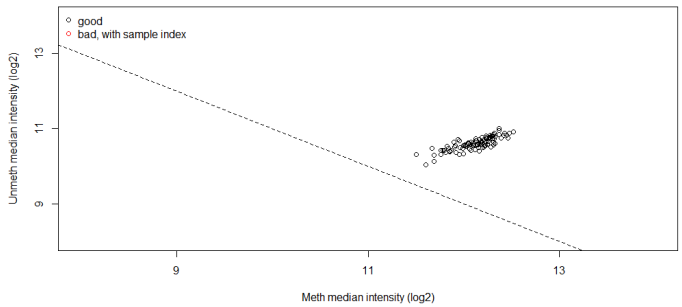

Figure 7: The plot of quality control with getQC, addQC, and plotQC functions.

PlotQC plot (Figure 7) demonstrates that bad samples do not exist in our data set.

Comparison of methylation in density plots (Figure 8) indicates high data quality because no notable deviations are visible from the rest of the samples. Also significant alterations between different diagnosis are not present.

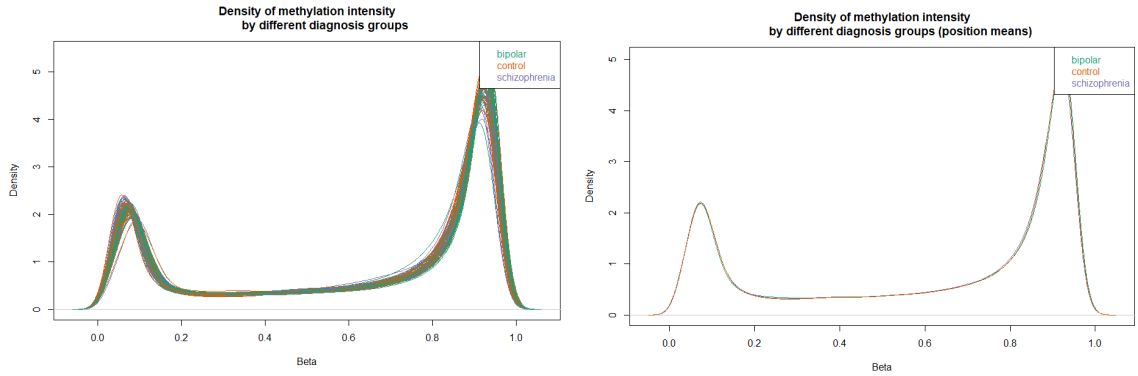

Figure 8: The density plot of methylation intensity of each sample(left) and density plot of all samples positions mean methylation intensity (right).

Saving data

Data was saved into *GSE112179_clear.rds* file after the processing.

Data clustering

For the second task it was required to perform data clustering, which was performed using `hclust` function with `ward.d` linkage method. This method takes into account variance of the clusters, thus it is considered to be the most eligible method for quantitative data sets.

The dendrogram (Figure 9) shows three distinguished groups after clusterisation. It is worth noting that NA values are marked as white color (such examples can be observed in race and post-mortem interval (PMI) colouring)

Clusters were analysed with the respect of sex, case, race, age, and PMI features. The main two clusters were distinguished based on sex. Within these two main clusters, samples were grouped into three smaller clusters.

The first (female) cluster is composed of 24 samples. This collection contains 15 samples of bipolar, 6 samples of schizophrenia, and 3 samples of control cases.

Second and third groups are found within male cluster separated from the first group. It is made of 29 samples, of which 8 are control, 8 are bipolar, and 13 are schizophrenia cases.

Meanwhile, the third group has got 44 samples: 21 belong to control, 8 to bipolar, and 15 to schizophrenia cases.

Regarding these clustering results, it could be stated that besides the feature of sex, there are no significant clusters for further investigation. If we compared numbers of each case within each cluster, we could state that the biggest portion of the first cluster is bipolar (15/24), of the second - schizophrenia (13/29), and the third - control samples (21/44), however only the bipolar cases in the first cluster make up more than a half of the cluster samples. Presumably, if more samples were included in the analysis, more confident statements could be made regarding clustering of methylation data.

The majority of samples had race 'white' (marked red). Race 'hispanic' is marked with a purple colour, and race 'black' is marked with blue colour. Due to the visible imbalance, clusterisation analysis in regards of 'race' feature cannot give any strong conclusions.

There are 3 age intervals that are noted with colours from red to blue: 23-40, 41-58, 59-77. The biggest cluster of the first age group can be observed within female samples. This cluster is composed of bipolar and

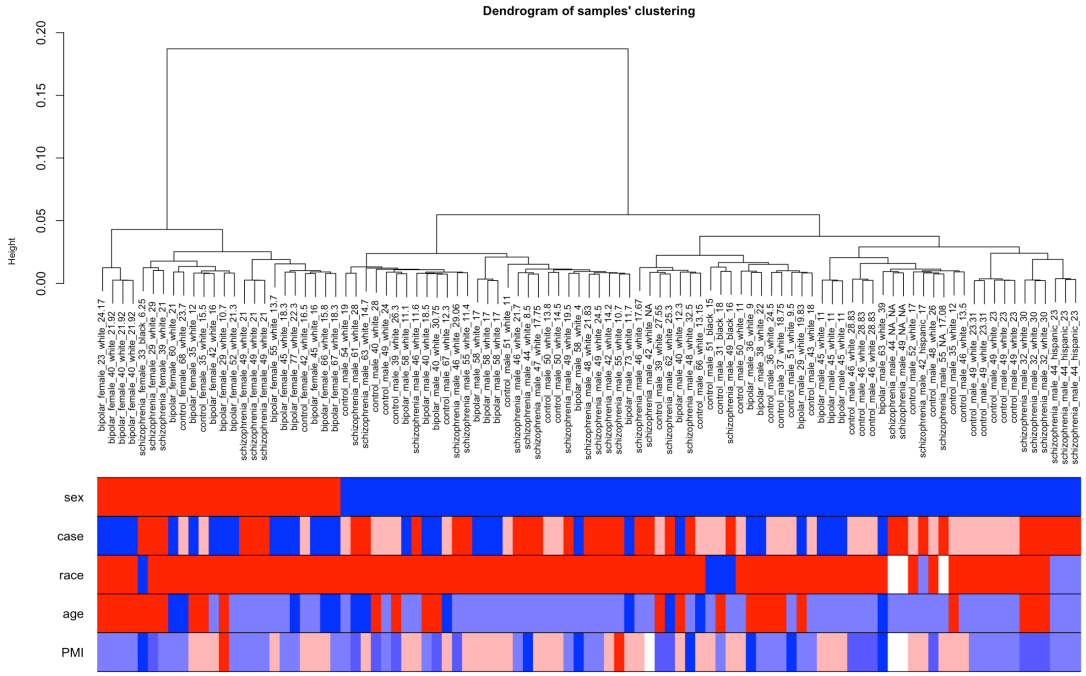

Figure 9: The dendrogram that shows three clusters after the hierarchical clustering using 1-cor() distance metric and Ward linkage method.

schizophrenia case samples. Another more significant cluster can be observed within males of the middle age group. This cluster is composed of all cases, however the majority of cases are schizophrenic. The samples of the last age group can be found in all three clusters of the dendrogram.

The post-mortem interval colouring (from youngest to oldest colour changes from red to blue) shows that most of the samples were collected within 18 time units after death (red, pink, and purple colors).

Heatmap plotting

The second part of the second task was to provide heatmaps for the most varying positions in the data set. The variability of each position was measured by calculating its variance within the samples.

Resources were not effective to reflect all modification positions in heatmap. Therefore, top 100 positions were selected with highest variances to picture data in heatmap.

For comparison, selected data was standardized so that positions would have modification mean 0 and standard deviation 1.

Generated heatmaps indicate that no significant data clustering is visible in our data set. It can be seen that the values were more evenly distributed after the standardization of the data, but this information did not reveal any important data areas, which would be important to investigate.

Several areas show more significant similarities as replicas are compared.

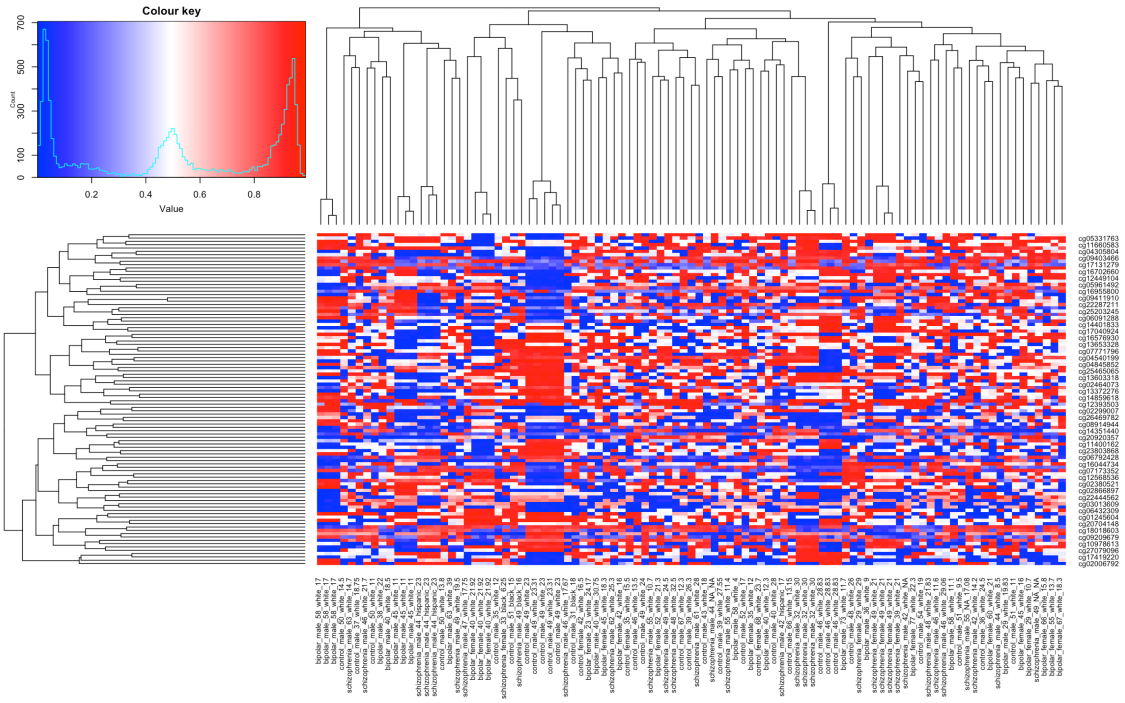

Figure 10: The heatmap of top 100 most variable modification positions with not standardized data.

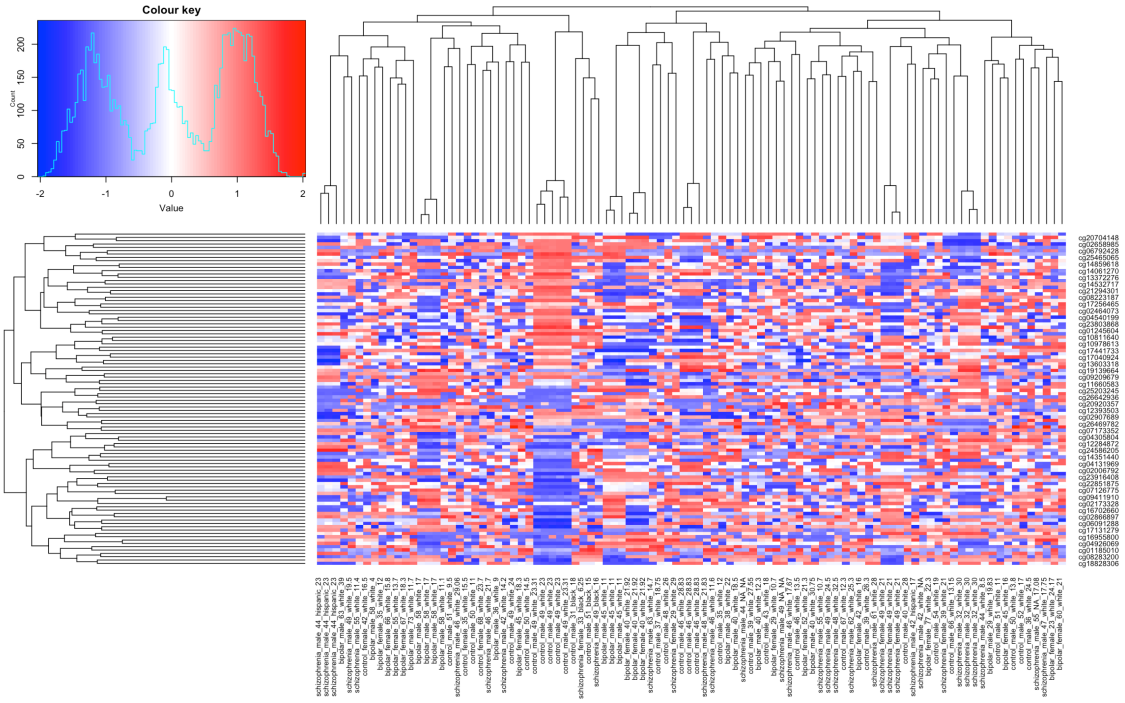

Figure 11: The heatmap of top 100 most variable modification positions with standardized data.

Heatmap plotting with replicas removed

Due to the similarity of the replicas, they were removed from the dataset and recalculations of variability was performed.

After the recalculations, it is not possible to draw any additional conclusions in the heatmap, as no additional clusters have emerged.

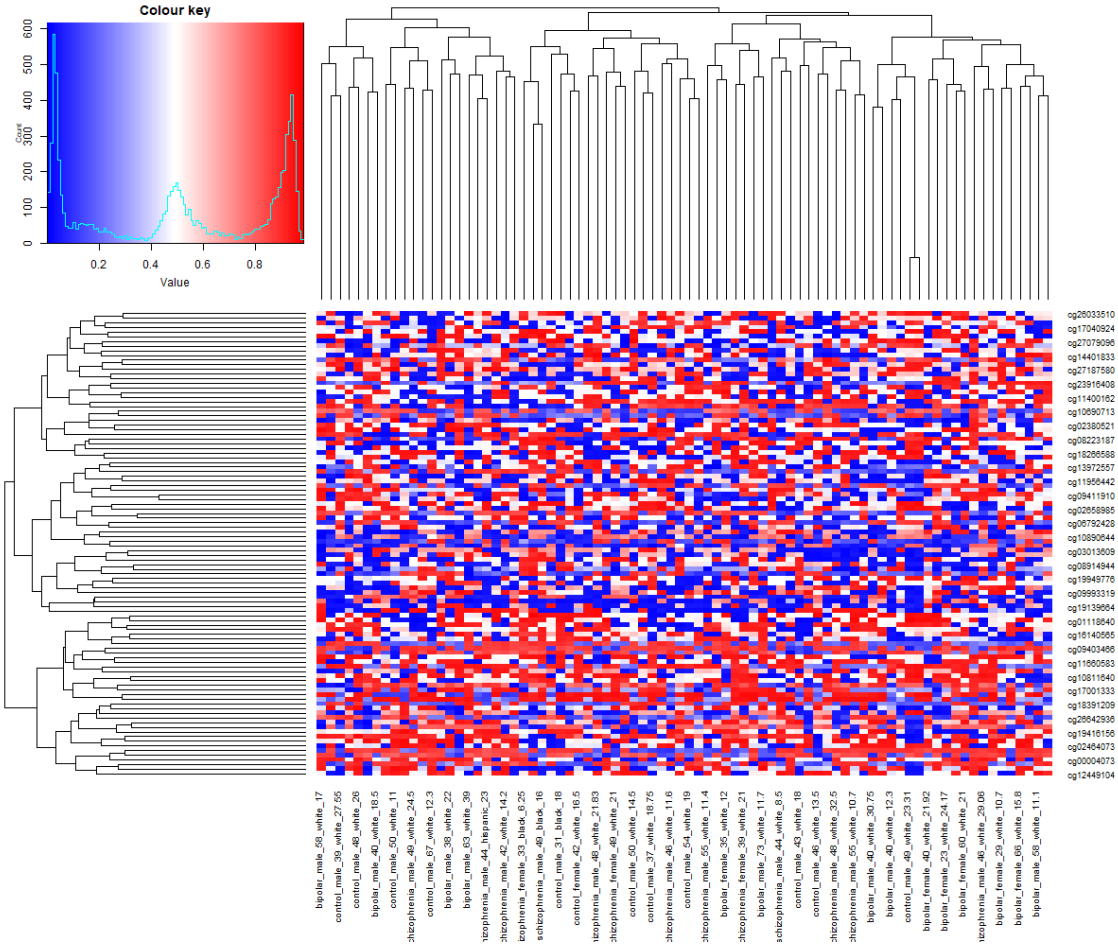

Figure 12: The heatmap of top 100 most variable modification positions with not standardized data after replicas removal.

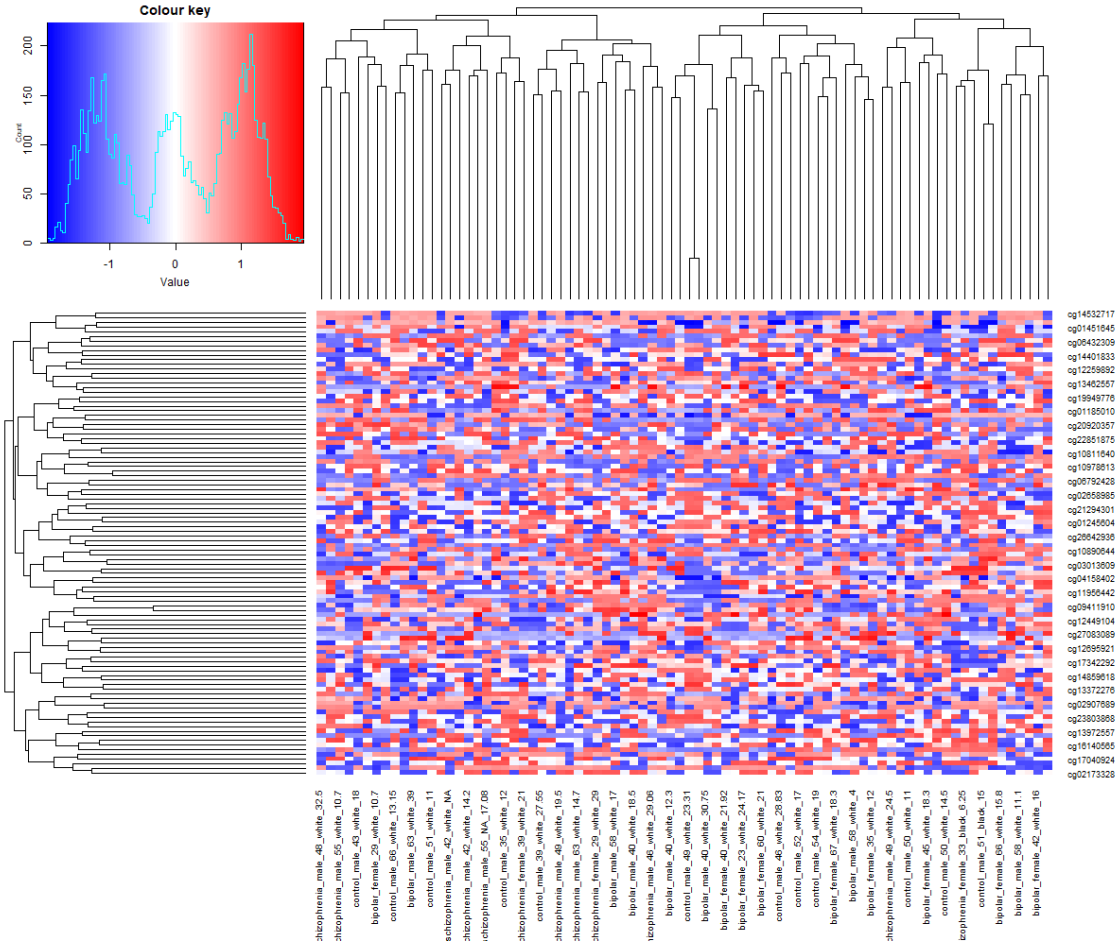

Figure 13: The heatmap of top 100 most variable modification positions with standardized data after replicas removal.

Clocks of DNA modification

The third part of the second task requires to predict age of patients from which the samples were taken and compare predictions with the real data. Furthermore, the next step for the analysis is to check, whether there is a significant difference of predicted age within each experimental group.

Firstly, it was checked, which methylation clocks can be computed for the given data set if the threshold is set to be 80 percent (as default) of required CpGs to compute each clock.

It was checked with `checkClocks` and `DNAmAge` functions that the only clock which could not be computed for the given data set was Bayesian Neural Network (BNN) (Alfonso & Gonzalez, 2020) (`DNAmAge` without age acceleration provided NAs in the output for each sample).

Finally, four methylation clocks were chosen for further workflow: Horwath, Hannum, Levine, and PedBE. Generally, these clocks were chosen because they were computed for the given data set and they predict chronological DNAm age in years. Horvath's clock was trained on samples from various tissues, thus it is a universal choice for any kind of samples. Hannum's clock was trained on blood samples. Nonetheless the tissue type does not match the neural one, from which our data set samples were taken, blood is often taken as an indicator of the state of the whole organism, thus it was decided to include this clock in the workflow. The same considerations can be applied to support the choice of Levine's clock.

Horvath's skin and blood clock was removed due to its incompatibility with the tissue type of our data set and its poor performance ($R^2 = 0.51$). On the contrary PdeBE clock was kept, since it explained the most of variability ($R^2 = 0.75$) in the data set when compared to other clocks, nonetheless the clock was trained on buccal epithelial swabs from patients 0-20 years old (Figure 12).

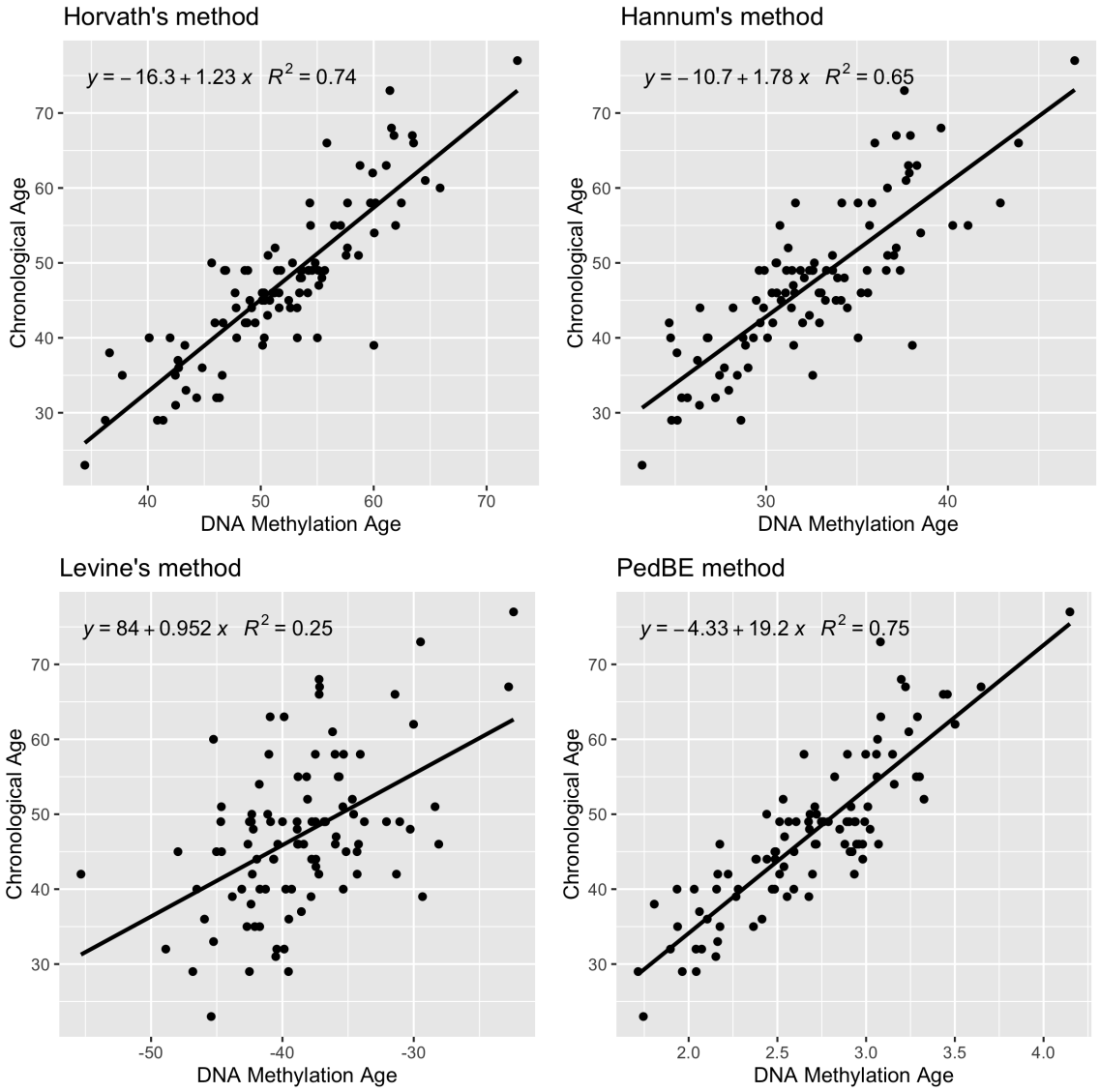

Figure 14: Correlation between biological age predicted by methylation clocks and the chronological age.

To conclude analysis of the predicted and chronological age correlation, the best clocks were Hovarth's and PdeBE, which had R^2 values equal to 0.74 and 0.75 respectively. The clock that had the worst accuracy was Levine's method that had $R^2 = 0.25$, which means that only 25% of the variability observed in the target variable (chronological age) is explained by this model.

Checking how predicted ages differ between experimental groups

The experimental groups (groups of interest) of our research are control, bipolar, and schizophrenia. Therefore, predicted age values will be analysed with regards of this kind of grouping of samples.

It was decided to run ANOVA tests for each methylation clock applied to our data set. The null hypothesis was set to be that there is no difference between predicted biological age means of each case sample group.

- Data: predicted DNA methylation age.
- Test: ANOVA.
- Hipotheses:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$
$$H_1 : \exists i, j \in \{1, 2, 3\} : i \neq j, \mu_i \neq \mu_j$$

- Significance level: $\alpha = 0.1$.
- p value $> \alpha \rightarrow H_0$ not rejected.

p-values of each test: Horvath's 0.806, Hannum's 0.966, Levine's 0.994, PedBE 0.648.

None of the ANOVA tests had p value lower than α, therefore there are no differences in predicted biological age means between control, bipolar, and schizophrenic sample groups.

Hypothesis testing

Grouping data

For the third task it was required to perform hypothesis testing for each position for samples grouped in several different ways. In our case, we decided to define these groups based on: sex, race, and diagnosis type.

Normality testing

Since there were more positions (554715 - up to 65% of all positions) that had p-value of Shapiro normality test higher than $\alpha = 0.05$, therefore it could be stated that the majority of positions had normal data distribution, thus t-test was chosen for further hypothesis testing.

Hypothesis testing

The first hypothesis testing was performed to compare sample groups collected based on sex. Tests were run for each position in the data set.

- Data: methylation values of the position.
- Test: Student t-test.
- Hipotheses:

$$H_0 : \mu_{i,male} = \mu_{i,female},\ where\ i\ -\ index\ of\ position$$
$$H_1 : \mu_{i,male} \neq \mu_{i,female}$$

- Significance level: $\alpha = 0.05$.

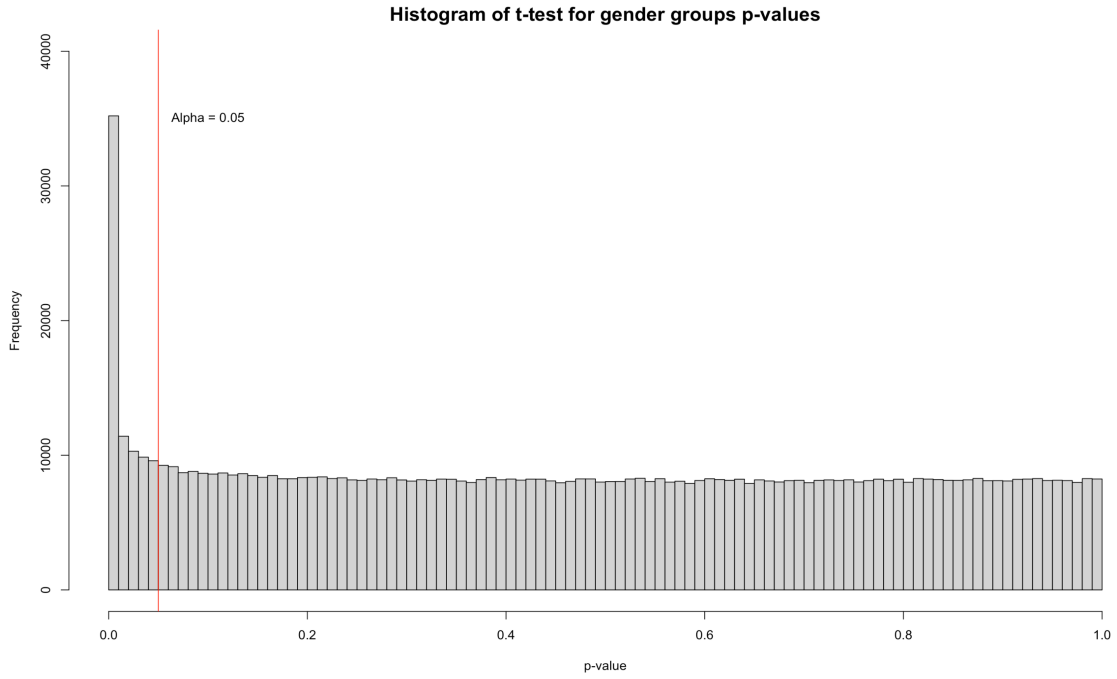

Figure 15: The histogram of p-values of t-test that compared sample groups of different sexes

The histogram of p-values (Figure 15) extracted from performed t-tests showed that there are approximately 80000 positions that have got statistically significant differences in methylation values between gender groups. Precisely, there were 76369 positions that had p-value lower than α.

Five positions that had the lowest p-values were: cg04462931, cg15183843, cg03572700, cg15228509, cg00399683. Their methylation profiles are presented in Figure 16. In all of these positions methylation values in female group were higher.

These positions belong to chromosome 7 (cg04462931 and cg00399683), Y (cg15183843), X (cg03572700), and 1 (cg15228509).

Manhattan plot

Since there were two positions from chromosome 7, it was decided to draw a Manhattan plot (Figure 17) with p-values of positions from this chromosome.

Volcano plot

It was required to draw volcano plots for each chromosome's effect sizes and p-values. Since collection of beta matrix's indeces of positions that belong to a particular chromosome is a time-consuming process, it was decided to plot one volcano plot for the 7th chromosome at first, and if there is enough time left, repeat the flow for other chromosomes

Correction of p-values

P-values from t-test for groups of sex

Number of statistically significant p-values (lower than $\alpha = 0.05$):

- without correction: 76369
- with FDR correction: 21783
- with Bonferroni correction: 15237

References

Alfonso, G., & Gonzalez, J. R. (2020). Bayesian neural networks for the optimisation of biological clocks in humans. *bioRxiv*.

Fortin, J.-P., & Hansen, K. D. (n.d.). Analysis of 450k data using minfi. *Dim*, *485512*, 6.

National Institutes of Health, N. C. I. at the. (n.d.). *Sentrix® BeadChip and BeadArray technology (illumina, inc.) | innovative molecular analysis technologies (IMAT)*. https://imat.cancer.gov/about-imat/outputs-and-achievements/individual-technologies-and-platforms/sentrix%C2%AE-beadchip-and

Pai, S., Li, P., Killinger, B., Marshall, L., Jia, P., Liao, J., Petronis, A., Szabó, P. E., & Labrie, V. (2019). Differential methylation of enhancer at IGF2 is associated with abnormal dopamine synthesis in major psychosis. *Nature Communications*, *10*(1), 1–12.

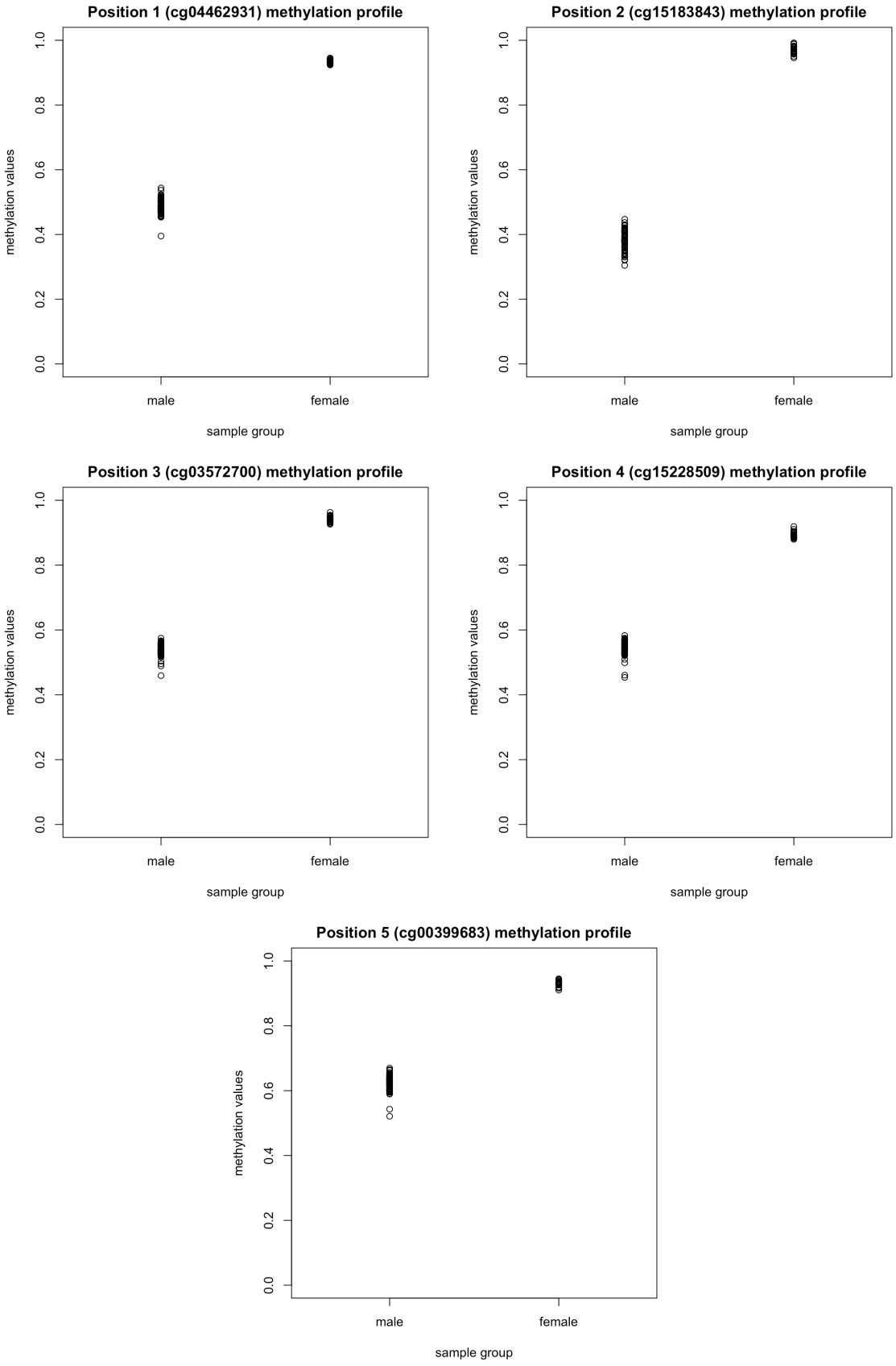

Figure 16: Methylation profiles of positions with top 5 smallest p-values

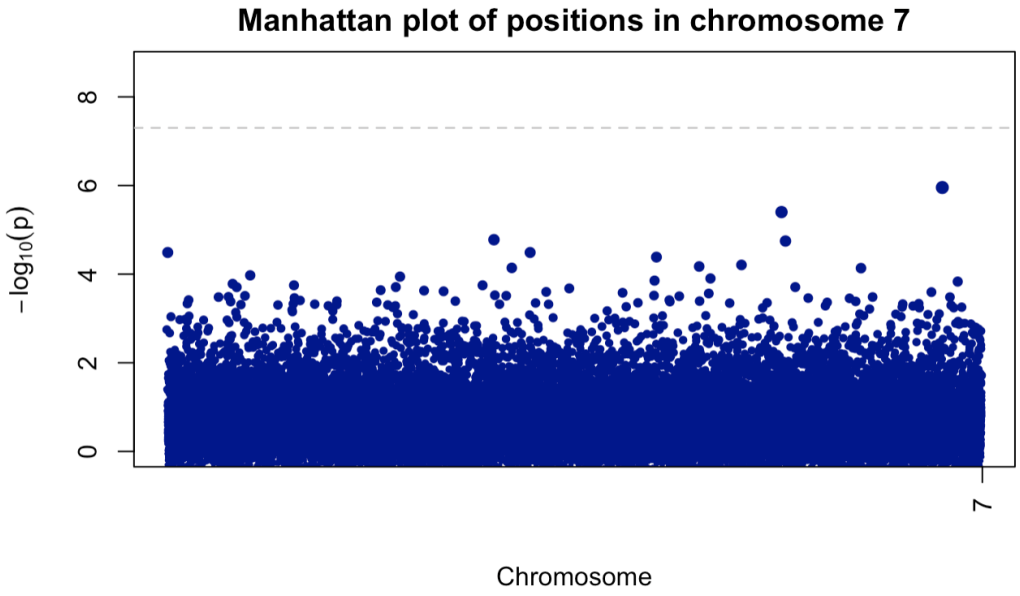

Figure 17: Manhattan plot for positions of chromosome 7