Second task

Akvilė Višniauskaitė ir Ieva Pudžiuvelytė

4/22/2022

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.

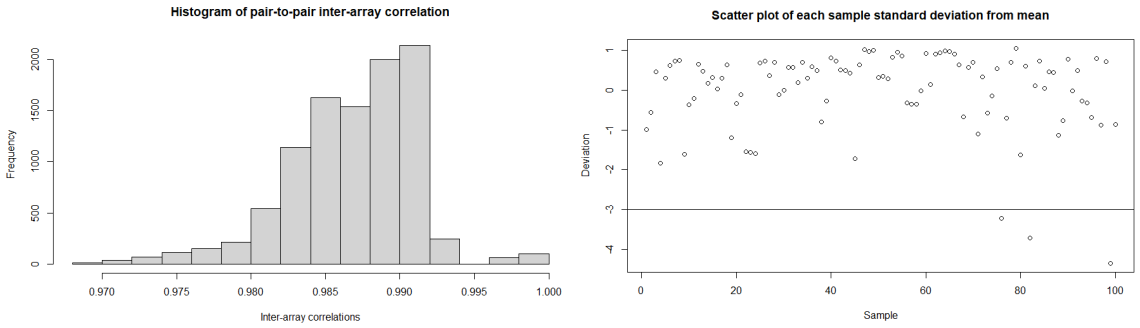

Figure 1: The correlation histogram (left) and scatter plot (right) to detect the outliers for elimination.

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

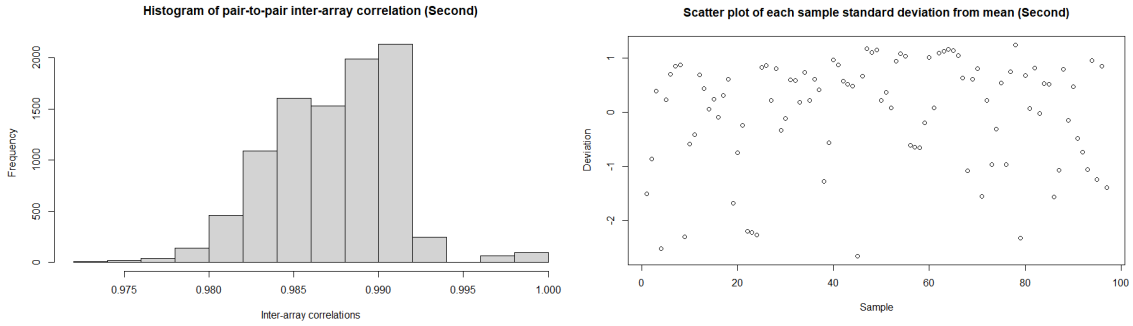

Figure 2: The correlation histogram (left) and scatter plot (right) of all dataset after the elimination of outliers.

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.

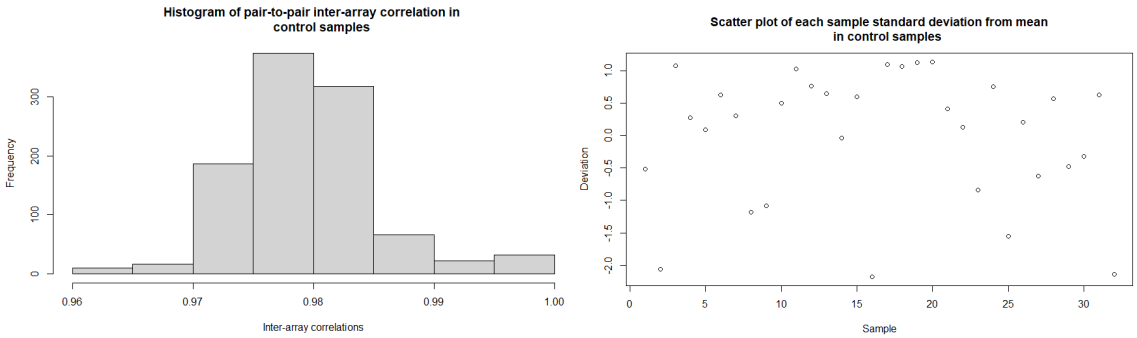

Figure 3: The pair-to-pair correlation histogram (left) and scatter plot (right) for control samples.

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.

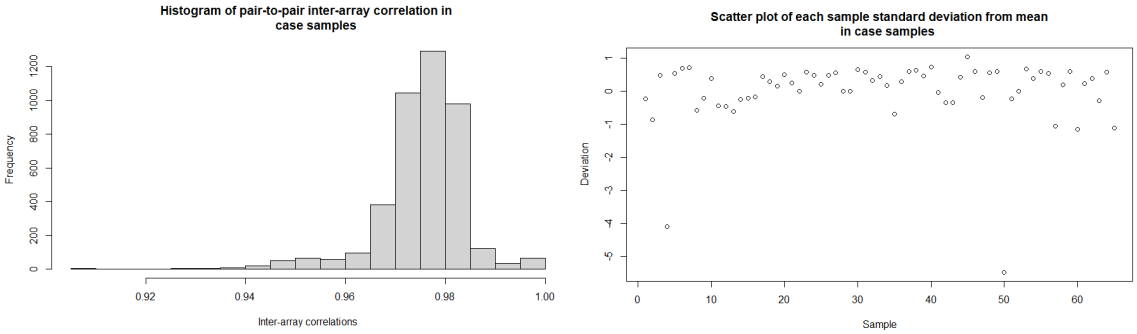

Figure 4: The pair-to-pair correlation histogram (left) and scatter plot (right) of all case samples.

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.

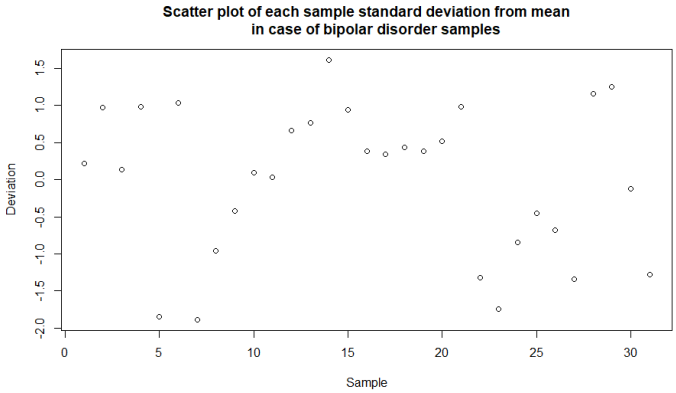

Figure 5: The scatter plot of the bipolar samples.

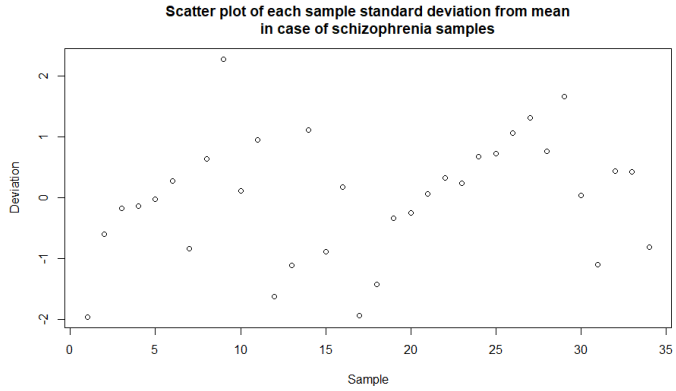

Figure 6: The scatterplot of the schizophrenia case samples.

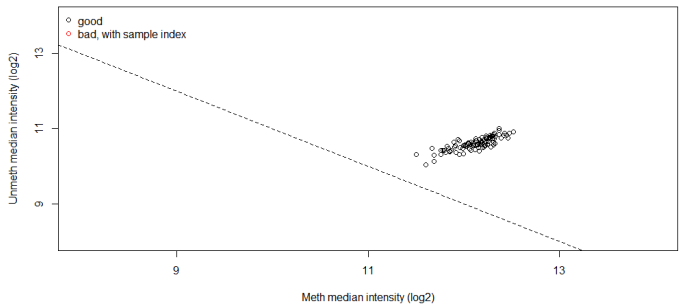

Figure 7: The plot of quality control with getQC, addQC, and plotQC functions.

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.

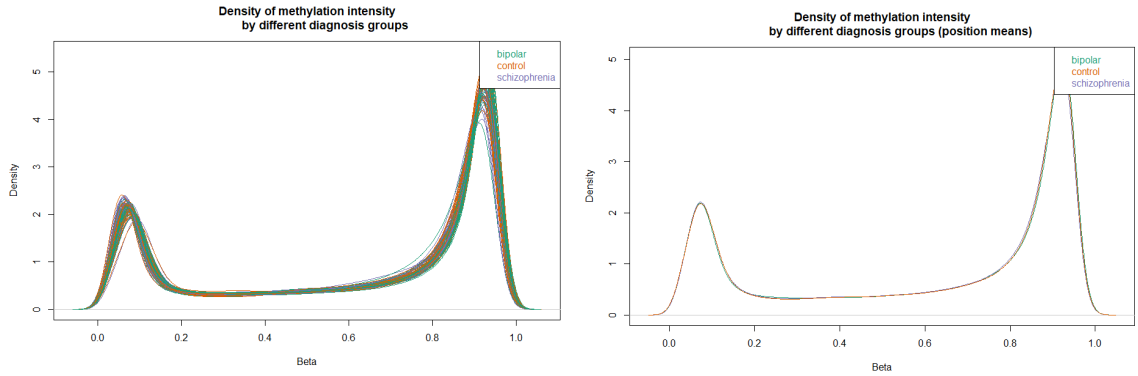

Figure 8: The density plot of methylation intensity of each sample(left) and density plot of all samples positions mean methylation intensity (right).

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.

Removing left-over samples from beta matrix

Calculation of distance matrix. `dist` function could not be used due to 'vector memory exhausted (limit reached?)' error message.

Data clustering was performed using `hclust` function with `ward.d` linkage method. This method takes into account variance of the clusters, thus it is said that it is the most eligible method for quantitative data sets.

The dendrogram (Figure 9) shows three distinguished groups after clusterisation. The groups are denoted by colours: dark, medium, and light gray.

The first group is overall composed of 24 samples. This collection contains 15 samples of bipolar cases, 6 samples of schizophrenia, and 3 samples of control.

Second and third groups are separated from the first group.

The second group is made of 29 samples, of which 8 are control, 8 are bipolar, and 13 are schizophrenia cases.

The third group has got 44 samples: 21 belong to control, 9 to bipolar, and 15 to schizophrenia cases.

Regarding these clustering results, it could be stated that within each of the clusters there is one dominant case. In the first cluster it is bipolar, in the third - control, and in the second - schizophrenia.

EXTRA (try the flow with Euclidean distance)

References

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

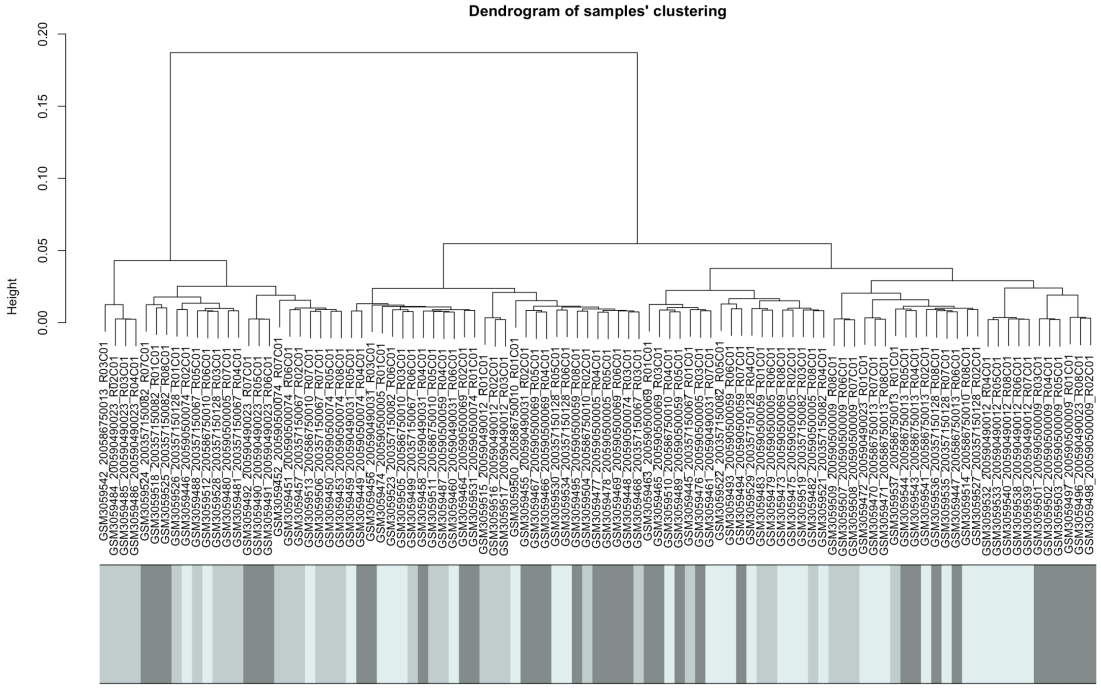

Figure 9: The dendogram that shows three clusters after the hierarchical clustering using 1-cor() distance metric and Ward linkage method.

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.