Research Article GU J Sci 39(2): 1177-1196 (2026) DOI: 10.35378/gujs.1710419

Gazi University

Journal of Science

http://dergipark.org.tr/gujs

Target Gene Prediction from Microarray Data Using Data Mining Methods

Zerrin YILDIZ CAVDAR[1,*], Tolga ENSARI[2], Ahmet SERTBAS[3], Muberra Nur AKCAMAN[4]

[1]*Business Intelligence Leader, Agesa, Istanbul, Türkiye*
[2]*Arkansas Tech University, Department of Engineering and Computing Science, Russellville, USA*
[3] *Istanbul University Cerrahpasa, Department of Computer Engineering, Istanbul, Türkiye*
[4]*Canakkale Onsekiz Mart University, Department of Computer Engineering, Canakkale, Türkiye*

Highlights

- This paper focuses on target gene prediction from microarray data.
- A method called SMATP (Smart Microarray Analyze and Target Prediction) has been developed.
- SMATP was validated on independent datasets and identified key target genes.

Article Info

Received: 30 May 2025
Accepted: 26 Apr 2026

Keywords

Microarray analyzing
Gene expression analyzing
Target gene prediction
Data mining

Abstract

Many different techniques such as microarray, microrna, RNA sequencing and parallel sequencing are used in biomedical research. Among these biotechnological approaches, microarray is widely used for analysing data such as DNA, RNA or proteins. Microarray technology offers advantages in many areas such as analysing gene expression, mutation analysis, epigenetic studies or biomarker discovery. The use of artificial intelligence methods in the analysis of large amounts of data, such as microarray data, offers a gain in accuracy and speed. In this study, gene expression analysis of microarray data is performed using data mining methods. Freely available datasets are used for the study. The first one is the microarray dataset investigating the effects of chronic hypoxia treatment on the brain of mice. The second is a microarray dataset that examines the changes in mouse neurons exposed to oxidative stress. The method we developed for analysing microarray data is applied separately to both data sets and led to successful results. In this work, after the datasets are made suitable for processing in a computer environment, the prediction process is developed using data mining methods. The study is concluded with the listing of the most affected genes among the result genes.

1. INTRODUCTION

Gene expression analysis is an important technique for understanding biological processes by examining how active certain genes are in the transcription process. This analysis has a wide range of applications, particularly in the field of disease mechanisms, biomarker discovery and personalised medicine. Gene expression analysis, one of the methods used to understand biological processes, is used to determine how dominant genes are in a cell or tissue. This technology, which is widely used in biomedical research, is preferred in many areas, such as drug development, biomarker discovery, preventive treatment processes, and understanding the mechanisms of diseases such as cancer [1].

Gene expression analysis are used on the main areas of Disease Research, Personalised Medicine and Biomarker Discovery. Disease Research method is used in the study of cancer or neurodegenerative diseases (such as Parkinson's, Alzheimer's, MS) [2]. Changing gene expression profiles are analysed by comparing normal cells with diseased cells. In this way, both the understanding of disease mechanisms and the development of treatment methods can be promoted. As the other area, Personalised Medicine allows the development of personalised drugs/treatments for differences in gene expression in different individuals. Depending on the genetic structure of the individual, the response to medication can vary [3]. With the help of gene expression analysis, it may be possible to reduce the side effects of drugs or offer a

*Corresponding author, e-mail: zerrintt@gmail.com

more effective treatment. As the third area, Biomarker Discovery can explore disease-specific biomarkers using gene expression analysis. Biomarkers that can contribute to diagnosis and treatment can be identified by considering the genes that are particularly highly expressed in some cancer cells or the gene expressions detected in blood samples.

Various methods such as "microarray analysis", "RNA sequencing", "qPCR - quantitative polymerase chaing reaction" can be used to analyze gene expression. Microarrays are a technology that allows the entire DNA information (genome) of a living organism can be visualized on a chip in order to detect differences in gene expression [4]. In microarray analysis, microarray data is commonly analyzed with bioinformatics techniques to evaluate differences in biological processes. In RNA sequencing (RNA-Seq), the transcriptome (the sum of all RNA molecules in a given state) in a cell or tissue is analyzed to measure which genes are activated and to what extent [5]. In qPCR, the amount of RNA or gene in the sample is measured precisely and accurately. Especially "real-time PCR" method is widely used in gene expression analysis [6].

One of the most preferred gene expression analysis methods, Microarray analysis technique provides high accuracy and speed of evaluation. Microarray gene expression analysis is used in many fields, such as genetic diseases, cancer research, preventive treatment and drug development processes [7]. Microarray technology, a gene or DNA fragment is placed on a glass or silicon chip in arrays with multiple replicates. These arrays are specifically designed for the target genetic material. For gene expression analysis with this method, the five steps are performed: First one, preparation of samples in the laboratory environment (The RNA is isolated from the sample to be analysed and converted into cDNA), second step, the labelling cDNA with a fluorescent dye, third step , hybridization process, ligation of the labelled cDNA to the sequences on the chip, each probe on it targets a specific gene and determines the expression level of that gene, the fourth step is Imaging, in that the hybridised samples are read and imaged, the last one is analysis, the gene expression level for each gene is determined and analysed using bioinformatic methods.

There are various software programs that can be used for gene expression analysis with microarray technology. The "Illumina BeadStudio/GenomeStudio" software developed by Illumina can be used to prepare microarray data for computerized processing [8]. It can be used for both gene expression and genotype analysis. The software "Affymetrix", which is specially developed for microarray data, enables the analysis of raw data and the extraction of gene expressions [9]. Another software "Partex" is developed for both microarray and RNA-Seq analysis [10]. Similarly, the software "GeneSpring" can be used for both analysis and visualization of microarray data [11].

In this study, the extraction of target genes from microarray data is performed using datasets generated with Illumina BeadStudio. The datasets are obtained from NCBI (National Centre of Biotechnology Information), an open access database platform. GEO (Gene Expression Omnibus) at NCBI contains a lot of open access genomic data. The datasets that are the subject of this work are GSE8262 and GSE10470, which can be accessed at GEO [12]. The datasets generated as a result of the studies performed on mice in the laboratory environment are preprocessed with Illumina BeadStudio software and transferred to the computer environment. The Bioconductor library of the R programming language is used for data mining studies on the datasets [13, 14].

2. MATERIAL METHOD

In this study, where gene expression analysis is performed, the microarray data is analysed using data mining methods. The datasets generated with Illumina BeadStudio are retrieved from the GEO platform. Our target gene prediction method (SMATP: Smart Microarray Analyze and Target Prediction) is created with the R programming language and the Bioconductor library. Two separate datasets are used to prove the reliability and accuracy of the method. The proposed method-SMATP is successful in both GSE8262 and GSE10470 datasets. It should be noted that SMATP is not a classification or prediction model but an analytical workflow. Therefore, performance is not evaluated using conventional accuracy metrics. Instead, the reliability of the results is assessed based on statistical significance (FDR-adjusted p-values), consistency across datasets, and biological relevance supported by the literature.

2.1. Datasets: GSE8262, GSE10470

The first dataset of the study is accessed from the NCBI GEO platform with the accession number GSE8262. The dataset can be accessed at ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8262 [15]. GSE8262 contains gene expression data from mice exposed to two different hypoxia treatments (CIH: Chronic Intermittent Hypoxia and CCH: Chronic Consistent Hypoxia). Microarray analysis is performed to investigate the effects of two different treatment models on cortical and hippocampal regions of the mouse brain. Treatment models used: "CIH: Intermittent exposure to low oxygen levels", "CCH: Continuous exposure to low oxygen levels". The GSE8262 dataset contains 18 samples: "CIH group: Includes 6 samples. It consists of cortical and hippocampal samples from mice subjected to chronic intermittent hypoxia treatment", "CCH group: Includes 6 samples. It consists of cortical and hippocampal samples from mice treated with chronic continuous hypoxia", "Control group: Includes 6 samples. It consists of 3 cortical and 3 hippocampal samples from untreated mice". The explanation of the abbreviations used for the examples in the microarray analysis is: "cx: cortex", "hc: hippocampus", "ctrl: control", "cih: chronic intermittent hypoxia", "cch: chronic consistent hypoxia".

When analysing the academic studies using the GSE8262 dataset, four studies were found. The study by Zhou et al investigated how chronic hypoxia alters gene expression in the brain of mice [16]. In particular, the role of the gene sarcospan (sspn) in the death of glial cells under hypoxia is investigated in detail. In the study, RNA is isolated and microarray analysis is performed after the selection of mice, hypoxia treatment and preparation of the examined tissues. It is found that 80 genes are affected by CIH (Chronic Intermittent Hypoxia) and 137 genes are affected by CCH (Chronic Consistent Hypoxia). This study provides important insights that may contribute to the treatment of neurological disorders associated with hypoxia. The results showed that chronic hypoxia conditions lead to significant changes in gene expression in the cerebral cortex and hippocampus regions. These data is available on the GEO platform with the code GSE8262. In another study by Wang et al, the role of the protein SENP1 (SUMO Specific Protease 1) in intermittent hypoxia-induced neuroinflammation and cognitive decline is investigated. The changes in microglia are investigated both in vitro (in the laboratory) and in vivo (in the living organism). The study shows that the SENP1 protein regulates intermittent hypoxia-induced neuroinflammation by modulating the SUMOlation of the NEMO protein and activating the NF-kB signalling pathway. These results suggest that modulation of SENP1 in microglia may be a potential target for the treatment of cognitive impairment induced by intermittent hypoxia [17]. Wang and colleagues also investigated the effect of SENP1 on neuroinflammation and neuronal damage caused by CIH associated with OSAS (obstructive sleep apnea syndrome) [18]. Microglial mouse models with SENP1 overexpression and SENP1 deficiency are used in this study. The analyses reported that CIH decreased SENP1 and TOM1 levels, increased SUMOlation of TOM1 protein and this had a promoting effect on microglial migration, neuroinflammation, Aβ42 (neuronal beta-amyloid42) accumulation and apoptosis. The effect of SENP1 on TOM1 suggests that it may develop new approaches for the treatment of cognitive dysfunction caused by CIH. In the last study by Cavdar et al [19] you will find information about the study on the prediction of target genes using this dataset. Information on the genes obtained as a result of the study was also provided.

The second dataset, which is the subject of the study, is accessed from the NCBI's GEO platform with the accession number GSE10470. The dataset can be accessed at ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19470 [20]. GSE10470 is created to investigate the effects of oxidative stress (HCA: homocysteate) on gene expression and the role of ATF4 (Activating Transcription Factor 4) in this process. Mouse embryonic neurons are used for the study. HCA is used as an experimental tool to understand the effect of ATF4, which is known to be activated upon oxidative stress. In the dataset, HCA is applied to samples from knock-out and wild-type mice lacking the ATF4 gene. The GSE10470 dataset contains 24 samples: "Wild-type HCA: contains 6 samples. It consists of neurons from normal mice treated with HCA", "Knock-out ATF4: Contains 6 samples. Consists of neurons from HCA-treated mice without the ATF4 gene", "Wild-type Control: Contains 6 samples. This is a control group with normal embryonic neurons from mice without HCA treatment", "Knock-out Control: Contains 6 samples. This is a control group with mouse embryonic neurons without ATF4 gene and without HCA". The explanation of the abbreviations used for the samples in the microarray analysis is: "wt: wild type, normal", "ko: knock-out, without the ATF4 gene", "ctrl: control", "hca: homocysteate".

Four studies were found using the GSE10470 dataset. In the study by Lange et al, the role of the protein ATF4 (Activating Transcription Factor 4) in neurons under oxidative stress is investigated [21]. It is emphasised that oxidative stress begins with an increase in free radicals and ATF4 is activated as a result of the depletion of antioxidants such as glutathione. This process leads to increased expression of genes that trigger cell death in neurons. The study used neuron cultures exposed to stress factors such as glutathione deficiency in mice and performed RNA isolation and microarray studies both in vivo and in vitro analyses. As a result, it is reported that ATF4 triggers oxidative stress-induced cell death in cells of the nervous system and that this process could be a potential target for the treatment of oxidative stress-related diseases. The data obtained in this study has been made available on the GEO platform with the code GSE10470. In another study by Woo et al., the neuroprotective effect of GRM8 (Metabotropic Glutamate Receptor 8) protein on inflammation in the central nervous system is investigated [22]. This study emphasises that increased glutamate levels in the pathogenesis of various neurodegenerative diseases (MS, Alzheimer's, Parkinson's) lead to excessive neuronal excitation and consequently to the death of neurons. It is assumed that the GRM8 protein has the potential to protect neurons by regulating glutamate levels. The study shows that GRM8 can be considered as a potential target in the treatment of neurodegenerative diseases. The study by Bloomingdale et al aimed to develop new strategies for the treatment of peripheral neuropathy caused by the chemotherapeutic agent bortezomib [23]. In the study, the GSE10470 dataset is used to measure nerve conduction velocities and pain behaviour in mice with neuropathic pain due to bortezomib treatment, and the effects of treatment with bortezomib and dexanabinol are examined in the laboratory. The results suggest that dexanabinol may have therapeutic potential in the treatment of chemotherapy-induced neuropathic pain. The GSE10470 dataset was also used in the study by Cavdar et al. The paper discusses how machine learning methods can be used to analyse this dataset and presents the result gene list [19].

All analyses were performed using RStudio 2024.04.1 Build 748 and Bioconductor (version 3.19). The primary packages used in this study include limma (version 3.60.6) for differential expression analysis and ggplot2 (version 3.5.1) for visualization.

The analyses were conducted on a system with an Intel(R) Core(TM) Ultra 7 155H processor, 16 GB RAM, running a Windows operating system.

2.2. Proposed Method: Smart Microarray Analyze and Target Prediction (SMATP)

In this work, an effective solution for target gene prediction is obtained by analysing microarray data with data mining method which we call SMATP. SMATP provided fast and accurate results by applying artificial intelligence technologies in both analysing and understanding microarray data and predicting target genes. The SMATP steps are shown in Figure 1.

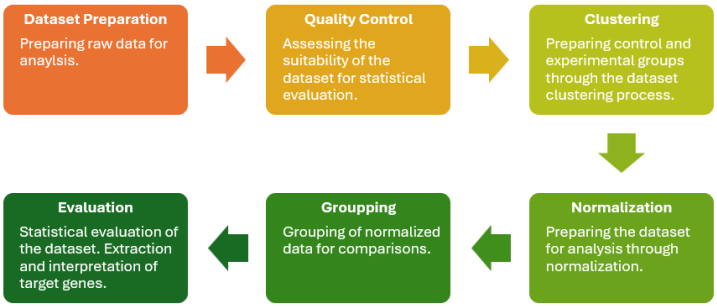


Figure 1. *Steps of the SMATP method*

Data preparation is the first step of our SMATP method, which we developed for analysing microarray data generated with Illumina BeadStudio/GenomeStudio. It involves accessing the dataset, transferring the data to the computer environment and creating a data frame.

*__Accessing the dataset:__ In this study, two independent datasets are used to prove the accuracy of the method. Both datasets are obtained from the GEO platform,

***Computerization of the data:** The delimited microarray raw data file accessed from the GEO platform is transferred to the computer environment using the R programming language. The transferred GSE8262 microarray raw data file contains 46120 data lines. A total of 18 samples are expressed by 145 columns. The GSE10470 microarray raw data contains 24048 data lines. A total of 24 samples are represented by 193 columns,

***Preparation of the data frame:** In order to analyse the dataset, the data frame is created by performing the necessary labelling operation. The data frames created with the data.frame() function of the R programming language are shown in Tables 1 and 2.

***Table 1.** Data Frame – GSE8262. cx:cortex, hc:hipocampus, ctrl:control, cih:chronic intermittent hypoxia, cch:chronic consistent hypoxia*

Array	Batch	Genotype	Replicate	Drug	ArrayCode
1	1	Cx	1	ctrl	1412023004_A
2	1	Cx	1	cih	1412023004_B
3	1	Cx	1	cch	1412023004_C
4	1	Hc	1	ctrl	1412023004_D
5	1	Hc	1	cih	1412023004_E
6	1	Hc	1	cch	1412023004_F
7	2	Cx	2	ctrl	1412023005_A
8	2	Cx	2	cih	1412023005_B
9	2	Cx	2	cch	1412023005_C
10	2	Hc	2	ctrl	1412023005_D
11	2	Hc	2	cih	1412023005_E
12	2	Hc	2	cch	1412023005_F
13	3	Cx	3	ctrl	1412023006_A
14	3	Cx	3	cih	1412023006_B
15	3	Cx	3	cch	1412023006_C
16	3	Hc	3	ctrl	1412023006_D
17	3	Hc	3	cih	1412023006_E
18	3	Hc	3	cch	1412023006_F

***Table 2.** Data Frame – GSE10470. wt: wild-type/normal, ko:knock-out/without ATF4 gene, ctrl:control, hca:homocysteate*

Array	Batch	Genotype	Replicate	Drug	ArrayCode
1	1	Wt	1	ctrl	1412066061_A
2	1	Wt	2	ctrl	1412066061_B
3	1	Wt	3	ctrl	1412066061_C
4	1	Wt	4	ctrl	1412066061_D
5	1	Ko	1	ctrl	1412066061_E
6	1	Ko	2	ctrl	1412066061_F
7	1	Ko	3	ctrl	1412066061_G
8	1	Ko	4	ctrl	1412066061_H
9	2	Ko	5	ctrl	1541554116_A
10	2	Wt	1	hca	1541554116_B
11	2	Ko	1	hca	1541554116_C
12	2	Wt	5	ctrl	1541554116_D
13	2	Ko	6	ctrl	1541554116_E
14	2	Wt	2	hca	1541554116_F
15	2	Ko	2	hca	1541554116_G
16	2	Wt	6	ctrl	1541554116_H
17	3	Wt	7	ctrl	1541554117_A
18	3	Ko	3	hca	1541554117_B
19	3	Wt	3	hca	1541554117_C
20	3	Ko	7	ctrl	1541554117_D
21	3	Wt	8	ctrl	1541554117_E
22	3	Ko	4	hca	1541554117_F
23	3	Wt	4	hca	1541554117_G
24	3	Ko	8	ctrl	1541554117_H

In the data frames shown in Tables 1 and 2; “Array: Sample sequence number information”, “Batch: It is the information in which batch group the process is performed”, “Genotype: Gene type information”, “Replicate: It is the information in which repeat group the process is performed”, “Drug: Treatment type information”, “ArrayCode: Microarray raw data sample code information".

Quality control is the second step of the study, i.e. the process of analysing the prepared data. The analysis process includes: "sorting the samples in the existing data", "preparing the data matrix", "checking the number of beads per beadset", "checking the detection values", "checking the signal distribution before normalisation".

The control processes carried out are described in detail and clearly below.
***Sorting the samples:** Using the data frame, 18 samples in GSE8262 and 24 samples in GSE10470 are sorted as "GSE8262: cx.ctrl.1, cx.cih.2, cx.cch.3, hc.ctrl.4, hc.cih.5, hc.cch.6, cx.ctrl.7, cx.cih.8, cx.cch.9, hc.ctrl.10, hc.cih.11, hc.cch.12, cx.ctrl.13, cx.cih.14, cx.cch.15, hc.ctrl.16, hc.cih.17, hc.cch.18" and "GSE10470: wt.ctrl.1, wt.ctrl.2, wt.ctrl.3, wt.ctrl.4, ko.ctrl.5, ko.ctrl.6, ko.ctrl.7, ko.ctrl.8, ko.ctrl.9, wt.hca.10, ko.hca.11, wt.ctrl.12, ko.ctrl.13, wt.hca.14, ko.hca.15, wt.ctrl.16, wt.ctrl.17, ko.hca.18, wt.hca.19, ko.ctrl.20, wt.ctrl.21, ko.hca.22, wt.hca.23, ko.ctrl.24".
***Preparation of the data matrix:** The data matrix after log2 transformation of the sorted data is shown in Tables 3 and 4.

Table 3. *Data matrix – GSE8262*

Example	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
cx.ctrl.1	5,440	6,322	6,591	7,141	7,345	15,634
cx.cih.2	5,472	6,307	6,559	7,106	7,276	15,445
cx.cch.3	5,443	6,309	6,580	7,155	7,379	15,896
hc.ctrl.4	5,354	6,235	6,492	6,999	7,159	15,569
hc.cih.5	5,347	6,285	6,536	7,069	7,229	15,534
hc.cch.6	5,385	6,236	6,492	7,013	7,173	15,501
cx.ctrl.7	5,336	6,194	6,458	7,020	7,212	15,531
cx.cih.8	5,329	6,081	6,326	6,856	6,992	15,290
cx.cch.9	5,350	6,064	6,331	6,883	7,085	15,481
hc.ctrl.10	5,365	6,117	6,370	6,982	7,190	15,576
hc.cih.11	5,293	6,057	6,326	6,866	7,058	15,532
hc.cch.12	5,382	6,133	6,382	6,928	7,077	15,396
cx.ctrl.13	5,202	5,986	6,248	6,853	7,080	15,346
cx.cih.14	5,304	6,053	6,311	6,883	7,086	15,637
cx.cch.15	5,202	5,991	6,252	6,791	6,995	15,473
hc.ctrl.16	5,244	5,928	6,146	6,680	6,779	14,860
hc.cih.17	5,190	5,931	6,190	6,784	7,012	15,289
hc.cch.18	5,396	6,081	6,334	6,856	7,015	15,228

Table 4. *Data matrix – GSE10470*

Example	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
wt.ctrl.1	6.317	7.008	7.365	8.279	9.196	15.828
wt.ctrl.2	6.162	6.833	7.167	7.971	8.633	15.681
wt.ctrl.3	6.081	6.844	7.225	8.119	9.033	15.847
wt.ctrl.4	6.217	6.898	7.265	8.158	9.046	15.895
ko.ctrl.5	6.042	6.806	7.187	8.094	9.023	15.795
ko.ctrl.6	6.154	6.896	7.266	8.159	9.049	15.816
ko.ctrl.7	6.276	6.908	7.237	8.084	8.863	15.778
ko.ctrl.8	6.208	6.848	7.204	8.087	8.946	15.478
ko.ctrl.9	6.164	6.809	7.075	7.666	8.050	15.070
wt.hca.10	6.119	6.823	7.100	7.657	8.067	15.129
ko.hca.11	6.184	6.941	7.234	7.901	8.411	15.330
wt.ctrl.12	6.242	6.948	7.259	7.920	8.486	15.373
ko.ctrl.13	6.115	6.812	7.089	7.697	8.117	15.133
wt.hca.14	6.129	6.837	7.117	7.686	8.084	15.211
ko.hca.15	6.168	6.800	7.064	7.635	8.002	15.003
wt.ctrl.16	5.711	6.651	6.924	7.459	7.799	14.924
wt.ctrl.17	6.396	6.993	7.271	7.803	8.142	15.237
ko.hca.18	5.982	6.770	7.029	7.564	7.853	14.982
wt.hca.19	6.482	7.075	7.361	7.983	8.450	15.449
ko.ctrl.20	6.254	6.939	7.307	8.051	8.773	15.425
wt.ctrl.21	6.127	6.869	7.146	7.725	8.137	15.299
ko.hca.22	5.741	6.654	6.918	7.467	7.806	15.084
wt.hca.23	5.963	6.731	7.013	7.613	8.040	15.300
ko.ctrl.24	5.044	6.416	6.681	7.227	7.541	14.812

***Bead number per beadset control:** The average number of beads contained in each bead cluster in each array is checked. Figures 2 and 3 show the average number of beads comprising each beadset in each array.

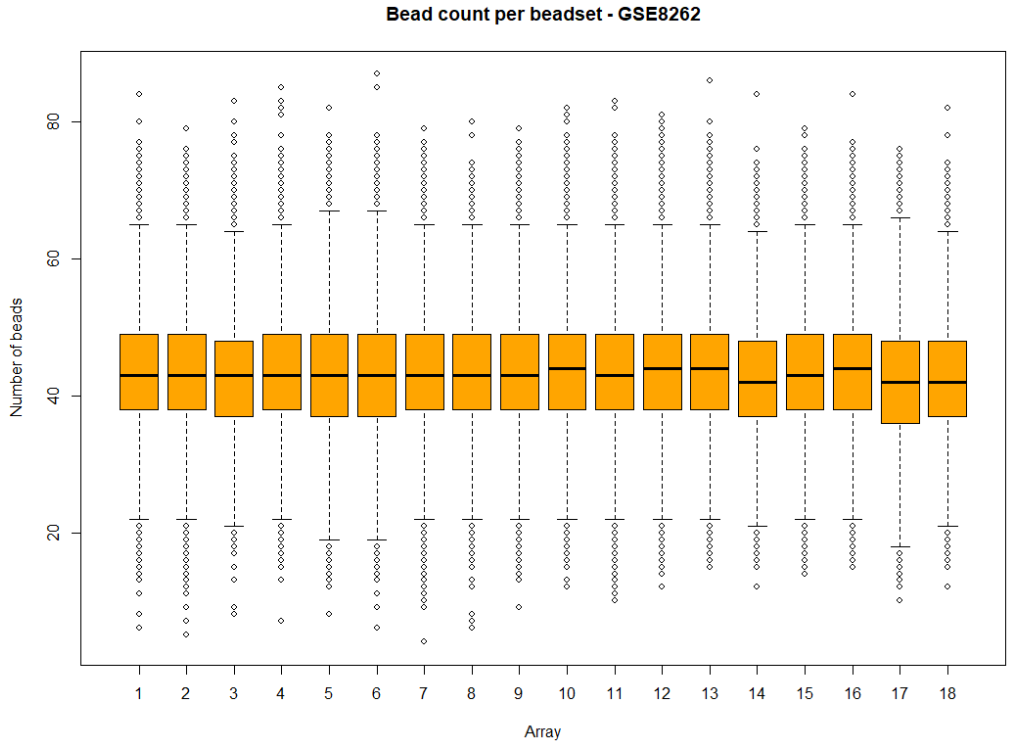


Figure 2. *Bead count per beadset – GSE8262*

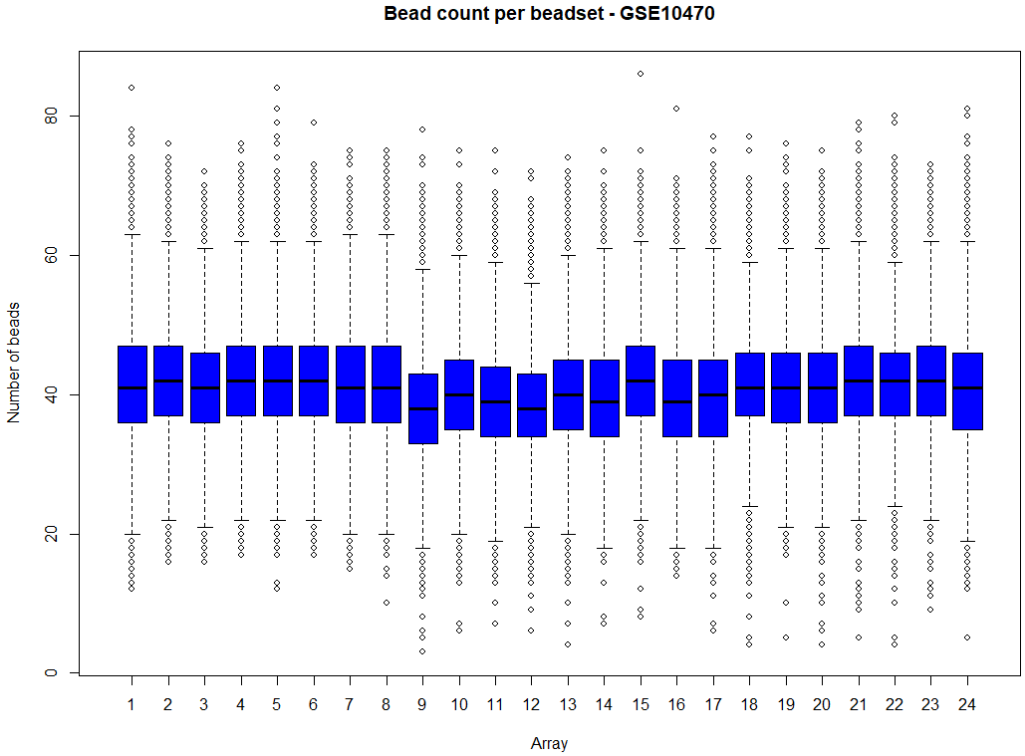


Figure 3. *Bead count per beadset – GSE10470*

***Detection score control:** Figures 4 and 5 show the detection scores generated by the BeadStudio software indicating the probability of genes being expressed between 0 and 1 (0: not expressed, 1: expressed).

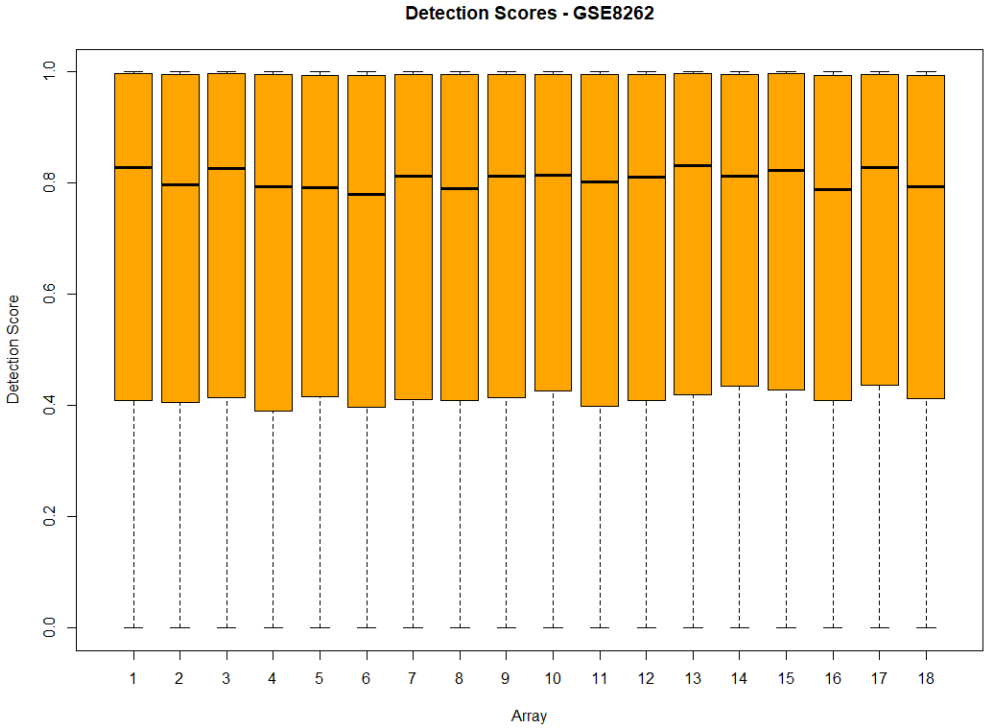


Figure 4. *Detection scores – GSE8262*

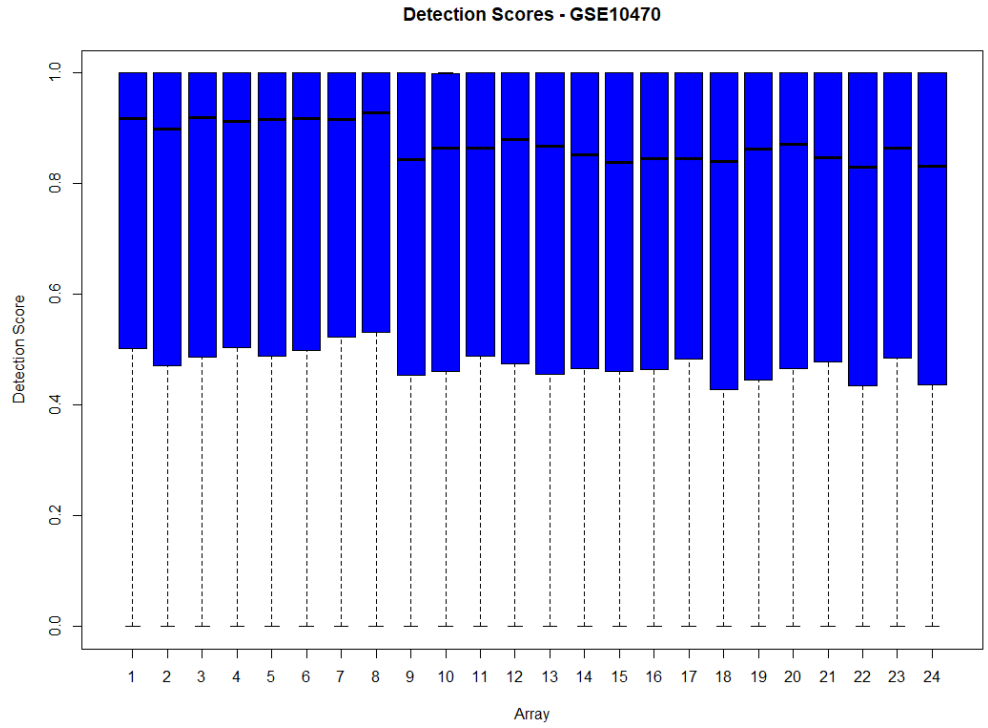


Figure 5. *Detection scores – GSE10470*

***Signal distribution control:** The signal distribution in the unprocessed raw data is checked in Figures 6 and 7.

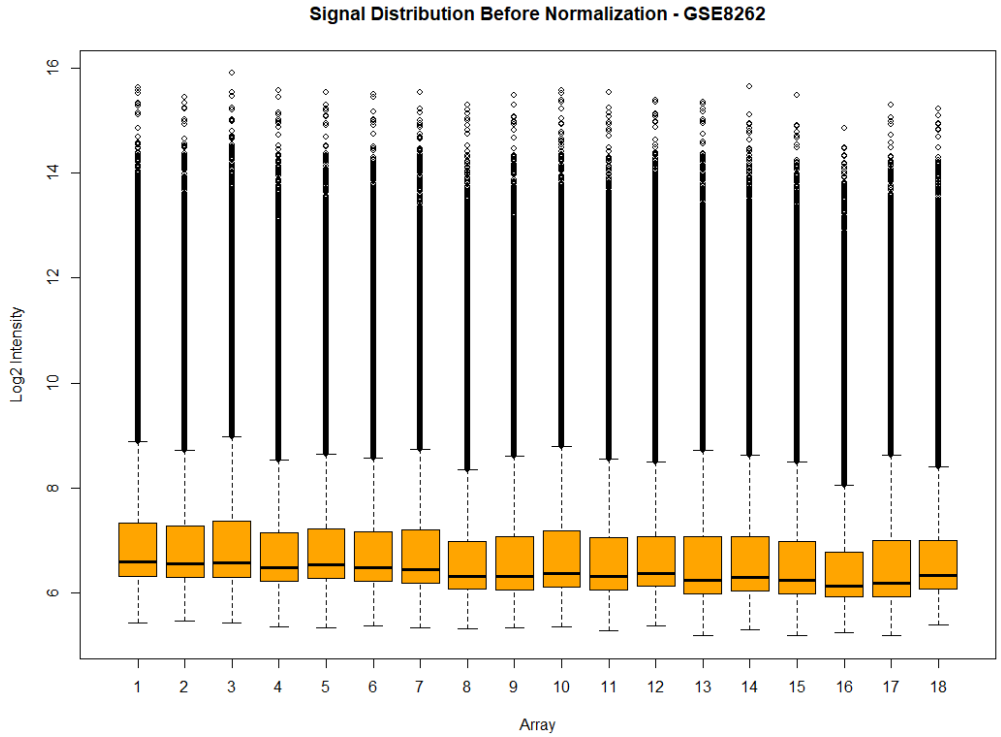


Figure 6. *Signal distribution before normalisation – GSE8262*

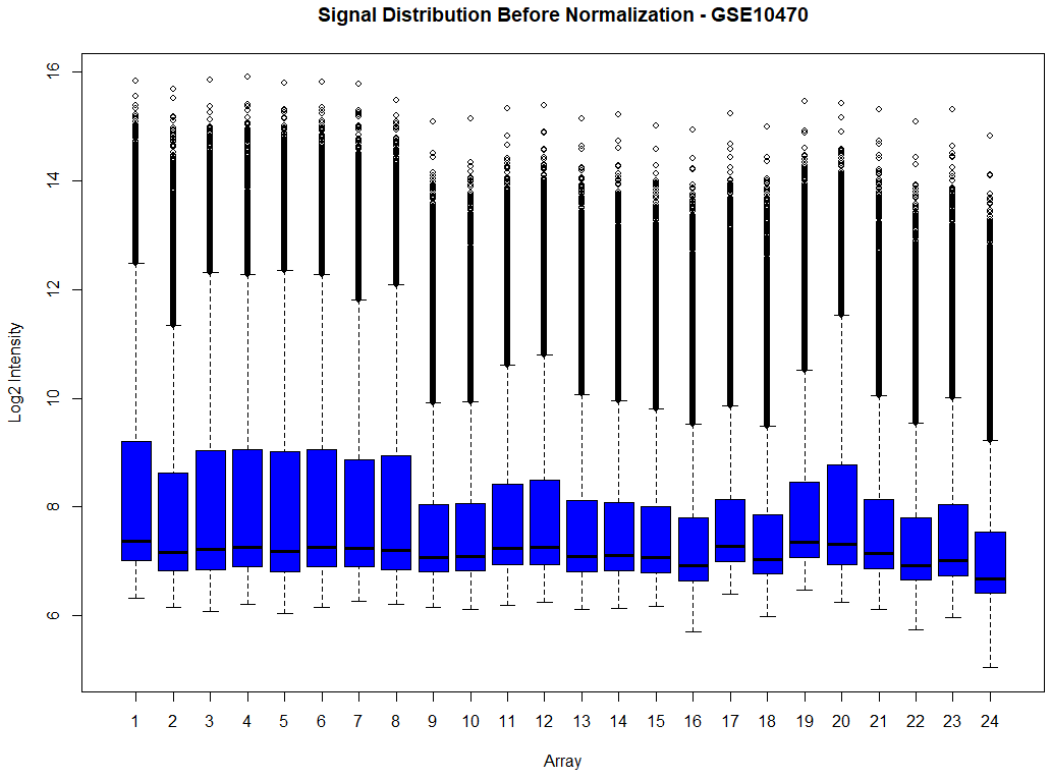


Figure 7. *Signal distribution before normalisation – GSE10470*

The gene expression data in the datasets whose control processes had been completed and are suitable for analysis are clustered and displayed in a heatmap. Clustering heatmaps facilitate the grouping process by clustering the samples. As clustering displays similar expression profiles, this facilitated the evaluation of

our datasets with multiple replicates and large amounts of data. Figures 8 and 9 show the heatmaps before the normalisation process.

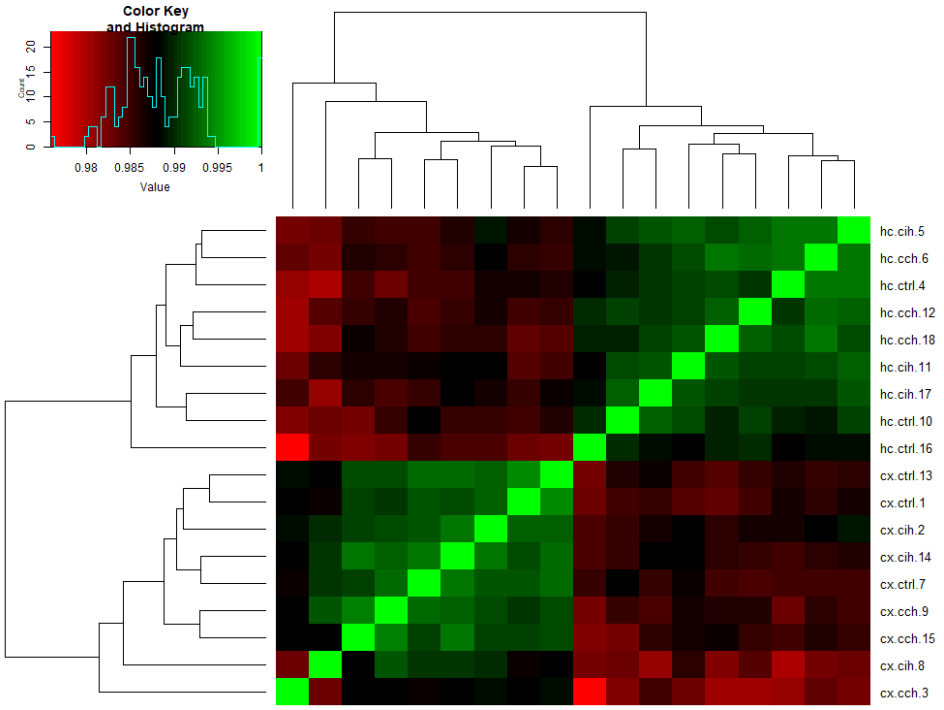


Figure 8. *Heatmap before normalisation – GSE8262*

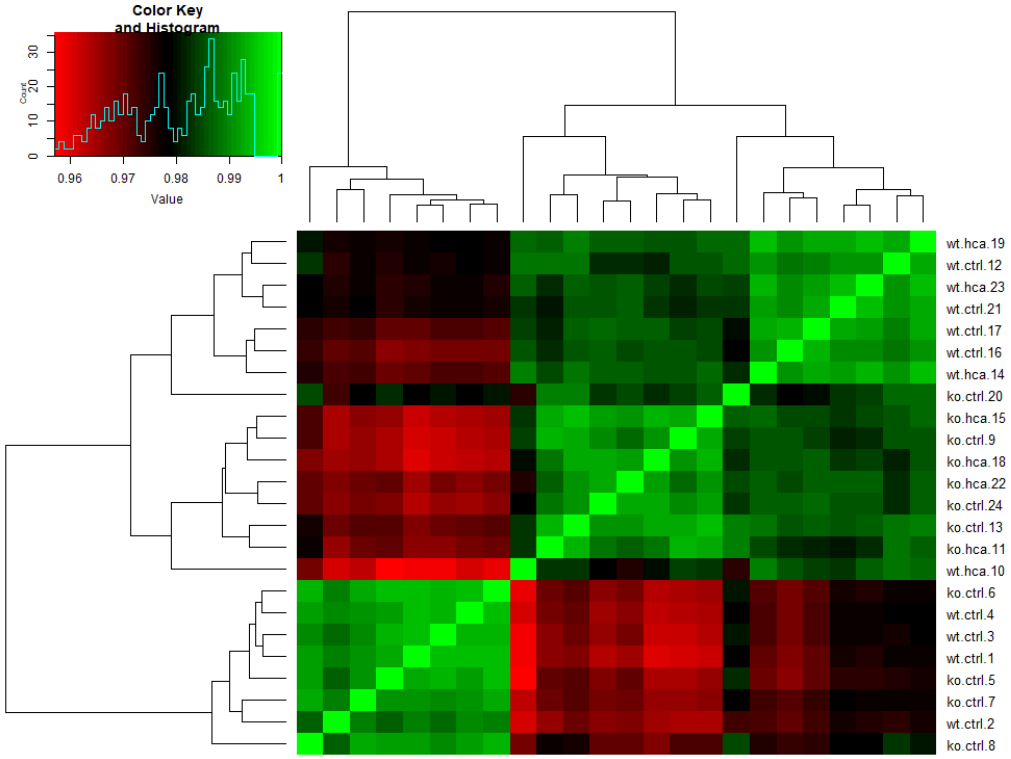


Figure 9. *Heatmap before normalisation – GSE10470*

During normalization, the features in the data set are converted into a specific range or scale. For features with large value differences in the data set, the values can be scaled between 0-1 and can be evaluated more accurately and easily. This makes it possible to compare data with different scales. Normalization processes

are applied to the data sets. The change in signal distribution after normalization is shown in Figures 10 and 11.

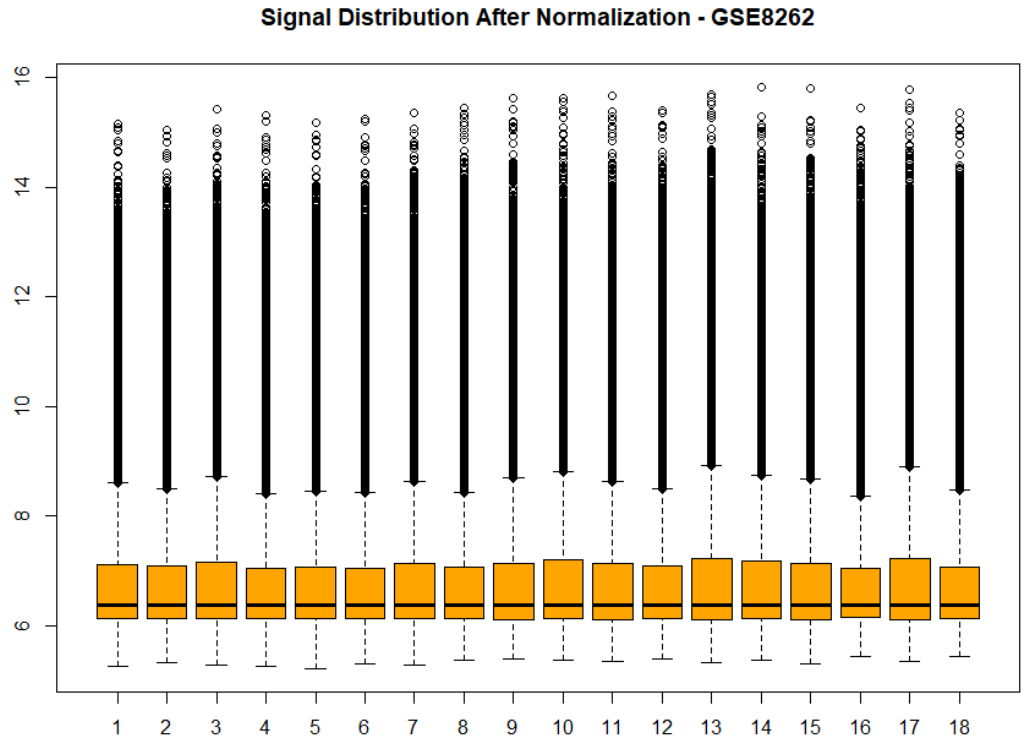


Figure 10. *Distribution after normalization – GSE8262*

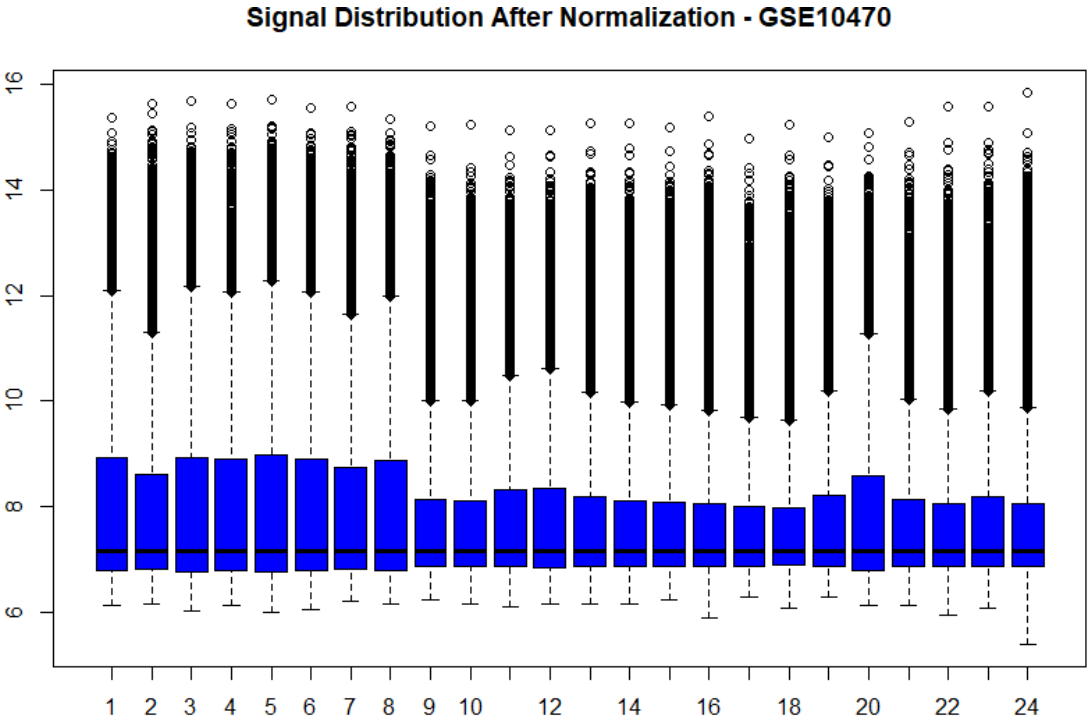


Figure 11. *Distribution after normalization - GSE10470*

After the normalization process was completed, the heatmaps also changed. The heatmaps after normalization are as shown in Figures 12 and 13.

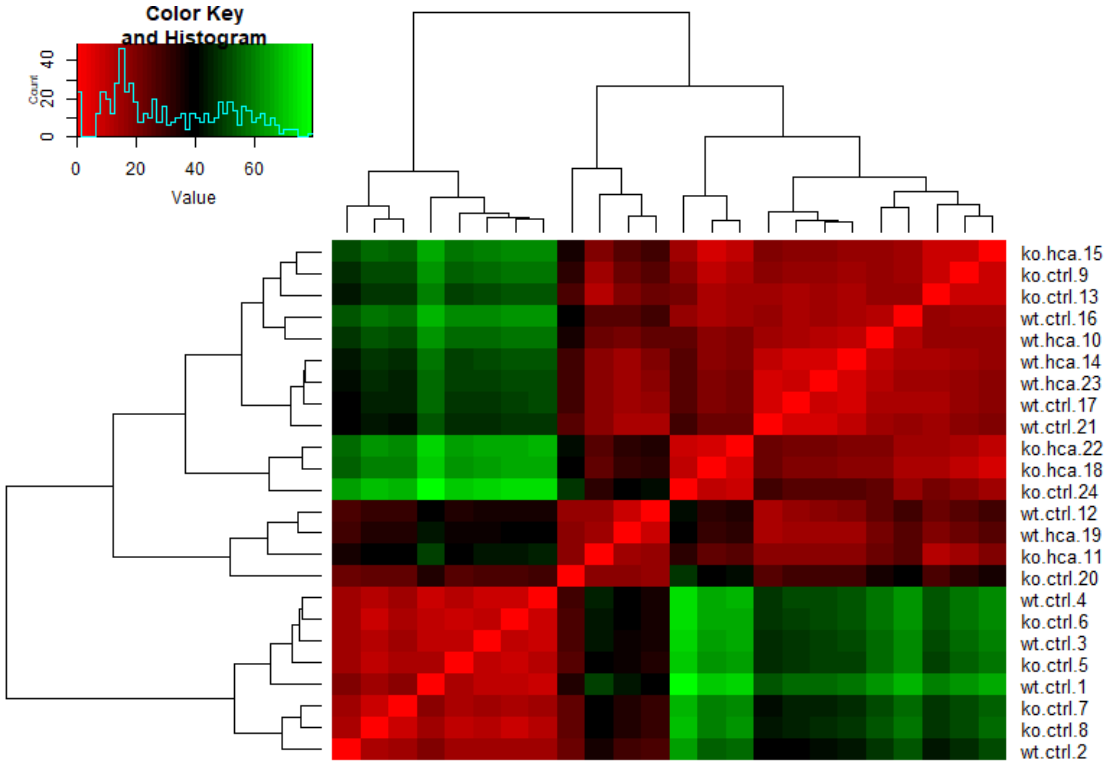


Figure 12. *Heatmap after normalization – GSE8262*

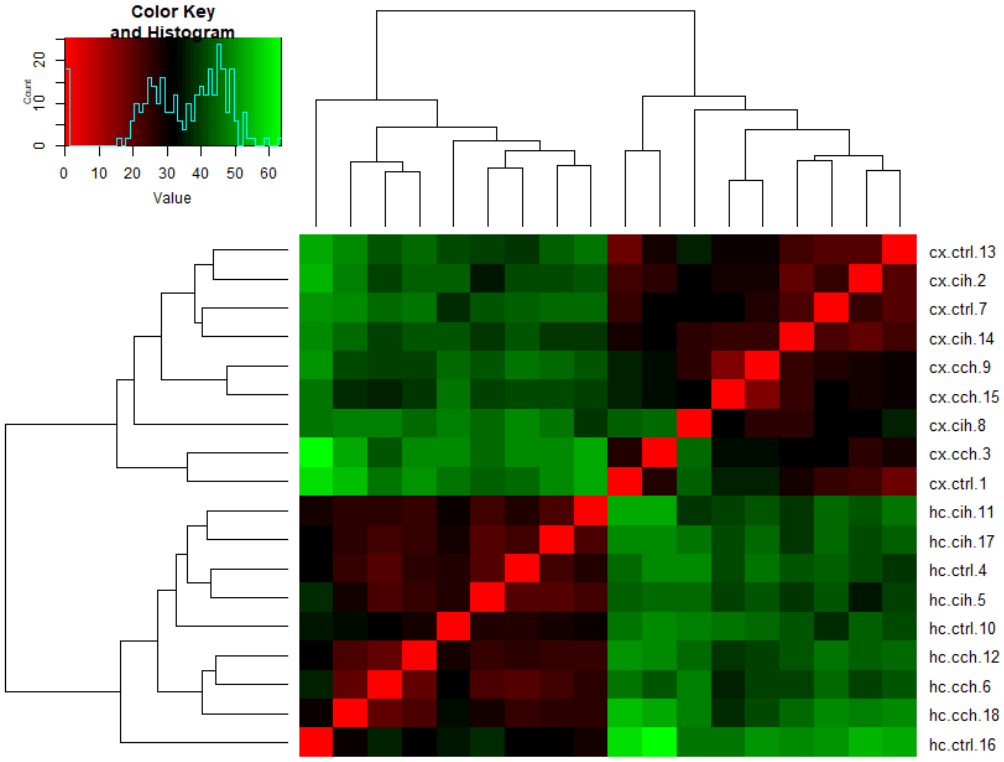


Figure 13. *Heatmap after normalization – GSE10470*

The contrast matrix is prepared to evaluate gene expression changes after normalization. After the design matrix (matrix of control and experimental groups) is prepared, the comparisons to be made are determined and finally a contrast matrix is prepared and comparison groups are prepared. With the help of this matrix; it is possible to test the differences in gene expression between two conditions, to determine the genes

affected by controlling the p-values in multiple comparisons and to determine which genes are down- or up-regulated. The contrast matrices prepared for both datasets are shown in Tables 5 and 6.

Table 5. *Contrast matrix – GSE8262*
CXwsHC: cx.ctrl vs hc.ctrl, cihHC: hc.cih vs hc.ctrl, cihCX: cx.cih vs cx.ctrl, cihHCnotcihCX: (hc.cih-hc.ctrl) vs (cx.cih-cx.ctrl), cihCXnotcihHC: (cx.cih-cx.ctrl) vs (hc.cih-hc.ctrl), cchHC: hc.cch vs hc.ctrl, cchCX: cx.cch vs cx.ctrl, cchHCnotcchCX: (hc.cch-hc.ctrl) vs (cx.cch-cx.ctrl), cchCXnotcchHC: (cx.cch-cx.ctrl) vs (hc.cch-hc.ctrl)

	CXwsHC	cihHC	cihCX	cihHC not cihCX	cihCX not cihHC	cchHC	cchCX	cchHC not cchCX	cchCX not cchHC
cx.ctrl	1	0	-1	1	-1	0	-1	1	-1
cx.cih	0	0	1	-1	1	0	0	0	0
cx.cch	0	0	0	0	0	0	1	-1	1
hc.ctrl	-1	-1	0	-1	1	-1	0	-1	1
hc.cih	0	1	0	1	-1	0	0	0	0
hc.cch	0	0	0	0	0	1	0	1	-1

Table 6. *Contrast matrix – GSE10470*
KOwsWT: ko.ctrl vs wt.ctrl, HCAwt: wt.hca vs wt.ctrl, HCAko: ko.hca vs ko.ctrl, HCAwtnotHCAko

	KOwsWT	**HCAwt**	**HCAko**	**HCAwtnotHCAko**
wt.ctrl	-1	-1	0	-1
ko.ctrl	1	0	-1	1
wt.hca	0	1	0	1
ko.hca	0	0	1	-1

The statistical analyses are performed with normalised data sets processed with a contrast matrix. The FRD (False Discovery Rate) statistical method is used for gene extraction. The p-value, which is frequently used in the FDR, is used to check the probability that the result is more extreme than the observed result. A low p-value indicates that the result is significant. Depending on the content of the data used, different thresholds can be considered significant.

To evaluate the results, the genes are first analysed for $p<0.05$ in the GSE8262 dataset and 1190 genes are classified as affected. When the genes are analysed for $p<0.01$ to increase sensitivity, the number of affected genes is 647. Due to the high number of genes reached, this time the change is analysed for $p<0.001$ and this time 279 genes are reached. That is 279 genes;
*For cchCX, 1 genes are downregulated, while 2 genes are upregulated,
*For cchHC, 1 genes are downregulated, while 2 genes are upregulated,
*For cihCX, 1 gene is down-regulated,
*For CXwsHC 81 genes are downregulated and 191 genes are upregulated.

The statistics of the genes accessed are shown in Figure 14. In the figure, up-regulated genes are shown in green and down-regulated genes are shown in red.

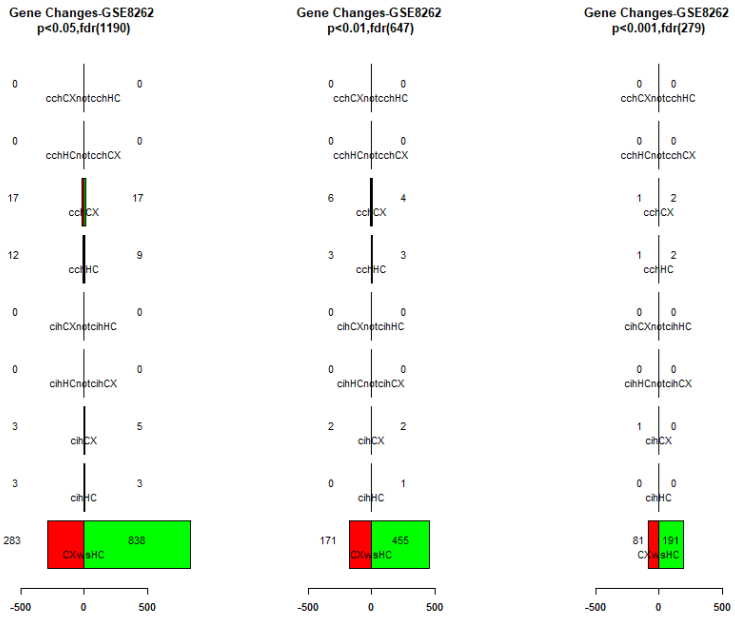


Figure 14. *Gene changes – GSE8262. Green bars indicate the number of up-regulated genes and red bars the number of down-regulated genes*

When the gene changes in the GSE10470 dataset are analysed for p<0.05, 18 genes are found to be affected. These 18 genes;
* For HCAwtnotHCAko, 1 genes upregulated,
* For HCAwt, 3 genes are up-regulated,
* For KOwsWT, 13 genes are down-regulated, while 1 genes are up-regulated.

The statistics of the genes accessed are shown in Figure 15. In the figure, up-regulated genes are shown in green and down-regulated genes are shown in red.

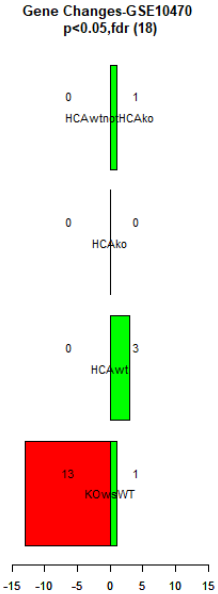


Figure 15. *Gene changes – GSE10470. Green bars indicate the number of up-regulated genes and red bars the number of down-regulated genes*

3. THE RESEARCH FINDINGS AND DISCUSSION

3.1. Pearson Correlation Based Clustering

In microarray analysis, clustering is one of the important steps in the detection of target genes. It is an effective method to see how similar samples in gene expression data are. As with the datasets used in this study, it facilitates the analysis of data with large datasets and many replicates. There are many correlation methods that can be used to analyse microarray data [24].

(a)Pearson Correlation: Measures the linear relationship between variables [25]. The Pearson correlation has 3 value ranges: "positive linear relationship (close to +1)", "negative linear relationship (close to -1)", "no linear relationship (close to 0)". In this study, Pearson correlation was selected for clustering due to its effectiveness in capturing linear relationships in gene expression data.
(b)Spearman Correlation: It measures the ordinal relationship between the variables.
(c)Kendall-Tau correlation: This method, which can provide more precise results than Spearman, measures the sequential relationship between the data.
(d)Hierarchical: It provides an evaluation by using various distance metrics such as Euclidean distance, Manhattan distance, cosine similarity to measure the relationship between data.
(e)K-Means: Measures the closeness of the data.

The SMATP method developed in the study involves a clustering procedure based on Pearson correlation. A correlation matrix is created using the prepared data frame and data matrix. This matrix is then converted into a vector and each element of the matrix became one-dimensional (a flat list). The clustering process is visualized using a heatmap.

3.2. Scale Normalization

Appropriate normalization of data is an important step in obtaining accurate results when analysing microarray data. Normalization can produce more reliable results by reducing technical and biological variation. Normalization ensures that the values are brought into a certain range during data analysis. This makes it possible to make healthy comparisons. There are several normalization methods that can be used when analysing microarray data generated with Illumina BeadStudio. Common normalization methods that can be used in the R programming language when analysing microarray data are as follows:

(a)Scale Normalization: It is used to reduce the signal intensity to the same scale by normalising the samples based on a specific scale.
(b)Quantile Normalization: It is a type of normalization that aims to equalise the intensity distributions in all samples. It makes the data comparable by reducing systematic variations in different samples
(c)Cyclic Loess Normalization: It provides normalization by pairwise comparisons in data that require high accuracy. Although it provides a more precise normalization, it has the disadvantage that the process takes longer.
(d)Robust Spline Normalization: Normalization is a form of normalization that combines background correction and variance stabilisation. It is an effective method to reduce noisy signals and technical variation.

In our SMATP method, scale normalization is preferred. This means that the data is reduced to a standard scale. After the normalization process, the signal distribution and clustering heatmap are taken again and the situation before and after normalization is evaluated by comparing both numerically and visually.

3.3. Statistical Evaluation with Emprical Bayes Method

There are various methods that can be used in the statistical evaluation of microarray analysis. Depending on the type and content of the data used, different methods may be more effective. Some methods that can be used in this process;

(a) eBayes (Empirical Bayes): This is a technique used in differential gene expression analysis. This method, prepared for linear models, uses Bayesian theory to model variations in gene expression. This method, which provides an effective solution with FDR (False Discovery Rate) control, can provide fast results with high-dimensional data [26]. The aim of FDR is to control the rate of false positives in multiple testing [27]. In the FDR statistic, the p-value indicates the probability that the obtained data provide a more extreme result than the observed result. Datasets with a low p-value indicate affected genes. Depending on the dataset, different threshold values can be used.
(b) T-Test: This is a simple and common method that is generally preferred for datasets with two (experimental- control).
(c) ANOVA: This is a preferred method for data with three or more groups.
(d) DESeq/DESeq2: One of the preferred methods for Illumina microarray data.
(e) EdgeR: Another preferred method for microarray and RNA-seq data.

Widely used methods such as DESeq2 and edgeR are primarily designed for RNA-Seq data and rely on count-based statistical models. In contrast, SMATP operates on normalized intensity values, making it more suitable for microarray datasets. The main contribution of SMATP lies not in proposing a novel statistical model, but in integrating multiple analytical steps into a coherent and reproducible workflow. In our SMATP method, the preferred eBayes method is used for the analysis of microarray data and controls are provided for different p-value thresholds depending on the characteristics of the datasets..

3.4. Target Prediction

As a result of the target gene extraction study using the SMATP method, affected genes are extracted for both datasets. The 10 most affected genes of both datasets are shared. The 10 most affected genes for GSE8262 are identified and shown with a heatmap in Figure 16.

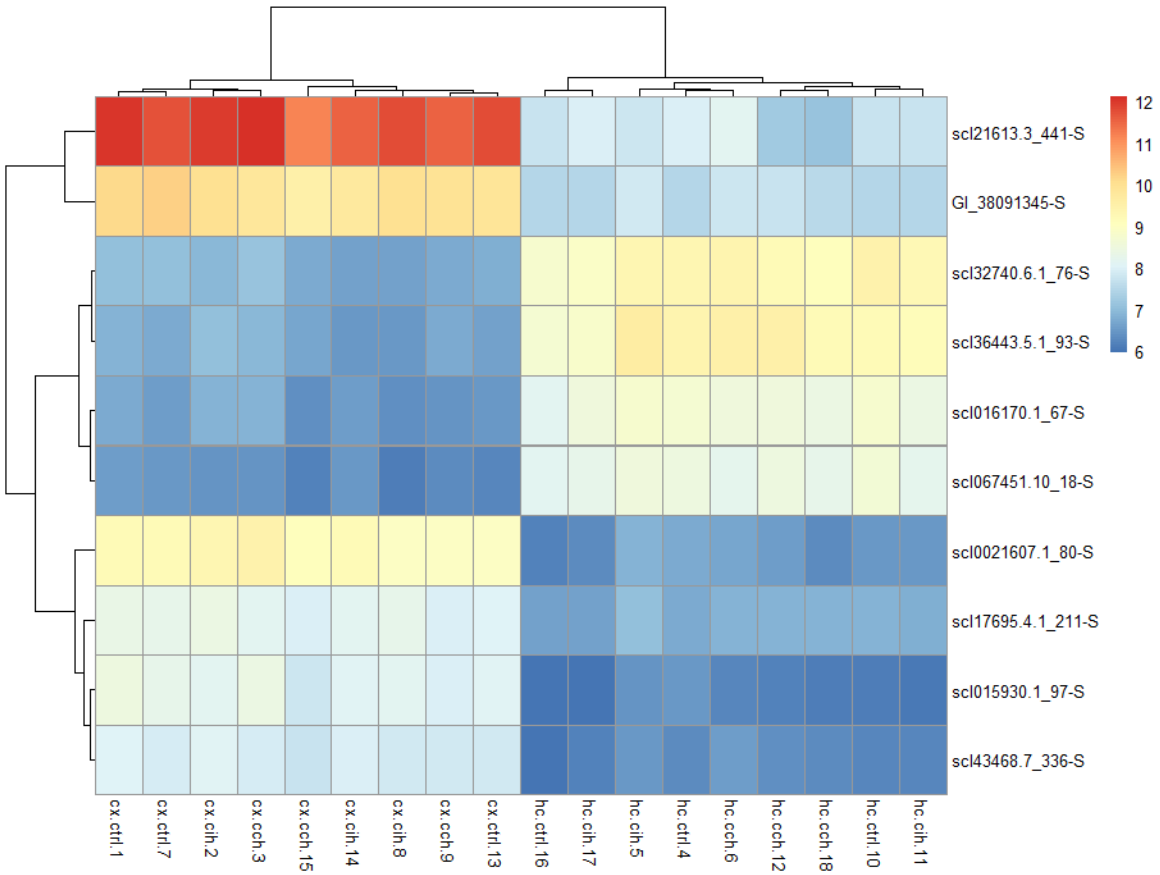


Figure 16. *Most affected genes – GSE8262*

The descriptions of the genes retrieved from the reference gene lists are shown in Table 7.

Table 7. *The 10 most affected genes' definitions – GSE8262*

Target	Symbol	Search Key	Definition
scl21613.3_441-S	Gpr88	NM_022427.1	Mus musculus G-protein coupled receptor 88 (Gpr88), mRNA.
Gl_38091345-S	Rasgef1c	scl41620.15_726	Mus musculus RasGEF domain family, member 1C (Rasgef1c), mRNA.
scl32740.6.1_76-S	Klk8	NM_008940.1	Mus musculus kallikrein related-peptidase 8 (Klk8), mRNA.
scl36443.5.1_93-S	Spink8	scl36443.5.1_93	Mus musculus serine peptidase inhibitor, Kazal type 8 (Spink8), mRNA.
scl016170.1_67-S	Il16	scl016170.1_67	Mus musculus interleukin 16 (Il16), mRNA.
scl067451.10_18-S	Pkp2	NM_026163.1	Mus musculus plakophilin 2 (Pkp2), mRNA.
scl0021607.1_80-S	Tcrb-V8.2	scl0021607.1_80	
scl17695.4.1_211-S	Neu2	NM_015750.1	Mus musculus neuraminidase 2 (Neu2), mRNA.
scl015930.1_97-S	Indo	NM_008324.1	Mus musculus indoleamine-pyrrole 2,3 dioxygenase (Indo), mRNA.
scl43468.7_336-S	Isl1	NM_021459.2	Mus musculus ISL1 transcription factor, LIM/homeodomain (Isl1), mRNA.

The 10 most affected genes for GSE10470 are identified and shown with a heatmap in Figure 17.

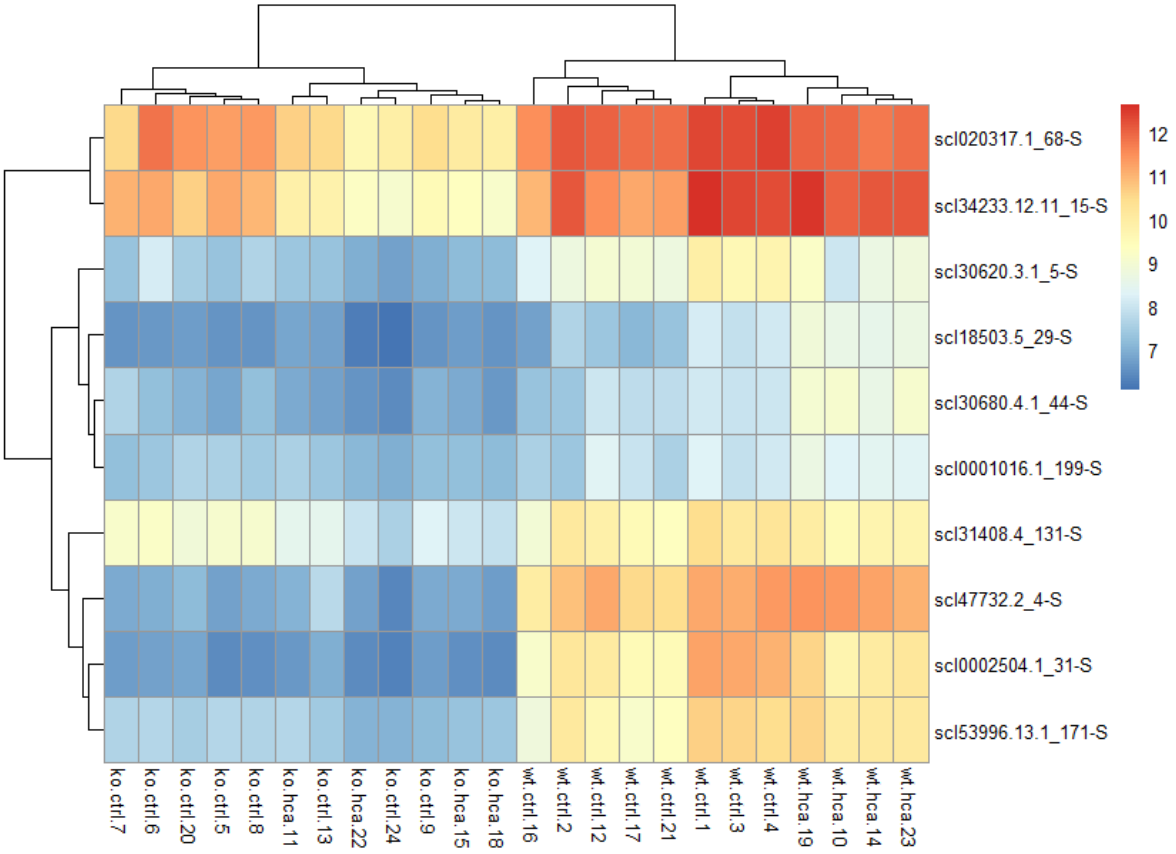


Figure 17. *Most affected genes – GSE10470*

The descriptions of the genes retrieved from the reference gene lists are shown in Table 8.

Table 8. *The 10 most affected genes' definitions – GSE10470*

Target	Symbol	Search Key	Definition
scl020317.1_68-S	Serpinf1	scl020317.1_68	Mus musculus serine (or cysteine) proteinase inhibitor clade F member 1 (Serpinf1) mRNA.
scl34233.12.11_15-S	Slc7a5	scl34233.12.11_15	Mus musculus solute carrier family 7 (cationic amino acid transporter y+ system) member 5 (Slc7a5) mRNA.
scl30620.3.1_5-S	Cox6a2	scl30620.3.1_5	Mus musculus cytochrome c oxidase subunit VI a polypeptide 2 (Cox6a2) mRNA.
scl18503.5_29-S	Trib3	scl18503.5_29	
scl30680.4.1_44-S	Nupr1	scl30680.4.1_44	Mus musculus nuclear protein 1 (Nupr1) mRNA.
scl0001016.1_199-S	Asns	scl0001016.1_199	Mus musculus asparagine synthetase (Asns) mRNA.
scl31408.4_131-S	Atf5	scl31408.4_131	Mus musculus activating transcription factor 5 (Atf5) mRNA.
scl47732.2_4-S	Atf4	scl47732.2_4	Mus musculus activating transcription factor 4 (Atf4) mRNA.
scl0002504.1_31-S	Atf4	scl0002504.1_31	Mus musculus activating transcription factor 4 (Atf4) mRNA.
scl53996.13.1_171-S	Slc7a3	scl53996.13.1_171	Mus musculus solute carrier family 7 (cationic amino acid transporter y+ system) member 3 (Slc7a3) mRNA.

4. RESULTS

In this study, gene expression analysis is performed on microarray data in Illumina BeadStudio format using the R programming language. A method called SMATP is developed to analyse the microarray data and applied to two different datasets. The details of the method are explained in the headings and the results obtained are presented using visualisations. The study can contribute to similar studies by summarising the steps to follow when analysing microarray data generated with Illumina BeadStudio.

The method is successfully applied in two different experiments. The gene information obtained from the study is shared in both datasets. It is expected that this information will make a positive contribution to the diagnosis, treatment, prevention and drug development of diseases. Some of the identified genes (e.g., Atf4 and Slc7a5) have been previously associated with oxidative stress and neuronal response mechanisms, supporting the biological relevance of the findings. A limitation of this study is that SMATP has been validated only on Illumina BeadStudio datasets. Future studies should evaluate its applicability on other microarray platforms and RNA-Seq datasets.

DECLARATION OF ARTIFICAL INTELLIGENCE(AI) USE

Generative AI agents or large language models (LLMs) were used only to improve clarity and quality of the English language presentation of the manuscript.

CONFLICTS OF INTEREST

No conflict of interest was declared by the authors.

REFERENCES

[1] Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., and Lander, E.S., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", Science, 531-537, (1999). DOI: https://doi.org/10.1126/science.286.5439.531

[2] Hanahan, D., and Weinberg, R.A., "Hallmarks of cancer: the next generation", Cell, 646-674, (2011). DOI: https://doi.org/10.1016/j.cell.2011.02.013

[3] Collins, F.S., and Varmus, H., "A new initiative on precison medicine", The New England Journal of Medicine, 793-795, (2015). DOI: https://doi.org/10.1056/NEJMp1500523

[4] Çavdar, Z.Y., Ensari, T., Sener, L.T., and Sertbaş, A., "Makine öğrenmesi tabanlı mikrodizi tekniği ile mikrorna hedef tahmini: araştırma çalışması", Avrupa Bilim ve Teknoloji Dergisi, 39-45, (2022). DOI: https://doi.org/10.31590/ejosat.1220962

[5] Wang, Z., Gerstein, M., and Snyder, M., "RNA-Seq: a revolutionary tool for transcriptomics", Nature reviews genetics, 57-63, (2009). DOI: https://doi.org/10.1038/nrg2484

[6] Bustin, S.A., "Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays", Journal of molecular endocrinology, 169-193, (2000). DOI: https://doi.org/10.1677/jme.0.0250169

[7] Schena, M., Shalon, D., Davis, R.W., and Brown, P.O., "Quantitative monitoring of gene expression patterns with a complementary DNA microarray", Science, 467-470, (1995). DOI: https://doi.org/10.1126/science.270.5235.467

[8] Illumina BeadStudio, https://www.illumina.com. Access Date: 23.09.2024

[9] Affymetrix, https://www.thermofisher.com/tr/en/home/life-science/microarray-analysis.html. Access Date: 16.09.2024

[10] Partex, https://www.partek.com/. Access Date: 16.09.2024

[11] GeneSpring, https://www.agilent.com/en/product/software-informatics/genomics-software-informatics/gene-expression/genespring-gx. Access Date: 16.09.2024

[12] National Center of Biotechnology Information, https://ncb.nlm.nih.gov. Access Date: 22.05.2024

[13] Bioconductor, https://www.bioconductor.org/. Access Date: 22.05.2024

[14] R Programming Language, https://www.r-project.org/. Access Date: 22.05.2024

[15] GSE8262, ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8262. Access Date: 05.06.2024

[16] Zhou, D., Wang, J., Zapala, M.A., Xue, J., Schork, N.J., and Haddad, G.G., "Gene expression in mouse brain following chronic hypoxia: role of sarcospan in glial cell death", Physiol Genomics, 370-379, (2008). DOI: https://doi.org/10.1152/physiolgenomics.00147.2007

[17] Wang, H., Yang, T., Sun, J., Zhang, S., and Liu, S., "SENP1 modulates microglia-mediated neuroinflammation toward intermittent hypoxia-induced cognitive decline through the de-SUMOylation of NEMO", Journal of Cellular and Molecular Medicine, 6463-7099, (2021). DOI: https://doi.org/10.1111/jcmm.16844

[18] Wang, H., Wang, X., Shen, Y., Wang, Y., Yang, T., Sun, J., and Liu, S., "SENP1 modulates chronic intermittent hypoxia-induced inflammation of microglia and neuronal injury by inhibiting TOM1 pathway", International Immunopharmacology, 1-13, (2023). DOI: https://doi.org/10.1016/j.intimp.2023.110230

[19] Çavdar, Z.Y., Sertbaş, A., and Ensari, T., "Hedef gen çıkarımı için makine öğrenmesi tabanlı mikrodizi verilerinin analizi", Bilgisayar Bilimleri ve Mühendisliği Dergisi, 174-183, (2024). DOI: https://doi.org/10.54525/bbmd.1510383

[20] GSE10470, ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE10470 (Access Date: 05.06.2024).

[21] Lange, P.S., Chavez, J.C., Pinto, J.T., Coppola, G., Sun. C.W., Townes, T.M., Geschwind, D.H., and Ratan, R.R., "ATF4 is an oxidative stress-inducible, prodeath transcription factor in neurons in vitro and in vivo", The Journal of Exprerimental Medicine, 1227-1242, (2008). DOI: https://doi.org/10.1084/jem.20071460

[22] Woo, M.S., Ufer, F., Rothammer, N., Liberto, G.D., Binkle, L., Haferkamp, U., Sonner, J.K., Engler, J.B., Hornig, S., Bauer, S., Wagner, I., Egervari, K., Raber, J., Duvoisin, R.M., Pless, O., Merkler, D., and Friese, M.A., "Neuronal metabotropic glutamate receptor 8 protects against neurodegeneration in CNS inflammation", The Journal of Exprerimental Medicine, 1-30, (2021). DOI: https://doi.org/10.1084/jem.20201290

[23] Bloomingdale, P., Meregalli, C., Pollard, K., Canta, A., Chiorazzi, A., Fumagalli, G., Monza, L., Pozzi, E., Alberti, P., Ballarini, E., Oggioni, N., Carlson, L., Liu, W., Ghandili, M., Ignatowski, T.A., Lee, K.P., Moore, M.J., Cavaletti, G., and Mager, D.E., "Systems pharmacology modeling identifies a novel treatment strategy for bortezomib-induced neuropathic pain", Frontiers in Pharmacology, 1-16, (2022). DOI: https://doi.org/10.3389/fphar.2021.817236

[24] Bishara, A.J., and Hittner, J.B., "Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches", Psychological Methods, 399, (2012). DOI: https://doi.org/10.1037/a0028087

[25] Bolstad, B.M., Irizarry, R.A., Åstrand, M., and Speed, T.P., "A comparison of normalization methods for high density oligonucleotide array data based on variance and bias", Bioinformatics, 185-193, (2003). DOI: https://doi.org/10.1093/bioinformatics/19.2.185

[26] Smyth, G.K., "Linear models and empirical bayes methods for assessing differential expression in microarray experiments", Statistical Applications in Genetics and Molecular Biology, (2004). DOI: https://doi.org/10.2202/1544-6115.1027

[27] Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., and Ploner, A., "False discovery rate, sensitivity and sample size for microarray studies", Bioinformatics, 3017-3024, (2005). DOI: https://doi.org/10.1093/bioinformatics/bti448