Application of R software for Microarray data analysis of pediatric versus adult Cornea

T.M. Desy [1], S. Adiga*[2], U. Adiga[3]

[1]ICMR, Nitte-DU, Department of Biochemistry, KS Hegde Medical Academy, Mangalore; [2] Computer Science department, Indian Institute of Information Technology, Hyderabad; [3] Nitte-DU, Department of Biochemistry, KS Hegde Medical Academy, Mangalore, India
* Corresponding author e-mail: shreyas.adiga@research.iiit.ac.in

Journal of Livestock Science (ISSN online 2277-6214) 14: 78-85
Received on 13/1/23; Accepted on 23/2/23; Published on 5/3/23
doi. 10.33259/JLivestSci.2023.78-85

Abstract

R, an open source and open development software project based on the R programming language for analysing and comprehending genetic data, is one of the most well-liked platforms. The workflow of analysing microarray data is shown in this work using the R analysis package, from different types of annotation, normalisation, expression index calculation, diagnostic plots, to pathway analysis for meaningful data visualisation and interpretation. Aim of the study was to analyse the microarray data of pediatric versus adult cornea using R software. Published microarray data gene expression patterns of pediatric and adult cornea were obtained from ncbi geo database. Integrated bioinformatics tools were used to analyze and compare the gene expression patterns. Data were processed using the software R, function and pathway enrichment analysis of differentially expressed genes (DEGs) was carried out using Gene ontology (GO) and KEGG database, protein protein interaction was studied by string database. A total of 33,297 differentially expressed genes (DEGs) were studied.There were no statistically significant differences in any gene between the groups. Genes enriched by more than 1.5 times were studied in more detail. Of the 20 major genes expressed, 11 were down-regulated and 9 were up-regulated. Volcano, mean difference plot, box plot, histogram, Venn's diagram compared the differentially expressed genes. None of the genes were significantly upregulated/ downregulated/ enriched in both the groups. On string analysis, 34 nodes, 8 edges were found. Average node degree was 0.471, average local cluster coefficient was 0.191 and p value for PPI enrichment was 0.326. The study displays the effectiveness of bioinformatics analysis methods in screening potential genes expressed in pediatric versus adult cornea using R software.

Key words: micro array; R software; annotation; normalization; cornea

Introduction

The human eye is one of the special sensory organs. From birth till adulthood, eyes go through a number of changes. Significant modifications are made to the orbital, neurological, and globe dimensions. In the first year of life, the intraocular structures go through developmental changes in order to have a crisp focus of the image on the retina, combined with neurological growth that enables processing of that retinal image. Along with intraocular expansion, the bone orbit and adnexa also expand extraocularly. These physiological changes must be recognised by an ophthalmologist caring for the paediatric patient in order to rule out any pathologic diagnoses. A select few illnesses can also obstruct these typical developmental changes. Cornea has claimed a lot of attention as the gene expression patterns may be different in pediatric age group as compared to adults.

To assess genome-wide mRNA transcript levels, microarrays are frequently used. There are now many databases with large microarray datasets, and new research is being done all the time. Notwithstanding this, there are not many tools available for quick and easy assessment of survey results. Microarray analysis can be difficult for researchers without the right skills and time-consuming for service providers with many users.

On the other hand, service providers who support users in data analysis face the challenge of communicating all generated results and comprehensively explaining the data analysis methods used at each step. DNA chips are now widely used in many life science laboratories. Despite their widespread adoption, it remains difficult to analyze the deluge of data generated by this powerful technology Microarray data analysis is a multi-step process with many published methods for each step. While the research community has yet to agree on a gold standard, certain methods have proven to be more appropriate in certain circumstances (Allison et al, 2006). For one thing, biologists who need to analyze their own microarray datasets may not have the computational and statistical knowledge needed to handle all aspects of a typical analysis workflow.

There are many commercial and freely available tools that can be used to perform the general steps for analyzing microarray data Some tools were created to perform a specific or very limited set of tasks (Saldanha, 2004; Al-Shahrour et al, 2004), while others attempt to cover many of the most important steps in data analysis (Herrero et al, 2003; Kapushesky et al, 2004; Hokamp 2004, Psarros et al, 2005; Romualdi et al, 2005; Sykacek et al, 2005). The former relies on tool-specific input and output data formats, while the latter lacks automatic pipelines. In both cases, several decisions were necessary and little documentation was provided to help interpret the results. Therefore, meaningful analysis with specific expertise is highly recommended. Additionally, some client-server tools (Oinn et al, 2004; Wilkinson et al, 2002) allow visual assembly of bioinformatics tools, as well as internet searches (via web services technology) of publicly available components. The complexity of software installation and maintenance indeed requires IT specialists, the price to pay for such a level of flexibility and computing power. Gene Publisher is currently the only solution available to solve most of the problems listed above (Knudsen et al, 2003). Unfortunately, it is only available as a web service and has strict limitations on the number of samples that can be analyzed simultaneously. Until now, to access a server with no input size limit, the input data had to be contained in a public database.

None of the above methods allow the user to adjust the selection of differentially expressed genes (DEGs) based on experimental design or the number of replicates available, although both criteria should be considered for a selection enlightened method. Moreover, none of the above methods allow comparison of the results obtained with different DEG selection methods Moreover, with the exception of the expression profiler (Kapushesky et al, 2004), none of them allow the incorporation of prior biological knowledge, for example in the form of user-defined gene lists However, the expression profiling tool is based on the assumption that genes from subject families are co-expressed. In practice, genes that do not meet this requirement are removed from the original list, while other co-expressed genes may be added. Finally, with the exception of GenePublisher (Knudsen et al, 2003), none of them guides the user in the interpretation of the results.

For all these reasons, R packages that perform fully automated analysis of microarray data are useful, using methods determined by sample size, experimental design, and the number of replicates available. It is based on standard or generally accepted methods (Allison et al, 2006) and depends on the R implementation of these methods It allows researchers with limited computer skills to easily analyze their own data or data from public repositories The tool can be used to quickly generate a first draft of the analysis, which can then be used to guide a more precise and specific analysis of the relevant results. R is a language and environment for statistical computing and graphics.

Objective of the present study was to analyse the microarray data of pediatric versus adult cornea, deposited in NCBI geo database so as to study their expression patterns using R software.

Methodology

Microarray data

Defining normal and age-dependent HCEnC transcriptomes will help us better understand the functional roles played by the corneal endothelium and provide a basis for the further development of gene-targeted comparisons of normal and dystrophic endothelial transcriptomes. Microarrays were used to

comprehensively characterize human corneal endothelial cell (HCEnC) gene expression, age-dependent differential gene expression, and identify expressed genes mapped to chromosomal loci associated with corneal endothelial dystrophy PPCD1, FECD4 and XECD.

In a study by Frausto et al, eleven corneas were obtained from different eye banks of six pediatric donors (4, 6, 10, 11, 17 and 18 years old) and five adults (53, 56, 57, 64 and 70 years old) (Frausto et al, 2014). Descemet's membrane and corneal endothelium were extracted from donor eye bank corneas as intact coverslips using our techniques adapted to ex vivo donor tissue Total RNA was extracted from the corneal endothelium using TriReagent and purified using the RNeasy cleaning kit. Isolated RNA was tested for integrity using the Agilent 2100 Electrophoresis Bioanalyzer System and found to be of sufficient quantity (approximately 1 microgram) and quality (RNA integrity numbers in the range of 8.3 to 9.0) to analyze the array using the Affymetrix 1.1ST gene chip (Frausto et al, 2014).

Gene Expression Omnibus is a public functional genomics database containing high throughput gene expression data, microarrays and microarrays. In this study, corneal transcriptome microarray data available at https://www.ncbi.nlm.nih.gov/geo/ were downloaded for analysis (Frausto et al, 2014).

Data processing

For processing raw data and screening for differentially expressed genes, statistical software R 4.0.1 (https://www.r-project.org/) and Bioconductor (http://bioconductor.org/biocLite. R) were used. The dataset was calibrated and normalized using the limma package. A central feature of Limma is the ability to fit nonlinear models to gene expression data to assess differential expression (Bhatt, S., et al 2013). Differentially expressed genes were then filtered out using the limma package. The screening thresholds were a p-value of 0.05 and a factor of variation of 1.5 The ggplot2 package was used to create volcano plots from DEGs, and the pheatmap package was used to aggregate large DEGs.

Function and pathway enrichment analysis of DEGs

Gene Ontology (GO, http://www.geneontology.org) is a community-created bioinformatics resource. It uses ontologies to improve biological knowledge by providing information on the function of genes and gene products (Guzman, M. G., et al 2015). The KEGG Database (https://www.kegg.jp/) is a resource for the qualitative interpretation of genome sequences and other biological data, including systematic, genomic, and chemical information, as well as categories human-specific health information (Shahen et al, 2018) Using GO/KEGG enrichment studies related to biological functions and signaling pathways, then re-examined using the Cluster Profiler software package, $p<0.05$ was considered statistically significant.

Protein-protein interaction networks are built, identified, and validated. The analysis of protein-protein interaction (PPI) networks of hub genes plays an important role in predicting the function of interacting proteins. It is a useful tool and can be used to gain a better understanding of cell function and disease mechanisms (Tsai et al, 2013). The STRING database (http://string-db.org) focuses on critical evaluations, predict protein-protein association data by integrating a large amount of known knowledge (Pei et al 2016, Biswal et al 2019). PPI network representation The STRING database is used to compile Cytoscape software. The statistical significance of these genes was validated using DNA microarrays. GEO2R provided the data. GEO2R is an interactive web tool that allows users to identify differences between two or more sets of samples in a GEO sequence, genes that are expressed (Moodie et al, 2018). $P < 0.05$ was considered statistically significant.

Results and Discussion

There were 33,297 differentially expressed genes (DEGs). There were no statistically significant differences in any gene between the groups.Genes enriched by more than 1.5 times were studied in more detail. Of the 20 major genes expressed, 11 were down-regulated and 9 were up-regulated.

According to the scatter plot plotted on the basis of the genes after the screening procedure, the gene expression of most of the genes was similar between the three groups and the control group from high expression level to low expression level. The Lima Volcano plot visualizes differentially expressed genes by showing statistical significance (P-value of -log10) versus magnitude of change (log2 fold change) Genes were significantly differentially expressed at an adjusted p-value threshold of 0.05. However, no genes were found to be significant (Fig 1).

Mean difference (MD) plots created using limma compare log2 fold changes to mean log2 expression values and can be used to identify differentially expressed genes. These genes were not significantly different between psychiatric disorders and controls (Fig 2). Visualize the distribution of values for a selected sample using a boxplot derived from the R boxplot. Code the samples according to the color of the group Viewing the distribution can help you determine if the chosen samples are suitable for differential expression analysis In general, the graph represents log-transformed and normalized data. The resulting data are centered on the median, indicating that they are normalized and comparable (Fig 3).

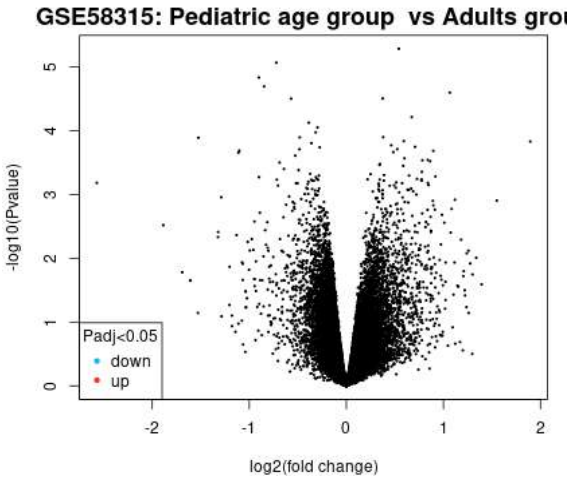

Fig 1: Volcano plot

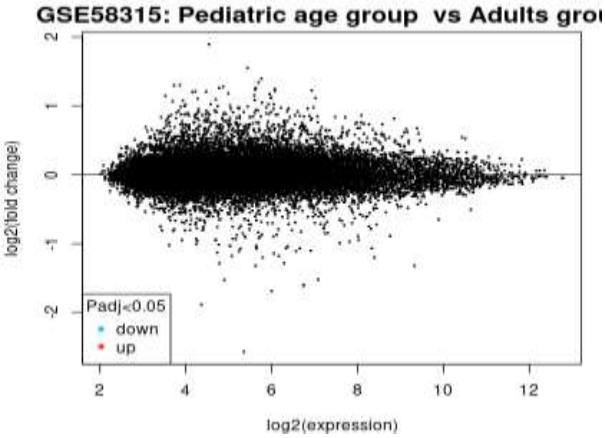

Fig 2: Mean difference plot

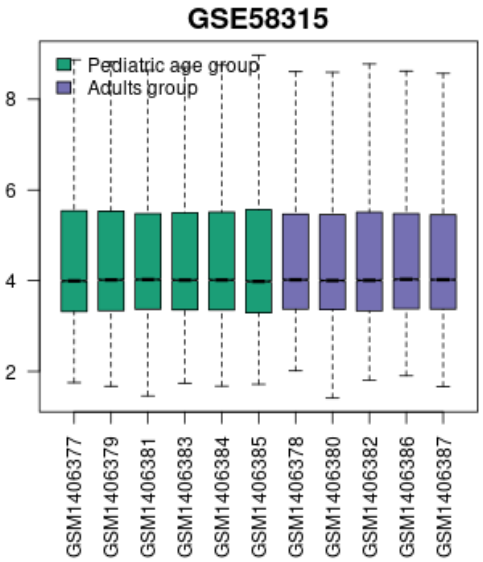

Fig 3: Boxplot

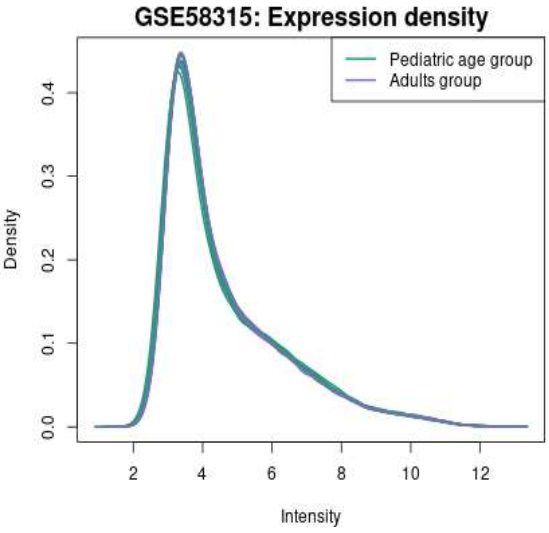

Fig 4: Expression density plot

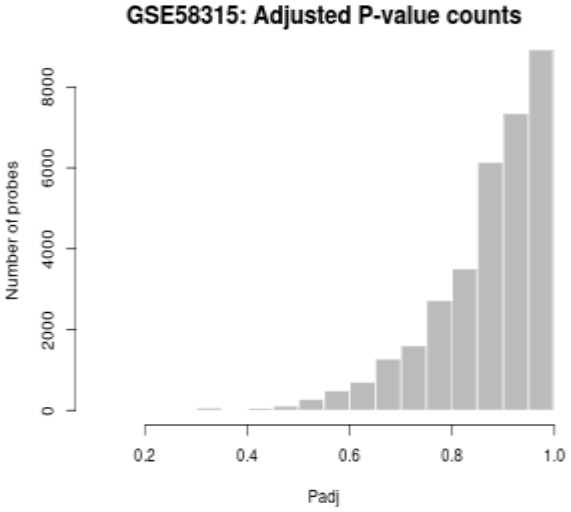

Fig 5: Histogram comparing depression Vs bipolar disorder Vs schizophrenia

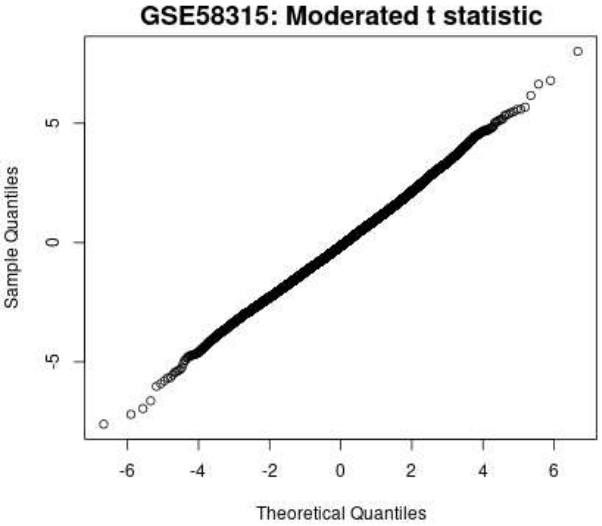

Fig 6: q-q plot

The distribution of selected sample values was visualized using expression density plots generated by R limma. The samples are stained in different groups. This plot is used in conjunction with boxplots to check data normalization before performing differential expression analysis. There was no statistically significant difference in expression density between the different conditions compared to the control (Fig 4).

The histogram created by hist is used to visualize the distribution of P values in the analysis results P values are calculated using all selected comparisons as listed in the most differentially expressed genes table. P values in both cases were not significant (Fig 5). Limma produces a quantile quantile (qqt) plot of the modest t statistic. The quantiles of a data sample was plotted against the theoretical quantiles of the Student's t-distribution. This graph is used to determine the quality of the limma test results (Fig 6). After fitting a linear model, use mean-variance plots generated by R limma (plotSA, vooma) to examine the mean-variance relationship of the expression data. This can indicate if there are large variances in the data This chart can help determine if the Precision Weights option is recommended for illustrating mean-variance trends. Precision weighing improves the accuracy of test results when there is a strong tendency for mean deviation. Each dot corresponds to a gene The red line indicates the mean-variance trend approach that can be used in differential gene expression analysis The blue line is an approximation of the constant variance (Fig 7). Uniform Manifold Approximation and Projection (UMAP), which is derived from umap, is a dimension reduction technique that can be used to visualise how samples are related to one another. The plot indicates the number of nearest neighbours used in the calculation (Fig 8). Venn diagram generated using limma (vennDiagram) is used to explore and download the overlap in significant genes between different groups. There was no gene that was common between the groups suggesting that genes involved are are different (Fig 9).

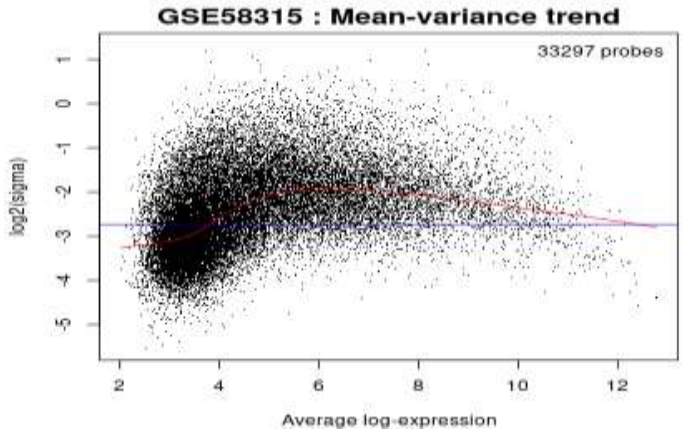

Fig 7: Mean variance trend plot

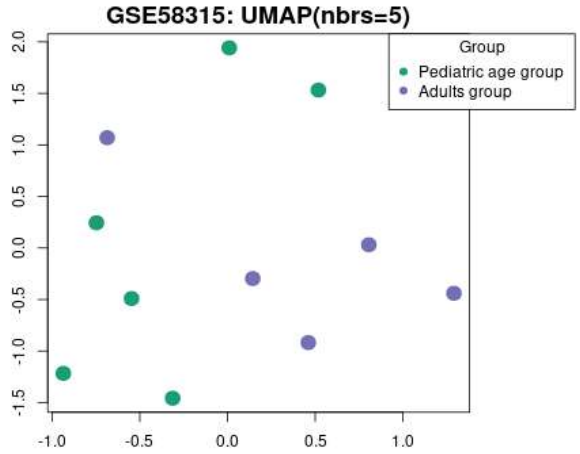

Fig 8: UMAP plot

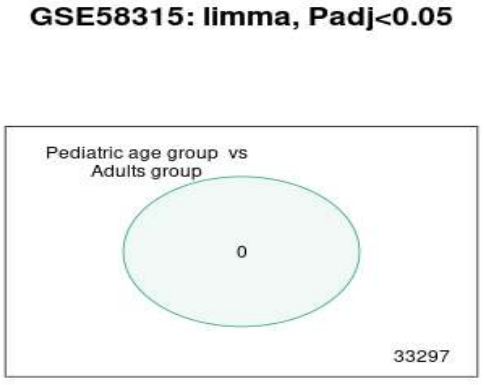

Fig 9: Venn diagram comparing the genes

Enrichment pathway analysis

Differentially expressed genes were defined based on the fold change (1.5- fold or greater) and the P-value ($P<0.05$) of the mean expression values. The number of genes that met this criterion and the pathways involving these genes is as in Table 1.

The majority of the genes exhibiting altered expression turned out to code for the transcription- or translation-related proteins.

STRING analysis

On string analysis, 34 nodes, 8 edges were found. Average node degree was 0.471, average local cluster coefficient was 0.191 and p value for PPI enrichment was 0.326 (Fig 10).

The interaction strength of genes enriched in GO was 0.91 and the false discovery rate (FDR) was 0.0373 Genes were enriched in MAPK signalling pathway, PI3K-Akt pathway, RAP1 signalling pathway, CAMP signalling pathway and various metabolic pathways The FDR is 0.018-0.0259, the intensity is 1.58-2.0, and the biological process is DNA damage response, p53 signal transduction, and intrinsic apoptotic signalling pathway in DNA damage response. With a potency of 0.91 and an FDR of 0.0373, enrichment in cellular components revealed transcriptional regulatory complexes.

KEGG enrichment pathway as represented in Fig 11, explains various pathways associated with the microarray data. Microarrays are an example of a high-throughput technology that allows exploration of genome-wide expression levels as well as identification of changes in gene expression using a zero-scale approach assumption. Over the past decade, these techniques have been used in numerous studies to identify changes in gene expression associated with psychiatric disorders. In addition to hypothesis-driven approaches that primarily rely on the analysis of candidate gene expression levels, transcriptomic studies can help identify novel biomarkers associated with various disorders, which may aid in the development of novel research strategies, intervention and the introduction of personalized medicine. There are not many studies using bioinformatics tools to analyze microarray data deposited by various studies This is an attempt to study the differences in gene expression patterns in the cornea of children and adults.

Gene expression profiling based on microarrays has become an effective technique for the classification, diagnosis, prognosis and treatment of cancer. Microarray-based gene expression profiling using R has become an important and promising method for cancer classification Using primarily two or more datasets as examples of the above findings on pediatric versus adult cornea, this project describes genes involved in the expression of various types of cancer, clearly delineating highly expressed genes that may be important biomarkers.

Conclusion

The study displays the effectiveness of bioinformatics analysis methods in screening potential genes and their expression patterns in pediatric versus adult cornea. R software may be useful in studying the genes and pathways in which those genes have a role may be predicted which may provide promising targets for the treatment of disorders to some extent.

Table 1: Top enriched genes

Genes	Description
IGFBP5	insulin like growth factor binding protein 5
SOD3	superoxide dismutase 3
GPNMB	glycoprotein nmb
CDK6	cyclin dependent kinase 6
RBM33	RNA binding motif protein 33
RRM2B	ribonucleotide reductase regulatory TP53 inducible subunit M2B
KIAA1217	KIAA1217
CSRP2	cysteine and glycine rich protein 2
ITGBL1	integrin subunit beta like 1
DCAF4	DDB1 and CUL4 associated factor 4
SPINT1	serine peptidase inhibitor, Kunitz type 1
HYKK	hydroxylysine kinase
PGPEP1	pyroglutamyl-peptidase I
ZNF146	zinc finger protein 146
MYH14	myosin heavy chain 14
DGCR6L	DiGeorge syndrome critical region gene 6 like
F8	coagulation factor VIII
HIST1H3A	NA
HIST1H4H	NA
KIAA1462	NA

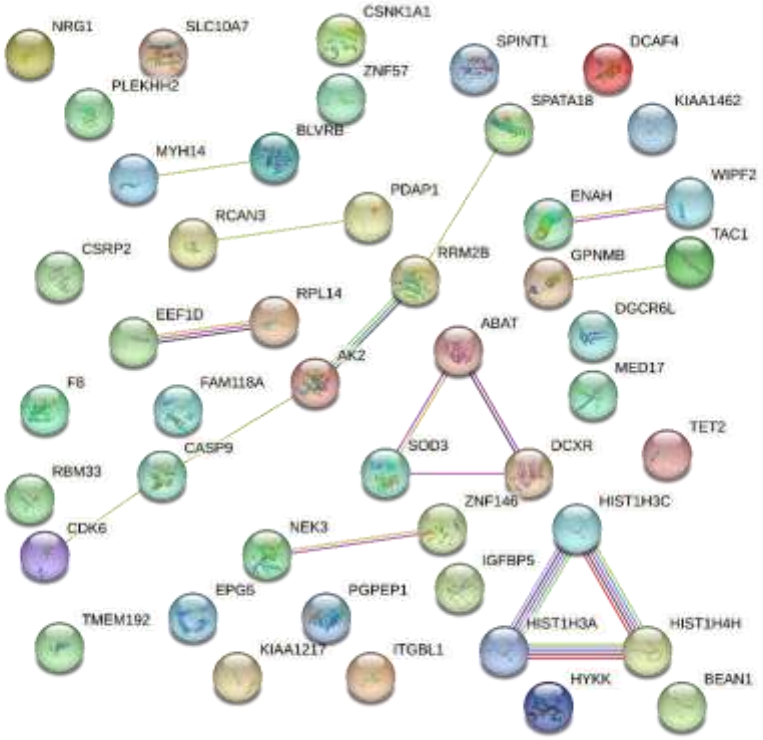

Fig 10: Protein-protein interaction by String analysis

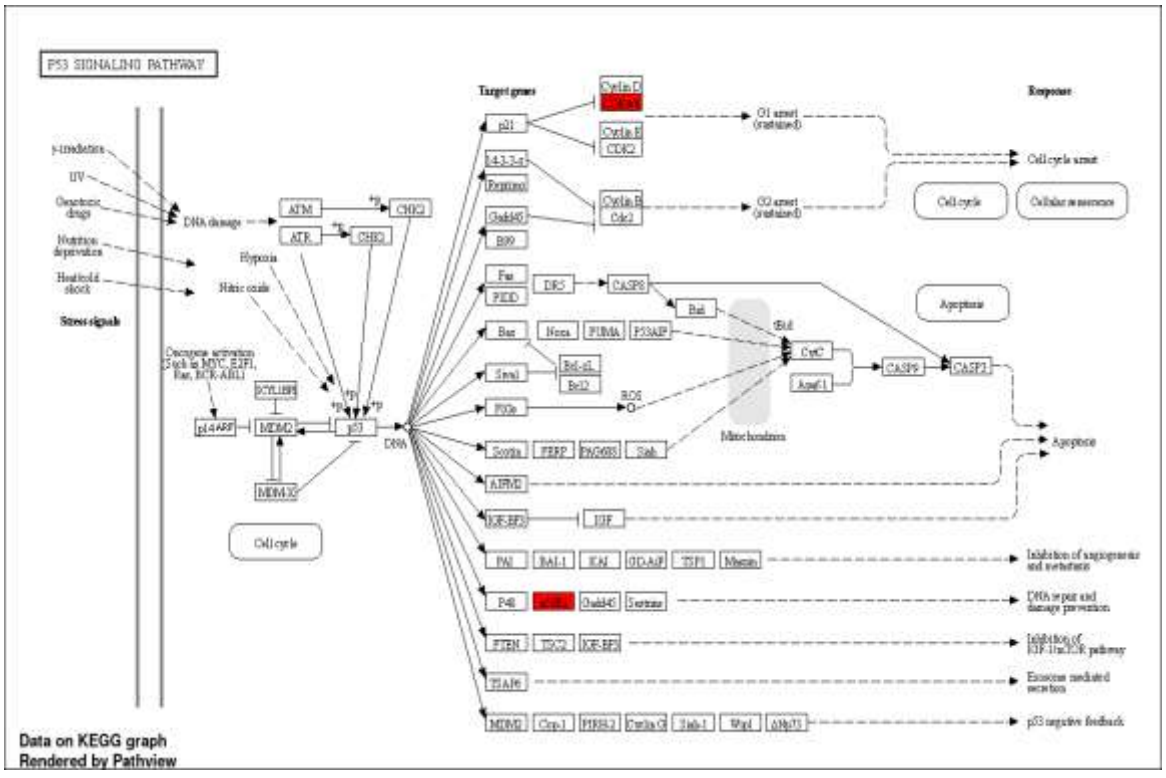

Fig 11: KEGG pathway

References

1) Allison, D.B., Cui, X., Page, G.P., & Sabripour, M. (2006). Microarray data analysis: from disarray to consolidation and consensus. Nature reviews genetics, 7(1), 55-65.
2) Al-Shahrour, F., Díaz-Uriarte, R., & Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics, 20(4), 578-580.
3) Bhatt, S., Gething, P.W., Brady, O.J., Messina, J.P., Farlow, A.W., Moyes, C.L., ... & Hay, S.I. (2013). The global distribution and burden of dengue. Nature, 496(7446), 504-507.
4) Biswal, S., Reynales, H., Saez-Llorens, X., Lopez, P., Borja-Tabora, C., Kosalaraksa, P., ... & Wallace, D. (2019). Efficacy of a tetravalent dengue vaccine in healthy children and adolescents. New England Journal of Medicine, 381(21), 2009-2019.
5) Frausto, R.F., Wang, C., & Aldave, A.J. (2014). Transcriptome analysis of the human corneal endothelium. Investigative ophthalmology & visual science, 55(12), 7821-7830.
6) Guzman, M.G., & Harris, E. (2015). Dengue. The Lancet, 385(9966), 453-465.
7) Herrero, J., Al-Shahrour, F., Diaz-Uriarte, R., Mateos, A., Vaquerizas, J. M., Santoyo, J., & Dopazo, J. (2003). GEPAS: a web-based resource for microarray gene expression data analysis. Nucleic acids research, 31(13), 3461-3467.
8) Hokamp, K., Roche, F.M., Acab, M., Rousseau, M.E., Kuo, B., Goode, D., ... & Brinkman, F.S. (2004). ArrayPipe: a flexible processing pipeline for microarray data. Nucleic acids research, 32(suppl_2), W457-W459.
9) Kapushesky, M., Kemmeren, P., Culhane, A.C., Durinck, S., Ihmels, J., Körner, C., ... & Brazma, A. (2004). Expression Profiler: next generation—an online platform for analysis of microarray data. Nucleic acids research, 32(suppl_2), W465-W470.
10) Knudsen, S., Workman, C., Sicheritz-Ponten, T., & Friis, C. (2003). GenePublisher: Automated analysis of DNA microarray data. Nucleic acids research, 31(13), 3471-3476.
11) Moodie, Z., Juraska, M., Huang, Y., Zhuang, Y., Fong, Y., Carpp, L.N., ... & Gilbert, P.B. (2018). Neutralizing antibody correlates analysis of tetravalent dengue vaccine efficacy trials in Asia and Latin America. The Journal of infectious diseases, 217(5), 742-753.
12) Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., ... & Li, P. (2004). Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics, 20(17), 3045-3054.
13) Pei, H., Zuo, L., Ma, J., Cui, L., Yu, F., & Lin, Y. (2016). Transcriptome profiling reveals differential expression of interferon family induced by dengue virus 2 in human endothelial cells on tissue culture plastic and polyacrylamide hydrogel. Journal of Medical Virology, 88(7), 1137-1151.
14) Psarros, M., Heber, S., Sick, M., Thoppae, G., Harshman, K., & Sick, B. (2005). RACE: remote analysis computation for gene expression data. Nucleic acids research, 33(suppl_2), W638-W643.
15) Romualdi, C., Vitulo, N., Favero, M.D., & Lanfranchi, G. (2005). MIDAW: a web tool for statistical analysis of microarray data. Nucleic acids research, 33(suppl_2), W644-W649.
16) Saldanha, A.J. (2004). Java Treeview—extensible visualization of microarray data. Bioinformatics, 20(17), 3246-3248.
17) Shahen, M., Guo, Z., Shar, A.H., Ebaid, R., Tao, Q., Zhang, W., ... & Wang, Y. (2018). Dengue virus causes changes of MicroRNA-genes regulatory network revealing potential targets for antiviral drugs. BMC systems biology, 12(1), 1-13.
18) Sykacek, P., Furlong, R.A., & Micklem, G. (2005). A friendly statistics package for microarray analysis. Bioinformatics, 21(21), 4069-4070.
19) Tsai, C.Y., Lee, K., Lee, C.H., Yang, K.D., & Liu, J.W. (2013). Comparisons of dengue illness classified based on the 1997 and 2009 World Health Organization dengue classification schemes. Journal of Microbiology, Immunology and Infection, 46(4), 271-281.
20) Wilkinson, M.D., & Links, M. (2002). BioMOBY: an open source biological web services proposal. Briefings in bioinformatics, 3(4), 331-341.