Exploratory Analysis – Differential gene expression analysis

Introduction

Transcriptomics changes have been observed in the developmental stages of organisms, with their life cycle, diseases such as viral infections. The main purpose of investigating transcriptomics of viral infection is studying the molecular mechanism of viral proteins that stimulates the expression of host proteins. These proteins which express under the perturbation of viral genes are mainly responsible for viral entry inside the cellular environment, viral genome replication and cellular apoptosis. The study of these proteins and their interactions with viral protein allows a researcher to develop new targeting therapy against viral infection as well as in the development of vaccines that could be used for long-term immunity.

Previously, transcriptomics studies were mainly performed by quantifying the expression of transcripts by selected genes by amplification methods using PCR which was replaced by microarray chips that can quantify almost all of the genes expressed from the human genome. Currently, the most accurate quantifiable method of gene expression is the RNA-Seq approach using next-generation sequencing technologies. In the current study, differential gene expression (DGE) analysis and enrichment analysis of the transcriptomic data has been performed that were collected from individuals who were characterized as healthy controls, convalescent, dengue fever, and dengue hemorrhagic fever. The aim of this work was to quantify and visualize the most significant genes in expression between these four populations.

Methods

For all of the transcriptomics and enrichment analysis and visualization, R-programming was used. First, transcriptomic data with an ID: GDS5093 was retrieved using *Bioconductors GEOquery* package. The dataset contained a total of 56 samples and 31654 features. The infected group of dengue virus was further divided into convalescent, dengue fever and dengue hemorrhagic fever. Additionally, a healthy control was examined. The expression of the probes in the raw data was plotted and summarized. Furthermore, different probes

corresponding to the same gene were removed by taking the average of the expression value to minimize the size of the dataset.

The samples were grouped by hierarchical clustering using a distance matrix by calculating the relationship among the samples using the “Euclidean” method. The distance matrix was used for hierarchical clustering and the methods; “complete” was used for constructing the dendrogram. The dendrogram was improved using the *dendextend* package of R.

For additional analysis, the top 100 genes were selected based on ranking performed by variance among the genes and a dendrogram was constructed. Furthermore, a heatmap of the top 100 genes was constructed by using the *pheatmap* package in R. Next, principal component analysis (PCA) was performed for samples by taking the genes as features.

For identification of DEGs, differential gene expression (DGE) analysis was performed by using the *limma* package in R. A total of six different studies were conducted for identification of DEGs between all four populations. A cutoff of adjusted P-value (adj. P) <0.05 and logarithmic fold change (logFC) > |0.5| was set to get the most differentially expressed genes. A cutoff of 1 in logFC resulted in graphs showing to not the main of the data. Therefore, a table with the top ranked significantly expressed genes was also created. Using the *EnhancedVolcano* package volcano plots for differential gene expression analysis have been created.

In addition, functional enrichment analysis with Reactome and KEGG pathway was separately performed for each of the relevant studies by using *reactomePA* package and *ClusterProfiler* package in R. For this analysis, only the most significant DEGs were selected using a logFC > |0.5|and adjusted P-value < 0.05 as ranking metrics.

Results

The plotting and summarizing of the data results in equally distributed data and no normalization is needed so the complete linkage can be used (Fig. 1).

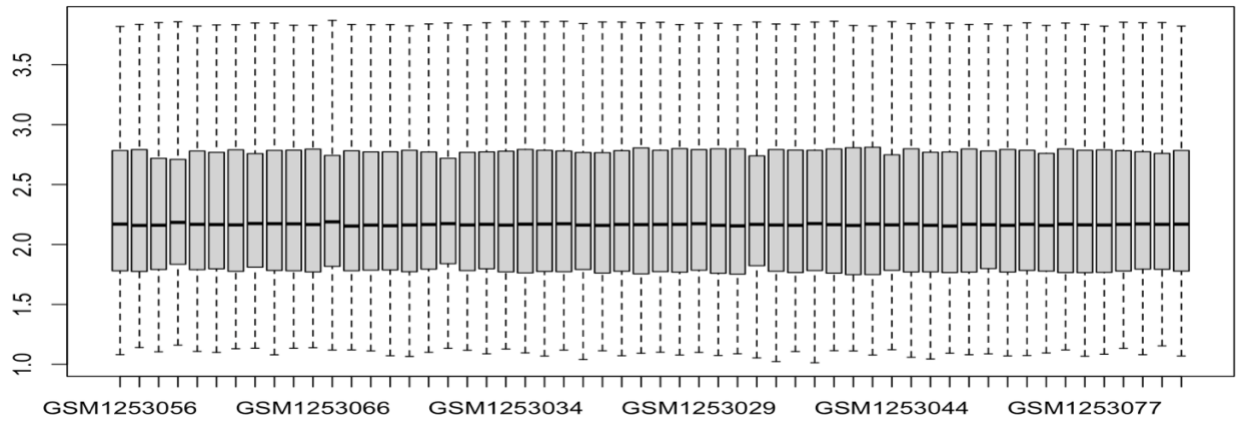

Figure 1 Boxplot summary of the data. Samples are showing normalized features and the complete linkage can be used.

Code Boxplot:

```
boxplot(exprs(eset),outline=FALSE)
```

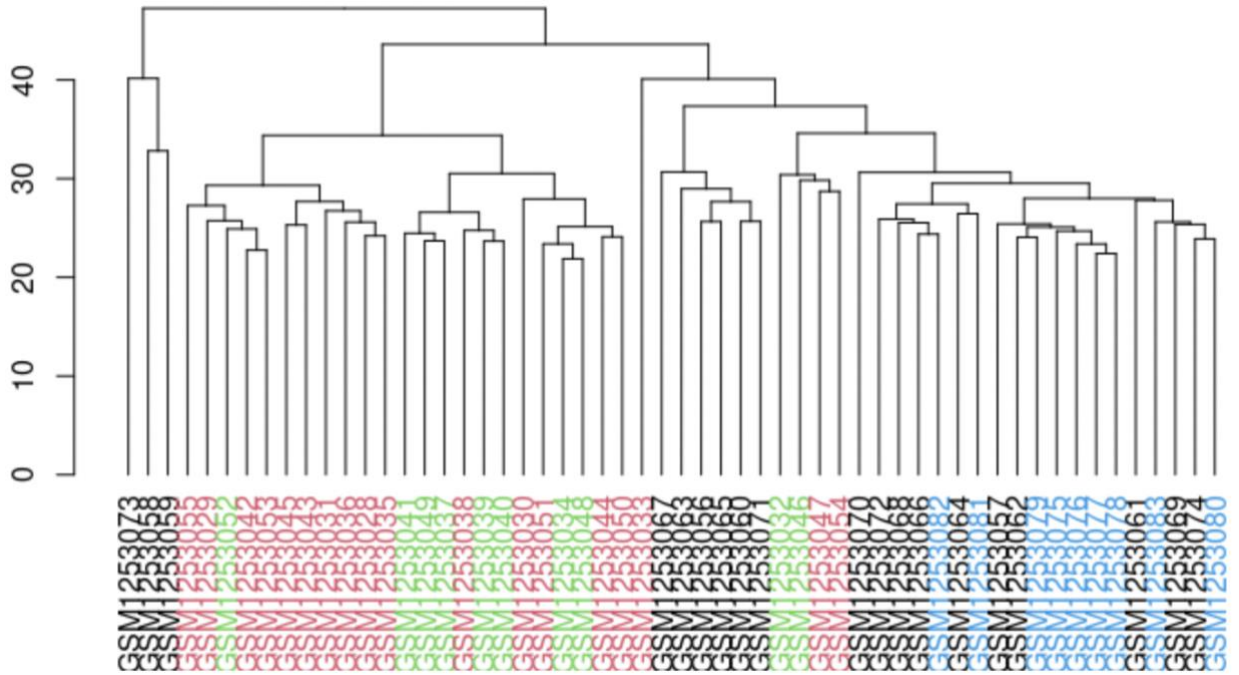

Figure 2 Dendrogram of the samples colored by disease status using "euclidean" method. Black (convalescent), 2 = red (dengue fever), 3 = green (dengue hemorrhagic fever), 4 = blue (healthy).

A Dendrogram clustered the main quantity of healthy and convalescent patients together and the main quantity of dengue hemorrhagic fever and dengue fever patients (Fig. 2). Additionally, a heatmap of the top 100 genes evaluated by highest standard deviation shows unique gene expression patterns for the groups of healthy control and convalescent patients in comparison to dengue fever and dengue hemorrhagic fever patients (Fig. 3). This implicates similar gene expression profiles for these populations for healthy and recovered patients in contrast to patients having the two types of dengue disease.

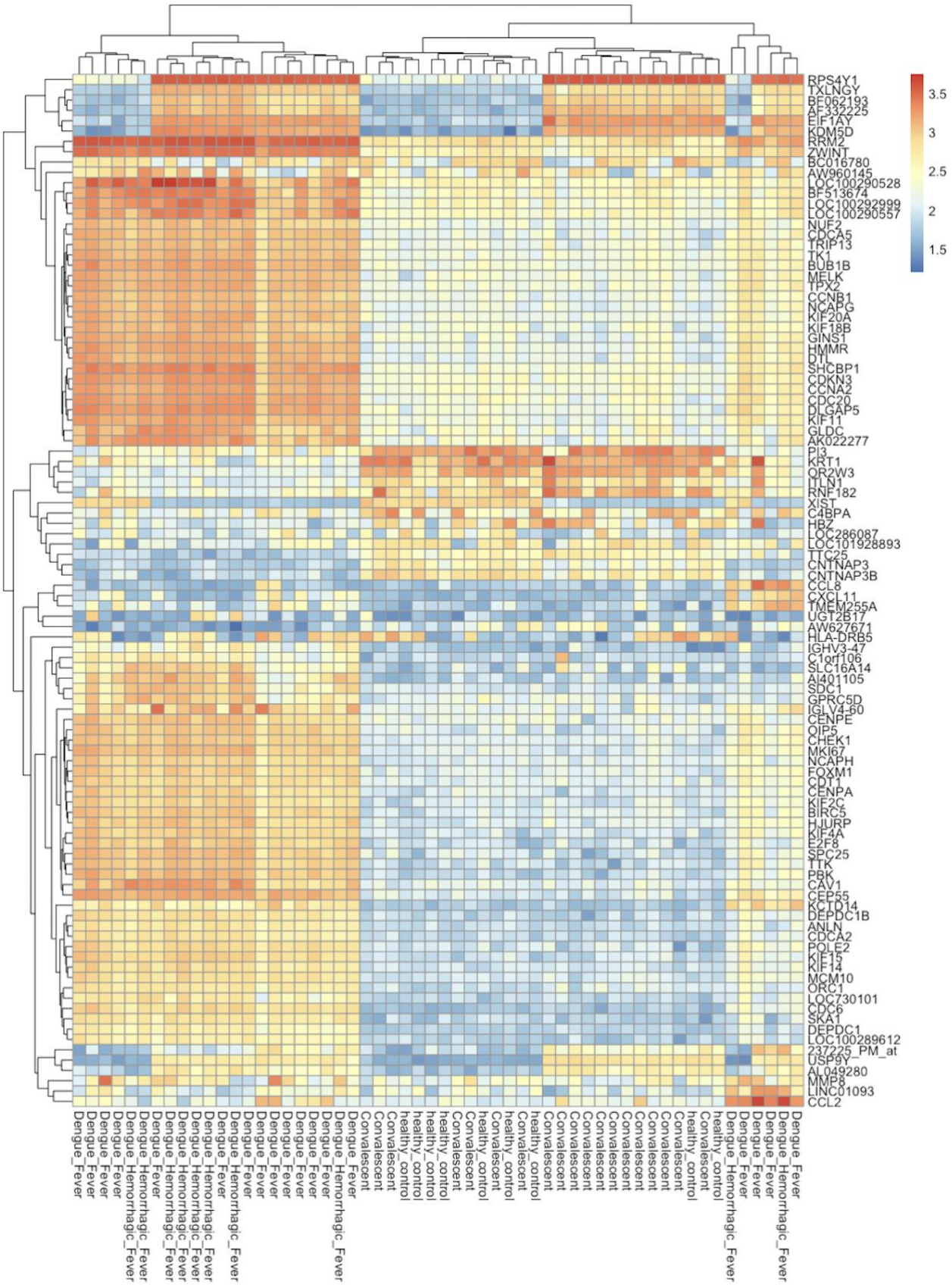

Figure 3 Heatmap of the top 100 genes sorted by highest standard deviation created with pheatmap package in R.

Furthermore, the PCA plot shows the clustering of healthy controls with convalescent samples and dengue fever with dengue hemorrhagic fever samples which also underlines similar gene expression in the underlying clusters (Fig. 4). However convalescent and healthy samples don't cluster completely together, showing that there still is a difference a gene expression. Same goes for dengue fever and dengue hemorrhagic fever, where a clustering is visible but a grouping in each group is still present. So differential gene expression should be considered.

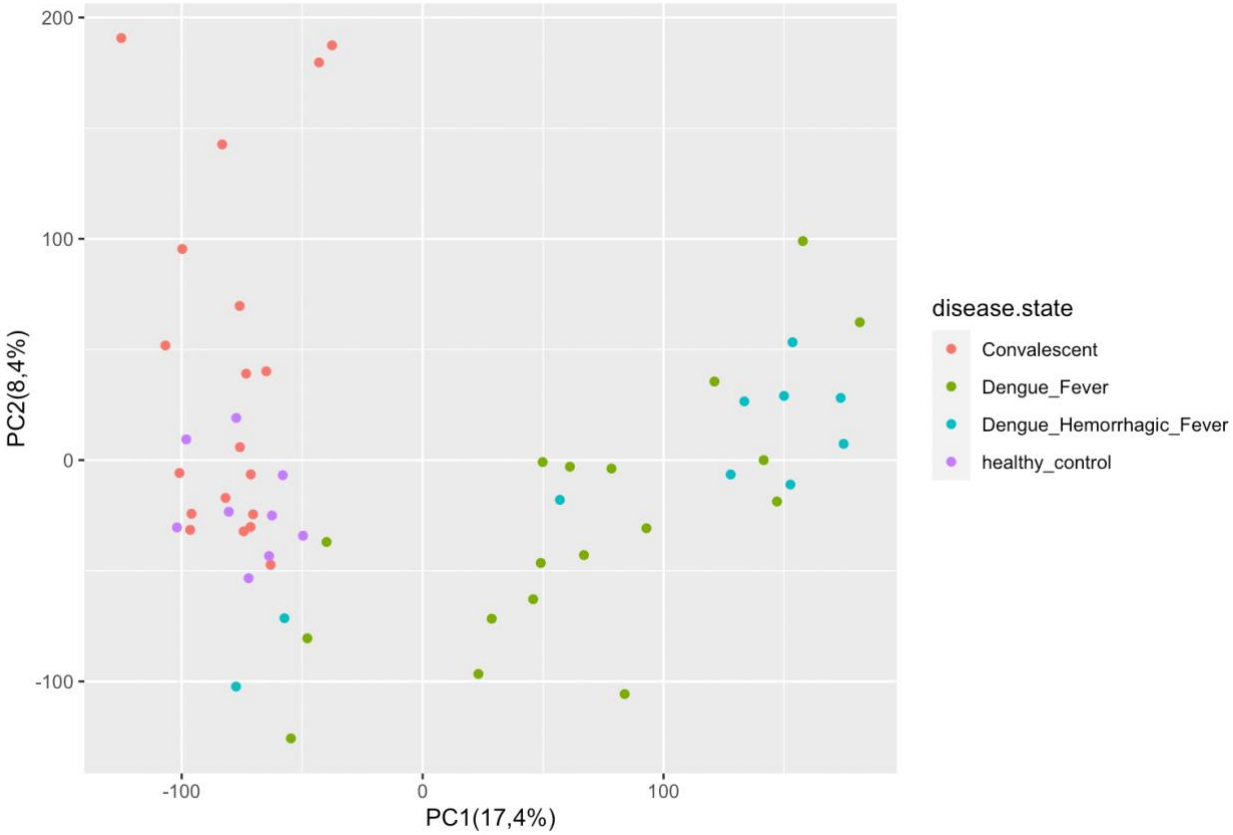

Figure 4 Principal component analysis of all four disease groups. Plotted with ggplot2 package in R

Differential gene expression analysis

The statistically most significant gene expression patterns are found out in the groups of dengue fever vs. healthy controls (Fig. 5B), dengue hemorrhagic fever vs. healthy controls (Fig. 5C) as well as in dengue fever vs. convalescent (Fig. 5D) and dengue hemorrhagic fever vs. convalescent patients (Fig. 5E). In the groups of convalescents vs. healthy controls (Fig. 5A) and dengue fever vs. dengue hemorrhagic fever (Fig 5F) no significant difference in gene expression has been found.

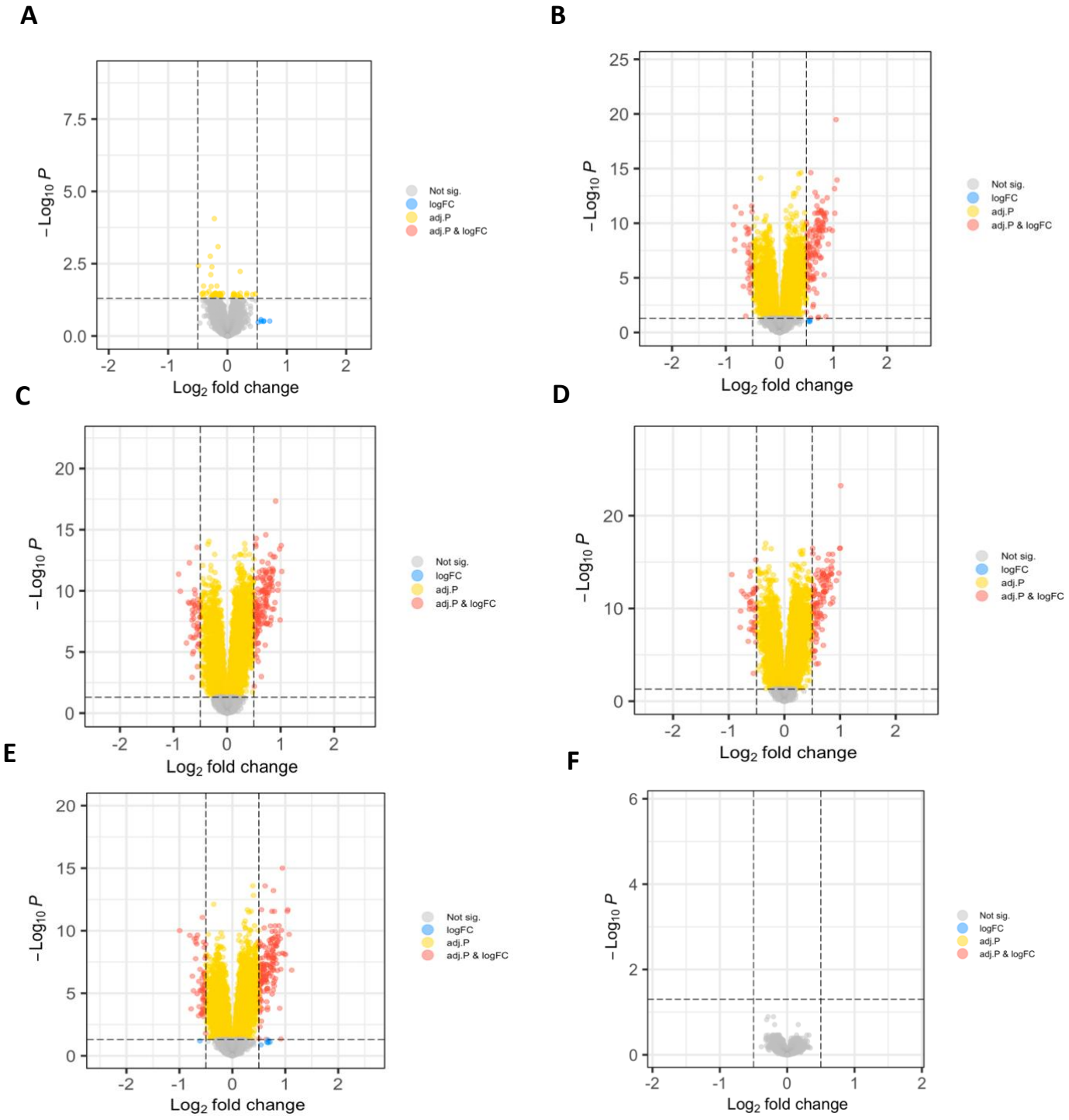

*Figure 5 The volcano plots of **(A)** convalescent vs healthy controls group. Here no differential gene expression is visible **(B)** Dengue fever vs healthy control. Showing significantly upregulated genes in patients with dengue feve. **(C)** Hemorrhagic dengue fever vs healthy control group. Significantly upregulated genes in patients with dengue hemorrhagic fever are visible. **(D)** Dengue fever vs convalescent group. Shows significant upregulated genes in dengue fever patients. **(E)** Dengue hemorrhagic fever vs convalescent. Here another time, hemorrhagic dengue fever inducts genes to be significantly upregulated. **(F)** Dengue fever vs dengue hemorrhagic fever. No statistically significant gene expressions have been measured. Info: Only the genes with the most statistically significant expression pattern have been plotted by colour (red) with a cutoff of adjusted P-value (adj. P) <0.05 and logarithmic fold change (logFC) > |0.5|. Plots have been created by enhancedVolcano package in R.*

Furthermore, the top genes showing the most difference in gene expression between these groups have been summarized in Table 1. Most significantly upregulated genes that were found in all different groups are *KCTD14*, *CDC6*, *CEP55*, *PBK*, *SPC25* and *AI401105*. The most significantly downregulated genes were found to be *CNTNAP3B* and *OR2W3*.

Table 1: List of most significantly differentially expressed genes between all groups

Genes signific. upregulated	**Log2FC DF vs. H**	**Log2FC DHF vs. H**	**Log2FC DF vs. Con**	**Log2FC DHF vs. Con**	**Biological function**
KCTD14	1.052	0.945	1.012	0.906	Protein homooligomerization
CDC6	1.024	1.039	0.995	1.010	Cell division, mitotic cell cycle
CEP55	1.069	1.047	1.000	0.979	Mitotic cytokineses
PBK	1.029	1.072	0.980	1.022	Mitotic cell cycle
SPC25	0.982	1.003	0.872	0.839	Cell cycle, cell division, mitosis
AI401105	0.830	1.121	0.700	0.990	
Genes signific. downregulated					
CNTNAP3B	-0.857	-0.995	-0.768	-0.905	Cell adhesion
OR2W3	-0.827	-0.755	-0.946	-0.874	Olfaction, sensory transduction

Table 1: Table of the most significantly differentially expressed genes detected by differential gene expression analysis. Upregulated genes were selected with pAdj > 0.05 and Log2FC > |1|. To expand the gene list and if genes were found in one of the groups with a LogFC over 1 like AI401105 they were searched as well in other groups if the deviation was not critically. Downregulated genes were selected with pAdj > 0.05 and Log2FC >|0.85|. When they were found in certain groups and the deviation was not critically, they were selected for the other groups as well. The groups of healthy vs. convalescent and dengue fever vs dengue hemorrhagic fever have not been considered since no differential gene expression was visible in the volcano plots. DF = dengue fever, DHF = dengue hemorrhagic fever, H = healthy control, Con = convalescent patients.

Functional Enrichment Analysis

Functional enrichment analysis between the groups of dengue fever and healthy controls (Fig 6) and dengue hemorrhagic fever (Fig 7) turned out to be very similar. Both groups present a gene set with main biological functions in cell cycle checkpoints including phases of mitotic cell cycle like prometaphase, anaphase, and metaphase as well as mitotic spindle checkpoints. In comparison to that the functional enrichment analysis of reactome pathway for the groups of dengue fever vs. convalescent and dengue hemorrhagic fever vs. convalescent additionally included high gene ratios for M phase in mitotic cycle (Fig. 8 and 9).

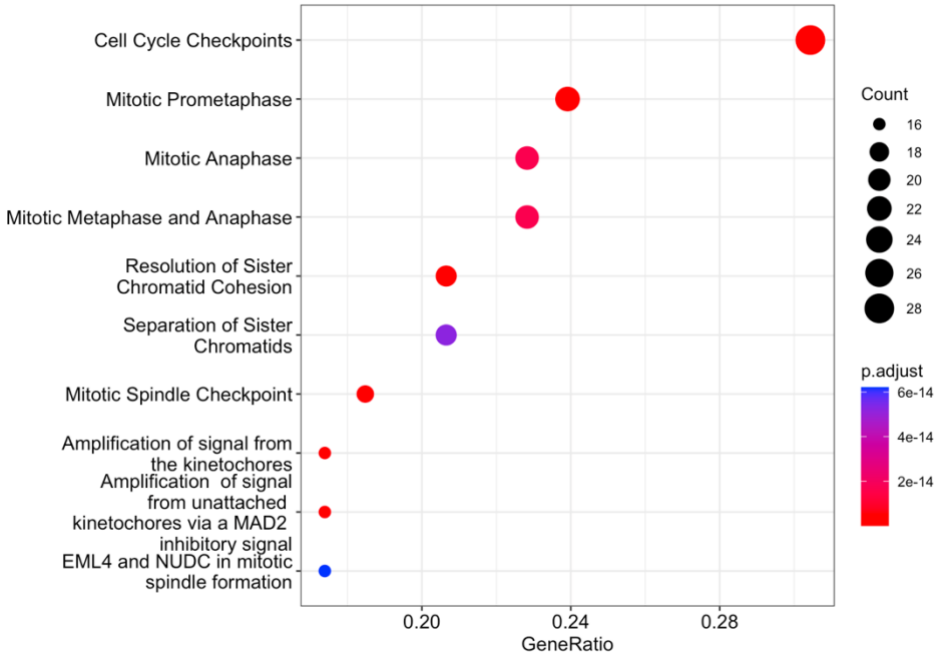

Figure 6 Functional enrichment analysis by reactome pathway of dengue fever vs healthy control group using ReactomePA package in R (Code in Supplemental 8).

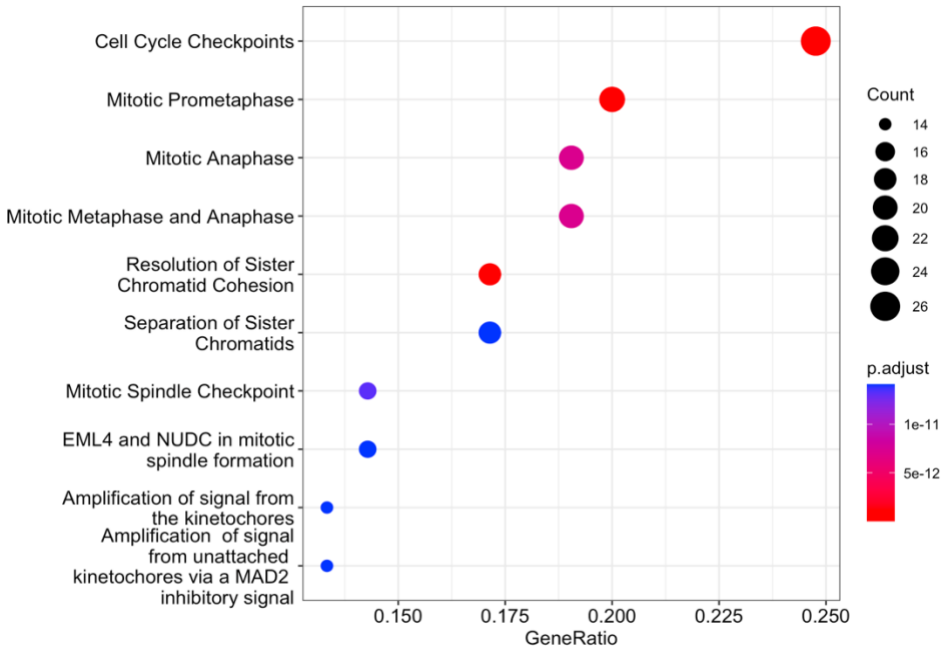

Figure 7 Functional enrichment analysis by reactome pathway of dengue fever vs healthy control group using ReactomePA package in R (Code in Supplemental 8).

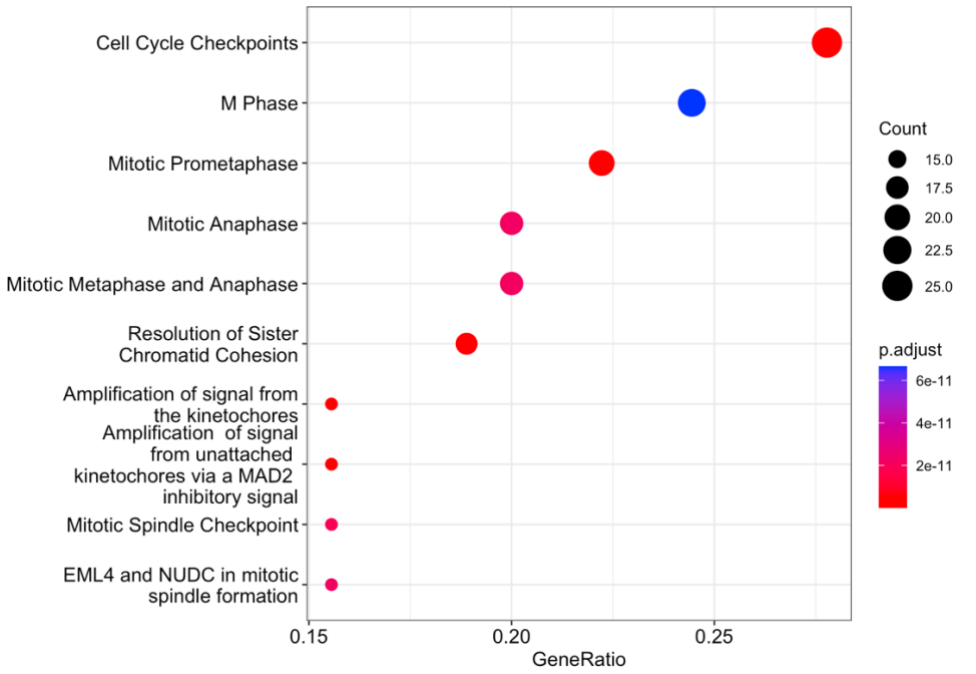

Figure 8 Functional enrichment analysis by reactome pathway of dengue fever vs convalescent group using ReactomePA package in R (Code in Supplemental 8).

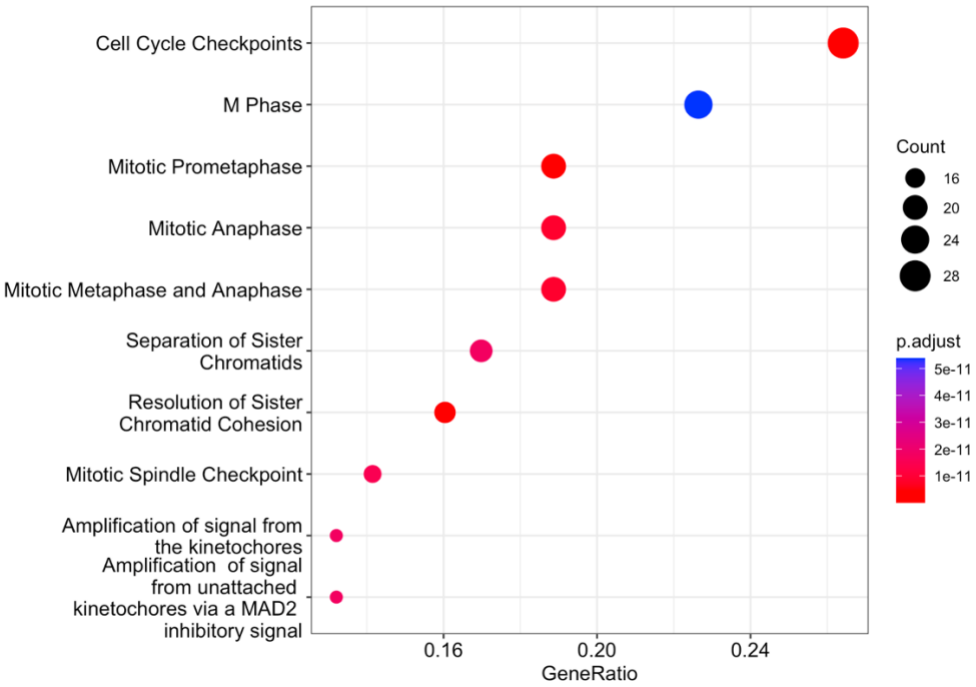

Figure 9 Functional enrichment analysis by reactome pathway of dengue hemorrhagic fever vs convalescent group using ReactomePA package in R (Code in Supplemental 8).

Moreover, a KEGG pathway analysis for functional enrichment analysis was performed giving similar results with most differentially expressed genes of the gene set were having functions in the cell cycle. However, a small set of genes was conducted to be part of cellular senescence and also Human T-Cell leukemia virus 1 infections (Figure 9-12). Other pathways the gene set

got described to included pyrimidine metabolism and p53 signaling pathway as well as oocyte meiosis. Similarly, to the reactome pathway analysis the functional enrichment analysis of the KEGG pathway mostly covers the same pathways for the different groups. But a difference could be investigated between dengue fever vs convalescent group where also functions in the pathway of viral carcinogenesis could be detected (Figure 12).

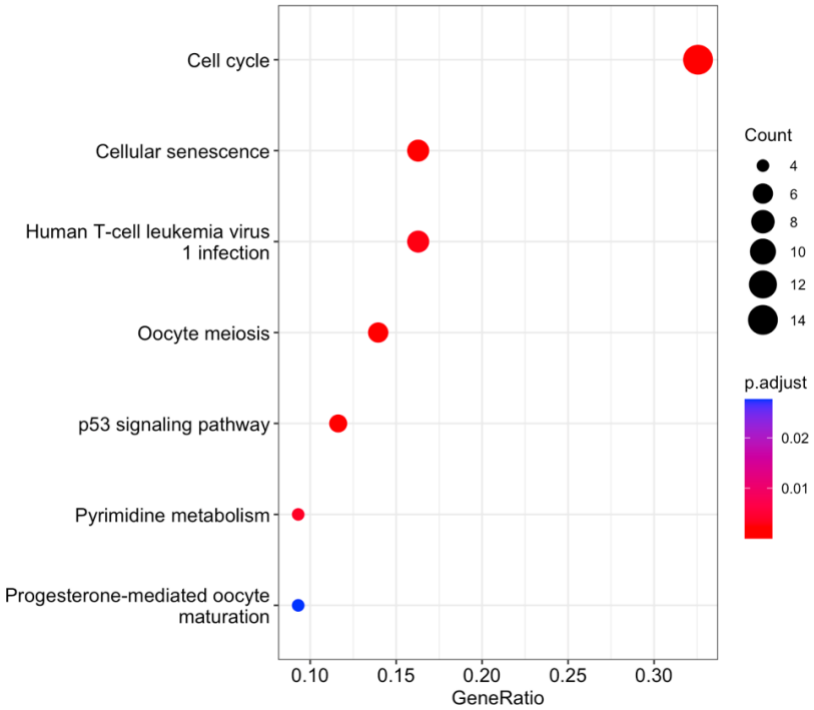

Figure 10 Functional enrichment analysis by KEGG pathway of dengue fever vs healthy group using Cluster Profiler package in R (Code in Supplemental 7).

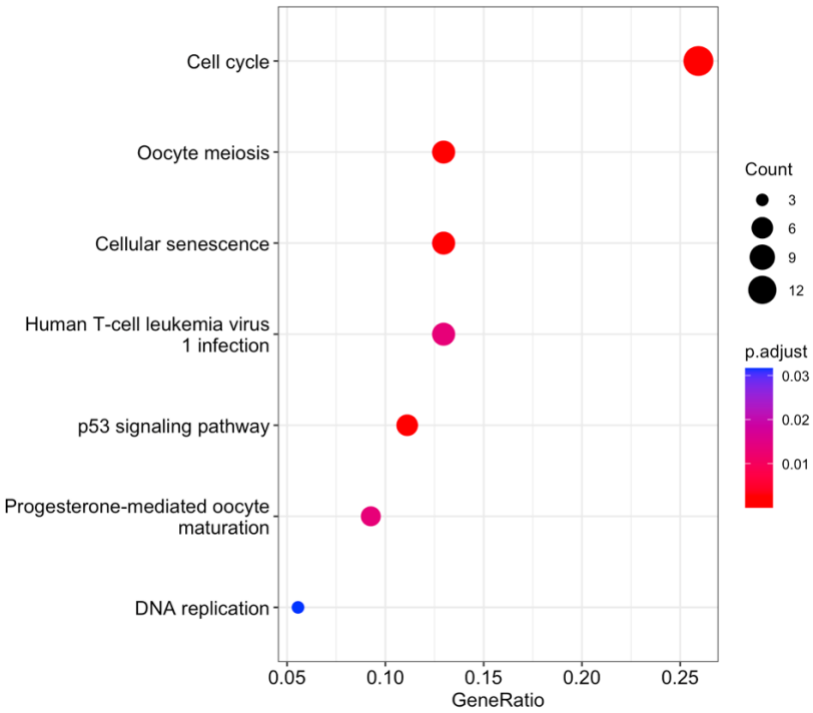

Figure 11 Functional enrichment analysis by KEGG pathway of dengue hemorrhagic fever vs healthy group using Cluster Profiler package in R (Code in Supplemental 7).

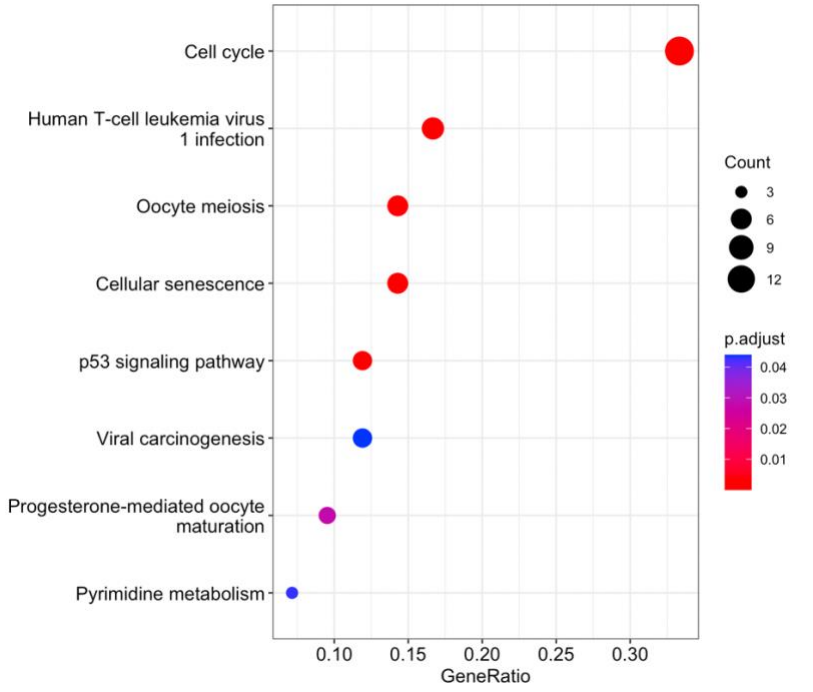

Figure 12 Functional enrichment analysis by KEGG pathway of dengue fever vs convalescent group using Cluster Profiler package in R (Code in Supplemental 7).

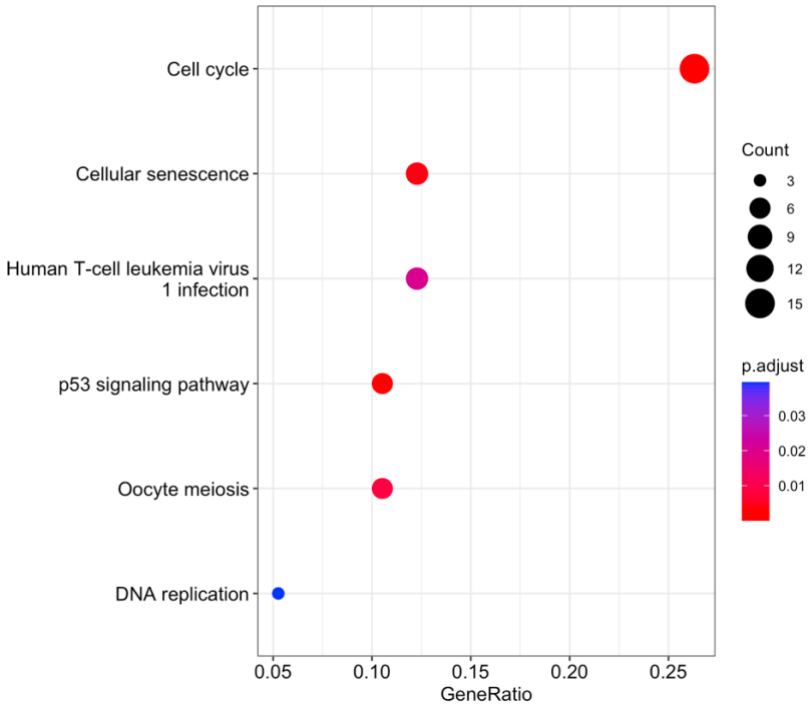

Figure 13 Functional enrichment analysis by KEGG pathway of dengue hemorrhagic fever vs convalescent group using Cluster Profiler package in R (Code in Supplemental 7).

Furthermore, a gene set enrichment analysis performed with the *GSEA* function was conducted and showed results in more specific proteins the virus RNA is responding to. Here, for all groups of dengue fever or convalescent fever vs healthy or convalescent patients the dotplots show significantly those patients gene profiles responding on proteins corresponding to DNA replication processes (ERF2 Targets) and supply of the cell (oxidative phosphorylation) as well as mitochondrial replication (MYC targets). Moreover, also protein synthesis (MTORC1 signaling), heme metabolism and as seen before the cell cycle (G2M checkpoints) are getting manipulated (Fig. 14-17).

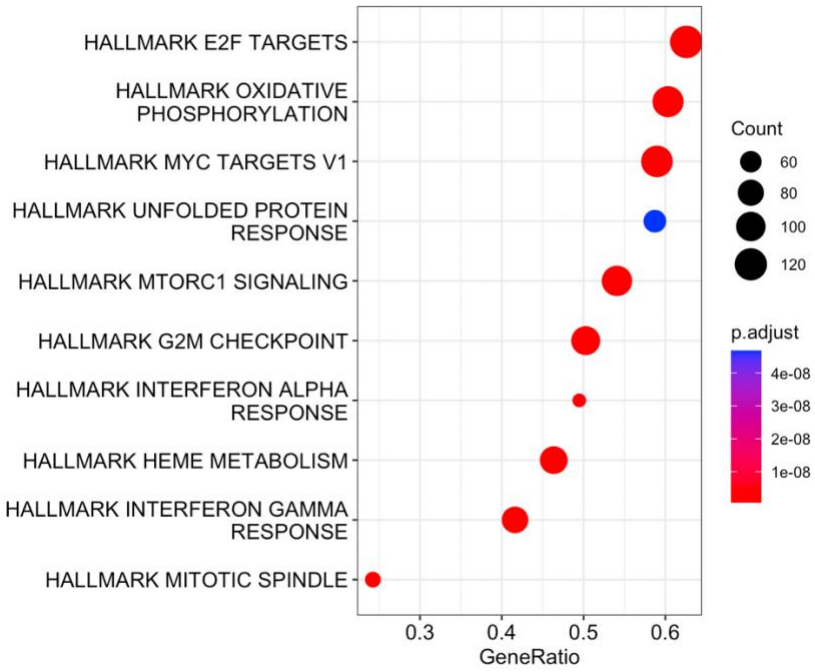

Figure 14 Functional enrichment analysis by GSEA function of dengue fever vs healthy group using Cluster Profiler package in R (Code in Supplemental 6)

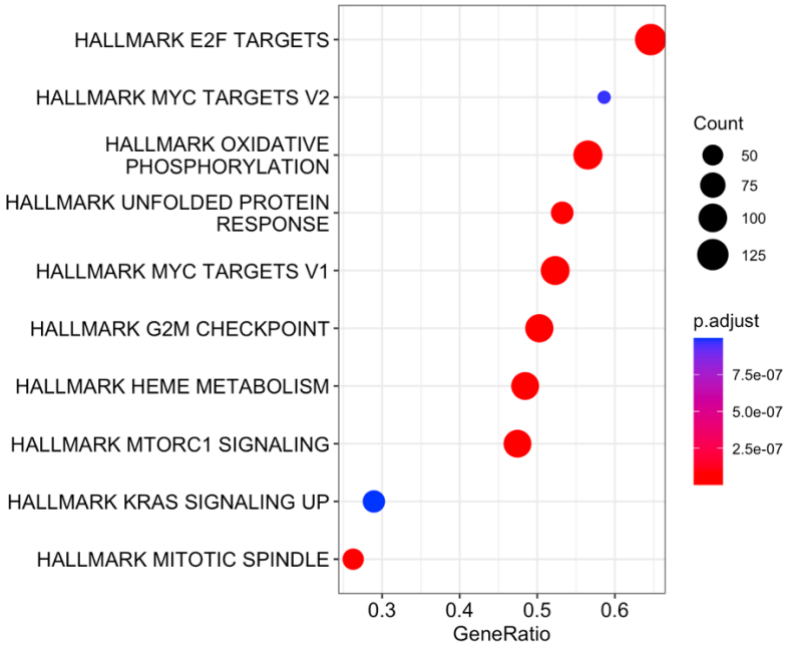

Figure 15 Functional enrichment analysis by GSEA function of dengue hemorrhagic fever vs healthy group using Cluster Profiler package in R (Code in Supplemental 6)

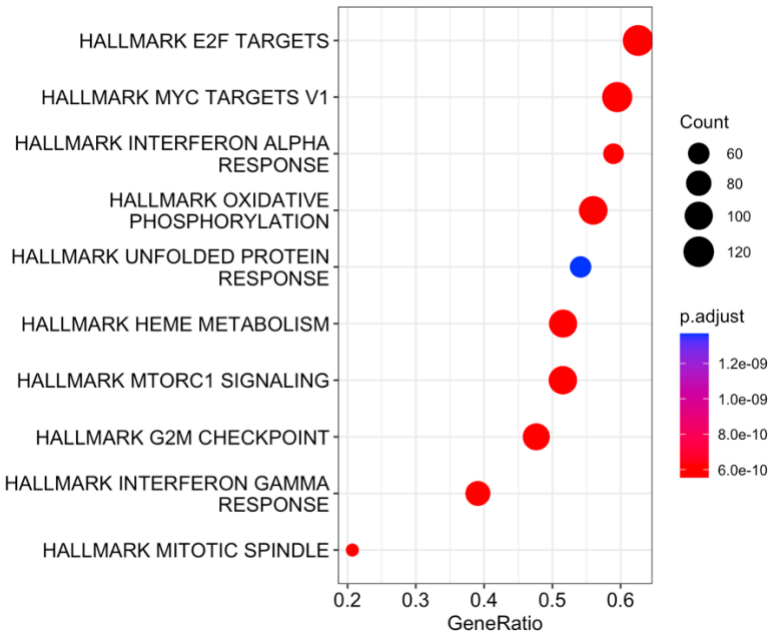

Figure 16 Functional enrichment analysis by GSEA function of dengue fever vs convalescent group using Cluster Profiler package in R (Code in Supplemental 6)

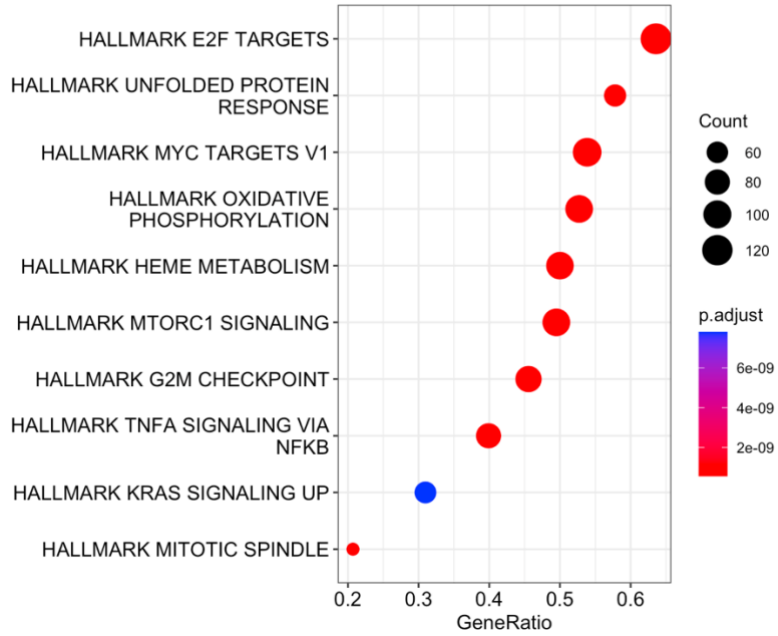

Figure 17 Functional enrichment analysis by GSEA function of dengue hemorrhagic fever vs convalescent group using Cluster Profiler package in R (Code in Supplemental 6)

Discussion

The exploratory analysis of the transcriptomic dataset GDS5093 has been found to give important insights in the gene expression profiles between dengue fever and dengue hemorrhagic fever patients in comparison to convalescent and healthy patients. The transcriptomics analysis of the dataset showed similar gene expression in convalescent and

healthy control groups, and dengue fever and hemorrhagic dengue fever. This might entail that the gene expression in a group, recovering from dengue infection, is very closely related to the gene expression in healthy controls. On the other hand, the gene expression in dengue fever and hemorrhagic dengue fever differentiates significantly from those in healthy individuals. Furthermore, the summary of the data with a dendrogram, heatmap and PCA gave clear explanation of the different gene expression profiles of the four different groups and showed a significant difference in dengue fever and dengue hemorrhagic fever patients towards convalescent and healthy patients since both subgroups mainly got clustered together. Furthermore, the differential gene expression analysis (Fig. 5 and Table 1) showed that the most significantly expressed genes were part of the cell cycle, subsequently mostly a part in mitosis. The functional enrichment analysis carried out by reactome and KEGG pathway also underlined these results since the most of the significantly expressed genes had functions inside the cell cycle. Referring this to the virus induced gene expression it shows that virus RNA is responding to the cell cycle which is biologically traceable since the virus wants to replicate itself to survive inside a host organism. Given the fact that a part of the differentially expressed genes also effects cellular senescence and for example the p53 signaling pathway, for manipulating homeostatic cellular processes is another measurement on which targets the virus is responding to and how it is changing gene expression for further development. Due to inducing a change in homeostasis the body response gets blocked stronger which is helping the virus. Furthermore, it has also been shown that patients infected with dengue virus have increased gene expression in metabolism, cell cycle, and cellular proliferation. The expression profiles for proteins corresponding to DNA-replication processes and supply of cells (Fig. 14-17) could be reasoned by the virus cells to infect as many host cells as possible.

The most statistically significant genes discovered in this context are *KCTD14* important for protein oligomerization. This mainly keeps the protein stable and offers protection against denaturation (The Uniprot Consortium, 2021). This represents another point of attack of the virus, since cell function and body reactions get strongly inhibited by this, and the virus development gets more difficult to stop. Next important virus induced genes that have been found were *CDC6, CEP55, PBK and SPC25*. *CDC6* is involved in the process of DNA replication thus overexpression helps the virus to infect the body as fast as possible. Additionaly like the other named genes all of them are tackling for the mitotic cell cycle. Considerably, this helps

to spread virus RNA in every cell of the host body (The Uniprot Consortium, 2021). All of these underlines the fact that cell-cycle is heavily manipulated under dengue fever and dengue hemorrhagic fever and shows that this is the main point of attack for this virus.

References

1. Blighe K, Rana S, Lewis M (2021). *EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling*. R package version 1.12.0

2. Orchestrating high-throughput genomic analysis with Bioconductor. W. Huber, V.J. Carey, R. Gentleman, ..., M. Morgan Nature Methods, 2015:12, 115.

3. Raivo Kolde (2019). pheatmap: Pretty Heatmaps. R package version 1.0.12.

4. The UniProt Consortium (2021) Uniprot: the universal protein knowledgebase in 2021. Nucleic Acid Res. 47:D506-515

5. Wickham H (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4

6. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu x, Liu S, Bo X, Yu G (2021). "clusterProfiler 4.0: A universal enrichment tool for interpreting omics data." *The*

7. Yu G, He Q (2016). "ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization." *Molecular BioSystems*, **12**(12), 477-479.