An Integrative Analysis Using Transcriptomics Data to Explore the Impact of Dengue Fever on Human Gene Expression

Abstract

Dengue virus (DENV) is a critical global health challenge, particularly in tropical regions. This study investigates the molecular impact of DENV infection on human gene expression using transcriptomic data from dataset GS5093 available on the Gene Expression Omnibus (GEO). The dataset encompasses a range of patient conditions, including convalescent, dengue fever, dengue hemorrhagic fever, and healthy controls. An integrative analysis employs Principal Component Analysis (PCA), Hierarchical Cluster Analysis (HCA), Differential Expression Analysis, and Gene Ontology (GO) to elucidate the molecular dynamics of dengue infection. Key upregulated genes identified in dengue fever include CEP55, KCTD14, PBK, and CDC6, while AI401105, PBK, CEP55, CDC6, and SPC25 are prominent in dengue hemorrhagic fever. Notably, CEP55 and PBK align with previous studies implicating them in flavivirus infections. GO analysis and pathway enrichment highlight significant biological implications, revealing shared molecular pathways such as the negative regulation of DNA replication associated with CDC6. These findings elucidate common molecular pathways involved in dengue pathogenesis, providing valuable insights into potential therapeutic targets and diagnostic markers. However, the study acknowledges limitations, including a small sample size, underscoring the need for larger datasets to enhance statistical power and the reliability of conclusions.

1. Introduction

Dengue virus (DENV) is a hyperendemic infectious disease that poses a major global health threat, notably prevalent in tropical regions of America, South-East Asia, and the West Pacific (WHO, 2023). Its primary mode of transmission is through the *Aedes* mosquito, giving rise to clinical manifestations ranging from mild asymptomatic dengue fever (DF) to severe dengue haemorrhagic fever (DHF). This comprehensive study aims to explore the impact of dengue fever on human gene expression from the transcriptomics dataset GS5093, accessible on the Gene Expression Omnibus (GEO). Particularly, this contains gene expression data from four distinct populations corresponding to different patient conditions: convalescent, dengue fever, dengue haemorrhagic fever, and healthy controls. There will be a multi-faceted approach incorporating exploratory analysis methods to assess the variation in gene expression among these populations. The additional identification of specific genes that demonstrate significant differences in expression across the disease states will provide further insight into biological implications.

2. Methods

2.1 Microarray Data

The initial gene expression dataset is based on DNA microarray analysis which examined blood samples from individuals during acute dengue virus (DENV) infection and convalescence (Kwissa *et al.*, 2014). This study categorised patients into 4 groups: convalescent (CONV), dengue fever (DF), dengue hemorrhagic fever (DHF), and healthy controls (HC). Specifically, this comprises gene expression data obtained from 56 individuals - 19 CONV, 18 DF, 10 DHF, and 9 HC. This dataset is readily accessible on the Gene Expression Omnibus (GEO) under accession number GDS5093. The pre-processing value is labelled as "transformed counts," with the assumption that established protocols were adhered after its publication.

2.2 Principal Component Analysis (PCA)

PCA is a multivariate statistical tool for dimensionality reduction. This was performed using the *scikit- learn* module, generating 10 principal components (PCs). A two-dimensional PCA plot (PC1 and PC2) illustrated the relationships between samples representing each disease state.

2.3 Hierarchical Cluster Analysis (HCA)

HCA was conducted in two phases: a sample- centric analysis of the entire dataset and a gene-centric analysis on the first 100 genes. This involved the *scipy.cluster.hierarchy* module, including the complete linkage method and Euclidean distance metric to produce dendrograms illustrating the hierarchal relationships among samples and genes, offering perspectives into their expression profiles and associations.

2.4 Heatmap

The *seaborn* package facilitated the creation of an annotated heatmap using the *clustermap* function, alongside the dendrograms representing the clustering of both genes and samples.

2.5 Volcano Plot

Differential expression analysis was carried to out to identify significant gene expression changes between various disease states and healthy controls. This involved the *scipy.stats* module with the *ttest_ind* function to calculate the t-statistic and associated p-values. The dataset was partitioned according to distinct disease states, in which the fold changes and p-values were computed for each gene. This was according to predefined significance thresholds of $\log_2$FC (1.0) and p-value (<0.05). The following results were stored into a data frame based on the identification of upregulated, downregulated, and non-significantly expressed genes. The outcome of the volcano plots were annotated with top upregulated genes, providing an overview of the differential expression patterns and the key genes associated with pathological conditions.

2.6 Gene Ontology (GO) and Pathway Enrichment

GO analysis was used to identified terms across cellular components, molecular functions, and biological processes of the key genes. This was followed by pathway enrichment, based on the Reactome database, which provides further insight into the involvement of genes in distinct biological pathways and functional implications.

3. Results and Discussion

3.1 PCA

Figure 1 presents a two-dimensional PCA plot, where PC1 accounts for 21.23% of the total variance and PC2 accounts for 8.59%. The clear clustering of the samples into distinct groups aligns with the clinical classification, demonstrating the validity of the analysis. The CONV samples are dispersed along the left region, reflecting diversity in gene expression profiles. The HC samples cluster tightly, indicating a consistent expression pattern that is easily distinguishable from other groups. Conversely, DF and DHF samples form a continuum from left to right, indicating genetic overlap, making it challenging to distinguish from using unsupervised learning alone. The complication in the pattern similarity between DF and DHF aligns with the original findings (Kwissa *et al.,* 2014).

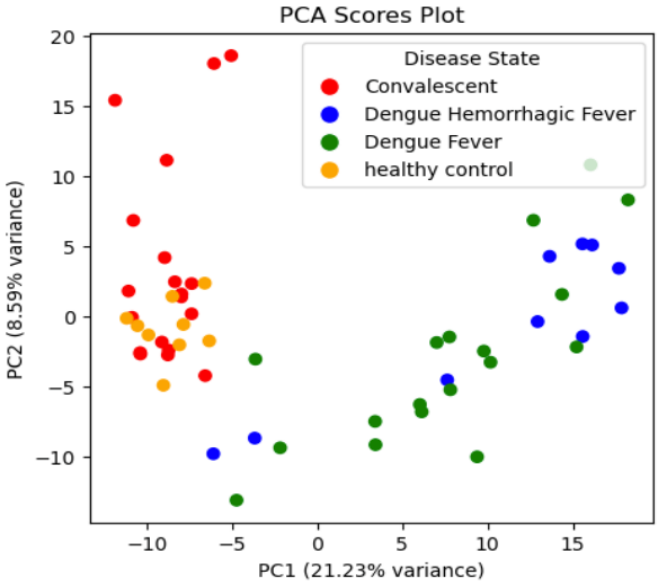


Figure 1. The two-dimensional plot illustrates the relationships between samples representing each disease state (Convalescent, Dengue Hemorrhagic Fever, Dengue Fever, and Healthy Control) based on PC1 (21.23%) and PC2 (8.59%) obtained through PCA.

3.2 HCA for sample-centric and gene-centric samples

Furthermore, HCA is used to visualise the intricate patterns of relationships within the gene expression dataset. Figure 2 shows initial phase of analysis was focused on sample-centric examination of the entire dataset. This demonstrates the relationships between 40 different dengue virus samples, based on their genetic sequences.

In Figure 3, the gene-centric dendrogram illustrates clustering of the top 100 genes based on their expression profiles. The dendrogram branches into two main clusters, with each further divided into smaller sub-clusters. These clusters indicate genes with similar expression profiles, reflecting shared molecular pathways or biological processes. This clustering aids in identifying key genetic markers that are differentially expressed in distinct disease states.

Overall, both dendrograms provide a comprehensive overview of the relationships within the gene expression dataset. The sample-centric dendrogram elucidates the genetic diversity across the different disease states, which is crucial for understanding disease progression, developing vaccines, and designing diagnostic tools. The gene-centric dendrogram offers valuable insights into the expression patterns of specific genes, laying the foundation for identifying potential diagnostic markers and therapeutic targets in dengue infection. These findings highlight the importance of further exploration and interpretation of biologically relevant patterns to enhance understanding of dengue pathogenesis and its clinical implications.

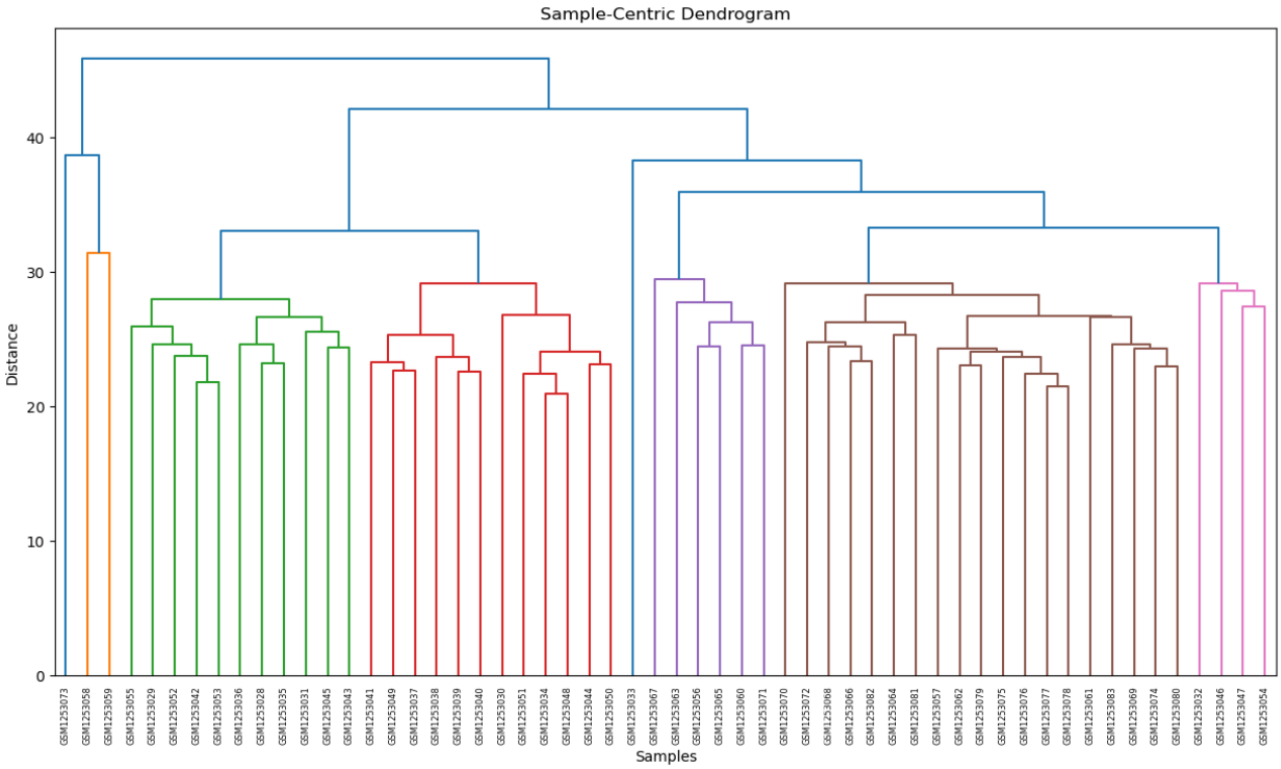


Figure 2. Sample-centric dendrogram.

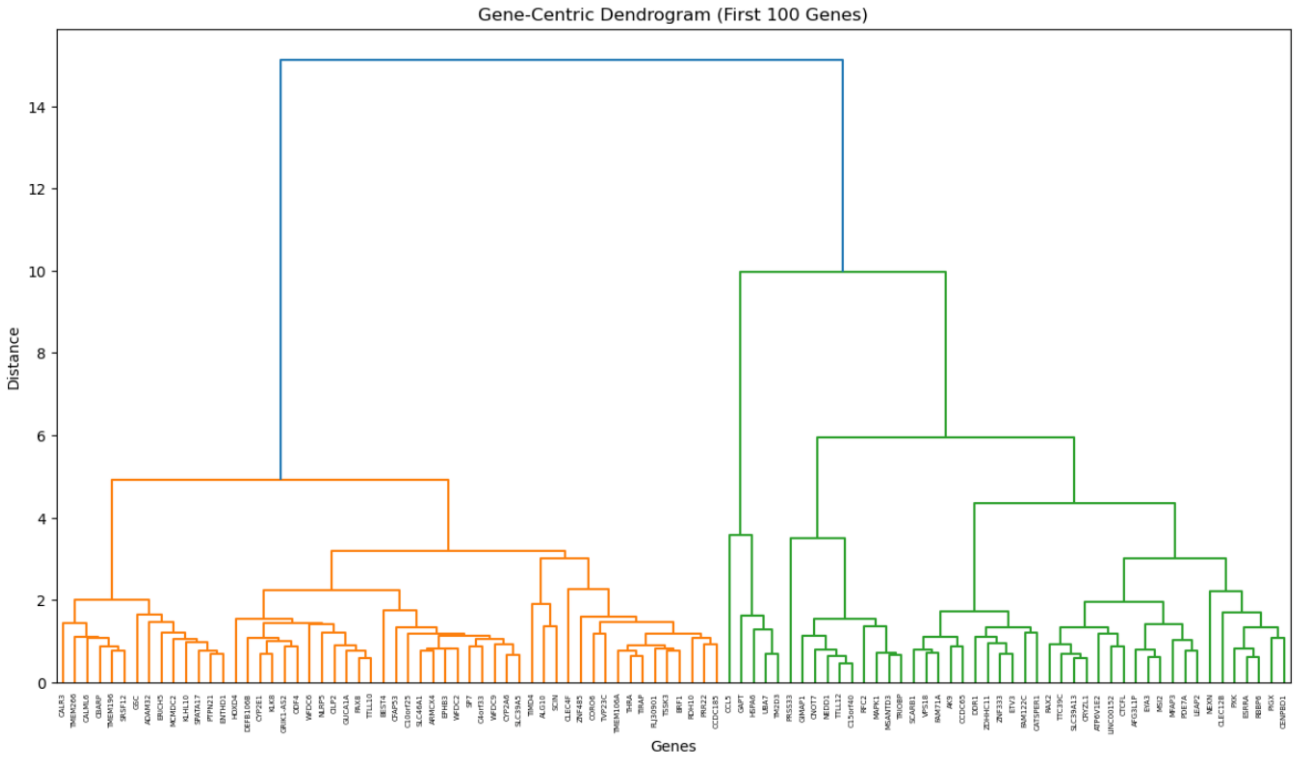


Figure 3. Gene-centric dendrogram of the first 100 genes.

3.3 Heatmap

Figure 4 provides a heatmap that demonstrates two distinct clusters of genes based on their expression profiles. The left cluster, characterized by lighter shades, represents genes with generally higher expression levels, while the right cluster, shown in darker shades, represents genes with lower expression levels. This clustering suggests that genes within the same group share similar expression patterns across different samples, likely indicating their involvement in related biological pathways or processes.

Along the top of the heatmap, samples are organized according to their expression profiles. The HC samples form a distinct and consistent cluster, highlighting their unique gene expression pattern, which differentiates them clearly from diseased samples. In contrast, the DF, DHF, and CONV samples display more diverse gene expression patterns, reflecting the variability of their responses to dengue infection.

The middle region of the heatmap, which includes the CC-chemokine receptor CCL5 gene, shows a noticeable shift in expression levels. This gene is known to be crucial for DENV-2 replication and disease progression, and its altered expression indicates its potential role as a biomarker for disease severity or as a therapeutic target (Marques *et al.,* 2015). The variation in expression patterns around this gene emphasises the molecular complexity of dengue infection.

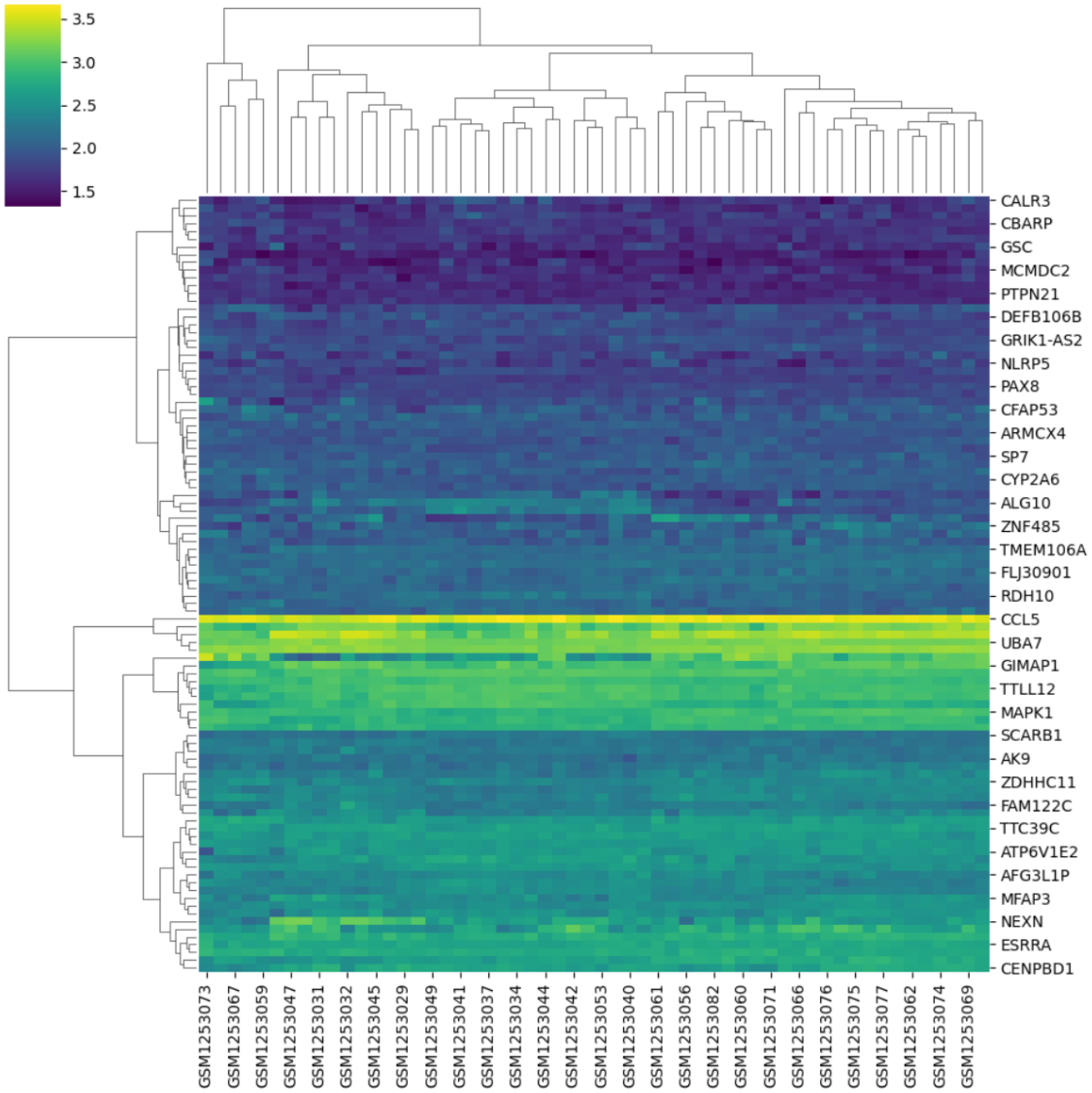


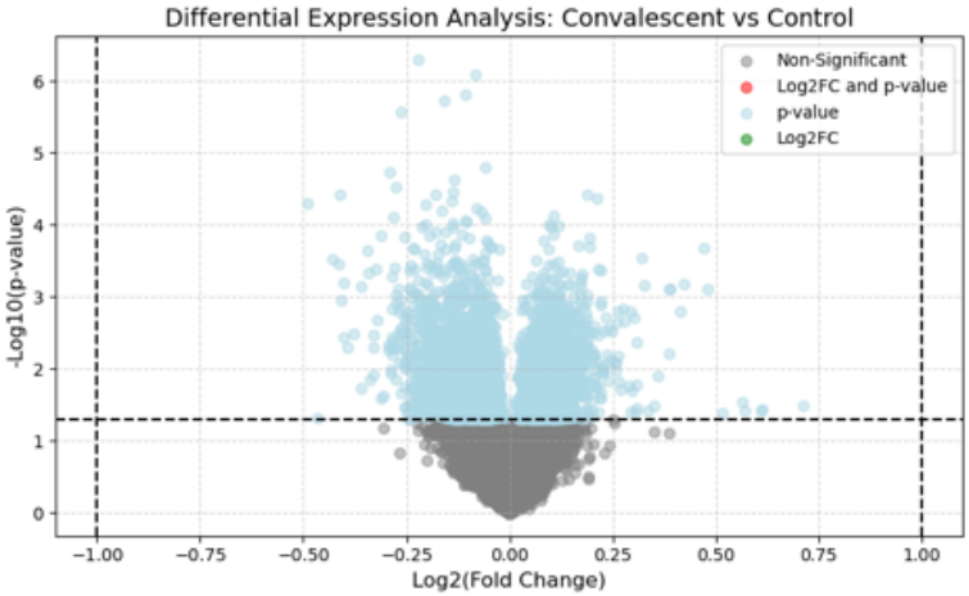
Figure 4. Heatmap depicting hierarchical clustering of gene expression patterns, with rows representing individual genes and columns representing different patient samples. The colour intensity reflects expression levels, with dark blue indicating low expression and light yellow indicating high expression. This visualisation highlights the complex relationships and distinct expression patterns across various disease states, providing insight into the molecular dynamics of dengue infection.

3.4 Differential Expression Analysis Using Volcano Plots

Figure 5 shows three volcano plots displaying the differential expression analysis of the HC against CONV, DHF, and DF. The x-axis shows the log2 fold change in gene expression between the two groups, and the y-axis shows the negative log10-transformed p-value of the differential expression test. The significance thresholds are established by log2FC of >1.0, and p-value of <0.05 to determine significant differentially expressed genes (Love *et al.,* 2014). The plotted points that meet these criteria are labelled as significant genes involved in dengue infection, and further summarised into Table 1 and Table 2.

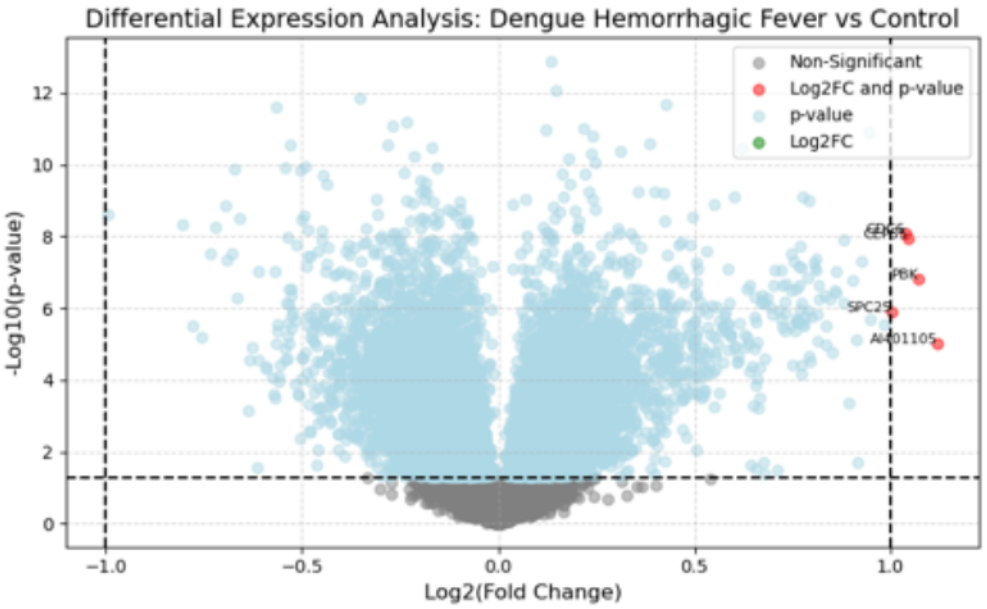
However, only a few significant results obtained for DF (CEP55, KCTD14, PBK, and CDC6) and for DHF (AI401105, PBK CEP55, CDC6, and SPC25). These genes are all significantly upregulated, with the common shared genes being CEP55, CDC6 and PBK. This is consistent with another study that suggested that upregulation of CEP55 and PBK are associated with cell differentiation and proliferation during flavivirus infection (Carlin *et al.,* 2018). Another recent study also supports the role of PBK as an important core protein in biological regulatory functions (Josyula *et al.,* 2023). Overall, there is limitations in this analysis due to the small sample size, holding limited statistical power to detect true differences in the gene expression.

A.



.

B.



C.

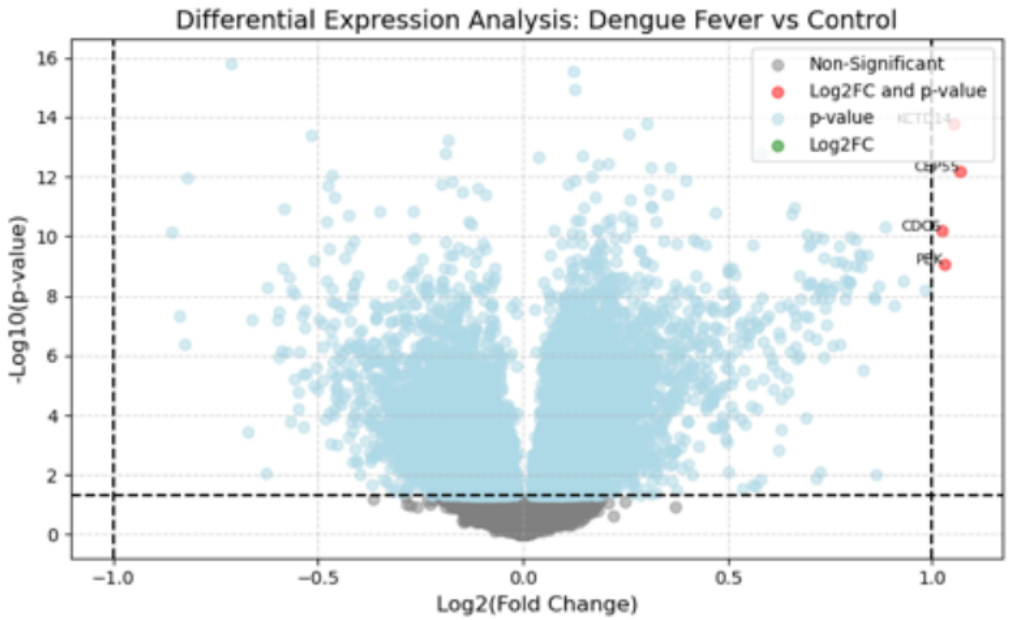


Figure 5. The volcano plots representing the Control versus Convalescent (A), Dengue Hemorrhagic Fever (B), and Dengue Fever (C). The dashes lines represent the p-value and Log2FC thresholds to define significant genes, which are further annotated on the plot.

Table 1. The Top Upregulated Genes for Dengue Fever (DF).

Rank	Gene	P-value	Fold Change
1	CEP55	6.72E-13	1.068829
2	KCTD14	1.64E-14	1.051681
3	PBK	8.95E-10	1.029902
4	CDC6	6.55E-11	1.024087

Table 2. The Top Upregulated Genes for Dengue Hemorrhagic Fever (DHF).

Rank	Gene	P-value	Fold Change
1	AI401105	9.85E-06	1.120596
2	PBK	1.52E-07	1.072204
3	CEP55	1.11E-08	1.047144
4	CDC6	7.92E-09	1.039302
5	SPC25	1.30E-06	1.003204

3.5 Gene Ontology (GO) and Reactome Pathway Enrichment Analysis

Both sets of upregulated genes identified in dengue fever (DF) and dengue hemorrhagic fever (DHF) were subject to Gene Ontology (GO) analysis to further elucidate their associated biological processes, molecular functions, and cellular components. It is important to note that both conditions share crucial upregulated genes, such as CEP55 and PBK, which have been previously implicated in flavivirus pathogenesis, underscoring their potential role in dengue infection. Subsequently, the Pathway Enrichment Analysis was performed to retrieve data from Reactome database (2022).

Table 3 represents the GO analysis results for DF, identifying "negative regulation of DNA replication" as the primary biological process. This indicates that during dengue fever, the virus may disrupt normal DNA replication to facilitate its replication and spread. The leading molecular function identified is "DNA replication origin binding," suggesting the involvement of proteins that recognise and bind to specific DNA sequences at the origin of replication. The top cellular component identified is the "mitotic spindle," a structure crucial for chromosome segregation during cell division. This association indicates that the viral infection may directly impact the cell cycle, possibly leading to abnormalities in cell division and genomic instability. Reactome analysis points to "CDC6 association with ORC (HAS-68689)" as the top pathway. This pathway involves CDC6, a key regulator of DNA replication initiation, suggesting that dengue infection may interfere with the initiation of DNA replication, potentially leading to disrupted cell cycle control.

Table 4 provides the GO analysis results for DHF, which similarly identifies the same top biological process, molecular function, and cellular components as found in DF. However, the Reactome analysis for DHF indicates that the top pathway is "cell cycle checkpoint R-HAS-69620," highlighting a critical checkpoint in the cell cycle that ensures the accuracy of cell division. This suggests that dengue hemorrhagic fever may significantly disrupt cell cycle regulation, potentially contributing to the severity of the disease by leading to unchecked cell proliferation or apoptosis.

Therefore, the similarity in GO results for DF and DHF indicates that both conditions likely involve common disruptions in DNA replication and cell cycle regulation. However, the differences in Reactome pathway analysis reflect the distinct pathogenic mechanisms of DF and DHF. While DF primarily involves interference with DNA replication initiation, DHF seems to further disrupt cell cycle checkpoints more significantly, potentially elucidating its increased severity. This analysis provides insight into the molecular underpinnings of dengue pathogenesis, highlighting the importance of DNA replication and cell cycle regulation in disease progression and emphasising potential therapeutic targets for intervention in severe dengue infections.

Table 3.1. GO Biological Process (DF)

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Negative Regulation Of DNA Replication (GO:0008156)	0.002997	0.02178	475.76	2764.28
2	DNA Replication Checkpoint Signaling (GO:0000076)	0.003396	0.02178	416.25	2366.46
3	Positive Regulation Of Chromosome Segregation (GO:0051984)	0.003595	0.02178	391.75	2204.79
4	Regulation Of Chromosome Segregation (GO:0051983)	0.005589	0.02178	246.53	1278.76
5	Positive Regulation Of G1/S Transition Of Mitotic Cell Cycle (GO:1900087)	0.006186	0.02178	221.84	1128.18
6	Regulation Of Metaphase/Anaphase Transition Of Cell Cycle (GO:1902099)	0.006385	0.02178	214.68	1084.94
7	Positive Regulation Of Cytokinesis (GO:0032467)	0.007778	0.02178	175.07	850.23
8	DNA Integrity Checkpoint Signaling (GO:0031570)	0.008374	0.02178	162.24	775.91
9	Positive Regulation Of Cell Division (GO:0051781)	0.008573	0.02178	158.37	753.68
10	Nucleus Organization (GO:0006997)	0.008772	0.02178	154.67	732.58

Table 3.2. GO Molecular Function (DF)

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	DNA Replication Origin Binding (GO:0003688)	0.002598	0.01039	555.11	3304.67
2	Protein Serine/Threonine Kinase Activity (GO:0004674)	0.06667	0.1185	19.21	52.03
3	Kinase Binding (GO:0019900)	0.08888	0.1185	14.19	34.34
4	Sequence-Specific Double-Stranded DNA Binding (GO:1990837)	0.1355	0.1355	9.00	17.99

Table 3.3. GO Cellular Component (DF)

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Mitotic Spindle (GO:0072686)	0.02830	0.1447	46.61	166.15
2	Spindle (GO:0005819)	0.04135	0.1447	31.56	100.54
3	Nucleolus (GO:0005730)	0.1455	0.2558	8.32	16.04
4	Nuclear Lumen (GO:0031981)	0.1471	0.2558	8.22	15.76
5	Intracellular Non-Membrane-Bounded Organelle (GO:0043232)	0.2184	0.2558	5.25	7.99
6	Nucleus (GO:0005634)	0.2192	0.2558	3.46	5.25
7	Intracellular Membrane-Bounded Organelle (GO:0043231)	0.2766	0.2766	2.87	3.68

Table 3.4. Reactome (2022) Pathway Enrichment (DF)

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	CDC6 Association With ORC:origin Complex R-HSA-68689	0.001599	0.02394	951.86	6128.34
2	Transcription Of E2F Targets Under Negative Control By DREAM Complex R-HSA-1362277	0.003795	0.02394	369.96	2062.22
3	G0 And Early G1 R-HSA-1538133	0.005389	0.02394	256.03	1337.30
4	G1/S-Specific Transcription R-HSA-69205	0.005788	0.02394	237.71	1224.71
5	Activation Of Pre-Replicative Complex R-HSA-68962	0.006385	0.02394	214.68	1084.94
6	Activation Of ATR In Response To Replication Stress R-HSA-176187	0.007181	0.02394	190.10	938.42
7	Orc1 Removal From Chromatin R-HSA-68949	0.01373	0.03581	97.69	418.90
8	CDK-mediated Phosphorylation And Removal Of Cdc6 R-HSA-69017	0.01432	0.03581	93.54	397.18
9	Switching Of Origins To A Post-Replicative State R-HSA-69052	0.01788	0.03743	74.56	300.03
10	Assembly Of Pre-Replicative Complex R-HSA-68867	0.02182	0.03743	60.82	232.62

Table 4.1. GO Biological Process (DHF)

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Negative Regulation Of DNA Replication (GO:0008156)	0.003745	0.02570	356.80	1993.61
2	DNA Replication Checkpoint Signaling (GO:0000076)	0.004243	0.02570	312.17	1705.22
3	Positive Regulation Of Chromosome Segregation (GO:0051984)	0.004492	0.02570	293.79	1588.07
4	Spindle Assembly Checkpoint Signaling (GO:0071173)	0.006484	0.02570	199.70	1006.18
5	Mitotic Spindle Assembly Checkpoint Signaling (GO:0007094)	0.006484	0.02570	199.70	1006.18
6	Mitotic Spindle Checkpoint Signaling (GO:0071174)	0.006484	0.02570	199.70	1006.18
7	Regulation Of Chromosome Segregation (GO:0051983)	0.006981	0.02570	184.89	917.89
8	Negative Regulation Of Mitotic Metaphase/Anaphase Transition (GO:0045841)	0.006981	0.02570	184.89	917.89
9	Positive Regulation Of G1/S Transition Of Mitotic Cell Cycle (GO:1900087)	0.007727	0.02570	166.38	809.09
10	Regulation Of Metaphase/Anaphase Transition Of Cell Cycle (GO:1902099)	0.007975	0.02570	161.00	777.86

Table 4.2 GO Molecular Function (DHF)

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	DNA Replication Origin Binding (GO:0003688)	0.003246	0.01298	416.31	2385.60
2	Protein Serine/Threonine Kinase Activity (GO:0004674)	0.08263	0.1465	14.41	35.93
3	Kinase Binding (GO:0019900)	0.1098	0.1465	10.64	23.50
4	Sequence-Specific Double-Stranded DNA Binding (GO:1990837)	0.1664	0.1664	6.75	12.11

Table 4.3. GO Cellular Component (DHF)

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Mitotic Spindle (GO:0072686)	0.03525	0.1799	34.95	116.93
2	Spindle (GO:0005819)	0.05141	0.1799	23.67	70.24
3	Nucleolus (GO:0005730)	0.1785	0.3157	6.24	10.76
4	Nuclear Lumen (GO:0031981)	0.1804	0.3157	6.17	10.56
5	Intracellular Non-Membrane-Bounded Organelle (GO:0043232)	0.2651	0.3654	3.94	5.23
6	Nucleus (GO:0005634)	0.3132	0.3654	2.31	2.68
7	Intracellular Membrane-Bounded Organelle (GO:0043231)	0.3856	0.3856	1.91	1.82

Table 4.4. Reactome (2022) Pathway Enrichment (DHF)

Index	Name	P-value	Adjusted p-value	Odds Ratio	Combined score
1	Cell Cycle Checkpoints R-HSA-69620	0.001781	0.03398	48.89	309.50
2	CDC6 Association With ORC:origin Complex R-HSA-68689	0.001999	0.03398	713.86	4436.86
3	Transcription Of E2F Targets Under Negative Control By DREAM Complex R-HSA-1362277	0.004741	0.03777	277.46	1484.80
4	Cell Cycle, Mitotic R-HSA-69278	0.006477	0.03777	24.92	125.58
5	G0 And Early G1 R-HSA-1538133	0.006732	0.03777	192.01	960.21
6	G1/S-Specific Transcription R-HSA-69205	0.007230	0.03777	178.28	878.83
7	Activation Of Pre-Replicative Complex R-HSA-68962	0.007975	0.03777	161.00	777.86
8	Activation Of ATR In Response To Replication Stress R-HSA-176187	0.008968	0.03777	142.57	672.09
9	Cell Cycle R-HSA-1640170	0.009998	0.03777	19.78	91.09
10	Orc1 Removal From Chromatin R-HSA-68949	0.01713	0.05155	73.26	297.93

4. Conclusion

This comprehensive study explored the impact of dengue infection on human gene expression and highlighted key upregulated genes, such as CEP55 and PBK. Both genes were found to be significantly upregulated in DF and DHF, indicating their potential role in the disease's progression. The pathway analysis revealed that pathways related to the negative regulation of DNA replication and CDC6 association were prominently affected. Specifically, "CDC6 association with ORC (HAS-68689)" was identified as the principal pathway impacted in DF, indicating disruptions in DNA replication initiation that could affect the cell cycle. For DHF, the "cell cycle checkpoint R-HAS-69620" pathway was most impacted, suggesting a deeper disruption in cell cycle regulation, which could explain the increased severity of the disease.

Despite a small sample size limiting statistical power, the study contributes significantly to understanding the molecular mechanisms of dengue infection. By identifying the shared biological processes and molecular pathways, this research provides a basis for future therapeutic targeting and improved diagnostics. The identified pathways, including DNA replication and cell cycle checkpoints, offer critical insights into disease progression and highlight potential targets for clinical interventions. Further research with larger datasets will be essential to validate these findings and enhance our understanding of the underlying pathogenesis.

5. References:

Carlin, A. F., Wen, J., Vizcarra, E. A., McCauley, M., Chaillon, A., Akrami, K., Kim, C., Ngono, A. E., Lara-Marquez, M. L., Smith, D. M., Glass, C. K., Schooley, R. T., Benner, C., & Shresta, S. (2018). A longitudinal systems immunologic investigation of acute Zika virus infection in an individual infected while traveling to Caracas, Venezuela. *PLoS neglected tropical diseases*, *12*(12), e0007053.

Josyula, J. V. N., Talari, P., Pillai, A. K. B., & Mutheneni, S. R. (2023). Analysis of gene expression profile for identification of novel gene signatures during dengue infection. *Infectious medicine*, *2*(1), 19–30.

Kwissa, M., Nakaya, H. I., Onlamoon, N., Wrammert, J., Villinger, F., Perng, G. C., Yoksan, S., Pattanapanyasat, K., Chokephaibulkit, K., Ahmed, R., & Pulendran, B. (2014). Dengue virus infection induces expansion of a CD14(+)CD16(+) monocyte population that stimulates plasmablast differentiation. *Cell host & microbe*, *16*(1), 115–127.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, *15*(12), 550.

Marques, R. E., Guabiraba, R., Del Sarto, J. L., Rocha, R. F., Queiroz, A. L., Cisalpino, D., Marques, P. E., Pacca, C. C., Fagundes, C. T., Menezes, G. B., Nogueira, M. L., Souza, D. G., & Teixeira, M. M. (2015). Dengue virus requires the CC-chemokine receptor CCR5 for replication and infection development. *Immunology*, *145*(4), 583–596.

World Health Organization. (2023). Dengue and severe dengue: Fact sheet. Available at: https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue [Accessed 10 December 2023].