Received: 15 January 2024 *Accepted: 15 May 2024*

Determining Differentially Expressed Genes in Dengue Patients During Disease Progression

Coorey, H.[1], Jayatillake, R.[1], Jayathilaka, N.[2], Ambanpola, N.[2]

hashinicoorey98@gmail.com, njayathi@kln.ac.lk, njayathi@kln.ac.lk, nimrothambanpola@gmail.com

[1]*Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka.*

[4]*Department of Chemistry, Faculty of Science, University of Kelaniya, Sri Lanka.*

Abstract

Gene expression studies on gene transcription to synthesize functional gene products have been used extensively to understand biological differences between different disease conditions. Thus, this study determines differentially expressed genes in dengue infection during disease progression following the three phases: Febrile, Defervescence and Convalescent. Integrative data analysis of two publicly available longitudinal datasets in Gene Expression Omnibus (GEO) database has been employed to accomplish the prime objective of exploring temporal gene expression patterns. The Friedman test was given more emphasis due to the non-normality distributions of data. Repeated measures analysis of variance (ANOVA) and linear mixed models were also implemented to examine the potential of detecting differentially expressed genes despite non-normality. The Friedman test revealed significant differences in gene expression levels across different phases in dengue disease over time. This led to a notably higher count of genes showing differential expression compared to the other two methods: Repeated measures ANOVA and linear mixed models. The pathway analysis approach consists of significant differentially expressed genes derived from the Friedman test. The results identified upregulated pathways with any significant change in the overall expression of genes within pathways over time for the Febrile and Defervescence phases considering the Convalescent phase as a baseline. Moreover, genes available in pathways were not identified by the two parametric tests for non-normal data implying that the parametric approaches resulted in the least significance for data with non-normal distributions.

Keywords: Dengue, Friedman test, Gene expression studies, Longitudinal data, Non-normality.

Introduction

In biology, the basic unit of heredity is known as a gene. It can be viewed from different aspects of its inheritance, biological function, molecular structure etc. Being a vital part of the genome, protein-coding genes encode the information for making proteins. To determine

which proteins and in which quantities are present in a cell, control of these mechanisms is essential. Although a gene is significant in gene expression, it does not function in isolation. Suites of genes are involved in performing biological functions. The structure of different gene functions in different sequential steps of a specific biological process referred to as the genetic pathway (Hejblum et al., 2015). The cruciality of determining a particular set of molecular functions in a bioloical process is evolved with cellular differentiation through differential gene expression. Thus, molecular signatures of various diseases provide information on developing drug candidates.

This study involves in determining differentially expressed genes in dengue infection over time and it provides important clues to the underlying transcriptional control mechanisms and network structure of a biological cell which aids in understanding the biological differences between different stages in dengue disease progression. Following the identification of significant differentially expressed genes (DEGs), i.e., biomarkers associated with the development of dengue infection vary over time, this study aims to identify metabolic pathway functions that significantly vary over time. This allows to gain insights into the functional working mechanism of cells beyond the detection of differentially expressed genes and develop drug candidates that can either target or avoid specific pathways or networks to develop new drugs.

Moreover, it would be surprising if considerable departures from normality were not identified given the nature of the underlying biology. De Torrenté et al. (2020) shows that the expressions of less than 50% of all genes were normally distributed and other genes consist of different distributions such as gamma, bimodal, lognormal, etc. However, the normality assumption has not been checked strictly in gene expression studies. Patino & Ferreira (2018) states a few reasons for that. Mainly researchers are unaware of statistical assumptions, standard approaches used to check assumptions and remedies for that, and many parametric tests have been applied without knowledge of underlying distributions. However, the good practice is to assess the feasibility of the utilized statistical tests. Hence, the study discussed in this paper will address this issue by employing both parametric and non-parametric tests with respect to the normality.

Objective of the study

On view of the above explanation based on past studies, the objectives of this study are (i) to identify differentially expressed genes in dengue infection over time and, functionally categorizing significant differentially expressed genes and (ii) to identify significant differences between parametric and non-parametric tests with respect to the normality in analysing DEGs.

Materials and Methods

Data description and preparation

The datasets required to accomplish the objective were acquired from publicly available microarray datasets in GEO database. Two keywords: “dengue expression” & “Homo sapiens” were used to search for gene expression studies related to dengue

disease in the GEO database. One hundred forty-five search terms appeared, and studies were selected from the database such that they follow the criteria: Studies with Homo sapiens which include the disease phase. Among them, two microarray gene expression datasets of whole blood or peripheral blood mononuclear cells (PBMCs): GSE28405, GSE43777 were chosen after thoroughly reviewing all the studies and datasets as other studies departed from the established criteria. In both studies blood samples were collected at three time points following the three stages of development of dengue: Febrile, Defervescence and Convalescent. Considering the available data, the prime focus was on Deoxyribonucleic acid (DNA) microarray rather than Ribonuclei acid (RNA) sequencing. Following table provides a summary on datasets used in this study.

Table 1.

A summary of the datasets.

Data Set	Country	Number of		Microarray platform
		Subjects	Genes	
Set 1	Singapore	31	23961	Illumina HumanRef-8 V1BeadChip
Set 2	Venezuela	18	54675	Affymetrix HG-U133 plus 2

The two datasets were normalized and analyzed independently as the two data sets derived from the two microarray platforms: Affymetrix and Illumina are different microarray technologies. Background correction, normalization and filtering were performed on both datasets using Bioconductor R packages (Silver et al., 2009). Quality control and pre-processing for dataset 1 were performed using the "BeadArray" package. Bead-averaged data was normalized using a quantile normalization method using the "Lumi" package (Tolfvenstam et al., 2011). Quality control of raw data in dataset 2 was done using the Robust Multi-chip Average (RMA) method in the 'Affy' package (Sun et al., 2013). Furthermore, gene signal normalization was done using housekeeping genes that do not respond to most treatments as references to compare to genes of interest (target genes) that do change. Glyceraldehyde 3-phosphate dehydrogenase (GAPDH), β-actin and Hypoxanthine-guanine phosphoribosyltransferase (HPRT) were chosen as reference genes as these three genes are the most stable genes for transcript normalizing in dengue infected studies (Kumar et al., 2018). After obtaining normalized gene expression values, the two datasets were checked for outliers separately by visualizing the intensity distributions of subjects of data. Then the filtering was applied to reduce the number of genes and increase the power to detect genes. Even with the multiple testing adjustment, it can result in low power since the number of hypothesis tests is still high in gene expression studies. Therefore, non-specific filtering method, i.e., filtering by variance was employed to further filter out genes.

Prior to any modelling, the normality was checked using the Shapiro-Wilk test on log transformed data. The results indicate that the dataset 2 satisfied the normality assumption for

majority of the genes while dataset 1 violated it for most of the genes. Both parametric & nonparametric tests were performed on two datasets separately to examine whether the results of both tests yielded a significant difference with respect to the normality.

Statistical tests

The Friedman test is an appropriate nonparametric test to check the differences between disease conditions, when there are more than two groups with repeated measures (Siegel & Castellan-Jr., 1988). In this study, it was applied to each gene considering 3 phases: febrile, defervescence and convalescent and subjects as blocks. In this study, p-values were obtained and adjusted to correct multiple testing issue. Obtained p values were adjusted for the between gene comparisons using the q-value procedure. For studies considering multiple genomes, a q-value as a false discovery rate (FDR) based measure was suggested because the FDR using The Benjamini-Hochberg procedure is too conservative for genomics applications (Storey & Tibshirani, 2003). This study used the following approach to obtain q-values using calculated p-values. If the q-value is less than 0.05, the null hypothesis that the groups coming from populations with the same median is rejected. Since, it does not reveal which phases differ for genes which can be found out using post hoc tests, the Wilcoxon-Nemenyi-McDonald-Thompson test (Pereira et al., 2015) was applied to compare the disease conditions: "Febrile", "defervescence" and "convalescent" to detect significant differences. Considering each gene at a time the test has been implemented to determine significantly differentially expressed genes between two different phases. Obtained p values were adjusted for the between gene comparisons using the q-value procedure. If the q-value is less than 0.05, it is declared significant.

Repeated measures ANOVA was performed to detect any overall differences between related means over time. In this study considering one gene at a time repeated measures ANOVA was performed to detect significant differences among "Febrile", "defervescence" and "convalescent". Then the p values obtained for calculated F statistics were adjusted using the q-value procedure to control the false discovery rate (FDR) due to multiple hypothesis testing (Storey & Tibshirani, 2003). If q values are less than the general threshold value of 0.05, then the null hypothesis that the related population means are not different was rejected and significantly differentially expressed genes among three conditions were identified. When significant differences were detected in the disease phases, a pairwise comparison of three phases was performed using a paired sample t-test to determine which pairs were significantly different for significant genes. Another parametric approach called the random intercept linear mixed model (Demidenko, 2013) was performed considering one gene at a time and the convalescent phase as the baseline. The most appropriate covariance structure with the smallest Akaike's Information Criteria (AICC) value was selected for the data. If the adjusted p-value for the fixed effect, i.e., the disease phase considered in this study, is less than the general threshold value, the gene was considered significant over time.

Once significantly differentially expressed genes are detected over time, it aims to gather

knowledge about relevant groups of genes or pathways to identify the underlying biological processes and mechanisms. Those groups of genes that share a common biological function are defined based on prior biological knowledge, e.g., biochemical pathways or coexpression in previous experiments (Subramanian et al., 2005). In this study, pathways related to the detected differentially expressed genes were identified using the "Reactome" pathway database. Following the identification of functionally linked gene sets or pathways, gene sets are analysed as they are useful in summarizing biological trends. The gene set analysis is more powerful than the gene-by-gene analysis for several reasons. Even though none of the genes in the group exhibit very significant absolute fold changes, it can detect changes in their expression levels. Further, changing all the genes in a particular pathway might have a more significant biological impact than a considerable increase of a single gene (Hejblum et al., 2015).

In the context of longitudinal microarray data, gene set analysis is burdensome as the dynamics of gene expressions within a gene set can be varied in a complicated way. Detecting such heterogeneity within a gene set has led to detect any change over time. Thus, the interest is focused on identifying any significant change in the overall expression of genes within gene sets i.e., pathways over time. Once pathways using the "Reactome" database are identified, considering a pathway at a time, gene expressions in each pathway were modelled using mixed models to examine any significant trend over time or heterogeneity between gene trends within the pathway. The trend is captured using linear polynomial functions. Then the significance of a pathway is determined by testing the significance of the time trend implying testing both random effects and fixed effects at once in a pathway. They are tested simultaneously using the likelihood ratio. If the computed p-value is less than 0.05, the null hypothesis that the genes within a pathway are stable over time is rejected. Since multiple gene sets are investigated at a time, Benjamini-Yekutieli correction procedure is used to adjust p values (Hejblum et al., 2015).

Results and Discussion

Univariate analysis

For dataset 1 and 2, 5455 and 5548 significant DEGS over time were declared respectively from the Friedman test. Considering one gene at a time, p values were obtained and adjusted using the q-value procedure to correct for the multiple testing issue. Significant genes were selected using the threshold value of 0.05 based on the q values. Since the Friedman test does not give information on which phases carried out differences for the significant genes, the Wilcoxon-Nemenyi-McDonald-Thompson test was applied. Considering one gene at a time, the test was applied to compare the rank sum of conditions of each comparison: "Febrile & Convalescent", "Defervescence & Convalescent" and "Febrile & Defervescence" to detect significant genes. The tables 2-4 represents the number of significant genes for each comparison for dataset 1 and dataset 2 separately. The three disease phases: febrile, defervescence, and convalescent are abbreviated as F, D, and C respectively.

Table 2.

Significant genes from the Wilcoxon-Nemenyi-McDonald-Thompson test followed by the Friedman test.

Group	Dataset 1	Dataset 2
F & C	1109	1540
D & C	2782	2511
F & D	2400	1149
	6291	6200

Table 3.

Significant genes from repeated measures ANOVA followed by the paired sample t test.

Group	Dataset 1	Dataset 2
F & C	794	1702
D & C	1918	2407
F & D	830	1454
	3542	5563

Table 4.

Significant genes from linear mixed models.

Group	Dataset 1	Dataset 2
F & C	1109	1540
D & C	2782	2511
	3891	4051

Note that considering all the facts significant DEGs identified from the Friedman test followed by the Wilcoxon-Nemenyi-McDonald-Thompson test were carried out into the pathway analysis as this test has not lost any information regarding significant genes. Hence before moving to the pathway analysis, significantly DEGs derived from the Friedman test followed by the Wilcoxon-Nemenyi-McDonald-Thompson test were reconsidered as shown below. Gene expression patterns in two groups: Febrile & Convalescent and Defervescence & Convalescent were explored by considering convalescent phase as the baseline since the interest is focused on identifying significant gene sets or pathways in the above-mentioned groups in the pathway analysis. Then DEGs were categorized into upregulated & downregulated genes by calculating log2 fold change which measures how much a quantity changes between the two phases for each gene. Here the threshold value of 1 or $|log2fold\ change| \geq 1$ was used. Summary results are shown in Table 5.

Table 5.

Upregulated & downregulated genes.

Group	Dataset 1		Dataset 2	
	Upregulated	Downregulated	Upregulated	Downregulated
F & C	420	432	214	92
D & C	255	1738	228	260

Pathway analysis

In the pathway analysis, it is essential to note that the significant genes of most interest are the ones identified using the Friedman test followed by the Wilcoxon-Nemenyi-

McDonald-Thompson test. That is due to the results of the parametric techniques for dataset 1 may be unstable due to violation of the normality assumption. As described in the univariate analysis, the significant DEGs for "Febrile & Convalescent" and "Defervescence & Convalescent" were explored further as upregulated and downregulated genes as they carried out into the pathway analysis. Detected significant upregulated and downregulated DEGs common to the two datasets are presented in table 6.

Table 6.

A summary of upregulated & downregulated genes.

Group	Upregulated genes	Downregulated genes
F & C	144	41
D & C	130	109

Reactome pathways for those upregulated and downregulated genes for the two groups were obtained. According to the results obtained, more importantly, no downregulated pathways have been discovered for either of the groups. However, 27 and 26 upregulated pathways were identified for "Febrile & Convalescent" and "Defervescence & Convalescent" groups respectively.

Time course gene set analysis

Tables 7 and 8 provide computed likelihood ratios, p values and adjusted p values for only the significant pathways over time. It should be noted that the significance of pathways was checked for only dataset 2 as it satisfied the normality assumption while dataset 1 violated it.

Table 7.

Significant pathways for upregulated genes in "Febrile & Convalescent" group.

Reactome pathway	Likelihood ratio	p value	Adjusted p value
Interferon Signaling (R-HSA-913531)	181.9051	$P < 0.05$	$P < 0.05$
Cytokine Signaling in Immune system (R-HSA-1280215)	176.8954	$P < 0.05$	$P < 0.05$
Innate Immune System (R-HSA-168249)	295.4940	$P < 0.05$	$P < 0.05$

Table 8.

Significant pathways for upregulated genes in "Defervescence & Convalescent" group.

Reactome pathway	Likelihood ratio	P value	Adjusted p value
Interferon Signaling (R-HSA-913531)	315.9065	$P < 0.05$	$P < 0.05$

Results in Table 7 indicate that only 3 pathways were declared significant over time among 27 pathways for the "Febrile & Convalescent" group in dataset 2. Even though all p values were less than the threshold of 0.05, the adjusted p values using the Benjamini-Yekutieli correction method was used to determine significance of the pathways. As depicted in Table 8, based on adjusted p values, only one gene set was identified as significant over time among 27 gene sets.

Comparison of parametric & nonparametric test results

Mainly three statistical models were implemented on the two datasets as explained in previous sections. The significant DEGs derived from those three methods vary for several reasons. Figure 1 and Figure 2 present comparisons between the implemented methods performed to detect those differences for the two datasets separately. The three implemented techniques: Friedman test, Repeated Measures ANOVA & Linear Mixed models are abbreviated as 'F', 'R', & 'L' respectively in the Figures 1 and 2.

Figure 1.

Venn diagram of implemented models for dataset 1.

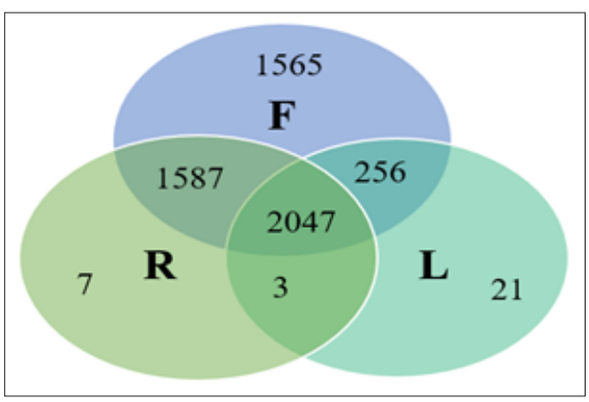
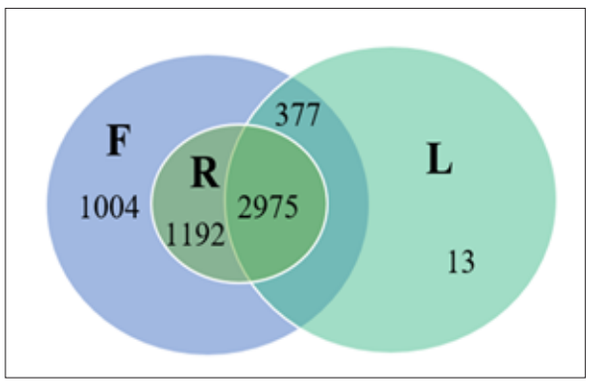


Figure 2.

Venn diagram of implemented models for dataset 2.



Out of 6029 genes 5455 and 5548 genes satisfy the Friedman test for dataset 1 and 2 respectively. However, 3644 and 4167 significant DEGs were identified using Repeated Measures ANOVA indicating that the number of significant DEGs have reduced by a large number in dataset 1 compared to the dataset 2. Which may be due to dataset 1 not satisfying the normality assumption and parametric tests on skewed data resulting in fewer genes being significant. However, the results of Repeated Measures ANOVA do not deviate considerably from the results of Friedman test for dataset 2. From the results obtained in the analysis, it was suggested that the Friedman test favored on dataset 1 while the linear mixed models favored on dataset 2.

The significant DEGs derived from the Friedman test followed by the Wilcoxon-Nemenyi-McDonald-Thompson test were used to obtain the gene pathways. However, it is noteworthy to investigate whether those genes in the pathways have become significant for other two implemented tests: Repeated Measures ANOVA and Linear Mixed Models. Failure in detecting those genes would indicate loss of important biomarkers if only parametric

approaches were considered. Surprisingly, all the genes derived from the Friedman test were also identified as DEGs by Repeated Measures ANOVA and Linear Mixed Models for dataset 2. However, for dataset 1, few of those genes were not identified as DEGs by the two parametric approaches. These facts established the importance of normality assumption as performing the parametric approaches on skewed data (dataset 1) resulted in losing significant DEGs. Moreover, it was seen that the parametric approaches did not fail to detect all the genes derived from the Friedman test for normally distributed data. However, it cannot be guaranteed that the significant DEGs derived from non-parametric approaches are always accurate as parametric tests are the most powerful approaches to detect differences for normally distributed data.

Conclusions

In conclusion, the analysis indicates that a considerable number of genes possess the ability to differentiate between the disease conditions "Defervescence" and "Convalescent." The application of the Friedman test yielded a higher number of significant DEGs over time compared to the repeated measures ANOVA and linear mixed models for both datasets. Notably, the parametric approaches exhibited the least number of significant DEGs when applied to data with non-normal distributions. Therefore, the assumption of normality plays a crucial role in identifying significant DEGs over time. These findings emphasize the importance of selecting appropriate statistical methods and considering the underlying distribution characteristics when analyzing gene expression data in relation to disease conditions. Considering pathway analysis, Twenty-seven and twenty-six upregulated pathways were identified for the significant DEGs derived from the Friedman test for "Febrile & Convalescent" and "Defervescence & Convalescent" groups respectively and no downregulated pathways have been discovered for either of the groups. Among them, three upregulated pathways: Interferon Signaling (R-HSA-913531), Cytokine Signaling in Immune system (R-HSA-1280215), Innate Immune System (R-HSA-168249) for "Febrile & Convalescent" and one upregulated pathway: Interferon Signaling (R-HSA-913531) for "Defervescence & Convalescent" group had significant change in the overall expression of genes within pathways over time.

References

De Torrenté, L., Zimmerman, S., Suzuki, M., Christopeit, M., Greally, J. M., & Mar, J. C. (2020). The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data. BMC Bioinformatics, 21(21), 1–18. https://doi.org/10.1186/S12859-020-03892-W/FIGURES/7.

Demidenko, E. (2013). Mixed models: Theory and applications with R: Second edition. Mixed Models: Theory and Applications with R: Second Edition, 1–717. https://doi.org/10.1002/9781118651537.

Hejblum, B. P., Skinner, J., & Thiébaut, R. (2015). Time-Course Gene

Set Analysis for Longitudinal Gene Expression Data. PLOS Computational Biology, 11(6), e1004310. https://doi.org/10.1371/JOURNAL.PCBI.1004310.

Kumar, V. E., Cherupanakkal, C., Catherine, M., Kadhiravan, T., Parameswaran, N., Rajendiran, S., & Pillai, A. B. (2018). Endogenous gene selection for relative quantification PCR and IL6 transcript levels in the PBMC's of severe and non-severe dengue cases. BMC Research Notes, 11(1), 1–6. https://doi.org/10.1186/S13104-018-3620-2/FIGURES/2.

Patino, C. M., & Ferreira, J. C. (2018). Meeting the assumptions of statistical tests: an important and often forgotten step to reporting valid results. Jornal Brasileiro de Pneumologia, 44(5), 353. https://doi.org/10.1590/S1806-37562018000000303.

Pereira, D. G., Afonso, A., & Medeiros, F. M. (2015). Overview of Friedmans Test and Post-hoc Analysis. Communications in Statistics: Simulation and Computation, 44(10), 2636–2653. https://doi.org/10.1080/03610918.2014.931971.

Siegel, S. & Castellan-Jr., N. J. (1988). Nonparametric Statistics for the Behavioral Sciences, International Edition. 262–272.

Silver, J. D., Ritchie, M. E., & Smyth, G. K. (2009). Microarray background correction: maximum likelihood estimation for the normal–exponential convolution. Biostatistics (Oxford, England), 10(2), 352. https://doi.org/10.1093/BIOSTATISTICS/KXN042.

Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences of the United States of America, 100(16), 9440. https://doi.org/10.1073/PNAS.1530509100.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545–15550. https://doi.org/10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG.

Sun, P., García, J., Comach, G., Vahey, M. T., Wang, Z., Forshey, B. M., Morrison, A. C., Sierra, G., Bazan, I., Rocha, C., Vilcarromero, S., Blair, P. J., Scott, T. W., Camacho, D. E., Ockenhouse, C. F., Halsey, E. S., & Kochel, T. J. (2013). Sequential waves of gene expression in patients with clinically defined dengue illnesses reveal subtle disease phases and predict disease severity. PLOS Neglected Tropical Diseases, 7(7). https://doi.org/10.1371/JOURNAL.PNTD.0002298.

Tolfvenstam, T., Lindblom, A., Schreiber, M. J., Ling, L., Chow, A., Ooi, E. E., & Hibberd, M. L. (2011). Characterization of early host responses in adults with dengue disease. BMC Infectious Diseases, 11(1), 1–7. https://doi.org/10.1186/1471-2334-11-209/TABLES/4.