2023
SICASH
SLIIT INTERNATIONAL CONFERENCE ON
ADVANCEMENTS IN SCIENCES & HUMANITIES

Determining Differentially Expressed Genes in Dengue Patients during Disease Progression

*[1]H. Coorey, [2]R. Jayatillake, [3]N. Jayathilaka, [4]N. Ambanpola

[1,2]Department of Statistics, Faculty of Science, University of Colombo, Sri Lanka
[3,4]Department of Chemistry, Faculty of Science, University of Kelaniya, Sri Lanka

Corresponding author - *hashinicoorey98@gmail.com

ARTICLE INFO

Article History:
Received: 10 September 2023
Accepted: 01 November 2023

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

Citation:

H. Coorey, R. Jayatillake, N. Jayathilaka, N. Ambanpola. (2023). Determining Differentially Expressed Genes in Dengue Patients during Disease Progression. Proceedings of SLIIT International Conference on Advancements in Sciences and Humanities, 1-2 December, Colombo, pages 309-314.

ABSTRACT

Gene expression studies on gene transcription to synthesize functional gene products have been used extensively to understand the 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 the 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. Since previous studies on gene expression have not primarily relied on normality assumption, repeated measures analysis of variance and linear mixed models were implemented to examine the potential of detecting differentially expressed genes despite non-normality. The Friedman test indicated that gene expression levels differentiate with different phases in dengue disease over time, resulting in a high number of significant differentially expressed genes compared to the other two techniques. The pathway analysis approach consists of

hashinicoorey98@gmail.com
https://doi.org/10.54389/AJRM6708

significant differentially expressed genes derived from the Friedman test. The results identified 27 and 26 upregulated pathways for the "Febrile and Convalescent" and "Defervescence and Convalescent" groups respectively. 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.

1. 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. In order 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 is referred to as the genetic pathway. The cruciality of determining a particular set of molecular functions in a biological process has evolved with cellular differentiation through differential gene expression. Thus, molecular signatures of various diseases provide information on developing drug candidates.

This study involves 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 the development of drug candidates that can either target or avoid specific pathways or networks to develop new drugs.

Moreover, it would be doubtful if considerable departures from normality were not identified given the nature of the underlying biology. de Torrenté et al. (2020) show 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) state a few reasons for that. Mainly researchers are unaware of statistical assumptions, standard approaches used to check assumptions and remedies for them, 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 normality.

The primary objective of this study is to identify differentially expressed genes in dengue infection over time. Along with that, functional categorization of significant differentially expressed genes i.e., performing a pathway analysis of differentially expressed genes is completed. Furthermore, the study also focuses to identify significant differences between parametric and non-parametric tests with respect to the normality.

2. MATERIALS AND METHODS

The datasets required to accomplish the objective were acquired from publicly available microarray datasets in the 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, and 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. The Following table provides a summary of datasets used in this study:

Table 1 *- A summary of the datasets*

	Country	No: of sub-jects	No: of genes	Microarray platform
Dataset1	Singapore	31	23961	Illumina HumanRef-8 V1BeadChip
Dataset2	Venezuela	18	54675	Affymetrix HG-U133 plus 2

The two data sets derived from the two microarray platforms were normalized and analyzed independently. Background correction, normalization and filtering were performed on both datasets using Bioconductor R packages. 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. Quality control of raw data in dataset 2 was done using the Robust Multi-chip Average (RMA) method in the ‘Affy’ package. The gene expression data were expressed as log2 values. Further, 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 by removing unwanted variations, the two datasets were checked for outliers separately. Then the filtering was applied to reduce the number of genes and increase the power to detect genes. Even with the multiple testing adjustments, it can result in low power since the number of hypothesis tests is still high in gene expression studies. Therefore, a 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 Dataset 2 satisfied the normality assumption for the majority of the genes while Dataset 1 violated it for most of the genes. Both parametric and 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.

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. 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 issues. Obtained p values were adjusted for the between gene comparisons using the q-value procedure. 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 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 Analysis of Variance (ANOVA) was performed to detect any overall differences between related means. 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. 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 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.

3. RESULTS AND DISCUSSION

3.1. UNIVARIATE ANALYSIS AND PATHWAY ANALYSIS

For datasets 1 and 2, 5455 and 5548 genes were declared significant respectively. The following table represents the number of significant genes from the Wilcoxon-Nemenyi-McDonald-Thompson test for each comparison for Dataset 1 and Dataset 2 separately.

Table 2 - Significant genes from the Wilcoxon-Nemenyi-McDonald-Thompson test

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

[Three disease phases are: febrile (F), defervescenece (D), and convalescent (C)]

Moreover, gene expression patterns in two groups: Febrile and Convalescent and Defervescence and 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 an 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.

Table 3 - Upregulated & downregulated genes

Group	Dataset 1		Dataset 2	
	Upregu-lated	Down-regulat-ed	Upregu-lated	Down-regulat-ed
F & C	420	432	214	92
D & C	255	1738	228	260

Detected upregulated and downregulated DEGs common to the two datasets were used to discover the pathways using the “Reactome” pathway database. 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 the “Febrile and Convalescent” and “Defervescence and Convalescent” groups respectively.

3.2. 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.(a) and Figure 1.(b) present comparisons between the implemented methods performed to detect those differences for the two datasets separately.

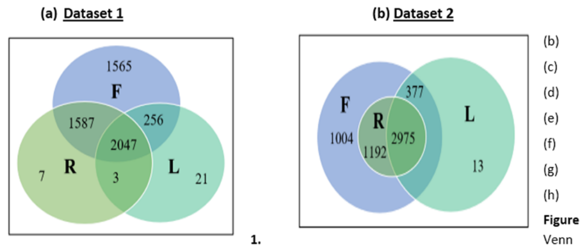

Figure 1. Venn Figure 1. diagram of implemented models (a) for data set 1, (b) for data set 2.

[F = Friedman test, R = Repeated Measures ANOVA and L = Linear mixed models].

Out of 6029 genes 5455 and 5548 genes satisfy the Friedman test for datasets 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. This 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 dataset 1 while the linear mixed models favored 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 the 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, a 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.

4. CONCLUSIONS

In conclusion, the analysis indicates that a considerable amount 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.

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

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