Effect of disease state on gene expression profile using Dengue dataset

By Bhavikhdeep Ghataore

Abstract:

This work aimed to identify the impact four populations have on the gene expression profile. Analysis was conducted on the Dengue dataset which displayed the expression of genes against samples from four different populations (Dengue fever, Dengue Haemorrhagic Fever, Convalescent and Healthy control). Any genes that were shown to be significantly expressed between the four states needed to be identified. Finally, this aimed to determine whether machine learning could be used to determine between patients with dengue fever and dengue haemorrhagic fever, when given unseen data. Explorative data analysis techniques of principal component analysis (PCA) and hierarchical cluster analysis (HCA) to visualise how the samples cluster together based on their disease state. To identify the genes that were significantly expressed between the states, volcano plots were used by plotting the 'Log2Foldchange' values against the '-Log2(p-values)' values. To identify whether machine learning could be used to classify between the two states, a Support Vector Machine (SVM) with bootstrapping was used. A confusion matrix was plotted to help visualise the accuracy of the classifier. The PCA and HCA results showed that convalescent and healthy control patients had similar gene expression profiles, whilst dengue fever and dengue haemorrhagic fever also had similar gene expression profiles. There were differences in genes that were expressed significantly between the disease states, the disease states dengue fever and haemorrhagic fever had a much larger number of significantly expressed genes in comparison to the convalescent state. The classifier was shown to have an accuracy level of 63.2%, which from a medical perspective may not be sufficient but adjustments to parameters within the model could help improve the mode and its performance. In conclusion, the disease state impacts the gene expression of a profile as well as the number of significant expressions. Finally, the use of machine learning in distinguishing between diseases depends on the accuracy level of the classifier. The level of accuracy can be improved using numerous factors such as a larger training dataset and optimised parameters.

Introduction:

A gene expression dataset and its metadata were provided, and the data was collected using DNA microarrays, from the blood of patients with acute dengue virus (DENV) infection and those recovering (convalescence). Dengue fever is a mosquito-borne disease and has a related life-threatening illness called dengue haemorrhagic fever. The metadata had shown the samples to be separated into four populations, healthy controls (CTRL), Dengue Fever (DENV), Dengue Haemorrhagic Fever (DHF) and convalescent (CONV). The original data layout contained the samples as columns and the genes as rows, the individual slots contained the expression of the gene for the sample.

This work aimed to identify how the disease state of an individual would affect the gene expression profile of that patient. Any genes that may significantly differ in expression between the four disease states would need to be identified and the significance they would have biologically would need explaining. The final aim of this work was to look from a treatment perspective to see whether it would be possible to use machine learning to produce a classifier that can distinguish between patients with dengue haemorrhagic fever or dengue fever when given unseen data.

Methods:

Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA):

The data provided was given in a gene-centric format and had to be transposed to a sample-centric format, to be able to visualise the effect of each population on the gene expression profile. The dataset contained high-dimensional data meaning it had more than 3 features. To visualise the high-dimensional data, a method of dimensionality reduction called principal component analysis was used. To validate conclusions made from PCA, Hierarchical cluster analysis (HCA) was used. It used to group similar objects into clusters, in this case, it would group them based on their gene expression profile. To visualise the clusters a hierarchical dendrogram with ward linkage was used, as shown in Figure 2. Ward linkage was used as it minimised the within-cluster variance for gene expression and this would allow higher accuracy of grouping, as well as easier identification of patterns. The square of the Euclidean distance is used as the distance metric for Ward linkage.

Volcano plots and statistical analysis:

Different genes may be expressed at various levels in comparison to a healthy individual. To visualise the significantly expressed genes between the four populations, volcano plots were plotted. Volcano plots are good to visualise the magnitude and significance of the expression of genes when comparing two populations. An independent t-test was used to calculate the p-values for the fold change between the two populations. The three volcano plots plotted were the ‘diseased’ states (convalescent, dengue fever and dengue haemorrhagic fever) against the healthy control samples, to display the effect those populations have on the expression of specific genes. The top five significantly expressed genes were annotated onto each volcano plot and a significance level of 0.05 was used.

SVM Machine Learning Classifier:

To build a classifier to distinguish between data that has come from patients with dengue fever or dengue haemorrhagic fever, we used a support vector machine (SVM) alongside bootstrapping. SVM was used due to its ability to manage high-dimensional data, bootstrapping allows for us to evaluate the performance of the model through calculating the average accuracy level. Lastly, a confusion matrix was produced to further analyse the performance of the SVM classifier.

Results and discussion:

Principal Component Analysis (PCA) and Hierarchical Cluster Analysis (HCA):

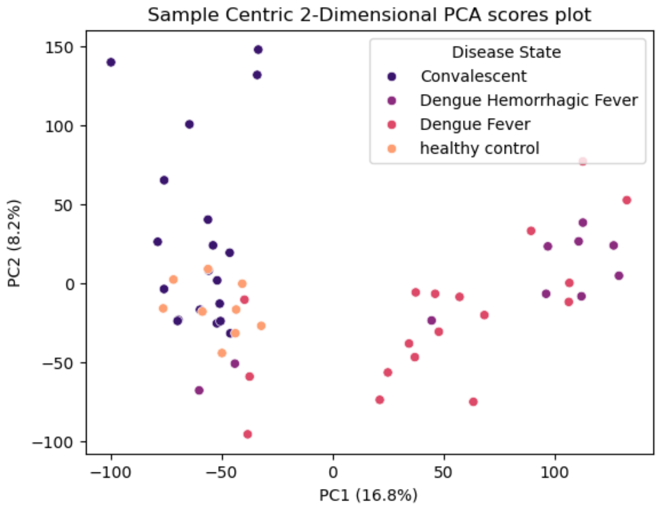


Figure 1: Sample Centric PCA scores plot displaying the variance of data from principal components 1 and 2.

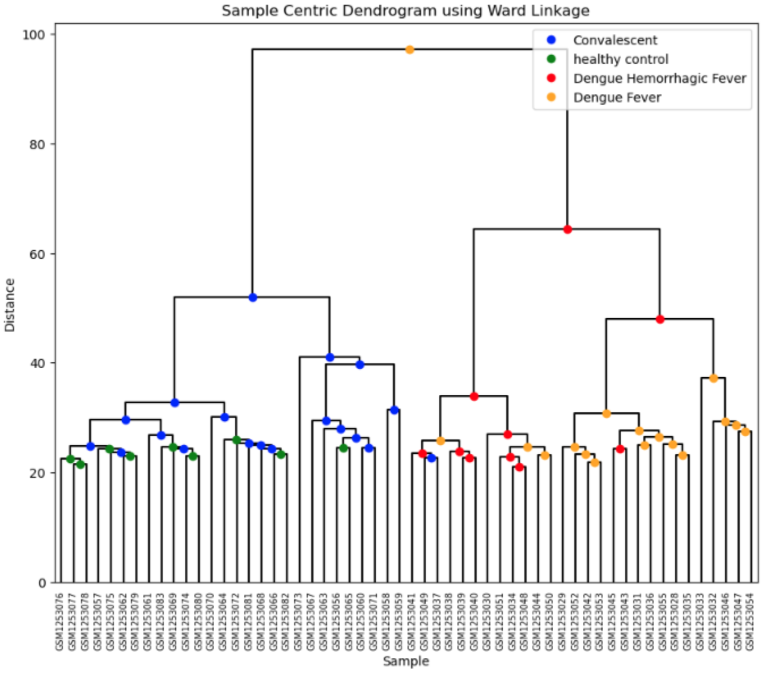


Figure 2: Sample Centric Hierarchical clustering dendrogram using ward linkage. Displays clustering pattern of the four populations.

Figure 1 is the result of sample-centric PCA with 2 components. The first 2 principal components were able to capture a total 25.6% variance of the original dataset. This is usually a low variance; however, a clear trend was still able to be seen in the PCA plot, when the 2 principal components were plotted (Figure 1). From the data, we can see that the population of the sample has a clear effect on the gene expression profile of the individual. Samples from individuals in a Convalescent disease state are shown to have a similar gene expression profile compared to the healthy control. An individual in a convalescent state in recovery would ideally have a much more similar gene expression in comparison to one who was diagnosed to have dengue fever or a dengue haemorrhagic fever state. This is clearly shown by a clear intersection present in Figure 1, as the convalescent samples cluster alongside the healthy control samples. Furthermore, the other two disease states are shown to have a similar gene expression profile due to it clustering together, and this can be since DHF is a related illness to that of dengue fever. For easier visual interpretation of the clustering that is present in the PCA plot, a sample-centric hierarchical clustering dendrogram using ward linkage was plotted, as shown in Figure 2. Ward linkage was used due to its ability to minimise the variance within clusters, and handle noise as well as outliers. Using the coloured plots, it is easier to visualise their convalescent samples (blue dots) and control (green dots) samples clustering together, whilst the other two populations do the same (red and orange dots), indicating similar gene expression profiles between the 2 populations.

Volcano plots Analysis:

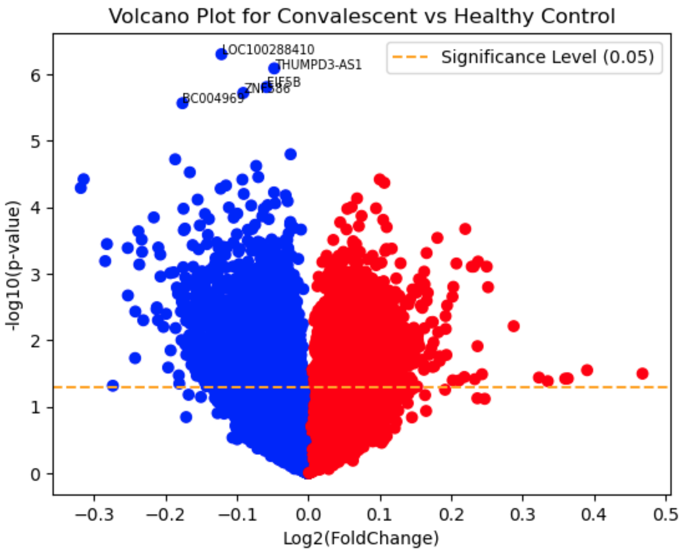


Figure 3: Volcano plot displaying differential gene expression between Convalescent and healthy control populations. The following significant genes were annotated, LOC100288410, THUMPD3-AS1, EIF5B, ZNF586 and BC004969.

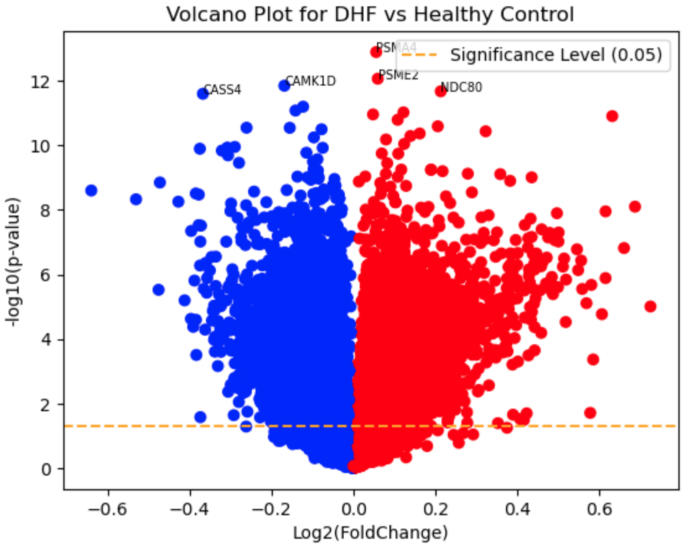


Figure 4: Volcano plot displaying differential gene expression between Dengue Haemorrhagic Fever and healthy control populations. The following significant genes were annotated, PSMA4, PSME2, CAMK1D, NDC80 and CASS4.

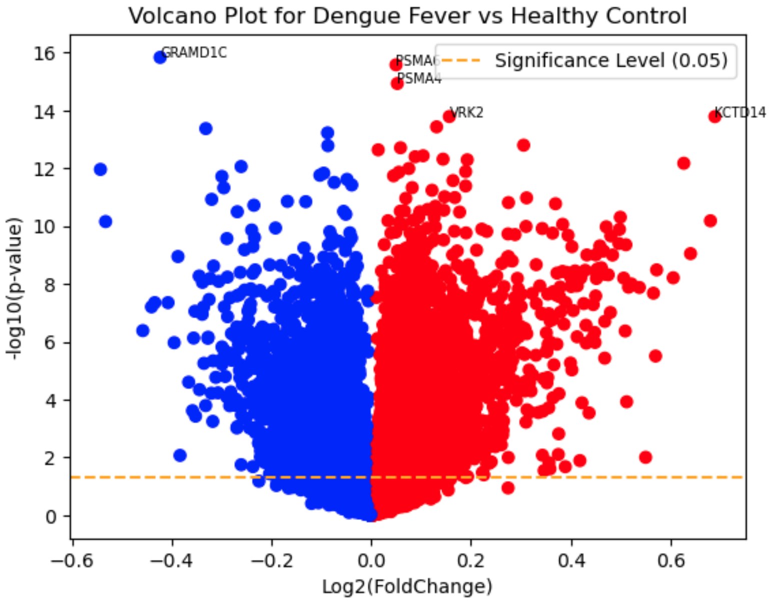


Figure 5: Volcano plot displaying differential gene expression between Dengue Fever and healthy control populations. The following significant genes were annotated, GRAMD1C, PSMA6, PSMA4, VRK2 and KCTD14.

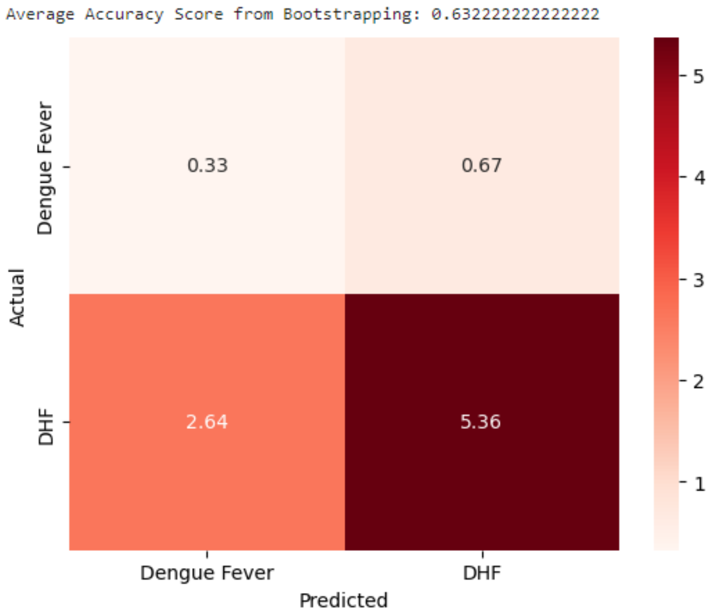
The Figures 3, 4 and 5 display volcano plots for convalescent, DHF and dengue fever populations against the healthy control, respectively. A threshold for up and down-regulated genes was not included due to the small dataset. Furthermore, implementing thresholds may cause us to miss genes that may have importance due to them falling outside of the expression threshold. Therefore, the top five significant genes were annotated based on their p-value and the higher expressed genes are shown as red and lower as blue.

Figure 3 displays the differential gene expression between the convalescent and healthy control populations. The fold-change range of -0.3 to 0.5, and the range of the y-axis can allow us to conclude that there were not as many over/under-expressed genes that were significant (above significance value -$\log_{10}$(0.05))for an individual in a convalescent state in comparison to an individual with DHF(many more significant genes as well genes that have a higher fold-change). This makes sense as an individual who is in convalescence is on his way to becoming healthy, and the comparison of the gene expression profiles should not be too different as shown in the PCA and HCA analysis. The top five significant genes for this volcano plot were LOC100288410, THUMPD3-AS1, EIF5B, ZNF586 and BC004969 and were all shown to be under-expressed.

The differential gene expression between Dengue Haemorrhagic Fever and the healthy control populations is shown in Figure 4. This diagram shows there to be many more significant genes as well as those that are over and under-expressed. The expression of numerous genes is affected when the individual is in a DHF disease state in comparison to a healthy person. This is in line with the conclusions made from the PCA and HCA analysis which is that these two populations have significantly different gene expression profiles. The top five significant genes for the data shown in Figure 4 were PSMA4, PSME2, CAMK1D, NDC80 and CASS4. CASS4 and CAMK1D were shown to be under-expressed whilst the remaining three were overexpressed.

The final volcano plot displayed the differential expression between dengue fever and the healthy control population. This volcano plot displayed similar patterns in terms of expression as DHF against the control population. A vast majority of genes were significantly under-expressed and overexpressed, the top five significant genes for this comparison were GRAMD1C, PSMA6, PSMA4, VRK2 and KCTD14. Two genes are shown to be extremely significant, one of them being overexpressed (KCTD14) and the other under-expressed (GRAMD1C). These genes may be involved in a pathway that leads to the progression of the disease, and because of it, they become differentially expressed.

SVM Machine Learning Classifier and Confusion Matrix:



```
Average Accuracy Score from Bootstrapping: 0.632222222222222
```

Figure 6: Displays the average score from Bootstrapping SVM as well as its confusion matrix

An SVM classifier was produced to distinguish between patients with dengue fever and dengue haemorrhagic fever. The data was split into 70% training data (19 samples) and 30% test data (9 samples). Bootstrapping was used to run the classifier with one hundred iterations and an average accuracy score of 63.2% was calculated. This result is broken down in the Average confusion matrix, which is shown in the confusion matrix, in Figure 6. The matrix shows that the classifier correctly predicts a DHF patient 5.36 times and incorrectly predicts it to be dengue fever 0.67 times, on average out of the one hundred iterations. Furthermore, it correctly predicts the disease state as dengue fever 0.33 times, whilst mistaking it for DHF 2.64 times. The percentages can be calculated by dividing each value by the test sample number of nine. An accuracy level of 63.2% is not ideal when looking from a medical perspective, and thus the accuracy of this classifier may be improved in several ways. The classifier can be optimised using hyperparameter tuning and choosing a different kernel type or regularisation factor (C) may improve performance. Furthermore, having a larger data set would allow a larger set of training data for the classifier, which should improve its accuracy in distinguishing between the two disease populations.

Conclusion:

In conclusion, the disease state of an individual was shown to influence the gene expression profile of an individual. Individuals in a convalescent state were shown to have a similar profile to the healthy controls, and the other two populations displayed similar gene expression profiles. These conclusions were supported by the PCA, HCA and volcano plot analysis. Being in a dengue fever or dengue haemorrhagic state resulted in higher differential expression of numerous genes, which may play a role in the progression of each condition. The SVM classifier was shown to have a 63.2% accuracy rate, from a medical perspective this may not be good enough, however, the accuracy of these classifiers can be improved by optimising certain parameters.