BIO720P: AI & DATA SCIENCE IN BIOLOGY 2023

Coursework

Gaurab Rana 170921895

Abstract:

This practical aims to investigate the effects of dengue infection on the human genomes. A dengue transcriptomics data set is provided of expression of genes amongst a sample of patients with different disease states (healthy, dengue fever, dengue haemorrhagic fever and convalescence).

This data set is then subjected to multiple analysis methods to extract key information such as principal component analysis (PCA), gene set enrichment pathway analysis, differential gene expression analysis and using hierarchical cluster analysis to visualize the data into heatmaps and cluster grams.

The key findings of this practical was that there are key genes which are differentially expressed in infected individuals (dengue fever and dengue haemorrhagic fever patients). These key genes are used to explain different cluster patterns in PCA's between healthy and convalescent individuals from the infected individuals.

The key genes undergo pathway analysis to observe which pathways they are involved in. The importance of these key genes is great as they can be studied to understand the effects of dengue and how the different pathways are affected. We can also use the key genes as drug targets for development of medicine and treatment.

Introduction:

Dengue is a viral infection that currently has no cure and the only treatment for it are drugs to alleviate symptoms such as pain. Dengue infections are contracted from contact with mosquitos and in some cases the infection can develop into severe dengue.[1] The aim of this practical is to investigate the effects of dengue infection on human gene expression. Samples were collected from different population groups to gain data from patients contracted with dengue at different stages of infection (dengue fever, dengue haemorrhagic fever and convalescence) and some healthy non-infected samples used as control[2]. The samples were then subject to gene microarrays to obtain a gene expression transcriptomics dataset of the samples.

The dataset then underwent further downstream analysis, such as principal component analysis (PCA) and hierarchical cluster analysis (HCA) to observe any patterns or clusters within samples, differential gene expression (DGE) analysis to extract any genes which are significantly expressed in different population groups, gene enrichment analysis to observe statistical significance of genes within certain groups, pathway analysis to observe the effects of certain genes on different biological pathways.

Methods:

The data in this practical went analysis using python and certain python packages. The following modules and libraries in python were used during this practical for data wrangling: pandas, matplotlib, numpy, pickle. For data visualization and methods for analysis, the following modules and libraries were used: sklearn and scanpy (PCA), scipy (HCA and dendrograms), seaborn (heatmaps and clustermaps), pydeseq2 (differential expression analysis) and gseapy (gene set enrichment analysis).

The meta data of our dengue data states that our patient samples are grouped into 4 different categories of disease state: convalescent, dengue haemorrhagic fever, dengue fever and a healthy disease-free control.

Firstly the dengue data set[2] is loaded into python and an unsupervised learning method such as principal component analysis is used to identify any patterns within our data. A principal component analysis (PCA) is a dimension reduction method of a data set to analyze large data sets in a lower dimensional space. It transforms the data into principal components of highest variance to cluster the data. The components are created by T=XW, where it uses the data matrix of weights matched with the corresponding column loadings[3].

In theory, the expected results of the dengue data PCA could be four different distinct clusters as represented by the four different population groups within our samples, as they might display different gene expression profiles in certain genes which allow them to cluster differently. Using the sklearn and scanpy libraries, a PCA was performed.

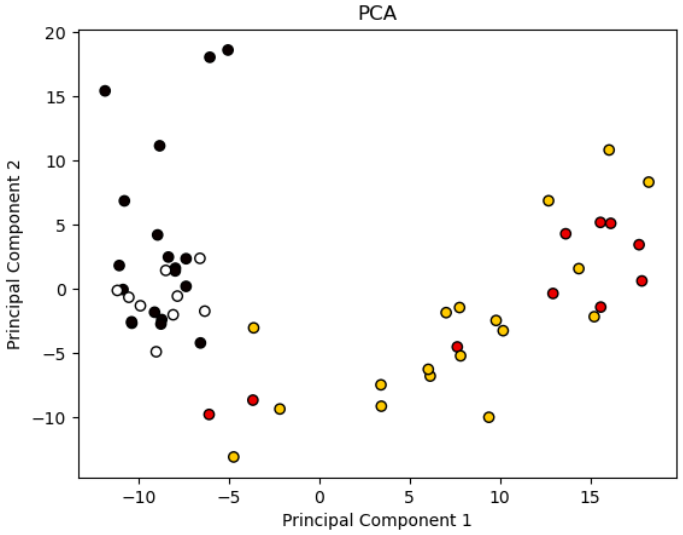

***Figure 1a**: Principal component analysis of the dengue data set using **scipy** library in python.*

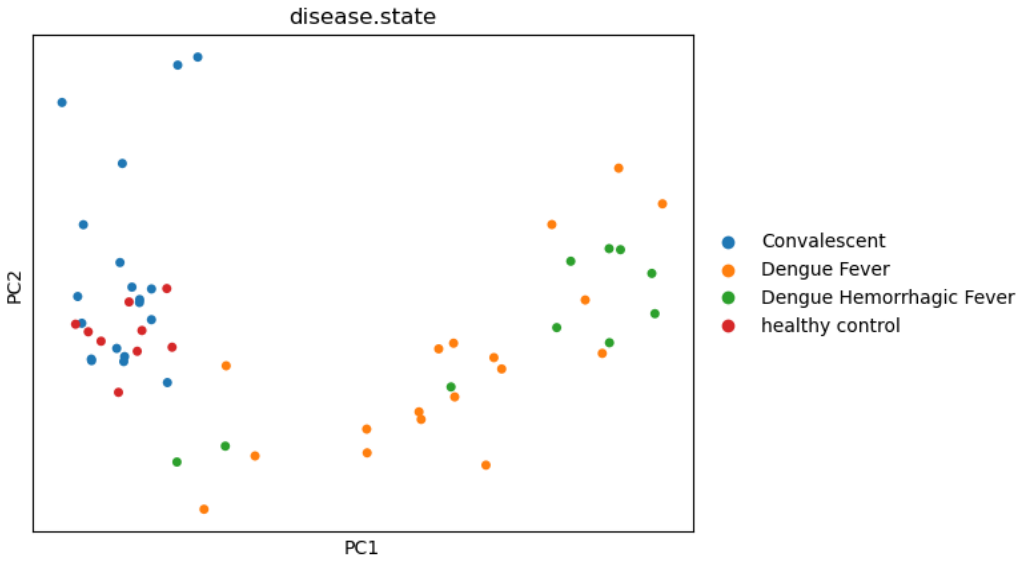

***Figure 1b**: Principal component analysis of dengue data set using **scanpy** library in python.*

The scanpy PCA (*Figure 1b*) labels the points to help identify which points belongs to which groups. A scanpy PCA was also performed because it uses the AnnData object which was extracted from another data analysis method of differential gene expression.

Next, PyDESeq2 is used as a python package for differential expression analysis of the genes within the samples of our dengue data set. Using the PyDESeq2 example[4], a DEA analysis was followed on the dengue data set. A single factor analysis was done using the disease state as the condition. A DeseqDataSet object is created from our dengue data set and the dengue meta data, to fit and store dispersion and log-fold change parameters. We specify DeseqDataSet method to have disease state as our design factor.
Then the deseq2() method is used to fit dispersions and log-fold changes. You can print the DeseqDataSet (DDS) to get information about our object. It shows that the dds extends on the AnnData class, which we used with scanpy to get a PCA.
A statistical analysis is then used on our DDS to gather p-values and adjusted p-values for differential expression of our genes. The method for this is DeseqStats, we can specify a contrast in this method to measure the log-fold changes of certain conditions(design factors which in our practical are the disease states) compared to other conditions. A summary method can be used which uses the Wald test to get p-values, cooks filtering and multiple test adjustments.

Figure 2a:
A statistical analysis result of a DeseqDataSet using the summary method on our dengue data set comparing log-fold change between CTRL and DENV.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
DDR1	2.353701	0.135180	0.019923	6.785160	1.159576e-11	5.322077e-10
RFC2	2.867595	-0.087704	0.017425	-5.033274	4.821728e-07	7.126894e-06
HSPA6	3.286650	0.044469	0.027550	1.614102	1.065054e-01	2.371352e-01
PAX8	1.837859	-0.006398	0.020221	-0.316417	7.516864e-01	8.609419e-01
GUCA1A	1.772470	-0.000276	0.019499	-0.014142	9.887168e-01	9.940705e-01
...	...	...	...	...	...	...
AW014299	2.858665	-0.000462	0.007697	-0.060021	9.521390e-01	9.760899e-01
AW025284	2.731239	0.006287	0.010177	0.617727	5.367556e-01	7.073445e-01
FAM86DP	2.134226	-0.130718	0.045311	-2.884919	3.915147e-03	1.739711e-02
AI424872	2.389028	0.086159	0.021987	3.918615	8.905921e-05	7.027539e-04
AI744451	2.470226	-0.066841	0.025706	-2.600234	9.316009e-03	3.552931e-02

These statistical analysis results of our genes can then be filtered further by subsetting p-value to <0.05 and log2FoldChange to >(+/-0.5) to return statistically significant genes which have high up or down regulation. This data can be visualized in a bar chart using matplotlib.

	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
BIRC5	2.438818	-0.522772	0.051169	-10.216670	1.669793e-24	9.321329e-22
CDC6	2.225591	-0.681235	0.051421	-13.248108	4.627231e-40	2.221441e-36
TTK	2.378845	-0.573300	0.062604	-9.157509	5.310920e-20	1.400916e-17
KDM5D	2.539530	-0.551068	0.193263	-2.851395	4.352784e-03	1.904561e-02
SPC25	2.415790	-0.606491	0.068509	-8.852658	8.545904e-19	1.881981e-16
CXCL11	1.946451	-0.512840	0.099435	-5.157527	2.502327e-07	4.001723e-06
SKA1	2.205787	-0.510417	0.071633	-7.125454	1.037383e-12	6.088345e-11
KIF4A	2.399308	-0.511591	0.050589	-10.112763	4.849701e-24	2.476856e-21
CEP55	2.519235	-0.627634	0.049505	-12.678264	7.804163e-37	2.081457e-33
DTL	2.588402	-0.500404	0.044024	-11.366745	6.122877e-30	8.165197e-27
PBK	2.384060	-0.641181	0.059274	-10.817273	2.851357e-27	2.581073e-24
KIF15	2.349688	-0.538681	0.058068	-9.276698	1.748125e-20	5.311646e-18
KCTD14	2.238563	-0.689878	0.036880	-18.705782	4.441699e-78	1.066185e-73
E2F8	2.356538	-0.566834	0.068667	-8.254884	1.520503e-16	2.097595e-14
CNTNAP3	2.141255	0.541462	0.050240	10.777600	4.391939e-27	3.765146e-24
CDCA2	2.261169	-0.516954	0.052239	-9.895895	4.337131e-23	1.964311e-20
DEPDC1B	2.267570	-0.507291	0.059410	-8.538855	1.355579e-17	2.213559e-15
AI401105	2.225654	-0.571089	0.087085	-6.557797	5.460832e-11	2.214220e-09
CNTNAP3B	2.294635	0.531244	0.061458	8.644055	5.425232e-18	9.791524e-16

Figure 2b: A total list of significant differentially expressed genes which highly up or down regulated when comparing between control and dengue fever populations

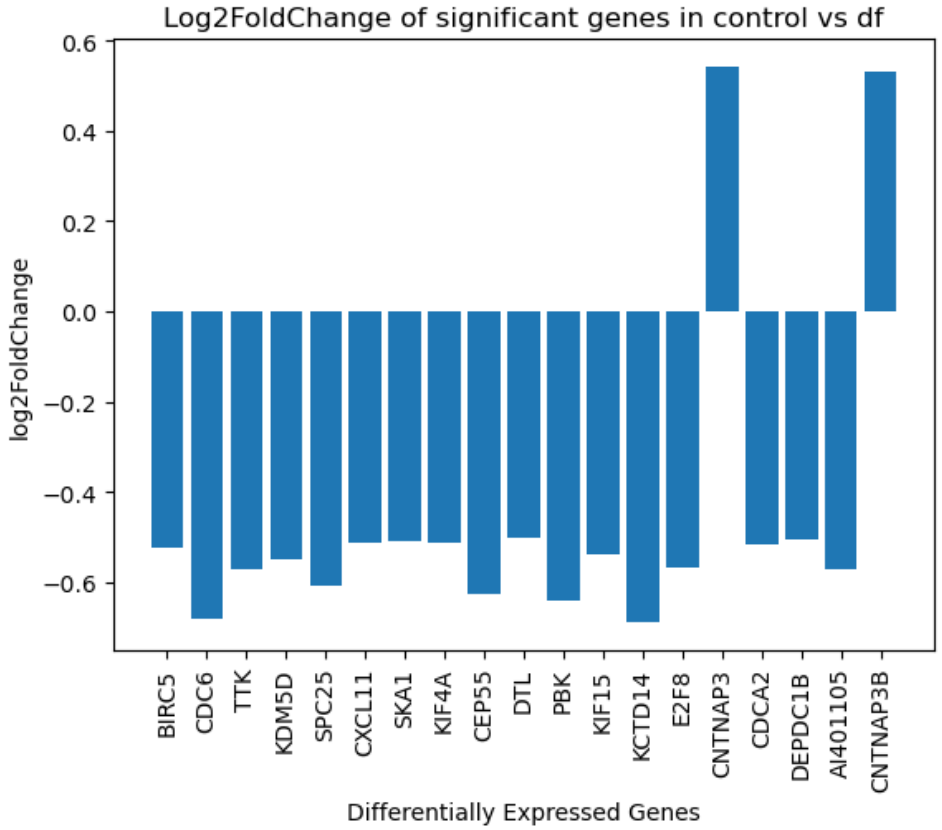

Figure 2c: a bar plot of log2foldchanges in our statistically significant differentially expressed genes when comparing log2foldchange values between control and dengue fever populations

These statistical analysis can be done on our other populations by changing the contrast to compare between them and getting their significant differentially expressed genes.

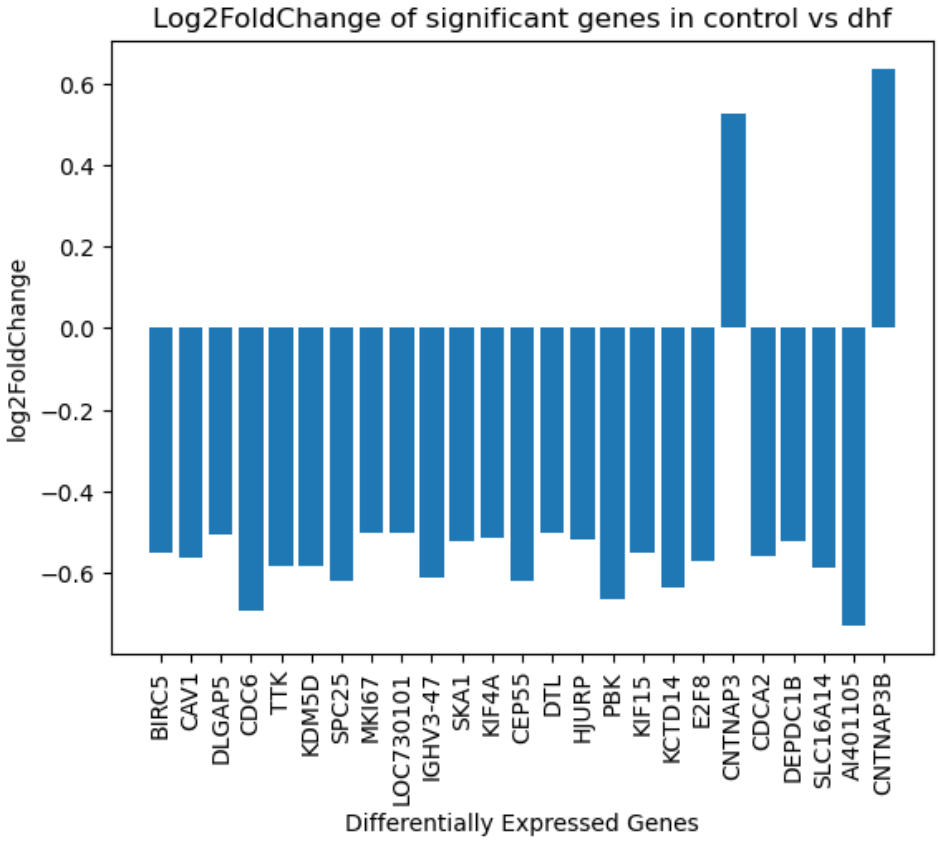

Figure 2d: a bar plot of log2foldchanges in our statistically significant differentially expressed genes when comparing log2foldchange values between control and dengue haemorrhagic fever populations.

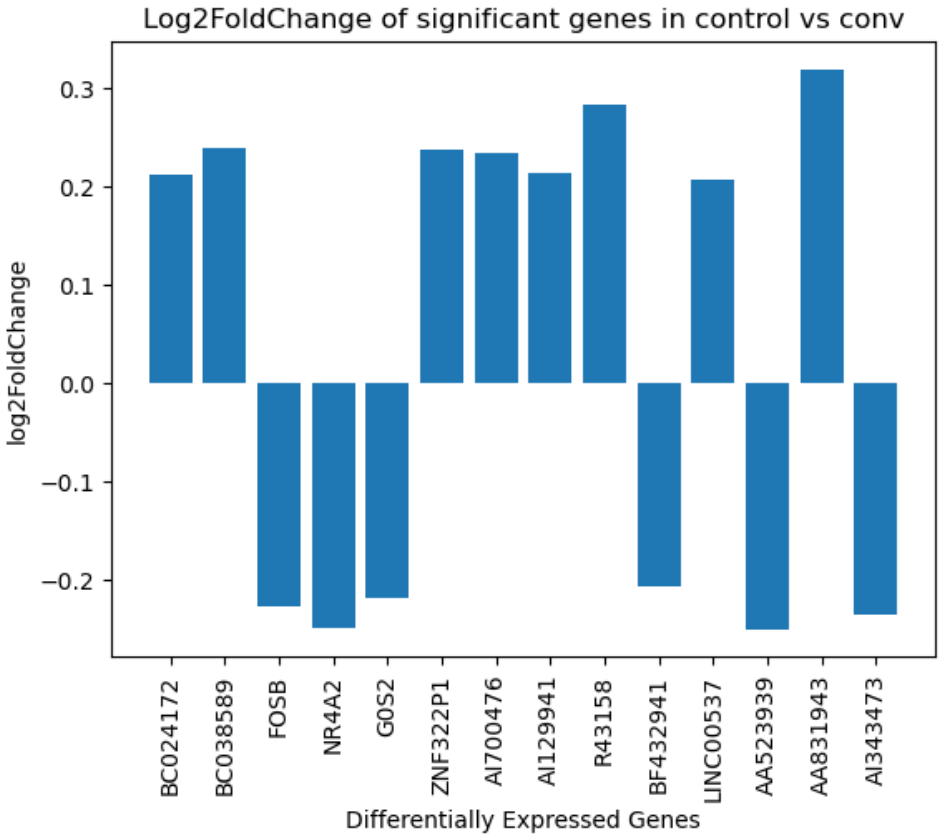

Figure 2e: a bar plot of log2foldchanges in our statistically significant differentially expressed genes when comparing log2foldchange values between control and convalescent populations.

To further analyze these differentially expressed genes, gene set enrichment analysis (GSEA) was used. GSEApy[5] is loaded into python and using the enrichr method which allows one to input your genes of interest as a gene list and search for those gene sets in a database for pathway analysis. For the the 'control vs df' population, the genes of interest are gathered from subsetting the log2foldchange and p-values to get a gene list of statistically significant differentially expressed genes, as shown in figure 2c. This gene list was then searched through the 'Reactome 2022' and 'KEGG_2021_human' database to return terms and results of pathways which our genes are involved in.

	Gene_set	Term	Overlap	P-value	Adjusted P-value	Old P-value	Old Adjusted P-value	Odds Ratio	Combined Score	Genes
0	Reactome_2022	Unattached Kinetochores Signal Amplification V...	3/93	0.000089	0.002854	0	0	41.439583	386.335699	BIRC5;SKA1;SPC25
1	Reactome_2022	EML4 And NUDC In Mitotic Spindle Formation R-H...	3/97	0.000101	0.002854	0	0	39.668218	364.851522	BIRC5;SKA1;SPC25
2	Reactome_2022	Cell Cycle Checkpoints R-HSA-69620	4/271	0.000109	0.002854	0	0	19.689388	179.675887	BIRC5;CDC6;SKA1;SPC25
3	Reactome_2022	Resolution Of Sister Chromatid Cohesion R-HSA-...	3/106	0.000132	0.002854	0	0	36.185680	323.287328	BIRC5;SKA1;SPC25
4	Reactome_2022	Mitotic Spindle Checkpoint R-HSA-69618	3/110	0.000147	0.002854	0	0	34.825935	307.316387	BIRC5;SKA1;SPC25
...	...	...	...	...	...	...	...	...	...	...
103	KEGG_2021_Human	Hepatitis B	1/162	0.143239	0.198082	0	0	6.839199	13.290218	BIRC5
104	KEGG_2021_Human	Hippo signaling pathway	1/163	0.144059	0.198082	0	0	6.796639	13.168688	BIRC5
105	KEGG_2021_Human	Chemokine signaling pathway	1/192	0.167535	0.204765	0	0	5.756254	10.283924	CXCL11
106	KEGG_2021_Human	Cytokine-cytokine receptor interaction	1/295	0.246077	0.270685	0	0	3.720144	5.216055	CXCL11
107	KEGG_2021_Human	Pathways in cancer	1/531	0.400404	0.400404	0	0	2.038889	1.866155	BIRC5

Figure 3a: The results of a GSEA enrich pathway analysis for statistically significant differentially expressed genes when comparing log2foldchange values between control and dengue fever populations.

The results can be visualized in a barplot as such:

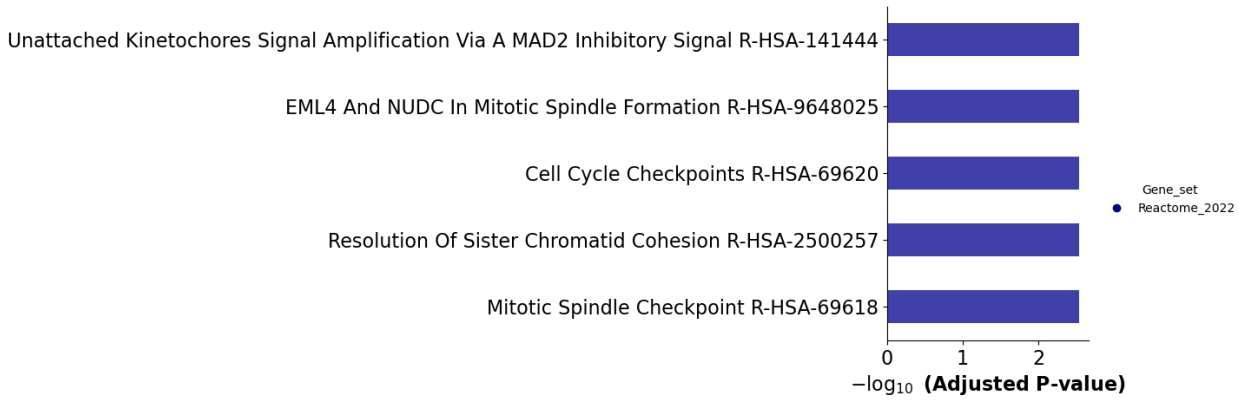

Figure 3b: the top 5 terms and pathways that our genes are involved in for statistically significant differentially expressed genes when comparing log2foldchange values between control and dengue fever populations.

The GSEA enrichr method is applied to the other comparison groups as well.

	Gene_set	Term	Overlap	P-value	Adjusted P-value	Old P-value	Old Adjusted P-value	Odds Ratio	Combined Score	Genes
0	Reactome_2022	Unattached Kinetochores Signal Amplification V...	3/93	0.000208	0.008454	0	0	30.128788	255.452241	BIRC5;SKA1;SPC25
1	Reactome_2022	EML4 And NUDC In Mitotic Spindle Formation R-H...	3/97	0.000235	0.008454	0	0	28.840909	240.945144	BIRC5;SKA1;SPC25
2	Reactome_2022	Resolution Of Sister Chromatid Cohesion R-HSA-...	3/106	0.000306	0.008454	0	0	26.308914	212.913973	BIRC5;SKA1;SPC25
3	Reactome_2022	Cell Cycle Checkpoints R-HSA-69620	4/271	0.000333	0.008454	0	0	14.059568	112.569629	BIRC5;CDC6;SKA1;SPC25
4	Reactome_2022	Mitotic Spindle Checkpoint R-HSA-69618	3/110	0.000341	0.008454	0	0	25.320306	202.156604	BIRC5;SKA1;SPC25
...	...	...	...	...	...	...	...	...	...	...
133	KEGG_2021_Human	Focal adhesion	1/201	0.223278	0.289156	0	0	4.119792	6.176957	CAV1
134	KEGG_2021_Human	Proteoglycans in cancer	1/205	0.227194	0.289156	0	0	4.038194	5.984407	CAV1
135	KEGG_2021_Human	Endocytosis	1/252	0.271809	0.313314	0	0	3.274236	4.265200	CAV1
136	KEGG_2021_Human	Prion disease	1/273	0.290935	0.313314	0	0	3.018229	3.726477	CAV1
137	KEGG_2021_Human	Pathways in cancer	1/531	0.489890	0.489890	0	0	1.528695	1.090838	BIRC5

Figure 3c: The results of a GSEA enrich pathway analysis for statistically significant differentially expressed genes when comparing log2foldchange values between control and dengue haemorrhagic fever populations.

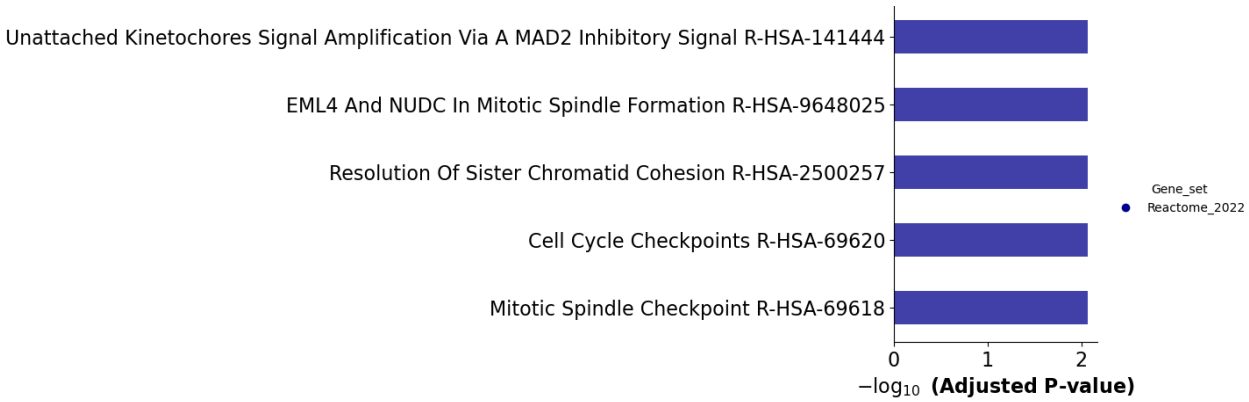

Figure 3d: the top 5 terms and pathways that our genes are involved in for statistically significant differentially expressed genes when comparing log2foldchange values between control and dengue haemorrhagic fever populations.

	Gene_set	Term	Overlap	P-value	Adjusted P-value	Old P-value	Old Adjusted P-value	Odds Ratio	Combined Score	Genes
0	Reactome_2022	SUMOylation Of Intracellular Receptors R-HSA-4...	1/29	0.020116	0.216495	0	0	54.829670	214.177558	NR4A2
1	Reactome_2022	NGF-stimulated Transcription R-HSA-9031628	1/39	0.026965	0.216495	0	0	40.380567	145.903491	FOSB
2	Reactome_2022	Nuclear Receptor Transcription Pathway R-HSA-3...	1/53	0.036479	0.216495	0	0	29.488166	97.635762	NR4A2
3	Reactome_2022	Nuclear Events (Kinase And Transcription Facto...	1/61	0.041877	0.216495	0	0	25.546154	81.058519	FOSB
4	Reactome_2022	Signaling By NTRK1 (TRKA) R-HSA-187037	1/114	0.076934	0.216495	0	0	13.528251	34.697431	FOSB
5	Reactome_2022	PPARA Activates Gene Expression R-HSA-1989781	1/116	0.078233	0.216495	0	0	13.291639	33.867968	G0S2
6	Reactome_2022	Regulation Of Lipid Metabolism By PPARalpha R-...	1/118	0.079530	0.216495	0	0	13.063116	33.070789	G0S2
7	Reactome_2022	Estrogen-dependent Gene Expression R-HSA-9018519	1/119	0.080179	0.216495	0	0	12.951760	32.683753	FOSB
8	Reactome_2022	Signaling By NTRKs R-HSA-166520	1/132	0.088566	0.216495	0	0	11.658837	28.261096	FOSB
9	Reactome_2022	SUMO E3 Ligases SUMOylate Target Proteins R-HS...	1/168	0.111424	0.227134	0	0	9.128973	20.032748	NR4A2
10	Reactome_2022	SUMOylation R-HSA-2990846	1/174	0.115181	0.227134	0	0	8.809693	19.039932	NR4A2
11	Reactome_2022	ESR-mediated Signaling R-HSA-8939211	1/188	0.123891	0.227134	0	0	8.144385	17.008326	FOSB
12	Reactome_2022	Signaling By Nuclear Receptors R-HSA-9006931	1/260	0.167442	0.283364	0	0	5.858925	10.470579	FOSB
13	Reactome_2022	Signaling By Receptor Tyrosine Kinases R-HSA-9...	1/496	0.296504	0.465935	0	0	3.028904	3.682224	FOSB
14	Reactome_2022	Metabolism Of Lipids R-HSA-556833	1/732	0.406777	0.596606	0	0	2.026202	1.822551	G0S2
15	Reactome_2022	Generic Transcription Pathway R-HSA-212436	1/1190	0.576456	0.754028	0	0	1.216083	0.669888	NR4A2
16	Reactome_2022	RNA Polymerase II Transcription R-HSA-73857	1/1312	0.613347	0.754028	0	0	1.095758	0.535634	NR4A2
17	Reactome_2022	Post-translational Protein Modification R-HSA-...	1/1383	0.633419	0.754028	0	0	1.035512	0.472838	NR4A2
18	Reactome_2022	Gene Expression (Transcription) R-HSA-74160	1/1449	0.651206	0.754028	0	0	0.984807	0.422413	NR4A2
19	Reactome_2022	Metabolism Of Proteins R-HSA-392499	1/1890	0.750980	0.817050	0	0	0.736939	0.211041	NR4A2
20	Reactome_2022	Metabolism R-HSA-1430728	1/2049	0.779911	0.817050	0	0	0.673753	0.167478	G0S2
21	Reactome_2022	Signal Transduction R-HSA-162582	1/2465	0.841516	0.841516	0	0	0.547015	0.094388	FOSB
22	KEGG_2021_Human	Cocaine addiction	1/49	0.033770	0.099573	0	0	31.951923	108.259234	FOSB
23	KEGG_2021_Human	Amphetamine addiction	1/69	0.047246	0.099573	0	0	22.531674	68.775176	FOSB
24	KEGG_2021_Human	IL-17 signaling pathway	1/94	0.063847	0.099573	0	0	16.454094	45.269591	FOSB
25	KEGG_2021_Human	Aldosterone synthesis and secretion	1/98	0.066478	0.099573	0	0	15.772403	42.757158	NR4A2
26	KEGG_2021_Human	Parathyroid hormone synthesis, secretion and a...	1/106	0.071719	0.099573	0	0	14.564835	38.378242	NR4A2
27	KEGG_2021_Human	Osteoclast differentiation	1/127	0.085349	0.099573	0	0	12.124542	29.838641	FOSB
28	KEGG_2021_Human	Alcoholism	1/186	0.122652	0.122652	0	0	8.233264	17.276717	FOSB

Figure 3e: The results of a GSEA enrich pathway analysis for statistically significant differentially expressed genes when comparing log2foldchange values between control and convalescent populations.

A barplot could not be achieved for the ‘control vs convalescent’ populations as the terms were not statistically significant.

Using the seaborn library[6], methods such as hierarchical clustering are used to create hatmaps and clustergrams for further visualization of data. The clustermap method from seaborn is used with the set of significantly differentially expressed genes during log2foldchange comparison of control against dengue fever groups to create a clustergram of those genes against our samples.

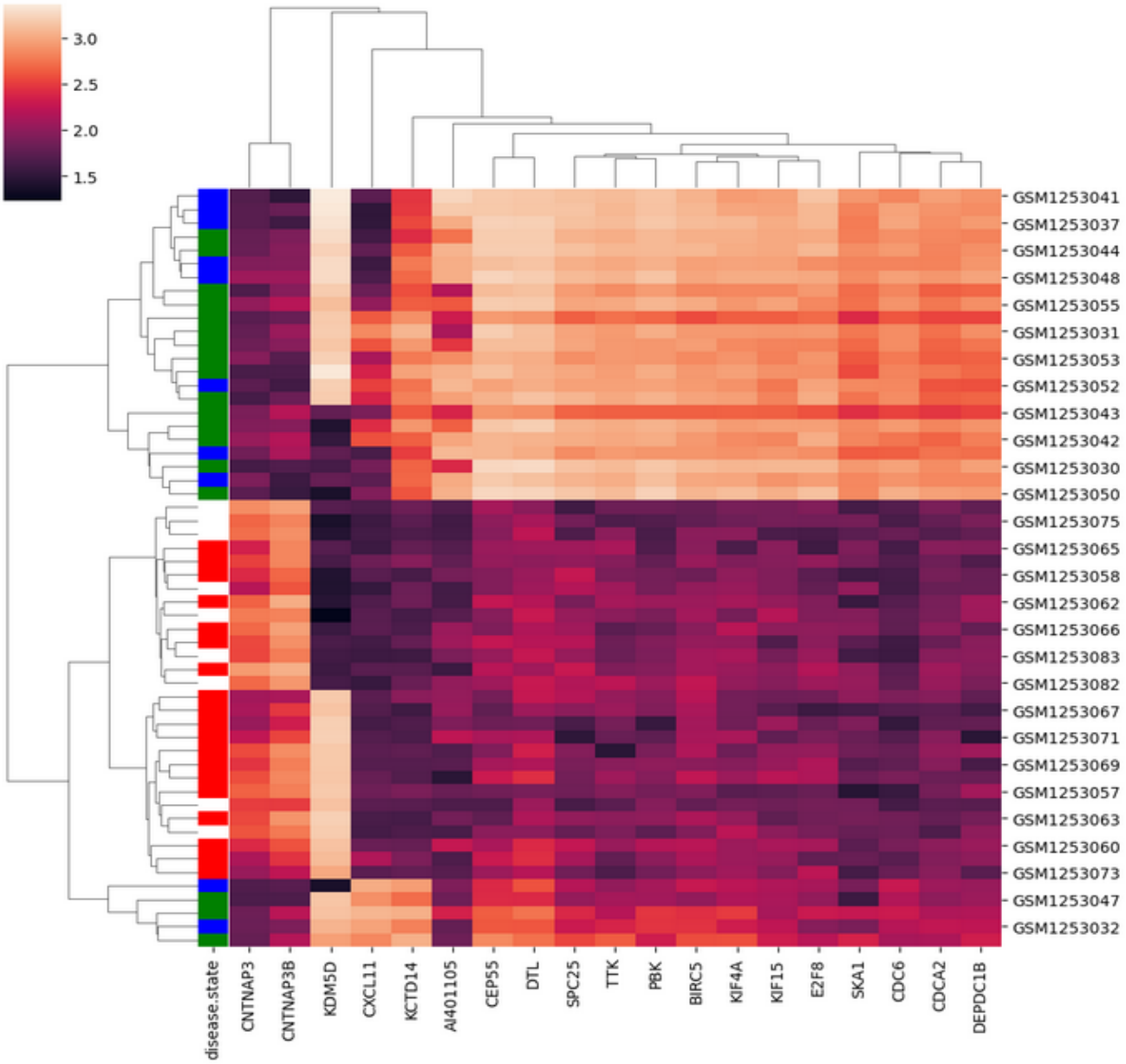

Figure 4a: clustergram of statistically significant differentially expressed genes when comparing log2foldchange values between control and dengue fever populations. (LEGEND: red=convalescent, white=healthy control, blue= dengue haemorrhagic fever, green=dengue fever)

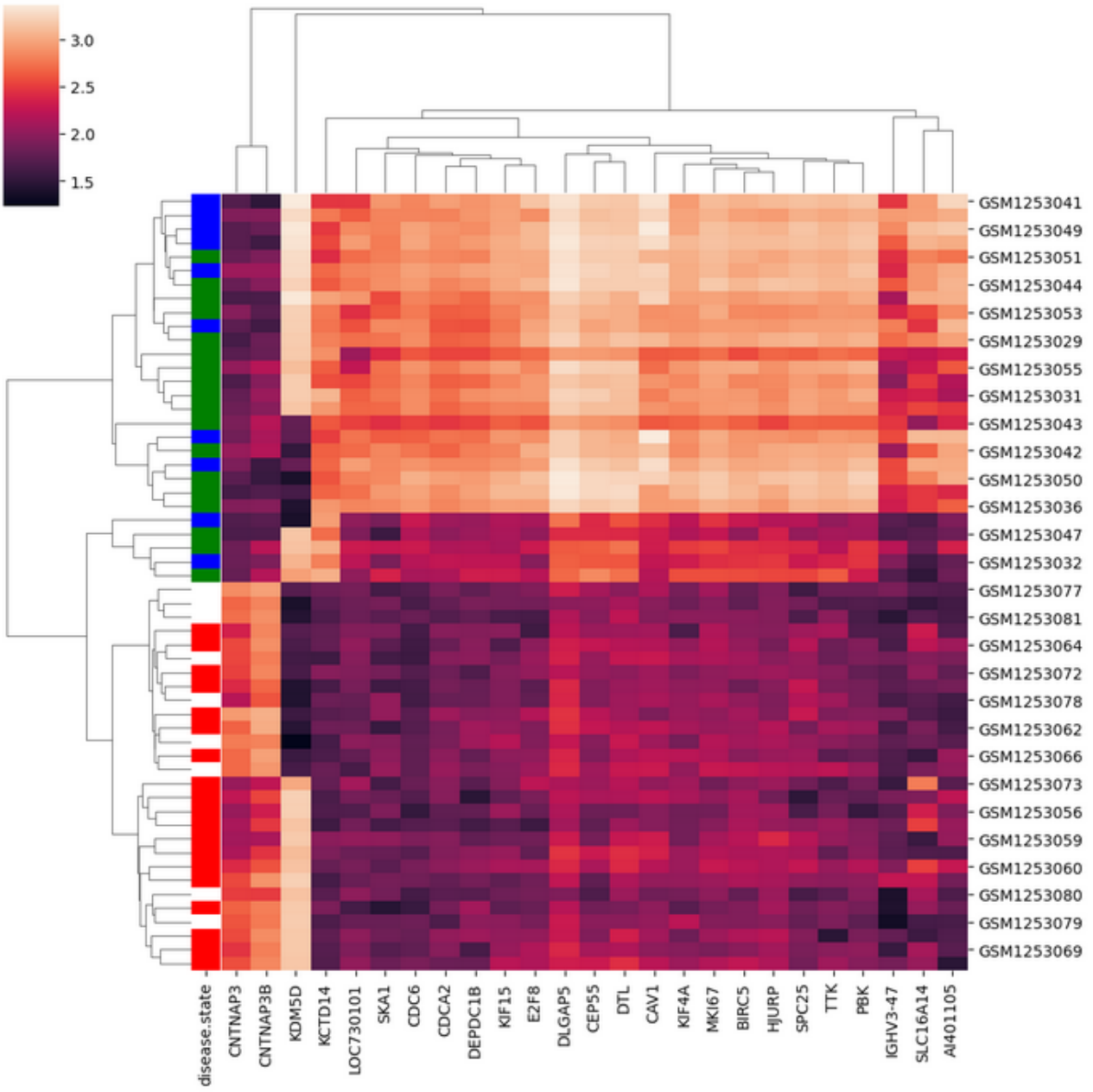

Figure 4b: clustergram of statistically significant differentially expressed genes when comparing log2foldchange values between control and dengue haemorrhagic fever populations. (LEGEND: red=convalescent, white=healthy control, blue= dengue haemorrhagic fever, green=dengue fever)

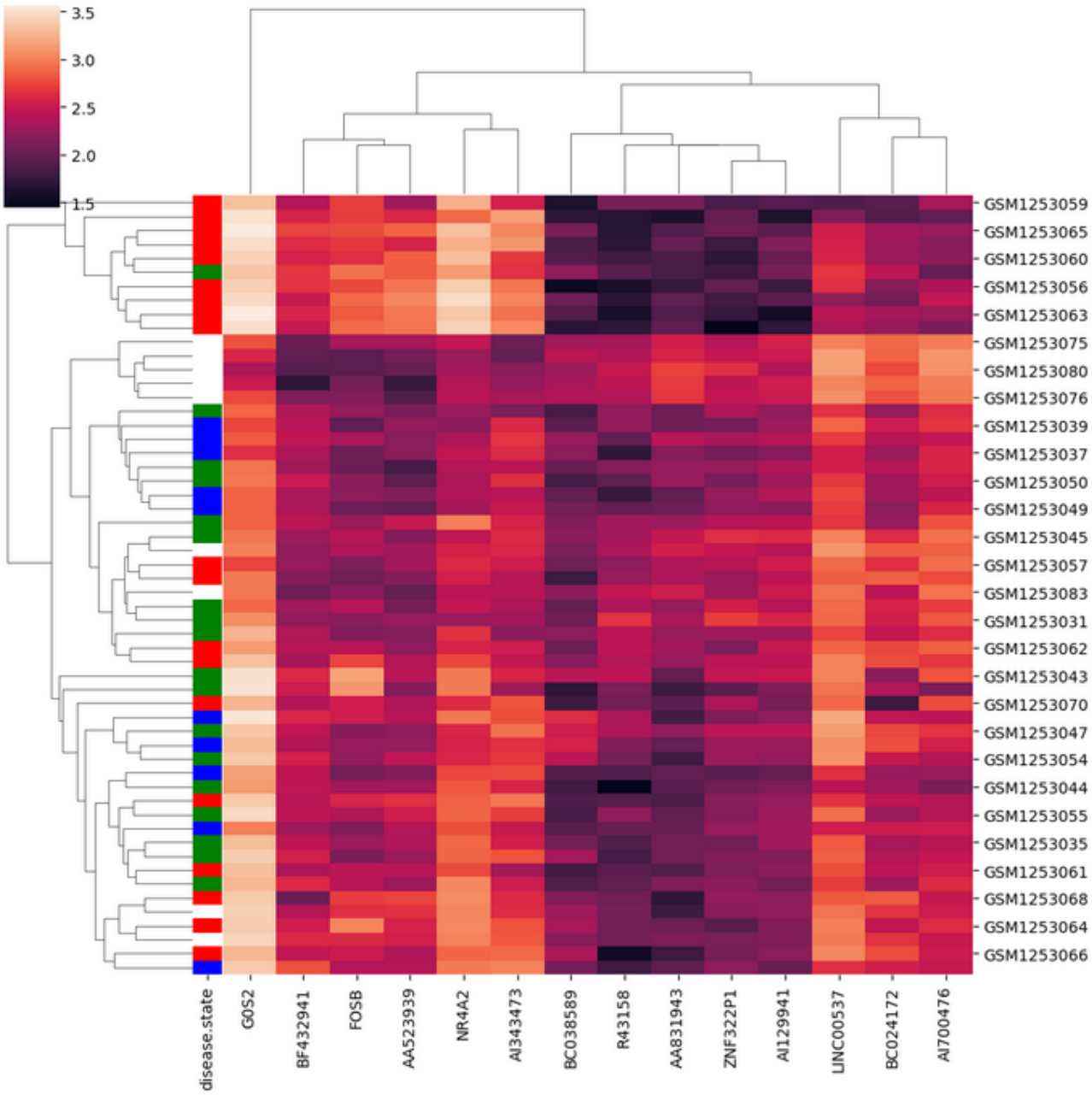

Figure 4c: clustergram of statistically significant differentially expressed genes when comparing log2foldchange values between control and convalescent populations. (LEGEND: red=convalescent, white=healthy control, blue= dengue haemorrhagic fever, green=dengue fever)

Results and discussion:

The dengue data set was firstly subjected to a data analysis of unsupervised learning method which includes principal component analysis. From the principal component analysis of our dengue data (*Figures 1a and 1b*), we can see from principal component 1 it seems that the healthy controls and convalescent populations within the samples form a cluster whereas the the dengue fever and dengue haemorrhagic fever populations are clustered more towards the higher values of PC1 and are also more widespread across the whole PCA.

The healthy populations and convalescent populations seem to cluster together, this could be because a convalescent patient is a recovering patient and as such contains less viral dna and so they share similar gene sets with similar gene expression.

The dengue fever and dengue haemorrhagic fever patients seem to be spread out more across the PCA, this could be because there is variation as some patients might have weaker immune systems and are affected worse by the dengue infection. Therefore this might cause fluctuations within gene expression profiles and explain the high variation within the dengue fever and dengue hemorrhagic fever data points in the PCA.

While we may not be able to distinguish clusters between populations of healthy control patients and convalescent patients, we may be able to distinguish the healthy controls and convalescent patients from patients with dengue fever (DF) or dengue haemorrhagic fever (DHF).

From the differential expression analysis (using PyDESeq2 package) and statistical tests to extract significantly high/low expressed genes between different populations we can see some trends and patterns. The subsetting of pvalue to less than 0.05 allows us to extract statistically significant genes from the stats analysis results (Figure 2a), however a further subset to the log2foldchange allows us to extract differentially expressed genes which have increased regulation or decreased regulation. As log2foldchange is in a log scale, a value of 1 shows a double in gene expression and so even small values such as 0.5 can show a significant increase in expression.

There were no huge changes in gene regulation as most of the log2foldchanges were around +/-0.3, and the highest changes in regulations were +/-0.5 therefore I chose to subset my genes of interest to this log2foldchange of +/-0.5.

However when comparing the log2foldchanges between control and convalescent populations, the highest log-fold changes (LFC) were seen about the +/- 0.25, and rest of the genes had very subtle changes to LFC's which were around 0.

From figures 2c and 2d, the log2foldchange 'control vs df' and 'control vs dhf' groups actually have similar genes with similar expression. In fact, all the genes from 'control vs df' except 'CXCL11' are also present in the significantly differentially expressed genes for 'control vs dhf' with similar log2foldchange gene expression. These shared differentially expressed genes could between both 'control vs df' and 'control vs dhf' could be the reason why we see a distinction between clusters of control population and dhf or df populations.

Gene set enrichment analysis (GSEA) is important allows us to identify key biological pathways which are affected or associated with dengue infection, this can be by showing us specific pathways which are up or downregulated when infected with the virus. This can also help in drug discovery by targeting those specific pathways.

The GSEA pathway analysis results pulled some interesting data. In figures 3b and 3d the top 5 pathways dengue fever and dengue haemorrhagic fever are involved in are the same. This makes sense as they share a similar set of significantly differentially expressed genes. These pathways can be investigated further to provide a better understanding of how dengue fever affects humans and can also be a guideline into treatment as these pathways could be involved in drug targeting for treatment of dengue.

While the control vs convalescent genes did not provide statistically significant results, we can still look at the pathways which the significant differentially expressed genes are involved in from figure 3e. It seems to affect a lot of transcription pathways from the reactome 2022 database.

The clustergrams of dengue fever and dengue haemorrhagic fever (figures4a and 4b) look similar as they do share similar genes in the differentially expressed gene analysis. From the gene expression profiles in the clustergram, it is possible to separate control and convlaescent populations (white and red) disease states from dengue fever and dengue haemorrhagic fever (blue and green) disease states as they form distinct patterns in the clustergram(fig4b).

These genes (fig 4b) can therefore be used to distinguish between healthy and convalescent populations from dengue fever and dengue haemorrhagic fever populations. Therefore further research can be made on these key genes to create a machine learning method to distinguish healthy or recovering patients to infected patients.

However it might prove difficult to distinguish between patients with dengue fever and dengue haemorrhagic fever as there seem to be no distinct clusters in our clustergrams (fig 4a and 4b).

Figure 4c, shows the clustergram for significant differentially expressed genes in convalescent populations, and the clusters for these gene sets are even more varied.

Some limitations of this practical is that the data sets had a relatively small sample size of only 56 samples, therefore outliers in the samples can heavily affect the data set and our results. With more samples, we could observe stronger and accurate findings in key genes or better clustering of populations in PCA's or clustergrams.

Another limitation is that the samples only had one readings, the findings of our results could be more strongly justified had there been multiple readings of the samples, for example take three blood tests of a patient and then provide transcriptomics data for all three readings to compare data between different readings.

Further steps for this practical is to build a machine learning model using different algorithms such as support vector machines or deep forest algorithms to observe how well our key genes

are used to create a pattern between infected (dengue fever and dengue haemorrhagic fever) and healthy or convalescent individuals.

Conclusion

While there is no clear contrast between convalescent (recovering individuals from dengue infection) and healthy individuals, there does seem to be a contrast between infected individuals and convalescent or healthy individuals. Exploratory analysis such as PyDESeq2 of dengue transcriptomics data allows us to gather key genes which are significantly differentially expressed in dengue fever and dengue haemorrhagic fever. The significance of these key genes are important as they can be used to build machine learning models for distinction between individuals infected with dengue and non infected individuals. Further downstream pathway analysis of these key genes can also be studied to provide further understanding of the effects of dengue on the human body, by looking at which pathways are affected by these key genes. The key genes can also be targets for drugs for treatment and therapies.

References:

[1] Dengue WHO
https://www.who.int/news-room/fact-sheets/detail/dengue-and-severe-dengue#:~:text=Overview,body%20aches%2C%20nausea%20and%20rash.

[2] Dengue data set
https://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS5093

[3] PCA
https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202#d1e242
https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

[4] PyDESeq2
https://pydeseq2.readthedocs.io/en/latest/auto_examples/plot_minimal_pydeseq2_pipeline.html#sphx-glr-auto-examples-plot-minimal-pydeseq2-pipeline-py
https://github.com/owkin/PyDESeq2#getting-started

[5] GSEAPY
https://gseapy.readthedocs.io/en/latest/gseapy_example.html

[6] Seaborn
https://seaborn.pydata.org/