Exercise 1: Exploratory analysis of transcriptomics data (35%)

Introduction

Dengue virus (DENV) is an arthropod borne single-stranded positive-sense ribonucleic acid virus which infects approximately 50-200 million people annually (Murugesan and Manoharan, 2020). There are four DENV serotypes and infection with these DENVs can cause a range of responses including dengue fever (DF), dengue haemorrhagic fever (DHF) and severe dengue shock syndrome (DSS) (Rothman and Ennis, 1999; Murugesan and Manoharan, 2020).

It is important that the molecular mechanisms determining host response to DENV infection are better understood to help prevent and treat severe cases. It is thought that there are currently approximately 22,000 annual fatalities caused by dengue infections (Byard, 2016), and a greater mechanistic understanding of what differentiates these patients from asymptomatic infections would help identify at-risk individuals and target genes for treatment.

This work investigates the gene expression differences in blood samples from patients with dengue fever (DF), the more severe and life-threatening dengue haemorrhagic fever (DHF), convalescent (COA) patients, and healthy control (HC) individuals. Here the gene expression differences between the patient populations will be visualised in order to identify key genes and pathways underlying the variance in host response to infection. The data being used for this investigation was collected using DNA microarrays for a study exploring whole blood samples of patients with acute dengue virus infections, and convalescence in Bangkok during the 2009 season (Kwissa *et al.*, 2014). Here, the sample data is analysed by using unsupervised clustering, differential gene expression analysis and functional enrichment analysis.

Methods

This analysis of DENV gene expression in human whole blood samples was carried out in the R programming language. There are 31,654 genes in each of the 56 samples in the data set including those with DF (n=18), DHF (n=10), COA (n=19) patients who were released from hospital at least four weeks prior, and HC samples (n=9). The data used for analysis was downloaded from the Gene Expression Omnibus (GEO) database (Barrett *et al.*, 2012) using the Bioconductor GEOquery package.

The imported DENV data was converted into an ExpressionSet object before the gene expression and phenotype data were extracted and tabulated. Exploratory analysis of the gene expression data discovered replicate gene expression measurements for many of the genes. The mean expression level was calculated and used for genes with repeated measurements.

Hierarchical cluster analysis (HCA) was conducted on the gene expression data and whole blood samples. For this analysis, distance between samples and genes was determined using

the Euclidean measure to calculate the absolute magnitudes of the distances between samples or genes, and to highlight any outliers. The complete linkage measure was selected to reduce the impact of the gene expression measurement scale. Cluster analysis was conducted to show the relationship between the blood samples from different disease states using the complete set of gene expression data to produce a dendrogram. Following this, the 100 genes with the highest standard deviation across all samples were selected and used for a second HCA.

Principal component analysis (PCA) by singular value decomposition (SVD) was performed on the data set. The results of this analysis were then used to produce a PCA plot of the first two principal components in order to visualise clustering of the blood samples.

Differential gene expression analysis was implemented to determine the quantity of genes differentially expressed in contrasts between each sample-type. For this analysis a p-value threshold of 0.05 was used to determine whether the differential expression of a gene was significant.

Finally, functional enrichment analysis and more specifically Gene Set Enrichment Analysis (GSEA) was used to pair differentially expressed genes to biological functions. The results of this analysis were then visualised in dot plots showing the gene ontology (GO) categories enriched in the data set.

Results and Discussion

Unsupervised cluster analysis

Cluster analysis was used to analyse the 56 whole blood samples used in this investigation and two primary clusters were observed. It was found that the healthy control (HC) and convalescent (COA) patient samples clustered differently to the dengue fever (DF) and dengue haemorrhagic (DHF) patient samples, as shown in Figure 1. Within the two observed clusters there is no obvious distinction between the sample types. In the cluster containing the HC and COA samples there appeared to be little differentiation between the samples, see Figure 1b and Figure 2. This similarity between HC and COA individuals indicates that blood gene expression levels for those recovering from acute dengue virus infection may return to near normal levels after four weeks. However, as visualised in Figure 1a, a subset of the COA samples appear to cluster separately to all other samples. When hierarchal cluster analysis was performed using only the 100 most differentially expressed genes, figure 2, this distinct COA cluster was not observed. This absence of the COA cluster in the second analysis suggests that the separation of these samples may be due to another variable unrelated to dengue virus infection or in genes less strongly impacted by infection. The small sample size and use of Euclidean distance for the analysis may also have caused the differentiation of the cluster to be more pronounced. The second of the two primary clusters consists of only DF and DHF samples with no apparent distinction between them. As a result, it could be suggested that this difference in response severity to DENV infection may not be a consequence of differential gene expression but other factors instead. Of the 28 acute DENV infection samples, 2 DHF and 3 DF samples were found to cluster with the HC and COA patients (Figure 1). A similar outcome was found in the analysis using the top 100

differentially expressed genes, in which there were 2 DHF and 4 DF samples clustering with the non-disease samples (Figure 2). However, the heatmap in Figure 2 shows that these samples seem to have gene expression values that are dissimilar to the other disease samples and the HC/COA samples. A proposed reason for the unexpected gene expression of these samples is their possible stage of DENV infection or individual differences in rate of recovery. Among the cohort of 28 individuals with acute DENV infection, the estimated day of illness range from day 2 to day 9 (Kwissa *et al.*, 2014) and depending on the rate of recovery for the individual, this means some patients may be transitioning from the gene expression profile of a disease state individual to that of a recovering or healthy individual. There are many other possible explanations for why these DF and DHF samples are clustering with HC and COA patient samples and a simple one may be that it is due to individual differences that are not accounted for in the small sample size.

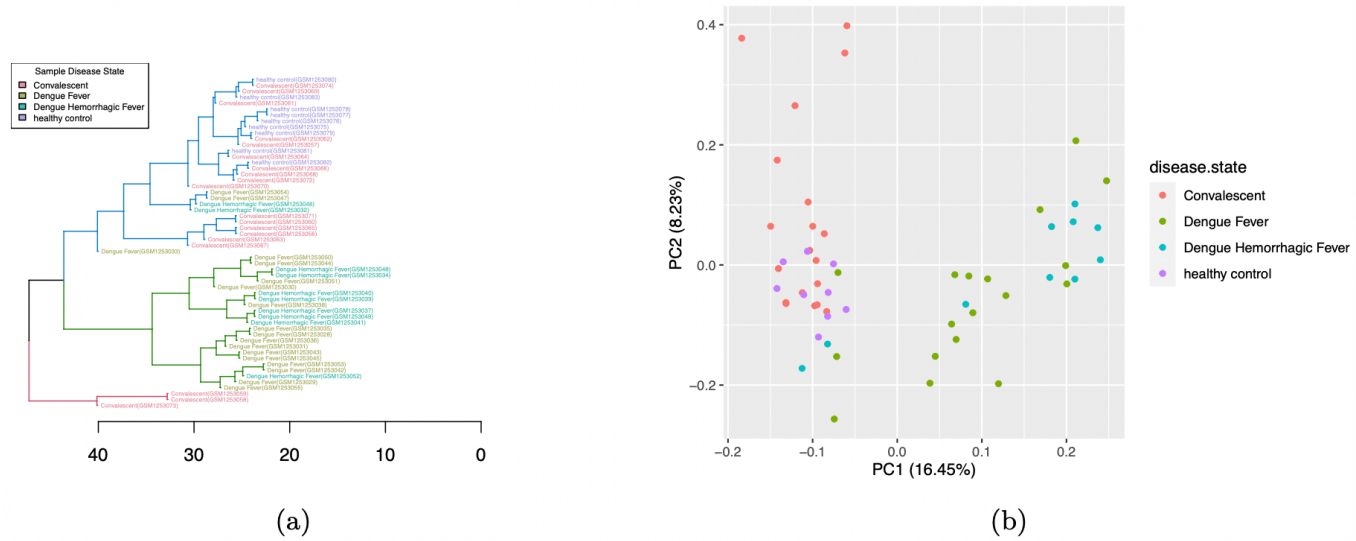

Figure 1. Unsupervised clustering of the entire gene expression profiles exhibited by each disease state. (a) A dendrogram showing the results of hierarchical cluster analysis of the DENV data set. For this analysis Euclidean distance and complete linkage were used. (b) A PCA plot of the first two principal components for each patient blood sample and the resultant clustering. The data used for these plots comes from the gene expression of 31,654 genes of 56 blood samples including healthy (n=9), recovering (n=19) or DENV (n=28) infected individuals. For the R code used to generate these plots see the appendix.

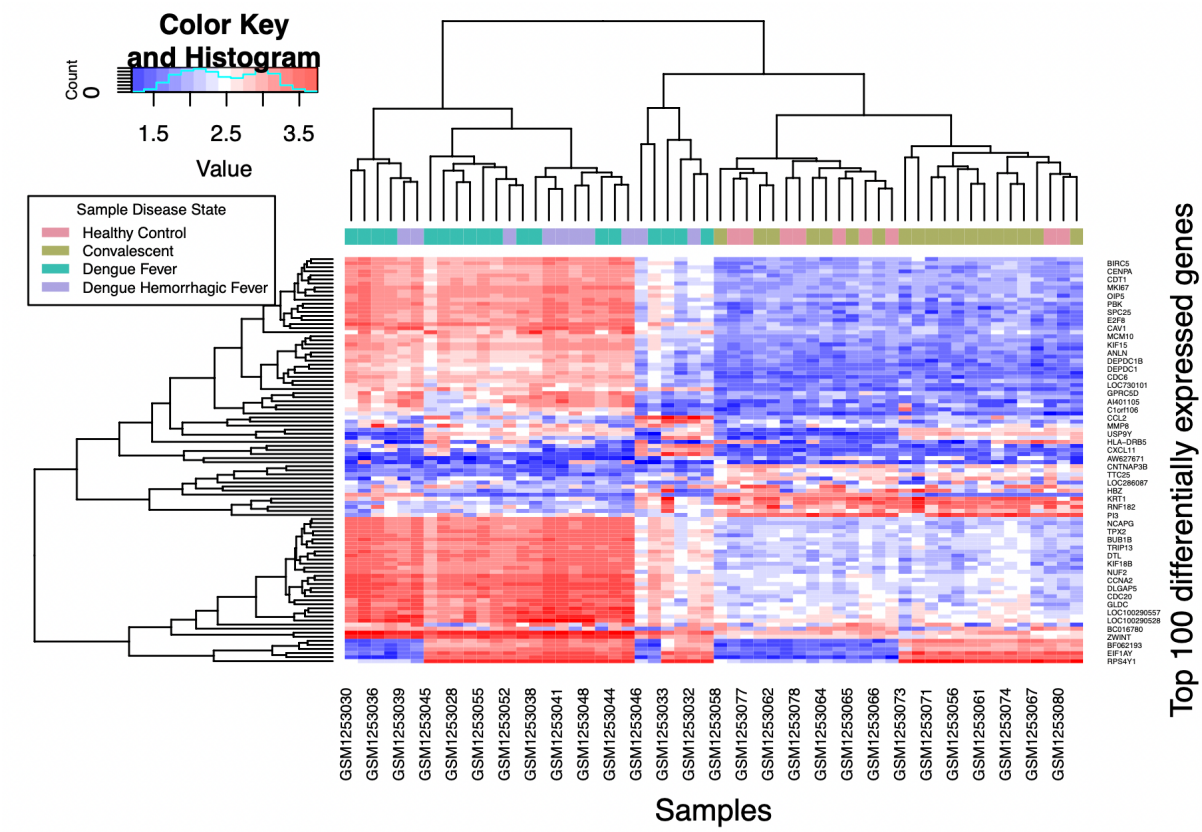

Figure 2. Hierarchical clustering and heatmap for genes with high differential expression levels. Hierarchical cluster analysis was conducted on the top 100 differentially expressed genes across all samples used for this investigation. The 100 genes with the highest standard deviation in expression across all samples were selected to represent the top 100 differentially expressed genes. The heatmap indicates the value of the expression level of each gene in each blood sample. The data used for this plot comes from the gene expression of 31,654 genes of 56 blood samples including healthy (n=9), recovering (n=19) or DENV (n=28) infected individuals. For the R code used to generate this plot see the appendix.

Differential gene expression analysis

The initial differential gene expression analysis looked at the number of significantly differentially expressed genes between all sample types examined in this work. As expected from the cluster analysis, there were large numbers of genes found to be differentially expressed between the acute DENV infected samples and the non-disease samples (Table 1). There were no genes significantly differentially expressed in a comparison of the DF and DHF disease types. This supports the lack of differentiation between the two disease states found in the cluster analysis and suggests differences in disease severity are a product of mechanisms other than transcriptional regulation. A small quantity of genes were differentially expressed between the samples from HC and COA individuals, with only 15 genes being down-regulated and 35 up-regulated in HCs when compared with COA samples (see Table 1). The minimal differences in gene expression between HC and COA samples provides further evidence for almost complete return to typical gene expression levels four weeks after DENV infection. However, when comparing the differential gene expression between COA and DF or DHF samples to HC and DF or DHF samples, there is a large difference in the number of differentially expressed genes. The COA samples have 4419 genes down-regulated when compared to the DF samples whereas the HC samples have 3282 genes down-regulated when compared to DF samples. The apparent disparity in results regarding the similarity of gene expression between COA and HC samples raises further questions about the recovery of gene expression following acute DENV infection. It is possible that in COA individuals, expression of genes impacted by DENV infection remains

on the boundary of typical expression levels observed in HC individuals. These boundary genes may still be correcting or overcorrecting as a consequence of infection, and this overcorrection could be responsible for the significant differences in gene expression levels observed.

Table 1. A table of the total differential gene expression between each sample type. This table shows the number of genes that are differentially expressed between each of the four sample types found in this investigation. Convalescent samples are represented by C, dengue fever by DF, dengue haemorrhagic fever by DHF and healthy control by HC. The p-value threshold for the significant differential expression was $p < 0.05$ and there was no logFC threshold. For the R code used to generate this table see the appendix.

	CvsDF	CvsDHF	HCvsC	DFvsDHF	HCvsDF	HCvsDHF
Down	4419	4311	15	0	3282	3612
NotSig	22929	23124	31604	31654	25194	24429
Up	4306	4219	35	0	3178	3613

Differential gene expression between each sample type was plotted on volcano plots (see Figure 3 and Figure 4) in which a $\log_2$FC threshold of 1 or -1 was set. These plots identify genes that show the highest levels of differential expression which are likely key in the response to DENV infection. The comparison of gene expression between HC samples and DHF samples produced the highest number of genes that exceeded the fold-change threshold and showed significant differential expression. The four genes significantly down-regulated in HC individuals when compared with DHF individuals are: USP30-AS1, NDC80, LOC662494 and VWCE. USP30-AS1 is a long non-coding RNA transcribed from the antisense strand of the USP30 gene (Chen *et al.*, 2021; Howe *et al.*, 2021). USP30-AS1 has been shown to be involved in genetic regulation and regulation of immune system genes (Zhou *et al.*, 2022). Although USP30-AS1 has not previously been implicated in DENV, it has been implicated in Glioblastoma (Wang *et al.*, 2021), acute myeloid leukaemia (Zhou *et al.*, 2022) and cervical cancer (Chen *et al.*, 2021), with increased expression being associated with a worse prognosis in each case. Consequently, it could be hypothesised that the increased severity of DHF when compared with DF may be partially due to the higher levels of USP30-AS1 expression that is shown in this dataset. The NDC80 gene encodes a component of the essential kinetochore-associated NDC80 complex (The UniProt Consortium *et al.*, 2021; Sayers *et al.*, 2022) where it interacts with other components of the complex to aid in chromosome congression and attachment of kinetochores to spindle microtubules. Overexpression of NDC80, as found here in DHF, has previously been associated with poor prognoses in pancreatic cancer and osteosarcoma, with proposed roles in DNA damage and proliferation (Meng *et al.*, 2015; Xu *et al.*, 2017). The VWCE gene, also known as the von Willebrand factor C and EGF domain-containing protein, encodes a protein involved in enabling calcium ion binding and has also been shown to be involved in the cellular response to viral infection (Barrett *et al.*, 2012; The UniProt Consortium *et al.*, 2021). Consequently, it is likely overexpressed in DHF individuals as part of the cellular response to the viral infection.

The genes overexpressed in DF individuals when compared with HC individuals are BRCA2 and NCAPG, as shown in Figure 3b. BRCA2 is a gene involved in double-strand break repair and homologous recombination that has previously been identified as a gene up-regulated in dengue fever (Loke *et al.*, 2010; The UniProt Consortium *et al.*, 2021). NCAPG is a gene that encodes for the condensin complex subunit 3 protein which is involved in the conversion of interphase chromatin into mitotic-like condense chromosomes (Xiao *et al.*, 2020). NCAPG is

another protein for which overexpression is implicated in cancer and cell proliferation (Zhang *et al.*, 2020). Many of the genes overexpressed in DF and DHF when compared to HC samples appear to be related to DNA repair or key DNA-related mechanisms, suggesting possible biomarkers or treatment targets for the disease.

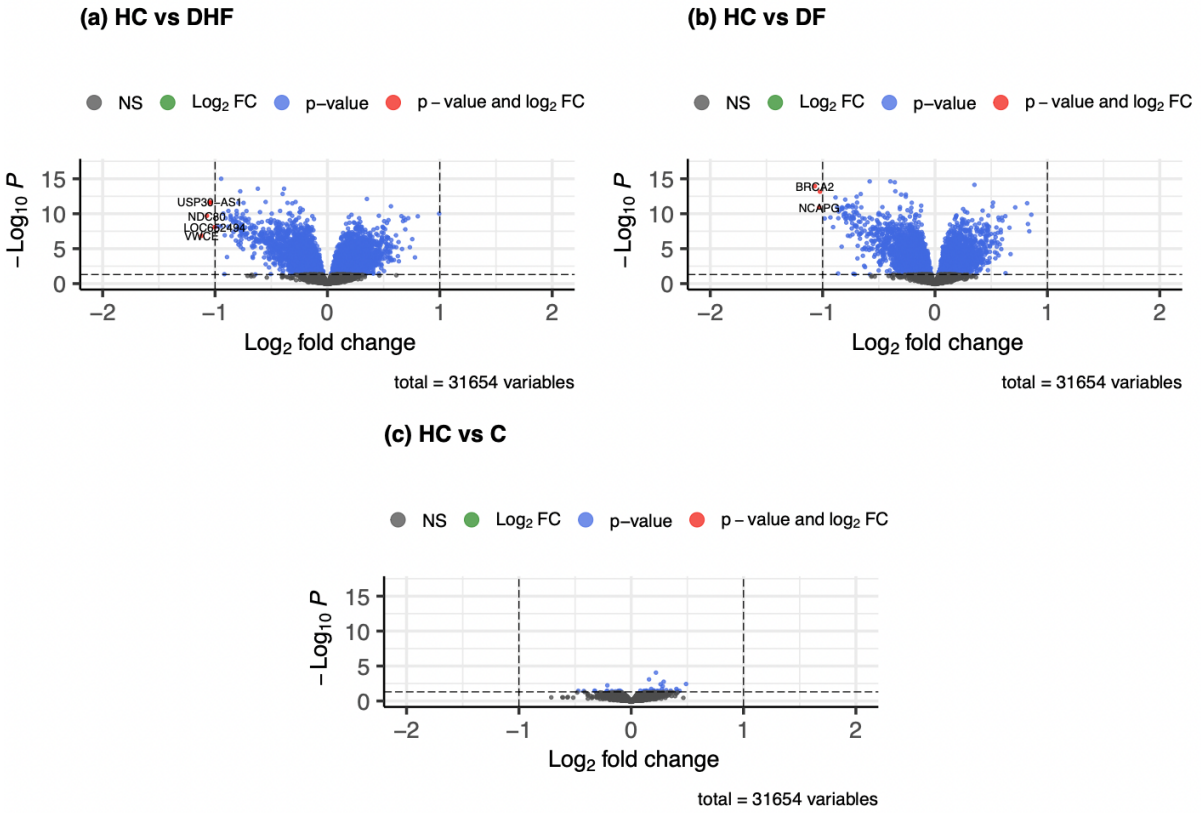

Figure 3. Differential gene expression of healthy control samples when compared with DENV samples. All volcano plots shown in this figure have a p-value threshold of $p < 0.05$ and a $\log_2$FC threshold of 1 or -1. The data used for these plots comes from the gene expression of 31,654 genes in each of 56 blood samples. (a) The gene expression of healthy control samples when compared with samples from individuals with dengue haemorrhagic fever. (b) The gene expression of healthy control samples when compared with samples from individuals with dengue fever. (c) The gene expression of healthy control samples when compared with samples from individuals recovering from DENV infection. For the R code used to generate these plots see the appendix.

The differential gene expression analysis between the COA samples and the DF/ DHF samples produced fewer genes exceeding the $\log_2$FC threshold than for HC samples. The most differentially expressed genes in the comparison of COA samples and DHF samples were CDC6 and PBK. The CDC6 gene encodes the cell division control protein 6 which is involved in the initiation of DNA replication and DNA replication checkpoint control (Lim and Townsend, 2020; The UniProt Consortium *et al.*, 2021). The PBK gene encodes the lymphokine-activated killer T-cell originated protein kinase which is involved in the MAP kinase phosphorylation, the activation of lymphoid cells and cell cycle regulation (Huang *et al.*, 2021; The UniProt Consortium *et al.*, 2021). There is only one gene which exceeds the $\log_2$FC threshold in the comparison of COA and DF gene expression profiles, and that gene is FAM72A. FAM72A is a strong indicator of viral infection due to the role it plays in the somatic hypermutation and class-switch recombination of B cell receptors during immune responses. Consequently, it could be asserted that the increased expression of FAM72A in

DF samples when compared with the COA samples may be a consequence of the immune response to infection rather than a product of the virus itself.

The comparison of DF and DHF samples shown in Figure 4c visualises the lack of differential gene expression found between the two disease states, providing further evidence for differences in DENV infection severity being caused by mechanisms other than different levels of gene expression. The comparison of gene expression between HC and COA samples shown in the volcano plot in Figure 3c shows that none of the small number of significantly differentially expressed genes exceed the log_2FC threshold of 1 or -1. However, some of these differentially expressed genes reach log_2FC values of 0.5, differences which could still be highly impactful in producing differing phenotypes between the two sample types.

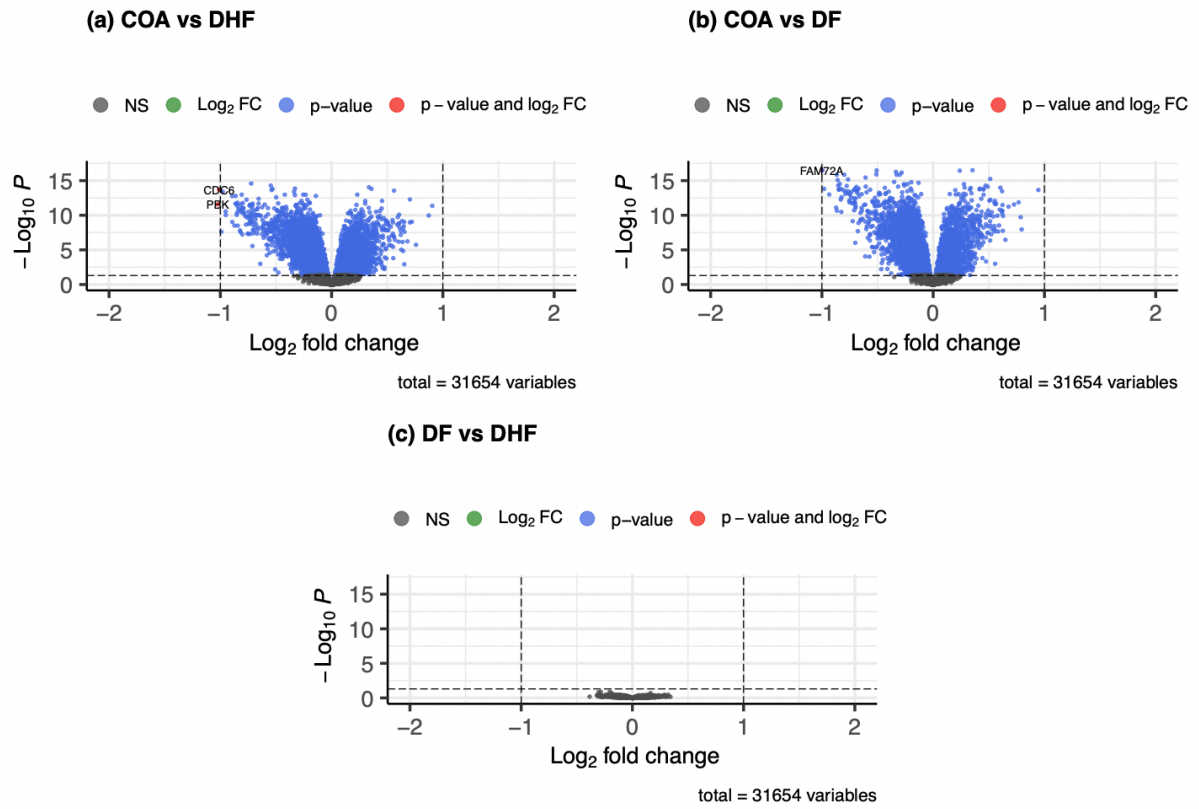

Figure 4. Differential gene expression between DENV samples. All volcano plots shown in this figure have a p-value threshold of $p < 0.05$ and a log_2FC threshold of 1 or -1. The data used for these plots comes from the gene expression of 31,654 genes in each of 56 blood samples. (a) The gene expression of convalescent samples when compared with samples from individuals with dengue haemorrhagic fever. (b) The gene expression of convalescent samples when compared with samples from individuals with dengue fever. (c) The gene expression of dengue fever samples when compared with samples from individuals with dengue haemorrhagic fever. For the R code used to generate these plots see the appendix.

Functional enrichment analysis

Gene set enrichment analysis (GSEA) was conducted in an attempt to elucidate categories of genes related to biological function that were repressed or activated at high frequencies in the gene expression comparisons. This analysis looked at the gene ontology (GO) categories that were enriched in each of the gene expression comparisons with high numbers of significantly differentially expressed genes. Enrichment patterns in these broader categories related to biological function can be used to indicate mechanistic properties or features of DENV infection and possibly infer causes of differences in disease severity. It was found that there

are 12 GO categories which are enriched in each comparison shown in Figure 5. This common gene expression pattern suggests that there is a general consistent gene expression response to DENV infection in all conditions. It is likely that the gene expression changes in the 12 common GO categories provide the basis for the infection responses to the virus with differences in disease symptoms/severity resulting from other distinct biological processes less consistently impacted. The GO categories enriched across all comparisons suggest that up-regulation of genes related to chromosomal and cell-cycle regulation is core to the gene expression response to DENV infection.

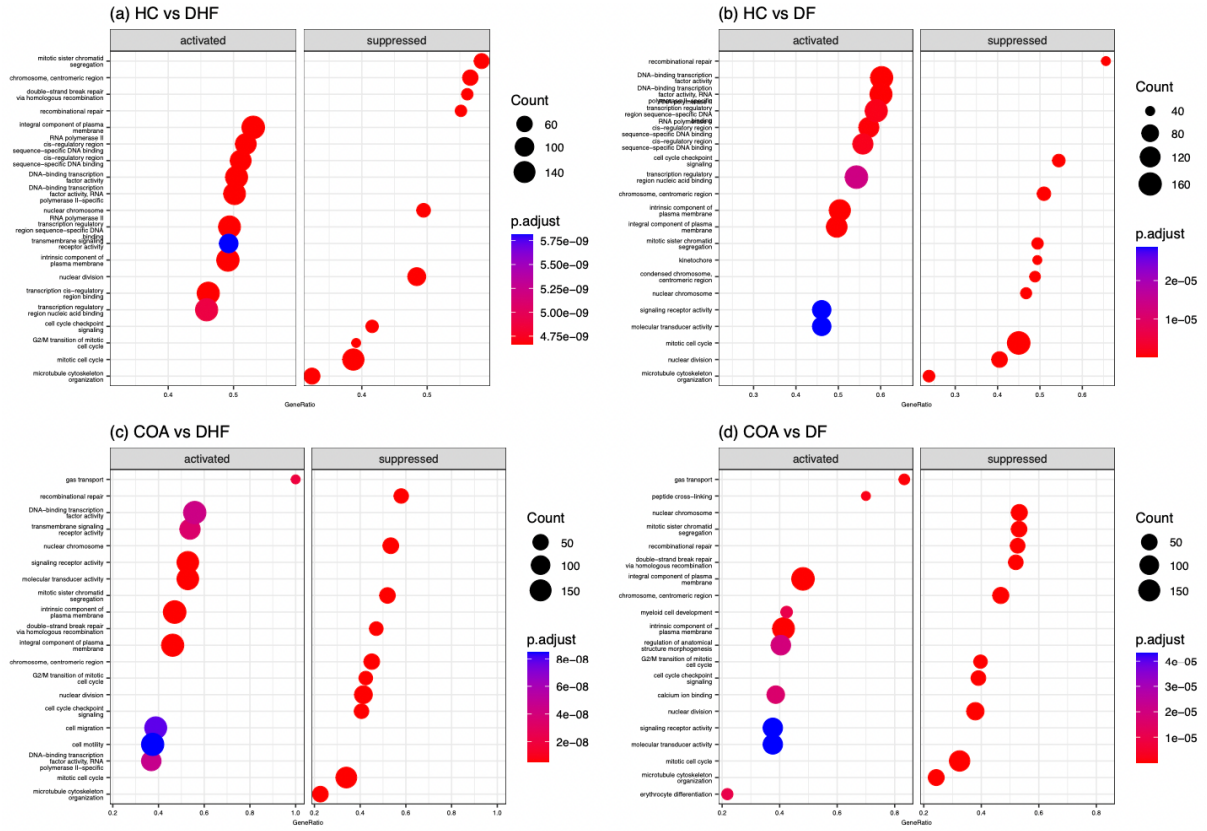

Figure 5. Gene set enrichment analysis of all sample type contrasts with differential gene expression. Each of the dot plots show the gene ontology (GO) categories enriched, a comparison of two sample types, the number of genes enriched and the gene ratio (count/ category size). (a) The GO categories enriched in a comparison of healthy control and dengue haemorrhagic fever samples. (b) The GO categories enriched in a comparison of healthy control and dengue fever samples. (c) The GO categories enriched in a comparison of convalescent and dengue haemorrhagic fever samples. (d) The GO categories enriched in a comparison of convalescent and dengue fever samples. For the R code used to generate these plots see the appendix.

The mechanisms by which a DENV infection develops into DHF are still poorly understood, but it is known that the DHF phenotype typically presents a higher viral load than DF and higher levels of circulating antigens have been found (Ubol *et al.*, 2008). Moreover, it has been hypothesised that dysfunction of immune cells and the complement system in DHF patients increase the viral load and tissue damage when compared to DF patients (Ubol *et al.*, 2008). The present study did not find evidence of major differences in these gene categories when comparing the DF and DHF enrichment analysis for both the COA and HC comparisons. This could be due to the low sample size and lack of differential expression between the two disease phenotypes. In both the COA and HC comparisons there were 5 enriched GO categories that were different between the DF and DHF phenotypes. This does indicate gene expression differences between the two disease phenotypes, but no general

patterns emerged that would implicate categories integral to the mechanisms of the more severe DHF phenotype. The lack of distinction between the GO categories of DF and DHF individuals found here, further supports the hypothesis that many of the phenotypic differences observed between DF and DHF individuals is a consequence of individual differences, particularly in the immune system, between patients.

The HC and COA contrasts with DHF appear to have some gene expression that appears inconsistent with each other. The gene expression profile of DHF samples, when compared with HC samples, shows up-regulation of genes in the 'Double strand break repair via homologous recombination' and 'Recombinational repair' GO categories. In contrast, the gene expression profile DHF samples, when compared with COA samples, shows down-regulation of genes in the 'DNA-binding double strand break transcription factor activity' GO category. A possible explanation for these results is that DENV infection in DHF individuals causes repression of a subset of genes involved in double-strand break binding which has the consequence of increased transcription of other DNA break genes in order to compensate for the repression. However, if this hypothesis was in fact true, all three enriched categories would be expected to be observed in both DHF GSEA. Instead, it is likely that the enrichment of these different GO categories is a result of the limited differential gene expression between COA and HC individuals or individual differences between patients in the study.

Conclusion

The results of this work indicate that the difference in phenotype observed in individuals with dengue fever and dengue haemorrhagic fever are primarily a result of factors other than levels of gene expression as this study found no significant differential gene expression between the two disease states. Furthermore, a large amount of significant gene expression was found between both healthy control / convalescent samples and disease states. Moreover, a set of gene ontology categories were consistently found to be enriched in all contrasts, suggesting a characteristic gene expression pattern of DENV infection. Finally, this work suggests that individuals are close to full recovery four weeks after DENV infection, with a very minimal amount of differential gene expression when compared to healthy control individuals.

Appendix

This section contains the R code used to generate each of the figures and tables in this report.

Figure 1a:

```
# Calculate distance matrix
dists <- dist(t(nr_gene_ex_df), method = 'euclidean')

# HCA
HC <- hclust(dists, method = 'complete', members = NULL)

# Disease type
Disease_type <- rev(levels(pheno[,3]))

# Disease labels
```

```
disease_labels <- pheno$disease.state

# Colours for disease types
species_col <- rev(rainbow_hcl(4))[as.numeric(disease_labels)]

# Create dendrogram object (dendextend package)
dend <- as.dendrogram(HC)

# Colour branches based on clusters - 3 main clusters
dend <- colour_branches(dend, k=3)

# Match the labels to the real classification
labels_colors(dend) <-
   rainbow_hcl(4)[sort_levels_values(
      as.numeric(pheno$disease.state)[order.dendrogram(dend)]
   )]

# Add the disease state to the labels
labels(dend) <- paste(as.character(pheno$disease.state)[order.dendrogram(dend)],
                           "(",labels(dend),")",
                           sep = "")

# We hang the dendrogram a bit:
dend <- hang.dendrogram(dend,hang_height=0.1)

# reduce the size of the labels:
dend <- set(dend, "labels_cex", 0.3)

#Plot
par(mar = c(3,3,3,7))
plot(dend,
     horiz =  TRUE,
     nodePar = list(cex = .007))

legend(x='topleft',
       title = 'Sample Disease State',
       legend = c('Convalescent','Dengue Fever', 'Dengue Hemorrhagic Fever', 'healthy control'),
       fill =rainbow_hcl(4),
       cex =0.5)
```

Figure 1b:

```
## PCA ##
# Data set where the repeats are averaged
nrgex_pca <- prcomp(t(nr_gene_ex_df), scale = TRUE)

# Summary object showing the proportion of variance
s <- summary(nrgex_pca)

## Plot PCA ##
# Plot the first two principle components for each disease state
autoplot(nrgex_pca, data = pheno, colour = 'disease.state')
```

Figure 2:

```
## Select the top 100 genes by standard deviation ##
# Find average sd for each gene
genesd <- apply(nr_gene_ex_df, 1, sd)

# Sort by sd (decreasing)
genesd <- sort(genesd, decreasing = TRUE)

# Select the top100 highest sd rows
tp100sd <- as.data.frame(genesd[1:100])

# Select the top 100 sd gene names
tp100sd_names <- rownames(tp100sd)

# Select/subest sample data for top 100 genes
df100 <- nr_gene_ex_df[tp100sd_names,]

## Produce Heatmap ##
# df100 = 100 most differentially expressed genes across all samples
```

```
heatmap.2(x=as.matrix(df100),
          trace = 'none',
          col = 'bluered',
          xlab = 'Samples',
          distfun = function(x) dist(x, method="euclidean"),
          hclustfun = function(x) hclust(x, method="complete"),
          ylab = 'Top 100 differentially expressed genes',
          ColSideColors = c(
            rep('#ABB065',19), # Conv
            rep('#ACA4E2',10), # DHF
            rep('#39BEB1',18), # DF
            rep('#E495A5',9)), # HC
          cexRow = 0.35,
          cexCol = 0.75,
          key = TRUE,
          margins = c(7,5))

# square line ends for the color legend
par(lend = 1)
#Turn off legend clipping margins
par(xpd=TRUE)
# Plot legend
legend(x=-0.15,
       y=0.93,
       legend = c("Healthy Control", "Convalescent","Dengue Fever",
                  "Dengue Hemorrhagic Fever"), # category labels
       col = c("#E495A5", "#ABB065", "#39BEB1", "#ACA4E2"),  # color key
       lty= 1,
       lwd = 5,
       cex = 0.5,
       title = 'Sample Disease State')
```

Table 1:

```
# Design matrix
design <- model.matrix(~0+pheno$disease.state)

#Assign column names
colnames(design) <- c('Convalescent',
                      'Dengue_Fever',
                      'Dengue_Hemorrhagic_fever',
                      'Healthy_Control')
#Contrast matrix
cont_matrix <- makeContrasts(CvsDF = Convalescent-Dengue_Fever,
                             CvsDHF = Convalescent-Dengue_Hemorrhagic_fever,
                             HCvsC = Healthy_Control-Convalescent,
                             DFvsDHF = Dengue_Fever-Dengue_Hemorrhagic_fever,
                             HCvsDF = Healthy_Control-Dengue_Fever,
                             HCvsDHF = Healthy_Control-Dengue_Hemorrhagic_fever,
                             levels=design)

# Fit expression matrix to a linear model
fit <- lmFit(nr_gene_ex_df, design)

# Compute contrast
fit_contrast <- contrasts.fit(fit, cont_matrix)

# Bayes statistics of differential expression
fit_contrast <- eBayes(fit_contrast)

# Summary of results (number of differentially expressed genes)
results <- decideTests(fit_contrast)
dfde <- summary(results)

kable(dfde, digits=1,
      align = 'l',
      caption = ' A table showing the total differential gene expression between each disease state.')
```

Figure 3/Figure 4:

```
# Generate a list of top 100 differentially expressed genes
top_genes <- topTable(fit_contrast, number = 100, adjust = "BH")
```

```
# Contrast data (each different) - used for volcano plot & functional enrichement
CvsDF <- topTable(fit_contrast, coef = 1, number = 100000, adjust = 'BH')
CvsDHF <- topTable(fit_contrast, coef = 2, number = 100000, adjust = 'BH')
HCvsC <- topTable(fit_contrast, coef = 3, number = 100000, adjust = 'BH')
DFvsDHF <- topTable(fit_contrast, coef = 4, number = 100000, adjust = 'BH')
HCvsDF <- topTable(fit_contrast, coef = 5, number = 100000, adjust = 'BH')
HCvsDHF <- topTable(fit_contrast, coef = 6, number = 100000, adjust = 'BH')

## Function for plotting volcano plots ##

volcano_func <- function(contrast_data, title){
  EnhancedVolcano(contrast_data,
                  lab = rownames(CvsDHF),
                  labSize = 3,
                  ylim = c(0,17),
                  xlim = c(-2,2),
                  x = 'logFC',
                  y = 'adj.P.Val',
                  pCutoff = 0.05,
                  FCcutoff = 1,
                  title = title,
                  subtitle = '',
                  pointSize = 1.0,
                  colAlpha = 0.75)

}
```

Figure 5:

```
## Function that carries out GSEA and plots dotplots of the results ##
GSEA_func <- function(contrast_data, title) {
  # Get the entrezIds for our genes
  entrezIds_Org = as.data.frame(mapIds(org.Hs.eg.db,
                                       keys = rownames(contrast_data),
                                       keytype = "SYMBOL",
                                       column = "ENTREZID",
                                       multiVals = "first"))
  # Set the column name for gene enetrez IDs
  colnames(entrezIds_Org) <- 'ENTREZID'
  # Add entrez ID column to new fit_contrast toptable
  contrast_data <- cbind(contrast_data, ENTREZID = entrezIds_Org$ENTREZID)
  # Select only significant genes - adj.p.val
  contrast_data <- contrast_data[contrast_data[,'adj.P.Val'] < 0.05,]
  # Remove NAs
  contrast_data <- na.omit(contrast_data)
  # Select the logFC
  logfc_list <- contrast_data$logFC
  #Name these rows with ENTREZID
  names(logfc_list) <- contrast_data$ENTREZID
  # sort the list in decreasing order (required for clusterProfiler)
  logfc_list = sort(logfc_list, decreasing = TRUE)
  # Perform GSEA
  GSEA <- gseGO(geneList = logfc_list,
                keyType = "ENTREZID",
                ont = 'ALL',
                OrgDb = 'org.Hs.eg.db')
  # Produce dotplot of results
  require(DOSE)
  dotplot(GSEA, showCategory=10,
          split=".sign",
          font.size = 4.5,
          title = title) + facet_grid(.~.sign)
}
```

Bibliography

Barrett, T. *et al.* (2012) 'NCBI GEO: archive for functional genomics data sets—update', *Nucleic Acids Research*, 41(D1), pp. D991–D995. doi:10.1093/nar/gks1193.

Byard, R.W. (2016) 'Lethal Dengue Virus Infection: A Forensic Overview', *The American Journal of Forensic Medicine and Pathology*, 37(2), pp. 74–78. doi:10.1097/PAF.0000000000000236.

Chen, M. *et al.* (2021) 'Long non-coding RNA USP30-AS1 aggravates the malignant progression of cervical cancer by sequestering microRNA-299-3p and thereby overexpressing PTP4A1', *Oncology Letters*, 22(1), pp. 1–12. doi:10.3892/ol.2021.12766.

Howe, K.L. *et al.* (2021) 'Ensembl 2021', *Nucleic Acids Research*, 49(D1), pp. D884–D891. doi:10.1093/nar/gkaa942.

Huang, H. *et al.* (2021) 'PBK/TOPK: An Effective Drug Target with Diverse Therapeutic Potential', *Cancers*, 13(9), p. 2232. doi:10.3390/cancers13092232.

Kwissa, M. *et al.* (2014) 'Dengue Virus Infection Induces Expansion of a CD14+CD16+ Monocyte Population that Stimulates Plasmablast Differentiation', *Cell host & microbe*, 16(1), pp. 115–127. doi:10.1016/j.chom.2014.06.001.

Lim, N. and Townsend, P.A. (2020) 'Cdc6 as a novel target in cancer: Oncogenic potential, senescence and subcellular localisation', *International Journal of Cancer*, 147(6), pp. 1528–1534. doi:10.1002/ijc.32900.

Loke, P. *et al.* (2010) 'Gene Expression Patterns of Dengue Virus-Infected Children from Nicaragua Reveal a Distinct Signature of Increased Metabolism', *PLoS Neglected Tropical Diseases*, 4(6), p. e710. doi:10.1371/journal.pntd.0000710.

Meng, Q.-C. *et al.* (2015) 'Overexpression of NDC80 is correlated with prognosis of pancreatic cancer and regulates cell proliferation', *American Journal of Cancer Research*, 5(5), pp. 1730–1740.

Murugesan, A. and Manoharan, M. (2020) 'Dengue Virus', *Emerging and Reemerging Viral Pathogens*, pp. 281–359. doi:10.1016/B978-0-12-819400-3.00016-8.

Rothman, A.L. and Ennis, F.A. (1999) 'Immunopathogenesis of Dengue Hemorrhagic Fever', *Virology*, 257(1), pp. 1–6. doi:10.1006/viro.1999.9656.

Sayers, E.W. *et al.* (2022) 'Database resources of the national center for biotechnology information', *Nucleic Acids Research*, 50(D1), pp. D20–D26. doi:10.1093/nar/gkab1112.

The UniProt Consortium *et al.* (2021) 'UniProt: the universal protein knowledgebase in 2021', *Nucleic Acids Research*, 49(D1), pp. D480–D489. doi:10.1093/nar/gkaa1100.

Ubol, S. *et al.* (2008) 'Differences in Global Gene Expression in Peripheral Blood Mononuclear Cells Indicate a Significant Role of the Innate Responses in Progression of Dengue Fever but Not Dengue Hemorrhagic Fever', *The Journal of Infectious Diseases*, 197(10), pp. 1459–1467. doi:10.1086/587699.

Wang, N. *et al.* (2021) 'USP30-AS1 contributes to mitochondrial quality control in glioblastoma cells', *Biochemical and Biophysical Research Communications*, 581, pp. 31–37. doi:10.1016/j.bbrc.2021.10.006.

Xiao, C. *et al.* (2020) 'NCAPG Is a Promising Therapeutic Target Across Different Tumor Types', *Frontiers in Pharmacology*, 11, p. 387. doi:10.3389/fphar.2020.00387.

Xu, B. *et al.* (2017) 'Elevated NDC80 expression is associated with poor prognosis in osteosarcoma patients', *European review for medical and pharmacological sciences*, 21, pp. 2045–2053.

Zhang, X. *et al.* (2020) 'NCAPG Induces Cell Proliferation in Cardia Adenocarcinoma via PI3K/AKT Signaling Pathway', *OncoTargets and Therapy*, 13, pp. 11315–11326. doi:10.2147/OTT.S276868.

Zhou, W. *et al.* (2022) 'LncRNA USP30-AS1 promotes the survival of acute myeloid leukemia cells by cis-regulating USP30 and ANKRD13A', *Human Cell*, 35(1), pp. 360–378. doi:10.1007/s13577-021-00636-7.

Exercise 2: Machine learning for disease diagnosis (40%)

Introduction

Crohn's disease (CD) is a chronic inflammatory disease of the gastrointestinal tract with a steadily increasing prevalence worldwide (Torres *et al.*, 2017). Presenting symptoms can vary between individuals, with many of the symptoms being symptoms of other diseases. Consequently, CD diagnosis is typically made with endoscopic and/or radiologic findings (Feuerstein and Cheifetz, 2017). This work aims to use machine learning models to diagnose CD from metabolomic data samples, an approach that could make disease diagnosis significantly more efficient and less expensive. The data used here is metabolomic data acquired with gas chromatography mass spectrometry (GC-MS) from healthy control individuals (HC) or individuals diagnosed with (CD).

Methods

The entire analysis of the metabolic GC-MS data outlined in this report was analysed in the R programming language. The GC-MS was pre-processed to extract relevant data before carrying out dimensionality reduction to make the data more manageable for the machine learning algorithms. Machine learning models were tested on four types of patient samples: faecal, blood, urine, and breath. The following methodology was repeated for each of the sample types so that they could be compared to determine the most useful sample type for CD disease prediction.

Dimensionality reduction was carried out with a centred and scaled principal component analysis (PCA) as this would still capture the variance between samples with many fewer variables. Following PCA, the number of principal components that would be retained for the machine learning models was determined. This was done by considering the proportion of variance contained within each component as shown in Figure 1.

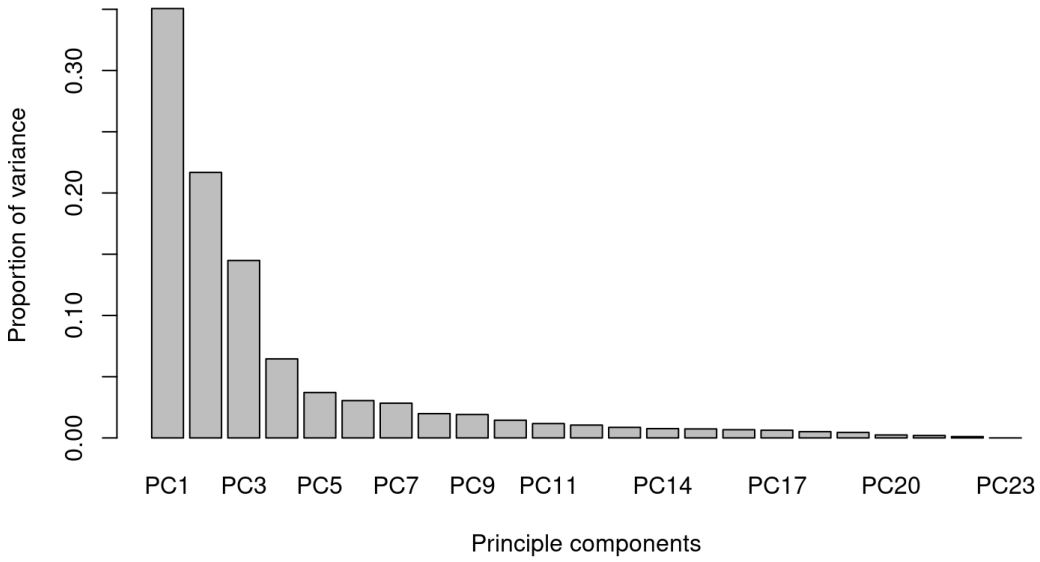

Figure 1. Distribution of the variance between CD and HC samples accounted for by each principal component. This graph shows the proportion of variance that can be explained by each of the principal components produced by a PCA.

The number of principle components were selected to retain as much variance as possible while reducing the amount of noise in the data. For example, the first four principal components which accounted for 77.68% of the total variance (Figure 1) were retained for the faecal data. The retained principal components were then used to build and test the machine learning models. The Classyfire R package (Chatzimichali and Bessant, 2016) was used to train ensembles of support vector machines (SVMs). Bootstrapping was repeated 100 times to avoid reliance on any single bootstrapping split and the number of ensembles used for each sample type was dependent upon the added accuracy that using more ensembles provided. An initial SMV model was trained, in which 100 ensembles were used and the results of this testing determined the number of ensembles used for the final model. Figure 2 shows the plot used to visualise the changes in accuracy of the faecal SVM model as the number of ensembles increased, the number of ensembles used corresponded to a plateau in average test accuracy or when adding more ensembles no longer increased model accuracy. For the faecal sample SVM model 49 ensembles were specified.

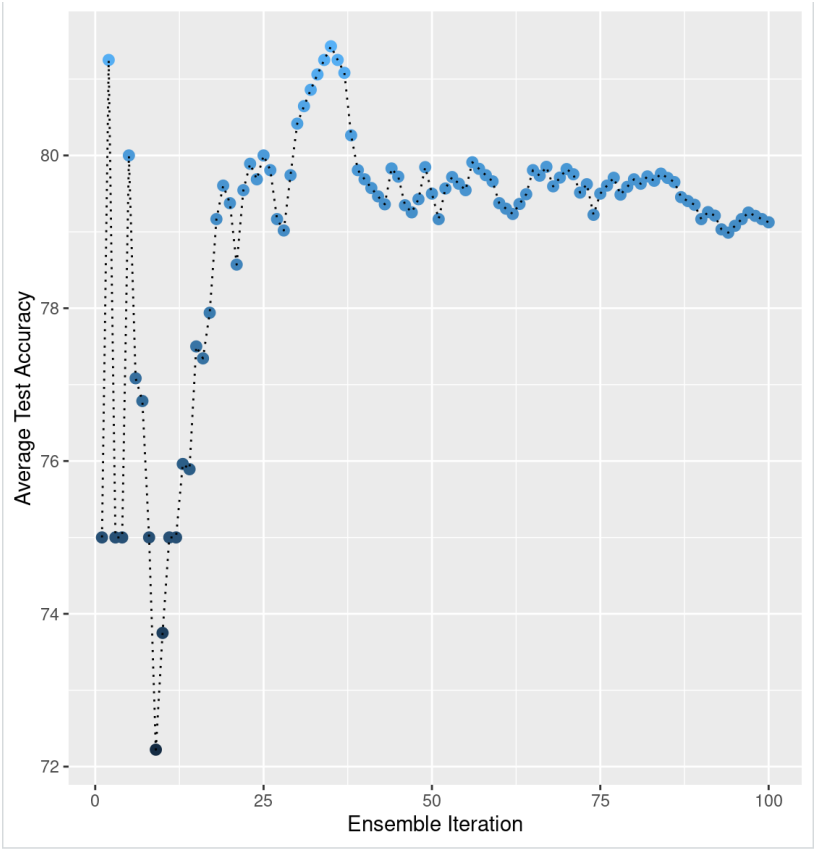

Figure 2. Effect of additional ensembles on the classification effectiveness of a faecal sample support vector machine (SVM). A plot used to determine the optimum number of ensembles for a SVM by showing how the average test accuracy varies with the number of ensembles.

Once the model parameters were determined, the final model for the sample type in question was constructed. The above process was repeated for each of the faecal, blood, breath, and urine sample types. The success of the SVM models in predicting CD was assessed with mean test accuracy scores.

The SVM model that was most successful in classifying CD patients was selected for permutation testing to further assess the accuracy of the model. The permutation testing was

repeated 100 times to ensure that a reliable null distribution was produced that could be compared with the SVM model.

Finally, a random forest machine learning algorithm was constructed and tested to compare with the most successful SVM model. The Caret R package (Kuhn, 2008) was used to build this random forest model and the cross-validation method of re-sampling was selected. The random forest model selected was based on the accuracy achieved with different numbers of predictors.

Results and Discussion

The four SVM models trained on the different sample types each had differing levels of effectiveness in diagnosing CD, as shown in Figure 3. The faecal sample type was shown to be the most effective in differentiating HC patients from those with CD, with an average accuracy of 79.85%. Urine samples were found to be the next most useful for CD diagnosis and were closely followed by breath samples, with average accuracies of 61% and 58.27% respectively. The least useful sample type for CD diagnosis by SVMs were the blood samples with an average accuracy of 53.61%. These results suggest that faecal matter contains more metabolites can be used to distinguish between HC and CD individuals that the other three sample types examined.

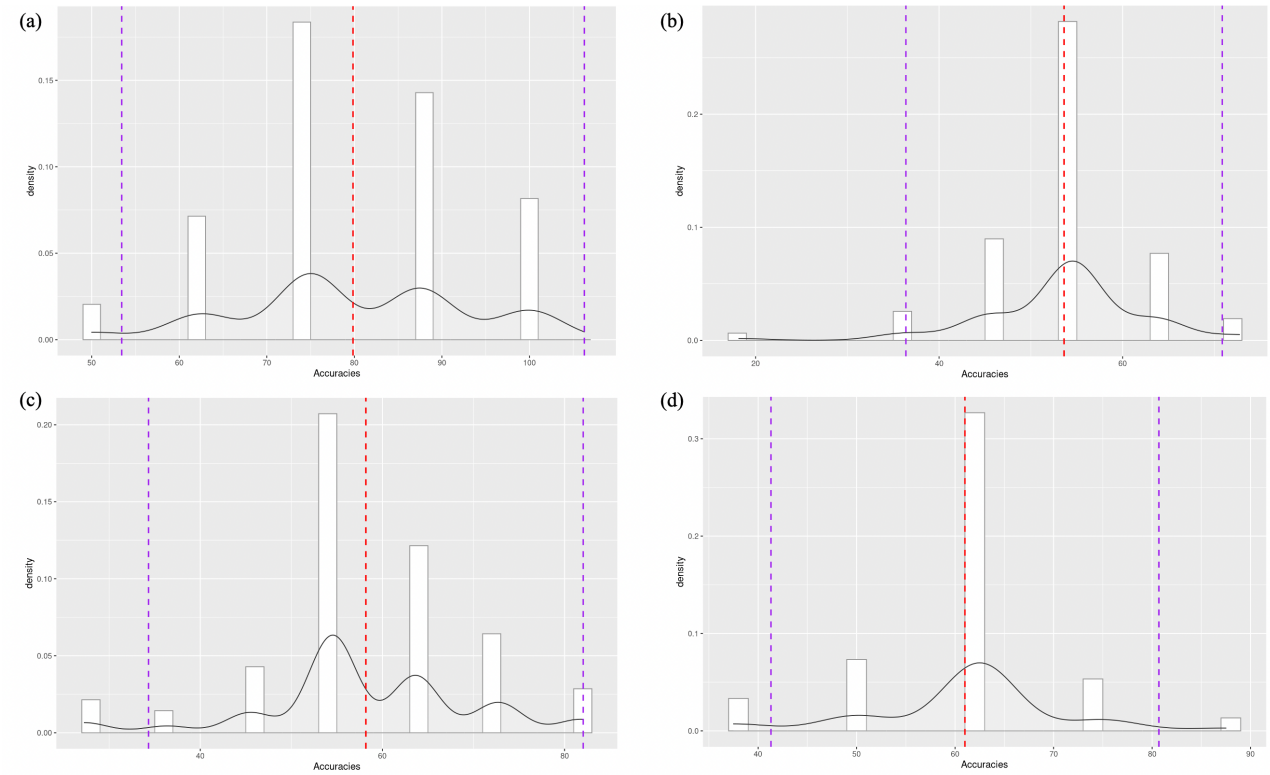

Figure 3. Classification accuracy of SVM models for CD across different sample types. For each of the plots in this figure, the purple lines represent the 95% confidence intervals and the red lines the mean density. (a) Test accuracy results for the faecal SVM. (b) Test accuracy results for the blood SVM. (c) Test accuracy results for the breath SVM. (d) Test accuracy results for the urine SVM.

Permutation testing was conducted to determine whether the accuracy observed in the faecal SVM was sufficiently higher than what you would observe by chance. A comparison between the 100 permutations and 49 ensembles indicated that the faecal SVM is sufficiently different

to be considered somewhat effective (Figure 4). The mean accuracy of the permutations was 50.90% compared with 79.85% for the faecal SVM. Moreover, the overall accuracies of the ensembles are sufficiently above the 95% confidence interval of the permutation distribution (Figure 4a). Consequently, we can have confidence that the predicted results of the faecal SVM are meaningful.

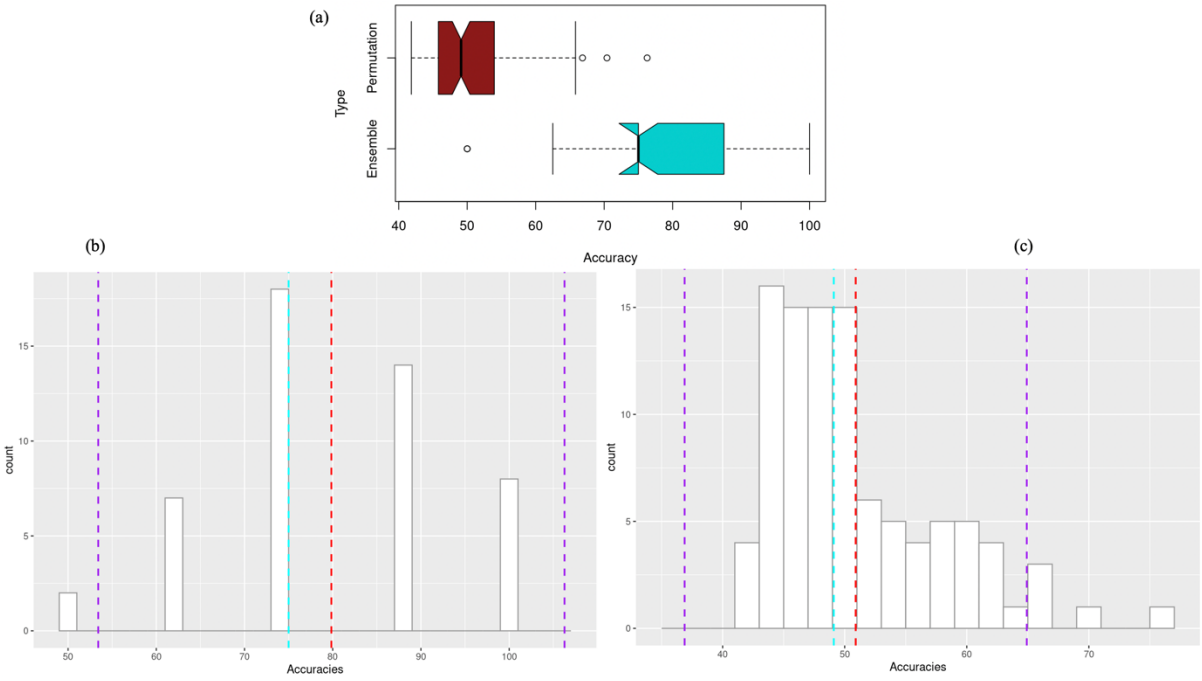

Figure 4. Differences in classification accuracy between permutations and faecal SVM ensembles. The red lines show the mean accuracy, blue lines the median accuracy, and purple lines the 95% confidence intervals. (a) A boxplot comparing the distributions of test accuracy between permutations and ensembles. (b) The accuracy distribution of the 49 ensembles from the faecal SVM model. (c) The accuracy distribution of the 100 permutations in the permutation testing.

Among the faecal samples there were differences in the classification accuracy of the SVM model. Figure 5b depicts the relatively small difference in classification accuracy between faecal CD and HC samples, with mean accuracies of 84% and 76% respectively. This result suggests that the constructed faecal SVM model is more effective at classifying CD samples than HC samples. However, more pronounced differences in classification accuracy can be found between the individual faecal samples themselves. This is visualised in Figure 5a where it is obvious that the SVM has difficulties classifying specific samples. One such sample is W1072_FA_CD which was incorrectly classified 20 times in the 49 ensembles. This would indicate that this sample closely resembles a HC sample and may be close to recovery from the disease, or that the data used to train the SVM may not be sufficient to classify certain types of CD samples. Another faecal sample that was mis-classified at a high rate was W303_FA_CD which was mistaken for a HC sample 16 times in the 49 ensembles, further suggesting that there may be a CD metabolic profile the SVM struggles to handle.

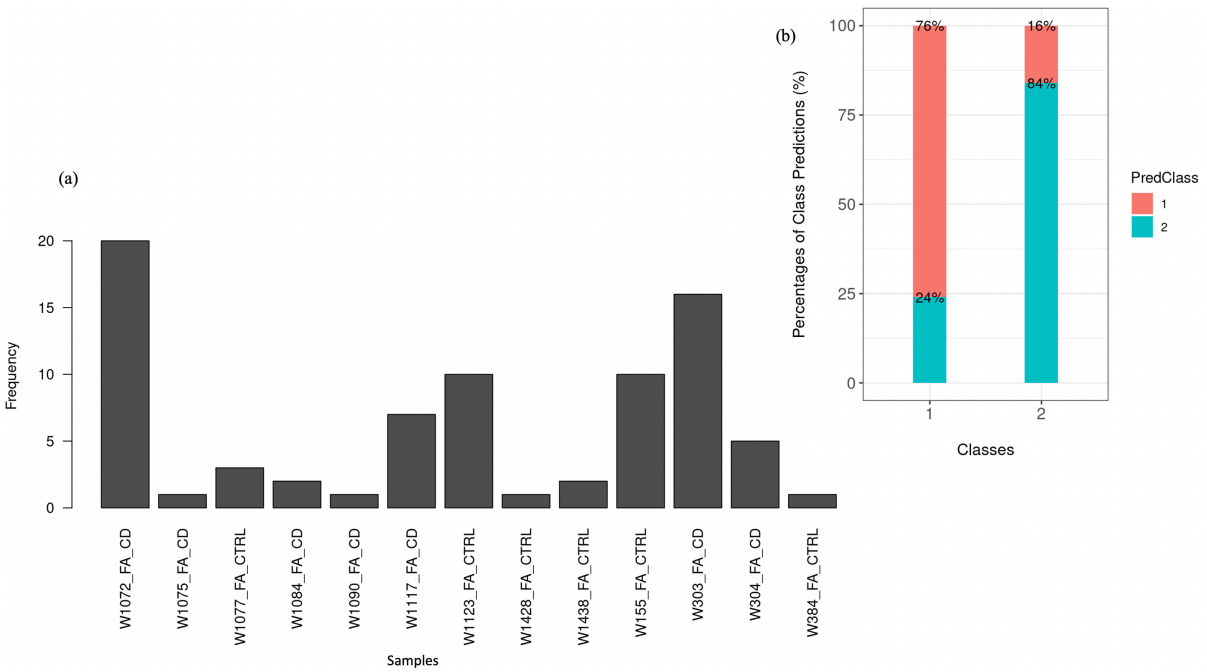

Figure 5. Differences in classification accuracy of faecal samples by the faecal SVM. (a) The frequency at which individual faecal samples were incorrectly classified by the faecal SVM. (b) Classification accuracy differences between disease classes in faecal samples, (class one corresponds with HC samples and class two with CD samples).

A random forest classification model was constructed using the faecal data to see if it could outperform the faecal SVM model. The random forest model used 3 randomly selected predictors and outperformed the SVM with an average accuracy of 83.33% (Figure 6). The random forest model had a kappa value of 0.65, which can be considered a good level of agreement between the predicted disease class and the actual disease class when accounting for the amount of agreement you would expect to see by chance.

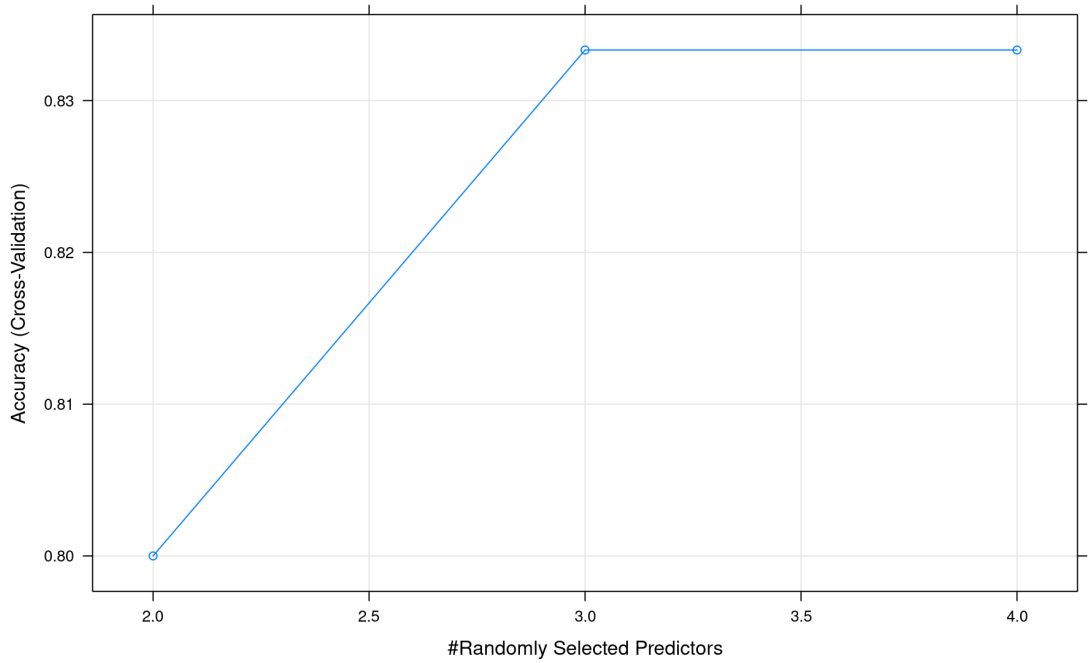

Figure 6. Differences in classification accuracy of a faecal sample random forest model with changes in the number of randomly selected predictors. The number of predictors used for the classification models is defined by the lowest number at which classification accuracy is highest, which is 3 in this case.

Conclusions

The results of this work suggest that Crohn's disease can be predicted from metabolic data with up to 83.33% accuracy and that the most useful type of metabolic data comes from faecal samples. It was also found that a random forest classification model was more effective at classifying metabolic data from faecal samples than a support vector machine model.

Bibliography

Chatzimichali, E.A. and Bessant, C. (2016) 'Novel application of heuristic optimisation enables the creation and thorough evaluation of robust support vector machine ensembles for machine learning applications', *Metabolomics*, 12(1), p. 16. doi:10.1007/s11306-015-0894-4.

Feuerstein, J.D. and Cheifetz, A.S. (2017) 'Crohn Disease: Epidemiology, Diagnosis, and Management', *Mayo Clinic Proceedings*, 92(7), pp. 1088–1103. doi:10.1016/j.mayocp.2017.04.010.

Kuhn, M. (2008) 'Building Predictive Models in R Using the caret Package', *Journal of Statistical Software*, 28, pp. 1–26. doi:10.18637/jss.v028.i05.

Torres, J. *et al.* (2017) 'Crohn's disease', *The Lancet*, 389(10080), pp. 1741–1755. doi:10.1016/S0140-6736(16)31711-1.

Exercise 3: Investigating the human kinase network (25%)

Introduction

There are approximately 518 human protein kinases which make up roughly 1.7% of human genes (Manning *et al.*, 2002; Duong-Ly and Peterson, 2013). Protein kinase phosphorylation activity is a ubiquitous mechanism for the temporal and spatial regulation of proteins involved in virtually all cellular processes (Pearlman, Serber and Ferrell, 2011). This phosphorylation consistently occurs amongst protein kinases, where they phosphorylate other protein kinases in networks known as signalling networks. These networks are essential to a large proportion of biological mechanisms and their disruption is associated with many disease phenotypes such as cancer (Smith *et al.*, 2020) and type-2 diabetes (Nandipati, Subramanian and Agrawal, 2017). The present work aims to characterise and visualise a protein kinase signalling network using prior knowledge of kinase interactions stored in the SIGNOR database (Perfetto *et al.*, 2016).

Methods

The complete set of phosphorylation data was downloaded from SIGNOR. The pandas python library (McKinney, 2010) was then used to manipulate key information from the downloaded data and remove information that was not required. Only a subset of all phosphorylation interactions stored on the SIGNOR database was used for this analysis in order to characterise a network small enough to be biologically meaningful. For interaction data to be retained, the interaction was required to be a direct phosphorylation interaction for which the residue and position of phosphorylation were known. Furthermore, only interactions between two protein kinases were retained and used to produce the visualisation of the signalling network.

Following the data manipulation, the curated data was imported into the Cytoscape software platform (Shannon *et al.*, 2003) for network visualisation. A signalling network was constructed from the data in Cytoscape, with protein kinases as nodes and phosphorylation interactions as edges. Further visual changes were made to the network to increase its interpretability.

The signalling network produced in Cytoscape was then analysed with the Cytoscape analyse network function to produce a table of summary statistics describing the signalling network. From these summary statistics a graph detailing the degree distribution of the network was created.

Results and Discussion

Visualisation and characterisation of the protein kinase signalling network

The protein kinase signalling network produced in this work and shown in Figure 1 is made up of 315 nodes and 1736 edges. There appear to be multiple hub genes with high numbers of edges connected to them. These hub genes are easily identifiable as they are the largest nodes in the network. It is probable that the protein kinases most important in cellular processes are

the hub nodes, such as SRC, visible in the network. The most obvious hub node shown in Figure 1 is the tyrosine kinase SRC, which has a degree of 90. Of the 90 edges linked to SRC, 78 of them are directed outwards and only 12 inwards. This means that in the described protein kinase network, SRC phosphorylates 78 tyrosine residues across many other genes in the network and is itself phosphorylated 12 times. Other protein kinases identified in this analysis which are likely to have important or essential cellular functions include MAPK1, PDPK1 and AKT1. A trend which is apparent in this network is that there are more high-importance tyrosine and serine kinases than threonine kinases, characterised by the relative absence of large green nodes in Figure 1.

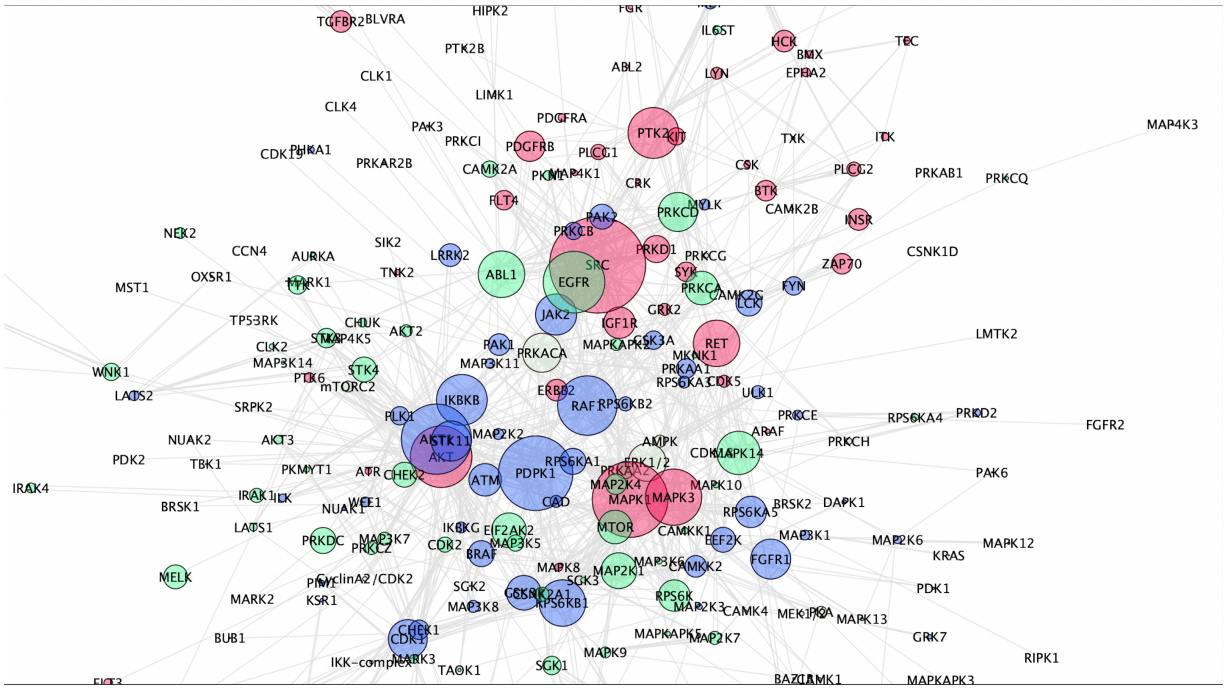

Figure 1. Visualisation of a protein kinase signalling network. Nodes in the network represent protein kinases of which there are 315 and the edges represent phosphorylation interactions of which there are 1736. The size of the node positively correlates with the degree of the node, and node colour represents the amino acid residue targeted for phosphorylation by the node. Pink nodes are tyrosine kinases, blue nodes are serine kinases and green nodes are threonine kinases. This signalling network was produced using Cytoscape.

How might the network change in response to MAPK1_HUMAN activity inhibition?

MAPK1 is a tyrosine kinase that has a degree of 71 in the outlined protein kinase signalling network. Inhibition of this protein would have a significant impact on the signalling network due to MAPK1's position as a core hub gene and the node with the second highest degree. Moreover, 62 of the edges connected to MAPK1 are directed outwards, meaning the inhibition would prevent the phosphorylation of up to 62 other kinases, which could in turn prevent them from phosphorylating their targets. This cascade effect of inhibition could have a drastic impact on the network in which the number of edges in the network would be greatly reduced. The reduction of activity in this network would likely have significant negative effects and possible negative disease outcomes.

What could be the effect on the network of a mutation on the 642nd residue of WEE1_HUMAN which changes serine to glutamine?

The change of serine to glutamine on the 642nd residue of WEE1 would prevent phosphorylation of WEE1 by BRSK1. This mutation would likely have a much smaller impact on the kinase signalling network than the inhibition of MAPK1 and there are several reasons for this. WEE1 is a kinase that has a relatively small role in the signalling network with a degree of 10, of which 2 edges are directed outwards and 8 inwards. This means fewer kinases will be impacted by the reduced level of WEE1 phosphorylation. The other reason that this mutation may only have small impact is that BRSK1 is the only kinase that phosphorylates WEE1 at the 642nd residue. This means that phosphorylation of WEE1 can still occur on other residues by the other source kinases. Moreover, there may be a level of redundancy in the kinase activity of WEE1 and phosphorylation at residues other than 642 may be enough to activate normal WEE1 phosphorylation activity. However, this mutation may still have a substantial impact on certain kinases in the network as phosphorylation of WEE1 by BRSK1 may be required to activate the phosphorylation activity of WEE1. Another mechanism by which this mutation could affect the network is through BRSK1 kinases being unable to unload phosphoryl groups by phosphorylating WEE1, which could prevent them from being phosphorylated. This again could have a cascading effect through the network, but this is unlikely as BRSK1 only has a degree of 2.

What effect may a mutation on the 474th residue of AKT1_HUMAN, changing tyrosine to phenylalanine, have on the network?

AKT1 has a large impact on the signalling which means disruption to its phosphorylation activity could have very large and consequential effects on the network. However, AKT1 does not get phosphorylated by any kinases at residue 474, which could mean that the mutation would have minimal impact on the network activity. AKT1 does get phosphorylated at residue 473 by several other kinases. Consequently, the impact of this mutation on the signalling network would depend on whether the mutation would change the structure of the protein sufficiently to inhibit phosphorylation at the 473rd residue. If phosphorylation at the 473rd residue of AKT1 was inhibited, then a drastic decrease in the kinase activity of the signalling network could be observed at a similar scale to inhibition of MAPK1.

What does the degree distribution of the network look like and how does it compare to other biological networks?

The degree distribution of this network is highly right skewed, with extremely high frequencies of nodes with very low degrees and very low frequencies of nodes with high degrees (figure 2). This degree distribution appears to show a power-law distribution in which a small number of highly connected hub-nodes impact the distribution. This form of distribution is common in biological networks, especially in protein and metabolite interaction networks (Jeong *et al.*, 2000; Yook, Oltvai and Barabási, 2004). Even in non-biological networks, distributions of this kind are not uncommon, with a famous example being the power-law degree distribution of the internet (Govindan and Tangmunarunkit, 2000).

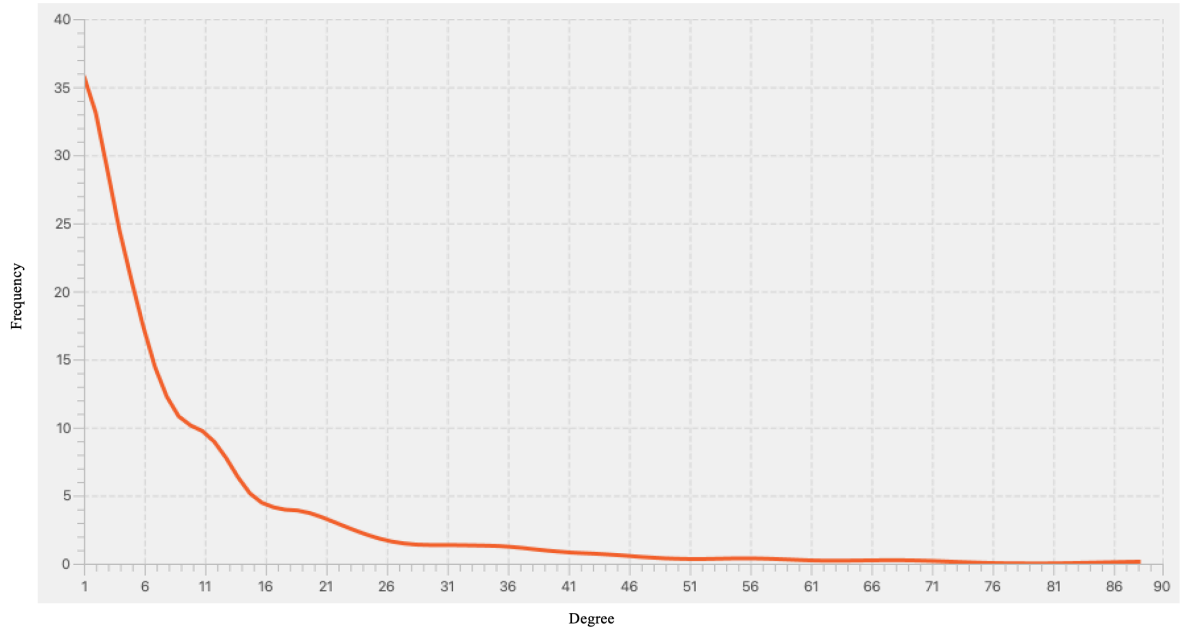

Figure 2. The degree distribution for kinases in the signalling network. This plot shows that very low frequencies of kinases have high degrees in the signalling network and that a very high frequency of kinases have a low degree.

Conclusions

This work was able to produce a protein kinase signalling network from phosphorylation data stored on the SIGNOR database. The importance of several kinases in cellular process was inferred by using the degree distribution as a measure of kinase importance. Moreover, the degree distribution observed was found to be in line with those of similar biological networks.

Bibliography

Duong-Ly, K.C. and Peterson, J.R. (2013) 'The Human Kinome and Kinase Inhibition as a therapeutic strategy', *Current protocols in pharmacology / editorial board, S.J. Enna (editor-in-chief) ... [et al.]*, 0 2, p. Unit2.9. doi:10.1002/0471141755.ph0209s60.

Govindan, R. and Tangmunarunkit, H. (2000) 'Heuristics for Internet map discovery', in *Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064). Proceedings IEEE INFOCOM 2000. Conference on Computer Communications. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies (Cat. No.00CH37064)*, pp. 1371–1380 vol.3. doi:10.1109/INFCOM.2000.832534.

Jeong, H. *et al.* (2000) 'The large-scale organization of metabolic networks', *Nature*, 407(6804), pp. 651–654. doi:10.1038/35036627.

Manning, G. *et al.* (2002) 'The Protein Kinase Complement of the Human Genome', *Science* [Preprint]. doi:10.1126/science.1075762.

McKinney, W. (2010) 'Data Structures for Statistical Computing in Python', in. *Python in Science Conference*, Austin, Texas, pp. 56–61. doi:10.25080/Majora-92bf1922-00a.

Nandipati, K.C., Subramanian, S. and Agrawal, D.K. (2017) 'Protein kinases: mechanisms and downstream targets in inflammation-mediated obesity and insulin resistance', *Molecular and Cellular Biochemistry*, 426(1–2), pp. 27–45. doi:10.1007/s11010-016-2878-8.

Pearlman, S.M., Serber, Z. and Ferrell, J.E. (2011) 'A Mechanism for the Evolution of Phosphorylation Sites', *Cell*, 147(4), pp. 934–946. doi:10.1016/j.cell.2011.08.052.

Perfetto, L. *et al.* (2016) 'SIGNOR: a database of causal relationships between biological entities', *Nucleic Acids Research*, 44(D1), pp. D548–D554. doi:10.1093/nar/gkv1048.

Shannon, P. *et al.* (2003) 'Cytoscape: a software environment for integrated models of biomolecular interaction networks', *Genome Research*, 13(11), pp. 2498–2504. doi:10.1101/gr.1239303.

Smith, H.L. *et al.* (2020) 'DNA damage checkpoint kinases in cancer', *Expert Reviews in Molecular Medicine*, 22, p. e2. doi:10.1017/erm.2020.3.

Yook, S.-H., Oltvai, Z.N. and Barabási, A.-L. (2004) 'Functional and topological characterization of protein interaction networks', *PROTEOMICS*, 4(4), pp. 928–942. doi:10.1002/pmic.200300636.