Using host "omic" datasets to better understand and diagnose paediatric infectious and inflammatory diseases

Heather Jackson, MSc., BSc. (Hons.)

Department of Infectious Disease, Imperial College London

Supervisors:

Dr. Myrsini Kaforou
Dr. Jethro Herberg
Prof. Michael Levin

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

July 2022

Declaration of Originality

I, Heather Jackson, declare that this thesis titled "Using host "omic" datasets to better understand and diagnose paediatric infectious and inflammatory diseases" and the work presented in it are my own. It is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy. The work described here has not been submitted for any degree, diploma, or any other qualification. This thesis is the result of my own work and the guidance of my supervisors, except what is described in the text below. These statements are repeated at the start of each respective chapter. Use of the first-person plural is for reasons of clarity.

The clinical databases and omic datasets were made available to me by my supervisors, Dr. Myrsini Kaforou, Dr. Jethro Herberg, and Prof. Michael Levin. The clinical recruitment of patients in chapter 3 was part of the EUCLIDS and PERFORM studies which were led by Prof. Mike Levin in addition to principal investigators from each clinical site. The patients with Kawasaki disease included in chapter 4 were either recruited through the University of California San Diego (UCSD) Kawasaki Disease Research Center which was led by Prof. Jane Burns, or through the "Genetic determinants of Kawasaki Disease for susceptibility and outcome" study which was led by Dr. Jethro Herberg and Rachel Galassini. The febrile control patients in chapter 4 were included in the EUCLIDS or PERFORM studies, or the IRIS or GENDRES studies which were led by Prof. Mike Levin and Dr. Federico Martinon-Torres, respectively. The clinical recruitment of patients included in chapter 5 was part of the DIAMONDS study which was led by Prof. Mike Levin, in addition to the EUCLIDS and PERFORM studies and the UCSD Kawasaki Disease Research Center.

Chapter 3:

Dr. Shea Hamilton performed the laboratory elements for the generation of the SomaScan dataset. Dr. Stephanie Menikou, Dr. Honglei Huang, Adam Thorne and Dr. Roman Fischer performed the laboratory elements and data generation for the liquid chromatography-tandem mass spectrometry (LS-MS/MS) experiments. Judith Zandstra and Prof. Taco Kuijpers performed the literature search to identify potential protein biomarkers, and they also performed protein quantification using ELISA and Luminex assays in cross-platform validation phase 1 and 2. Dr. Andrew McArdle processed the LC-MS/MS data using MaxQuant. FS-PLS was developed by Prof. Lachlan Coin.

Chapter 4:

Dr. Shea Hamilton performed the laboratory elements for the generation of the SomaScan dataset. Dr. Stephanie Menikou, Dr. Honglei Huang, Adam Thorne and Dr. Roman Fischer performed the laboratory elements and data generation for the LS-MS/MS experiments. Dr. Andrew McArdle processed the LC-MS/MS data using MaxQuant. Dr. Victoria Wright performed the laboratory

elements and data generation of the transcriptomic sequencing experiments. Quality control and normalisation of the transcriptomic validation datasets was performed by Dr. Clive Hoggart and Dr. Victoria Wright.

Chapter 5:

Giselle d'Souza, Dr. Samuel Nichols, Oliver Powell and Dr. Victoria Wright performed the laboratory work for the RNA Sequencing (RNA-Seq). Dominic Habgood-Coote processed the raw RNA-Seq data, generating gene and exon counts that I then used. Luca Miglietta, Giselle d'Souza and Dr. Jesus Rodriguez Manzano performed the laboratory work for the reverse transcription-quantitative polymerase chain reaction (RT-qPCR) quantification of the gene targets. FS-PLS was developed by Prof. Lachlan Coin.

Heather Jackson

London, July 2022

Abstract

The presence of an infectious or inflammatory disease leads to changes in gene expression and protein abundance which are detectable in the host's blood. Through studying the host omic profiles of individuals with infectious and inflammatory diseases, we can identify molecular changes that occur due to disease. Sparse combinations of genes or proteins that distinguish between disease groups of interest can be determined using feature selection. These sparse combinations, which are known as host diagnostic signatures, can be developed into point-of-care tests for rapid and accurate disease diagnosis. Host omic profiles can also be used to uncover important genes, proteins, and biological pathways implicated in disease progression and manifestation.

In this thesis, I have used host transcriptomic and proteomic profiles to address three major challenges in paediatric infectious and inflammatory disease. I have firstly described a 6-protein host diagnostic signature that can accurately differentiate between bacterial and viral infections in febrile children, with area under the receiver operating characteristic curves (AUCs) between 89.4%-93.6% in an independent validation set. Secondly, Kawasaki disease (KD) is a paediatric inflammatory syndrome without a known cause. I have used host transcriptomic and proteomic profiles to determine whether the host response to KD detectable on the molecular level has more in common with the host response to bacterial or viral infections, or indeed neither. These analyses have revealed considerable heterogeneity in the host response to KD. Finally, I have discovered a 5-gene host diagnostic signature that accurately distinguished multisystem inflammatory syndrome in children from KD, bacterial infections, and viral infections with AUCs between 91.7%-93.2% in an independent validation set.

In summary, this thesis has demonstrated that the combination of high-dimensional omic datasets and statistical and machine learning tools provides invaluable insights into the biology of paediatric infectious and inflammatory diseases and offers alternative diagnostic approaches.

Acknowledgements

There are many people who I would like to thank for having made this thesis possible. I owe a special thanks to my main supervisor, Dr. Myrsini Kaforou, who has taught me so much during the past four years. Myrsini, your unwavering support and friendship have allowed me to thrive in my PhD. You have been a constant source of inspiration and I will always be grateful to have worked with you. Many thanks to my other supervisors, Prof. Mike Levin and Dr. Jethro Herberg, for your invaluable contributions and help throughout the years. Thank you to all three of my supervisors for including me wholeheartedly in the team and for providing me with the support and opportunities to grow as a scientist.

I have been extremely fortunate to have worked with fantastic people throughout my PhD. I am grateful for the Levin Research Group, the whole Section of Paediatric Infectious Diseases at Imperial College, and Prof. Aubrey Cunnington for fostering such a positive working environment within the section. The Bioinformatics group has been a great sounding board throughout the years, and I am thankful to all of its members. I am particularly grateful that my path crossed with Isobel Parkes whilst she worked at Imperial and for her friendship ever since.

I would like to thank the members of the PERFORM and DIAMONDS consortia as well as Prof. Jane Burns and her team at UCSD for welcoming me into your research projects and for allowing me to use your valuable data. Thank you to Dr. Jesus Rodriguez-Manzano and Luca Miglietta for imparting your diagnostics and RT-qPCR wisdom on to me. Many thanks to Dr. Myrsini Kaforou, Dr. Shea Hamilton, Dr. Andrew McArdle, Dominic Habgood-Coote, Isobel Parkes, Tara Hameed and Jonathan Ish-Horowicz for reading sections of my thesis and for providing me with extremely valuable feedback.

I would like to thank The Wellcome Trust for their generous funding throughout my master's and PhD. Their funding has enabled me to attend multiple scientifically stimulating conferences and has allowed me to enjoy living in London comfortably, a privilege that few PhD students experience. For your friendship and support over the past five years: Jonathan Ish-Horowicz, Tara Hameed, Maddy Hall, Léonie Strömich. I am so grateful that we went through this journey together and I wouldn't have had it any other way.

To my friends in London and elsewhere: thank you for always bringing a smile to my face, for distracting me from my PhD and for the unforgettable memories during the past five years. Jordan, thank you for putting up with the ups and downs of PhD life and for always being there for me.

Finally, and most importantly, thank you to my family, without whom none of this would have been possible. To my parents, grandparents, and sisters: thank you for your unconditional love and support and for always believing in me. This thesis is for you.

Table of Contents

Declaration of Originality **3**
Abstract **5**
Acknowledgements **6**
Table of Contents **7**
List of Figures **10**
List of Tables **10**
Abbreviations **16**
Copyright Declaration **21**
Publications **22**
1 Introduction **25**
1.1 Host response to infection 25
1.2 Transcription and translation 28
1.3 An overview of omic data and its applications 30
1.3.1 Transcriptomics 32
1.3.1.1 Overview 32
1.3.1.2 Low-throughput RNA quantification 33
1.3.1.3 High-throughput RNA quantification 35
1.3.2 Proteomics 38
1.3.2.1 Overview 38
1.3.2.2 Low-throughput protein detection methods 38
1.3.2.3 High-throughput protein detection methods 39
1.4 Use of host omics in paediatric infectious and inflammatory disease 43
1.4.1 Host omics for improving understanding of pathogenesis and disease manifestation 43
1.4.2 Host diagnostic signatures 44
1.4.3 Diagnosing bacterial and viral infections in febrile children 45
1.4.4 Improving diagnosis and understanding of Kawasaki Disease 47
1.4.5 Multisystem inflammatory syndrome in children 48
1.5 Thesis contributions 50
2 Methods **54**
2.1 Overview 54
2.2 Raw data processing 54
2.2.1 Processing raw RNA Sequencing data 55
2.2.2 Processing raw liquid chromatography tandem mass spectrometry data 56
2.3 Quality control, transformation, and normalisation 57
2.3.1 Quality control and transformation of microarray data (Figure 5A) 61
2.3.2 Quality control and transformation of RNA Sequencing data (Figure 5B) 62
2.3.3 Quality control and transformation of SomaScan® data (Figure 5C) 63
2.3.4 Quality control and transformation of liquid chromatography tandem mass spectrometry data (Figure 5D) 63
2.4 Data analysis 63
2.4.1 Clustering 63
2.4.2 Differential expression and abundance analysis 64
2.4.3 Feature selection 67
2.4.4 Signature performance evaluation 68
2.4.5 Cross validation 70

2.4.6 Pathway enrichment analysis 71
3 Using a multi-platform approach to identify a host protein signature for distinguishing between bacterial and viral infections in febrile children 73
3.1 Chapter aims and contributions 73
3.2 Cohorts and study design 76
3.2.1 Case definitions 82
3.2.2 Data generation 83
3.2.2.1 Discovery datasets 83
3.2.2.2 Cross-platform validation phase 1 84
3.2.2.3 Cross-platform validation phase 2 85
3.2.3 Dataset exploration 85
3.3 Results 88
3.3.1 Discovery phase 88
3.3.2 Cross-platform validation phase 1 93
3.3.3 Cross-platform validation phase 2 98
3.3.3.1 Comparison to the MeMed BV™ 100
3.3.3.2 The 6-protein signature in uncertain diagnostic groups 102
3.3.4 Biological functions of the 6 proteins 103
3.4 Summary 104
4 Kawasaki Disease patient stratification and pathway analysis based on host transcriptomic and proteomic profiles 107
4.1 Chapter aims and contributions 107
4.2 Cohorts and study design 109
4.2.1 Case definitions 112
4.2.2 Data generation 112
4.2.3 Dataset exploration 113
4.3 Results 117
4.3.1 Comparing Kawasaki disease to bacterial and viral infections 117
4.3.1.1 Differential expression/abundance analysis 117
4.3.1.2 Pathway enrichment analysis 119
4.3.1.3 Clustering with bacterial and viral infections 122
4.3.1.4 Classifying Kawasaki disease as bacterial or viral 125
4.3.2 Clustering Kawasaki disease patients alone 129
4.4 Summary 130
5 Diagnosis of multisystem inflammatory syndrome in children by a whole-blood transcriptional signature 134
5.1 Chapter aims and contributions 134
5.2 Cohorts and study design 136
5.2.1 Case definitions 140
5.2.2 Data generation 141
5.2.2.1 Discovery RNA Sequencing dataset 141
5.2.2.2 Reverse transcription-quantitative polymerase chain reaction validation dataset 142
5.2.3 Dataset exploration 144
5.3 Results 145
5.3.1 Discovery phase 146
5.3.1.1 Exon analysis 151
5.3.2 Validation phase 152
5.3.2.1 Assay optimisation 152
5.3.2.2 4-gene FS-PLS signature performance 154

5.3.2.3 Improvement of the 4-gene signature using the 6 significantly differentially expressed genes 156
5.3.2.4 Performance of the 5-gene RT-qPCR signature 159
5.3.2.5 Performance of the 5-gene RT-qPCR signature using cross validation 161
5.3.3 Biological relevance of the 5 genes in the 5-gene RT-qPCR signature 162
5.4 Summary 163
6 Discussions and future directions 166
6.1 Conclusion 168
Bibliography 170

List of Figures

Figure 1: Cells of the innate and adaptive immune systems. Image created with BioRender.....27
Figure 2: Overview of the central dogma of molecular biology, i.e., the conversion of DNA to RNA to protein. DNA = deoxyribonucleic acid; RNA = ribonucleic acid; tRNA = transfer RNA; mRNA = messenger RNA. Image created with BioRender. 30
Figure 3: Four of the principal levels of biological organisation and their corresponding omic level. DNA = deoxyribonucleic acid; RNA = ribonucleic acid. Image created with BioRender. 32
Figure 4: The World Health Organisation (WHO) criteria for diagnosing MIS-C. Information taken from: https://www.who.int/news-room/commentaries/detail/multisystem-inflammatory-syndrome-in-children-and-adolescents-with-covid-19 (World Health Organisation, 2020). ECHO = echocardiogram; NT-proBNP = N-terminal pro b-type natriuretic peptide; PT = prothrombin time; PTT = partial thromboplastin time; ESR = erthryocyte sedimentation rate; COVID-19 = Coronavirus Disease 2019. 50
Figure 5: Overview of quality control and normalisation processes carried out for each type of omic data presented in this thesis. **A**: microarray; **B**: RNA-Seq; **C**: SomaScan; **D**: LC-MS/MS. RNA-Seq = RNA Sequencing; LC-MS/MS = liquid chromatography tandem mass spectrometry; PCA = principal component analysis. 61
Figure 6: A schematic representing a receiver operating characteristic (ROC) curve. The black line represents a test fit, whilst the red line represents a random guess. The ideal test is shown in the top left corner of the plot.....70
Figure 7: An overview of the biomarker discovery and validation phases carried out and described in chapter 3. The discovery phase involved identifying a shortlist of protein biomarker candidates from high-dimensional proteomic datasets. The candidates were then validated in cross-platform validation phase 1. A sparse protein signature was identified from the proteins that underwent cross-platform validation phase 1. This sparse signature was subjected to another round of cross-platform validation (cross-platform validation phase 2). LC-MS/MS = liquid chromatography-tandem mass spectrometry; FS-PLS = forward selection-partial least squares; NSDB = non-sterile definite bacterial. Image created with BioRender.....77
Figure 8: The clinical phenotyping algorithm used to classify patients included in the PERFORM and DIAMONDS studies. Patient groups with * indicate that these patients could have a viral coinfection. This phenotyping algorithm was used in the PERFORM and DIAMONDS studies. The figure was adapted from [160].....83
Figure 9: Principal component analysis (PCA) plots for principal components (PCs) 1 and 2 (**A**-**B**) and 3 and 4 (**D**-**F**) for the SomaScan dataset. This dataset was used as a discovery dataset for the identification of candidate protein biomarkers for distinguishing between bacterial and viral infections in febrile children. Points are coloured according to disease group (**A**, **D**), age in months (**B**, **E**), and sex at birth (**C**, **F**).....86
Figure 10: Principal component analysis (PCA) plots for principal components (PCs) 1 and 2 (**A**-**B**) and 3 and 4 (**D**-**F**) for the MS-A dataset. This dataset was used as a discovery dataset for the identification of candidate protein biomarkers for distinguishing between bacterial and viral infections in febrile children. Points are coloured according to disease group (**A**, **D**), age in months (**B**, **E**), and sex at birth (**C**, **F**).....87
Figure 11: Principal component analysis (PCA) plots for principal components (PCs) 1 and 2 (**A**-**B**) and 3 and 4 (**D**-**F**) for the MS-B dataset. This dataset was used as a discovery dataset for the identification of candidate protein biomarkers for distinguishing between bacterial and viral infections in febrile children. Points are coloured according to disease group (**A**, **D**), age in months (**B**, **E**), and sex at birth (**C**, **F**).....88
Figure 12: Volcano plots showing the log2 fold-change (logFC) values and the -log10 BH p-values for proteins significantly differentially abundant between bacterial vs. viral infections in febrile children. Results are shown for the SomaScan (**A**), MS-A (**B**), and MS-B (**C**) datasets. Red points are proteins with BH p-values <0.05 and absolute logFC values >1; gold points are proteins with BH p-values <0.05 and absolute logFC <1; green points are proteins with absolute logFC >1 and

BH p-values >0.05; black points are not significant (NS). Positive and negative logFC indicate proteins increase and decrease in bacterial patients, respectively. 90
Figure 13: Logged abundance values of the proteins measured by Luminex and ELISA in cross-platform validation phase 1 coloured according to disease group. These proteins were identified in the discovery phase as potential discriminatory biomarkers for bacterial vs. viral infections. ns: p-value >0.05; *: p-value <0.05; **: p-value ≤0.01; ***: p-value ≤0.001; ****: p-value ≤0.0001. All p-values are adjusted for multiple testing using Benjamini-Hochberg. 95
Figure 14: The ROC curves of the protein signatures identified by running FS-PLS on either Luminex (**A**) or Luminex and ELISA (**B**) proteins measured in cross-platform validation phase 1. The Luminex signature is: SELE, IL18, NCAM1, NGAL, IFN-γ. The Luminex + ELISA signature is: SELE, IL18, NCAM1, SAA1, ANGPT2. The AUCs (95% CI) are printed on the plot. CI = confidence interval. 97
Figure 15: ROC curves showing the performance of the 5-protein signature in the cross-platform validation phase 2 dataset. The 5-protein signature was identified in cross-platform validation phase 1 (section 3.3.2) for distinguishing between bacterial and viral infections. The ROC curves are calculated using (**A**) original logistic regression weights calculated from cross-platform validation phase 1 ("original model" in Table 9) (**B**) logistic regression weights retrained using cross-platform validation phase 2 data ("retrained model" in Table 9), and (**C**) the simple DRS ("simple DRS" model in Table 9). AUCs and 95% confidence intervals (CI) are printed on the plots. 99
Figure 16: Comparison of the AUC (95% CIs) between the 5-protein signature (**A**) vs. the 5-protein signature + LG3BP (**B**) in distinguishing between bacterial and viral infections. ROC curves were generated using cross-platform validation phase 2 data and using the simple DRS. The 5-protein signature was identified by FS-PLS and LG3BP was identified by Limma in cross-platform validation phase 1 (section 3.3.1). CI = confidence interval. AUCs and 95% confidence intervals (CI) are printed on the plots. 100
Figure 17: The weighted disease risk scores for the 6-protein signature calculated using retrained model weights for individuals in all phenotypic groups in the cross-platform validation phase 2 dataset. The dashed line shows the threshold of classification, i.e., the disease risk score values above which a sample was classified as bacterial. This threshold was calculated using Youden's index [152]. to optimise sensitivity (88.1%) and specificity (70.9%). 103
Figure 18: BH-adjusted p-values obtained from the Dunn's test contrasting in silico immune cell proportion estimates between each of the phenotypic groups. Immune cell proportions were estimated using CIBERSORTx. KD = Kawasaki disease; HC = healthy controls. Point size represents -log10 (BH-adjusted p-values). 114
Figure 19: PCA biplots for transcriptomic discovery dataset used in chapter 4. Points are coloured by disease group (**A**, **D**), age in months (**B**, **E**) and sex at birth (**C**, **F**). PCs 1-2 and 3-4 are shown. KD = Kawasaki disease; F = female; M = male; PC = principal component. 116
Figure 20: PCA biplots for the proteomic discovery dataset used in chapter 4 with points coloured by disease group (**A**, **D**), age in months (**B**, **E**) and sex at birth (**C**, **F**). PCs 1-2 and 3-4 are shown. KD = Kawasaki disease; F = female; M = male; PC = principal component. 117
Figure 21: Cross plots contrasting the differential abundance analysis results from KD vs. HC to those from bacterial vs. HC (**A** for transcriptomics; **C** for proteomics) and viral vs. HC (**B** for transcriptomics; **D** for proteomics). LFC values for each feature are plotted against each other to show concordance (same LFC direction) and discordance (mismatched LFC direction). Each point represents a feature (gene or protein). KD = Kawasaki Disease; HC = healthy controls; LFC = log2 fold-change. Significance is considered as BH-adjusted p-values <0.05. 119
Figure 22: The top pathways identified as significantly enriched from the lists of genes (**A**) and proteins (**B**) SDA between KD, bacterial, and viral infections vs. healthy controls. Pathways were identified by g:Profiler2 followed by filtering with REVIGO. Points represent pathways with colour indicating directionality and size representing BH-adjusted p-value (larger = lower p-value). Direction = "Both" indicates that the pathway was identified as enriched from the list of genes that increased and decreased in the disease group vs. healthy controls. 121
Figure 23: The proportion of patients from each disease group in each cluster for transcriptomics (**A**) and proteomics (**B**). Clusters were identified used k-means clustering with k=3 for

transcriptomics and k=2 for proteomics. DB = definite bacterial; DV = definite bacterial; KD = Kawasaki disease.....122
Figure 24: Pathways upregulated and downregulated between KD patients in clusters 1, 2 and 3 for the transcriptomic dataset. Clusters were identified using k-means applied to KD, bacterial and viral patients. Pathways were identified by g:Profiler2 followed by filtering with REVIGO. Points represent pathways with colour indicating directionality and size representing BH-adjusted p-value (larger = lower p-value). Direction = "Both" indicates that the pathway was identified as enriched from the list of genes that increased and decreased in each comparison. BH-adjusted p-values <0.05 were considered significant. KD = Kawasaki disease. 125
Figure 25: Bacterial DRS (DB-DRS) plotted against viral DRS (DV-DRS) for KD (discovery and validation), bacterial (from the validation cohort) and viral (from the validation cohort) patients from the (**A**) transcriptomic and (**B**) proteomic datasets. Boxplots are shown for each disease group. Dotted lines represent the thresholds of classification obtained by Youden's index. KD = Kawasaki disease; DRS = disease risk score. 128
Figure 26: The proportion of KD transcriptomic samples assigned to clusters KD1-T, KD2-T and KD3-T coloured according to the cluster they were assigned previously in section 4.3.1.3. Clustering was performed by k-means using only the KD patient transcriptomes. 130
Figure 27: An overview of the key steps and patient groups included in this chapter. The study was split into two phases: 1) the discovery phase, in which biomarker candidates for diagnosing MIS-C were identified from RNA sequencing; and 2) the validation phase, in which the biomarker candidates were quantified in an independent set of samples using RT-qPCR. KD = Kawasaki disease; HC = healthy controls; FS-PLS = forward selection-partial least squares; RT-qPCR = reverse transcription-quantitative polymerase chain reaction. 136
Figure 28: The World Health Organisation (WHO) case definition for MIS-C. Information taken from: https://www.who.int/news-room/commentaries/detail/multisystem-inflammatory-syndrome-in-children-and-adolescents-with-covid-19 (World Health Organisation, 2020). ECHO = echocardiogram; NT-proBNP = N-terminal pro b-type natriuretic peptide; PT = prothrombin time; PTT = partial thromboplastin time; ESR = erthryocyte sedimentation rate; COVID-19 = Coronavirus Disease 2019. 141
Figure 29: Principal Component Analysis (PCA) biplots from PCA performed on the discovery RNA-Seq dataset after normalisation. Principal components (PC) 1-2, PC3-4 and PC5-6 are shown. Points represent samples and are coloured by phenotype (**A**), sex at birth (**B**), age in months (**C**), and RNA-Seq batch (**D**). F = female; M = male; KD = Kawasaki disease; MIS-C = multisystem inflammatory syndrome in children.....145
Figure 30: Volcano plots showing log2 fold change (LFC) values and adjusted -log10 p-values from differential expression analysis comparing **A**: MIS-C vs. Kawasaki Disease; **B**: MIS-C vs. viral infections; **C**: MIS-C vs. bacterial infections; **D**: MIS-C vs. KD, viral and bacterial infections combined. Data is from the RNA-Seq discovery dataset. Points represent genes. Points are coloured as followed: red - genes with adjusted p-values <0.05 and absolute LFC >1; yellow – genes with adjusted p-values <0.05 and absolute LFC <1; green – absolute LFC >1; and black = not significant (NS). Genes to the right of 0 are increased in MIS-C patients. MIS-C = multisystem inflammatory syndrome in children. 147
Figure 31: ROC curves for the performance of the 4-gene signature identified by FS-PLS in the RNA-Seq discovery dataset. The 4-gene signature is composed of HSPBAP1, VPS37C, TGFB1 and MX2. **A**: performance of the signature in distinguishing between MIS-C vs. KD+viral+bacterial combined; **B**: performance of the signature for MIS-C vs. KD; **C**: performance of the signature for MIS-C vs. viral; and **D**: performance of the signature for MIS-C vs. bacterial. AUCs and 95% confidence intervals (CI) are printed on the plots. KD = Kawasaki disease; MIS-C = multisystem inflammatory syndrome in children.....149
Figure 32: Gene expression levels in the RNA-Seq data of the 10 genes identified as candidate biomarkers for distinguishing between MIS-C and KD, viral and bacterial infections from the RNA-Seq discovery dataset. HSPBAP1, VPS37C, TGFB1 and MX2 were selected by FS-PLS. TRBV11-2, FOSB, NCK2, PER1, HROB and CHKA were selected as they were the top SDE genes between MIS-C vs. KD, viral and bacterial. KD = Kawasaki disease; MIS-C = multisystem inflammatory syndrome in children.....150

Figure 33: Boxplots showing normalised CT values for each of the candidate biomarker genes taken forward to validation. Data is shown for the full RT-qPCR validation dataset. HSPBAP1, VPS37C, TGFB1 and MX2 were originally selected by FS-PLS in the discovery dataset, whilst TRBV11-2, FOSB, NCK2, PER1, HROB and CHKA were selected through differential expression analysis contrasting MIS-C to KD, viral and bacterial infections. KD = Kawasaki disease; COVID-19 = Coronavirus Disease 2019; MIS-C = multisystem inflammatory syndrome in children. 154
Figure 34: Performance of the 4-gene signature identified by FS-PLS with performance shown for the RT-qPCR validation dataset. **A-B**: ROC curve for evaluating the performance of the signature in contrasting MIS-C from all comparator groups including or excluding COVID-19. **C-F**: ROC curve showing performance for MIS-C vs. individual comparator groups. AUCs and 95% confidence intervals (CI) are printed on the plots. KD = Kawasaki disease; MIS-C = Multisystem Inflammatory Syndrome in Children. 155
Figure 35: Performance of the 5-gene RT-qPCR signature (HSPBAP1, VPS37C, TGFB1, MX2, TRBV11-2) in the RT-qPCR validation set. The 5-gene signature was derived from adding TRBV11-2 to the 4-gene signature from FS-PLS. **A-B**: ROC curve for evaluating the performance of the 5-gene RT-qPCR signature in contrasting MIS-C from all comparator groups without (**A**) and with (**B**) COVID-19. **C-F**: performance in contrasting MIS-C from individual comparator groups. AUCs and 95% confidence intervals (CI) are printed on the plots. KD = Kawasaki disease; MIS-C = multisystem inflammatory syndrome in children; COVID-19 = Coronavirus Disease 2019..... 160
Figure 36: Boxplots showing the weighted disease risk score for patients from the different disease groups. The dashed line represents the threshold of classification calculated using the Youden's index by maximising both sensitivity and specificity of MIS-C vs. KD, viral and bacterial. Sensitivity = 91.7% and specificity = 81.7% for MIS-C vs. KD, viral and bacterial. MIS-C = multisystem inflammatory syndrome in children; KD = Kawasaki disease; COVID-19 = Coronavirus Disease 2019. 160
Figure 37: ROC curves for the 5-gene RT-qPCR signature calculated using the MIS-C predictions returned from 10-fold cross validation. Probability of MIS-C was estimated for each patient in the left-out fold. **A**: performance for MIS-C vs. KD, viral and bacterial. **B**-**D**: performance for MIS-C vs. each comparator disease group separately. AUCs (95% confidence intervals) are printed on the plots. KD = Kawasaki disease; MIS-C = multisystem inflammatory syndrome in children. 162

List of Tables

Table 1: Clinical and laboratory features of patients whose samples were included in the discovery of novel protein biomarkers for distinguishing between bacterial and viral infections in febrile children. The SomaScan dataset was generated using the SomaScan platform. MS-A and MS-B were generated from liquid chromatography tandem mass spectrometry.DB = definite bacterial; DV = definite viral; IQR = interquartile range; RSV = respiratory syncytial virus; EBV = Epstein-Barr virus; CRP = C-reactive protein. 79
Table 2: Clinical and laboratory features of patients whose samples were included in cross-platform validation phase 1. The levels of protein biomarker candidates identified in the discovery phase were tested using these samples. ELISA and Luminex immunoassays were used for protein quantification. DB = definite bacterial; DV = definite viral; IQR = interquartile range; CRP = C-reactive protein; RSV = respiratory syncytial virus. 80
Table 3: Clinical and laboratory features of patients whose samples were included in cross-platform validation phase 2. Luminex immunoassays were used for protein quantification. The protein signature identified in cross-platform validation phase 1 was tested in these individuals. DB = definite bacterial; NSDB = non-sterile definite bacterial; DV = definite viral; CRP = C-reactive protein; IQR = interquartile range; RSV = respiratory syncytial virus. 82
Table 4: The shortlist of proteins candidates for distinguishing between bacterial and viral infections in febrile children. These proteins were identified from the three datasets used in the discovery phase: SomaScan, MS-A and MS-B datasets. SDA = significantly differentially abundant. Protein symbols with * indicate that the ELISA/Luminex assays were available for this protein. 92
Table 5: Proteins identified as potential biomarkers for distinguishing between bacterial and viral infections through literature searches. 93
Table 6: The 29 proteins biomarker candidates with ELISA or Luminex immunoassays available taken forward into the first phase of cross-platform validation for distinguishing between bacterial and viral infections in febrile children. P-values were calculated using Mann-Whitney U tests comparing bacterial to viral infections and corrected for multiple testing using the Benjamini-Hochberg (BH) procedure. * next to the protein names indicates that the protein was identified during the discovery phase. BH-adjusted p-values <0.05 are significant. ns = not significant. 94
Table 7: The 2 signatures identified by iterative FS-PLS using cross-platform validation phase 1 measurements of either Luminex proteins or Luminex + ELISA proteins for distinguishing between bacterial and viral infections in febrile children. Robustness indicates the number of times the protein was included in any signature across the 25 iterations of FS-PLS/number of total iterations. Proteins with names italicised and underlined increase and decrease in bacterial infections, respectively. pAUC = partial AUC; CI = confidence interval. 97
Table 8: The weights of the 5 proteins included in the Luminex signature identified in cross-platform validation phase 1 by FS-PLS. "Cross-platform validation phase 1" weights were used in the "Original" model and "cross-platform validation phase 2" weights were used in the "Retrained" model. Weights were calculated using a GLM model contrasting bacterial to viral infections. Positive and negative weights indicate that the protein increases and decreases in bacterial infections, respectively. 99
Table 9: The performance of the 5-protein signature in the cross-platform validation phase 2 dataset. The performance was calculated ROC models either using original weights from cross-platform validation phase 1 ("Original model", weights shown in Table 8), re-estimated model weights using cross-platform validation phase 2 data ("Retrained model", weights shown in Table 8) or with the simple DRS. CI = confidence interval; DRS = disease risk score; pAUC = partial AUC. 99
Table 10: Information about the datasets used in chapter 4. KD = Kawasaki disease; DB = definite bacterial; DV = definite viral; HC = healthy control; LC–MS/MS = liquid chromatography with tandem mass spectrometry; * = not used in analysis. 110
Table 11: Clinical and demographic patient information for patients included in each dataset used in chapter 4 split by disease group. Ages are in months with age range in brackets. 111

Table 12: Clinical information for the patients with Kawasaki disease whose samples were included in any of the four datasets used in the analysis in chapter 4. IQR = interquartile range. 111
Table 13: The features selected by the LASSO models for the DB and DV classifiers for the transcriptomic and proteomic discovery datasets. Feature names in italics are those that increase in the classifier group of interest (i.e., bacterial in the DB classifiers), and those in bold decrease in the classifier group of interest (i.e., viral and HC in the DB classifiers). DB = definite bacterial; DV = definite viral.....127
Table 14: Demographic and clinical characteristics of the patients with MIS-C, KD, bacterial and viral infections used in the case-control discovery RNA-Seq dataset. This dataset was used to identify potential protein biomarkers for distinguishing between bacterial and viral infections in febrile children. Values are median (IQR) unless stated. IQR: interquartile range; KD = Kawasaki disease; PICU = paediatric intensive care unit; RSV = respiratory syncytial virus; CRP = C-reactive protein.....139
Table 15: Demographic and clinical characteristics of the patients with MIS-C, KD, bacterial and viral infections, and COVID-19 used in the RT-qPCR validation dataset. This dataset was used to evaluate the performance of the biomarkers identified from the RNA-Seq discovery dataset. Values are median (IQR) unless stated. IQR = interquartile range; KD = Kawasaki disease; CRP = C-reactive protein; RSV = respiratory syncytial virus; PICU = paediatric intensive care unit; COVID-19 = Coronavirus Disease 2019; SARS-CoV-2 = severe acute respiratory syndrome coronavirus 2. 140
Table 16: The top SDE genes between MIS-C vs. KD, viral and bacterial combined that were taken forward to RT-qPCR validation. The genes were identified by DESeq2 on the RNA-Seq discovery dataset. BH = Benjamini-Hochberg; LFC = log-fold change. 150
Table 17: The exons selected for each of the candidate biomarker genes identified from the RNA-Seq discovery dataset. Metrics used to select the exons are shown in the table. Mean exon counts are shown. CI = confidence interval. MIS-C = multisystem inflammatory syndrome in children. ..151
Table 18: Primer sets developed in this study. These primers were used for quantification of the genes selected for validation using RT-qPCR. Primers were developed for the optimal assays identified in the optimisation stage. GAPDH was used as the reference gene. 153
Table 19: The performance of the new signatures composed of different combinations of the 6 SDE genes in addition to the 4-gene signature. Only signatures that led to significant BH-adjusted roc.test p-values are shown. Total change of performance is the improvement in AUC for MIS-C vs. KD, viral and bacterial, MIS-C vs. KD, MIS-C vs. viral, and MIS-C vs. bacterial. roc.test p-values are BH-adjusted. SIS = signature improvement score (calculated by total change in performance/number of genes added). BH = Benjamini-Hochberg; MIS-C = multisystem inflammatory syndrome in children; KD = Kawasaki disease. 157

Abbreviations

Abbreviation	Full name
A2GL	leucine-rich alpha-2-glycoprotein
AACT	alpha 1-antichymotrypsin
AFM	afamin
AHA	American Heart Association
AMR	antimicrobial resistance
ANGPT2	angiopoetin 2
APC	antigen presenting cell
APOC1	apolipoprotein C1
APOH	apolipoprotein H
AT3	anti-thrombin 3
AUC	area under the curve
BCA	bicinchoninic acid
BH	Benjamini-Hochberg
CAA	coronary artery aneurysm
CD163	scavenger receptor cysteine-rich type 1 protein M130
CD25	interleukin-2 receptor alpha chain
cDNA	complementary DNA
CERU	ceruloplasmin
CHKA	choline kinase alpha
CI	confidence interval
CLUS	clusterin
CNTN5	Contactin-5
CO7	complement component C7
COCONUT	combat co-normalization using controls
COVID-19	Coronavirus Disease 2019
CPM	counts per million
CRP	C-reactive protein
CSF	cerebrospinal fluid
CT	cycle threshold
CX3CL1	fractalkine
DAMP	damage-associated molecular pattern
DARC	Duffy antigen gene
DB	definite bacterial
DC	dendritic cells
DIAMONDS	Diagnosis and Management of Febrile Illness using RNA Personalised Molecular Signature
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
DOC	deoxycholate

DRS	disease risk score
DV	definite viral
EB	empirical Bayes
ELISA	enzyme-linked immunosorbent assay
ESI	electrospray ionisation
EU	European Union
EUCLIDS	European Union Childhood Life-Threatening Infectious Disease Study
FA5	coagulation factor V
FBLN3	EGF-containing fibulin-like extracellular matrix protein 1
FDR	false discovery rate
FETUA	alpha-2-HS-glycoprotein
FOSB	FosB proto-oncogene
FPKM	fragments per kilobase per million fragments mapped
FS	forward selection
FS-PLS	forward selection-partial least squares
GAPDH	glyceraldehyde 3-phosphate dehydrogenase
GC	guanine cytosine
GCSF	granulocyte colony-stimulating factor
GEO	gene expression omnibus
GLM	generalised linear model
GO	gene ontology
HC	healthy control
HCG	human chorionic gonadotropin
HIV-1	human immunodeficiency virus type 1
HNE-α1AT	human neutrophil elastase-α1-antitrypsin
HRG	histidine-rich glycoprotein
HROB	homologous recombination factor with OB-fold
HSPBAP1	HSPB1 Associated Protein 1
IFC	integrated fluidic circuit
IFN-γ	interferon gamma
IGFBP3	insulin-like growth factor-binding protein 3
IL1	interleukin 1
IL18	interleukin 18
IP10	interferon gamma-induced protein 10
IPSP	plasma serine protease inhibitor
IQR	interquartile range
ISG15	interferon-stimulating gene 15
KAIN	kallistatin
KD	Kawasaki disease
LASSO	least absolute shrinkage and selection operator
LBP	liposaccharide binding protein
LC	liquid chromatography

LC-MS	liquid chromatography-mass spectrometry
LC-MS/MS	liquid chromatography-tandem mass spectrometry
LFC	log2 fold-change
LFQ	label free quantification
LG3BP	galectin-3-binding protein
LPS	lipopolysaccharide
m/z	mass to charge ratio
MALDI	matrix-assisted laser desorption/ionisation
MASP1	mannan-binding lectin serine protease 1
MASP2	mannan-binding lectin serine protease 1
MHC	major histocompatibility complex
MIS-C	multisystem inflammatory syndrome in children
MLE	maximum likelihood estimation
MPIF1	C-C motif chemokine 23
MPO	myeloperoxidase
mRNA	messenger RNA
MS/MS	tandem mass spectrometry
MS1	first mass spectrometer
MS2	second mass spectrometer
MSD	Meso Scale Discovery
MX2	MX dynamin like GTPase 2
NB	Northern blot
NCAM1	neural cell adhesion molecule 1
NCK2	NCK adaptor protein 2
ncRNA	non-coding RNA
NGAL	neutrophil gelatinase-associated lipocalin
NGS	next-generation sequencing
NK	natural killer
NRP1	neuropilin-1
NSDB	non-sterile definite bacterial
PAMP	pathogen-associated molecular pattern
pAUC	partial AUC
PC	principal component
PCA	principal component analysis
PCR	polymerase chain reaction
PCT	procalcitonin
PER1	period circadian regulator 1
PERFORM	Personalised Risk assessment in febrile illness to optimise Real-life Management across the European Union
PICU	paediatric intensive care unit
PIMS-TS	paediatric inflammatory multisystem syndrome temporally associated with SARS-CoV-2

PLG	plasminogen
PLS	partial least squares
PRR	pattern recognition receptor
PRTN3	proteinase 3
PTM	post-translational modification
QC	quality control
qPCR	quantitative polymerase chain reaction
RBP4	retinol binding protein 4
RIN	RNA integrity number
RNA	ribonucleic acid
RNA-Seq	RNA Sequencing
ROC	receiver operating characteristic
RPKM	reads per kilobase per million reads mapped
rRNA	ribosomal RNA
RSN	robust spline normalisation
RSV	respiratory syncytial virus
RT	reverse transcription
RT-qPCR	reverse transcription-quantitative polymerase chain reaction
S100A12	S100 calcium binding protein A12
S100A8	S100 calcium binding protein A8
SAA1	serum amyloid A1
SAA2	serum amyloid A2
SARS-CoV-2	severe acute respiratory syndrome coronavirus 2
SBS	sequencing by synthesis
SDA	significantly differentially abundant
SDE	significantly differentially expressed
SDS	sodium dodecyl sulphate
SDS-PAGE	sodium dodecyl sulphate–polyacrylamide gel electrophoresis
SELE	E-Selectin
SELEX	systematic evolution of ligands by exponential enrichment
SIS	signature improvement score
SNP	single nucleotide polymorphism
SOMAmer	slow off-rate modified aptamer
SPP24	secreted phosphoprotein 24
ST2	interleukin-1 receptor-like 1
STAR	spliced transcripts alignment to a reference
SVD	singular value decomposition
TCA	trichloroacetic acid
TCR	T cell receptor
TGFB1	transforming growth factor beta 1
TIMP1	TIMP metallopeptidase inhibitor 1

Tm	melting temperature
TMPRSS2	transmembrane protease, serine 2
TNF sR-I	tumour necrosis factor receptor superfamily member 1A
TNF-α	tumour necrosis factor-α
TOF	time of flight
TPM	transcripts per million
TRAIL	TNF-related apoptosis-inducing ligand
TRBV11-2	T cell receptor beta variable 11-2
tRNA	transfer RNA
TSS	toxic shock syndrome
ULBP3	UL16 binding protein 3
VPS37C	vacuolar protein sorting-associated protein 37C
WB	Western blot
Zmax	maximum coronary artery z score
ZPI	protein Z-dependent protease inhibitor

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Publications

Publications arising from the work described in this thesis at the time of submission:

1. **Jackson, H.**, Menikou, S., Hamilton, S., McArdle, A., Shimizu, C., Galassini, R., Huang, H., Kim, J., Tremoulet, A., Thorne, A., Fischer, R., de Jonge, M. I., Kuijpers, T., Wright, V. J, Burns, J. C., Casals-Pascual, C., Herberg, J., Levin, M., Kaforou, M. (2021) Kawasaki Disease Patient Stratification and Pathway Analysis Based on Host Transcriptomic and Proteomic Profiles, *International Journal of Molecular Science,* **22**(11).

Publications arising from the work that is not described in this thesis but was performed in the duration of this PhD:

1. **Jackson, H.**, Calle, I. R., Broderick, C., Habgood-Coote, D., D'Souza, G., Nichols, S., Gómez-Rial, J., Rivero-Valesco, C., Rodríguez-Núñez, N., Barbeito-Castiñeiras, G., Pérez-Freixo, H., Acosta, M. B-D., Cunnington, A. J., Herberg, J., Wright, V. J., Gómez-Carballa, A., Salas, A., Levin, M., Martinon-Torres, F., Kaforou, M. (2022) Characterisation of the blood RNA host response underpinning severity in COVID-19 patients, *Scientific Reports*, **12**(12216).
2. Mehta, R., Chekmeneva, E., **Jackson H.**, Sands, C., Mills, E., Arancon, D., Li, H. K., Arkell, P., Rawson, T. M., Hammond, R., Amran, M., Haber, A., Cooke, G. S., Noursadeghi, M., Kaforou, M., Lewis, M. R., Takats, Z., Sriskandan, S. (2022) Antiviral metabolite 3'-deoxy-3',4'-didehydro-cytidine is detectable in serum and identifies acute viral infections including COVID-19, *Med*, **3**(3), 204-215.
3. Wang, X., Nijman, R., Camuzeaux, S., Sands, C., **Jackson, H.**, Kaforou, M., Emonts, M., Herberg, J., Maconochie, I., Carrol, E. D., Paulus, S. C., Zenz, W., Van der Flier, M., de Groot, R., Martinon-Torres, F., Schlapbach, L. J., Pollard, A. J., Fink, C., Kuijpers, T., Anderson, S., Lewis, M. R., Levin, M., McClure, M. (2019) Plasma lipid profiles discriminate bacterial from viral infections in febrile children, *Scientific Reports*, **9**(17714).

Articles that have been submitted for publication that have arisen from the work described in this thesis at the time of submission:

1. **Jackson, H.,** Miglietta, L., Habgood-Coote, D., D'Souza, G., Shah, P., Nichols, S., Vito, O., Powell, O., Davidson, M., Shimizu, C., Agyeman, P., Beudeker, C., Brengel-Pesce, K., Carrol, E., Carter, M., De, T., Eleftheriou, I., Emonts, M., Epalza, C., Georgiou, P., de Groot, R., Fidler, K., Fink, C., van Keulen, D., Kuijpers, T., Moll, H., Papaptheodorou, I., Paulus S., Pokorn, M., Pollard, A., Calle, I. R, Rojo, P., Secka, F., Schlapbach, L., Tremoulet, A., Tsolia, M., Usuf, E., van der Flier, M., von Both, U., Vermont, C., Yeung, S., Zavadska, D., Zenz, W., Coin, L., Cunnington, A., Burns, J. C., Wright, V., Martinon-Torres, F., Herberg, J., Rodriguez-Manzano, J., Kaforou, M., Levin, M. (2022) Diagnosis of multisystem inflammatory syndrome in children

by a whole-blood transcriptional signature. ***Under review*** *at the Journal of the American Medical Association (JAMA).*

1 Introduction

1.1 Host response to infection

Humans encounter countless microbes on a daily basis, including those in the surrounding environment and those living on and inside of us. Some microbes can lead to disease, and these are referred to as pathogens. Humans have developed various internal and external mechanisms to protect themselves against pathogens. The aim of external mechanisms is to prevent pathogens from entering the human body. If external barriers are breached and a pathogen can enter the organism's body, the immune system is responsible for eliminating the pathogen and preventing it from establishing disease. The human immune system, which is a complex network of biological processes that protect people from disease, is composed of two major sub-systems: the innate and adaptive (or acquired) immune systems. The innate and adaptive immune systems undertake distinct tasks that contribute towards protecting the human body against disease. The innate immune system is sometimes referred to as the "non-specific" immune system, whilst the adaptive immune system is referred to as the "specific" immune system. Despite their differences, the innate and adaptive immune systems are highly interlinked and act in coordination with one another.

The innate immune response acts rapidly following the identification of a foreign organism in the body, using defensive mechanisms that are present from birth and are encoded in the host's germline. There are multiple components of the innate immune system. First, there are physical barriers such as the skin and epithelial layers, alongside secreted mucus layers that coat various cavities in the body and trap pathogens, preventing them from entering the body. Many types of white blood cells, or leukocytes, are involved in the innate immune system (Figure 1), including granulocytes, monocytes, macrophages, dendritic cells, and natural killer (NK) cells. Some innate immune cells, notably dendritic cells (DC) and macrophages, express pattern recognition receptors (PRRs). PRRs can detect pathogens including bacteria, virus, and fungi, through recognising conserved molecular structures of pathogens, also known as pathogen-associated molecular patterns (PAMPs), or through recognising molecules released by damaged cells (damage-associated molecular patterns, DAMPs). Examples of PAMPs include bacterial lipopolysaccharide, bacterial flagellin, and nucleic acids. Detection of PAMPs or DAMPs leads to cellular innate immune responses including inflammation, phagocytosis (ingestion of foreign substances by cells), and cytotoxicity (elimination of infected cells). Soluble proteins and small molecules form key components of innate immunity, including complement proteins, cytokines, and chemokines, which mediate general inflammatory immune responses, and attract and regulate immune system cells [1]. Whilst the innate immune system can recognise pathogens through PRRs, this recognition and the downstream immune response it triggers is generic, which may be inadequate for

clearance of some pathogens. For many pathogens, a more specific immune response is needed, and for this, the adaptive immune system is required.

The adaptive immune system is activated through the innate immune system. The adaptive immune system is much slower acting than the innate immune system, taking between 4-7 days to mount a response against a pathogen. The response mounted is specific to the target antigens detected. An antigen is a molecule that either originates from within the body (self) or external to the body (non-self, i.e., a pathogen) that can stimulate an immune response. It is essential that the immune system can correctly differentiate between self and non-self to avoid the destruction of host molecules and harmless or potentially beneficial organisms or particles by the adaptive immune system. The innate immune system ensures that the adaptive immune system is only recruited in instances where harmful non-self-molecules are detected.

Following recognition of harmful non-self-antigens, the adaptive immune responses are mediated by immune cells called B and T lymphocytes (Figure 1) which lead to antibody and cell-mediated immune responses, respectively. B lymphocytes secrete antibodies (or immunoglobulins) which are Y-shaped proteins with regions called antigen-binding sites that recognise and bind to a specific antigens, such as viral surface molecules. Following secretion from B lymphocytes, antibodies circulate in the bloodstream and bind specifically to their complementary non-self-antigen, either directly neutralising the associated pathogen, or targeting the pathogen or infected cells for destruction by other components of the immune response [2]. Foreign antigens are presented on the surface of infected host cells and are recognised by specific T cell receptors (TCR). In addition to the TCR, T lymphocytes express either CD4 or CD8 receptors (either CD4+ T lymphocytes, or CD8+ T lymphocytes, respectively) and function to activate other immune cells including T and B cells, remove pathogens and infected host cells, and regulate the immune response.

One of the most important functions of the adaptive response is its ability to develop immunological memory. In some instances, once we gain adaptive immunity to a pathogen either through infection or vaccination, the immunity lasts a lifetime. For some other pathogens however, such as influenza or severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), immunity is not lifelong and we can become re-infected, although the severity of infection is sometimes reduced. Immunological memory is highly advantageous as it not only leads to a much faster immune response against pathogens, but the immune response mounted is far more effective, which can prevent infection transmission, reduce severity of infection, and sometimes eliminate disease manifestation completely.

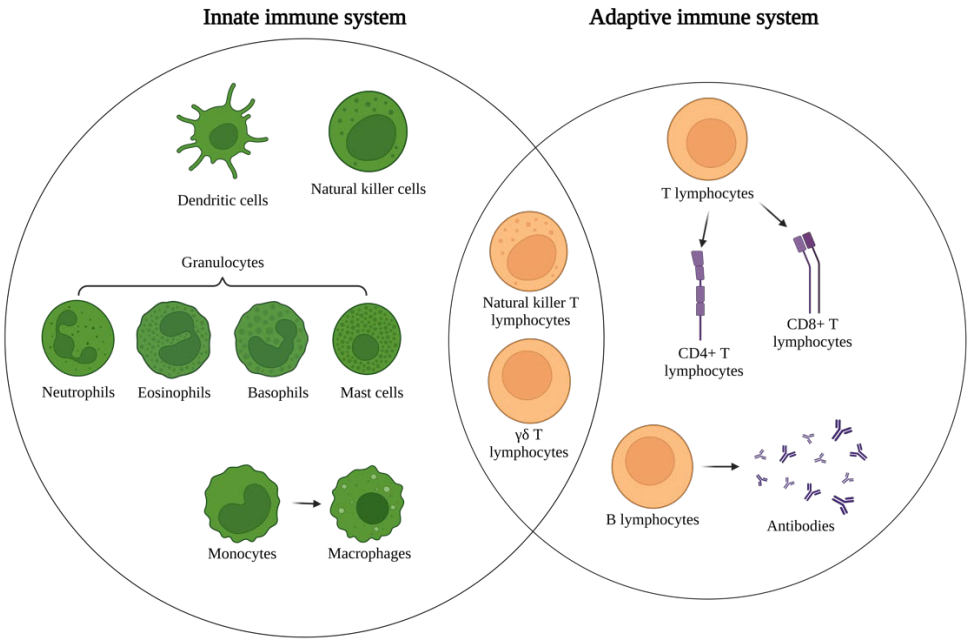


Figure 1: *Cells of the innate and adaptive immune systems. Image created with BioRender.*

The host-pathogen interaction refers to how disease-causing microbes survive and proliferate within their host. Pathogens and their hosts are very closely intertwined, and after millions of years of co-existing, they have very interrelated evolutionary relationships. Pathogens must overcome host immune defences to establish and propagate themselves and, as such, they have evolved many complex immunomodulatory mechanisms that enable them to escape the host's innate and adaptive immune responses [3]. Pathogens are equally a major source of evolutionary pressure on host populations, including humans. Humans have adapted through evolution; gaining or losing genetic variants that reduce or increase their susceptibility to certain pathogens, respectively. For example, the malarial parasite *Plasmodium vivax* uses the Duffy antigen gene (DARC, or FY) to enter red blood cells, and a mutation that disrupts DARC gene expression has reached fixation in much of sub-Saharan Africa. As a result, *P. vivax* infection is rare in sub-Saharan Africa despite favourable environmental conditions for the parasite's survival [4, 5]. This is an example of hosts evolving to avoid infection.

The details of the host-pathogen relationship differ between pathogens. Individual pathogens use specific mechanisms to establish and maintain infection, which in turn affect the host response to infection with a given pathogen. For example, viruses depend on host cells for survival and propagation of their genetic material. Viruses invade host cells and hijack their normal cellular processes. The cells of the immune system could struggle to identify the viral particles (virions) which hide within host cells; however over time, a system has evolved in which infected cells display what is inside of them on their surface molecules. Class I major histocompatibility complex (MHC) proteins on the surface of cells display fragments of what is found within the cell, and if the cell is infected with a virus, this will include fragments of viral proteins or antigens. Cells that display antigenic fragments

on their surfaces are known as antigen presenting cells (APC). Cytotoxic T-cells scan the antigens presented by MHC, and if the antigens presented are complementary to the cytotoxic T cells' TCRs, they will respond by releasing cytotoxic factors to kill the infected cell. Furthermore, virally infected cells will release interferons which prevent replication of viruses and warn other proximal cells of the viruses' presence.

Bacteria use different strategies to viruses to establish infection and propagate themselves and as such, the host immune response varies to reflect this. Many types of bacteria do not depend on host cells for their survival, meaning that they can be found extracellularly or circulating in blood and tissue fluids. The complement system, which directly kills bacteria, is essential for the rapid clearance of bacterial pathogens. Many proteins of the complement system are found in the blood and tissue fluids and are activated following recognition of microbial components. The extracellular localisation of the complement proteins ensures that bacteria can be rapidly removed. Some bacteria have overcome these host defences and are intracellular, such as *Chlamydia* and *Rickettsia.* Many intracellular bacteria have evaded the host immune response by colonising non-professional APCs, or by dwelling in APCs but interfering with the process of antigen presentation [6].

Widespread changes in gene expression and protein production occur following activation of the immune system because of pathogen colonisation. This is a result of changes in production of certain proteins and small molecules, such as chemokines, cytokines and complement proteins, expansion of immune cell populations, and overall regulation of the immune response. The exact immune response elicited by the human host differs according to the type of pathogen. Consequently, the changes in gene expression and protein production that occur following infection will also vary based on the type of pathogen. Host molecules such as genes or proteins that are involved in the immune response can be quantified and contrasted between individuals with certain infectious or inflammatory diseases. This can reveal differences between healthy and diseased states, or between one disease and another.

1.2 Transcription and translation

DNA (deoxyribonucleic acid) is the macromolecule that contains genetic instructions for the development, functioning, growth, and reproduction of all organisms. DNA is composed of two polynucleotide strands that coil around each other to form a double helix. The polynucleotide chains are composed of a sequence of nucleotide bases which are made up of one of four nucleobases (cytosine [C], guanine [G], adenine [A], thymine [T]), a deoxyribose sugar, and a phosphate group. The nucleotides on each strand are joined together via bonds between the deoxyribose sugar and phosphate groups of neighbouring nucleotides. The two strands are joined via covalent bonds between complementary bases (thymine and adenine; guanine and cytosine). The sequence of the four nucleotides encodes genetic information about functional products. DNA must undergo transcription

and translation for cells to produce functional products which are RNA (ribonucleic acid) and proteins, respectively. Although all cells present in an organism contain the full set of DNA, not all products are required in all cells at every given moment. Therefore, genes are only expressed and proteins are only produced if the cell requires the functional products encoded in the corresponding DNA section. Transcription and translation are the two processes used by cells to convert genes encoded in DNA into functional proteins (Figure 2). DNA must first be converted to RNA via transcription, which is then converted into proteins vial translation; a process known as the "central dogma of molecular biology".

Transcription is the first step in this process. In transcription, a section of DNA, i.e., a gene, is copied to produce RNA. RNA is a macromolecule made up of nucleotides. Unlike DNA, RNA is generally single-stranded and contains the uracil (U) base instead of thymine and the sugar ribose instead of deoxyribose. RNA polymerase is the enzyme that is responsible for the synthesis of RNA from DNA. Once the RNA product has been synthesised, it undergoes polyadenylation (the addition of a poly-A tail) and splicing, in which non-protein-coding sections (known as introns) of the RNA sequence are removed, leaving just the protein-coding regions (exons). The final product is known as messenger RNA (mRNA). Transcription is tightly regulated through transcription factors which either activate the transcription of a gene by recruiting RNA polymerase or suppress the transcription of a gene by preventing RNA polymerase from binding to the promoter region. Transcription factors ensure that genes are only transcribed if their products are required.

mRNA is then decoded in structures called ribosomes to produce a specific amino acid chain. Amino acids are organic compounds that join to form polypeptide chains which ultimately form proteins. Trios of nucleotides in mRNA are known as codons, and codons correspond to certain amino acids. Transfer RNAs (tRNA) are essential for translation. On one end of each tRNA is an anticodon, a set of 3 nucleotides complementary to a given mRNA codon, and on the other end there is the specific amino acid that corresponds to the codon. Within the ribosome, tRNAs bind to their complementary codons and the corresponding amino acids are linked together to form a growing polypeptide chain. A start and stop codon are present at the beginning and end of the mRNA sequence, respectively. There are regulatory mechanisms in place at all stages of translation to ensure that protein products are only produced when necessary.

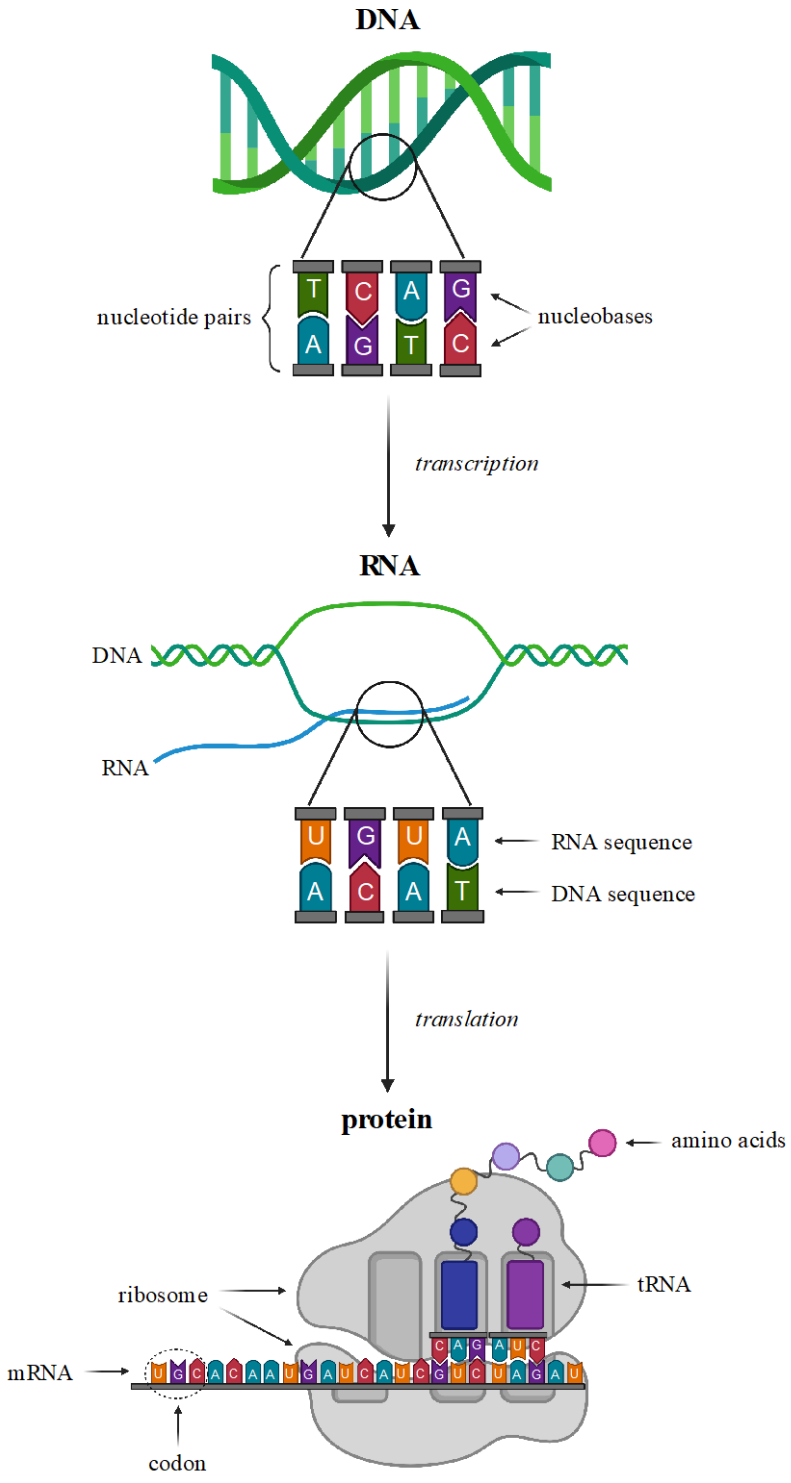


Figure 2: *Overview of the central dogma of molecular biology, i.e., the conversion of DNA to RNA to protein. DNA = deoxyribonucleic acid; RNA = ribonucleic acid; tRNA = transfer RNA; mRNA = messenger RNA. Image created with BioRender.*

1.3 An overview of omic data and its applications

The "omic" fields of biology are disciplines of science with the suffix "-omic" at the end of their name. "-Omic" refers to the complete characterisation of a set of molecules. Genomics, transcriptomics,

proteomics, and metabolomics are all examples of omic fields of study and refer to the universal detection of all DNA (the genome), all RNA (the transcriptome), all proteins (the proteome), and all metabolites (the metabolome), respectively, in a given sample (Figure 3). Recent technological advances and rapid decreases in the cost of sequencing and high-throughput technologies have made it considerably easier for researchers to generate and study omic datasets.

Traditionally, molecules or sets of molecules are measured with the aim of validating or disproving a given hypothesis. This approach to experimental biology is "hypothesis-driven". Contrastingly, omic approaches are "hypothesis-generating" or "hypothesis-free" meaning that all possible data is utilised to explore all possibilities and ultimately define a hypothesis that can be further tested [7]. The basic premise of omic approaches is that a biological system can be more thoroughly understood when it is considered as a whole, ensuring that potentially important elements of a system are not excluded. For example, genomics attempts to characterise all genetic variants in an individual (i.e., a genome) which differs from traditional genetics that focuses on fewer variants, for example those selected due to a certain hypothesis [8]. Omic technologies have been pivotal for the progression of the field of systems biology. Systems biology aims for systems-level understanding of biological systems rather than exploring a system's individual components (e.g., individual genes, proteins, pathways) [9].

The remarkable progression in sequencing technologies means that there are now various platforms available for sequencing omic datasets on each omic level. Furthermore, huge advances in computing performance and computational hardware and software have enabled the development of countless analytical methods, dramatically increasing the efficiency and scope of applications of omic data. Whilst all types of omic data have extensive applications, this thesis will focus primarily on transcriptomics and proteomics and their applications.

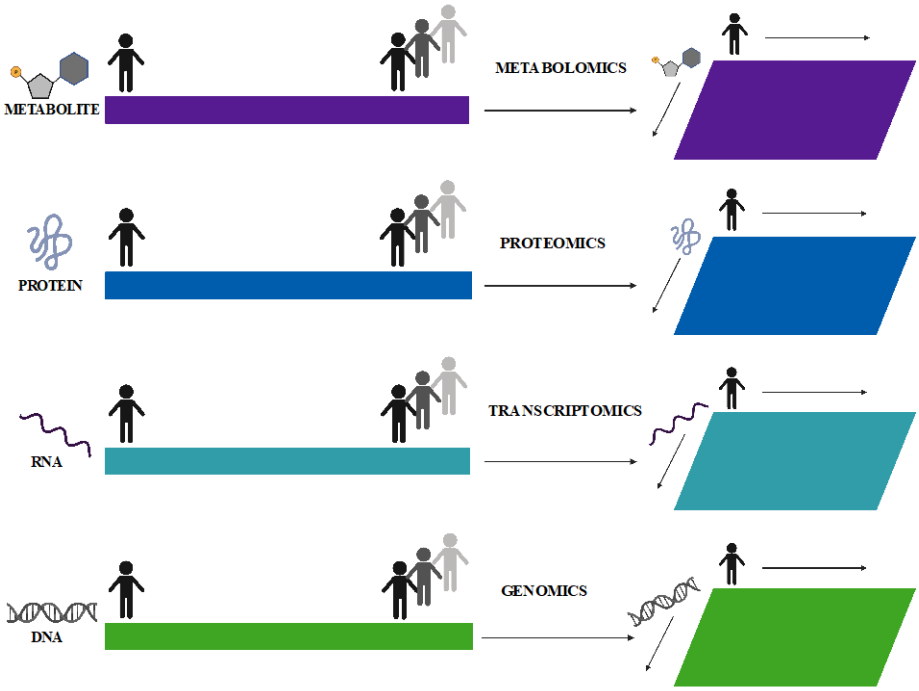


Figure 3: *Four of the principal levels of biological organisation and their corresponding omic level. DNA = deoxyribonucleic acid; RNA = ribonucleic acid. Image created with BioRender.*

1.3.1 Transcriptomics

1.3.1.1 Overview

Transcriptomics is the study of the transcriptome, which is the full set of RNA transcripts expressed in a sample. The transcriptome can be characterised for a specific cell type, tissue, or organ. Transcriptomics aims to detect and quantify all transcripts present in a sample. The exact set of genes expressed in an individual varies between each cell, tissue, and organ. Furthermore, the individual's transcriptome will react to external stimuli and their physiological condition. Transcriptomics provides a snapshot of the genes expressed in an individual and is a powerful approach for exploring how gene expression varies according to internal or external conditions. Whilst proteins are translated from RNA transcripts, not all RNA transcripts are translated into proteins. As such, transcriptome profiling facilitates the detection of non-coding RNAs (ncRNAs), portions of the genome that are transcribed but not translated [10]. Furthermore, transcriptomics can reveal the start sites, 5' and 3' ends, splicing patterns, and post-transcriptional modifications of genes; information that is important for understanding transcriptional and translational dynamics in healthy and diseased states [11].

Many methods have been developed for characterising and quantifying RNA transcripts. These methods can be broadly categorised into approaches for characterising a small number of genes (low-throughput), or a large number of genes (high-throughput). The two key transcriptomic methods are

microarray and RNA Sequencing (RNA-Seq). These are high-throughput sequencing methods that can handle many samples simultaneously whilst also quantifying multiple RNA transcripts in parallel.

1.3.1.2 Low-throughput RNA quantification

1.3.1.2.1 Northern blot

Northern blot (NB) is a low-throughput method for RNA separation [12]. RNA is separated according to size on a denaturing gel. Once separated, the RNA is transferred to a membrane, and a reporter probe that is complementary to the gene of interest is hybridised to the bound RNA. The abundance of the bound RNA of interest can then be analysed, in addition to the size which can offer insights into splicing variants. In instances where the RNA transcript of interest is present at low levels, or if the amount of RNA sample available is limited, a quantification method such as a polymerase chain reaction (PCR)-based method may be more appropriate.

1.3.1.2.2 Reverse transcription-quantitative polymerase chain reaction (RT-qPCR)

PCR is a fundamental tool in biological research and medicine in which a specific DNA fragment is amplified from a complex DNA sample [13]. PCR takes place across 20-40 rounds of repeated temperature changes (thermal cycles), with three key steps performed within each round: denaturing, annealing, and extension. Denaturing involves separation of the target double-stranded DNA into two single-stranded DNA molecules. Annealing then occurs, in which two short DNA molecules that are complementary to the DNA of interest (known as primers) join to the 3′ end of each single-stranded DNA molecule at the complementary location. Hydrogen bonds are formed between base pairs. Annealing is followed by extension which involves the enzyme DNA polymerase and free nucleotides. DNA polymerase synthesises a new DNA strand which is an extension of the primer, adding complementary nucleotides to the target DNA sequence. At the end of each cycle, the DNA target has been amplified and the process begins again, this time with the amplified product also undergoing denaturing, annealing and extension.

Quantitative PCR (qPCR) couples amplification of a DNA target via PCR with "real-time" measurement of the concentration of the target sequence [14, 15]. End-point measurement of the target DNA can be misleading as the reaction will almost certainly reach a plateau in which the amplification of the target DNA is decreased due to depletion of the reaction mixture (i.e., free nucleotides). For target sequences present in high concentrations in the sample, this plateau will occur at an earlier cycle of the PCR experiment than for target sequences present in low concentrations. For both sequences, the amount of target DNA amplified at the end of the cycles will be the same despite the differences in their starting concentrations [14]. Measuring target DNA accumulation in real-time as the reaction progresses – as is performed in qPCR - is a much more sensitive and accurate approach to quantification

of starting target sequence concentration. qPCR uses fluorescent reporter molecules whose levels increase with accumulating amounts of product DNA. The fluorescence intensity is measured across cycles, and the cycle (known as the cycle threshold, CT) in which the fluorescence intensity is greater than the background fluorescence is used as a measure of the target DNA's starting concentration [16].

Whilst qPCR was developed for DNA quantification, it has become a routine technique for measuring RNA transcripts, or gene expression [15]. Prior to using qPCR for RNA quantification, DNA must be synthesised from the target RNA to form complementary DNA (cDNA). cDNA synthesis is performed via reverse transcription (RT) using enzyme reverse transcriptases [15, 16].

The process of combining qPCR with reverse transcription is known as reverse transcription qPCR (RT-qPCR). RT-qPCR can either be performed in a one-step or two-step assay. In a one-step assay, reverse transcription and qPCR are performed in the same tube with all reagents added simultaneously (i.e., reverse transcriptase alongside DNA polymerase). Whilst this approach is faster and leads to less experimental variation, it can be less sensitive than two-step assays since the reaction conditions are not optimised for each individual step, leading to a compromise between the two reactions [16, 17]. Two-step RT-qPCR assays involve the RT and qPCR stages being carried out in separate tubes as distinct stages. Two-step assays are advantageous as the two distinct experimental stages can be optimised, and the cDNA pool that is generated from the first step can be stored and used for multiple RT-qPCR assays [16]. This enables replicates to be included in the experimental design. Despite these advantages, the risk of contamination can be higher than with one-step RT-qPCR [16].

RT-qPCR is a highly sensitive, accurate and reproducible method which is often used as the gold-standard for validating the findings of high-throughput gene profiling approaches. RT-qPCR is a suitable choice if the number of genes of interest is small because it is a laborious process if the number of target transcripts is high. High-throughput RNA quantification methods including microarray and RNA-Seq are more appropriate choices in these instances.

High-throughput PCR has recently been developed, which enables quantification of multiple gene targets in parallel via PCR. Platforms include the Microfluidic Dynamic Assay™, TaqMan® Open Array® and SmartChip [18]. The SmartChip assay, which is the most recently developed high-throughput PCR method, can quantify over 1000 genes in a single run, however this can only be performed on a single sample at a time. As a result, the main use of these high-throughput PCR methods is for quantification of a reasonably small number of genes.

1.3.1.3 High-throughput RNA quantification

1.3.1.3.1 Nanostring nCounter®

There are multiple approaches for high-throughput RNA quantification, i.e., quantification of many RNA transcripts in parallel. The Nanostring nCounter® system profiles nucleic acid molecules, including RNA [19]. The nCounter® system uses panels of known gene targets, allowing for up to 800 genes to be profiled simultaneously. Pre-formed panels related to certain biological pathways or diseases can be used, or custom panels can be designed. nCounter® uses multiplexed gene probes, with a unique pair of probes (capture and reporter probes) for each gene of interest in the array [19]. The capture probe contains a sequence complementary to the gene of interest coupled to an affinity tag, such as biotin. The reporter probe contains a second sequence complementary to the gene of interest, this time coupled to a colour-coded tag which is made up of a single-stranded DNA molecule annealed to a series of labelled complementary RNA segments. The RNA segments, which are all labelled with a specific fluorophore, are in a linear order which produces a unique colour code for each RNA target [19].

The Nanostring nCounter® platform depends on a pre-selected range of 800 mRNA targets. Therefore, although the nCounter® platform is high-throughput, it is not an unbiased or hypothesis-free approach to gene expression analysis as the gene targets have been chosen due to their biological function or involvement in certain processes or diseases.

1.3.1.3.2 Microarray

DNA microarray and RNA-Seq are the two main hypothesis-free gene expression profiling approaches. The DNA microarray uses the principle that a single-stranded sequence of nucleic acids will form a double stranded hybrid if presented with a complementary sequence, the same guiding principle used in NB and PCR experiments [20]. In the DNA microarray, this principle is expanded with the aim of measuring thousands of RNA targets. The DNA microarray was first developed in 1995 [21], but it rapidly became a fundamental method for gene expression profiling following the completion of the Human Genome Project in 2001 [22, 23]. The Human Genome Project provided the raw material, i.e., knowledge of the DNA sequences that encode genes, for the gene targets to include on the microarray [24].

A DNA microarray is composed of a microscopic slide printed with over 30,000 spots at pre-defined positions [25]. At each spot, a specific nucleic acid sequence is bound to the slide, with the type of nucleic acid determining the class of microarray (oligonucleotide microarrays or cDNA-based microarrays). In oligonucleotide microarrays, the nucleic acid sequences bound to the slide will be single-stranded oligonucleotides (short single strands of synthetic DNA or RNA), whilst in cDNA-

based microarrays, the nucleic acid sequences will be cDNA PCR products [20]. Oligonucleotide arrays are the preferred microarray approach for high-throughput gene expression profiling as they are faster, more specific, and more reproducible than their cDNA counterparts, and as such, all further description of the "microarray" will refer to the oligonucleotide microarray [26].

Each spot on the microarray corresponds to a specific transcript of interest [20]. Millions of copies of oligonucleotide sequences complementary to the target RNA transcript (probes) are found on each spot [27]. cDNA is obtained from the mRNA samples of interest and is fluorescently labelled. Microarray uses either one-colour or two-colour approaches. One-colour approaches hybridise one RNA sample at a time, which they label with a single fluorophore (usually yellow) [28]. In two-colour approaches, two samples are labelled with two different fluorophores (normally red and green) and are hybridised together on the microarray. The two samples should be from different phenotypic groups, enabling comparison, such as disease group *vs.* healthy control [28].

The labelled samples are incubated on the microarray, leading to hybridisation between complementary cDNA fragments and oligonucleotide probes [20]. Following incubation, non-specific hybrids are washed away, and a laser is used to excite the fluorescent dyes, producing light which is detected and quantified by a confocal scanner and specialised software [20]. The exact location and shape of each spot is detected, followed by the total intensities from all oligonucleotide probes in the specific spot. The intensity from the fluorescent dyes will increase with the number of cDNA fragments hybridised to oligonucleotide probes, and this intensity is used to determine the transcript abundance. The output from two-colour microarrays is a ratio which reports the fluorescence intensity following laser excitation relative to the starting fluorescence of the cDNA [20]. For one-colour microarrays, the intensity of the single fluorophore is returned.

The microarray enables transcriptome profiling of many samples simultaneously for tens of thousands of gene targets, resulting in a huge number of data points for a single experiment. As a result, the protocols required for processing, normalising and analysing microarray data are complex and computationally very demanding. Whilst the microarray revolutionised transcriptomic profiling, it relies on probes from known gene targets, meaning that it cannot be used for detection of novel transcripts.

1.3.1.3.3 RNA Sequencing (RNA-Seq)

Next-generation sequencing (NGS), or massively parallel sequencing is an approach to DNA sequencing that uses massively parallel processing. Massively parallel processing is a term used to describe when many computer processers are used simultaneously to carry out a set of computations. Prior to the emergence of NGS it took over a decade to sequence the human genome, and with NGS technology, this time was reduced to a single day [29]. The basic principle of NGS is that millions of

fragments of DNA are sequenced in parallel and mapped to the human reference genome [30]. NGS not only revolutionised genomics, but it also transformed transcriptomics as it led to the development of RNA-Seq. In RNA-Seq, NGS technologies are used to sequence cDNA, enabling the quantification of all transcribed RNAs [11, 31].

The first step of RNA-Seq is to isolate RNA from each sample and determine its quality through calculating the RNA integrity number (RIN) [31]. Library preparation follows, in which the desired RNA molecules are isolated and reverse transcribed into cDNA. At this point, ribosomal RNAs (rRNAs) are removed, or polyadenylated RNAs are selected to enrich the samples for mRNA rather than rRNA. It is important to remove rRNAs as otherwise they will consume most of the sequencing reads, limiting the detection of less abundant reads [31]. A library of cDNA fragments is then generated from the cDNA with sequencing adaptors attached to one end (single-end sequencing), or to both ends (paired-end sequencing) [11]. Sequencing adaptors are short oligonucleotides that are ligated to cDNA fragments. Not only are they fundamental for immobilising the cDNA fragments on the sequencing flow cell, but sequencing adaptors also contain barcoding sequences and forward and reverse primers for paired-end sequencing. Barcoding is used to allow for multiplexing in which the cDNA libraries from multiple samples are pooled. A 6-base pair barcode sequence, unique to each sample in the experiment, is included in the adaptor enabling identification of the sample that each read originates from [30]. The final step of library preparation is amplification using PCR.

The cDNA fragments are then immobilised onto the flow cell surface and sequencing can begin. Sequencing in NGS occurs through the sequencing by synthesis (SBS) technology. SBS uses 3′ blocked fluorescently labelled nucleotides named single labelled deoxynucleoside triphosphates (dNTP). In each sequencing cycle, a single dNTP binds to the cDNA fragment through base pair complementarity. The fluorescent tag on the dNTP indicates which nucleotide has been added to the fragment and the terminator prevents the next base from joining. Once the fluorescence signal has been emitted, the terminator is removed through cleavage and the next sequencing cycle begins, with the dNTP complementary to the next base joining to the cDNA fragment [32]. In paired-end sequencing, this process is repeated twice: once for the forward cDNA strand, and once for the backwards cDNA strand. Once the cDNA fragments have been sequenced base-by-base, the sequences are mapped to a reference genome, and the number of reads mapping to specific regions of the genome (i.e., genes or transcripts) is quantified. If a reference genome is unavailable, transcriptomes can be assembled *de novo*.

The key advantage of RNA-Seq over microarray is that it does not depend upon known nucleic acid sequences. Whilst microarrays use probes that target specific genes, RNA-Seq returns the exact sequence of nucleic acids in the transcriptome. As a result, RNA-Seq can detect novel transcripts, indels (small insertions and deletions), and can provide information about sequence variations such as single nucleotide polymorphisms (SNPs), splicing variations and exon usage [11, 33]. RNA-Seq also offers a

much wider dynamic range than microarray, enabling detection of lower abundance transcripts, in addition to much higher resolution [11]. As with other high-throughput omic technologies, RNA-Seq demands a complicated data analysis workflow, with many steps essential for transforming the output of the sequencer to analysis-ready counts (described in chapter 2 sections 2.2.1 and 2.3.2).

1.3.2 Proteomics

1.3.2.1 Overview

Proteomics is the study of the proteome which is the set of proteins produced in an organism. Much of the functional information about genes is characterised by the proteome, with the proteome often considered as closer to the phenotype than the genome or transcriptome [34]. Protein levels not only depend on gene expression, but also on complex host translational control and regulation, including post-translational modifications (PTMs). PTMs are chemical modifications of proteins following their translation through the addition of modifying groups. Examples of PTMs include phosphorylation (attachment of phosphate group), glycosylation (addition of a glycan group), and ubiquitination (attachment of ubiquitin).

Proteomics quantifies the abundance of proteins in a sample obtained from an individual and provides information about protein structure including PTMs. Not only will the proteome differ extensively between individuals, but it also varies between tissues, organs and cells in an individual's body. Furthermore, the proteome will differ based on the host's surrounding environmental and physiological states, such as whether the host is infected with a pathogen or has a disease. Proteomics aims to capture and characterise all of these changes. There are many different methods for detecting and quantifying the proteins present in a sample. These can be broadly classified into either low- or high-throughput protein detection methods.

1.3.2.2 Low-throughput protein detection methods

Low-throughput protein detection methods are designed to detect and quantify a small number of proteins. By definition, low-throughput methods can handle and process less samples than high-throughput methods. Low-throughput protein detection approaches are used when the number of variables (i.e., proteins) to be detected and quantified is small, for example proteins that are included in a specific hypothesis.

Within the category of low-throughput protein detection methods are singleplex assays that can measure a single protein at a time. Immunoassays are tests that detect the presence of a molecule through use of an antibody. Many protein detection methods are immunoassays, meaning that they

use antibodies that are complementary and highly specific to the target protein or antigen. These types of protein detection methods are known as antibody-based methods.

1.3.2.2.1 Western blot

Western blot (WB) is a singleplex antibody-based protein detection method designed for the quantification of individual proteins. The proteins in the sample are mixed with sodium dodecyl sulphate (SDS) which coats the proteins with a net negative charge and causes them to unfold. The proteins are then separated according to their molecular weight through sodium dodecyl sulfate–polyacrylamide gel electrophoresis (SDS-PAGE). To enable antibody detection, the proteins are transferred onto a blotting membrane and then incubated with the primary antibody which binds to the target protein. Any unbound primary antibody is washed away prior to incubation with the secondary antibody, which binds to the bound primary antibody and is linked to a reporter enzyme that produces colour or light, enabling protein detection and quantification [35].

1.3.2.2.2 Enzyme-linked immunosorbent assay (ELISA)

One of the most popular singleplex tools for antibody-based protein quantification is the enzyme-linked immunosorbent assay (ELISA). ELISAs detect the presence of specific protein targets by using highly specific antibodies directed against the protein. The antibodies are conjugated to a reporter enzyme and added to the sample. If the complementary antigen to the antibodies (i.e., the protein) is present in the sample, the antibodies will bind to the antigen. Unbound antibodies are washed away prior to the addition of the enzyme's substrate to the plate. If binding occurred between the antibodies and the antigen/target protein, there will be a detectable enzymatic reaction which indicates presence of the target protein [36]. ELISAs are popular tools not only due to their high specificity, but also since they can be fully automated and are relatively fast and cheap [37]. Despite this, ELISAs can only quantify a single protein at a time so are only suitable if the number of target proteins is low.

1.3.2.3 High-throughput protein detection methods

ELISAs and WBs are well suited towards measuring individual or a small number of proteins, however they are impractical for measuring many analytes simultaneously [38]. High-throughput protein detection methods are much faster and can handle larger sample numbers and detect multiple proteins in parallel, known as multiplexing. Multiplexing proteins requires a lower sample volume per protein, making it more practical in cases where sample availability is lower, for example in human studies. Furthermore, multiplex protein detection methods allow inferences to be made about how proteins interact with each other since the proteins are measured simultaneously. Overall, they provide a more holistic snapshot [38].

1.3.2.3.1 Meso Scale Discovery

Multiplex immunoassays involve the multiplexing, or combining, of multiple immunoassays. In multiplex immunoassays, a selection of capture antibodies is immobilised on a solid surface (spatially separate assays) or on beads or particles (bead-based immunoassays). In spatially separate assays, the various antibody pairs share the same reaction volume, but the capture antibodies are isolated with each spot on the assay capturing a target protein. Spot coordinates are used to determine which protein is being detected and the entire assay is read simultaneously. The Meso Scale Discovery (MSD) platform is an example of a spatially separate multiplex immunoassay that can detect up to 71 unique proteins [38]. The MSD technology uses electrochemiluminescence which is a type of luminescence generated during electrochemical reactions taking place in solutions. In MSD, capture antibodies are printed on the plate at pre-determined precise positions, a process known as multispot technology. On the bottom of the plate are high-binding carbon electrons. The sample is then incubated on the plate and target proteins bind to their capture antibodies. A detection antibody with an electrochemiluminescent label (SULFO-TAG labels) is added and the carbon electrodes are energised, leading to light emission by the SULFO-TAG labels. The light intensity at each position on the plate is measured, leading to quantification of each of the target proteins. The use of electrochemiluminescence enables MSD to detect proteins with high sensitivity and specificity within a wide dynamic range [38].

1.3.2.3.2 Luminex®

In bead-based immunoassays, capture antibodies are immobilised on individual beads which bind to protein targets. Protein targets are identified through combinations or levels of dyes that correspond to the analyte-specific capture bead [38]. The Luminex® platform is a popular bead-based multiplex immunoassay which can encode up to 100 different protein targets. The Luminex® assay is made up of 100 different beads, each with a specific colour made from using red and infrared fluorophores at distinct ratios. Each bead corresponds to a particular position on the fluorescent spectrum. The beads are then coated with antibodies complementary to target antigens and are incubated with the sample. Following the binding of a target protein to an antibody, secondary antibodies with fluorescent tags are added that also bind to the target protein. Flow cytometry and a pair of lasers are then used, firstly to determine the exact spectral identity of each bead so that the target protein can be identified, and then to determine the quantity of the target protein bound to each bead.

1.3.2.3.3 SomaScan®

The SomaScan® assay is another example of a bead-based multiplexed assay, however it does not use antibodies. The SomaScan® assay uses modified aptamers for detection and quantification of up to 7000 proteins. Aptamers are short single-stranded molecules that bind with high affinity and specificity to target molecules, in a similar way to antibodies. Aptamers are selected *in vitro* using selection

methods such as systematic evolution of ligands by exponential enrichment (SELEX). SELEX is a combinatorial chemistry technique for producing oligonucleotides that bind to target ligands [39]. SomaScan® uses slow off-rate modified aptamers (SOMAmers®) which are DNA-based aptamers with protein-like properties due to the addition of functional groups that mimic amino acid side chains. Furthermore, SOMAmers have "slow off-rates", obtained by selecting aptamers with slow dissociation rates, a process that resolves non-specific aptamer-protein binding [40]. The SomaScan® assay uses three serial dilutions of every sample which leads to the detection of proteins across an 8- to 10-log dynamic range (~100 fM–1 μM), enabling identification of both low and high abundance proteins [38].

The sample is incubated with a mixture of SOMAmers which all contain a fluorescent tag, a biotin, and a photocleavable group. SOMAmers bind to their target proteins and bound SOMAmer-protein complexes are captured on streptavidin beads via SOMAmer-bound biotin with any unbound proteins washed away. Biotin molecules then bind to the bead-bound target proteins. Ultraviolet light causes bound SOMAmer-protein complexes to be released from the beads back into solution when they are then challenged with an anionic competitor, dextran sulphate. As a result of the kinetic challenge by dextran sulphate, non-cognate complexes (i.e., complexes with non-specific binding) dissociate, leaving only cognate complexes (i.e., complexes with specific binding). SOMAmer-protein complexes that survived the kinetic challenge are now captured onto beads via their target protein biotin tag. The SOMAmers that remain are released and quantified using single-stranded DNA probes that are complementary to SOMAmer DNA sequences. The fluorescent tags attached to the SOMAmers are then used for detection and quantification via DNA microarray [40].

1.3.2.3.4 Mass spectrometry-based proteomics

Not all high-throughput protein detection methods rely on antibodies. Mass spectrometry (MS) based proteomic methods are widely used antibody-free protein detection methods. Antibody-free, or untargeted, protein detection methods can be advantageous as they can detect and quantify proteins for which antibodies do not exist. Furthermore, they can elucidate the sequence of a peptide or protein, enabling discovery of novel proteins and shedding light on protein structure.

MS is an analytical technique that measures the mass-to-charge ratio (m/z) of molecules present in a sample. Molecules measured by a mass spectrometer could be peptides, proteins, metabolites, or lipids. Mass spectrometers can not only quantify known compounds, but they can also determine a compound's molecular weight, structural and chemical properties, as well as identify previously unknown molecules. MS-based proteomic approaches can either be "top-down" or "bottom-up", meaning that proteins are either analysed by MS as intact proteins (top-down) or are enzymatically fragmented into peptides prior to MS (bottom-up) [41].

MS is combined with liquid chromatography (LC) to form liquid chromatography-mass spectrometry (LC-MS), a "bottom-up" MS-based proteomic approach [42]. Proteins within a sample solution are first separated using LC, a type of separation method. In LC, a sample is dissolved in a "mobile phase" which then passes through a "stationary phase". The components of the sample will pass through the stationary phase at different rates, leading to separation. Once the protein sample has been separated into peptide fragments, the fragments are introduced into the mass spectrometer for detection and quantification.

Mass spectrometers are made up of three key components: an ion source, a mass analyser, and a detector. Analyte molecules in the sample are converted from the liquid phase to the gas phase by the ion source in a process known as ionisation. Electronspray ionisation [43] (ESI) and matrix-assisted laser desorption/ionisation (MALDI) [44] are two key ionisation techniques that are central to MS proteomics. ESI ionises the analytes out of solution whilst MALDI sublimates and ionises the samples out of a dry crystalline matrix using laser pulses [45].

Once ionised, the molecules are separated by the mass analyser based on their m/z [46]. There are different types of mass analysers that use different approaches for molecule separation, with the most common mass analysers being quadrupoles, time-of-flight (TOF) or orbitrap analysers [46]. Quadrupole mass analysers are composed of four cylindrical rods in a parallel arrangement. An oscillating electrical field is applied to the rods which leads to ion separation. In a TOF mass analyser, ions are accelerated and are separated based on the differences in their velocities and subsequent arrival times at the detector. TOF is paired with MALDI to form the MALDI-TOF MS method. The Orbitrap mass analyser is an ion trap mass analyser made up of an outer barrel-like electrode and a central spindle-like electrode. Ions are separated as they oscillate at different frequencies around the central electrode and between the two outer electrodes. The separated ions are then measured, and their m/z ratios are recorded alongside their relative abundance by the ion detector.

Tandem mass spectrometry (MS/MS) involves performing mass analysis at least twice with two or more mass analysers coupled together. Molecules of a sample are ionised to form precursor ions, which are separated using the first spectrometer (MS1). Precursor ions of particular m/z are fragmented into product or fragment ions which are introduced into the second spectrometer (MS2). MS2 separates the product ions according to their m/z. MS/MS greatly improves specificity compared to single MS, and is often also coupled with LC to form LC-MS/MS.

The data output of LC-MS/MS proteomics is extremely complicated and requires specific analytical pipelines. The untargeted nature of LC-MS/MS proteomics means that all detectable peptides will be characterised, a major benefit of LC-MS/MS over targeted proteomic methods. It is considerably more challenging, however, to perform data analysis of LC-MS/MS data in comparison to antibody-based or targeted high-throughput methods, such as SomaScan®, Luminex® and MSD, since the peptide

fragments must first be identified. Many software tools have been developed for analysis of LC-MS data, including the popular tool MaxQuant [47]. The pipeline for processing and normalising raw LC-MS/MS data is described in chapter 2 in sections 2.2.2 and 2.3.4.

1.4 Use of host omics in paediatric infectious and inflammatory disease

As outlined in chapter 1 section 1.1, the immune response differs according to the type of pathogen that is causing infection, leading to specific changes in levels of immune mediator molecules such as cytokines and immune cell subsets, amongst other changes. Changes in gene expression and protein production arise to ensure that the necessary protein or metabolite products are available for the molecular processes of the immune response. Changes in gene expression and protein production in response to infection can be observed in omic profiles generated from samples obtained from infected individuals. Transcriptomics captures the changes in gene expression, and proteomics captures the changes in protein production. The concept of "host omics", i.e., using omic profiles obtained from the host to gain insights into infectious and inflammatory diseases, has gained popularity as the cost of sequencing has reduced. Host omics offers many opportunities to improve understanding and management of paediatric infectious and inflammatory diseases. In this thesis, host omics generally refers to the use of either transcriptomic or proteomic data.

1.4.1 Host omics for improving understanding of pathogenesis and disease manifestation

Through quantifying all observable components, host omics can capture extensive information about the host and the pathogen. Parallel advances in sequencing and computational technologies have expanded the toolkit of potential data analysis options available to researchers, enabling a spectrum of insights to be gained from omic datasets. Through contrasting omic profiles between groups of individuals, it is possible to see which molecules (i.e., genes, proteins, metabolites) are disrupted in certain diseases compared to other diseases or compared to the healthy state. This knowledge can then be used to infer which biological processes are altered following infection through identifying the biological functions that disrupted features are associated with. Knowing which biological pathways differ between disease groups can be used to decipher information about the host response to infection. For example, Aschenbrenner et al. used whole blood transcriptomes to dissect the host immune response to COVID-19, revealing extensive immune dysregulation in COVID-19 compared to healthy controls [48].

Considerable heterogeneity in disease manifestation is observed between individuals infected with the same pathogen. Often, infection leads to a spectrum of symptoms of varying severity, and some of this can be explained by variations between individuals in their molecular backgrounds. Host omics is a fundamental tool for uncovering an individual's molecular background and can explain the heterogeneity observed between individuals exposed to the same pathogen. In the same paper, Aschenbrenner et al. [48] contrasted transcriptomic profiles from patients with varying levels of severity of COVID-19 and found that severe COVID-19 was characterised by a major upregulation of neutrophil-associated genes and pathways. This information could be used to identify novel targets for treatment interventions for preventing the development of severe disease.

Currently, infection is treated in a "one-size-fits-all" approach, despite the known variations in disease manifestation between individuals. Sub-groups of patients can be identified using information about clinical phenotypes or underlying molecular profiles, a process known as subtyping [49]. Knowledge of patient subtypes can then be used to guide treatment decisions on an individual level. Stratifying patients into disease subtypes will be an essential component of precision medicine, in which patients receive targeted prevention and intervention strategies according to their clinical manifestations and molecular make-up in addition to their environment and lifestyle. Precision medicine has the potential to revolutionise healthcare as it not only considers an individual's current symptoms but also their likelihood of transitioning between disease stages and progressing to severe disease [49].

Host omics can also shed light on pathogenesis and how a pathogen establishes and maintains infection. For example, Neogi et al. [50] used host RNA-Seq and untargeted proteomics to unveil the specific host pathways hijacked by the tick-borne virus that causes Crimean-Congo hemorrhagic fever to replicate in the host cell. When these pathways were blocked *in vitro*, viral reproduction was inhibited, demonstrating the importance of the pathways in viral propagation.

1.4.2 Host diagnostic signatures

Traditionally, infectious diseases are diagnosed through isolation of the causative agent by using a spectrum of laboratory culture-based tests. Whilst these approaches can lead to the successful identification of pathogens in some instances, they have their shortcomings. Firstly, the disease-causing organism may reside in an inaccessible site, such as the lung, or its levels may be too low to detect even though the pathogen is causing disease. Laboratory tests may detect microbes that are not necessarily causing disease, including organisms that the body has come into contact with, but the immune system is actively clearing prior to disease onset. Organisms that the body has already cleared may also be detected, for example some people continue to test positive on SARS-CoV-2 PCR assays weeks or even months following SARS-CoV-2 infection and active coronavirus disease 2019 (COVID-19). In some diseases or syndromes, there are no gold-standard laboratory-based assays for pathogen detection.

Multisystem inflammatory syndrome in children (MIS-C) is one such example, as although it occurs following SARS-CoV-2 infection, detection of SARS-CoV-2 does not mean that an individual has MIS-C. Indeed, in many individuals with MIS-C, SARS-CoV-2 PCR assays will no longer be positive. Similarly, the causative pathogen of Kawasaki disease (KD) is unknown, so laboratory tests for pathogen isolation would not help in the diagnosis of KD.

Host diagnostic signatures approach infectious disease diagnosis from a different angle: the host. Rather than aiming to isolate the disease-causing organism, host diagnostic signatures use host omic profiles to identify host molecules (i.e., genes, proteins, or metabolites) with sufficiently different levels across conditions to enable diagnosis. A diagnostic signature is usually composed of a handful of features, or biomarkers, that are selected through applying machine learning and statistical methods to host omic profiles. Using an easily accessible fluid such as blood, host molecules such as proteins or RNA transcripts included in a diagnostic signature can be quantified leading to accurate diagnosis. This approach has various benefits. Using an easily accessible fluid makes the diagnostic process far less invasive and disruptive especially for children. By measuring levels of circulating molecules, the systemic changes induced by infection are captured. This means that a host molecular diagnostic approach can identify an infection irrespective of site of infection. Equally, by capturing systemic changes induced by active infection, host molecular diagnostic approaches are less likely to return false positive test results, such as when culture-based methods identify microbes that are not causing active disease (e.g., pathogens that have already been cleared). Furthermore, the results from a host molecular diagnostic approach would take far less time to become available than complex culture-based diagnostic methods.

Host diagnostic signatures have already been explored and applied extensively, not only within the field of infectious diseases, but also for cancer and neurodegenerative disease diagnosis and prognosis [51, 52]. The aim of identifying host diagnostic signatures is for them to be implemented into point-of-care tests. Point-of-care tests are rapid tests that are performed at or near the site of the patient and which provide swift diagnoses and enable appropriate treatment to be implemented in a timely manner.

1.4.3 Diagnosing bacterial and viral infections in febrile children

Managing children with fever (febrile children) is a common and important problem facing healthcare providers. The majority of febrile children attending healthcare have self-resolving viral infections; however, a small minority suffer from life-threatening bacterial infections [53]. If left untreated, some bacterial infections can have serious downstream implications as they can lead to the escalation of an exaggerated immune response, sometimes resulting in sepsis. Sepsis causes life-threatening organ dysfunction and can sometimes lead to septic shock, a condition in which an individual's blood

pressure drops dangerously low. Bacterial infections can cause septic shock as toxins produced by some bacteria damage the small blood vessels, causing them to leak fluid into surrounding tissues. This affects the heart's ability to pump blood to organs which leads to a drop in blood pressure and limits the amount of blood reaching vital organs. Sepsis and septic shock are serious medical complications of bacterial infections with potentially fatal consequences. Sepsis and septic shock have considerable mortality, with estimates of 6% and 10% for sepsis and septic shock, respectively, amongst children in European paediatric intensive care units (PICU). For children who survive, approximately 1 in 3 are discharged with a disability, with some requiring amputation or skin grafts [54]. A study carried out in a paediatric emergency department in a UK hospital reported that 41% (n=2,130) of children presenting to the paediatric emergency department (n=18,104) displayed one or more sepsis warning signs [55], highlighting just how common sepsis is in the UK.

Clinical features alone do not reliably distinguish between bacterial and viral infections. Culture-based methods are the gold-standard diagnostic approaches; however, they have various shortcomings. Obtaining results from bacterial cultures can take a long time, and their sensitivity can vary depending on antibiotic administration, stage of disease, and the type or amount of fluid collected [56-59]. C-reactive protein (CRP) and procalcitonin (PCT) are host peptides whose levels are often used as biomarkers of bacterial infections [60]. Despite their frequent use, CRP and PCT are imperfect biomarkers for distinguishing bacterial from viral infections as elevated levels of both biomarkers have not only been observed in the plasma of patients with confirmed bacterial infections but also viral infections, including SARS-CoV-2 and adenovirus [61-65].

Antibiotics are often prescribed unnecessarily to children infected with viruses out of fear of missing a life-threatening bacterial infection [66]. This unnecessary use of antibiotics contributes to the spread of antimicrobial resistance (AMR) and is costly to healthcare providers. Conversely, severe bacterial infections can be missed which can have drastic, and sometimes fatal, consequences. As a result, an alternative approach for diagnosing bacterial and viral infections in febrile children is urgently required.

Diagnostic signatures derived from host omic data are a promising alternative approach for differentiating between bacterial and viral infections in febrile children. Multiple host gene signatures have been identified for differentiating between bacterial and viral infections in children, including signatures containing as few as two genes [67-70]. Host protein biomarkers for diagnosing bacterial and viral infections have also been explored, however currently only one protein signature (i.e., combination of proteins) has been proposed [71]. Oved et al. [72] proposed a 3-protein signature composed of CRP, interferon gamma-induced protein 10 (IP10) and TNF-related apoptosis-inducing ligand (TRAIL) for distinguishing between bacterial and viral infections which has since been developed into a point-of-care diagnostic device, MeMed BV™. Whilst this signature has been further

validated in independent paediatric and adult cohorts [73, 74], the 3-protein signature was identified from a targeted selection of protein biomarker candidates. These biomarker candidates were selected following literature searches and bioinformatic screening exploring genes differentially expressed across different infectious and inflammatory diseases in microarray datasets. As such, the signature discovery process by Oved et al. was not a bias-free approach, rather it was a hypothesis-driven approach to signature identification. At present, there is a gap in the literature of host protein signatures for diagnosing bacterial and viral infections in febrile children that have been identified from high-dimensional proteomic profiles using a hypothesis-free approach. In addition, the protein selection process for the MeMed BV™ was performed across humans and animals of all ages, meaning that there is currently no protein diagnostic signature specifically identified for diagnosing febrile illness in children.

The EU-funded PERFORM (Personalised Risk assessment in febrile illness to optimise Real-life Management across the European Union) study aimed to develop a comprehensive management plan for febrile patients across Europe. One of the major goals of PERFORM was to improve diagnosis of febrile illness by identifying host diagnostic signatures from transcriptomic and proteomic profiles obtained from children with bacterial or viral infections. Chapter 3 of this thesis addresses the question of identifying and validating host protein signatures for diagnosing febrile illness in children.

1.4.4 Improving diagnosis and understanding of Kawasaki Disease

Kawasaki disease (KD) is an acute paediatric inflammatory disorder that occurs most frequently in children under five years old [75]. KD was first described in Japan over 50 years ago by Dr. Tomisaku Kawasaki [76]. KD is the most common cause of acquired heart disease in children in Europe, Japan, and North America as it causes formation of coronary aneurysms (CAAs) in some children if left untreated [77]. KD leads to an intense inflammatory immune response, including changes to the levels of neutrophils, monocytes, lymphocytes and cytokines [78].

The cause of KD is unknown. KD displays epidemicity and seasonality, with higher case numbers observed in winter in the Northern Hemisphere. These characteristics suggest that KD could be caused by an infectious trigger [79, 80]. KD has high incidence amongst Japanese people [79, 80], which leads to speculation that genetic predispositions may elevate an individual's susceptibility to KD. Genetic studies have identified various genes that could be associated with elevated susceptibility to KD, and KD-induced CAAs [81]. Any pathogen suggested as the cause of KD must be able to explain the epidemiological characteristics of the disease, i.e., high rates in Japanese individuals, in addition to the symptoms and complications such as CAA formation.

Various viral infectious triggers have been suggested as the cause of KD, including cytomegalovirus, chikungunya and respiratory syncytial virus (RSV), amongst others. Multiple bacterial infectious

triggers have also been suggested, with a focus on bacteria that produce superantigens [81]. Superantigens are toxins secreted by bacteria that induce strong immune system activation, including a pro-inflammatory cytokine storm. In one study, superantigen genes were detected in 70% of stool samples from patients with KD, which contrasts to 26.9% in a group of healthy controls [82]. Despite the many theories attempting to explain the cause of KD, none have been independently confirmed [83]. The current consensus is that in some children with a genetic predisposition, an infectious trigger manifests in disease, causing KD [79, 81].

Since the causative agent of KD is not known, there is no diagnostic test for KD. Instead, KD is diagnosed based on the presence of a persistent fever in addition to a selection of 4 out of 5 clinical features (polymorphous rash, abnormal enlargement of lymph nodes in the head and neck, conjunctival injection, changes of the lips and oral cavity, and changes in extremities) [84]. Due to the lack of a diagnostic test, KD can be missed or confused with other inflammatory syndromes. An improved understanding of KD, including its trigger and pathogenesis, are required to not only develop an accurate diagnostic test, but also to ensure appropriate treatment and patient management. A 13-transcript host RNA signature for diagnosing KD *vs*. bacterial and viral infections has been proposed by Wright et al. [85], demonstrating that KD has a unique transcriptomic profile. This suggests that further clues could be found within the host molecular profiles of children with KD that could provide information surrounding the cause and pathogenesis of KD. In chapter 4 of this thesis, I contrast the host transcriptomes and proteomes of children with KD to those from children with bacterial and viral infections to determine whether the host response to KD is more similar to the host response to bacterial or viral infections, or indeed neither.

1.4.5 Multisystem inflammatory syndrome in children

Children infected with SARS-CoV-2 are generally asymptomatic or experience mild symptoms of disease. Some children develop a severe hyperinflammatory syndrome 2-6 weeks after SARS-CoV-2 infection called multisystem inflammatory syndrome in children (MIS-C) [86], or paediatric inflammatory multisystem syndrome temporally associated with SARS-CoV-2 (PIMS-TS) [87]. MIS-C emerged around April 2020, with multiple reports of a KD-like inflammatory syndrome occurring in areas experiencing high SARS-CoV-2 rates [86-88].

A systematic review of literature surrounding MIS-C found that the average age of children who developed MIS-C was 9 years old, and that it occurred more frequently in male children than female children. Furthermore, it was observed that MIS-C occurred more in children of Hispanic (33%) and Black or Afrodescendents ethnicity (32%) compared to White ethnicity (19%) [89]. Immune disruption has been observed in patients with MIS-C, with lower-than-normal levels of subsets of B cells and T cells alongside elevated levels of cytokines and neutrophils [90, 91]. The exact reason why some

children develop MIS-C following SARS-CoV-2 infection is unknown, however it is thought to be caused by an excessive immune response to infection with SARS-CoV-2. Autoantibodies, which are antibodies that recognise self-antigen, had been previously suggested as a key element in the pathogenesis of KD [92]. Autoantibodies have since been suggested as a potential factor driving the emergence of MIS-C due to self-reactive antibodies produced in response to SARS-CoV-2 [93, 94].

MIS-C is clinically distinct from paediatric COVID-19, however children with MIS-C display symptoms that are highly overlapping with those displayed by children with KD and bacterial infections, including streptococcal and staphylococcal toxic shock. Symptoms include a persistent fever, mucocutaneous changes, multisystem inflammation, meningism and high levels of inflammatory biomarkers [87]. MIS-C was originally thought to be an atypical form of KD, however some patients with MIS-C display symptoms that are rarely seen in KD, including shock, coagulopathy and gastrointestinal symptoms [87, 94, 95]. Furthermore, the epidemiological features of MIS-C differ to those of KD (typically older children of Black or Hispanic descent *vs*. younger children of Asian descent). As a result, MIS-C is indeed considered as a distinct syndrome to KD and COVID-19.

SARS-CoV-2 is not always detectable by PCR assays in children presenting with MIS-C, and even if it was, SARS-CoV-2 seropositivity does not indicate that a child has MIS-C. Given the highly overlapping clinical characteristics between MIS-C and other infectious and inflammatory conditions, diagnosing MIS-C based on clinical symptoms is a major challenge. An accurate diagnostic test is urgently required to enable timely diagnosis of MIS-C. Currently, most countries follow the World Health Organisation (WHO) diagnostic criteria for MIS-C [96] which requires fulfilment of various criteria (Figure 4).

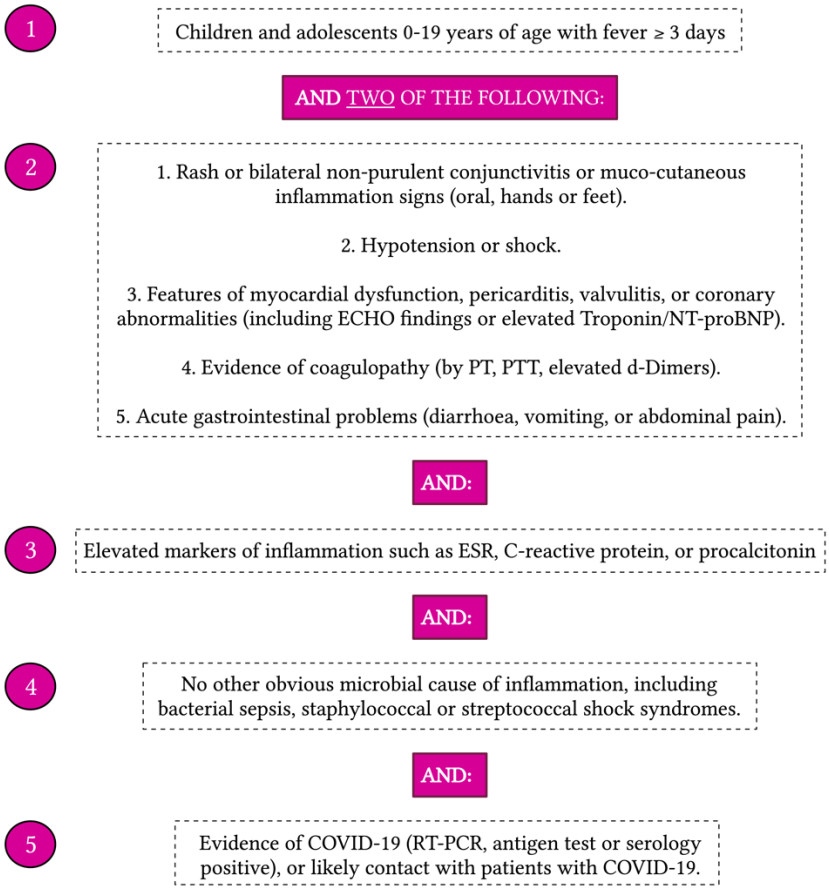


Figure 4: *The World Health Organisation (WHO) criteria for diagnosing MIS-C. Information taken from: https://www.who.int/news-room/commentaries/detail/multisystem-inflammatory-syndrome-in-children-and-adolescents-with-covid-19 (World Health Organisation, 2020). ECHO = echocardiogram; NT-proBNP = N-terminal pro b-type natriuretic peptide; PT = prothrombin time; PTT = partial thromboplastin time; ESR = erthryocyte sedimentation rate; COVID-19 = Coronavirus Disease 2019.*

Host diagnostic signatures may overcome the shortcomings in traditional diagnostic processes for MIS-C as they do not depend on the isolation of a causative pathogen. Host transcriptomic and proteomic profiles have been explored in the context of MIS-C, however all studies have focused on the molecular processes underpinning the disease, looking to gain biological insights rather than to identify diagnostic signatures [93, 94, 97-99]. To date, there have been no host gene or protein signatures proposed for diagnosing MIS-C. In chapter 5 of this thesis, I describe the processes undertaken for the discovery and validation of a diagnostic gene signature for differentiating between MIS-C and other paediatric infectious and inflammatory conditions.

1.5 Thesis contributions

Diagnosing paediatric infectious and inflammatory diseases with certainty is a major challenge. The exact nature of the challenge differs depending on the type of disease. For diseases with a known causative pathogen, such as those caused by bacteria and viruses, gold-standard culture-based diagnostic tests can be used, however these tests have a slow turnaround time and are often unreliable.

Diagnosis is made considerably more challenging when the cause of a disease is unknown, as is the case with Kawasaki disease (KD). In this instance, there are many more hurdles to overcome before accurate diagnosis is possible. The emergence of multisystem inflammatory syndrome in children (MIS-C) provided paediatricians with yet more diagnostic challenges due to the highly overlapping clinical presentation of MIS-C with other paediatric infectious and inflammatory conditions. With MIS-C, the causative pathogen is known (SARS-CoV-2), yet isolation of SARS-CoV-2 does not indicate that a child has MIS-C.

Host omic profiles have been shown to aid in diagnosis of paediatric infectious and inflammatory diseases [67, 68, 72, 100-103] and have uncovered important information surrounding the biology of these diseases [83, 97, 98, 104, 105]. This thesis has used high-throughput host omic profiles to address some of the ongoing challenges in paediatric infectious and inflammatory diseases. The key contributions of this thesis to the scientific field of paediatric infectious and inflammatory diseases will be discussed in this section.

Chapter 3 presents the discovery and subsequent validation of host protein biomarkers for diagnosing bacterial and viral infections in febrile children. Distinguishing between viral infections and life-threatening bacterial infections in children with fever is a major challenge for paediatricians. Culture-based diagnostic approaches are the gold-standard; however, they are slow and often return unreliable or inaccurate results. Alternative approaches are required for accurate diagnosis of bacterial and viral infections in febrile children. In chapter 3, I have provided considerable contributions to the ongoing effort to improve diagnosis of febrile children. Chapter 3 is the first time high-dimensional host proteomic datasets have been used to identify biomarkers for diagnosing bacterial and viral infections specifically in children. Chapter 3 contributes a novel 6-protein signature that can accurately differentiate between bacterial and viral infections in febrile children and could be developed into a rapid point-of-care diagnostic test. The 6-protein signature presented in chapter 3 outperforms the only other protein diagnostic signature for bacterial and viral infections (MeMed BV™ by [72]) and does so without depending on the flawed host protein biomarker CRP. In chapter 3, I have compiled a shortlist of 34 candidate protein biomarkers for diagnosing febrile children, including many novel proteins that have not been identified as potential biomarkers before, and I have identified hundreds of proteins with significantly different levels between bacterial and viral infections that could be used by other researchers to guide future studies.

Chapter 4 contains extensive comparisons of the host transcriptomic and proteomic profiles of children with KD, a paediatric inflammatory disease without a known cause. Despite the various viral and bacterial pathogens proposed as the trigger of KD, none have been confirmed. Further knowledge surrounding the biological pathways implicated in the manifestation and host response to KD is required to improve understanding of the pathogenesis and aetiology of the disease. In chapter 4, I have performed the first comprehensive comparison of host gene expression and protein abundance

profiles between children with KD and children with bacterial and viral infections. Chapter 4 contributes many novel biological findings in the context of KD, including various genes, proteins and pathways that are altered in KD in addition to bacterial and viral infections, or altered uniquely in KD. Chapter 4 shows, for the first time, that the host molecular profiles from children with KD are highly heterogeneous, revealing subtypes of disease. This is an intriguing finding that raises the question as to whether KD is indeed a single disease or rather a spectrum of diseases. Importantly, chapter 4 reveals that despite the similarities between the molecular profiles of children with KD and those with bacterial and viral infections, KD can be distinguished on the host transcriptional level, suggesting that KD may not be caused by bacterial or viral triggers. The work presented in chapter 4 has been published in the *International Journal of Molecular Science* [83].

In chapter 5, I present the discovery and validation of sparse gene signatures for distinguishing MIS-C from KD, bacterial and viral infections. MIS-C is a newly emerged paediatric inflammatory syndrome that occurs 2-6 weeks after infection with SARS-CoV-2. The highly overlapping clinical presentation of MIS-C and other paediatric infectious or inflammatory diseases and the lack of a gold-standard diagnostic test make accurate diagnosis of MIS-C extremely challenging. The major contribution of chapter 5 is a 5-gene host diagnostic signature that can accurately distinguish MIS-C from KD, bacterial and viral infections. This diagnostic signature of MIS-C is the first to be reported, and could be quickly developed into a point-of-care test, revolutionising the diagnosis of MIS-C. Chapter 5 also contains the first comparison of host transcriptomes of children with MIS-C to those from children with KD and a range of bacterial and viral infections, providing novel insights into the genes that are altered between MIS-C and each of the other disease groups.

Chapter 2

2 Methods

2.1 Overview

Omic datasets are high-dimensional datasets, meaning that they quantify many features across multiple samples. Often, they have more features (p) than samples (n), i.e., $p > n$. Inevitably, not all features will be relevant to the research question at hand with many features considered as "noise". Therefore, statistical methods must be employed to identify the most important features and reduce the effects of technical factors and noise.

The computational and statistical methodologies used in this thesis will be described in this chapter. Methods applicable to the 4 types of data presented in this thesis will be discussed. Pipelines for processing raw data are outlined in section 2.2, followed by quality control (QC) and normalisation processes in section 2.3. Analytical approaches are discussed in section 2.4. Four types of omic data were used for the results presented in this thesis: microarray (Illumina HumanHT-12), RNA Sequencing (RNA-Seq), SomaScan®, and liquid chromatography tandem mass spectrometry (LC-MS/MS).

2.2 Raw data processing

For all omic datasets presented here, the first key step is to transform the raw data, or the output from the apparatus used for omic profiling (see chapter 1 section 1.3), into a format that can be parsed by computers. The complexity of this step varies between omic profiling approaches, with some platforms returning data in a format that can be easily parsed with the help of software tools, and with other platforms requiring various computationally demanding steps that require specialist knowledge.

The microarray and SomaScan® platforms quantify the levels of known RNA and protein targets, respectively. Contrastingly, RNA-Seq and untargeted LC-MS/MS quantify all RNAs or proteins identifiable in a sample, respectively, meaning that they do not depend on known target molecules. The steps required to process the raw data files obtained from SomaScan® and microarray are considerably more straightforward as the target features do not need to be identified. For both technologies, image processing software transforms the images received from the microarray scanners into raw data files by numerically encoding the measured intensity for each bead for SomaScan® and each probe for microarray [40, 85]. Whilst commercial software tools can be used for downstream analyses of the raw data files, in this thesis, open-source software packages for the R Statistical

Computing Language [106] were used to read in the files returned from the sequencing technologies. The lumi [107] and readat [108] packages were used to read in the data from the Illumina HumanHT-12 microarray and SomaScan® platforms, respectively. The output from the lumi and readat packages is a set of abundance values for each feature (i.e., transcript or protein) for each sample in the dataset.

2.2.1 Processing raw RNA Sequencing data

Processing raw RNA-Seq data into an analysis-ready format requires multiple time-consuming and computationally expensive steps. At each step, various choices are available to researchers regarding the tools or exact analytical approaches taken. The series of software tools described here are those used to process the RNA-Seq data used in this thesis.

RNA-Seq processing pipelines are either based on the alignment of reads to a reference sequence (genome or transcriptome) or are alignment-free methods. Whilst alignment-free methods are faster and computationally cheaper, alignment-based methods have been shown to outperform alignment-free methods for quantification of small and lowly-expressed ribonucleic acids (RNAs) [109]. The aim of alignment is to map the reads to a reference (either genome or transcriptome) to determine which features (this could be transcripts, genes, or exons) are expressed [110]. Reads can be aligned to the reference genome or transcriptome. The RNA-Seq data presented in this thesis was aligned to the version 89 ensembl GCh38 human genome [111] using STAR (spliced alignment to a reference; [112]). Genomic alignment was selected as it offers more flexibility in downstream analyses, including detection of post-transcriptional regulation events and quantification of the RNA decay rate [113, 114]. Furthermore, if the transcriptome reference was used, a considerable proportion of reads would remain unaligned as the transcriptome reference would not contain intronic sequences and would miss RNAs that are unannotated.

There are multiple options of aligners available. The STAR aligner was used to align the RNA-Seq dataset presented in chapter 5 due to its speed and high alignment accuracy [115]. The STAR aligner follows a two-step process made up of seed searching, followed by clustering, stitching, and scoring. In seed searching, STAR searches for the longest sequence within a read that exactly matches to one or more locations on the reference genome. The portion of the read that has been mapped to the reference is known as seed1. The process then repeats with STAR matching the longest sequence of the unmapped portion of the read to the reference, with the next portion of the read that is mapped to the reference known as seed2. The process of seed searching continues for the forward and reverse directions of the read sequence. In clustering, stitching, and scoring, the separate seeds are stitched together by clustering around sets of anchor seeds. A scoring scheme guides the stitching process by applying scores and penalties for mismatches, insertions, deletions, and junction gaps. For each read,

the scores of the stitched combinations are compared and the combination with the highest score is reported as the alignment of the read [112].

After alignment, gene expression is quantified through feature counting which involves counting the number of aligned reads that overlap with each gene in the annotation. Feature counting is essential for biological interpretation of the aligned reads. For the RNA-Seq data used in this thesis, featureCounts was used for feature counting, ultimately returning a $p \times n$ matrix where p are the RNA transcripts and n are the samples [116].

Quality control (QC) steps are performed throughout the RNA-Seq processing pipeline. FastQC [117] summarises various QC metrics from the FASTQ files which are returned from the sequencer. FastQC provides information about the per base and per sequence quality scores, per sequence GC content (percentage of guanine or cytosine bases per sequence), per base N content (percentage of bases at each position with no base call, i.e., "N"), and whether certain sequences are overrepresented (such as contaminants), amongst other analyses. MultiQC [118] returns a single summary report containing QC information from various other tools, including FastQC, featureCounts and STAR, offering insights into how successfully the RNA-Seq processing pipeline has been performed.

2.2.2 Processing raw liquid chromatography tandem mass spectrometry data

Data analysis is one of the principal challenges of modern mass spectrometry (MS)-based proteomics [119]. MaxQuant is a popular and widely used quantitative proteomics software package for analysing MS shotgun proteomic data generated from a range of machines and was used for the liquid chromatography-tandem mass spectrometry (LC-MS/MS) dataset presented in this thesis in chapters 3 and 4 [119]. MaxQuant involves a series of algorithms that perform distinct steps, including peptide/protein detection, identification, and quantification.

In LC-MS/MS, peptides are fragmented into product ions. Therefore, the first step in processing raw LC-MS/MS data is to identify the peptides (features) that the product ions originate from in a process called feature detection. Mass spectrometers return many 3D spectra. MaxQuant uses the mass to charge ratios (m/z), retention times and peak intensities for feature detection. Feature detection determines which 3D peaks originate from isotopes of the same peptide. Once features are detected, the mean and standard deviation of the mass of each peptide are calculated across all isotopes that are detected from the same peptide. The first search then takes place, which includes a database search using the Andromeda database [120]. This identifies the peptides that have produced the features detected using the feature mass standard deviation previously calculated across isotopes of a given peptide. Some peptides are successfully labelled and identified in the first search, but many cannot be

identified from the masses estimated due to errors introduced from technical factors. Recalibration then occurs in which the known database masses of the identified peptides are used to recalibrate the rest of the spectra, leading to re-estimated mass means and standard deviations, resulting in more accurate values. Features that could not be identified in the first search are now at the correct mass position, enabling them to be successfully identified. Peptide sequences can be determined by comparing the m/z of fragments of the same peptides (known as MS2), with differences in the m/z used to infer the amino acids that differ between fragments. MaxQuant can also characterise post-translational modifications (PTMs) through contrasting the identified and unidentified MS/MS spectra, positioning hypothetical residues on the unidentified spectra and using a novel algorithm to determine whether the unidentified spectra are indeed modified forms of the already-identified spectra [119]. This is an unbiased way of identifying PTMs since all modifications are considered and explored.

MaxQuant uses various quantification strategies, suitable for multiple MS-based proteomic approaches. Label free quantification (LFQ) is a MS proteomic approach that quantifies the relative amount of protein in two or more samples and does not require the proteins to be labelled, for example with stable isotope compounds, a process that is complex and expensive. MaxLFQ [121] works within the MaxQuant software suite to perform label-free quantification and normalisation across samples, resulting in relative protein quantification across all samples [119, 121]. MaxLFQ returns an *n x p* matrix where *n* represents samples and *p* represents protein groups.

2.3 Quality control, transformation, and normalisation

Once omic datasets have been transformed from raw data into matrices, QC, transformation, and normalisation steps are essential to identify and account for bias and errors introduced through technical factors [122]. Although each omic approach will have slight variations in the exact methods required, the overall aims and steps are the same. The principles of each key step in the QC, transformation and normalisation pipeline will be outlined and the variations and specific methods applicable to each of the 4 key omic approaches addressed in this thesis (microarray, RNA-Seq, SomaScan®, LC-MS/MS) will be described (Figure 5).

Principal component analysis (PCA) is a fundamental tool that is used for QC of high-dimensional omic data. PCA is a dimensionality reduction method that reduces the number of variables and extracts the most important information from the dataset, improving interpretability whilst also minimising information loss [123]. PCA transforms a dataset into a set of principal components (PCs) which are linear combinations of the original variables *p*, with up to *p-1* PCs. The first PC, Y_1, which captures the most variance in the dataset is calculated by a linear combination of variables across all samples $X_1, X_2, \ldots, X_p$:

Equation 1

$$Y_1 = a_{11}X_1 + a_{12}X_2 + \cdots a_{1p}X_p$$

where $a_{11}, a_{12}, \ldots a_{1p}$ represent weights, known as loadings, for each variable across all samples, with the sum of squares of the weights constrained to 1. Loadings describe the contribution of each variable to each PC, with larger absolute loadings indicating a strong relationship between the feature and the PC, and positive and negative loadings indicating positive and negative correlations between the feature and PC, respectively. The second PC, Y_2, is calculated in the same way, capturing the highest amount of variance possible whilst being orthogonal to the previous PC:

Equation 2

$$Y_2 = a_{21}X_1 + a_{22}X_2 + \cdots a_{2p}X_p$$

The process continues for all subsequent PCs. Each PC is associated with an eigenvalue which indicates the variance that the PC explains in the dataset, with the sum of all eigenvalues across PCs = 1. Each sample, *n,* has a score for each PC, *k*. Per-sample, per-PC scores, Y_{nk}, are linear combinations of the original variable for each sample $x_{n1}, x_{n2}, \ldots x_{np}$ and the loadings, $a_{11}, a_{12}, \ldots a_{1p}$, calculated as:

Equation 3

$$Y_{nk} = a_{1k}x_{n1} + a_{2k}x_{n2} + \cdots + a_{pk}x_{np}$$

Scores for individual PCs, Y_k, can be visualised by biplots contrasting PCs, i.e., PC1 *vs*. PC2; PC3 *vs*. PC4; PC5 *vs*. PC6. Not only can outlier samples be identified, but the association between samples' PC scores and metadata values can be calculated, revealing important factors that contribute to the variance in the dataset. Furthermore, if unexpected clustering patterns are observed on biplots, loadings for the respective PCs can be inspected to gain insights into the variables that are driving the observed patterns. In this thesis, PCA was performed at multiple stages throughout the QC processes to identify the key drivers of variation in the datasets and decipher the impact of processing and QC steps.

During the QC process, features that are either missing in a large proportion of samples or have levels lower than a certain threshold are removed. Extremely low levels of features across all samples indicates that they are not present at biologically meaningful levels and thus are not of interest. Not only can these features introduce noise into the dataset, their presence leads to an unnecessary number of statistical tests being performed in downstream analyses [124].

Normalisation methods aim to remove experimental and technical bias whilst retaining biological variation [125]. Variations in underlying data distributions mean that the exact normalisation strategy

performed differs between datasets from different omic platforms. Despite these differences, they are typically always applied to all samples present in a dataset to make samples more comparable. Specific normalisation methods used in this thesis will be described in sections 2.3.1 to 2.3.4 of this chapter. Some variance introduced by technical or experimental factors cannot be removed by normalisation, such as batch effects. A batch effect is when non-biological factors introduce variability in measured data. Batch effects are a common problem since omic datasets are often generated in multiple batches due to logistical and practical limitations including the size of assays and the number of samples that can be handled in a day [126, 127]. Batch effects can also be introduced due to variations in the experimental conditions, such as the reagents used or even the individual carrying out the experiments [126]. Due to their frequency in omic datasets, various methods are available for identifying and removing batch effects, and the choice of method typically reflects the omic approach used.

ComBat [128, 129] is a popular batch effect correction method that was originally designed for correcting batch effects in microarray datasets but has since been applied to various data modalities including proteomic [130] and metabolomic [131] datasets. ComBat uses empirical Bayes (EB), a class of methods for statistical inference in which the prior distribution is estimated from the data rather than fixed before the data is observed (as in standard Bayesian methods). EB methods are well suited to omics due to their ability to handle high-dimensional datasets with small numbers of samples *n* [128]. EB methods can "borrow information" across features and experimental conditions, with the borrowed information resulting in more stable and reliable inferences.

ComBat first defines a model, Y_{ijf}, which represents the abundance value for feature *f* for sample *j* from batch *i*:

Equation 4

$$Y_{ijf} = \alpha_f + X\beta_f + \gamma_{if} + \delta_{if}\varepsilon_{ijf}$$

where α_f is the overall feature abundance, X is a design matrix for sample conditions, β_f is the vector of regression coefficients for X, and γ_{if} and δ_{if} are the additive and multiplicative batch effects of batch *i* for feature *f*, respectively. ε_{ijf} represents the error terms which are normally distributed with mean zero and variance σ_{f^2}.

Data is standardised to account for different magnitudes in abundances across features. Model parameters α_f, β_f, γ_{if} and δ_{if} are estimated as $\hat{\alpha}_f$, $\hat{\beta}_f$, $\hat{\gamma}_{if}$ and $\hat{\delta}_{if}$ using feature-wise ordinary least-squares. The standardised data, Z_{ijf}, is calculated as:

Equation 5

$$Z_{ijf} = \frac{Y_{ijf} - \hat{\alpha}_f - X\hat{\beta}_f}{\hat{\sigma}_f}$$

The EB estimates for batch effect parameters γ_{if} and δ_{if} which represent additive and multiplicative batch effects of batch i for feature f, respectively, are estimated using hyperparameters estimated empirically from the standardised data. The calculated batch effect estimators are then used to adjust the data P_{ijf} as:

Equation 6

$$P_{ijf} = \frac{\hat{\sigma}_f}{\hat{\delta}_{if}}\left(Z_{ijf} - \hat{\gamma}_{if}\right) + \hat{\alpha}_f + X\hat{\beta}_f$$

Various modifications of ComBat have been developed, including COCONUT (COmbat CO-Normalization Using conTrols [132]) which uses control/healthy samples to estimate $\hat{\alpha}_f$, $\hat{\beta}_f$, $\hat{\sigma}_f$, $\hat{\delta}_{if}$ and $\hat{\gamma}_{if}$ which are then applied to the remaining samples, i.e., disease group samples. COCONUT was developed to circumvent the assumption used in ComBat that all phenotypic groups (i.e., diseased and healthy) have the same distribution. COCONUT applies the same background distribution to diseased groups whilst maintaining a distance between these groups and the healthy controls [132]. The major drawback of COCONUT is that it depends on control samples that are assumed to be from the same distribution which is not always the case.

ComBat-seq is another modified version of ComBat that is optimised RNA-Seq count data [127]. ComBat assumes Gaussian distributions for the underlying distribution of the data. This is inappropriate for RNA-Seq count data which is better fitted to a negative binomial distribution. Furthermore, RNA-Seq counts are in integer format meaning that downstream tools for RNA-Seq data analysis expect integers. ComBat-seq ensures that integers are returned following batch effect correction, unlike the original ComBat method [127]. ComBat-seq obtains estimates directly from a negative binomial regression model. Further differences between the model includes the final adjustment. In ComBat, batch effects are removed by subtracting the mean batch effect estimate and scaling by variance estimates. This differs from ComBat-seq which maps the empirical distribution of count data to an expected distribution that is "batch-free", retaining the integer counts that are necessary for downstream RNA-Seq analysis tools [127].

The specific QC and normalisation processes applied to the four types of omic data used in this thesis will now be described.

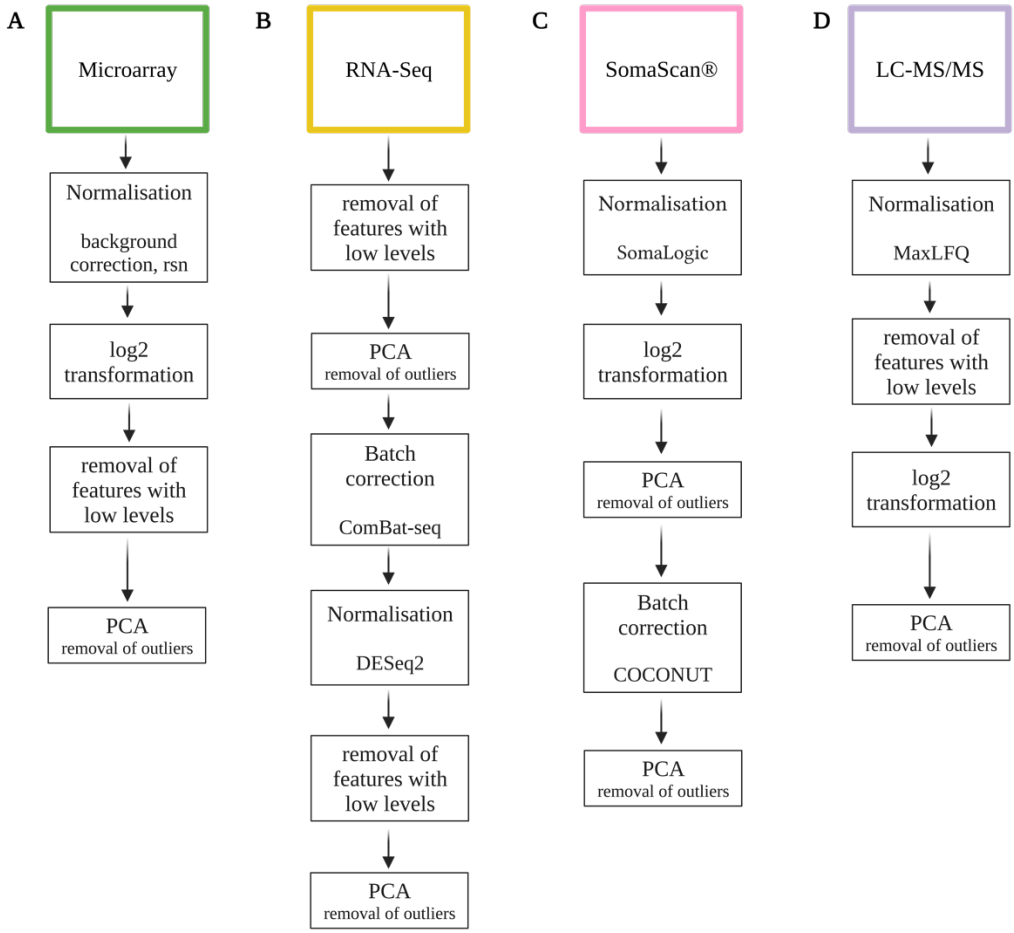


Figure 5: *Overview of quality control and normalisation processes carried out for each type of omic data presented in this thesis.* ***A****: microarray;* ***B****: RNA-Seq;* ***C****: SomaScan;* ***D****: LC-MS/MS. RNA-Seq = RNA Sequencing; LC-MS/MS = liquid chromatography tandem mass spectrometry; PCA = principal component analysis.*

2.3.1 Quality control and transformation of microarray data (Figure 5A)

Normalisation of the microarray datasets presented in this thesis involved background correction and robust spline normalisation (RSN) followed by log2 transformation using the R package lumi [107]. Background correction was used to adjust for fluctuations in ambient fluorescence in the neighbourhood of each microarray, reducing noise [133]. RSN, which was identified as the most appropriate normalisation method for Illumina microarray data following a systematic comparison [134], combines the strengths of two other normalisation methods – quantile and loess normalisation [135, 136]. RSN applies quantile normalisation first followed by loess normalisation and combines the positives of both methods (preservation of the feature rank order and continuous mapping of the values, respectively) whilst overcoming the drawbacks of both methods (discontinuous mapping of values and no rank preservation, respectively) [134]. GenomeStudio, the software associated with Illumina array platforms, calculates a detection *p*-value which indicates the confidence that a probe is expressed above the background defined by negative control probes. Probes were retained if at least 80% of samples in each disease group had detection *p*-value <0.01. PCA was performed to identify

outliers which were subsequently removed and to evaluate the association between PC scores and disease group, age, and sex at birth.

2.3.2 Quality control and transformation of RNA Sequencing data (Figure 5B)

Genes without counts ≥10 in at least three samples were removed alongside ribosomal genes and genes in the globin gene depletion kit (*HBB*, *MT-RNR2*, *MT-RNR1*, *HBA2*, *HBG2*, *HBG1*, *HBA1*). PCA was performed to identify outliers which were subsequently removed and to evaluate the association between PC scores and disease group, age, plate, sequencing batch, and sex at birth. Plate and batch effects were removed using ComBat-seq [127] and counts were normalised using DESeq2's median of ratios method [137]. Whilst various normalisation methods exist for RNA-Seq, DESeq2's median of ratios method was selected since it accounts for sequencing depth and RNA composition, enabling comparisons of gene levels between samples. This renders it appropriate for differential expression analysis, unlike counts per million (CPM), transcripts per kilobase million (TPM), and reads/fragments per kilobase per million reads/fragments mapped (RPKM/FPKM). Sequencing depth measures the total number of usable reads from the sequencing machine. Some samples may have been sequenced at varying depths, meaning that accounting for sequencing depth is essential for enabling between sample comparisons.

DESeq2 expects data as an $n\ x\ m$ matrix of counts, k_{ij}, with $i = 1,..,n$ genes and $j = 1,..,m$ samples. Size factors s_j enable samples sequenced at different depths to be comparable and are calculated by:

Equation 7

$$\hat{s}_j = median \frac{k_{ij}}{(\Pi_{v=1}^{m} k_{iv})^{1/m}}$$

where the denominator of the expression represents a pseudo-reference sample obtained by taking the geometric mean across samples. Each size factor estimate, s_j, is the median ratio of gene counts relative to the geometric mean per gene. Normalised data, P_{ij}, is obtained by dividing counts by the s_j for each sample:

Equation 8

$$P_{ij} = \frac{k_{ij}}{\hat{s}_j}$$

Following normalisation, further filtering of lowly expressed genes was performed to remove genes without counts ≥20 in at least three samples.

2.3.3 Quality control and transformation of SomaScan® data (Figure 5C)

SomaLogic returned pre-normalised values for the SomaScan® dataset. Normalisation was performed by scaling the data using multiple scale factors that correct for variations in aptamer hybridisation efficiency, inter- and intra-assay variability, variability in the starting quantities of proteins, and plate effects. A hybridisation normalisation scale factor was calculated for each sample using the median values of control SOMAmers. Samples with a hybridisation normalisation scale factor outside of the standard range 0.4-2.5 were flagged for removal. Following normalisation, protein abundances were log2 transformed and batch effects were removed by COCONUT [132]. PCA was performed to identify outliers which were subsequently removed and to evaluate the association between PC scores and disease group, age, and sex at birth.

2.3.4 Quality control and transformation of liquid chromatography tandem mass spectrometry data (Figure 5D)

For the two LC-MS/MS datasets presented in this thesis, the MaxLFQ algorithm [121] was used for relative quantification and normalisation. Protein groups identified as contaminants or those missing in over 90% of samples overall or within in each disease group were removed. Protein abundances were log2 transformed and PCA was performed to identify outliers which were subsequently removed and to evaluate the association between PC scores and disease group, age, and sex at birth.

2.4 Data analysis

Once data has been processed and normalised and QC checks have occurred, downstream data analysis can be performed. The output of data pre-processing is a matrix of features and samples. The statistical analysis methods used to analyse the datasets presented in this thesis will be described below. Many of the statistical methods are machine learning approaches. Machine learning approaches are stratified into unsupervised and supervised approaches depending on whether they used labelled data (i.e., disease group) with unsupervised methods (e.g., PCA, clustering) using unlabelled data and supervised methods (e.g., classification, regression) using labelled data. All statistical analyses were performed in R Statistical Computing Language [106].

2.4.1 Clustering

Clustering is an unsupervised approach for grouping data points according to their similarities. Clustering is a powerful tool for the analysis of complex omic datasets as it uses all available information to detect hidden patterns and reveal natural groups of samples present in data. Whilst many clustering algorithms exist, *k*-means was selected as the clustering algorithm to use for analysis

of transcriptomic (microarray) and proteomic (LC-MS/MS) data from patients with Kawasaki disease (KD), bacterial and viral infections in chapter 4. *k*-means was selected due to its simplicity, flexibility, and speed to implement, including for large datasets. *k*-means partitions n samples into k clusters, ensuring that each sample belongs to the cluster with the closest mean. The aim is to minimise the within-cluster sum of squares which is synonymous with the cluster variance.

One of the disadvantages of *k*-means clustering is that it requires the number of clusters k to be specified in advance. Whilst this can introduce bias, there are various statistical indices that can be used to determine the optimal k, such as the Silhouette coefficient that measures how similar a sample is to its own cluster compared to other clusters [138], and the Elbow method that measures the distance between each sample to its cluster centroid. NbClust [139] is an R package that gathers all indices together and simultaneously applies them to a dataset across a range of k. In chapter 4, NbClust was applied prior to *k*-means clustering to identify the optimal value of k for the datasets, with k=2, k=3, …, k=10 tested.

2.4.2 Differential expression and abundance analysis

Differential expression analysis is a very popular analytical approach that identifies the genes that show differences in expression levels between conditions, such as disease groups [140]. Differential abundance analysis is synonymous with differential expression analysis; however, it compares the abundances of protein or metabolite levels between conditions. Differential abundance/expression analysis is a fundamental analytical tool used in all chapters of this thesis. Whilst the most appropriate differential expression or abundance analysis method may differ according to the type of data that is being analysed, the basic principles of differential expression analysis remain the same, which is to use generalised linear models (GLM) to compare the levels of features between groups.

Differential expression/abundance analysis is performed on a matrix of expression or abundance values composed of $p \times n$ where p represents genes (or other features, such as peptides/proteins) and n represents samples. For each row (feature), a linear model is fitted as follows:

Equation 9

$$\mathrm{E}[y_j] = X\alpha_j$$

where y_j is the expression data for feature j, X is the design matrix and α_j is the matrix of coefficients. Each row of X corresponds to a sample and the columns contain the covariates of interest. The design matrix includes the covariate or characteristic to be tested, for example disease group in a differential expression analysis model that aimed to compare disease group A *vs*. disease group B, in addition to any other important covariates that could affect expression/abundance values, such as age and sex.

Linear models are very flexible, enabling complex design matrices to be used in differential expression analysis. The vector of coefficients α_j contains the differences in expression according to the covariates included in X.

In differential expression analysis, entire experiments are analysed as an integrated whole rather than by considering each feature in isolation through using empirical Bayes (EB) methods [141]. The highly parallel nature of sequencing experiments lends itself to use with EB methods as information can be borrowed between genes. Since the design of the linear models fitted across all features in the experiment is the same, the global variability across all features is used to estimate the variance for each feature, in addition to the variance estimated from the feature-specific linear model. This provides a more accurate estimation than if the feature-specific models were used in isolation [137]. This approach ensures reliable and stable inferences even in experiments with small sample sizes or low numbers of replicates [141].

Differential expression/abundance analysis involves performing statistical tests on up to tens of thousands of features simultaneously which increases the false positive rate and thus the risk of incorrect rejection of the null hypothesis (type 1 error). This phenomenon is known as the "multiple testing problem" and is a key consideration in differential expression/abundance analysis. Typically, features are only considered as significantly differentially expressed (SDE) or significantly differentially abundant (SDA) if they remain significant following multiple testing correction. Two main multiple testing correction approaches are used: the Bonferroni and the Benjamini-Hochberg (BH, [142]; also known as the false discovery rate - FDR) corrections. In the Bonferroni correction, adjusted *p*-values are calculated by dividing the original *p*-value by the number of tests performed. The Bonferroni correction is very stringent and does not take into account the distribution of the original *p*-values. The BH adjustment is more appropriate for differential expression analysis as it attempts to control the proportion of false positive tests rather than controlling the probability of at least one positive test, which the Bonferroni correction attempts to do [143]. The first step of the BH adjustment is to rank the *p*-values in ascending order, returning a rank *i* for each feature. The feature-wise BH *p*-value $p_{(j)}^{BH}$ is then calculated as follows for each gene:

Equation 10

$$p_{(j)}^{BH} = \frac{mp_{(i)}}{i}$$

where *m* is the total number of tests, and *i* is the rank of the feature.

Differential expression/abundance analysis returns the adjusted *p*-value (typically BH-adjusted) in addition to the $\log_2$ fold-change (LFC) value for each feature. The LFC is the degree of change between the groups of interest, with positive and negative values indicating increasing and decreasing levels in

the case group, respectively. If "adjusted *p*-values" are referred to in this thesis, these *p*-values will have been adjusted using the BH adjustment.

Two differential expression analysis methods were used in this thesis. Limma [141, 144] was applied to the proteomic datasets (SomaScan® and LC-MS/MS) presented in chapters 3 and 4 in addition to the microarray dataset presented in chapter 4, and DESeq2 [137] was applied to the transcriptomic (RNA-Seq) dataset presented in chapter 5. The key differences between Limma and DESeq2 and their appropriate applications will be described in more detail.

Limma was originally designed for differential expression analysis of microarray data but has been extensively developed. Following data pre-processing and normalisation, Limma is suitable for data from all omic technologies including protein quantification approaches [145-147]. Limma includes a data transformation step – "voom" transformation – that transforms the data (i.e., gene or transcript levels, protein abundances) to the logarithmic (base 2) scale and estimates each feature's mean-variance relationship [144].

DESeq2 was specifically developed for differential expression analysis of RNA-Seq count data. It is less flexible than Limma as it only accepts integers, meaning that it can only be applied to raw RNA-Seq counts. RNA-Seq is count data measuring the number of reads aligned to a gene, hence it cannot be continuous and modelled using the normal distribution. The Poisson or Negative Binomial distributions are suitable for count data; however, the Poisson distribution assumes that the mean and variance are equal for each feature, which is rarely the case in RNA-Seq. Using the Poisson distribution for differential expression analysis would underestimate variability – leading to overdispersion – resulting in high false positive rates [148]. The Negative Binomial is suitable for when mean < variance which is typically the case with RNA-Seq count data.

Both Limma and DESeq2 use EB frameworks to borrow information between features when estimating the feature variance [137, 141]. Both methods estimate the gene-wise dispersion values first, and then model the relationship between the expected gene-wise dispersion and mean across all genes in the dataset. Gene-wise dispersion estimates are then shrunk towards the expected dispersion values using an EB approach in which the degree of shrinkage is determined by a) the similarity between true dispersions and the fitted line and b) the degrees of freedom in the dataset; with increases in sample size leading to decreases in the degree of shrinkage [137, 149]. This process of shrinkage reduces the impact of under-estimates of dispersion, thus reducing the false positive rate and providing more accurate estimates of the degree of change between groups [137].

Differential expression/abundance analysis returns a list of *p*-values and LFC values for all features in the analysis which should be filtered, typically using *p*-value thresholds and/or LFC thresholds. The resulting list of SDE or SDA features can then be used for downstream analyses and exploration.

2.4.3 Feature selection

The aim of feature selection is to reduce the number of variables present in a dataset by selecting those that are the most important. Feature selection algorithms reduce the dimensionality of the dataset with the aim of producing simpler models that are easier to handle. Whilst unsupervised feature selection methods exist, supervised feature selection methods will be the focus of this thesis. In supervised feature selection, the most relevant features for distinguishing between the groups of interest are identified. This results in a signature composed of a combination of features that, when combined, can differentiate between the groups of interest.

In chapters 3 and 5, feature selection was performed with the aim of identifying a small number of features (proteins and genes, respectively) that could be translated into rapid, point-of-care tests for diagnosing bacterial *vs*. viral infections, and MIS-C, respectively. An in-house method called Forward Selection-Partial Least Squares (FS-PLS [67], https://github.com/lachlancoin/fspls.git) was used in both instances to identify protein (chapter 3) and gene signatures (chapter 5). FS-PLS combines forward selection (FS), a stepwise regression method that begins with an empty model and adds the most important variables sequentially, with partial least squares (PLS), a regression method that reduces the set of variables to a smaller set of uncorrelated variables. FS-PLS combines the model simplicity and interpretability of FS with the dimensionality reduction strength of PLS to result in small models composed of non-correlated features [150].

FS-PLS starts by fitting N univariate regression models, where N is the number of features. The maximum likelihood estimation (MLE) function is used to estimate the regression coefficient of each model and a t-test is used to assess the model's goodness of fit. SV1 corresponds to the first variable selected. SV1 has the highest MLE and smallest p-value. Using singular value decomposition (SVD), FS-PLS projects out the variation explained by SV1 enabling it to add additional features that are not correlated to SV1.

In chapter 4, least absolute shrinkage and selection operator (LASSO) was used to identify a selection of features (genes and proteins separately) that could be used to classify bacterial and viral infections. In this instance, LASSO was selected over FS-PLS as the aim was not to obtain a small signature for diagnostic purposes so there were less constraints surrounding the number of features selected. LASSO is a regression analysis model that performs variable selection and regularization simultaneously [151]. Through using regularization, LASSO shrinks the regression coefficients (shrinkage) by imposing a penalty on their size. This results in sparse models with few features with non-zero coefficients. LASSO uses L1 regularization which means that not all coefficients are retained. L2 regularization, which is used by Ridge regression, occurs when the coefficients are shrunk but all are retained. Elastic Net sits between LASSO and Ridge regression as it selects variables like LASSO, whilst jointly shrinking the

coefficients of correlated predictors, like Ridge regression [151]. LASSO obtains feature coefficients $\hat{\beta}^{lasso}$ as follows:

Equation 11

$$\hat{\beta}^{lasso} = argmin\left\{\left(y_i - \beta_0 - \sum_{j=1}^{p}(x_{ij}\beta_j)\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

where y_i is the response variable, β_0 is the intercept term, x_j are the features, β_j is the original vector of feature coefficients, and λ is the non-negative complexity parameter that controls the amount of shrinkage. In LASSO regression $\lambda = 1$ whilst in Ridge regression $\lambda = 0$.

2.4.4 Signature performance evaluation

Feature selection methods including FS-PLS and LASSO select features (i.e., a signature) that have been identified due to their discriminatory abilities for a specific response variable (i.e., disease group). In this thesis, the expression or abundance values of features included in signatures were combined into a single score for each sample – a disease risk score (DRS) [67]. This is to evaluate the performance of the signature and apply it to independent samples. Two variations of the DRS were used: the simple DRS and the weighted DRS.

In the simple DRS, the features in the signature were first stratified according to their direction of change. The total expression/abundance of features that decreased was subtracted from the total expression/abundance of features that increased for each individual, i, as follows:

Equation 12

$$Simple\ DRS^i = \sum_{k=o}^{n} value_k^i - \sum_{l=0}^{m} value_l^i$$

where n and m are the features that increase and decrease, respectively, in the disease of interest compared to the comparator groups(s). The simple DRS is highly translatable and simple to implement and was first described in [67].

The weighted DRS was calculated as follows for each individual, i:

Equation 13

$$Weighted\ DRS^i = value_j^i \times \beta_j$$

where j represents the features included in the signature and β represents the feature coefficients which were obtained through GLMs that modelled the relationship between the features and the disease classification of interest. These feature coefficients were either calculated from a previous experiment (i.e., during the signature discovery phase), or were obtained from the dataset on which the weighted DRS was calculated.

Once the DRS was calculated for each sample, the performance of the signature was evaluated through using the area under the receiver operating characteristic (ROC) curve (AUC), sensitivity and specificity. These metrics use information about the rates of:

- true positives: gold-standard "positive" individuals that are classified as "positive";
- false positives: gold-standard "negative" individuals that are classified as "positive";
- true negatives: gold-standard "negative" individuals that are classified as "negative";
- false negatives: gold-standard "positive" individuals that are classified as "negative".

The sensitivity, also known as the true positive rate, is the probability of an actual positive testing positive and is calculated as follows:

Equation 14

$$sensitivity = \frac{true\ positive}{true\ positive + false\ negative}$$

The specificity, also known as 1-false positive rate, is the probability of test being negative when the disease is absent and is calculated as follows:

Equation 15

$$specificity = \frac{true\ negative}{true\ negative + false\ positive}$$

In ROC curves, the sensitivity is plotted against 1-specificity (Figure 6). The AUC is the area under the ROC curve and is used as a global assessment of the diagnostic accuracy of the test, considering both specificity and sensitivity. The AUC of a perfect classifier would be 1, as shown by the blue circle on Figure 6. The red line on Figure 6 shows a random guess test which would have an AUC of 0.5 (50%).

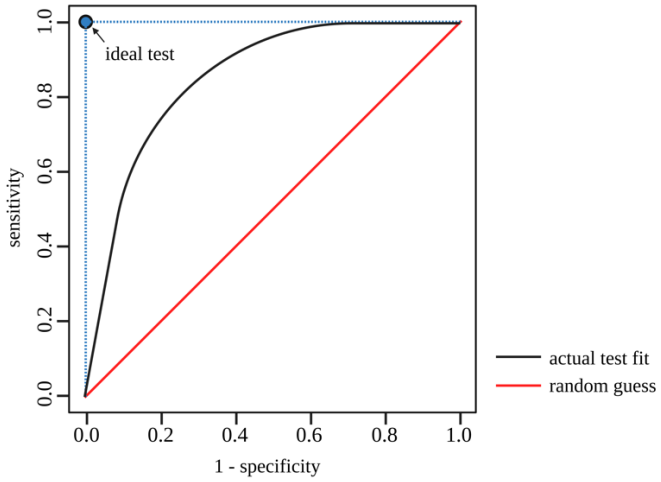


Figure 6: *A schematic representing a receiver operating characteristic (ROC) curve. The black line represents a test fit, whilst the red line represents a random guess. The ideal test is shown in the top left corner of the plot.*

Youden's index [152] (J) is a statistic that captures the performance of a dichotomous test by summarising sensitivity and specificity as follows:

Equation 16

$$J = sensitivity + specificity - 1$$

Youden's index was used to calculate sensitivity and specificity for the gene and protein signatures and classifiers. The sensitivity and specificity values that maximised Youden's index were selected for the threshold of classification (i.e., a DRS value) that was used for classifying samples into the disease groups of interest.

2.4.5 Cross validation

Cross validation is a resampling method for training and testing a model on different portions of a dataset. Cross validation is a powerful technique that is particularly useful for estimating the predictive ability of a model on unknown or previously unseen data. Cross validation helps to reduce the risk of overfitting which occurs when a predictive model fits exactly against its training data and does not perform well against unseen data. In *k*-fold cross validation, the dataset is split into *k* folds or partitions. The model is performed iteratively for *k* iterations, with each *k* fold being sequentially left out and used as test data and the remaining data used to train the model. *k*-fold cross validation can be used to identify which features perform the best across all partitions of samples (chapter 3), for accurate parameter estimation (chapter 4), or to predict the classifications of patients by a diagnostic signature (chapter 5).

2.4.6 Pathway enrichment analysis

Often, very long lists of SDE/SDA features are returned from differential expression/abundance analysis. Whilst some genes can be explored through literature searches, it is generally impractical to manually identify the biological functions of all SDE/SDA features. Pathway enrichment analysis refers to a selection of tools that aim to give biological meaning to the results from high-throughput sequencing experiments and downstream analyses such as differential abundance/expression analysis. Pathway analysis tools detect groups of related features that are altered and provide information about the biological processes that these features are associated with, offering mechanistic insights [153].

A common statistical test used in pathway enrichment analysis is the Fisher's exact test [154] which determines whether the proportion of features of interest in the pathway (i.e., the list of SDE genes returned from differential expression analysis) is greater than the proportion of features outside the pathway (the background set). This analysis is known as over-representation analysis. Use of the background set is important to reduce the frequency of false positives and inappropriately small *p*-values. Over-representation analysis tests can be ranked or non-ranked, with ranked tests using a ranked list of features, e.g., ranked according to absolute LFC returned from differential expression analysis. For ranked over-expression analysis, pathways whose features are enriched at the top of the ranked list are searched for first.

Whilst many pathway enrichment analysis tools exist, g:Profiler2 [155] was selected for use in chapter 4 as it can be applied to both gene- and protein-level data. g:Profiler2 uses gene ontology (GO) terms which are either biological processes (e.g., DNA repair), molecular functions (e.g., transport), or cellular components (e.g., ribosome). Each GO term is associated with a selection of genes that are known to be involved with that specific term. g:Profiler2 performs statistical enrichment analysis which determines whether the genes in a given GO term are enriched in the input data at a rate higher than would be expected normally.

Sometimes, lists of GO terms returned by pathway enrichment analysis can be large and highly redundant due to the redundancy in the GO database itself. For example, the GO term "leukocyte activation" is the "parent term" of GO terms including "leukocyte activation involved in immune response", "leukocyte activation involved in inflammatory response", and "myeloid leukocyte activation", amongst many others. This information is redundant since many of the genes included in these GO terms will be overlapping. REVIGO is a webserver that summarises long lists of significant GO terms and reduces redundancy [156]. REVIGO uses a simple clustering algorithm to determine the semantic similarities between all pairs of GO terms provided. Highly similar groups of GO terms are identified, and *p*-values are used to determine which term from the group of GO terms should be

retained. REVIGO was used in chapter 4 to filter the lists of significant pathways identified by g:Profiler2.

Chapter 3

3 Using a multi-platform approach to identify a host protein signature for distinguishing between bacterial and viral infections in febrile children

Managing children with fever (febrile children) is a common and important problem facing healthcare providers. The majority of febrile children attending healthcare have self-resolving viral infections; however, a small minority suffer from life-threatening bacterial infections. If left untreated, some bacterial infections can have serious downstream implications as they can lead to the escalation of an exaggerated immune response, resulting in sepsis. Antibiotics are frequently prescribed to children with viral infections out of fear of missing a bacterial infection. Current gold-standard culture-based diagnostic approaches are slow and can fail to detect the disease-causing pathogen, for example if it is residing in a non-sterile site such as the lung. Infection gives rise to changes in protein production, with specific changes induced by certain types of pathogens. We can detect these changes through studying host proteomes obtained from individuals with infections. We used high-dimensional proteomic profiling to detect key protein biomarkers that can distinguish between bacterial and viral infections in febrile children. This chapter describes the discovery and validation of these biomarkers and evaluates their diagnostic performance in independent validation datasets.

3.1 Chapter aims and contributions

In this chapter, I present the analysis of a large, prospective case-control study to identify optimal protein biomarkers for distinguishing between bacterial and viral infections in febrile children. Differentiating between self-resolving viral infections and severe bacterial infections in febrile children is a common challenge. Clinical features alone do not reliably distinguish between bacterial and viral infections. Culture-based methods are the gold-standard diagnostic approaches; however, they have various shortcomings. Obtaining results from bacterial cultures can take a long time, and their sensitivity can vary depending on antibiotic administration, stage of disease, and the type or amount of fluid collected [56-59]. Studying the host response to infection through host proteomic profiling can provide useful insights and can lead to the identification of host biomarkers of infection with diagnostic potential [157]. Diagnostic protein signatures are combinations of host proteins that, when combined, can distinguish between disease groups of interest. The work presented in this chapter forms one of the major components of the PERFORM (Personalised Risk Assessment in Febrile Illness to optimise Real-life Management across the European Union) project which aimed to identify biomarkers that could be

developed into an accurate and rapid point-of-care diagnostic test for distinguishing between bacterial and viral infections in febrile children.

Various biomarkers have been proposed for distinguishing between bacterial and viral infections in febrile children [71], including C-reactive protein (CRP). CRP is a widely used yet flawed and non-specific biomarker that is upregulated in bacterial infections in addition to various inflammatory diseases and viral infections [63, 64]. There is one commercially available rapid diagnostic test for distinguishing between bacterial and viral infections - MeMed BV™ - that measures three host proteins: CRP, interferon gamma-induced protein 10 (IP10) and TNF-related apoptosis-inducing ligand (TRAIL) [72]. The 3-protein signature included in the MeMed BV™ was identified through hypothesis-driven literature searches and targeted screening of biomarker candidates. The MeMed BV™ test was developed and optimised for use in populations of all ages (i.e., adults and children). Although it displays promising performance in paediatric populations [73], there may be alternative protein biomarkers that are optimal for diagnosing bacterial and viral infections in children.

Given the vast number of proteins in the human body, it is likely that there are many promising host protein biomarkers for distinguishing between bacterial and viral infections in febrile children that have not yet been identified and characterised in the literature. In this chapter, I use high-dimensional proteomic datasets to show that there are many host protein biomarker candidates that can distinguish between bacterial and viral infections. Using a hypothesis-free approach, I identify a shortlist of optimal protein biomarker candidates and evaluate their performance in two independent validation datasets. I show that bacterial and viral infections in febrile children can be distinguished by a sparse host protein signature.

The primary aims of this chapter are:

1. To explore the host proteomic profiles of children with bacterial and viral infections using an unbiased, hypothesis-free approach;

2. To use the proteomic profiles to identify a shortlist of protein biomarkers with diagnostic potential for discriminating between bacterial *vs*. viral infections in febrile children;

3. To reduce the shortlist of protein biomarkers through further feature selection performed in an independent cross-platform validation set, leading to the identification of a protein diagnostic signature;

4. To evaluate the performance of the protein diagnostic signature identified in Aim 3 in a further validation set composed of independent samples.

The structure of this chapter is as follows. Section 3.2 describes the clinical cohorts, the data generation pipelines, and initial dataset exploration. The results are introduced in section 3.3. The discovery phase is described in 3.3.1, including differential abundance analysis to identify lists of proteins that are significantly differentially abundant (SDA) between bacterial *vs*. viral infections, and feature selection to identify small diagnostic signatures (i.e., combinations of proteins with diagnostic potential). A shortlist of protein biomarkers was assembled in the discovery phase which underwent cross-platform validation, as described in section 3.3.2. Section 3.3.2 describes our attempts to narrow down the list of protein biomarkers and identify a small optimal protein diagnostic signature using feature selection. The small signature is taken forward to further validation as described in section 3.3.3. The results are discussed and summarised in section 3.4.

Diagnosing bacterial and viral infections in febrile children is a widespread and important challenge facing healthcare providers globally. Various host biomarkers have been proposed previously and one host diagnostic signature has been developed into a commercially available test. The work presented in this chapter is the first time that high-dimensional omic datasets have been used to identify key biomarkers and signatures for distinguishing between bacterial and viral infections using an unbiased approach. This chapter contributes a 6-protein signature that can accurately distinguish between bacterial and viral infections in febrile children. The signature has undergone extensive validation, meaning that it could be rapidly developed into a point-of-care diagnostic test with far-reaching benefits. The host protein diagnostic signature contributed in this chapter not only outperforms the only other known host protein diagnostic signature (MeMed BV™), but it also does not include CRP, a widely used yet flawed biomarker that is included in the MeMed BV™. This work is the first time that a protein diagnostic signature has been identified and validated specifically for use in paediatric populations.

Declaration of originality statement for chapter 3:

I carried out all of the work presented in this chapter, except for the following components. Dr. Shea Hamilton performed the laboratory elements for the generation of the SomaScan dataset. Dr. Stephanie Menikou, Dr. Honglei Huang, Adam Thorne and Dr. Roman Fischer performed the laboratory elements and data generation for the liquid chromatography-tandem mass spectrometry (LS-MS/MS) experiments at the Discovery Proteomics Facility, Oxford. Dr. Judith Zandstra and Prof. Taco Kuijpers performed the literature search to identify potential protein biomarkers, and they also performed protein quantification using ELISA and Luminex assays in cross-platform validation phase 1 and 2. Dr. Andrew McArdle processed the LC-MS/MS data using MaxQuant. FS-PLS was developed by Prof. Lachlan Coin.

3.2 Cohorts and study design

This study is composed of three main phases: the discovery phase; cross-platform validation phase 1; and cross-platform validation phase 2. In brief, a shortlist of potential protein biomarkers for distinguishing between bacterial and viral infections in febrile children was identified through high-throughput host protein sequencing in the discovery phase. The list of protein candidates identified in the discovery phase were validated in an independent set of samples using Luminex® and ELISA immunoassays in cross-platform validation phase 1, and feature selection was performed to reduce the number of protein biomarker candidates and identify a small diagnostic signature, i.e., a selection of proteins that can accurately distinguish between bacterial and viral infections. The proteins included in the small diagnostic signature underwent an additional round of cross-platform validation (cross-platform validation phase 2) in a further set of independent samples to evaluate the performance of the signature (Figure 7). Two rounds of cross-platform validation were performed because the protein signature identified in cross-platform validation phase 1 required a further validation study to evaluate its performance in a dataset composed of independent samples.

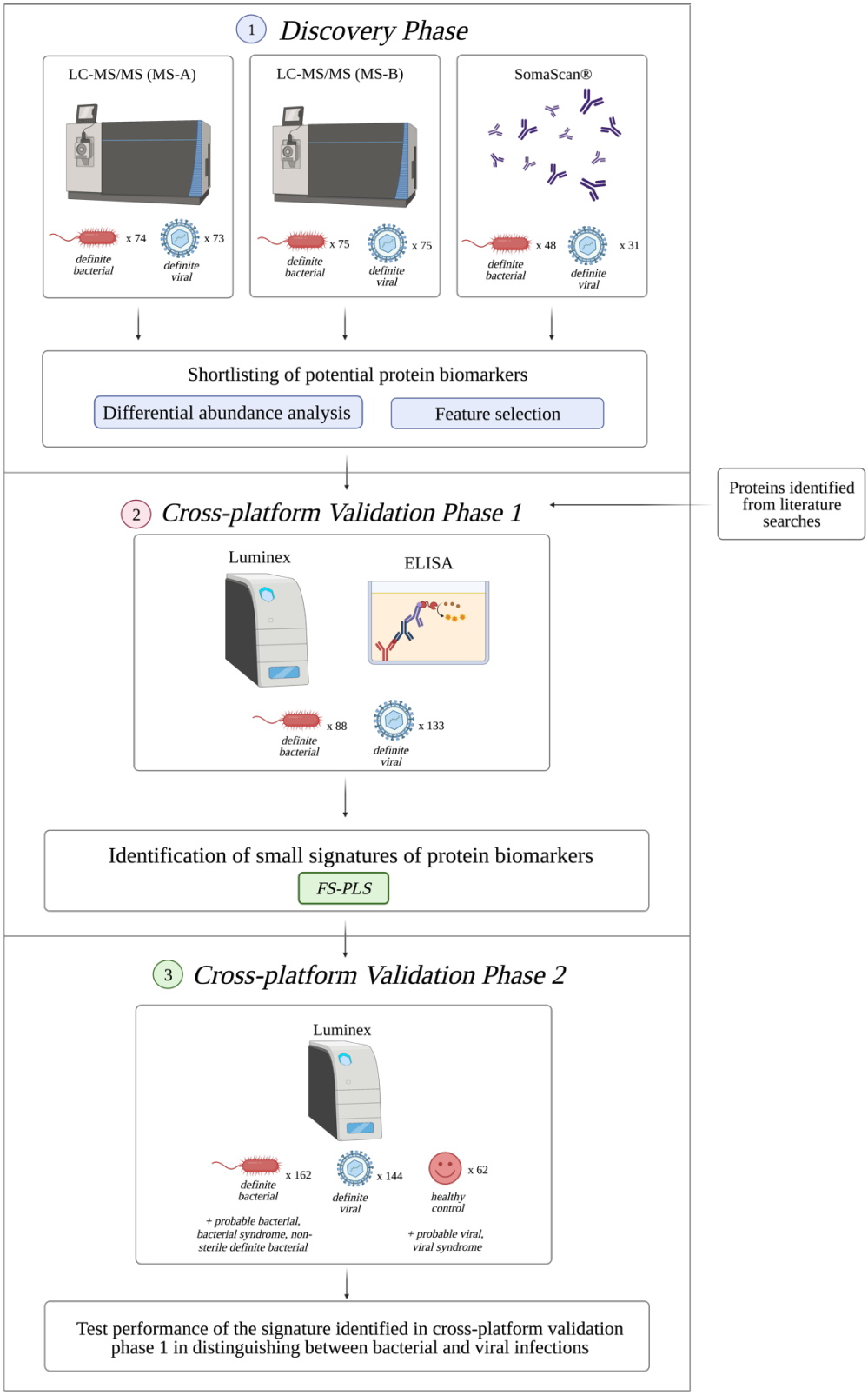


Figure 7: *An overview of the biomarker discovery and validation phases carried out and described in chapter 3. The discovery phase involved identifying a shortlist of protein biomarker candidates from high-dimensional proteomic datasets. The candidates were then validated in cross-platform validation phase 1. A sparse protein signature was identified from the proteins that were quantified in cross-platform validation phase 1. This sparse signature was subjected to another round of cross-platform validation (cross-platform validation phase 2). LC-MS/MS = liquid chromatography-tandem mass spectrometry; FS-PLS = forward selection-partial least squares; NSDB = non-sterile definite bacterial. Image created with BioRender.*

The discovery phase was made up of three proteomic datasets known as the "SomaScan", "MS-A", and "MS-B" datasets. The SomaScan and MS-A datasets were composed of patients that were recruited into the EUCLIDS study (European Union Childhood Life-Threatening Infectious Disease Study; NRES Committee London – Fulham, 11/LO/1982) [158] which was a prospective, multi-centre, European Union (EU)-funded study carried out in six countries in Europe between 2012-2015. The EUCLIDS study focused on life-threatening bacterial infections in children across Europe. EUCLIDS aimed to access the outcomes and burden of life-threatening bacterial infections through genomic and transcriptomic sequencing. Patients between the ages of 1 month – 18 years with (suspected) sepsis or severe focal infections were recruited into the EUCLIDS study. There was no overlap between the EUCLIDS patients included in the SomaScan and MS-A datasets.

The MS-B dataset was composed of patients recruited into the PERFORM study (perform2020.org/; London – Central Research Ethics Committee, 16/LO/1684) [60, 159, 160]. The PERFORM study was a prospective, multi-centre EU-funded study carried out in multiple European countries between 2016-2021. PERFORM aimed to develop a comprehensive management plan for febrile patients, including through identifying protein and gene biomarkers for diagnosing febrile children. Children were recruited into the PERFORM study if they were between 1 month to 18 years old and had a temperature ≥38.0°C or fever <72 hours before presenting to hospital. Samples in the cross-platform validation phase 1 and 2 datasets were also from patients recruited into the PERFORM study, however no samples were overlapping between any of the MS-B, cross-platform validation phase 1 and cross-platform validation phase 2 datasets.

The three independent proteomic datasets (SomaScan, MS-A, MS-B) were all used for the discovery of protein biomarkers because there were important differences between each of the datasets that meant that using all three in parallel would provide a more heterogeneous comparison of bacterial *vs*. viral infections. The datasets were generated using two very different platforms. SomaScan®, which was used to generate the SomaScan dataset, is a targeted proteomic quantification platform. Liquid chromatography-tandem mass spectrometry (LC-MS/MS), which was used to generate the MS-A and MS-B datasets, is an untargeted approach to protein quantification. Through using the SomaScan dataset, we are benefiting from a) the large number of proteins that it quantifies and b) its highly accurate and sensitive approach to quantification. However, SomaScan is dependent on the set of protein targets included in the assay. Some potentially highly discriminatory markers that are detectable in the patients but not included in the assay would be missed. Henceforth, the LC-MS/MS (MS-A and MS-B) datasets were also used. Furthermore, as outlined above, the patient cohorts that make up the three datasets differ due to the studies that they were recruited into. As such, the range of pathogens that are causing disease and the spectrum of disease severity varies between datasets. By using multiple datasets with a wide range of severity and cause of disease, we are more likely to

identify protein biomarkers that are effective irrespective of these factors and are not driven by confounding variables such as severity.

Table 1: *Clinical and laboratory features of patients whose samples were included in the discovery of novel protein biomarkers for distinguishing between bacterial and viral infections in febrile children. The SomaScan dataset was generated using the SomaScan platform. MS-A and MS-B were generated from liquid chromatography tandem mass spectrometry. DB = definite bacterial; DV = definite viral; IQR = interquartile range; RSV = respiratory syncytial virus; EBV = Epstein-Barr virus; CRP = C-reactive protein.*

	SomaScan		**MS-A**		**MS-B**	
	DB (*n*=48)	DV (*n*=31)	DB (*n*=74)	DV (*n*=73)	DB (*n*=75)	DV (*n*=75)
Age, median (IQR), months	25 (9-77)	11 (4-35)	35 (10-60)	9 (2-23)	57 (12-101)	37 (10-83)
Females, no. (%)	24 (50%)	17 (55%)	47 (64%)	35 (48%)	33 (44%)	33 (44%)
Duration of symptoms, median (IQR), days	3 (2-7)	3 (1-4)	2 (1-4)	3 (2-7)	2 (1-5)	3 (1-6)
CRP, median (IQR), mg/L	124 (68-226)	15 (5-35)	108 (39-221)	18 (6-27)	73 (21-147)	10 (3-22)
Platelets, median (IQR), $\times 10^9$/L	216 (128-334)	249 (205-376)	228 (158-351)	306 (228-388)	263 (207-356)	274 (199-339)
Lymphocytes, median (IQR), $\times 10^9$/L	1.8 (1.1-3.6)	2.0 (1.4-3.1)	1.5 (0.6-2.7)	2.2 (1.7-3.1)	1.4 (1.0-3.0)	2.5 (1.6-4.1)
Neutrophils, median (IQR), $\times 10^9$/L	8.7 (3.6-14.1)	6.3 (2.9-10.2)	4.7 (1.8-12.9)	3.7 (2.7-6.4)	11.2 (6.1-17.3)	4.7 (2.7-7.3)
Monocytes, median (IQR), $\times 10^9$/L	0.8 (0.4-1.3)	1.0 (0.7-1.2)	1.5 (0.7-3.0)	0.9 (0.4-1.2)	1.1 (0.6-1.4)	0.8 (0.5-1.1)
Top causative pathogens or disease	*N. meningitidis* (*n*=11) *S. aureus* (*n*=8) *S. pneumoniae* (*n*=5) *S. pyogenes* (*n*=5)	Enterovirus (*n*=6) RSV (*n*=5) Rhinovirus (*n*=4)	*N. meningitidis* (*n*=28) *S. pneumoniae* (*n*=13) *S. aureus* (*n*=9) *E. coli* (*n*=7)	RSV (*n*=22) Enterovirus (*n*=13) Rhinovirus (*n*=9) Influenza (*n*=6)	*E. coli* (*n*=21) *S. pyogenes* (*n*=14) *S. aureus* (*n*=9) *N. meningitidis* (*n*=5)	Influenza virus (*n*=19) Rhinovirus (*n*=10) Adenovirus (*n*=9) EBV (*n*=7)
Study of origin	EUCLIDS	EUCLIDS	EUCLIDS	EUCLIDS	PERFORM	PERFORM
Sample type	Serum	Serum	Plasma	Plasma	Plasma	Plasma

Clinical and laboratory features of the patients whose samples were included in any of the discovery datasets are shown in Table 1. Patients with definite bacterial infections (DB group) were older than the patients with definite viral infections (DV group) across all three datasets. The median duration of symptoms at the time of sampling was similar across all disease groups and datasets at either 2 or 3 days. Across all datasets, CRP and neutrophil levels were higher in patients with bacterial infections than those with viral infections.

The most common causative pathogens differed between datasets. The DB patients in the SomaScan and MS-A dataset were recruited into the EUCLIDS study which focused on severe bacterial infections. As a result, the top bacterial pathogens in the SomaScan and MS-A datasets differ to the top bacterial pathogens in the patients in the MS-B dataset who were recruited into the PERFORM study. For example, 23% (*n*=11) and 38% (*n*=28) of the DB patients in the SomaScan and MS-A datasets,

respectively, were infected with *N. meningitidis* in comparison to just 7% (*n*=5) of the DB patients in the MS-B dataset.

Table 2: *Clinical and laboratory features of patients whose samples were included in cross-platform validation phase 1. The levels of protein biomarker candidates identified in the discovery phase were tested using these samples. ELISA and Luminex immunoassays were used for protein quantification. DB = definite bacterial; DV = definite viral; IQR = interquartile range; CRP = C-reactive protein; RSV = respiratory syncytial virus.*

	Cross-platform validation phase 1	
	DB (*n*=88)	DV (*n*=113)
Age, median (IQR), months	50 (13-115)	38 (13-76)
Females, no. (%)	45 (51%)	51 (45%)
Duration of symptoms, median (IQR), d	2 (1-4)	3 (1-5)
CRP, median (IQR), mg/L	103 (33-202)	10 (4-24)
Platelets, median (IQR), $\times 10^9$/L	264 (166-395)	270 (201-349)
Lymphocytes, median (IQR), $\times 10^9$/L	2.1 (1.1-3.5)	2.8 (1.7-3.8)
Neutrophils, median (IQR), $\times 10^9$/L	11.4 (5.5-17.6)	5.1 (2.7-6.5)
Monocytes, median (IQR), $\times 10^9$/L	1.0 (0.6-1.6)	0.8 (0.5-1.1)
Causative pathogens or disease	*E. coli* (*n*=16) *S. pyogenes* (*n*=11) *S. aureus* (*n*=11) *N. meningitidis* (*n*=11) *S. pneumoniae* (*n*=9) Other (*n*=30)	Adenovirus (*n*=19) Influenza (*n*=18) Rhinovirus (*n*=15) RSV (*n*=12) Enterovirus (*n*=12) Other (*n*=37)

The patients with bacterial infections in cross-platform validation 1 (Table 2) and cross-platform validation phase 2 (Table 3) were older than the patients with viral infections, and they also had higher CRP and neutrophil levels. The median duration of symptoms when the samples were taken was similar at either 2 or 3 days for bacterial and viral patients across both datasets. These trends were also observed in the discovery datasets (Table 1).

The top causative bacterial pathogens in cross-platform validation phases 1 and 2 were very similar to those present in the MS-B dataset. For example, *E. coli* was the most common bacterial pathogen in all three datasets, with 18% (*n*=16) of the DB group in the cross-platform validation phase 1 dataset, 43% (*n*=70) of the DB group in cross-platform validation phase 2, and 28% (*n*=21) of the DB group in the MS-B discovery dataset. Despite the differences in frequency of pathogens across datasets, *S. aureus* was consistently amongst the top causative bacterial pathogens in all 5 datasets (all three discovery datasets; cross-platform validation phase 1; cross-platform validation phase 2). Similarly, rhinovirus was consistently amongst the top causative viral pathogens in all 5 datasets.

Clinical and laboratory parameters for the "uncertain" phenotypic groups are shown in Table 3. Overall, the patients with suspected bacterial infections (probable bacterial, bacterial syndrome) and suspected viral infections (probable viral, viral syndrome) followed a similar trend to the DB and DV

groups, respectively. The suspected bacterial patients were older than the suspected viral patients, and they had higher CRP and neutrophil levels than the probable viral patients.

Table 3: *Clinical and laboratory features of patients whose samples were included in cross-platform validation phase 2. Luminex immunoassays were used for protein quantification. The protein signature identified in cross-platform validation phase 1 was tested in these individuals. DB = definite bacterial; NSDB = non-sterile definite bacterial; DV = definite viral; CRP = C-reactive protein; IQR = interquartile range; RSV = respiratory syncytial virus.*

	DB (*n*=162)	**NSDB (*n*=31)**	**Probable bacterial (*n*=64)**	**Bacterial syndrome (*n*=2)**	**DV (*n*=144)**	**Probable viral (*n*=75)**	**Viral syndrome (*n*=12)**
Age, median (IQR), m	53 (12-120)	116 (35-188)	47 (19-101)	71 (47-71)	44 (10-92)	29 (9-64)	27 (10-44)
Females, no. (%)	73 (45%)	11 (36%)	28 (44%)	0 (0)	72 (50%)	31 (41%)	4 (33.3%)
Duration of symptoms, median (IQR), d	3 (1-5)	2 (2-7)	3 (2-5)	4 (3-4)	3 (1-5)	2 (1-4)	3 (2-6)
CRP, median (IQR), mg/L	89 (33-175)	79 (25-129)	107 (54-179)	8 (8-8)	7 (3-18)	4 (2-11)	9 (3-30)
Platelets, median (IQR), ×10⁹/L	284 (230-396)	295 (215-365)	285 (220-397)	265 (259-270)	250 (192-337)	285 (222-371)	353 (309-407)
Lymphocytes, median (IQR), ×10⁹/L	2.3 (1.2-3.3)	1.8 (1.5-2.6)	2.2 (1.3-3.3)	3.2 (1.9-4.4)	2.3 (1.3-3.8)	3.1 (1.8-4.8)	3.6 (2.8-5.4)
Neutrophils, median (IQR), ×10⁹/L	10.0 (6.7-16)	5.6 (4.2-8.8)	11.7 (7.4-17)	7.0 (6.9-7.1)	3.6 (2.2-6.3)	3.8 (2.4-6.7)	13.3 (8.5-16.6)
Monocytes, median (IQR), ×10⁹/L	1.2 (0.7-1.6)	0.9 (0.6-1.1)	1.3 (0.8-2)	0.8 (0.8-0.8)	0.6 (0.4-1.0)	0.7 (0.4-1.0)	1.7 (1.0-2.0)
Causative pathogens or disease	*E. coli* (*n*=70) *S. aureus* (*n*=31) *S. pneumoniae* (*n*=23) *S. pyogenes* (*n*=11) Other (*n*=27)	*Salmonella spp.* (*n*=11) *Mycoplasma* (*n*=8) *Campylobacter* (*n*=6) *M. tuberculosis* (*n*=3) Other (*n*=3)	*Streptococcus* (*n*=10) *Staphylococcus* (*n*=7) *C. difficile* (*n*=3) *Campylobacter* (*n*=3) Other (*n*=41)	NA	Influenza virus (*n*=45) EBV (*n*=21) Enterovirus (*n*=19) RSV (*n*=18) Other (*n*=41)	Rhinovirus (*n*=7) Measles (*n*=7) Enterovirus (*n*=4) EBV (*n*=4) Other (*n*=53)	Rhinovirus (*n*=3) RSV (*n*=2) Adenovirus (*n*=2) Influenza (*n*=2) Other (*n*=3)

3.2.1 Case definitions

Patients were classified based on the phenotyping algorithm described in [67, 160] and shown in Figure 8. A definite bacterial (DB) classification required isolation of a sterile-site pathogenic bacteria matching the syndrome whilst a definite viral (DV) classification required isolation of a virus matching the syndrome in addition to CRP levels ≤60 mg/L and neutrophil levels ≤12×10⁹/ml. Healthy control (HC) patients were enrolled in the EUCLIDS and PERFORM studies. Cross-platform validation phase 2 included patients in the non-sterile bacterial (NSDB), probable bacterial, bacterial syndrome, probable viral, viral syndrome groups, which were classified as shown in Figure 8. NSDB classification required isolation of a pathogen that is normally only detected on mucosal surfaces/non-sterile sites, or a pathogen that is diagnosed serologically, e.g., *M. tuberculosis, B. pertussis*, *M. pneumoniae*, *Borrelia* species, *Campylobacter* and *Salmonella*. Patients in the NSDB group in addition to the patients in the "uncertain groups" (e.g., probable bacterial, probable viral, bacterial syndrome, viral syndrome) were

only included in cross-platform validation phase 2 to test the performance of the signature in these groups. They were not used for signature discovery as the aim was to identify a "pure" signature derived from DB and DV cases.

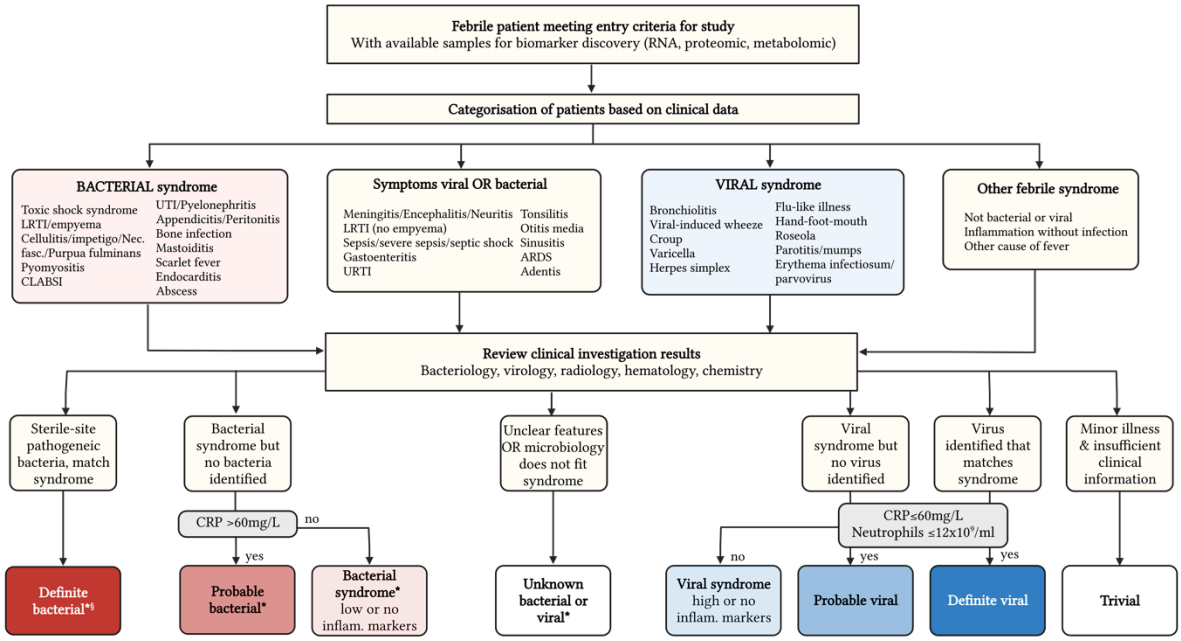


Figure 8: *The clinical phenotyping algorithm used to classify patients included in the PERFORM and DIAMONDS studies. Patient groups with * indicate that these patients could have a viral coinfection. This phenotyping algorithm was used in the PERFORM and DIAMONDS studies. The figure was adapted from [160].*

3.2.2 Data generation

3.2.2.1 Discovery datasets

3.2.2.1.1 The SomaScan dataset

Proteomic profiling for 1,301 human proteins was carried out in serum samples using the SomaScan® 1.3K multiplexed assay (SomaLogic, Inc.; 1.3K Assay) [40] based on protein capture using modified aptamers (slow off-rate modified aptamers, SOMAmers®). SomaScan is a targeted protein quantification assay. All samples were diluted into three concentrations (40%, 1%, 0.05%) to detect proteins with low, medium and high abundances. Samples were split across three plates with DB and DV samples present on each plate in relative proportions. SomaLogic, the company that perform the SomaScan® assay, return the data in a normalised format, however further batch effect correction and quality control (QC) steps were performed, including log2 transformation, principal component analysis (PCA), and batch effect removal by COCONUT [161], a batch effect correction tool that is described in chapter 2 section 2.3.

3.2.2.1.2 MS-A and MS-B datasets

The proteomic profiles for the MS-A and MS-B cohorts were quantified from plasma samples using liquid chromatography tandem mass spectrometry (LC-MS/MS). LC-MS/MS is an untargeted approach to protein quantification that separates proteins using liquid chromatography (LC) and analyses the mass of peptide fragments using two mass spectrometers. Prior to protein quantification, a top 12 abundant protein depletion spin column (Catalog no. 85165, ThermoFisher Scientific) was used to remove 12 highly abundant proteins: α1-acid glycoprotein, α1-antitrypsin, α2-macroglobulin, albumin, apolipoprotein AI, apolipoprotein AII, fibrinogen, haptoglobin, Immunoglobulin A, Immunoglobulin G, Immunoglobulin M and transferrin. Without this step, these proteins would dominate the proteins quantified. After peptide clean up, 0.5µg of peptides was injected into a LC-MS/MS (LUMOS Fusion, Thermo Fisher scientific) at the Discovery Proteomics Facility, Oxford (https://www.tdi.ox.ac.uk/research/research/tdi-mass-spectrometry-laboratory/mass-spectrometry). Samples were ordered according to disease group. To avoid bias by system error, the performance of LC-MS/MS (sensitivity of detection, retention time shifting, number of proteins identified) was monitored by running a quality control sample (pool of all individual samples) every 10 samples within batch. A blank in between runs was included across the entire batch which cleans the column and avoids sample carry over.

The raw dataset files for MS-A and MS-B were processed separately by MaxQuant (1.6.10.43) [47] based on default settings with matching between runs activated (20 min tolerance) and contaminants included. The false-discovery rate (FDR) cut-off for proteins and peptides was 1% with razor protein FDR enabled and the variable modifications were methionine oxidation and protein N-terminal acetylation. Maximum peptide mass was 4600Da. Second peptide search was enabled. Relative quantification was performed using the MaxLFQ algorithm [121]. Following MaxLFQ, the datasets underwent QC processes including removing contaminant proteins and lowly expressed proteins, log2 transforming the datasets and performing PCA.

3.2.2.2 Cross-platform validation phase 1

Proteins identified in the discovery phase as having diagnostic potential were taken forward to cross-platform validation phase 1, in which their levels were quantified in plasma samples using either Luminex® or ELISA immunoassays, two protein detection methods described in chapter 1 sections 1.3.2.3.2 and 1.3.2.2.2, respectively. CRP, human neutrophil elastase-α1-antitrypsin (HNE-α1AT) complexes, Mannan-binding lectin serine protease 1 (MASP1; Cloud-Clone Corp., product no: SEB895Hu), interferon-stimulating gene 15 (ISG15; FineTest, product no. EH1673), secreted phosphoprotein 2 (SPP24; Cusabio, product no. CSB-EL022604HU), anti-thrombin 3 (AT3; R&D biosystems, product no. DY1267-05) and serum amyloid A1 (SAA1; R&D Biosystems, product no. DY3019-05) concentrations were measured by ELISA according to the manufacturer's protocol.

Proteins were quantified using ELISA if they were not available via Luminex®. The other selected biomarkers were analysed using a Customized Luminex® human cytokine multiplex panel with custom analytes according to the manufacturer's protocol (R&D Systems, Inc, Bio-Techne, Minneapolis MN, USA). Protein dilutions were 10-fold across proteins and samples were analysed across 96-well plates. The fluorescence responses and concentrations of analytes were obtained using a Bio-Plex 200 system and the accompanying Bio-Plex Manager Software 6.0. The concentration values and detection limits were determined from standard curves generated from each kit's standards using the Bio-Plex Software Manager 6.0 (BioRad).

Extrapolated protein measurements above or below the standard range were excluded to avoid artificially skewing the data. Values were scaled and plate effects from the Luminex® assays were removed using ComBat [129]. PCA was performed and values were log2 transformed. During cross-platform validation phase 1, a small protein signature was identified using feature selection and these proteins included in the signature were taken forward to cross-platform validation phase 2.

3.2.2.3 Cross-platform validation phase 2

Proteins that were selected in the signature derived during cross-platform validation phase 1 were measured in cross-platform validation phase 2. The same experimental protocols and QC/normalisation steps were performed as described above in section 3.2.2.2.

3.2.3 Dataset exploration

PCA was performed on each of the discovery datasets (Figure 9 - Figure 11) prior to identifying protein biomarkers for distinguishing between bacterial and viral infections. For the SomaScan dataset (Figure 9), disease group is captured on PC1, with the viral patients forming a tight cluster (Figure 9A). PC3 and PC4 also worked together to capture disease group (Figure 9D), with the viral patients clustering towards the top of PC4 and the left of PC3 (i.e., lower PC3 values).

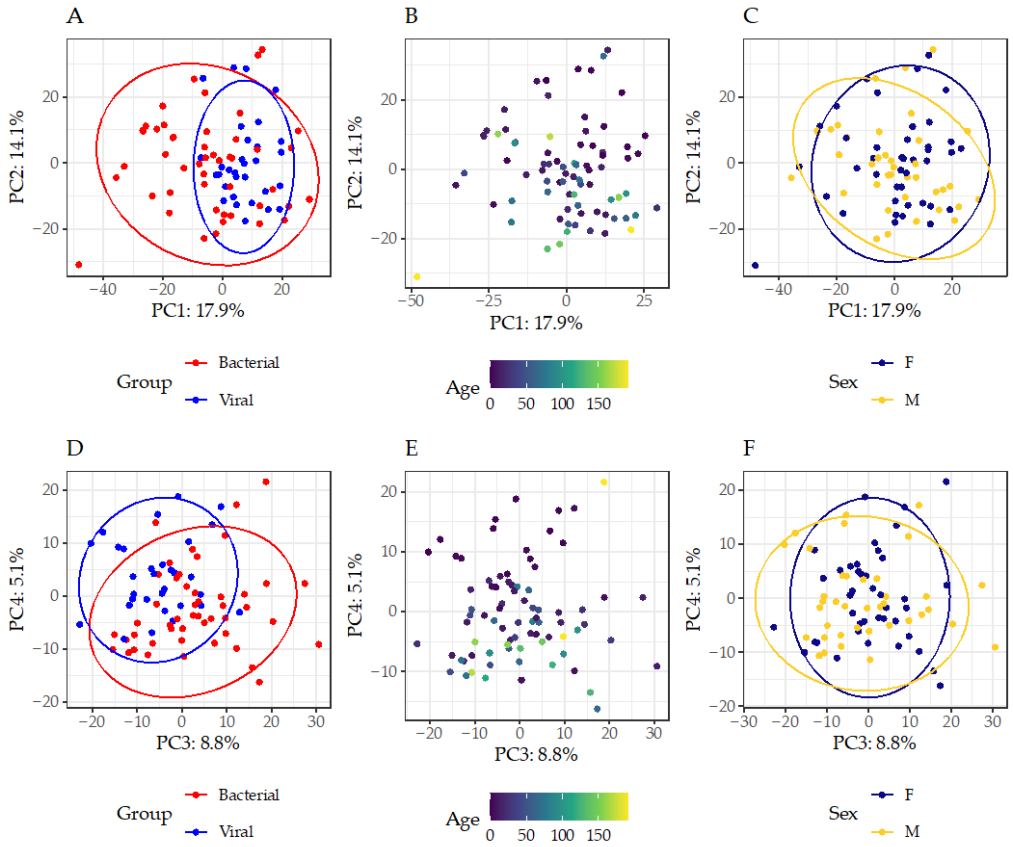


Figure 9: *Principal component analysis (PCA) plots for principal components (PCs) 1 and 2 (**A**-**B**) and 3 and 4 (**D**-**F**) for the SomaScan dataset. This dataset was used as a discovery dataset for the identification of candidate protein biomarkers for distinguishing between bacterial and viral infections in febrile children. Points are coloured according to disease group (**A**, **D**), age in months (**B**, **E**), and sex at birth (**C**, **F**).*

For the MS-A dataset, disease group was captured on PC2 with the bacterial patients clustering towards the top of PC2 (Figure 10A). There are outlier patients towards the right of PC1 (i.e., higher PC1 values). Upon further investigation of these patients, three of the patients with PC1 values >25 eventually died, suggesting that PC1 is capturing severity of disease. For reference, a total of 10 patients died overall in the MS-A dataset. Given the biological significance of these patients, they were not removed. Disease group is captured on PC3 and PC4 with visible stratification between the bacterial and viral patients (Figure 10D). PC3 is capturing age, with older patients having higher PC3 and lower PC4 scores (Figure 10E).

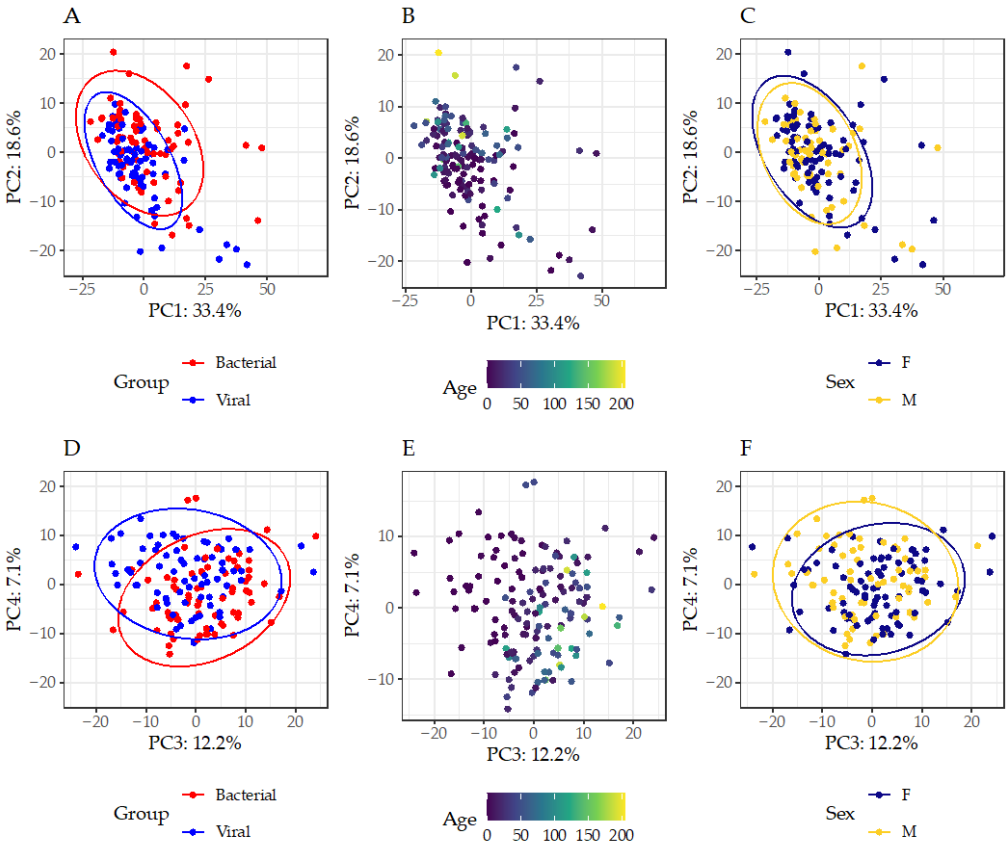


Figure 10: *Principal component analysis (PCA) plots for principal components (PCs) 1 and 2 (**A**-**B**) and 3 and 4 (**D**-**F**) for the MS-A dataset. This dataset was used as a discovery dataset for the identification of candidate protein biomarkers for distinguishing between bacterial and viral infections in febrile children. Points are coloured according to disease group (**A**, **D**), age in months (**B**, **E**), and sex at birth (**C**, **F**).*

For the MS-B dataset, age is captured on PC1 with the older patients towards the left of PC1 (i.e., with small PC1 values) (Figure 11B). Disease group was also captured on PC1 with the bacterial patients having lower PC1 values. This fits in with the findings in Table 1 that the bacterial patients were older than the viral patients in this cohort. Disease group is also captured on PCs 3 and 4.

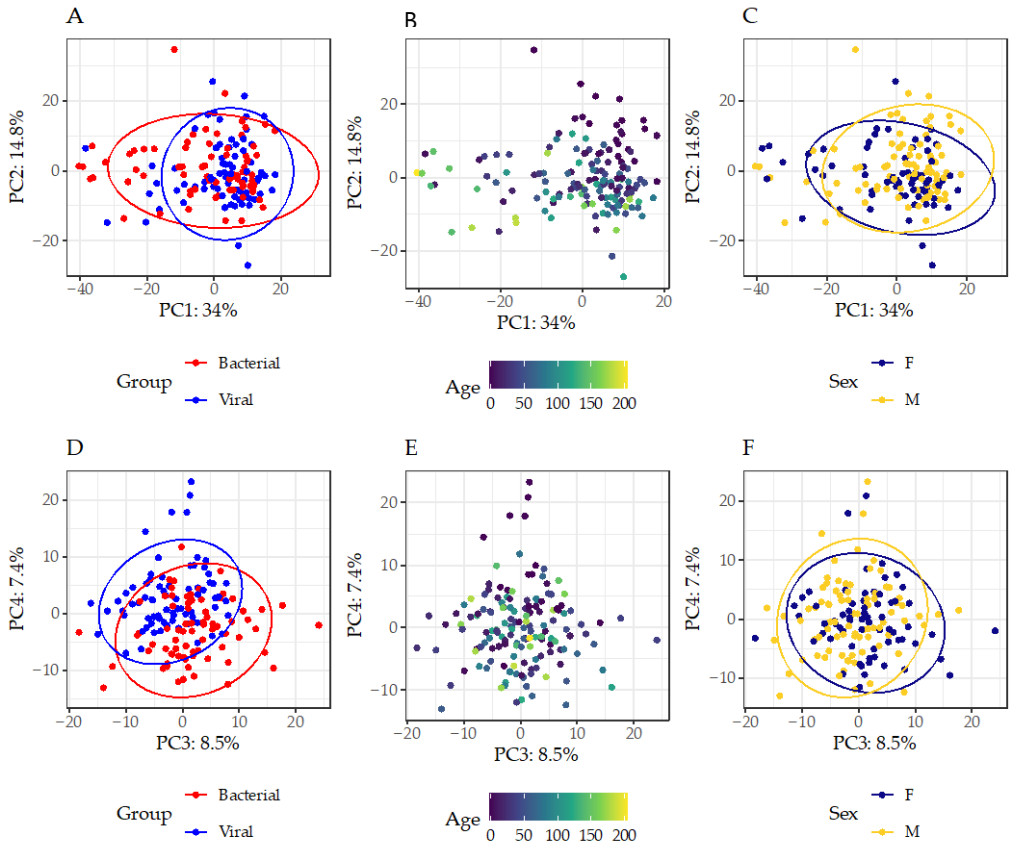


Figure 11: *Principal component analysis (PCA) plots for principal components (PCs) 1 and 2 (**A-B**) and 3 and 4 (**D-F**) for the MS-B dataset. This dataset was used as a discovery dataset for the identification of candidate protein biomarkers for distinguishing between bacterial and viral infections in febrile children. Points are coloured according to disease group (**A**, **D**), age in months (**B**, **E**), and sex at birth (**C**, **F**).*

3.3 Results

The PERFORM study aimed to develop a comprehensive plan for the management of febrile children through identifying and validating protein biomarkers for distinguishing between patients with confirmed bacterial and viral infections. We have performed a robust protein biomarker identification process using three independent multi-platform discovery datasets (Figure 7) which is described in section 3.3.1. The top candidates identified in the discovery process underwent the first round of cross-platform validation in an independent set of patients (section 3.3.2), in which further feature selection was performed to identify a sparse host protein signature. The sparse host protein signature was further validated (section 3.3.3) in an additional independent set of patients with confirmed bacterial and viral infections alongside patients with unconfirmed or suspected infection.

3.3.1 Discovery phase

Three independent proteomic datasets were used for the discovery of protein biomarkers (Table 1). There were important differences between the three discovery datasets that led to the decision to

perform analyses on each dataset separately. Not only were the datasets generated using different platforms (SomaScan dataset: SomaScan platform; MS-A and MS-B: LC-MS/MS), the SomaScan dataset was composed of serum samples whilst the MS-A and MS-B datasets were composed of plasma samples. Despite the similarities between the MS-A and MS-B datasets, they were not merged due to the differences in the cohorts of bacterial and viral patients in the datasets (Table 1). The SomaScan and MS-A datasets included bacterial and viral patients recruited into the EUCLIDS study which was focused on severe bacterial infections. These patients were typically more severely unwell than the bacterial and viral patients included in the MS-B dataset from the PERFORM study, making it difficult to merge the two datasets.

Differential abundance analysis was performed to identify proteins that were significantly differentially abundant (SDA). In differential abundance analysis, generalised linear models (GLM) are applied to each feature to compare the abundance of the feature between two groups of interest. Limma [141] was used for differential abundance analysis contrasting bacterial *vs*. viral infections in each dataset independently, with Benjamini Hochberg [142] (BH)-adjusted *p*-values <0.05 considered significant. Age and sex were included in the Limma models for all datasets, and plate was included as an additional covariate in the SomaScan dataset since multiple plates were used. A total of 431 proteins were SDA between bacterial *vs*. viral infections in the SomaScan dataset (proteins=1,300), with 198 proteins more abundant in bacterial infections and 233 proteins less abundant in bacterial infections *vs*. viral infections (Figure 12A). In the MS-A dataset (proteins=368), 54 proteins were SDA between bacterial *vs*. viral infections with 20 proteins more abundant in bacterial infections and 34 proteins less abundant in bacterial infections *vs*. viral infections (Figure 12B). In the MS-B dataset (proteins=410), 97 proteins were SDA between bacterial *vs*. viral infections with 28 proteins more abundant in bacterial infections and 69 proteins less abundant in bacterial infections *vs*. viral infections (Figure 12C). 16 proteins were SDA between bacterial *vs*. viral infections in all three datasets with concordant log-fold change (LFC) directions. 6 of the 16 proteins increased and 9 decreased in bacterial infections compared to viral infections (Table 4).

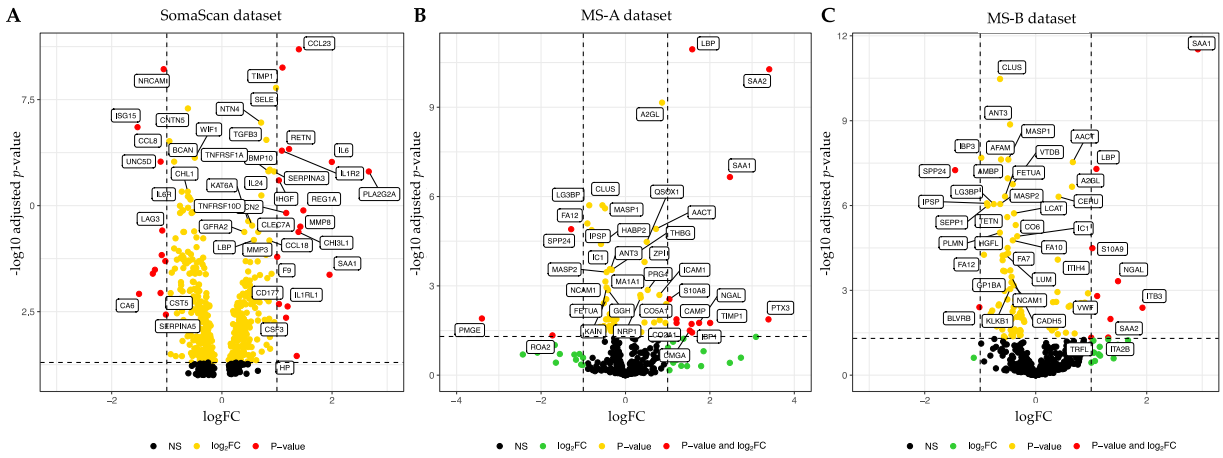


Figure 12: *Volcano plots showing the log2 fold-change (logFC) values and the -log10 BH p-values for proteins significantly differentially abundant between bacterial vs. viral infections in febrile children. Results are shown for the SomaScan (**A**), MS-A (**B**), and MS-B (**C**) datasets. Red points are proteins with BH p-values <0.05 and absolute logFC values >1; gold points are proteins with BH p-values <0.05 and absolute logFC <1; green points are proteins with absolute logFC >1 and BH p-values >0.05; black points are not significant (NS). Positive and negative logFC indicate that proteins increase and decrease in bacterial patients, respectively.*

Whilst it is highly encouraging that, given the differences between the three discovery datasets, there are features that are consistently SDA between bacterial *vs*. viral infections, we also wanted to identify the most important combinations of features in each of the datasets individually. Forward selection-partial least squares (FS-PLS) is a feature selection method that is described in detail in chapter 2 section 2.4.3. FS-PLS was applied to each dataset to reduce the search space and identify the optimal combinations of proteins that can distinguish between bacterial and viral infections. FS-PLS was applied to each dataset across 100 iterations, with a different test-training split (at 30:70) used each time. Within each iteration, potential signatures were discovered from the training set with signature performance evaluated using the test set. The signature with the highest area under the receiver operating characteristic (ROC) curve (AUC) in the test set for bacterial *vs*. viral was selected within each iteration. At the end of the 100 iterations, the number of times each signature was selected across the 100 iterations was quantified, as well as the number of times each individual protein was included in any of the 100 signatures. The latter was divided by the total number of iterations performed to return a "robustness" score. This approach – using cross validation with feature selection – was performed to ensure that the signature and features selected for each dataset were robust across the dataset and not driven by a subset of samples.

When iterative FS-PLS was applied to the SomaScan dataset, a 3-protein signature was selected the most frequently (5/100 iterations), composed of ISG15 (robustness: 0.92), TIMP Metallopeptidase Inhibitor 1 (TIMP1; robustness: 0.42), and UL16 Binding Protein 3 (ULBP3; robustness: 0.05). When iterative FS-PLS was applied to the MS-A dataset, a four-protein signature was selected the most

frequently (5/100 iterations), composed of liposaccharide binding protein (LBP; robustness: 0.55), clusterin (CLUS; robustness: 0.38), apolipoprotein H (APOH; robustness: 0.32), and histidine-rich glycoprotein (HRG; robustness: 0.08). When iterative FS-PLS was applied to the MS-B dataset, a five-protein signature was selected the most frequently (7/100 iterations), composed of AT3 (robustness: 0.95), ceruloplasmin (CERU; robustness: 0.49), SPP24 (robustness: 0.29), apolipoprotein C1 (APOC1; robustness: 0.17) and alpha 1-antichymotrypsin (AACT; robustness: 0.09).

We put together a shortlist of potential protein biomarkers for distinguishing between bacterial and viral infections (Table 4) using results from each of the three discovery datasets. Proteins were included in the shortlist based on the following criteria:

1. Proteins identified as SDA between bacterial *vs*. viral infections in all three datasets;

2. The top 5 most SDA proteins between bacterial *vs*. viral infections from each dataset individually;

3. Proteins included in the most frequently selected signature across the 100 iterations for each dataset individually;

4. The top 5 proteins with the highest robustness scores across the 100 iterations for each dataset individually.

16 proteins fulfilled criteria 1: AACT, CLUS, CO7, FA5, FBLN3, FETUA, IPSP, LBP, LG3BP, MASP1, NCAM1, NGAL, NRP1, PLG, SAA1, ZPI. 11 proteins fulfilled criteria 2: A2GL, AT3, CLUS, CNTN5, IGFBP3, LBP, MPIF1, SAA1, SAA2, SELE, TIMP1. 12 proteins fulfilled criteria 3: AACT, AT3, APOC1, APOH, CERU, CLUS, HRG, ISG15, LBP, SPP24, TIMP1, ULBP3. 14 proteins fulfilled criteria 4: A2GL, AT3, APOH, CERU, CLUS, ISG15, KAIN, LBP, MASP1, MASP2, SELE, SPP24, TIMP1, TNF sR-I. Since some proteins fulfilled >1 of the criteria, the shortlist was composed of 35 proteins, with 18 that increased and 17 that decreased in bacterial infections *vs*. viral infections (Table 4).

Table 4: *The shortlist of proteins candidates for distinguishing between bacterial and viral infections in febrile children. These proteins were identified from the three datasets used in the discovery phase: SomaScan, MS-A and MS-B datasets. SDA = significantly differentially abundant. Protein symbols with * indicate that the ELISA/Luminex assays were available for this protein.*

Symbol	Protein Name	Uniprot	Up in…	Reason(s) for inclusion
A2GL	Leucine-rich alpha-2-glycoprotein	P02750	Bacterial	MS-A Limma top 5; MS-A most robust
AACT	Alpha-1-antichymotrypsin	P01011	Bacterial	SDA in all 3 datasets; MS-B top signature
AFM	Afamin	P43652	Viral	MS-B top 5 SDA
AT3*	Antithrombin-III	P01008	Viral	MS-B top signature; MS-B top 5 SDA; MS-B most robust
APOC1	Apolipoprotein C-I	P02654	Bacterial	MS-B top signature
APOH	Beta-2-glycoprotein 1	P02749	Bacterial	MS-A top signature; MS-A most robust
CERU	Ceruloplasmin	P00450	Bacterial	MS-B top signature; MS-B most robust
CLUS*	Clusterin	P10909	Viral	SDA in all 3 datasets; MS-A top signature; MS-A top 5 SDA; MS-B top 5 SDA; MS-A most robust
CNTN5	Contactin-5	O94779	Viral	SomaScan top 5 SDA
CO7	Complement component C7	P10643	Viral	SDA in all 3 datasets
FA5	Coagulation factor V	P12259	Viral	SDA in all 3 datasets
FBLN3	EGF-containing fibulin-like extracellular matrix protein 1	Q12805	Bacterial	SDA in all 3 datasets
FETUA*	Alpha-2-HS-glycoprotein	P02765	Viral	SDA in all 3 datasets
HRG	Histidine-rich glycoprotein	P04196	Bacterial	MS-A top signature
IGFBP3	Insulin-like growth factor-binding protein 3	P17936	Viral	MS-B top 5 SDA
IPSP	Plasma serine protease inhibitor	P05154	Viral	SDA in all 3 datasets
ISG15*	Ubiquitin-like protein ISG15	P05161	Viral	SomaScan top signature; SomaScan most robust
KAIN	Kallistatin	P29622	Viral	MS-B most robust
LBP*	Lipopolysaccharide-binding protein	P18428	Bacterial	SDA in all 3 datasets; MS-A top signature; MS-A top 5 SDA; SomaScan most robust; MS-A most robust
LG3BP*	Galectin-3-binding protein	Q08380	Viral	SDA in all 3 datasets
MASP1	Mannan-binding lectin serine protease 1	P48740	Viral	SDA in all 3 datasets; MS-A most robust
MASP2*	Mannan-binding lectin serine protease 2	O00187	Viral	MS-B most robust
MPIF1	C-C motif chemokine 23	P55773	Bacterial	SomaScan top 5 SDA
NCAM1*	Neural cell adhesion molecule 1	P13591	Viral	SDA in all 3 datasets
NGAL*	Neutrophil gelatinase-associated lipocalin	P80188	Bacterial	SDA in all 3 datasets
NRP1	Neuropilin-1	O14786	Bacterial	SDA in all 3 datasets
PLG	Plasminogen	P00747	Viral	SDA in all 3 datasets
SAA1*	Serum amyloid A-1	P0DJI8	Bacterial	SDA in all 3 datasets; MS-A top 5 SDA; MS-B Limma top 5
SAA2	Serum amyloid A-2	P0DJI9	Bacterial	MS-A top 5 SDA
SELE*	E-Selectin	P16581	Bacterial	SomaScan top 5 SDA; SomaScan most robust
SPP24*	Secreted phosphoprotein 24	Q13103	Viral	MS-B top signature; MS-B most robust
TIMP1*	Metalloproteinase inhibitor 1	P01033	Bacterial	SomaScan top signature; SomaScan top 5 SDA; SomaScan most robust
TNF sR-I	Tumor necrosis factor receptor superfamily member 1A	P19438	Bacterial	SomaScan most robust
ULBP3	UL16-binding protein 3	Q9BZM4	Bacterial	SomaScan top signature
ZPI	Protein Z-dependent protease inhibitor	Q9UK55	Bacterial	SDA in all 3 datasets

The proteins identified in the discovery phase and shown in Table 4 were selected for cross-platform validation. Following the discovery phase, we planned to validate the proteins using Luminex® or

ELISA immunoassays. A total of 13 of the proteins identified had either ELISA or Luminex® assays available (indicated by * in Table 4).

In parallel to the high-dimensional, unbiased signature discovery process, a literature search was performed to identify potential protein biomarkers for distinguishing between bacterial and viral infections (performed by Judith Zandstra and Prof. Taco Kuijpers, Sanquin/Amsterdam UMC, The Netherlands). 16 proteins were identified during this process (Table 5) with 9 proteins increasing and 6 proteins decreasing in bacterial *vs.* viral infections. Only secreted proteins detectable in the blood were considered during this process due to the nature of the samples available and the aims of the study. Proteins identified from literature searches were validated in addition to proteins identified from the high-dimensional proteomic datasets to a) enable fair comparison between established biomarkers of bacterial *vs.* viral infection from the literature and the biomarkers identified here and b) to increase the possibility of identifying promising discriminatory markers.

Table 5: *Proteins identified as potential biomarkers for distinguishing between bacterial and viral infections through literature searches.*

Protein	Protein full name	Main cell type source	Disease group with raised levels	Reference
ANGPT2	Angiopoetin 2	Endothelial cells	Bacterial	[162, 163]
CD163	Scavenger receptor cysteine-rich type 1 protein M130	Macrophages	Bacterial	[164]
CD25	Interleukin-2 receptor alpha chain	Leukocytes	Viral	[165]
CRP	C-reactive protein	Hepatocytes	Bacterial	[166]
CX3CL1	Fractalkine	Endothelial cells	Viral	[167, 168]
GCSF	Granulocyte colony-stimulating factor	Monocytes	Bacterial	[169]
HNE-αAT1	Human neutrophil elastase-α1-antitrypsin	Neutrophils	Bacterial	[170]
IFN-γ	Interferon gamma	Leukocytes	Viral	[171]
IL18	Interleukin 18	Macrophages	Bacterial	[172]
IP10	C-X-C motif chemokine 10	T cells, macrophages	Viral	[173]
MPO	Myeloperoxidase	Neutrophils	Bacterial	[174]
PRTN3	Proteinase 3	Neutrophils	Bacterial	[170]
S100A12	S100-A12	Neutrophils	Bacterial	[175]
S100A8	S100-A8	Neutrophils	Bacterial	[175]
ST2	Interleukin-1 receptor-like 1	Endothelial cells, T cells	Viral	[176]
TRAIL	Tumor necrosis factor ligand superfamily member 10	Leukocytes, endothelial cells, various tissues	Viral	[173]

3.3.2 Cross-platform validation phase 1

ELISA and Luminex® immunoassays were used to measure the levels of proteins identified in the discovery phase (n=13) in addition to the literature derived proteins (n=16). Levels were measured in

an independent set of plasma samples from patients with bacterial (DB) infections (*n*=88) and viral (DV) infections (*n*=113) (Table 2). The aim of cross-platform validation 1 was to narrow down the shortlist of the 29 protein biomarkers identified in the discovery phase and from the literature search. We aimed to select an optimal combination (i.e., a protein signature) for distinguishing between bacterial and viral infections in febrile children.

Boxplots for each protein are shown in Figure 13. Mann-Whitney U tests were performed to establish whether the levels of each protein were significantly different between bacterial infections *vs*. viral infections. Of the 13 proteins identified during the discovery phase from the proteomic datasets, 10 had significantly different levels between bacterial and viral infections. Of the 16 proteins identified from the literature, 11 had significantly different levels between bacterial and viral infections (Table 6).

Table 6: *The 29 proteins biomarker candidates with ELISA or Luminex immunoassays available taken forward into the first phase of cross-platform validation for distinguishing between bacterial and viral infections in febrile children. P-values were calculated using Mann-Whitney U tests comparing bacterial to viral infections and corrected for multiple testing using the Benjamini-Hochberg (BH) procedure. * next to the protein names indicates that the protein was identified during the discovery phase. BH-adjusted p-values <0.05 are significant. ns = not significant.*

Protein symbol	**Platform used**	***P*-value**
ANGPT2	Luminex	4.832×10^{-05}
AT3*	ELISA	1.469×10^{-05}
CD163	Luminex	*ns*
CD25	Luminex	*ns*
CLUS*	ELISA	1.643×10^{-06}
CRP	ELISA	4.396×10^{-17}
CX3CL1	Luminex	1.196×10^{-07}
FETUA*	Luminex	3.227×10^{-03}
G-CSF	Luminex	3.673×10^{-04}
HNE-αAT1	ELISA	2.417×10^{-03}
IFN-γ	Luminex	3.239×10^{-04}
IL18	Luminex	*ns*
IP10	Luminex	*ns*
ISG15*	ELISA	*ns*
LBP*	Luminex	*ns*
LG3BP*	Luminex	1.263×10^{-06}
MASP1*	ELISA	*ns*
MPO	Luminex	*ns*
NCAM1*	Luminex	1.284×10^{-05}
NGAL*	Luminex	2.115×10^{-11}
PRTN3	Luminex	2.342×10^{-03}
S100A12	Luminex	3.534×10^{-06}
S100A8	Luminex	2.442×10^{-05}
SAA1*	ELISA	4.385×10^{-08}
SELE*	Luminex	3.774×10^{-08}
SPP24*	ELISA	1.108×10^{-02}
ST2	Luminex	4.385×10^{-08}
TIMP1*	Luminex	1.127×10^{-04}
TRAIL	Luminex	*ns*

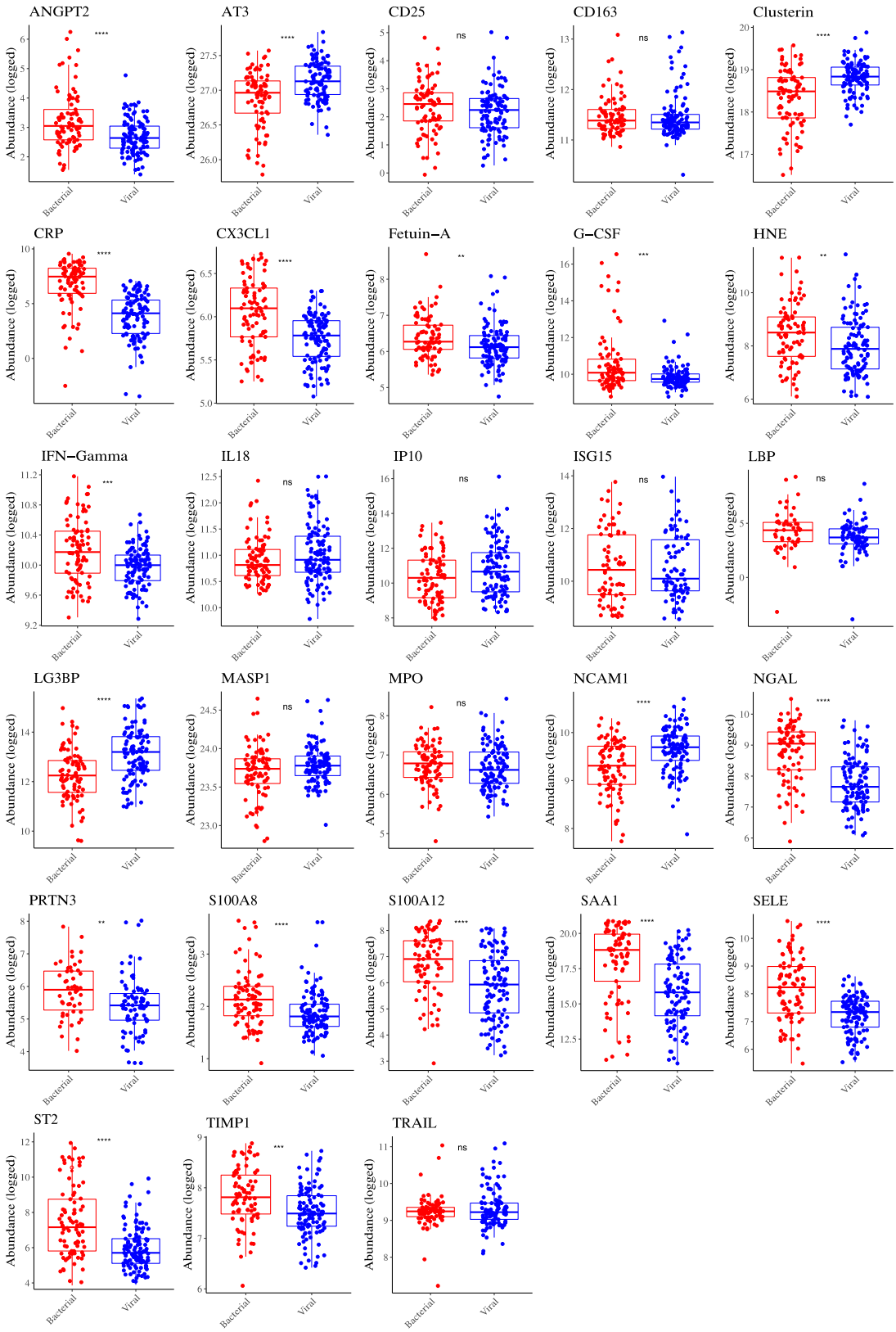


Figure 13: *Logged abundance values of the proteins measured by Luminex and ELISA in cross-platform validation phase 1 coloured according to disease group. These proteins were identified in the discovery phase as potential discriminatory biomarkers for bacterial vs. viral infections. ns: p-value >0.05; *: p-value <0.05; **: p-value ≤0.01; ***: p-value ≤0.001; ****: p-value ≤0.0001. All p-values are adjusted for multiple testing using Benjamini-Hochberg.*

The aim of cross-platform validation phase 1 was to narrow down the shortlist of 29 protein candidates. I therefore applied iterative FS-PLS to the proteins quantified in cross-platform validation phase 1 identify a small protein signature for distinguishing between bacterial *vs.* viral infections. FS-PLS was ran twice, with and without the proteins measured using ELISA. This distinction was made since it is more time-consuming and expensive to measure proteins using ELISA. The ultimate aim of this project was to identify a protein signature that could be measured rapidly in all healthcare settings using a cheap and easy-to-use platform in a rapid point-of-care test. ELISA assays can only readily measure one protein at a time, meaning that it would be considerably more complicated to develop a rapid point-of-care test that includes proteins measured by ELISA. Luminex® is more suitable in situations where multiple proteins are quantified simultaneously, meaning that there would be fewer steps from a Luminex® multiplex panel containing proteins in the diagnostic signature to a rapid point-of-care protein diagnostic test ready for clinical use.

When FS-PLS was applied to the proteins measured using Luminex® assays (total proteins=21), the most frequently selected signature was a 5-protein signature composed of NCAM1, IL18, SELE, NGAL, and IFN-γ, which was selected in 7/25 iterations (the Luminex signature). When FS-PLS was applied to the proteins measured by ELISA assays (except CRP since it was used in the classification of DV patients) in addition to the Luminex proteins (total proteins=28), the most frequently selected signature was a 5-protein signature composed of SELE, IL18, NCAM2, SAA1, ANGPT2, which was also selected in 7/25 iterations (the Luminex + ELISA signature). For both signatures, GLMs contrasting bacterial *vs.* viral infections were used to estimate coefficients for each gene which were used as the weights in the weighted disease risk score (DRS) calculations. The weighted DRS, which is described in detail in chapter 2 section 2.4.4, is an approach for combining all features included in a signature into a single score. It is calculated by multiplying the feature abundances by feature weights or coefficients (i.e., estimated by GLM) that indicate the degree and direction of change between the disease groups of interest.

The weighted DRS values were used for the receiver operating characteristic (ROC) models and subsequent performance evaluation metrics. The Luminex signature achieved an overall area under the ROC curve (AUC) of 89.1% (95% confidence interval [CI]: 84.3%-94.0%; Table 7; Figure 14A). The Luminex + ELISA signature achieved an AUC of 88.7% (95% CI: 84.2%-93.3%; Table 7, Figure 14B). When the performance of the Luminex signature (NCAM1, IL18, SELE, NGAL, IFN-γ) was contrasted against the Luminex + ELISA signature (SELE, IL18, NCAM1, SAA1, ANGPT2), the Luminex signature had a slightly higher overall AUC in addition to higher maximum sensitivity, specificity and partial AUC (pAUC) specificity > 90% (Table 7, Figure 14).

Table 7: *The 2 signatures identified by iterative FS-PLS using cross-platform validation phase 1 measurements of either Luminex proteins or Luminex + ELISA proteins for distinguishing between bacterial and viral infections in febrile children. Robustness indicates the number of times the protein was included in any signature across the 25 iterations of FS-PLS/number of total iterations. Proteins with names italicised and underlined increase and decrease in bacterial infections, respectively. pAUC = partial AUC; CI = confidence interval.*

Signature	Proteins (robustness)	AUC (95% CI)	pAUC sensitivity > 90%	pAUC specificity > 90%	Sensitivity	Specificity
Luminex	*SELE* (0.92), IL18 (0.92), NCAM1 (0.88), *NGAL* (0.8), IFN-γ (0.48)	89.1% (84.3%-94.0%)	3.5%	6.1%	90.4%	67.0%
Luminex + ELISA	*SELE* (0.92), IL18 (0.72), NCAM1 (0.8), *SAA1* (0.72), *ANGPT2* (0.44)	88.7% (84.2%-93.3%)	4.5%	5.3%	89.7%	64.4%

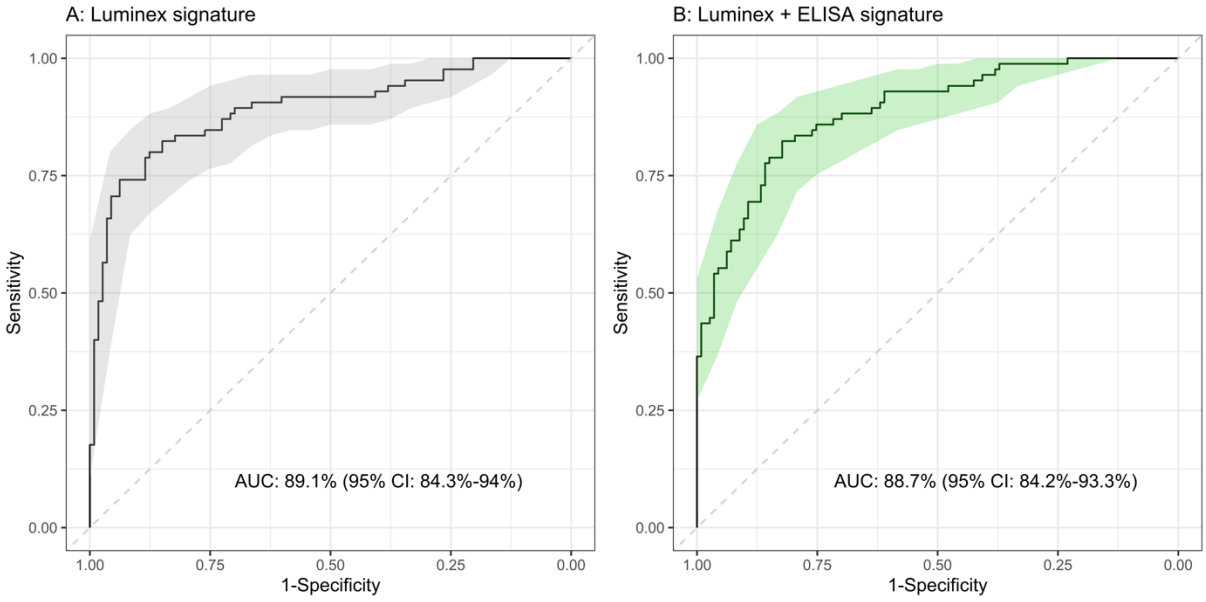


Figure 14: *The ROC curves of the protein signatures identified by running FS-PLS on either Luminex (**A**) or Luminex and ELISA (**B**) proteins measured in cross-platform validation phase 1. The Luminex signature is: SELE, IL18, NCAM1, NGAL, IFN-γ. The Luminex + ELISA signature is: SELE, IL18, NCAM1, SAA1, ANGPT2. The AUCs (95% CI) are printed on the plot. CI = confidence interval.*

CRP levels, which are often used in clinical care for diagnosing bacterial infections, were used for DV classification (Figure 8). CRP, however, is not a perfect biomarker as it is sometimes low in patients with bacterial infections. As a result, we wanted to identify proteins that could distinguish between bacterial and viral infections in instances where CRP was ≤60 mg/L for bacterial patients. Differential abundance analysis was performed contrasting patients with viral infections to patients with bacterial infections with CRP levels ≤60mg/L (the threshold used for DV classification). LG3BP was the top SDA protein, with an BH-adjusted *p*-value of 0.013 (LFC: -0.706).

The Luminex signature performed better than the Luminex + ELISA signature. Therefore, I took forward the 5-proteins included in the Luminex signature into another round of cross-platform

validation in addition to LG3BP due to its discriminatory abilities when CRP is low. Another round of cross-platform validation was required since the 5-protein signature identified in cross-platform validation phase 1 could be overfitting to the cross-platform validation phase 1 dataset.

3.3.3 Cross-platform validation phase 2

We tested the levels of the proteins identified in cross-platform validation phase 1 as promising biomarker candidates using an independent set of plasma samples from patients with bacterial (DB; *n*=162) and viral (DV; *n*=144) infections, in addition to samples from patients in the following phenotypic groups: non-sterile DB infections (NSDB; *n*=31), probable bacterial (*n*=64), bacterial syndrome (*n*=2), probable viral (*n*=91), viral syndrome (*n*=12) and healthy controls (*n*=61). All patients were recruited into the PERFORM study. Clinical and laboratory features of the patients are in Table 3.

The performance of the 5-protein Luminex signature (Table 7: NCAM1, IL18, SELE, NGAL, IFN-Gamma) in classifying bacterial *vs.* viral infections was tested through calculating weighted DRS that were used to estimate the AUCs. Original weights estimated from the GLM in cross-platform validation phase 1 (shown in Table 8) were used for the "Original" model. For the "Retrained" model, retrained model weights were obtained from gene coefficients from a GLM contrasting bacterial *vs.* viral infections using the protein abundance values measured in cross-platform validation phase 2 (weights shown in Table 8). The AUC calculated using the original model weights was 79.9% (95% CI: 75.1%-84.8%; Table 9, Figure 15), which was further improved by 9.3% to 89.2% (95% CI: 85.7%-92.7%; Table 9, Figure 15) when retrained model weights were used. It is unsurprising that the signature performed better with the retrained model weights since this process involves re-estimating the weights using the cross-platform validation phase 2 data. It is important to note however, that the performance with the retrained weights could be an over-estimate, since it is likely that the model is overfitting to the data.

The measurements of the proteins included in the signature were combined into a single score for each sample – the simple DRS. The simple DRS was calculated by adding the total log2 transformed abundance values of proteins expected to increase and subtracting the total log2 transformed abundance values of proteins expected to decrease for each sample. The direction (i.e., whether the proteins are expected to increase or decrease) was determined from the direction of the weights calculated by FS-PLS in cross-platform validation phase 1 (Table 8). The AUC using the simple DRS was slightly lower than the AUC using the retrained model weights at 87.6% (95% CI: 83.6%-91.5%; Table 9, Figure 15). The benefit of the simple DRS approach is that it does not require weights so is much more straightforward to calculate and translate across platform. Equally, it is less prone to overfitting because it does not refit the model weights.

Table 8: *The weights of the 5 proteins included in the Luminex signature identified in cross-platform validation phase 1 by FS-PLS. "Cross-platform validation phase 1" weights were used in the "Original" model and "cross-platform validation phase 2" weights were used in the "Retrained" model. Weights were calculated using a GLM model contrasting bacterial to viral infections. Positive and negative weights indicate that the protein increases and decreases in bacterial infections vs. viral infections, respectively.*

Protein name	Cross-platform validation 1 weights	Cross-platform validation 2 weights
NCAM1	-0.2863086	-0.1180592
SELE	0.2881772	0.1868027
IL18	-0.2386278	-0.1655074
NGAL	0.1463522	0.1903396
IFN-γ	-0.4386233	-0.0499154

Table 9: *The performance of the 5-protein signature in the cross-platform validation phase 2 dataset. The performance was calculated ROC models either using original weights from cross-platform validation phase 1 ("Original model", weights shown in Table 8), re-estimated model weights using cross-platform validation phase 2 data ("Retrained model", weights shown in Table 8) or with the simple DRS. CI = confidence interval; DRS = disease risk score; pAUC = partial AUC.*

Model	AUC	95% CI	pAUC (100-90% sensitivity)	pAUC (100-90% specificity)	Sensitivity	Specificity
Original	79.9%	75.1%-84.8%	2.6%	3.9%	80.5%	65.2%
Retrained	89.2%	85.7%-92.7%	4.4%	6.1%	93.1%	70.4%
Simple DRS	87.5%	83.7%-91.3%	4.0%	5.5%	84.0%	70.9%

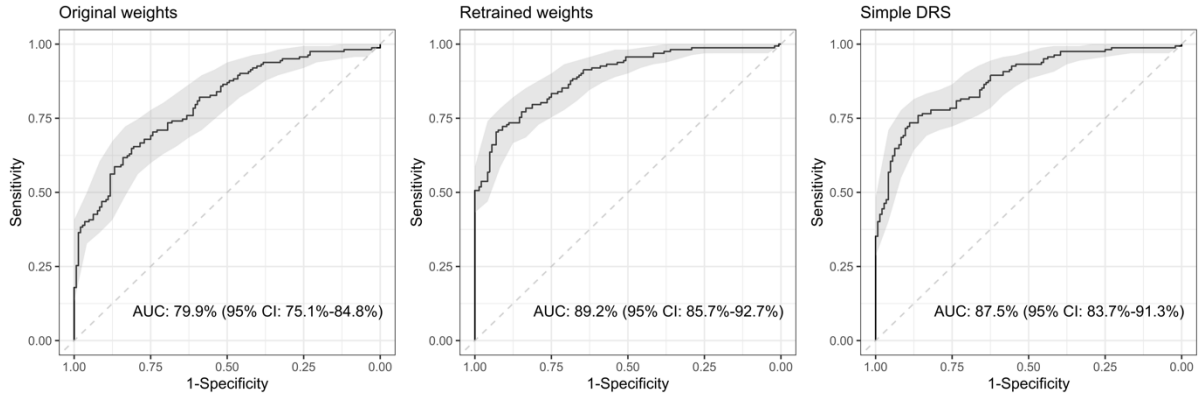


Figure 15: *ROC curves showing the performance of the 5-protein signature in the cross-platform validation phase 2 dataset. The 5-protein signature was identified in cross-platform validation phase 1 (section 3.3.2) for distinguishing between bacterial and viral infections. The ROC curves are calculated using (**A**) original logistic regression weights calculated from cross-platform validation phase 1 ("original model" in Table 9) (**B**) logistic regression weights retrained using cross-platform validation phase 2 data ("retrained model" in Table 9), and (**C**) the simple DRS ("simple DRS" model in Table 9). AUCs and 95% confidence intervals (CI) are printed on the plots.*

We also took forward LG3BP to cross-platform validation phase 2 due to its discriminatory ability in distinguishing between viral patients and bacterial patients with CRP levels ≤60 mg/L. We combined LG3BP with the 5-protein signature using the cross-platform validation phase 2 dataset. This combination of 6 genes led to an improved AUC of 89.4% (95% CI: 85.8%-92.9%; Figure 16) when the simple DRS was used. When model weights were retrained with LG3BP included in the signature, the combination of 6 genes achieved an AUC of 93.6% (95% CI: 90.9%-96.3%). The addition of LG3BP led to statistically significantly differences between the ROC models for the 5-protein and 6-protein

signatures calculated using retrained model weights (p-value: 1.5×10^{-4}) but not for the models calculated using the simple DRS. The models were compared using the roc.test function in the pROC package.

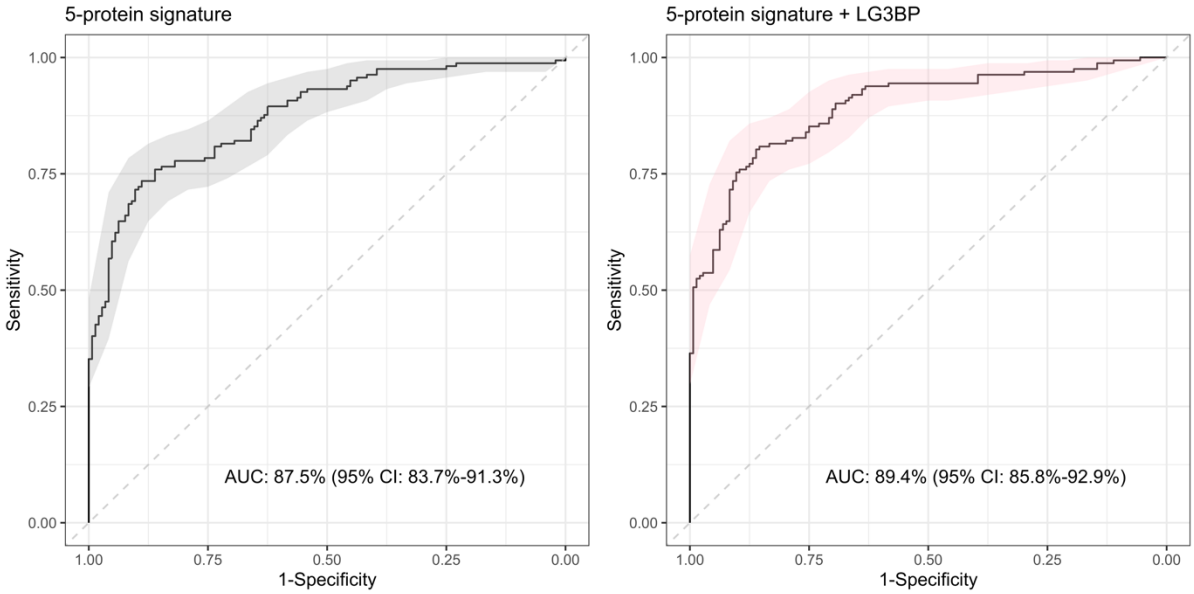


Figure 16: *Comparison of the AUC (95% CIs) between the 5-protein signature (**A**) vs. the 5-protein signature + LG3BP (**B**) in distinguishing between bacterial and viral infections. ROC curves were generated using cross-platform validation phase 2 data and using the simple DRS. The 5-protein signature was identified by FS-PLS and LG3BP was identified by Limma in cross-platform validation phase 1 (section 3.3.1). CI = confidence interval. AUCs and 95% confidence intervals (CI) are printed on the plots.*

Although LG3BP led to an improved AUC, it did not reduce the number of bacterial patients misclassified (16/162 for both the 5-protein signature and the 5-protein signature + LG3BP). It did however lead to a decrease in the number of viral patients misclassified when retrained model weights were used (50/144 for the 5-protein signature *vs.* 26/144 for the 5-protein signature + LG3BP) and when the simple DRS was used (50/144 for the 5-protein signature *vs.* 44/144 for the 5-protein signature + LG3BP). This suggests that LG3BP improved the signature's specificity. Indeed, the specificity of the 6-protein (5-proteins + LG3BP) signature was 84.9% and 76.8% for ROC models with the retrained weights and simple DRS, respectively, compared to a specificity of 70.9% and 70.4% for the 5-protein signature with the retrained weights and simple DRS, respectively. The sensitivity of the 6-protein signature was 88.1% and 86.2% for ROC models with retrained weights and simple DRS, respectively. The 6-protein signature (5-protein signature + LG3BP) was used for downstream analyses given its higher specificity in classifying viral patients.

3.3.3.1 Comparison to the MeMed BV™

The 6-protein signature identified here was contrasted to the 3-protein signature identified by Oved et al. [72] which is composed of CRP, IP10 and TRAIL. This 3-protein signature has been developed into a commercial diagnostic test, with reported sensitivity and specificity of 93.7% and 94.2%, respectively,

in a paediatric cohort [73]. Oved et al. [72] selected their three proteins through literature searches and bioinformatic screening, exploring the genes differentially expressed across different infectious and inflammatory diseases in microarray datasets, although limited details of the screening process are provided. Our approach is a more unbiased method for identifying host protein biomarkers for diagnosing bacterial and viral infections.

The Oved et al. [72] signature is the only other host protein signature reported for diagnosing bacterial and viral infections. To enable comparisons between the signature identified in our analyses and the 3-protein Oved et al. signature, we also measured the levels of CRP, IP10 and TRAIL in the cross-platform validation phase 2 dataset. Using the cross-platform validation phase 2 dataset, the CRP+IP10+TRAIL signature achieved an AUC of 87.6% (95% CI: 82.9%-87.1%) when the three proteins were combined into a single simple DRS for each sample, with a sensitivity and specificity of 86.4% and 77.7%, respectively. Overall, the 3-protein signature displayed lower performance in our dataset than the performance reported by Papan et al. [73]. The lower performance of the 3-protein signature in classifying the bacterial and viral samples presented here could reflect important differences in the methods of protein quantification between the MeMed BV™ in comparison to the Luminex® and ELISA (for CRP) assays presented here. For example, the monoclonal antibody clones used may be different, which could affect the epitopes recognised and could lead to different binding affinities. The 3-protein signature was not selected specifically for use in paediatric populations. Oved et al. [72] performed a generic literature search looking for protein biomarkers reported in any age group of humans in addition to animals. Therefore, it is not entirely surprising that their signature displays lower performance in a completely paediatric human population. Furthermore, in the study by Papan et al. [73], a board of clinicians assigned the "true" bacterial or viral phenotypes using a range of clinical and laboratory variables. Of note, CRP levels were available to these clinicians, raising the possibility of incorporation bias. This could potentially over-estimate the diagnostic test's accuracy since CRP was used in the host protein diagnostic test and the "gold-standard" diagnostic process.

CRP levels were used in the initial classification of the viral patients presented here, which means that the performance of CRP+IP10+TRAIL in our dataset is likely to be an overestimation, again due to incorporation bias. Therefore, to perform a fair comparison with our signature, we calculated the performance of the 3-protein signature in the absence of CRP, i.e., IP10+TRAIL, which achieved an AUC of 72.5% (95% CI: 66.6%-78.2%), with sensitivity and specificity of 42.1% and 29.6%, respectively. Furthermore, we added CRP to our 6-protein signature which resulted in an AUC of 92.8% (95% CI: 90%-95.7%) and sensitivity and specificity of 87.7% and 83.0%, respectively. AUCs reported here were calculated using the simple DRS. The performance of the 6-protein signature was contrasted against the 3-protein signature with and without CRP using the roc.test function in pROC [177]. In both instances, the 6-protein signature outperformed the 3-protein signature with statistically significant

differences in AUC (signatures including CRP p-value: 5.92×10^{-4}; signatures excluding CRP p-value: 2.86×10^{-08}).

CRP is an imperfect marker with elevated levels in bacterial infections in addition to various other infectious and inflammatory conditions, including SARS-CoV-2, multisystem inflammatory syndrome in children (MIS-C), influenza, and severe adenovirus [64, 178, 179]. A major benefit of the 6-protein signature presented in this chapter is that it does not depend on CRP. This means that it may be able to diagnose patients that CRP does not successfully classify, i.e., bacterial patients with CRP levels <60mg/L and viral patients with CRP levels >60mg/L.

3.3.3.2 The 6-protein signature in uncertain diagnostic groups

The patients presented here and included in the EUCLIDS, PERFORM and DIAMONDS studies have undergone extensive phenotyping. For the patients in the DB (and NSDB) and DV classes, there was adequate evidence to say with certainty that these patients had bacterial and viral infections, respectively. Individuals in the "certain" phenotypic groups (DB, DV, NSDB) would greatly benefit from a point-of-care test using host protein measurements as diagnosis would be considerably faster and much less invasive. Furthermore, the DB, DV and NSDB classifications reported here are not the initial diagnoses from the clinicians who cared for the patients, but are the conclusions from a rigorous, retrospective investigation of all available information. The groups of patients that this type of test could arguably benefit the most are those in the "uncertain" categories (probable bacterial, probable viral, bacterial syndrome, viral syndrome) that were also included in cross-platform validation phase 2 (Table 3).

The 6-protein signature was applied to the additional phenotypic groups present in the cross-platform validation phase 2 dataset, including NSDB, probable bacterial, bacterial syndrome, probable viral, and viral syndrome to determine whether they would be classified as bacterial or viral (Figure 17). A threshold of classification was identified from the ROC model that used the retrained model weights (Table 8, Table 9) using Youden's index [152]. Of the bacterial-related phenotypes, 77.4% (n=24) of the 31 NSDB samples, 81.2% (n=52) of the 64 probable bacterial samples, and 50% (n=1) of the 2 bacterial syndrome samples were classified as DB. For the viral-related groups, 25% (n=3) of the 12 viral syndrome samples and 60% (n=45) of the 75 probable viral samples were classified as DV. These rates of classification for the uncertain groups are highly encouraging since these patients (except for those in the NSDB group) were those in which a diagnosis with certainty was missing. For the probable viral patients, 44% (n=33) were administered antibiotics, and of these 33 patients, 66.6% (n=22) were classified as DV using the 6-protein signature. If the 6-protein signature was available as a point-of-care test when these patients were being treated, they may not have received antibiotics, thus helping to limit the spread of antimicrobial resistance (AMR) and reduce the rates of unnecessary antibiotic prescription.

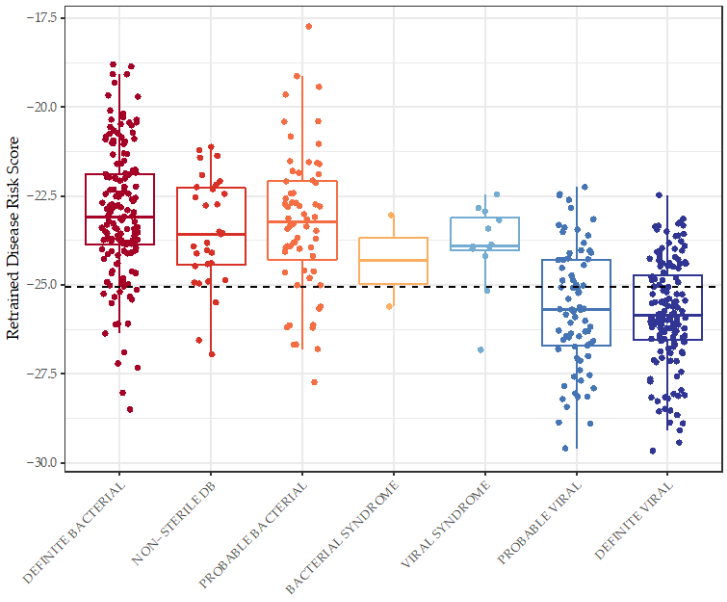


Figure 17: *The weighted disease risk scores for the 6-protein signature calculated using retrained model weights for individuals in all phenotypic groups in the cross-platform validation phase 2 dataset. The dashed line shows the threshold of classification, i.e., the disease risk score values above which a sample was classified as bacterial. This threshold was calculated using Youden's index [152]. to optimise sensitivity (88.1%) and specificity (70.9%).*

3.3.4 Biological functions of the 6 proteins

The 6 proteins included in the 6-protein signature have biological functions that make sense in the context of bacterial and viral infections. Of the 6 proteins in the signature, 3 were elevated in bacterial patients (SELE, NGAL, and IFN-γ), and 3 were elevated in viral patients (IL18, NCAM1, and LG3BP). SELE, which was elevated in bacterial patients, is a glycoprotein expressed on endothelial cells after activation by interleukin 1 (IL1), tumour necrosis factor-α (TNF-α) or bacterial lipopolysaccharide (LPS) [180]. SELE mediates leukocyte rolling and is involved in neutrophil, monocyte, and T cell recruitment to inflammatory foci, and the immune response to bacterial infections [180]. NGAL, which displayed elevated levels in bacterial patients, has a fundamental role in control of bacterial infections. NGAL prevents bacterial iron acquisition through sequestering iron-loaded bacterial siderophores which provide essential iron nutrients to the bacteria, thus preventing their survival [181]. IFN-γ is a proinflammatory cytokine with critical roles in innate and adaptive immunity, including a protective role against bacterial infections [182].

IL18 is a proinflammatory cytokine that induces other inflammatory cytokines. IL18 promotes cell activation of Th1 cells, which are involved in the host defence against viral infections and enhances cytotoxic activity of CD8+ T cells and NK cells [183]. IL18 is elevated following various viral infections including murine coronaviruses and influenza A infection [184-186]. NCAM1, which was also elevated in viral patients is a glycoprotein with various functions, including the expansion of T and B lymphocytes and NK cells. NCAM1 has also been identified as a potential viral receptor for both rabies

virus [187] and Zika virus [188]. LG3BP, which increased in viral patients, is a soluble scavenger receptor that has been associated with various viral infections including human immunodeficiency virus (HIV) [189] and Dengue virus [190].

3.4 Summary

Novel diagnostic methods for rapid and accurate diagnosis of bacterial and viral infections are a necessity if we are to reduce the amount of antibiotics prescribed needlessly. Host protein biomarkers such as CRP and procalcitonin (PCT) are already extensively used in clinical practise, however they are generic inflammatory markers with elevated levels observed in both bacterial and viral infections [157]. Traditional culture-based diagnostic procedures are time-consuming and can be unreliable, returning positive culture tests for microbes that are not causing disease. Furthermore, these methods can fail to detect disease-causing pathogens if they reside in a site that is difficult to access such as the lung. A host protein diagnostic signature would provide many benefits. Firstly, it would be much faster at providing results compared to culture-based approaches. In addition, because a host protein diagnostic signature would capture the changes in protein abundances brought about by infection, it would be able to detect infection regardless of infection site. Molecular rapid diagnostic tests - such as host protein signatures – are more cost effective than traditional culture-based methods [191], which means that a test would be highly advantageous in resource-limited settings.

In this chapter, I have used high-dimensional proteomic datasets to discover novel protein biomarker targets for distinguishing between bacterial and viral infections in febrile children. These biomarker targets were subsequently measured and validated using Luminex® and ELISA immunoassays. I performed further feature selection of these candidate biomarkers to identify a novel 5-protein signature for distinguishing between bacterial and viral infections. Further validation of these 5 proteins in an independent cohort has revealed AUCs ranging between 79.9%-89.2% when distinguishing between confirmed bacterial and viral infections. The performance of the signature improved significantly following the addition of LG3BP, with notable improvements in specificity in addition to AUC. Using a simple performance evaluation metric – the simple DRS – the combination of 6 proteins (SELE, IL18, NCAM1, NGAL, IFN-γ, LG3BP) achieved an AUC of 89.4% (95% CI: 85.8%-92.9%) which was improved to 93.6% (95% CI: 90.9%-96.3%) following retraining of model weights.

This work presented in this chapter is not without limitations. Coinfection of bacterial and viral pathogens can occur. Since DV patients required CRP levels ≤60mg/L, it is unlikely that they were infected with a co-incident bacterial infection, however it is possible that some DB patients were infected with a co-incident viral infection. Despite the possibility of coinfections in some patients, the 6-protein signature can accurately classify 90.1% of DB and 81.9% of DV patients, meaning that it is expected to be robust to coinfections.

The multi-platform, multi-cohort study presented in this chapter has identified a host signature with clinical diagnostic potential for distinguishing between bacterial and viral infections in febrile children. Further steps will now be required to develop this 6-protein signature into a point-of-care diagnostic test. Firstly, the optimal way of measuring all 6 proteins in combination with each other must be explored so that a prototype test can be developed. Point-of-care tests measuring proteins are already used in clinical practise, such as SARS-CoV-2 rapid lateral flow tests which detect the presence of SARS-CoV-2 proteins in nasopharyngeal swabs, and pregnancy tests which identify the host-produced human chorionic gonadotropin (HCG) hormone in urine. The MeMed BV™ is another example of a diagnostic test using host protein measurements. Once a prototype test has been developed, the performance of the 6-protein signature must then be further explored through a clinical trial, contrasting the results of the point-of-care test to the gold-standard culture tests or the conclusions reached by the clinician(s). Whilst the next steps will be costly and time-consuming, the signature presented in this chapter represents an important "first step" towards being able to diagnose febrile children with host protein measurements and is the first study to use high-dimensional proteomic quantification methods to identify potential protein biomarker candidates specifically for use in children.

To conclude, the 6-protein signature that I have presented here displays excellent performance in differentiating between bacterial and viral infections in febrile children. In doing so, it outperforms the only other host protein diagnostic test, and is able to achieve high performance without depending on CRP, a flawed yet widely used biomarker for bacterial infections.

Chapter 4

4 Kawasaki Disease patient stratification and pathway analysis based on host transcriptomic and proteomic profiles

Kawasaki disease (KD) is an acute inflammatory disorder of childhood which leads to the formation of coronary artery aneurysms (CAA) in some children if left untreated, making it the leading cause of acquired heart disease in Europe, Japan, and North America. Various triggers of KD have been proposed; however the aetiology of KD remains unknown. Identification of the causative agent of KD would greatly improve diagnosis of KD, leading to fewer missed diagnoses. Furthermore, it would enable rapid and appropriate treatment administration, reducing disease morbidity. There are variations in the host immune response according to the type of pathogen that an individual is infected with. These variations are observable on the host molecular level through omic datasets. Omic profiles obtained from individuals with infectious and inflammatory diseases provide insights into the host response to disease by revealing genes, proteins, and pathways that are disrupted in disease. This approach is particularly powerful when datasets from multiple omic levels are considered in parallel. In this chapter, host omic profiles, including transcriptomes and proteomes, from children with KD are contrasted to those from children with bacterial and viral infections to determine whether the host response to KD more closely resembles the response to bacterial or viral infections, or indeed neither. Through contrasting the host response observable on the molecular level, we hope to uncover clues regarding the aetiology and pathogenesis of KD, contributing to the ongoing search for the causative agent of KD.

4.1 Chapter aims and contributions

This chapter presents the analysis of the transcriptomic and proteomic profiles of children with KD, contrasting them to the transcriptomic and proteomic profiles of children with bacterial and viral infections. KD is an acute paediatric inflammatory disorder that most commonly occurs in children under five years old [75]. The aetiology of KD is unknown. Despite the many bacterial and viral pathogens proposed as the cause of KD, none have been independently confirmed. Using the knowledge that levels of host ribonucleic acid (RNA) and proteins can be used to distinguish children with KD from those with febrile diseases [85, 103], we hypothesised that host transcriptomics and proteomics could be used to decipher whether the inflammatory response during KD is more similar to the response to bacterial or viral infections, or neither. We hypothesised that if KD was caused by a viral or bacterial pathogen, the host molecular profiles from children with KD would resemble the host

molecular profiles of children with viral or bacterial infections, respectively. The work presented in this Chapter is based on an article published in the *International Journal of Molecular Science (IJMS)* by Jackson et al. in 2021 [83].

The primary aims of the original paper are:

1. To contrast the host transcriptomic and proteomic profiles from children with KD to those from children with bacterial and viral infections using a combination of supervised and unsupervised machine learning and statistical methods;

2. To decipher whether the host molecular changes occurring following KD have more in common with the changes occurring due to bacterial or viral infections, or neither;

3. To explore the heterogeneity within the transcriptional and translational response of patients with KD using unsupervised machine learning approaches in the absence of bacterial and viral patients.

The structure of this chapter is as follows. Section 4.2 describes the clinical cohorts, data generation pipelines, and initial exploration of the datasets. Section 4.3 presents the findings of the study, with analyses contrasting KD patients to bacterial and viral patients found in section 4.3.1, including differential expression/abundance analysis (4.3.1.1), pathway enrichment analysis (4.3.1.2), *k*-means clustering (4.3.1.3) and classification (4.3.1.4). The analyses exploring KD patients in the absence of bacterial and viral patients are found in section 4.3.2. Section 4.4 contains a summary and discussion of the results.

As outlined above, KD is a paediatric inflammatory syndrome without a known cause, with various bacterial and viral pathogens proposed as the causative agent. Previous studies exploring host omic profiles of children with KD have studied one omic level in isolation. Therefore, this chapter contributes the first parallel analysis of host proteomes and host transcriptomes of children with KD. Furthermore, this chapter contributes the first extensive analysis and comparison of host omic profiles from children with KD to those from children with bacterial or viral infections. Small signatures for diagnosing KD *vs*. febrile controls have been identified elsewhere [85, 103], however previous studies have not focused on the biological differences of these conditions observable on the host molecular level. The biological findings of this chapter are novel contributions; the considerable heterogeneity that we observed in the host omic profiles of children with KD has not been shown before.

Declaration of originality for chapter 4:

I carried out all of the work presented in this chapter, except for the following components. Dr. Shea Hamilton performed the laboratory elements for the generation of the SomaScan dataset. Dr. Stephanie

Menikou, Dr. Honglei Huang, Adam Thorne and Dr. Roman Fischer performed the laboratory elements and data generation for the liquid chromatography-tandem mass spectrometry (LS-MS/MS) experiments at the Discovery Proteomics Facility, Oxford. Dr. Andrew McArdle processed the LC-MS/MS data using MaxQuant. Dr. Victoria Wright performed the laboratory elements and data generation of the transcriptomic sequencing experiments. Quality control and normalisation of the transcriptomic validation datasets was performed by Dr. Clive Hoggart and Dr. Victoria Wright.

4.2 Cohorts and study design

Four datasets were used in this chapter, including two proteomic datasets (1 x discovery; 1 x validation) and two transcriptomic datasets (1 x discovery; 1 x validation) (Table 10). Samples were from patients recruited into multiple studies in the UK, Europe, and the United States. All samples were obtained from patients with written parental informed consent.

The KD samples included in the transcriptomic datasets were from patients recruited in the United States as part of the ongoing US-based Kawasaki Disease Research Center Program (https://medschool.ucsd.edu/som/pediatrics/research/centers/kawasaki-disease/pages/default.asp) which was set up to foster excellence in care for patients with KD and to support clinical, laboratory, and epidemiologic investigation into the aetiology, pathophysiology, and natural history of the disease. All patients enrolled in the Kawasaki Disease Research Center Program were recruited at Rady Children's Hospital San Diego. This cohort of patients is subsequently referred to as the "UCSD cohort".

The KD samples included the proteomic datasets were from patients recruited into the ongoing UK Kawasaki study "Genetic determinants of Kawasaki Disease for susceptibility and outcome" (NRES Committee London – Fulham; 13/LO/0026) which recruits acutely unwell children with KD during hospital admission in participating hospitals around the UK with the aim of unpicking the genetic determinants of KD. This cohort is subsequently referred to as the "UK KD cohort".

All four datasets included samples from healthy control (HC) children and children with febrile illnesses including definite bacterial (DB) and definite viral (DV) infections. DB, DV and HC samples included in the proteomic discovery dataset were from patients recruited into the PERFORM study (Personalised Risk assessment in Febrile illness to Optimise Real-life Management across the European Union; perform2020.org; London – Central Research Ethics Committee 16/LO/1684) [60, 159, 160]. For the proteomic validation dataset, DB, DV and HC patients were recruited into the EUCLIDS (European Union Childhood Life-Threatening Infectious Disease Study; 11/LO/1982) study [158]. For the transcriptomic discovery and validation dataset, the DB, DV and HC samples were obtained from patients recruited in the United Kingdom and Spain as part of the IRIS study (Immunopathology of Respiratory, Inflammatory and Infectious Disease; NIHR ID 8209) and GENDRES study (Genetic,

Vitamin D, and Respiratory Infections Research Network; gendres.org) [104]. The GENDRES study aimed to explore the genetic basis of individual susceptibility to respiratory infections and the IRIS study explored the immunopathology of respiratory infectious diseases. The EUCLIDS and PERFORM studies have been described in detail in chapter 3.

All of the analyses were performed on the discovery datasets. The validation datasets were only used to test the performance of machine learning classifiers.

Table 10: *Information about the datasets used in chapter 4. KD = Kawasaki disease; DB = definite bacterial; DV = definite viral; HC = healthy control; LC–MS/MS = liquid chromatography with tandem mass spectrometry; * = not used in analysis; GEO = Gene Expression Omnibus.*

Dataset Name	GEO Accession(s)	Platform(s) Used for Generation	Cohort				Citation(s)
			KD (*n*)	DB (*n*)	DV (*n*)	HC (*n*)	
Transcriptomic discovery	GSE73461	*Microarray*: HumanHT-12 version 4.0	UCSD (77)	IRIS, GENDRES (31)	IRIS, GENDRES (92)	IRIS, GENDRES (62)	[85]
Transcriptomic validation	GSE73462 GSE73463	*Microarrays*: 1 × HumanHT-12 version 3.0 1 × HumanHT-12 version 4.0	UCSD (101)	IRIS, GENDRES (23)	IRIS, GENDRES (28)	IRIS, GENDRES (16)*	[104, 192]
Proteomic discovery	*NA*	LC–MS/MS	UK KD (26)	PERFORM (73)	PERFORM (75)	PERFORM (25)	[83]
Proteomic validation	*NA*	SomaScan® [40]	UK KD (26)	EUCLIDS (48)	EUCLIDS (31)	EUCLIDS (25)	[83]

The age distribution of KD patients was relatively similar across datasets at a median of 26 months (range: 3-178) and 30 months (range: 3-210) for the transcriptomic discovery and proteomic discovery datasets, respectively, and 37 months (range: 4-182) and 16 months (range: 0-210) for the transcriptomic validation and proteomic validation datasets, respectively (Table 11). Over 50% of KD patients were male in all four datasets. Similarly, for the DB and DV patient groups, over 50% of patients were male in all datasets except for the transcriptomic validation dataset for DB (44% DB were male) and the proteomic validation dataset for DV (45% of DV were male; Table 11).

Table 11: *Clinical and demographic patient information for patients included in each dataset used in chapter 4 split by disease group. Ages are in months with age range in brackets.*

	Kawasaki disease			**Definite bacterial**			**Definite viral**			**Healthy controls**		
	N	Median age	Males	*N*	Median age	Males	*N*	Median age	Males	*N*	Median age	Males
Transcriptomics												
Discovery	77	26 (3-178)	42 (55%)	31	22 (0-171)	16 (52%)	92	14 (0-175)	65 (70%)	62	40 (0-182)	31 (50%)
Validation	101	37 (4-182)	61 (60%)	23	22 (7-132)	10 (44%)	28	18 (0-172)	17 (60%)	*NA*	*NA*	*NA*
Proteomics												
Discovery	26	30 (3-210)	18 (69%)	73	57 (0-205)	41 (56%)	75	37 (0-180)	42 (56%)	25	58 (1-155)	17 (68%)
Validation	26	16 (0-210)	18 (69%)	48	25 (1-191)	24 (50%)	31	11 (0-97)	14 (45%)	25	101 (3-191)	12 (48%)

Table 12 contains the clinical features of the KD patients included in the four datasets analysed. The range of C-reactive protein (CRP) levels were similar for the KD patients across all four datasets with median values of 119 mg/L, 87 mg/L, 92 mg/L and 92 mg/L for the transcriptomic discovery, transcriptomic validation, proteomic discovery, and proteomic validation, respectively. CAA formed in less than 50% of KD patients in all four datasets (Table 12) and the median duration of fever at the time of sample collection was 5 and 6 days for the transcriptomic discovery and validation datasets, respectively, and 7 for both proteomic datasets. The range of the duration of fever at sample collection was larger for the patients in the proteomic datasets.

Table 12: *Clinical information for the patients with Kawasaki disease whose samples were included in any of the four datasets used in the analysis in chapter 4. IQR = interquartile range.*

	Transcriptomics		***Proteomics***	
	Discovery (*n*=77)	Validation (*n*=101)	Discovery (*n*=26)	Validation (*n*=26)
Day of fever at sample collection, median (range)	5 (2 to 7)	6 (2 to 10)	7 (3 to 20)	7 (4 to 22)
Coronary artery abnormalities, No. (%)	33 (43%)	27 (27%)	8 (30%)	6 (23%)
Laboratory values, median (IQR)				
Hemoglobin z-score	-1.3 (-2.0 to -0.3)	-1.2 (-2.0 to -0.4)	-0.22 (-0.24 to -0.19)	-0.3 (-0.31 to -0.26)
C-reactive protein mg/L	119 (48 - 192)	87 (59 - 173)	92 (43 - 167)	92 (59 - 167)
Platelet count $x10^3/\mu l$	352 (303 - 448)	408 (324 - 474)	319 (252 - 407)	303 (250 - 425)
White blood cell count /µl	14,200 (10,400 - 18,300)	13,900 (11,000 - 19,000)	13,900 (12,150 - 18,275)	13,300 (10,500 - 18,200)
Neutrophil count /µl	9,000 (6,600 - 12,400)	10,000 (7,300 - 12,600)	10,500 (7,125 - 14,200)	8,800 (6,600 - 11,950)

4.2.1 Case definitions

KD was diagnosed using American Heart Association (AHA) criteria of KD [84]. The AHA criteria includes presence of fever for ≥5 days alongside ≥4 of 5 of the principal clinical features (bilateral nonpurulent conjunctivitis, oral mucosal changes, cervical lymphadenopathy >1.5 cm, peripheral extremity changes and rash). Maximum coronary artery z score (Zmax) was used to classify CAA as follows: normal (Zmax <2.5), small (Zmax 2.5 to <5.0), or large (Zmax ≥5.0).

Febrile control patients had a diagnostic work-up as directed by the clinical team, including blood count; blood chemistry; CRP; blood, urine, and throat swab cultures; cerebrospinal fluid (CSF) analysis; and chest radiographs. Multiplex polymerase chain reaction (PCR) was performed to detect common respiratory viruses in nasopharyngeal aspirates or throat swabs, and common viruses in blood [85]. Following diagnostic work up, DB classification was conditional upon the isolation of a pathogenic bacteria from a sterile site (blood, CSF, pleural space, joint, urine), and a clinical syndrome in keeping with the identified bacterial species. Identification of bacteria in a sterile-site sample was taken as conclusive evidence for a confirmed bacterial infection, omitting the need for thresholds of inflammatory markers. DV classification was conditional upon the identification of a virus matching the syndrome and no evidence of bacterial coinfection. CRP levels ≤60mg/L were required for DV classification.

4.2.2 Data generation

The proteomic discovery dataset was generated by liquid chromatography tandem mass spectrometry (LC-MS/MS). LC-MS/MS is an untargeted approach to protein quantification that separates proteins using liquid chromatography (LC) and analyses the mass of peptide fragments using two mass spectrometers. This dataset was the same dataset as the MS-B dataset described in chapter 3. Full details of the data generation process are found in chapter 3, however in brief, 0.5µg of peptide for each sample was injected into a LC-MS/MS (LUMOS Fusion, Thermo Fisher scientific) at the Discovery Proteomics Facility, Oxford. The raw dataset files were processed separately by MaxQuant (1.6.10.43 [47]) and relative quantification was performed using the MaxLFQ algorithm [121]. Following MaxLFQ, the datasets underwent quality control (QC) processes including removing contaminant proteins and lowly expressed proteins, log2 transforming the datasets and performing principal component analysis (PCA).

The proteomic validation dataset was generated by SomaScan®. This is the same dataset that was used in chapter 3 (the SomaScan dataset). The dataset was generated using the 1.3K SomaScan® assay and the data was normalised by SomaLogic (full details in chapter 3). Batch effects were removed using COCONUT [161] and protein abundances were log2 transformed.

The transcriptomic discovery dataset was generated using the Human HT-12 v.4 Expression BeadChip microarrays (Illumina) [85]. Whole blood was collected into PAXgene blood RNA tubes (PreAnalytiX, Germany), incubated for 2 hours, and frozen at -20°C within 6 hours of collection before storage at -80°C. Ribonucleic acid (RNA) was extracted using PAXgene blood RNA kits (PreAnalytiX, Germany) according to the manufacturer's instructions. The integrity and yield of the total RNA was assessed using an Agilent 2100 Bioanalyser and a NanoDrop 1000 spectrophotometer. After quantification and QC, biotin-labeled complementary DNA (cDNA) was prepared using Illumina TotalPrep RNA Amplification kits (Applied Biosystems) from 500ng RNA. Labelled cDNA was hybridized, and the arrays were scanned using an Illumina BeadArray Reader according to the manufacturer's instructions. Using Genome Studio software, the microarray images were inspected for artifacts and QC parameters were assessed. The expression data is available on Gene Expression Omnibus (GEO; accession: GSE73461). The dataset was normalised using background correction and robust spline normalisation (RSN). The probe measurements were log2 transformed. GenomeStudio, the software associated with Illumina array platforms, calculates a detection *p*-value which indicates the confidence that a probe is expressed above the background defined by negative control probes. Probes were retained if at least 80% of samples in each disease group had detection *p*-value <0.01. PCA was performed to identify outliers which were subsequently removed and to evaluate the association between PC scores and disease group, age, and sex at birth.

The transcriptomic validation dataset was created by merging two existing microarray datasets with one containing samples from acute and convalescent KD patients (GEO accession: GSE73462) [192] and the other containing samples from patients with bacterial and viral infections (GEO accession: GSE73463) [104]. For both datasets, whole blood was extracted using PAXgene Blood RNA extraction tubes (Qiagen, Sussex, UK). After quantification and QC, biotin-labelled cDNA was prepared using Illumina TotalPrep RNA Amplification kits according to the manufacturer's instructions (Applied Biosystems). For the dataset containing acute and convalescent KD samples, cDNA was hybridized to Human HT-12 v.4 Expression BeadChip microarrays (Illumina) and for the dataset containing bacterial and viral samples, cDNA was hybridized to Human HT-12 v.3 BeadChip microarrays (Illumina). The chips were scanned with Illumina Bead Array Reader confocal scanners and QC checks were performed using the Genome Studio software. Normalisation and QC processes described above for the transcriptomic discovery dataset were applied to both of the validation datasets separately, and the datasets were merged and batch effects were removed by ComBat [129].

4.2.3 Dataset exploration

Following QC and pre-processing, exploration of the datasets was performed. All analyses presented in this chapter were performed on the transcriptomic and proteomic discovery datasets. The validation datasets were only used to test the performance of classifiers; therefore dataset exploration was done

only on the discovery datasets. Henceforth, "the transcriptomic dataset" and "the proteomic dataset" will refer to the discovery datasets.

First, *in silico* estimates of immune cell proportions were calculated for the transcriptomic dataset using CIBERSORTx [193], a machine learning tool that infers immune cell proportions by using expression levels of certain immune cell-associated genes. The proportions of neutrophils, lymphocytes, monocytes and mast cells were compared across all disease groups (KD, DB/bacterial, DV/viral, HC) using the Kruskal-Wallis test with *p*-values adjusted using the Benjamini Hochberg (BH) adjustment [142]. The levels of lymphocytes (BH *p*-value: 1.449×10^{-17}), monocytes (BH *p*-value: 1.361×10^{-14}), mast cells (BH *p*-value: 1.066×10^{-10}) and neutrophils (BH *p*-value: 9.202×10^{-11}) were significantly different across the phenotypic groups. Pairwise Dunn's tests were performed for each cell type with significant Kruskal-Wallis *p*-values (Figure 18). We observed significant differences between HC and each of the phenotypic groups (KD, bacterial, viral) for all of the cell types with significant Kruskal-Wallis *p*-values. Whilst levels of lymphocytes and neutrophils were significantly different between KD *vs.* viral infections with levels decreasing and increasing in KD, respectively, levels of monocytes were significantly different between KD *vs.* bacterial infections with levels decreasing in KD. From this analysis, it is evident that there are considerable differences in the *in silico* estimated immune cell proportions between the disease groups explored in this chapter. Since CIBERSORTx can only be used for gene expression data, and complete laboratory measurements for immune cell proportions were not available for the patients in the proteomic dataset, it was not possible to replicate this analysis using the proteomic dataset.

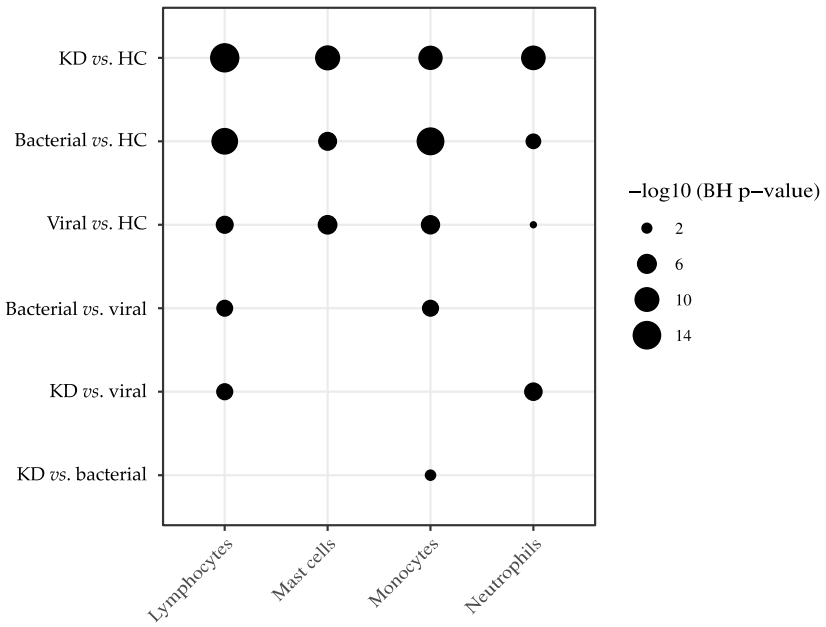


Figure 18: *BH-adjusted* p*-values obtained from the Dunn's test contrasting* in silico *immune cell proportion estimates between each of the phenotypic groups. Immune cell proportions were estimated using CIBERSORTx. KD = Kawasaki disease; HC = healthy controls. Point size represents -log10 (BH-adjusted* p*-values).*

PCA was performed to identify the major contributors of variance in the dataset. In the transcriptomic dataset, principal component (PC) 1 (29.24%) captured disease group, with the KD patients located between the bacterial and viral groups (Figure 19A). On PC4 (Figure 19D) viral samples were separated from the other phenotypic groups suggesting that PC4 is driven by viral-associated genes. The correlations between PC1 and the *in silico* immune cell proportions was tested, revealing that PC1 was positively correlated with neutrophils (Pearson's $r = 0.79$), mast cells (Pearson's $r = 0.52$) and monocytes (Pearson's $r = 0.65$) and negatively correlated with lymphocytes (Pearson's $r = -0.92$). These findings suggest that immune cell proportions are correlated with a considerable proportion of the variance captured in PC1 and are important contributors to the variance in this dataset.

In the proteomic dataset, PC1 (29.18%) captured variation caused by age differences, while PC2 (13.39%) and PC3 (10.56%) strongly capture the disease group effects, with the KD patients grouped together between the clearly separated bacterial and viral groups (Figure 20A, D).

PCA revealed that there are visible differences between the host transcriptomic and proteomic profiles from children with KD and those from children with either bacterial or viral infections. These findings motivated the further exploration of the host omic profiles, using a selection of statistical and machine learning tools, as described in subsequent sections.

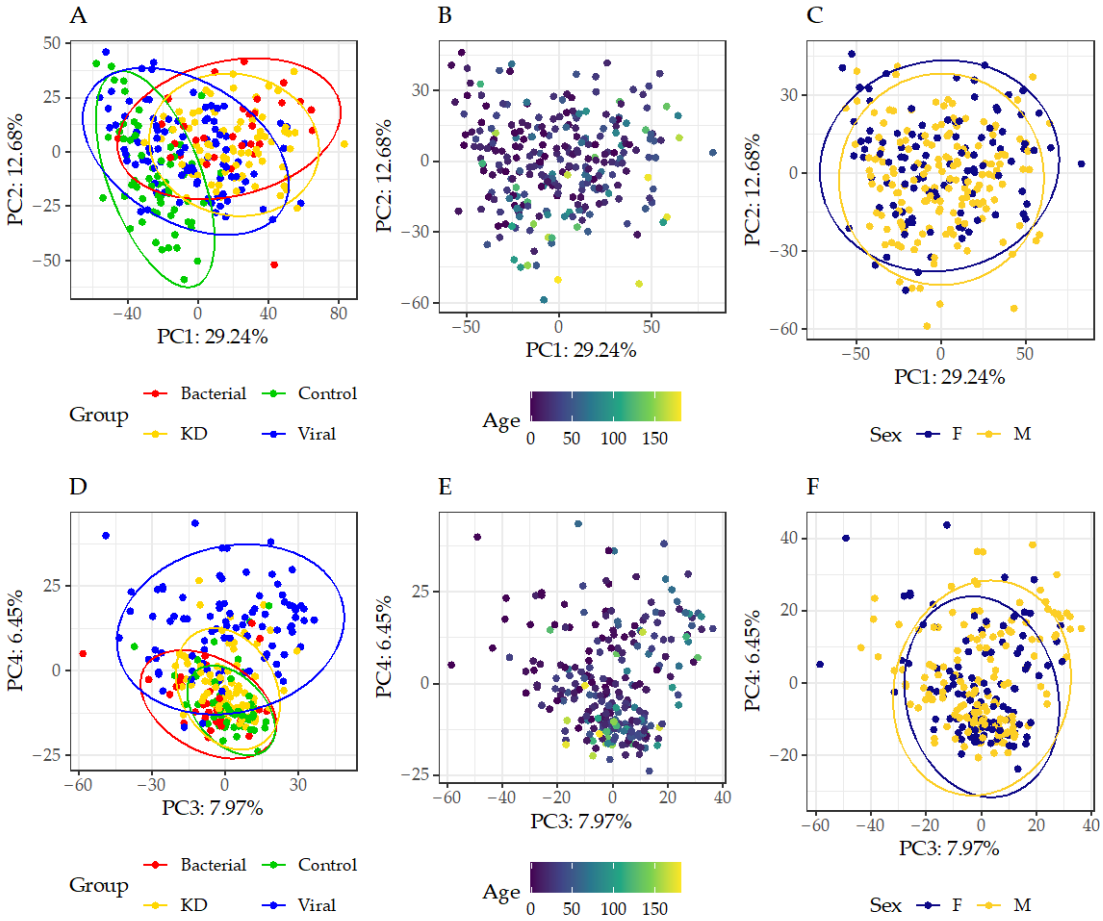


Figure 19: *PCA biplots for transcriptomic discovery dataset used in chapter 4. Points are coloured by disease group (**A**, **D**), age in months (**B**, **E**) and sex at birth (**C**, **F**). PCs 1-2 and 3-4 are shown. KD = Kawasaki disease; F = female; M = male; PC = principal component.*

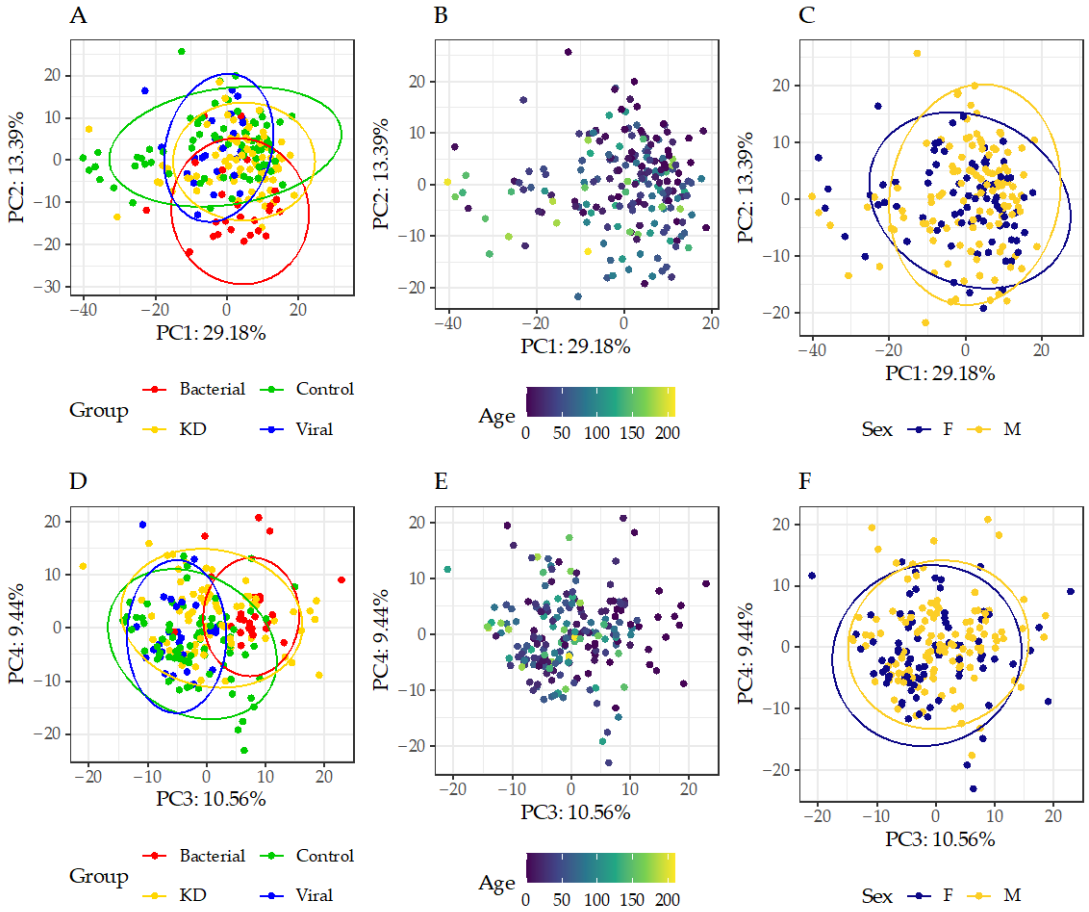


Figure 20: *PCA biplots for the proteomic discovery dataset used in chapter 4 with points coloured by disease group (**A**, **D**), age in months (**B**, **E**) and sex at birth (**C**, **F**). PCs 1-2 and 3-4 are shown. KD = Kawasaki disease; F = female; M = male; PC = principal component.*

4.3 Results

4.3.1 Comparing Kawasaki Disease to bacterial and viral infections

One of principle aims of this chapter was to contrast the host transcriptomes and proteomes of children with KD to those of children with bacterial and viral infections. There was no overlap between the patients included in the proteomic datasets and those included in the transcriptomic datasets, hence the analyses were performed independently on each type of omic data. All results presented are shown for the discovery datasets. The validation datasets were only used to test the performance of the classifier in section 4.3.1.4.

4.3.1.1 Differential expression/abundance analysis

Differential expression/abundance analysis (henceforth referred to as differential abundance analysis in this chapter) was performed to identify the genes and proteins that were disrupted in disease. Differential abundance analysis identifies the features that show differences in expression or

abundance levels between conditions, such as disease groups. Differential abundance analysis uses generalised linear models (GLM) to compare the levels of features between groups, returning a *p*-value and log2 fold-change (LFC) which represent the significance and the degree of change between the groups of interest, respectively. Each of the disease groups (KD, bacterial, viral) were contrasted against HC samples using Limma [141] whilst accounting for age and sex. Given the correlations between PC1 and immune cell proportions in the transcriptomic dataset, immune cell proportions were also included in the Limma model for the transcriptomic dataset. It was not possible to include immune cell proportions in the Limma model for the proteomic dataset since CIBERSORTx only accepts transcriptomic data and there is not an alternative method for protein abundances. Features with BH-adjusted *p*-values <0.05 were considered significantly differentially abundant (SDA).

For the transcriptomic dataset, 3,213 genes were SDA between KD *vs*. HC out of a total of 13,035 genes. A total of 3,124 genes were SDA between bacterial *vs*. HC, and 4,663 genes were SDA between viral *vs*. HC. For the proteomic dataset, 114 proteins were SDA between KD *vs*. HC out of a total of 344 proteins. A total of 128 proteins were SDA between bacterial *vs*. HC, and 79 were SDA between viral *vs*. HC. The lists of SDA genes and proteins were contrasted between each of the pairwise comparisons to identify which features were uniquely involved in the host response during KD, and which were overlapping with the host response to bacterial and viral infections (Figure 21). Of the 3,213 genes SDA between KD *vs*. HC, 1,589 (49%) were also SDA between bacterial *vs*. HC with concordant LFC directions (black points in Figure 21A), and 1,747 (54%) were also SDA between viral *vs*. HC with concordant directions (black points in Figure 21B). A total of 1,047 (33%) genes were consistently SDA between KD *vs*. HC, bacterial *vs*. HC, and viral *vs*. HC, representing the genes involved in the conserved immune response to infection/inflammation. Despite the overlapping genes, 886 (27%) of the genes SDA between KD *vs*. HC were unique to KD, representing the genes only involved in the immune response to KD.

Of the 114 proteins SDA between KD *vs*. HC, 94 (82%) were also SDA between bacterial *vs*. HC with concordant LFC directions (i.e., direction of change is the same; black points in Figure 21C), and 59 (52%) were also SDA between viral *vs*. HC with concordant directions (black points in Figure 21D). A total of 57 (50%) proteins were consistently SDA between KD *vs*. HC, bacterial *vs*. HC, and viral *vs*. HC, representing the proteins involved in the generic immune response to infection/inflammation. Only 17 (15%) of the proteins SDA between KD *vs*. HC were unique to KD, representing the proteins only involved in the immune response to KD.

The results on the transcriptomic level clearly indicate that, whilst there are similarities between KD and bacterial and viral infections, KD elicits a distinct host immune response detectable on the gene expression level. The pattern is not quite as clear on the proteomic level, with the majority of proteins overlapping between KD and either bacterial or viral infections. The higher proportion of overlapping genes between KD and bacterial infections suggests that the host response to KD detectable on the

protein abundance level has more in common with the response to bacterial infections in comparison to viral infections.

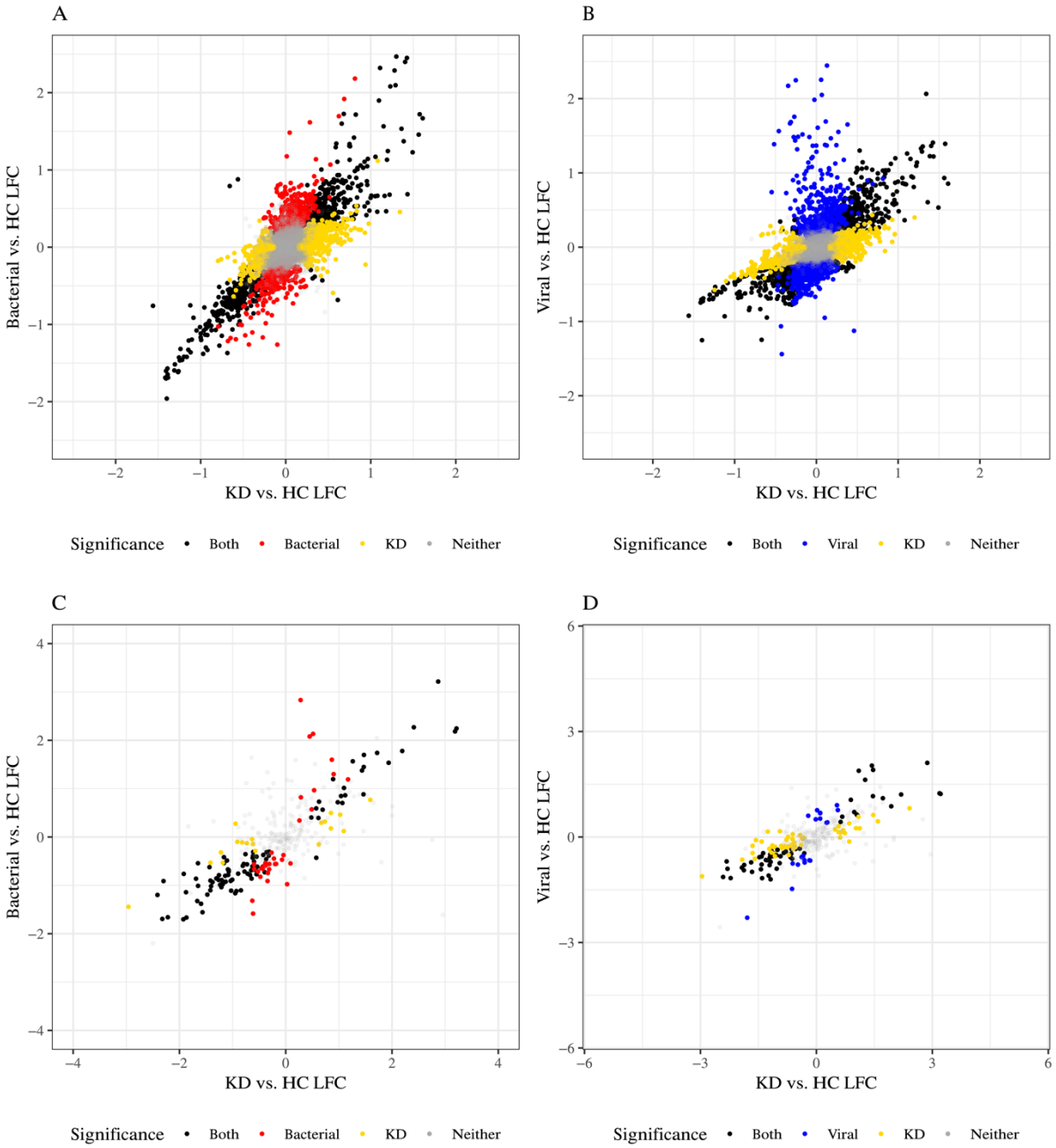


Figure 21: *Cross plots contrasting the differential abundance analysis results from KD* vs. *HC to those from bacterial* vs. *HC (**A** for transcriptomics; **C** for proteomics) and viral* vs. *HC (**B** for transcriptomics; **D** for proteomics). LFC values for each feature are plotted against each other to show concordance (same LFC direction) and discordance (mismatched LFC direction). Each point represents a feature (gene or protein). KD = Kawasaki Disease; HC = healthy controls; LFC = log2 fold-change. Significance is considered as BH-adjusted* p*-values <0.05.*

4.3.1.2 Pathway enrichment analysis

Pathway enrichment analysis was performed on the lists of features identified as SDA between KD *vs.* HC, bacterial *vs.* HC, and viral *vs.* HC. Pathway analysis was performed using g:Profiler2 [155], a

pathway enrichment analysis tool that uses a modified Fisher's test to test whether a set of features in a pathway (i.e., SDA genes) are present in a greater proportion than the features outside the pathway (i.e., the background genes or all genes in the transcriptomic dataset). g:Profiler2 tests whether features are enriched in gene ontology (GO) terms and due to the overlapping nature of GO, there is often redundancy in the sets of enriched pathways identified. As a result, REVIGO [156] – a webserver that summarises long lists of significant GO terms and reduces redundancy – was applied to the significant pathway lists to reduce redundancy. The identified pathways were contrasted across disease groups to determine which pathways were unique to the host response to KD and which were shared between the host response to KD and bacterial and/or viral infections. Pathways with a BH-adjusted *p*-value <0.01 were considered significant. The directionality of the pathway was inferred from whether the pathway was identified from the list of features that increased or decreased in the disease group of interest *vs*. HC.

On the transcriptomic level, 336 pathways were identified as significantly enriched from the list of genes SDA between KD *vs*. HC. For bacterial *vs*. HC and viral *vs*. HC, 352 and 210 pathways were significantly enriched, respectively. Of the 336 pathways from KD *vs*. HC, 64% (*n*=216) were also enriched in bacterial *vs*. HC and 39% (*n*=132) were enriched in viral *vs*. HC, with 35% (*n*=116) enriched in all three comparisons. A total of 104 (31%) pathways were unique to KD *vs*. HC.

Figure 22A, which shows the top pathways for the three comparisons on the transcriptomic level, demonstrates that amongst the top pathways, only one was significant across all comparisons: "regulation of response to stimulus". "Antigen presentation via MHC class I" was one of the top pathways identified from the lists of genes increasing in KD *vs*. HC and viral infections *vs*. HC. Expression of MHC molecules - molecules on cell surfaces that present antigenic peptides to T cells - is induced by a range of immune activators including interferons [194, 195]. "Neutrophil degranulation" and "vesicle-mediated transport" were both upregulated in KD *vs*. HC and bacterial infections *vs*. HC. In fact, neutrophil degranulation was the most significant pathway for KD *vs*. HC and bacterial *vs*. HC, suggesting that neutrophils play a fundamental role in the host response to KD and bacterial infections. Neutrophils are an important component of the early response to infection. In KD, it is known that neutrophil leucocytosis occurs in the initial days of illness and elevated levels of neutrophil elastase – a molecule secreted by neutrophils during inflammation – have been reported in KD patients, in addition to elevated levels of the neutrophil-activating chemokine IL-8 [196, 197]

On the proteomic level, 359 pathways were identified as significantly enriched from the list of proteins SDA between KD *vs*. HC. This contrasts to 283 and 258 for the lists of proteins from bacterial *vs*. HC and viral *vs*. HC. Of the 359 pathways from KD *vs*. HC, 71% (*n*=255) were also enriched in bacterial *vs*. HC and 60% (*n*=217) were also enriched in viral *vs*. HC, with 40% (*n*=143) enriched in all three comparisons, representing the conserved host immune response to KD and both bacterial and viral

infections visible on the translational level. A total of 77 (21%) pathways were unique to KD *vs*. HC, representing the proteins involved in the unique host response to KD visible on the translational level.

Figure 22B shows the top pathways for each of the three comparisons on the proteomic level. Many of the pathways enriched in all three comparisons are involved in the immune response, such as "acute inflammatory response", "leukocyte mediated immunity", and "platelet degranulation". Interestingly, for the proteomic data, all of the overlapping pathways have concordant directions, suggesting high similarity in the translational response between disease groups. Furthermore, there are fewer pathways that are unique to KD on the protein level in comparison to the gene expression level.

On both omic levels, many of the features (genes or proteins) and pathways significant between KD *vs*. HC were overlapping between KD and bacterial and/or viral infections *vs*. HC. This suggests that there are many similarities in the immune response to KD and bacterial and/or viral infections that are detectable at the host molecular level. It is evident from the overlapping pathways between KD *vs*. HC and either bacterial *vs*. HC or viral *vs*. HC that some elements of the host response during KD appear more "bacterial-like" and some elements of the host response appear more "viral-like". Despite the many pathways overlapping between KD *vs*. HC and either bacterial or viral infections *vs*. HC, there are clear sets of pathways that are uniquely enriched in KD *vs*. HC, as was observed in the differential abundance analysis. This indicates that KD elicits a specific host immune response observable at the gene expression and protein abundance level.

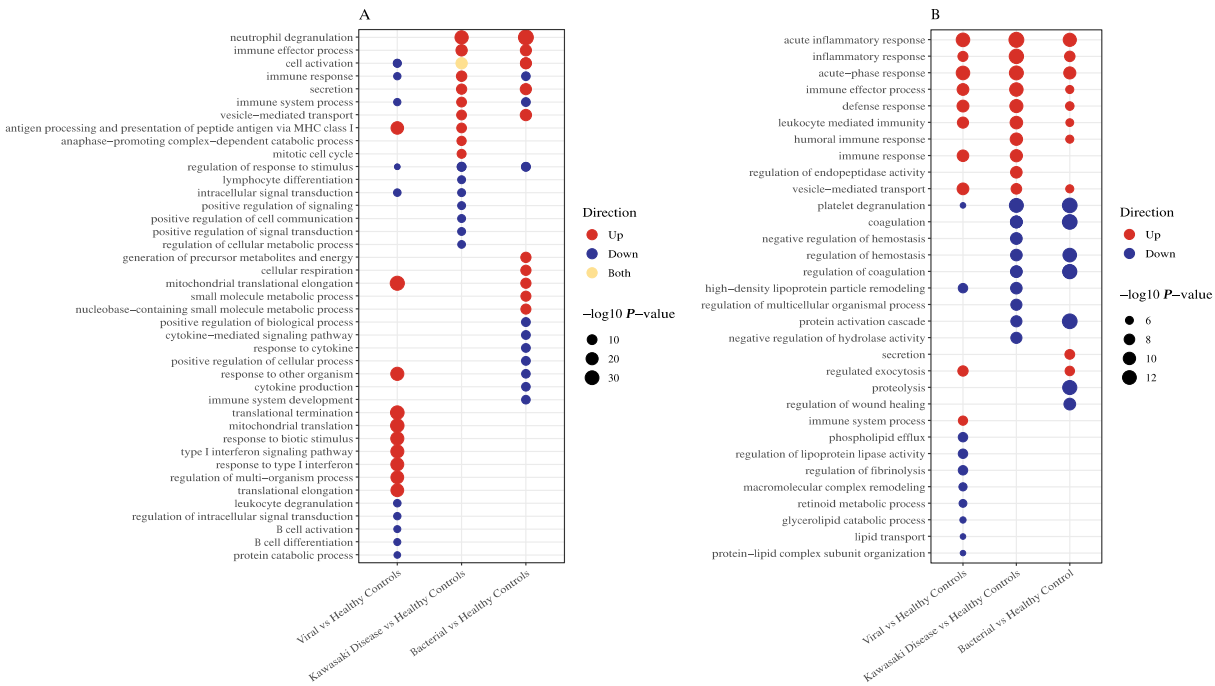


Figure 22: *The top pathways identified as significantly enriched from the lists of genes (**A**) and proteins (**B**) SDA between KD, bacterial, and viral infections vs. healthy controls. Pathways were identified by g:Profiler2 followed by filtering with REVIGO. Points represent pathways with colour indicating directionality and size representing BH-adjusted* p*-value*

(larger = lower p-value). Direction = "Both" indicates that the pathway was identified as enriched from the list of genes that increased and decreased in the disease group vs. healthy controls.

4.3.1.3 Clustering with bacterial and viral infections

Sections 4.3.1.1 and 4.3.1.2 demonstrated that there are considerable similarities between the host response to KD and bacterial and/or viral infections detectable at the transcriptional and translational level when pairwise comparisons were made between disease groups and HC children. The host response during KD and the host response to bacterial and viral infections was further explored in an unsupervised manner using *k*-means clustering to determine whether the KD patients were more likely to cluster with bacterial or viral patients. *k*-means clustering was used to identify any hidden patterns in the data that were not detectable through supervised pairwise comparisons.

As described in chapter 2 section 2.4.1, *k*-means clustering is an unsupervised clustering method that aggregates data points (i.e., host transcriptomic profiles) according to their similarities to each other. Age and sex have both been shown affect the human transcriptome [198, 199] and proteome [200, 201]. To mitigate the impact of sex and age on both omic levels, and to avoid inadvertently clustering the patients by these variables, gene expression and protein abundance values were corrected for sex and age by taking the residual gene expression and protein abundance values, respectively, after removing the contribution of these variables through Limma [144]. Given the high correlations between PC1 and the CIBERSORTx estimated *in silico* immune cell proportions discussed in section 4.2.3, the gene expression values were also corrected for immune cell proportions.

Using the R tool NbClust [139] the optimal number of clusters (*k*) was identified, which was *k*=3 for the gene expression data and *k*=2 for the protein abundance data. As outlined in chapter 2 section 2.4.1, NbClust tests various metrics in parallel to determine the optimal *k* for a dataset. Clustering was performed, with 53 (27%), 60 (30%), and 87 (43%) patients in each cluster on the transcriptomic level, and 88 (51%) and 86 (49%) patients in each cluster on the proteomic level (Figure 23).

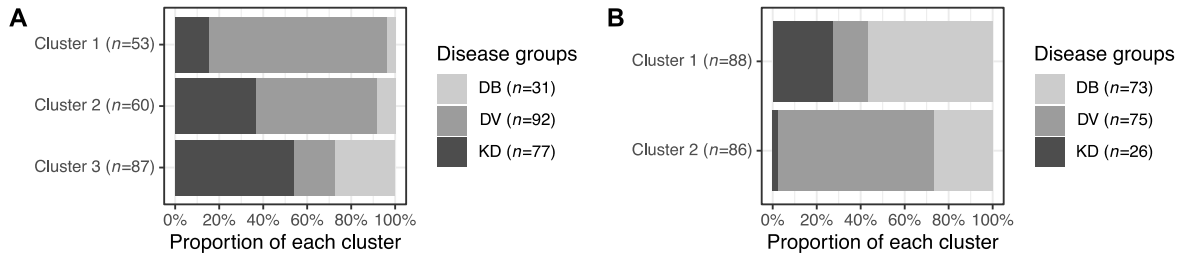


Figure 23: *The proportion of patients from each disease group in each cluster for transcriptomics (**A**) and proteomics (**B**). Clusters were identified used k-means clustering with k=3 for transcriptomics and k=2 for proteomics. DB = definite bacterial; DV = definite bacterial; KD = Kawasaki disease.*

On the gene expression level (Figure 23A), cluster 1 was heavily viral dominated, containing 47% (*n*=43) of the viral patients and the lowest proportion of KD patients (10%; *n*=8) and bacterial patients (7%;

n=2) out of any of the clusters. Contrastingly, the majority (61%; *n*=47) of the KD patients belonged to cluster 3 alongside the majority (77%; *n*=24) of the bacterial patients and the lowest proportion of viral patients across any of the clusters (17%; *n*=16). Whilst cluster 2 was not as extreme as either cluster 1 or 3, it was dominated with viral patients (36%; *n*=33), alongside 29% (*n*=22) of the KD patients and 16% (*n*=5) of the bacterial patients. The KD and bacterial patients followed the same trend, with their proportions increasing across clusters from 1 to 3, whilst the viral patients showed the reverse of this trend.

The two clusters identified on the protein level (Figure 23B) were clearly split into a bacterial dominated cluster (cluster 1) and a viral dominated cluster (cluster 2) which contained 68% (*n*=50) of the bacterial patients and 81% (*n*=61) of the viral patients, respectively. The majority (92%; *n*=24) of the KD patients clustered with the bacterial patients in cluster 1, with the remaining 8% (*n*=2) clustering with the viral patients, strongly suggesting that the host proteome of patients with KD resembles the host proteome of patients with bacterial infections more closely than those with viral infections, as was observed in the differential abundance and pathway analyses.

The associations between cluster membership for KD patients and various clinical variables associated with KD presentation were tested using the Fisher's Exact test for categorical variables and the one-way ANOVA for continuous variables. The following parameters were tested: strawberry tongue (yes/no/unknown); lymph node swelling (yes/no/unknown); presence of CAA (yes/no/unknown); peeling (yes/no/unknown); levels of CRP; month of year; and the duration of fever at sampling. On the transcriptomic level, there were significant associations between CRP levels (Fisher's exact *p*-value: 4.8×10^{-2}), lymph node swelling (ANOVA *p*-value: 4×10^{-2}) and peeling (ANOVA *p*-value: 5×10^{-2}). Levels of CRP were highest amongst KD patients in the bacterial dominated cluster 3, alongside the highest proportion of patients that reported lymph node swelling (17/21) and peeling (38/55). No clinical variables were significantly associated with cluster membership of the KD patients in the proteomic dataset.

The omic profiles of the KD patients were compared between clusters, with differential abundance analysis performed using Limma [141] followed by pathway enrichment analysis on the lists of SDA features using g:Profiler2 [155] and REVIGO [156].

When KD patients in cluster 1 were contrasted to KD patients in clusters 2 and 3 on the transcriptomic level, 503 SDA genes were identified alongside 156 significantly enriched pathways. 454 genes were SDA between KD patients in cluster 2 *vs*. KD samples in clusters 1 and 3, and 118 pathways were significantly enriched. Finally, 651 genes were SDA between KD patients in cluster 3 *vs*. KD patients in clusters 1 and 2, and 274 pathways were significantly enriched.

Pathway enrichment analysis of the genes SDA between KD patients in cluster 1 *vs.* KD patients in clusters 2 and 3 revealed an up-regulation of pathways associated with the antiviral response such as "type I interferon signalling pathway" and "cytokine-mediated signalling pathway" (Figure 24), which reflects the composition of the majority viral cluster 1. Furthermore, three pathways – "response to biotic stimulus" (i.e., a stimulus caused or produced by a living organism), "response to other organism" and "type I interferon signalling" were enriched in the KD patients in cluster 1 in addition to the viral transcriptomic patients when viral patients were contrasted to HC (Figure 22A in section 4.3.1.2). As outlined in section 4.3.1.2, one of the top pathways upregulated in both KD and viral patients *vs*. HC children was "antigen presentation via MHC class I", a process that that is induced by interferon amongst other mediators. Taken together, these findings suggest that interferon-induced activation is an element of the host response that is shared between KD and viral infections. Five pathways that were downregulated in the bacterial transcriptomic patients *vs*. HC (Figure 22A in section 4.3.1.2) were upregulated in the KD patients in cluster 1 in comparison to KD patients in clusters 2 and 3. Pathways associated with the inflammatory immune response, such as "inflammatory response" and "myeloid leukocyte activation" were enriched in KD patients in the majority bacterial cluster 3, indicating that these KD patients are displaying a more bacterial-like host response to KD.

When the KD patients identified in cluster 1 (n=24) on the proteomic level were contrasted to the KD patients in cluster 2 (n=2), only two proteins were SDA: serum amyloid A1 (SAA1) and retinol binding protein 4 (RBP4). The low number of SDA proteins is likely to reflect the small number of KD patients in cluster 2. Both RBP4 and SAA1 have been previously associated with KD, with decreased levels of RBP4 [202] and elevated levels of SAA1 [203] reported in KD; a pattern that the 24 KD patients in cluster 1 followed, with the opposite trend observed in the two KD patients in cluster 2. Since the causative agent of KD is unknown, there is no diagnostic test for KD. As such, some KD patients presented here may not be "true" KD and could have unrecognised alternative diagnoses. The two patients included in cluster 2 may be examples of this, or they could have viral coinfections that are causing the more viral-like host response. Alternatively, they could have been misclassified by the k-means algorithm.

Despite some pathways overlapping between transcriptomic KD patients from different clusters, a unique set of pathways was enriched for each of the clusters (72 unique pathways for cluster 1; 56 unique for cluster 2; and 135 unique for cluster 3). This suggests that although KD is considered as a single disease, there is considerable heterogeneity in the host response to KD that is detectable at the transcriptomic level. The KD patients did not form a single KD-dominated cluster independent of bacterial or viral patients on either of the omic levels, suggesting that gene expression and protein abundance is highly variable across patients with KD, with more similarities between patients with KD and either bacterial or viral infections, than within the group of patients with KD.

The overlap between pathways enriched in the KD patients in different clusters on the transcriptomic level (Figure 24) and those enriched in either bacterial or viral samples *vs.* HC (Figure 22A) provides further evidence that some KD patients elicit a host response visible at the transcriptomic level that appears more viral-like (i.e., KD patients in cluster 1) and some patients elicit a host response that appears more bacterial-like (i.e., KD patients in cluster 3).

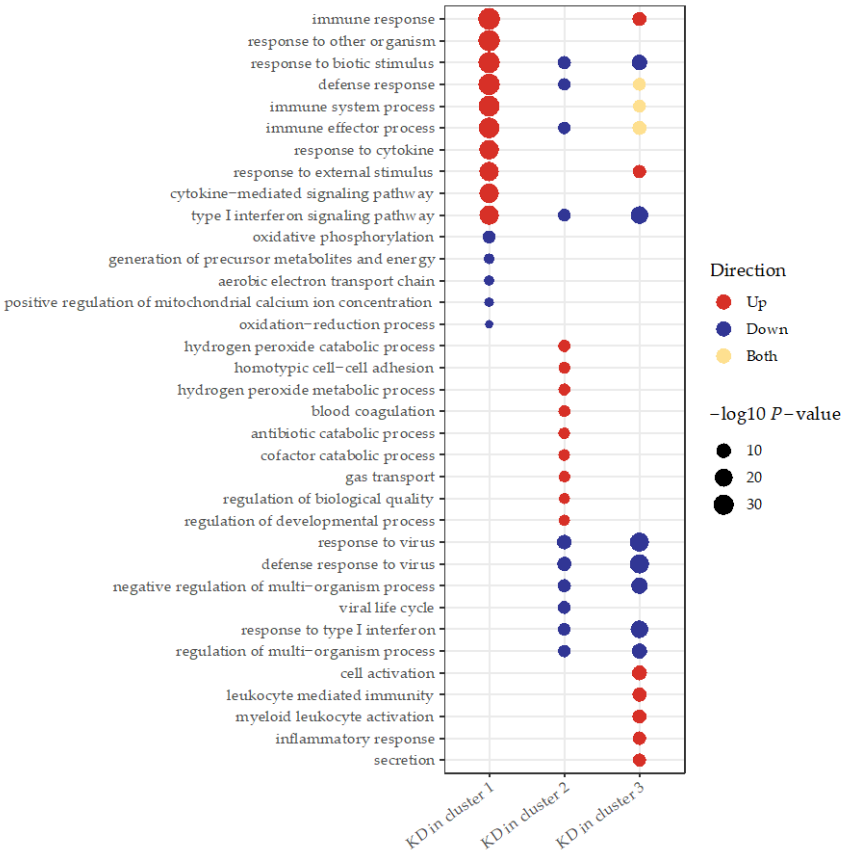


Figure 24: *Pathways upregulated and downregulated between KD patients in clusters 1, 2 and 3 for the transcriptomic dataset. Clusters were identified using k-means applied to KD, bacterial and viral patients. Pathways were identified by g:Profiler2 followed by filtering with REVIGO. Points represent pathways with colour indicating directionality and size representing BH-adjusted* p*-value (larger = lower* p*-value). Direction = "Both" indicates that the pathway was identified as enriched from the list of genes that increased and decreased in each comparison. BH-adjusted* p*-values <0.05 were considered significant. KD = Kawasaki disease.*

4.3.1.4 Classifying Kawasaki Disease as bacterial or viral

Contrasting the entire host transcriptomic and proteomic profiles of children with KD to those of children with bacterial and viral infections in sections 4.3.1.1 to 4.3.1.3 revealed that there were elements of the host response to KD that appeared more viral-like and elements that appeared more bacterial-like. As outlined in the chapter 1 section 1.4.2 and shown in chapter 3, small combinations of genes and proteins can be used to classify paediatric infectious and inflammatory diseases, including KD and bacterial and viral infections [67, 85, 192]. In chapter 3, I use host proteomics to identify a small signature for distinguishing between bacterial and viral infections in febrile children, with the ultimate

aim of developing this signature into a rapid point-of-care diagnostic test. I wanted to determine whether KD patients would be classified as either bacterial or viral infections if such a test was available. To achieve this, I built two machine learning classifiers that each returned the probability of a patient being either bacterial or viral. A two-way classifier approach was selected rather than a single, binary classifier (as was used in chapter 3) so that patients could be classified as neither, or indeed both, bacterial and viral. Classification was used in addition to clustering as it is a supervised approach that uses disease group labels to predict the classification of an individual patient rather than contrast the patient against others in the cohort/dataset.

Least absolute shrinkage and selection operator (LASSO) regression was performed using the R package glmnet [151] to develop the classifiers. LASSO regression was selected rather than FS-PLS since the number of features was not a key constraint, as was the case in chapter 3. On both omic levels, two LASSO models were used – the "DB classifiers" and the "DV classifiers". The DB classifiers were trained to identify bacterial patients from viral and HC patients, and the "DV classifiers" were trained to identify viral patients from bacterial and HC patients. For the DB classifiers, the LASSO models were trained using features SDA between bacterial *vs*. a combined group of viral and HC patients. For the DV classifiers, the LASSO models were trained using features SDA between viral *vs*. a combined group of bacterial and HC patients. The omic datasets that were corrected for age, sex (and *in silico* immune cell proportions for the transcriptomic dataset) were used. The LASSO models were trained on the discovery transcriptomic and proteomic datasets separately. 10-fold cross validation was performed to select the optimal value of lambda (λ) within the discovery datasets, which controls the degree of shrinkage. In LASSO, shrinkage refers to the degree to which the regression coefficients for features in the model are forced towards zero, thus resulting in a small number of features with non-zero coefficients and achieving feature selection. The features selected by LASSO were combined into a single score for each patient, using the simple DRS method outlined in detail in chapter 2 section 2.4.4. In brief, the simple DRS is calculated through subtracting the total expression/abundance of features that decreased from the total expression/abundance of features that increased for each individual. The "DB-DRS" was the DRS for the DB classifiers, and the "DV-DRS" was the DRS for the DV classifiers.

The performance of the classifiers was tested using the bacterial and viral patients from the transcriptomic and proteomic validation datasets using Youden's index [152] to determine the optimal DB- and DV-DRS thresholds above which a sample was classified as bacterial or viral, respectively. As outlined in chapter 2 section 2.4.4, Youden's index identifies the point in the receiver operating characteristic (ROC) curve where sensitivity and specificity are both maximised, providing a threshold of classification (i.e., a simple DRS value). The gene expression and protein abundance values of the transcriptomic and proteomic validation datasets were also corrected for age and sex (and *in silico* immune cell proportions for the transcriptomic dataset) prior to prediction of the bacterial and viral patients. Finally, the DB and DV classifiers were applied to the KD patients from the discovery and

validation datasets using the same thresholds of classifications applied to the bacterial and viral patients from the validation datasets.

On the gene expression level, the LASSO model selected 38 genes for the DB classifier, with 26 increasing and 12 decreasing in bacterial patients *vs*. viral patients and HC children (Table 13). The classifier achieved an area under the ROC curve (AUC) of 93.5% (95% confidence interval [CI]: 86.9%-100%) when classifying bacterial patients in the transcriptomic validation dataset. The DV classifier was composed of 32 genes with 13 increasing and 19 decreasing in viral patients *vs*. bacterial patients and HC children (Table 13). The DV classifier achieved an AUC of 93.5% (95% CI: 85.6%-100%) when classifying the viral patients in the transcriptomic validation dataset.

On the protein abundance level, 17 proteins were selected by the LASSO model for the DB classifier, with 5 increasing and 12 decreasing in bacterial patients *vs*. viral patients and HC children (Table 13). This classifier achieved an AUC of 92.7% (95% CI: 87.4%-98.1%) when classifying the bacterial patients in the proteomic validation dataset. The viral classifier was composed of 20 proteins, with 13 increasing and 7 decreasing in viral patients *vs*. bacterial patients and HC children (Table 13). The viral classifier achieved an AUC of 93.6% (95% CI: 88.5%-98.6%) when classifying the viral patients in the proteomic validation dataset.

Table 13: *The features selected by the LASSO models for the DB and DV classifiers for the transcriptomic and proteomic discovery datasets. Feature names in italics are those that increase in the classifier group of interest (i.e., bacterial in the DB classifiers), and those in bold decrease in the classifier group of interest (i.e., viral and HC in the DB classifiers). DB = definite bacterial; DV = definite viral.*

Classifier	Dataset	Features in the classifier
DB classifier	Transcriptomic	*ACBD6, HPSE, PTP4A3, TRAPPC9, CYBRD1, RTN2, TDRD9, LOC647104, SREBF1, LOC643332, IQGAP1, CYFIP2, PGCP, CBARA1, C1orf83, S100A4, VWA5A, LOC645387, MYO5A, FAM107B, SERPINB8, EIF4ENIF1, GPR137B, LTA4H, FAM55C, FLJ46552,* **PTGDS, MTMR11, GPR109B, RALGDS, NLRP3, TTC5, CLK3, NLRP3, E2F4, SSR2, CACNA2D3, NOD2**
DV classifier	Transcriptomic	*CCDC28B, NASP, SLC43A3, IFI27, TNS3, PCCB, DBT, HIST1H4H, LHFPL2, C4orf18, PRMT5, LOC729535, CCR2,* **LRIG1, CYBRD1, SIGLEC10, LST1, ARHGAP4, MIAT, LFNG, P2RY10, DUSP14, VHL, ARHGEF6, CYB5R1, DEF8, CMTM2, TMC6, LST1, ECOP, GOLGA8A, TMEM185B**
DB classifier	Proteomic	*FA9, CFAB, AACT, SAA1, ITIH4,* **NCHL1, PLMN, ANT3, APOE, FETUA, FA11, VTNC, IC1, CLUS, FA5, AFAM, MASP1**
DV classifier	Proteomic	*NCHL1, PLMN, ANT3, SAMP, FETUA, FA11, VTNC, IC1, THBG, CLUS, AFAM, MASP1, LG3BP,* **FA9, AACT, ANGT, S10A9, PON1, ITIH4, CBPB2**

The classifiers were applied to the KD patients from the discovery and validation datasets, with DB-DRS and DV-DRS calculated. On the gene expression level, 10% (*n*=18) of the 178 KD patients from the discovery and validation datasets had DB-DRS high enough to be classified as bacterial, 9% (*n*=16) had DV-DRS high enough to be classified as viral, and one sample had both DB-DRS and DV-DRS high enough to be classified as both bacterial and viral. 81% (*n*=145) of the KD patients did not have sufficiently high DB-DRS or DV-DRS to be classified as either bacterial or viral on the transcriptomic level (Figure 25A).

On the protein abundance level, 77% (*n*=40) of the 52 KD patients from the discovery and validation datasets had DB-DRS high enough to be classified as bacterial, 13% (*n*=7) had DV-DRS high enough to be classified as viral, and 6% (*n*=3) had DB-DRS and DV-DRS high enough to be classified as both bacterial and viral. 4% (*n*=2) of the KD patients did not have sufficiently high DB-DRS or DV-DRS to be classified as either bacterial or viral on the proteomic level (Figure 25B).

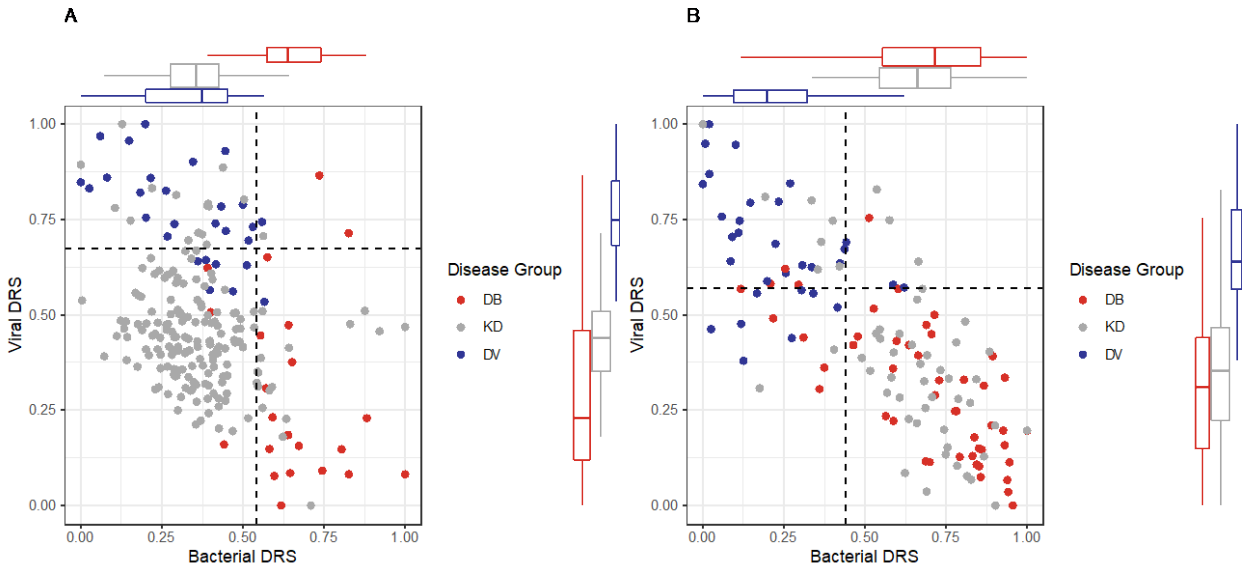


Figure 25: *Bacterial DRS (DB-DRS) plotted against viral DRS (DV-DRS) for KD (discovery and validation), bacterial (from the validation cohort) and viral (from the validation cohort) patients from the (**A**) transcriptomic and (**B**) proteomic datasets. Boxplots are shown for each disease group. Dotted lines represent the thresholds of classification obtained by Youden's index. KD = Kawasaki disease; DRS = disease risk score.*

As observed in Figure 23B when *k*-means clustering was applied to the proteomic dataset, the KD patients appeared more like bacterial patients on the protein abundance level, with almost all of the KD patients (83%) classified as bacterial (or bacterial and viral) on the proteomic level (Figure 25B). These findings further corroborate what we had already observed on the host proteomic level through differential expression analysis (section 4.3.1.1), pathway analysis (section 4.3.1.2), and clustering analysis (section 4.3.1.3): that the host proteome of children with KD very closely resembles the host proteome of children with bacterial infections.

The trend is not so clear on the transcriptomic level, with the majority of the KD patients classified as neither bacterial nor viral (81%; Figure 25A). The 145 KD patients classified as neither bacterial nor viral on the transcriptomic level were further explored through performing differential abundance analysis followed by pathway enrichment analysis, contrasting the transcriptome of these KD patients to HC children. Amongst the top pathways enriched were "defence response to fungus" (BH-adjusted *p*-value: 7×10^{-08}) and "response to fungus" (BH-adjusted *p*-value: 7×10^{-07}), hinting that KD in these patients may not be triggered by a process typical to common bacterial or viral infections, but possibly a novel process, such as one typical to fungal infections. A fungal trigger of KD has been suggested elsewhere [204, 205], however these findings require further investigation. The two-way classifier

approach performed on the transcriptomic level hints that it may not be a simple dichotomous question as to whether KD is more similar to bacterial or viral infections, motivating further analysis of the host omic profiles of children with KD.

4.3.2 Clustering Kawasaki Disease patients alone

The analyses presented in section 4.3.1 show that, whilst there are similarities between the host response to KD and the host response to both bacterial and viral infections that are visible on the transcriptomic and proteomic level, KD patients are clearly distinct from both bacterial and viral infections, especially concerning gene expression levels. On the protein level, despite the close similarities between host proteomes of children with KD and those of children with bacterial infections, there were some unique proteins and pathways that were only significant in KD *vs*. HC. Therefore, I performed further analyses using the transcriptomes and proteomes of children with KD without the bacterial and viral comparator groups to determine whether additional patterns could be observed. *k*-means clustering was performed as described in section 4.3.1.3, with k=3 identified as optimal by NbClust [139] for both omic levels. As in section 4.3.1.3, clustering was performed using the values that were corrected for age, sex and, for the transcriptomic dataset, *in silico* immune cell proportions estimates.

On the gene expression level, 41% (n=32) of the 77 KD patients were in cluster 1 (referred to as cluster KD1-T), 30% (n=23) were in cluster 2 (cluster KD2-T), and 29% (n=22) were in cluster 3 (cluster KD3-T). Of the 32 KD patients in cluster KD1-T, 31 (97%) were found previously in the bacterial dominated cluster 3 described in section 4.3.1.3 (Figure 23 and Figure 26), suggesting that these patients displayed a more bacterial-like response visible on the transcriptomic level. Over half (63.6%; n=14) of the 22 KD patients found in cluster KD3-T were previously found in cluster 2 (Figure 23 and Figure 26), with the remaining 8 patients found in the bacterial dominated cluster 3. Cluster KD2-T contained all except 1 of the 8 KD patients found in the viral dominated cluster 1, 8 of the KD patients previously found in cluster 2, and 8 of the KD patients found in cluster 3 (Figure 23 and Figure 26).

On the protein abundance level, 15% (n=4) of the 26 proteomic KD patients were in cluster 1 (cluster KD1-P), 27% (n=7) were in cluster 2 (cluster KD2-P) and 58% (n=15) were in cluster 3 (cluster KD3-P) (Figure 23). Previously in section 4.3.1.3, all patients except for two were in the bacterial dominated cluster 1. When KD patients were clustered without the comparator patients, the two patients that were clustering alone previously were no longer assigned to the same cluster.

The relationship between various clinical variables and cluster membership was tested using the following clinical variables: strawberry tongue (yes/no/unknown); lymph node swelling (yes/no/unknown); presence of CAA; peeling (yes/no); levels of CRP; month of year; and the duration of fever at sampling. The only significant association was between CRP levels and cluster membership

on both omic levels (transcriptomics p-value: 4.0×10^{-02}, proteomics p-value: $1.0\text{x}10^{-02}$). On the proteomic level, CRP was highest in KD patients from cluster KD1-P and cluster KD2-P. On the transcriptomic level, the highest CRP levels were found in cluster KD1-T which contained the majority of the KD patients that previously clustered with bacterial patients, suggesting that these patients displayed a more inflammatory profile. Furthermore, there was a significant association between cluster membership on the transcriptomic level and samples' transcriptomic DB-DRS (p-value: 6×10^{-3}) with the DB-DRS highest in cluster KD1-T and decreasing across clusters.

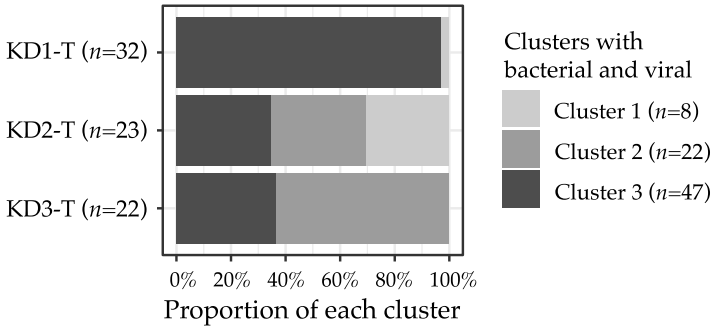


Figure 26: *The proportion of KD transcriptomic samples assigned to clusters KD1-T, KD2-T and KD3-T coloured according to the cluster they were assigned previously in section 4.3.1.3. Clustering was performed by k-means using only the KD patient transcriptomes.*

k-means performed on the KD patients without the bacterial and viral comparator samples added weight to the idea that the host response to KD detectable at the transcriptomic and proteomic level is highly heterogeneous, as discussed in sections 4.3.1.3 and 4.3.1.4. On both omic levels, 3 clusters were identified as optimal, indicating that there are subgroups of KD patients that can be distinguished using their molecular profiles. Only one significant association was found between the KD patients' cluster membership and various clinical symptoms (CRP levels). It is striking that the clusters presented here cannot be described by any of these clinical symptoms except for CRP levels, suggesting that the host molecular heterogeneity observed does not simply reflect clinical symptoms.

4.4 Summary

In this chapter, I have applied many of the data analysis tools described in chapter 2 section 2.4 to transcriptomic and proteomic datasets obtained from children with KD, bacterial and viral infections. KD is a paediatric inflammatory disease that shares many clinical features with bacterial and viral infections. The aetiology of KD is unknown, with bacterial and viral triggers suggested. I have performed extensive comparisons between the host response to KD that is visible on the transcriptional and translational level and the host response to bacterial and viral infections.

Differential abundance analysis and pathway analysis contrasting KD, bacterial and viral infections to HC revealed that there are sets of common features and pathways involved in the host response to KD

in addition to the response to bacterial and/or viral infections. This indicated that some elements of the host response to KD are more similar to the host response to bacterial infections, and some elements are more similar to the host response to viral infections. Equally, there are unique sets of features and pathways that are only visible in KD patients, suggesting that KD elicits a unique host response. This was further corroborated through the two-way classifier approach that revealed that it is not a simple dichotomous question regarding whether the response to KD more closely resembles the response to bacterial or viral infections, since the majority of the KD patients from the transcriptomic datasets were classified as neither bacterial nor viral infections. Further analysis of the KD patients that were classified as neither bacterial nor viral revealed pathways related to the response to fungal infection, raising the possibility that KD could be caused by a fungal trigger.

KD appeared more similar to bacterial infections on the protein level, with almost all KD patients clustering with the bacterial patients, and with many KD patients classified as bacterial in the two-way classifier approach (Figure 23B and Figure 25B). These similarities could be interpreted as following the striking similarities between KD and bacterial streptococcal and staphylococcal toxic shock syndrome (TSS) [206, 207] and could be consistent with idea of the proteome being closer to the observed phenotype than the transcriptome [208]. Alternatively, these similarities could reflect the fact that the proteomic dataset (n=344 proteins) contained fewer features than the transcriptomic dataset (n=13,035 genes). This difference in dimensionality could lead to greater granularity in the transcriptomic dataset, thus revealing differences that may have been missed on the protein level which may capture more generic changes.

Our analyses have revealed that the host response to KD is highly heterogeneous, a finding that has not been reported elsewhere. Not only did the KD patients fail to form a unique cluster when k-means was applied to KD, bacterial and viral patients on both omic levels, they also formed three clusters when k-means was applied to KD patients alone, again on both omic levels. This heterogeneity could be explained by differences in the host condition, such as genetic or epigenetic variations, or differences in prior pathogen exposure and the host environment. These factors may lead to different manifestations of the same pathogen observable on the molecular level. It is already known that differences in host genetics can affect KD manifestation since the formation of CAAs are known to be impacted by the host genetic background [209]. Another explanation for the heterogeneity in the host response to KD could be that KD is triggered by multiple pathogens as suggested by Rypdal et al. [210], each giving rise to indistinguishable clinical phenotypes with changes that are only that are visible on the molecular level. This idea would fit in with our findings since there were no associations between the key KD symptoms and the clusters identified in section 4.3.2 when k-means was applied to the KD patients alone. Furthermore, there were sets of genes, proteins and pathways that overlapped between KD *vs*. HC and either bacterial *vs*. HC or viral *vs*. HC, suggesting that KD elicits a more viral-like

response in some individuals, and a more bacterial-like response in others. This finding could also be interpreted as suggesting that there are multiple triggers of KD.

As described in chapter 1 section 1.4.1, disease tends to be treated in a "one-size-fits-all" despite the known variations in disease manifestation. Personalised medicine aims to use host omic data to deliver patient care that is tailored around the host's underlying molecular make-up in addition to their environment and lifestyle [49]. The heterogeneity observed in the host response to KD described in this chapter provides evidence of how subgroups of patients with the same diseases can be detected on the molecular level. This heterogeneity in the host response to KD warrants further exploration. An even larger cohort of KD patients could be used and a larger number of associations with clinical and laboratory parameters could be explored. Furthermore, performing metagenomic sequencing on the KD patients within each cluster would be intriguing to determine whether the microbial composition found in patients differs between clusters, possibly providing a clue to the aetiology of KD.

The transcriptomic and proteomic datasets presented in this chapter were not obtained from the same sets of patients. Therefore, it was not possible to determine whether the clusters of KD patients identified were consistent across omic levels. It would be intriguing to repeat this analysis using transcriptomic and proteomic datasets obtained from the same sets of samples to determine whether the clusters are identical regardless of the biological molecules characterised.

The aim of this chapter was to perform an extensive comparison between the host molecular profiles from children with KD and those from children with bacterial and viral infections, aiming to decipher whether the host response to KD resembles the host response to bacterial and/or viral infections. In doing so, we hoped that the findings could be used to guide further studies surrounding the cause of KD. The key finding from this chapter is that the host transcriptomic and proteomic responses during KD are highly heterogenous, with some elements resembling the host response to viral infections, some elements resembling the host response to bacterial infections, and some elements resembling neither. Especially on the transcriptome level, KD elicited a response that is distinguishable on the host molecular level from bacterial and viral infections. This provides motivation for further exploration of alternative potential triggers of KD that are neither bacterial nor viral, such as those caused by a fungus. The extraordinary heterogeneity observed in the host response to KD should be explored further; if KD is indeed caused by multiple triggers, or represents a spectrum of clinically indistinguishable diseases, this could ultimately have implications on the best way to treat and diagnose patients with KD.

Chapter 5

5 Diagnosis of multisystem inflammatory syndrome in children by a whole-blood transcriptional signature

Multisystem inflammatory syndrome in children (MIS-C) occurs 2-6 weeks following infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in some children. The highly overlapping clinical symptoms between MIS-C and other paediatric infectious or inflammatory diseases, such as Kawasaki disease (KD) and bacterial and viral infections, in addition to the lack of a gold-standard diagnostic test, make accurate diagnosis of MIS-C extremely challenging. Currently, MIS-C is diagnosed by fulfilment of a selection of clinical symptoms and criteria. A reliable diagnostic test is urgently required to ensure that children receive the appropriate medical interventions to treat MIS-C and other clinically similar diseases. Whole blood transcriptomics performed using samples from patients with infectious and inflammatory diseases can reveal changes in gene expression that are disease specific. Sparse combinations of RNA transcripts, otherwise known as gene signatures, can be used to differentiate between disease groups. In this chapter, I contrast host transcriptomes of children with MIS-C obtained from RNA Sequencing (RNA-Seq) to those from children with KD, viral and bacterial infections, with the aim of identifying a gene signature for diagnosing MIS-C.

5.1 Chapter aims and contributions

In this chapter, I present the analysis of a prospective study to identify host gene expression biomarkers for diagnosing MIS-C. MIS-C is a severe hyperinflammatory syndrome that occurs in some children 2-6 weeks after SARS-CoV-2 infection. Reports of MIS-C first occurred in April 2020 in areas with high SARS-CoV-2 rates [86-88]. There were reports early in the pandemic from several countries documenting an upsurge in KD-like inflammatory syndromes, which with hindsight, were likely to be MIS-C [211]. There are many clinical similarities between MIS-C and other paediatric infectious and inflammatory diseases, such as KD, sepsis, severe bacterial and viral infections, and staphylococcal and streptococcal toxic shock syndromes [212]. There is currently no diagnostic test for MIS-C, and due to the overlapping clinical and laboratory features between MIS-C and other conditions, diagnosing MIS-C is a major challenge. Infectious and inflammatory diseases are characterised by unique patterns of host ribonucleic acid (RNA) abundance in whole blood [150]. Sparse gene signatures, based on small numbers of transcripts have been reported for several diseases including tuberculosis disease [100, 102, 105, 213], malaria [214], bacterial and viral infections [67, 101, 104, 215], and KD [85]. In this chapter, I show that MIS-C can be distinguished from KD and a wide range of bacterial and viral infections by a

sparse RNA signature that could form the basis of a diagnostic test for clinical use. The work presented in this chapter is based on an article submitted to the *Journal of the American Medical Association (JAMA)* by Jackson et al. in July 2022.

The primary aims of the chapter are:

1. To explore the whole blood host transcriptomic profiles of children with MIS-C and contrast them to the whole blood host transcriptomic profiles from children with KD, viral and bacterial infections;

2. To identify a sparse host gene diagnostic signature or combination of host gene biomarkers that can be used to distinguish MIS-C from KD, viral and bacterial infections;

3. To validate any candidate biomarkers identified in Aim 2 using reverse transcription-quantitative polymerase chain reaction (RT-qPCR) in an independent cohort.

The structure of this chapter is as follows. Section 5.2 describes the clinical cohorts, the data generation pipelines, and initial dataset exploration. The results are introduced in section 5.3 starting with the discovery phase in section 5.3.1. The discovery phase includes differential expression analysis performed on the RNA-Seq discovery dataset to identify genes that are significantly differentially expressed (SDE) between MIS-C and KD, viral and bacterial infections, as well as feature selection using forward selection-partial least squares (FS-PLS). The candidate biomarkers identified by FS-PLS and differential expression analysis are then validated using RT-qPCR which is described in section 5.3.2. Section 5.4 discusses the results in the context of the existing literature.

As outlined above, MIS-C is a newly emerged paediatric inflammatory disease without a diagnostic test. Diagnosis of MIS-C is achieved through fulfilment of various clinical symptoms which overlap with multiple other paediatric infectious and inflammatory diseases. This chapter contributes the first molecular diagnostic test for MIS-C. The signature presented in this chapter has been successfully translated across platforms to the simpler RT-qPCR platform, with the signature displaying similar performance in RT-qPCR as in RNA-Seq. This paves the way for development of a rapid point-of-care test to distinguish MIS-C from other phenotypically similar diseases. Furthermore, this chapter also contributes the first comprehensive analysis of whole blood host transcriptomes from a reasonably large group of patients with MIS-C, with the only other MIS-C whole blood transcriptomic dataset including 8 MIS-C cases [97]. Finally, this chapter is the only study to contrast the whole blood transcriptomes of children with MIS-C to those from children with a range of other paediatric infectious and inflammatory diseases including KD, viral infections, and bacterial infections.

Declaration of originality for chapter 5:

I carried out all of the work presented in this chapter, except for the following components. Giselle d'Souza, Dr. Samuel Nichols, Oliver Powell and Dr. Victoria Wright performed the laboratory work for the RNA-Seq. Dominic Habgood-Coote processed the raw RNA-Seq data, generating gene and exon counts that I then used. Luca Miglietta, Giselle d'Souza and Dr. Jesus Rodriguez Manzano performed the laboratory work for the RT-qPCR quantification of the gene targets. FS-PLS was developed by Prof. Lachlan Coin.

5.2 Cohorts and study design

This study was composed of two phases: the discovery phase and the validation phase (Figure 27). Gene biomarker candidates for distinguishing between MIS-C and KD, bacterial and viral infections were identified in the discovery phase using host whole blood transcriptomic profiles from RNA-Seq, a high-throughput method for quantifying and characterising RNA molecules present in a sample. Levels of the candidate biomarkers were tested in the validation phase using RT-qPCR, a method that enables rapid detection of the levels of a small number of target RNAs. Further details about RNA-Seq and RT-qPCR are found in chapter 1 section 1.3.1.

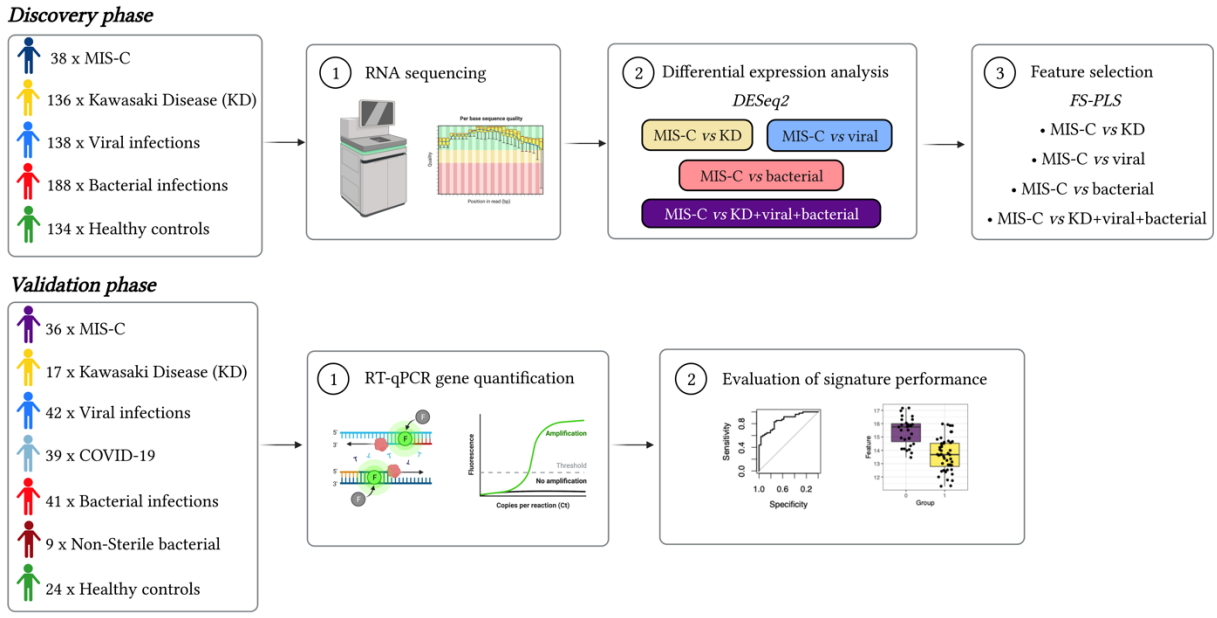


Figure 27: *An overview of the key steps and patient groups included in this chapter. The study was split into two phases: 1) the discovery phase, in which biomarker candidates for diagnosing MIS-C were identified from RNA sequencing; and 2) the validation phase, in which the biomarker candidates were quantified in an independent set of samples using RT-qPCR. KD = Kawasaki disease; HC = healthy controls; FS-PLS = forward selection-partial least squares; RT-qPCR = reverse transcription-quantitative polymerase chain reaction.*

The discovery and validation sets included patients with MIS-C, KD, confirmed bacterial infections (DB), and confirmed viral infections (DV), in addition to healthy controls (HC) which were only used for normalisation of the RNA-Seq counts.

Children with MIS-C included in the discovery and validation sets were recruited into the DIAMONDS study (Diagnosis and Management of Febrile Illness using RNA Personalised Molecular Signature Diagnosis, London – Dulwich Research Ethics Committee: 20/HRA/1714) between April 2020-April 2021. The DIAMONDS study commenced recruitment in 2020 and is ongoing. The DIAMONDS study is a European Union (EU)-funded, multi-centre study that aims to develop a molecular test for the rapid diagnosis of serious infectious and inflammatory diseases using personalised gene signatures. The predecessor of the DIAMONDS study is the PERFORM study (Personalised Risk Assessment in Febrile Illness to optimise Real-life Management across the European Union) which was described in detail in chapter 3. Whilst the PERFORM study aimed to diagnose bacterial *vs*. viral infections, the DIAMONDS study aims to diagnose other diseases such as KD, malaria, and MIS-C. The children with MIS-C were recruited in hospitals in the UK, Greece, and Spain (Imperial College London Teaching Hospitals, London, UK; Alder Hey Hospital, Liverpool, UK; Newcastle University Teaching Hospitals, Newcastle, UK; Evelina Children's Hospital, London, UK; Great Ormond Street Hospital, London, UK; Hospital National and Kapodistrian University of Athens, Greece; Panagiotis & Aglaia Kyriakou Children's Hospital, Athens, Greece; and Hospital 12 de Octubre, Madrid, Spain).

The bacterial, viral and HC patients included in the RNA-Seq discovery dataset were recruited into the EUCLIDS and PERFORM studies. The EUCLIDS study, which is outlined in chapter 3, explored severe bacterial infections in children. The bacterial (including non-sterile DB; NSDB), viral, KD and HC patients in the validation dataset were recruited into either the PERFORM study or the DIAMONDS study, and the children with coronavirus disease 2019 (COVID-19) in the validation dataset were recruited into the DIAMONDS study. The children with KD in the discovery dataset were recruited in the United States through the Kawasaki Disease Research Center at the University of California San Diego (UCSD; see chapter 4 for further details). All patients with KD in the discovery set were recruited before the onset of the SARS-CoV-2 pandemic, ensuring that they were not misclassified MIS-C cases.

Patients were recruited at the earliest time point following presentation to hospital and blood samples for research were taken at the same time as routine clinical investigations where possible. The majority of patients were recruited in hospital accident and emergency departments or on transfer to wards or intensive care units.

In chapter 3, a 6-protein signature was tested on patients with NSDB infections, with this group treated separately to the patients in the DB class. The NSDB group includes pathogens that are only normally identified on mucosal surfaces (e.g., stomach, lungs) or by non-culture methods (e.g., *Campylobacter*, *Salmonella*, *Borrelia burgdorferae*, *Mycoplasma pneumoniae* and *Bordetella pertussis*). In this chapter, the

NSDB patients were included in the DB group due to the predominance of gastrointestinal symptoms in MIS-C cases, and the close resemblance between enteric infections and MIS-C. We aimed to identify a diagnostic signature that could distinguish MIS-C from all bacterial infections, including those from non-sterile sites (i.e., the NSDB group).

Table 14 and Table 15 show the demographic and clinical characteristics of the patients included in the discovery and validation datasets, respectively. MIS-C patients in the discovery dataset were older than the comparator disease groups (median age 126 months *vs*. 30 months for KD, 7 months for viral, and 38 months for bacterial) and had a longer median duration of symptoms prior to presentation (6 days) than the bacterial infection groups (2 days for bacterial), but similar to KD (6 days) and viral infections (5 days) (Table 14). In the validation dataset the patients with viral infections were older (median age 42 months) than those in the discovery dataset (median age 7 months). This is likely to reflect the composition of causative pathogens in each dataset since 54% (*n*=75) of the viral patients in the discovery cohort had respiratory syncytial virus (RSV) (Table 14), which is common in children under the age of 1. This contrasts to just 14% (*n*=6) of the viral patients in the validation cohort (Table 15).

In both datasets (Table 14 and Table 15), MIS-C patients had lower lymphocyte counts than any of the other comparator groups, which fits in with reports of lymphopenia (low lymphocytes) in MIS-C [216]. Monocytes and platelets were also the lowest in the MIS-C group. Levels of C-reactive protein (CRP), which is often used for classifying bacterial patients, were highest in the MIS-C group *vs*. any other comparator group.

A higher proportion of MIS-C patients in the discovery dataset were admitted to paediatric intensive care units (PICU; 60.6%; *n*=23; Table 14) than in the validation set (42%; *n*=14; Table 15). The proportion of MIS-C patients requiring inotropes was similar for the discovery and validation datasets at 52% (*n*=20) and 42% (*n*=15), respectively. Slightly more MIS-C patients required ventilation in the discovery dataset *vs*. the validation dataset. 13.2% (*n*=5) of the patients in the discovery dataset required non-invasive ventilation compared to 6% (*n*=2) of the patients in the validation dataset. 10.5% (*n*=4) of the patients in the discovery dataset required invasive ventilation compared to 8% (*n*=3) of the patients in the validation datasets.

In the RNA-Seq discovery dataset, samples from 134 HC children were included with a median age of 92 months (IQR: 29-154). 47% (*n*=63) of the HC in the discovery cohort were female. HC patients were used for normalisation of the data but not for differential expression analysis or feature selection. In the RT-qPCR validation dataset there were samples from 24 HC children with a median age of 78 months (IQR: 27-119) and 38% (*n*=9) were female.

Table 14: *Demographic and clinical characteristics of the patients with MIS-C, KD, bacterial and viral infections used in the case-control discovery RNA-Seq dataset. This dataset was used to identify potential protein biomarkers for distinguishing between bacterial and viral infections in febrile children. Values are median (IQR) unless stated. IQR: interquartile range; KD = Kawasaki disease; PICU = paediatric intensive care unit; RSV = respiratory syncytial virus; CRP = C-reactive protein.*

	***MIS-C* (n=*38*)**	***KD* (n=*136*)**	***Viral* (n=*138*)**	***Bacterial* (n=*188*)**
Age in months	126 (65-151)	30 (16-52)	7 (2-20)	38 (11-96)
Sex at birth (female, %)	13 (39%)	59 (43%)	45 (33%)	85 (45%)
Days since symptom onset	6 (5-7)	6 (5-8)	5 (2-6)	2 (1-4)
Required inotropic support (n, %)	20 (53%)	2 (1%)	3 (3%)	50 (27%)
Required non-invasive ventilation (n, %)	5 (13%)	2 (1%)	6 (4%)	46 (24%)
Required invasive ventilation (n, %)	4 (11%)	0	7 (5%)	16 (9%)
Admitted to PICU (n, %)	23 (61%)	1 (1%)	25 (18%)	84 (45%)
White blood cells (10^9/L)	9.7 (8.1-11.5)	13.2 (10.4-18.0)	9.9 (6.0-14.0)	13.2 (8.1-20.6)
Neutrophils (10^9/L)	8.1 (4.5-11.0)	9.2 (6.3-11.5)	3.6 (1.8-6.5)	14.1 (6.0-47)
Lymphocytes (10^9/L)	1.2 (0.8-1.8)	2.8 (1.7-4.9)	4.7 (2.8-6.7)	2.8 (1.5-8.6)
Monocytes (10^9/L)	0.3 (0.2-0.4)	0.7 (0.5-1.1)	0.8 (0.5-1.4)	1.6 (0.7-4.8)
Platelets (10^9/L)	209 (144-301)	343 (257-426)	365 (245-509)	243 (157-329)
CRP (mg/L)	173 (99-250)	73 (43-162)	13 (4-38)	117 (40-212)
Most common causative pathogen	NA	NA	RSV (*n*=75) Rhinovirus (*n*=16) Influenza (*n*=7) Rotavirus (*n*=7)	*N. meningitidis* (*n*=46) *S. aureus* (*n*=30) *S. pneumoniae* (*n*=30) Group A strep. (*n*=17)

Table 15: *Demographic and clinical characteristics of the patients with MIS-C, KD, bacterial and viral infections, and COVID-19 used in the RT-qPCR validation dataset. This dataset was used to evaluate the performance of the biomarkers identified from the RNA-Seq discovery dataset. Values are median (IQR) unless stated. IQR = interquartile range; KD = Kawasaki disease; CRP = C-reactive protein; RSV = respiratory syncytial virus; PICU = paediatric intensive care unit; COVID-19 = Coronavirus Disease 2019; SARS-CoV-2 = severe acute respiratory syndrome coronavirus 2.*

	***MIS-C* (n=36)**	***KD* (n=17)**	***Viral* (n=42)**	***Bacterial* (n=50)**	***COVID-19* (n=39)**
Age in months	103 (64-145)	26 (15-46)	42 (18-112)	46 (7-114)	16 (2-162)
Sex at birth (female, %)	12 (33%)	7 (41%)	16 (38%)	25 (50%)	14 (36%)
Days since symptom onset	6 (4-7)	5 (4-8)	4 (2-6)	2 (1-5)	2 (1-5)
Required inotropic support (n, %)	15 (42%)	0	1 (2%)	5 (10%)	1 (3%)
Required non-invasive ventilation (n, %)	2 (6%)	1 (6%)	1 (2%)	3 (6%)	2 (5%)
Required invasive ventilation (n, %)	3 (8%)	0	2 (5%)	12 (24%)	0
Admitted to PICU (n, %)	15 (42%)	1 (6%)	4 (10%)	18 (36%)	2 (5%)
White blood cells (10^9/L)	8.2 (6.5-12.3)	12.3 (8.9-13.9)	10.1 (6.5-13.2)	12.5 (8.1-21.6)	7.9 (5.6-9.3)
Neutrophils (10^9/L)	6.5 (4.5-8.7)	6.7 (4.9-11.6)	5.1 (3.4-7.6)	7.0 (3.4-15.2)	3 (1.9-4.4)
Lymphocytes (10^9/L)	1.1 (0.6-1.6)	2.4 (1.7-5.2)	2.6 (1.7-4.9)	1.9 (0.9-2.9)	2.8 (1.6-4.9)
Monocytes (10^9/L)	0.2 (0.2-0.4)	0.7 (0.5-1.0)	0.9 (0.5-1.3)	0.9 (0.5-1.3)	0.7 (0.5-1.2)
Platelets (10^9/L)	169 (124-251)	382 (279-434)	261 (194-326)	289 (235-332)	294 (230-355)
CRP (mg/L)	193 (80-230)	51 (42-146)	6 (3-22)	57 (13-129)	3 (1-8)
Most common causative pathogen	NA	NA	Adenovirus (*n*=10) Epstein-Barr virus (*n*=7) RSV (*n*=6) Measles (*n*=4)	*E. coli* (*n*=12) *S. pneumoniae* (*n*=7) Group A strep. (*n*=5) S. aureus (*n*=5) *N. meningitidis* (*n*=5)	SARS-CoV-2 (*n*=39)

5.2.1 Case definitions

MIS-C classification was based on patients meeting the full World Health Organisation (WHO) MIS-C case definition [96] as shown in Figure 28. KD diagnosis was based on the American Heart Association (AHA) diagnostic criteria as outlined in chapter 4 section 0 [84]. DV classification was as stated in chapter 3 section 3.2.1 and chapter 4 section 0, which was that patients required identification of virus matching the syndrome in addition to CRP levels ≤60 mg/L and neutrophil levels ≤12×10^9/ml. The DB group included patients in whom an appropriate bacterial pathogen was isolated from a normally sterile site, and clinical presentation was consistent with the pathogen isolated, as described in chapter 3 section 3.2.1 and chapter 4 section 0. In addition, pathogens normally only detected on mucosal surfaces or diagnosed serologically (*M. tuberculosis*, *B. pertussis*, *M. pneumoniae*, *Borrelia* species, *Campylobacter* and *Salmonella*) were included. This is the same as the group known as NSDB infections in chapter 3. As MIS-C frequently presents with gastrointestinal symptoms and can resemble enteric

infections in addition to Lyme disease and mycoplasma infections, we decided to include the NSDB group in the DB group since it was important to distinguish MIS-C from all bacterial infections, including those from non-sterile sites. COVID-19 was diagnosed in children presenting to hospital with febrile or respiratory illness with PCR confirmation of SARS-CoV-2 in respiratory secretions. HCs were otherwise healthy children attending hospitals for non-infectious or inflammatory conditions or for pre-surgery screening who were recruited to EUCLIDS, PERFORM or DIAMONDS. These control children were used in the RNA-Seq dataset solely for normalisation of sequencing runs.

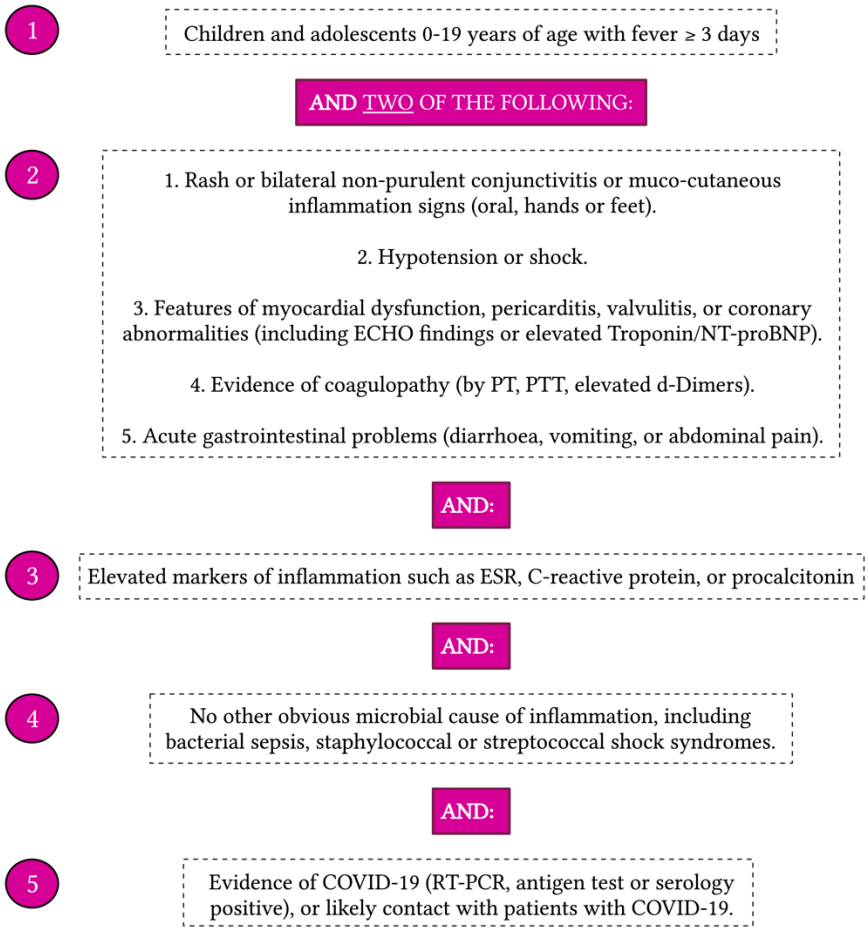


Figure 28: *The World Health Organisation (WHO) case definition for MIS-C. Information taken from: https://www.who.int/news-room/commentaries/detail/multisystem-inflammatory-syndrome-in-children-and-adolescents-with-covid-19 (World Health Organisation, 2020). ECHO = echocardiogram; NT-proBNP = N-terminal pro b-type natriuretic peptide; PT = prothrombin time; PTT = partial thromboplastin time; ESR = erthryocyte sedimentation rate; COVID-19 = Coronavirus Disease 2019.s*

5.2.2 Data generation

5.2.2.1 Discovery RNA Sequencing dataset

The discovery dataset was generated using RNA-Seq on whole blood. Venous blood was collected from patients at the earliest time point after recruitment and prior to treatment with intravenous immunoglobulin (IVIG) or immunomodulators into PAXgene Blood RNA tubes (PreAnalytiX,

Germany) and stored at -80°C until extraction. Total RNA was isolated using recommended methodology (including miRNA for PAXgene blood RNA vacutainers and after additional DNAse treatment) and sent for sequencing at The Wellcome Centre for Human Genetics in Oxford, United Kingdom.

The RNA-Seq protocol was as follows: Genomic material was quantified using RiboGreen (Invitrogen) on the FLUOstar OPTIMA plate reader (BMG Labtech) and the size profile and integrity analysed on the 2200 TapeStation (Agilent, RNA ScreenTape). Input material was normalised, and strand specific library preparation was completed using NEBNext® Ultra™ II mRNA kit (NEB) and NEB rRNA/globin depletion probes following manufacturer's instructions. Libraries were prepared on a Tetrad (Bio-Rad) using in-house unique dual indexing primers (based on [217]). Individual libraries were normalised using Qubit and pooled together. The pooled library was diluted to ~10 nM for storage and denatured and further diluted prior to loading on the sequencer. Sequencing was performed at the Oxford Genomics Centre using a Novaseq6000 platform at 150 base-pair paired-end configuration, generating a raw read count of 30 million reads per sample. Following sequencing, the raw RNA-Seq reads were processed as outlined in detail in chapter 2 section 2.2.1.

The discovery dataset was composed of two batches of RNA-Seq known as EUCLIDS RNA-Seq and PERFORM RNA-Seq due to the majority of the samples originating from EUCLIDS and PERFORM, respectively. In the EUCLIDS RNA-Seq dataset, there were samples from patients with KD, bacterial and viral infections, and HCs. In the PERFORM RNA-Seq, there were samples from patients with MIS-C, bacterial infections, viral infections including COVID-19, and HCs. Prior to merging the two datasets, pre-processing was performed on each dataset separately. Ribosomal genes, genes in the depletion kit (*HBB*, *MT-RNR2*, *MT-RNR1*, *HBA2*, *HBG2*, *HBG1*, *HBA1*), and genes without counts ≥10 in at least three samples were removed. Principal component analysis (PCA) was performed to identify and remove outliers. Outliers were determined using the Hotelling's *t*-squared test with a confidence level of 0.99. Plate and site effects and effects of sampling tube (PAXgene tubes were used for the majority but a small proportion of patients in the EUCLIDS study had samples collected into Tempus tubes) were removed using ComBat-seq from the sva package [129]. ComBat-seq is a batch effect correction method that has been adapted for use with RNA-Seq count data (further details of ComBat-seq are in chapter 2 section 2.3). The counts from the two datasets were merged using only genes present in both datasets. ComBat-seq was used to correct for the experimental technical effects between the two batches and counts were normalised using DESeq2 [137] (details in chapter 2 section 2.3.2). Genes without counts ≥20 in at least three samples were removed from the normalised dataset.

5.2.2.2 Reverse transcription-quantitative polymerase chain reaction validation dataset

Genes identified as candidate biomarkers were quantified by RT-qPCR. A housekeeping gene - glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) – was also quantified for normalisation

purposes. Each gene was quantified using 4 replicates. Primers were designed for the genes included in the diagnostic signature. Gene sequences were downloaded from Ensembl Genome Browser (https://www.ensembl.org/index.html) using as reference human genome the assembly GRCh38.p13. Primer design and *in silico* analysis were conducted using GENEious Prime 2022.0.1 (https://www.geneious.com/). Primer characteristics were evaluated through IDT OligoAnalyzer software (https://eu.idtdna.com/pages/tools/oligoanalyzer) using the J. SantaLucia thermodynamic table for melting temperature (Tm) calculation, hairpin, self-dimer, and cross-primer formation (https://www.thermofisher.com/uk/en/home/brands/thermo-scientific/molecular-biology/molecular-biology-learning-center/molecular-biology-resource-library/thermo-scientific-web-tools/multiple-primer-analyzer.html). Intercalating dye-based PCR assays were designed, developed and purchased from IDT Integrated DNA Technologies (IDT, Coralville, IA).

The performance of the signature was assessed using a Biomark HD high-throughput qPCR platform and 192.24 Dynamic Array integrated fluidic circuit (IFC) following manufacturer's instructions. The experimental workflow consisted of 4 steps: reverse transcription (RT), pre-amplification, on-chip gene expression, and data analysis.

The real-time data from RT-qPCR was pre-processed by performing baseline correction and cycle thresholds (CT) values were extracted using a global threshold, as determined by Fluidigm's Real-Time qPCR Software (version 4.1.2). RT-qPCR is described in more detail in chapter 1 section 1.3.1.2, however a brief explanation of CT values is required. In qPCR, fluorescent reporter molecules are added to the reaction and their levels increase with accumulating amounts of product DNA as the PCR cycles progress. The fluorescence intensity is measured across cycles, and the cycle (i.e., CT) in which the fluorescence intensity is greater than background fluorescence is used as a measure of the target DNA's starting concentration. Lower CT values indicate that fewer qPCR cycles were required for the fluorescence intensity to exceed that of the background fluorescence, indicating that there is a greater quantity of product DNA. Higher CT values are indicative of lower levels of product DNA as more cycles were required for the fluorescence intensity to exceed background fluorescence. In the RT-qPCR experiment presented in this chapter, 40 cycles of qPCR were performed and as such, CT values of 40 were treated as missing. Samples missing measurements for ≥3 of the 4 *GAPDH* assay replicates were excluded. For each sample, the standard deviation for each gene across replicates was calculated. For genes with a standard deviation ≤3 across replicates, the mean CT value was taken for a given sample. For genes with standard deviations across replicates >3, CT values were manually examined and either set to NA or the mean across a subset of replicates was taken if they had a standard deviation ≤3. For each sample and gene, the averaged *GAPDH* values were subtracted from the averaged gene values to obtain normalised gene measurements.

5.2.3 Dataset exploration

PCA was performed on the discovery RNA-Seq dataset to identify the major sources of variation in the dataset (Figure 29). When principal component (PC) 1 was plotted against PC2, the patients cluster by disease group with clear stratification between the viral and bacterial patients and HC (Figure 29A, left). The MIS-C patients are split between the viral, bacterial and KD groups, reflecting the similarities between MIS-C and each of these groups. PC4 clearly captured sex, with two distinct groups corresponding to male and female patients (Figure 29B, right). Younger patients had higher PC2 scores, as did viral patients (Figure 29A, C, left), reflecting the lower age of the viral group (Table 14). Figure 29D shows patients coloured by RNA-Seq batch. It is encouraging to see that there is no clear clustering by batch in any of the PC biplots shown (Figure 29D).

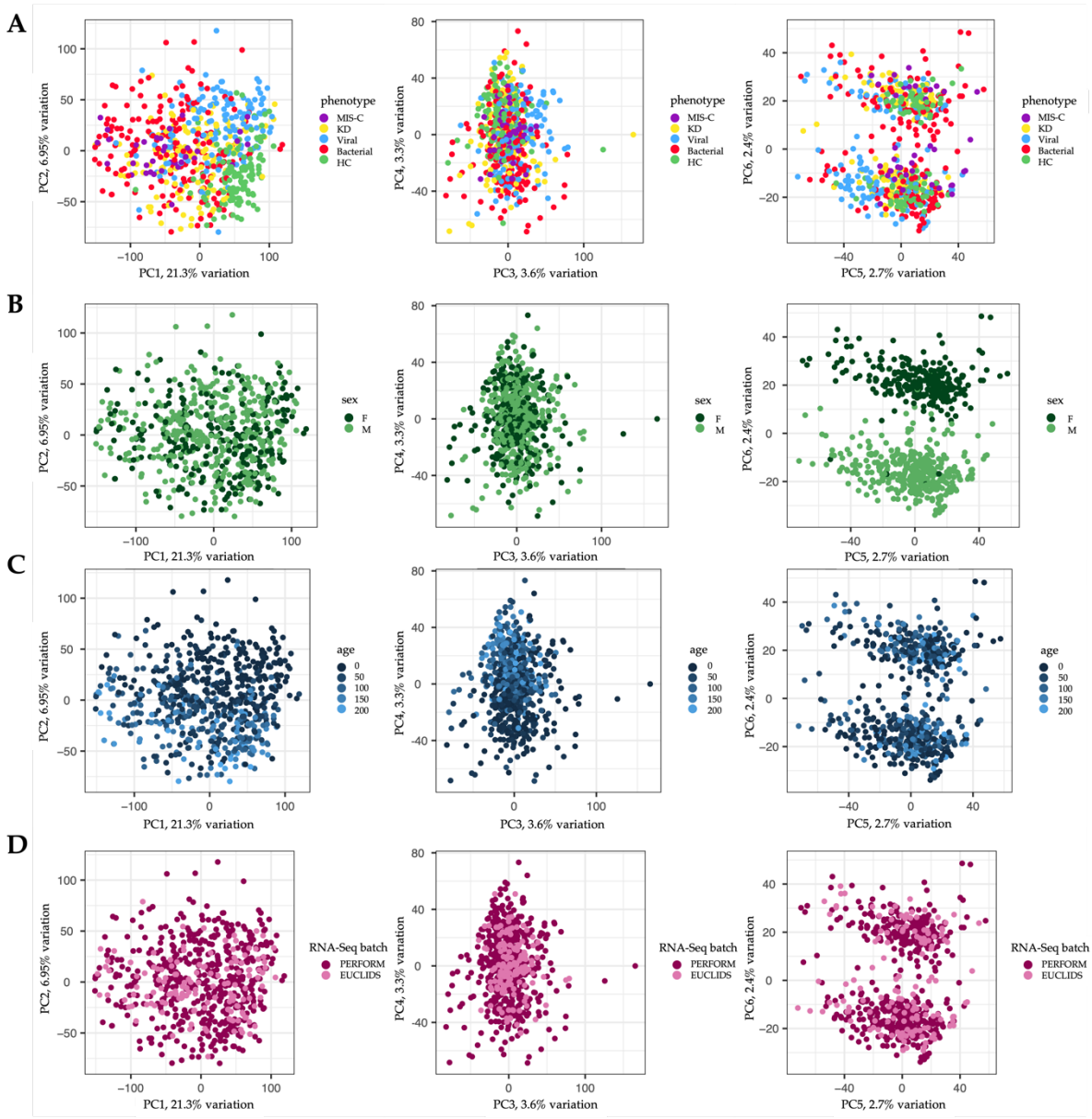


Figure 29: *Principal Component Analysis (PCA) biplots from PCA performed on the discovery RNA-Seq dataset after normalisation. Principal components (PC) 1-2, PC3-4 and PC5-6 are shown. Points represent samples and are coloured by phenotype (**A**), sex at birth (**B**), age in months (**C**), and RNA-Seq batch (**D**). F = female; M = male; KD = Kawasaki disease; MIS-C = multisystem inflammatory syndrome in children.*

5.3 Results

The aim of this study was to identify host gene expression candidate biomarkers for distinguishing between MIS-C and KD, bacterial and viral infections. Host whole blood RNA-Seq data was used to identify the biomarker candidates, which were then validated using RT-qPCR in an independent dataset. A 4-gene signature was identified by feature selection using forward selection-partial least squares (FS-PLS) and was taken forward to validation (*HSPBAP1, VPS37C, TGFB1, MX2*) alongside the top significantly differentially expressed (SDE) genes for MIS-C *vs*. KD, viral and bacterial infections (*TRBV11-2, FOSB, HROB, NCK2, PER1, CHKA*). The discovery process is described in section 5.3.1,

followed by the process of quantifying all 10 biomarker candidate genes using RT-qPCR in section 5.3.2. The performance of the 4-gene FS-PLS signature using RT-qPCR is described first (section 5.3.2.2), followed by section 5.3.2.3 that outlines how the performance of the signature was improved through addition of the 6 SDE genes also validated. The performance of the final signature – the 5-gene RT-qPCR signature - is outlined in section 5.3.2.4 and the results are summarised and discussed in section 5.4.

5.3.1 Discovery phase

Whole blood transcriptomes of MIS-C patients (n=38) and patients with KD (n=136), viral infections (n=138), and bacterial infections (n=188), and healthy controls (n=134) were included in the discovery cohort. The first step was to perform differential expression analysis to identify transcriptomic differences between MIS-C and bacterial infections, viral infections, and KD (Figure 30). Differential expression analysis was performed using DESeq2 [137] with age, sex and RNA-Seq batch included in the model. Genes with Benjamini-Hochberg (BH) [142] adjusted p-values <0.05 were considered SDE. For MIS-C (n=38) *vs.* KD (n=136) 4,786 genes were SDE with 2,681 and 2,105 genes over- and under-expressed in MIS-C, respectively. For MIS-C *vs.* viral infections (n=138) 10,654 genes were SDE, with 5,973 and 4,681 genes over- and under-expressed in MIS-C, respectively. For MIS-C vs. bacterial infections (n=188) 3,718 genes were SDE, with 1,776 and 1,942 genes over- and under-expressed in MIS-C, respectively. When MIS-C was contrasted against KD, viral and bacterial infections as a combined group, 5,696 genes were SDE, with 3,250 and 2,446 genes over- and under-expressed in MIS-C, respectively. The considerable amount of SDE genes between MIS-C and each of the comparator groups highlights that, despite the clinical similarities, these conditions are indeed distinguishable at the transcriptomic level, implying that the immune response to each disease group is unique.

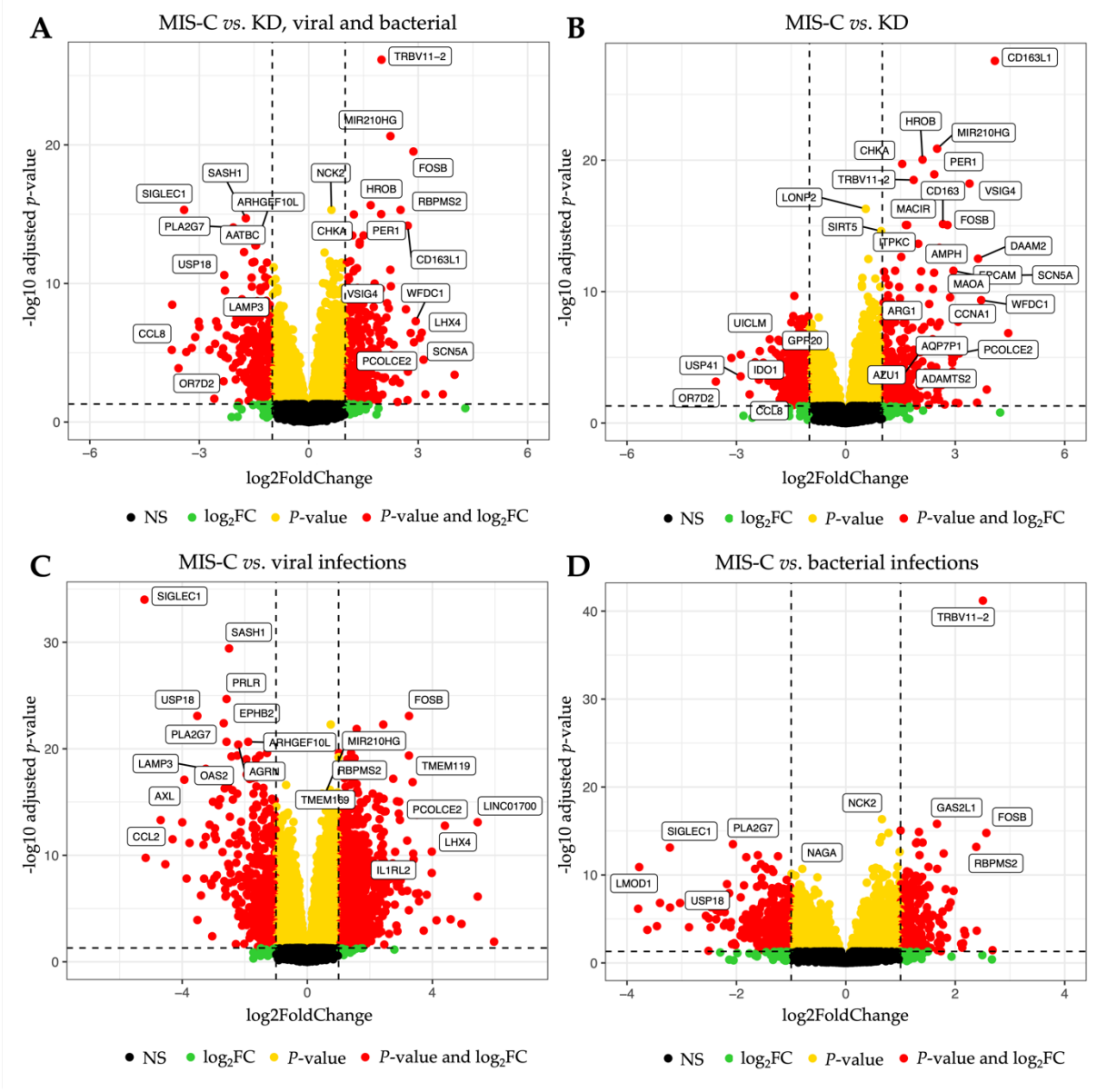


Figure 30: *Volcano plots showing log2 fold change (LFC) values and adjusted -log10 p-values from differential expression analysis comparing* ***A****: MIS-C vs. Kawasaki Disease;* ***B****: MIS-C vs. viral infections;* ***C****: MIS-C vs. bacterial infections;* ***D****: MIS-C vs. KD, viral and bacterial infections combined. Data is from the RNA-Seq discovery dataset. Points represent genes. Points are coloured as followed: red - genes with adjusted p-values <0.05 and absolute LFC >1; yellow – genes with adjusted p-values <0.05 and absolute LFC <1; green – absolute LFC >1; and black = not significant (NS). Genes to the right of 0 are increased in MIS-C patients. MIS-C = multisystem inflammatory syndrome in children.*

FS-PLS was applied to identify a small combination of genes – a diagnostic gene signature. FS-PLS is a feature selection algorithm that uses singular value decomposition (SVD) to project out the residual variance captured in features already included in the signature. As a result, FS-PLS selects a small signature composed of non-correlated features. FS-PLS was also applied to the proteomic datasets presented in chapter 3 section 3.3.1, however a modified version of FS-PLS was used in this chapter. The modified version of FS-PLS enables simultaneous comparison of multiple disease groups. FS-PLS was run with the following comparisons included: MIS-C *vs.* KD+DV+DB (both RNA-Seq datasets); MIS-C *vs.* KD+DV+DB (EUCLIDS RNA-Seq only); MIS-C *vs.* DV+DB (PERFORM RNA-Seq only; no

KD samples present in PERFORM RNA-Seq), MIS-C *vs*. DB; MIS-C *vs*. DV; and MIS-C *vs*. KD. FS-PLS was run using genes with BH adjusted *p*-values <0.001 and absolute log-fold change (LFC) >0.5 in at least one of MIS-C *vs*. KD, *vs*. DV, or *vs*. DB, and with normalised mean counts >50 in all disease groups and >100 in at least one disease group. The mean counts threshold was included to a) improve the chance of translation across to RT-qPCR assays and b) to act as an additional control for batch effects since the two RNA-Seq batches were sequenced at different depth (EUCLIDS RNA-Seq was sequenced at lower depth). FS-PLS was run with leave-10-out cross-validation across 50 iterations and with beam=20, meaning that 20 signatures were identified in parallel, with the highest performing signature selected. The performance of the signature in the whole dataset was evaluated through calculating the area under the receiver operating characteristic (ROC) curve (AUC), using the weighted disease risk score (DRS; see chapter 2 section 2.4.4). The performance in differentiating MIS-C from each individual disease group was also tested.

Variable selection using FS-PLS identified a 4-gene signature that distinguished MIS-C from KD, viral and bacterial infections: *HSPBAP1* (HSPB1 Associated Protein 1), *VPS37C* (Vacuolar Protein Sorting-Associated Protein 37C), *TGFB1* (transforming growth factor beta 1) and *MX2* (MX Dynamin Like GTPase 2). The gene expression values of the four genes were combined into a single score for each patient – the weighted DRS. The weighted DRS was calculated by multiplying the expression values of the genes in the signature by their weights/coefficients which were obtained from a generalised linear model (GLM) contrasting MIS-C to KD, viral and bacterial infections. The weighted DRS values were then used for ROC models and to calculate the AUC. The 4-gene signature achieved an AUC of 95.4% (95% confidence interval [CI]: 92.5%-98.4%; Figure 31A) in distinguishing between MIS-C *vs*. KD, viral and bacterial in the discovery dataset. When the individual comparator groups were considered separately, the 4-gene signature achieved an AUC of 92.5% (95% CI: 87.7%-97.3%; Figure 31B) for MIS-C *vs*. KD, 98.6% (95% CI: 97%-100%; Figure 31C) for MIS-C *vs*. viral infections and 95.2% (95% CI: 92%-98.3%; Figure 31D), for MIS-C *vs*. bacterial infections.

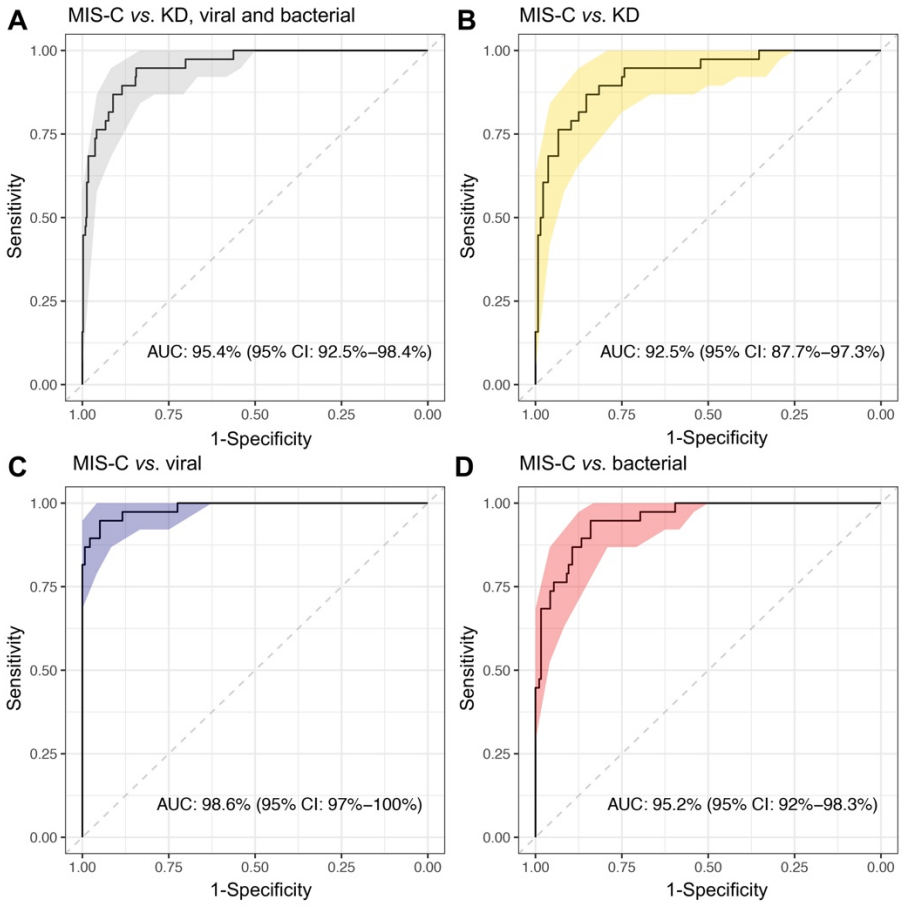


Figure 31: *ROC curves for the performance of the 4-gene signature identified by FS-PLS in the RNA-Seq discovery dataset. The 4-gene signature is composed of* HSPBAP1, VPS37C, TGFB1 *and* MX2. ***A****: performance of the signature in distinguishing between MIS-C* vs. *KD+viral+bacterial combined;* ***B****: performance of the signature for MIS-C* vs. *KD;* ***C****: performance of the signature for MIS-C* vs. *viral; and* ***D****: performance of the signature for MIS-C* vs. *bacterial. AUCs and 95% confidence intervals (CI) are printed on the plots. KD = Kawasaki disease; MIS-C = multisystem inflammatory syndrome in children.*

The genes included in the 4-gene signature were taken forward to RT-qPCR validation. Many gene targets fail to translate across platforms from RNA-Seq to low-throughput quantification methods such as RT-qPCR due to various factors. This occurs when the discovery dataset contains latent factors that are driving the selection of the gene targets, e.g., batch effects. The RNA-Seq data presented here originated from two batches, with all MIS-C patients coming from the RNA-Seq PERFORM batch and all KD patients coming from the RNA-Seq EUCLIDS batch. Whilst we attempted to remove the batch effects, it is possible that they still persist due to the confounding between disease group and batch. As a result, we decided to introduce redundancy into the gene set for subsequent transfer to RT-qPCR assays to improve the chances of successful translation across platforms. Therefore, we selected the top 6 SDE genes between MIS-C and KD, viral and bacterial combined (Figure 30A) with base mean counts >10 and positive LFCs, as shown in Table 16.

Table 16: *The top SDE genes between MIS-C vs. KD, viral and bacterial combined that were taken forward to RT-qPCR validation. The genes were identified by DESeq2 on the RNA-Seq discovery dataset. BH = Benjamini-Hochberg; LFC = log-fold change.*

Gene short name	Gene long name	BH-adjusted *p*-value	LFC
TRBV11-2	T Cell Receptor Beta Variable 11-2	7.14×10^{-27}	1.99
FOSB	FosB Proto-Oncogene	2.99×10^{-20}	2.86
HROB	Homologous Recombination Factor With OB-Fold	2.24×10^{-16}	1.70
NCK2	NCK Adaptor Protein 2	4.91×10^{-16}	0.62
PER1	Period Circadian Regulator 1	9.87×10^{-16}	1.98
CHKA	Choline Kinase Alpha	1.05×10^{-15}	1.23

Overall, 10 genes were taken forward to validation (Table 17) with the expression levels of 8 genes increasing in MIS-C *vs.* the comparator groups (*TRBV11-2*, *HSPBAP1*, *TGFB1*, *FOSB*, *HROB*, *NCK2*, *CHKA*, *PER1*), and the levels of 2 genes decreasing in MIS-C (*VPS37C*, *MX2*). The 10 genes identified as candidate biomarkers in the RNA-Seq discovery phase are shown in Figure 32.

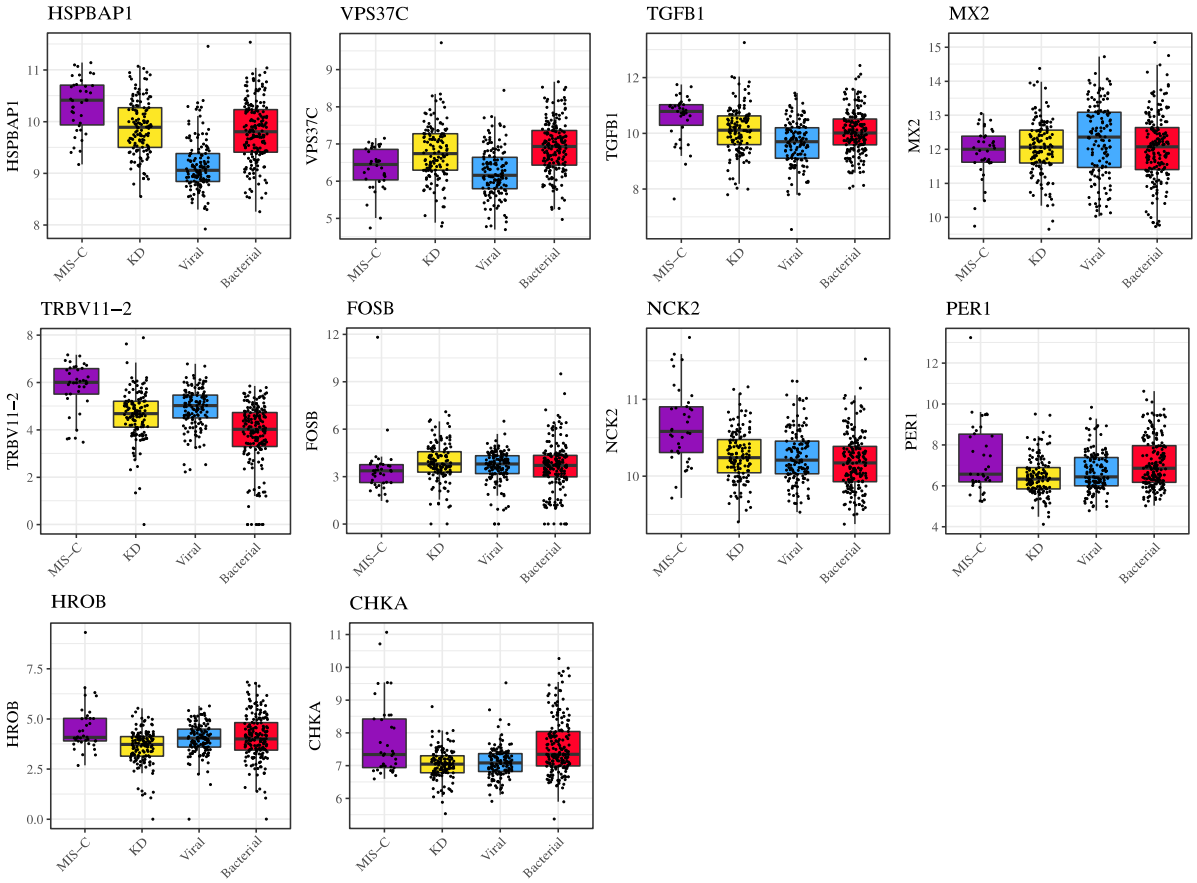


Figure 32: *Gene expression levels in the RNA-Seq data of the 10 genes identified as candidate biomarkers for distinguishing between MIS-C and KD, viral and bacterial infections from the RNA-Seq discovery dataset.* HSPBAP1, VPS37C, TGFB1 *and* MX2 *were selected by FS-PLS.* TRBV11-2, FOSB, NCK2, PER1, HROB *and* CHKA *were selected as they were the top SDE genes between MIS-C* vs. *KD, viral and bacterial. KD = Kawasaki disease; MIS-C = multisystem inflammatory syndrome in children.*

5.3.1.1 Exon analysis

The discovery of candidate biomarkers using RNA-Seq data was performed using gene counts which capture both introns and exons for a given gene. Only exons are subsequently transformed into complementary DNA (cDNA), which is the starting material for RT-qPCR gene quantification. As a result, it was necessary to explore the coding regions of the target genes to identify the optimal region of the gene to target during RT-qPCR. Since most genes contain multiple exons, we examined the exon counts of the 10 genes identified during the RNA-Seq discovery phase and selected the optimal exon for each gene based on: the LFC between MIS-C *vs.* viral and bacterial; AUC for MIS-C *vs.* viral and bacterial; the mean exon counts overall and within the MIS-C, viral and bacterial groups individually; and the exon length. Exon counts for the PERFORM RNA-Seq batch were used since this batch included MIS-C patients, however it did not include KD. The exons chosen for each candidate biomarker and the values of the metrics described above are shown in Table 17. Primers were designed based on the exons selected.

Table 17: *The exons selected for each of the candidate biomarker genes identified from the RNA-Seq discovery dataset. Metrics used to select the exons are shown in the table. Mean exon counts are shown. CI = confidence interval. MIS-C = multisystem inflammatory syndrome in children.*

Gene name	ID	AUC (95% CI)	Mean exon counts	MIS-C mean counts	Bacterial and viral mean counts	Exon length	Exon fold change
HSPBAP1	ENSE00001900976	85.7% (77.5%-93.9%)	49.9	74.8	34.4	438	1.12
VPS37C	ENSE00003510659	62.9% (52.0%-73.8%)	20.0	17.1	21.9	171	-0.36
TGFB1	ENSE00001196164	50.5% (38.8%-62.3%)	1752.4	1785.5	1731.7	887	0.04
MX2	ENSE00003494762	75.9% (66.5%-85.2%)	269.4	149.7	343.9	254	-1.20
TRBV11-2	ENSE00002493270	85.9% (77.8%-94.1%)	48.5	87.4	24.3	297	1.84
FOSB	ENSE00002819874	57% (4.6%-68.4%)	5.2	9.2	2.8	277	1.73
NCK2	ENSE00001750070	69.1% (58.2%-80%)	93.2	109.7	83.0	238	0.40
PER1	ENSE00003631592	56.3% (43.9%-68.8%)	11.4	16.7	8.0	189	1.06
HROB	ENSE00001390690	61.9% (50.3%-73.5%)	18.2	25.3	13.8	1169	0.87
CHKA	ENSE00002177366	59.3% (47.6%-70.9%)	34.6	40.1	31.1	227	0.37

5.3.2 Validation phase

5.3.2.1 Assay optimisation

Ten candidate biomarker genes in addition to *GAPDH* were taken forward to validation using RT-qPCR. *GAPDH* was selected as a housekeeping gene. RT-qPCR was selected as the gene quantification platform for validation since it is closer to the type of test that could be used as a rapid point-of-care diagnostic test. In fact, various devices that can provide RT-qPCR results rapidly are now available, including several platforms already in use in clinical settings worldwide [218-220].

For each candidate biomarker gene, four assays were developed initially with each assay targeting different regions of the gene. These assays were run on a small subset of 48 patients with the aim of identifying the optimal assay for each gene which would correspond to the region of the gene that translated best across platforms. The following patient groups were included: MIS-C (n=13); KD (n=5); viral infections (n=4); bacterial infections (n=8); COVID-19 (n=4); and HC (n=8). For each assay, the mean normalised CT value was quantified across all samples and within each disease group, in addition to the AUC and LFC for MIS-C *vs.* KD, viral and bacterial infections. Assays with lower mean CT value (i.e., higher expression), higher AUCs and LFC with concordant directions to the directions reported in the RNA-Seq discovery dataset were selected.

Once the optimal assays were selected for each gene, the primers were designed (Table 18) and the selected assays were run using all patients in the validation dataset (Table 15). The ten candidate biomarker genes were quantified using RT-qPCR on samples from patients with MIS-C (n=36), KD (n=17), viral infections (n=42), bacterial infections (n=50), and COVID-19 (n=39) (Figure 33). CT values were normalised using the *GAPDH* housekeeping gene (gene-*GAPDH*). Due to the nature of CT values, trends are reversed, and larger CT values indicate lower expression and vice versa.

Table 18: *Primer sets developed in this study. These primers were used for quantification of the genes selected for validation using RT-qPCR. Primers were developed for the optimal assays identified in the optimisation stage. GAPDH was used as the reference gene.*

Gene name	Exon Ensembl ID	Forward primer name	Forward primer sequence	Reverse primer name	Reverse primer sequence
CHKA	ENSE00003515000	CHKA_forward	GGCGGCCTTAG CAACATG	CHKA_reverse	TCTGCAAAATCG CTCCATACA
FOSB	ENSE00002819874	FOSB_forward	GACAGCGTACT TTGAGGACTC	FOSB_reverse	AACATTTCCCTG GGCACAG
HROB	ENSE00001390690	HROB_forward	TGTGCCTGCCAT CCACAA	HROB_reverse	GTGGGAACTGTT AAGGCAGTT
HSPBAP1	ENSE00001900976	HSPBAP1_forward	CTCCTGTTTCTA GCAAGACTCAA	HSPBAP1_reverse	TTCCCAATTAAG TTCAGCCACT
MX2	ENSE00003494762	MX2_forward	GGACGCCTTCA CAGAATATGA	MX2_reverse	ATTTCAGTAAAG GAGGAAACCGA
NCK2	ENSE00001750070	NCK2_forward	CAGGAGCTGGA CATCAAGAA	NCK2_reverse	AGTGTGTCCTTC AGGTTCTTC
PER1	ENSE00003631592	PER1_forward	CCAGGTGATTA AGTACGTGCT	PER1_reverse	GGCACCTGGTAG GTCATCAT
TGFB1	ENSE00001196164	TGFB1_forward	ATTTATTGAGC ACCTTGGGCA	TGFB1_reverse	TCAGAGTGTTGC TATGGTGAC
TRBV11-2	ENSE00002493270	TRBV11-2_forward	ATCTGGCCATG CTACCCTTTA	TRBV11-2_reverse	AAGCTTTGCAGG TTGGATCTT
VPS37C	ENSE00001119066	VPS37C_forward	AATTTTCTTCAG CACTGCAGC	VPS37C_reverse	GACTCTTCTTCG ATCTTCATGC
GAPDH	ENSE00000111640	GAPDH_forward	GACAGTCAGCC GCATCTTC	GAPDH_forward	ACTCCGACCTTC ACCTTCC

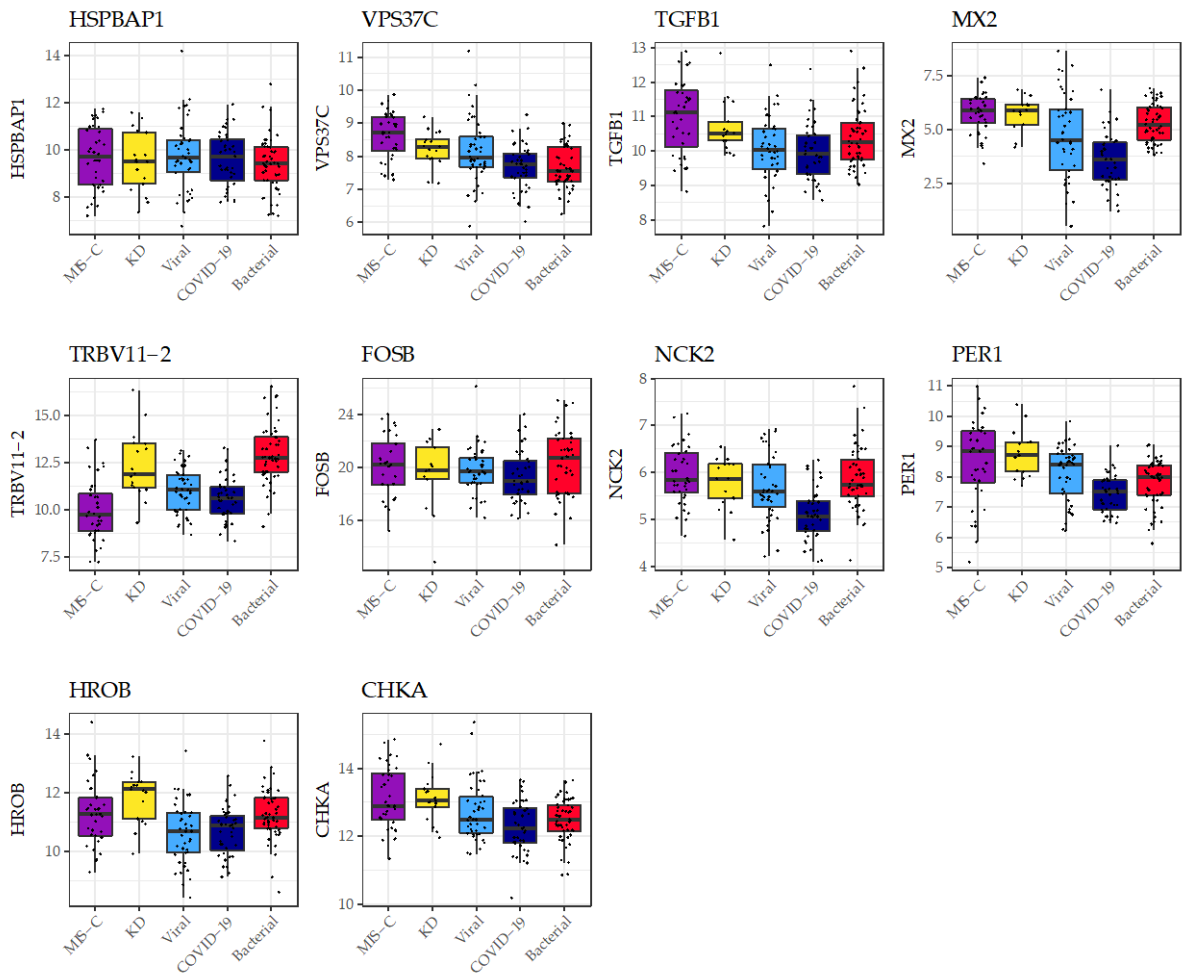


Figure 33: *Boxplots showing normalised CT values for each of the candidate biomarker genes taken forward to validation. Data is shown for the full RT-qPCR validation dataset.* HSPBAP1, VPS37C, TGFB1 *and* MX2 *were originally selected by FS-PLS in the discovery dataset, whilst* TRBV11-2, FOSB, NCK2, PER1, HROB *and* CHKA *were selected through differential expression analysis contrasting MIS-C to KD, viral and bacterial infections. KD = Kawasaki disease; COVID-19 = Coronavirus Disease 2019; MIS-C = multisystem inflammatory syndrome in children.*

5.3.2.2 4-gene FS-PLS signature performance

First, the performance of the 4-gene signature (*HSPBAP1, VPS37C, TGFB1, MX2*) identified by FS-PLS in the discovery RNA-Seq dataset was evaluated. Whilst 6 additional genes were quantified using RT-qPCR, we wanted to study the performance of the 4-gene signature alone prior to adding in any of the 6 additional SDE genes (*TRBV11-2, FOSB, NCK2, PER1, HROB, CHKA*). A weighted DRS was calculated for each patient. The weighted DRS was calculated through multiplying the normalised CT values of the features in the signature by weights (or coefficients) obtained from a GLM contrasting MIS-C to KD, viral and bacterial infections. The weighted DRS were used for ROC models contrasting MIS-C to the comparator groups together or individually and AUCs were calculated (Figure 34). Since the signature was not discovered using COVID-19 patients, the MIS-C *vs.* "all" comparison was performed with and without COVID-19 to evaluate the performance of the signature using the groups that the signature was discovered on. The 4-gene signature achieved an AUC of 79.2% (95% CI: 70.6%-87.7%;

Figure 34A) when distinguishing MIS-C from all groups including COVID-19 and an AUC of 77.2% (95% CI: 68%-86.3%; Figure 34B) when distinguishing MIS-C from all groups excluding COVID-19. When broken down into the individual groups, the 4-gene signature performed the best when distinguishing MIS-C from COVID-19 (84.8%, 95% CI: 75.8%-93.7%; Figure 34F), despite not having been discovered on COVID-19 patients. The second-best performance was for MIS-C *vs*. bacterial infections, with an AUC of 82.4% (95% CI: 73%-91.8%; Figure 34E). For MIS-C *vs*. KD and MIS-C *vs*. viral infections, the performance was lower at 63.1% (95% CI: 47.3%-78.8%; Figure 34C) and 76.7% (95% CI: 65.9%-87.4%; Figure 34D), respectively. Overall, despite the promising performance in distinguishing MIS-C from COVID-19 and bacterial infections, the 4-gene signature did not perform as well as it did in the discovery dataset. For the MIS-C *vs*. viral comparison, this drop in performance could be explained by the different viral cohorts included in the discovery and validation datasets (Table 14 and Table 15). Young patients with RSV dominated the viral patients in the discovery dataset, whilst the viral patients in the validation dataset were older and had a larger range of causative pathogens. The 4-gene signature may have been optimised for distinguishing MIS-C from RSV, rather than a wide range of viral infections.

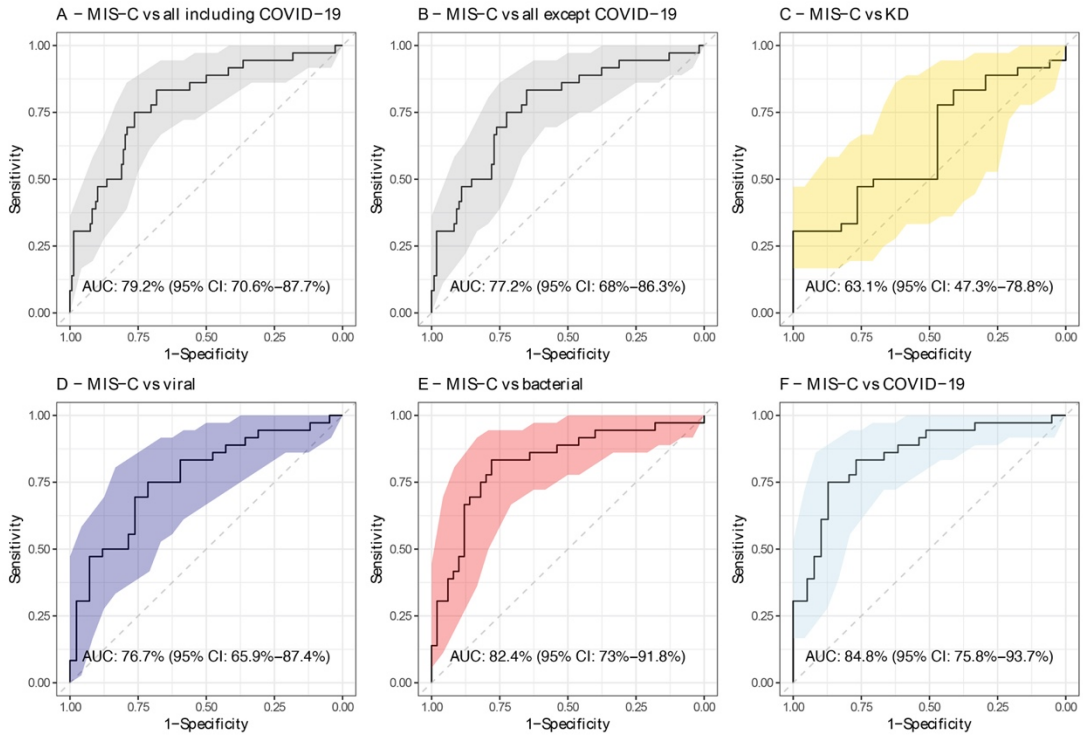


Figure 34: *Performance of the 4-gene signature identified by FS-PLS with performance shown for the RT-qPCR validation dataset.* ***A-B****: ROC curve for evaluating the performance of the signature in contrasting MIS-C from all comparator groups including or excluding COVID-19.* ***C-F****: ROC curve showing performance for MIS-C* vs. *individual comparator groups. AUCs and 95% confidence intervals (CI) are printed on the plots. KD = Kawasaki disease; MIS-C = Multisystem Inflammatory Syndrome in Children.*

5.3.2.3 *Improvement of the 4-gene signature using the 6 significantly differentially expressed genes*

A total of 10 genes were taken forward to validation by RT-qPCR (4 x genes in FS-PLS signature; 6 x SDE genes). Given the lower performance of the 4-gene signature in distinguishing MIS-C from viral infections and KD (Figure 34C, D), we examined the discriminatory performance of sets of genes combining the RNA-Seq 4-gene signature with the 6 additional top SDE genes (*TRBV11-2*, *FOSB*, *HROB*, *NCK2*, *PER1*, *CHKA*). All 6 SDE genes were added to the 4-gene signature individually, then pairwise combinations of the genes were added, followed by combinations of 3, 4, and 5 genes, and finally adding all 6 genes together. The weighted DRS was calculated for each patient using the "new signature", and then used for ROC models and AUC calculations. The improvements in performance were evaluated by contrasting the ROC model for MIS-C *vs.* KD, viral and bacterial between the 4-gene signature to each new signature using the roc.test function in the pROC package [177]. Out of the 63 new signatures generated by using different combinations of the 6 genes in addition to the 4-gene signature, 32 led to significant improvements (BH-adjusted *p*-value <0.05) and are shown in Table 19.

The combination that led to the largest improvement based on BH-adjusted *p*-values was when all 6 SDE genes were added to the 4-gene signature, leading to a 10-gene signature and an AUC of 94.6% for MIS-C *vs*. KD, viral and bacterial. In host signature diagnostics however, the size of the signature is important for translation with a preference towards smaller signatures. This is because it is experimentally easier and cheaper to quantify a smaller number of features and some platforms may have limits on the number of features that can be quantified in parallel. A signature improvement score (SIS) for each signature was calculated to identify the combinations of additional genes that were the most efficient in improving signature performance relative to the number of genes added. The SIS was calculated as follows:

$$Signature\ improvement\ score_i = \frac{\sum \Delta\ AUC_{ci}}{no.additional\ genes}$$

where *i* represents each new signature, Δ represents the change in AUC as a decimal; and c represents each comparison for MIS-C *vs*. KD, viral and bacterial, MIS-C *vs*. KD, MIS-C *vs.* viral, MIS-C *vs*. bacterial.

The highest SIS was obtained for a 5-gene signature composed of the FS-PLS 4-gene signature combined with *TRBV11-2* which had a score of 0.714 (Table 19). The addition of the single gene led to AUC improvements of 16.1% for MIS-C *vs*. KD, viral and bacterial, 27.8% for MIS-C *vs*. KD, 12.4% for MIS-C *vs*. viral, and 15.2% for MIS-C *vs*. bacterial. As a result, this combination of 5 genes (*HSPBAP2*, *VPS37C*, *TGFB1*, *MX2* and *TRBV11-2*) was explored in further analyses. All of the signatures that led to significant performance improvements contained *TRBV11-2*, highlighting the importance of this gene

in MIS-C. *TRBV11-2* was not selected in the original FS-PLS signature due to the mean counts threshold that we introduced. *TRBV11-2* counts were below the threshold for the non-MIS-C disease groups.

Table 19: *The performance of the new signatures composed of different combinations of the 6 SDE genes in addition to the 4-gene signature. Only signatures that led to significant BH-adjusted roc.test* p*-values are shown. Total change of performance is the improvement in AUC for MIS-C vs. KD, viral and bacterial, MIS-C vs. KD, MIS-C vs. viral, and MIS-C vs. bacterial. roc.test* p*-values are BH-adjusted. SIS = signature improvement score (calculated by total change in performance/number of genes added). BH = Benjamini-Hochberg; MIS-C = multisystem inflammatory syndrome in children; KD = Kawasaki disease.*

Gene(s) added	MIS-C *vs.* KD, viral, bacterial	MIS-C *vs.* KD	MIS-C *vs.* viral	MIS-C *vs.* bacterial	roc.test *p*-value	Number of genes added	Total change in performance	SIS
TRBV11-2	93.2%	90.8%	89.0%	97.6%	6.55×10^{-04}	1	0.714	0.714
CHKA, TRBV11-2	93.8%	89.5%	91.6%	97.0%	7.56×10^{-04}	2	0.726	0.363
FOSB, TRBV11-2	93.1%	90.8%	88.9%	97.3%	8.45×10^{-04}	2	0.709	0.354
HROB, TRBV11-2	94.4%	90.7%	91.3%	98.3%	3.68×10^{-04}	2	0.753	0.377
NCK2, TRBV11-2	93.7%	92.8%	89.6%	97.6%	3.68×10^{-04}	2	0.744	0.372
PER1, TRBV11-2	93.9%	89.9%	90.8%	97.8%	3.68×10^{-04}	2	0.731	0.366
CHKA, FOSB, TRBV11-2	93.8%	90.2%	91.3%	97.2%	6.47×10^{-04}	3	0.732	0.244
CHKA, HROB, TRBV11-2	94.2%	90.2%	91.7%	97.7%	3.68×10^{-04}	3	0.745	0.248
CHKA, NCK2, TRBV11-2	94.3%	92.2%	91.5%	97.4%	2.39×10^{-04}	3	0.761	0.254
CHKA, PER1, TRBV11-2	94.0%	89.5%	91.7%	97.4%	3.72×10^{-04}	3	0.734	0.245
FOSB, HROB, TRBV11-2	94.3%	90.2%	91.7%	97.9%	3.68×10^{-04}	3	0.749	0.250
FOSB, NCK2, TRBV11-2	94.0%	92.3%	90.1%	97.8%	2.39×10^{-04}	3	0.750	0.250
FOSB, PER1, TRBV11-2	93.8%	90.0%	90.7%	97.6%	4.18×10^{-04}	3	0.728	0.243
HROB, NCK2, TRBV11-2	94.7%	93.1%	91.7%	97.8%	2.39×10^{-04}	3	0.780	0.260
HROB, PER1, TRBV11-2	94.3%	90.4%	91.4%	98.1%	2.39×10^{-04}	3	0.748	0.249
NCK2, PER1, TRBV11-2	94.4%	91.7%	91.1%	98.1%	2.39×10^{-04}	3	0.760	0.253
CHKA, FOSB, HROB, TRBV11-2	94.1%	90.0%	91.8%	97.5%	3.91×10^{-04}	4	0.742	0.185
CHKA, FOSB, NCK2, TRBV11-2	94.5%	91.7%	91.9%	97.6%	1.92×10^{-04}	4	0.763	0.191

CHKA, FOSB, PER1, TRBV11-2	93.9%	89.7%	91.5%	97.2%	4.41×10^{-04}	4	0.730	0.183
CHKA, HROB, NCK2, TRBV11-2	94.9%	92.2%	92.5%	97.9%	1.69×10^{-04}	4	0.782	0.196
CHKA, HROB, PER1, TRBV11-2	94.3%	90.2%	91.8%	97.8%	2.24×10^{-04}	4	0.748	0.187
CHKA, NCK2, PER1, TRBV11-2	94.3%	91.5%	91.7%	97.4%	1.69×10^{-04}	4	0.757	0.189
FOSB, HROB, NCK2, TRBV11-2	94.8%	92.5%	92.3%	97.6%	1.69×10^{-04}	4	0.778	0.195
FOSB, HROB, PER1, TRBV11-2	94.4%	89.9%	91.9%	98.1%	2.44×10^{-04}	4	0.749	0.187
FOSB, NCK2, PER1, TRBV11-2	94.2%	92.0%	91.1%	97.6%	1.69×10^{-04}	4	0.756	0.189
HROB, NCK2, PER1, TRBV11-2	94.7%	93.0%	91.7%	97.8%	1.69×10^{-04}	4	0.778	0.195
CHKA, FOSB, HROB, NCK2, TRBV11-2	94.8%	92.2%	92.4%	97.7%	8.13×10^{-05}	5	0.777	0.155
CHKA, FOSB, HROB, PER1, TRBV11-2	94.2%	90.4%	91.8%	97.5%	2.66×10^{-04}	5	0.746	0.149
CHKA, FOSB, NCK2, PER1, TRBV11-2	94.5%	91.7%	91.9%	97.6%	9.07×10^{-05}	5	0.764	0.153
CHKA, HROB, NCK2, PER1, TRBV11-2	94.8%	92.5%	92.1%	97.8%	8.13×10^{-05}	5	0.779	0.156
FOSB, HROB, NCK2, PER1, TRBV11-2	94.6%	93.0%	91.9%	97.5%	8.13×10^{-05}	5	0.778	0.156
CHKA, FOSB, HROB, NCK2, PER1, TRBV11-2	94.6%	92.3%	92.0%	97.5%	5.02×10^{-05}	6	0.771	0.129

5.3.2.4 Performance of the 5-gene RT-qPCR signature

This 5-gene combination, now referred to as the 5-gene RT-qPCR signature, was selected as the optimal diagnostic signature out of all combinations of the 4-gene signature and the 6 SDE genes. ROC curves and AUCs for the 5-gene RT-qPCR signature, calculated using the weighted DRS, are shown in Figure 35. The AUC for the 5-gene signature for MIS-C *vs*. all groups including COVID-19 was 90.8% (95% CI: 85.8%-95.8%; Figure 35A), and when COVID-19 was excluded, the AUC was 93.2% (95% CI: 88.8%-97.7%; Figure 35B). The performance improved considerably for MIS-C *vs*. KD and MIS-C *vs*. viral, leading to AUCs of 90.8% (95% CI: 82%-99.7%; Figure 35C) and 89.0% (82.0%-96.0%; Figure 35D), respectively. MIS-C *vs*. bacterial had the highest AUC, at 97.6% (95% CI: 94.8%-100%; Fig. 8E). The weighted DRS values are shown for each patient in Figure 36.

The boxplot in Figure 36 and ROC curves, especially Figure 35C-E demonstrate superior performance of the 5-gene RT-qPCR signature for MIS-C *vs*. KD or MIS-C *vs*. bacterial than MIS-C *vs*. viral. The high performance of the 5-gene RT-qPCR signature in distinguishing MIS-C from KD is highly encouraging given the particular difficulty in distinguishing MIS-C from KD. Early in the pandemic, there were reports from several countries documenting an upsurge in KD-like inflammatory syndromes [211, 221], which with hindsight were likely misdiagnoses of MIS-C as KD. Equally, being able to differentiate MIS-C from bacterial infections is extremely important so that children are administered the most appropriate treatment, i.e., antibiotics in the case of bacterial infections. Similarities between MIS-C and sepsis, septic shock, staphylococcal, and streptococcal shock syndromes present considerable difficulties in terms of managing patients [87, 222, 223], as administering immunomodulating agents (the usual treatment for MIS-C) to patients with severe infections such as sepsis may be deleterious. Additionally, many patients who turn out to have MIS-C receive prolonged unnecessary courses of broad-spectrum antibiotics before the diagnosis is made. Therefore, it is particularly encouraging that the 5-gene RT-qPCR signature achieves such good performance in distinguishing between MIS-C *vs*. bacterial infections.

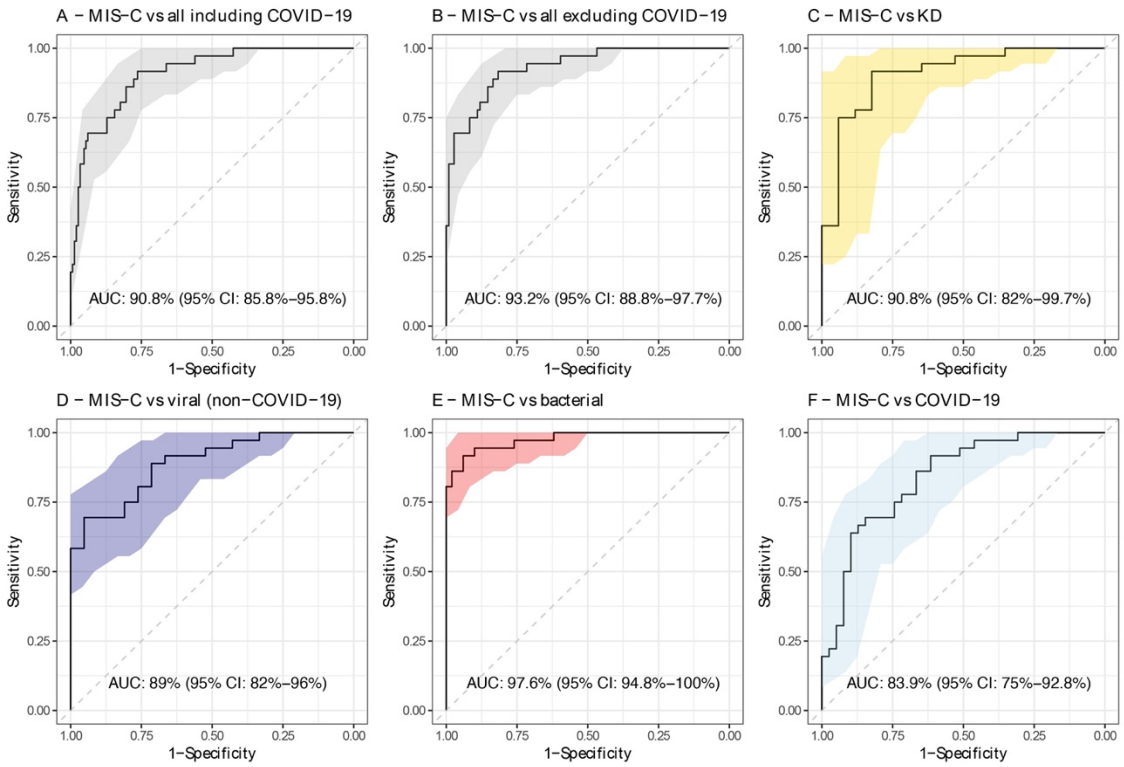


Figure 35: *Performance of the 5-gene RT-qPCR signature (*HSPBAP1, VPS37C, TGFB1, MX2, TRBV11-2*) in the RT-qPCR validation set. The 5-gene signature was derived from adding* TRBV11-2 *to the 4-gene signature from FS-PLS.* ***A-B****: ROC curve for evaluating the performance of the 5-gene RT-qPCR signature in contrasting MIS-C from all comparator groups without (****A****) and with (****B****) COVID-19.* ***C-F****: performance in contrasting MIS-C from individual comparator groups. AUCs and 95% confidence intervals (CI) are printed on the plots. KD = Kawasaki disease; MIS-C = multisystem inflammatory syndrome in children; COVID-19 = Coronavirus Disease 2019.*

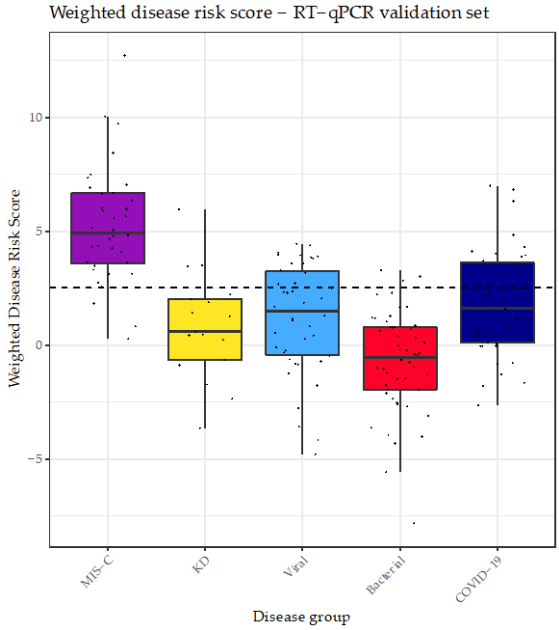


Figure 36: *Boxplots showing the weighted disease risk score for patients from the different disease groups. The dashed line represents the threshold of classification calculated using the Youden's index by maximising both sensitivity and specificity of MIS-C vs. KD, viral and bacterial. Sensitivity = 91.7% and specificity = 81.7% for MIS-C vs. KD, viral and bacterial. MIS-C = multisystem inflammatory syndrome in children; KD = Kawasaki disease; COVID-19 = Coronavirus Disease 2019.*

5.3.2.5 Performance of the 5-gene RT-qPCR signature using cross validation

The performance estimates of the 5-gene RT-qPCR signature shown in ROC plots in Figure 35 were calculated using the weighted DRS with gene weights estimated using the RT-qPCR data. As such, it is possible that these are over-estimations since the model is likely to be overfitting to the data. In the absence of an additional independent validation cohort, it is challenging to obtain a true estimate of the 5-gene RT-qPCR signature. Since this combination of 5 genes was not identified by FS-PLS, we cannot use the FS-PLS weights as these were estimated for the 4-gene signature in the absence of *TRBV11-2*. To overcome this issue, we used 10-fold cross validation (using the caret package [224]), with probabilities of class membership (MIS-C or not MIS-C) returned for every patient in the left-out data. The probability of MIS-C was used to calculate ROC models and AUCs.

Using 10-fold cross validation, the 5-gene RT-qPCR signature maintained excellent performance, with an overall AUC of 91.7% (95% CI: 85.6%-96.9%) for MIS-C *vs*. KD, viral and bacterial (Figure 37A). As with the models using weighted DRS, the MIS-C *vs*. bacterial comparison had the highest AUC at 96.4% (95% CI: 92.8%-99.9%; Figure 37D), followed by the MIS-C *vs*. KD comparison (Figure 37B) with an AUC of 88.9% (95% CI: 78.5%-99.3%) and finally the MIS-C *vs*. viral comparison at 87.2% (95% CI: 79.5%-95%; Figure 37C). Sensitivity and specificity were calculated using Youden's index [152], a method described in chapter 2 section 2.4.4 that optimises both sensitivity and specificity. The sensitivity of the 5-gene RT-qPCR signature using the simple DRS was 86.1% and the specificity was 85.3%. Using the threshold of classification from Youden's index, disease classes were estimated (MIS-C or not MIS-C), with 86.1% (*n*=31) MIS-C patients correctly classified as MIS-C. 17.6% (*n*=3) of KD patients, 26% (*n*=11) of viral patients, and 4% (*n*=2) of bacterial patients were incorrectly classified as MIS-C. These results using 10-fold cross validation are encouraging since they suggest that the 5-gene signature presented here is likely to display good performance in diagnosing MIS-C in further independent validation cohorts.

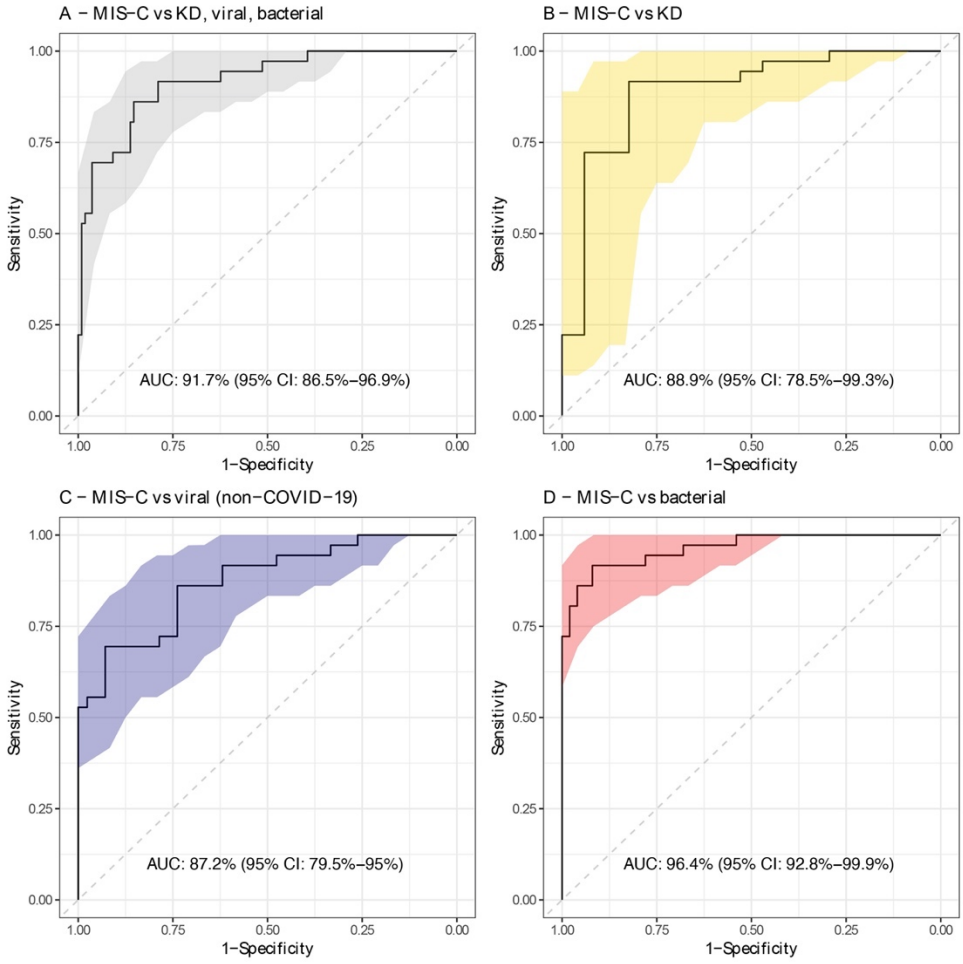


Figure 37: *ROC curves for the 5-gene RT-qPCR signature calculated using the MIS-C predictions returned from 10-fold cross validation. Probability of MIS-C was estimated for each patient in the left-out fold.* ***A****: performance for MIS-C vs. KD, viral and bacterial.* ***B-D****: performance for MIS-C vs. each comparator disease group separately. AUCs (95% confidence intervals) are printed on the plots. KD = Kawasaki disease; MIS-C = multisystem inflammatory syndrome in children.*

5.3.3 Biological relevance of the 5 genes in the 5-gene RT-qPCR signature

The genes included in the 5-gene RT-qPCR signature have biological functions that may provide insights into pathogenesis of MIS-C. *HSPBAP1* regulates expression of *TMPRSS2*, the gene encoding the enzyme transmembrane protease, serine 2, which is used by SARS-CoV-2 for S protein priming [225]. *TGFB1* encodes a ligand that, amongst other functions, modulates expression and activation of cytokines including IFN-γ and tumour necrosis factor alpha (*TNF-a*). Elevated expression of *TGFB1* in CD4+ T cells and higher numbers of TGFB1-expressing T and CD14-positive cells have been observed in a small set of patients with severe COVID-19 [226]. *VPS37C* is involved in endosomal sorting and has been shown to play a role in viral budding, specifically of human immunodeficiency virus, type 1 (HIV-1) by the HIV-1 Gag protein [227]. *MX2* encodes a GTPase and is part of the antiviral response induced by type I and III interferons. Elevated levels of *MX2* have been identified in individuals with

SARS-CoV-2 [228], and other viral infections [229]. Our finding that *TRBV11-2* is the most SDE gene is in keeping with previous reports, where expansion of peripheral T cells carrying TRBV11-2 has been correlated with disease severity and serum cytokine levels [230, 231]. Furthermore, it has been postulated that the selective expansion of specific T-cell V beta subsets is being driven by a super-antigen mediated process [99, 222, 230-232].

5.4 Summary

In this chapter, I have used whole blood RNA-Seq to compare the blood transcriptomes of patients with MIS-C to those with KD, viral and bacterial infections. Despite overlapping clinical and laboratory features between these conditions, there were clear differences in the blood transcriptome, with thousands of genes differentially expressed between MIS-C and each disease group. Using feature selection tools, I identified a minimal gene signature distinguishing MIS-C from the other disease groups, alongside 6 additional candidate biomarkers. Cross-platform validation of the genes was performed in an independent patient set using RT-qPCR, a methodology more suited for clinical use of the signature as a diagnostic test. Cross-platform validation showed that a combination of 5 genes can effectively distinguish between MIS-C and KD, viral and bacterial infections.

The translation of gene targets from high-throughput RNA-Seq to low-throughput RT-qPCR assays is challenging, with many genes failing to translate. Genes fail to translate for multiple reasons including inappropriate primer design (i.e., primers targeting a lowly expressed gene region), technical factors in the discovery dataset that cause spurious findings, and non-comparable clinical cohorts in the discovery and validation datasets. The 5 genes included in the 5-gene RT-qPCR signature have successfully translated across platforms, suggesting that it would be possible to include them in a rapid point-of-care diagnostic test. Such a diagnostic test would be extremely clinically useful due to the aforementioned challenges in diagnosing MIS-C. As the SARS-CoV-2 pandemic has progressed, multiple viral variants have emerged with slightly different clinical presentations. The risk of MIS-C occurrence appears to be lower with the Omicron variant, especially amongst vaccinated children, suggesting that MIS-C may become rarer over time [233]. Paradoxically, this may mean that an accurate test for MIS-C is more important, as if MIS-C is rarer, it will become more challenging to identify and will be more likely to be misdiagnosed as KD. Whilst the data presented here has originated from the United Kingdom, European Union and United States, a rapid point-of-care test for MIS-C may be the most useful in low- and middle-income countries globally, notably due to the wider range of prevalent infections that share features with MIS-C.

The work presented in this chapter is not without limitations. Most importantly, an additional independent validation cohort is required to further examine the performance of the 5-gene RT-qPCR signature since this combination of the 4-gene RNA-Seq signature + *TRBV11*-2 was derived from the

RT-qPCR dataset itself. The MIS-C patients included in the signature discovery and validation datasets were recruited during the period of April 2020-April 2021. As such, the SARS-CoV-2 variants that triggered the illness are limited to those that were circulating during this period and notably, would not include the Omicron variant that was first detected in November 2021. Validation of the 5-gene RT-qPCR signature in a cohort composed of MIS-C patients who were infected with a wider range of SARS-CoV-2 variants, in addition to patients who have received SARS-CoV-2 vaccinations is required.

Furthermore, as mentioned in chapter 4, a limitation of studies related to KD is that there is not a diagnostic test for KD, and the situation is the same for MIS-C. In the absence of a diagnostic test, it is impossible to say with certainty that the MIS-C and KD patients presented here are indeed MIS-C and KD, respectively. However, the KD patients included in the discovery set were recruited prior to the COVID-19 pandemic. For MIS-C, we made every effort to only include in both the discovery and validation patients meeting the WHO definition, including evidence of recent SARS-CoV-2 infection detected by antibody or PCR. Despite these limitations, our data provides evidence that MIS-C is characterised by a unique transcriptomic signature, and that a diagnostic test based on a small number of RNA transcripts is achievable.

In conclusion, despite their clinical similarities, considerable transcriptomic differences were observed between MIS-C and KD, bacterial and viral infections. MIS-C can be distinguished from KD, bacterial and viral infections using a sparse combination of gene biomarkers. The sparse combination of biomarkers identified here could be developed into a rapid point-of-care test that could revolutionise diagnosis and treatment of MIS-C as well as KD, bacterial and viral infections.

Chapter 6

6 Discussions and future directions

This thesis has explored using host omic profiles to better diagnose and understand paediatric infectious and inflammatory diseases. Chapter 1 has provided an overview of omic technologies and how they can be used to overcome some of the major challenges in the field of paediatric infectious and inflammatory diseases. Chapter 2 outlined the statistical and machine learning approaches that are used to process and analyse omic data, describing various powerful techniques for extracting important information. In chapters 3 to 5, three research projects are presented that explore three challenges in paediatric infectious and inflammatory diseases. Chapter 3 addresses the problem of diagnosing bacterial and viral infections in febrile children. In chapter 4, the unknown aetiology of Kawasaki disease is addressed. Chapter 5 addresses the challenge of differentiating between multisystem inflammatory syndrome in children (MIS-C) and KD, bacterial and viral infections. Whilst these chapters could be considered as stand-alone projects, they cover several closely related themes, and all shared a common aim: to improve diagnosis and understanding of paediatric infectious and inflammatory diseases.

This thesis has explored bacterial infections, viral infections, KD, and MIS-C. The key overarching theme of this thesis has been whether we can distinguish between these groups of patients using their omic profiles. In chapters 3 to 5, I have used various statistical and machine learning methods to show that, despite the clinical similarities between each of these disease groups, they are distinguishable at the host molecular level. I have shown that bacterial and viral infections can be diagnosed using a combination of six protein biomarkers (chapter 3), and that MIS-C can be distinguished from KD, bacterial and viral infections using just five host gene biomarkers (chapter 5). These are important findings that pave the way for the development of rapid point-of-care diagnostic tests that would ensure that more children receive the treatment that they require in a timely manner. In chapter 4, I performed extensive comparisons between the host transcriptomes and proteomes of children with KD and those from children with bacterial and viral infections. I show that there are similarities between KD and bacterial and viral infections observable on the host molecular level. Equally, however, a unique set of gene expression and protein abundance changes were brought about in response to KD, suggesting it is not a simple dichotomous question as to whether the response to KD is more similar to bacterial or viral infections. Not only have I contrasted the omic profiles between disease groups, but I also contrasted the host omic profiles within the KD patient groups. In doing so, I found that there were subgroups of patients that were clinically indistinguishable but displayed very different host transcriptomic and proteomic profiles. This has important implications for the care of KD patients going forward as previously it has been treated with a one-size-fits-all-approach, which may not be

suitable given these differences. It is possible that we are incorrectly assuming that KD is a single disease, and rather it may be a spectrum of clinically indistinguishable conditions.

Omic datasets are a rich source of information and feature selection tools are required to extract the most important information for the biological question at hand. In every chapter of this thesis, supervised feature selection methods have been to identify sparse combinations of genes or proteins with diagnostic purposes. In chapter 3, I used the feature selection algorithm forward selection-partial least squares (FS-PLS) to identify a sparse 6-protein signature that achieved AUCs between 89.4%-93.6% in an independent validation set. I used FS-PLS again in chapter 5 to identify a 5-gene diagnostic gene signature that could accurately distinguish MIS-C from KD and bacterial and viral infections with AUCs between 91.7%-93.2% in an independent validation set. In both instances, FS-PLS identified host diagnostic signatures that not only achieved excellent discriminatory performance, but also successfully translated across platforms to low-throughput quantification methods that are closer to the type of quantification approaches that would be used in a rapid point-of-care diagnostic test. In chapter 4, I used LASSO regression to determine whether KD patients would be classified as either bacterial or viral. By using a two-way classifier approach, I showed that the majority of KD patients were diagnosed as neither bacterial nor viral based on their host transcriptomic profiles, raising the possibility that KD is caused by neither bacterial nor viral triggers.

When using supervised feature selection algorithms such as FS-PLS and LASSO to identify diagnostic signatures, the accuracy of the initial phenotyping of the patients is crucial. As captured in the phrase "garbage in, garbage out", if the patients' phenotypes are unreliable, the diagnostic signature identified is unlikely to be accurate. In the case of the bacterial *vs*. viral protein signature presented in chapter 3, the definite bacterial (DB) and definite viral (DV) patients included in the cohorts used to identify the signature were well phenotyped patients in whom bacterial/viral diagnoses were confirmed by gold-standard, culture-based methods and following a retrospective, extensive examination of all of the available clinical data. This explains the signature's high performance in independent cohorts who were phenotyped using the same algorithm (Figure 8) and is a major strength of the study. Paradoxically, however, the groups of patients in whom a diagnosis is not 100% certain are the ones who may benefit the most from a host molecular diagnostic test. For example, the individuals in "probable" bacterial or viral groups as shown in chapter 3 are those for whom there was insufficient evidence to say whether they had a bacterial or infection, even after extensive clinical review. A rapid diagnostic test would be extremely useful for these individuals as it would act as another component of the diagnostic toolkit, potentially providing the evidence that a clinician required to administer or withhold antibiotics. However, without including these groups of patients in the discovery cohorts used to identify the diagnostic signatures, we are unlikely to obtain signatures that are optimised to classify these individuals. A future direction of the work presented in chapter 3 would be to determine whether any additional proteins could be added to the signature that can distinguish between bacterial

and viral infections when the unconfirmed phenotypic groups are included. This would lead to a more heterogeneous comparison of bacterial infections (suspected and confirmed) *vs.* viral infections (suspected and confirmed) and could produce a diagnostic test that could lead to more certain diagnoses.

The diagnostic signatures presented in chapters 3 and 5 are binary classifiers (i.e., bacterial *vs.* viral, and MIS-C *vs.* not MIS-C, with "not MIS-C" consisting of KD, bacterial and viral). In a post-COVID-19 era, the spectrum of paediatric infectious and inflammatory diseases has become even more complex. Diagnostic decisions are unlikely to always be binary and performing multiple dichotomous diagnostic tests could be impractical. For example, if the 5-gene MIS-C signature presented in chapter 5 was developed into a rapid point-of-care test and returned a negative test result (i.e., not MIS-C), a further diagnostic test would be still required to determine whether the individual had KD, a bacterial infection, a viral infection or something completely different. An obvious extension of the work presented in chapters 3 and 5 would be to identify multi-class diagnostic signatures. A multi-class diagnostic signature would be a combination of host molecules that work together to diagnose more than two distinct diseases. A multi-class signature that could diagnose MIS-C, KD, bacterial infections, and viral infections in parallel would be particularly useful, or this could be expanded to include other diseases such as malaria or tuberculosis.

In chapter 4, I have presented the parallel analysis of host transcriptomes and proteomes from children with KD. There was no overlap between the individuals included in either of the omic levels, meaning that the omic datasets were treated in isolation. Multi-omic approaches are rapidly growing in popularity, with various powerful software tools available for analysing omic datasets obtained from the same set of individuals [234-236]. It would be intriguing to repeat this analysis using transcriptomic and proteomic datasets obtained from the same set of individuals. Performing clustering across both omic levels simultaneously would reveal whether the patient subgroups are consistent irrespective of data type. The KD patients included in the proteomic dataset in chapter 4 were very similar to the bacterial patients. If transcriptomic profiles were also available for these individuals, we could unpick whether this similarity is conserved on the gene expression level, or whether this similarity is driven by the data type.

6.1 Conclusion

In this thesis, I have used host transcriptomic and proteomic datasets obtained from children with various paediatric infectious and inflammatory diseases, including bacterial infections, viral infections, KD, and MIS-C. Using statistical and machine learning approaches, I have contrasted omic profiles both between (chapters 3-5) and within disease groups (chapter 4). I have shown that despite the clinical similarities between the paediatric infectious and inflammatory diseases studied, each of these diseases

leads to distinct host gene expression and/or protein abundance changes. I have shown that bacterial and viral infections in febrile children can be differentiated from each other using a 6-protein signature that achieves AUCs of up to 93.6% in an independent cross-platform validation dataset. Similarly, MIS-C can be distinguished from KD, bacterial and viral infections using a 5-gene signature that achieves AUCs of up to 93.2% in an independent cross-platform validation dataset. Both of these signatures could be developed into rapid point-of-care diagnostic tests, revolutionising the diagnosis of infectious and inflammatory diseases in children. Parallel analysis of host proteomic and transcriptomic profiles from children with KD has shown that the host response during KD is highly heterogeneous, with some elements of the host response appearing more "viral-like", some elements appearing more "bacterial-like", and some elements sharing no similarities with the host response to either bacterial or viral infections. These findings raise the possibility that KD may be a spectrum of diseases rather than a single condition. In summary, the combination of high-dimensional omic datasets and statistical and machine learning tools can provide extensive insights into paediatric infectious and inflammatory diseases, including providing alternative diagnostic approaches and biologically important insights.

Bibliography

1. Chaplin, D.D., *Overview of the immune response.* J Allergy Clin Immunol, 2010. **125**(2 Suppl 2): p. S3-23.
2. Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P., *Chapter 24, The Adaptive Immune System*, in *Molecular Biology of the Cell. 4th edition*. 2002, Garland Science: New York.
3. Finlay, B.B. and G. McFadden, *Anti-immunology: evasion of the host immune system by bacterial and viral pathogens.* Cell, 2006. **124**(4): p. 767-82.
4. Karlsson, E.K., D.P. Kwiatkowski, and P.C. Sabeti, *Natural selection and infectious disease in human populations.* Nat Rev Genet, 2014. **15**(6): p. 379-93.
5. Kwiatkowski, D.P., *How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria.* The American Journal of Human Genetics, 2005. **77**(2): p. 171-192.
6. Kaufmann, S.H.E., H.L. Collins, and U.E. Schaible, *25 - Immune responses to intracellular bacteria*, in *Clinical Immunology (Third Edition)*, R.R. Rich, et al., Editors. 2008, Mosby: Edinburgh. p. 389-409.
7. Horgan, R.P. and L.C. Kenny, *'Omic' technologies: genomics, transcriptomics, proteomics and metabolomics.* The Obstetrician & Gynaecologist, 2011. **13**(3): p. 189-195.
8. Hasin, Y., M. Seldin, and A. Lusis, *Multi-omics approaches to disease.* Genome Biol, 2017. **18**(1): p. 83.
9. Kitano, H., *Systems biology: toward system-level understanding of biological systems.* Foundations of systems biology, 2001: p. 1-36.
10. Palazzo, A.F. and E.S. Lee, *Non-coding RNA: what is functional and what is junk?* Frontiers in Genetics, 2015. **6**.
11. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
12. He, S.L. and R. Green, *Northern blotting.* Methods in enzymology, 2013. **530**: p. 75-87.
13. Garibyan, L. and N. Avashia, *Polymerase chain reaction.* The Journal of investigative dermatology, 2013. **133**(3): p. 1-4.
14. Dymond, J.S., *Chapter Twenty Three - Explanatory Chapter: Quantitative PCR*, in *Methods in Enzymology*, J. Lorsch, Editor. 2013, Academic Press. p. 279-289.
15. VanGuilder, H.D., K.E. Vrana, and W.M. Freeman, *Twenty-five years of quantitative PCR for gene expression analysis.* Biotechniques, 2008. **44**(5): p. 619-26.
16. Wong, M.L. and J.F. Medrano, *Real-time PCR for mRNA quantitation.* BioTechniques, 2005. **39**(1): p. 75-85.
17. Battaglia, M., et al., *Epithelial tumour cell detection and the unsolved problems of nested RT-PCR: a new sensitive one step method without false positive results.* Bone Marrow Transplantation, 1998. **22**(7): p. 693-698.
18. Lamas, A., Franco, C. M. , Regal, P. , Miranda, J. , Vázquez, B. , Cepeda, A, *High-Throughput Platforms in Real-Time PCR and Applications*. Polymerase Chain Reaction for Biomedical Applications. 2016: IntechOpen.
19. Geiss, G.K., et al., *Direct multiplexed measurement of gene expression with color-coded probe pairs.* Nat Biotechnol, 2008. **26**(3): p. 317-25.
20. Trevino, V., F. Falciani, and H.A. Barrera-Saldana, *DNA microarrays: a powerful genomic tool for biomedical and clinical research.* Mol Med, 2007. **13**(9-10): p. 527-41.
21. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
22. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.
23. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
24. Bumgarner, R., *Overview of DNA microarrays: types, applications, and their future.* Current protocols in molecular biology, 2013. **Chapter 22**: p. Unit-22.1.
25. Conzone, S.D. and C.G. Pantano, *Glass slides to DNA microarrays.* Materials Today, 2004. **7**(3): p. 20-26.
26. Pandey, S.C., et al., *Chapter 5 - DNA microarray analysis of Leishmania parasite: strengths and limitations*, in *Pathogenesis, Treatment and Prevention of Leishmaniasis*, M. Samant and S. Chandra Pandey, Editors. 2021, Academic Press. p. 85-101.
27. Jordan, B., *Historical background and anticipated developments.* Ann N Y Acad Sci, 2002. **975**: p. 24-32.
28. Oberthuer, A., et al., *Comparison of performance of one-color and two-color gene-expression analyses in predicting clinical endpoints of neuroblastoma patients.* Pharmacogenomics J, 2010. **10**(4): p. 258-66.
29. Barba, M., H. Czosnek, and A. Hadidi, *Historical perspective, development and applications of next-generation sequencing in plant virology.* Viruses, 2014. **6**(1): p. 106-136.
30. Behjati, S. and P.S. Tarpey, *What is next generation sequencing?* Archives of disease in childhood. Education and practice edition, 2013. **98**(6): p. 236-238.
31. Kukurba, K.R. and S.B. Montgomery, *RNA Sequencing and Analysis.* Cold Spring Harbor protocols, 2015. **2015**(11): p. 951-969.
32. Stark, R., M. Grzelak, and J. Hadfield, *RNA sequencing: the teenage years.* Nat Rev Genet, 2019. **20**(11): p. 631-656.

33. Cummings, B.B., et al., *Improving genetic diagnosis in Mendelian disease with transcriptome sequencing.* Sci Transl Med, 2017. **9**(386).
34. Aslam, B., et al., *Proteomics: Technologies and Their Applications.* Journal of Chromatographic Science, 2017. **55**(2): p. 182-196.
35. Kurien, B.T. and R.H. Scofield, *Western blotting.* Methods, 2006. **38**(4): p. 283-93.
36. Sakamoto, S., et al., *Enzyme-linked immunosorbent assay for the quantitative/qualitative analysis of plant secondary metabolites.* J Nat Med, 2018. **72**(1): p. 32-42.
37. Partridge, M.A., et al., *High-throughput antibody-based assays to identify and quantify radiation-responsive protein biomarkers.* Int J Radiat Biol, 2010. **86**(4): p. 321-8.
38. Ren, A.H., E.P. Diamandis, and V. Kulasingam, *Uncovering the Depths of the Human Proteome: Antibody-based Technologies for Ultrasensitive Multiplexed Protein Detection and Quantification.* Mol Cell Proteomics, 2021. **20**: p. 100155.
39. Singer, B.S., et al., *Libraries for genomic SELEX.* Nucleic Acids Res, 1997. **25**(4): p. 781-6.
40. Gold, L., et al., *Aptamer-based multiplexed proteomic technology for biomarker discovery.* PLoS One, 2010. **5**(12): p. e15004.
41. Han, X., A. Aslanian, and J.R. Yates, 3rd, *Mass spectrometry for proteomics.* Curr Opin Chem Biol, 2008. **12**(5): p. 483-90.
42. Karpievitch, Y.V., et al., *Liquid Chromatography Mass Spectrometry-Based Proteomics: Biological and Technological Aspects.* Ann Appl Stat, 2010. **4**(4): p. 1797-1823.
43. Fenn, J.B., et al., *Electrospray ionization for mass spectrometry of large biomolecules.* Science, 1989. **246**(4926): p. 64-71.
44. Karas, M. and F. Hillenkamp, *Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons.* Anal Chem, 1988. **60**(20): p. 2299-301.
45. Aebersold, R. and M. Mann, *Mass-spectrometric exploration of proteome structure and function.* Nature, 2016. **537**(7620): p. 347-55.
46. Sinha, A. and M. Mann, *A beginner's guide to mass spectrometry–based proteomics.* The Biochemist, 2020. **42**(5): p. 64-69.
47. Cox, J. and M. Mann, *MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification.* Nat Biotechnol, 2008. **26**(12): p. 1367-72.
48. Aschenbrenner, A.C., et al., *Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients.* Genome Med, 2021. **13**(1): p. 7.
49. Saria, S. and A. Goldenberg, *Subtyping: What It is and Its Role in Precision Medicine.* IEEE Intelligent Systems, 2015. **30**(4): p. 70-75.
50. Neogi, U., et al., *Multi-omics insights into host-viral response and pathogenesis in Crimean-Congo hemorrhagic fever viruses for novel therapeutic target.* Elife, 2022. **11**.
51. Wang, H., et al., *A 3-Gene-Based Diagnostic Signature in Alzheimer's Disease.* Eur Neurol, 2022. **85**(1): p. 6-13.
52. Qian, Y., et al., *Prognostic Cancer Gene Expression Signatures: Current Status and Challenges.* Cells, 2021. **10**(3).
53. Girodias, J.-B. and B. Bailey, *Approach to the febrile child: A challenge bridging the gap between the literature and clinical practice.* Paediatrics & child health, 2003. **8**(2): p. 76-82.
54. Boeddha, N.P., et al., *Mortality and morbidity in community-acquired sepsis in European pediatric intensive care units: a prospective cohort study from the European Childhood Life-threatening Infectious Disease Study (EUCLIDS).* Critical Care, 2018. **22**(1): p. 143.
55. Nijman, R.G., et al., *Management of Children With Fever at Risk for Pediatric Sepsis: A Prospective Study in Pediatric Emergency Care.* Frontiers in Pediatrics, 2020. **8**.
56. Isaacman, D.J., et al., *Effect of number of blood cultures and volume of blood on detection of bacteremia in children.* J Pediatr, 1996. **128**(2): p. 190-5.
57. Zadroga, R., et al., *Comparison of 2 blood culture media shows significant differences in bacterial recovery for patients on antimicrobial therapy.* Clin Infect Dis, 2013. **56**(6): p. 790-7.
58. Rosenblum, J., et al., *Repeating blood cultures in neutropenic children with persistent fevers when the initial blood culture is negative.* Pediatric blood & cancer, 2013. **60**(6): p. 923-927.
59. Iroh Tam, P.Y., S.K. Obaro, and G. Storch, *Challenges in the Etiology and Diagnosis of Acute Febrile Illness in Children in Low- and Middle-Income Countries.* J Pediatric Infect Dis Soc, 2016. **5**(2): p. 190-205.
60. Hagedoorn, N.N., et al., *Development and validation of a prediction model for invasive bacterial infections in febrile children at European Emergency Departments: MOFICHE, a prospective observational study.* Archives of Disease in Childhood, 2021. **106**(7): p. 641.
61. Gautam, S., et al., *Severe respiratory viral infection induces procalcitonin in the absence of bacterial pneumonia.* Thorax, 2020. **75**(11): p. 974-981.
62. Houghton, R., et al., *C-reactive protein-guided use of procalcitonin in COVID-19.* JAC Antimicrob Resist, 2021. **3**(4): p. dlab180.

63. Heiskanen-Kosma, T. and M. Korppi, *Serum C-reactive protein cannot differentiate bacterial and viral aetiology of community-acquired pneumonia in children in primary healthcare settings.* Scand J Infect Dis, 2000. **32**(4): p. 399-402.
64. Liu, F., et al., *Prognostic value of interleukin-6, C-reactive protein, and procalcitonin in patients with COVID-19.* J Clin Virol, 2020. **127**: p. 104370.
65. Zhu, J., et al., *Clinical characteristics of 3062 COVID-19 patients: A meta-analysis.* J Med Virol, 2020. **92**(10): p. 1902-1914.
66. Hagedoorn, N.N., et al., *Variation in antibiotic prescription rates in febrile children presenting to emergency departments across Europe (MOFICHE): A multicentre observational study.* PLOS Medicine, 2020. **17**(8): p. e1003208.
67. Herberg, J.A., et al., *Diagnostic Test Accuracy of a 2-Transcript Host RNA Signature for Discriminating Bacterial vs Viral Infection in Febrile Children.* JAMA, 2016. **316**(8): p. 835-45.
68. Ramilo, O., et al., *Gene expression patterns in blood leukocytes discriminate patients with acute infections.* Blood, 2007. **109**(5): p. 2066-77.
69. Hu, X., et al., *Gene expression profiles in febrile children with defined viral and bacterial infection.* Proc Natl Acad Sci U S A, 2013. **110**(31): p. 12792-7.
70. Mahajan, P., et al., *Association of RNA Biosignatures With Bacterial Infections in Febrile Infants Aged 60 Days or Younger.* JAMA, 2016. **316**(8): p. 846-57.
71. Leticia Fernandez-Carballo, B., et al., *Distinguishing bacterial versus non-bacterial causes of febrile illness - A systematic review of host biomarkers.* J Infect, 2021. **82**(4): p. 1-10.
72. Oved, K., et al., *A novel host-proteome signature for distinguishing between acute bacterial and viral infections.* PLoS One, 2015. **10**(3): p. e0120012.
73. Papan, C., et al., *A host signature based on TRAIL, IP-10, and CRP for reducing antibiotic overuse in children by differentiating bacterial from viral infections: a prospective, multicentre cohort study.* Clin Microbiol Infect, 2021.
74. Paz, M., et al., *1213. A TRAIL/IP-10/CRP Signature Distinguishes between Viral and Bacterial Infection in Chronic Obstructive Pulmonary Disease Patients.* Open Forum Infectious Diseases, 2020. **7**(Supplement_1): p. S628-S628.
75. Patel, R.M. and S.T. Shulman, *Kawasaki disease: a comprehensive review of treatment options.* J Clin Pharm Ther, 2015. **40**(6): p. 620-5.
76. Kawasaki, T., *[Acute febrile mucocutaneous syndrome with lymphoid involvement with specific desquamation of the fingers and toes in children].* Arerugi, 1967. **16**(3): p. 178-222.
77. Singh, S., P. Vignesh, and D. Burgner, *The epidemiology of Kawasaki disease: a global update.* Arch Dis Child, 2015. **100**(11): p. 1084-8.
78. Menikou, S., P.R. Langford, and M. Levin, *Kawasaki Disease: The Role of Immune Complexes Revisited.* Front Immunol, 2019. **10**: p. 1156.
79. Nagata, S., *Causes of Kawasaki Disease-From Past to Present.* Front Pediatr, 2019. **7**: p. 18.
80. Burns, J.C., et al., *Seasonality of Kawasaki disease: a global perspective.* PLoS One, 2013. **8**(9): p. e74529.
81. Dietz, S.M., et al., *Dissecting Kawasaki disease: a state-of-the-art review.* Eur J Pediatr, 2017. **176**(8): p. 995-1009.
82. Suenaga, T., et al., *Detection of multiple superantigen genes in stools of patients with Kawasaki disease.* J Pediatr, 2009. **155**(2): p. 266-70.
83. Jackson, H., et al., *Kawasaki Disease Patient Stratification and Pathway Analysis Based on Host Transcriptomic and Proteomic Profiles.* Int J Mol Sci, 2021. **22**(11).
84. McCrindle, B.W., et al., *Diagnosis, Treatment, and Long-Term Management of Kawasaki Disease: A Scientific Statement for Health Professionals From the American Heart Association.* Circulation, 2017. **135**(17): p. e927-e999.
85. Wright, V.J., et al., *Diagnosis of Kawasaki Disease Using a Minimal Whole-Blood Gene Expression Signature.* JAMA Pediatrics, 2018. **172**(10): p. e182293-e182293.
86. Feldstein, L.R., et al., *Multisystem Inflammatory Syndrome in U.S. Children and Adolescents.* N Engl J Med, 2020. **383**(4): p. 334-346.
87. Whittaker, E., et al., *Clinical Characteristics of 58 Children With a Pediatric Inflammatory Multisystem Syndrome Temporally Associated With SARS-CoV-2.* JAMA, 2020. **324**(3): p. 259-269.
88. Dufort, E.M., et al., *Multisystem Inflammatory Syndrome in Children in New York State.* N Engl J Med, 2020. **383**(4): p. 347-358.
89. Santos, M.O., et al., *Multisystem inflammatory syndrome (MIS-C): a systematic review and meta-analysis of clinical characteristics, treatment, and outcomes.* J Pediatr (Rio J), 2021.
90. Carter, M.J., et al., *Peripheral immunophenotypes in children with multisystem inflammatory syndrome associated with SARS-CoV-2 infection.* Nat Med, 2020. **26**(11): p. 1701-1707.
91. Syrimi, E., et al., *The immune landscape of SARS-CoV-2-associated Multisystem Inflammatory Syndrome in Children (MIS-C) from acute disease to recovery.* iScience, 2021. **24**(11): p. 103215.
92. Sakurai, Y., *Autoimmune Aspects of Kawasaki Disease.* J Investig Allergol Clin Immunol, 2019. **29**(4): p. 251-261.

93. Consiglio, C.R., et al., *The Immunology of Multisystem Inflammatory Syndrome in Children with COVID-19.* Cell, 2020. **183**(4): p. 968-981 e7.
94. Gruber, C.N., et al., *Mapping Systemic Inflammation and Antibody Responses in Multisystem Inflammatory Syndrome in Children (MIS-C).* Cell, 2020. **183**(4): p. 982-995 e14.
95. Cheung, E.W., et al., *Multisystem Inflammatory Syndrome Related to COVID-19 in Previously Healthy Children and Adolescents in New York City.* JAMA, 2020. **324**(3): p. 294-296.
96. Organisation, W.H., *Multisystem inflammatory syndrome in children and adolescents with COVID-19.* 2020.
97. Beckmann, N.D., et al., *Downregulation of exhausted cytotoxic T cells in gene expression networks of multisystem inflammatory syndrome in children.* Nat Commun, 2021. **12**(1): p. 4854.
98. Diorio, C., et al., *Proteomic profiling of MIS-C patients indicates heterogeneity relating to interferon gamma dysregulation and vascular endothelial dysfunction.* Nat Commun, 2021. **12**(1): p. 7222.
99. Sacco, K., et al., *Immunopathological signatures in multisystem inflammatory syndrome in children and pediatric COVID-19.* Nat Med, 2022.
100. Gliddon, H.D., et al., *Identification of Reduced Host Transcriptomic Signatures for Tuberculosis Disease and Digital PCR-Based Validation and Quantification.* Frontiers in Immunology, 2021. **12**.
101. Kaforou, M., et al., *Diagnosis of Bacterial Infection Using a 2-Transcript Host RNA Signature in Febrile Infants 60 Days or Younger.* JAMA, 2017. **317**(15): p. 1577-1578.
102. Kaforou, M., et al., *Detection of Tuberculosis in HIV-Infected and -Uninfected African Adults Using Whole Blood RNA Expression Signatures: A Case-Control Study.* PLOS Medicine, 2013. **10**(10): p. e1001538.
103. Zandstra, J., et al., *Biomarkers for the Discrimination of Acute Kawasaki Disease From Infections in Childhood.* Front Pediatr, 2020. **8**: p. 355.
104. Herberg, J.A., et al., *Transcriptomic profiling in childhood H1N1/09 influenza reveals reduced expression of protein synthesis genes.* J Infect Dis, 2013. **208**(10): p. 1664-8.
105. Sweeney, T.E., et al., *Genome-wide expression for diagnosis of pulmonary tuberculosis: a multicohort analysis.* Lancet Respir Med, 2016. **4**(3): p. 213-24.
106. Team, R.C., *R: A Language and Environment for Statistical Computing*. 2022.
107. Du, P., W.A. Kibbe, and S.M. Lin, *lumi: a pipeline for processing Illumina microarray.* Bioinformatics, 2008. **24**(13): p. 1547-1548.
108. Cotton, R.J. and J. Graumann, *readat: An R package for reading and working with SomaLogic ADAT files.* BMC Bioinformatics, 2016. **17**(1): p. 201.
109. Wu, D.C., et al., *Limitations of alignment-free tools in total RNA-seq quantification.* BMC Genomics, 2018. **19**(1): p. 510.
110. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome Biol, 2016. **17**: p. 13.
111. Cunningham, F., et al., *Ensembl 2022.* Nucleic Acids Res, 2022. **50**(D1): p. D988-D995.
112. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
113. Alkallas, R., et al., *Inference of RNA decay rate from transcriptional profiling highlights the regulatory programs of Alzheimer's disease.* Nat Commun, 2017. **8**(1): p. 909.
114. Gaidatzis, D., et al., *Analysis of intronic and exonic reads in RNA-seq data characterizes transcriptional and post-transcriptional regulation.* Nat Biotechnol, 2015. **33**(7): p. 722-9.
115. Baruzzo, G., et al., *Simulation-based comprehensive benchmarking of RNA-seq aligners.* Nat Methods, 2017. **14**(2): p. 135-139.
116. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-30.
117. Andrews, S., *Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data.* 2022. **5**: p. 47-81.
118. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report.* Bioinformatics, 2016. **32**(19): p. 3047-8.
119. Tyanova, S., T. Temu, and J. Cox, *The MaxQuant computational platform for mass spectrometry-based shotgun proteomics.* Nat Protoc, 2016. **11**(12): p. 2301-2319.
120. Cox, J., et al., *Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment.* Journal of Proteome Research, 2011. **10**(4): p. 1794-1805.
121. Cox, J., et al., *Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ.* Mol Cell Proteomics, 2014. **13**(9): p. 2513-26.
122. O'Rourke, M.B., et al., *What is Normalization? The Strategies Employed in Top-Down and Bottom-Up Proteome Analysis Workflows.* Proteomes, 2019. **7**(3).
123. Jolliffe, I.T. and J. Cadima, *Principal component analysis: a review and recent developments.* Philos Trans A Math Phys Eng Sci, 2016. **374**(2065): p. 20150202.
124. Law, C.W., et al., *RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR.* F1000Res, 2016. **5**.
125. Liu, X., et al., *Normalization Methods for the Analysis of Unbalanced Transcriptome Data: A Review.* Front Bioeng Biotechnol, 2019. **7**: p. 358.
126. Leek, J.T., et al., *Tackling the widespread and critical impact of batch effects in high-throughput data.* Nat Rev Genet, 2010. **11**(10): p. 733-9.

127. Zhang, Y., G. Parmigiani, and W.E. Johnson, *ComBat-seq: batch effect adjustment for RNA-seq count data.* NAR Genomics and Bioinformatics, 2020. **2**(3).
128. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-127.
129. Leek, J.T., et al., *The sva package for removing batch effects and other unwanted variation in high-throughput experiments.* Bioinformatics, 2012. **28**(6): p. 882-883.
130. Goh, W.W.B. and L. Wong, *Protein complex-based analysis is resistant to the obfuscating consequences of batch effects --- a case study in clinical proteomics.* BMC Genomics, 2017. **18**(2): p. 142.
131. Liu, Q., et al., *Addressing the batch effect issue for LC/MS metabolomics data in data preprocessing.* Scientific Reports, 2020. **10**(1): p. 13856.
132. Sweeney, T.E., H.R. Wong, and P. Khatri, *Robust classification of bacterial and viral infections via integrated host gene expression diagnostics.* Science translational medicine, 2016. **8**(346): p. 346ra91-346ra91.
133. Silver, J.D., M.E. Ritchie, and G.K. Smyth, *Microarray background correction: maximum likelihood estimation for the normal-exponential convolution.* Biostatistics, 2009. **10**(2): p. 352-63.
134. Schmid, R., et al., *Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3.* BMC Genomics, 2010. **11**: p. 349.
135. Bolstad, B.M., et al., *A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.* Bioinformatics, 2003. **19**(2): p. 185-93.
136. Cleveland, W.S., *Robust Locally Weighted Regression and Smoothing Scatterplots.* Journal of the American Statistical Association, 1979. **74**(368): p. 829-836.
137. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 2014. **15**(12): p. 550.
138. Rousseeuw, P.J., *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis.* Journal of Computational and Applied Mathematics, 1987. **20**: p. 53-65.
139. Charrad, M., et al., *NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set.* Journal of Statistical Software, 2014. **61**(6): p. 1 - 36.
140. Soneson, C. and M. Delorenzi, *A comparison of methods for differential expression analysis of RNA-seq data.* BMC Bioinformatics, 2013. **14**: p. 91.
141. Ritchie, M.E., et al., *limma powers differential expression analyses for RNA-sequencing and microarray studies.* Nucleic Acids Res, 2015. **43**(7): p. e47.
142. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing*. 1995, WileyRoyal Statistical Society. p. 289-300.
143. Xu, C., et al., *Exploring the potential benefits of stratified false discovery rates for region-based testing of association with rare genetic variation.* Front Genet, 2014. **5**: p. 11.
144. Smyth, G.K., *limma: Linear Models for Microarray Data*, in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, R. Gentleman, et al., Editors. 2005, Springer New York: New York, NY. p. 397-420.
145. Kammers, K., et al., *Detecting Significant Changes in Protein Abundance.* EuPA Open Proteom, 2015. **7**: p. 11-19.
146. van Ooijen, M.P., et al., *Identification of differentially expressed peptides in high-throughput proteomics data.* Brief Bioinform, 2018. **19**(5): p. 971-981.
147. Agioutantis, P.C., H. Loutrari, and F.N. Kolisis, *Computational Analysis of Transcriptomic and Proteomic Data for Deciphering Molecular Heterogeneity and Drug Responsiveness in Model Human Hepatocellular Carcinoma Cell Lines.* Genes (Basel), 2020. **11**(6).
148. Seyednasrollah, F., A. Laiho, and L.L. Elo, *Comparison of software packages for detecting differential expression in RNA-seq studies.* Brief Bioinform, 2015. **16**(1): p. 59-70.
149. Phipson, B., et al. *Empirical Bayes in the presence of exceptional cases, with application to microarray data*. 2013.
150. Gliddon, H.D., et al., *Genome-wide host RNA signatures of infectious diseases: discovery and clinical translation.* Immunology, 2018. **153**(2): p. 171-178.
151. Friedman, J.H., T. Hastie, and R. Tibshirani, *Regularization Paths for Generalized Linear Models via Coordinate Descent.* Journal of Statistical Software, 2010. **33**(1): p. 1 - 22.
152. Youden, W.J., *Index for rating diagnostic tests.* Cancer, 1950. **3**(1): p. 32-5.
153. Garcia-Campos, M.A., J. Espinal-Enriquez, and E. Hernandez-Lemus, *Pathway Analysis: State of the Art.* Front Physiol, 2015. **6**: p. 383.
154. Reimand, J., et al., *Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap.* Nat Protoc, 2019. **14**(2): p. 482-517.
155. Kolberg, L., et al., *gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler.* F1000Res, 2020. **9**.
156. Supek, F., et al., *REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms.* PLoS ONE, 2011. **6**(7): p. e21800-e21800.
157. Zandstra, J., I. Jongerius, and T.W. Kuijpers, *Future Biomarkers for Infection and Inflammation in Febrile Children.* Front Immunol, 2021. **12**: p. 631308.

158. Martinon-Torres, F., et al., *Life-threatening infections in children in Europe (the EUCLIDS Project): a prospective cohort study.* Lancet Child Adolesc Health, 2018. **2**(6): p. 404-414.
159. Hagedoorn, N.N., et al., *Shock Index in the early assessment of febrile children at the emergency department: a prospective multicentre study.* Arch Dis Child, 2021.
160. Nijman, R.G., et al., *A Novel Framework for Phenotyping Children With Suspected or Confirmed Infection for Future Biomarker Studies.* Front Pediatr, 2021. **9**: p. 688272.
161. Sweeney, T., *COCONUT: COmbat CO-Normalization Using conTrols (COCONUT). R package version 1.0.2.* 2017.
162. Giuliano, J.S., Jr., et al., *Admission angiopoietin levels in children with septic shock.* Shock, 2007. **28**(6): p. 650-654.
163. Giuliano, J.S., Jr. and D.S. Wheeler, *Excess circulating angiopoietin-2 levels in sepsis: harbinger of death in the intensive care unit?* Crit Care, 2009. **13**(1): p. 114.
164. Fabriek, B.O., et al., *The macrophage scavenger receptor CD163 functions as an innate immune sensor for bacteria.* Blood, 2009. **113**(4): p. 887-92.
165. Zhou, Y., *Regulatory T cells and viral infections.* Front Biosci, 2008. **13**: p. 1152-70.
166. Sproston, N.R. and J.J. Ashworth, *Role of C-Reactive Protein at Sites of Inflammation and Infection.* Front Immunol, 2018. **9**: p. 754.
167. García-Álvarez, M., et al., *High plasma fractalkine (CX3CL1) levels are associated with severe liver disease in HIV/HCV co-infected patients with HCV genotype 1.* Cytokine, 2011. **54**(3): p. 244-8.
168. Bolovan-Fritts, C.A., R.N. Trout, and S.A. Spector, *Human cytomegalovirus-specific CD4+-T-cell cytokine response induces fractalkine in endothelial cells.* J Virol, 2004. **78**(23): p. 13173-81.
169. Fischer, J.E., et al., *Diagnostic accuracy of G-CSF, IL-8, and IL-1ra in critically ill children with suspected infection.* Intensive Care Med, 2002. **28**(9): p. 1324-31.
170. Korkmaz, B., T. Moreau, and F. Gauthier, *Neutrophil elastase, proteinase 3 and cathepsin G: physicochemical properties, activity and physiopathological functions.* Biochimie, 2008. **90**(2): p. 227-42.
171. Russell, C.D., et al., *The Human Immune Response to Respiratory Syncytial Virus Infection.* Clin Microbiol Rev, 2017. **30**(2): p. 481-502.
172. Sahoo, M., et al., *Role of the inflammasome, IL-1β, and IL-18 in bacterial infections.* TheScientificWorldJournal, 2011. **11**: p. 2037-2050.
173. Srugo, I., et al., *Validation of a Novel Assay to Distinguish Bacterial and Viral Infections.* Pediatrics, 2017. **140**(4).
174. Ciragil, P., E.B. Kurutas, and M. Miraloglu, *New markers: urine xanthine oxidase and myeloperoxidase in the early detection of urinary tract infection.* Dis Markers, 2014. **2014**: p. 269362.
175. Austermann, J., C. Spiekermann, and J. Roth, *S100 proteins in rheumatic diseases.* Nature Reviews Rheumatology, 2018. **14**(9): p. 528-541.
176. Fares-Gusmao, R., et al., *Differential Pattern of Soluble Immune Markers in Asymptomatic Dengue, West Nile and Zika Virus Infections.* Scientific Reports, 2019. **9**(1): p. 17172.
177. Robin, X., et al., *pROC: an open-source package for R and S+ to analyze and compare ROC curves.* BMC Bioinformatics, 2011. **12**: p. 77.
178. Appenzeller, C., et al., *Serum C-reactive protein in children with adenovirus infection.* Swiss Med Wkly, 2002. **132**(25-26): p. 345-50.
179. Shahbaznejad, L., et al., *Clinical characteristics of 10 children with a pediatric inflammatory multisystem syndrome associated with COVID-19 in Iran.* BMC Pediatr, 2020. **20**(1): p. 513.
180. Mantovani, A. and E. Dejana, *Endothelium*, in *Encyclopedia of Immunology (Second Edition)*, P.J. Delves, Editor. 1998, Elsevier: Oxford. p. 802-806.
181. Nasioudis, D. and S.S. Witkin, *Neutrophil gelatinase-associated lipocalin and innate immune responses to bacterial infections.* Med Microbiol Immunol, 2015. **204**(4): p. 471-9.
182. Czarniecki, C.W. and G. Sonnenfeld, *Interferon-gamma and resistance to bacterial infections.* Apmis, 1993. **101**(1): p. 1-17.
183. Van Der Sluijs, K.F., et al., *Enhanced viral clearance in interleukin-18 gene-deficient mice after pulmonary infection with influenza A virus.* Immunology, 2005. **114**(1): p. 112-20.
184. Julkunen, I., et al., *Molecular pathogenesis of influenza A virus infection and virus-induced regulation of cytokine gene expression.* Cytokine Growth Factor Rev, 2001. **12**(2-3): p. 171-80.
185. Pirhonen, J., et al., *Virus infection activates IL-1 beta and IL-18 production in human macrophages by a caspase-1-dependent pathway.* J Immunol, 1999. **162**(12): p. 7322-9.
186. Zalinger, Z.B., R. Elliott, and S.R. Weiss, *Role of the inflammasome-related cytokines Il-1 and Il-18 during infection with murine coronavirus.* J Neurovirol, 2017. **23**(6): p. 845-854.
187. Thoulouze, M.I., et al., *The neural cell adhesion molecule is a receptor for rabies virus.* J Virol, 1998. **72**(9): p. 7181-90.
188. Srivastava, M., et al., *Chemical proteomics tracks virus entry and uncovers NCAM1 as Zika virus receptor.* Nature Communications, 2020. **11**(1): p. 3896.
189. Rodríguez-Gallego, E., et al., *Proteomic Profile Associated With Loss of Spontaneous Human Immunodeficiency Virus Type 1 Elite Control.* The Journal of Infectious Diseases, 2019. **219**(6): p. 867-876.

190. Liu, K.T., et al., *Serum Galectin-9 and Galectin-3-Binding Protein in Acute Dengue Virus Infection.* Int J Mol Sci, 2016. **17**(6).
191. Pliakos, E.E., et al., *The Cost-Effectiveness of Rapid Diagnostic Testing for the Diagnosis of Bloodstream Infections with or without Antimicrobial Stewardship.* Clin Microbiol Rev, 2018. **31**(3).
192. Hoang, L.T., et al., *Global gene expression profiling identifies new therapeutic targets in acute Kawasaki disease.* Genome Med, 2014. **6**(11): p. 541.
193. Newman, A.M., et al., *Determining cell type abundance and expression from bulk tissues with digital cytometry.* Nature Biotechnology, 2019. **37**(7): p. 773-782.
194. Tanaka, K., et al., *Role of the major histocompatibility complex class I antigens in tumor growth and metastasis.* Annu Rev Immunol, 1988. **6**: p. 359-80.
195. Danese, S., *Nonimmune cells in inflammatory bowel disease: from victim to villain.* Trends Immunol, 2008. **29**(11): p. 555-64.
196. Asano, T. and S. Ogawa, *Expression of IL-8 in Kawasaki disease.* Clin Exp Immunol, 2000. **122**(3): p. 514-9.
197. Biezeveld, M.H., et al., *Sustained activation of neutrophils in the course of Kawasaki disease: an association with matrix metalloproteinases.* Clin Exp Immunol, 2005. **141**(1): p. 183-8.
198. Peters, M.J., et al., *The transcriptional landscape of age in human peripheral blood.* Nature Communications, 2015. **6**(1): p. 8570.
199. Bongen, E., et al., *Sex Differences in the Blood Transcriptome Identify Robust Changes in Immune Cell Proportions with Aging and Influenza Infection.* Cell Rep, 2019. **29**(7): p. 1961-1973.e4.
200. Tanaka, T., et al., *Plasma proteomic biomarker signature of age predicts health and life span.* eLife, 2020. **9**: p. e61073.
201. Miike, K., et al., *Proteome profiling reveals gender differences in the composition of human serum.* Proteomics, 2010. **10**: p. 2678-91.
202. Kimura, Y., et al., *Identification of candidate diagnostic serum biomarkers for Kawasaki disease using proteomic analysis.* Scientific Reports, 2017. **7**(1): p. 43732.
203. Whitin, J.C., et al., *A Novel Truncated Form of Serum Amyloid A in Kawasaki Disease.* PLoS One, 2016. **11**(6): p. e0157024.
204. Manlhiot, C., et al., *Environmental epidemiology of Kawasaki disease: Linking disease etiology, pathogenesis and global distribution.* PLoS One, 2018. **13**(2): p. e0191087.
205. Rodó, X., et al., *Tropospheric winds from northeastern China carry the etiologic agent of Kawasaki disease from its source to Japan.* Proc Natl Acad Sci U S A, 2014. **111**(22): p. 7952-7.
206. Curtis, N., et al., *Evidence for a superantigen mediated process in Kawasaki disease.* Arch Dis Child, 1995. **72**(4): p. 308-11.
207. Han, S.B. and S.Y. Lee, *Antibiotic use in children with Kawasaki disease.* World J Pediatr, 2018. **14**(6): p. 621-622.
208. Diz, A.P., M. Martínez-Fernández, and E. Rolán-Alvarez, *Proteomics in evolutionary ecology: linking the genotype with the phenotype.* Mol Ecol, 2012. **21**(5): p. 1060-80.
209. Burgner, D., et al., *A genome-wide association study identifies novel and functionally related susceptibility Loci for Kawasaki disease.* PLoS Genet, 2009. **5**(1): p. e1000319.
210. Rypdal, M., et al., *Clustering and climate associations of Kawasaki Disease in San Diego County suggest environmental triggers.* Scientific Reports, 2018. **8**(1): p. 16140.
211. Verdoni, L., et al., *An outbreak of severe Kawasaki-like disease at the Italian epicentre of the SARS-CoV-2 epidemic: an observational cohort study.* The Lancet, 2020. **395**(10239): p. 1771-1778.
212. Levin, M., *Childhood Multisystem Inflammatory Syndrome — A New Challenge in the Pandemic.* New England Journal of Medicine, 2020. **383**(4): p. 393-395.
213. Anderson, S.T., et al., *Diagnosis of childhood tuberculosis and host RNA expression in Africa.* N Engl J Med, 2014. **370**(18): p. 1712-1723.
214. Omar, M., et al., *Host Blood Gene Signatures Can Detect the Progression to Severe and Cerebral Malaria.* Frontiers in cellular and infection microbiology, 2021. **11**: p. 743616-743616.
215. Li, H.K., Kaforou, M., Rodriguez-Manzano, J., Channon-Wells, S., Moniri, A., Habgood-Coote, D., Gupta, R., Mills, E. A., Arancon, D., Lin, J., Chiu, Y., Pennisi, I., Miglietta, L., Obaray, N., Mehta, R., Herberg, J., Wright, V. J., Georgiou, P., Shallcross, L., Mentzer, A. J., Levin, M., Cooke, G., Noursadeghi, M. and Sriskandan, S. , *Discovery and Validation of a 3-Gene Transcriptional Signature to Distinguish COVID-19 and Other Viral Infections from Bacterial Sepsis in Adults; A Case-Control then Observational Cohort Study.*
216. Bar-Meir, M., et al., *Characterizing the differences between multisystem inflammatory syndrome in children and Kawasaki disease.* Scientific Reports, 2021. **11**(1): p. 13840.
217. S. Lamble, E.B., M. Attar, D. Buck, R. Bowden, G. Lunter, D. Crook, B. El-Fahmawi & P. Piazza *Improved workflows for high throughput library preparation using the transposome-based nextera system.* BMC Biotechnology, 2013. **13**(104).
218. Hirotsu, Y., et al., *Direct comparison of Xpert Xpress, FilmArray Respiratory Panel, Lumipulse antigen test, and RT-qPCR in 165 nasopharyngeal swabs.* BMC Infectious Diseases, 2022. **22**(1): p. 221.

219. Hong, L.Z., et al., *Systematic evaluation of multiple qPCR platforms, NanoString and miRNA-Seq for microRNA biomarker discovery in human biofluids.* Scientific Reports, 2021. **11**(1): p. 4435.
220. Pabinger, S., et al., *A survey of tools for the analysis of quantitative PCR (qPCR) data.* Biomolecular Detection and Quantification, 2014. **1**(1): p. 23-33.
221. Ouldali, N., et al., *Hyper inflammatory syndrome following COVID-19 mRNA vaccine in children: A national post-authorization pharmacovigilance study.* The Lancet Regional Health – Europe, 2022. **17**.
222. Noval Rivas, M., et al., *COVID-19-associated multisystem inflammatory syndrome in children (MIS-C): A novel disease that mimics toxic shock syndrome-the superantigen hypothesis.* The Journal of allergy and clinical immunology, 2021. **147**(1): p. 57-59.
223. Riphagen, S., et al., *Hyperinflammatory shock in children during COVID-19 pandemic.* Lancet, 2020. **395**(10237): p. 1607-1608.
224. Kuhn, M., *Building Predictive Models in R Using the caret Package.* Journal of Statistical Software, 2008. **28**(5): p. 1 - 26.
225. Hoffmann, M., et al., *SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor.* Cell, 2020. **181**(2): p. 271-280.e8.
226. Ferreira-Gomes, M., et al., *SARS-CoV-2 in severe COVID-19 induces a TGF-β-dominated chronic immune response that does not target itself.* Nature Communications, 2021. **12**(1): p. 1961.
227. Eastman, S.W., et al., *Identification of human VPS37C, a component of endosomal sorting complex required for transport-I important for viral budding.* J Biol Chem, 2005. **280**(1): p. 628-36.
228. Bizzotto, J., et al., *SARS-CoV-2 Infection Boosts MX1 Antiviral Effector in COVID-19 Patients.* iScience, 2020. **23**(10): p. 101585.
229. Staeheli, P. and O. Haller, *Human MX2/MxB: a Potent Interferon-Induced Postentry Inhibitor of Herpesviruses and HIV-1.* J Virol, 2018. **92**(24).
230. Hoste, L., et al., *TIM3+ TRBV11-2 T cells and IFNgamma signature in patrolling monocytes and CD16+ NK cells delineate MIS-C.* J Exp Med, 2022. **219**(2).
231. Porritt, R.A., et al., *HLA class I-associated expansion of TRBV11-2 T cells in multisystem inflammatory syndrome in children.* J Clin Invest, 2021. **131**(10).
232. Ramaswamy, A., et al., *Immune dysregulation and autoreactivity correlate with disease severity in SARS-CoV-2-associated multisystem inflammatory syndrome in children.* Immunity, 2021. **54**(5): p. 1083-1095 e7.
233. Holm, M., et al., *Risk and Phenotype of Multisystem Inflammatory Syndrome in Vaccinated and Unvaccinated Danish Children Before and During the Omicron Wave.* JAMA Pediatrics, 2022.
234. Wang, B., et al., *Similarity network fusion for aggregating data types on a genomic scale.* Nature Methods, 2014. **11**(3): p. 333-337.
235. Rohart, F., et al., *mixOmics: An R package for 'omics feature selection and multiple data integration.* PLOS Computational Biology, 2017. **13**(11): p. e1005752.
236. Argelaguet, R., et al., *Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets.* Molecular Systems Biology, 2018. **14**(6): p. e8124.