Plasma proteomic landscape and patient stratification in response to severe infection

Yuxin Mi

University College

University of Oxford

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

Hilary Term, 2022

Abstract

Plasma proteomic landscape and patient stratification in response to severe infection

Yuxin Mi, University College, Hilary Term 2022

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

Rationale: A dysregulated host response to infection can cause life-threatening organ dysfunction, presenting as the clinical syndrome of sepsis. Sepsis is an unmet global health challenge, with substantial clinical and molecular heterogeneity hindering the development of targeted therapies. Subphenotypes in sepsis have been identified at the clinical and molecular levels, especially using the leukocyte transcriptome. Due to the limited throughput of proteome profiling, there has not been a comprehensive analysis of the sepsis plasma proteome that enables understanding of the individual response. In severe COVID-19, a maladaptive immune response to SARS-CoV-2 infection can also lead to organ failure and adverse outcome. This thesis characterises the proteomic landscape in sepsis and COVID-19 to investigate disease mechanisms and inter-individual variation in the response.

Methods: The plasma proteome was profiled using high-throughput mass spectrometry for 2622 samples from 1612 individuals, including ICU sepsis patients with community acquired pneumonia or faecal peritonitis and non-septic controls. Higher-depth proteomics and cytokine data were also generated for subsets of the sepsis patients to complement the analysis. For understanding the COVID-19 blood proteome, 395 plasma and serum samples from patients and controls were analysed.

Results: Various aspects of the immune response distinguished sepsis from control conditions, highlighting alterations in innate immunity, acute-phase response, neutrophil function and extracellular matrix organisation, with sepsis-specific proteins identified. Three sepsis patient subgroups based on the plasma proteome were identified and validated in independent samples. The subgroups were associated with differential clinical severity and distinct molecular characteristics, including one subgroup showing greater activity in immune pathways and predictive of higher mortality. The proteomic and previously defined leukocyte transcriptomic patient subgroups exhibited significant interaction, shared and distinct molecular mechanisms, and improved risk stratification when combined. Proteomic features of COVID-19 severity were identified, including acute-phase response, complement cascade, tissue necrosis and shifted lipoprotein metabolism.

Conclusion: This thesis has identified key biological processes of the sepsis response and clinically informative patient subgroups at the plasma proteome level, highlighting opportunities for patient stratification and development of a precision medicine approach.

Acknowledgements

First and foremost, I want to thank my two supervisors Julian Knight and Katie Burnham. Julian has steered every aspect of my projects with his wisdom and insight, and provided a wide variety of opportunities through the study. This thesis would not have been possible without his long-standing guidance and patience. I am also privileged to have Katie's company throughout the years, not only for all her constructive advices and considerate support on things big and small, but also for having a role model to learn from in almost all aspects of the study.

All members of the Knight group deserve my gratitude for their help at various stages. Particularly, I want to thank Cyndi Goh, Andrew Kwok, Justin Whalley, Eddie Cano-Gamez and all others working on the sepsis project for the inspiring discussions. This thesis has by no doubt benefited from the multidisciplinary knowledge in the group. I am also grateful to the GAinS investigators, in particular Charles Hinds, Emma Davenport and David Antcliffe, for valuable feedback on this work and for involving me in wider efforts of collaboration.

I would like to thank my friends who have enriched my life in Oxford, especially Jinlin, Sylvanna, and Xijin, as well as my friends who supported me from afar, especially Doudou and He. Despite the differences in endeavours we are undertaking, stories shared by you and time spent with you have always been motivational and encouraging. Finally, my mum and dad have raised and fostered my optimism in overcoming any difficulties in the journey to date and to come. Cheng who has been there for me all the time has never failed to be my endless source of faith and strength.

Declarations

I declare that, unless otherwise stated, all work presented in this thesis is my own. Several aspects of the study relied upon collaboration where part of the work was conducted with or by others.

Study recruitment: The Genomics Advances in Sepsis (GAinS) study began recruiting patients in 2005 from 34 intensive care units across the UK. From October 2017 until the study concluded in 2020, I have been responsible for liaising with research nurses, providing supplies, maintaining sample records, and collecting and processing samples, with joint effort from Andrew Kwok since October 2018. Patient recruitment and sample collection for the COvid-19 Multi-omics Blood ATlas (COMBAT) study were performed by the COMBAT consortium. Samples from other seven studies that had concluded and published prior to this thesis are also included, with details for each cohort described in Methods. Samples from these cohorts were put together for this thesis especially with help from Charles Hinds, David Antcliffe, Gareth Ackland, Hew Torrance, and Matthew Neville.

TimsTOF mass spectrometry for sepsis proteome: I randomised and aliquoted the 2622 samples, with help from Giuseppe Scozzafava. Sample preparation, mass spectra acquisition and protein quantification were performed by Roman Fischer, Raphael Heilig, Iolanda Vendrell, Philip Charles, and Georgina Berridge (Target Discovery Institute, Oxford).

QE-HF mass spectrometry for sepsis proteome: I randomised and aliquoted the 192 samples, with help from Katie Burnham. Sample preparation, mass spectra acquisition and protein quantification were performed by Raphael Heilig and Roman Fischer. I also participated in sample preparation and protein identification.

TimsTOF mass spectrometry for COVID-19 proteome: Sample aliquoting and preparation, mass spectra acquisition and protein quantification were performed by Roman Fischer, Iolanda Vendrell, Georgina Berridge and Philip Charles. Alberto Santos Delgado (Big Data Institute, Oxford) pre-processed the plasma dataset. I pre-processed the serum dataset. The Luminex data in COMBAT was generated by Luzheng Xue, Yi-ling Chen and Jian Luo.

Gene expression: RNA sample processing was performed by Katie Burnham, Emma Davenport, Jayachandran Radhakrishnan, myself, Alice Allcock, Narelle Magueri, and Ashley Thorpe. Microarray data was generated by the Wellcome Centre for Human Genetics Core Genomics facilities and previously described by Davenport et al. (2016) and Burnham et al. (2017). Batches of microarray data have been cleaned-up and combined by Katie Burnham, and processed to gene-level quantifications by Eddie Cano-Gamez. RNA-seq data was generated by the Wellcome Sanger Institute and processed by Katie Burnham. The SRS (sepsis response signature) transcriptomic endotypes were identified, validated, and further developed by Davenport et al. (2016), Burnham et al. (2017) and Cano-Gamez et al. (2022).

Associated Publications

An integrated multi-omics blood atlas reveals signatures and drivers of severe COVID-19
Cell (2022), Volume 185, Issue 5, Pages 916-938.e58
The COMBAT Consortium

High-throughput mass spectrometry maps the sepsis plasma proteome and differences in response
Manuscript in submission
Y Mi , KL Burnham, PD Charles, R Heilig, I Vendrell, J Whalley, HD Torrance, DB Antcliffe, SM May, MJ Neville, G Berridge, P Hutton, C Goh, J Radhakrishnan, A Nesvizhskii, F Yu, GAinS Investigators, EE Davenport, S McKechnie, R Davies, DJP O'Callaghan, P Patel, F Karpe, AC Gordon, GL Ackland, CJ Hinds, R Fischer, JC Knight

Inflammatory sub-phenotypes in sepsis: relationship to outcomes, treatment effect and transcriptomic sub-phenotypes
Manuscript in submission
DB Antcliffe, **Y Mi**, S Santhakumaran, KL Burnham, AT Prevost, JK Ward, T Marshall, C Bradley, F Al-Beidh, P Hutton, S McKechnie, EE Davenport, CJ Hinds, CM O'Kane, DF McAuley, M Shankar-Hari, AC Gordon, JC Knight

Contents

List of Figures

List of Tables

Abbreviations

95% CI	95% confidence interval
ALT	Aspartate transaminase
APACHE	Acute Physiology and Chronic Health Evaluation
APP	Acute phase proteins
ARDS	Acute respiratory distress syndrome
AST	Alanine transaminase
AUC	Area under the curve
B-H	Benjamini-Hochberg
CAP	Community acquired pneumonia
CDC	Centers for disease control and prevention
CDF	Cumulative distribution function
COMBAT	COVID-19 multi-omics blood atlas
COPD	Chronic Obstructive Pulmonary Disease,
COVID/COVID-19	Coronavirus disease 2019
Cox PH	Cox proportional hazards
CPAP	Continuous positive airway pressure
DA	Differentially abundant
DAMP	Damage-associated molecular pattern
DE	Differentially expressed
DIA	Data-independent acquisition
ECM	Extracellular matrix
ECMO	Extra-corporeal membrane oxygenation
eCRF	Electronic case report form
ED	Emergency department
eQTL	Expression quantitative trait locus
FC	Fold change
FDR	False discovery rate
FP	Faecal peritonitis
GAinS	Genomic advances in sepsis
GLM	Generalised linear model
GOBP	Gene ontology biological process
GOCC	Gene ontology cellular component
GPCR	G protein-coupled receptor
GWAS	Genome-wide association study
HLA	Human leukocyte antigen
HR	Hazard ratio
ICU	Intensive care unit
IGF	Insulin-like growth factor
IL	Interleukin
INR	The international normalised ratio
IQR	Interquartile range
KNN	K-nearest neighbours
LOD	Limit of detection
LOQ	Limit of quantification

LPS	Lipopolysaccharide
MHC	Major histocompatibility complex
MS	Mass spectrometry
NK	Natural killer
No.	Number
PAMP	Pathogen-associated molecular pattern
PCA	Principal component analysis
PCC	Pearson's correlation coefficient
PLS-DA	Partial least squares-discriminant analysis
PMN	Polymorphonuclear cells
pQTL	Protein quantitative trait locus
PRR	Pattern recognition receptor
PVE	Proportion of variance explained
QC	Quality control
QE-HF	Q-Exactive HF
qPCR	Quantitative polymerase chain reaction
s.d.	Standard deviation
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
SGUL	St George's, University of London
SIRS	Systemic inflammatory response syndrome
SNP	Single nucleotide polymorphism
SOFA	Sequential organ failure assessment
SRCC	Spearman's rank-order correlation coefficient
SRS	Sepsis response signature
TDI	Target Discovery Institute
timsTOF	Trapped Ion Mobility Spectrometry - Time-of-Flight
TLR	Toll like receptor
TNF	Tumour necrosis factor
VANISH	Vasopressin vs Norepinephrine as Initial Therapy in Septic Shock
VSN	Variance Stabilizing Normalization
WHO	World Health Organization

1

Introduction

This chapter presents the aims of this thesis, and provides an overview of the context to these research questions. In particular, blood proteomics approaches are introduced, and existing knowledge on the host response to sepsis is reviewed.

The overall objective of this thesis is to investigate the individual host response to severe infections, primarily in the context of sepsis, using proteomics and other omics approaches in order to facilitate patient stratification and biomarker discovery. Sepsis is a clinical syndrome currently defined as life-threatening organ dysfunction caused by a dysregulated host response to infection (Singer et al. 2016). Functional

genomics studies have revealed potential disease subtypes with different underlying biology, suggesting that molecular profiling could be a powerful approach towards understanding the host response and improving treatment for individual patients.

In this thesis I describe the first comprehensive analysis of the sepsis plasma proteome covering a wide dynamic range, for a large-scale, well-defined ICU sepsis cohort and comparator groups, including finely-characterised phenotypes. I will describe the sepsis specific blood proteome, and investigate the heterogeneity in sepsis patients primarily from the standpoint of plasma proteins but also utilising information from whole blood leukocyte transcriptomic endotypes. As part of an Oxford-led consortium contributing to the pandemic response to understand COVID-19, I also investigated the plasma and serum proteome following SARS-CoV-2 infection for different disease severities and comparing with all-cause sepsis and influenza. The work described in Chapters 3 and 4 of this thesis was conducted as part of the UK Genomic Advances in Sepsis study (GAinS). The work described in Chapter 5 was conducted as part of the COvid-19 Multi-omics Blood ATlas (COMBAT) consortium.

In this introductory chapter, I will provide an overview of the pre-existing knowledge and approaches relevant to these aims in order to put the work described in this thesis in the current context of these topics. Firstly, I will introduce the omics approaches to study human complex diseases, with an emphasis on current understanding of the blood proteome and the methods that can be used to study this composition. I will then describe the maladaptive immune response to infection in the context of sepsis, and review the functional genomics studies in sepsis, particularly at the blood protein level. I will also introduce the background of investigating the blood proteomic response to SARS-CoV-2 infection. Finally, I will detail my specific thesis aims.

1.1 Multi-omics approaches in multifactorial diseases

In contrast to monogenic disease, multifactorial diseases such as cancer, cardiovascular diseases, immune and autoimmune disorders, and neurodegenerative diseases are caused by the simultaneous action of multiple factors including variation in multiple genes, gene expression, proteins, metabolites, environmental factors such as lifestyle, and in the case of infectious diseases the pathogen and pathogen-host interaction. An "omics" approach refers to a global and simultaneous assessment of all measurable molecules or characteristics of the same type in a certain system (adapted from Hasin et al. 2017). The promise of omics approaches is that by measuring these factors as comprehensively as possible in an individual, it might be possible to identify the variation and interactions that are causing the diseases, and to provide targets for therapeutic intervention or markers for patient stratification. Such disciplines often include genomics, transcriptomics, proteomics, metabolomics, as well as emerging types such as glycomics, lipidomics, metagenomics, and phenomics.

Among omics approaches, nucleic-acid based technologies are the most advanced in terms of standardised protocols, analytical tools, and public databases (Misra et al. 2019). The rapid advancements in next-generation sequencing (NGS) technologies in the last two decades (Goodwin et al. 2016) have equipped biomedical researchers with high-throughput data generation methods for genomes (e.g., single nucleotide polymorphisms and copy number variants), epigenomes (e.g., DNA methylation, histone modifications, and chromatin accessibility), transcriptomes (e.g., gene expression, splicing variants, non-coding RNAs), as well as with single-cell sequencing profiles in the most recent years. On the other hand, proteomics is an approach to quantify the collection of proteins in certain sample types using both shotgun and targeted approaches, as will be described in Section 1.3.

These technologies have been used for various research and clinical applications, illustrated by the following examples. Zierer et al. (2016) used a mixed graphical

model to integrate selected age-associated markers from four omics datasets along with disease phenotypes from 510 individuals, identifying seven modules that represent distinct aspects of aging, and demonstrating the interconnectivity of age-related diseases. Krishnan et al. (2018) integrated genotyping and liver and adipose transcriptome data in ~100 diverse inbred strains of mice with non-alcoholic fatty liver disease (NAFLD), used gene network modeling to predict regulatory genes which were then validated in knockdown experiments, highlighting mitochondria dysfunction as a key mechanistic driver of NAFLD. Niu et al. employed mass spectrometry-based proteomics to reveal novel plasma proteins associated with NAFLD (2019), and further proposed biomarker panels for identifying early stages of liver fibrosis, inflammation and steatosis in alcohol-related liver disease using a paired liver-plasma proteomics approach (2020).

Integration of genomic evidence and other omics datasets also informs the discovery of novel drug targets or repurposing opportunities, improving the chance of selecting therapeutically valid targets for clinical trials. It is estimated that genetically supported mechanisms could double the success rate of a drug target to proceed from phase I trial to approval, compared with those without it (Nelson et al. 2015). To facilitate this process, both Open Targets and the Priority Index are sate-of-the-art pipelines developed for prioritising drug target genes from genetic evidence (Ghoussaini et al. 2021; Fang et al. 2019). The Priority Index additionally leverages the knowledge of molecular interactions, enabling identification of a network of highly-rated genes that mediate pathway crosstalk in immune-mediated diseases, and has been demonstrated to successfully recover experimentally or clinically verified targets (Fang and Knight 2022).

Terminology in patient subgrouping

One major objective of omics approaches is to stratify patients with a variable clinical phenotype into more homogeneous subgroups with treatable traits. Terminology for

such subgroups has been interchangeably and inconsistently used by researchers, as discussed by Seymour et al. (2017) and DeMerle et al. (2021). According to one plausible definition consistently given by Seymour et al. (2017) and Reddy et al. (2020), there are usually three stages of advancing the understanding of such patient subgroupings through research and clinical trials, starting from a phenotype: within the lexicon of precision medicine, a phenotype is a set of clinical features or traits that identify a group of patients with a common syndrome or condition; a subphenotype is a set of features, e.g. gene expression or biomarker profiles, that are associated with disease characteristics and distinguish groups of patients that share a presenting phenotype, without proof of mechanism and causality, with "sub" referring to either "not visible clinically" or "a further division" (Seymour et al. 2017); an endotype is a subphenotype with proven differences in the biological mechanism, often associated with an anticipated response to treatment; if a treatment targeted to the mechanism is proven clinically successful, the endotype becomes a treatable trait. On the other hand, a subgroup, subtype, or subpopulation are broader terms for the division of a patient population by any observable characteristics, without specific clinical or mechanistic inference.

Biomarker discovery

Another major objective of omics approaches is biomarker discovery. A biomarker, as defined by the FDA-NIH biomarker working group (2016), is "a defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions". Types of biomarkers include specific analytes or molecules (e.g., fibrinogen), anatomic or radiographic features (e.g., joint angle), or physiologic characteristics (e.g., blood pressure). One successful example of biomarkers guiding immunotherapy is that patients with a colorectal tumor bearing K-ras (a G-protein downstream of EGFR) mutations did not benefit from cetuximab (a monoclonal antibody against EGFR),

whereas patients with a wild-type K-ras tumor did benefit from cetuximab (Karapetis et al. 2008). As another established example, at least six endotypes have been identified or proposed in asthma, including aspirin-sensitive asthma (ASA), allergic asthma, and severe late-onset hypereosinophilic asthma (Lötvall et al. 2011). Among these, the ASA endotype has been recognised for decades. In addition to a distinct clinical presentation, ASA is associated with a mechanism of increased cysteinyl leukotriene production and overexpression of leukotriene C4 synthase (Cowburn et al. 1998), which may be caused by gene polymorphisms in the leukotriene synthesis pathway. Accordingly, ASA is characterised by a positive response to anti-leukotriene treatment especially 5-lipoxygenase inhibitors (Dahlén et al. 1998), and can be distinguished by biomarkers including increased urinary leukotrienes and peripheral blood eosinophilia.

In the scope of this thesis, biomarkers revealed in omics approaches more narrowly refer to certain protein abundance or gene expression levels that could be measured to stratify patients and thus identify those who might benefit from a specific intervention, or monitor the effects of targeted therapies.

1.2 The human blood secretome and proteome

Of the various tissue types used in clinical sampling, blood is most commonly assayed and has a central role in healthy or pathological human physiology. The relatively non-invasive nature of blood sampling compared with taking biopsy at infected sites make it an ideal source of tissue sampling in clinical settings. The cellular components of blood include the erythrocytes (red blood cells), thrombocytes (platelets), and lymphocytes (white blood cells). The liquid component of blood which constitutes about 55% of the total volume is called plasma when all soluble components are retained, or serum when the coagulation cascade has been activated. Most cells of the body have direct or indirect communications with plasma. The main functions of

plasma include transportation of blood cells/ nutrients/ gases/ wastes, homeostatic regulation, as well as mediating many aspects of the immune response. Many of these functions are carried out through proteins secreted to the blood and acting as enzymes, hormones, complement components, protease inhibitors, coagulation factors, immune-active molecules, and transporters of lipids/ hormones/ vitamins/ minerals.

Secretome

The secretome is a collection of proteins secreted to a compartment or organ system such as the blood, the digestive tract, the brain, as well as certain intracellular compartments like Golgi and ER lumen. Human cells such as endocrine cells and B-lymphocytes are specialized in protein secretion, but other cell types also secrete proteins at a varying level. Liver is the tissue of origin for more than half of the actively secreted blood proteins, including serum albumin (ALB), fibrinogen (FGA/FGB/FGG), transferrin (TF), many of the apolipoproteins and complement related proteins.

Plasma proteome

The human plasma proteome comprises three main categories of proteins (Anderson and Anderson 2002). The most abundant are the classical plasma proteins with functional roles in the blood. These include: ALB (~45mg/mL) that maintains osmotic pressure and binds and transports various substances; apolipoproteins(e.g., APOA1), mediating lipid transport and metabolism; acute phase proteins (e.g., CRP), with crucial roles in innate immunity; and components of the coagulation cascade including C2, F12, and fibrinogens. The second category are the tissue leakage products, including aspartate aminotransferase, alanine aminotransferase, and the troponins (e.g., TNNT2). The third category are signalling molecules including protein hormones and cytokines (including chemokines, interferons, interleukins, growth factors). These molecules usually have low baseline abundances in steady state and are up-regulated

in relevant conditions. For example, the baseline levels of IL-6 and IL-1 are below 5pg/mL. Most of the proteins in the first and third categories are actively transported out of the cells so are secretory proteins.

The Human Protein Atlas (HPA) project predicted 14% (2793 genes) of the human protein-coding genes to have at least one secreted protein product, and 784 proteins to be actively secreted to peripheral blood (Uhlén et al. 2015). Among the secreted protein mass in the plasma, albumin is the most abundant and accounts for 55% of blood proteins; 38% are made up by globulins that transport ions/ hormones/ lipids and immunoglobulins; fibrinogens account for 7%; and the remainder are mainly regulatory proteins including cytokines, enzymes, proenzymes, and hormones. Leakage products from cell death or damage and proteins present in the blood due to diseases have largely variable levels of abundance in the circulation. By collecting from published studies and publicly available databases, HPA presented plasma concentrations for 435 proteins based on immunoassays, and for 4072 proteins based on mass spectrometry. HPA also measured the longitudinal plasma levels for 1456 proteins based on proximity extension assay. Together this is a useful resource to understand blood protein concentrations and the main functional groups (Uhlén et al. 2019; proteinatlas.org).

Clinical importance

Components of the plasma including electrolytes, small molecules, drugs and proteins are routinely measured for clinical diagnostics purposes. The aims of such tests include disease diagnosis or confirmation, risk prediction, evaluating effectiveness of treatments or monitoring prognosis. Acknowledging that differences could exist between hospital systems, estimation based on tests performed in University Hospital Munich in 2016 showed that among the blood components, proteins are the most frequently assayed analytes in clinical practice (42% of analyses), followed by small molecules (35%) and cells (17%) (Geyer et al. 2017). More protein-based diagnostics

are being developed with the rapid improvement of protein quantification techniques.

In addition to being a rich source of blood diagnostic tests, blood proteins are also widely recognised as therapeutic targets. The majority of drugs target proteins. Out of the 812 protein targets of FDA-approved drugs that are directly related to pharmacological mechanisms, 80 are secreted proteins, 443 are integral membrane or single pass trans-membrane proteins, and 29 have membrane and secreted isoforms (Uhlén et al. 2015; Wishart et al. 2018). Therefore, a comprehensive analysis of the blood proteome in disease conditions is important not only for understanding the pathophysiology, but also for implicating potential therapeutic and diagnostic targets.

1.3 Mass spectrometry-based proteomics

A typical clinical test of blood proteins targets a single or a few analytes with either an enzymatic assay or an antibody-based immunoassay. In contrast to these assays, untargeted mass spectrometry (MS) can in principle measure all the proteins, and is hypothesis-free, thus suitable for exploratory approaches. It also does not suffer from the specificity issue of affinity-based assays since the mass and sequence of peptides are unique (Geyer et al. 2017).

1.3.1 Principle of mass spectrometry

Protein identification and quantification through tandem MS is based on accurately measuring the mass of fragmentation spectra of peptides derived from protein digestions, separated according to their mass-to-charge ratio (m/z). For example, in a typical data-dependent acquisition workflow of LC-ESI-MS/MS, proteins in samples are first enzyme-digested and the peptides are separated with high-performance liquid chromatography (LC). The successive sets of eluted peptides gain charge through electrospray ionisation (ESI). The ionised peptides are then separated by m/z in the first mass analyser (MS1) and suitable precursor ions are isolated and entered into

the collision cell where chemical bonds break at specific positions. The successive peaks differ by the mass of particular amino acids in the original peptide, thus the sequence of the peptide are revealed by m/z detected in the second mass analyser (MS2). The abundance of a certain protein can then be determined by the area-under-curve, totalling the intensity of its representative peptides over MS scanning time in the time-serial spectra in MS1 or MS2.

Challenges in measuring the blood proteome

In complex matrices like the plasma, the depth of the analysis is thus limited by the peptide-sequencing speed and sensitivity of the mass spectrometer. Peptides with higher abundance are more likely to be ionised, and also more likely to enter from MS1 to MS2. Thus, the more abundant proteins are generally more likely to be identified and quantified. The presence of proteins with high abundance like ALB will prevent the low-abundance proteins from being assayed. A key challenge in measuring the blood proteome is the extraordinary dynamic concentration range spanning more than 10 orders of magnitude, from the scale of pg/mL to mg/mL (Schiess et al. 2009). The detection limit of untargeted mass spectrometry experiments most often covers classical plasma proteins and tissue leakage products. The low-abundance cytokines and other signalling peptides can only be detected with affinity-based assays, or in MS assays that are targeted, or fractionated, or with an extended acquisition time for each sample.

Considering the discrepancy between MS detection depth and the requirements for biomarker discovery, in my study of the sepsis blood proteome I applied the antibody based Luminex xMAP technology to complement the MS measurements and to achieve an improved protein coverage in a subset of samples. The Luminex assay detects protein analytes with a high sensitivity in a form of bead-based ELISA (Reslova et al. 2017). I chose a highly-multiplexed panel with 65 targets (mostly cytokines) to allow an optimal coverage at the low-abundance end.

1.3.2 Technological advancements

Mass spectrometry-based proteomics has been an essential tool for clinical and biomedical research. However, challenges remain in the proteome discovery depth in clinical specimen, the throughput in assaying larger number of samples in a single experiment, and the comparability between batches. The technology has rapidly improved over the last decade, approaching the sensitivity, dynamic range, and throughput required for clinical biomarker studies (Aebersold and Mann 2016). For example, the research team led by Matthias Mann developed a streamlined workflow "Plasma Proteome Profiling" for rapid, shotgun, label-free proteome quantification (Geyer et al. 2016a), with successful applications in depicting proteome changes in weight loss (Geyer et al. 2016b), in Roux-en-Y gastric bypass surgery (Albrechtsen et al. 2018), and in non-alcoholic fatty liver disease (Niu et al. 2019). The research team led by Markus Ralser proposed a data-independent acquisition (DIA) method, Scanning SWATH that substantially accelerates MS duty cycles and demonstrated the utility in identifying biomarkers for COVID-19 severity (Messner et al. 2021).

The throughput in clinical mass spectrometry has been particularly improved with the timsTOF system, which couples Trapped Ion Mobility Spectrometry (TIMS) to ultra-high resolution Quadrupole Time-of-Flight (QTOF) technology. In TIMS, ions are carried forward by a gas flow while being held back by an electric field, thus accumulating at a position in the tunnel where the two forces balance. Ion species separated by ion mobility are then released from the device into the QTOF mass analyser as a function of their collisional cross section, therefore adding a fourth dimension of separation on top of m/z, retention time, and intensity. The parallel accumulation-serial fragmentation could improve the peptide-sequencing speed up to ten folds without compromising sensitivity (Meier et al. 2015).

Our collaborators Roman Fischer et al. at the Target Discovery Institute (Oxford) were among the first groups of scientists to apply the timsTOF system to clinical samples

from mid-2019, along with developing an automated sample preparation platform and investigating computational methods to cope with the protein inference and quantification from the high volume of mass spectra acquired (Kosinski et al. 2019). This has provided us the opportunity to profile a large number of clinical samples in one batch in order to be sufficiently powered to answer our disease-specific questions.

Meanwhile, affinity-based proteomics assays have also significantly advanced in recent years. The Olink platform combines proximity extension assays with next generation DNA sequencing and allows for simultaneous analysis of around 1500 protein targets (Zhong et al. 2021). The dual-binding to more than one epitope on the target protein by antibodies reduces cross-reactivity; amplification and sequencing of DNA barcodes overcome the limitation in number of distinguishable targets in conventional fluorescence readouts. The SOMAscan platform utilises DNA aptamers to bind to protein targets and then to be quantified by standard DNA techniques, allowing quantification of from 1300 to over 5000 proteins in complex matrices (Gold et al. 2010). These two platforms have a much higher per-sample cost than mass spectrometry, and target large pre-defined panels of protein targets selected based on prior knowledge on the disease area e.g. inflammation and immune response, respiratory, oncology, or metabolic diseases. These assays have been successfully applied in many health and disease research areas, especially in mapping the genetic determinants of circulating proteins (pQTLs) in large cohorts (Macdonald-Dunlop et al. 2021; Suhre et al. 2017; Sun et al. 2018).

1.4 Immune response and sepsis

The immune system is the body's defence against pathogenic microorganisms including viruses, bacteria and archaea, fungi, and parasites (Murphy and Weaver 2016). It is made up of a variety of effector cells and molecules and functions mainly through three lines of protection: the anatomic barriers, innate and adaptive

immunity. Other than immunity initiated by exogenous pathogens, traumatic insults (major surgery, injury, burns) can also induce innate immunity and inflammation following the disruption of macrobarriers (e.g., skin) and microbarriers (e.g., cell membrane) and exposure of the immune system to "self" damage-associated molecular patterns (DAMPs) as well as "non-self" pathogen-associated molecular patterns (PAMPs) (Huber-Lang et al. 2018).

1.4.1 Immune diversity, maladaptive response and sepsis

The human immune system is extremely diverse from individual to individual. This intrinsic variability is rooted in the fact that immunity is controlled by the most polymorphic genes, including the MHC and immunoglobulin gene regions. Up to half of the observed inter-individual immune variation is driven by cumulative effects of thousands of genome loci, while the residue is potentially explained by variations in age, sex, diet, environmental exposure, and the microbiome (Liston et al. 2021). As examples, farm-animal exposure was shown to protect from asthma (Stein et al. 2016); increased TH17 cells was shown as a potential mechanism linking high-salt diet to autoimmunity (Kleinewietfeld et al. 2013); gut-microbiome transfer in mice was shown to promote immune biases in T cell differentiation (Atarashi et al. 2013, 2015, 2017). As each variant in the immune system can be both beneficial and detrimental depending on the context (e.g., specific pathogens or autoimmune disorders), the inter-individual diversity has been both an evolutionary advantage as well as the substrate for various immune-associated diseases among the population.

Following the recognition of PAMPs or DAMPs by the pattern recognition receptors (PRRs), gene expression activation lead to the induction of inflammation and innate immunity, including cytokine production. Ideally, a balanced pro-inflammatory and anti-inflammatory reaction in an immune response is obtained in order to rapidly clear pathogens or damaged tissues and restore the cell and tissue structure. However, in both sterile and infectious inflammatory conditions, for a subset of the patients the

immune response can become dysregulated with a failure to return to homeostasis (a maladaptive immune response). The cause of this dysregulation is not well-understood but risk factors can include both individual factors and environmental factors (e.g., haemorrhage, secondary infection, or extended surgical intervention (Huber-Lang et al. 2018)). Escalation of the innate immune response can lead to further barrier disturbance, compromised defence, and changes in metabolic and hypoxic states, generating more DAMPs and PAMPs, and thus forming a vicious cycle of the innate immune response. The systemic inflammatory response syndome (SIRS) describes this exaggerated defence response to noxious stressors (usually an infection but also include sterile insults). Clinically SIRS is defined by satisfaction of any two of the criteria including: (1) body temperature >38°C or <36°C; (2) heart rate >90bpm; (3) respiratory rate >20 breaths/min or $PaCO_2$ <32 mmHg; (4) leucocyte count >12000 or >4000/mm^3, or >10% immature bands (Chakraborty and Burns 2021).

The first consensus definition of sepsis described it as a SIRS caused by infection (Sepsis-1; Bone et al. 1992). The second consensus described sepsis as a clinical syndrome of infection complicated by acute organ dysfunction, kept the diagnostic criteria from Sepsis-1 but provided more detailed signs and symptoms (Sepsis-2; Levy et al. 2003). Kaukonen et al. (2015) demonstrated that 12.1% patients (n=13,278) with severe infection and organ failure did not meet the SIRS criteria on presentation but showed similar characteristics including mortality with those who met the SIRS criteria. Improved understanding of the pathobiology recognises sepsis as a multifaceted host response involving both excessive inflammation and immunosuppression along with major modifications in non-immunologic pathways. Acknowledging these, the most recent consensus criteria (Sepsis-3) removed the requirement of meeting SIRS criteria and took a broader perspective, defining sepsis as a life-threatening organ dysfunction that is caused by a dysregulated host response to infection (Singer et al. 2016).

1.4.2 Sepsis epidemiology

Sepsis is a global health priority highlighted by the WHO (Reinhart et al. 2017). Based on records from 195 countries and territories, the estimated burden of sepsis in 2017 is 48.9 million incident cases and 11.0 millon deaths worldwide, representing 19.7% of all global deaths (Rudd et al. 2020). Sepsis is the primary cause of admission to and death in an intensive care unit (ICU) in the UK and the US. For example in the US, sepsis is the most common cause of in-hospital deaths and costs more than US$24 billion annually. Substantial differences in burden and case-fatality remain between regions, with higher burden observed in areas with a lower socio-demographic index. Therefore, despite a trend of decreasing burden, sepsis clearly imposes a substantial global concern in terms of morbidity and mortality and warrants greater attention from healthcare and research communities (Genga and Russell 2017; Mayr et al. 2014; Rudd et al. 2020).

Symptoms of sepsis include difficulty in breathing, increased heart rate, low urine volume, low blood pressure, swelling, high or low body temperature, and confusion, reflecting dysfunction in respiratory/ cardiovascular/ renal/ neural functions, as well as vasodilation and a dysregulated inflammation. An increase in the sequential organ failure assessment (SOFA) score of 2 points or more was defined as the central diagnostic component (Singer et al. 2016). Sepsis-3 also recommends using the quick SOFA score which does not require laboratory testing to identify patients with suspected infection who are likely to have poor outcomes. A subset of sepsis patients develop septic shock, which is defined as hypotension despite adequate intravenous fluid resuscitation (Levy et al. 2003; Sepsis-2). Septic shock is associated with substantially higher mortality. Risks factors of developing sepsis from the infections include being very young or old, diabetes, weakened immune system, major surgery, and giving birth.

The most common infection sources of sepsis are respiratory (including community

acquired pneumonia (CAP)), gastrointestinal (including fecal peritonitis (FP)), genitourinary and musculoskeletal. An epidemiology study based on adult ICUs in England from 2011 to 2015 reported 50%, 25%, 6% and 5% for these four sources, respectively (tocite Shankar-Hari et al. 2017). No definitive source is found in 7% of the cases. Sepsis can be caused by infection from bacteria, virus, fungi, or mixed pathogens. The most commonly identified pathogens for CAP include *Streptococcus pneumoniae*, influenza and other viruses. FP is caused by colonic perforation resulting from cancer, trauma, or stitch breakage after abdominal surgery, which leads to infection of the peritoneum area by the gut microbe. To improve pathogen identification from the enormous number of possible infectious agents, metagenomics sequencing-based methods such as *Castanet* proposed by Goh et al. (2019) are designed to be applied in high-throughput settings and augment the pathogen detection of conventional clinical microbiology methods like culture and nucleic acid amplification.

1.4.3 Sepsis immunopathology

Although pathogens are the initial causes of sepsis, it is the dysregulated host response to the pathogen that leads to the sepsis syndrome. The host immune failure leads to multiple organ dysfunction both indirectly through uncontrolled pathogen activities and directly through endothelial injury, reactive oxygen species, and hypercoagulation (Rubio et al. 2019). Excessive inflammation used to be considered as the main driver of mortality and morbidity in sepsis, but no anti-inflammatory drug trials had been successful (Marshall 2014) prior to the emergence of COVID-19. It is now believed that in sepsis, the immune response initiated by an invading pathogen fails to return to homeostasis and instead culminates in a pathological syndrome that is characterised by both sustained excessive inflammation and immune suppression. It is argued that there is also a fundamental reorganisation of immune and metabolic cell processes (Van Der Poll et al. 2017). These alterations in immune and metabolic pathways may

persist after the infection has been resolved.

Complements and coagulopathy

Activation of the complement system, the coagulation system, and vascular endothelium are interconnected hallmark inflammatory responses in sepsis (Van Der Poll et al. 2017). The complement cascade enhances the inflammatory response through binding and opsonising pathogens for engulfment by phagocytes, recruiting more phagocytes, and forming membrane-attack complexes. However, uncontrolled activation of complement factors can cause damage to tissues. Blockade of C5a and C3a signalling showed beneficial evidence in sepsis animal models but had inconsistent effects in clinical trials (Silasi-Mansat et al. 2010; Shao et al. 2015). Coagulation and the initiation of innate immunity share co-dependent pathways. As the main driver of coagulation in sepsis, tissue factor (F3) along with the clotting factors, thrombin, and fibrin can induce pro-inflammatory cell signalling via protease-activated receptors (Nieman 2016). A strong activation of the coagulation system including disseminated intravascular coagulation has been associated with sepsis, which could lead to microvascular thrombosis and haemorrhage.

Lymphopenia and lymphocyte exhaustion

Immunosuppression observed in sepsis features multiple defects in innate and adaptive immunity, including a massive decrease of circulating conventional T cells and B cells especially the activated memory B cells, and the upregulation of regulatory T cells. The persistence of sepsis-associated lymphopenia in a substantial subgroup of patients correlates with higher mortality (Drewry et al. 2014). Potential mechanisms include apoptosis through the mitochondrial and death receptor pathways, and excessive extravasation and accumulation of the lymphocytes at sites of inflammation and damaged endothelia, but data remain unclear (Chang et al. 2007; Hotchkiss et al. 2013). B cells surviving the accelerated apoptosis show an exhausted phenotype

with decreased MHC II expression and increased IL-10 production (Gustave et al. 2018). Multiple dysfunctions in T cell metabolism and activation are reported in sepsis (Kumar 2018). The T-cell-denpendent peripheral maturation of B cells is also impaired following T cell abnormalities.

Neutrophil extracellular traps

Immature myeloid cells can be generated by emergency granulopoiesis in infections and become myeloid-derived suppressor cells composed of a granulocytic and a monocytic subset. The enhanced generation and release of immature granulocytes in sepsis is associated with clinical deterioration (Daix et al. 2018). Neutrophils are the most abundant type of granulocytes and the first-line defence cells in innate immunity. The immature neutrophils have substantially impaired function but show an increased production or delayed clearance of neutrophil extracellular traps (NETs) which is a network of DNA and histone and granular proteins including MPO and S100A9. The prolonged presence of NETs in the vasculature can lead to endothelial injury and excessive coagulation (Camicia et al. 2014). The vascular leakage and sequestration of leukocytes in thrombus can lead to leukocyte loss and contribute to immune failure.

Metabolic shift

Inflammatory and cancer cells demonstrate a metabolic shift from oxidative phosphorylation to aerobic glycolysis in order to more rapidly produce ATP and metabolic intermediates for protein synthesis. The excessive activation of this process in some sepsis patients leads to succinate accumulation, which in turn leads to stabilisation of HIF-1α and increased transcription of IL-1β and other cytokines. Increased mitochondrial oxidation of succinate also induces the production of reactive oxygen species and subsequently pro-inflammatory gene expression (Mills et al. 2016). A generalised defect in energy metabolism was reported in leukocytes with LPS-induced tolerance and from patients who showed signs of immunosuppression (Cheng

et al. 2016)

Treatments

The current treatments for sepsis include antibiotics, surgery to remove areas of infection, and supportive treatments including ventilation, intravenous fluids and vasopressors. There is a relative failure in the development of targeted therapies in sepsis despite a variety of pathophysiological alterations having been described (Marshall 2014). For example, although there is clear rationale to supplement the anticoagulation pathways in treating sepsis, clinical trials for antithrombin, recombinant TFPI (tissue factor pathway inhibitor), or recombinant activated protein C failed to show consistent benefits (Van Der Poll et al. 2017). Approximately 200 biomarkers have been evaluated, but none of these is routinely used as diagnostic or prognostic markers clinically.

The relative failure in clinical trials is partly due to the fact that patient inclusion was mainly based on clinical criteria such as severity of disease, which fails to resolve the heterogeneity in underlying pathological processes. Sepsis is a highly heterogeneous and dynamic syndrome in terms of the site and pathogen of infection, the organ impaired, and importantly how the host response is disturbed at the molecular level. How the host response should be manipulated is highly controversial, regarding treatment with immunomodulators including anti-inflammatory agents or immune stimulants. Both may be appropriate approaches, depending on the activity of targetable immunological pathways in each individual patient.

In order to resolve this heterogeneity, understanding specific characteristics of the host response in individual patients on the basis of biochemical and/or immunological profiles is advocated (Van Der Poll et al. 2017). Omics approaches could help to link molecular changes at the RNA, protein, or metabolite levels with specific complications, and to identify patient endotypes. Biomarkers could be identified to reflect the pathophysiological events. It has also been demonstrated that susceptibility

to and outcome from infectious diseases has strong genetic associations. Therefore, there is great potential for developing precision medicine in sepsis, where the genomic variation and immune state of the patient should be both taken into account to apply appropriate intervention.

1.5 Genomics approaches in sepsis

1.5.1 Sepsis genetics

It was recognised as early as in 1988 that outcome from infection has a strong genetic background, as risk of premature death from infection was 5.8 fold higher in adoptees if a biological parent had died from the same cause, while no such effect was observed for the adoptive parents (Sørensen et al. 1988). In the last decades, genetics and genomics approaches have enabled better understanding of molecular mechanisms for infectious diseases. For example, susceptibility to invasive disease following *Streptococcus pneumoniae* infection has been associated with single-gene variants in *NEMO*, *IRAK4* and *MYD88*, through a probable mechanism of impaired TLR-IL-1R signalling (Kwok et al. 2020).

To understand sepsis from a genomics and functional genomics view, the UK Genomic Advances in Sepsis (GAinS) study was established in 2005 by the UK Critical Care Genomics group (co-chief investigators Hinds and Knight). This has enabled recruitment of one of the largest such cohorts worldwide, associated with a database of high quality phenotypic and clinical information. From 2005 to 2020, a total of 1313 adult patients with sepsis due to either community acquired pneumonia (CAP) or faecal peritonitis (FP) were enrolled from 34 intensive care units (ICU) in 27 sites across the UK. Buffy coat samples for DNA extraction were taken for all the GAinS patients. Serial samples of blood leukocytes (for RNA extraction), plasma, and urine were taken on days 1, 3, and/or 5 from ICU admission when available.

The first genome-wide association study (GWAS) of sepsis outcome was performed in 2534 patients from GAinS, the Genetics of Sepsis and Septic Shock in Europe (GenOSept) study and two other independent studies (Rautanen et al. 2015). The only significant association with survival in genome-wide meta-analysis was identified in common variants in the *FER* gene ($p = 9.7 \times 10^{-8}$) which is a tyrosine kinase involved in regulation of actin cytoskeleton, cell adhesion, and migration. Subsequent work in a murine model of pneumonia by Dolgachev et al. (2018) showed that *FER* overexpression enhanced innate immunity and accelerated bacterial clearance, highlighting the potential pathophysiological importance of the gene. A second GWAS reported significant associations between 28-day mortality and variants in the *VPS13A* and *CRISPLD2* genes (Scherag et al. 2016). To confirm or refute other suggestive associations from the GWA studies requires a larger sample size, a more homogeneous patient population, and a more precise definition of phenotypes.

1.5.2 Sepsis transcriptomic endotypes

A wide range of genome-wide gene expression analyses have been reported in sepsis cohorts with aims including: early diagnosis of sepsis and distinguishing from SIRS (Scicluna et al. 2015; Maslove et al. 2019); expression profiles that associate with specific phenotypes e.g. sepsis-induced acute respiratory distress syndrome (ARDS) (Kangelaris et al. 2015; Zhang et al. 2019); prediction of adverse outcome (Wong et al. 2015; Sweeney et al. 2018b; Stanski and Wong 2019); identifying disease endotypes (as described below); and more recently the immune cell states associated with the disease using single cell genomics (Reyes et al. 2020). Whole-blood transcriptomic profiling from these studies suggested that more-severe sepsis is associated with overexpression of neutrophil proteases, exhaustion in adaptive immunity, and overall profound immune dysregulation.

Work within the GAinS study has also sought to use omics approaches to resolve the functional genomics landscape of the disease with a focus on genome-wide

gene expression profiling. Davenport et al. (2016) used unsupervised clustering of peripheral blood leukocyte microarray transcriptomic data to identify two distinct groups in patients with sepsis due to CAP. The two sepsis response signature (SRS) groups were then also identified and verified in patients with sepsis due to FP, and the major predictor of variation in gene expression between sepsis patients was shown to be SRS group rather than the source of infection (Burnham et al. 2017). SRS1 patients had significantly higher mortality compared with SRS2 within 14 days of ICU admission, and showed transcriptomic features consistent with immunosuppression including endotoxin tolerance, T cell exhaustion, and HLA class II downregulation. The relatively immune competent SRS2 group showed increased mortality in response to corticosteroids on post hoc analysis of the VANISH trial (Antcliffe et al. 2018). This indicates that measuring gene expression in sepsis might be a powerful approach for identifying different underlying biology, predicting specific outcomes and drug response.

To be able to validate SRS patient subgroups in independent cohorts and to test interaction with treatment response, a set of seven genes predictive of SRS membership was identified in CAP (Davenport et al. 2016) and verified in FP (Burnham et al. 2017). By integrating data from 3149 sepsis or healthy samples, further investigation by Cano-Gamez et al. (2022) extended the SRS stratification to be scalable to other infectious aetiologies (e.g., H1N1 influenza and COVID-19) and technology platforms (qPCR and RNA-seq) and to include not only sepsis patients but also less severe patients with suspected infection, with SRS3 representing a small group of relatively healthy individuals. In addition, a quantitative score (SRSq) was derived to better reflect the continuity in the differential molecular characteristics of the immune dysfunction, which improved correlation to clinical outcomes. SRS and SRSq can be reliably and conveniently assigned using the SepstratifieR package (Cano-Gamez 2022) based on a 7-gene or 19-gene signature to stratify new patients with severe infection.

To better understand the cellular basis of the differential immune response in SRS groups, Kwok et al. (2022) applied single-cell gene expression and surface protein profiling together with mass cytometry to profile leukocytes from fresh whole blood samples. Expansion of a specific immature *IL1R2*+ neutrophil subset was highlighted in SRS1 involving alterations in *STAT3*/*CEBPB* related gene programs. More general changes in gene expression of the neutrophil compartment observed in sepsis patients persisted into convalescence.

Considering the accumulating knowledge in the biological mechanism and treatment interaction of the SRS classification, SRS patient subgroups are referred to as SRS endotypes in this thesis, although a deeper understanding of the mechanistic basis of these endotypes and whether they represent treatable traits, namely that they can be successfully targeted by an intervention, is desirable.

Researches in other sepsis cohorts have also utilised leukocyte gene expression to stratify patients into more homogeneous groups with shared biological features. For example, Scicluna et al. (2017) reported four sepsis molecular endotypes (Mars1–4) that significantly improved risk prediction combined with clinical data. The higher risk Mars1 endotype was reliably identified by *BPGM* and *TAP2*. Samples from 265 CAP patients from GAinS were used as a validation cohort in this study and the differences in SOFA scores and mortality were replicated. Sweeney et al. (2018) identified three subtypes they termed as "Inflammopathic, Adaptive, and Coagulopathic", in multiple datasets of bacterial sepsis. Baghela et al. (2022) included sepsis patients from both ICU and emergency rooms, and reported five mechanistic endotypes named "Neutrophilic-Suppressive, Inflammatory, Innate-Host-Defense, Interferon, and Adaptive". The "Adaptive" subtype in both reports are relatively benign. In paediatric sepsis, Wong et al. showcased the utility of their model for consistent risk stratification and further as a predictive enrichment tool that identified a subgroup of sepsis mice that could be rescued by higher doses of antibiotics (Wong et al. 2019).

Efforts have been made to analyse and understand the overlap among these classifications (Stanski and Wong 2019). The critical next step is to develop a consensus sepsis subgrouping unifying the existing systems, as has been achieved for colorectal cancer through large-scale international collaboration (Guinney et al. 2015). In addition to gene expression-based patient subgroups, other omics data and clinical measurements should also be utilised to derive predictive and prognostic models towards a targeted sepsis treatment.

1.6 Sepsis blood proteome

In understanding the sepsis response, it is important to investigate if leukocyte gene expression differences are propagated to the proteomic level and consider the wider sepsis proteome reflecting tissue function and dysfunction. Compared with regulation at the gene expression level, protein abundance is subject to more complexities including translational modification, post-translational modification, and protein degradation. Circulating protein levels in the peripheral blood are further affected by regulated secretion and tissue leakage. Individual blood proteins can be measured with lower cost and higher throughput than gene expression in clinical settings, which means proteins are potentially better biomarkers translatable to bedside diagnosis. Since proteins comprise the majority of drug targets, protein studies may yield implications on potential therapeutic targets.

This section summarises existing knowledge of the sepsis blood proteome, focused on the untargeted proteome discovery approach primarily using tandem mass spectrometry.

With gradual development of proteomics technologies over the last two decades, studies of the sepsis blood proteome have also seen a change from using two-dimensional electrophoresis (Swathi Raju et al. 2016; Hayashi et al. 2019) to applying labelled or label-free mass spectrometry, and aptamer- (Shubin et al. 2020) or

antibody-based multiplex protein quantification, with the supplement of ELISA (Van Vught et al. 2017; Sharma et al. 2019) or fluorescence imaging (Sharma et al. 2017) for validation.

Compared with healthy volunteers, various differentially abundant proteins and regulated pathways have been identified in both patients (Sharma et al. 2017; Sharma et al. 2019) and murine sepsis models (Toledo et al. 2019; Hohn et al. 2018; Pimienta et al. 2019). The altered biological processes identified in these studies include vascular homeostasis, coagulation cascades, acute phase response, complement, inflammatory response, and lipid metabolism.

One focus of sepsis proteomics research has been improving diagnosis of sepsis to be earlier and more accurate, particularly distinguishing sepsis from sterile inflammations (Kiehntopf et al. 2011; Tong et al. 2019; Li et al. 2022). For example, Papafilippou et al. (2020) identified 67 plasma proteins based on 19 patients (including OTUD7A, IGKV1D-13, NECAB2, FLOT2) that could potentially differentiate sepsis from SIRS based on a total of only 19 patients, demonstrating the benefit of using lipid-based nanoparticles to enrich for low-abundance plasma proteins. Thavarajah et al. (2020) reported that increases in peptide intensity from common plasma proteins (e.g., ITIH3, SAA2, FN1) as well as intracellular proteins (e.g., COL24A1, POTEB, ADAMTS7) were associated with sepsis, and highlighted sepsis-associated variation in SAA1 processing. Pierrakos et al. (2020) have reviewed the use of single targeted proteins as biomarkers in sepsis. The most frequently studied markers are C-reactive protein (CRP), procalcitonin (PCT), IL-6, presepin, and CD64. They reported 9 proposed biomarkers to have a better diagnostic value for sepsis than either or both of PCT and CRP, concluded that most of the 258 identified biomarkers had not been well-studied, emphasising the need for evaluating the role of biomarkers in sufficiently sized clinical studies.

To identify biomarker candidates predictive of outcomes, differential abundance analysis has been performed between sepsis survivors and non-survivors using

mass spectrometry-based protein quantifications (De Coux et al. 2015; Swathi Raju et al. 2016; Sharma et al. 2017). Pathways such as extrinsic coagulation, complement cascades, and actin cytoskeleton were found to associate with survival. Among these cohorts, the study from Langley et al. (2013) included a relatively large sample size of 121, 52, and 61 sepsis patients in the derivation and two validation cohorts, respectively. The authors generated an integrated dataset of clinical features, MS-based plasma metabolome (214 annotated metabolites on emergency department admission), and proteome (195 proteins identified after affinity depletion). Metabolic differences between survivors and non-survivors included relatively impaired fatty acid transport and β-oxidation, gluconeogenesis, and the citric acid cycle. Interestingly no difference was observed between shock and non-shock patients among survivors. The authors identified a predictive model of sepsis mortality that out-performed clinical scores, consisting of 3 clinical variables and 5 blood metabolites. This study illustrated the utility of an integrative omics approach to understand the metabolic alternations and to identify predictive markers, although the proteome detection depth could be improved to gain a systemic view, and a wider sample coverage would allow inclusion of more diversified aetiologies and better power to compare between the clinical or molecular subgroups of the complex trait.

The sepsis proteomic response has also been characterised longitudinally (Sharma et al. 2017; Harberts et al. 2020), or between patients with different disease severity (Tong et al. 2019). For example, Tong et al. used data-independent acquisition-based MS to assay peripheral blood mononuclear cells (PBMCs) and identified 17 proteins (including HMGB1, ACSL1, LCN2, LTF) out of 122 proteins that differentiated between 11 septic shock and 27 non-shock sepsis patients.

However, due to the limitation in mass spectrometry throughput and sample availability, existing sepsis proteomics studies in patient samples have often been restricted by a small sample size. With the exception of Langely et al. (2013),

most studies presented <40 patients in each comparator group, or reaching <90 by Kiehntopf et al. (2011) and Ng et al. (2010). Due to the high heterogeneity in sepsis pathophysiology, a limited sample size will hinder the ability to include appropriate controls with sufficient power, to detect differences between the variable clinical phenotypes, or to distinguish between the individual sepsis response at the molecular level. Furthermore, multi-omics data (e.g. transcriptome and proteome) in the same subset of patients would provide the opportunity to understand interaction between regulations at the different molecular levels and more mechanistic insight into the subphenotypes.

1.7 Respiratory failure and immune dysfunction in COVID-19

Coronavirus disease 2019 (COVID-19) was declared a pandemic in March 2020 by WHO. It has been an enormous global burden and disruption to normal life, with a total of 520 million confirmed cases and 6.2 million confirmed deaths globally reported to WHO by May 2022 (WHO COVID-19 Dashboard). The disease is caused by infection by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Although most patients infected only develop mild respiratory illness or are asymptomatic, a small proportion of patients develop severe and critical illness with lung injury or even signs of other organ dysfunction which could lead to death. These proportions vary largely by vaccination status, age, variant strain, geography and ethnicity. A systemic analysis in the pre-vaccine era estimated a decrease in median infection-fatality ratio globally from 0·466% to 0·314% between April 2020 and January 2021 (COVID-19 Forecasting Team, 2022). Cohort studies have shown the risk factors for poor outcome include older age and existing conditions like cardiovascular disease, diabetes , cancer, chronic respiratory disease and severe obesity (Mantovani et al. 2020; Fresán et al. 2021).

SARS-CoV-2 infection can cause sepsis

Although differences have been observed between clinical manifestations of COVID-19 and the more common forms of sepsis in e.g. duration, coagulopathy and single- or multi- organ failure, COVID-19 also shares many common features with viral sepsis at the clinical and immunopathological level. Li et al. (2020) reported that many severe or critically ill COVID-19 patients developed typical clinical manifestations of shock (although in the absence of overt hypotension), severe metabolic acidosis, disseminated intravascular coagulation, and impaired liver and kidney function in addition to lung injury, thus fulfilling the diagnostic criteria for sepsis clinically but whether this was caused by a dysregulated host response remain to be verified. Other characteristics shared with sepsis of respiratory origin included diffuse alveolar damage, increased pulmonary inflammation, altered mental state, low blood pressure, and high lactate (Beltrán-García et al. 2020). Giamarellos-Bourboulis et al. (2020) reported that COVID-19 patients with severe respiratory failure showed patterns of immune dysregulation including low HLA-DR expression, depletion of CD4 and CD19 lymphocytes and NK cells, and macrophage activation syndrome (a hyperinflammatory condition also known as secondary hemophagocytic lymphohistocytosis). Arunachalam et al. (2020) observed significantly higher bacterial DNA and LPS in plasma from severe and critically ill COVID-19 patients along with increased levels of pro-inflammatory mediators, and impaired mTOR signalling, suggesting a sepsis-like condition.

Therefore, acknowledging that sepsis is a multifactorial disorder, SARS-CoV-2 infection can be considered as a specific cause of sepsis (Shappell et al. 2020). In other words, the dysregulated immune response to SARS-CoV-2 infection at least partially mediates the organ failure and adverse outcome in severe COVID-19 cases. Therefore, approaches and implications from multi-omics studies in the host response to sepsis could potentially be applied to understand the heterogeneity in COVID-19 response.

In this perspective, Sweeney et al. (2021) considered severe COVID-19 as viral sepsis and reported that the three transcriptomic endotypes they defined in bacterial sepsis also recapitulated the immune phenotypes in COVID-19. Reyes et al. (2020) reported an expanded, immunosuppressive, CD14+ monocyte state from bacterial sepsis, similar to monocytic myeloid-derived suppressor cells. They showed that expression of this program is also up-regulated in monocytes from severe COVID-19 patients, and can be induced in healthy bone marrow progenitor cells in vitro by blood plasma from bacterial sepsis or severe COVID-19 patients in an IL-6 and IL-10-dependent way (Reyes et al. 2021). Therefore, the expansion of suppressive myeloid cells may play an important role in response to both sepsis and severe COVID-19.

The COVID-19 Multi-omics Blood ATlas (COMBAT) consortium

To form a rapid response in understanding the disease, the COvid-19 Multi-omics Blood ATlas (COMBAT) project was initiated in March 2020. The COMBAT project focused mainly on the question of why a subset of patients with COVID-19 develop severe illness and whether better targeted therapies can be informed. In a large collaborative effort, work by the consortium delineated the underlying immune dysfunction through multimodal deep phenotyping of COVID-19 patients of different disease severity and progression stage, with comparator groups including healthy controls, flu patients, and non-COVID-19 sepsis patients. The cellular and molecular hallmarks of disease severity, the proteome-based subphenotypes, and the feature groupings revealed by integrative analysis have been described in detail by the consortium (COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium 2022). Work described in Chapter 5 of this thesis has contributed to understanding the alterations in the plasma proteome part of the COMBAT consortium.

1.8 Aims and objectives

The aims and specific objectives of this thesis are to:

1. **Describe the sepsis-specific plasma proteome (Chapter 3)**

   I aim to:

   (a) optimise the workflow for pre-processing the large-scale clinical mass spectrometry dataset

   (b) define the sepsis response proteome through comparison with healthy controls and sterile inflammation including patients following surgery and with critical illness

   (c) understand the plasma proteome network in inflammatory conditions

   (d) elucidate the association between plasma proteins and clinical phenotypes

2. **Investigate heterogeneity in the individual sepsis response at the plasma proteome level (Chapter 4)**

   I aim to:

   (a) elucidate how the sepsis response proteome varies between patients by identifying patient subtypes

   (b) assess the clinical significance of the proteomic subtypes as a new way of understanding heterogeneity in sepsis

   (c) describe the molecular characteristics of the subtypes and the indicated pathways

   (d) understand how the proteomic subtypes relate to transcriptomic endotypes and change over time

   (e) validate the subtypes in independent samples or cohorts

   (f) build prediction models for the subtypes to assess whether they can be distinguished with a large or minimal panel of protein biomarkers

3. **Characterise the host response proteome of COVID-19 (Chapter 5)**

   I aim to:

(a) understand changes in inflammatory and immune states during severe SARS-CoV-2 infection from the standpoint of the plasma proteome

(b) compare between patients with different levels of severity

(c) compare between the blood proteome response to COVID-19, to severe influenza infection, and to all-cause sepsis

2

Methods

This chapter describes the patient cohorts, experimental procedures, and statistical methods used in this thesis.

2.1 Patient recruitment

2.1.1 UK Genomic Advances in Sepsis

The Genomic Advances in Sepsis (GAinS) Study was an observational study of patients admitted to adult ICU with sepsis due to either community acquired pneumonia or faecal peritonitis, as previously described (Davenport et al. 2016; Rautanen et al. 2015). The study was initiated in 2005 by the UK Critical Care Genomics group, initially recruiting adult patients (>18yr) in 34 participating UK ICUs, four of which were still recruiting when I joined the study in 2017 until the recruitment was concluded in 2020 (Figure 2.1). Ethics approval was granted nationally (REC Reference Number 05/MRE00/38 and 08/H0505/78) and for individual participating centres. Written, informed consent was obtained from all patients or a legal representative.

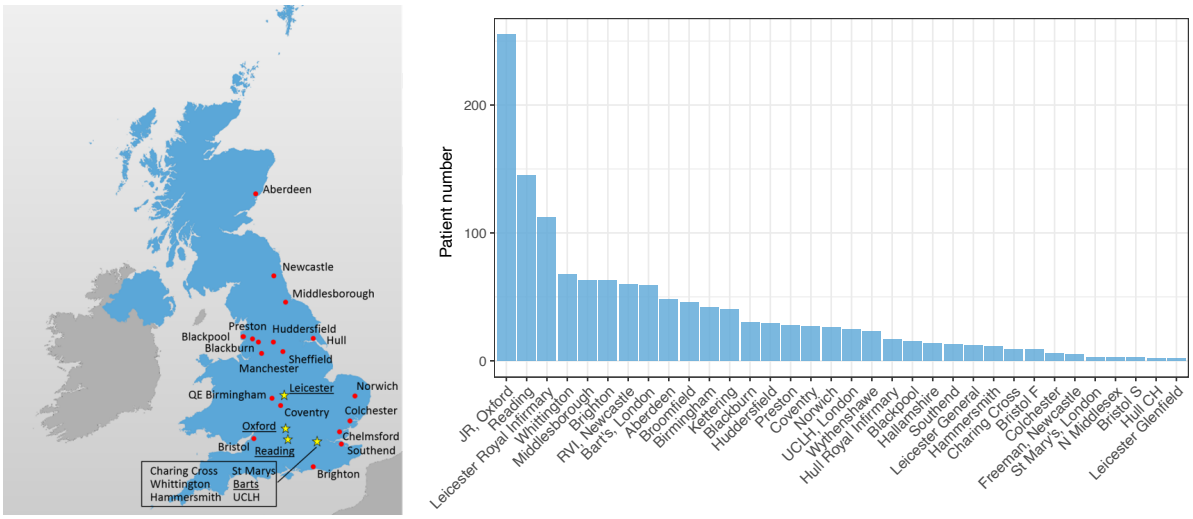


Figure 2.1: GAinS recruitment sites. Map shows the 22 cites for 34 ICUs across the UK that collected serial samples for GAinS. Four ICUs still actively recruiting in the year 2017 were underlined and indicated in yellow. Barplot shows the numbers of patients recruited at each ICU after recruitment was concluded.

In the GAinS study, sepsis was diagnosed according to the ACCP/SCCM guidelines (Sepsis-2; the 2001 American College of Chest Physicians / Society of Critical Care Medicine international sepsis definitions conference; Levy et al. 2003) as infection with

signs of systemic inflammation with all patients showing organ dysfunction during ICU admission. Community acquired pneumonia (CAP) was defined as a febrile illness associated with a cough, sputum production, breathlessness, leukocytosis and radiological features of pneumonia which was acquired in the community or within two days of hospital admission (Angus et al. 2002; Walden et al. 2014). Faecal peritonitis (FP) was diagnosed at laparotomy as inflammation of the peritoneal membrane secondary to large bowel perforation and faecal contamination (Tridente et al. 2014). Exclusion criteria included: patient or legal representative unwilling or unable to give consent; patient <18 years of age; pregnancy; an advanced directive to withhold or withdraw life sustaining treatment; admission for palliative care only; or immune-compromise (Davenport et al. 2016).

More than 1300 patients have been recruited with serial samples taken for functional genomics analysis. Samples for DNA extraction were collected for an additional >1000 patients. Serial samples (blood leukocytes, plasma, and urine) were collected by the research nurses on the first, third, and fifth day of ICU admission where possible. For the plasma component used for quantitative proteomics, blood was collected into one 5mL EDTA vacutainer. The tube was inverted gently 10 times, and then centrifuged at 1600RCF for 10min at 4°C. Four 500µL aliquots of the plasma layer was then transferred into cryotubes and stored at -20°C until being transferred to Oxford in batches and stored at -80°C.

A variety of genomic and functional genomic data have been generated for overlapping subsets of the GAinS patients, including genome-wide genotyping (Rautanen et al. 2015), exome sequencing, microarray and RNA-seq -based gene expression (Davenport et al. 2016; Burnham et al. 2017), microRNA, and metagenomics Goh et al. 2019.

Clinical phenotyping

Comprehensive demographic and clinical phenotypic data including treatments and outcomes were recorded in electronic case report forms (eCRF). Some of the clinical

measurements including total white cell counts, blood pressure, heart rate, blood creatinine and bilirubin were recorded as the highest and lowest value measured on the corresponding day. Outcome in terms of death or survival was followed up for 6 months following ICU admission.

The Acute Physiology and Chronic Health Evaluation II (APACHE II) scores on admission and the Sequential Organ Failure Assessment (SOFA) scores on Day1/3/5 were calculated from data on eCRF by the GAinS investigators. Septic shock was defined as hypotension despite adequate intravenous fluid resuscitation (Sepsis-2), so any patient-timepoint with vasopressors or inotropes administered, or with the lowest mean arterial pressure <65mmHg was classified as septic shock. Acute respiratory distress syndrome (ARDS) phenotypes were assigned according to the Berlin definition (Ferguson et al. 2012). Microbiology information for CAP patients were determined by Cyndi Goh, considering results from both clinical microbiology methods and metagenomics sequencing (Goh et al. 2019).

2.1.2 Covid-19 Multi-omics Blood Atlas

The COvid-19 Multi-omics Blood ATlas (COMBAT) study was designed for deep molecular, multi-omic and immunological profiling of COVID-19 in peripheral blood, comparing between varying severities of COVID-19 and comparator disease or health states. Patients admitted to Oxford University Hospitals NHS Foundation Trust were recruited through the Sepsis Immunomics study if they showed a syndrome consistent with COVID-19 and a positive RT-PCR test for SARS-CoV-2 from an upper respiratory tract (nose/throat) swab. Patients were recruited during the pandemic between 13th March and 28th April 2020 and had whole blood sampled on days 1, 3 and 5 of either hospital or ICU admission where possible. Ethical approval was given by the South Central Oxford C Research Ethics Committee (REC) in England for Sepsis Immunomics (REC reference 19/SC/0296).

The Sepsis Immunomics study is a prospective observational cohort study applying

an integrated immune-omic approach to understand why some patients have a severe response to infection. Patients were recruited either from ICU if they had symptoms and signs of established sepsis (suspected infection with an acute change in total SOFA score of ⩾2 points), or from the emergency department and medical wards if they had a change in quick SOFA score by ⩾2 points and a NEWS2 score of ⩾7 or intensive care review requested. Exclusion criteria included: patients or consultees unwilling or unable to give consent or advice declaration; advanced directive to withhold or withdraw life sustaining treatment; admission for palliative care only; pregnancy and 6 weeks post-partum; or severe acquired immunodeficiency (Kwok et al. 2022).

In addition to the hospitalised COVID-19 patients, four comparator groups were included: (1) patients with all-cause sepsis from sources other than COVID-19 recruited to Sepsis Immunomics prior to the pandemic (hospitalised encompassing severe and critical disease); (2) healthcare workers with symptoms consistent with mild COVID-19 and a positive test for SARS-CoV-2 were recruited and sampled at or after 7 days from the start of symptoms; (3) volunteers 55 years or over and self-reporting as healthy; (4) patients critically unwell with influenza (managed in ICU for ⩾24hr requiring ventilator support); (5) patients with critical illness due to COVID-19 where serum samples available (managed in ICU for ⩾24hr requiring ventilator support). Patient groups (4) and (5) were recruited at one of three sites (St George's University Hospitals NHS Foundation Trust, Guy's and St Thomas' NHS Foundation Trust and King's College Hospital NHS Foundation Trust) with PCR diagnostics for influenza or SARS-CoV2 performed by accredited laboratories; patients were recruited through the Aspergillosis in patients with severe influenza (AspiFlu ISRCTN51287266) study (Wales REC 5, reference 19/WA/0310). More details on patient recruitment and clinical phenotyping have been described (COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium 2022).

2.1.3 Other cohorts

In addition to GAinS, samples from seven other studies were included in the MS2019 dataset described in Chapters 3 and 4 as non-sepsis comparator groups or a validation sepsis cohort.

VANISH (Vasopressin vs Norepinephrine as Initial Therapy in Septic Shock; REC reference 12/SC/0014) was a randomised clinical trial enrolling adult patients who had septic shock requiring vasopressors despite fluid resuscitation (Gordon et al. 2016). The study reported no significant difference in outcome between vasopressin and norepinephrine groups, or between hydrocortisone and placebo groups. Up to four timepoints following recruitment during the acute admission were included for each of the 45 VANISH patients assayed in MS2019. The baseline timepoint (TP0) was mostly taken within 6 hours after the onset of shock and before the study drugs were given. Time from diagnosis of shock to study drug1 administration was within 6hrs for 43 of the patients, and had a median of 4hrs. Samples were also taken at subsequent timepoints (TP1/2/3) roughly at 24hr/48hr/96hr following TP0, respectively.

The Oxford BioBank (OBB; REC reference 18/SC/0588) is a random, population-based biobank of healthy participants initially recruited between the ages of 30 and 50 years from the Oxfordshire general population (Karpe et al. 2018). Individuals with: previous diagnosis of myocardial infarction or heart failure currently on treatment; untreated malignancies; or other systemic ongoing disease, and pregnant women were excluded. One sample was included for each of the 152 OBB participants included in MS2019. To match the demographics of other cohorts as closely as possible, older OBB participants were selected.

BIONIC (Biomarker based Identification Of Nosocomial Infective Complications; REC reference 14/EM/1223) was a prospective observational study at the Royal London Hospital that recruited patients undergoing elective major abdominal surgeries encompassing colorectal, upper GI (gastrointestinal) and HPB (Hepato-Pancreato-

Biliary) surgeries. For each of the 43 patients included in MS2019, three samples were included, taken either immediately before induction of anaesthesia ("BIONIC_preop"), or 24hr post-operation ("BIONIC_24hr"), or 48hr post-operation ("BIONIC_48hr"). Ten patients also had a 2–6hr post-operation sample included.

XMIN (Acquired loss of cardiac vagal activity is associated with myocardial injury in patients undergoing non-cardiac surgery; REC reference 16/LO/0635) was an observational mechanistic cohort study at the University College London, recruiting adult patients undergoing major elective noncardiac surgeries including orthopaedic, upper GI and colorectal surgeries. Researchers of this study had examined whether serial measures of cardiac vagal dysfunction was associated with perioperative myocardial injury and noncardiac morbidity (May et al. 2019), and had measured microRNA concentrations in extracellular vesicles(May et al. 2020). Two samples for each of the 106 XMIN patients were included in MS2019, one taken pre-operation ("XMIN_0") and one taken within 24hr post-operation ("XMIN_1").

The MOTION (Methylnaltrexone for the Treatment of Opioid Induced Constipation and Gastrointestinal Stasis in Intensive Care Patients; REC reference 14/LO/2004) trial recruited adult ICU patients who were mechanically ventilated, receiving opioids and were constipated (Patel et al. 2020). For the 50 patients included in MS2019, reasons for ICU admission included non-operative medical causes (n=34) and emergency(n=13) or elective (n=3) operative procedures. Patients admitted to ICU for infection or respiratory causes were not included. One sample from each patient, taken before the study drug was given, was included for MS2019.

The MONOGRAM (REC reference 15/LO/0933) study was set up for diagnosing and monitoring infection in critically ill patients using metabolic and immunological signatures. Baseline samples taken within 48hr of ICU admission for 14 mechanically ventilated patients with SIRS but without identified infection were included in MS2019.

TACE (REC reference 10/H0709/77) is a general abbreviation for observational studies on the mechanisms of monocyte priming and tolerance in vitro and in vivo to ascertain TNF-α converting enzyme (TACE) activity and metabolic signatures of patients with direct and indirect acute lung injury (ALI). Adult patients with or at risk of ALI, admitted to ICU within Imperial College Healthcare NHS Trust, were recruited within 48hr of onset of ALI or intubation. More details were described in associated publications (O'Callaghan et al. 2015; Antcliffe et al. 2017; Antcliffe et al. 2018). The baseline samples on study entry for twelve patients with noninfectious SIRS (ten with brain injury, one with cardiac arrest, one with motor neurone disease) were included in MS2019.

2.2 Luminex assay for cytokine measurement in GAinS

2.2.1 The Luminex technology

To comprehensively profile circulating cytokines in the sepsis response as a complement of the mass spectrometry analysis. the Luminex xMap (multi-analyte profiling) technology is utilised. This is a bead-based multiplexed immunoassay system in a microplate format (Reslova et al. 2017). The beads are coded by different spectral addresses created by internally labelling the microspheres with different ratios of two fluorophores in the red and far-red wavelengths. Target-specific antibodies are covalently attached to the beads, giving each target a unique spectral address. For targets that are present in the sample, a sandwich is generated that is composed of the bead with attached target-specific antibody, the target molecule, the biotinylated target-specific antibody, and the streptavidin reporter linked to the biotin. Using a dual laser system in fluorescent flow cytometry or quantitative fluorescent microscopy, identity of the target is given by the spectral address of the bead and the intensity is given by the fluorescent label attached to streptavidin. Absolute concentration of the target can be derived by mapping to standard curves as in a

conventional ELISA (enzyme-linked immunosorbent assay). The largest preconfigured panel of Luminex assays available at the time the work was used, namely the Human ProcartaPlex^{TM} Immune Monitoring Panel (ThermoFischer Scientific, Waltham, MA, USA) which simultaneously targets 65 analytes of cytokines including chemokines, interferons, interleukins and growth factors, as well as MMP-1 which is a collagenase. For simplicity, these 65 analytes are referred to as "cytokines" or "inflammatory mediators" in this thesis.

2.2.2 The Luminex assay

Plasma samples from the GAinS study were measured across three 96-well plates. Each plate also included two blank wells and duplicates of seven gradient dilutions of standards. Samples were randomised between plates to minimise the potential influence of batch effect between the plates. Serial samples from the same patient were kept together. Only visually non-haemolytic plasma samples that had never been thawed before were used. Samples were prepared according to manufacturer's instructions. 25μL was pipetted for each sample into the plates. Capture beads were incubated overnight at 4°C. I acquired the data on a Luminex 100 system at the Kennedy Institute of Rheumatology (Oxford). Minimum counts triggering a warning message was set to 100 bead reads per bead region. The last four columns of the second plate were read twice (along with the standards) due to machine failure in the first run, thus were considered as a separate plate in quality control.

2.2.3 Quality control

Raw median fluorescence intensity (MFI) was acquired for non-replicated samples across 4 batches. To correct for background fluorescence, Median Fluorescence Intensity (MFI) of each analyte in each well was divided by the average MFI for the analyte in the two blank wells on the corresponding plate. Absolute concentration levels of measured analytes were determined by mapping the net MFI back to the

plate- and analyte- specific standard curves with R package nCal (Fong et al. 2013). Logarithmic transformed values were input to 5-parameter logistic curve fitting, with the standard points equally weighted. For each analyte, the upper or lower limit of quantification (ULOQ or LLOQ) was determined as the concentration of the least or most diluted standard that had a median backfit bias no larger than 30% (According to manufacturer's instructions). For most of the analytes, only small percentages of samples were outside the quantification ranges (11 analytes had >10%, none had >45%; Table A.1). For each analyte, an overlay of the four standard curves was examined to show that there was no batch effect and that the concentrations were mostly obtained from the linear quantifiable range, for example for IL-6 and IFN-γ (Fig. A.1). Sample concentrations beyond the LOQs cannot be precisely quantified and thus were replaced with LOQ values. Infinite values (n=4) generated in estimated concentrations were replaced by median values of the corresponding analytes.

Outlying samples of poor quality were identified by considering results from multiple quality control measurements, including bead counts for each spectral address, density distribution and boxplots per sample, hierarchical clustering, principal component analysis, and calculating sample correlations. Sample distribution on the first two principal components post-QC confirmed no further outliers and no batch effect across the four plates. In the post-QC dataset, 65 cytokine concentrations were available for 204 samples from 146 sepsis patients in GAinS.

2.3 Mass spectrometry

2.3.1 TimsTOF mass spectrometry

For the mass spectrometry dataset described in Chapters 3 and 4 (MS2019), the 2622 clinical plasma sample tubes from the 8 cohorts (including 1889 from GAinS) were randomised across 28 96-well plates on dry ice. Sample tubes for each plate were then thawed together at 4°C and 50μL of each sample aliquoted to the plate, which was

then stored at -80°C before being processed.

All samples were processed in one batch utilising a BravoAssaymap liquid handler robot (Agilent) and injected into LC-MS/MS acquisition in succession. QC plate pools were ran to monitor the performance of trypsin digestion in each plate. An additional pool was created at peptide level (QC system pools) and run every 24 samples to monitor the performance of liquid chromatography (LC) across the dataset. No affinity depletion was applied to the samples. A plasma library was generated combining top 64 depletion with high pH fractionation to increase the detection range of the proteins. Samples were analysed using the high-throughput Evosep One LC system connected to the TimsTOF (Trapped Ion Mobility Spectrometry-Time-of-Flight) Pro mass spectrometer (Bruker Daltonics). Peptides were analysed using the pre-built 100 samples per day method and acquisition ran in PASEF (Parallel accumulation-serial fragmentation) mode.

The mass spectra acquired was analysed by the Fragpipe (v13.0; Kong et al. 2017) pipeline. Data was searched against a fused target/decoy database generated by Philosopher (v3.2.9), consisting of human UniProt SwissProt sequences plus common contaminants. Label free quantitation (LFQ) was conducted with IonQuant and Match-Between-Runs enabled and using Top-3 quantitation. For matching the ions, ion, peptide and protein FDRs were relaxed to 0.1 and minimum correlation set to 0 to allow pre-fractionated library samples to be included.

The plasma and serum samples in COMBAT were also assayed using the Evosep One-TimsTOF Prot platform and followed a similar procedure as for the GAinS samples, except that the peptide database searched against also included UniProt SARS-CoV-2 sequences. Experimental details have been previously described (COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium 2022).

2.3.2 QE-HF mass spectrometry

192 plasma samples from 123 GAinS and non-septic control patients were assayed on a LC-MS/MS system comprised of a Dionex Ultimate 3000 nano-liquid chromatograph and a Thermo Q-Exactive HF tandem mass spectrometer. For this dataset, samples were randomised across the two plates and aliquoted by centrifugation for 5 min at 10,000 RCF and then transferring 50μL below the top lipid layer to the plate wells. Sample preparation involved delipidation, depletion of Top 12 most abundant proteins using the Pierce^{TM} Top 12 Abundant Protein Depletion Spin Columns (ThermoFisher), protein precipitation, digestion, desalting, drying, and peptide quantification. Concentrations were normalised across the samples and approximately 200ng of each sample were injected into the system with 1 hour gradient time. A pool of all the samples was created for data alignment and comparing performance consistency across the acquisition time. Spectra data were analysed using Progenesis QI software combined with PEAKS (v8.5) as a data search engine using standard settings. This dataset is referred to as "MS192" in this thesis.

2.4 Bioinformatic and statistical analysis

2.4.1 Pathway and network analysis

Pathway enrichment analysis was performed using the XGR R package (Fang et al. 2016) with annotations from Gene Ontology Biological Process (GOBP), Gene Ontology Cellular Component (GOCC), or the Reactome pathway database. Annotation items with a size between 5 and 2000 and a minimum overlap with input data of 5 were included for the enrichment tests, unless otherwise specified. Significantly enriched terms were defined by FDR <0.05 in hypergeometric tests with all proteins or genes detected (including in fractionated and depleted library samples for proteins) as the background.

Protein-protein interaction (PPI) data was retrieved from the STRING v11 database with a confidence score cut-off of 0.7 and zero additional interactors. The PPI network was constructed and visualized through the Cytoscape v3.8.2 platform (Shannon et al. 2003) using perfuse force directed layout. The main network was then divided into clusters by the Markov cluster algorithm applied in the "clusterMaker" plugin (Morris et al. 2011). In MS2019, proteins in each clusters were tested for GOBP enrichment, using a minimum overlap of three for the six smaller clusters.

The protein co-expression network was constructed using the WGCNA R package (Langfelder and Horvath 2008) and visualised using Cytoscape. The soft thresholding power of 3 was chosen based on the lowest power for which the scale-free topology fit index curve reaches a high value (>0.85). For network construction I used a signed hybrid type of network, unsigned type of topological overlap matrix, biweight mid-correlations, and a maximum 5% of samples that can be considered as outliers on either side of the median. The minimum module size was set to 5 considering the smaller number of features in proteomics compared with transcriptomics datasets.

In single sample gene set enrichment analysis (ssGSEA), the enrichment scores of genes sets were calculated based on walking down a list of proteins ranked by their intensity in each single sample. Through the ssGSEA projection the matrix of protein intensities of genes (proteins) in samples is converted to a matrix of enrichment scores of gene sets in samples. This projection was performed using the Gene Pattern platform (Reich et al. 2006), using GOBP annotations. Gene sets with less than 10 overlaps with the input proteins were excluded.

2.4.2 Statistical tests

All statistical analysis were performed in R (R Core Team 2015). Principal component analysis was performed using the prcomp R function without scaling. Differential abundance analysis for proteins and differential expression analysis for genes were performed by fitting the intensities in linear models using the limma R package

(Ritchie et al. 2015), using only the first available sample of each patient and including age and sex as covariates. In COMBAT, the one earliest sample at the maximal severity of each patient were used. The Benjamini-Hochberg procedure was used to adjust for multiple testing. Significance for downstream analysis was defined as FDR <0.05 and fold change (FC) >1.5 unless otherwise specified. Limma fits a linear model to the level of each gene/protein and applies an empirical Bayes smoothing to the standard errors of the estimated log-fold changes to account for the overall variance. Comparisons between the pre- and post- operation samples in MS2019 were paired and with no additional covariates. All tests were two-sided.

Cytokine concentrations measured by Luminex assay were compared by Wilcoxon rank-sum tests (i.e. Mann-Whitney tests) using the first available sample of each patient. Categorical clinical variables were compared using χ^2 tests without Yate's correction. Numerical clinical variables were compared between two groups using Wilcoxon rank-sum tests. For comparing numerical clinical variables between three groups, Kruskal–Wallis test (i.e. one-way analysis of variance on ranks) was used to determine whether there was any significant difference (FDR<0.05) between the groups. For variables where the null hypothesis in Kruskal–Wallis tests were rejected, Dunn's post-hoc tests were performed to compare between each pair of the three groups, the significance levels of which were labelled on boxplots after adjusting for multiple testing within the variable. SOFA scores and ARDS levels were considered as numerical variables. Only the first available samples of each patients were included in comparing the clinical variables.

Survival differences were assessed and Kaplan-Meier curves plotted using the R packages survival and survminer (Kassambara et al. 2021; Therneau and Grambsch 2000). The input data is a data frame specifying the time to event from the sampling day, the event (death or end of 28-day or 6-month observation) and the patient groups. For patients with multiple timepoints sampled, the last sample within five days of ICU admission was used. The p values were given by log-rank tests. Hazard ratios

and the confidence intervals were calculated using uni-variate or multi-variate Cox proportional hazard models.

2.4.3 Data visualisation

On all boxplots, hinges show the first and third quartiles on both sides of the median, and the whiskers extend to the highest or lowest values within 1.5×IQR of the hinges. Data points are plotted on horizontally jittered positions on the boxplots. On all volcano plots, proteins more abundant in Group A of the contrast "Group A vs B" are plotted on the right-hand side of the plot. Differentially abundant proteins (FDR<0.05 and |FC|>1.5) are plotted in red points. Trajectory of patients' movements between the clusters were visualized using the ggalluvial package (Brunson 2020), restricted to patients with a baseline (Day1) sample available. Width of the flows on the y-axis are scaled to the number of patients in each movement type.

2.4.4 Cluster identification and prediction

Unsupervised consensus clustering was applied using the ConsensusClusterPlus R package (Wilkerson and Hayes 2010). In each iteration, unsupervised hierarchical clustering was performed based on Euclidean distance for dissimilarity between samples and Ward's method as linkage for cluster agglomeration.

In Chapter 4, I tested eight models in total for ConC prediction using three statistical learning methods: partial least squares discriminative analysis (PLS-DA), generalized linear models (GLM), and class prediction by random forest through the R packages ropls, glmnet, caret and randomForest (Friedman et al. 2010; Thévenot et al. 2015; Kuhn 2021; Liaw and Wiener 2002). For lasso and ridge regressions, the value of λ was determined by minimizing the prediction error rate in 10-fold cross validation (CV) within the training set. When there was not a great increase in CV misclassification rate between the λ that gives the smallest error, and the largest λ such that the error is within the smallest error plus one standard error, the latter value of λ was chosen

so that the model was less heavily based on the training set. For elastic net regression ($0 < \alpha < 1$), the best values for λ and α were selected by testing a grid of 20×20 combinations, and selecting the combination that gives the highest prediction accuracy in cross validation.

3

A Large-scale Proteomic Atlas of Host Response in Sepsis and Sterile Inflammation

This chapter aims to define the sepsis response proteome through comparison with healthy controls and sterile inflammation including patients following surgery and with other critical illness.

3.1 Introduction

3.1.1 Limitation in existing sepsis proteomics studies

Clinical proteomics strategies have traditionally involved measuring with high depth in a few samples and validating in larger cohorts by targeted approaches. This would potentially miss out interesting targets in the discovery step due of lack of power. In sepsis proteomics studies using patient samples, several groups have compared sepsis to health or other inflammatory conditions to improve understanding of disease mechanisms and diagnosis at the proteome level (Kiehntopf et al. 2011; Sharma et al. 2017; Thavarajah et al. 2020; Tong et al. 2019). Another focus has been to identify biomarker candidates predictive of mortality (De Coux et al. 2015; Langley et al. 2013). However, technological limitations have restricted these studies to either small sample sizes (generally <100 in each comparator group, with the exception of Langley et al.) or focusing on selected proteins, mainly plasma cytokines (Fjell et al. 2013). Due to the highly heterogeneous pathophysiology in sepsis, and the wide range of possible

appropriate comparator groups, this has also precluded the resolution of variation in the individual sepsis response using a discovery approach.

The combination of rich clinical sample resources and high-throughput MS that will be described in this chapter would enable investigation of the sepsis blood proteome at the largest scale to date. This could ensure not only sufficient power in both the discovery and validation stage of the study but also the inclusion of a panel of appropriate controls to derive robust signals. This strategy is also advocated by Geyer et al. (2017) as a "rectangular strategy" where as many proteins as possible are measured for as many individuals and conditions as possible. This could lead to a higher confidence in revealing any protein patterns that differentiate between the known conditions and unknown subgroups of individuals.

3.1.2 Mass spectrometry data pre-processing

Understanding mass spectrometry output

In mass spectrometry (MS) based protein assays, after peptide spectra are acquired in a LC-MS/MS system, protein identities and quantifications are obtained and false discoveries controlled through computational platforms like Progenesis (Nonlinear Dynamics), PEAKS (Bioinformatics Solutions Inc.), MaxQuant (Tyanova et al. 2016a), or FragPipe (Kong et al. 2017). Quantifications can be reported at either proteins or protein groups level, as LFQ (label-free quantification) intensities or spectral counts. Taking protein-level intensity reports as an example, unique peptides are peptides unique to a certain protein. Razor peptides are peptides found in more than one protein identified. In 'razor' peptide assignment, the peptide is assigned to the protein with the larger number of identified peptides, and will only contribute to the level of one protein. Accordingly, the razor/unique/total intensities reported for each protein species is based on razor peptides, or unique peptides, or any peptides with sequence corresponding to this protein. Both razor intensity and unique intensity are commonly

used as the raw data in the form of a matrix with proteins as rows and samples as columns.

Before statistical analysis to infer any biological findings, pre-processing should be performed with steps typically including data filtering, visual inspection of protein distributions, normalisation of intensities, and imputation of missing values (Sinitcyn et al. 2018). The Perseus platform developed by Tyanova et al. (2016) is widely applied for this aim but is designed for smaller scale application and lacks customisation in pre-processing, analysis, and visualisation.

In addition to the non-complete detections which are common to most mass spectrometry datasets, the heterogeneous sample composition and long acquisition process in this experiment described in this chapter have introduced more variation in sample conditions and acquisition conditions, which need to be carefully accounted for during data pre-processing. In addition to the basic steps as applied in Perseus, for MS2019 I customised each step of pre-processing considering the nature of the data, taking into account other potential technical bias, and adapting applicable algorithms developed for transcriptomics analysis.

Normalisation

Mass-spectrometry based proteome measurements need to be normalised to account for systematic bias from non-biological variation and make samples more comparable. Normalising to average intensity is a simplistic and the most widely applied method in platforms like Perseus but this is not necessarily the best approach.

Using spike-in and experimental datasets, Valikangas et al. (2016) systematically evaluated seven normalisation methods in quantitative label-free proteomics, including linear regression, local regression, Variance Stabilizing Normalization (VSN), quantile normalisation, median normalisation, Progenesis, and EigenMS. They reported that VSN reduced variation the most between technical replicates and also performed consistently well in the differential expression analysis. Karp

et al. (2010) demonstrated that VSN is able to address the variance heterogeneity in iTRAQ (isobaric tags for relative or absolute quantitation), and suggested that VSN could increase the capabilities of other mass spectrometry based quantitations as well. I applied the R package Normalyzer (Chawade et al. 2014) on MS192 as a smaller sepsis mass-spectormetry dataset (described in Section 4.4.1) to compare between common methods including linear regression, local regression, total intensity, average intensity, median intensity, VSN, and quantile normalisation, using log_2 intensities as a baseline. I also found that VSN reduced intragroup variation to the largest extent, improved correlation within group, and most effectively stabilised the variance in low-abundance proteins.

Therefore, I chose VSN as the most optimal approach to normalise MS-based proteomic datasets, and applied it to MS2019 raw intensities in the result section 3.3.5.

Missingness and imputation

Proteome profiling with affinity-based assays like SOMAscan or Olink most often use the coefficients of variation (CV) or the limits of detection (LOD) as quality control for the variables measured (Sun et al. 2018; Filbin et al. 2020). Unlike in these assays, in mass spectrometry based proteome profiling it is the nature of technology to have certain levels of missingness in the proteins detected. For example, Messner et al. (2020) measured 245 uniquely identified proteins with 87% data completeness and at least five peptides, over 409 acquisitions. It is common practice to either set a detection number threshold to filter proteins, or in smaller sample sets to keep all proteins detected (Shen et al. 2020; Nie et al. 2021).

Proteins could be either missing not at random (MNAR) i.e. the lower abundance below a certain detection limit is missing; or missing not completely at random (MNCAR) i.e. proteins with lower abundance are less likely to be detected but there is also randomness derived from sampling ionised peptides from MS1 to enter the collision cell. Accordingly, the missing values in the data matrix can be imputed by

the minimal value measured, or by zero, or by random-draw from a down-shifted distribution (Rieckmann et al. 2017; Geyer et al. 2016b; Coscia et al. 2016). The last approach simulates a low-level distribution and is widely applied including being the default method in Perseus (Tyanova et al. 2016b). In cases where random effects including technical factors instead of low-abundance play a larger role in missingness, these approaches to replace the missing value will be less ideal since sample similarities or protein similarities in the downstream analysis will be identified based on the random missingness resulted from the high similarity in imputed values.

An alternative non-parametric algorithm called k-nearest-neighbours (KNN) to replace missing values is applicable regardless of the assumptions on the cause of missingness, as long as there is a relationship between the variable to be imputed and the other variables. Co-expression relationships is a sensible assumption to make in gene expression or protein abundance measurements. For a protein with missing values to be imputed, KNN finds the k most similar proteins (neighbours) based on the non-missing values, and fills in the missing value by the mean observed for the neighbour proteins in this sample (Hastie et al. 2021). It is most suitable for sparse matrix that is also not too sparse.

3.1.3 Aims

The overall aim of this chapter is to characterise the sepsis-specific plasma proteome using large-scale mass spectrometry profiling of cohorts of sepsis patients and related conditions, with the hypothesis that plasma proteomic features differentiate the host response to sepsis from control conditions and correlate with sepsis severity. Specifically, I will:

1. pre-process the large-scale clinical mass spectrometry dataset by developing a customised workflow
2. construct the plasma proteome interaction or co-expression network in

inflammatory conditions

3. elucidate the association between plasma proteins and clinical phenotypes
4. identify proteins that are specifically differentially abundant in sepsis through comparison with non-sepsis groups including healthy volunteers, patients following major surgery, and with critical illness other than sepsis
5. identify key pathways at the plasma proteome level that differentiates the sepsis response from related conditions
6. validate the findings in a validation cohort with independent samples

3.2 Results: Patient cohorts, power and design

3.2.1 Patient and control cohorts

In understanding the host response in sepsis, there has not been a consensus on what a gold-standard control cohort should comprise. The pre-sepsis samples of the same patients (prior to infection and/or before evidence of organ dysfunction) would be ideal, however, this is not clinically practical since it cannot typically be effectively predicted which subset of patients are going to develop sepsis. To address the issue in this mass spectrometry study, instead of including a single control group, a panel of non-sepsis controls that comprised healthy volunteers (OBB), patients undergoing major surgery(XMIN and BIONIC), and ICU patients without identified infection (MOTION, TACE, MONOGRAM) were included (Table 3.1). These different phenotypes (Fig. 3.1) enable consideration of the "sepsis-specific" response to be distinguished from the potential effect of differences in demographics, co-morbidities, sterile inflammation caused by surgery or injury, and individual differences.

This plasma proteomics dataset in sepsis and controls is referred to as "MS2019" since the mass spectra acquisition was performed in the year 2019. A total of 2622 samples from 1612 individuals were assayed in this experiment, composed as shown

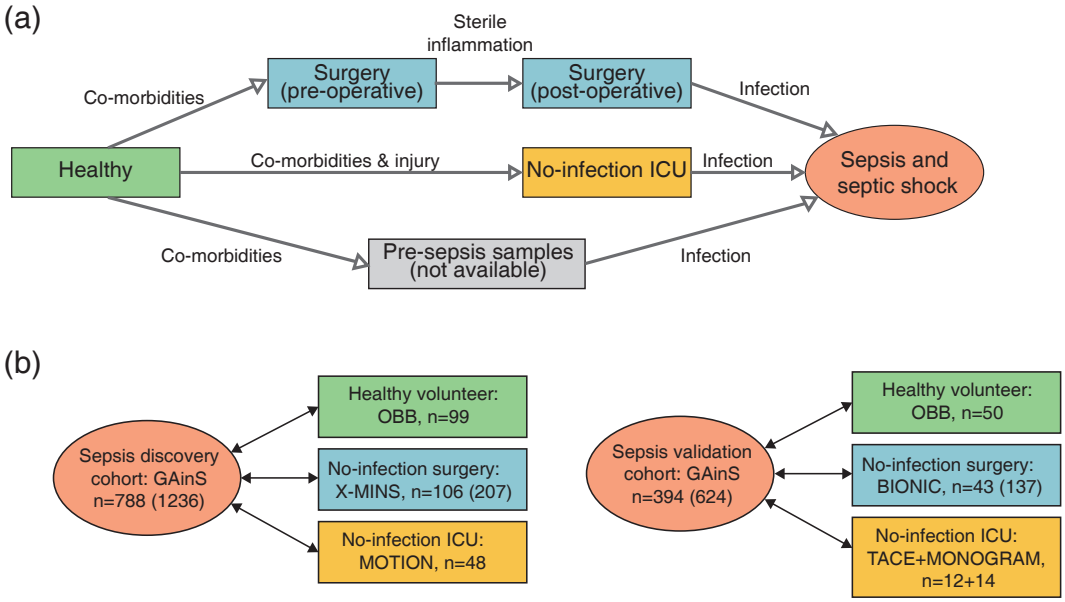


Figure 3.1: The sepsis patients and control cohorts. (a) An illustration of the relation between the sepsis and control groups. Arrows indicate a putative direction of increasing illness driven by the factors annotated, with no indication of disease progression. Samples from each of these groups were included in MS2019 except that no pre-sepsis sample was available for the sepsis patients. **(b)** Pairs of comparator groups that were tested in differential abundance analysis. Arrows indicate the group comparisons. The post-operative surgery samples and the no-infection ICU samples represent a sterile inflammation state that has the key difference with sepsis samples of having no identified infection. Post-QC instead of pre-QC patient and sample numbers are shown to reflect the numbers used in group contrasts. The numbers correspond to Table 3.5. Sample numbers were labelled in parentheses where more than one sample per patient were in the dataset.

Table 3.1: MS2019 cohort composition before quality control. Table shows the study names, general types of cohort, number of patients, number of samples, anti-coagulants that samples in each study were collected with, and the timepoints of sampling. Each study cohort and the day or timepoints of the samples are described in Methods. Nine samples in GAinS were assayed in duplicates. Abbreviations: ICU: intensive care unit; TP: timepoint; pre-op: pre-operation; 2-6hr/24hr/48hr: 2-6hr/24hr/48hr post-operation; VANISH: Vasopressin vs Norepinephrine as Initial Therapy in Septic Shock; OBB: Oxford BioBank; BIONIC: Biomarker based Identification Of Nosocomial Infective Complications; XMIN: Acquired loss of cardiac vagal activity is associated with myocardial injury in patients undergoing non-cardiac surgery; MOTION: Methylnaltrexone for the Treatment of Opioid Induced Constipation and Gastrointestinal Stasis in Intensive Care Patients; MONOGRAM: Diagnosing and monitoring infection in critically ill patients using metabolic and immunological signatures; TACE: Mechanisms of monocyte priming and tolerance in vitro and in vivo involving TNF-α converting enzyme.

Study	Type	N(patient)	N(sample)	Sample collected with	Sampling timepoints
GAinS	ICU sepsis	1190	1889	EDTA	Day 1/3/5
VANISH	Septic shock clinical trial	45	154	EDTA	TP 0, 1, 2, 3
BIONIC	Abdominal surgery	43	139	citrate	Pre-op, 2-6hr, 24hr, 48hr
XMIN	Non-cardiac surgery	106	212	EDTA	TP 0, 1
TACE	Brain injury ICU	12	12	EDTA/heparin	-
MOTION	No-infection ICU	50	50	citrate	-
MONOGRAM	No-infection ICU	14	14	EDTA	-
OBB	Healthy volunteer	152	152	EDTA	-
Total		**1612**	**2622**		

in Table3.1. Among these, 1190 patients were included from GAinS, an observational study that recruited adult ICU patients with sepsis due to CAP or FP. In addition, sepsis patients were also included from VANISH, a clinical trial enrolling adult patients with septic shock that collected samples at baseline (TP0) and subsequent timepoints (TP1/2/3). More details on each cohort have been described in Methods.

Samples from these studies were divided into a discovery cohort and a non-overlapping validation cohort in downstream analysis in order to identify signals that can be verified. The GAinS patients and OBB patients were separated by a split of 2:1 with random draws while keeping serial samples from the same patient in the same discovery or validation cohort. Other than the sepsis or healthy samples, the discovery cohort also included the XMIN surgery samples and MOTION no-infection ICU samples. The validation cohort also included the BIONIC surgery samples and no-infection ICU samples from TACE and MONOGRAM. This separation and random split was performed on the cleaned up dataset after pre-processing, with the numbers

given in Section 3.3.6.

3.2.2 Statistical power calculation

Because of the limitations in throughput of proteomics technologies and in availability of clinical samples, clinical proteomics studies are often criticised for using too few biologically distinct samples to draw reliable conclusions. Identification or validation of protein biomarkers from mass spectrometry measurements relies on a good experimental design ensuring adequate statistical power. Therefore, it is essential to estimate the sample size required to detect differences at a specific statistical power and false discovery rate before performing a new experiment. To this aim, the relation between sample size and statistical power in one sepsis-control contrast and one within-sepsis contrast was calculated based on another smaller-scale sepsis proteomics dataset (MS192) that was generated on a Q Exactive HF LC-MS/MS system which was more commonly used than timsTOF. This dataset is described in more detail in Section 4.4.1. The calculation took four steps: (1) Estimate the proportion of variables with true differential abundance signal; (2) Select appropriate level of type I and type II error to control FDR; (3) Estimate the variance of proteins measured; (4) Calculate the curve between size and power for proteins with different levels of variance, the result from which is shown in Fig. 3.2. The equations and methods used, and estimations from each of the steps are detailed in the Appendix.

A fold change cut-off of 1.5 was used to be consistent with the group comparisons. For the sepsis-control contrast (CAP vs post-operation), a minimum of 171 biological replicates is needed in each group to detect difference at 80% power and FDR$\leqslant$0.05 significance, for 70% of the analytes. In the same conditions, 269 biological replicates is needed for the within-sepsis (SRS) contrast.

Compared with the actual group sizes in MS192, results of the sample size calculation showed that the study design is under-powered to detect 80% of the truely differential signals. The scale of MS192 is already not modest-sized in clinical proteomics studies

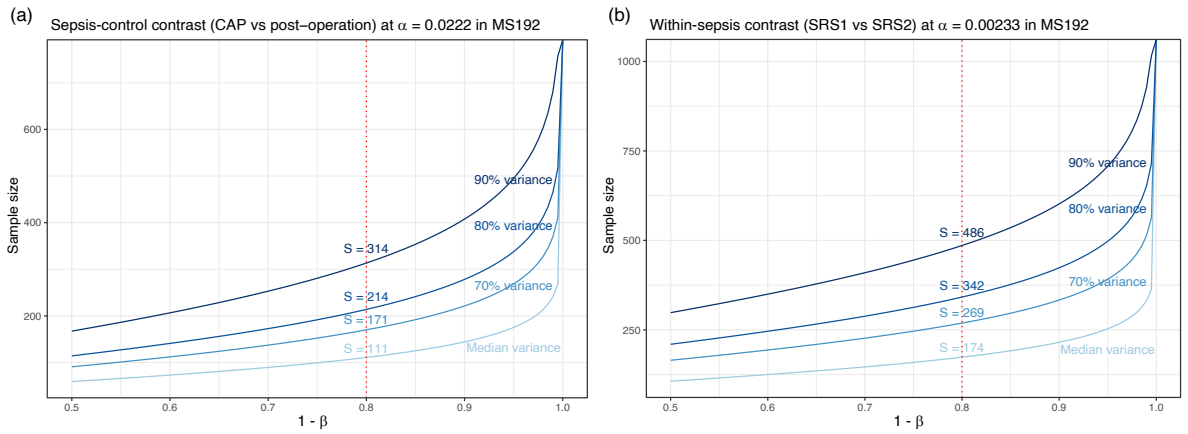


Figure 3.2: Relation of samples size and statistical power using α (type I error) values indicated from Fig. B.1 to control FDR $\leqslant$0.05. Separate curves are given for proteins with different variance at the percentiles stated, with more variable proteins at higher percentiles. The point of intersection between the blue curves and the red dotted line (1-β (power)=80%) denote the minimum sample size (S) required to achieve the desired power and significance level for 50%/70%/80%/90% of the analytes.

and is among the largest ones in existing reports on MS-based sepsis proteomics. This highlights the value of a large-scale dataset as MS2019 to detect most true signals with confidence, and to understand the more detailed structure in sepsis subtypes with sufficient statistical power. This also indicates that some of the sepsis-control contrasts in MS2019 (as in Table 3.1) could still be under-powered due to limited sizes in the appropriate control cohorts. As power calculation is largely affected by the technology platform to run the experiment and the biological nature of the samples (Cairns et al. 2009), these indications on sample size were also re-evaluated based on the contrasts of interest using MS2019 data in discussion Section 4.7.1.

3.3 Results: Data pre-processing

In addition to the 2647 injections from 2622 samples, mass spectra were acquired for a variety of quality control injections including pools and blanks, totalling to 4553 injections. No affinity depletion of the top-abundant proteins was applied to the samples. Mass spectra were acquired on a TimsTOF mass spectrometer, and processed using Fragpipe to generate the label-free quantifications. More details are described in

Methods. Figure 3.3 shows the steps of pre-processing applied to this dataset, which are detailed below.

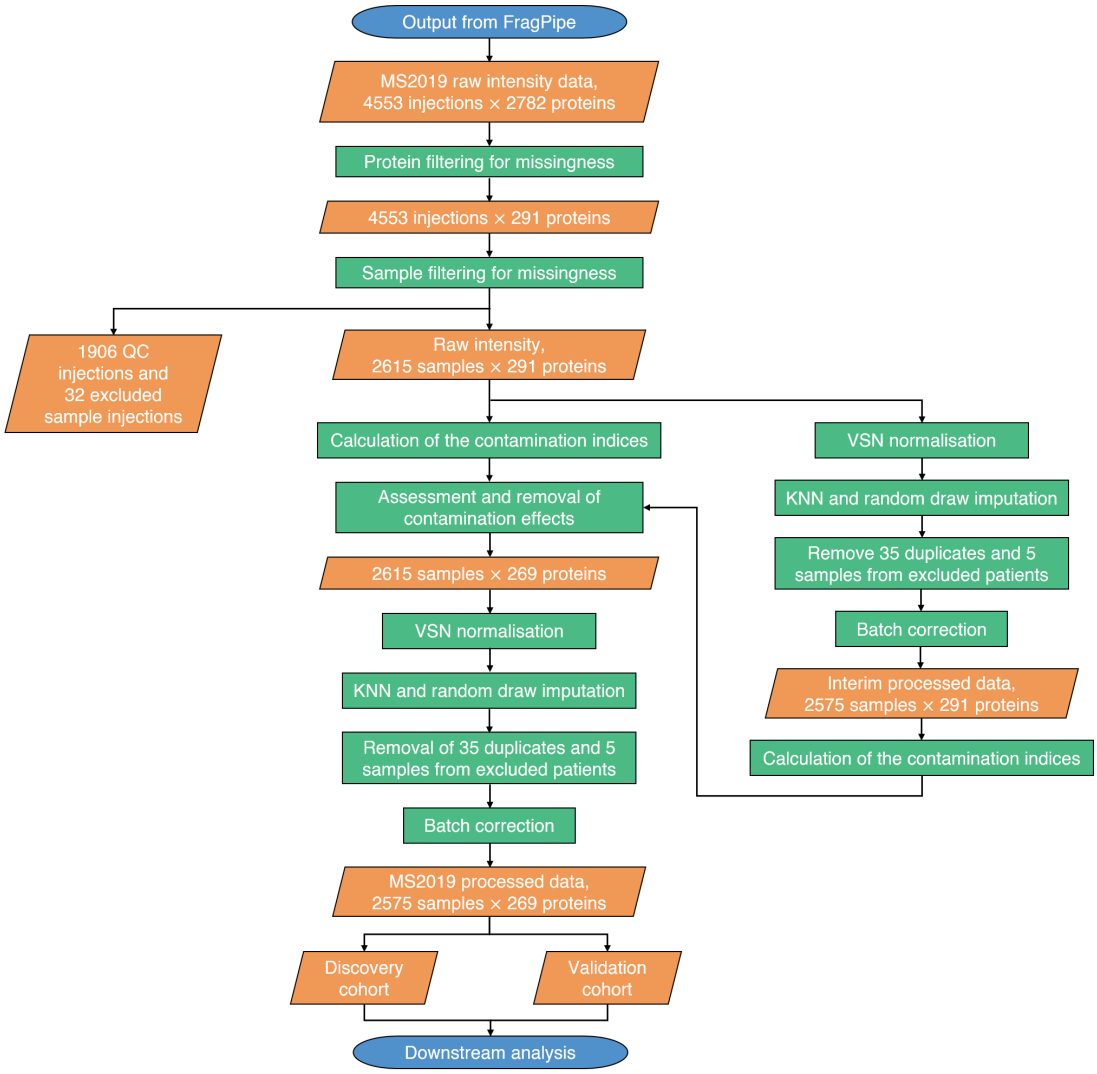


Figure 3.3: A flowchart of MS2019 pre-processing steps. The processes, data versions, and starting or terminating points are denoted in green, orange, and blue shapes, respectively. Abbreviations: VSN – Variance Stabilizing Normalization; KNN – k-nearest neighbours; QC injections – quality control injections, here referring to all the injections that did not come from an actual sample. Samples from five patients were excluded for withdrawal of consent or having no clinical information available.

3.3.1 Raw intensities

Starting from the Fragpipe output of 4553 injections, I subsetted and used razor intensities of 2782 proteins identified by at least one unique peptide. This was a sparse matrix with 92.7% values missing since the majority of proteins were only

detected in the fractionated libraries instead of in actual samples. Zeroes reported by Fragpipe were taken as missing values. Gene names were retrieved by mapping Accession IDs on the UniProt protein knowledgebase (Bateman et al. 2021). Histogram of log_2-transformed raw intensities showed an approximate normal distribution, with two smaller peaks at the lower- and higher- abundance end (Fig. 3.4). A long tail at the lower end of the distribution could be considered as random machine noise and removed, which was not observed in this dataset. In all the detected but not the missing values, the 0.5%, 50%, and 99.5% quantiles of the raw intensities were 1.2×10^3, 1.0×10^5, and 1.6×10^7, with a minimum intensity of 89. The difference in magnitude was similar from the median to either ends of the distribution, thus the low-end values were considered as true protein abundance measured instead of machine noise and thus were not removed.

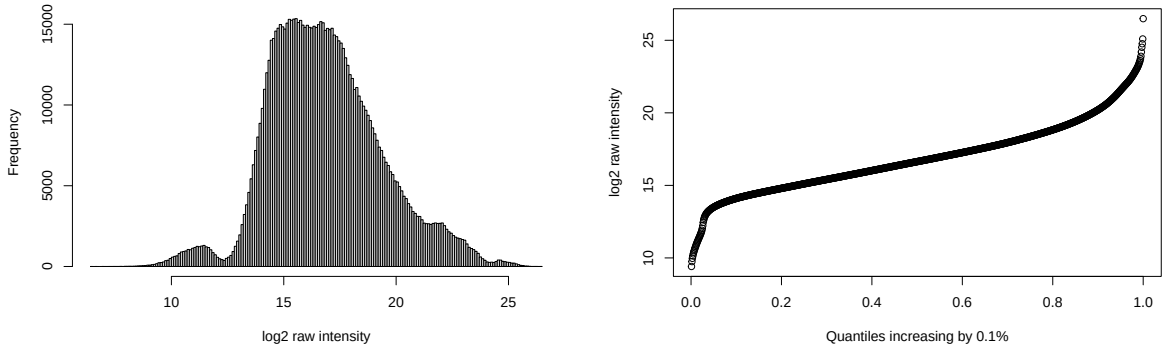


Figure 3.4: Histogram and quantiles of raw intensity after log_2 transformation, including all injections and non-filtered proteins.

3.3.2 Protein filtering

To generate a matrix of reliable measurements, proteins need to be reduced to a list that is convincingly detected in at least one of the biological groups. This was performed before filtering on samples so that sample quality was only inferred from the better detected proteins. One common approach for filtering gene expression probes is to keep probes that are detected in at least the number of the smallest biological group included. This approach is not biased by how sample subgroups

are defined. However, MS2019 is composed of cohorts with varying size, with the smallest group size of only 74 (the no-infection ICU group). This approach would keep 558 proteins but many of these were only sparsely detected across the groups and not specific to any of the groups. Imputed values for the very low-detected proteins will exhibit random variations and would add further noise but not signal to the dataset.

To keep the proteins that reach a detectable level in only certain biological conditions, detection was counted within each of the five biological groups including: the healthy volunteers of OBB; the pre-operation samples of XMIN and BIONIC; the post-operation samples of XMIN and BIONIC; the no-infection ICU samples of MOTION, MONOGRAM and TACE; the sepsis and septic shock samples of GAinS and VANISH. A total of 225 proteins were detected in at least 50% injections in at least one of the five groups. Among the cohorts, the GAinS study is the majority group constituting >70% samples. As sepsis is a highly heterogeneous disease there could be biological sample subgroups with distinct protein profiles. Filtering based on sepsis as one group could potentially miss out the proteins that really differentiate between these unknown subgroups. Therefore, subgroups were further defined within GAinS samples for calculating protein detection. A few approaches for defining the subgroups were considered.

One approach is to separate between sepsis patients based on clinical phenotypes including source of infection, timepoint after ICU admission, the presence of ARDS or septic shock, SOFA and APACHE scores, or clusters on clinical variables. However, the clinical phenotypes may not be sufficient for separating subgroups with underlying biological difference, which may be captured better by features at the molecular level, as showcased by the differentiated immune response captured by the SRS transcriptomic endotypes (Davenport et al. 2016; Burnham et al. 2017). I tested dividing GAinS samples into 6 severity groups by total SOFA scores, filtering in which groups added only 10 more proteins post-filtering.

The approach I applied was to define protein detection groups within the GAinS

samples by unsupervised clustering. For this aim, protein intensity or missingness in sample injections were converted to a binary state of detected (1) or missing (0) and input for the clustering. This approach will discover adequately-sized sample subsets with shared patterns in detection states for a group of proteins instead for individual proteins.

Consensus clustering was used to determine the protein detection groups since this method provides both evidence for selecting an optimal cluster number, and cluster results robust to outlying poor-quality samples or features. Only proteins with ⩾10 detections across all injections (n=1374) were input to the clustering. The consensus index matrices and the increase in area under the CDF (cumulative distribution function) curves showed clearly an optimal partitioning at a cluster number of 4 (Fig.B.3). After breaking up the sepsis groups into VANISH and the four GAinS detection groups, the number of proteins post-filtering increased to 291 with the threshold of detection in ⩾50% injections in at least one of the 9 groups. 138 proteins have ⩾50% detection in all the 10 groups plotted in the group detection heatmap Fig.3.5(a). There were 16/3/5/32 proteins unique (i.e. ⩾50% detection in only this group) to GAinS detection groups 1/2/3/4. This filtering-by-group strategy kept more proteins compared to setting one total detection threshold as lenient as 30% but not differentiating between cohorts (Fig.3.5(b)).

3.3.3 Sample filtering

After reducing proteins to the 291 more robustly measured species, sample-wise protein detection numbers and medians per sample were plotted to represent the quality across the injections (Fig.3.6). According to the lab record by Raphael Heilig, the chromatographic column was blocked and changed twice at injections 5105 and 7528. Injection 5105 right before plate 16 did correspond to a change in sample quality. The pools injected among the samples ("QC (system)") showed similar levels but less variation compared with the samples. The lowest detections were observed in

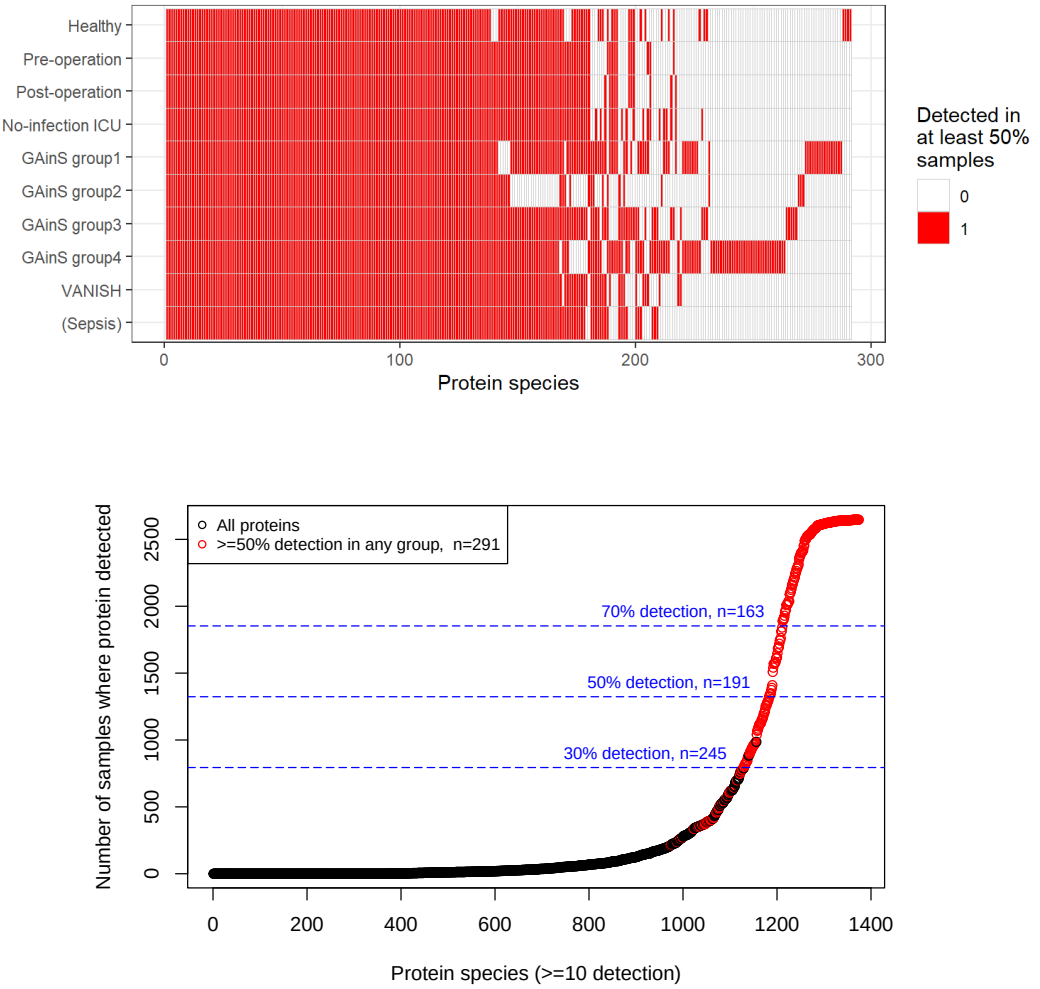


Figure 3.5: MS2019 protein filtering. **(a)** Heatmap showing whether the 291 proteins kept in the filtering had ⩾50% detection in each of the groups. The sepsis group as a whole is also plotted but not used in determining whether a protein is kept. **(b)** Number of sample injections where each protein was detected, labeling the 291 proteins kept and the numbers of proteins if using one detection threshold across all cohorts. Only protein species with ⩾10 injections were plotted.

plate 16 (injections 5163-5268). Samples separated into two groups between injections 6000 and 6600 without a clear mechanism. Plotting detection numbers in pre-filtering proteins showed similar patterns across the acquisition as with plotting post-filtering proteins here.

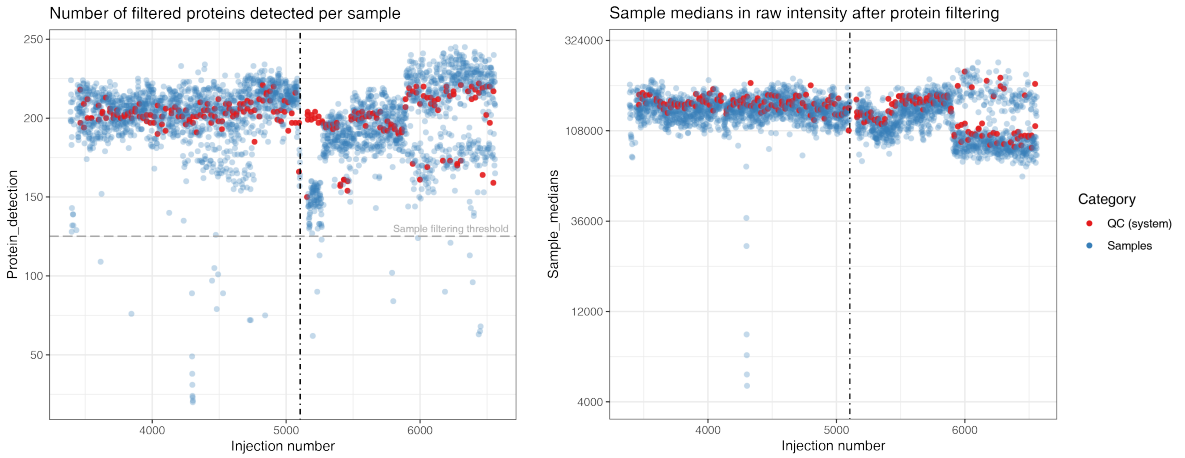


Figure 3.6: MS2019 sample filtering. Number of filtered proteins detected and the median intensities were plotted for each injection along the acquisition process, used as evidence for sample filtering. The injection numbers correspond to the machine record in a chronological order, starting at 3390 as the first injection of this experiment. Only the injection period containing the sample injections are plotted. The chromatographic column was blocked and changed at injection 5105 (vertical dash-dotted line). The horizontal grey dashed line on the detection numbers indicate the sample filtering threshold (detection$\geqslant$125) used.

As the variation was more clearly observed in the detection number, injections were filtered so that the ones with less than 125 (43% of 291) proteins detected were removed. This threshold was chosen as it was relatively clear from the scatter plot (Fig.3.6) and the histogram (Fig.B.4) that injections with even less detections were outlying. Most samples from plate 16 were kept as these were not outlying from the other samples in PCA performed on processed data. Among the injections removed there were 32 samples injections, leaving 2615 sample injections for further processing.

3.3.4 Correction for cell residue contaminations

After sample filtering, the data can be normalised, missing values imputed and batches corrected to give the "interim processed data", as shown in the flowchart (Fig. 3.3).

However, it is first necessary to assess the impact of some known technical effects. Although all plasma samples were isolated by centrifugation of blood samples taken with anticoagulants, samples from different studies recruited in a variety of clinical settings could potentially have differences in pre-analytical handling. Even following the same protocol in one study, in the interim processed data the GAinS samples, recruited over 15 years in 33 ICUs, showed a subset of 12% samples that were significantly higher in proteins including actins and actin-binding proteins that point to higher platelet residues in plasma. The variation in sample handling instead of in biological conditions of the patients confound the analysis. Therefore, biases caused by blood sample collection should be corrected for during pre-processing as much as possible.

Geyer et al. (2019) generated a mass-spectrometry based reference proteome measured in platelet-free plasma obtained after a 4-step centrifugation from healthy volunteer blood samples, and compared it against the common contamination cell types the platelet or the erythrocyte component, or serum samples after coagulation. They proposed three panels of marker proteins (n=29/29/12) to assess plasma sample contamination by erythrocyte lysis, platelet contamination, or partial coagulation. In the report, they calculated the erythrocyte index or the platelet index as the sum of the marker protein intensities (which are higher in the cellular components) divided by total intensity in the sample, and the coagulation index as the total intensity divided by the sum of the marker proteins (which are less abundant in serum). I applied this approach to quantify the three types of potential contamination in all samples in MS2019, by first taking the overlaps of the marker panels with our list of 291 proteins detected post-filtering, lists of which are shown in Table 3.2.

The contamination indices could be calculated based on either raw intensities or the interim processed data, but neither is optimal. The former is prone to inaccuracy from non-corrected batch effect (as will be discussed in section 3.3.6), while the latter is subject to bias during missing value imputation where the nearest neighbours could

Table 3.2: Table of marker proteins that overlap between the panels proposed by Geyer et al. and the filtered protein list in MS2019. These marker proteins were used for the calculation of the three contamination indices in MS2019.

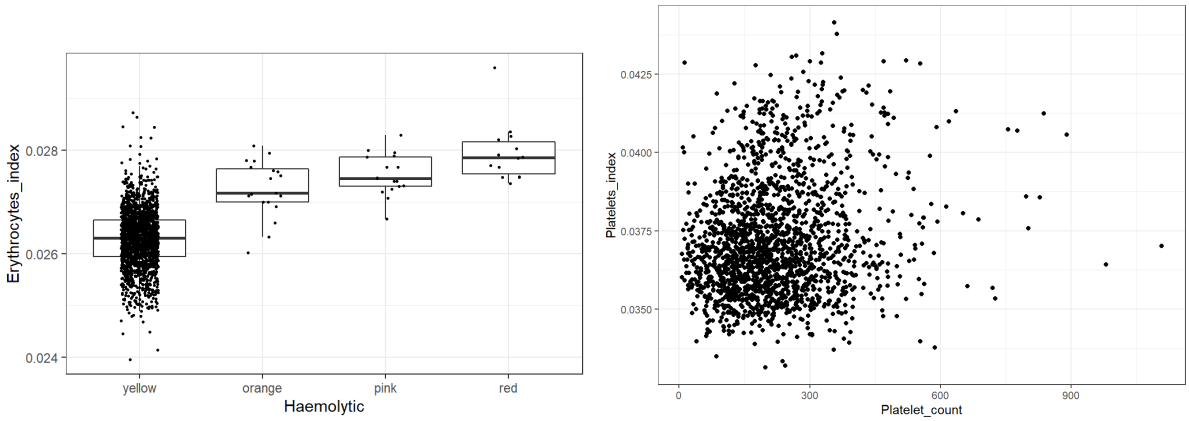
Contamination Index	Number of overlaps	Overlapping marker proteins
Erythrocyte index	8	ALDOA, GAPDH, LDHB, HBA1, ACTB, ACTG1, HBB, HBD
Platelet index	12	PPBP, GAPDH, GSN, PFN1, ACTN1, VCL, CFL1, TMSB4X, YWHAZ, FERMT3, TAGLN2, ACTB
Coagulation index	9	F13A1, F2, SERPINC1, APOC3, FGA, FGB, FGG, SERPINA5, ECM1

have been found based on sample similarity due to technical effects. I therefore calculated and considered both of them. There was significant correlation between the three contamination indices calculated in raw or in processed data. However, the correlation was much stronger in the platelet index than in the other two (Fig.B.5(a-c)), indicating that the other two may be more affected by batch effects. Batch effects may also drive the patterns observed when correlating the two cellular residue indices in raw data, which are absent in the interim processed data. (Fig.B.5(e-f)).

Taking the version from interim processed data as an example, the samples that were visually haemolytic i.e. were red/pink/orange in color were higher in erythrocyte index, although there was only a small number of haemolytic samples (Fig. 3.7(a)). The platelet index does not have a strong correlation with the clinically measured platelet counts on the day (Fig.3.7(b), Pearson's r=0.13), indicating that the variation in platelet index was not due to a difference in patient condition.

Comparing across the studies included, the erythrocyte index and platelet index had the highest levels in OBB and a subset of the GAinS samples (Fig.3.8). The outlying GAinS samples would likely form a distinct cluster driven by those technical effects. With the most strict blood collection protocol, XMIN samples had the lowest level and least variation in the platelet index. EDTA plasma was reported to be significantly higher in erythrocyte-specific proteins than heparin or citrate plasma (Geyer et al. 2019), which may explain why the three citrate groups (BIONIC pre-/post-operation and MOTION) showed a similar level of erythrocyte index as XMIN.

The coagulation index also spanned a wide range in GAinS, but with less variation



(a) (b)

Figure 3.7: Correlation of contamination indices with sample colors (a) or clinically measured platelet counts on the day (b). The visually-inspected sample colours (yellow/orange/pink/red) reflect whether samples are haemolytic. Only GAinS samples were included in both plots. One sample with platelet count of 3385 was omitted in plotting.

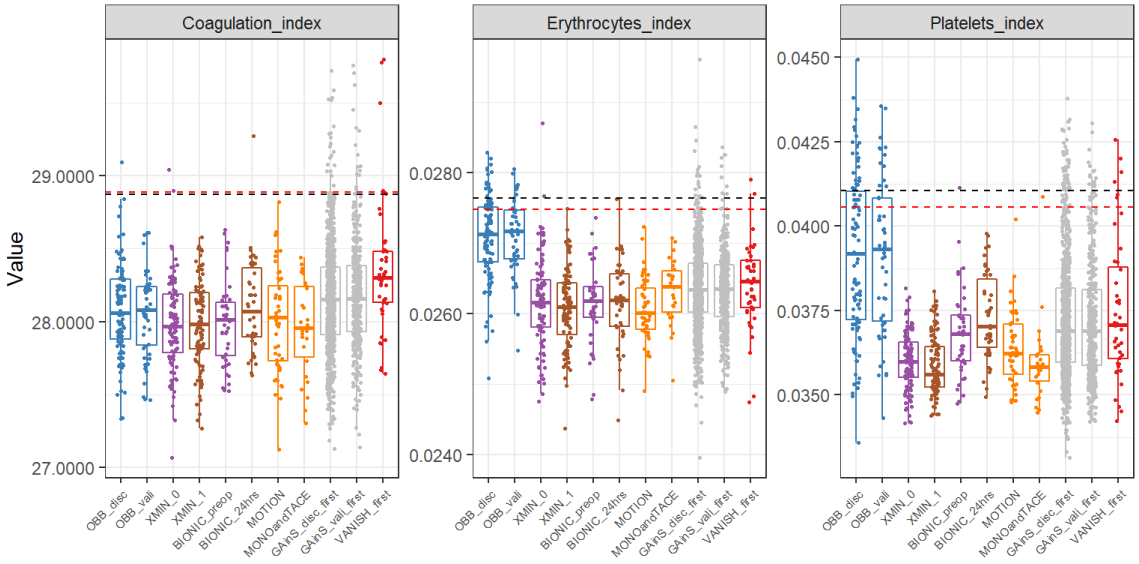


Figure 3.8: Boxplots of the three contamination indices across the studies. Contamination indices were calculated in processed data version where the contaminations were not corrected. For patients with serial sample points, only the first available sample was used. Red dashed lines indicate the sample outlier cut-off (mean + 2 s.d.) calculated in all MS2019 samples, excluding OBB in the erythrocyte index and platelet index. Black dashed lines were calculated in the plotted samples.

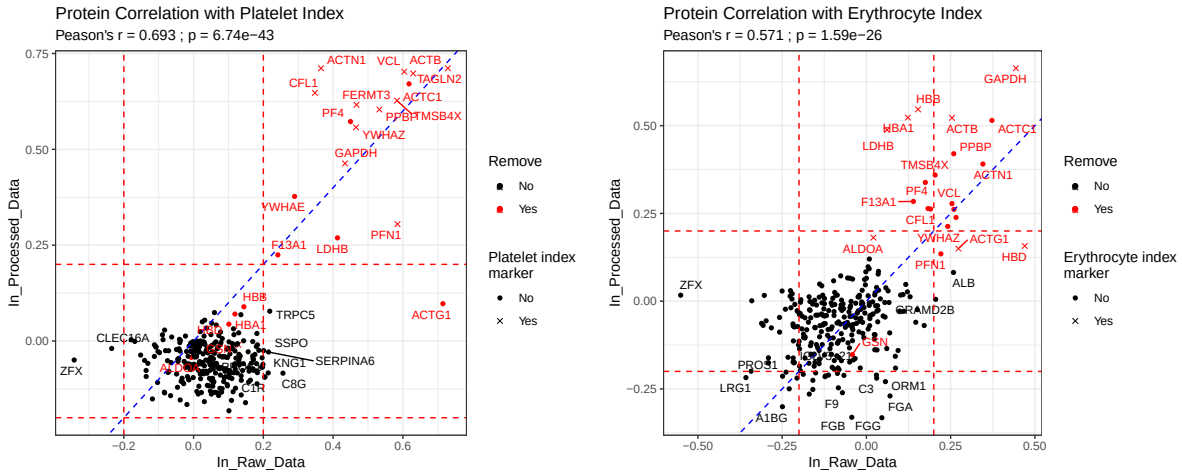
across the non-sepsis studies. As dysregulated thrombosis is part of sepsis pathophysiology and that fibrinogens are known to change during inflammatory conditions, the difference in coagulation index was taken as true signal and not requiring correction. The median level of the coagulation index was indeed higher in sepsis samples and highest in VANISH septic shock samples (Fig.3.8).

Comparing across the GAinS recruitment sites, the pattern of erythrocyte index or platelet index across the sites was similar, with more variation in the latter (Fig.B.6). There were higher values and more outliers in sites UK14, UK30, and UK43. This supports that the difference in the marker proteins were related to a difference in sample handling despite the same protocol was used across the sites.

Because sample quality indicated from the contamination indices showed a spectrum instead of having clear outliers, it is questionable at what extent the quality-associated proteins can be reliably interpreted across the dataset. Therefore, I corrected for the cell residue contaminations by identifying and removing the affected proteins from the dataset. This needs to be performed before the normalisation or imputation, so that the contamination proteins would not skew the distribution. The GAinS detection groups used for protein filtering were not driven by the contamination proteins, since proteins unique to any of the detection groups did not associate with the contamination proteins, and that the 12% high-contamination samples identified from the interim processed data were not enriched in any of the detection groups. Therefore, the identification and removal of the contamination proteins could be performed on the cleaner data matrix obtained after protein filtering.

Proteins affected by platelet residues should be co-regulated and have a high correlation with each other, which is the same for proteins affected by erythrocyte residues. Therefore, additional proteins affected by contamination could be identified by their correlation to the marker proteins, or to the contamination index as a whole. I calculated the correlations between individual protein levels and the two contamination indices both in raw log_2 intensity and in the interim processed

data. Correlations calculated on all samples were strongly correlated with those calculated on a single biological group, the GAinS study (Pearson's r >0.98 for both indices). Based on correlations on all samples, I removed any proteins that either had consistently high correlation with one of the indices in both data versions, or was one of the marker proteins (Fig.3.9). According to these criteria, there were 16 proteins with Pearson's correlation coefficient (PCC) >0.20 with the platelet index in both raw and interim processed data (Table 3.3); the 10 proteins with high correlations with erythrocyte index were included in these 16. The cut-off of 0.20 was chosen for being a clear cut-off in the distribution of the correlations (Fig.B.7). There were 6 proteins with PCC lower than -0.20 with erythrocyte index in both raw and processed data. However, these six were not removed since they were not the top correlated proteins in either data version, and that many of them have immune-related functions with five being lower in healthy volunteer. Five proteins in the erythrocyte index marker panel (HBA1, HBB, ALDOA, ACTG1, HBD, all have relatively high correlations) and one in the platelet marker panel (GSN, low correlation) were also removed. In total 22 proteins were removed, leaving 269 to be processed further.



(a) (b)

Figure 3.9: Pearson correlation coefficients of protein levels with the platelet index (a) or the erythrocyte index (b), calculated in either raw log_2 intensity after protein filtering (along x axis), or in processed data version where the contaminations were not corrected for (along y axis). Red dashed lines are the 0.20 cut-off. 22 proteins to be removed are labelled in red.

To assess if any further proteins need to be removed, protein correlations with

Table 3.3: Proteins removed for being potential cell residue contaminations

Contamination index	Criterion of removal	Proteins removed
Platelet index	PCC>0.20 in both raw and processed data	ACTB, VCL, TAGLN2, ACTC1, ACTN1, FERMT3, TMSB4X, PPBP, PF4, CFL1, YWHAZ, GAPDH, YWHAE, LDHB, PFN1, F13A1
	Other proteins on the marker panel	GSN
Erythrocyte index	PCC>0.20 in both raw and processed data	GAPDH, ACTC1, ACTB, ACTN1, PPBP, TMSB4X, VCL, YWHAE, TAGLN2, YWHAZ
	Other proteins on the marker panel	HBA1, HBB, ALDOA, ACTG1, HBD

the contamination indices were also calculated for the remaining protein in the fully processed data generated after batch correction in section 3.3.6. Scatter plots comparing the correlation of removed or remaining proteins with the indices (Fig.B.8) showed that there was only weak correlation of the remaining proteins so these did not need to be further removed.

One remaining question is whether samples with high contamination indices should be removed as well. Because of the potential batch effects observed in calculating the indices from raw data (Fig.B.5(c-d)), samples outliers were defined using the processed uncorrected data. Potentially contaminated samples were defined as those with an index more than two standard deviations above the mean as suggested in the report by Geyer et al. This requires that most samples have low levels of contamination so that outliers of the distribution are clearly apparent, which is not the case in MS2019 but the threshold is still useful for identifying the most contaminated samples. By this cut-off, samples outlying in either platelet or erythrocyte index were not outlying in protein detection numbers, or in PCA on raw intensity or on fully processed data after batch correction. Intensities of the 22 removed proteins did not constitute the highest or a large proportion of the total proteins detected in the outlying samples. Therefore, the cell residue contaminations reflected in the indices or the removed proteins did not dominate the other protein detections in the outlying samples thus samples do not need to be further removed.

3.3.5 Normalisation and imputation

For reasons described in the Introduction section, I chose to use Variance Stabilising Normalisation (VSN; Huber et al. 2002) to account for systematic bias and applied it to the raw intensities after protein filtering and removal of the 22 contamination proteins. Histogram of the protein intensities was not largely altered (Fig.3.10). Smoothing curve on the mean-s.d. plot after VSN (Fig.3.11) indicated that low abundance proteins still tend to have larger variation, but the individual protein points did not show a strong negative correlation.

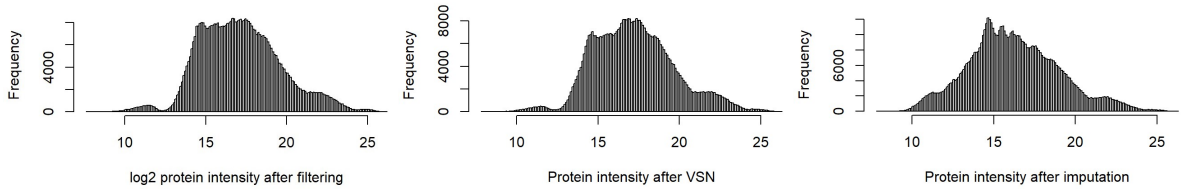


Figure 3.10: Distributions of protein intensity in pre-processing. The three histograms are generated after protein and sample filtering, or after VSN normalisation, or after hybrid imputation, respectively.

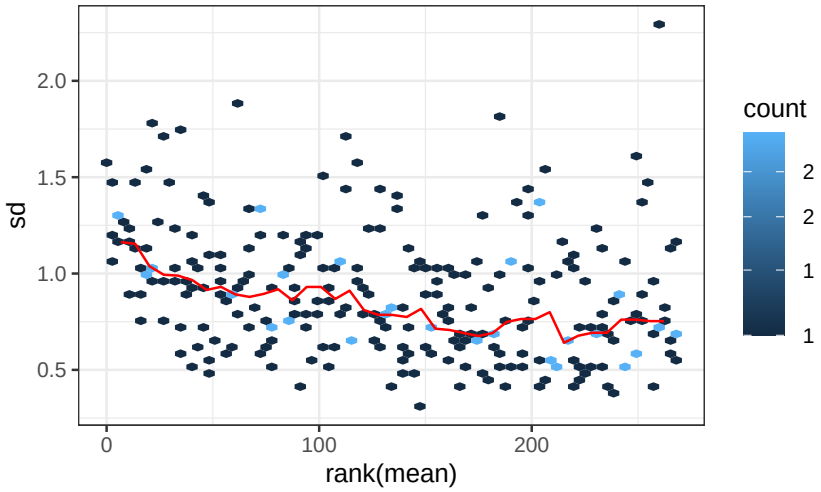


Figure 3.11: Mean-SD plot: Protein standard deviations plotted against ranks of protein means, after VSN normalisation. Proteins with smaller ranks are the less abundant ones. Red solid line shows the running median estimator with window width of 10%.

Under the assumption of missing not at random (MNAR), there should be a relatively clear cut-off in the intensity distribution beyond which there are few data points.

The histograms of this dataset (Fig.3.10) show both fewer data points on the lower abundance end but also a smaller peak in the low-level distribution, where it is more appropriate to make the assumption of missing not completely at random (MNCAR). Therefore, a hybrid imputation approach was applied here. For 170 proteins detected in ⩾60% of samples, I used K-nearest neighbours which is more tolerant on the assumption of the causes of missingness (as described in the introduction section 3.1.2). When there is higher proportion of missingness in a protein, KNN is not suitable since neighbour proteins cannot be accurately located. For the 99 proteins detected in <60% of samples, missing values were imputed by random draw from a protein-specific normal distribution with a standard deviation (s.d.) of 30% s.d. of the measured values and a down-shift of the mean by 1.8 s.d. The histogram of all intensities (Fig.3.10) showed that the distribution on the low-level end was visually smoother after the imputation.

3.3.6 Batch correction

As the 28 plates had the mass spectra acquired over a long duration, distributions should be checked among the plates to see if there was any batch effect within this one experiment. The plates did show a gradual change in scores along PC1–6 especially in PC1, with plates 1/2/15/16/17 outlying to the largest extent (Fig.B.9), indicating a batch effect along the acquisition that need to be corrected. This was not related to the cell residue contaminations, as there was no correlation between PC1–6 and the erythrocyte index or the platelet index of the samples (PCC of erythrocyte index and PC2 = -0.13, |PCC|<0.07 for the rest).

The samples were fully randomised so the study groups should be equally represented in the plates. Thus the pattern of PC1 scores between the plates indicate the technical batches. There were repeated pool injections between the samples, with spike-in of isotope labelled reference peptides at a 10% or 100% concentration. The spectral yield (PSM/MS2 ratio) and the number of proteins detected in those spiked-in pools

were also used to monitor the technical variation and thus to determine batches. PSM (peptide spectrum matches) is the number of spectra which could be assigned to a peptide sequence. For each injection, the proportion of PSM out of the total number of MS2 scans (PSM/MS2) represents the spectral yield. As the binding capacity of the chromatographic column decrease over time, the spectral yield can change but the direction of change is not clear in a noisy matrix like plasma.

Examining the three metrics together (Fig.3.12), there was a pattern corresponding to the experiment record where the chromatographic column was blocked and changed after injections 5105 or 7528, and where the ion-transfer capillary was cleaned after injection 6647. Therefore instead of using each plate as a relatively small batch, seven plates (plates 3-7, 19-20) that were the most consistent between each other in the three plots were used as the reference batch (batch "B") where measurements would not be altered. For the rest of the plates, a new batch was defined if a plate was visually different from the previous plate in any of the three metrics. Plate 16 was treated as a separate batch for having low detection in samples as shown in Fig.3.6. In total, 17 batches were defined and corrected for accordingly using the ComBAT function from the sva package (Leek et al. 2021). After the correction, there were no plates consistently outlying in PC1–6 scores (Fig.B.9), indicating the batch effect had been effectively removed.

As a quality check on the cleaned up data after pre-processing, distributions of sample colours, storage time, age, and sex were plotted along PC1 to PC6 after batch correction (Fig.B.10) to see if any known factors in sample quality or demographics is driving the differentiation between samples. Although samples of orange/pink/red colours show a shift in the distribution, the direction and amount of shift compared with non-haemolytic samples were not consistent. Samples from females or males showed a small shift on PC1. There was no correlation between storage time or age with the PC scores.

For group comparisons, the GAinS patients and OBB patients were separated into a

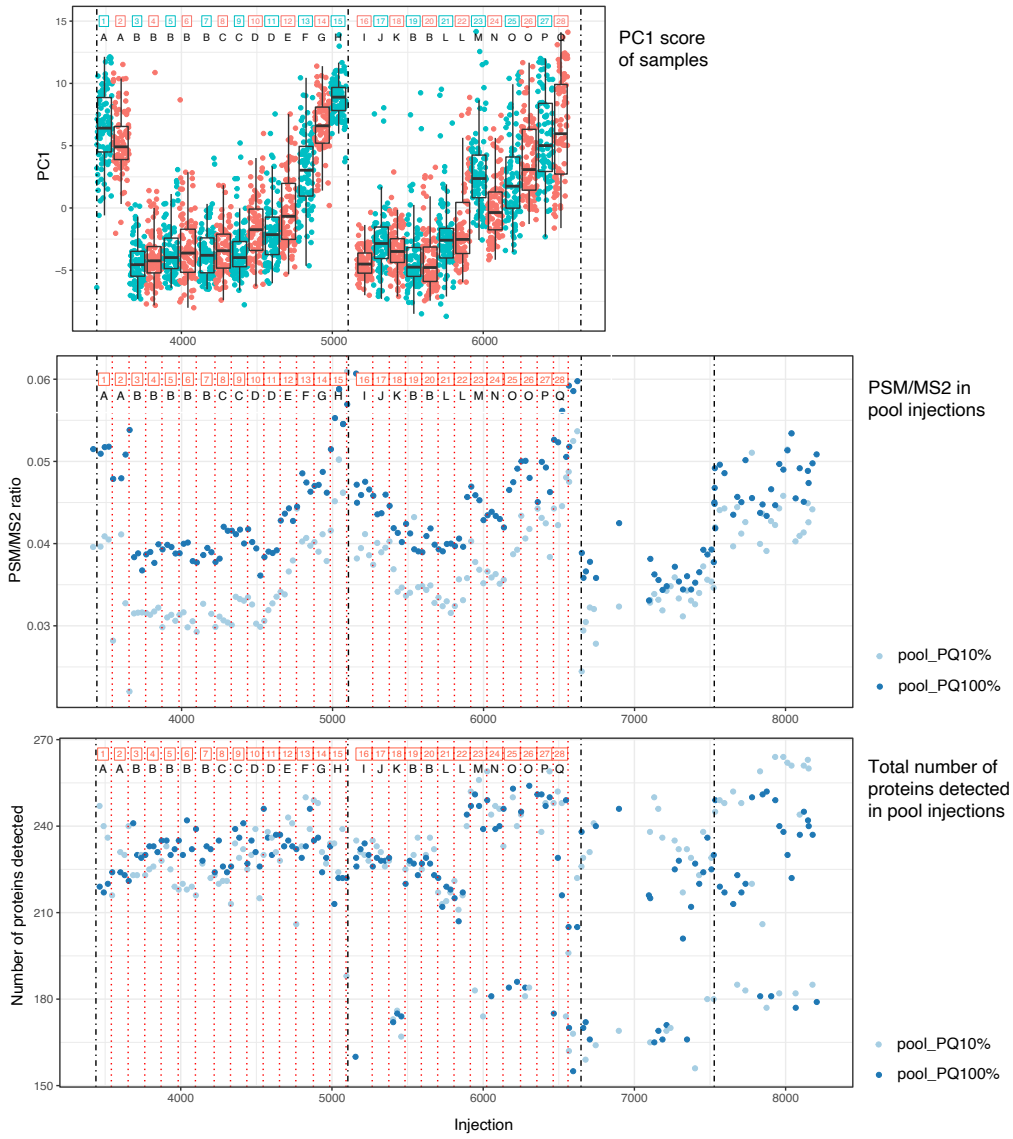


Figure 3.12: Batches across the plates were determined based on the PC1 scores in samples from each plate, the PSM/MS2 ratio in pool injections, and the total protein detection number in pool injections. Plate numbers 1-28 were labelled in squares. Seventeen batches determined were labelled A to Q, with batch B being the reference batch. Dashed lines indicate injection 3390 which is the first injection, injections 5105 and 7528 where the chromatographic column was changed, and injection 6647 where the ion-transfer capillary was cleaned. On the PC1 scores, samples from plates were labelled with alternating colour. PSM: peptide spectral matches; MS2: total number of MS2 scans; poo_PQ10%: pool injection with spike-in of PQ500 isotope labelled reference peptides at 10% concentration; pool_PQ100%: pool injection with spike-in at 100% concentration.

Table 3.4: Sample composition of cleaned-up MS2019 data. Total n(samples)=2575. Total n(patients)=1598. Abbreviations: preop, pre-operation; postop, post-operation; BIONIC_24hrs are samples taken 24hrs after surgery; BIONIC_others are samples taken either 2–6hrs or 48hrs after surgery.

Discovery cohort								
Study	GAinS	OBB	XMIN_preop	XMIN_postop	MOTION			
n(samples)	1236	99	104	103	48			
n(patients)	788	99	104	103	48			
Validation cohort								
Study	GAinS	OBB	BIONIC_preop	BIONIC_24hrs	BIONIC_others	MONOGRAM	TACE	VANISH
n(samples)	624	50	43	42	52	14	12	148
n(patients)	394	50	43	42	42	14	12	44

discovery and a validation cohort by a split of 2:1 with random draws. Serial samples from the same patient were kept in the same discovery or validation cohort. Other than the sepsis or healthy samples, the discovery cohort also included the XMIN surgery samples and MOTION no-infection ICU samples. The validation cohort also included the BIONIC surgery samples and no-infection ICU samples from TACE and MONOGRAM. The cleaned-up data contained quantifications of 269 proteins in 2575 samples constituted as in Table 3.4.

3.4 Results: Clinical characteristics of comparator groups

The study cohorts and the initial research aims of each study have been introduced in Methods. Table 3.5 shows the summary demographics and sources of infection or causes of hospital admission for each cohort, among all patients with a sample included in the processed data. The clinical characteristics including comorbidities, physiological measurements, treatments and outcomes, were listed in detail for the discovery and validation cohorts in GAinS (Table 3.6), or for all patients in GAinS (Table B.2). Consistent with effective random splitting, there was no significant difference found between the GAinS discovery or validation cohort. For the two surgery cohorts, clinical characteristics taken on the corresponding sampling

Table 3.5: Comparison of demographics of cohorts included, post-QC. The post-operative samples of the XMIN and BIONIC studies and samples from the MOTION, TACE and MONOGRAMS studies are referred to as a sterile inflammation state but are not used collectively as a single comparator group. In MOTION, non-operative medical causes included post cardiac arrest, cardiovascular, multiple trauma, neurologic, and others. HPB: Hepato-Pancreato-Biliary. Upper GI: upper gastrointestinal. *total: total number of patients with information available.

Study	GAinS		VANISH	OBB		XMIN
Description	ICU sepsis		Septic shock	Healthy volunteer		Surgery
Cohort	Discovery	Validation	Validation	Discovery	Validation	Discovery
No. patients (No. samples)	788 (1236)	394 (624)	44 (148)	99 (99)	50 (50)	106 (207)
Age median (IQR), years	65 (53-75)	65 (54-74)	64.5 (54-77)	50 (50-51)	50 (50-51)	67(61-74)
Men, No./total (%)	431/788 (55)	210/394 (53)	32/44 (73)	48/99 (48)	33/50 (66)	62/106 (58)
Caucasian ethnicity, No./total* (%)	740/784 (94)	370/387 (96)	34/44 (77)	99/99 (100)	50/50 (100)	94/106 (89)
Source of infection, No./total* (%)	CAP: 516/784 (66); FP: 268/784 (34).	CAP: 253/392 (65); FP: 139/392 (35)	Lung infection	-	-	-
Type of surgery or medical cause for admission, No./total (%)	-	-	-	-	-	Colorectal: 36/106 (34); Orthopaedic: 28/106 (26); Upper GI: 42/106 (40).

Study	BIONIC	MOTION	TACE	MONOGRAMS
Description	Abdominal surgery	No-infection ICU	No-infection ICU	No-infection ICU
Cohort	Validation	Discovery	Validation	Validation
No. patients (No. samples)	43 (137)	48 (48)	12 (12)	14 (14)
Age median (IQR), years	68 (57-74.5)	58(46-71)	62.5 (52-70)	59.5 (54.5-67)
Men, No./total (%)	23/43 (53)	35/48 (73)	5/12 (42)	12/14 (86)
Caucasian ethnicity, No./total* (%)	39/43 (91)	31/45 (69)	11/11 (100)	Unknown
Source of infection, No./total* (%)	-	-	-	-
Type of surgery or medical cause for admission, No./total (%)	Colorectal: 16/43 (37); HPB: 22/43 (51); Upper GI: 5/43 (12).	Non-operative causes: 32/48 (67); Emergency surgery: 13/48 (27); Elective surgery: 3/48 (6).	Brain injury: 10/12 (83); Others 2/12 (17).	Brain injury: 6/14 (43); Cardiac arrest: 3/14 (21); Trauma: 2/14 (14); Others 3/14 (21).

timepoints were listed in Tables 3.7 and 3.8.

Table 3.6: Summary of clinical characteristics in the GAinS cohorts. N (disc.) and N (vali.) denote the numbers of patients with the measurement available, in the discovery or validation cohort respectively. No significant difference in any variable listed was found between the two cohorts (FDR<0.05, correction for multiple testing performed in numerical or in categorical variables separately). *For clinical measurements on the basis of day rather than patient, values recorded here are those measured on the day of the first available sample included in MS2019 for the patient. [1]Microbiology information is only available for patients with CAP. Patients with mixed bacterial-viral or fungal infections were not counted in the bacterial or viral infections. [2]Septic shock is defined as hypotension despite adequate intravenous fluid resuscitation, according to the Sepsis-2 definition (Levy et al. 2003). So any patient timepoint with vasopressors or inotropes administered, or with the lowest mean arterial pressure <65mmHg was counted as septic shock. [3]ARDS phenotypes assigned according to the Berlin definition (Ferguson et al. 2012). [4]Renal support includes haemofiltration, diafiltration and dialysis. Abbreviations: IQR=interquartile range, SOFA= Sequential Organ Failure Assessment, APACHE=Acute Physiology and Chronic Health Evaluation, CAP=community acquired pneumonia, FP=faecal peritonitis, ARDS= Acute Respiratory Distress Syndrome.

	Discovery cohort	Validation cohort	N (disc.)	N (vali.)
No. patients	788	394	-	-
Age median (IQR), years	65 (53-75)	65 (54-74)	788	394
Men, No./total (%)	431/788 (55)	210/394 (53)	788	394
Pre-existing conditions, No./total (%)				
Heart/vascular diseases	321/787 (41)	165/394 (42)	787	394
Respiratory diseases	374/787 (48)	178/394 (45)	787	394
Current or ex- smoker	264/762 (35)	120/379 (32)	762	379
Malignancy or immune disease	137/787 (17)	74/394 (19)	787	394
Diabetes	151/787 (19)	66/394 (17)	787	394
Estimated time from disease onset, median (IQR), days				
Patients with CAP	5 (3-7)	5 (3-7)	323	171
Patients with FP	2 (1-4)	3 (1-5)	265	135
Microbiology[1], No./total (%)				
Bacterial	164/209 (78)	70/92 (76)	209	92
- Streptococcus pneumoniae	68/477 (14)	29/239 (12)	477	239
Viral	35/209 (17)	20/92 (22)	209	92
- Influenza	27/477 (6)	12/239 (5)	477	239
APACHE II score at day 1, median (IQR)	14 (11-18)	15 (11-19)	376	179

Table 3.6 continued from previous page

	Discovery cohort	Validation cohort	N (disc.)	N (vali.)
SOFA scores*, median (IQR)				
Cardiovascular	1 (0-4)	1 (0-4)	781	390
Respiratory	2 (2-2)	2 (2-2)	742	372
Kidney	0 (0-1)	0 (0-1)	781	390
Liver	0 (0-0)	0 (0-0)	754	371
Hematological	0 (0-1)	0 (0-1)	779	388
Neurological	0 (0-0)	0 (0-0)	781	390
Total	5 (3-8)	6 (3-8)	718	355
Physiological variables*, median (IQR)				
Lowest mean arterial pressure, mmHg	66 (60-74)	66 (60-74)	774	387
Lowest systolic blood pressure, mmHg	98 (88-110)	98 (85-110)	779	389
Highest heart rate, beats/min	110 (96-125)	110 (95-124)	779	389
Lowest heart rate, beats/min	79 (70-90)	80 (69-90)	779	389
Arterial pH	7.37 (7.29-7.43)	7.38 (7.28-7.44)	403	183
Respiratory rate	26 (20-33)	27 (20-34)	778	388
Partial pressure of oxygen (PaO_2), kPa	8.9 (7.9-10.3)	9 (7.9-10.3)	737	369
Fraction of inspired oxygen (FiO_2)	0.4 (0.3-0.6)	0.4 (0.3-0.55)	779	388
PaO_2/FiO_2, kPa	21.5 (14.7-31.3)	22.8 (15.4-30.0)	737	369
Partial pressure of CO_2, kPa	5.4 (4.6-6.5)	5.4 (4.7-6.4)	737	368
Lactate, mmol/L	1.7 (1.3-2.6)	1.6 (1.2-2.4)	491	246
Bicarbonate, mmol/L	24 (20.85-28)	24 (21-27)	747	367
Highest urea, mmol/L	9 (5.45-14.35)	9 (6-14.05)	775	383
Urine volumne, mL/24hrs	1615 (978-2465)	1582 (1008-2340)	773	388
Highest creatinine, μmol/L	86 (60-143.25)	91 (62-139.5)	780	388
Lowest creatinine, μmol/L	79 (56-122)	85 (59.75-125.25)	780	388
Highest bilirubin, μmol/L	10 (6-18)	10 (7-18)	753	369

Table 3.6 continued from previous page

	Discovery cohort	Validation cohort	N (disc.)	N (vali.)
Alanine transaminase (AST), units/L	46 (26-70)	38 (21-76)	69	41
The international normalized ratio	1.2 (1.1-1.4)	1.2 (1-1.3)	356	156
Prothrombin time, seconds	14.7 (12.6-17.2)	15 (12.9-18.0)	265	159
Aspartate transaminase (ALT), units/L	30 (18-55)	27 (17-50)	461	229
AST/ALT ratio	1.3 (0.9-1.7)	1.6 (1.2-2.2)	46	30
Highest temperature, C	37.4 (36.9-38)	37.4 (36.9-38)	779	389
Lowest temperature, C	36.2 (35.8-36.7)	36.2 (35.7-36.7)	779	389
Lowest platelets, $\times 10^3/\mu L$	205 (140.5-287)	198 (132-275)	779	387
Highest white cell count, $\times 10^3/\mu L$	12.7 (9.1-17.5)	12.75 (8.6-18.8)	780	388
Lowest white cell count, $\times 10^3/\mu L$	11.3 (7.4-16.1)	11.4 (7.6-16.7)	780	388
Haematocrit, %	34 (30-39)	34 (29-39)	338	157
Lymphocyte count, $\times 10^3/\mu L$	0.84 (0.5-1.3)	0.84 (0.5-1.2)	758	378
Monocyte count, $\times 10^3/\mu L$	0.59 (0.31-0.9)	0.6 (0.3-1.0)	759	376
Polymorphonucleocyte count, $\times 10^3/\mu L$	10.06 (6.6-15)	10.2 (7.0-15.4)	760	378
Vasopressors or inotropes*, No./total (%)	317/781 (40)	169/390 (43)	781	390
Duration of inotrope/vasopressor support, median(IQR), days	2 (0-4)	2 (0-4)	786	394
Shock*[2], No./total (%)	448/773 (58)	231/387 (60)	773	387
Mechanical ventilation/CPAP*, No./total (%)	500/781 (64)	251/390 (64)	781	390
Duration of mechanical respiratory support, median(IQR), days	3 (1-9)	4 (1-11)	785	394
ARDS[3], No./total (%)				
No ARDS	579/787 (74)	275/394 (70)	787	394
Mild	16/787 (2)	9/394 (2)	787	394
Moderate	75/787 (10)	38/394 (10)	787	394
Severe	117/787 (15)	72/394 (18)	787	394

Table 3.7: Clinical characteristics of the XMIN patients post-operation. Measurements were taken within 24hr post-operation for 103 XMIN patients with a 24hr post-operation sample included in MS2019 processed data. N denotes the number of patients with the measurement available.

Measurement	Median (IQR)	N
Mean heart rate, beats/min	71.0 (62.4-78.4)	94
Systolic blood pressure (lying), mmHg	119 (110-134)	96
Diastolic blood pressure (lying), mmHg	62 (56-70)	96
Blood hemoglobin, g/dL	11.5 (10.3-12.6)	100
Blood creatinine, μmol/L	70.5 (60.0-82.2)	100
Blood white cell count, $\times 10^3/\mu L$	11.1 (9.0-12.7)	100
Blood neutrophil count, $\times 10^3/\mu L$	9.0 (7.2-10.9)	100
Blood lymphocyte count, $\times 10^3/\mu L$	1.0 (0.7-1.5)	100

Table 3.6 continued from previous page

	Discovery cohort	Validation cohort	N (disc.)	N (vali.)
Acute renal failure*, No./total (%)	162/781 (21)	87/390 (22)	781	390
Renal replacement therapy*, No./total (%)	85/781 (11)	40/390 (10)	781	390
Duration of renal support 4, median(IQR), days	0 (0-0)	0 (0-0)	786	393
Treated with activated protein C, No./total (%)	29/787 (4)	17/394 (4)	787	394
Corticosteroids, No./total (%)	191/787 (24)	86/394 (22)	787	394
28-day mortality, No./total (%)	133/780 (17)	65/392 (17)	780	392
6-month mortality, No./total (%)	189/780 (24)	99/392 (25)	780	392

For the pairs of groups for which protein abundance is compared in the following section 3.6.1, the matching clinical characteristics including demographics were compared in each cohort pair with the multiple-test correction performed within each pair. Comparing the first available samples from the GAinS discovery cohort against the post-operation samples in XMIN, the "mean heart rate" in XMIN was lower than the "lowest heart rate" in GAinS (FDR<0.0001); "blood creatinine" in XMIN was lower than the "lowest creatinine" in GAinS (FDR=0.03); blood lymphocyte count was higher

Table 3.8: Clinical characteristics of the BIONIC patients post-operation. Measurements were taken within 24hr post-operation for the 42 BIONIC patients with a 24hr post-operation sample included in MS2019 processed data. N denotes the number of patients with the measurement available.

Measurement	Level	N
Length of hospital stay, median (IQR), days	10 (7.25-14)	42
Malignancy, No./total (%)	38/42 (90.5)	42
Post-operation location, No./total (%)		42
- Ward	1/42 (2.4)	
- High-dependency unit	33/42 (78.6)	
- Intensive care unit	8/42 (19.0)	
Blood hemoglobin, median (IQR), g/dL	11.6 (9.9-12.2)	41
Blood white cell count, median (IQR), $\times 10^3 \mu L$	11.7 (9.7-16.2)	41
Blood neutrophil count, median (IQR), $\times 10^3 \mu L$	9.3 (8.1-13.3)	41
Blood lymphocyte count, median (IQR), $\times 10^3 \mu L$	1.3 (0.9-1.6)	41
Blood monocyte count, median (IQR), $\times 10^3 \mu L$	1 (0.7-1.3)	41
Blood creatinine, median (IQR), μmol/L	67 (61.25-80.75)	42
Blood C-reactive protein, median (IQR), mg/L	117.5 (73.25-152.5)	42
Respiratory support, No./total (%)	2/42 (4.8)	42
Current antibiotics, No./total (%)	29/41 (70.7)	41
Temperature $\geqslant$38°C, No./total (%)	4/40 (10)	40

in XMIN than in GAinS (FDR=0.01); age was higher in XMIN (FDR=0.02) and there was more men in XMIN (FDR=0.01), although the differences were only marginal.

Comparing the first available samples from the GAinS validation cohort against the 24hr post-operation samples in BIONIC, only 19% BIONIC patients were in ICU, while all GAinS patients were critically ill; "blood creatinine" in BIONIC was lower than the "lowest creatinine" in GAinS (FDR=0.01); lymophocyte and monocyte counts were lower in GAinS than in BIONIC (FDR<0.0001 for both); age and sex were not different between GAInS and BIONIC.

Comparing the first available samples in GAinS vs OBB, in the discovery cohort OBB participants had lower age (p<0.0001) and lower proportion of males (p=0.02); in the validation cohort, OBB participants had lower age (p<0.0001) but higher proportion of males (p=0.046), due to a random difference that occurred in the random split of OBB participants. In all the six sepsis-control comparisons in the following sections,

Table 3.9: Comparison of clinical characteristics between MOTION and GAinS discovery cohort patients, using the first available samples. Numerical variables were compared by Mann-Whitney tests. Categorical variables were compared by Chi-squared tests or Fisher's exact tests.

Measurement	Level in MOTION	Level in GAinS disc.	FDR
Numerical variables (median (IQR))			
Age	58 (46-71)	65 (53-75)	0.067
Highest temperature, C	37.8 (37.2-38.2)	37.4 (36.9-38)	**0.0367**
Lowest temperature, C	36.3 (36-36.7)	36.2 (35.8-36.7)	0.509
Lowest mean arterial pressure, mmHg	66 (58-74.2)	66 (60-74)	0.993
Highest heart rate, beats/min	91.5 (81.8-108.5)	110 (96-125)	**3.4E-06**
Lowest heart rate, beats/min	65 (58-72)	79 (70-90)	**5.8E-09**
Highest creatinine, μmol/L	65.5 (55-94.5)	86 (60-143.2)	0.059
Highest white blood cell, x10^9/L	10.9 (8.2-14.3)	12.7 (9.1-17.5)	0.069
Lowest white blood cell, x10^9/L	9.1 (8.8-9.3)	11.3 (7.4-16.1)	0.509
Total APACHE II	16.5 (13-22)	14 (11-18)	**0.0180**
Lactate, mmol/L	0.9 (0.8-1.1)	1.7 (1.3-2.6)	**1.4E-16**
Partial pressure of oxygen, kPa	10.8 (9.7-12.2)	8.9 (7.9-10.3)	**3.7E-07**
Fraction of inspired oxygen	0.3 (0.2-0.4)	0.4 (0.3-0.6)	**0.0010**
PaO_2/FiO_2 ratio, kPa	35.5 (26.6-44.1)	21.5 (14.7-31.3)	**6.9E-08**
Categorical variables (No./total (%))			
Men	35/48 (73)	431/788 (55)	**0.0090**
Acute renal failure	1/48 (2)	162/781 (21)	**0.0057**
Moderate or severe ARDS	3/48 (6)	192/787 (24)	**0.0192**
Renal replacement therapy	0/48 (0)	85/781 (11)	**6.6E-17**

age and sex were included as covariates.

Comparing the first available samples in the GAinS discovery cohort against the MOTION samples (Table 3.9), MOTION patients had higher highest temperature and higher APACHE scores, while GAinS patients had higher heart rates, higher lactate, lower PFratio, and higher occurrence of ARDS or renal failure, suggesting that the MOTION patients may be experiencing more acute inflammation while the sepsis patients had more severe cardiovascular/respiratory/renal dysfunctions. A full list of clinical characteristics for MOTION patients were summarised in Table B.3.

3.5 Results: The plasma proteome network in sepsis and controls

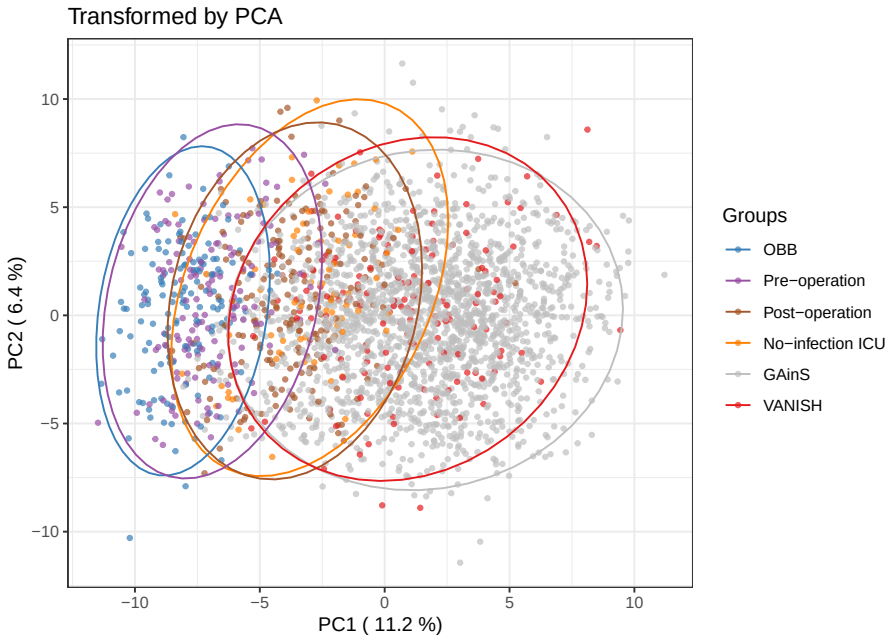
3.5.1 The overall differentiation of samples

To first examine the overall sample structure, I performed a PCA including all comparator groups in the processed data (Fig.3.13(a)). The proportions of variance explained by the first two PCs were relatively small, indicating that in the multi-dimensional space of protein intensities, samples were dispersed among the space instead of having higher density surrounding a certain plane. This reflected a limited amount of correlation between the proteins and large variations in a highly heterogeneous sample set analysed. Nonetheless, the biological differences between the groups were captured on the first principal component. There was a gradual separation along PC1, from the relatively healthy OBB and pre-operation samples, to the post-operation and no-infection ICU samples, and then to the sepsis and septic shock samples with the highest PC1 scores. The second PC separated within each cohort. Within GAinS patients, PC1 scores had significant positive correlations with clinical characteristics including lactate, SOFA scores, occurrence of shock, mortality, and significant negative correlations with variables including blood pressure, platelets,

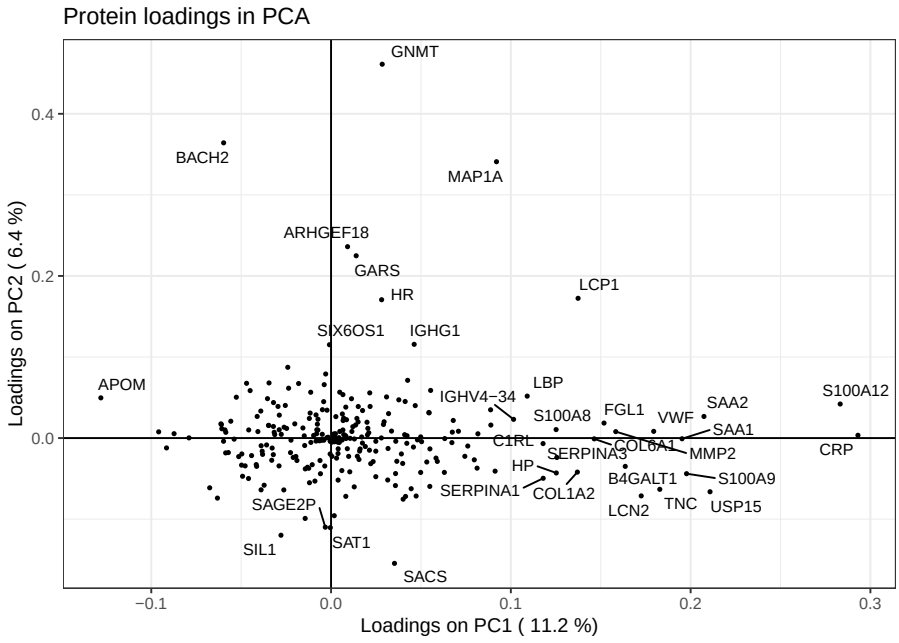
and lymphocyte proportion (Fig. 3.14). Thus, a higher PC1 score was associated with higher sepsis severity.

Proteins with loadings >0.1 on PC1 (Fig.3.13(b)) included: Acute-phase proteins (CRP, SAA1, SAA2, SERPINA1, SERPINA3, HP, C1RL), pro-inflammatory proteins in the S100 family (S100A8, S100A9, S100A12), other proteins with innate immune or anti-bacterial functions (LCN2, LBP, USP15), and extracellular matrix (ECM) proteins (TNC, MMP2, COL1A2, COL6A1). Among these, tenascin C (TNC), S100 proteins, and serum amyloid A are damage-associated molecular patterns (DAMPs), which are molecules released from damaged or dying cells due to trauma or infection, and promote an innate immune response (Roh and Sohn 2018). VWF (von Willebrand factor) and FGL1 (Fibrinogen-like protein 1) which is indicated to suppress T cell activation also had high positive loadings on PC1. APOM (Apolipoprotein M) implicated for lipid transport had negative loading on PC1 < -0.1. Intracellular proteins GNMT, BACH2, MAP1A and SACS had the largest loadings on PC2, indicating a variable level of tissue damage and leakage leading to their detection in circulation.

I then used single sample gene set enrichment analysis (ssGSEA) to understand which biological processes were separating among the cohorts. Simply put, the enrichment score in GSEA is calculated for each gene set considered, by walking down a ranked gene list and increasing or decreasing a running-sum statistic when encountering a gene in the set or not in the set. The magnitude of the increment depends on e.g. correlation of the gene with the phenotype of interest. The enrichment score is then the maximum deviation from zero in the random walk (Subramanian et al. 2005). In ssGSEA that I applied here, proteins were ranked by their abundance within a single sample. One enrichment score is calculated for each gene set in each sample, instead of for the whole sample set. Therefore, through an ssGSEA projection the matrix of protein intensities of genes (proteins) in samples is converted to a matrix of enrichment scores of gene sets in samples. PCA on the projected matrix revealed the gene sets that



(a)



(b)

Figure 3.13: Principal component analysis in MS2019. (a) Scatter plot of sample scores on the first two PCs. Pre- and post- operation groups are those from the XMIN and BIONIC studies. No-infection ICU included samples from MOTION, TACE, and MONOGRAM. Data ellipses at a 95% confidence level assuming a multivariate normal distribution were plotted for each group. The proportions of variance explained by each PC were stated in brackets. **(b)** Protein loadings in PCA. Proteins with the absolute value of loadings >0.1 on PC1 or >0.15 on PC2 were labelled.

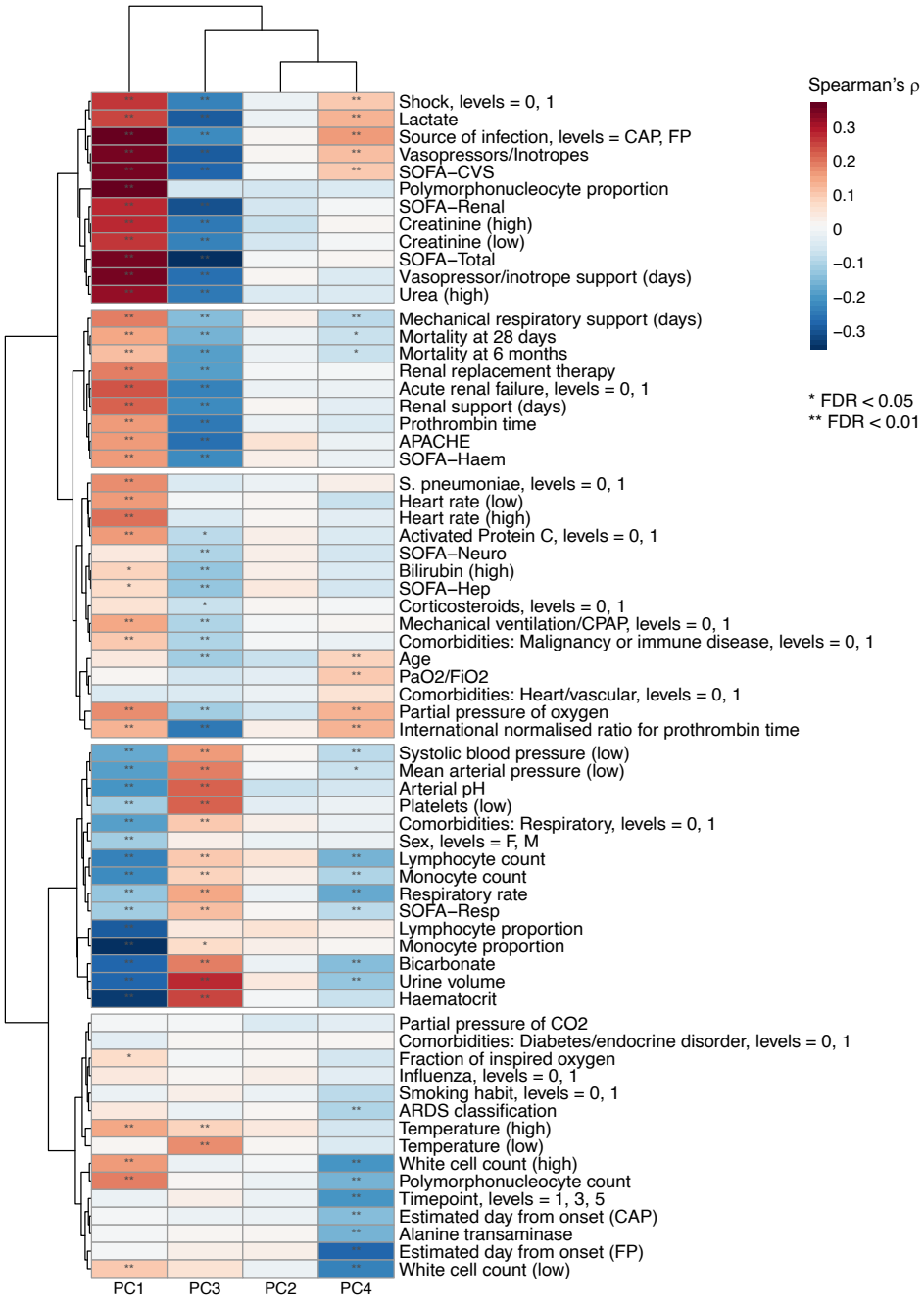


Figure 3.14: Correlations between clinical characteristics and PC1-4 scores in first available samples of GAinS patients. Heatmap shows Spearman's rank correlation coefficients, labelled by FDR corrected within all tests. Clinical variables detected in less than 30% of patients were excluded. Variables only available in certain groups of the patients (e.g. estimated days from CAP/FP onset) were retained. *FDR<0.05; **FDR<0.01. Level 0 – absence of the event; level 1 – presence of the event.

differentiated among the samples.

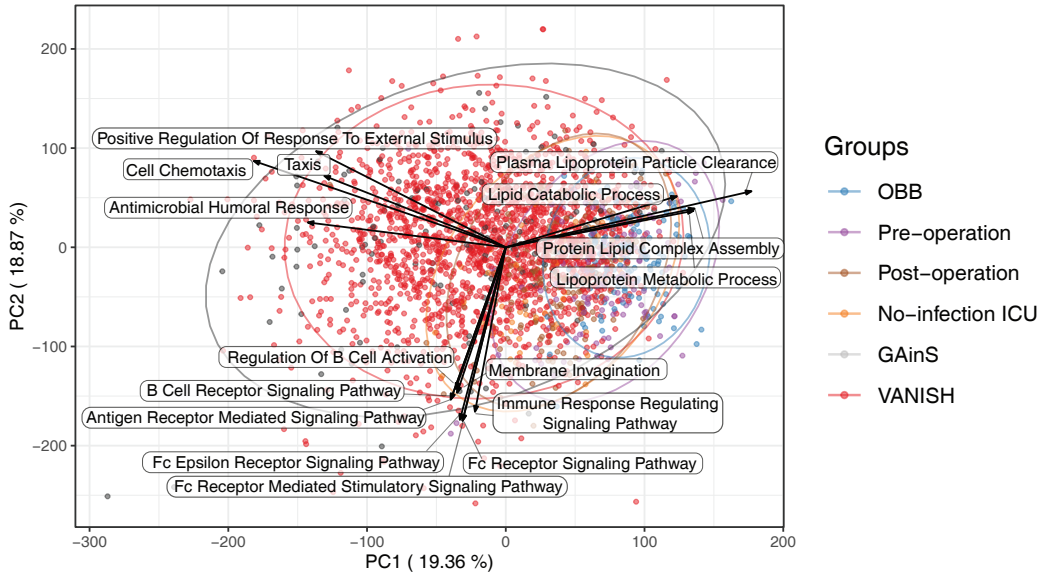


Figure 3.15: Separation of samples by biological processes shown in ssGSEA. A biplot of PCA performed on the matrix of gene set enrichment scores in single samples. GOBP annotations were used to define gene sets. GOBP terms with top 8 loadings on PC1 or top 8 loadings on PC2 were shown in arrows. Lengths of the arrows are scaled to the loadings.

Using GOBP annotations, the PCA on enrichment scores also separated the comparator groups along PC1 (Fig. 3.15). Among the gene (protein) sets with the highest loadings, the main biological processes associated with a more severe response towards the sepsis end included antimicrobial humoral response, cell chemotaxis, and positive regulation of response to external stimulus. Proteins annotated with lipoprotein metabolic process and lipid catabolic process were more abundant towards the healthy end on PC1. A group of immune response processes almost perpendicular to the sepsis-healthy axis separated samples within each cohort, including regulation of B cell activation, and Fc (crystallisable fragment) receptor signalling pathway.

Plasma protein abundance was widely associated with clinical characteristics in the GAinS patients (Fig.3.16). A group of proteins (including PTGDS, B2M, CFD, LCN2, VWF, COL6A1, USP15, MMP2, COL1A2, CD14, PLTP and CRP) had the highest positive correlations with clinical variables where a higher level indicates being more severely ill, including SOFA-Total, APACHE, occurrence of shock or renal failure, and

prothrombin time. This group of proteins also had the highest negative correlations with clinical variables where a lower level indicates higher severity, including bicarbonate, blood pressure, haematocrit, lymphocyte and monocyte proportions. A separate group of proteins exhibited the strongest associations in the opposite directions, including SERPIND1, C3, APOA1, HRG, KNG1, APOC4, KLKB1, VTN, and HABP2. There were also protein groups that appeared to indicate particular aspects of organ function. For example, proteins PTGDS, B2M, CFD and CST3 had the highest correlations with variables suggesting a more impaired renal function including blood creatinine, urea, urine volume, and level of renal support.

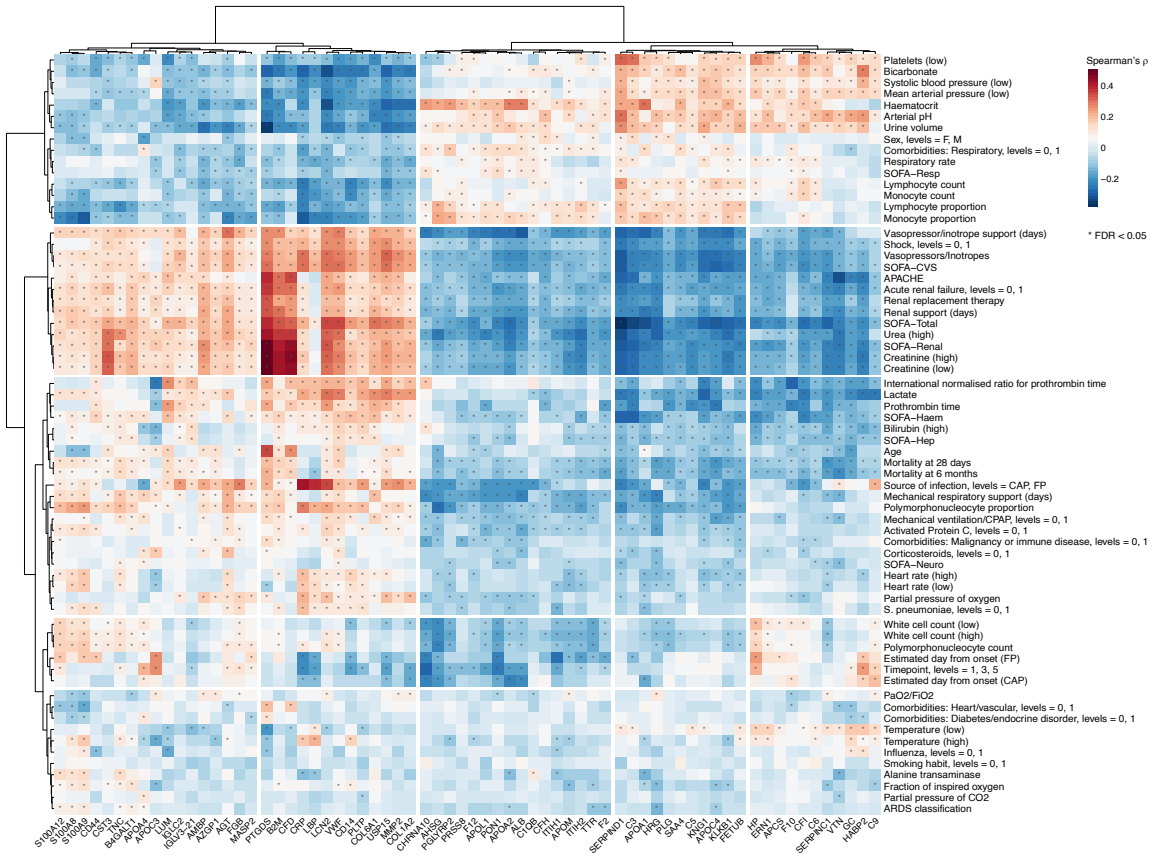


Figure 3.16: Correlations between clinical characteristics and protein abundance in first available samples of GAinS patients. Only the one-fourth of proteins (n=68) with the highest number of significant associations (*FDR<0.05) were displayed. Heatmap shows Spearman's rank correlation coefficients, labelled by FDR corrected within all tests. Clinical variables should be interpreted the same as described in Table 3.6. In variables with "levels": level 0 – absence of the event; level 1 – presence of the event.

3.5.2 Protein interaction network

To understand the functional network structure of the proteins reliably detected in this dataset, I retrieved protein-protein interaction (PPI) data from the STRING v11 database (Szklarczyk et al. 2019) and visualised the strengths of interactions using the Cytoscape platform (Shannon et al. 2003). 141 proteins were connected in the main network (Fig. B.11), including proteins either positively and negatively correlated with PC1 as in Fig. 3.13, higher scores of which were associated with higher disease severity towards sepsis. 81 proteins shown as singletons did not have an interaction that passed the confidence score cut-off with any other protein in the dataset.

To understand the functional clusters within the proteins with interactions identified, I used Markov clustering applied in the clusterMaker plugin to isolate more stable complexes from the proteins in the main network. I identified seven protein clusters with at least five nodes (Fig. 3.17), which were enriched respectively for biological processes including coagulation and fibrinolysis, complement activation, cholesterol metabolic process, acute-phase response, extracellular matrix organisation for two clusters, and neutrophil degranulation. These were examples of the most actively regulated processes that could be reflected in the plasma proteome network uncovered by an untargeted discovery approach, in a comprehensive sample set including inflammatory conditions with or without infection. Other than the seven functional clusters, there were also smaller clusters or singletons from the main network with a variety of functions in the innate immune response, including two 3-member clusters constituted of immunoglobulin-related proteins or calprotectin components.

3.5.3 Protein co-expression network

Weighted gene co-expression network analysis (WGCNA) is a systems biology method for describing the gene correlation patterns across samples, which has been successfully applied in various biological and disease contexts (Langfelder and

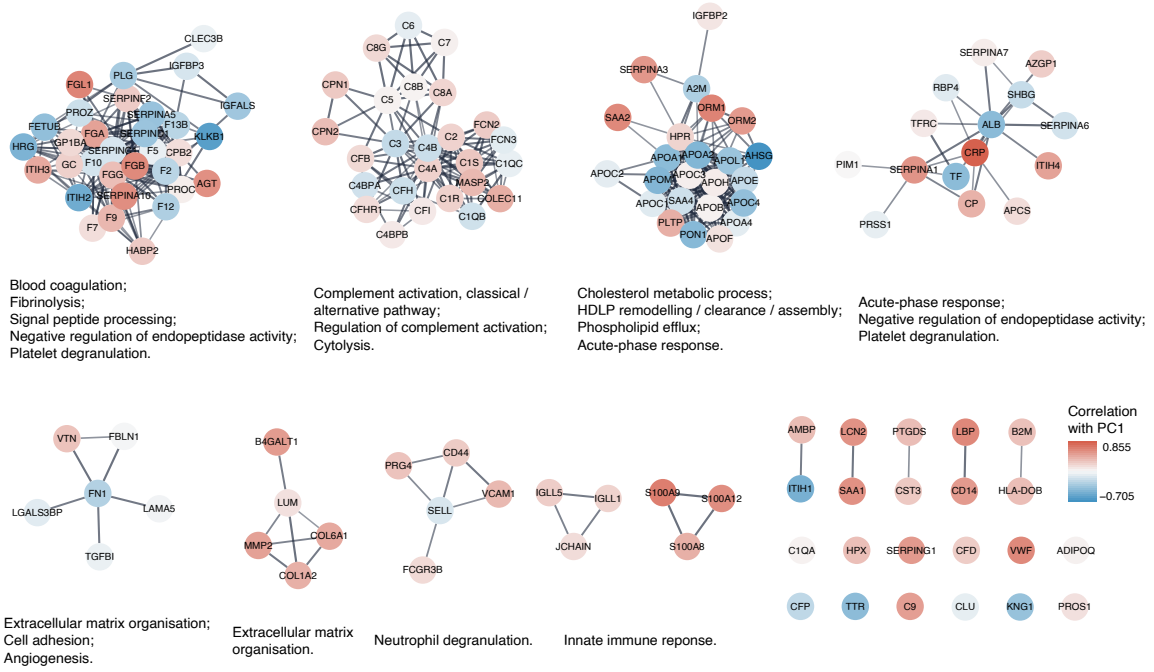


Figure 3.17: Clustered protein-protein interaction network of proteins detected in MS2019. Node colour was mapped to the Pearson's correlation coefficient between the protein level and PC1 score across the samples. GO biological processes enriched in member nodes were listed below each cluster, identified with XGR.

Horvath 2008). Assuming that proteomics data in addition to microarray data can also be more completely represented by considering the pairwise relationships between measured protein species, I applied WGCNA in MS2019 to understand the plasma proteome co-expression network.

The nodes of such a network are abundance profiles of a protein. The edges represent pairwise correlations between two proteins. The network is "weighted" by raising the correlation to a power $\beta \geqslant 1$ (soft thresholding), so that the network construction emphasises high correlations over low correlations. I applied a signed network where only positive but not negative correlations were considered for adjacency between two proteins. Using a soft thresholding power of 3 and other parameters selected as described in Methods, 184 out of 269 proteins were grouped into 16 protein co-expression modules. Compared with application of WGCNA in gene expression datasets, there was a relatively large proportion of features (85/269=32%) that did not belong to any of the co-expression modules, which is probably due to the much lower

number of proteins detected compared with gene probes thus the incomplete coverage of the proteome.

The WGCNA package calculates the similarity between the protein species by transforming the correlation-based adjacency into the Topological Overlap Matrix (TOM) as shown in the heatmap Fig. B.12. The 85 proteins that did not belong to any of the co-expression modules had low levels of correlation with other proteins. Proteins in the 16 co-expression modules had higher correlation and formed larger clusters at higher hierarchy.

I also visualised pairwise similarity between the 184 proteins in a network with the weight of edges mapped to the pairwise topological overlap (Fig.3.18). Proteins assigned to the same module were closer together with higher overlap. The co-expression network also made it visually clearer that certain proteins with closely related functions are also highly co-expressed in the plasma. Examples include the acute-phase proteins in the blue module, the complement C1q complex subunits in the cyan module, the S100 family proteins in the tan module, the apolipoproteins in the yellow module, the fibrinogens in the brown module, and the immunoglobulins in the closely connected black, purple and magenta modules. Labelling the modules on the clustered protein-protein interaction network also supported the relation between reported functional interactions and co-expression patterns identified here from plasma (Fig.3.19). More examples include the complements in the red module, and the extracellular proteins in the greenyellow module.

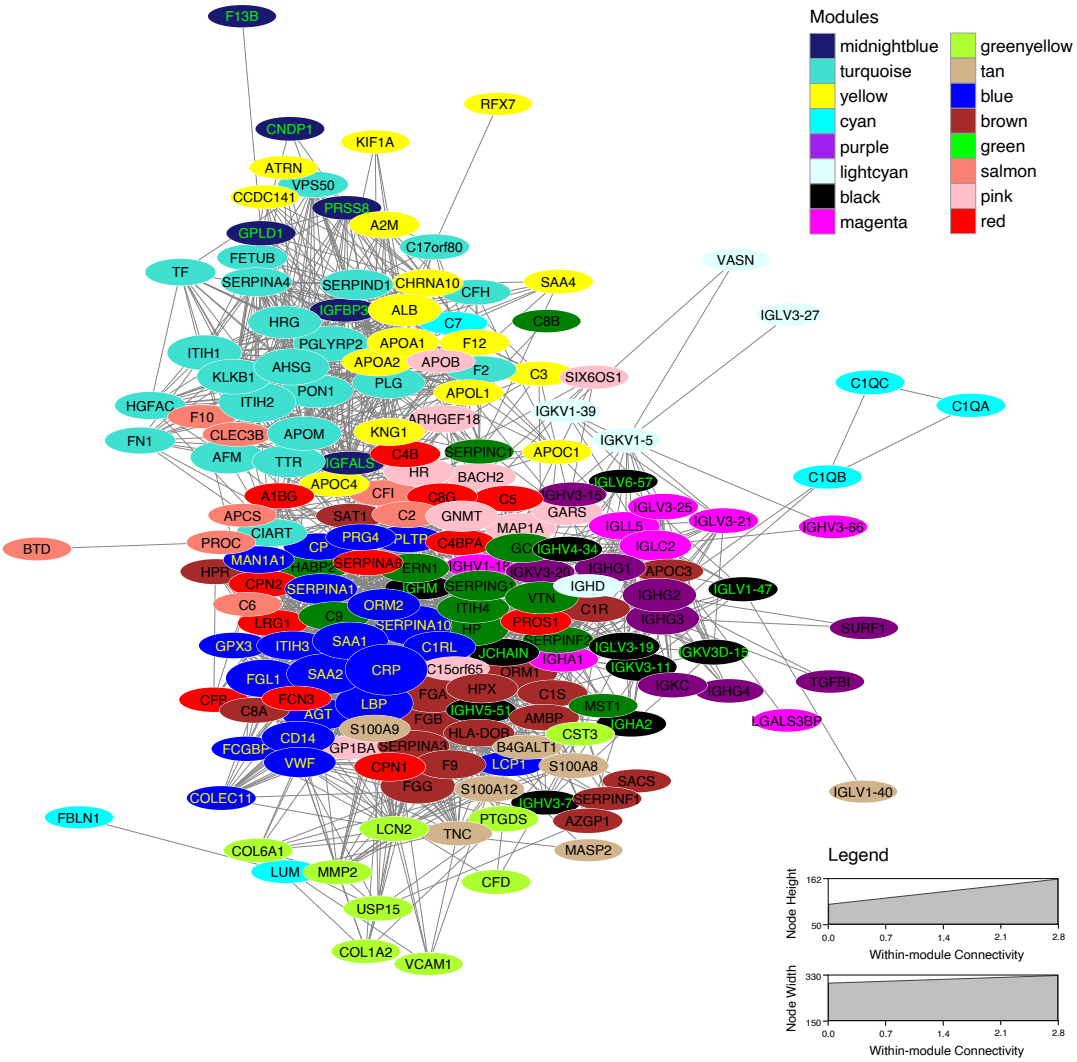


Figure 3.18: Protein co-expression network visualised using the Cytoscape platform. Node colour are mapped to module colours. Only edges with topological overlap ⩾0.02 are included thus 180 out of 184 nodes are plotted. Weight of the edges are mapped by applying a perfuse force directed layout. Wider and shorter edges denote higher topological overlap between two connected nodes. Nodes with larger width and height have higher intramodular connectivity, which is the sum of adjacency with all other nodes in the same module. Exact position of the nodes have been manually adjusted to display all the labels.

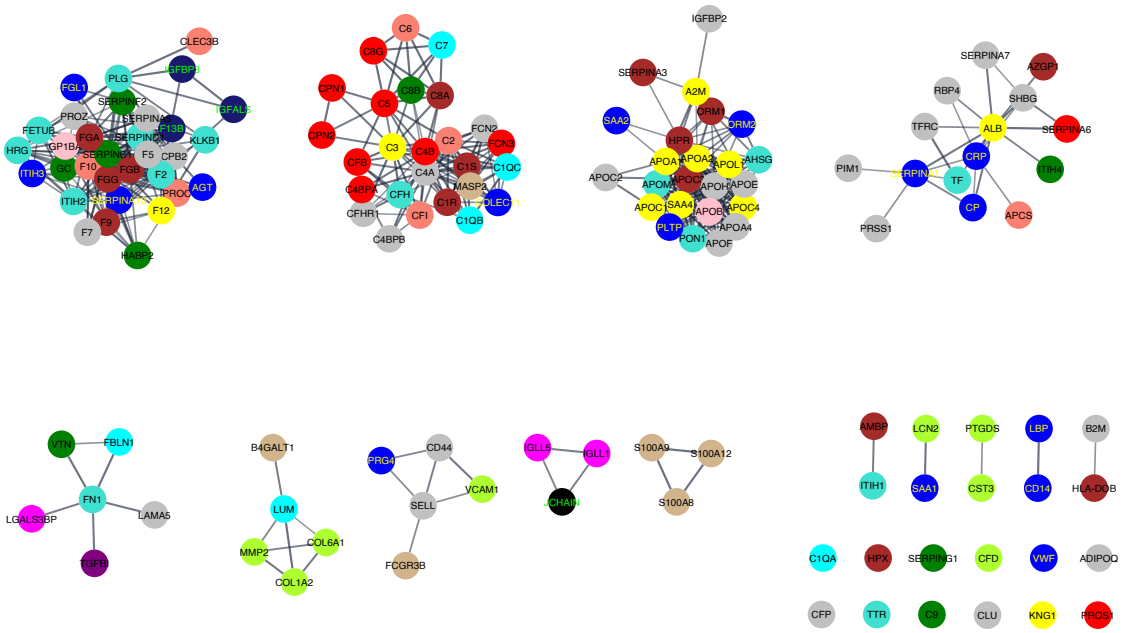


Figure 3.19: The clustered protein-protein interaction network with node colours mapped to co-expression modules. Generation of the clustered network is the same as described for Fig. 3.17, with the only difference in node colours. Grey nodes are proteins not co-expressed with other proteins in the dataset and should not be interpreted as a co-expression module.

3.6 Results: Characterisation of sepsis-specific proteomic response

3.6.1 Group comparisons

I then compared each of the three non-sepsis biological groups against the sepsis samples in GAinS to understand the biological difference between the groups from the perspective of differentially abundant proteins. For each type of control, two contrasts were performed between cohorts either both assigned to the discovery cohort or both assigned to the validation cohort and the results were compared. For GAinS patients with serial samples, only the first available sample of each patient was included in the contrasts.

Sepsis versus healthy volunteers

There were 61 and 61 proteins (53 in overlap) differentially abundant (FDR<0.05 and |FC|>1.5) between GAinS and OBB in the randomly split discovery or validation cohort (Fig. 3.20). Proteins more abundant in GAinS had larger fold change and significance, and showed a large overlap with the proteins with large positive loadings on PC1, including proteins implicated with the acute-phase response (e.g. CRP, SAA1, SAA2), coagulation process (VWF, FGB, FGA), and immune or immune-regulatory functions (LBP, S100A9, FGL1). Among these, C-reactive protein (CRP) is an extensively used marker for inflammation, secreted by the liver and induced by IL-6; The von Willebrand factor (VMF) promotes adhesion of platelets to the sites of vascular injury by forming a molecular bridge and thus enhances coagulation; S100A9 with S100A8 form the heterodimer calprotectin which is constitutively expressed in the cytosol of neutrophils and monocytes as a Ca^{2+} sensor, is released actively during inflammation and stimulates leukocyte recruitment and cytokine secretion (Wang et al. 2018). ORM1 (Alpha-1-acid glycoprotein 1) was also higher in GAinS. ORM1 functions as a transport protein in the blood stream and also appears to function in modulating the activity of the immune system during acute-phase reaction (UniProt).

The monocyte differentiation antigen CD14 was also higher in GAinS, which could be in the soluble form sCD14. CD14 serves as a high-affinity receptor for LPS (lipopolysaccharides) and LBP (LPS-binding protein) complexes, which are released from the phagocyte cell membrane into the circulation creating sCD14, and activate TLR4-specific pro-inflammatory signalling cascade. sCD14 could also be secreted from liver in response to inflammation and infection, induced by IL-6. Elevated levels of sCD14 have been found in sepsis and were evaluated of diagnostic and prognostic values in sepsis (Zhang et al. 2015; Ríos-Toro et al. 2017).

Proteins lower in GAinS included: apolipoproteins (APOA1, APOA2, APOM), alpha-2-HS-glycoprotein (AHSG, possesses opsonic properties and influences the mineral

phase of bone), hepatocyte growth factor activator (HGFAC), plasma serine protease inhibitor (SERPINA5, acts as both procoagulant and anticoagulant), transthyretin (TTR, a thyroid hormone-binding protein with the level inhibited by inflammation), and transcription regulator protein BACH2 (regulates apoptosis and adaptive immunity).

Sepsis versus surgery response

There were 23 proteins and 43 proteins (14 in overlap) that were differentially abundant between GAinS and the post-operation samples from XMIN or BIONIC, respectively (Fig. 3.20). As all cohorts had EDTA plasma except BIONIC, MOTION and 3 patients in TACE, the larger differences observed in the GAinS-BIONIC contrast partly reflected the effect of different anti-coagulants used on the plasma proteome measured, which is discussed in detail in section 3.7.5. The 14 overlapping proteins were all also differentially abundant in the two GAinS-OBB contrasts, except for ficolin-2 (FCN2) which are lectin-type soluble PRRs and activate the lectin complement pathway (Matsushita 2010).

Top proteins higher in GAinS included HP, B4GALT1, TNC, and S100A12, among the others. HP (haptoglobin) is a liver-secreted protein that is induced by IL-6 and modulates many aspects of the acute phase response, including being indicated to suppress lymphocyte function (Sadrzadeh and Bozorgmehr 2004). It has a wide range of anti-inflammatory functions and acts as an antioxidant to capture oxygenated, free hemoglobin and facilitating its removal from the circulation (Theilgaard-Mönch et al. 2006). B4GALT1 is a galactosyltransferase that is ubiquitously expressed and plays a critical role in the processing of N-linked oligosaccharide moieties in glycoprotein substrates including apolipoproteins, fibrinogen, immunoglobulin G, and transferrin (Montasser et al. 2021). It is a membrane-bound protein and a secreted protein in processed form. The cell surface form of B4GALT1 is also a recognition molecule during cell to cell and cell to matrix interactions. TNC

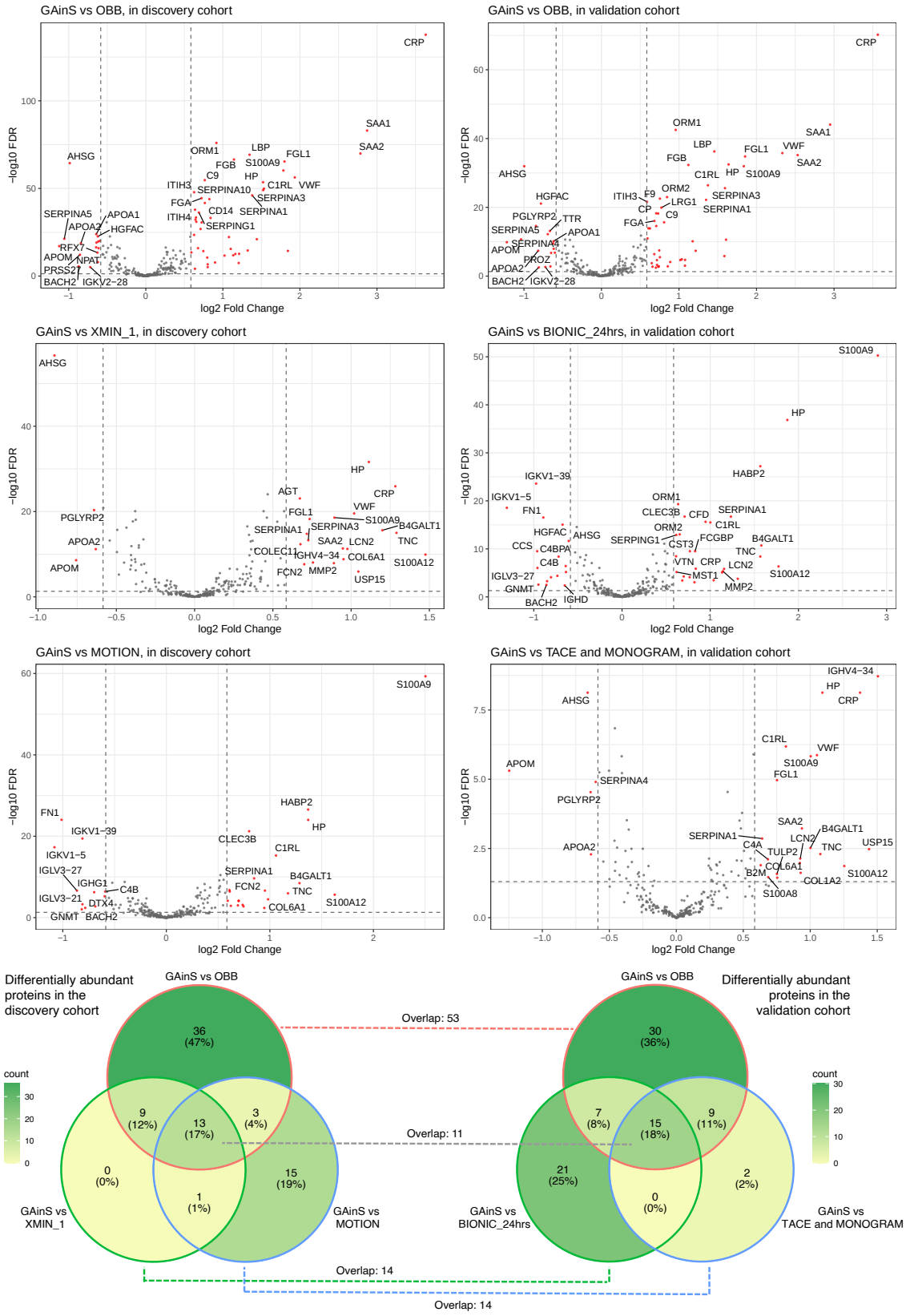


Figure 3.20: Group comparisons between GAinS and one of the non-sepsis cohorts. The GAinS discovery cohort and the GAinS validation cohort were used for contrasts on the left or on the right, respectively. In the volcano plots, differentially abundant proteins (FDR<0.05 and |FC|>1.5) are shown in red. The ones with the top significance had the protein names labelled. Venn diagrams show the numbers of overlapping differentially abundant proteins.

(tenascin) is an ECM protein highly expressed during embryonic development, in tissue injury, chronic inflammation and cancer. The expression can be induced by pro- and anti-inflammatory cytokines, hypoxia or mechanical stress. By interacting with other ECM molecules and cell surface receptors, TNC guides neuronal development, modulates cell migration, and mediates tissue repair (Midwood and Orend 2009). Both TNC and S100A12 are damage associated molecular pattern (DAMP) proteins that stimulate innate immune cells. Protein S100-A12 is a member of the S100 family of low molecular weight calcium-binding proteins, mainly expressed in neutrophil cytoplasm. Upon release, S100A12 functions in pro-inflammatory processes of recruiting leukocytes, promoting cytokine and chemokine production, and inducing oxidative stress (Meijer et al. 2012).

Top proteins higher in post-operation samples included AHSG and APOM. PGLYRP2 was lower in XMIN_1. PGLYRP2 is an amidase that may play a scavenger role by digesting biologically active peptidoglycan into inactive fragments (UniProt).

Sepsis versus no-infection ICU cases

In comparing samples from ICU sepsis patients or ICU patients without known infection, samples from the TACE and MONOGRAM studies were combined because of the small sizes and the relative similarity in the causes of admission. There were 32 proteins and 26 proteins (14 in overlap) differentially abundant between GAinS and MOTION or TACE plus MONOGRAM, respectively (Fig. 3.20). The difference in protein levels between no-infection ICU and sepsis observed here shared a lot in common with the difference between surgery response and sepsis. Among the 14 overlapping proteins from the two no-infection ICU contrasts, all except C1RL, SAA2, or IGHV4-34 were also among the 14 overlapping proteins from the two sepsis-surgery contrasts.

C1RL is a serine hydrolase that cleaves prohaptoglobin into their active forms. SAA2 (serum amyloid A-2) is a major acute-phase reactant secreted by the liver. It is known

to elevate rapidly under acute inflammatory conditions and has chemoattractant activity. IGHV4-34 (Immunoglobulin heavy variable 4-34) has an inherent ability to encode autoreactive antibodies. B cells expressing BCR immunoglobulin using the IGHV4-34 gene are expanded following infections by microbial pathogens including e.g. cytomegalovirus (CMV), Epstein-Barr virus (EBV), Mycoplasma pneumoniae, as well as in certain autoimmune disorders e.g. lupus or chronic lymphocytic leukemia (Xochelli et al. 2017).

AHSG, PGLYRP2, APOM and APOA2 were lower in TACE plus MONOGRAM as well as in XMIN_1, compared with GAinS. S100A9, FN1, C4B, HABP2, CLEC3B, GNMT and four immunoglobulin variable regions (IGKV1-39, IGKV1-5, IGLV3-27, IGLV3-21) were differentially abundant in comparing the citrate plasma cohorts (BIONIC or MOTION) with GAinS, which could be partially (e.g. S100A9) or potentially completely attributed to the difference in anti-coagulants, as discussed in section 3.7.5.

3.6.2 Sepsis-specific changes in the plasma proteome

Overlap at pathway level

In either the discovery cohort or the validation cohort, GAinS samples have been compared with three non-sepsis controls. To understand at the pathway level what are collectively reflected from the differentially abundant proteins, I tested each set of proteins in pathway enrichment analysis applied through XGR (Fang et al. 2016), using both GOBP and REACTOME annotations (Fig. 3.21). Although fold change and significance of individual proteins differ by the exact pair of cohorts being compared, the pathways enriched in all three comparisons indicate a core set of biological processes that are specifically altered in sepsis instead of being identified due to a more random factor in one pair of the comparator groups.

In the discovery cohort, sepsis differed from all control conditions in processes

relating to acute-phase response, neutrophil degranulation, regulation of IGF (insulin-like growth factor) transport and uptake by IGF binding proteins, innate immune system, and post-translational protein modification. All these enriched terms were replicated in the validation cohort comparisons, where sepsis differed from other conditions also in processes relating to extracellular matrix organisation, platelet activation/signalling/aggregation, and cytokine signalling in immune system. As expected, the most number of enrichments were identified in the GAinS-OBB contrast, including toll-like receptor signalling pathway, blood coagulation, integrin cell surface interactions, and GPCR signalling.

I also tested for cellular component enrichment of the differentially abundant proteins against the background of all proteins detected in this dataset without the protein filtering. Although many intracellular proteins have been indicated from the group comparisons, the enriched components were mostly extracellular regions including extracellular exosome and extracellular matrix (Fig. B.13). Proteins differentially abundant between sepsis and healthy volunteers were also enriched in high-density lipoprotein particle, endoplasmic reticulum lumen, and secretory granule lumen.

Overlap at individual protein level

Results from the six group comparisons were summarised in a heatmap Fig. 3.22. The 94 proteins that were differentially abundant in any of the 6 contrasts were clustered into blocks by the mean level of the protein in each of the 11 comparator groups. The seven protein blocks roughly corresponded to different patterns across the groups: from the top to the bottom of the figure, proteins in block 1 were lowest in sepsis; blocks 4 and 5 were higher in sepsis, with the differences in block 5 more prominent; blocks 2 and 6 were the highest and the lowest in healthy cohort, respectively; blocks 3 and 7 are proteins lower or higher in citrate samples, respectively.

Varying protein levels observed in each of the contrasts could reflect not only the biological differences but also intrinsic or random difference between the study

Figure 3.21: Pathway enrichment of differentially abundant proteins. Terms significantly enriched in all three contrasts in either the discovery or validation cohort were shaded in yellow. Horizontal bars indicate 95% confidence intervals of log_2(odds ratio). IGF: insulin-like growth factor; IGFBP: IGF binding protein.

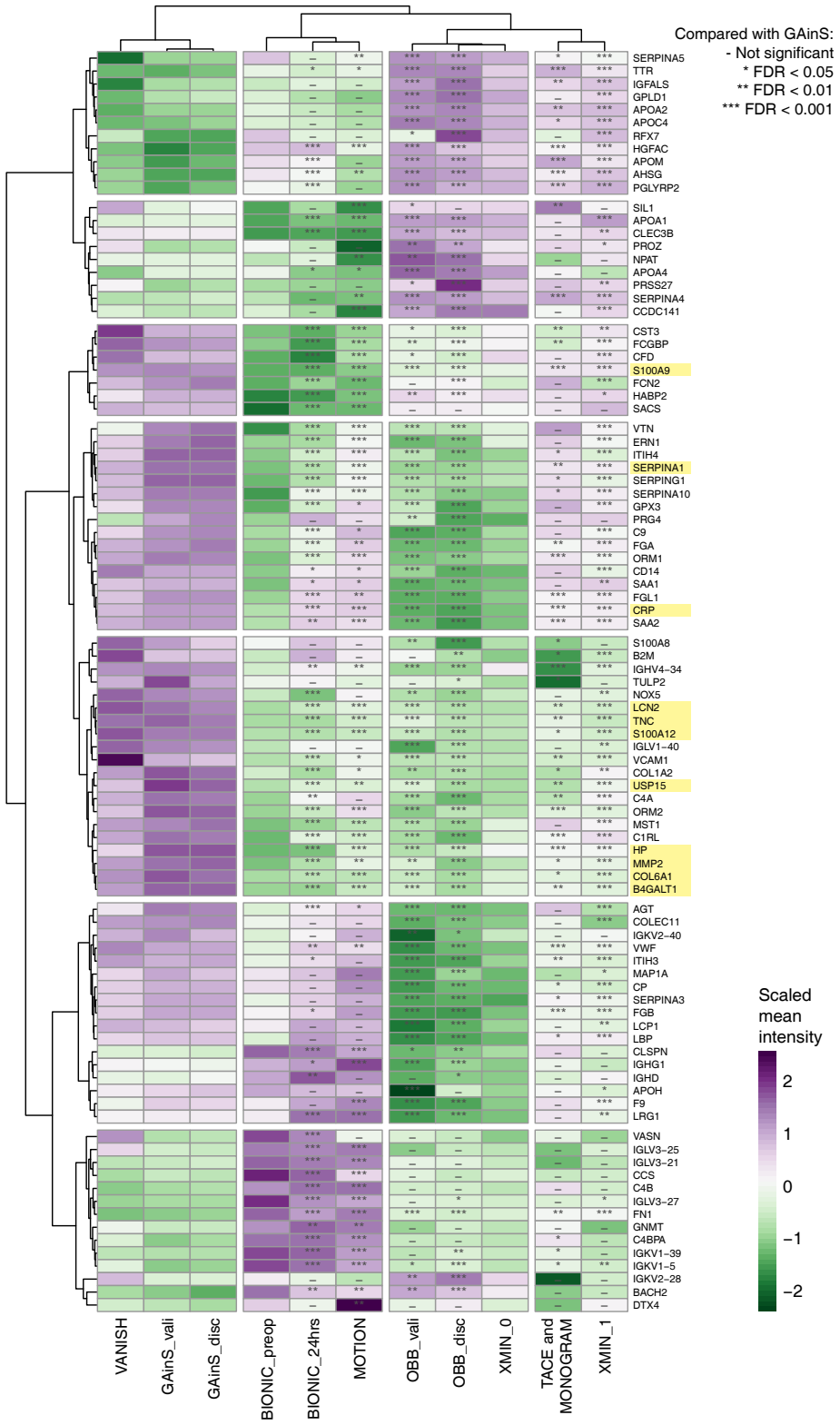


Figure 3.22: Summary heatmap showing the mean protein levels per group. Mean levels were scaled across the groups (rows) for plotting and clustering. 94 proteins that were differentially abundant in any of the six GAinS-control comparisons described were included. For six of the control groups, asterisks or dashes indicate the significance level in comparing this group with the GAinS discovery ("_disc") or validation ("_vali") cohort. 11 proteins that were differentially abundant (FDR<0.05 and |FC|>1.5) in all six contrasts were shaded in yellow. For GAinS or VANISH patients, only the first available samples per patient were included.

populations in e.g. underlying conditions, demographics, a specific reason for admission, and pre-analytical sample handling. Therefore, taking the overlap between multiple comparisons could avoid as much as possible reporting a differential abundance not related to sepsis. There were 13 and 15 differentially abundant proteins, respectively, that overlapped in the three comparisons in either the discovery or validation cohort (Fig. 3.20 venn diagrams), both enriched for neutrophil degranulation. 11 proteins were differentially abundant in all the six contrasts (labelled in Fig. 3.22). The higher abundance of these 11 proteins in sepsis had been repeated in 6 cohorts so was unlikely to reflect any random difference in study populations. Protein abundance across the groups were shown in boxplots for these 11 proteins higher in sepsis, together with eight of the proteins lower in sepsis (Fig. 3.23).

Furthermore, the paired pre-operation samples available for the two surgery cohorts made it possible to understand the plasma proteome change in surgery. Using paired tests between baseline and day 1 samples, there were 6 proteins (SAA1, CRP, LBP, SAA2, FGL1, VMF) and 8 proteins (the six above, plus SERPINA3 and S100A9) that were differentially abundant (FDR<0.05 and |FC|>1.5) in the surgery responses in BIONIC and in XMIN, respectively. The 8 proteins overlapped with the 11 proteins described above on CRP and S100A9. Excluding these two, there were 9 proteins specifically higher in sepsis, which were repeated in the six contrasts and not signficantly different in surgery responses. This included: HP, SERPINA1, B4GALT1, S100A12, TNC, MMP2, COL6A1, LCN2, and USP15.

LCN2 (lipocalin-2) is a multipotent 25kDa protein mainly secreted by neutrophils but also expressed in macrophages, hepatocytes, epithelia, and adipocytes. The best characterised function is that LCN2 obstructs the siderophore iron-acquiring strategy of bacteria and thus inhibits bacterial growth. Besides, LCN2 was shown in mice model to be necessary for neutrophil homeostasis and migration, and for macrophages to induce inflammatory cytokines and phagocytose bacteria (Wang et al. 2019). LCN2 is

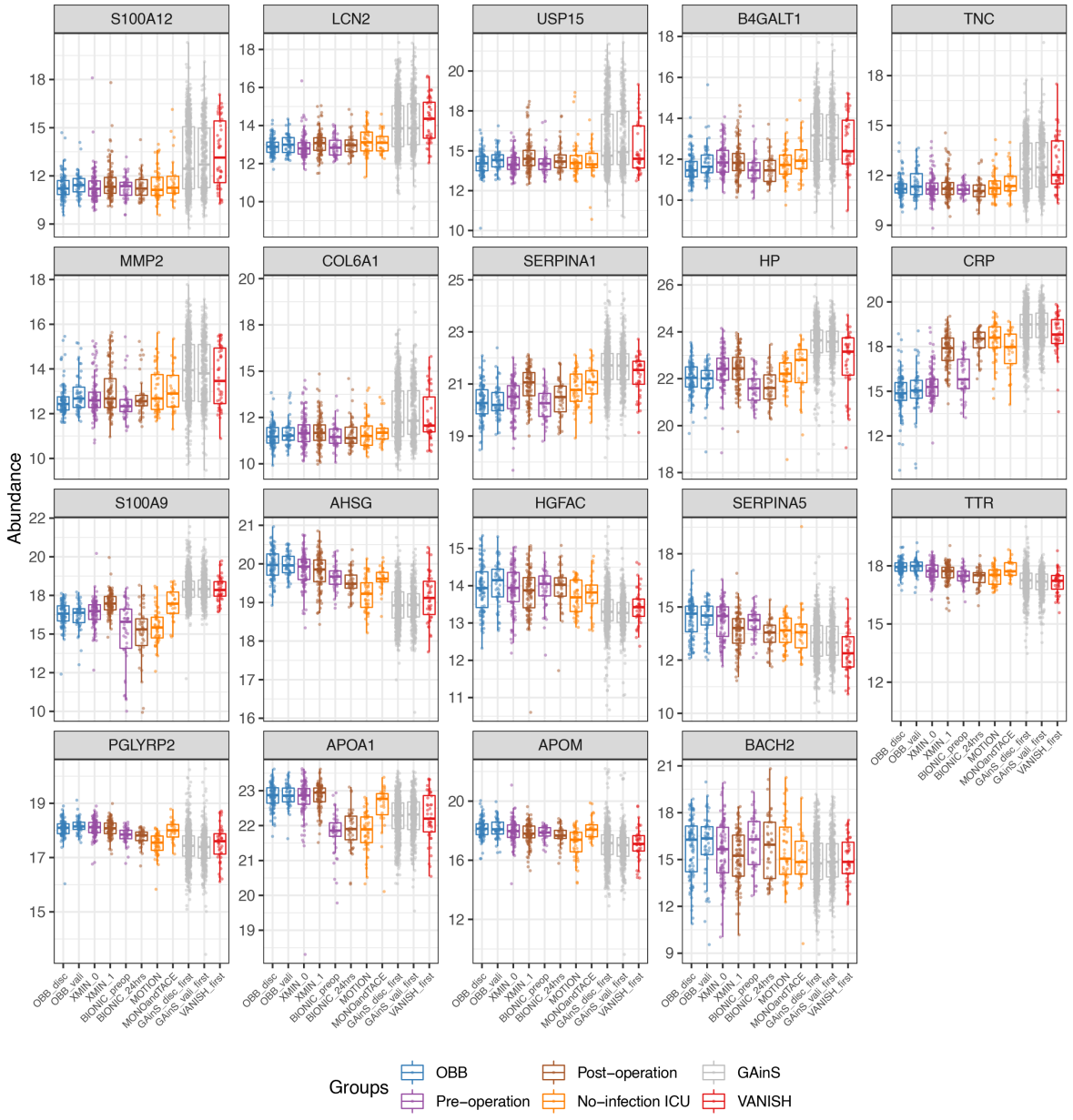


Figure 3.23: Boxplots showing distribution of 19 representative proteins across the comparator groups. Abbreviations in x axis labels: disc – discovery cohort; vali – validation cohort; first – only the first sample of each patient is used; preop – pre-operation; postop – post-operation; MONOandTACE – samples from MONOGRAM and TACE studies combined.

an acute-phase protein with elevated levels in active inflammatory disease. It is also an early marker of renal injury (Mishra et al. 2005).

USP15 (Ubiquitin carboxyl-terminal hydrolase 15) is a positive regulator in TNFα- and IL-1β- induced NF-κB activation (Zhou et al. 2020), which is a transcription factor with pivotal roles in mediating multiple aspects of innate and adaptive immune functions. USP15 is also shown to inhibit anti-tumor T cell responses (Zou et al. 2014). Another protein in the USP family, USP25, reduces the susceptibility of mice to LPS-induced septic shock, although it is not clear whether the direction of regulation of USP15 or USP25 is the same (Zhong et al. 2013). Notably, USP15 is located mainly in cytosol and also in nucleoplasm and nuclear bodies, shown by staining in human cell lines (Human Protein Atlas, Thul et al. 2017). In MS2019 USP15 was detected in plasma from only 3% of the healthy volunteer samples but 40% of sepsis samples.

MMP2 (72kDa type IV collagenase) is a ubiquitinous metalloproteinase that is involved in diverse functions such as angiogenesis, tissue repair, and inflammation. Matrix metalloproteases (MMPs) are a family of zinc-dependent endoproteases that can degrade the extracellular matrix (ECM) proteins. The cleavage of ECM collagen in basement membrane might help neutrophils cross blood and lymph circulation into sites of infection (Parks et al. 2004). MMPs are stimulated by cytokines and also release membrane-bound cytokines such as TNF-α. MMP -3, -8, -9, -19 were also detected in the fractionised library injections in MS2019 but with low detection rates across the samples (<1%).

COL6A1 (Collagen alpha-1(VI) chain) is one of the three major subunits of collagen VI, which is an important ECM protein that interacts with other ECM molecules and cell membranes, providing structural support for cells. The COL6A1 subunit is involved in multiple signalling pathways that regulate apoptosis, proliferation, angiogenesis, fibrosis, and inflammation, reported in a variety of tissue types (Zhu et al. 2015).

SERPINA1 (Alpha-1-antitrypsin) is an acute phase glycoprotein mainly synthesized

in liver and also a granule protein of neutrophils. It is a major blood protein after albumin and the immunoglobulins. SERPINA1 is a well-recognized inhibitor of human neutrophil serine proteases, particularly neutrophil elastase (Janciauskiene et al. 2018), and therefore protects some tissues such as the lower respiratory tract against proteolytic destruction. It is a vital regulator of various neutrophil functions, including inhibition of caspase-3 which is essential in cellular apoptosis. A C-terminal fragment of SERPINA1 was also reported as a discriminatory marker in sepsis with immunomodulatory functions (Blaurock et al. 2016).

Functions of HP, B4GALT1, S100A12, and TNC have been described in section 3.6.1. Overall, all the 9 sepsis-specific proteins except for B4GALT1 have clear functions in inflammation and the innate immune response. Three are acute phase proteins (LCN2, SERPINA1, HP); three are ECM proteins (MMP2, COL6A1, TNC); three have protective role in tissue damage (SERPINA1, HP, TNC); four are closely related to neutrophil function (LCN2, MMP2, SERPINA1, S100A12); seven are either induced by cytokines or induce cytokine production (LCN2, MMP2, USP15, SERPINA1, HP, TNC, S100A12). It should be noted that a same protein-coding gene could show different functions depending on the protein isoform produced (which cannot be distinguished from this dataset), the cell type of expression, and the tissue context of the protein activity. Therefore, the exact part in sepsis pathophysiology played by each of the proteins identified should still be investigated in detail for each individual protein.

3.7 Discussion

3.7.1 Detection limit of the dataset

In mass spectrometry based proteomics datasets, it is important to acknowledge that only a subset of the proteome is being measured and reported. Proteins more likely to be detected are those present in higher abundance or have higher “flyability”, which is a term used to cover the relative efficiencies of ionisation, transfer, and detection in

tandem mass spectrometry (Sinitcyn et al. 2018). Without data for spike-in peptides, it is difficult to know the actual detection range for a specific mass spec experiment, but a rough range for the lower detection limit could be estimated. In MS2019, in the 291 proteins post-filtering, the five proteins with the smallest median intensity were: IGHV3-74, MASP2, PF4, IGHV3-9, and F7. According to the Human Protein Atlas (Uhlén et al. 2019), the reference concentrations measured by immunoassays in healthy plasma for MASP2, PF4, and F7 are 340ng/mL, 7.7ng/mL, and 350 ng/mL, respectively. Considering the study sample sizes, the median level of a protein measured in MS2019 should be within the range in sepsis. MASP2 and F7 were not significantly different between GAinS and OBB. Plasma level of PF4 in acute phase was reported to be 27.1+/-20.1 ng/mL (mean+/-s.d.) (Lorenz and Brauer 1988). Therefore, a rough estimation is that proteins abundant in at least the 10^2 ng/ml magnitude could be detected in this specific experiment.

Many of the known biomarkers lie below this detection limit so will not be detected in this kind of experiments. For example, troponins are one of the most commonly tested biochemical markers for tissue injury for example after surgery. The normal range of troponin in circulation is below 0.04ng/mL. Fong et al. (2021) reported a 5.13 and 4.02 fold increase of troponin I2 and T2 in plasma after elective surgeries. It could then be estimated that the post-surgery circulation level of troponin is in the magnitude of 0.2 ng/mL, which is roughly two magnitudes lower than the estimated lower detection limit of MS2019. Thus changes in troponions would not be expected to be detectable in this dataset.

3.7.2 Applying algorithms from transcriptome data analysis

Compared to proteome quantification, gene expression profiling is a much more mature field with a wide variety of algorithms developed. Because of the similarity of microarray intensities and mass spectrometry intensities, many of the algorithms could be well applied to proteome data to facilitate the processing and data analysis. In

this chapter I demonstrated the successful applications of methods/packages like VSN, KNN, Normalyzer, limma, WGCNA, and XGR. Proteomics downstream analysis could benefit greatly from methods developed for transcriptomics analysis, so researchers should pay close attention to new transcriptomics analysis methods and any works demonstrating the applicability of the transcriptomics methods to certain types of proteomics data.

One example is the application of co-expression network analysis to proteomics datasets. Other than examining individual protein variations, protein co-expression modules can serve as a useful link between function interpretations based on pathway enrichments, and the module-level differentiation between biological groups or correlations to clinical traits. In MS2019, the small sizes of the co-expression modules I identified (median size = 10) had added more randomicity to the enrichment tests and thus limited the gain in switching from protein-based analysis to module-based analysis. In proteomics datasets with potentially higher depth and fewer samples, WGCNA may be a promising approach in utilising the co-expression relations to combat the multiple testing issue, while in my analysis in MS2019 the approach was applied with an emphasis on understanding the plasma proteome network and supporting the relation between co-expression patterns and protein interactions.

At the same time, not all algorithms developed for gene expression can be readily applied to mass spectrometry data. For example, variable filtering and imputation in a sparse matrix (e.g. >80% zero values) with unknown subsets is also a common issue in single-cell RNA sequencing (scRNA-seq) data, where some genes are detected in only certain cell subsets but not the others. Markov affinity-based graph imputation of cells (MAGIC) is a method commonly applied in scRNA-seq, that uses diffusion of values between similar cells along an affinity-based graph structure to impute missing gene counts, correct and denoise the cell-count matrix, and recovers gene-gene relations and cell subsets (Dijk et al. 2018). With the application of MAGIC to MS2019 we might be able to apply a very lenient filter to retain as many as 500+ proteins, followed

by sensible imputation to the missingness. However, we decided that we should not apply MAGIC on our mass spectrometry dataset because (a) the mechanisms behind missing values in scRNA-seq and mass spec are intrinsically different; (b) data input of RNA read counts or mass spec intensities take different forms of distribution; (c) the emphasis of downstream analysis is different in discovering cell subsets or an accurate differential abundance analysis, whereas a more strict filtering strategy should be applied in the latter case.

3.7.3 Implications for plasma sample collection

Our study showed that even when following the same protocol in one study, clinical plasma samples could exhibit large variation in cellular contaminations when the sample size is large enough. While it is widely recognised in extracellular miRNA measurements that the presence of residual cells can have a significant influence (Mitchell et al. 2016), our study showed the same case for proteins measurements. This potential contamination by cellular residues is often overlooked in plasma protein studies, as the details of plasma collection and storage is rarely reported, and that the sample quality-associated proteins are frequently reported as candidate biomarkers (Geyer et al. 2019).

Is there a way then to avoid the variation by modifying the collection protocol, to make the most out of the samples collected and the plasma proteins measured?

In our cohort, aside from OBB and XMIN, the other studies all used similar protocols of the more standard plasma sample collection, which was a one-step centrifugation at 1000–1861g for 10–15 min at 4°C. Aliquots of the supernatant plasma layer were then taken and immediately stored at -20°C or -80°C freezer. In XMIN, whole blood was centrifuged at 3500g for 10min which was roughly twice the centrifugal force as in other studies. The plasma layer was then sterile filtered through a 0.8μm membrane filter to ensure any platelets or erythrocyte lysis were removed and will not interfere with their extracellular miRNA analysis. The level of platelet index was

indeed the lowest and least variable in XMIN. This suggested that aside from the 4-step centrifugation process applied by Geyer et al., one high-speed centrifugation plus a filtering step could also ensure a cell-free plasma component. However, both protocols take much longer in a clinical setting than the standard one-step centrifugation protocol.

Geyer et al. made practical recommendations including to centrifuge blood to generate plasma immediately after collection, to harvest plasma immediately after centrifugation, and to discard the last 500 μL of plasma above the platelet bed or to use a second centrifugation step to generate platelet-poor plasma (PPP, <10000 platelets/μL). Recommendation for miRNA studies also call for PPP, which is generated by centrifuging whole blood collected in anticoagulants twice before taking the plasma aliquot, with the first centrifugation to remove the bulk of circulating cells, and the second to remove residual platelets (Mitchell et al. 2016). There was not a clear protocol in what the speed and duration of the two centrifugations should be, but a rough range of 1500g for 10min for both rounds were believed to likely have removed platelets. To establish a most optimal protocol to both minimise the influence of cellular contaminations in plasma and minimise sample collection efforts, further work should be performed to compare the sample quality-associated proteins in plasma samples collected in the same individuals with the different protocols described.

Even with a refined plasma collection protocol, it is necessary in pre-processing to calculate the contamination indices or to look at individual marker protein levels. This is to add more confidence that any signal observed in downstream analysis reflect true biology but not variation in sample handling.

When samples are already collected and contamination is observed, how should this effect be corrected for in the plasma protein measurements? When there are only a small number of samples clearly outlying from the distribution, the outlying samples which are potentially mis-handled could be removed, as recommended by Geyer et al.

When the distribution is not clear-cut or a larger number of samples are affected, it could be considered to adjust the affected proteins as in MS2019. Notably, those proteins enriched in the cellular components could still be at detectable baseline levels in cell-free plasma due to the life cycle of these cells. Geyer et al. also reported detecting the marker proteins in pure plasma. In studies where the sample quality is more consistent, these proteins could also be proposed as biomarkers with thorough validation. The more stringent approach was applied here by removing the proteins most affected by the cell residue contaminations. These proteins could also potentially be biologically interesting. For example, among the removed proteins, PF4 and PPBP are also released from platelet granules so the plasma levels could reflect platelet activation; changes in secreted F13A1 may reflect activation of the coagulation cascade. However, in contaminated samples, the true biological signal cannot be disentangled from the technical variation so these measurements need to be removed.

3.7.4 Cause of the batch effect

The pattern on PC1 scores before batch correction was not fully explained by the point where the chromatographic column was changed, thus there would be at least two mechanisms causing the batch effect. One is column health which should gradually decay instead of dropping sharply between plates. Proteins with large loadings on PC1 tend to be the lower-abundant ones especially the immunoglobulins, suggesting PC1 before batch correction may reflect a sensitivity issue resulted from the column gradually blocking. The second mechanism could be certain factors in sample preparation or chromatography issues that could affect proteins with certain characteristics. I calculated the GRAVY (grand average of hydropathicity) index based on the protein sequences to represent protein hydrophobicity, which was not correlated with protein loadings on PC1 before batch correction. The mechanism causing the batch effect remains not fully clear although the technical variation can be effectively removed by determining batches based on randomised sample PC1 scores

and repeated pool injections as described in section 3.3.6.

3.7.5 EDTA or citrate as anti-coagulant

In blood sampling for proteomic studies, EDTA and citrate have been widely used as calcium chelators to inhibit coagulation in plasma samples and were also recommended by the Human Proteome Organization in 2005 (Omenn et al. 2005). Among the two, it is known that different anti-coagulants used can introduce variation on the quantitative results of blood protein measurements. However, it is not clear which specific protein types are affected and how they are affected. Also there is not a protein with consistent abundance in the blood stream across biological conditions that could be used as internal controls, as GAPDH for gene expression and actin for intracellular protein expression. In the cohorts included in MS2019, citrate was used in BIONIC and MOTION; heparin was used in three patients in TACE; EDTA was used in GAinS and other cohorts. Therefore the difference in EDTA or citrate used would mostly affect the differential abundance observed in the GAinS-BIONIC and GAinS-MOTION contrasts.

To understand which proteins are affected by the sample types, I used a published dataset containing EDTA and citrate plasma from the same six individuals analysed in one mass spec experiment (Ilies et al. 2017). Assuming that any difference observed between the 6 and 6 samples were due to anti-coagulants but not technical variations, I extracted lists of proteins that were only detected in EDTA or citrate plasma, or were detected in both but were significantly different between the two (student t-tests, FDR <0.05). I counted for the proteins that overlaped between this list and the proteins I identified in MS2019 to be differentially abundant between GAinS (EDTA) and BIONIC_24hr (citrate), or between XMIN_0 (EDTA) and BIONIC_preop (citrate), highlighting 12 proteins with a matching direction between the two datasets so the difference could have been driven by anticoagulants (Table 3.10). This included one protein that was higher in EDTA samples (S100A9), and 11 proteins that were higher

in citrate samples (IGKV1-5, VASN, IGLV3-21, FN1, AHSG, C4BPA, APOM, C4B, SERPINA3, IGHG1, IGLC2).

Many MS2019 proteins were not detected in Ilies et al. because the levels should be low in healthy individuals. To account for these proteins as well, proteins that overlapped between these two MS2019 contrasts (GAinS vs BIONIC_24hr, and XMIN_0 vs BONIC_preop) were also included for consideration of the sample-type effect. In addition to the 12 proteins listed above, this included 5 proteins that should be interpreted with care if found higher in EDTA samples (CLEC3B, HABP2, HP, CFD, SACS), and 6 proteins if found higher in citrate samples (IGKV1-39, CCS, CLSPN, IGLV3-27, IGHD, GNMT). There was a moderately strong correlation (Pearson's r=0.59, Fig.B.14) between fold changes of these two contrasts, indicating the differentiation of proteins was affected by anti-coagulants.

Table 3.10: Numbers of overlapping proteins between the dataset published by Ilies et al. and the differentially abundant proteins in MS2019. Numbers in bold are the total counts. Numbers in italics show the proteins where the differential abundance observed in MS2019 is in the same direction as the difference between citrate and EDTA samples in Ilies et al. Abbreviations: E – EDTA samples; C – citrate samples.

		Total in Ilies et al.	MS2019 (t-test) GAinS vs BIONIC_24hr		MS2019 (rank test) GAinS vs BIONIC_24hr		MS2019 (t-test) BIONIC_preop vs XMIN_0	
			Up in GAinS (E)	Up in BIONIC (C)	Up in GAinS (E)	Up in BIONIC (C)	Up in XMIN (E)	Up in BIONIC (C)
Total in MS2019			**27**	**16**	**73**	**104**	**15**	**19**
Detected in MS2019 but not in Ilies et al		**119**	12	7	30	46	5	11
Only detected in EDTA in Ilies et al.		**90**	*1* (S100A9)	0	*5*	0	*1* (S100A9)	0
Only detected in citrate in Ilies et al.		**73**	0	*3* (IGKV1-5, VASN, IGLV3-21)	1	*4*	0	*3* (IGKV1-5, VASN, IGLV3-21)
Detected in both EDTA and citrate in Ilies et al.	Higher in E	**22**	0	0	*1*	0	0	0
	Higher in C	**159**	10	*5* (FN1, AHSG, C4BPA, APOM, C4B)	34	*49*	7	*5* (FN1, C4BPA, SERPINA3, IGHG1, IGLC2)
	Not different	**40**	4	1	7	5	2	0

Ilies et al. calculated correlation in ranks instead of LFQ intensities between sample types, thus I tested to see whether using ranks could improve the correlation in differential abundance analysis. I converted their published data into protein ranks within each sample and compared between EDTA/citrate using rank-sum tests. Using ranks in their dataset improved the correlations between EDTA/citrate samples of each individual from ~0.76 to ~0.84, and decreased the number of proteins significantly

different (FDR <0.05) from 181/221 to 106/221. Thus I also tested the GAinS-BIONIC_24hr contrast using ranks. However, as there was not an efficient cut-off on fold changes, the number of differentially abundant proteins was inflated. There were 49 proteins up in BIONIC that overlapped with the proteins higher in citrate from the Ilies et al. dataset. Therefore, converting to ranks did not alleviate the sample-type-effect in MS2019. For comparing cohorts with different anti-coagulants, I decided to still use t-tests applied through limma which is consistent with the other EDTA-EDTA contrasts, taking note of the 12 plus 11 proteins that were more likely to be affected by the sample-type-effect.

Notably, although HP, AHSG, and APOM were among this list to be interpreted with care, their differential abundance as described in the result section were also among the top significant ones in the EDTA-EDTA contrast between GAinS and XMIN_1, indicating true biological difference.

3.7.6 Considerations in non-sepsis control cohorts

A key factor in selecting control samples is that they should be free from infections. Notably, 13 of the 43 BIONIC patients later developed pneumonia with identified microbial infection. Understanding the factors predisposing to post-surgery infection is another research question in BIONIC itself. The symptom onset of pneumonia were no earlier than 5 days post-operation. The actual time of infection is unknown, however, clinically it is considered unlikely that infections had already developed at the 24hrs or 48hrs sampling timepoint. Using all BIONIC samples for PCA also showed a separation by sampling timepoint along PC2 but no separation in samples from patients who later developed infection or not (Fig. B.15). Therefore, the 24hrs post-operation samples for all 43 BIONIC patients were included as the sterile inflammation control for sepsis.

Compared with sepsis patients, the two surgery cohorts had much higher occurrence of malignancy (38/42 (90.5%) for BIONIC, 51/103 (49.5%) for XMIN). For many patients

this were what the surgeries were performed for. To understand whether this would be a major confounder, I compared samples from XMIN patients with or without known malignancy, in either the pre-operation or the post-operation timepoint. No significantly different proteins were identified in either contrasts, suggesting that malignancy had little effect on the plasma proteome as measured in this dataset. I also performed the GAinS-XMIN_1 contrast including malignancy as a covariate, which produced strongly correlated fold changes (Pearson's r=0.99) and a similar group of differentially abundant proteins as when performing the contrast without accounting for malignancy.

Because of the largely unequal sample sizes in comparing GAinS vs. TACE plus MONOGRAM (394 vs 26), I also used Welch t-tests to perform this contrast which is more conservative when the group size and variance is different between the two groups. The Welch t-tests and student t-tests (applied in limma) produced strongly correlating fold changes (Pearson's r=0.99) and arrived at the same group of differentially abundant proteins.

3.7.7 Conclusion

In this chapter, I characterised the sepsis plasma proteome by constructing the protein interaction and co-expression networks, and comparing between sepsis and a panel of control cohorts. The sepsis response is distinguished by various aspects of the immune response especially in innate immunity, with nine sepsis-specific proteins identified highlighting alterations in acute-phase response, neutrophil function, and ECM organisation. These findings show the value of understanding sepsis pathophysiology on the level of plasma proteome, and provide implications for monitoring the molecular state in sepsis by measuring plasma proteins. In addition, I also presented a refined workflow for pre-processing in mass spectrometry based plasma proteomics datasets, which could be applied to future studies.

4

Heterogeneity in the individual sepsis proteomic response

This chapter elucidates how the sepsis response proteome varies between patients through the identification and characterisation of patient subgroups with distinct molecular profiles and clinical features. Prediction models on the subgroup membership were built for validation in independent cohorts.

4.1 Introduction

Since outcome in sepsis is an overall result from a complex interplay between pathogen, host and environmental factors (Goh and Knight 2017), the differentiation between survivors and non survivors may not best capture the heterogeneity in immune response or host state in sepsis. It has been shown in sepsis and various other disease conditions that patient stratification using omics data could group patients into more homogeneous disease subtypes with different prognosis and potentially different underlying biology, facilitating the development of personalised treatment and improving the outcome of clinical trials (Davenport et al. 2016; Burnham et al. 2017; Antcliffe et al. 2019; Neyton et al. 2022; Niu et al. 2021). In sepsis, despite extensive studies on leukocyte transcriptomic endotypes, no such study has

investigated the molecular patient subgroups at the proteome level with a sufficient sample and untargeted proteome coverage.

Plasma proteins, in particular, have been demonstrated to be important as prognostic or monitoring biomarkers reflecting the physiological status, pathogenic processes, or treatment responses (FDA-NIH: Biomarker Working Group 2016). Proteins and enzymes account for the largest proportion of *in vitro* diagnostics at relatively high throughputs and low costs (Geyer et al. 2017). However, only a limited number of conditions have specific biomarkers available, and few new protein biomarkers have been approved in recent decades (Anderson et al. 2013). This could be attributed to common issues including the limited protein detection depth, varying sample quality, underpowered study designs, and insufficient validation studies, highlighting the demand and potential for protein biomarker discovery in a well-defined dataset with appropriate clinical questions.

In the previous chapter I have demonstrated the informativeness of using the large-scale plasma proteomics dataset MS2019 for characterising the sepsis-specific response. Among the 1860 patient samples from the GAinS study, there was large variation in protein abundance e.g. in Figure 3.23, spanning from levels overlapping with the range of control cohorts, to more extreme levels that suggest a more severe state towards the sepsis end of the spectrum. This indicates that protein levels, as well as gene expression, may be meaningful indicators differentiating patient subtypes in sepsis. The large sample size in MS2019 empowers sepsis patient stratification based on the blood proteome with a high level of granularity. The availability of transcriptome, plasma proteome and cytokine data on overlapping sets of samples provides the opportunity to understand interaction between the molecular subphenotypes.

4.1.1 The complementary GAinS proteomics datasets

MS192 is a mass spectrometry (QE-HF) based sepsis proteomics dataset that included 192 plasma samples from 96 CAP sepsis patients, 6 FP sepsis patients, and 21 non-sepsis control patients undergoing pre-elective cardiac surgeries (not used in analysis). Although a smaller scale of samples was assayed compared with MS2019, this dataset was still well-sized and had a high protein detection depth, which could be attributed to both the top-abundant proteins depletion applied, and a longer machine time for each sample. This dataset was used to characterise a wider range of proteins to complement the analysis in MS2019. More details are in section 4.4.

Cytokines are important mediators of the inflammation and immune response but the majority are below the detection range of the MS approaches applied. To cover the quantification of these low-abundance proteins, I used the Luminex assay to measure 65 cytokines and other signalling molecules in 204 samples from 146 GAinS sepsis patients. The Luminex dataset was also used to complement the analysis in section 4.4.

4.1.2 GAinS transcriptomics datasets

Recruitment of patients to GAinS included the collection of not only plasma samples but also leukocytes and other specimens on the Day1, 3, and/or 5 of ICU admission. Genome-wide gene expression from the leukocytes has been analysed in multiple microarray and RNA-seq datasets, including the work leading to the identification of the two SRS transcriptomic endotypes. There are a large number of patient-timepoints that overlap between the transcriptomics and proteomics datasets, providing the opportunity both for integrative analysis and for understanding the molecular subtypes at another omics layer.

Microarray datasets were acquired using Illumina Human-HT-12 v4 Expression BeadChips, full details of which have been described in the previous reports

(Davenport et al. 2016; Burnham et al. 2017). The batch-normalised cleaned-up data measured 22130 genes in 676 samples. Among these, 642 serial samples from 518 patients also had the corresponding patient-timepoints assayed in MS2019.

The RNA-seq dataset was generated on NovaSeq with 100bp paired end reads across 12 lanes. After QC and feature filtering, the final set consisted of 20416 features for 864 samples, including 837 samples from 649 patients overlapping with MS2019. Gene expression was represented as TMM(trimmed mean of M-value)-normalised log-transformed counts per millon.

The SRS (sepsis response signature) transcriptomic endotypes analysed in this chapter were assigned based on these gene expression datasets (Cano-Gamez et al. 2022). The approach is also described in SepstratifieR (Cano-Gamez 2022), a package to stratify patients with suspected infection into groups with different molecular characteristics based on the expression level of a small set of genes. Briefly, SRS classifications derived from the original microarray datasets as well as gene expression profiles from healthy individuals (labelled SRS3) were used as a reference set, then the new data was aligned to the reference set by the mutual nearest neighbours algorithm. SRS classifications were then predicted using a 7-gene set random forest model. For the 11 border-line patient-timepoints that had different assignments from the platforms, assignments were taken in the order of priority as microarray before RNA-seq before qPCR.

A small number (48) of GAinS patient-timepoints included in MS2019 were assigned as SRS2 in the original two-group classification, and as SRS3 in the three-group classification derived from this approach. SRS3 contained a small proportion of individuals in the low severity/recovery spectrum that are transcriptionally closer to healthy volunteers. In analysis of this chapter, SRS3 was combined with SRS2 as a "non-SRS1" category.

4.1.3 Aims

The overall aim of this chapter is to understand the heterogeneity in sepsis response at the plasma proteome level, with the hypothesis that clinically and molecularly distinct proteomic subphenotypes can be identified and replicated in the ICU sepsis cohort. Specifically, I will:

1. use unsupervised approaches to define plasma proteome-based sepsis patient subgroups in the GAinS discovery cohort
2. identify the clinical and molecular characteristics of the proteomic subtypes, including association to outcome
3. build prediction models for cluster membership to validate the clusters in multiple independent sample sets
4. propose lists of proteins as a first step towards biomarkers
5. investigate the interaction between the plasma proteome- or leukocyte transcriptome- based patient subgroups.

4.2 Results: Unsupervised subgroup identification in discovery cohort

4.2.1 Cluster identification

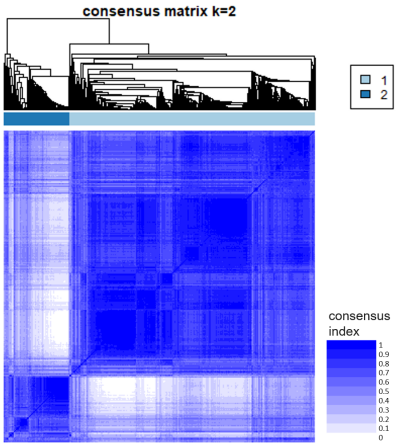
To identify potential subgroups in the GAinS patients and samples in MS2019, I used unsupervised clustering based on protein intensities in the GAinS discovery cohort, which was composed of samples from a randomly selected $\frac{2}{3}$ subset of all GAinS patients in MS2019 as described in Section 3.3.6. The sample composition of MS2019 was shown in Table 3.4.

I applied consensus clustering to generate unsupervised clusters from 1236 samples in

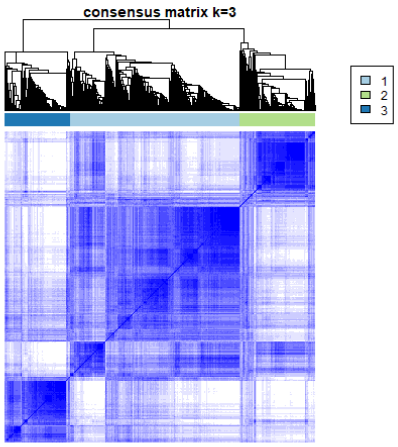
the discovery cohort. Compared with a single-run hierarchical clustering performed on all available samples and features, the consensus clustering approach has the advantage of cluster results being robust to small variations in sample or feature composition. For each cluster number k tested from two to ten, 500 iterations were performed, randomly sampling 80% of samples and 90% of proteins in each iteration (Fig. 4.1a-d). The cumulative distribution function (CDF) curves (Fig. 4.1e) were inspected to assess cluster stability. The increases in area under the CDF were the largest when k increased from 2 to 3, and reached an elbow point at k=4 (Fig. 4.1f). I also assessed how model accuracy improved as cluster number increased by calculating the total within-cluster variation in 10-fold cross-validation with k-means clustering. The fit of test points to the nearest centres improved gradually as the cluster numbers increased from 1 to 10 (Fig. C.1). The step widths were the largest from k=1 to k=3. Therefore, k=3 was chosen as the optimal number of partitioning for this dataset (Fig. 4.1b), balancing the need for explaining most of the variability and for the simplicity of describing the major patient subgroups.

Protein abundance across the clusters was visualised on a heatmap with the samples organised by the dendrogram from consensus clustering at k=3 (Figure 4.2). The difference in pattern for certain branches of proteins was visually observable between the clusters. By this partition, 260/301/675 samples were classified into proteomic clusters 1/2/3 respectively, constituting 21.0%, 24.4%, and 54.6% of the samples (Table 4.1). The order of clusters 1, 2, and 3 was determined based on the order of the branch separating out each cluster from the majority of samples. Henceforth, the three proteome-based patient clusters defined here by consensus clustering are referred to as “ConC1/ConC2/ConC3”.

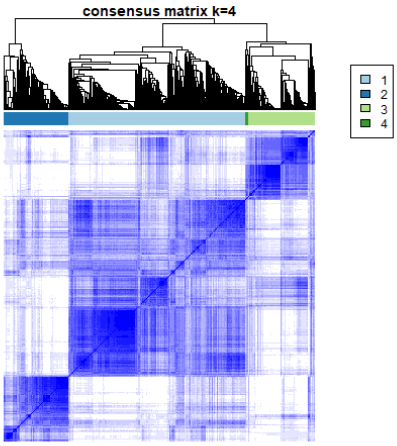
In protein filtering in pre-processing, some proteins had been retained for being detected in subsets of GAinS samples from the raw data (“protein detection groups”). To assess whether these proteins were driving the classification, the three consensus clusters were compared with the four protein detection groups. The two ways of



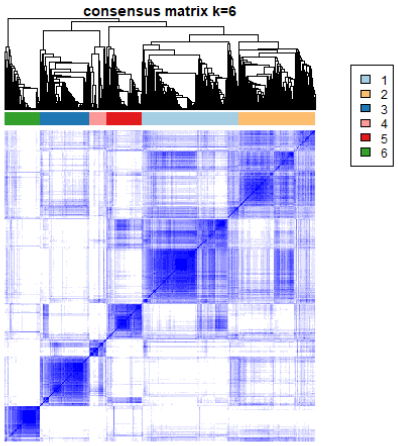
(a) Consensus matrix at k=2.



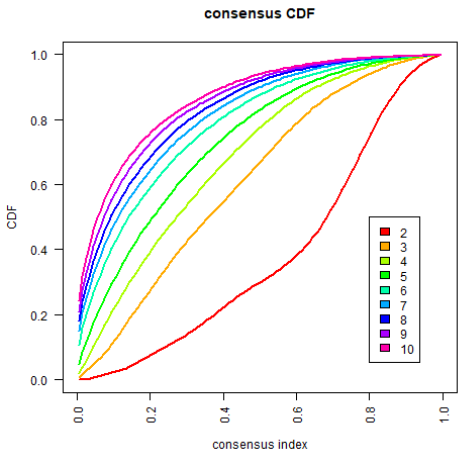
(b) Consensus matrix at k=3.



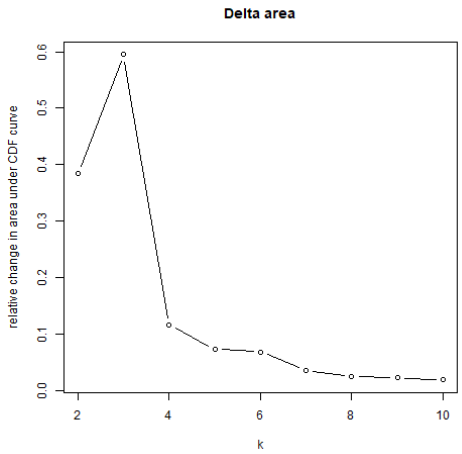
(c) Consensus matrix at k=4.



(d) Consensus matrix at k=6.



(e) CDF of consensus index.



(f) ΔAUCDF

Figure 4.1: Consensus clustering in discovery cohort samples (n=1236). **(a-d)** Cluster dendrogram and heatmap of consensus index between each sample pair are shown for cluster numbers 2, 3, 4, and 6, using the same colour scale for consensus index. A darker blue colour in the consensus matrix indicates a higher frequency of two samples falling in the same cluster. **(e)** Cumulative distribution function (CDF) curve of the consensus index, with an increasing number of clusters (k). **(f)** Relative increase in the area under CDF at each increase of k from 2 to 10.

Table 4.1: Sample numbers of the proteomic clusters in GAinS discovery cohort.

	ConC1	ConC2	ConC3
N(samples)	260	301	675
%samples	21.0%	24.4%	54.6%
N(patients)	170	183	435

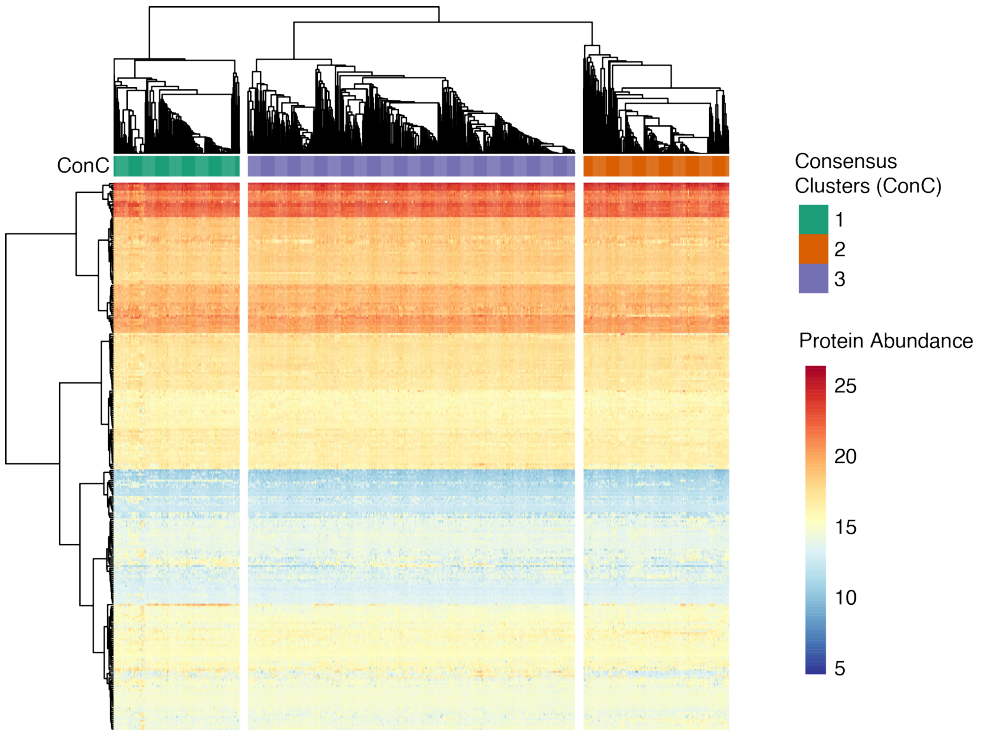
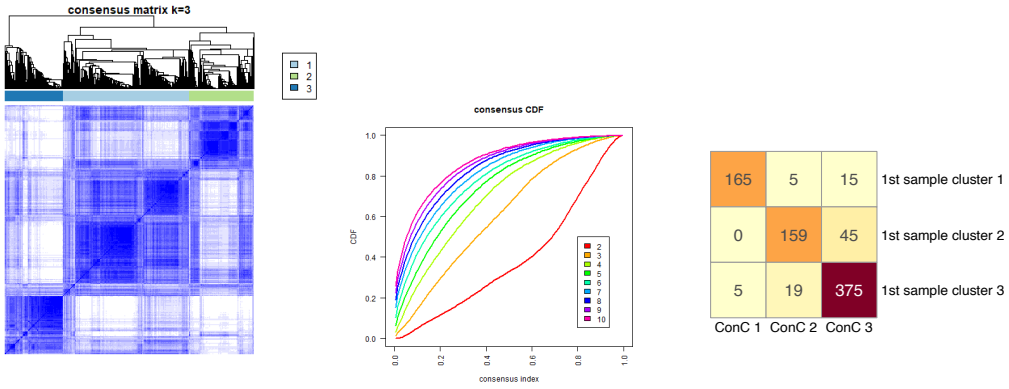


Figure 4.2: Heatmap of protein abundance across clusters. Rows show proteins (n=269) and columns show samples (n=1236). ConC1/2/3 membership is shown on the colour bar below the dendrogram. Colour gradient shows the scale of protein intensities. No extra scaling is performed on the protein intensities in plotting the heatmap.

grouping samples were significantly not independent (χ^2 test p <0.0001), but the overlap was only 48.8% at the greatest. Thus the detection-group-specific proteins contributed to but did not dominate the consensus clusters in processed data.

In stratifying patients based on the molecular profiles, it has been debated whether the subgroups should be defined based on patients (i.e. the discovery set can only consist of single independent samples from each patient), or based on patient-timepoints (i.e. the discovery set can include multiple samples per patient at different timepoints). Both approaches have been applied in the discovery of sepsis transcriptome endotypes (Davenport et al. 2016; Burnham et al. 2017; Scicluna et al. 2017). Restricting to one sample per patient could prevent patients with serial samples having higher weights in the discovery set. However, our aim is to derive a classification that could be applied not only on admission to ICU but on later days as well, and we expect the classification to change for some patients during disease progression. Thus, to achieve an approximation of the true composition of sampled timepoints in clinical settings, I did not restrict the discovery set by the timing of the samples. Furthermore, I applied the same consensus clustering approach to a set consisting of only the first available samples per patient following ICU admission (n=788) and demonstrated that the three-cluster structure also existed at patient level (Fig. 4.3). A high proportion of 88.7% of samples had the same classification as when clustering was performed at timepoint level. Therefore, sample classifications derived from all serial samples were used for further characterisation, and are referred to as "patient subgroups" or "patient clusters" as these classifications stratify patients at a specific timepoint. The patient subgroup membership could change over the stay in ICU.



(a) **(b)** **(c)**

Figure 4.3: Consensus clustering in discovery cohort, first available samples only (n=788). **(a)** Consensus matrix at k=3. Colour scale for consensus index is the same as in Fig. 4.1. **(b)** Cumulative distribution function (CDF) with increasing number of clusters (k). **(c)** Contingency table of the two classifications. The diagonal which summed up to 699 shows the number of samples that had the same classification when consensus clustering was performed at patient level (rows) or at patient-timepoint level (columns).

4.2.2 Comparison of clinical characteristics between clusters

Comparing all measurements

To understand whether the three proteome-based clusters represent patients that are clinically distinct, 66 clinical variables were compared between the three clusters. For patients with multiple timepoints sampled, only the first available timepoint was included in the comparisons. For categorical variables, independence of variable levels and the cluster assignments were tested by χ^2 tests. For numerical variables including the SOFA scores and ARDS classification, the three clusters were compared using the Kruskal-Wallis test by ranks. 32 out of 50 numerical variables had the null hypothesis of Kruskal-Wallis test rejected with FDR <0.05. For these, Dunn's post-hoc test was used to compare each pair of clusters, and the significance of these comparisons are labelled on the box plots (Fig. 4.4) after B-H correction. Significance and sample numbers for all the clinical variables tested are summarised in Table C.1 and Table C.2.

As shown in Figure 4.4 (numerical variables), Figure 4.5 (categorical variables), and Table 4.2 (summary metrics), the signal observed in most of the variables was that

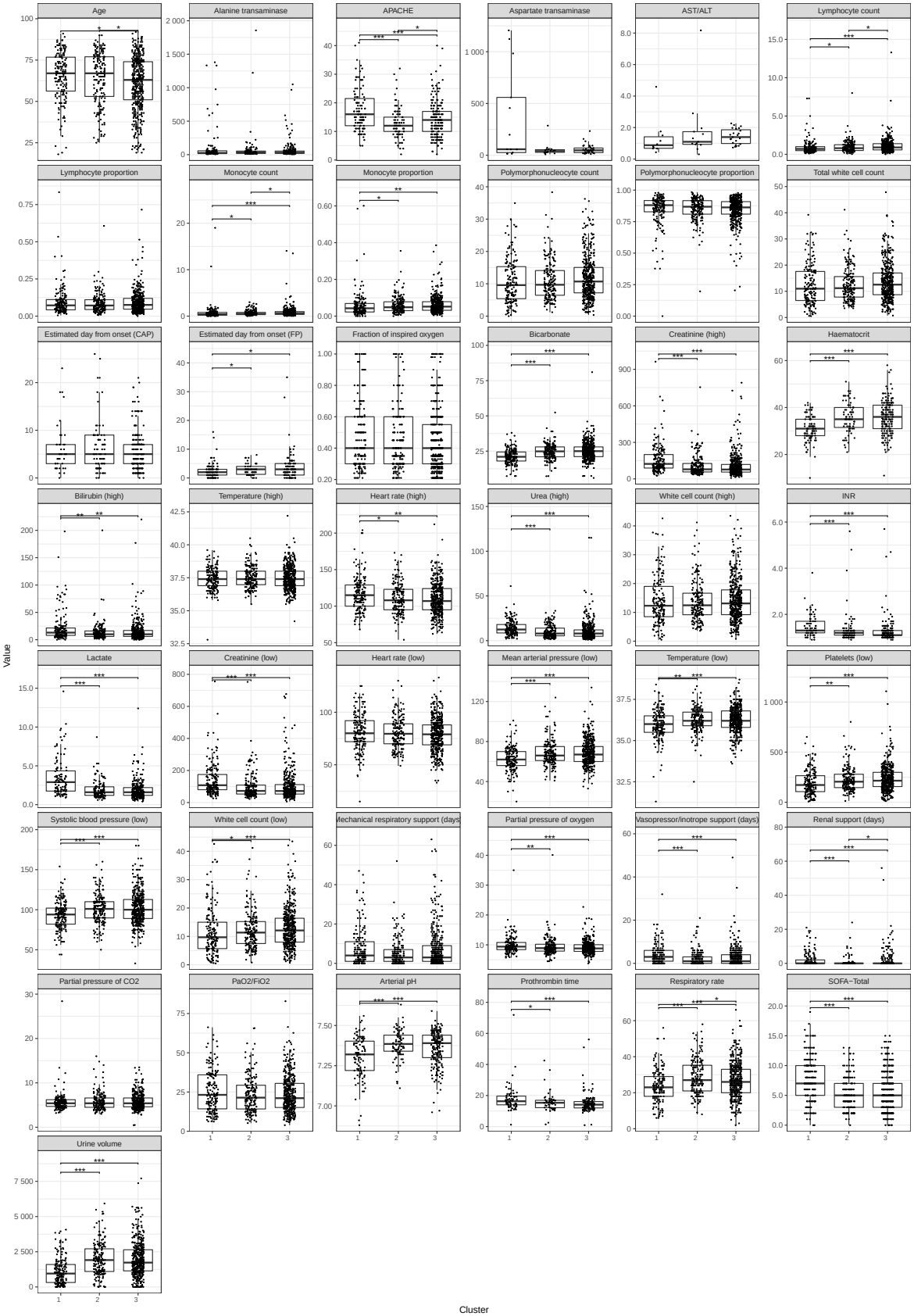


Figure 4.4: Box plots of numerical clinical variables between the three discovery cohort clusters (ConC). Asterisks indicate significance from Dunn's post-hoc tests (* FDR <0.05; ** FDR <0.01; *** FDR <0.001). For visualisation, two ConC3 sample points with INR=14 and 99 were omitted. INR: international normalised ratio for prothrombin time. PaO_2/FiO_2: partial pressure of oxygen divided by fraction of inspired oxygen.

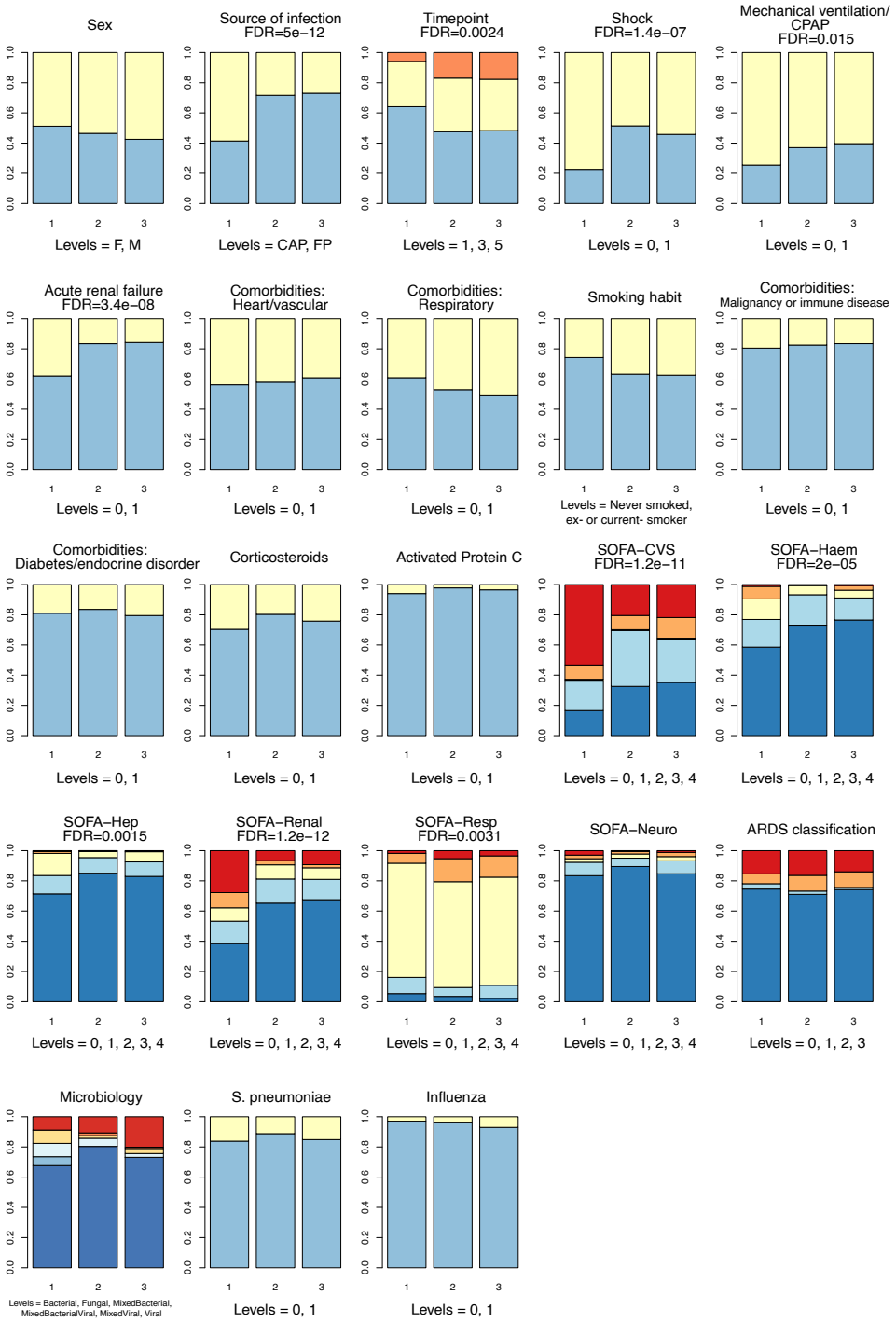


Figure 4.5: Bar plots of categorical clinical variables between the three discovery cohort clusters (ConC). FDR values labelled are from χ^2 tests, except for SOFA organ scores and ARDS classification for which the significance is derived from Kruskal-Wallis tests. Clinical variables without FDR labelled were not significantly different between the clusters by FDR<0.05. For each variable, colours from the bottom to the top of the bars are in the order of the levels stated. For example, in the "Source of infection" variable, blue and yellow stand for proportion of CAP or FP patients in the cluster, respectively. Colours in "Microbiology" correspond to: bacterial (dark blue), fungal (medium blue), mixed bacterial (light blue), mixed bacterial-viral (light orange), mixed viral (dark orange), and viral (red). For variables with two levels 0 and 1, level 0 corresponds to the absence and level 1 the presence of the event stated.

Table 4.2: Summary metrics of the levels of categorical and numerical clinical variables across the three discovery cohort clusters (ConC). Only variables with a significant difference between the clusters were listed. Significance between the cluster pairs is labelled on Fig 4.4 and Fig 4.5, and listed in TableC.1 and Table C.2. Annotation and units of the clinical variables are the same as in Table 3.6.

Clinical variable	ConC1	ConC2	ConC3
	No./total (%)		
Diagnosis - FP	99/169 (59)	51/180 (28)	117/434 (27)
Timepoint - Day1	109/170 (64)	87/183 (48)	210/435 (48)
Timepoint - Day5	10/170 (6)	31/183 (17)	77/435 (18)
Shock	130/168 (77)	87/179 (49)	231/426 (54)
Mechanical ventilation/CPAP	126/169 (75)	114/181 (63)	260/431 (60)
Acute renal failure	64/169 (38)	30/181 (17)	68/431 (16)
	Median (IQR)		
Age	67 (56-77)	67 (53-77)	63 (51-74)
Estimated day from onset (FP)	2 (1-3)	3.0 (1.2-4.0)	3 (1-5)
Systolic blood pressure (low)	94 (82-102)	101 (90-110)	100 (89-113)
Mean arterial pressure (low)	62 (56-70)	66 (61-75)	67 (60-75)
Heart rate (high)	115 (100-129)	108 (95-123)	107 (95-124)
Inotropic support (days)	3 (1-6)	1 (0-3)	1 (0-4)
Arterial pH	7.3 (7.2-7.4)	7.4 (7.3-7.4)	7.4 (7.3-7.4)
Respiratory rate	23 (18-29)	27 (21-35)	26 (20-33)
Partial pressure of oxygen	9.4 (8.4-10.8)	8.9 (7.9-10.2)	8.8 (7.8-10.1)
Lactate	2.9 (1.8-4.3)	1.6 (1.2-2.3)	1.6 (1.2-2.2)
Bicarbonate	21 (18-24)	25 (21-28)	25 (21-28)
Urea (high)	12.4 (8.6-18.0)	8.0 (5.4-14.0)	8.0 (4.9-12.0)
Urine volume	953 (318-1595)	1910 (1100-2710)	1730 (1142-2642)
Creatinine (high)	123 (89-200)	78 (57-126)	77 (56-118)
Creatinine (low)	108 (80-175)	72 (53-109)	72 (53-112)
Renal support (days)	0 (0-2)	0 (0-0)	0 (0-0)
Prothrombin time	16 (14-20)	15 (12-17)	14 (12-16)
The international normalized ratio	1.3 (1.2-1.7)	1.2 (1.1-1.3)	1.2 (1.1-1.3)
Bilirubin (high)	13 (8-22)	10 (6-16)	10 (6-17)
Temperature (low)	36 (36-36)	36 (36-37)	36 (36-37)
Platelets (low)	172 (106-265)	205 (146-280)	215 (156-298)
White cell count (low)	9.7 (5.7-15.0)	11.4 (7.5-15.2)	12 (8-16)
Haematocrit	31 (28-35)	35 (32-40)	36 (31-41)
Lymphocytes (raw)	0.70 (0.48-1.00)	0.8 (0.5-1.2)	0.90 (0.58-1.39)
Monocytes (raw)	0.4 (0.2-0.8)	0.50 (0.33-0.80)	0.64 (0.36-0.98)
Monocytes (proportion)	0.043 (0.023-0.069)	0.048 (0.030-0.078)	0.052 (0.032-0.079)
APACHE	16 (12-22)	12 (10-15)	14 (10-17)
SOFA-Total	7 (5-10)	5 (3-7)	5 (3-7)
SOFA-CVS	4 (1-4)	1 (0-3)	1 (0-3)
SOFA-Haem	0 (0-1)	0 (0-1)	0 (0-0)
SOFA-Hep	0 (0-1)	0 (0-0)	0 (0-0)
SOFA-Renal	1 (0-4)	0 (0-1)	0 (0-1)
SOFA-Resp	2 (2-2)	2 (2-2)	2 (2-2)

ConC1 was the clinically most severe group of patients. For example, compared with both ConC2 and ConC3, patients with the first available sample assigned to ConC1 had: higher APACHE and SOFA scores; lower lymphocyte, monocyte, or white cell counts; lower bicarbonate, arterial pressure, platelets, or urine volume; higher creatinine, bilirubin, heart rate, lactate, or prothrombin time; higher occurrence of shock or acute renal failure; and more days on vasopressors or renal support.

Among the other two clusters, ConC2 patients compared with ConC3 had lower APACHE score, lower lymphocyte or monocyte count, higher proportion with renal support, and higher respiratory rate. ConC1 and ConC2 patients had higher age than ConC3. The earlier timepoints were enriched in ConC1 (FDR=0.0024), which was also significant when testing all samples instead of restricting to the first available samples (p=0.012).

Interaction of clusters with aetiology

These clinical characteristics indicated that ConC1 is the overall more severe group of patients, regarding almost all organ functions except for respiratory functions which could be partly explained by the fact that FP patients were enriched in ConC1. Compared with both ConC2 and ConC3, ConC1 patients had higher blood partial pressure of oxygen, lower respiratory rate, lower proportion with respiratory support, and lower SOFA-respiratory score; although the difference was not significant in PaO_2/FiO_2 ratio or number of days on respiratory support.

The enrichment of FP patients in ConC1, however, cannot explain the differentiation in clinical severity across the clusters. Among the GAinS patients included, CAP and FP patients differed in many aspects of the clinical phenotype but the comparison of overall severity was not clear: CAP patients had higher APACHE score (Mann-Whitney FDR=0.0024), more severe respiratory failure as expected, higher aspartate transaminase, higher 28-day mortality (χ^2 FDR=0.0017); but on the other hand FP patients had lower urine volume, longer prothrombin time,

lower lymphocyte/monocyte counts, tended to have higher 6-month mortality (χ^2 FDR=0.062); and there was no difference in SOFA-total score (full list of variables compared detailed in Table C.3). At the protein level, FP patients as a whole were associated with protein features suggesting higher sepsis severity, including higher abundance of CRP, LCN2, USP15, COL1A2, and MMP2, compared with CAP patients as a whole.

When the numerical or categorical clinical phenotypes were compared within CAP patients only, most of the differences between the clusters were replicated (Fig. C.2, Fig. C.3). ConC1 CAP patients also had longer duration of respiratory support (Fig. 4.6), suggesting that ConC1 represented the more severe group of patients also accounting for respiratory functions.

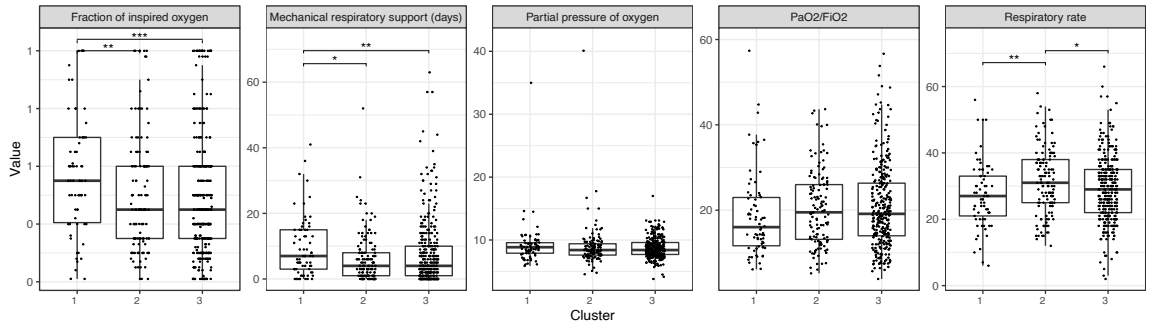


Figure 4.6: Comparison of clinical variables reflecting respiratory functions in CAP patients (n=516), using the mixed aetiology ConC assignment.

To further investigate the question of whether the three-cluster structure is aetiology-specific, I applied the same consensus clustering approach to only the CAP samples (n=811) or only the FP samples (n=420) within the discovery cohort. A four-cluster structure was found in CAP at the elbow point of the AUCDF curve. 87% samples had the same assignments as in the three clusters defined with mixed aetiology, assuming a cluster correspondence based on the sequence of clusters separating out from the majority, and combining the third and fourth CAP clusters. A three-cluster structure was found in FP samples, with 92% overlap in the cluster assignments. In both the CAP clusters and the FP clusters, the clusters were significantly different in 28-day mortality, and the two clusters corresponding to ConC1 (from mixed aetiology

Table 4.3: Comparison of 28-day mortality between the 3 clusters or between each cluster and the other 2 clusters combined, using the Cox proportional hazard model with cluster membership as the only variable. For patients with multiple samples, the cluster assignment from the latest available sample of the patient (among Day 1/3/5) was used.

Comparison	N (patients)	HR (95% CI)	log-rank p
ConC1 vs ConC3	123 vs 451	3.0 (2.0-4.6)	1.30E-07
ConC2 vs ConC3	205 vs 451	1.7 (1.1-2.5)	0.013
ConC1 vs ConC2	123 vs 205	1.8 (1.2-2.8)	0.0081
ConC1 vs others	123 vs 656	2.5 (1.7–3.7)	1.30E-06
ConC2 vs others	205 vs 574	1.2 (0.83-1.7)	0.33
ConC3 vs others	451 vs 328	0.47 (0.33-0.66)	1.50E-05

assignment) had the highest mortality. The high correspondence between clusters assigned restricted or not restricted by aetiology indicated that the three-cluster sample structure is stable to the composition of CAP or FP patients within the cohort. For simplicity, the cluster assignments based on all-cause sepsis (CAP and FP) were used for further characterisations.

Difference in mortality

In comparing outcome across the clusters, patients were significantly different in both acute mortality (28-day after sampling, $p<0.0001$) and late mortality (176-days i.e. roughly 6 months, $p<0.0001$) post-sampling, see Figure 4.7 and Table 4.3. At both timescales ConC1 exhibited the highest mortality and ConC3 the lowest. Comparing the two smaller clusters against the majority group (ConC3) at 28-day, patients from ConC1 had an increased hazard ratio of 3.0 (HR (95%CI) = 3.0 (2.0-4.6), log-rank p <0.0001), patients from ConC2 had an increased HR of 1.7 (HR (95%CI) = 1.7 (1.1-2.5), log-rank p=0.013). Compared with ConC2, ConC1 also had significantly higher risk (HR (95%CI) = 1.8 (1.2-2.8), log-rank p=0.008). Therefore, consensus clustering on the plasma proteome has identified a subgroup of patients (ConC1) that exhibited both higher clinical severity and poorer outcome, plus another subgroup of patients (ConC2) that were less severe than ConC1 but had higher mortality than the majority group (ConC3).

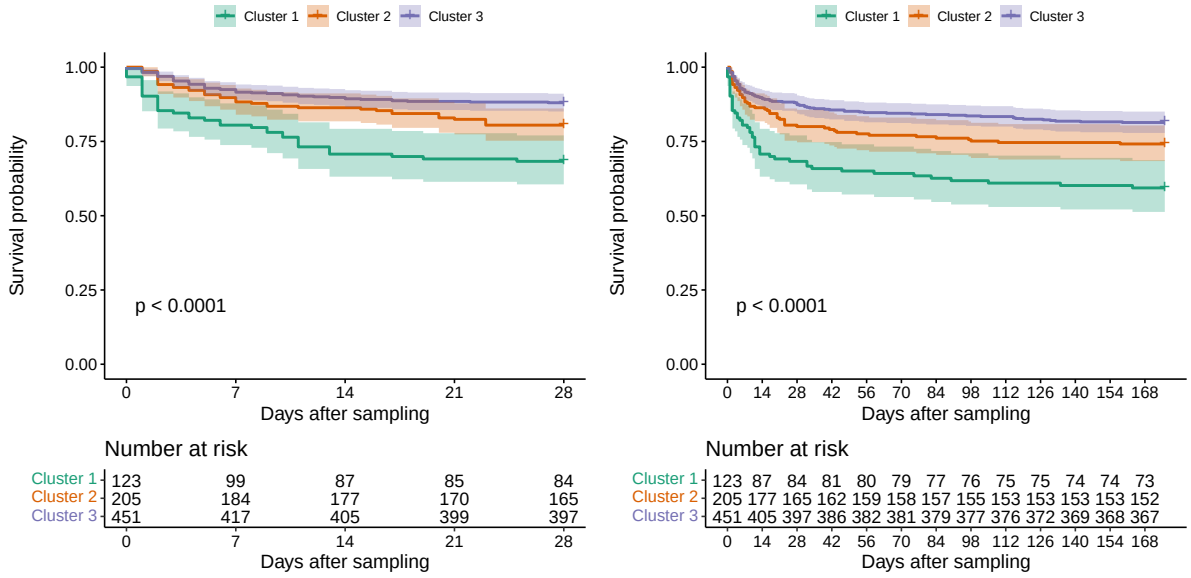


Figure 4.7: Kaplan-Meier curves of consensus clusters in discovery cohort, comparing survival probability at 28 days or 176 days post-sampling. For patients with multiple samples, the cluster assignment from the latest available sample of the patient (among Day 1/3/5) was used. P values by log-rank tests are labelled. Shades on the curves show 95% confidence intervals.

Among non-survivors at 6 months, the recorded causes of death for patients with their latest available sample (among Day1/3/5 of ICU admission) assigned to ConC1 were less due to unrelated cardiac/pulmonary events (0.0%, 5.7%, 2.5% for ConC1/2/3) or other unrelated causes (6.3%, 17.0%, 19.8%), compared with patients assigned to ConC2 or ConC3 (Table C.4). Other recorded causes mainly include failure to resolve organ dysfunction and persistent or recurrent sepsis.

4.2.3 Comparison of protein profiles between clusters

Distribution of clusters after dimension reduction

I then sought to understand the difference between the clusters at the protein level. Samples from the three clusters were visualised alongside non-sepsis control samples after PCA transformation. Although there was considerable overlap between these sample groups, along the first principal component, there was a transition from OBB healthy volunteer samples, to the pre- and post- operation samples from XMIN, and

then the sepsis clusters with ConC1 at the farthest end (Fig. 4.8). Such a spread was also observed after the non-linear transformation of UMAP. On UMAP there was less overlap between the sepsis clusters, which is as expected since the algorithm is also cluster-based and is optimised for separating the clusters it identifies.

Comparison to healthy volunteer

To understand the differences in protein levels in each cluster relative to a consistent control group, I contrasted samples from each cluster with the healthy volunteer (HV) samples in the discovery cohort. For contrasts of Cluster 1/2/3 vs HV, 81/74/55 proteins were differentially abundant, respectively (FDR<0.05 and |FC|>1.5). Thirty-eight of these differentially abundant proteins were shared between the three contrasts (Figure 4.9(a)). The ConC1-HV contrast had the largest number of unique proteins (n=28), including 4 proteins for immunoglobulin constant regions, 3 for variable regions, and 3 apolipoproteins. GOBP terms enriched in these 28 proteins included phagocytosis and positive regulation of B cell activation (Figure 4.9(b)), suggesting that ConC1 may be associated with altered activity of B cell signalling.

Although the differentially abundant proteins were largely shared between the three clusters, the magnitude of difference varied. Log fold changes of 111 proteins differentially abundant in any of the contrasts were plotted as a heatmap (Figure 4.9(c)). A group of proteins that were shown to be higher in sepsis compared with controls (in the previous chapter) had the highest levels in ConC1, including USP15, COL1A2, CRP, S100A12, VWF, COL6A1, MMP2, and LCP1. A prominent feature of ConC2 was the lower abundance of MAP1A, GNMT and IGLV3-21, which were higher or not significant in ConC1 or ConC3 compared with HV. Among the 16 proteins unique to the ConC2-HV contrast, there were also 4 immunoglobulins, which were lower in ConC2 relative to HV, indicating that specific aspects of humoral immunity were altered in this subgroup of patients.

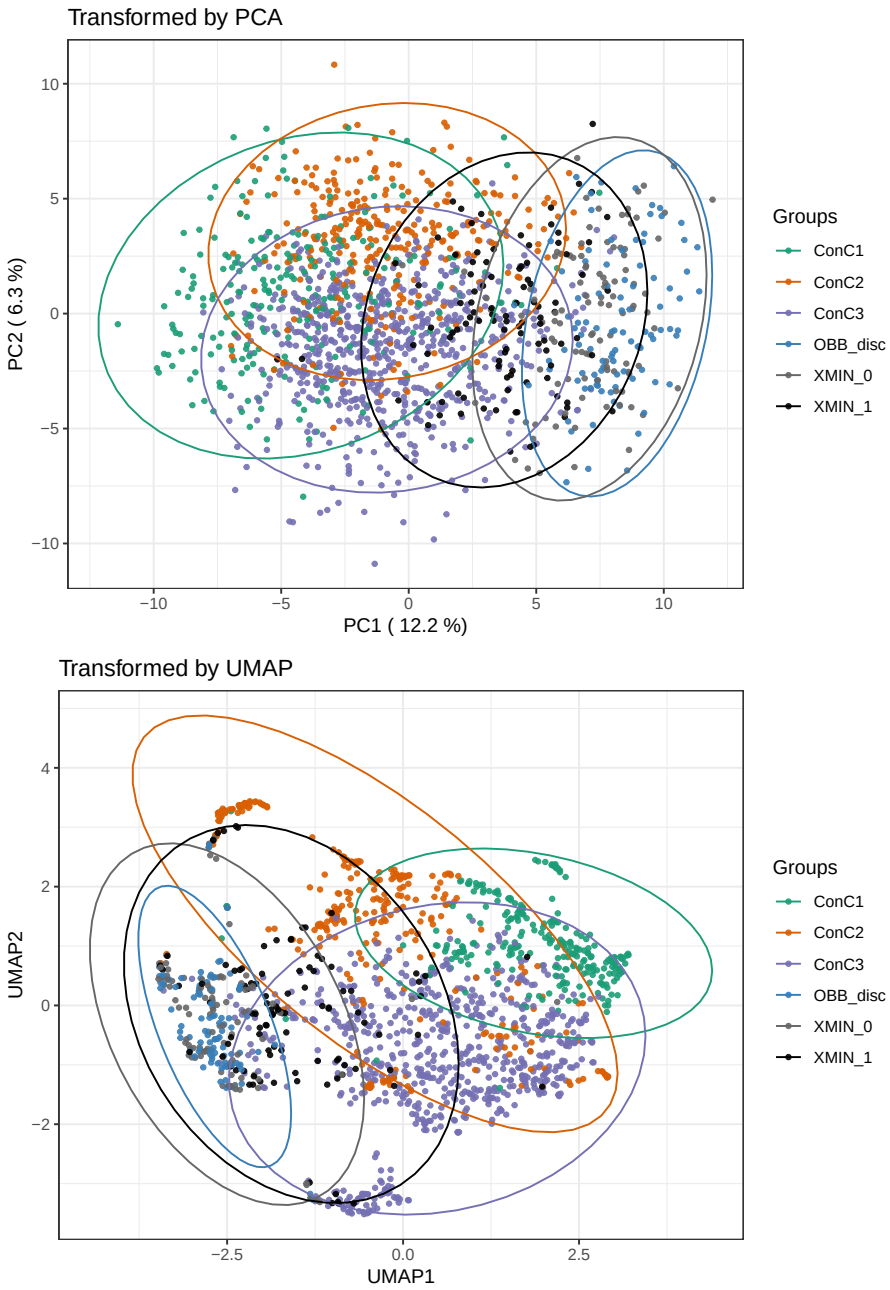
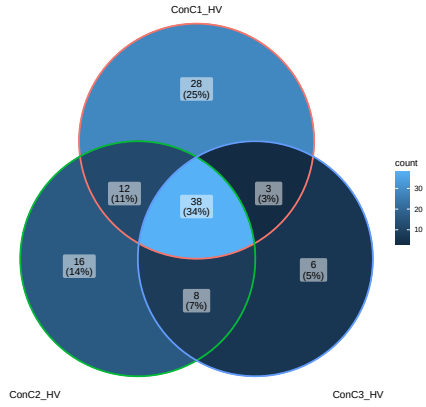
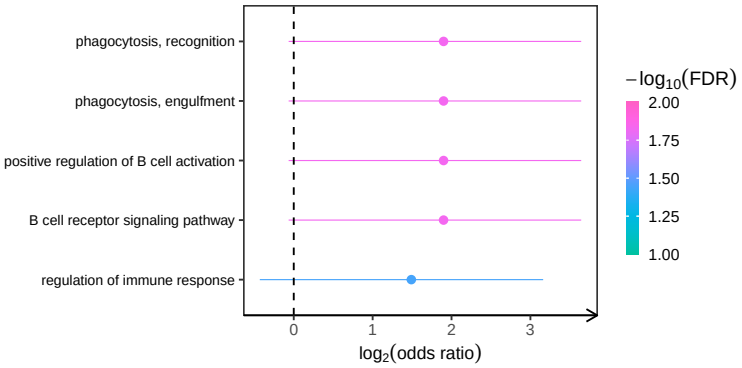


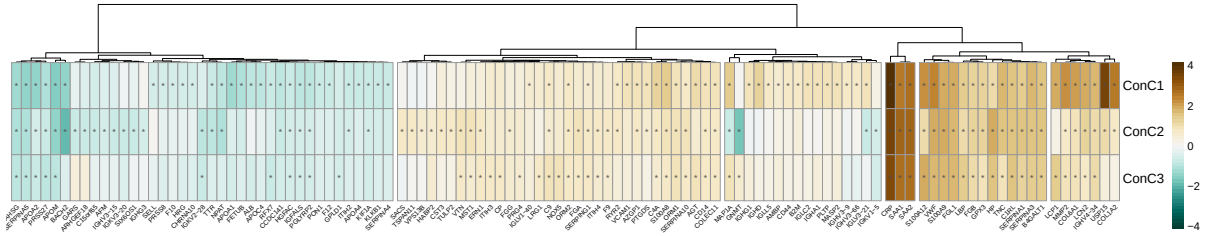
Figure 4.8: Consensus clusters visualised after transformation by PCA or UMAP, in discovery cohort. Data ellipses based on multivariate normal distributions are shown for each comparator group.



(a) Overlap between the differentially abundant proteins.



(b) Pathways enriched in 28 proteins unique to the ConC1-HV contrast.



(c) LogFC of contrasts of each cluster with HV.

Figure 4.9: Comparison of protein profiles of each cluster with HV, in discovery cohort. Only first available samples of each patient were used in the contrasts. Age and sex were included as covariates. Percentage of protein numbers out of the total 111 differentially abundant proteins (FDR<0.05 and |FC|>1.5) are shown in (a). GOBP annotations were tested in (b). Asterisks in (c) denote FDR<0.05 & |FC|>1.5 for the contrast. Order of columns and rows in the heatmap were rearranged by clustering on the logFC values.

ConC1 vs ConC3

To understand the unique protein profile pattern of each cluster in the context of sepsis, I compared each of the two smaller clusters ConC1 and ConC2 against the majority cluster ConC3 which comprised more than half of the samples or patients. There were 20 proteins (most with higher abundance) and 13 proteins (most with lower abundance) that were differentially abundant (DA) in the ConC1-ConC3 and the ConC2-ConC3 contrast, respectively. (FDR<0.05 and |FC|>1.5, Figure 4.10 (a, b)). Only three DA proteins overlapped: BACH2 and ARHGEF18 in the same direction between the two contrasts, and LCP1 in the opposite direction.

Proteins more abundant in ConC1 vs ConC3 were enriched for multiple immune-related pathways, including interleukin signalling (COL1A2, LCN2, LCP1, MMP2, S100A12), Fc-γ or Fc-ϵ receptor signalling pathway, leukocyte migration, complement activation, and ECM organisation (Fig. 4.10 (d, e)). For example, although LCP1 (lymphocyte cytosolic protein 1) is well known as an actin-binding protein that modulates actin dynamics and induce cell adhesion, it is also indicated to play a role in T cell activation and modulates cell surface expression of IL-2RA (interleukin-2 receptor alpha chain) (Wabnitz et al. 2007).

USP15 and COL1A2 were the top two proteins with higher abundance in ConC1. USP15 (Ubiquitin carboxyl-terminal hydrolase 15) is a positive regulator in TNFα- and IL-1β- induced NF-κB activation. In analysis described in the previous chapter, USP15 was among the nine proteins that were differentially abundant in all sepsis-control contrasts but not in the surgery response. COL1A2 (Collagen type I α-2 chain) is a fibril-forming collagen found in most connective tissues. COL1A2 was also significantly higher (FDR<0.05) in sepsis compared with the controls, but the fold changes did not reach the threshold.

Proteins less abundant in ConC1 included: APOA2, Apolipoprotein A-II, which affects HDL (high density lipoprotein) metabolism; BACH2, a transcription regulator that

modulates apoptosis and adaptive immunity, which was low in sepsis compared with healthy control; and ARHGEF18, which acts as a guanine nucleotide exchange factor, induces formation of actin stress fibers and production of reactive oxygen species, and regulates epithelial morphogenesis.

ConC2 vs ConC3

MAP1A and GNMT were the top two proteins with lower abundance in ConC2 compared with ConC3 (Fig. 4.10 (b)). MAP1A (Microtubule-associated protein 1A) is a structural protein that mediates filamentous cross-bridging between microtubules and other skeletal elements; GNMT (Glycine N-methyltransferase) has a possible crucial role in methionine metabolism. Notably, many of the proteins less abundant in ConC2 (MAP1A, GNMT, BACH2, ARHGEF18, LCP1, HR) are annotated to have intracellular locations like the cytoplasm, nucleus, cell membrane, and cytoskeleton (UniProt, 2021).

There were 36 proteins less abundant in ConC2 vs Conc3 using a less stringent threshold of FDR<0.05 and |FC|>1.2, enriched for GOBP terms involving phagocytosis, B cell activation, complement activation, and defence response to bacterium, which were all also enriched in proteins higher in ConC1 vs ConC2. There was a modest negative correlation (Pearson's r = -0.52) between the fold changes in the two contrasts ConC1-ConC2 and ConC2-ConC3.

ConC1 vs ConC2

There were 23 proteins more abundant and 10 proteins less abundant in ConC1 compared with ConC2 (FDR<0.05 and |FC|>1.5, Fig. 4.10 (c)). Proteins higher in ConC1 were mostly immune function related proteins and immunoglobulins, reflecting the feature of the two smaller clusters when compared with ConC3. Proteins lower in ConC1 included among others albumin, apolipoproteins, and proteins mediating cell adhesion (VTN, TSPAN11), enriched for lipoprotein metabolic process

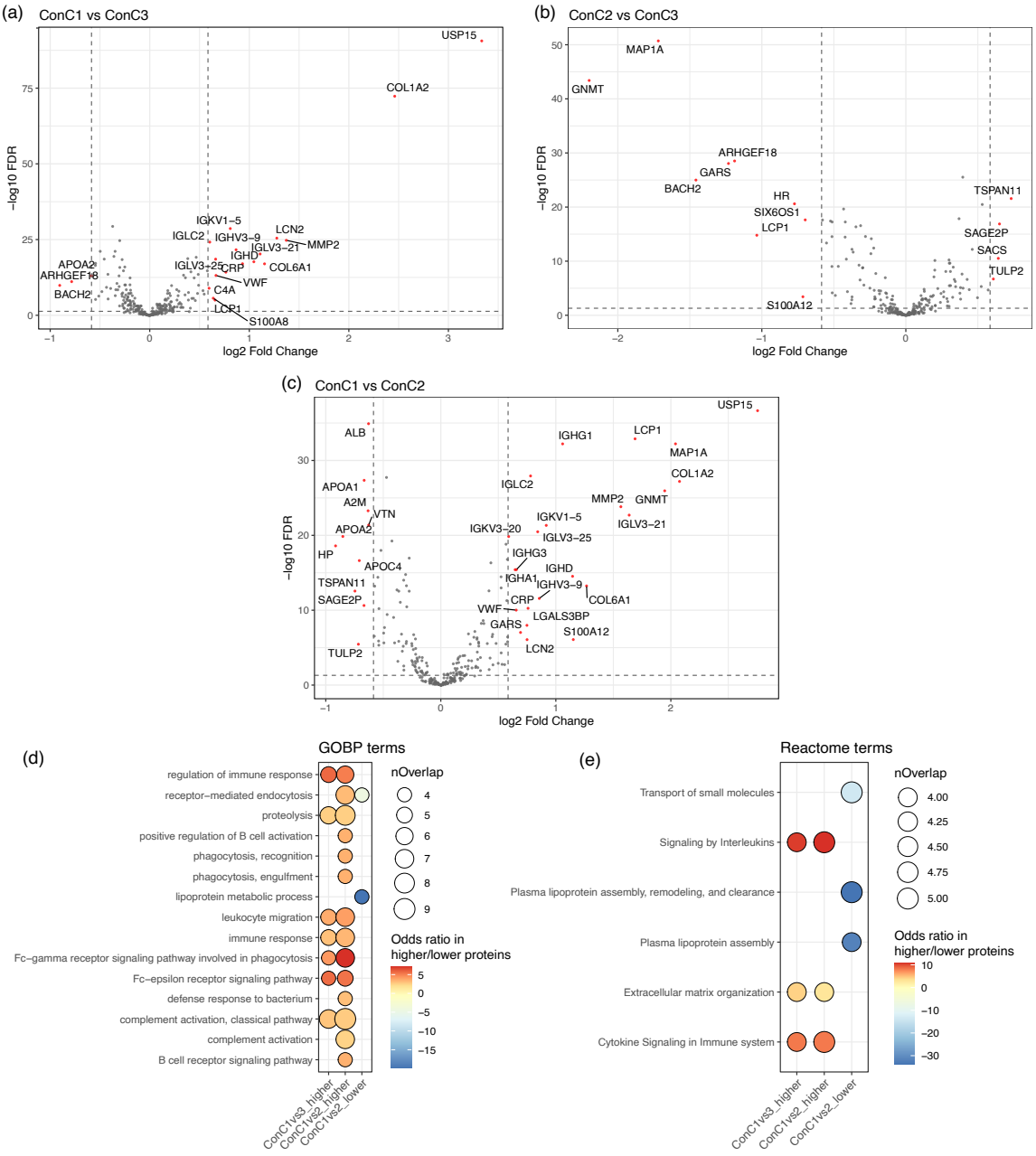


Figure 4.10: Comparison of protein profiles between the 3 clusters, in discovery cohort. (a-c) Volcano plots of the contrasts between each cluster pair. Proteins on the right-hand side of each plot are more abundant in ConC1/2/1, respectively. Red points labelled by gene names denote differentially abundant proteins (FDR<0.05 and |FC|>1.5, grey dashed lines indicate these thresholds). For patients with more than one samples, only the first available samples were included in the contrasts. Age and sex were included as covariates. **(d, e)** GOBP or Reactome terms enriched in the differentially abundant proteins with either higher or lower abundance in each of the three contrasts. A term needs to have a minimum of 4 overlapping proteins with the data input to be tested. Odds ratios are shown on a scale from white to red for more abundant proteins in the corresponding clusters; and on a scale from white to blue for less abundant proteins. Only protein sets with any significant terms detected in the corresponding annotation are shown. For example, 9 proteins less abundant in ConC2 vs ConC3 (FDR<0.05 and |FC|>1.5) were not enriched for any term, so "ConC2vs3_lower" is not shown on (d) or (e).

at the pathway level (Fig. 4.10 (d, e)). Interestingly, HP (haptoglobin) was higher in all sepsis-control contrasts but was less abundant in ConC1 vs ConC2. HP is induced by IL-6 and regulates the acute phase response. There was a relatively strong correlation between the fold changes in the ConC1-ConC2 and ConC1-ConC3 contrasts (Pearson's r=0.73). Distribution of the aforementioned proteins across the three clusters are available in Figure 4.11.

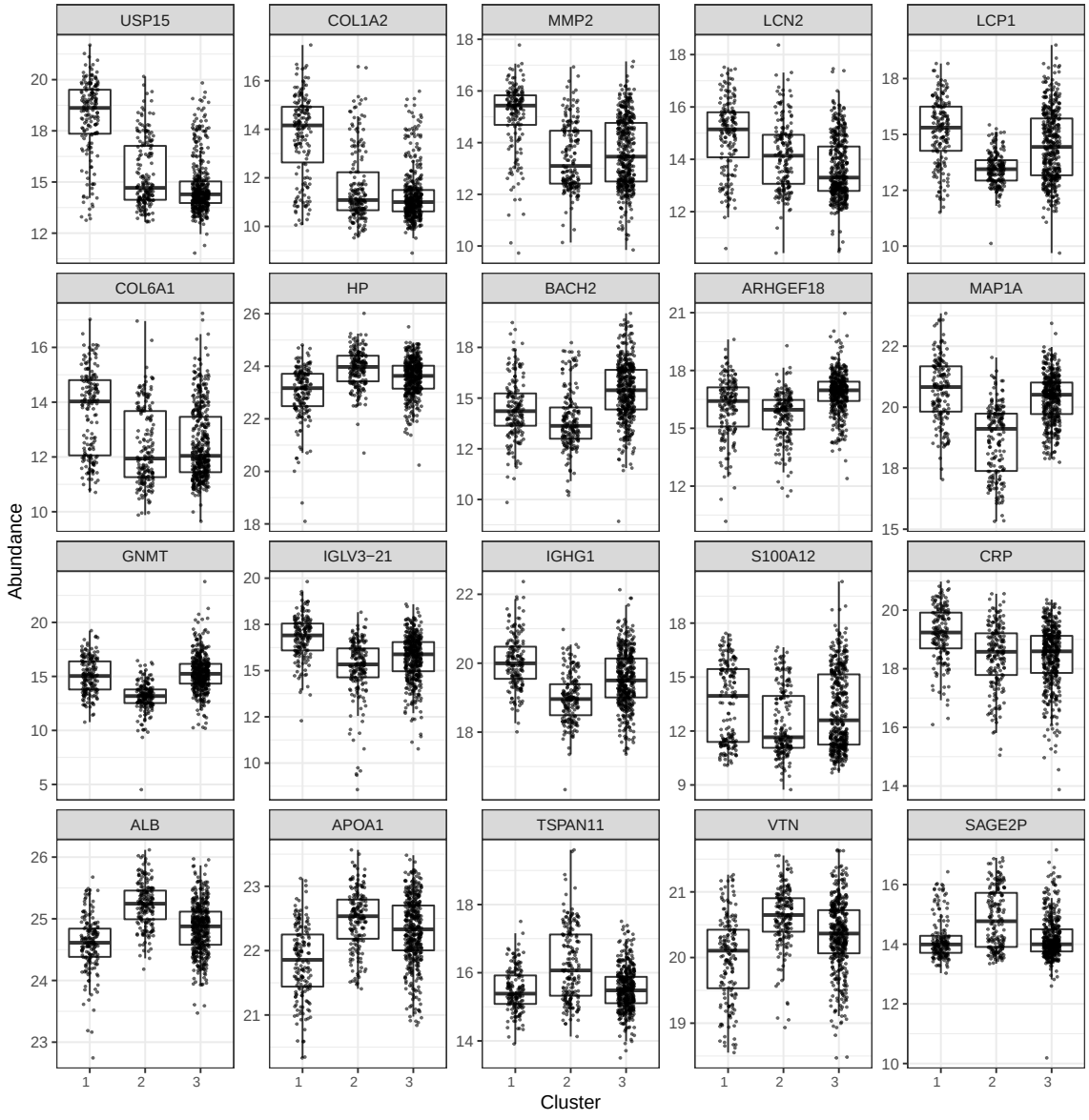


Figure 4.11: Boxplots of protein abundance across the three clusters, in discovery cohort. Proteins mentioned in the text are plotted. Distributions of certain proteins (e.g. LCN2, S100A12) appear bi-modal. These are the proteins with missingness in more than 40% samples in raw data and thus a random-draw from a down-shifted distribution was performed during imputation in data pre-processing.

4.3 Results: Predicting discovery cohort based subgroups in the validation cohort

To validate the three patient clusters in independent samples, cluster prediction models were built based on the discovery cohort samples, and the model with the highest accuracy was applied to the validation cohort. The predicted clusters in validation cohort were then characterised to see if the distinctions in molecular and clinical phenotypes could be replicated.

In evaluating the prediction model performances, accuracy is the sum of the diagonal of the contingency table divided by the total number; sensitivity for a particular event is the proportion of events correctly predicted, out of the number of true events; precision is the proportion of events correctly predicted, out of the number of predicted events.

4.3.1 Developing three-cluster prediction models

To find the prediction model that has the best performance on this dataset, I tested three statistical learning methods including partial least squares discriminative analysis (PLS-DA), generalised linear models (GLM), and class prediction by random forest. For the training and testing of each of the models, the GAinS discovery cohort samples were randomly split into a training set (80% samples, n=992) for building the predictive model, and a test set (20% samples, n=244) for evaluating the model.

As there was not a need to restrict the number of proteins included in the model, first I tried PLS-DA with all the 269 proteins as predictors. Plotting the proportion of variation explained and y-label permutation suggested that 6 orthogonal predictive components had a significant and moderate-high proportion of variation explained in $\frac{6}{7}$ samples of the training set (R2Y value = 0.642, pR2Y=0.01) or in the remaining $\frac{1}{7}$ samples in 7-fold cross-validation (Q2Y value = 0.543, pQ2Y=0.01) (Figure C.4).

This model had a relatively high training set accuracy of 91.3% samples with a correct 3-cluster prediction. However, the test-set accuracy showed a drop to 84.0%, which could potentially be improved by restricting the number of predictors to prevent overfitting. Thus I then restricted the input for PLS-DA model to 138 proteins that were differentially abundant (DA) between any pair of the 3 discovery cohort clusters at a less stringent threshold (FDR<0.05 and |FC|>1.2). The proportion of variation explained in cross-validation, and the training or test set accuracy were all close to the 269-protein PLS-DA model (Table 4.4).

Therefore, as well as restricting the number of predictors, I then tested including an overfitting penalty by applying generalised linear models, with three ways of calculating the penalty (lasso, ridge, and elastic net) to enable variable selection. For lasso and ridge models, the input candidate proteins were restricted to the 138 DA proteins. For the more flexible elastic net models, 138 or 269 input proteins were tested as well as 41 proteins DA between either pair of the 3 clusters at a larger fold change (FDR<0.05 and |FC|>1.5). Symmetric multinomial models were fitted so that each of the 3 classes were represented by a GLM. Two parameters were optimised as described in Methods: the tuning parameter λ which controls the overall strength of the penalty by defining the amount of coefficient shrinkage; and the elastic net mixing parameter α which controls whether the penalty behaves more towards a lasso (α=1) or ridge (α=0) regression for correlated predictors.

The best tuning parameter of α for the three elastic net models were all less than 0.15 so more inclined to ridge regression where more variables are retained. As expected, the tuned lasso model included the smallest number of proteins (46/138) and the three tuned elastic net models produced similar and the highest accuracies. The elastic net model with the most number of input proteins had the best test set accuracy of 91.4% samples predicted correctly into one of the 3 clusters. 181 out of 269 proteins were included as predictors in this model, suggesting a marginal benefit in including the non-DA proteins to improve the model performance in this dataset. The parameters

Table 4.4: Summary on three-cluster prediction models. The fourth column shows the numbers of protein predictors included in the models out of the numbers input to train the models. ConC assignments based on consensus clustering in the GAinS discovery cohort were taken as the ground truth. The training set accuracy and test set accuracy are based on three-cluster predictions. ConC1 sensitivity is the number of true ConC1 samples successfully predicted out of the total number of true ConC1, in test set.

Algorithm	α	λ	N(predictors)/N(input)	Training set accuracy	Test set accuracy	ConC1 sensitivity
lasso	1	0.033	46/138	896/992 = 90.3%	208/244 = 85.2%	42/55 = 76.4%
ridge	0	0.214	138/138	907/992 = 91.4%	216/244 = 88.5%	44/55 = 80.0%
elastic net	0.100	0.0232	41/41	894/992 = 90.1%	221/244 = 90.6%	45/55 = 81.8%
elastic net	0.100	0.0559	124/138	923/992 = 93.0%	221/244 = 90.6%	46/55 = 83.6%
elastic net	0.147	0.0559	181/269	941/992 = 94.9%	223/244 = 91.4%	47/55 = 85.5%
PLS-DA	-	-	138/138	894/992 = 90.1%	209/244 = 85.7%	42/55 = 76.4%
PLS-DA	-	-	269/269	906/992 = 91.3%	205/244 = 84.0%	42/55 = 74.6%
random forest	-	-	269/269	965/992 = 97.3%	210/244 = 86.1%	37/55 = 67.3%

selected and the accuracy of each model are listed in Table 4.4.

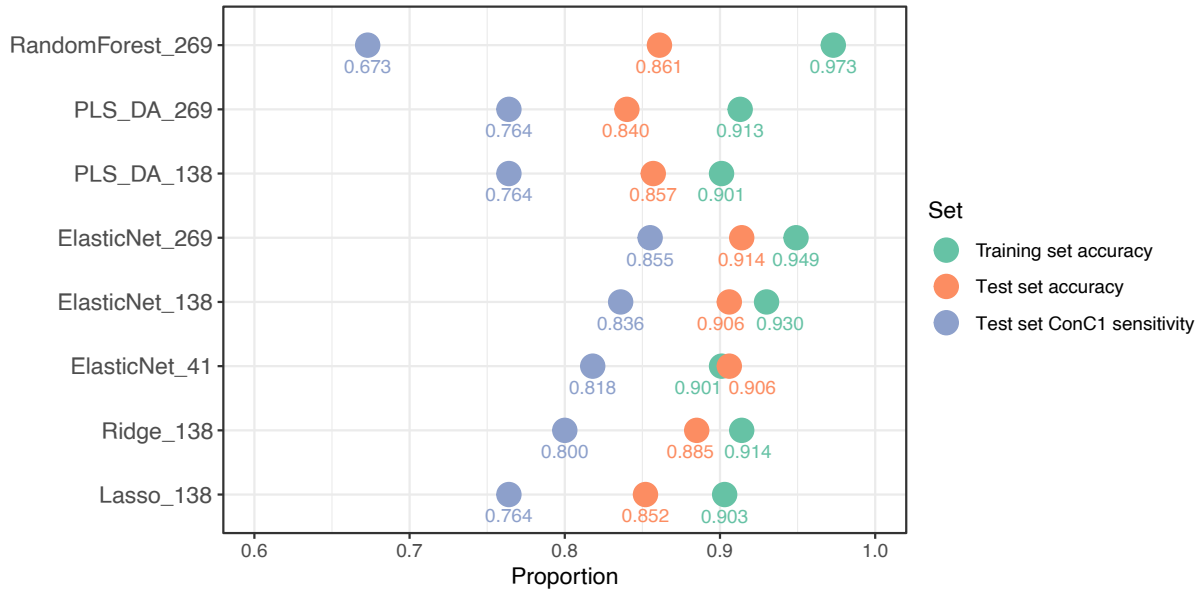


Figure 4.12: Comparison of 3-cluster prediction model performances. This is a scatter plot of the 8 models and three ways of calculating the accuracy as described in Table 4.4. Numbers in y axis labels indicate the numbers of protein candidates input to train each model.

Compared with linear regression models, the random forest algorithm has the potential advantages that it: performs better when the data contain non-linear associations; does not make assumptions for predictor distribution; can handle multicollinearity better; and utilises both bagging of samples and random selection of variables. Therefore, I also tried random forest to see if it fitted the data structure better and could further improve the predictions. Random forest chooses a random

subset of features to build each decision tree, and then predicts the class that get the most votes from the trees. Since the elastic net model with 269 protein candidates had the best performance so far, 269 proteins were used as candidate inputs to train the random forest. The 3-cluster accuracy from 10-fold cross validation was used as the optimising metric for tuning the parameters one-by-one. In the optimised model, 35 variables are randomly sampled for each iteration; the maximum number of terminal nodes of each tree is 42; 350 trees will be trained; the other parameters were kept as default values in the package. Tuning the parameters overall had only a small effect on the accuracy obtained (<5% difference). The optimised random forest model had a high train-set accuracy of 97.3%, but the test set accuracy dropped to 86.1%.

The test-set cluster assignments across the 8 models were shown on a heatmap (Figure C.5), alongside the true classification derived from unsupervised consensus clustering. Assignments were consistent across the models for most samples while being variable for the minority. The elastic net model with 269 protein candidates giving 181 protein predictors ("ElasticNet_269", or "the 181-protein 3-cluster model") had the highest test-set accuracy (Figure 4.12) so was selected as the best-performance model and applied to the validation cohort. With this model, 125 samples (20.0%), 127 samples (20.4%), and 372 (59.6%) samples in the validation cohort were assigned to ConC1/2/3, respectively (Table 4.5), which were similar proportions to those in the discovery cohort (Table 4.1). The model coefficients show a variable level of contribution from the 181 predictor proteins. The top proteins contributing to each of the 3 clusters are plotted in Fig. 4.13, and were not necessarily the top differentially abundant proteins in the between-cluster contrasts.

Table 4.5: Sample numbers of predicted proteomic clusters in the GAinS validation cohort.

Predicted:	**ConC1**	**ConC2**	**ConC3**
N(samples)	125	127	372
%samples	20.0%	20.4%	59.6%
N(patients)	89	74	231

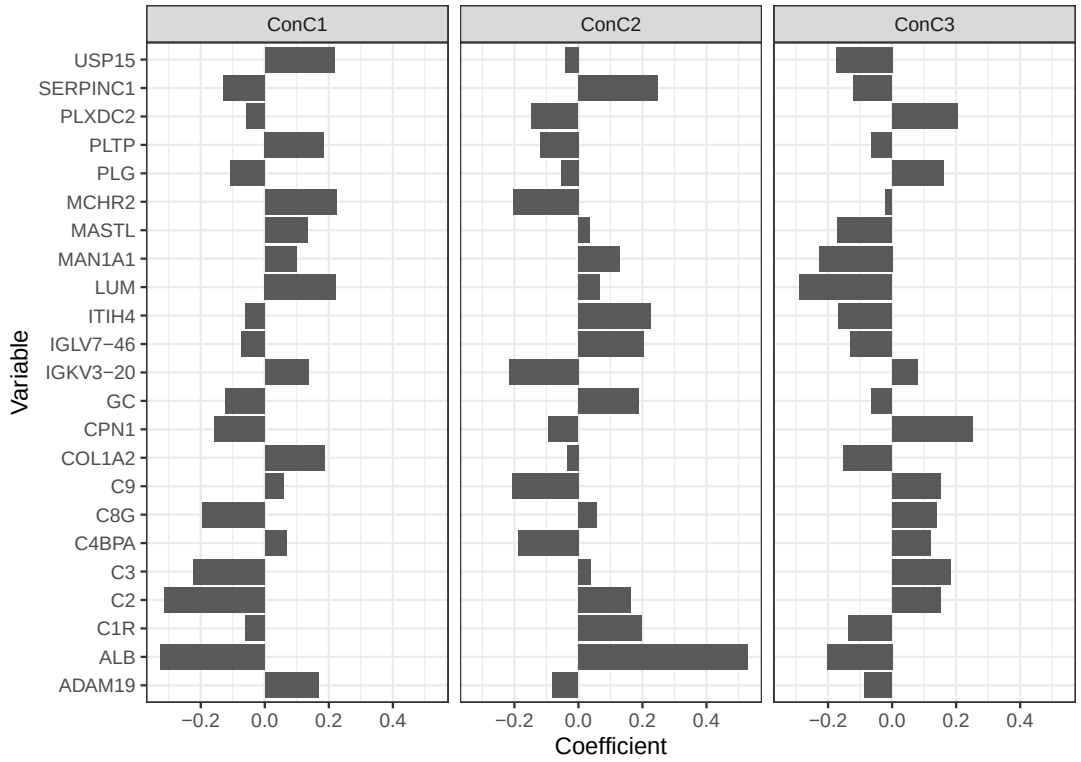


Figure 4.13: Top protein contributions to the best-performance elastic net model. Top 10 proteins with the largest absolute values of coefficients were included for each of the 3 models, consolidating to a list of 23 proteins. A positive coefficient indicates that a higher abundance of the corresponding protein predict the sample to be in the cluster; while a negative coefficient indicates that a lower abundance predict the samples to be in the cluster.

4.3.2 Characterisation of the three predicted clusters in validation cohort

To assess whether the predicted validation cohort clusters capture the same distinctions in clinical and molecular profiles as the discovery cohort clusters, I compared the clinical variables and protein abundance between the three predicted clusters using the same approach as in the discovery cohort. Only clusters assignments for the first available sample of each patient were included.

Comparison of clinical characteristics

The distribution of the numerical or categorical clinical variables across the three predicted clusters (ConC) are shown in Fig. C.7 and Fig. C.8, with the significance and group sizes detailed in Table C.5 and Table C.6.

Table 4.6 summarises which differences in clinical phenotypes between the clusters were replicated in the predicted clusters. Consistent with the discovery cohort, predicted ConC1 was the most clinically severe group of patients, for example in being higher in lactate and creatinine, lower in bicarbonate and blood pressure, having higher SOFA scores and being enriched for patients with shock or with acute renal failure. The enrichment of FP patients and earlier timepoints in ConC1 was also replicated. It was observed in both cohorts that ConC2 patients had higher lactate than ConC3 patients so could potentially be more severe.

Partly due to the reduced sample size, in some variables the difference between the predicted clusters showed the same trend as in the discovery cohort but did not reach significance: ConC1 tended to have longer prothrombin time and higher heart rate compared with both ConC2 and ConC3. ConC2 tended to have higher age and lower lymphocyte count compared with ConC3. It was not clear how respiratory functions compared between the mixed aetiology clusters.

In addition, both predicted ConC1 and ConC2 had higher proportion of

Table 4.6: Summary of differences in clinical phenotypes across the proteomics clusters in either the discovery or validation cohort. Significance was taken from Dunn's post-hoc tests or χ^2 tests. In "ConC A vs ConC B", the direction of "lower" or "higher" is indicated in ConC A relative to ConC B. Directions in non-significant contrasts were inferred from visual inspection of the boxplots. NS – not significant; PMN – polymorphonuclear cells; % – proportion. Full names of other abbreviated clinical variables are as described in Table 3.6.

	ConC1 vs ConC2	ConC1 vs ConC3	ConC2 vs ConC3
Significant in both discovery and validation cohort, same direction			
		higher age, lower lymphocyte count, lower respiratory rate	higher lactate
	lower monocyte count and proportion, lower bicarbonate, higher creatinine (high and low), higher urea (high), higher lactate, lower LMAP, lower temperature (low), lower platelets (low), lower LSBP, more days on inotropes, more days on renal support, lower arterial pH, lower urine volume; higher proportion of FP, more earlier timepoints, more on shock, more on mechanical ventilation/CPAP, more with acute renal failure, higher SOFA-Total/CVS/Haem/Renal scores		
Significant in both discovery and validation cohort, opposite direction			
	none		
Significant in discovery cohort, NS in validation cohort but in same direction			
	lower LWCC, lower respiratory rate		higher age, lower lymphocyte count
	lower haematocrit, higher bilirubin (high), higher heart rate (high), higher INR, higher PaO2, higher PT; higher SOFA-Hep, lower SOFA-Resp		
Significant in discovery cohort, NS in validation cohort and the direction not clear			
	lower lymphocyte count	lower LWCC	lower APACHE, lower monocyte count, higher NDRS, higher respiratory rate
	lower estimated days from onset for FP patients		
Significant in validation cohort, NS in discovery cohort but in same direction			
		higher PMN %	higher PMN %
Significant in validation cohort, NS in discovery cohort and the direction not clear			
		lower lymphocyte %	lower lymphocyte %, higher urea (high)
	more with comorbidity or malignancy of immune disease		

polymorphonucleocytes and lower proportion of lymphocytes compared with predicted ConC3, which was not observed in the discovery cohort.

In summary of the clinical phenotypes, the overall higher multi-organ severity (except for respiratory functions) and enrichment of FP patients in ConC1 were verified in the validation cohort. Variables show varying in levels of significance but the trend was the same for most of the variables. ConC2 were also suggested to be more severe than ConC3 patients.

The clinical difference between ConC1 and others was also reflected in a difference in outcome. The three predicted clusters were significantly different in survival at both 28 days (log-rank p=0.00046) and 6 months (log-rank p=0.00016) post-sampling (Fig. 4.14). As was seen in the discovery cohort, predicted ConC1 had an elevated risk of 28-day mortality compared with ConC3 (HR (95 % CI)=2.9 (1.7-4.9), p=0.000168), with ConC2 (2.0 (1.0-4.0), p=0.0405), or with ConC2 and ConC3 combined (2.6 (1.6-4.4), p=0.0002). The elevated risk of ConC2 vs ConC3 observed in the discovery cohort was not validated (1.4 (0.76-2.6), p=0.278).

Comparison of protein profiles

The distribution of the predicted clusters was visualised alongside the surgery and OBB control groups in the validation cohort, after transformation by PCA or UMAP (Fig. 4.15). As in the discovery cohort, there was a gradient on PC1 from the healthy, pre-operation, post-operation, to the GAinS clusters, with predicted ConC1 having highest PC1 scores, and predicted ConC2 or ConC3 separating on PC2. The three clusters also separated well after transformation by UMAP, indicating that the subgroups predicted by the protein panel also corresponded to the sub-structure from an unsupervised approach in UMAP.

I then contrasted samples from each predicted cluster with the healthy volunteer samples in the validation cohort as a shared control. For contrasts of ConC 1/2/3, 84/80/54 proteins were differentially abundant, respectively (FDR<0.05 and |FC|>1.5),

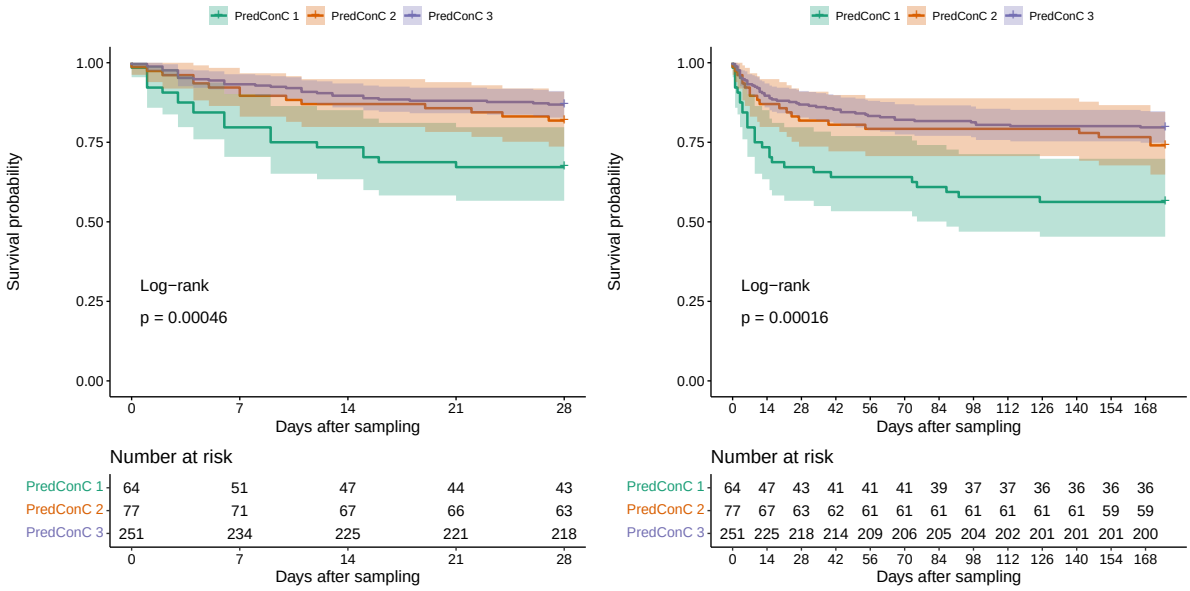


Figure 4.14: Kaplan-Meier curves of predicted clusters in validation cohort, comparing survival probability at 28 days or 176 days post-sampling. For patients with multiple samples, the cluster assignment from the latest available sample of the patient (among Day 1/3/5) was used. P values by log-rank tests are labelled. Shades on the curves show 95% confidence intervals.

which were similar numbers to the contrasts in the discovery cohort. Of these differentially abundant proteins, 35 were shared between the three contrasts. The 21 proteins unique to the ConC2-HV contrast were enriched for innate immune response and complement activation in the classicial pathway (FDR<0.05). The association with B cell signalling observed in the discovery cohort was not observed in the validation cohort.

The magnitude of differences versus HV varied between the three predicted clusters. Log fold changes of the 114 proteins differentially abundant (DA) in any of the contrasts were plotted as a heatmap (Fig. 4.16). Consistently with the discovery cohort, a group of proteins associated with sepsis severity had the highest levels in ConC1 including CRP, VWF, USP15, COL1A2, COL6A1, MMP2, and LCP1; Proteins MAP1A, GNMT and IGLV3-21 were lower in ConC2 relative to HV while being higher in ConC1 or ConC3. All 10 immunoglobulins differentially abundant between ConC1-HV were more abundant in ConC1, while for ConC2-HV, 5 were lower and 3 were higher in ConC2.

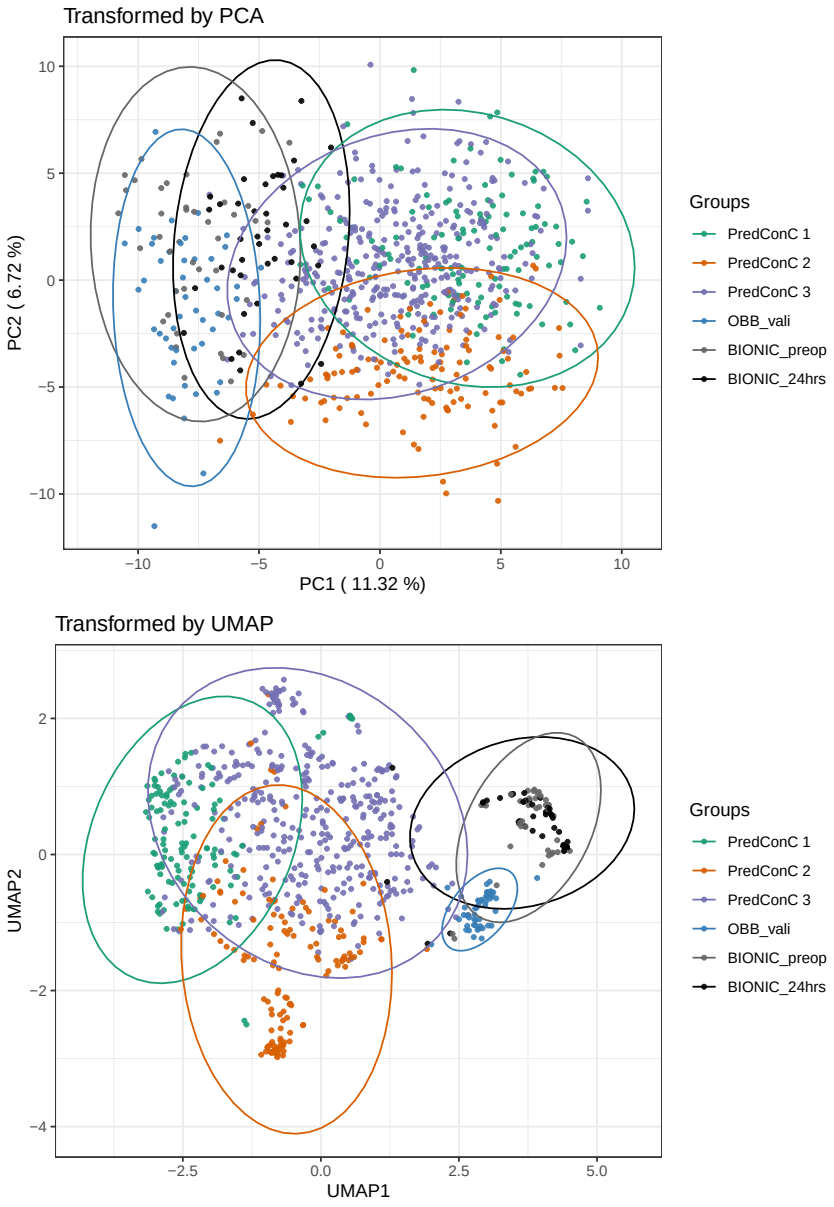


Figure 4.15: Predicted clusters visualised after transformation by PCA or UMAP, in validation cohort. Data ellipses based on multivariate normal distributions are shown for each comparator group.

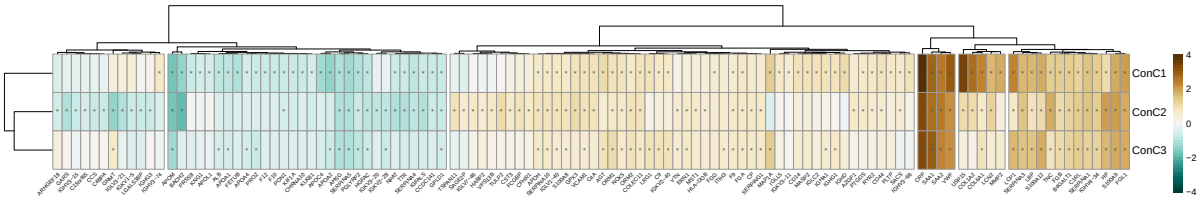


Figure 4.16: LogFC of contrasts of each cluster with HV, in validation cohort. Only first available samples of each patient were used in the contrasts. Age and sex were included as covariates. Asterisks denote FDR<0.05 & |FC|>1.5 for the contrast. Order of columns and rows in the heatmap were rearranged by clustering on the logFC values.

To see if the protein abundance characteristics that distinguished the discovery cohort clusters could be replicated in the predicted clusters, I compared samples from each pair of the predicted clusters. There were 19/22/39 proteins differentially abundant (FDR<0.05 and |FC|>1.5) respectively in the ConC1-ConC3, ConC2-ConC3, and ConC1-ConC2 contrasts (Fig. 4.17 (a-c)). The three contrasts all had a strong correlation between the log fold changes obtained in contrasting the discovery or validation cohort clusters (Pearson's r >0.92). These correlations were not restricted to the proteins with higher coefficients in the prediction model but were observed across all proteins.

Most of the top proteins and pathways that distinguished the discovery cohort clusters as discussed in section 4.2.3 were also identified in the validation cohort clusters: Proteins more abundant in ConC1 vs ConC3 were enriched in interleukin signalling (COL1A2, LCN2, LCP1, MMP2) and ECM organisation; USP15 and COL1A2 were two top proteins in ConC1 vs ConC3 while BACH2 and ARHGEF18 were lower; in validation cohort HP was not only lower in ConC1 vs ConC2 but also reached significance in ConC1 vs ConC3; MAP1A and GNMT were top two proteins lower in ConC2 vs ConC3, alongside with BACH2, ARHGEF18, and LCP1; proteins more abundant in ConC1 vs ConC2 were enriched for multiple pathways in immunity including phagocytosis, complement activation and B cell signalling; proteins lower in ConC1 vs ConC2 included among others albumin, apolipoproteins, and proteins mediating cell adhesion (TSPAN11, SELL), enriched for lipoprotein metabolic process. The distribution of each of these proteins across the predicted clusters were plotted in Figure C.9.

4.3.3 Protein signature panels

For a prediction model to be validated in other cohorts or ultimately to be applied as a bed-side test, the number of proteins to be measured in the samples needs to be restricted to the minimum. In this section I tested whether the clusters could be

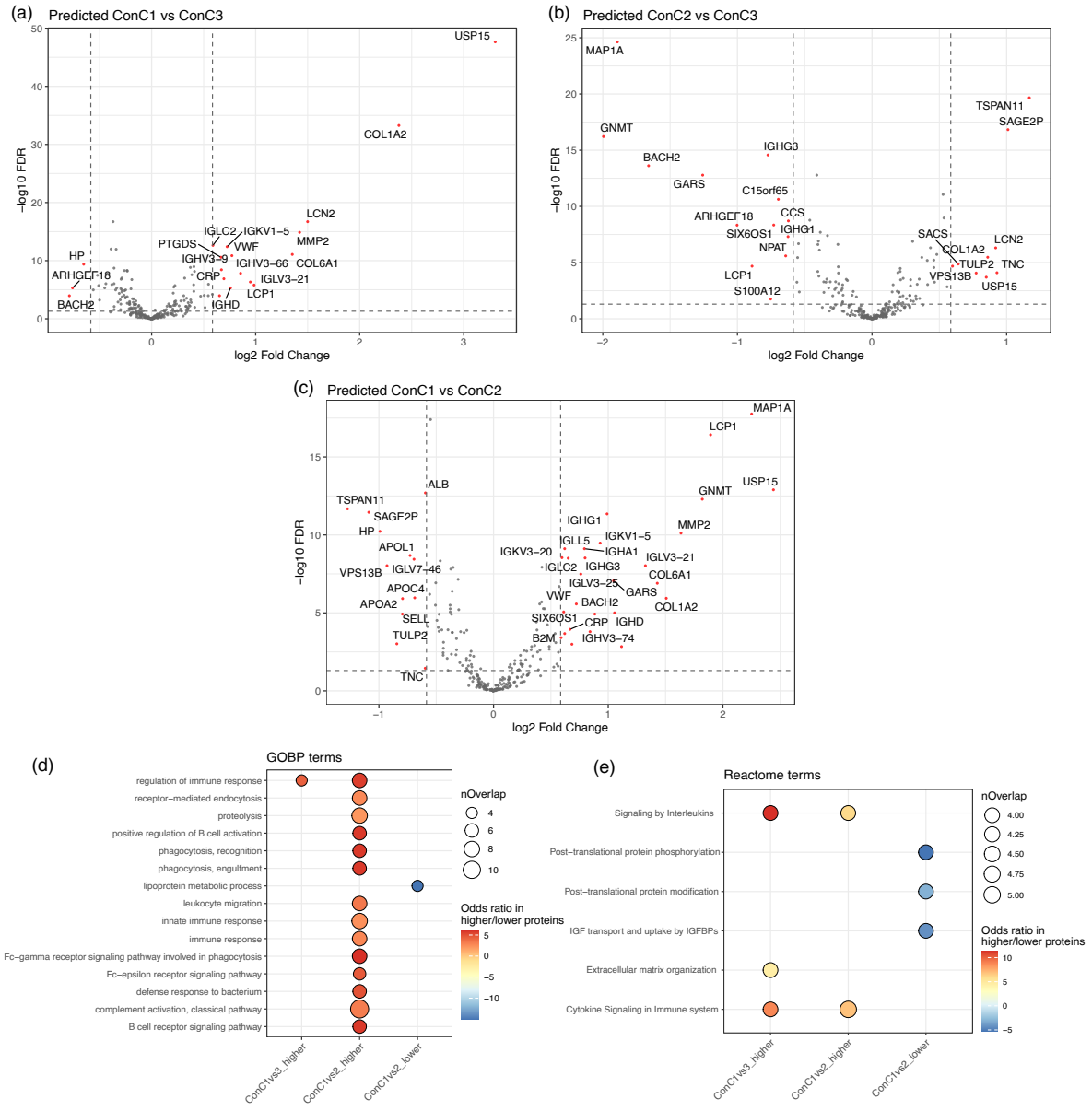


Figure 4.17: Comparison of protein profiles between the three predicted clusters, in validation cohort. (a-c) Volcano plots of the contrasts between each pair of clusters. Red points labelled by gene names denote differentially abundant proteins (FDR<0.05 and |FC|>1.5). For patients with more than one samples, only the first available samples were included in the contrasts. Age and sex were included as covariates. **(d-e)** GOBP or Reactome terms enriched in the differentially abundant proteins with either higher or lower abundance in each of the clusters compared with the others. A term needs to have a minimum of 4 overlapping proteins with the data input to be tested. Odds ratios are shown on a scale from white to red for higher-abundant proteins in the corresponding clusters; and on a scale from white to blue for less abundant proteins. Only protein sets with any significant terms detected in the corresponding annotation are shown.

Table 4.7: 3-cluster minimal models performance. Models were fitted using elastic net penalty. "Coefficients" refer to protein coefficients in the ElasticNet_138 model. "Significance" refers to FDR calculated in contrasting discovery cohort clusters.

proteins per cluster	picked by	α	λ	N (predictors)	Training set accuracy	Test set accuracy	ConC1 sensitivity in test set	Overlap with ElasticNet_269 predictions in validation cohort
3	coefficients	1	0.00402	7	769/992 = 77.5%	181/244 = 74.2%	42/55 = 76.4%	481/624 = 77.1%
5	coefficients	0.242	0.0150	11	840/992 = 84.7%	191/244 = 78.3%	42/55 = 76.4%	536/624 = 85.9%
3	significance	0.716	0.00623	8	815/992 = 82.2%	194/244 = 79.5%	40/55 = 72.7%	515/624 = 82.5%

effectively predicted by protein signature panels with only a small number of proteins.

Three-cluster prediction with a minimal protein panel

Proteins to be included in the minimum panel were selected based on their contributions in the elastic net model with 138 proteins input (Fig. C.6), which is restricted to differentially abundant (FDR<0.05 and |FC|>1.2) proteins to allow for better interpretation of the potential protein markers. In each of the three clusters, proteins were ranked by their absolute value of coefficient, then the three proteins with the top coefficients were selected for each cluster, ensuring that at least one higher-abundant protein and one lower-abundant protein were included for each cluster. Immunoglobulin variable region proteins were excluded since the high sequence similarity makes them less distinguishable from each other. Uncharacterised proteins (e.g. C15orf65) were also excluded. This yielded a panel with 7 unique proteins including USP15, C8G, C2, ALB, C4BPA, GC, and C3. I then used the package caret to train an elastic net model with the best accuracy. The 7-protein 3-cluster model had a training set accuracy of 77.5% and a test set accuracy of 74.2% (Table 4.7).

Increasing to selecting the top five proteins per cluster yielded a panel with 11 unique proteins. This included the 7 proteins listed above plus PLTP, SERPINC1, MAP1A, and SERPIND1. The 11-protein model had a training set accuracy of 84.7% and a test set accuracy of 78.3% (Table 4.7). Ten more test set samples were correctly assigned with the extra predictor proteins, and 55 more validation cohort samples had consistent assignment with the best-performing elastic net model described in section 4.3.1.

Aside from selecting protein candidates with the highest coefficients in the larger

Table 4.8: Contingency table of 8-protein prediction model performance in the test set. Rows show the number of samples predicted by the 8-protein model. Columns show the number of samples from consensus clustering i.e. the true classification.

	ConC1	**ConC2**	**ConC3**
Predicted ConC1	40	2	9
Predicted ConC2	2	32	6
Predicted ConC3	13	18	122

models, I also tried selecting proteins based on their significance in the cluster contrasts. I selected three proteins with the highest -$\log_{10}$(FDR) values for each of the three contrasts between the three discovery cohorts clusters, ensuring at least one proteins with higher abundance and one protein with lower abundance were included for each contrast. After removing redundant proteins, this gave an 8-protein model including USP15, COL1A2, APOA2, MAP1A, GNMT, TSPAN11, LCP1, and ALB. This best tuned elastic net model on differentially abundant proteins had a similar performance to the two models that selected proteins based on model coefficients (Table 4.7). The contingency table compared predicted and true ConC classification in the test set samples (Table 4.8) and showed that there was not an unbalanced misclassification, except that sensitivity for ConC2 was relatively low (61.5%).

Comparing across the three minimal models showed that an increasing number of protein predictors could improve model performance, although the 8-protein panel is potentially already a good minimal panel with four-fifth of the samples correctly assigned. Future work should be undertaken to locate an optimal balance of protein number in order to both restrict the cost of measuring protein markers and to obtain a model performance that is clinically informative.

Two-cluster predictions

As most differences in clinical measurements or outcome were observed between ConC1 and either ConC2 or ConC3, but not between ConC2 and ConC3, the greatest clinical value of the proteomic clusters may lie in distinguishing ConC1 patient-

Table 4.9: 2-cluster minimal models performance. Classification of ConC1 vs ConC2/3 were predicted using two approaches.

Algorithm	lambda	N(predictor)/ N(input)	Train set accuracy	Test set accuracy	Test set ConC1 sensitivity	Test set ConC1 precision
lasso	0.0185	24/38	937/992 = 94.4%	222/244 = 91.0%	37/55 = 67.3%	37/41 = 90.2%
elastic net	0.00259	10/10	925/992 = 93.2%	220/244 = 90.2%	41/55 = 74.5%	41/51 = 80.4%

timepoints from the rest. Therefore I tested whether two clusters (ConC1 vs others) can be separated better than three clusters using a smaller panel of proteins. The protein profiles also supported the grouping of ConC2 and ConC3 against ConC1, since ConC1 samples separated from the rest in unsupervised clustering (Fig. 4.2) as well as on PC1 (Fig. 4.8).

I tested two approaches for deriving a small panel: using the 38 proteins differentially abundant (FDR<0.05 and |FC|>1.5) in either the ConC1-ConC2 or ConC1-ConC3 contrast as input to a lasso model, so inputting a larger number of protein candidates to a model that tends to exclude the correlated variables; or, manually picking a more restricted number of proteins (10 in this case) based on the model coefficients, and running it through a more flexible elastic net model, which is similar to the approach of deriving the 7-protein 3-cluster prediction model described in the previous section.

In the first approach, 24 out of the 38 proteins were retained in the lasso model, producing a test set ConC1 sensitivity of 67.3% and ConC1 precision of 90.2% (Table 4.9). Proteins with the highest coefficients in this model included ALB, USP15, VTN, APOA1, COL1A2 and immunoglobulin variable region proteins.

For the second approach, the 10 proteins with the highest absolute values of ConC1 model coefficients in the ElasticNet_138 model (Fig. C.6) were selected as input to elastic net, excluding immunoglobulin variable region proteins, and including at least 3 proteins higher and 3 proteins lower in ConC1. These included USP15, PLTP, COL1A2, CD14, C8G, ALB, C2, C3, SERPIND1, and GC. This 10-protein elastic net model had a test set ConC1 sensitivity of 74.5% and precision of 80.4% (Table 4.9), which showed an increase in sensitivity and decrease in precision compared with the

first approach with more protein predictors.

Accuracy metrics were not comparable between 3-cluster or 2-cluster models, but ConC1 sensitivity was similar in the 10-protein 2-cluster model and the 8-protein 3-cluster model (72.7%). Predicted ConC1 vs others in the validation cohort also showed a similarly high hazard ratio of 28-day mortality in the former model (HR (95% CI)=2.4 (1.4-4.1), log-rank p=0.000916) and in the latter model (HR (95% CI)=2.4 (1.4-4.1), log-rank p=0.000907). Therefore, I did not observe a clear benefit by switching from 3-cluster prediction to 2-cluster prediction for the aim of either restricting the predictor number or improving model performance.

Comparison with prediction model on mortality

To help illustrate the utility of patient stratification using proteomic clusters, for comparison I also tested whether mortality can be directly modelled based on either proteins or clinical variables.

For input to the mortality model on proteins, I used 26 proteins that were differentially abundant between samples from 28-day survivors or non-survivors in the discovery cohort (FDR<0.05 and |FC|>1.2, using the last available sample of each patient). 23 out of the 26 proteins were included in the trained elastic net model. Only a small number of patients were predicted to have 28-day mortality, giving high precision but very low sensitivity (6.9%, 4.0%, or 4.8% in the training set, test set, or validation cohort).

Another mortality model was built based on the 60 numerical and categorical clinical variables. Eight variables not available at the time of sampling were not included, such as duration of respiratory support or microbiology identified by metagenomics testing. The random forest algorithm was applied so that no assumption is made about the predictor distributions. Missing clinical data was first imputed by the rfImpute function from the randomForest package, which was performed separately in the training set, test set, and validation cohort. After tuning the parameters, the

model used 35 variables for each iteration, a maximum of 20 terminal nodes in each tree, and 300 trees were trained. Variables with the highest importance included age, SOFA-neuro score, APACHE score, ARDS classification, and SOFA-total score. This model showed better prediction in the training set and test set, with sensitivity of 35.8% (59/165) and 13.3% (6/45), and precision of 100% (59/59) and 85.7% (6/7). However, in the validation cohort the model had the same issue as the mortality model on proteins: only 8 samples were predicted with 28-day mortality, yielding a low sensitivity of 2.9% (3/105). These 8 samples were from only 3 out of 392 patients (18/392=4.6%) so mortality was not compared between the two predicted groups.

Therefore, in this dataset mortality cannot be directly modelled and predicted from protein markers or clinical variables with a consistently reasonably good sensitivity, without further model optimisation. On the other hand, the proteomic clusters can consistently stratify patients into groups of different risks of mortality. Applying models constituted of either 181, 10, or 8 proteins in the validation cohort, the predicted ConC1 membership identified a subgroup of 16% patients that had significantly lower 28-day and 6-month survival (Fig. 4.14, Fig. C.10).

4.4 Results: Characterisation of the deeper proteome profile in the patient subgroups

Two subsets of the samples (corresponding to specific patient-timepoints) included in this dataset (MS2019) also had plasma proteins measured on two other technology platforms: QE-HF mass spectrometer, and Luminex xMAP assay, which both detected lower abundance proteins better in these subsets of patients. Using the proteomic cluster assignments from MS2019 and the protein measurements from these two platforms, I characterised the clusters further in the protein range that was not fully covered in the undepleted timsTOF platform in MS2019.

4.4.1 MS192: a depleted mass spectrometry dataset with higher detection depth

Processing of the MS192 dataset

During sample preparation of MS192, twelve high-abundance proteins were targeted in affinity-depletion to enable better detection of proteins with lower abundance within a certain machine time (see Table C.7 for the list of the top 12 proteins depleted). In the raw intensity output, there were 1356 proteins identified by at least one unique peptide. The raw intensity was pre-processed using similar principles as in MS2019. The 0.5% lowest measurements were considered to be machine noise and cut out from the lower tail of the distribution. Proteins with more than 30% missing values in all samples were then filtered out, leaving 1123 proteins in the cleaned-up dataset, which is a good detection depth in a non-fractionated label-free MS experiment like this one, with less than one hour gradient time. All samples passed quality control. The dataset was then normalised by VSN and imputed by KNN. There were 180 proteins that overlapped between the 1123 measured in MS192 and the 269 measured in MS2019.

Comparison of the ConC clusters in MS192

There were 148 sepsis samples that were included in both MS192 and MS2019. For comparing between the clusters, ConC assignments from both the discovery and validation cohort clusters in MS2019 were used. Among the overlapping samples, there were 22/36/90 samples in the three ConC clusters. By only including the first available samples following ICU admission for each patient, there were 20/21/59 patients in ConC1/2/3, respectively. I then compared protein measurements from MS192 in the overlapping samples between each pair of the three clusters. No differential abundance was detected between ConC1-ConC2 or ConC2-ConC3. None of the four top differentially abundant proteins in ConC2 in timsTOF (MAP1A, GNMT,

TSPAN11, SAGE2P) were also detected in MS192 after the protein filtering. In comparing ConC1-ConC3 in MS192, there were 144 proteins more abundant and 63 proteins less abundant in ConC1 (FDR<0.05 and |FC|>1.5) (Fig. 4.18 (a-c)).

Compared with the pathways identified in the timsTOF data, the enrichments of immune response related pathways, ECM organisation, and lipoprotein metabolic process were also observed in the QE-HF (MS192) data; enrichment in interleukin signalling was more specific, namely to IL-4 and IL-13 signalling; proteolysis was enriched but in the opposite direction. In addition, proteins higher in ConC1 were enriched for multiple pathways involving collagen organisation, or processes related to the cell cycle (Fig. 4.18 (d-g)). Therefore, protein measurements from another platform verified that Cluster 1 and 3 represent patient subgroups with distinct pathways indicated from the protein profiles, while the pathways differentiating Cluster 2 from the rest was not clear in this QE-HF dataset.

Because of the differences between the two platforms, including the top protein depletion performed in MS192 but not in MS2019, there was only limited correlation in the overlapping protein levels measured (Pearson's r >0.3 for 45/180 proteins). Therefore, the value of the MS192 dataset was more in adding to the understanding of the protein profiles of the clusters, rather than replicating the differences in individual proteins identified from the timsTOF data. The differences between the two datasets are elaborated upon in the Discussion section of this chapter.

4.4.2 Cytokines differentially abundant between the clusters

Complementary to the protein range covered by untargeted MS experiments, Luminex assays use an antibody-based approach to target the lower-abundance cytokines. This Luminex dataset as described in Methods included 204 samples overlapping with MS2019, enabling the characterisation of cytokine abundance across the ConC clusters. The Luminex data included all patient-timepoints in MS192, plus 53 more samples from FP patients. Among the 204 overlapping samples, 52/44/108 samples

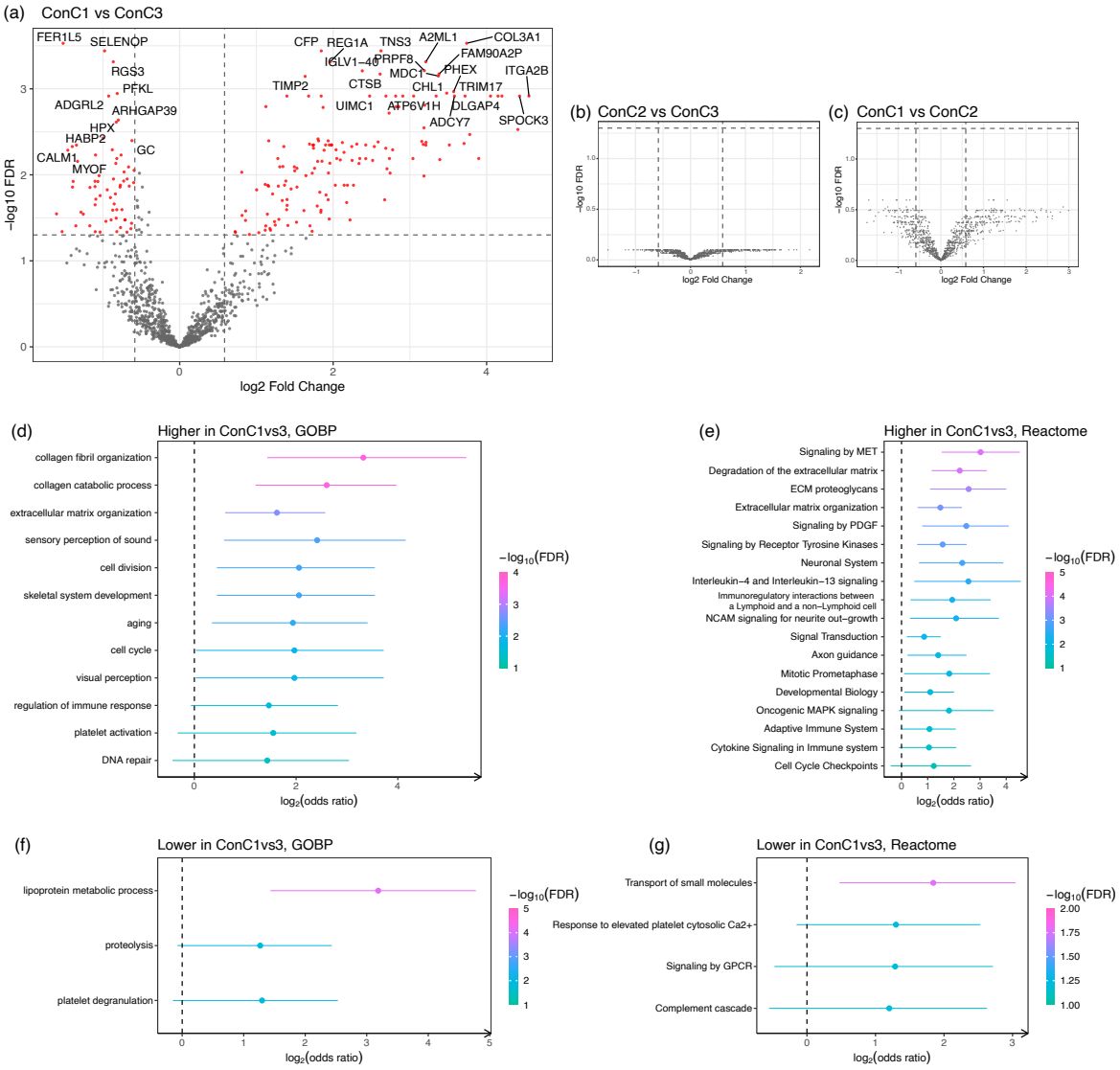


Figure 4.18: Comparison of protein profiles between ConC clusters in MS192. (a-c) Volcano plots of the contrasts between each pair of the clusters in the overlapping samples. Red points denote differentially abundant proteins (FDR<0.05 and |FC|>1.5, grey dashed lines indicate these thresholds). For patients with more than one samples, only the first available samples were included in the contrasts. Age and sex were included as covariates. **(d-g)** GOBP or Reactome terms enriched in the differentially abundant proteins with either higher or lower abundance in ConC1 compared against ConC3. Redundant terms have been removed by the xEnrichConciser function from the XGR package. The terms plotted were significantly enriched with FDR<0.05 except for the lower three terms in (g) where FDR=0.064. Horizontal lines in the forest plots indicate 95% confidence intervals of $\log_2$(odds ratio).

were assigned to ConC1/2/3 based on MS2019 data. By only including cluster assignments from the first available samples per patient, there were 48/27/71 patients in each cluster. In the Luminex data, abundance of 65 cytokines was compared between each pair of the clusters, using Mann-Whitney rank-sum tests on the first available samples.

There was no difference in cytokines observed between ConC2 and either ConC1 or ConC3. In comparison of ConC1 vs ConC3, most cytokines had a higher value of sampled median in ConC1, with 8 analytes reaching significance (Fig. 4.19, Fig. 4.20). These included 3 chemokines (MCP-1, IL-8, MIP-3α), two cytokines involved in more active B cell proliferation (APRIL, IL-6), two cytokines with inhibitory functions on the immune response (IL-2R, LIF), and MMP-1.

MMP-1 is an interstitial collagenase involved in ECM breakdown, which is not a cytokine but an enzyme included in the Luminex panel. MCP-1 (CCL2) is chemotactic for T cells (T_H2>T_H1), monocytes, and basophils. Target cells of IL-8 (CXCL8) include neutrophils, basophils, CD8 cell subsets, and endothelial cells. Target cells of MIP-3α (CCL20) include T cells (memory T cells, T_H17 cells), monocytes, immature dendritic cells, activated B cells, and NKT cells.

APRIL (A Proliferation Inducing Ligand) is produced by activated T cells, and promotes B cell proliferation. IL-6 is promptly and transiently produced in response to infections and tissue injuries. IL-6 has a wide range of roles in T- and B-cell growth and differentiation, and in acute phase protein production. LIF (leukemia inhibitory factor) is a highly pleiotropic cytokine in the IL-6 superfamily. Produced by bone marrow stroma and fibroblasts, LIF is known to control cell differentiation and growth in a time and tissue-dependent manner, to create an immunosuppressive microenvironment, and to suppress the development of IL-6-induced T_H17 lineages (Wrona et al. 2021). IL-2R is measured as the soluble form of IL-2Rα. It is reported that the binding of sIL-2Rα to IL-2 promotes IL-2 signalling instead of blocking the function. The complex promotes T-cell differentiation towards inhibitory T_{reg}

cells rather than T_H1 or T_H17 cells, leading to downstream T_{reg} proliferation and potentially development of immune tolerance (Yang et al. 2011). The cytokine functions were summarised using Janeway's Immunobiology (9th edition, Murphy and Weaver 2016) as well as other cited publications as references.

Overall, the Luminex data suggested that ConC1, compared with the majority of patients represented by ConC3, showed features of more active chemotaxis processes and more actively regulated IL-6-involved pathways including proliferation and differentiation of lymphocytes. It could also be concluded that ConC1 and ConC3 were characterised by a global higher- or lower- abundance of cytokines representing a high-inflammation or low-inflammation phenotype in the two patient subgroups.

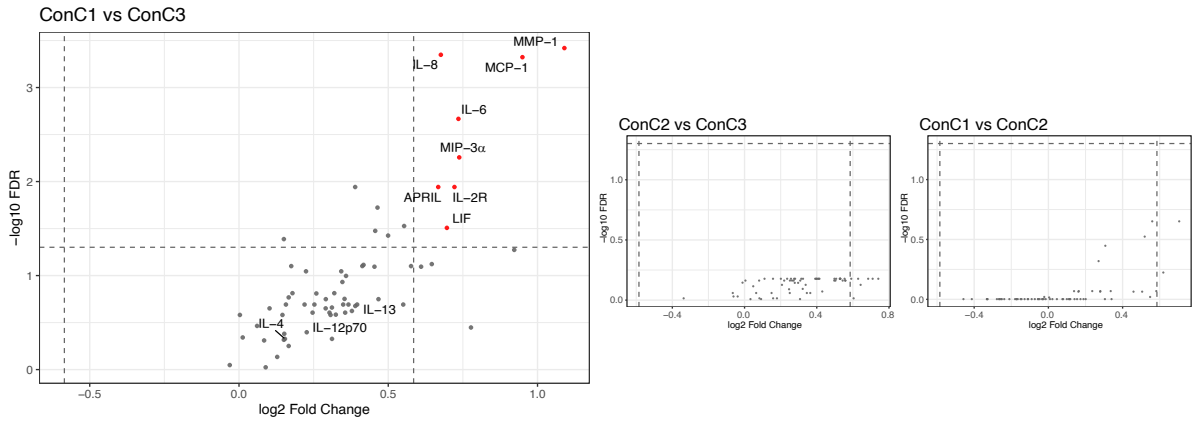


Figure 4.19: Comparison of cytokine concentrations between the ConC clusters, in Luminex data. Volcano plots of the contrasts between each cluster and the three others combined, in overlapping samples. Red points labelled denote differentially abundant proteins (FDR<0.05 and |FC|>1.5, grey dashed lines indicate these thresholds). FDRs were the B-H adjusted p values from Mann-Whitney rank-sum tests using the first available samples per patient. Log_2 fold changes were calculated as the ratios of median levels in the two groups compared, followed by a logarithmic transformation with base 2.

4.4.3 Summary of molecular characteristics of each cluster

Figure 4.21 summarises the molecular characteristics that distinguish the ConC clusters, based on the between-cluster contrasts in the three datasets described above (MS2019 discovery and validation cohort, MS192, and Luminex). Each term was derived from functions of the individual differentially abundant (DA) proteins or

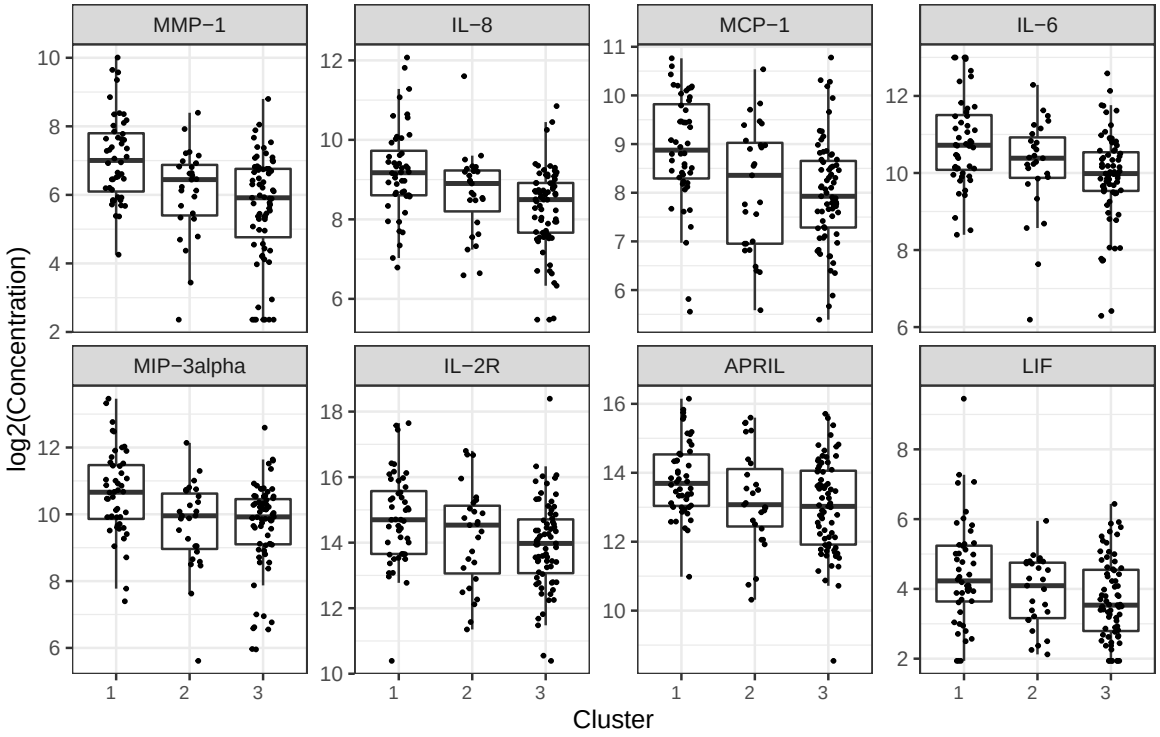


Figure 4.20: Boxplots showing the distribution of 8 differentially abundant Luminex analytes across the clusters, in overlapping samples between MS2019 and Luminex data, using only first available samples per patient. Concentrations are in the unit of pg/mL.

pathways enriched in DA proteins. To verify whether the signal is from only one of the three clusters, or is high in one cluster while low in the other one, the distribution of proteins overlapping with the pathway terms were visually examined, in each of the dataset separately. For example, distributions of proteins annotated for "ECM organisation" or "Interleukin signalling" in MS2019 showed these processes were more active in ConC1, without a clear difference between ConC2 and ConC3; distributions of more proteins in MS192 showed lower abundance in ConC3 (Fig.C.11, Fig.C.12). Only terms replicated across at least two of the datasets or discovery/validation cohorts, or with a clear direction indicated from the deeper proteome, were included in the summary graph. For each of the terms listed, the dataset, contrast and approach where the evidence was derived from were detailed in Table C.8.

In summary, ConC1 was characterised by higher abundance of proteins implicated in many aspects of the immune pathways, including cytokines and immunoglobulins.

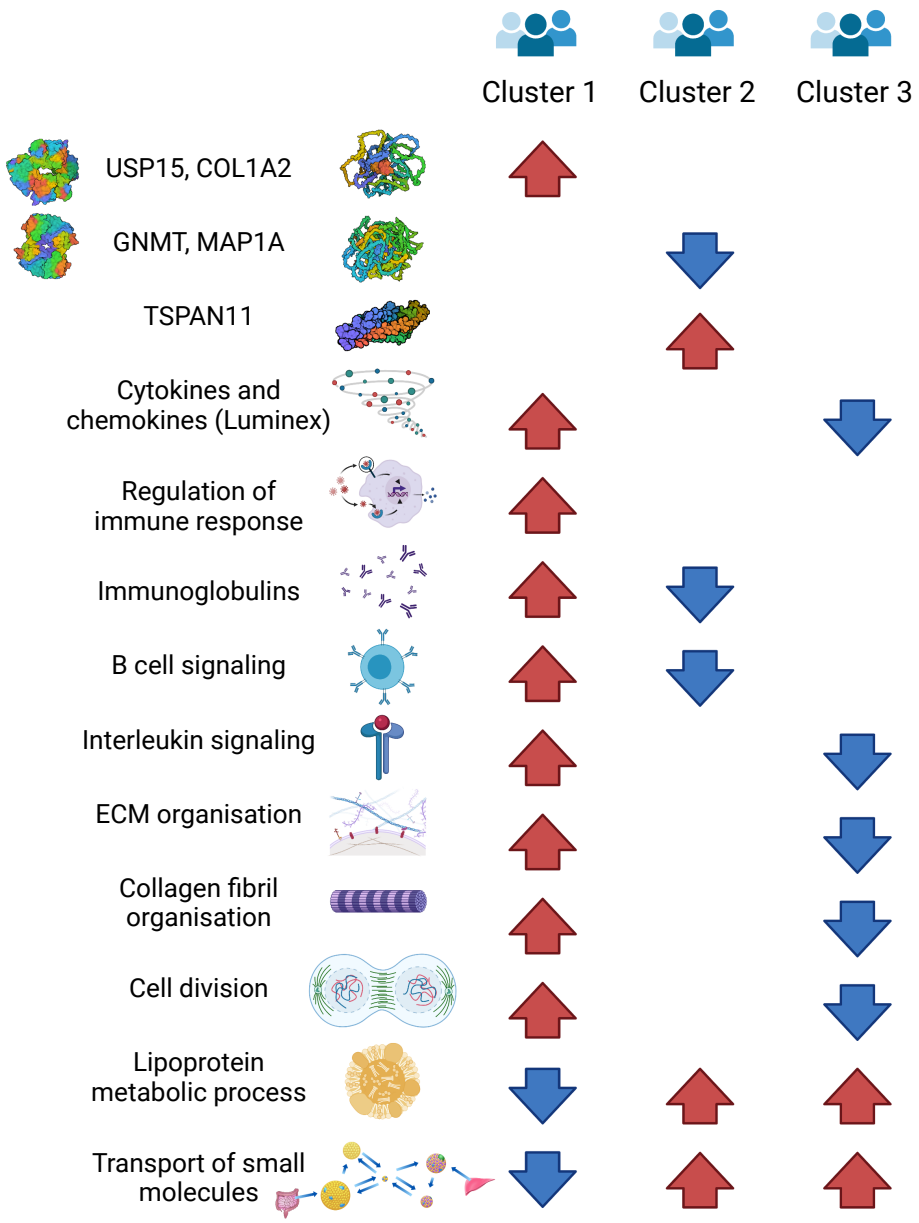


Figure 4.21: Summary graph of molecular characteristics for each ConC cluster, identified from between-cluster contrasts from the MS2019, MS192 and Luminex datasets. Evidence for each term is detailed in Table C.8. Red and blue arrows indicate that the corresponding proteins are with higher or lower abundance in the clusters, respectively.

ConC1 also exhibited higher collagens and other ECM components in the blood circulation, which may indicate a higher level of tissue damage. ConC2 showed lower abundance of various immunoblogulins or proteins indicated in some of the immune regulation pathways including B cell signalling. In ConC3 there was lower abundance of cytokines or proteins involved in interleukin signalling, or the ECM. Components of the lipoproteins metabolic or transportation process were shown to negatively correlate with sepsis severity, and here were lower in ConC1 while higher in the other two.

4.5 Results: Interaction of the proteomic and transcriptomic patient subgroups

Of the 1860 GAinS patient-timepoints (1182 patients) included in MS2019, 1354 (1010 patients) also have whole-genome gene expression data available measured by either microarray or RNA-seq as described in the Introduction section; 1361 (1016 patients) had the SRS transcriptomic endotypes assigned based on these two data types as well as from qPCR (Cano-Gamez et al. 2022). The aim of this section is to understand the interaction between the patient subgroups defined from either plasma proteome (ConC) or leukocyte transcriptome (SRS) in terms of how the group assignments overlap, what the molecular profiles underpinning the two classifications are, and whether the two can complement each other.

4.5.1 Overlap of SRS and ConC classifications

SRS1 was enriched in ConC1

For patient-timepoints with both the SRS and ConC group memberships available, the two classifications were compared at the patient-level using the first available samples per patient (4.10). In the discovery cohort, SRS1 patients were significantly enriched in ConC1 (χ^2 test p<0.0001), with SRS1 constituting 70% of ConC1 patients, and <37%

Table 4.10: Patient numbers in overlap between SRS and ConC classifications, in GAinS discovery or validation cohorts. Only classifications from the first available sample of each patient were used.

	Discovery cohort		Validation cohort	
	SRS1	non-SRS1	SRS1	non-SRS1
ConC1	105	44	50	20
ConC2	60	104	31	33
ConC3	126	242	63	138

of the other two clusters (Fig. 4.22). In the validation cohort there was also enrichment of SRS1 in ConC1 (χ^2 test p<0.0001), with SRS1 constituting 71% of ConC1 patients, 48% of ConC2 patients, and 31% of ConC3 patients (Fig. 4.22). Therefore, the patient subgroups identified from the two molecular levels were significantly related, but one cannot replace the other.

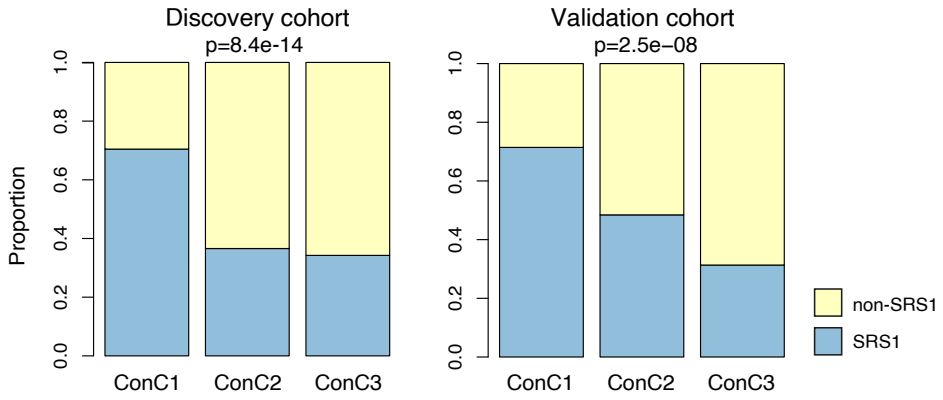


Figure 4.22: Patient proportions in overlap between SRS and ConC classifications. Bar plots show proportions of SRS1 (blue) or non-SRS1 (yellow) patients in proteomic clusters ConC1/2/3 (x axis), in discovery or validation cohort. P values are given by χ^2 tests on overlapping patient numbers. Only classifications from the first available sample of each patient were used.

ConC cluster movement

The three proteomic ConC clusters were significantly associated with the three timepoints of ICU sampling following admission (χ^2 p=0.0012), with ConC1 enriched in day1 and ConC3 enriched in day5 (Fig.C.13). Of the 1182 GAinS patients assayed in MS2019, 526 had samples in at least two of the time points. 302 of the 526 patients

Table 4.11: Number of patients who moved between specified proteomic clusters at specified sampling timepoints, in GAinS discovery and validation cohorts. Patients with samples available at both specified timepoints are counted in each table.

	From day1 to day3			From day3 to day5			From day1 to day5		
	To ConC1	To ConC2	To ConC3	To ConC1	To ConC2	To ConC3	To ConC1	To ConC2	To ConC3
From ConC1	43	19	24	50	15	25	21	17	29
From ConC2	7	28	26	2	36	35	2	20	18
From ConC3	10	26	108	5	36	128	6	16	78

(57.4%) showed at least one transition in cluster membership across days 1-3-5. There was a higher proportion of patients who moved from ConC1 to ConC2 or ConC3, than in the opposite directions (Table 4.11, Fig.4.23). There was also a higher proportion of patients who moved from ConC2 to ConC3 than in the opposite direction. This suggested that many of the patients who were available for further serial sampling were recovering and moving to proteomic clusters representing less severe states.

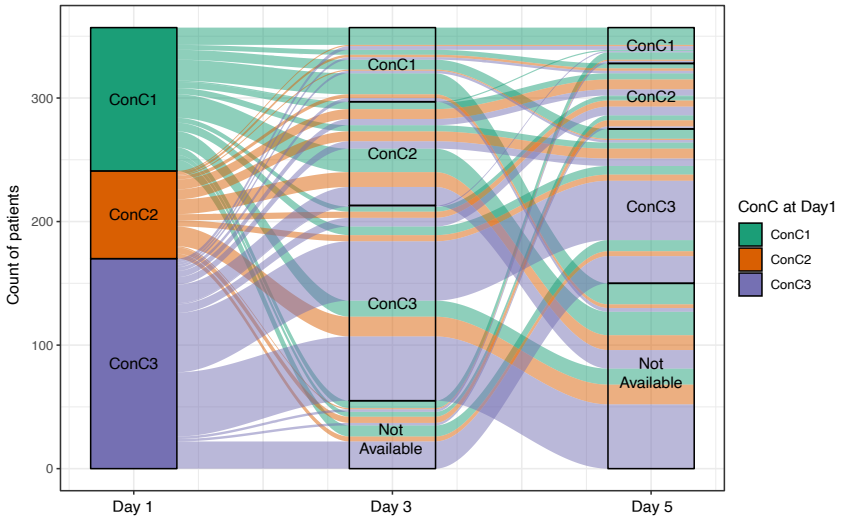


Figure 4.23: Proteomic cluster movement during Day1 to Day 5 of ICU admission, in 346 discovery or validation cohort GAinS patients with a Day1 sample and at least one subsequent sample available. The width of the flows are proportional to the number of patients with the corresponding ConC transition/persistence.

Overlap between SRS and ConC cluster movement

The overlaps and differences in classifications by SRS or ConC suggested that there should also be both overlap and distinctions in the mechanisms underlying risk stratification by SRS or by the proteomic clusters. If so, the recovery or deterioration

of patients reflected by movements in SRS or ConC memberships should also show an interaction.

258 GAinS patients in MS2019 had both ConC assignments available for ⩾2 timepoints and SRS available for ⩾2 (not necessarily the same) timepoints among day1/3/5. The numbers of overlap in the membership movements within day1-5 are shown in the contingency table (Fig.4.24(a)) and the percentages are shown in the stacked bar plot (Fig.4.24(b)). Since SRS enrichment was only consistently observed in ConC1 vs other clusters and as ConC2 and ConC3 were more similar in protein profiles, ConC2/3 were combined into one group for analysing the movement. There was a significant overlap between the four movement types in either SRS or ConC (χ^2 test $p<0.0001$). For each of the SRS movement group, comparing across the percentages it constituted within each ConC movement group, in Fig.4.24(b), patients who stayed in nonSRS1 (nonSRS1_stay) made up the largest proportion of patients who stayed in non-ConC1 proteomic clusters (ConC2/3_stay). Patients who stayed in SRS1 (SRS1_stay) had the largest percentage within patients who stayed in ConC1 (ConC1_stay), or moved to ConC1 from less severe clusters (ConC2/3_to_ConC1). Patients who were getting better in terms of transcriptomic endotypes (SRS1_to_nonSRS1) showed the largest percentage within patients who were getting better in terms of plasma proteomic subgroups (ConC1_to_ConC2/3). There was a small number of unfortunate cases where patients who deteriorated in terms of transcriptomic endotypes (nonSRS1_to_SRS1) also showed the greatest tendency to deteriorate in terms of plasma proteomic subgroups (ConC2/3_to_ConC1).

This supported the argument that there is correspondence between the transitions from and persistence of the clinically less/more severe proteomic or transcriptomic clusters, suggesting that there may be shared components of the mechanisms underlying the two classifications leading to a consistent effect over time.

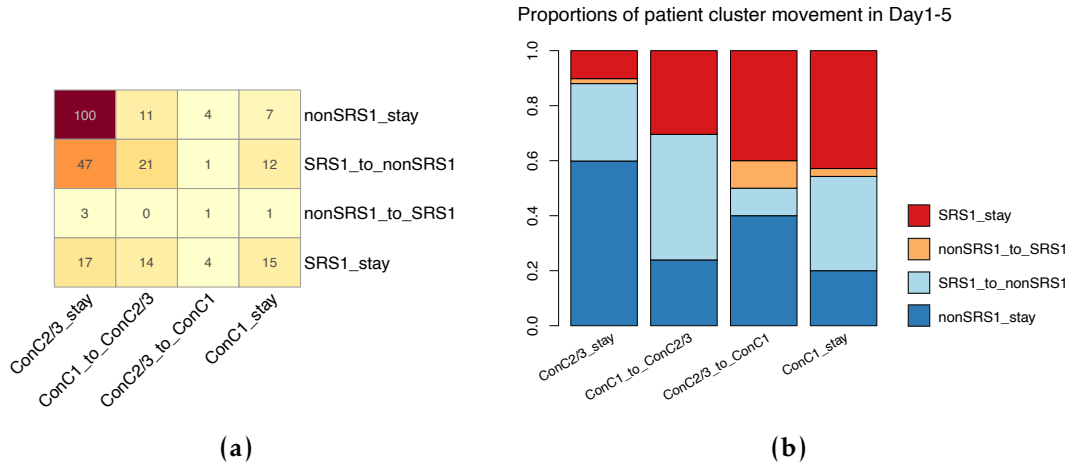


Figure 4.24: Overlap between SRS and ConC cluster movement. Class movement observed among two or three of the day1/3/5 timepoints were used for comparison. Patients with only one timepoint in either SRS or ConC assignments were not included. Patients with 3 timepoints who moved in different directions (for example, non-SRS1 at day1, SRS1 at day3, then non-SRS1 at day5) were also excluded. **(a)** Contingency table. **(b)** Stacked barplots. In each ConC movement group (x axis labels), proportions of each SRS movement group were represented by the height of the corresponding stacked bars.

4.5.2 Molecular profiles that underpin SRS or ConC clusters

To understand the shared and distinctive aspects of the molecular profiles that underpin the differentiation in SRS or ConC clusters, I tested for differentially abundant proteins and differentially expressed genes between the classifications. It needs to be noted that the two factors had significant correlation and thus collinearity, which will lead to uncertainty in estimating the coefficients if both are included in one linear model. As a rule of thumb, a variance inflation factor (VIF) that exceeds 5 or 10 indicates a problematic amount of collinearity (James et al. 2013, page 101). SRS or ConC membership alone do not have collinearity with the other covariates (age or sex) included in the model (VIF<1.1 for all variables). However, including both SRS and ConC in one model introduces a larger extent of collinearity, regardless of whether age and sex were included as covariates or not (without age/sex: VIF for SRS= 5.5, VIF for ConC =1.7, VIF for the interaction term SRS:ConC =7.0). Therefore, in separate linear models I compared SRS1 vs non-SRS1, or ConC1 vs non-ConC1, at both the protein and gene expression levels, using the MS2019 discovery and validation cohort samples

combined, restricted to patient-timepoints with both assignments available. Pathway enrichment analysis was performed for each contrast and the signals were compared.

At the protein level

In comparing SRS1 (n=435) with non-SRS1 (n=581), the top proteins with higher abundance in SRS1 included CRP, LCN2, USP15, COL1A2, SAA2, MMP2, S100A8, TNC and S100A12 (Fig.4.25), which showed a good overlap with proteins higher in ConC1. Testing the top 10% proteins (i.e. $10\% \times 269 = 27$ proteins, ranked by significance) higher in SRS1 or ConC1 with FDR<0.05 showed enrichment in cytokine signalling for both contrasts (Table 4.12). The proteins lower in SRS1 or ConC1 were not tested for enrichment, since the signal was less strong in both significance and fold change compared with the more abundant proteins.

There were also some proteins that did not show a change in the same direction in SRS1 and ConC1. A group of 8 immunoglobulins were more abundant in ConC1 but there was no immunoglobulin signal in SRS. HP and APOA2 were lower in ConC1 with no difference in SRS1 (green dots on Fig. 4.25 correlation plot). Among the proteins with higher abundance in SRS1, SAA2 (serum amyloid A-2, an acute-phase protein) and TNC (Tenascin, an ECM protein) were not different in ConC1, while S100A8 tended to be higher in ConC1 (yellow dots on Fig. 4.25 correlation plot).

At the gene expression level

At the gene expression level, contrasts and pathway enrichment tests were performed in microarray and RNA-seq data separately. Samples overlapping between the two data types were removed from the microarray dataset. Leukocyte gene expression data were compared in 837 samples (649 patients) with both RNA-seq and proteomics data, and in 517 samples (366 patients) with both microarray and proteomics data. The overlapping terms identified from both subsets were listed in Table 4.12. The top differentially expressed (DE) genes and the proportion called DE corresponded well

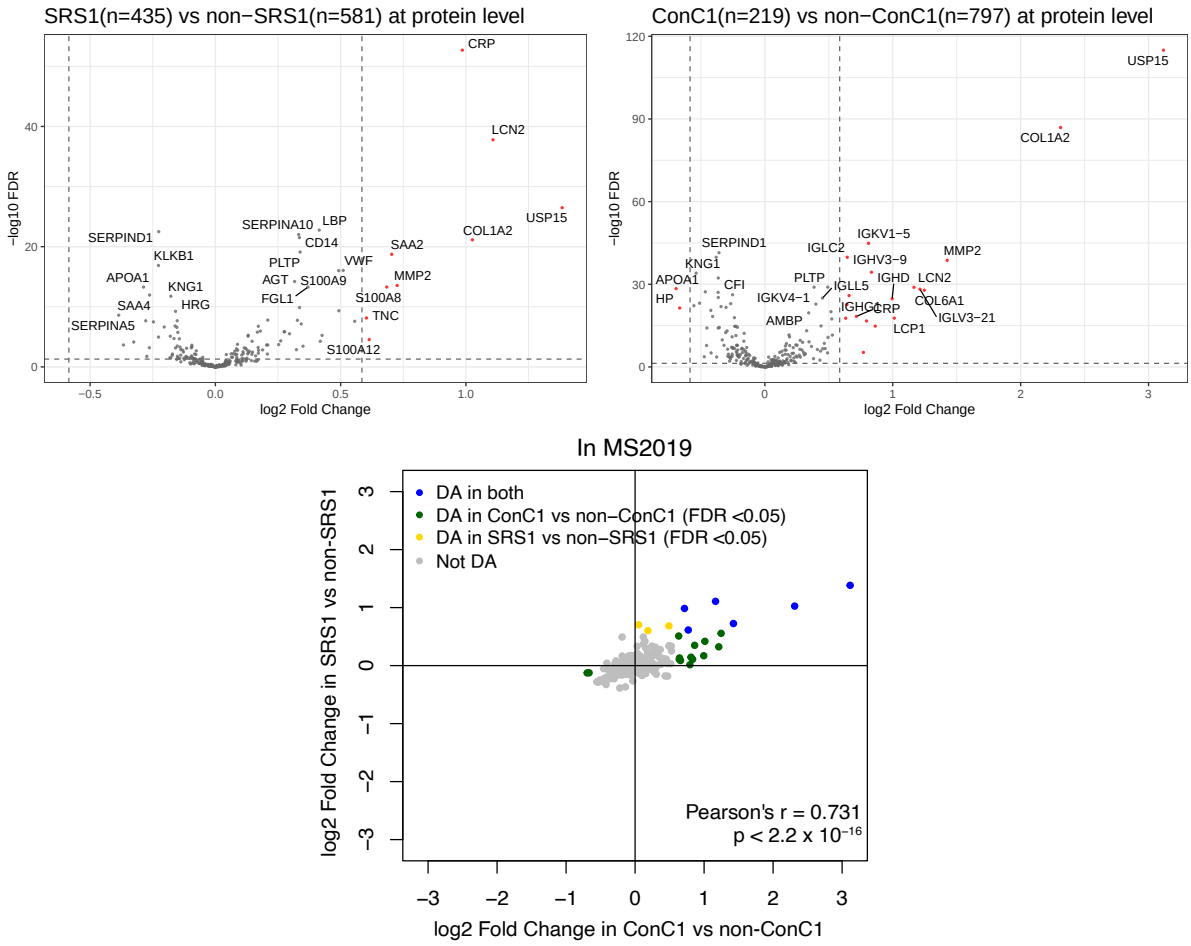


Figure 4.25: Protein abundance was compared between SRS1 vs non-SRS1, and between ConC1 vs non-ConC1 using MS2019 discovery and validation cohorts combined. Proteins on the right-hand side of each plot were more abundant in SRS1 or ConC1. Red points labelled by gene names denote differentially abundant proteins (FDR<0.05 and |FC|>1.5), which were labelled as "DA" on the correlation plot. For patients with more than one samples, only the first available samples were included in the contrasts. Age and sex were included as covariates.

between microarray and RNA-seq data (Fig.4.26(a-d)). No difference was observed between ConC2 and ConC3, thus the two clusters were combined to compare with ConC1. Unlike at the protein level, genes were up- or down- regulated in both directions in SRS1 and in ConC1. As expected, both the fold changes and the significance were larger in the SRS1 contrast than in the ConC1 contrast. Fold changes showed a strong correlation between the two contrasts, with more genes DE in the SRS contrasts and no genes DE in opposite directions (Fig.4.26(e-f)).

Considering the large number of genes tested, enrichments in genes DE in SRS1 or ConC1 were tested using only the top 2% genes (ranked by significance) in either direction against all genes tested as the background, using only GOBP annotations. There was also a large overlap between the pathways enriched in the top genes (Table 4.12).

The strong correlation in fold changes and the large overlap in enriched pathways suggested that most biological processes were altered in the same direction in SRS1 or in ConC1, but at different magnitudes, suggesting that there are both shared and distinct molecular mechanisms that differentiate the SRS or ConC clusters. As expected from the data type from which the clusters were derived, the magnitudes were larger at protein level for the ConC1 contrast, and larger at gene expression level for the SRS1 contrast.

Both genes higher in SRS1 and those higher in ConC1 were enriched for biological processes including neutrophil degranulation and oxidation-reduction process. Both genes lower in SRS1 and those lower in ConC1 were enriched for pathways including: adaptive or innate immune response, T cell costimulation, and cell surface receptor signalling pathway. These indicate that an active regulation of immune responses is a distinctive factor for both the SRS and ConC clusters; oxidation-reduction process was up-regulated in both SRS1 and ConC1, while T cell related signalling pathways were down-regulated in both SRS1 and ConC1. Meanwhile, there were differences in pathways distinguishing the SRS or ConC clusters: MHC II antigen presentation and

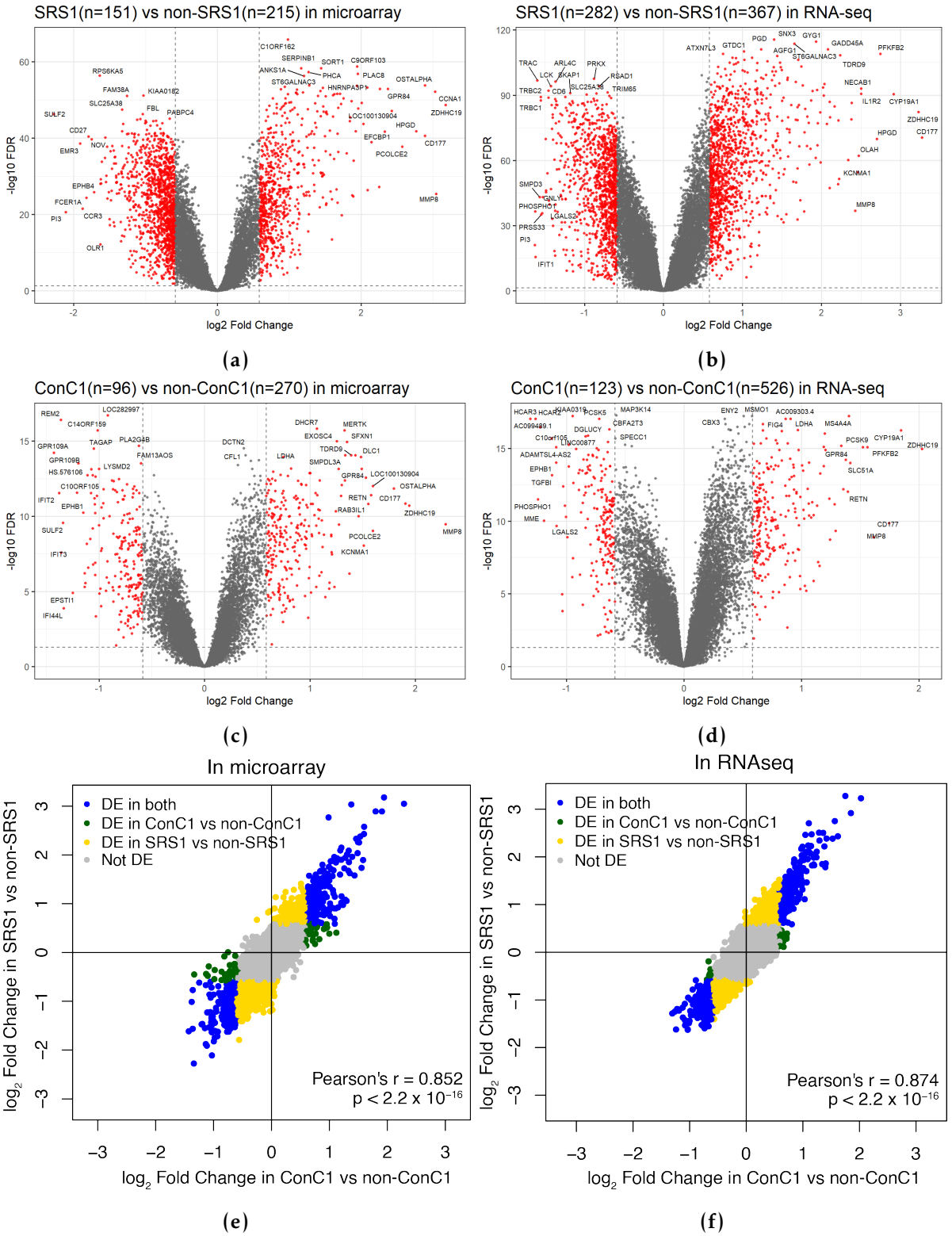


Figure 4.26: Comparison of gene expression between SRS or ConC clusters. In the volcano plots **(a-d)**, gene probes on the right-hand side of each plot were higher expressed in SRS1 or ConC1. Red points labelled by gene names denote differentially expressed genes (FDR<0.05 and |FC|>1.5), which were labelled as "DE" on the correlation plots **(e-f)**. For patients with more than one samples, only the first available samples were included in the contrasts. Age and sex were included as covariates.

Table 4.12: Comparison of enriched pathways in SRS1 or ConC1 contrasts. Significantly enriched terms (FDR<0.05) are listed in the table. Terms shared between the SRS1 and ConC1 contrasts were underlined. **At protein level,** top 10% proteins by significance were tested for enrichment against the total 269 proteins as background, using both GOBP and Reactome annotations. A minimum of 5 overlaps between input data and genes in the terms was required for the term to be tested. **At gene expression level,** top 2% genes by significance were tested for enrichment against the total genes as background, using only GOBP annotations. A minimum of 10 overlaps between input data and genes in the terms was required for the term to be tested. Redundant terms have been removed by the function xEnrichConciser. Only terms identified in both microarray and RNA-seq data are listed.

	Higher in SRS1 or ConC1	Lower in SRS1 or ConC1
SRS1 vs non-SRS1, at protein level	neutrophil degranulation; inflammatory response; acute-phase response; ECM organization; innate immune response; cytokine signalling; TLR cascades.	Not tested
ConC1 vs non-ConC1, at protein level	defense response to bacterium; regulation of immune response; Fc-gamma receptor signalling involved in phagocytosis; cytokine signalling; leukocyte migration; complement activation, classical pathway; receptor-mediated endocytosis; innate immune response.	Not tested
SRS1 vs non-SRS1, at gene expression level	innate immune response; neutrophil degranulation; oxidation-reduction process.	adaptive immune response; antigen processing and presentation of exogenous peptide antigen via MHC class II; cell surface receptor signalling pathway; immune response; inflammatory response; innate immune response; leukocyte migration; positive regulation of ERK1 and ERK2 cascade; regulation of immune response; T cell activation; T cell costimulation; T cell receptor signalling pathway.
ConC1 vs non-ConC1, at gene expression level	cell division; innate immune response; mitotic cell cycle; neutrophil degranulation; oxidation-reduction process; response to drug.	adaptive immune response; apoptotic process; cell adhesion; cell surface receptor signaling pathway; defense response to virus; G-protein coupled receptor signalling pathway; immune response; inflammatory response; innate immune response; interferon-gamma-mediated signalling pathway; regulation of immune response; response to virus; T cell costimulation; T cell receptor signalling pathway; type I interferon signalling pathway.

ERK1/2 cascade were down-regulated in SRS1; cell division and interferon signalling pathway differentiated the ConC clusters.

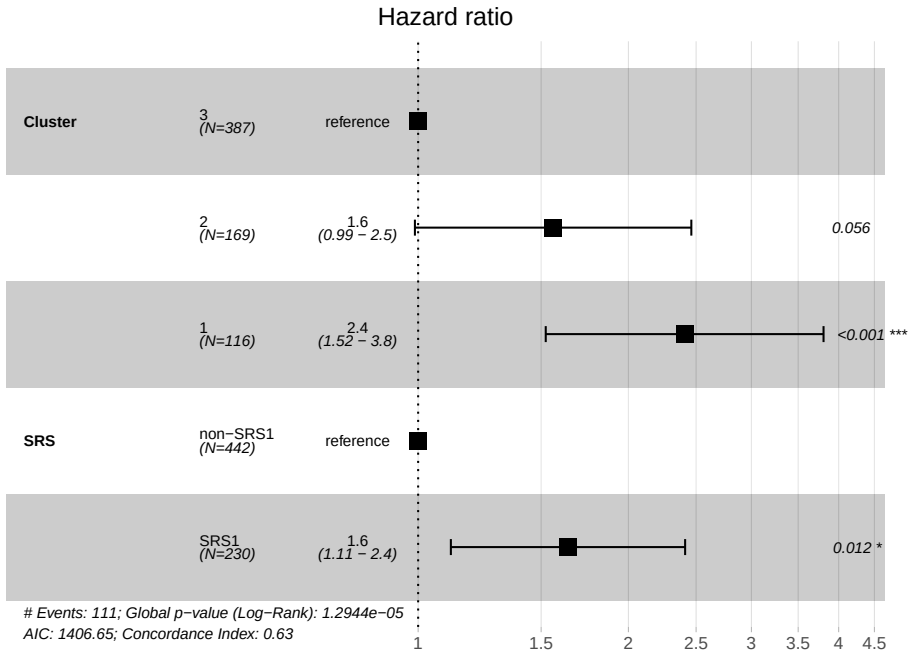
4.5.3 Combining the two classifications for risk stratification

Both SRS and ConC stratify sepsis patients into subgroups with significantly different mortality. There were both shared and distinct features between the two classifications in terms of cluster membership, cluster movement, and molecular characteristics. Therefore, I then analysed whether risk stratification could be further improved by using both classifications together.

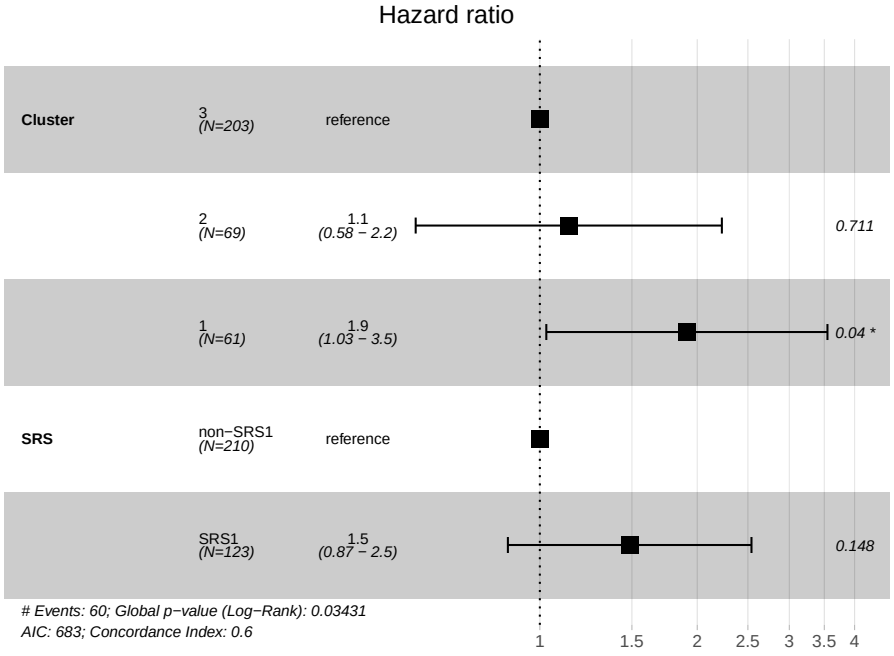
Including SRS and ConC in one multivariate Cox proportional hazard model indicated both to be independent risk predictors of 28-day mortality in the discovery cohort (Fig. 4.27), with a higher hazard ratio for ConC1 (HR (95%CI)=2.4 (1.5-3.8), p<0.001) than for SRS1 (1.6 (1.1-2.4), p=0.012). In the validation cohort, a higher hazard was also observed for ConC1 (1.9 (1.0-3.5), p=0.04), while SRS was not an independent risk predictor (1.5 (0.87-2.5), p=0.15), which could be partly attributed to the smaller sample size and the interaction between the two classifications.

In the discovery cohort, out of the six combinations, the 10.3%(69/672) patients assigned to both ConC1 and SRS1 had the lowest survival at both 28-day and 6-month, followed by the two groups of patients assigned to one of the more severe clusters (ConC2_SRS1 and ConC1_non-SRS1) (Fig. 4.28(a, b)). Compared with the lowest risk ConC3_non-SRS1 group which constituted 41.5% of the patients, the three higher risk groups had elevated 28-day mortality risks of 3.9, 2.6, and 2.4 folds, respectively (Fig. C.14(a)).

In the validation cohort, out of the six combinations, only the more severe combination of 12.3% (41/333) ConC1_SRS1 patients had a significantly higher risk (3.0 (1.5-6.0), p=0.002) compared with the reference ConC3_non-SRS1 group (Fig. 4.28(c, d), Fig. C.14(b)). The ConC1_non-SRS1 group had an increased HR of 2.0 but did not reach

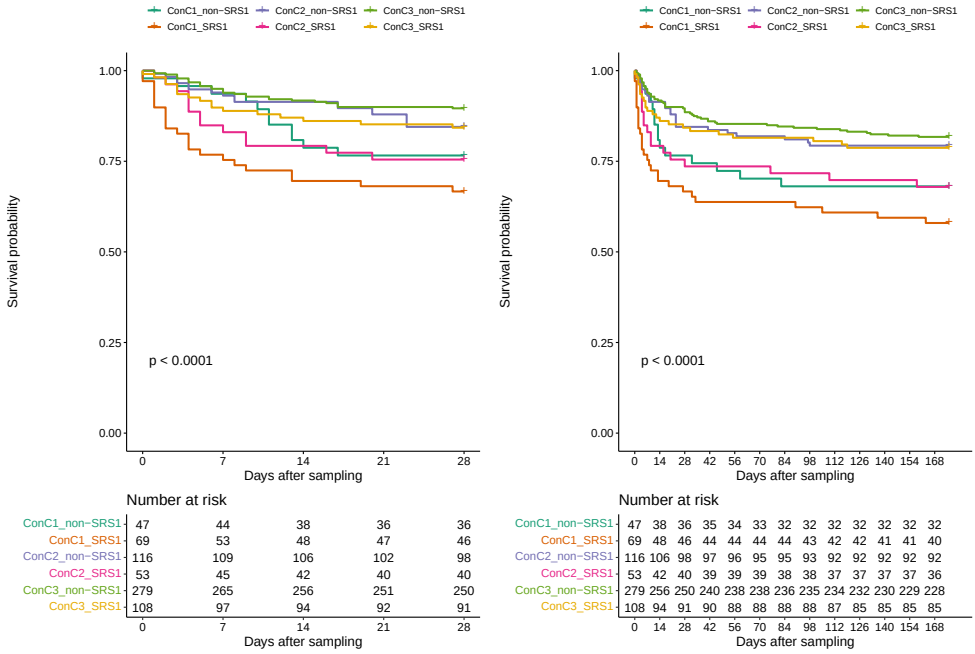


(a) In discovery cohort



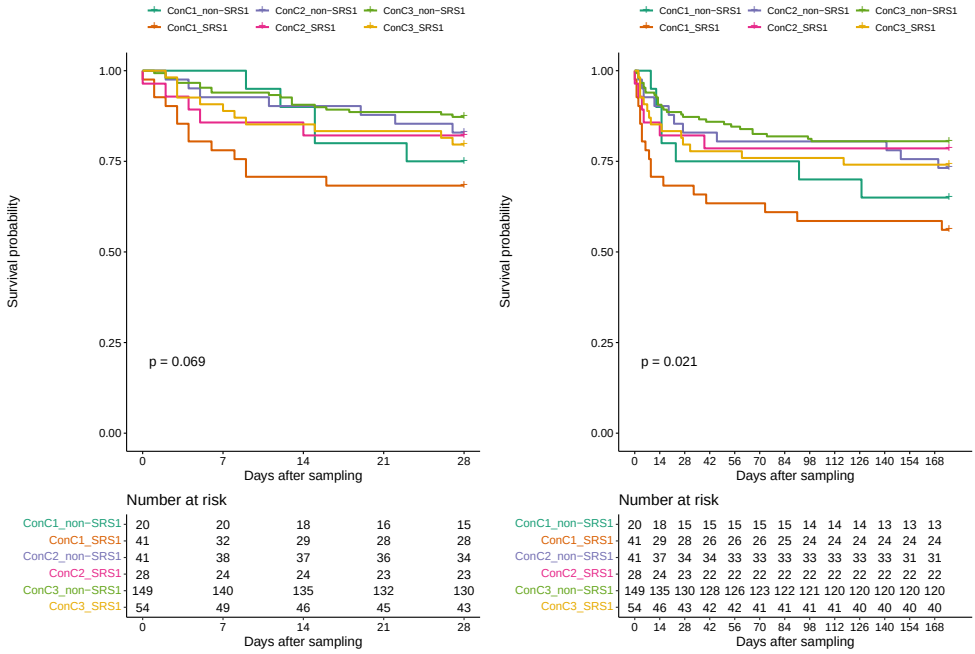
(b) In validation cohort

Figure 4.27: Multivariate Cox proportional hazard regression on 28-day mortality considering both ConC and SRS classifications. Model is built on 672 or 333 patients with both ConC and SRS assignments available in the discovery or validation cohort, respectively. Cluster assignments of the last available samples per patient with both SRS and ConC assignments were used. Hazards at 28-day post-sampling were compared.



(a) In discovery cohort, 28-day

(b) In discovery cohort, 6-month



(c) In validation cohort, 28-day

(d) In validation cohort, 6-month

Figure 4.28: Kaplan-Meier curves comparing survival between six groups of patients with a combined ConC and SRS classification. Group assignments of the last available samples per patient were used.

significance (p=0.18). In conclusion, the combination of ConC1 and SRS1 consistently identified a group of one-tenth of the patients that had an increased HR of 3 to 4 compared with the reference low-risk group. The higher-risk classifications (ConC1, ConC2, SRS1) on their own also contributed to patient stratification but with a varying level of significance.

4.6 Results: Validation of the clusters in VANISH

As well as GAinS, sepsis patients from the VANISH study were also included in MS2019. VANISH was a randomised clinical trial comparing between vasopressin and norepinephrine, or between hydrocortisone and placebo, enrolling adult septic shock patients (Gordon et al. 2016). I predicted and characterised the proteomics clusters in VANISH to see whether the differences in clinical characteristics including outcome could be replicated, as well as to have a preliminary understanding of whether there is interaction between proteomic clusters and the trial drugs.

4.6.1 VANISH samples

There were 44 VANISH patients included in MS2019 after QC, each with a baseline (TP0) sample and a varying availability of later timepoints (TP1/2/3). TP0 samples were mostly taken within 6hr of onset of septic shock and before the study drugs were given. TP1/2/3 sample were taken roughly at 24hr/48hr/96hr following TP0, respectively. The baseline clinical characteristics and outcome data were summarised in Table C.10. All VANISH patients included here had lung infections and required vasopressors despite adequate intravenous fluid resuscitation.

The proteomic clusters identified in GAinS were predicted in the 148 samples using the 181-protein 3-cluster model as described in Section 4.3.1. Considering all timepoints, there were 25/23/100 samples assigned to ConC1/2/3, respectively, with ConC1 constituting 25% of baseline samples but only 7% of TP3 samples (Table

Table 4.13: VANISH patient numbers assigned to each of the ConC clusters in each of the timepoints.

	TP0	TP1	TP2	TP3	Total
ConC1	11	9	3	2	25
ConC2	4	7	6	6	23
ConC3	29	22	27	22	100
Total	44	38	36	30	148

Table 4.14: Numbers and proportions of patients assigned to the three clusters at baseline, in VANISH or in GAinS subsets.

	ConC1	ConC2	ConC3	Total
GAinS CAP&FP, Day1	155 26.2%	129 21.8%	308 52.0%	592
GAinS CAP, Day1	56 14.7%	94 24.7%	230 60.5%	380
GAinS CAP with vasopressors, Day1	32 25.8%	25 20.2%	67 54.0%	124
VANISH, TP0	11 25.0%	4 9.1%	29 65.9%	44

4.13, Fig. C.15). A subset of GAinS patients who had sepsis due to CAP and also required vasopressors was most comparable to the VANISH cohort. A closely-matched proportion of 25.8% of the Day 1 samples of this GAinS subset was assigned to ConC1 (Table 4.14). Hierarchical clustering (Fig. 4.29) and PCA transformation (Fig. C.16) showed similarity in protein profiles between serial samples of the same patient, but not within samples of the same timepoint. ConC1 or ConC2 samples tend to cluster in certain separate branches.

4.6.2 Comparing clinical characteristics and outcome

Baseline clinical characteristics

Patient demographics, comorbidities, baseline clinical measurements as well as outcome were compared between the three patient clusters using baseline assignments. As there were only 4 patients assigned to ConC2 at TP0, only the

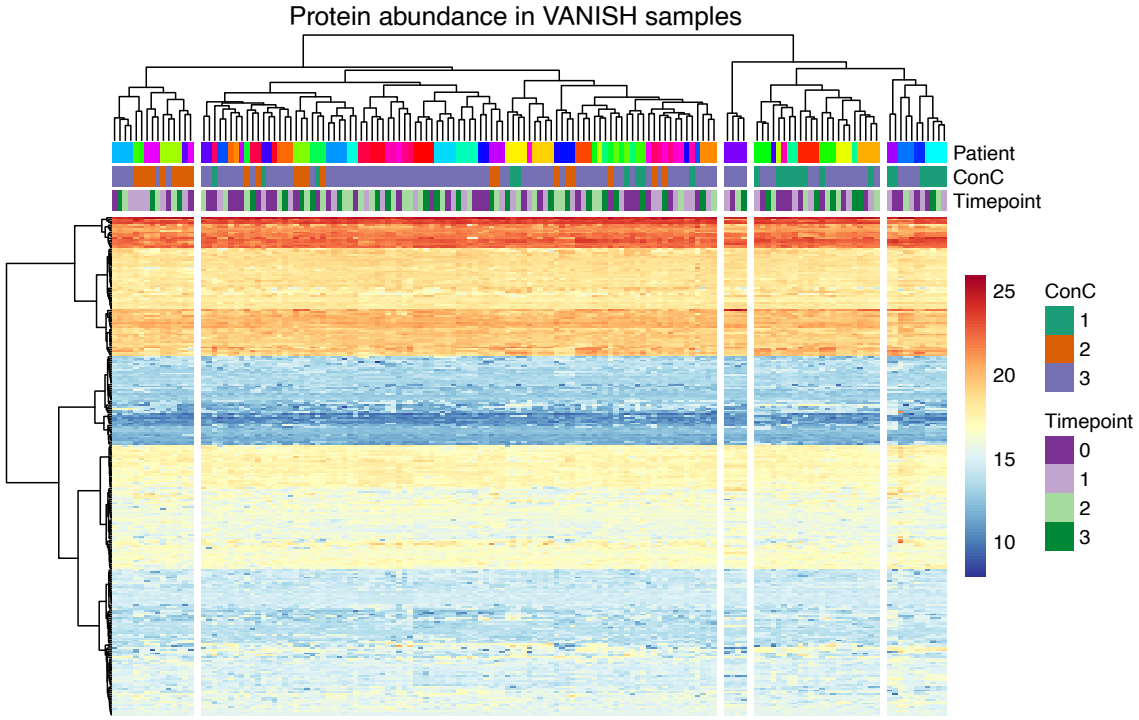


Figure 4.29: Hierarchical clustering of 148 VANISH samples (columns) on abundance of 269 proteins (rows).

comparison between ConC1 and ConC3 was meaningful to interpret. There was no significant difference (Kruskal-Wallis tests FDR <0.05) observed in either the numerical or categorical clinical variables (Fig. 4.30, Fig. C.17). Without correction for multiple testing in K-W tests, ConC1 patients had higher heart rate than ConC3 (K-W p=0.029, Dunn's post-hoc test FDR=0.029), and tended to have lower platelets (K-W p=0.058) or higher pre-existing condition of cirrhosis (χ^2 test p=0.070) than ConC3.

The survival curves

There was no difference in 28-day mortality observed between the three clusters using either baseline or TP3 assignments (Fig. 4.31). Only 1 out of the 4 patients assigned to ConC1 at TP3 survived at 28 days.

The trial reported no difference in outcome between noradrenaline or vasopressin treatment (drug1), or between hydrocortisone or placebo treatment (drug2) (Gordon et al. 2016). The SRS transcriptomic endotypes were shown to interact with the drug2

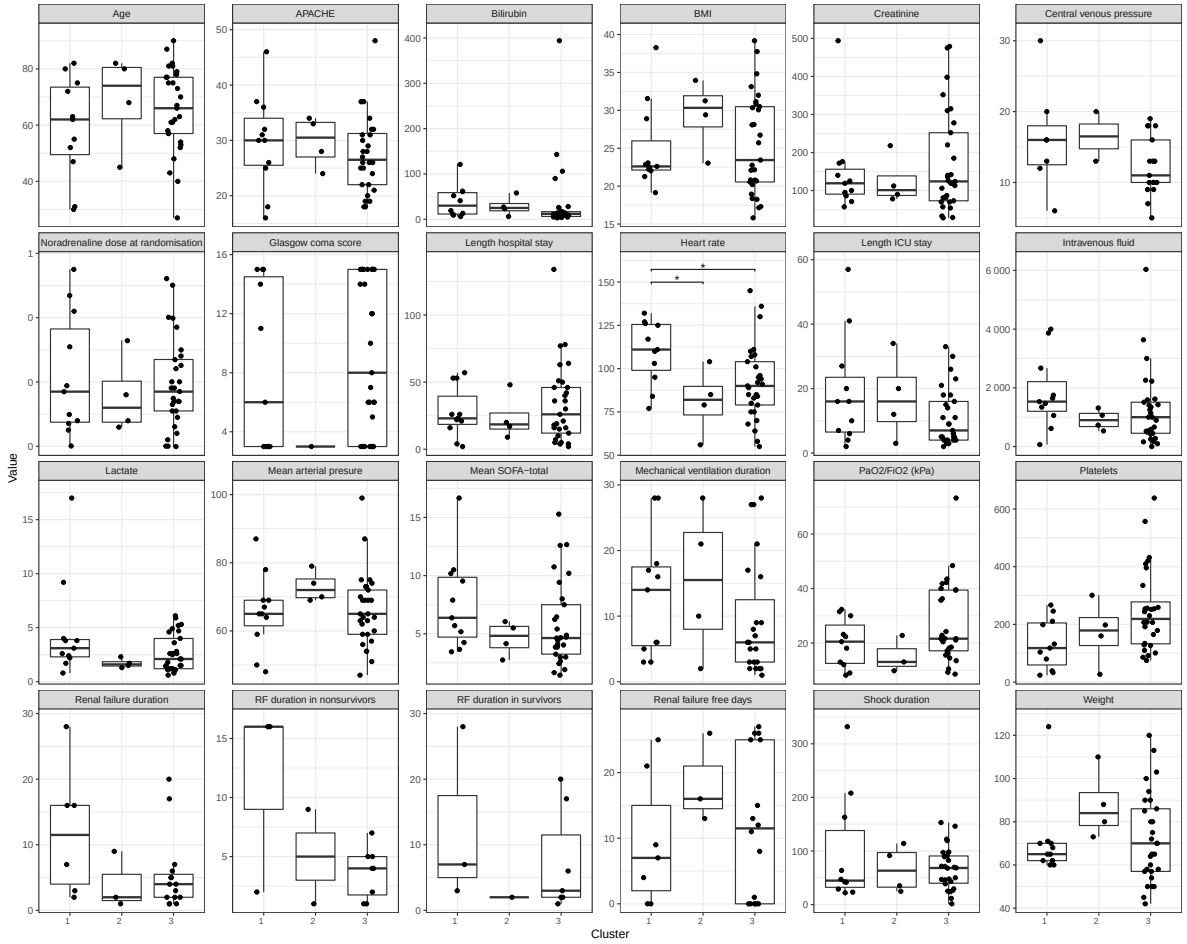


Figure 4.30: Comparison of numerical clinical variables between the ConC clusters in VANISH, using baseline measurements and baseline cluster assignments. Patient numbers were at most 11/4/29 for ConC1/2/3 respectively, subject to availability. Significance was labelled by Dunn's post-hoc test p values after correcting within the 3 contrasts. The clinical variables should be interpreted the same as detailed in Table C.10.

treatments. In SRS2 but not in SRS1 group, patients receiving hydrocortisone had significantly worse outcome compared with placebo (Antcliffe et al. 2019). Working with a much more restricted number, I did not see a significant difference between drug2 groups, within patients assigned to either ConC1 or ConC3 at baseline (Fig. 4.32). However, the survival rate was lower in hydrocortisone compared with placebo (without statistical significance) in ConC3 but not in ConC1 group, consistent with the enrichment of SRS1 patients in ConC1.

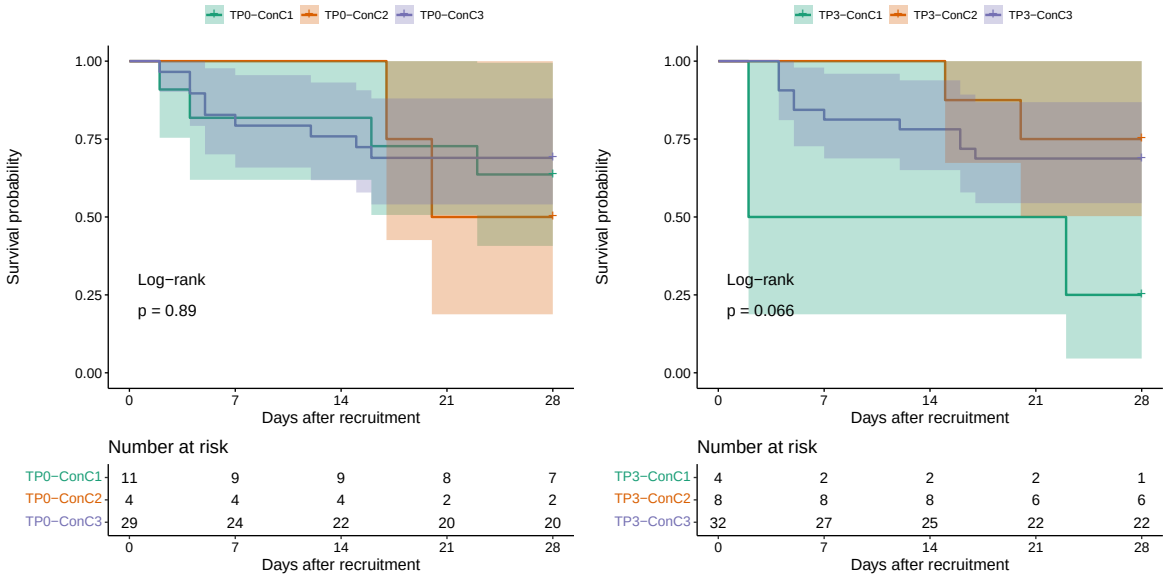


Figure 4.31: Kaplan-Meier curves to compare 28-day survival between the three proteomic clusters (ConC), using cluster assignments at either TP0 or TP3. Patient groups by trial drugs were not distinguished in this analysis. Shades on the curves show 95% confidence intervals.

4.6.3 Cluster membership movement

The serial samples available also enabled me to understand the trajectories of proteomic cluster membership within the VANISH patients. Considering the movement between successive timepoints, most of the ConC3 patients stayed in ConC3; most of ConC1 patients either stayed in ConC1 or moved to ConC2 in TP0 to TP1, and moved to ConC3 in later intervals (Table 4.15). Dynamics in the proteomic cluster assignments was visualised in all movements or in only patients with the four complete timepoints sampled (Fig. 4.33). ConC2 had a similar patient number

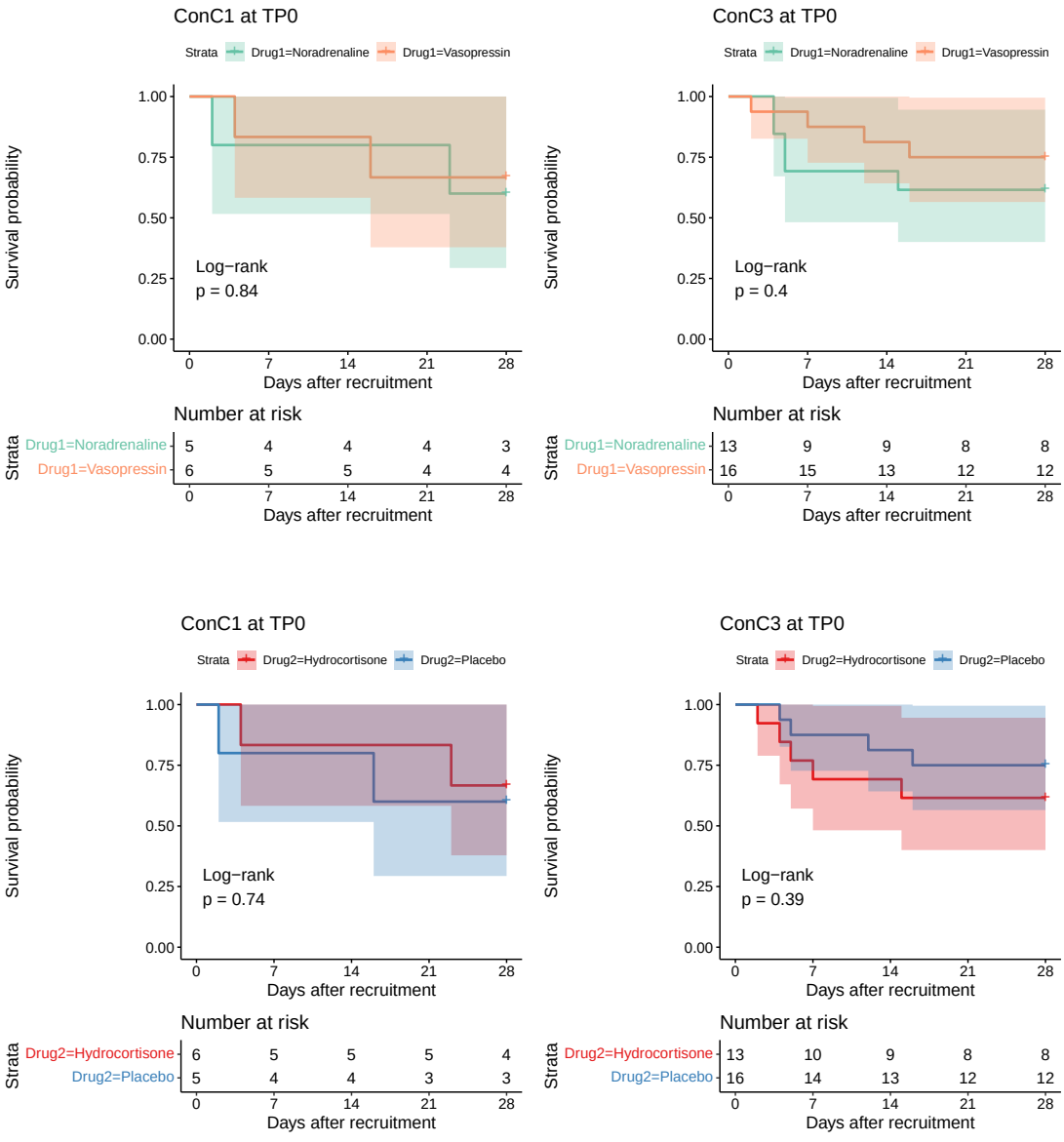


Figure 4.32: Kaplan-Meier curves comparing survival between patients receiving different trial drugs, restricted to patients assigned to ConC1 or ConC3 at the baseline timepoint.

Table 4.15: Patient numbers of proteomic cluster movement in VANISH. Only movements between two samples available at successive timepoints were included.

TP0 to TP1	To ConC1	To ConC2	To ConC3
From ConC1	4	3	1
From ConC2	0	1	2
From ConC3	5	3	19
TP1 to TP2	To ConC1	To ConC2	To ConC3
From ConC1	2	0	4
From ConC2	0	1	5
From ConC3	1	3	16
TP2 to TP3	To ConC1	To ConC2	To ConC3
From ConC1	1	0	2
From ConC2	0	3	1
From ConC3	0	3	15

between TP0 and TP3. Only two of the patients with TP3 sampled stayed in ConC1. ConC3 constituted larger proportions of patients at TP3 or TP2, than at TP1 or TP0. When cluster movements were calculated separately within patients who received either hydrocortisone or placebo, there was no difference observed between the two drug2 treatment groups.

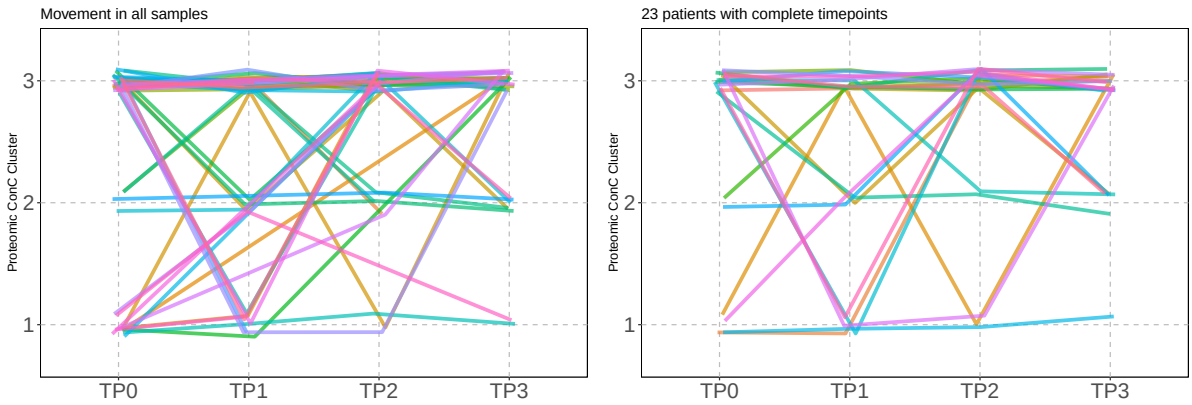


Figure 4.33: Trajectory of proteomic cluster movements in VANISH. Each coloured line represents one patient. Positions of the line endpoints are jittered vertically and horizontally. There is not an ordered relation between the three ConC clusters on y-axis.

4.7 Discussion

4.7.1 A retrospective power calculation

The statistical power calculation performed in section 3.2.2 was based on a pre-existing dataset (MS192). To understand the relation between sample size and power based on the form of data in MS2019 and for the questions explored in this chapter, I performed a retrospective power calculation taking the discovery cohort contrasts of GAinS (n=788) vs XMIN_1 (post-operation, n=103) and ConC1 (n=170) vs ConC3 (n=435) as examples. Using both the FDR<0.05 and |FC|>1.5 cut-offs, there were 23 and 20 out of 269 proteins that were differentially abundant in the two contrasts, respectively, giving estimates of π (proportion of true signal) to be 8.6% and 7.4%.

To achieve 80% power and FDR<0.05 significance for 90% of the analytes, 108 biological replicates are needed in each group for the sepsis-surgery contrast, while 130 are needed for the within-sepsis (ConC1 vs ConC3) contrast (Fig.4.34). These numbers are close to the actual sample numbers included in the contrasts, and giving confidence that the larger sample sizes could enable better identification of true signals and the discovery of within-sample structures with greater granularity. The sample sizes calculated here are smaller than those estimated based on MS192 mainly because of a smaller variable variance observed in MS2019. Notably, since the effect size and variance are estimated based on MS2019, this calculation cannot be used for describing the power of this study itself, but rather is informative for future validation studies to be designed.

4.7.2 Molecular characteristics of the three clusters

In Section 4.4.3 I summarised the molecular characteristics of the three clusters as a summary table. Some of the characteristics listed are further discussed here.

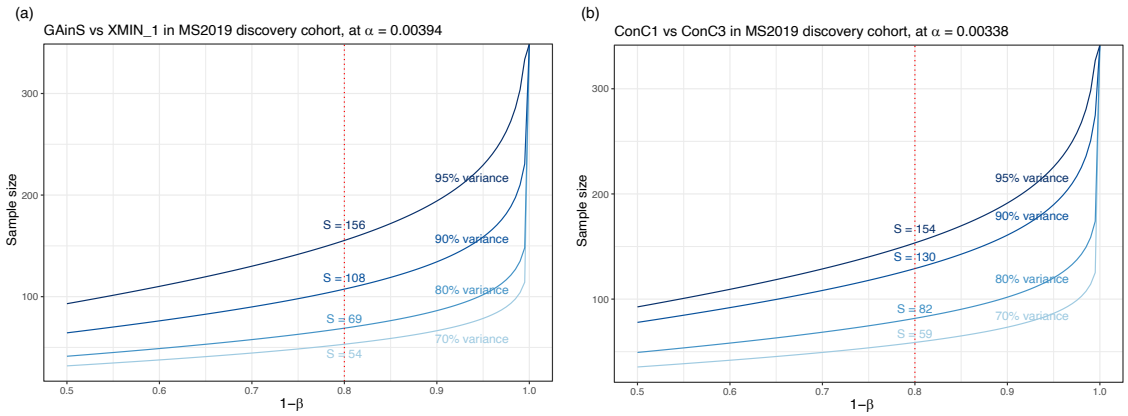


Figure 4.34: Relation of sample size and statistical power calculated based on MS2019 data, using α values to control FDR⩽0.05. Separate curves are given for proteins with different variance at the percentiles stated, with more variable proteins at higher percentiles. The point of intersection between the blue curves and the red dotted line (power=80%) denote the minimum sample size (in each group) required to achieve the desired power and significance level for 70%/80%/90%/95% of the analytes.

Interleukin signalling in ConC1 and ConC3

Interleukin or cytokine signalling was enriched in proteins with higher abundance in ConC1 in both the MS2019 discovery and validation cohorts (Fig. 4.10, Fig. 4.17). More specifically, IL-4 and IL-13 signalling was enriched in proteins higher in ConC1 or lower in ConC3 in MS192 (Fig. 4.18). In terms of the individual protein functions, higher COL1A2, LCN2, and MMP2 in ConC1 indicate up-regulation of IL-4 and IL-13 signalling through STAT3 signalling; higher LCP1 in ConC1 may indicate down-regulation of IL-12 signalling which leads to up-regulation of LCP1 in $CD4^+$ T cell junction (Rosengren et al. 2005); higher S100A12 in ConC1 secreted by neutrophils will activate the MAP-kinase and NF-κB signalling pathways, leading to production of pro-inflammatory cytokines. The signalling pathways were retrieved through the Reactome database (Fabregat et al. 2018).

In the Luminex data, the inflammatory cytokines were higher in ConC1, but IL-4, IL-13 and IL-12p70 were not significantly different in ConC1 compared with either ConC1 or ConC3 (Fig.4.19). Therefore, although the term interleukin signalling was indicated from the more abundant proteins, there was limited support from the actual cytokine

levels measured. It is possible that the altered interleukin levels could be observed in the intercellular micro-environments, but not at the total circulation level reflected in the Luminex data. Measuring more proteins and genes in the interleukin pathways in the relevant tissue or cell types could help clarify this in future studies.

B cell signalling in ConC1

Higher activity of B cell signalling in ConC1 was indicated from the ConC1vs2 contrast as well as the proteins unique to comparing ConC1 but not other clusters against HV in the discovery cohort. DA proteins annotated for this term were mostly immunoglobulins (IGHA1, IGHD, IGHG1, IGHG3, IGHV3-74, IGLL5). This group of proteins were also enriched for phagocytosis or positive regulation of B cell activation.

At the cytokine level, APRIL was significantly more abundant in ConC1 as compared to ConC3. APRIL (A proliferation-inducing ligand, also known as TNFSF13) is important for the survival and maturation of B cells. Upon secretion, APRIL interacts with two common receptors on B cells and plasma cells: the B cell–maturation antigen which promotes the survival of plasma cells, and the transmembrane activator and cyclophilin ligand interactor which supports T cell-independent B cell antibody responses and isotype switching (Abe et al. 2015). APRIL-deficient mice exhibited impaired mucosal humoral immunity (Huard et al. 2008).

USP15 in ConC1

Although USP15 is not a secreted protein, it is detected in plasma, more in the inflammation groups (in ~40% samples in sepsis,~20% samples in non-sepsis controls) and less in the healthy/pre-operation groups (<10% detection). In the complete quantifications after imputation, it was more abundant in sepsis compared with all the other control groups.

The protein profile of ConC1 has a feature of high abundance of two proteins USP15 and COL1A2 with high fold change and significance. COL1A2 (Collagen alpha-2(I)

chain) is a major structural protein in the ECM, while USP15 (Ubiquitin carboxyl-terminal hydrolase 15) is known to have important roles in regulating inflammation: USP15 is a positive regulator in TNFα- and IL-1β-induced NF-κB activation (Zhou et al. 2020); it inhibits anti-tumour T cell responses (Zou et al. 2014), and negatively regulates virus-induced RIG-I-dependent type I interferon induction pathway (Zhang et al. 2015); another similar deubiquitinating enzyme in the family, USP25, reduces the susceptibility of mice to LPS-induced septic shock (Zhong et al. 2013).

At the gene expression level, there was no difference observed in USP15 between the three clusters; compared with non-SRS1, it was higher expressed in SRS1 but the fold change did not reach 1.5-fold cut-off (Table C.9).

Therefore, USP15 is a potential marker protein with plausible biological functions in regulating the immune functions, although evidence at gene expression was less clear. To establish it as a biomarker for sepsis subgroups, more studies are needed to elucidate the function in the context of sepsis or in LPS-response of primary immune cells, and to validate the measurement of USP15 in plasma in independent cohorts potentially with targeted approaches.

4.7.3 Tissue/cell type origin of plasma proteins

Based on the clinical data, the more severe illness in patients from ConC1 was not restricted to a certain organ but was observed across the organs or organ systems. On the protein level, it is not clear whether the plasma proteins had a different composition of tissue origins between patient subgroups. Plasma proteins are generally products of liver reticuloendothelial cells, with contributions from the bone marrow, degenerating blood cells, spleen, B lymphocytes, and general body tissues, with a varying level of tissue enrichment or tissue specificity. In disease conditions or in certain patient subgroups, it is meaningful to investigate whether organs have altered contributions to the blood stream, and whether there is a varying level of tissue damage and cell necrosis. Most proteins measured in MS2019 had at least one

annotation in extracellular regions based on the GO Cellular Component annotations. Proteins with only intracellular annotations was not observed to correlate with sepsis severity.

For a similar aim of determining the tissue origin but for the genes instead of proteins detected, Aran et al. (2017) inferred immune and stromal cell type compositions from transcriptome profile of tumour microenvironments. Following a similar approach, it may be possible to determine tissue or cell type contributions to each sample's plasma proteome based on tissue or cell type specific gene expression and protein abundance reference datasets (Uhlén et al. 2015; Jiang et al. 2019). This calculation would be based on two assumptions: the tissue specificity of the proteins do not change in disease conditions, and that across the tissues, the efficiency from intracellular protein abundance (or gene expression) to secretion and protein detected in blood are the same. However, the MS2019 data measured plasma levels instead of a combination of the solid tissues or blood cells available from these datasets, so the calculation of contributions would be heavily caveated and discordance between protein abundance allocated by protein or RNA reference datasets may be observed. Correlation with clinical data, for example between PMN counts and calculated PMN contribution to protein abundance, could be a way to validate the approach.

Following similar assumptions, Malmstrom and colleagues (2016) constructed a distribution map of the tissue proteomes in healthy mice across ten highly vascularised organs or blood cell types or plasma, using DIA-MS (data-independent acquisition mass spectrometry). Based on this map they predicted a primary tissue origin (including "common") for the plasma proteins and analysed the overlap between the origins and protein clusters that showed different dose-dependent patterns in septic mouse models injected with different doses of *S. pyogenes* bacteria. They showed that a cluster of proteins only up-regulated in the highest dose group almost all had the origin from surrounding organs and cells but not from plasma, and were enriched for essential intracellular functions (e.g. cell cycle, glycolysis, mitosis). This suggested

cell necrosis and vascular endothelial damage in the most severe animals, which aligned with our observations in ConC1 patients. However, the accuracy of their prediction of primary tissue origin for plasma proteins remain to be validated, and tissue-enrichment in addition to tissue-specificity should be considered to determine the organ/cell type contributions in a more quantitative approach.

Further investigation on this topic would require a better understanding of which tissues or cell types each plasma protein is secreted from in disease conditions, as well as considering cell necrosis from the most relevant tissues. Furthermore, proteomics datasets differentiating between protein isoforms may better suit the purpose, since proteoforms (that capture alternative splicing and posttranscriptional or posttranslational processing) have higher cell-type specificity than their corresponding proteins so can better indicate the tissue/cell-type origins, as reported by Melani et al. (2022).

4.7.4 Differences between the MS2019 and MS192 dataset

The MS192 dataset was used in this chapter to compare a wider range of protein species between the ConC clusters in 148 samples overlapping between MS2019 and MS192. For this aim, the ConC assignments used were based on MS2019 data, instead of being predicted in MS192 as performed for the MS2019 GAinS validation cohort. This was due to intrinsic differences in the technical platforms used for the two datasets, leading to only limited correlation between protein levels measured in the two datasets across the overlapping samples. PCC across samples were >0.3 for only 45 out of the 177 overlapping proteins.

To understand from a technical standpoint why proteins have a variable range of correlations across the two platforms, I tested the correlation of Pearson's r (PCC) with sequence coverage and the number of peptides contributing to the identification of each protein, both reported by the Fragpipe software in MS2019. There was a weak correlation with coverage (PCC=0.252), with most proteins with PCC >0.3 being

Table 4.16: Differences in analytical workflow between the two sepsis mass spectrometry datasets.

	MS192	**MS2019**
Depletion	Top12 high-abundance proteins depleted	Not depleted
Mass spectrometer	QE-HF (Thermo Fisher)	TimsTOF (Bruker)
Gradient time	50min	15min
Protein identification platform	Progenesis + PEAKS	Fragpipe
Batches	Two different batches	
Protein filter	⩾70% in all samples	⩾50% in any sample group
Protein number post-filtering	1123	269
Sample number post-filtering	192	2575

detected with coverage >70%. No correlation was observed with the number of unique stripped peptides (peptides with no post-translational modification).

Among the technical differences listed between the two datasets (Table 4.16), depletion of the 12 high-abundance proteins performed in MS192 but not in MS2019 should have the largest influence on protein correlations. Biological samples like blood plasma and cerebrospinal fluid usually have a complex protein composition and a high dynamic range of concentrations including several highly abundant proteins like albumin and immunoglobulins. Affinity depletion is a commonly applied pre-analytical process used to reduce the complexity of such biological samples by removing the most abundant proteins, thus enhancing the detection of the less abundant ones. Studies have shown that depletion could greatly increase the number of proteins detected in MS, but at the same time a considerable number of non-targeted proteins could also be bound to the depletion material and thus not be identified in the depleted samples (Jankovska et al. 2019) or show an altered relative abundance (Tu et al. 2010). In-house experiments from the TDI showed a good correlation (PCC across proteins >0.98) when one non-depleted sample was ran on either TimsTOF or QE-HF, while the correlation was much lower (PCC across proteins ∼0.77) when one Top12 depleted and one non-depleted sample were ran in a same workflow. The depletion greatly altered greatly the intensity of the non-targeted proteins as well. The varying level of correlation between proteins in depleted or non-depleted samples could be partly attributed to a different level of interaction with the targeted depleted proteins,

or with the carrier proteins interacting with the twelve targets. The level of interaction could also be modulated in disease settings.

These differences in analytical workflow, especially in depletion, resulted in a different pool of proteins being characterised in the two datasets: 1123 with depletion, or 269 without depletion (post-filtering numbers). There was no interaction between unsupervised clusters defined from the two datasets, and only limited accuracy in predicting the MS2019-based proteomic clusters in MS192 data. Future validation of the clusters should also be performed based on non-depleted datasets.

4.7.5 Dynamic interaction between SRS and ConC groupings

The actual corresponding timescale between SRS and ConC group membership movements is yet to be determined. There should be a time-lapse between gene expression and protein secretion, and a time-lapse between an adverse host response reflected by gene expression and an altered tissue bed reflected in the circulation. The exact timescale of these processes are not clear and could differ across genes as well. The available sampling timepoints in GAinS are day1/3/5 after ICU admission. In the main result I looked at any movements observed within day1-5 without making an assumption of the corresponding timescale of moving in or out of SRS1 or ConC1. As an alternative approach I also compared the class movements of SRS vs ConC in day1 to day3, SRS vs ConC in day 3 to day 5, and SRS day1 to day3 vs ConC day 3 to day5. Significant interactions were found for the first two contingency tables (χ^2 tests $p<0.001$; patient numbers n=142, 170).

Another approach would be to compare the change in SRS1 or ConC1 membership in a quantitative way, instead of dividing patients into four groups based on switching between the classes. To enable this, a quantitative score for ConC1 membership needs to be generated at a comparable scale. A possible approach is to use diffusion maps as for generating the SRSq scores as described by Cano-Gamez et al. (2022).

4.7.6 Towards selecting a useful protein signature panel

Both 3-cluster prediction models and 2-cluster prediction models have been investigated in Section 4.3.3. Because of the much clearer indication of differences in clinical severity and molecular profiles at both the protein abundance and gene expression level in ConC1 compared with the two other clusters, being able to distinguish ConC1 from the rest is more meaningful both for validation with more targeted protein measurement approaches and for potential clinical applications. However, the 7-protein or 11-protein 3-cluster models had better ConC1 sensitivity compared with their 2-cluster model counterparts, as well as distinguishing more clusters with fewer proteins. This indicated that ConC2 and ConC3 had distinct protein profiles that made them inappropriate to group together in predictions. Protein signature panels for sepsis proteomic response should be aimed at a 3-cluster prediction.

Proteins with the highest coefficients from the prediction models were not necessarily those with the highest significance or fold change in differential abundance analysis. For example, among proteins in the 11-protein 3-cluster model, USP15, ALB, and MAP1A were among the top proteins in contrasts between the clusters, while the other predictors did not reach a fold change of 1.5 in any of the three contrasts. Many of the proteins with the highest coefficients in the best-performance ElasticNet_269 model were also not differentially abundant in any contrasts with a less stringent cut-off (FDR<0.05 and |FC|>1.2), such as LUM, MCHR2, C1R, ITIH4, and MAN1A1. Before a protein signature panel can be more confidently presented and advocated to be measured in independent cohorts, it needs to be better understood why there was this discrepancy between top DA proteins, and the proteins with the highest coefficients in the prediction model. Each protein candidate listed here should also be examined in more detail in how well the biological functions are understood, and the accuracy and feasibility of bed-side plasma protein measurement, to determine which proteins

should constitute an optimal patient stratification panel.

The performance of three-cluster prediction models were assessed based on the accuracy. A technical note is that a random 3-cluster classification keeping the same proportions in each cluster should have an expected accuracy of $(21\%)^2 + (24\%)^2 + (55\%)^2 = 40\%$, based on the cluster proportions from the discovery cohort. Thus the three-cluster accuracy of the prediction models should be compared with this value to assess the improvement over a random model, instead of being compared with 50%. For two-cluster predictions of ConC1, a random classification will produce an expected accuracy of $(21\%)^2 + (79\%)^2 = 67\%$, which appears high due to the unbalanced group proportions. Therefore, sensitivity and precision are used to describe the ConC1 model performances instead of using a 2-cluster accuracy.

Models directly predicting mortality

Results in Section 4.3.3 showed that while proteomic clusters consistently stratify patients into groups of different risk, directly modelling mortality based on either proteins or clinical variables produced very low sensitivity, using similar algorithms. A similar scenario was encountered in former research in the group where leukocyte gene expression was used to predict mortality and high inaccuracy was observed in one direction of the prediction (personal communications). These observations are probably due to the fact that in sepsis patients there is not a consistent mechanism leading to mortality, so separating patients directly by mortality cannot stratify patients into more homogenous groups that can be reproducibly identified. On the other hand, stratifying patients by unsupervised patterns in protein levels or gene expression can effectively divide patients into groups of more shared molecular mechanisms and/or physiological conditions, which lead to different risks of mortality, although it is also not possible to predict the outcome with certainty at an individual basis.

As an extra note on model parameter tuning, it could be argued that accuracy may

not be the best optimising metric for unbalanced groups like a mortality model or the ConC1 model. Sensitivity for predicting mortality may be a better metric for evaluating the mortality models, since it is preferable to be less tolerant towards false negatives than false positives, so that more clinical care can be diverted to the higher risk patients identified. However, as there was only a small change in accuracy with the different tuning parameters, I would expect the limited success in directly modelling mortality to be inherent to the aim and less affected by the model tuning process.

One of the largest studies modelling sepsis mortality on clinical data (Zhang et al. 2021) proposed a Sepsis Mortality Risk Score including 13 clinical variables selected by Lasso, measured at 24hrs of ICU entry for patients fulfilling sepsis-3 criteria. They reported an increasing probability of death with an increasing score, and effectively separated patients into 4 groups of increasing risk of 30-day mortality (3% to 68%), in both the discovery (n=5443) and validation cohort (n=5658). Among one of the more successful models directly predicting mortality as a binary classification, Langley et al. (2013) generated prediction models consisting of 3 clinical variables (age, lactic acid, haematocrit) and 5 blood metabolites, and reported better performance in predicting 28-day sepsis survival than widely used clinical scores. The overall best model Langley et al. reported was built using support vector machine learning, based on samples within 24hrs of arrival at the emergency department (ED). The model had a relatively good test set accuracy of 74.6% (compared to 59% by a random classification), while the improvement was most prominent over other models on a NPV (negative predictive value, sepsis non-survivor prediction) of 55%. However, the size of test set was relatively small (n=86) and I noticed that the proportion of 28-day-post-ED mortality in this dataset (49/173=28%) was higher than the 28-day-post-ICU mortality observed in MS2019 (17%), and this less-unbalanced grouping could facilitate a better model performance. Nonetheless, this study indicated the possibility of better mortality prediction in sepsis using a combination of clinical and omics data, as opposed to the approach presented in my analysis of patient stratification based on

the molecular characteristics, which may better relate with biological processes that could potentially be targeted by immunomodulatory therapies.

4.7.7 Validation in VANISH

The ConC proteomic clusters were also predicted in VANISH patients with sensible proportions in each cluster. There was no contradictory evidence found in VANISH clusters compared with GAinS clusters, in terms of similarity between samples, clinical characteristics, or cluster movement. At the same time, the lack of clinical difference observed between the VANISH clusters can be mostly attributed to the small patient numbers, which limited the ability to derive clear indications from this validation cohort. It also needs to be noted that VANISH patients, being all diagnosed with septic shock, were already the more severe subgroup of ICU sepsis patients, thus reducing the level of heterogeneity in comparison with GAinS.

In both GAinS and VANISH, there were more patients moving to the less severe cluster (ConC3) than in the opposite direction, in line with the expectation that patients with later timepoints sampled were enriched for patients who were recovering. Future work could be performed on time-series analysis of individual protein levels across the timepoints, including to see whether protein levels relate with clinical measurements, or differ between treatment groups or ConC clusters.

To understand how the proteome-based patient subgroups could inform more stratified patient treatments, clusters should be assigned in more clinical trial cohorts to analyse the interaction with not only steroids but also other immunomodulatory therapies such as IL-6R antagonists, interferons, GM-CSF inhibitors, intravenous immunoglobulins, and JAK kinase inhibitors. Outcome of the therapies could be monitored not only by mortality or organ support requirements, but also by molecular markers indicating dynamic alternations in certain signalling pathways and cellular programs.

4.7.8 Conclusion

In this chapter I defined three sepsis patient subgroups from unsupervised clustering on the plasma proteome, including one subgroup of patients (ConC1) that were clinically more severe and indicated to be more active in immune pathways and with higher cytokine levels. Significant interaction was detected with patient subgroups based on leukocyte gene expression. The two classifications exhibited both shared and distinct molecular mechanisms, and a better combined risk stratification. The proteomic clusters were identified from the MS2019 GAinS discovery cohort, predicted in the validation cohort, and characterised in MS2019 as well as in two alternative platforms for subsets of patient-timepoints. The proteomic clusters can be effectively predicted with a small number of plasma proteins.

Future work is required to validate the distinctions in biological processes between the clusters in relevant tissues, to elucidate the function of the marker proteins in the context of sepsis, and to develop an optimised protein signature panel. Assigning the patient subgroups in clinical trial cohorts could facilitate the understanding of the interaction between subgroups and immunomodulatory therapies in the aim of developing a more precise and effective treatment in sepsis.

5

The COVID-19 Blood Proteome

This chapter explores the host response proteome of COVID-19 patients at different levels of severity and in comparison with related non-COVID-19 conditions.

5.1 Introduction

5.1.1 The COMBAT proteomics modality

Patients with COVID-19 have a heterogeneous clinical presentation with many patients asymptomatic while others progressing to respiratory failure and hypoxia, to symptoms of severe coagulopathy, and even to multi-organ failure and death from the disease. The COvid-19 Multi-omics Blood ATlas (COMBAT) consortium investigated cell, mediator and pathway signatures of COVID-19 severity from peripheral blood, specific features that differentiate COVID from related conditions (flu and sepsis), and potential biomarkers of individual response (COMBAT consortium, 2022). Within the consortium, the proteomics primary analysis compared the plasma and serum proteome of COVID-19 patients and controls, with aims including to understand the pathophysiology reflected from the blood proteome, and to identify blood biomarkers towards a personalised medicine approach.

Work in this chapter is presented as two parts. The first part analyses the COVID response between different severities and in contrast to non-COVID sepsis patients and healthy volunteers, using a timsTOF-based plasma proteome dataset. The second part focuses on understanding the differences in molecular characteristics between critically ill COVID-19 and flu patients, using a timsTOF- and Luminex- based serum proteome dataset. Between the modalities within the COMBAT consortium (e.g., mass cytometry, total RNA-seq, single cell RNA-seq, B/T cell repertoire), the proteomics modality had the largest sample size especially covering the flu patients, thus a more detailed comparison of the clinical characteristics is presented in this chapter between

the COVID and flu patients.

5.1.2 Differentiating COVID-19 from sepsis or flu

Acknowledging that sepsis is a highly heterogeneous disease, evidence were showing that some COVID-19 patients display processes typically related to sepsis and thus a shared mechanism, including microvascular thrombosis, dysregulated endothelium, lymphocyte exhaustion, and molecular cascades like cytokine hyperactivation (Diao et al. 2020; Connors and Levy 2020; Shappell et al. 2020), as described in more detail in Introduction. It is also interesting to investigate the shared and different mechanisms between COVID-19 and pre-pandemic all-cause sepsis at the plasma proteome level.

SARS-CoV-2 and influenza viruses both cause contagious respiratory illnesses that have very similar clinical manifestations including fever, continuous cough, fatigue, headache and body aches. SARS-CoV-2 is more contagious and has a longer incubation period (2–14+ days) than influenza (1–4 days). Both viruses are enveloped RNA viruses. SARS-CoV-2 enters the host cell through binding of its spike proteins with ACE2 receptor on the host cell surface. Influenza viruses enter host cells through binding of hemagglutinin with sialic acid. COVID-19 patients in COMBAT all had the original wild type strain given the time window of recruitment. For influenza viruses there are four strains, among which human influenza A and B viruses cause the seasonal epidemics. Influenza A viruses are further divided into subtypes based on two proteins on the surface of the virus: hemagglutinin (H) and neuraminidase (N). Among the 18 flu patients analysed in Section 5.4, six had influenza A H1N1, four had influenza A H3N2, seven had unsubtyped influenza A, and one had influenza B.

Because of the similarities and differences between the two viral infections, it is important to understand both clinically and at the molecular level how the host response to COVID-19 compares with that to flu. In Section 5.4 I explored this at the circulating proteome level using serum samples from 37 critically ill COVID-19 patients and 18 critically ill flu patients, both recruited as part of the COMBAT project.

Proteomics data for these serum samples were pre-processed and analysed separately from the larger dataset of plasma samples recruited locally in Oxford that were used to compare between different disease severity of COVID-19 and to compare with sepsis.

5.1.3 Aims

The overall aim of this chapter is to characterise the host response proteome of COVID-19 in the context of related conditions, with the hypothesis that the plasma proteome profiles are distinct between COVID-19 patients of different severity and in comparison with all-cause sepsis or critically ill flu patients. Specifically, I will:

1. identify plasma proteins differentially abundant between COVID-19 patients with different levels of clinical severity and healthy controls
2. identify changes in inflammatory, immune, and metabolism processes reflected from the plasma proteome
3. understand the distinct plasma proteome response to COVID-19 compared with non-COVID-19 sepsis
4. understand the differences between clinical phenotypes including secondary infections in the critically ill COVID-19 and flu patients with serum samples included
5. understand the differences in molecular phenotypes in response to severe COVID-19 and flu reflected from the serum proteome

5.2 Results: The plasma proteome of COVID-19, sepsis and healthy volunteers

5.2.1 Data pre-processing

The timsTOF-based plasma proteome data used in this chapter and deposited in the COMBAT consortium was pre-processed with the following steps: proteins were filtered such that proteins with ⩾50% valid values in any group were kept; samples were filtered such that samples with >50% of missing values were removed; data was normalised by $\log_2$ transformation and median-centering; missing values were imputed with KNN when ⩾60% valid values were present for the protein, otherwise these were randomly drawn from a Gaussian distribution (downshift of the mean=1.8, s.d.=0.3). The resulting data matrix contained 353 samples and 105 proteins. Another version of pre-processing using approaches more closely matched to MS2019 (VSN normalisation, random-draw from protein-specific shifted distributions) showed that the two versions were strongly correlated and indicated similar differentiation of samples.

5.2.2 Cohort composition

From the pre-processed data, I further excluded 13 samples from analysis for malignancy, immunosuppression, or being repeat samples. The sample composition for downstream analysis is shown in Table 5.1. Based on the first-released WHO categorical criteria of severity, this included 25 inpatients (IP) with no requirement for supplemental oxygen (defined by the consortium as COVID_IP_mild), 49 inpatients with oxygen saturation $SaO_2 \leq 93\%$ on air but not requiring mechanical ventilation (COVID_IP_severe) and 23 inpatients requiring mechanical ventilation (COVID_IP_critical). The hospitalised patients were compared with community COVID-19 cases never admitted to hospital (COVID_community), or patients

Table 5.1: Sample composition for COVID-19 plasma proteome dataset. For the hospitalised COVID-19 patients, "Number of individuals" shows the number of patients whose max severity sampled fell in the corresponding category, while "Number of individuals sampled at max severity" further restricts to patients with a sample taken at the most severe clinical category observed.

Group	Number of samples	Number of individuals	Number of individuals sampled at max severity
Healthy	22	22	22
COVID_community	121	100	100
COVID_IP_mild	33	25	15
COVID_IP_severe	78	49	42
COVID_IP_critical	34	23	23
Sepsis_IP	52	38	38
Total	340	257	240

hospitalised (both in ward and in ICU) with all-cause sepsis recruited prior to the pandemic in the Sepsis Immunomics study (Sepsis_IP), together with healthy volunteers ≥ 55yr (Healthy). The community cases were healthcare workers who were younger than the other comparator groups and tended to be in recovery from day 7 after the symptom onset. Work by the consortium showed that clusters of hospitalised COVID-19 patients based on the clinical features showed broad concordance to the WHO categorical criteria (mild/severe/critical) (COMBAT consortium, 2022), therefore the criteria is used as primary severity comparator groups.

5.2.3 The overall proteome and samples structure

I first investigated whether there is a broad relation between the plasma proteome profiles and the clinically defined patient categories. There were generally high correlations between all plasma samples, with a median Pearson's correlation coefficient of 0.874. Samples from the same clinical category formed into more strongly correlated blocks, with more variation shown within the COVID-19 mild and sepsis groups (Fig. D.1). Unsupervised hierarchical clustering of the samples corresponded well with the disease and severity conditions, first separating the healthy and community COVID-19 individuals from the hospitalised COVID-19 and sepsis

patients, and then separating between different clinical states at lower merging heights of the dendrogram (Fig. 5.1). Comparing between the two clusters of community cases that clustered with the healthy volunteers or did not, the former had higher abundances of 56 proteins enriched for complement activation and processed related to coagulation, while age did not differ between the two groups.

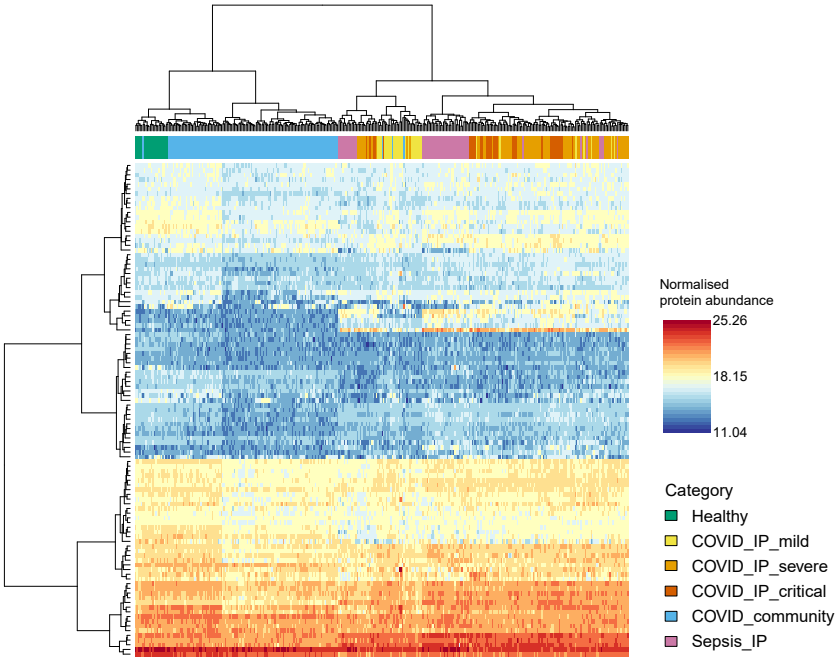
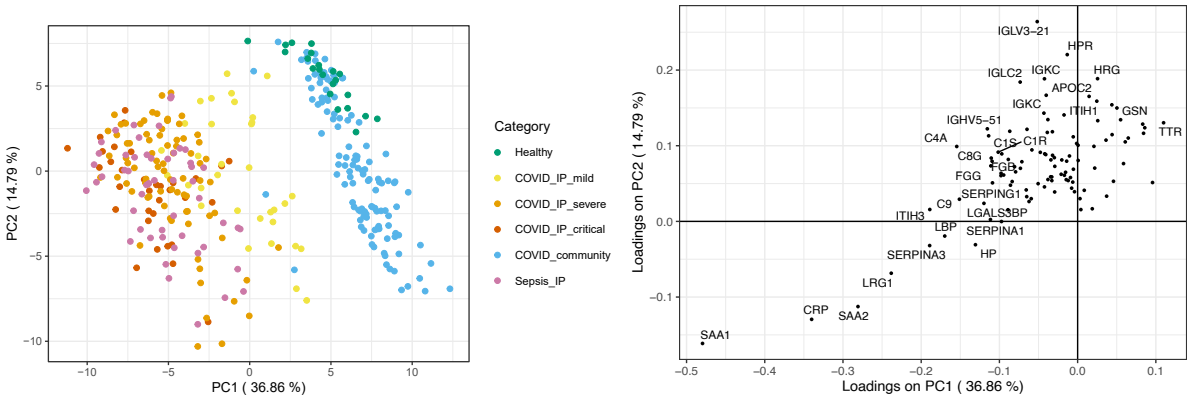


Figure 5.1: Hierarchical clustering and protein abundance heatmap in COMBAT using all samples (columns, n=340) and proteins (rows, n=105).

Samples of different disease and clinical states showed clear separation on the first two principal components (Fig. 5.2a). There was a significant overlap between the healthy volunteers and community cases. Therefore, both PCA and unsupervised clustering suggested that the differentiation between sample groups is associated with a variation in the overall proteome profile. Acute-phase proteins (SAA1, CRP, SAA2) and TTR contributed most to the separation of samples on PC1. Proteins with high loadings on PC1 or PC2 included serum amyloids (SAA1, SAA2, APCS), protease inhibitors (SERPINA3, SERPINA1, SERPING1, ITIH1, ITIH3), immunoglobulins (IGLV3-21, IGLC2, IGKC, IGHV5-21), complements (C9, C4A, C1S), and other proteins that

mediate the anti-pathogen immunity (LBP, HPR, LRG1) (Figure 5.2b).



(a) Sample structure **(b)** Protein contributions to PC1 and PC2.

Figure 5.2: PCA on 105 proteins in 340 COMBAT samples. Proteins with PC1 loadings >0.1 or PC2 loadings >0.15 are labelled.

The 105 proteins analysed clustered according to their relative abundance in HCA, with no major difference in biological functions observed between the four clusters (Fig. 5.1). On the other hand, clustering on the protein-protein interaction (PPI) network identified protein functional clusters including two small clusters enriched for cytolysis and complement activation, both positively correlating with disease severity (which negatively correlated with PC1) and one large cluster that showed positive and negative correlations with severity, enriched for biological processes involving cholesterol transport and fibrin blood clots (Fig. 5.3).

5.3 Results: The differentiated proteomic response

5.3.1 Association of plasma proteome with COVID-19 severity

For group comparisons involving hospitalised COVID-19 patients, I used only samples that were taken at the patient's maximum severity, restricted to the earliest such sample per patient, in order to obtain a clear interpretation of the assayed proteome reflecting the disease severity instead of the process of worsening or recovering of symptoms. Accounting for age and sex, I performed pairwise comparisons

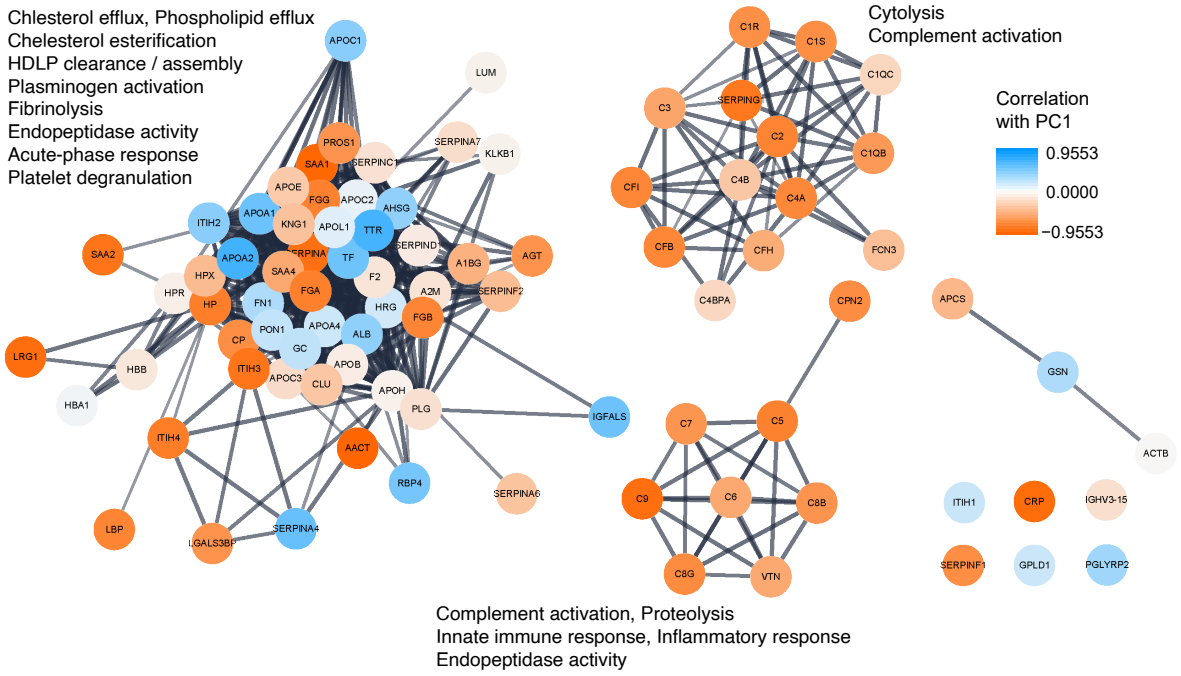
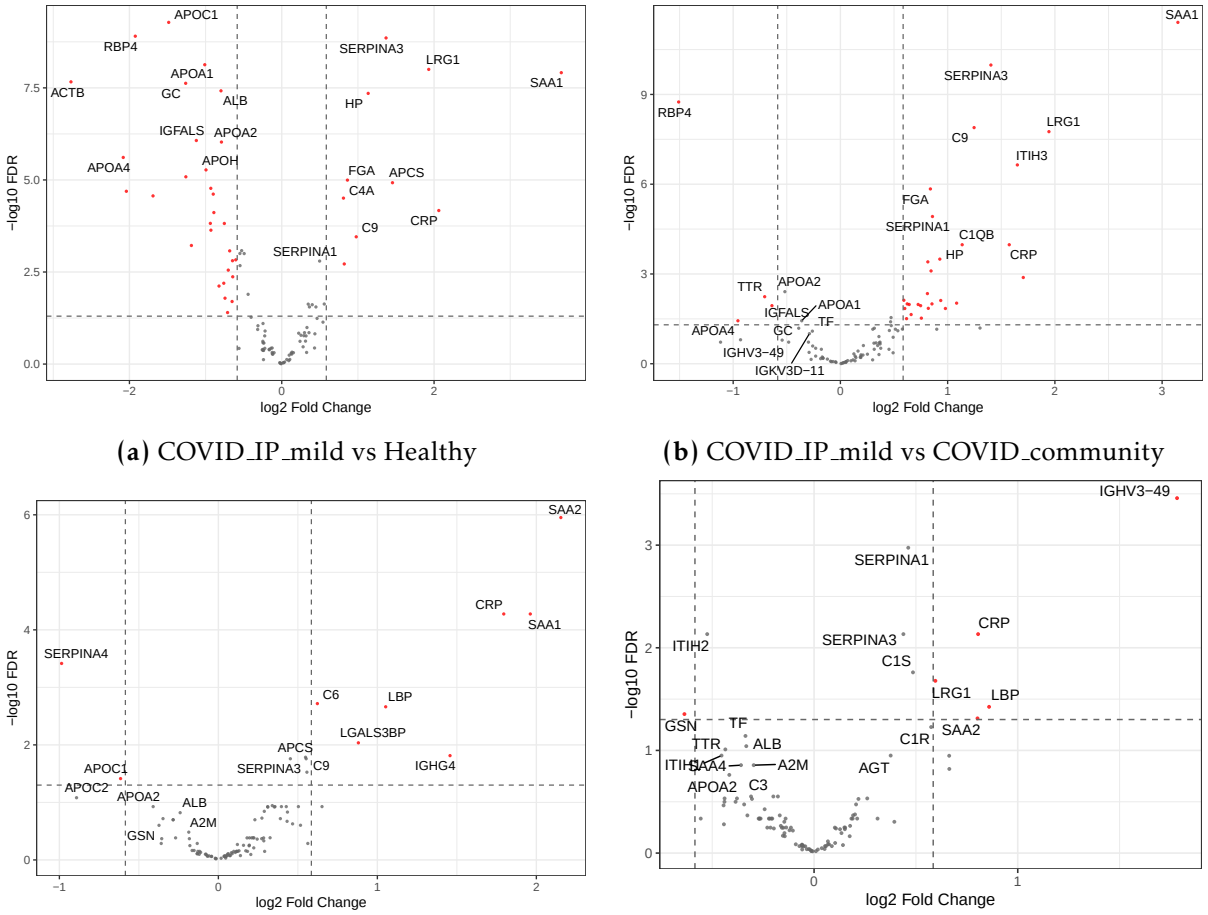


Figure 5.3: Clusters based on the interaction network between measured proteins. Node colour was mapped to Pearson's correlation coefficients between PC1 scores and the protein level across samples, as lower PC1 score was shown to correlate with higher disease severity. GOBP terms enriched in members of the three clusters were labelled.

between COVID_IP_mild vs. either Healthy or COVID_community, and between COVID_IP_severe vs. either COVID_IP_mild or COVID_IP_critical. Larger differences were observed between the hospitalised cases and non-hospitalised individuals, and smaller differences between the COVID-19 inpatients with different severity (Fig. 5.4).

The differentially abundant proteins in all four contrasts were enriched for inflammation and immune processes. The COVID-19 mild hospitalised patients and healthy controls had similar age and comorbidities but showed differences in metabolic processes or vesicle transport of retinoid, cholesterol, lipoproteins, and fat-soluble vitamins(Fig. 5.5), which was evident from proteins APOA1, APOA2, APOA4, APOC1, APOC2, RBP4, TTR, and PON1. Compared with the community cases, the most profound change in the mild hospitalised patients were higher levels of proteins annotated for complement activation, blood coagulation and acute-phase response (Fig. 5.5).



(a) COVID_IP_mild vs Healthy

(b) COVID_IP_mild vs COVID_community

(c) COVID_IP_severe vs COVID_IP_mild

(d) COVID_IP_critical vs COVID_IP_severe

Figure 5.4: Differential protein abundance between COVID-19 patients of different severity. Red dots denote the differentially abundant proteins (FDR<0.05 and |FC|>1.5). Proteins with the highest significance are labelled.

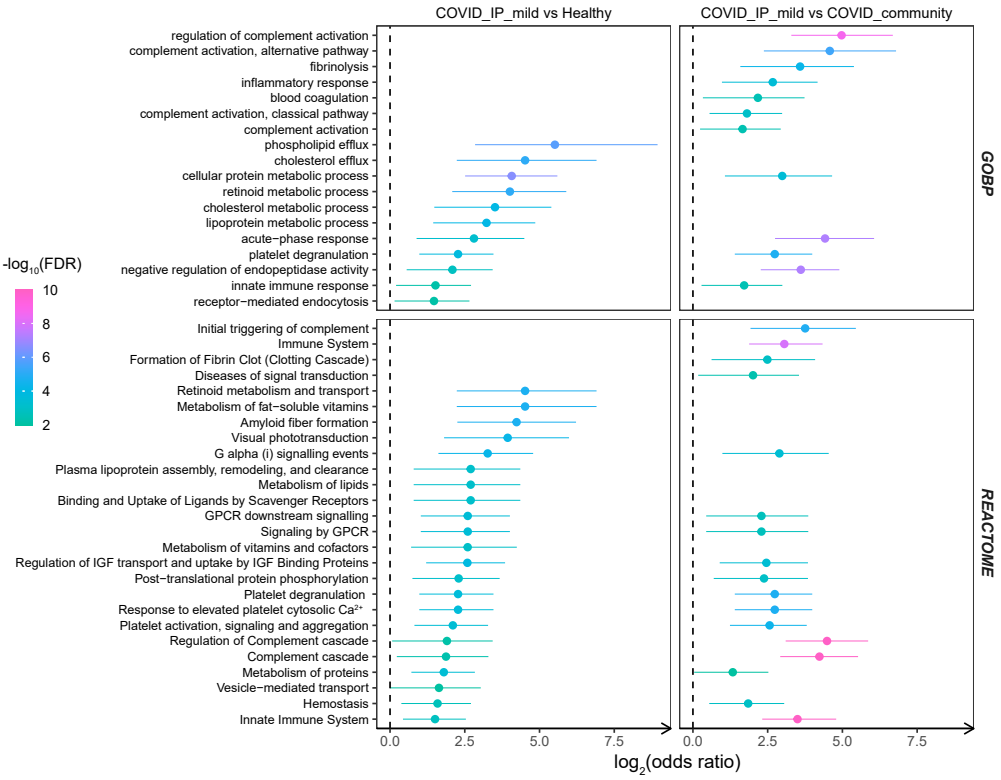


Figure 5.5: Pathway enrichment for mild hospitalised COVID patients. Forest plot shows GOBP terms or Reactome pathways significantly enriched (FDR<0.01) in proteins differentially abundant contrasting samples from mild hospitalised COVID-19 patients with those from healthy controls or from mild community COVID-19 cases. IGF=Insulin-like Growth Factor.

Among hospitalised COVID-19 patients, the severe patients differed from the mild or critically ill patients in processes relating to platelet degranulation (evident from proteins ITIH3, LGALS3BP, SERPINA3, SERPINA4) or neutrophil degranulation (evident from proteins GSN, LRG1, SERPINA1, SERPINA3), respectively (Fig. 5.6). This suggested that the adverse host response in the more severe COVID-19 patients is partly immune-mediated, which is typically associated with the pathophysiology of all-cause sepsis.

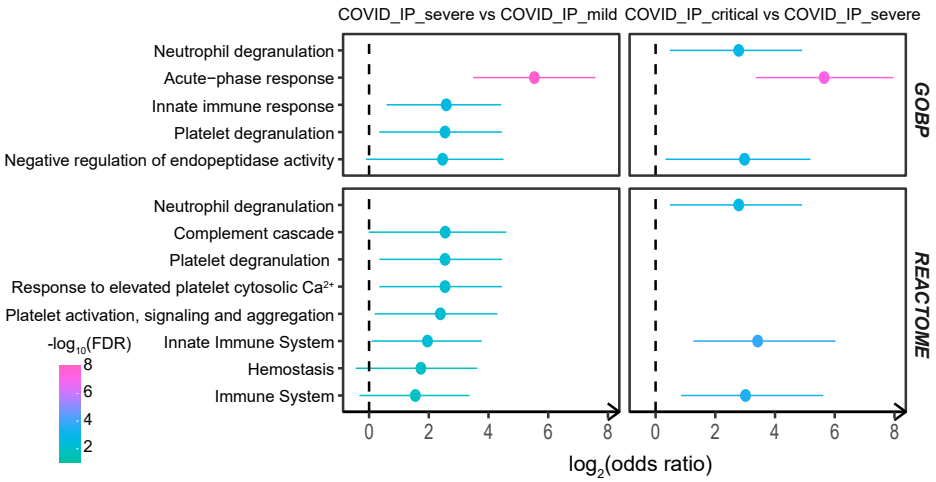


Figure 5.6: Pathway enrichment between COVID-19 severities. Forest plot shows pathways significantly enriched in differentially abundant proteins from contrasting patients at three levels of COVID-19 severity.

5.3.2 Molecular changes reflected in the COVID-19 plasma proteome

A few specific aspects of molecular changes in COVID-19 pathophysiology could be identified from comparing the plasma proteome of COVID-19 patients of different severity and the controls. Distributions across the groups of representative proteins supporting these indications are shown in Figure 5.7.

Acute-phase proteins and complements

The most prominent difference I observed was significantly increased levels in the more severe patients of acute-phase proteins and complement system proteins, including recognised biomarkers of inflammation (SAA1, SAA2, CRP), components of the complement membrane attack complex (C5, C6, C9, CFB), and functionally related protein families like protease inhibitors (SERPINA3, SERPINA1, ITIH3) and serum amyloid P-component (APCS).

Tissue damage and elevated coagulation

Upon acute complement activation, more cytokines and chemokines are produced and released, recruiting more macrophages to the site of infection which in excessive inflammation conditions could lead to lung injury. I observed a significantly lower level of plasma gelsolin (GSN) in mild COVID-19 inpatients vs healthy controls, and in critical vs severe COVID-19 inpatients. The depletion of gelsolin as part of the extracellular actin scavenger system suggested an elevated activity of removing toxic F-actin filaments released from necrotic cells. I also observed elevated levels of fibrinogen alpha, beta, and gamma chain (FGA, FGB, FGG), suggesting elevated coagulation following local tissue damage although no difference was found for prothrombin (F2) or kininogen 1 (KNG1).

IL-6-mediated inflammation

Interleukin-6 mediated pro-inflammatory processes interact closely with the processes described above, including APP production, complement activation, platelets degranulation, and virus binding to alveolar macrophages which leads to cytokine release as a start of the inflammation circle. In this cohort I observed elevated levels of galectin-3-binding protein (LGALS3BP), a pro-inflammatory factor which induces IL-6 expression, and other proteins implicated in IL-6-mediated inflammation like leucine-rich alpha-2-glycoprotein (LRG1), LPS-binding-protein (LBP), haptoglobin

Figure 5.7: Distribution of representative protein abundance across COMBAT comparator groups. Only one sample at max severity per individual is plotted.

(HP), and inter-alpha-trypsin inhibitor heavy chain 4 (ITIH4). Among these, HP is an acute-phase response protein; ITIH4 plays an important role in extracellular matrix organisation and is implicated in inflammation; LRG1 promotes cell proliferation and angiogenesis and is implicated in skin and lung fibrosis (UniProt); LBP forms into a complex with CD14 as part of the anti-bacterial defence response and sensitises TLR-mediated LPS recognition. Interestingly, aside from an expected higher abundance in sepsis patients where bacterial infections were more likely, LBP was also significantly increased in COVID-19 patients with higher severity. This suggests that either there was also a higher incidence of bacterial infections in more severe COVID patients in this cohort, or sensitization to LPS may also play a role in this specific viral infection and may contribute to excessive lung inflammation.

Lipoprotein metabolism

Lastly, I observed decreased levels of a number of apolipoproteins and hormone-binding proteins including apolipoproteins A-I, A-II, C-I, C-II (APOA1, APOA2, APOC1, APOC2) and transthyretin (TTR), suggesting a major change in lipoprotein and cholesterol metabolism. On the other hand, APOA1 is a major component of the high-density lipoprotein (HDL) complex, and has well-documented anti-inflammatory properties through interacting with macrophages (Iqbal et al. 2016). The lower abundance of apolipoproteins observed could either be a result of their downregulation in systemic inflammation, or reflect the intrinsic individual difference in metabolic conditions which may associate with different risks towards more severe symptoms.

5.3.3 Comparison with all-cause sepsis

Changes observed in the plasma proteome of COVID-19 patients can be attributed to both a general deterioration in organ functions due to a maladaptive response to severe infections, and also pathophysiological changes specific to SARS-CoV-2 infection.

Comparing either group with healthy volunteers, the biological processes reflected by differentially abundant proteins were largely shared between all-cause sepsis and COVID-19 (Fig. 5.9). To tease out a COVID-19 specific proteomic response, COVID-19 inpatients (severe and critical) were compared with sepsis patients due to other infections at similar clinical severity. 19 out of 105 proteins showed changes specific to COVID-19 (FDR<0.05 and |FC|>1.5, Fig. 5.8), enriched in acute-phase response, complement activation, and receptor-mediated endocytosis. Among these proteins, serum amyloid P-component (APCS) was significantly more abundant in COVID-19 patients, suggesting a role for clearing nuclear material released from damaged circulating cells, specific to SARS-CoV-2 infection. A few immunoglobulins (IGHG4, IGKV1-39, IGHG2) were at higher levels in COVID-19 but not in all-cause sepsis patients, suggesting their involvement in specific antigen recognition and binding process. Fibronectins (FN1) bind cell surfaces and various compounds including collagen, fibrin and actin, and are involved in cell adhesion and motility (UniProt). The lower abundance of FN1 only in sepsis but not in the COVID or Healthy groups (Fig. 5.7) may relate to processes not expected in the viral infections or healthy volunteers like bacterial adhesion or anti-bacterial defence. LGALS3BP has been demonstrated by affinity pulldown to be binding to SARS-CoV-2 spike glycoproteins and thus may reduce the virus infectivity (Gutmann et al. 2021). Overexpression of LGALS3BP reduced spike-mediated syncytia formation, and decreased spikepseudoparticle entry *in vitro* . Both the work by Gutmann et al. and data described here (Fig. 5.8, Fig. 5.7) observed a rise in circulating LGALS3BP in COVID-19 but not in other hospitalised sepsis patients, highlighting the specificity for this viral infection.

5.3.4 Samples during deterioration

Samples taken at a patient's maximum severity best reflect the plasma proteome at a specific clinical status (as analysed in the sections above), while samples taken during deterioration or recovery can help to address two questions: whether samples

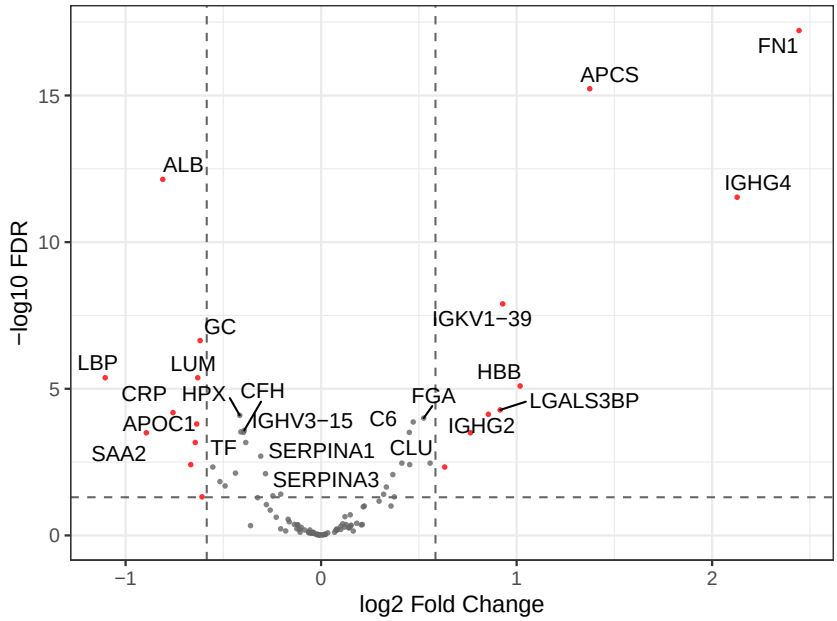


Figure 5.8: Differential protein abundance between COVID_IP_severe and COVID_IP_critical (combined) vs Sepsis. A positive $\log_2$FC refers to higher levels in severe and critical COVID-19 in-patients. Red dots denote significantly different proteins (FDR<0.05 and |FC|>1.5). Ten proteins with highest significance in either direction of fold change are labelled.

represent more the severity at sampling, or the maximum severity a patient will reach; and, whether certain proteins can be used to predict which patients would progress to a more severe status.

Considering sample availability and corresponding analytical power, the transition between severe and critically ill COVID-19 patients was taken as an example. In this cohort, there were 33 "transitional" samples taken at a severe clinical status from patients who became critically ill or died in later days. These were compared with 44 samples taken at a severe clinical status as the max severity of the patient (Fig. 5.10(a)), or with 34 samples taken at a critical clinical status (Fig. 5.10(b)). For these analyses serial samples were included in linear mixed models and between-patient variation was accounted for as a random factor. There were no proteins significantly differentially abundant (FDR<0.05) in the former contrast, while seven including IGHV3-49 and C1R were significantly more abundant in critical samples (FDR<0.05) in the latter contrast. This suggested that the blood proteome "in transition" more

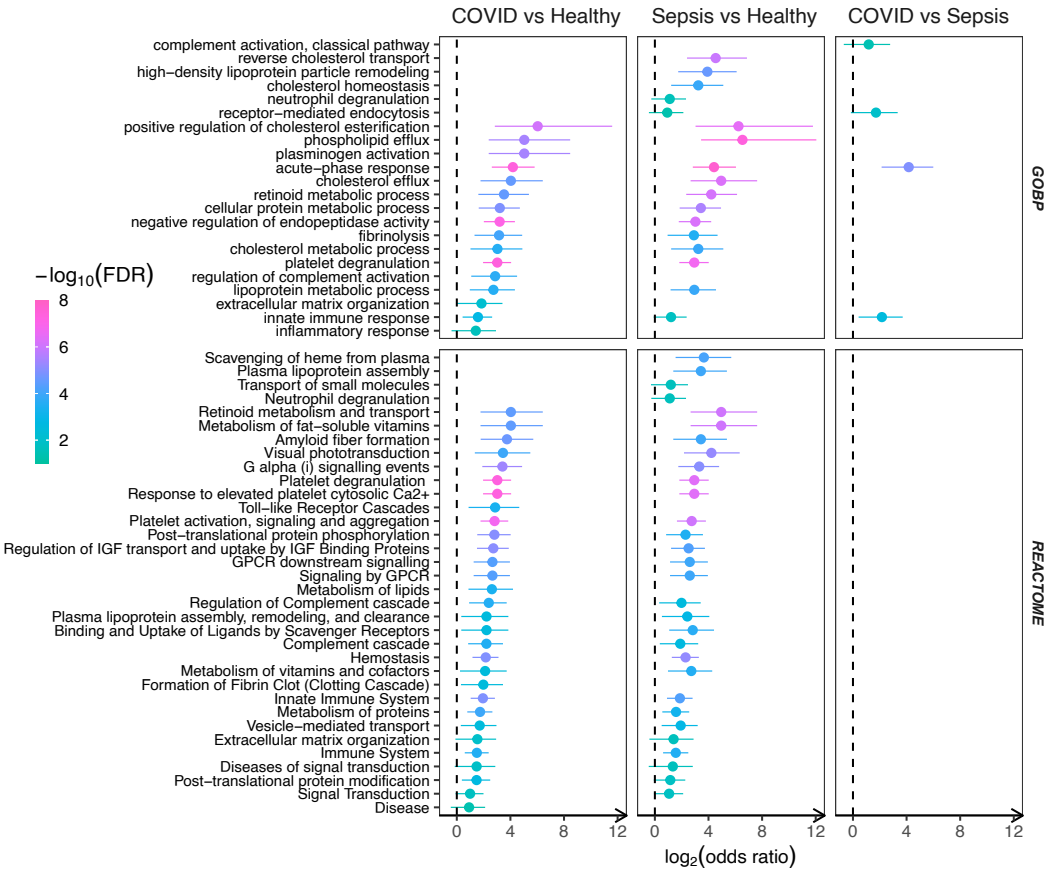
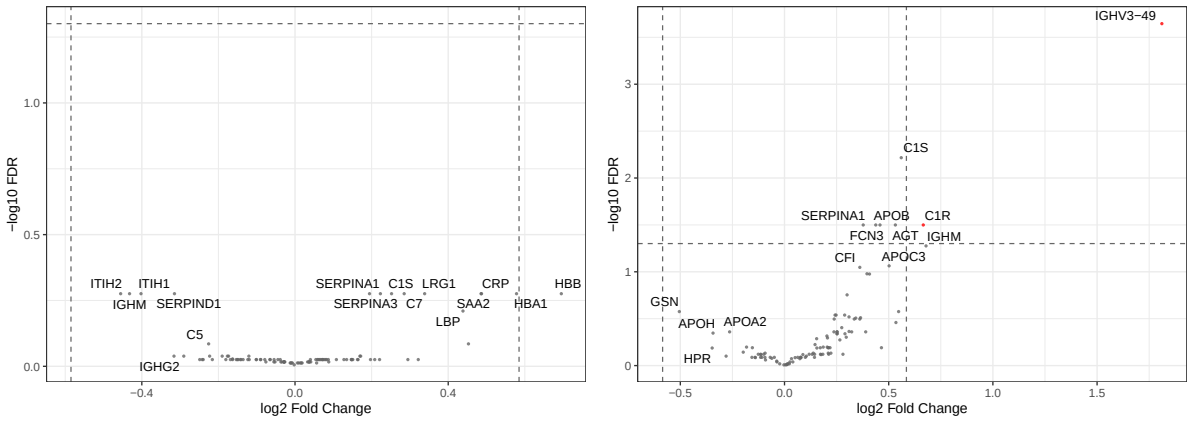


Figure 5.9: Pathway enrichment for COVID vs Sepsis. Forest plot shows GOBP terms or Reactome pathways significantly enriched (FDR<0.05) in proteins differentially abundant contrasting samples from severe and critical hospitalised COVID-19 patients combined, or sepsis patients, or healthy volunteers. IGF=Insulin-like Growth Factor.

closely resembles the clinical status at the time of sampling rather than a future status, and that the 105 proteins measured here would only provide limited if any power in distinguishing patients that would deteriorate clinically. This also aligned with the observation that there was a relatively strong correlation (Pearson's r=0.69, $p=3.1 \times 10^{-16}$) in $\log_2$FC between contrasting the critical vs transitional samples, and contrasting critical vs severe samples taken at max severity of the patients.



(a) Severe to critical transitional samples (n=33) vs severe at max severity samples (n=44).

(b) Critical state samples (n=34) vs severe to critical transitional samples (n=33)

Figure 5.10: Differential protein abundance in transitional samples. Contrasts were made between samples taken at the max severity of the COVID-19 patients and samples taken in states when patients were deteriorating from severe to critical states. Dashed lines indicate the thresholds for differentially abundant proteins (FDR<0.05 and |FC|>1.5). Proteins with the highest significance are labelled.

5.4 Results: Difference in response to COVID-19 versus flu

Serum protein abundance was measured in parallel on the timsTOF and Luminex platform for 37 ICU COVID-19 patients and 18 ICU flu patients (post-QC numbers that overlapped between two platforms) recruited independently of the COVID and control patients analysed above by plasma samples. TimsTOF data from the serum and plasma samples formed distinct clusters in preliminary analysis so the two sample types were separated in pre-processing and downstream analysis. Further filters were applied on the serum Luminex data before combining it with the TimsTOF-measured

proteins for analysis.

5.4.1 Serum data pre-processing

Protein quantifications by Luminex assay was obtained from the data deposition CBD-PRT-00001 within the consortium. This dataset included the raw fluorescence intensities (FI) and mapped concentrations of 51 cytokines using the Luminex xMap assay. Experimental details of the assay were described in the consortium report (COMBAT consortium, 2022). The mapped concentrations instead of raw FI were chosen to represent the quantification with further filtering. For each analyte I used the concentrations of the lowest and highest standards as the lower and upper limits of detection and censored the outlying sample concentrations at the corresponding limits. Considering that cytokines could be up-regulated in only subsets of patients with specific disease conditions, I kept any analytes that had $\geq$ 10% detectable measurements in at least one of the sample groups and excluded four analytes. The full list of analytes and proportion of sample values detectable are listed in Table D.1.

TimsTOF data of the serum samples was separated out from the FragPipe output including plasma samples. This included one sample per patient from 42 patients admitted to ICU due to COVID-19 and 22 patients admitted to ICU due to flu, plus 19 pools from these samples. I used the same pre-processing strategies as described in section 5.2.1. The resulting data matrix contained 87 proteins measured in 55 samples (37 COVID-19, 18 flu). There were 18 proteins detected in plasma COMBAT samples but not in the serum samples, including ACTB, FGB, FGG, IGHG4 among others. The only protein that was detected in serum samples by both TimsTOF and Luminex was complement C5.

After being pre-processed separately, the two data types were either relative quantifications centered around zero (TimsTOF) or absolute concentrations (Luminex). After a $\log_2$ transformation of the Luminex data to make the distribution approximately normal, the histogram of all the protein values showed a clear

difference between the two datasets as expected (Fig. D.2). To make the samples comparable after combining the two datasets, I used quantile normalisation by samples which makes the distribution identical across all 55 samples. Because of the difference between the two assay platforms, the relative abundance and variation (including fold changes) should be compared within but not between TimsTOF- and Luminex- measured proteins. Proteins need to be scaled to constant variance in analysing the between-sample structure in PCA or hierarchical clustering described in the following section.

5.4.2 Comparison of clinical phenotypes

To better understand the clinical conditions represented by the samples, I compared the patient demographics and clinical phenotypes between the COVID-19 and flu patients (Table 5.2). Demographics and pre-existing conditions were similar between the two groups, except that non-Caucasians made up a higher proportion of the COVID-19 group. The flu patients were more severe on ICU admission as manifested by higher SOFA and APACHE scores, and tended to have lower lymphocyte count on admission. In terms of treatments received, all patients were endotracheal-intubated, and all but one were on vasopressors. Half of the flu and none of the COVID-19 patients were on extracorporeal membrane oxygenation (ECMO). Despite the fact that flu patients were more ill on admission and more of them received ECMO, mortality in hospital was significantly lower ($p=0.014$). It is also worth noticing that samples from the flu patients were taken closer to the symptom onset and also closer to ICU admission. I then inspected how the two groups differ at the molecular level reflected from the blood proteome.

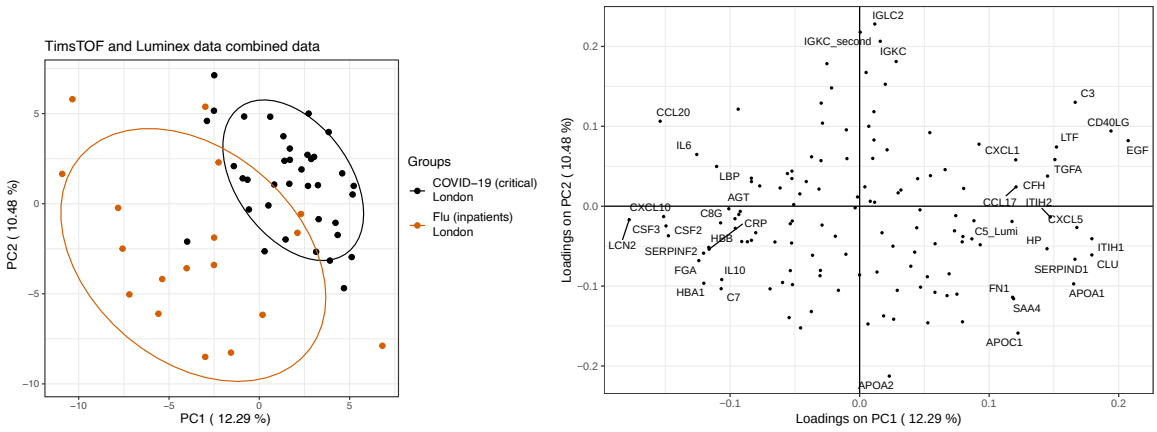
Table 5.2: Comparison of clinical phenotypes between COVID-19 and flu patients.
All patients had endotracheal intubation. Day 1 refers to the day of study ICU admission. Continuous variables were compared with Mann-Whitney test and categorical variables with χ^2 test or Fisher's exact test (if No.<10). P values <0.05 are considered significant and labelled in bold. No adjustment has been made for multiple testing.

	COVID-19 (ICU)	**Flu (ICU)**	**p-value**
Number of patients	37	18	-
Age median (IQR), years	57 (47-65)	56 (48-62)	0.73
Men, No./total (%)	24/37 (65)	14/18 (78)	0.37
Pre-existing conditions, No./total (%)			
- Hypertension	13/37 (35)	3/18 (17)	0.21
- Current or ex- smoker	12/37 (32)	8/18 (44)	0.38
- Renal disease	3/37 (9)	0/18 (0)	0.54
- Liver cirrhosis	0/37 (0)	1/18 (6)	0.33
- Diabetes	9/37 (24)	2/18 (11)	0.31
- Haematological cancer	0/37 (0)	0/18 (0)	-
- Other cancer	0/37 (0)	0/18 (0)	-
Ethnicity, No./total (%)			
- Caucasian	15/37 (41)	13/18 (72)	**0.044**
- Asian	10/37 (27)	4/18 (22)	1
- African	8/37 (22)	1/18 (6)	0.24
- Other	4/37 (11)	0/18 (0)	0.29
Body mass index, median (IQR)	27.8 (25.2-30.0)	25.8(23.0-28.3)	0.090
APACHE score on day 1, median (IQR)	12 (9-15)	20 (17-25)	$\mathbf{1.30 \times 10^{-7}}$
SOFA score on day 1, median (IQR)	6 (4-7)	10 (9-13)	$\mathbf{4.00 \times 10^{-7}}$
WHO ordinal score at sampling	All = 7	NA	-
Lowest lymphocyte count on day 1, $\times 10^3/\mu L$	0.7 (0.5-0.9)	0.45 (0.3-0.8)	0.088
Lowest total neutrophil count on day 1, $\times 10^3/\mu L$	7.3 (5.4-9.3)	6.8 (3.4-10.5)	0.71
Neutrophil-lymphocyte ratio	9.1 (5.5-16.0)	11.6 (6.9-21.3)	0.30
Highest total neutrophil count in ICU, $\times 10^3/\mu L$	17.0 (12.1-21.8)	21.4 (14.0-31.8)	0.17
Ever had vasopressors, No./total (%)	36/37 (97)	18/18 (1)	1
Duration of ventilation, median(IQR), days	17 (10-27)	18 (10-30)	0.88
ECMO, No./total (%)	0/37 (0)	9/18 (50)	$\mathbf{7.60 \times 10^{-6}}$
Ever had renal replacement therapy, No./total (%)	13/37 (35)	11/18 (61)	0.087
Length of ICU stay, median(IQR), days	19 (11-30)	24 (15-33)	0.45
Deceased in hospital, No./total (%)	18/37 (49)	2/17 (12)	**0.014**
Days from symptom to sampling, median(IQR), days	14 (12-21)	9.5 (7-11)	**0.00086**
Days from ICU admission to sampling, median(IQR), days	6 (4-10)	2.5 (1-4)	**0.00069**

5.4.3 The COVID-flu distinction at the proteome level

Sample structure

First I inspected the overall structure of the 55 samples in the combined dataset. The COVID and flu samples separated well on the first and second principal components (Fig. 5.11a) and in unsupervised hierarchical clustering (Fig. 5.12), suggesting that critically ill flu and COVID patients may have distinct blood proteomic responses despite the similarity in clinical symptoms. Hierarchical clustering on the combined dataset separated the patient groups better than clustering on only TimsTOF data in the same set of samples, as the branch containing mostly flu patients separated from the rest in the first branching event, instead of being separated at a lower merging height when only TimsTOF data was used (Fig. 5.12, Figure D.3).



(a) Flu and COVID-19 samples on PCA

(b) PCA loadings

Figure 5.11: Serum sample distribution on PCA, using combined data. Data was centred and scaled by proteins. Data ellipses are drawn at a 95% confidence level assuming a multivariate normal distribution

Proteins with large loadings on PC1 (>0.1) or on PC2 (>0.2) included proteins from both datatypes (Fig. 5.11b[1]), indicating that the sample structure was driven by both low-abundance signalling molecules and high-abundance classical plasma proteins.

[1] C5 measured by TimsTOF is labelled "C5" while C5 measured by Luminex is labelled "C5_Lumi". "IGKC_second": this protein is not mappable to a specific gene because the Uniprot ID returned from the protein search does not differentiate between the variable or constant region of the immunoglobulin kappa light chain. Thus I used the gene name for the constant region (IGKC) to represent it and used "second" to differentiate it from the IGKC protein returned from the protein search.

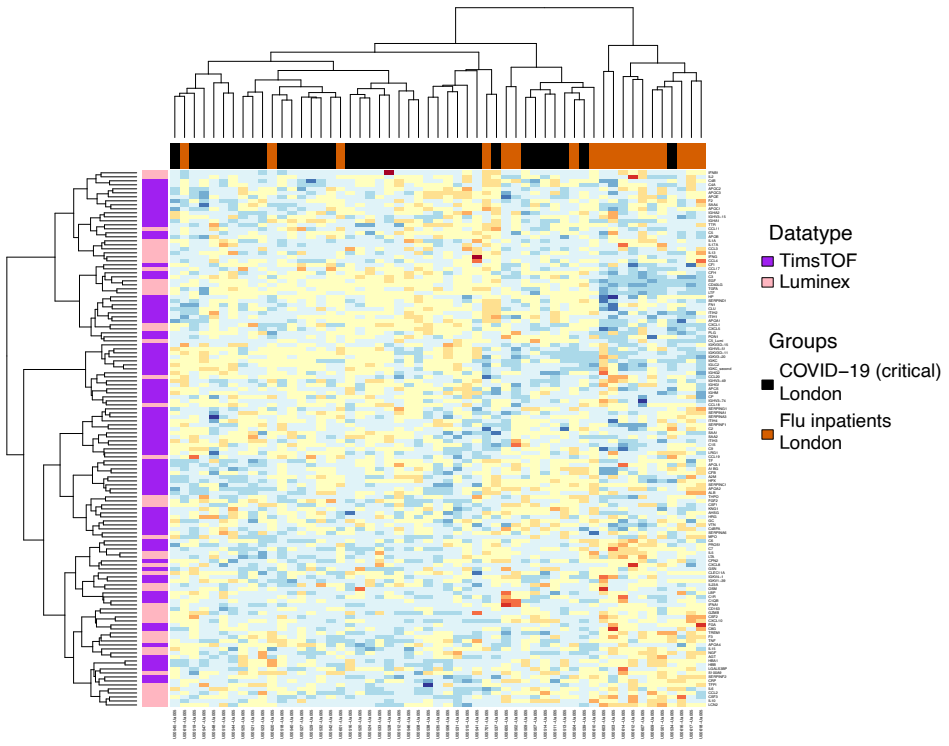


Figure 5.12: Heatmap and hierarchical clustering in COVID and flu serum samples using combined TimsTOF and Luminex data. Data was centered and scaled by proteins.

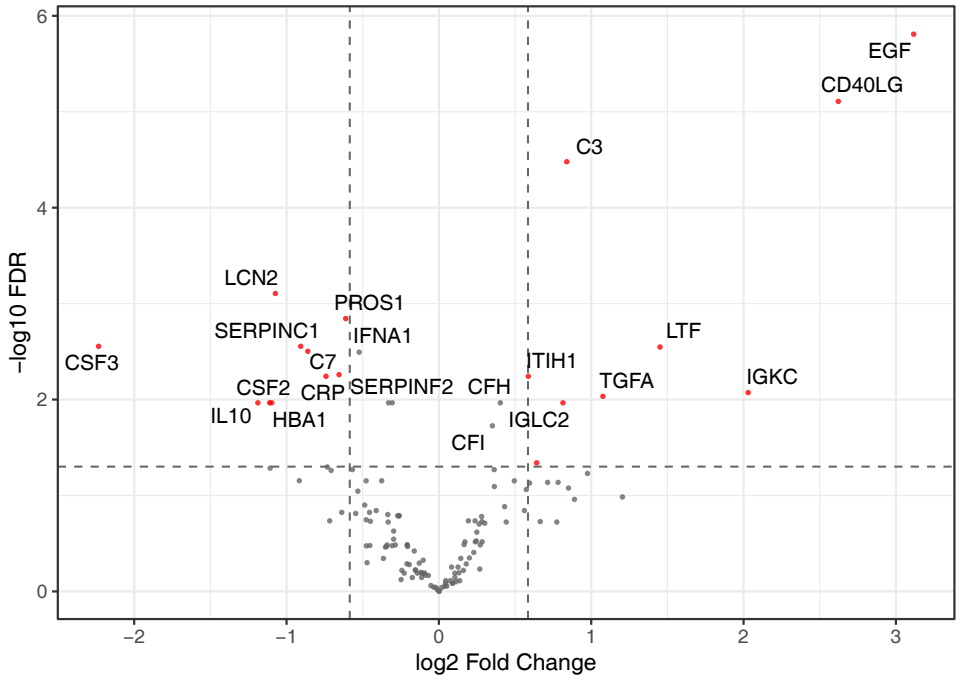
Notably, proteins with large loadings included apolipoproteins A1, A2, C1; acute-phase proteins: CRP, IL-6, SAA4; immunoglobulins: IGKC, IGLC2; complement components: C3, C5, C7, C8G; a large number of cytokines with both positive and negative loadings on PC1; other proteins with a role in innate immunity like HP, LBP, LCN2; proteins related to coagulation: FGA, SERPIND1, SERPINF2; inter-alpha-trypsin inhibitor heavy chain ITIH1 and ITIH2, which may act as a carrier of hyaluronan; and more general proteins like haemoglobin (HBA1, HBB) and fibronectin (FN1).

The COVID-flu contrast

To identify differential abundance at the individual protein level, I contrasted COVID-19 and flu samples using a linear model including age and sex as covariates. 24 proteins were significantly different between the two groups (FDR <0.05), with 19 reaching fold change >1.5 (Fig. 5.13). Proteins more abundant in COVID-19 included:

Epidermal growth factor (EGF) which stimulates growth of epidermal and epithelial cells; CD40LG, a cytokine that costimulates B-cell proliferation, promotes cytokine production by macrophages and dendritic cells, and is involved in immunoglobulin class switching; complement component 3 (C3); immunoglobulin kappa constant (IGKC); and lactoferrin (LTF) which has antimicrobial activity and is part of the innate defence mainly in mucus.

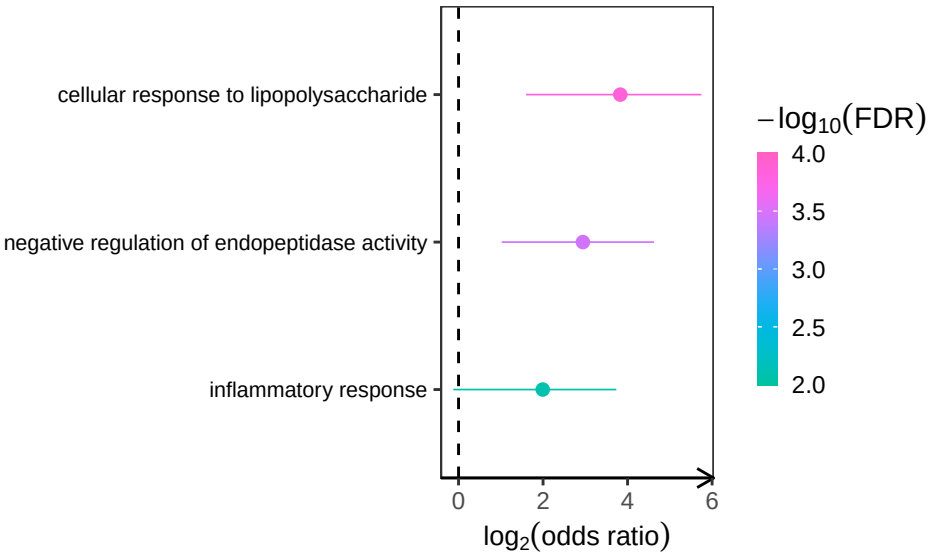
Figure 5.13: Comparison of protein abundance in COVID-19 and flu samples using combined data. A positive $\log_2 FC$ indicates higher abundance in COVID-19. Proteins with FDR <0.05 and fold change >1.5 are plotted in red. The most significant proteins in both direction of fold change are labelled.



Among the proteins more abundant in flu, lipocalin-2 (LCN2, alternative name neutrophil gelatinase-associated lipocalin) is involved in the innate immune response to bacterial infection and is a marker of acute kidney injury. LCN2 was among the proteins more abundant in sepsis compared with all control groups as described in Chapter 3. Granulocyte-macrophage colony-stimulating factor (GM-CSF, CSF2) has been tested as a vaccine adjuvant and local supplementation against Influenza A, and its overexpression in mice after influenza A virus infection was reported to prevent mortality (Halstead et al. 2018). The 19 differentially abundant proteins

were enriched in GOBP terms including cellular response to LPS which suggests a differential response to gram-negative bacteria; negative regulation of endopeptidase activity which was attributed to the protease inhibitors in the SERPIN and ITIH families; and inflammatory response (Fig. 5.14).

Figure 5.14: Biological processes enriched in proteins differentially abundant (FDR<0.05 and |FC|>1.5) between COVID-19 and flu. GOBP annotation terms with minimal size of 5 and minimal overlap with data of 3 were considered in the test.

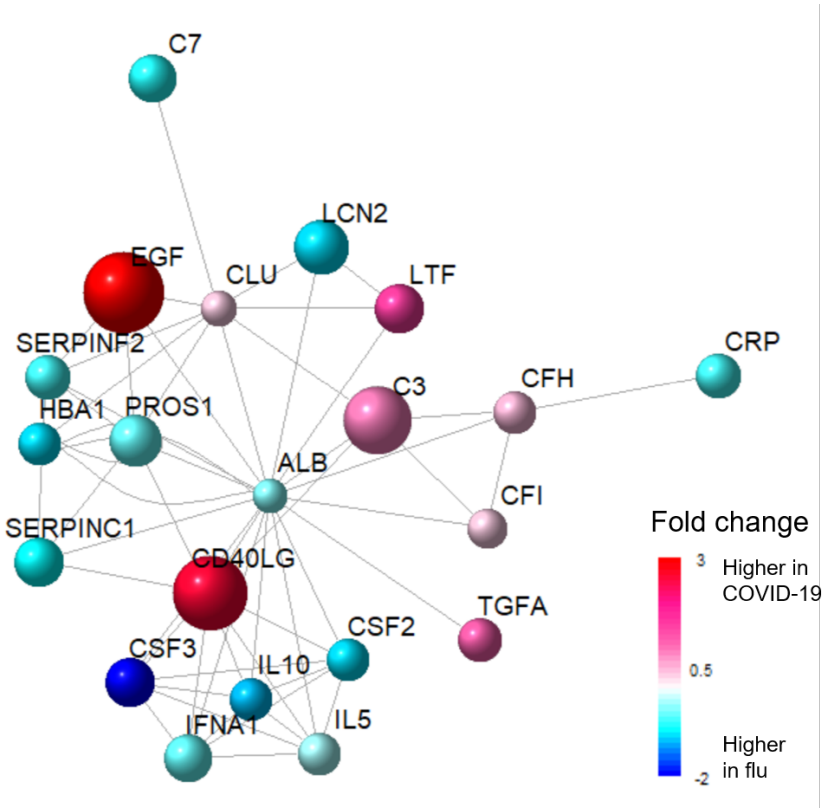


To confirm that the differential abundance analysis was robust to the transformation applied in combining the datatypes, I plotted the correlation between fold changes obtained for the TimsTOF proteins when performing the COVID-flu contrast in the combined data or only in TimsTOF data (Fig. D.4). There was a strong correlation (Pearson's r=0.971) although a smaller set of proteins reached significance in the combined data partly due to the larger number of texts corrected for multiple testing . Our collaborators at SGUL also independently measured cytokines from this cohort of patient samples and verified our findings that CSF3 (G-CSF), CSF2 (GM-CSF), and IL-10 were higher in flu, together with TNFα which tended to be higher in flu in our assay.

Network analysis

To understand the differential proteome network, I identified the most significant subnetwork of the COVID-flu contrast using XGR. The algorithm applied in the *xSubneterGenes* function is usually suitable for scoring and identifying a most significant subnetwork from a pool of a large number of genes passing a significance threshold. In this case there were only 24 significant proteins in total so the subnet size identified was restrained by the FDR=0.05 cutoff. The most significant protein-protein interaction network identified comprised 19 of the 24 proteins with FDR<0.05, plus two linker proteins CLU and ALB (Fig. 5.15).

Figure 5.15: The most significant PPI network identified from the COVID-flu contrast. Nodes are coloured by fold changes in the contrast, with higher abundance in COVID-19 represented in red. Node size is mapped to $-\log_{10} FDR$.



In the PPI network there was a closely-connected group of cytokines (CSF2, CSF3, IL-10, CD40LG), which were all more abundant in flu except for CD40LG, indicating a higher inflammation and immune response in these flu patients, although humoral

immunity appeared to be more profound in COVID as reflected by the upregulation of CD40LG and immunoglobulins. A group of proteins involved in coagulation regulation (PROS1, SERPINF2, SERPINC1) were also more abundant in the flu patients, suggesting a more active coagulation process in flu which could be related to damage in the blood vessel. Other proteins in the network point to aspects of the innate immune response (LTF, LCN2, CRP, C3, C7, CFI, CFH) and angiogenesis (EGF, CLU, TGFA). Insterestingly, both C3 and inhibitor of activated C3 (CFH, CFI) were higher or tended to be higher in COVID.

5.4.4 Interpretation of the differential proteins

I then aimed at understanding the differences in protein abundance in the context of the two diseases. By searching PubMed and Google Scholar for the keywords "COVID-19" (or "SARS-CoV-2"), "flu" (or "influenza") and "comparison" in February 2021, I did not find an informative review that compared the molecular pathology between the two viral infections. Thus I summarised some of the known pathological and molecular changes in both infections that had either been reported in others' studies or reflected from our analysis of the COVID-19 blood proteome. Most of the changes are shared in both severe SARS-CoV-2 and influenza infections, including: a diffuse alveolar damage (Cevik et al. 2020; Gill et al. 2010); local and systemic excess pro-inflammatory cytokine production and release; acute-phase response and complement activation including membrane attack complexes released by liver; recruitment of immune cells including T-cells, monocytes and neutrophils; tissue injury and cell necrosis especially involving epithelial cell death and intra-alveolar haemorrhage (Cevik et al. 2020; Kalil and Thomas 2019); exposure of endothelial cells to antigen and cytokines and thus dysregulated endothelium which further amplifies inflammation. Obese individuals are more susceptible to severe flu which may be attributed to defects in adaptive immunity (Paich et al. 2013), while our plasma proteome data suggested altered cholesterol metabolism in COVID. Three aspects where reported works or our

protein data suggested a difference between COVID and flu are discussed below.

Secondary infections

It has been reported in both infections that severe cases are predisposed to secondary bacterial infections which are associated with greater treatment requirement and higher mortality, with a similar proportion of positive bloodstream culture in severe flu (one-fourth, MacIntyre et al. 2018) and hospitalised COVID-19 patients (31%, Bhatt et al. 2020). This increased susceptibility has been suggested to be a result of depletion of alveolar macrophages, the suppression of lung immune cells to allow for tissue repair, and immune dysregulation in both viral infections (Kalil and Thomas 2019). In our cohort, individual-level co-infection data measured by the clinical team (Jonathan Youngs and others, St George's, University of London (SGUL)) were requested and the incidence rate compared between the flu and COVID-19 patients included (Table 5.3). For candida or viral coinfections, there were very few cases and no difference between the two groups. Evidence from blood culture or deep respiratory speciman (bronchoalveolar lavage (BAL) or non-directed BAL) showed more bacterial coinfection in the flu patients ($p=0.0011$). Although evidence of detection from sputum and deemed clinically significant (i.e. treated) indicated the opposite direction of effect, this could reflect upper respiratory tract colonisation and contamination and thus was less convincing for representing active coinfections than evidence from blood culture or deep respiratory specimen.

At the molecular level, among the proteins differentially abundant in the COVID-flu contrast, four proteins (LCN2, CSF2, CSF3, IL-10) are annotated with "cellular response to LPS" and they all had higher abundance in the flu samples. In accordance with this observation, flu patients also tended to have more gram-negative coinfections than COVID patients, among the other bacterial infections ($p=0.072$, Table 5.3). However, as differences in clinical management may substantially confound the comparison, a clear conclusion cannot be reached in whether it is the virology that

leads to different rates of coinfections.

Table 5.3: Secondary infections in the COVID-19 and flu patients.
BAL = bronchoalveolar lavage, NBL = non-directed bronchoalveolar lavage. *Total* reports indication of any types of infection from any of the detection methods. P values are from χ^2 tests. Bacterial species detected included *Streptococcus pneumoniae*, *Staphylococcus aureus*, *Streptococcus pyogenes*, *Haemophilus influenzae*, other gram-negative rods, and *Enterococcus*.

	COVID-19 (ICU)	Flu (ICU)	Test p-value
Number of patients	37	18	-
Secondary infections, No./total (%)			
- Bacterial, detected from blood culture/BAL/NBL	9/37 (24)	13/18 (72)	**0.0011**
- Gram-negative bacteria, detected from blood culture/BAL/NBL	9/37 (24)	9/18 (50)	0.072
- Bacterial, detected from sputum and treated	24/37 (65)	5/18 (28)	**0.020**
- Candida, detected from blook culture	2/37 (5)	0/18 (0)	1
- Viral	1/37 (3)	0/18 (0)	1
Total	26/37 (70)	14/18 (78)	0.75

Angiogenesis

Differences in pathophysiology between the two infections have been reported, based on an autopsies comparing of 7 lungs from patients who died from COVID-19 against 7 who died from ARDS secondary to influenza A (H1N1) infection (Ackermann et al. 2020). The authors reported that the amount of new vessel growth from COVID lungs was 2.7 times as high as that from patients with influenza, predominantly through intussusceptive angiogenesis. They also reported widespread thrombosis with microangiopathy from histologic analysis of pulmonary vessels in COVID-19 lungs, and showed that alveolar capillary microthrombi were 9 times as prevalent in COVID-19 as in influenza. Following on from this report, I then investigated if there was any evidence for differential angiogenesis or coagulation between the two infections at the blood proteome level.

EGF, TGF-α, and ITIH1 were among the proteins more abundant in COVID vs flu. Both EGF (epidermal growth factor) and TGF-α (transforming growth factor alpha, a mitogenic polypeptide) are EGFR ligands, activation of which has a critical role in tumor angiogenesis, and up-regulates VEGF expression in human cancer cell lines. Gefitinib was reported to exert anti-angiogenic effects by blocking EGF-induced upregulation of VEGF and IL-8 in human cancer cell lines (Hung et al.

2016). ITIH1 (Inter-alpha-trypsin inhibitor heavy chain H1) may act as a carrier of hyaluronan, which is a main ECM component with a key role in tissue regeneration, inflammation response, and angiogenesis (UniProt). Besides, clusterin (CLU, also called apolipoprotein J) tended to be more abundant in COVID (FDR=0.054, FC=1.29). Antisense oligonucleotides to clusterin were shown to inhibit angiogenesis and induce apoptosis of capillary cells *in vitro* (Jackson et al. 2005). Therefore, the observed differential protein abundance aligned with more active angiogenesis reported in COVID compared with flu.

Coagulation

Proteins PROS1, SERPINF2, SERPINC1 were closely connected in the PPI network (Fig. 5.15), all having higher abundance in flu, and are all implicated in regulation of coagulation. PROS1 (Vitamin K- dependent protein S) is a cofactor to activated protein C in the degradation of coagulation factors Va and VIIIa. It helps to prevent coagulation and stimulate fibrinolysis. The major targets of SERPINF2 (alpha-2-antiplasmin) are plasmin (an important enzyme that that degrades many blood plasma proteins, including fibrin clots) and trypsin. SERPINC1 (antithrombin-III) prevents thrombus formation by inhibiting thrombin, matriptase-3, factors IXa, Xa and XIa (UniProt).

In addition, both FGA (FDR=0.050, FC=1.66) and TFPI (FDR=0.055, FC=1.63) tended to be more abundant in flu. TFPI inhibits the tissue factor pathway, which is the primary cellular initiator of blood coagulation through thrombin formation. Our collaborator (Jonathan Youngs, SGUL) also observed higher concentrations of TFPI and F3 in their independent measurements. Therefore, these proteins indicated that both the formation and degradation of fibrin blood clots were more actively regulated in the flu patients.

However, as serum samples are obtained after blood clotting is activated, blood proteins that are either actively involved in clot formation (e.g., fibrinogens,

coagulation factors F2 and F3) or non-specifically adsorbed to clotting proteins will have their concentrations greatly altered inconsistently across the samples. Levels for these proteins should not be interpreted in data analysis. I will give a discussion in Section 5.5.5 on whether broader proteins involved in the regulation of blood clotting could still be interpreted.

5.5 Discussion

5.5.1 Proteome detection

Although both used non-depleted samples in a similar proteome discovery pipeline and strategy for protein filtering, a smaller number of proteins were detected in the COMBAT plasma proteome dataset than in MS2019, both in raw (702 and 2782 proteins) and processed (105 and 269 proteins) data. This could be partly attributed to the fact that there was a much greater extent of sample diversity in MS2019, which contributed to more proteins being identified with the match-between-runs feature enabled although this was computationally intense and took weeks of time and >150TB of disk space to complete in MS2019 by Roman Fischer et al.

In both datasets, a larger number of proteins were available in the raw data. Although most of these proteins were only sparsely detected in the actual samples (rather than in library fractions), other proteins could be informative in comparing protein detection instead of quantification across the samples. For example in COMBAT, some proteins involved in tissue damage (LDHA, LDHB, ACTC1, ACTN2) or in the coagulation cascade (FGL1, coagulation factors 5/7/9/10/11/13, VWF, and SERPINE1) were not included in the processed data but it is possible to compare these for differential detection if assuming the missingness is at least partly dues to low-abundance (i.e. not completely at random).

5.5.2 Candidate blood biomarkers for COVID-19 severity from COMBAT

One aim of the full proteome analysis not addressed in this chapter was to identify blood protein biomarkers for sub-phenotyping within hospitalised COVID-19 patients. When directly comparing the survivors (n=18) vs non-survivors (n=54) within the severe and critical COVID-19 patients, I observed no signal close to significance in the 105 plasma proteins. Spectral clustering on the similarity matrix also could only discriminate a few cases as shown by Alberto Santos Delgado. However, when he applied Similarity Network Fusion on networks from both mass spectrometry and Luminex datasets, two clusters associated with different disease severity and 28-day mortality were identified. The main discriminatory features between the clusters included 11 proteins: IL-6, IL-8, CCL2, CCL19, CCL20, CXCL10, S100A9, SAA1, SERPINA3, GM-CSF, and CLEC11A. He further validated the model in an independent COVID proteomics dataset measured by Olink. Therefore, a predictive set of plasma proteins could be used to stratify acute COVID-19 patients, which would be informative for the response state and could potentially help to resolve patient heterogeneity in clinical trials for targeted immune therapies.

5.5.3 Other studies of the COVID-19 blood proteome

Molecular changes that associate with COVID severity

With a globally concentrated scientific effort against COVID-19 in the year 2020, independent researchers were on the same frontline in understanding the COVID-19 blood proteome. Shen et al. (July 2020) measured 894 proteins and 941 metabolites in serum from 65 COVID-19 and 53 control individuals, identified 93 proteins that show differential expression in severe COVID-19, and trained a model with random forest including 22 proteins and 7 metabolites to distinguish severe–non-severe patients. They highlighted dysregulation of macrophage, platelet degranulation, complement

system pathways, and massive metabolic suppression as key altered pathways in COVID-19 patients, providing complementary molecular insights for the pathogenesis of SARS-CoV-2 infection. Messner et al. (July 2020) applied an ultra-high-throughput clinical proteomics platform on serum samples from a cohort of 48 hospitalised COVID-19 cases and separate healthy controls, and identified 27 potential biormarkers that were differentially abundant depending on the WHO severity grades, including A1BG, CD14, LGALS3BP, LRG1, HP, APOC1, and GSN. Neither study compared COVID-19 against sepsis or influenza.

Most of the molecular changes described in Section 5.3.2 corresponded well with these two reports on COVID-19 serum or plasma proteome available at the time. Shen et al. also reported that acute-phase proteins were among the most significantly upregulated proteins in the serum of the severe COVID-19 patients. Although the non-COVID-19 control groups included had different conditions, their data also showed an increased level of APCS (alternative name SAP) in COVID-19 patients but not in patients with similar symptoms that tested negative for COVID-19, consistent with our result in non-COVID-19 sepsis patients. Consistent with our data, Messner et al. reported depletion of GSN and increased levels of FGA/FGB/FGG suggesting elevated levels of tissue damage and coagulation in more severe COVID conditions. Shen et al. reported downregulation of proteins involved in platelet degranulation, including pro-platelet basic protein (PPBP, alternative name CXCL7) and platelet factor 4 (PF4), which could reflect a depletion of platelets in more severe disease thus the lower abundance of these platelet-derived chemokines observed.

However, there were a few proteins where our data cannot replicate findings in the two reports. We did not observe an increase of SAA4 in severe COVID-19 patients, as described by Shen et al. Actins (ACTB) had lower abundance in COVID-19 patients, instead of higher abundance as reported by Messner et al.

There are several reasons why our measurement was different from other reports. The cohorts we included were different in levels of COVID-19 severity, patient ethnicity,

pathology condition of control groups, and potentially SARS-CoV-2 strains. Different mass spectrometry platforms were used which could lead to different isoforms being detected and differing quantification accuracy along the dynamic range. For example, Messner et al. did not distinguish between cytoplasmic beta-actin and gamma-actin (ACTB; ACTG1), while in our data ACTB but not ACTG1 was detected and quantified.

Time-resolved molecular phenotyping of disease progression

Other than capturing severity-associated molecular signatures, it is also critical to understand how the molecular phenotype develops over time, and whether sampling at early time points has prognostic value for predicting outcome and treatment requirements. To this aim, Demichev et al. (August 2021) measured 86 diagnostic parameters and 309 plasma protein groups in 687 sampling points from a cohort of 139 hospitalised COVID-19 patients.

As a general pattern, they reported that proteins and clinical parameters correlated with higher COVID-19 severity were downregulated over time and vice versa, which means at the molecular level there was a reversion to baseline although patients were still being classified at the same WHO grade. These included downregulation over time of components of the coagulation cascade, clinical markers of inflammation, and upregulation of ECM proteins and proteins involved in lipid metabolism. The authors interpret these changes as an initial spike in the systemic inflammatory response, which was gradually alleviated, followed by a protein signature indicative of tissue repair, metabolic reconstitution and immunomodulation. The authors also identified panels of proteins and diagnostic parameters for predicting remaining time in hospital and future clinical worsening for mildly ill patients, and for predicting survival chances in critically ill patients with good performance in an independent validation cohort of 24 patients.

This study provided valuable and timely insight into COVID-19 disease progression and the potential prognostic value of plasma proteins. However, there were only an

average of 5 timepoints per patient and the trajectory analysis was further restricted to samples during the peak severity i.e. the max WHO grade of the patient. Sampling of more frequent timepoints at a wider disease progression scale would enable finer resolution of the molecular trajectory, including the order of regulation of the aforementioned pathways, how change in proteins relate with clinical deterioration or recovery, or the administration of immune-modulating medicine like dexamethasone which had not become part of standard care at the time of the study.

5.5.4 The acute-phase and liver-derived proteins in COVID-19 and sepsis response

The higher abundance of acute-phase proteins in circulation is one of the most prominent characteristics comparing sepsis to control cohorts, as described in Section 3.6.2. These included SERPINA1, HP, LCN2, and CRP being significantly different in all six contrasts, plus other proteins including LBP, SAA1 and SAA2 in some of the contrasts. The lower abundance of apolipoproteins were also observed in a number of contrasts. Most acute-phase proteins and many of the apolipoproteins are actively synthesized in the liver, which is the tissue of origin for more than half of the blood secretome. Although a different range of proteins were measured in the COVID-19 dataset described in this chapter, these liver-derived proteins also differentiated COVID-19 patients of different severity, reflected in large fold changes in proteins including SAA1, SAA2, CRP, APCS, and apolipoproteins A-I, A-II, C-I and C-II.

Consistent with these results, longitudinal analysis of serum proteins in ICU COVID patients (Gutmann et al. 2021) showed a correlated cluster of liver-derived acute-phase proteins (including CRP, SERPINA1, SERPINA3, SAA1, SAA2, and ORM1) to have significantly increased levels at week 1 compared with ICU admission or at week 2, with no difference in trajectories observed in patients that died or survived. On the other hand, a cluster comprising many other liver-derived proteins (apolipoproteins linked to lipid metabolism, biotinidase, complement and coagulation factors) showed

a consistent increase over the 2 weeks and had significantly higher abundance in survivors compared with non-survivors, linking the recovery of these liver-derived proteins to survival.

In addition, there is evidence that acute-phase proteins produced by tissues other than the liver are also associated with a differential response in sepsis. The long pentraxin PTX3 is an acute-phase protein and member of the pentraxin family (together with CRP and APCS) that has an important role in mediating the innate immune response by acting as a PRR and contributing to opsonization. The oligomerisation of PTX3 under oxidant exposure is required for this biological activity (Inforzato et al. 2008). Secreted PTX3 can bind to pathogens and induce classical complement activation, with the absence of oligomerisation associated with less complement activation and less cardiac damage (Cuello et al. 2014).

Unlike many other acute phase proteins, PTX3 in circulation does not have a main liver origin. The expression is induced by LPS in mouse heart and lung but not in liver (Lee et al. 1994; Introna et al. 1996). In plasma of healthy volunteers injected with low-dose LPS, Burnap et al. (2021) observed a rapid degranulation response at 6hr post-injection containing PTX3 and other proteins of known neutrophil, platelet and endothelial origins, prior to a later acute phase response of primarily liver-derived plasma proteins at 24hr (including SAA1 SAA2, LBP and CRP), suggesting a temporal contribution of the organ and cell types to the endotoxemia plasma proteome. To illustrate the origin of PTX3 multimer, their work in mouse plasma and aortic tissue suggested that PTX3 may be stored in multimeric form in neutrophil granules and that PTX3 multimer deposition in the vessel wall resulted from leukocyte infiltration, particularly neutrophils, instead of *de novo* synthesis in the vasculature.

The redox-dependent multimerization state of PTX3 in sepsis patient plasma consistently associated with a differential outcome, with the transformation of octameric to monomeric PTX3 as early as 2 days post-ICU admission associated with greater survival (Cuello et al. 2014).

The sepsis and COVID-19 datasets (MS2019 and COMBAT) described in this thesis emphasized a broader patient coverage over an optimal protein detection depth and PTX3 did not pass the protein filtering thresholds in either datasets thus was not included in processed data. In MS2019, PTX3 was only measured in 2.9% samples but the detection rate was higher in sepsis samples than in controls.

5.5.5 Limitations in the COVID-19 vs flu analysis

Interpreting coagulation regulation in serum samples

While both plasma and serum samples are commonly used in investigating the blood proteome, the difference in blood clotting make data from the two sample types not comparable. Considerable differences in the proteome composition between these sample types have been reported in both mass spectrometry- and antibody- based quantitative proteomics studies (Schwenk et al. 2010; Ignjatovic et al. 2019;).

Although using serum samples, the COVID-flu contrast revealed signals of more active coagulation regulation in flu. While levels for proteins actively depleted in clot formation (fibrinogen and the clotting factors) should not be interpreted based on serum samples, it is also questionable whether proteins more broadly involved as regulators of the coagulation pathway can be reliably interpreted from serum.

Therefore, I considered multiple aspects of evidence: Firstly, Schwenk et al. (2010) showed that both PROS1 and FGB (among other proteins) were less abundant in serum compared with plasma samples. Secondly, in our data with both plasma and serum samples but in different patients, the 18 proteins only detected in the plasma dataset included FGB and FGG but no other proteins functionally related to coagulation. Thirdly, using either plasma or serum samples from critically ill patients, the histogram of quantifications for the 5 proteins involved in coagulation regulation (PROS1, SERPINF2, SERPINC1, PLG, and TFPI) was down-shifted compared with the other proteins in serum but not in plasma samples (Fig. D.5). Lastly, of the 18 flu

patients included, half (n=9) were on ECMO so should be expected to be on anti-coagulants and thus have different coagulation protein profiles. However, no signal was observed in contrasting in the serum samples despite a more lenient correction for multiple testing applied considering the small sample size. Altogether these observations suggest that it may not be reliable to interpret thrombus dynamics from analysing serum samples, even when the protein of interest is not actively depleted in clotting.

In addition, a few other proteins that are not directly related to thrombus formation (C3, C7, LTF) were also reported by Schwenkd et al. to have different levels in serum and plasma. Thus, caution should be taken in interpreting these proteins as well in analysing serum samples.

Disease progression

Another limitation in using this cohort to understand the different molecular phenotypes of the two viral infections is that the contrast was confounded by a different stage of the diseases being sampled. The flu samples were taken closer to the symptom onset ($p<0.001$) and closer to ICU admission ($p<0.001$), so may more closely reflect an acute phase response. Inflammation reflected by CRP and a few cytokines was indeed more active in the flu patients at the time of sampling.

Notably, the flu patients were more ill on ICU admission (higher SOFA and APACHE scores, $p<1 \times 10^{-6}$ for both) but had better outcome in hospital (fewer deceased, $p=0.014$). It would be useful to understand the disease progression around admission and sampling, if serial sampling and clinical phenotyping like sequential SOFA score or other measures of oxygenation were available on days following ICU admission.

5.6 Conclusion

Mass spectrometry-based plasma proteomics revealed features of COVID-19 severity, including acute-phase response, complement cascade, tissue necrosis, fibrin clots, inflammation pathways involving IL-6, and shifted lipoprotein metabolism. COVID-19 was also associated with differential protein abundance compared with all-cause sepsis patients at similar severity, including fibronectins, serum amyloid P and immunoglobulins. In comparing critically ill COVID-19 and flu patients, a higher abundance of cytokines indicated a higher inflammation response in these flu patients except for humoral immunity. More active vascular angiogenesis was indicated for the COVID patients, while stronger response to LPS and more gram-negative bacterial coinfections were observed in this specific set of flu patients. Future investigations could further advance the understanding of the pathophysiology and individual response to COVID-19 by expanding to relevant tissues in addition to peripheral blood, by improving the proteome measurement depth, and by a more detailed temporal profiling and delineation of the course of immune response and disease progression .

6

General Discussion

This chapter outlines the broader conclusions and future areas suggested by the work described in this thesis.

This thesis has characterised the blood proteome and patient heterogeneity in response to severe infections, specifically in sepsis and in severe SARS-CoV-2 infection. For this aim, one immunoassay-based and three mass spectrometry-based proteomics datasets have been generated for a large cohort of sepsis or COVID-19 patients with detailed clinical phenotyping. This has allowed me to explore the individual variation in the blood proteomic response using supervised and unsupervised approaches, to understand the interaction with transcriptomic endotypes, and to validate the results in independent samples and cohorts. This chapter discusses the broad findings of this thesis in the context of the current state of research, outlines the limitations of this work, and suggests potential areas of future work.

6.1 Clinical proteomics in sepsis

Taking advantage of a combination of a rich source of clinical patient samples and cutting-edge high-throughput mass spectrometry, we were able to generate a large-scale blood proteomics dataset including 2622 samples from 8 cohorts. This has provided the statistical power to establish the shared and specific proteomic features of sepsis and related conditions at a scale not achieved in existing studies (Chapter 3) as well as to identify subphenotypes within sepsis (Chapter 4). In addition, from a technical point of view, my analysis of this dataset has presented a customised pre-processing workflow for clinical proteomics datasets of this scale, and contributed to understanding the impact of sample handling decisions, including the importance of collecting platelet-poor plasma and aligning the choice of anti-coagulants between comparator cohorts, and the potential large alteration to the proteome caused by affinity depletion.

A variety of molecular changes were inferred from analysing the patient blood proteome, including globally elevated cytokine levels in a subset of sepsis patients and various biological pathways that associate with sepsis severity. Because of the nature of blood proteome composition and the detection limit of mass spectrometry, these pathways most often indicate aspects of innate immunity (acute-phase response, inflammatory response, complement activation), tissue damage (extracellular matrix organisation, coagulation and fibrinolysis), the secretion processes (neutrophil degranulation, platelet activation, immunoglobulins), and metabolism carried out in the circulation (lipid metabolism and transport). Another barrier to drawing mechanistic conclusions based on the blood proteome is that the origins of the proteins detected cannot be reliably determined, as discussed in Section 4.7.3. Therefore, to obtain a more comprehensive view of the patient response, it is important to integrate transcriptomics and other data types such as metabolomics and deep clinical phenotyping (e.g. radiomics), and to investigate the expression and proteomic profile

of specific solid tissue or blood cell subsets, which could potentially pinpoint the disease-associated alterations and suggest targets for modulating the immune response or cell metabolism.

To this aim, Hohn and colleagues (2018) re-analysed previous studies on rat septic models and showed that the blood and organ proteomes highlighted alterations in different pathways. Rodrigues and colleagues (2021) measured the monocyte proteome in 9 septic shock patients and revealed molecular features in energy metabolism and inflammatory pathways. A larger cohort would enable investigation of heterogeneous responses across patients. More recently, studies from the COMBAT consortium (2022), the PA-COVID-19 study group (Georg et al. 2022) and in sepsis (Kwok et al. 2022; Reyes et al. 2020) used combinations of mass cytometry, single-cell RNA-seq, VDJ-sequencing and mechanistic studies to provide valuable insights into specific T cell or myeloid cell subsets that associate with COVID-19 or sepsis severity. Intracellular protein activities would provide further information on the metabolic and immunological state of these cell populations. Furthermore, in future studies it may be essential to differentiate between protein isoforms since post-transcriptional and post-translational modifications (e.g. phosphorylation and methylation) could markedly alter the protein function and weaken the relation to the transcriptome.

6.2 Implications for patient stratification

As the final product of gene expression regulation, proteins can not only reflect the biological processes in the patient response but also serve as biomarkers with predictive and prognostic values. In Chapter 4, I identified three plasma proteomic subphenotypes (ConC1/2/3) with distinct molecular and clinical characteristics, and evaluated prediction models including an 8-protein 3-cluster model with 79.5% test-set accuracy. However, the protein signature panel for ConC assignment needs to be further optimised before being applied to additional validation cohorts.

Future work: Optimisation of proteomic cluster prediction

First, the model accuracy and sensitivity may be further improved by conducting a more thorough feature selection. In addition to the generalised linear models and random forest applied in Chapter 4, an optimal minimal number of protein predictors could be determined for example through the forward selection algorithm developed by Herberg et al. (2016), or supervised machine learning using e.g. the SIMON software (Tomic et al. 2019). Second, the feasibility and accuracy of measuring candidate proteins with lower-plex methods needs to be considered. In contrast to untargeted proteome discovery using mass spectrometry, a minimal panel of proteins can be more rapidly and cost-effectively measured in targeted approaches, for example by customisable panels using the ProcartaPlex (ThermoFisher, bead-based immunoassay), Olink (proximity extension), or SOMAscan (DNA aptamer) platforms. The antibody/apatamer specificity and compatibility of the candidate protein targets, together with the cost for the number of analytes would need to be evaluated when choosing the optimal protein panel and the platform. Lastly, proteins with plausible biological functions would be more easily interpreted as biomarkers. For example, although USP15 (Ubiquitin carboxylterminal hydrolase 15, higher abundance in ConC1) has known functions in regulating inflammation, further functional characterisation may elucidate the mechanism of its presence in plasma and the role it plays in the differential response to sepsis.

In addition to the categorical cluster assignments, individual differences at the blood proteome level could also be evaluated as a continuous trait, as has been developed for transcriptome-level subphenotypes (the quantitative SRS scores, Cano-Gamez et al. 2022). Comparable quantitative scores for the proteomic clusters would allow understanding of the interaction with SRS and with disease progression at better granularity. Besides, as significant differences between SRS were observed at the protein level (Fig. 4.25), a prediction model for SRS could potentially be built based on blood proteins which are more convenient to measure clinically than gene expression.

The added value from the large number of serial samples included in MS2019 also has not been fully exploited in this thesis. In addition to analysing the cluster movements, time-serial analysis at the individual protein level and the protein co-expression module level would provide valuable insights into biological processes in the disease course and association with clinical recovery or deterioration.

Inflammatory subphenotypes in sepsis and other critical illness

Mass spectrometry-based blood proteomics often characterises the high-abundance proteins but not the low-abundance ones such as cytokines which play essential roles in mediating the inflammatory and immune response. In critical illness with infective and non-infective aetiologies, patient classification based on circulating cytokines, with or without clinical or gene expression data, has yielded patient subphenotypes and traits suggestive of specific processes underlying the immune dysregulation (Reddy et al. 2020). For example, Calfee and colleagues (Calfee et al. 2014; 2018) identified a hyperinflammatory and a hypoinflammatory subphenotype in ICU patients with ARDS, using latent class analysis on baseline clinical and cytokine data. Compared with the other group, the hyperinflammatory group had fewer ventilator-free days and showed a differential treatment response of improved survival with simvastatin (an immunomodulator used to control hypercholesterolemia) compared with placebo.

In sepsis, using cytokines measured in GAinS (65 cytokines in 124 CAP Day1 patients, described in this thesis) and in VANISH (21 cytokines in 155 baseline patients), I and colleagues (David Antcliffe and others) identified a high-cytokine and a low-cytokine patient cluster (collaborative manuscript in preparation). Findings in sepsis from our studies as well as from others (Fjell et al. 2013; Seymour et al. 2019) showed similarity to reports in ARDS that a hyper-inflammatory subphenotype is associated with higher levels of cytokines, more severe organ dysfunction, and higher mortality, compared with the hypo-inflammatory subphenotype, although no significant association with

differential treatment response was observed between the two clusters in VANISH. In our study, the inflammatory and transcriptomic (SRS) subphenotypes showed similarity at the molecular level but only limited interaction in cluster assignments. In GAinS there was not a significant overlap between the two cytokine clusters and the three ConC subgroups in overlapping patients, suggesting that subphenotypes defined from these data types may capture different aspects of the individual response and may be considered together for better patient stratification.

Future work: A consensus patient classification

In addition to the blood proteome angle deployed in this thesis, a wide range of efforts using different data types have been applied to derive patient endotypes in sepsis, including clustering based on primarily clinical phenotyping (Seymour et al. 2019; Zador et al. 2019) and the leukocyte transcriptome (summarised in Section 1.5.2). The aims of these classifications are often not to predict mortality on an individual patient level, but to stratify patients into more homogeneous groups which may relate better with distinct sepsis pathophysiological features and potentially improve the efficacy of immunomodulatory therapies.

Additional work is required to derive a consensus classification system to test for interaction with treatments in clinical trials. Across the same data type, for example in gene expression, existing sepsis subgroups should be harmonised through collaboration to derive consensus groupings (Stanski and Wong 2019). Analysis in Chapter 4 across different data types (plasma proteome and leukocyte transcriptome) showed shared and distinct features of these classifications and indicated complementary roles in differentiating the immune response and finding relationships with outcome. Additionally, across different presentations of critical illness (e.g., sepsis, ARDS, severe COVID-19, acute pancreatitis), evidence has suggested generalisable patterns in the maladaptive response (Neyton et al. 2022; Sinha et al. 2021; Reyes et al. 2021) thus it may be valuable to test endotypes

and approaches defined in one disease in another, or to derive treatable traits across different types of critical illness (Maslove et al. 2022). For example, the three proteomic clusters proposed in this thesis could be applied to the sterile inflammation ICU cohorts or critically ill COVID-19 patients to understand whether more homogeneous subgroups could also be obtained.

Clinical translation

Despite the lack of progress in targeted treatments for sepsis, the recent success of immune-modulatory therapies in severe COVID-19 (Kalil et al. 2021; Gordon et al. 2021; Horby et al. 2021) demonstrates the potential of such approaches in a relatively homogeneous patient group with immune dysregulation. Accumulating evidence has also indicated that there may be differential responses to treatment between patient endotypes, for example to corticosteroids in sepsis (Antcliffe et al. 2019) and in COVID-19-related ARDS (Sinha et al. 2021).

In order to derive sepsis endotypes that enable targeted therapy, mechanistic evidence is needed to link the subphenotypes with specific immune responses that could be targeted; the subphenotypes need to be validated in independent cohorts to demonstrate reproducibility; serial sampling and clinical phenotyping should be obtained to discriminate inter-individual variation from disease progression, and to understand the different individual trajectories; the prognostic or predictive value of the subphenotypes needs to be demonstrated through clinical trials or post-hoc analysis; and a consensus classification needs to be derived, acknowledging challenges that stem from different study types, statistical methods, and timing of sampling across the studies, incomplete coverage of the patients across data types, and marginal patients with uncertain class memberships (Demerle et al. 2021). The diverse mechanisms revealed by multi-omic phenotyping in COVID-19 suggested the value of such approaches in identifying measurable traits suggestive of processes driving pathogenesis, and also illustrated the strength and challenges of multi-modal data

integration (COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium 2022).

6.3 Integration of multi-omics data

Multiple genomic and functional genomic datasets have been generated for overlapping subsets of the GAinS cohort, including Illumina genotyping (n=1168), exome sequencing (n>500), metagenomics (n>500), microarray- or RNAseq- based transcriptomics (n=1044), TCR and BCR repertoire sequencing (n~70), and mass spectrometry- or Luminex- based proteomics (n=1182). The molecular layers are interdependent but also reflect different aspects of the individual response. Integration of two or more data types can not only provide valuable insights into the inter-connectivity of the molecular phenotypes, but also capture complementary aspects of the biological state and thus refine patient stratification.

To understand shared and distinct mechanisms at the molecular layers, in Chapter 4 I characterised the proteomic and transcriptomic patient clusters at both levels and identified common and distinct pathways that were differentially regulated in each classification (Table 4.12). Future investigation using integrative methods such as matrix decomposition (Hore et al. 2016) or multi-omics factor analysis (MOFA, Argelaguet et al. 2020) to identify latent components/factors that link the RNA and protein features could disentangle the shared and single-modality variation, describe shared mechanisms between the layers, and therefore elucidate the contribution from each layer to the differentiation across patients or association with clinical phenotypes. Applying MOFA on 36 sepsis patients included in COMBAT, Kwok and colleagues (Kwok et al. 2022) showed that bulk gene expression and cell composition but not plasma proteins had large contributions to the same latent factors separating the SRS groups. The incomplete correspondence across modalities shown in this analysis as well as in comparing the ConC/SRS class memberships should be partly attributed to the difference in tissue types sampled. While SRS groups primarily describe the

dysregulation in leukocyte gene expression, the plasma proteome may more closely reflect the function of other organ systems through the secretome, tissue leakage products, and signalling molecules. Another factor is that since the higher-abundance proteins are more likely to be detected by mass spectrometry, the measurement does not comprehensively capture the whole plasma proteome while gene expression profiling achieves a more complete coverage of the transcriptome.

To achieve a more comprehensive patient classification, in Chapter 4 I combined ConC and SRS classifications and identified ~11% patients with an increased mortality hazard ratio of 3~4 compared with the majority reference group (Section 4.5.3). Network-based methods could further facilitate the inference of patient connectivity and community detection. In COMBAT (2022), Alberto Santos Delgado integrated TimsTOF and Luminex data using similarity network fusion (Wang et al. 2014) and further subdivided COVID-19 patients clinically assigned as severe; the clusters were not identified using the single modalities. Piotr Sliwa integrated proteomics, bulk RNAseq and CyTOF cell composition data in COMBAT using multi-layer patient similarity networks (Kivelä et al. 2014), uncovered unknown relationships, and identified patient clusters that associated with clinical severity and were enriched for specific pathways (unpublished work). Of the modalities used, the timsTOF proteins had the best predictive power for the clinical categories, suggesting the utility of plasma proteins in identifying more homogeneous states of disease. It would be interesting to apply such network-based approaches in the GAinS cohort and compare the outputs with patient clusters derived from single data types.

Lastly, integration with genomic information could uncover genetic drivers of the differential host response. In GAinS, gene expression in sepsis was mapped as a quantitative trait and 3795 cis- and 171 trans- acting loci (eQTL) were identified for key immune and metabolic response genes in sepsis response networks (Davenport et al. 2016). This analysis has revealed genetic modulation involving key mediators implicated in the hypoxic response, switch to glycolysis, and endotoxin tolerance,

including effects specific to SRS group. Further work is needed to understand if genetic variation can also predispose patients to distinct sepsis endotypes. A preliminary pQTL mapping has identified cis-associations for 23 proteins (including CFH, SERPINA1 and ITIH3) and trans-associations for 6 proteins (including CFB and TF) (I and Nikhil Milind and colleagues), providing the opportunity to understand the propagation of genetic regulation by comparison with the eQTL results, and to investigate context-specific regulation at the plasma proteome level.

6.4 Conclusion

The work described in this thesis has greatly expanded the resource of functional genomics data in sepsis to the proteomics level, including a high-throughput mass spectrometry dataset with >2500 clinical patient samples. Extensive analysis of the blood proteome has revealed key features in innate immunity, tissue necrosis and metabolism that distinguish the sepsis response from related conditions and define clinically informative sepsis subphenotypes. Future investigations should focus on understanding the cell type- or tissue- specific proteome in combination with other molecular levels especially the transcriptome. Clinical translation of patient stratification using omics approaches calls for mechanistic evidence, consensus classifications, optimisation of biomarker panels, and integration of multi-modal data. Together these findings and future directions highlight opportunities for a shift from the current syndrome-based single-disease paradigm to a mechanism-targeted personalised medicine approach.

A

Appendix to Chapter 2

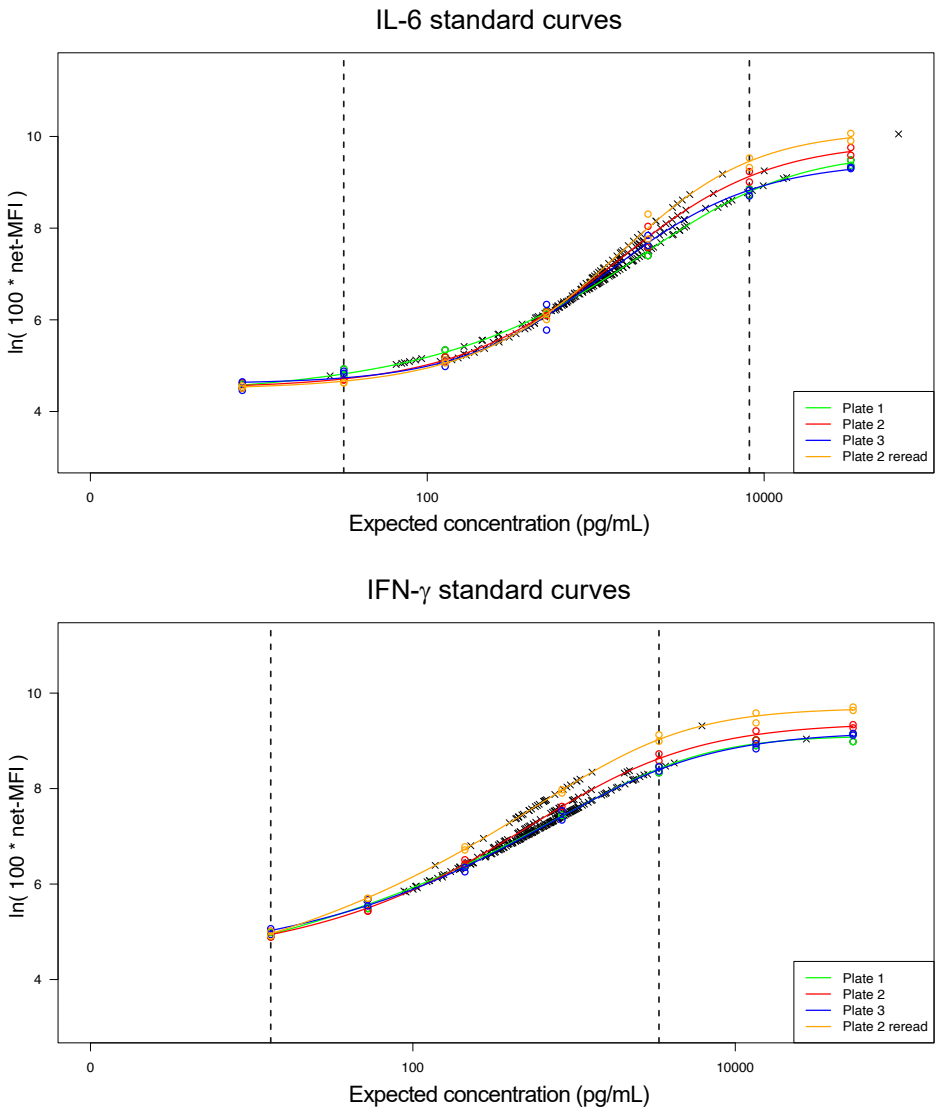


Figure A.1: Representative standard curves for two plasma cytokines assayed using Luminex. Coloured circles denote the gradient standards and black crosses denote the samples fitted on the curves. Vertical dashed lines indicate the lower and upper limits of quantifications (LOQs) for each analyte, sample values beyond which were censored at the LOQs in quality control.

Table A.1: Percentage of total samples assayed that were above or below the limits of quantification. LLOQ, lower limit of quantification; ULOQ, upper limit of quantification.

Analyte	Percentage <LLOQ	Percentage >ULOQ	Analyte	Percentage <LLOQ	Percentage >ULOQ	Analyte	Percentage <LLOQ	Percentage >ULOQ
MCP-1	0.0	0.0	IL-7	0.0	0.4	MMP-1	7.5	0.4
IL-6	0.8	2.5	BLC	0.0	0.0	IL-15	38.3	0.0
IL-8	2.1	0.4	Eotaxin-2	5.4	1.3	M-CSF	0.0	10.0
IL-10	0.8	0.4	Eotaxin	0.0	0.0	MCP-3	1.7	0.0
IL-18	0.0	0.4	IL-13	0.0	0.0	MIG	6.3	0.0
CCL3	4.2	0.0	IL-31	25.8	0.0	IL-16	0.4	0.0
IP-10	0.0	24.2	SCF	0.0	0.0	IL-21	4.2	0.0
IFN-γ	0.0	2.1	G-CSF	30.4	2.9	IL-3	9.2	0.0
IL-1β	7.5	0.0	GM-CSF	1.3	0.0	CD40-ligand	2.1	0.0
IL-2	37.5	0.0	HGF	0.0	2.9	FGF-2	6.7	0.4
IL-17	0.0	0.0	MIP-1β	1.7	0.0	IL-22	5.8	2.5
TNF-α	0.0	0.0	Eotaxin-3	0.8	0.4	VEGF-A	4.6	0.8
IFN-α	41.3	1.3	IL-9	1.7	0.0	TSLP	0.0	0.0
IL-12p70	5.4	0.0	MIF	0.8	1.7	IL-20	0.0	1.3
IL-4	0.0	0.4	TNF-β	7.5	2.1	ENA-78	0.0	0.4
IL-1α	22.9	1.7	bNGF	0.4	0.0	CD30	0.0	0.0
MCP-2	0.4	1.3	MIP-3α	0.0	0.0	TNF-RII	0.0	0.0
IL-2R	2.5	0.4	I-TAC	1.7	0.0	BAFF	12.9	0.0
SDF-1α	0.0	5.0	TRAIL	0.0	0.0	MDC	0.4	1.3
IL-27	0.0	0.0	Fractalkine	9.2	0.0	APRIL	0.0	0.0
LIF	10.8	0.0	GRO-α	9.6	1.7	Tweak	0.0	0.0
IL-5	11.3	0.0	IL-23	0.0	0.4			

B

Appendix to Chapter 3

Statistical power calculation

Relation between sample size and statistical power was calculated for one sepsis-control contrast and one within-sepsis contrast based on the MS192 dataset. Methods and the estimations are detailed below.

How to calculate the sample size

For testing one analyte in a simple linear regression which is equivalent to a t-test, the required sample size S in each group to detect a difference of magnitude Δ is given by:

$$S = 2\left(\frac{z_{\alpha/2} + z_{\beta}}{\Delta}\right)^2 \Sigma_p^2 \tag{B.1}$$

as shown in standard text books (for example in Chow et al. 2017), where

- $z_{\alpha/2}$ and z_{β} are the $100\frac{\alpha}{2}th$ and $100\beta th$ percentiles of the standard normal distribution.
- α is the probability of type I error i.e. rejecting H_0 when H_0 is true.
- β is the probability of type II error i.e. failing to reject H_0 when H_0 is false. The statistical power is given by $1 - \beta$.
- Σ_p^2 is the total variance within the group. The larger one of the variance calculated from each of the two groups being compared should be used. In MS2019 this variance only came from biological but not technical replications.
- Δ is the difference in group means that the study is being powered to detect, and is equivalent to the selected cut-off on $\log_2$(fold change) with values 1.25, 1.5, and 2.0 commonly used for the fold change.

Therefore, the calculation of S is simplified to the estimation of sample variance of a specified dataset, and the selection of appropriate α and β levels. In experiments where more than several analytes are tested, it is essential to select a stringent value of α to reduce the chance of false positive results. One definition of the FDR (false discovery rate), as used by Cairns et al. (2009), is given by:

$$FDR = E\left(\frac{\#FD}{\#FD + \#TD}\right) \tag{B.2}$$

where #FD is the number of false discoveries, and #TD is the number of true

discoveries. Assuming each protein analyte is either differentially abundant by some amount Δ or not differentially abundant, we have $E(\#FD) = \alpha(1-\pi)P$ and $E(\#TD) = (1-\beta)\pi P$, where π is the proportion of truely differentially abundant analytes, and P is the total number of analytes. Thus the expected FDR can be approximated by

$$\hat{E}(FDR) \approx \frac{\alpha(1-\pi)}{\alpha(1-\pi)+(1-\beta)\pi} \tag{B.3}$$

This was then used to select appropriate combinations of α and β to control FDR at a desirable level.

Calculation based on an existing dataset

Both the variance Σ_p^2 and the differential proportion π are specific to the dataset and groups being compared, while estimates could be obtained from existing data of similar nature. For understanding the sample sizes needed to study MS-based proteomic response in sepsis, I used another MS dataset (MS192) which included 192 samples from 102 sepsis patients and 21 surgery control patients, generated with collaborators on a Q Exactive HF LC-MS/MS system (Thermo Fisher) and the proteins identified using software Progenesis and PEAKS.

Step 1: Estimate π. I performed two sample size calculations respectively for a sepsis-control contrast and a within-sepsis contrast. Out of the 1123 proteins in the pre-processed MS192 dataset, there were 388 proteins ($\hat{\pi}$ = 34.6%) that were differentially abundant (FDR<0.05) between CAP sepsis patients (n=96) and pre-operation patients (n=21), and 59 proteins ($\hat{\pi}$ = 5.25%) differentially abundant (FDR<0.05) between the two sepsis transcriptome endotypes SRS1 (n=48) and SRS2 (n=48).

Step 2: Select appropriate α **and** β **to control for FDR**. Given the estimates of π from the two contrasts and following equation (B.3), the estimated FDR along α is shown in Fig. B.1 for a range of values for the statistical power $(1-\beta)$. For the CAP-post-operation contrast, to control FDR at $\leqslant 0.05$, type I error (α) needs to be no larger than 0.022 when 80% power is achieved, or no larger than 0.025 when 90% power is achieved. For the SRS contrast within sepsis patients where a much more subtle difference was detected, to control FDR at $\leqslant 0.05$, α needs to be no larger than 0.0023 when 80% power is achieved, or no larger than 0.0026 when 90% power is achieved. Across the commonly considered ranges, α had a much larger effect on FDR than $1-\beta$ had, and the critical values of α for each contrast were close. For calculating the relation between sample size and power in Step 4, the more stringent values of α were used.

Step 3: Estimate the variance. For each of the four comparator groups from the two contrasts, variance of 1123 proteins were calculated within the group and the percentiles shown in Table B.1. A larger sample size is needed to achieve a desirable power in proteins with larger variance, thus in Step 4 the sample sizes were calculated for the different percentiles of variance. For each contrast, the larger variance between the two groups were used.

Step 4: Calculate the relation between size and power. Following equation (B.1),

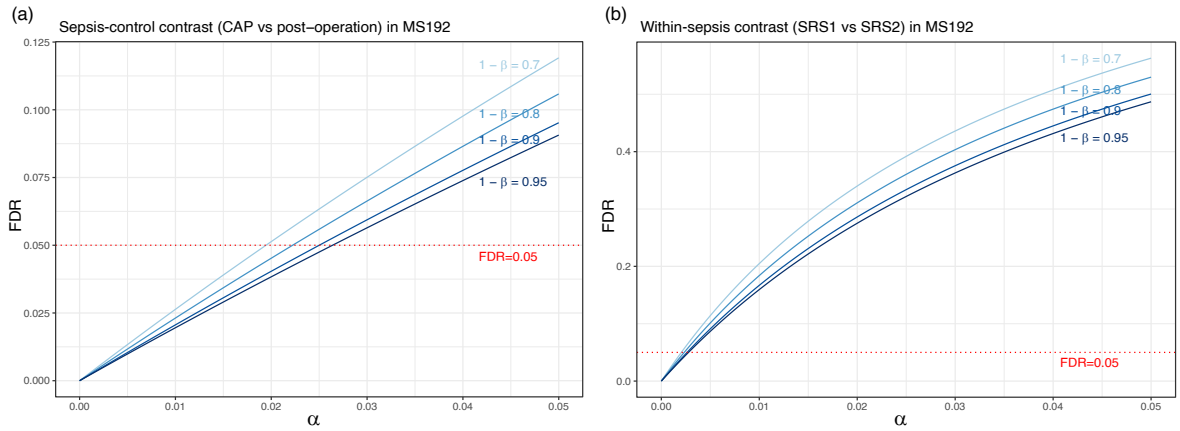


Figure B.1: Estimation of FDR from α and β, using the proportion of differential analytes estimated from the two contrasts in MS192. Given a certain β, the acceptable maximum value of α is given by the x position of the point of intersection between the corresponding blue curve and the red dotted line (FDR=0.05).

Table B.1: Percentiles of the 1123 values of variance for the 1123 proteins, calculated within each the four comparator groups in MS192. For power calculation for the CAP-post-operation contrast or the SRS1-SRS2 contrast, the larger variance of each contrast (i.e. those calculated from CAP or SRS1) was used. CAP: sepsis patients with community-acquired pneumonia.

	Post-operation	**CAP**	**SRS1**	**SRS2**
Median	1.39	1.94	1.97	1.75
70%	2.10	2.98	3.05	2.82
80%	2.67	3.74	3.87	3.60
90%	3.75	5.48	5.50	5.26
100%	12.40	22.22	21.60	22.32

the relation between sample size and statistical power was calculated as in Fig. B.2. A fold change cut-off of 1.5 was used to be consistent with the group comparisons in the following result sections. For the CAP-post-operation contrast, a minimum of 171 biological replicates is needed in each group to detect difference at 80% power and FDR<0.05 significance, for 70% of the analytes. In the same conditions, 269 biological replicates is needed for the SRS contrast. Other combinations of α and β are also sensible but the minimum sample sizes should be at a similar range.

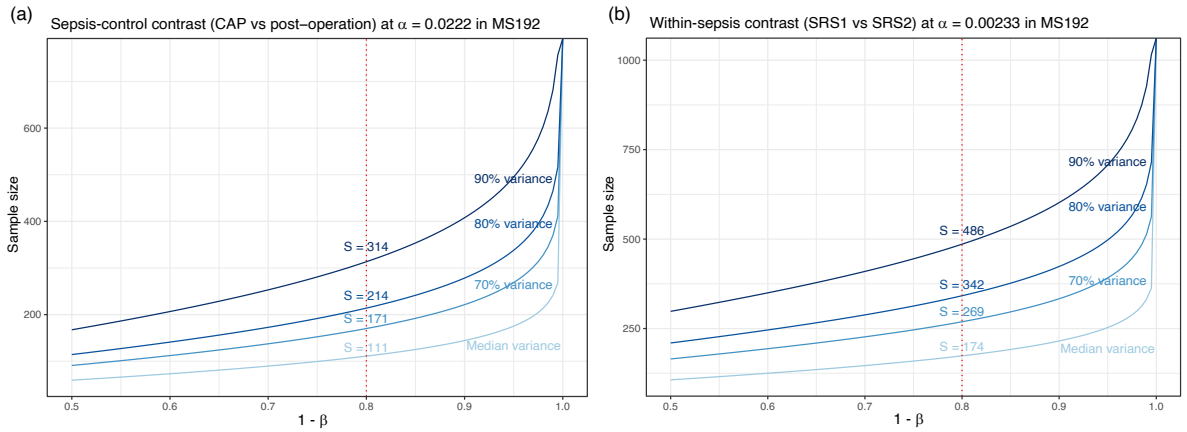


Figure B.2: Relation of samples size and statistical power using α values indicated from Fig. B.1 to control FDR $\leqslant$0.05. Separate curves are given for proteins with different variance at the percentiles stated, with more variable proteins at higher percentiles. The point of intersection between the blue curves and the red dotted line (power=80%) denote the minimum sample size required to achieve the desired power and significance level for 50%/70%/80%/90% of the analytes.

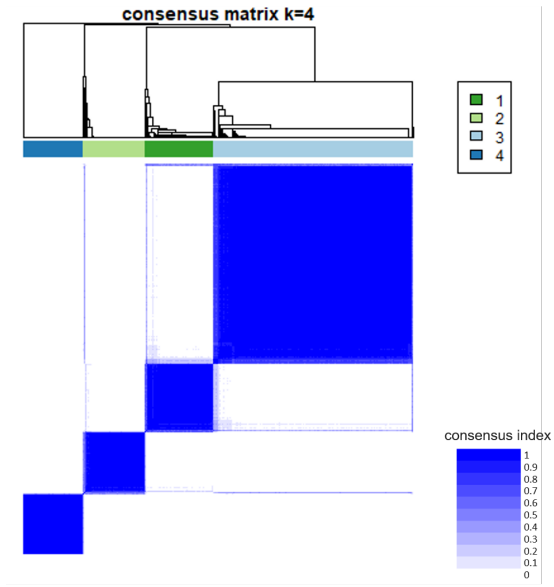
Appendix figures and tables

Table B.2: Summary of clinical characteristics in all GAinS patients in processed data. N denotes the number of patients with the measurement available. *For clinical measurements on the basis of day rather than patient, values recorded here are those measured on the day of the first available sample included in MS2019 for the patient. [1]Microbiology information is only available for patients with CAP. Patients with mixed bacterial-viral, or fungal infections were not counted in the bacterial or viral infections. The definitions of septic shock, ARDS phenotypes, and the abbreviations are the same as in Table 3.6.

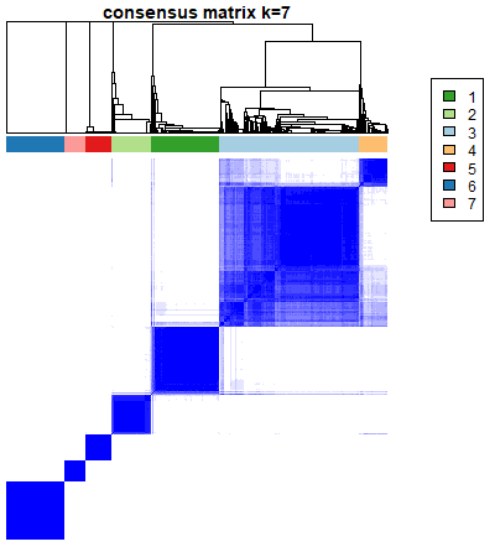
	All GAinS patients	N
No. patients	1182	-
Age median (IQR), years	65 (53-75)	1182
Men, No./total (%)	641/1182 (54)	1182
Pre-existing conditions, No./total (%)		
Heart/vascular diseases	486/1181 (41)	1181
Respiratory diseases	552/1181 (47)	1181
Current or ex- smoker	384/1141 (34)	1141
Malignancy or immune disease	211/1181 (18)	1181
Diabetes	217/1181 (18)	1181
Estimated time from disease onset, median (IQR), days		
Patients with CAP	5 (3-7)	494
Patients with FP	3 (1-4)	400
Microbiology[1], No./total (%)		
Bacterial	234/301 (78)	301
- Streptococcus pneumoniae	97/716 (14)	716
Viral	55/301 (18)	301
- Influenza	39/716 (5)	716
APACHE II score at day 1, median (IQR)	14 (11-18)	555
SOFA scores*, median (IQR)		
Cardiovascular	1 (0-4)	1171
Respiratory	2 (2-2)	1114
Kidney	0 (0-1)	1171
Liver	0 (0-0)	1125
Hematological	0 (0-1)	1167
Neurological	0 (0-0)	1171
Total	6 (3-8)	1073
Physiological variables*, median (IQR)		
Lowest mean arterial pressure, mmHg	66 (60-74)	1161
Lowest systolic blood pressure, mmHg	98 (87-110)	1168
Highest heart rate, beats/min	110 (96-125)	1168
Lowest heart rate, beats/min	80 (70-90)	1168
Arterial pH	7.37 (7.29-7.43)	586
Respiratory rate	26 (20-33)	1166
Partial pressure of oxygen (PaO_2), kPa	8.96 (7.9-10.3)	1106
Fraction of inspired oxygen (FiO_2)	0.4 (0.3-0.6)	1167
PaO_2/FiO_2, kPa	22 (15-31.1)	1106
Partial pressure of CO_2, kPa	5.4 (4.61-6.5)	1105
Lactate, mmol/L	1.7 (1.2-2.5)	737
Bicarbonate, mmol/L	24 (21-27.78)	1114
Highest urea, mmol/L	9 (5.82-14.18)	1158
Urine volumne, mL/24hrs	1600 (983-2428)	1161

Table B.2 continued from previous page

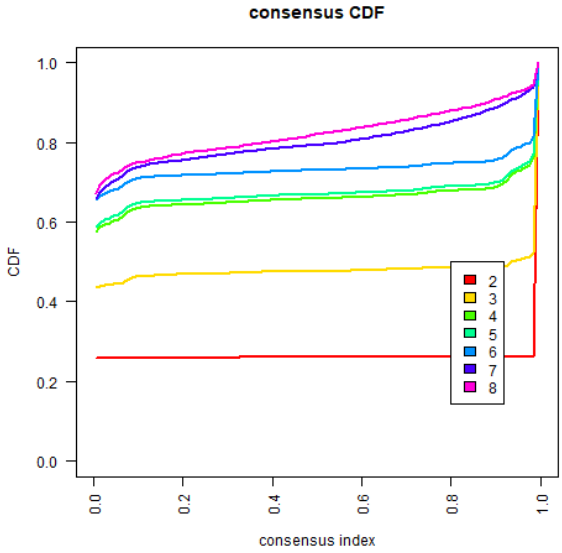
	All GAinS patients	N
Highest creatinine, μmol/L	88 (60-142)	1168
Lowest creatinine, μmol/L	81 (57-123)	1168
Highest bilirubin, μmol/L	10 (7-18)	1122
Alanine transaminase (AST), units/L	42.5 (26-73)	110
The international normalized ratio	1.2 (1.1-1.4)	512
Prothrombin time, seconds	14.75 (12.8-17.3)	424
Aspartate transaminase (ALT), units/L	29 (18-54)	690
AST/ALT ratio	1.4 (1.0-2.0)	76
Lowest platelets, $\times 10^3/\mu L$	204 (137.25-282)	1166
Highest temperature, C	37.4 (36.9-38)	1168
Lowest temperature, C	36.2 (35.8-36.7)	1168
Highest white cell count, $\times 10^3/\mu L$	12.7 (9-17.8)	1168
Lowest white cell count, $\times 10^3/\mu L$	11.3 (7.5-16.2)	1168
Haematocrit, %	34 (29.5-39)	495
Lymphocyte count, $\times 10^3/\mu L$	0.84 (0.5-1.28)	1136
Monocyte count, $\times 10^3/\mu L$	0.6 (0.3-0.9)	1135
Polymorphonucleocyte count, $\times 10^3/\mu L$	10.1 (6.7-15.1)	1138
Vasopressors or inotropes*, No./total (%)	486/1171 (42)	1171
Duration of inotrope/vasopressor support, median(IQR), days	2 (0-4)	1180
Shock*, No./total (%)	679/1160 (58)	1160
Mechanical ventilation/CPAP*, No./total (%)	751/1171 (64)	1171
Duration of mechanical respiratory support, median(IQR), days	4 (1-9)	1179
ARDS, No./total (%)		
No ARDS	854/1181 (72)	1181
Mild	25/1181 (2)	1181
Moderate	113/1181 (10)	1181
Severe	189/1181 (16)	1181
Acute renal failure*, No./total (%)	249/1171 (21)	1171
Renal replacement therapy*, No./total (%)	125/1171 (11)	1171
Duration of renal support, median(IQR), days	0 (0-0)	1179
Treated with activated protein C, No./total (%)	46/1181 (4)	1181
Corticosteroids, No./total (%)	277/1181 (24)	1181
28-day mortality, No./total (%)	198/1172 (17)	1172
6-month mortality, No./total (%)	288/1172 (25)	1172



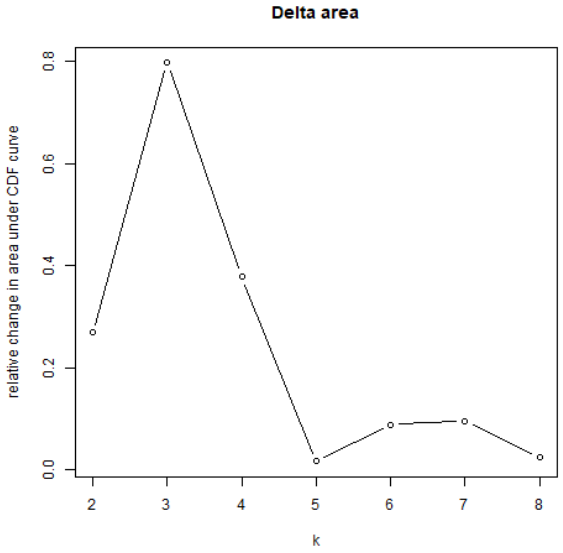
(a) Consensus matrix at k=4.



(b) Consensus matrix at k=7.



(c) CDF of consensus index.



(d) ΔAUCDF

Figure B.3: Consensus clustering of GAinS injections (n=1900) on binary states of unfiltered protein detection. **(a, b)** Clustering dendrogram and heatmap of consensus index between each sample pair are shown for cluster numbers 4 and 7, using the same colour scale for consensus index. A darker blue colour in the consensus matrix indicates a higher frequency of two samples falling in the same cluster. **(a)** Number of injections in clusters 4/2/1/3 are 292/300/334/974, respectively. **(c)** Cumulative distribution function (CDF) of the consensus index, with increasing number of clusters (k). **(f)** Relative increase in area under CDF at each increase of k from 2 to 8.

Table B.3: Summary of clinical characteristics for the 48 MOTION patients in the pre-processed dataset. "N" denotes the number of patients with information available.

Measurement	Displayed as	Level	N
Highest temperature, C	median (IQR)	37.8 (37.2-38.2)	48
Lowest temperature, C	median (IQR)	36.3 (36-36.7)	48
Highest Mean Arterial Pressure, mmHg	median (IQR)	103 (94-113)	48
Lowest Mean Arterial Pressure, mmHg	median (IQR)	66 (58-74.2)	48
Highest Heart Rate, beats/min	median (IQR)	91.5 (81.8-108.5)	48
Lowest Heart Rate, beats/min	median (IQR)	65 (58-72)	47
Highest Respiratory Rate, resp/min	median (IQR)	23 (19.5-26)	48
Lowest Respiratory Rate, resp/min	median (IQR)	14 (11-16)	45
Highest pH	median (IQR)	7.5 (7.4-7.5)	48
Lowest pH	median (IQR)	7.4 (7.3-7.4)	48
Highest Creatinine, umol/L	median (IQR)	65.5 (55-94.5)	48
Acute Renal Failure	No./total (%)	1/48 (2)	48
Highest hemoglobin, g/L	median (IQR)	97 (88.2-109.2)	48
Lowest hemoglobin, g/L	median (IQR)	93 (79-99)	9
Highest white blood cell, $x10^9$/L	median (IQR)	10.9 (8.2-14.3)	48
Lowest white blood cell, $x10^9$/L	median (IQR)	9.1 (8.8-9.3)	10
Lowest GCS off sedation	median (IQR)	10 (3.2-14.8)	46
Severe COPD	No./total (%)	1/48 (2)	48
Cirrhosis	No./total (%)	0/48 (0)	48
Immunocompromised	No./total (%)	0/48 (0)	48
Total APACHE II	median (IQR)	16.5 (13-22)	48
Oxygen saturation, %	median (IQR)	98 (97-98.2)	48
Lactate, mmol/L	median (IQR)	0.9 (0.8-1.1)	48
Vasoactive drugs	No./total (%)	34/48 (71)	48
Partial pressure of oxygen, kPa	median (IQR)	10.8 (9.7-12.2)	48
Fraction of inspired oxygen	median (IQR)	0.3 (0.2-0.4)	48
PaO2/FiO2 ratio, kPa	median (IQR)	35.5 (26.6-44.1)	48
Moderate or severe ARDS	No./total (%)	3/48 (6)	48
Renal replacement therapy	No./total (%)	0/48 (0)	48
Bilirubin, umol/L	median (IQR)	14 (7-23)	41
Richmond Agitation Sedation Score	median (IQR)	-5 (-5–4)	48
Traumatic Brain injury	No./total (%)	11/48 (23)	48
On muscle-relaxant	No./total (%)	4/48 (8)	48
Number of advanced respiratory support days	median (IQR)	11.5 (7-22.2)	48
Number of basic respiratory support days	median (IQR)	1 (0-2)	48
Number of advanced cardiovascular support days	median (IQR)	0 (0-3)	48
Number of basic cardiovascular support days	median (IQR)	12 (6-19)	48
Number of renal support days	median (IQR)	0 (0-0)	48
Number of neurological support days	median (IQR)	2 (0-9.2)	48
Number of gastrointestinal support days	median (IQR)	14 (6-24.5)	48
Number of dermatological support days	median (IQR)	0 (0-0)	48
Number of liver support days	median (IQR)	0 (0-0)	48
BMI	median (IQR)	24.8 (22.9-28)	48

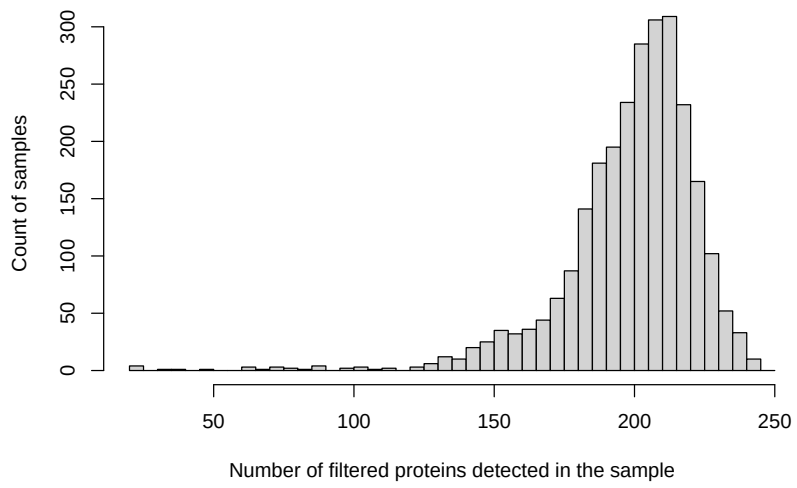
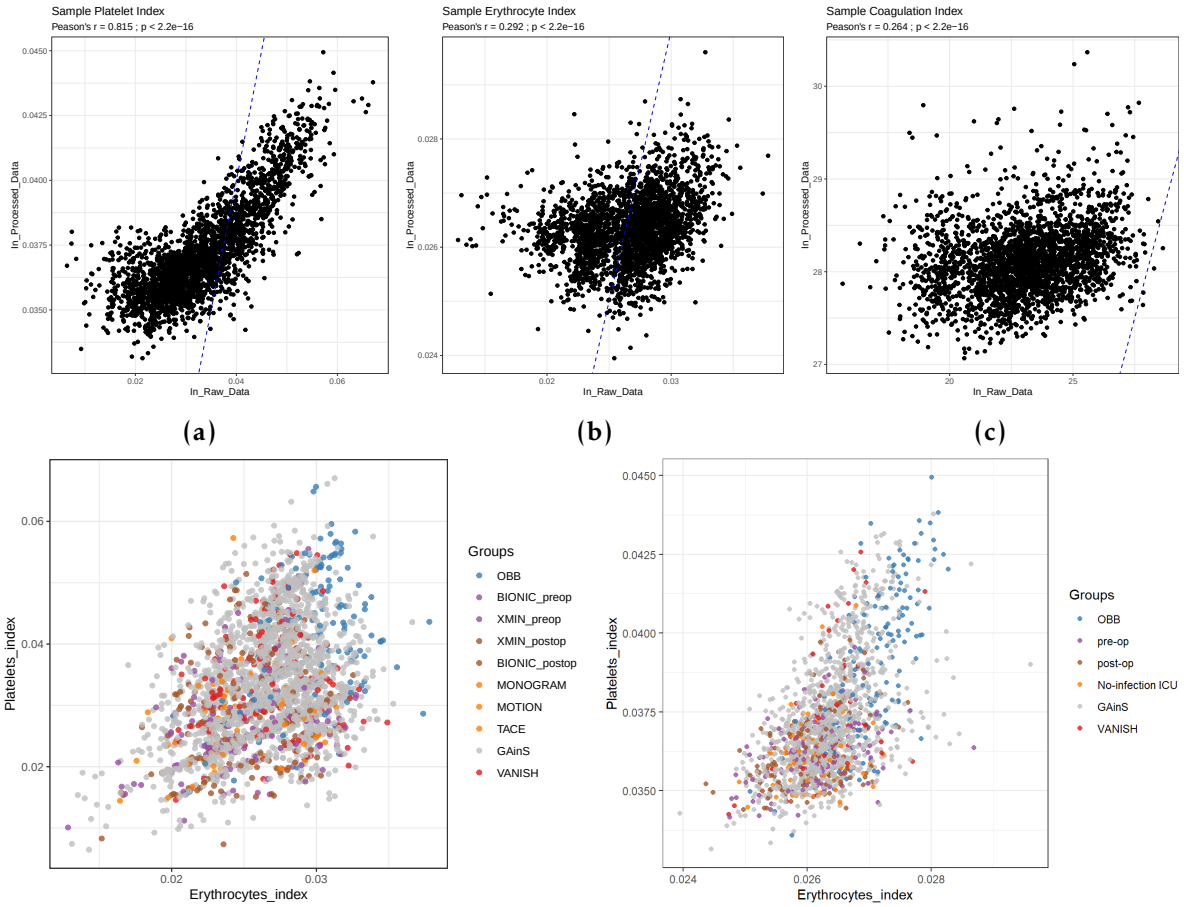


Figure B.4: Histogram of number of filtered proteins detected per sample. Only injections from actual samples (n=2647) were included.



(a) **(b)** **(c)**

(d) In raw data **(e)** In processed data

Figure B.5: Scatter plots of the three contamination indices showing correlation between those calculated from raw or interim processed data **(a-c)** or showing the correlation between the two cellular contamination indices in either raw **(e)** or processed **(e)** data. Blue dashed lines in (a-c) indicate the reference line $y = x$.

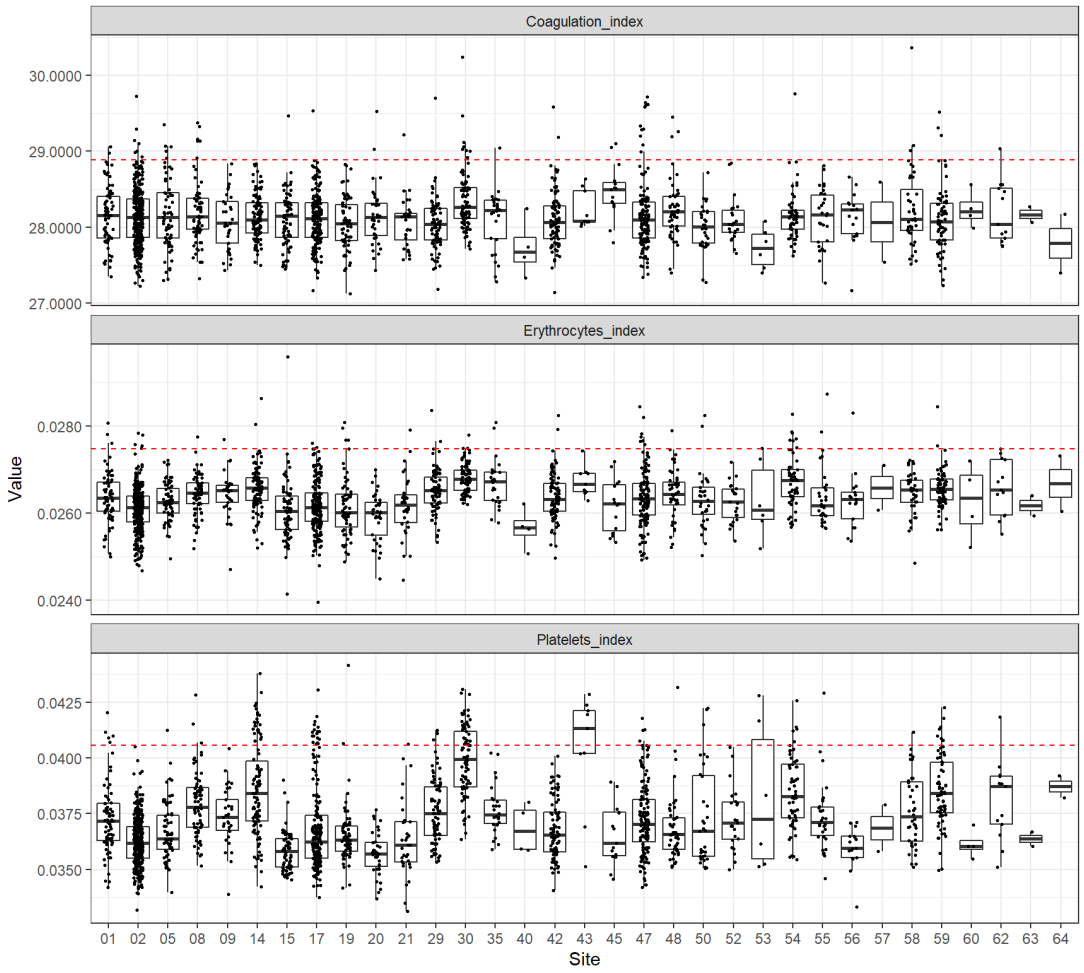
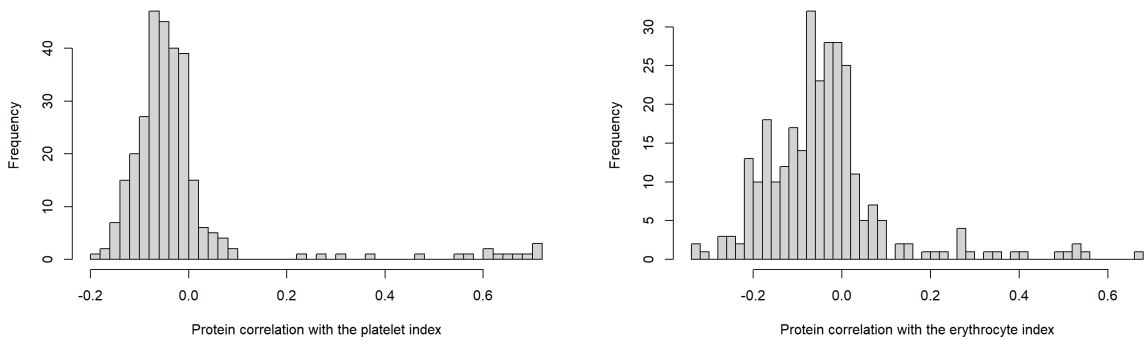


Figure B.6: Contamination indices across GAinS recruitment sites. Contamination indices were calculated in processed data version where the contaminations were not corrected. Red dashed lines indicate the sample outlier cut-off (mean + 2 s.d.) calculated in all MS2019 samples, excluding OBB in the erythrocyte index and platelet index.



(a) **(b)**

Figure B.7: Histogram of Pearson's correlation coefficients of 291 proteins with the platelet index (a) or the erythrocyte index (b), calculated in all samples in processed data version where the contaminations were not corrected.

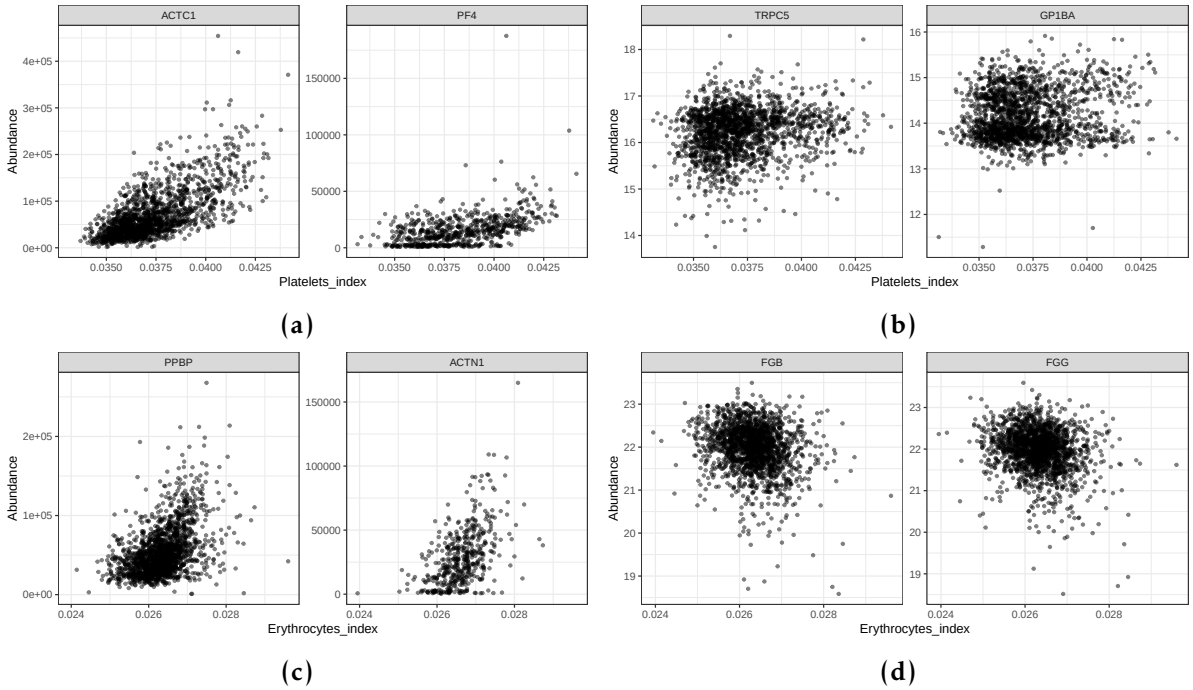


Figure B.8: Scatter plots of the removed (a, c) or remaining (b, d) protein intensities with the platelet index (a, b) or the erythrocyte index (c, d), in GAinS samples. Removed proteins with the top correlations but not on the marker panels were chosen to be plotted. Proteins remaining with top correlations in the fully processed data (after section 3.3.6) were chosen to be plotted. The fully processed protein intensities were used for the proteins remaining, while the raw log_2 intensities were used for the removed proteins. The contamination indices were calculated on the processed data version without correction for the cellular residue proteins.

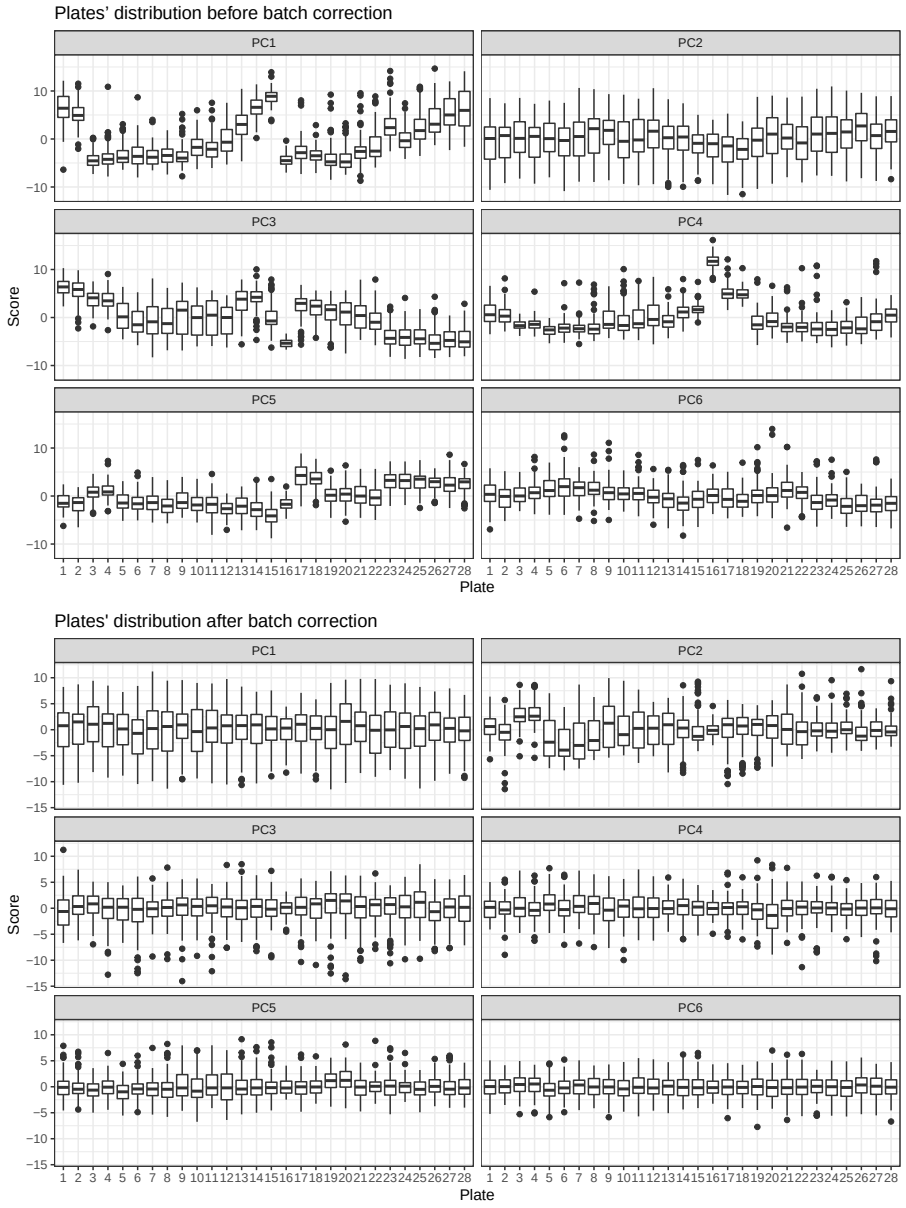


Figure B.9: Boxplots of PC1–6 scores across sample plates before or after batch correction.

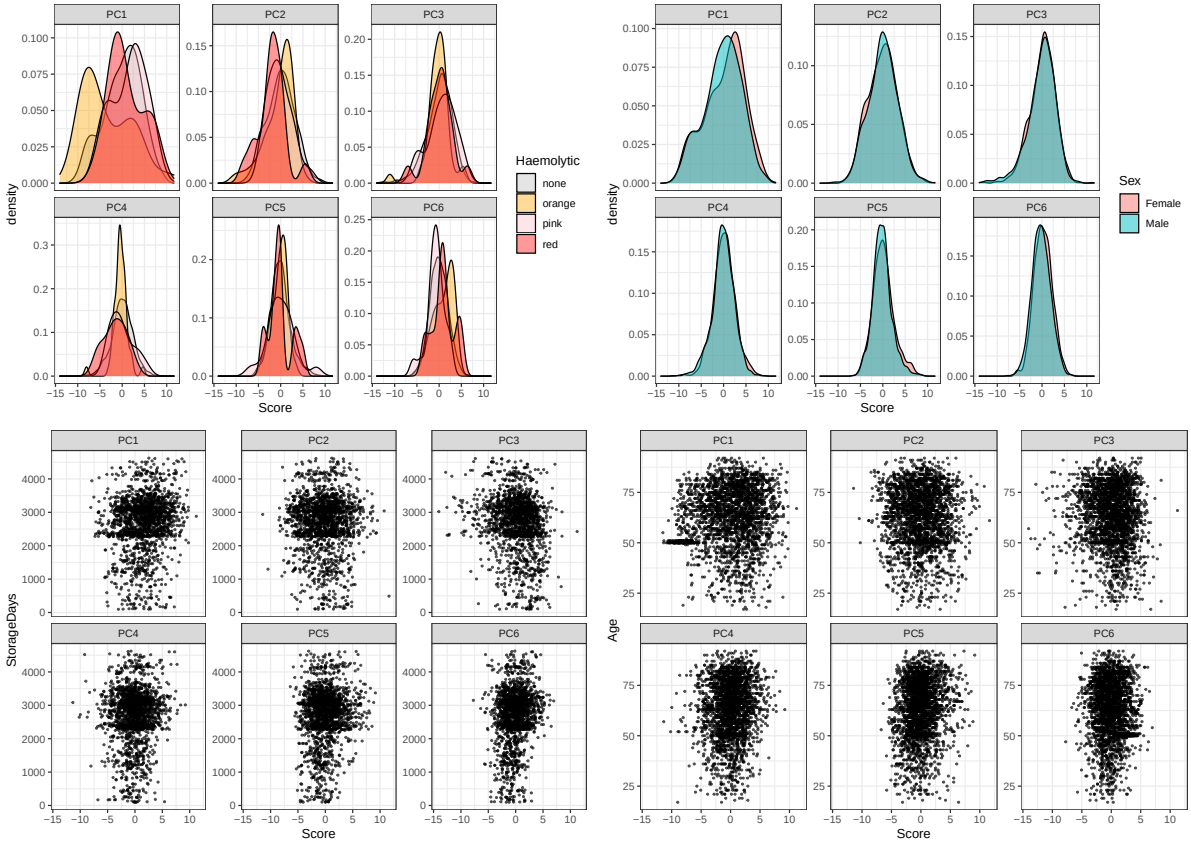


Figure B.10: Distribution on PC1–6 of samples with different levels of haemolysis, storage days, age or sex of patients. Sample numbers are: n(non-haemolytic)=2494, n(orange)=43, n(pink)=21, n(red)=17; n(female)=1112, n(male)=1463.

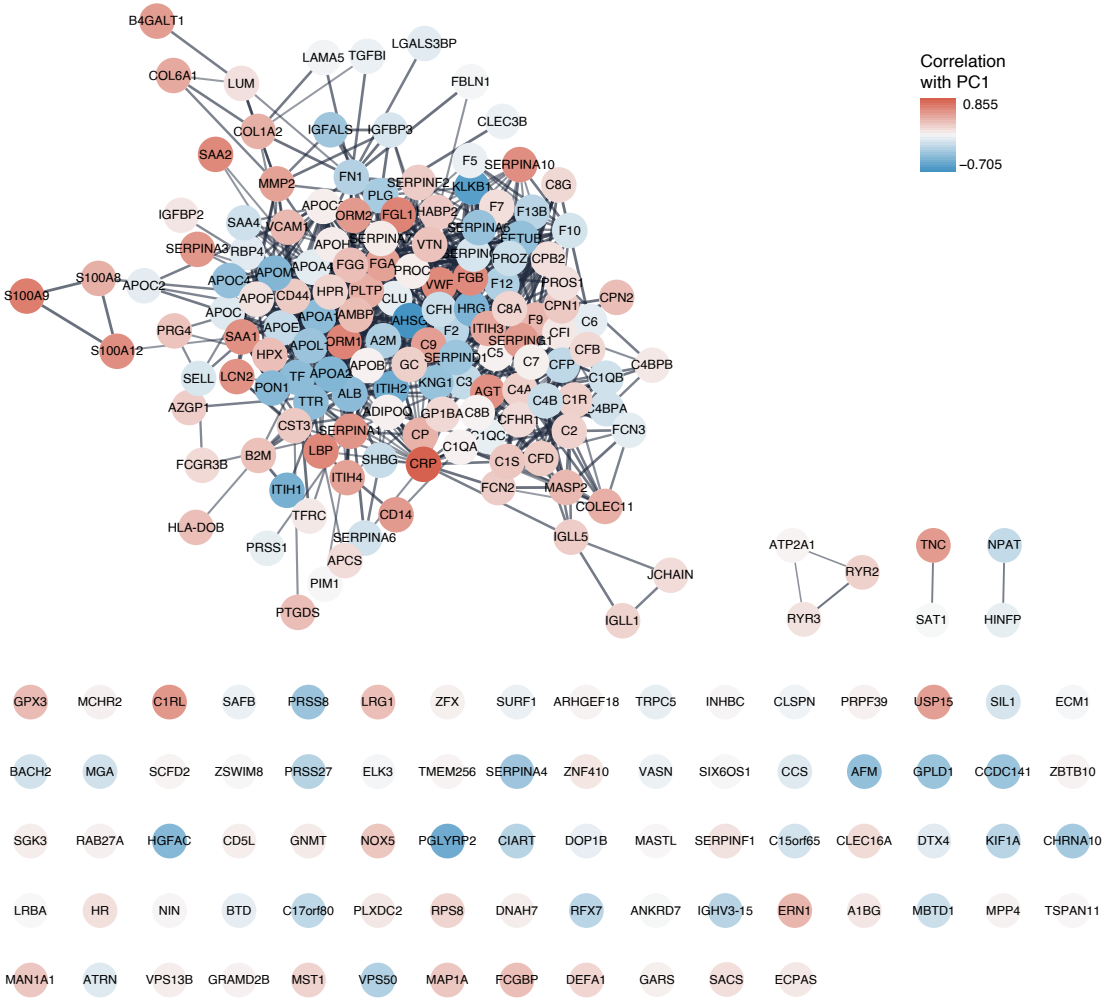


Figure B.11: Protein-protein interaction network of proteins detected in MS2019. Among the 269 proteins in the processed data, 229 mappable in the STRING database were shown in the figure. PPI network was constructed with a relatively high confidence score cut-off of 0.7 on the edges. Node colour was mapped to the Pearson's correlation coefficient between the protein level and PC1 score across the samples.

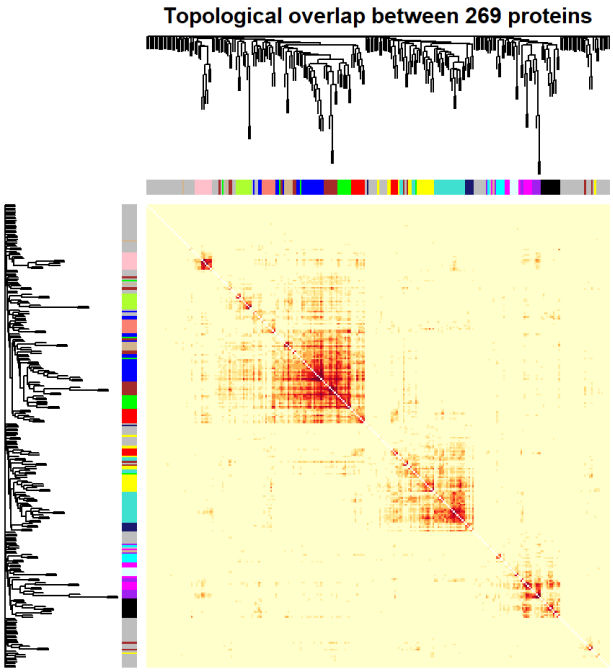


Figure B.12: Heatmap and dendrogram of the topological overlap matrix (TOM) based on protein co-expression. Rows and columns are the 269 protein species. Distance used in the cluster dendrogram is based on topological overlap between two proteins nodes, which is a function of the adjacency between the two nodes and between either node with all other nodes. For visualising in the heatmap, dissimilarity (1-TOM) was raised to the power of 15 to minimize effects of noise and spurious associations. A darker colour represents a higher similarity between two nodes. The colour bars for the rows and columns indicate the module membership of the 269 protein species, with the grey colour indicating 85 proteins that do not belong to any of the co-expression modules.

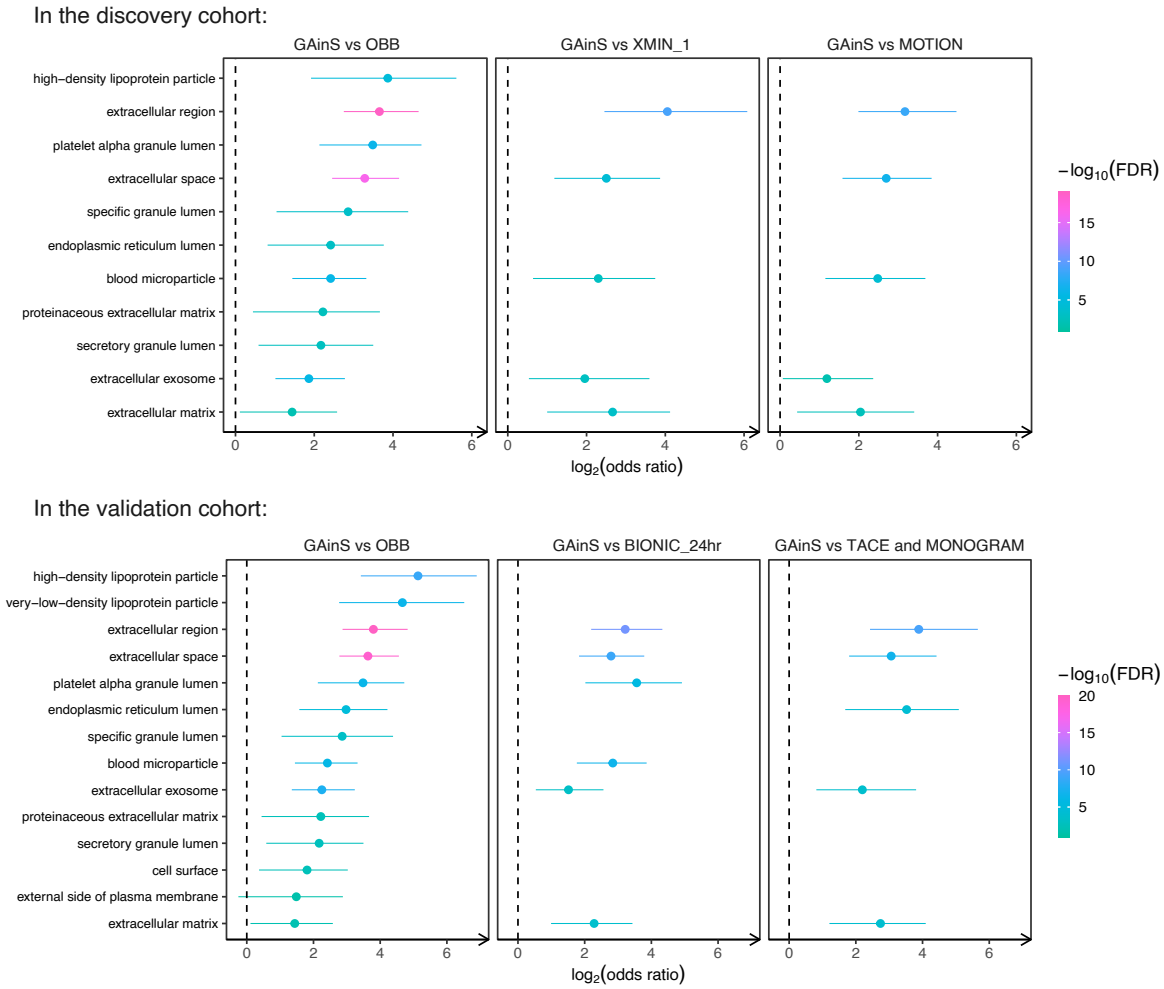


Figure B.13: Enrichment of differentially abundant proteins in Gene Ontology cellular components. Horizontal bars indicate 95% confidence intervals of log_2(odds ratio).

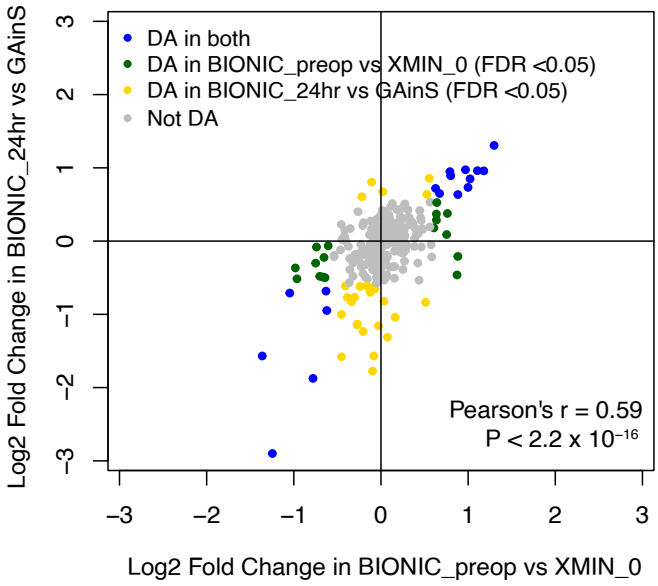


Figure B.14: Correlation between two citrate-EDTA contrasts in MS2019. DA: differentially abundant.

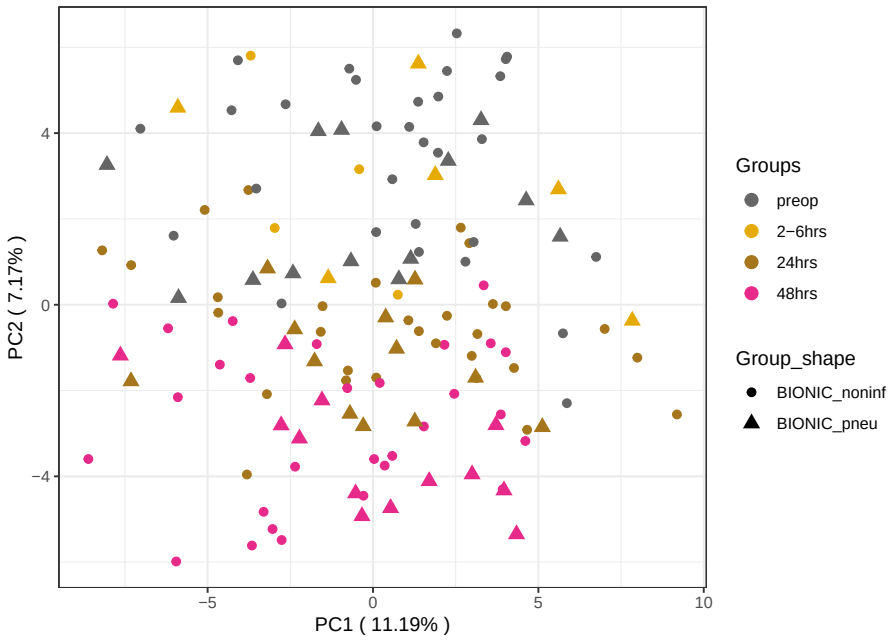


Figure B.15: PCA of plasma proteome in serial samples in BIONIC. Colour of the data points denote the timepoint relative to operation. Shape of the data points denote whether the patient had later developed pneumonia or did not have identified infection. PVE of PC1 or PC2 are shown in brackets.

C

Appendix to Chapter 4

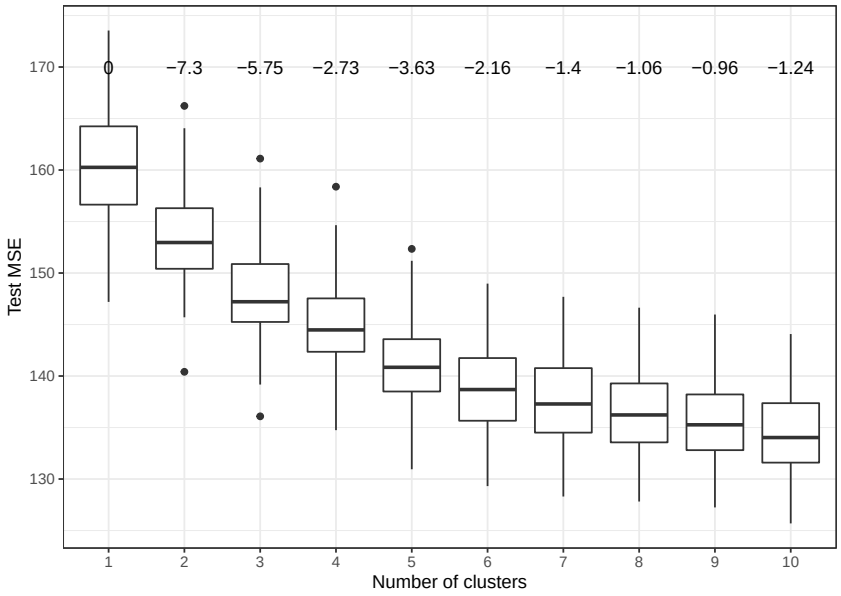


Figure C.1: K-means cross validation in discovery cohort. Test mean squared error (MSE) shows the mean sum-of-squared distances of test points to the nearest centre. A smaller test MSE indicates that a partitioning can be better generalised to samples out of the training sets. Fifty random draws were performed for each cluster number.

Table C.1: Comparison of continuous clinical variables between the clusters, in discovery cohort. More description on the clinical variables can be found in the Methods chapter. KW: Kruskal-Wallis test. Dunn's: Dunn's post-hoc test. "Dunn's FDR" shows the p values corrected within the three pairs tested for each variable, in the order of "Dunn's test pairs".

Clinical variable	KW FDR	Dunn's FDR	Dunn's test pairs	N(patients) with information		
				ConC1	ConC2	ConC3
Age	0.014	0.012, 0.020	ConC1vs3, ConC2vs3	170	183	435
Estimated day from onset (FP)	0.062	0.049, 0.028	ConC1vs2, ConC1vs3	99	50	116
Estimated day from onset (CAP)	0.82			41	80	202
Systolic blood pressure (low)	9.6E-06	5.1e-05, 1.4e-06	ConC1vs2, ConC1vs3	169	180	430
Mean arterial pressure (low)	4.6E-06	2.7e-05, 6.6e-07	ConC1vs2, ConC1vs3	168	180	426
Heart rate (high)	0.014	0.0275, 0.0035	ConC1vs2, ConC1vs3	169	180	430
Heart rate (low)	0.19			169	180	430
Vasopressor/Inotrope support (days)	2.3E-06	1.6e-06, 1.5e-06	ConC1vs2, ConC1vs3	169	183	434
Arterial pH	0.0001	4.5e-05, 1.7e-04	ConC1vs2, ConC1vs3	108	86	209
Respiratory rate	0.000026	4.5e-06, 1.0e-04, 4.1e-02	ConC1vs2, ConC1vs3, ConC2vs3	168	180	430
Partial pressure of oxygen	0.00049	5.1e-03, 6.3e-05	ConC1vs2, ConC1vs3	166	168	403
Fraction of inspired oxygen	0.57			169	181	429
PaO2/FiO2	0.5			166	168	403
Partial pressure of carbon dioxide	0.76			166	168	403
Mechanical respiratory support (days)	0.078			169	183	433
Lactate	3.1E-13	5.5e-09, 4.4e-15	ConC1vs2, ConC1vs3	111	115	265
Bicarbonate	1.6E-12	3.1e-09, 1.2e-13	ConC1vs2, ConC1vs3	168	170	409
Urea (high)	1.2E-11	2.0e-08, 9.7e-13	ConC1vs2, ConC1vs3	168	180	427
Urine volume	7.4E-16	3.1e-13, 9.1e-17	ConC1vs2, ConC1vs3	169	177	427
Creatinine (high)	3.4E-13	4.1e-09, 7.5e-15	ConC1vs2, ConC1vs3	169	180	431
Creatinine (low)	1.3E-11	1.4e-08, 1.4e-12	ConC1vs2, ConC1vs3	169	180	431
Renal support (days)	2.8E-09	1.6e-09, 1.9e-08, 4.9e-02	ConC1vs2, ConC1vs3, ConC2vs3	169	183	434
Aspartate transaminase	0.21			13	19	37
Alanine transaminase	0.79			109	106	246
AST/ALT	0.26			10	13	23
Prothrombin time	0.0012	0.03199, 0.00017	ConC1vs2, ConC1vs3	68	56	141
The international normalized ratio	0.000026	1.1e-04, 3.3e-06	ConC1vs2, ConC1vs3	67	88	201
Bilirubin (high)	0.0054	0.0033, 0.0020	ConC1vs2, ConC1vs3	163	174	416

Table C.1 continued from previous page

Clinical variable	KW FDR	Dunn's FDR	Dunn's test pairs	N(patients) patients with information		
				ConC1	ConC2	ConC3
Temperature (high)	0.9			169	180	430
Temperature (low)	0.0025	0.00222, 0.00075	ConC1vs2, ConC1vs3	169	180	430
Platelets (low)	0.00034	3.7e-03, 4.1e-05	ConC1vs2, ConC1vs3	169	179	431
White cell count (high)	0.71			169	180	431
White cell count (low)	0.0038	4e-02, 7e-04	ConC1vs2, ConC1vs3	169	180	431
Haematocrit	3.4E-08	1.8e-06, 9.4e-09	ConC1vs2, ConC1vs3	92	71	175
Lymphocytes (raw)	0.00082	0.03499, 0.00014, 0.04981	ConC1vs2, ConC1vs3, ConC2vs3	160	177	421
Monocytes (raw)	0.00012	2.3e-02, 1.6e-05, 2.7e-02	ConC1vs2, ConC1vs3, ConC2vs3	160	177	422
Polymorphonucleocytes (raw)	0.19			161	177	422
Total white cells (raw)	0.1			160	177	421
Lymphocytes (proportion)	0.64			160	177	421
Monocytes (proportion)	0.0097	0.032, 0.002	ConC1vs2, ConC1vs3	160	177	421
Polymorphonucleocytes (proportion)	0.27			160	177	421
ARDS classification	0.73			169	183	435
APACHE	0.000029	6.8e-06, 2.4e-04, 4.0e-02	ConC1vs2, ConC1vs3, ConC2vs3	107	84	185
SOFA-Total	1.2E-11	2.3e-10, 1.2e-11	ConC1vs2, ConC1vs3	163	164	391
SOFA-CVS	1.2E-11	6.8e-10, 5.0e-12	ConC1vs2, ConC1vs3	169	181	431
SOFA-Haem	0.00002	2.0e-04, 2.1e-06	ConC1vs2, ConC1vs3	169	179	431
SOFA-Hep	0.0015	0.00089, 0.00055	ConC1vs2, ConC1vs3	164	174	416
SOFA-Renal	1.2E-12	5.5e-10, 1.1e-13	ConC1vs2, ConC1vs3	169	181	431
SOFA-Resp	0.0031	0.0013, 0.0019	ConC1vs2, ConC1vs3	168	170	404
SOFA-Neuro	0.26			169	181	431

Table C.2: Comparison of categorical clinical variables between the clusters, in discovery cohort. CPAP – continuous positive airway pressure.

Clinical variable	χ^2 FDR	N(patients) with information		
		ConC1	ConC2	ConC3
Sex	0.24	170	183	435
Diagnosis	5.00E-12	169	180	434
Timepoint	0.0024	170	183	435
Shock	1.40E-07	168	179	426
Mechanical ventilation/CPAP	0.015	169	181	431
Acute renal failure	3.40E-08	169	181	431
Comorbidities: Heart/vascular	0.56	169	183	435
Comorbidities: Respiratory	0.059	169	183	435
Comorbidities: Malignancy or immune disease	0.057	167	180	415
Comorbidities: Diabetes/endocrine disorders	0.69	169	183	435
Smoking habit	0.56	169	183	435
Corticosteroids	0.17	169	183	435
Activated Protein C	0.24	169	183	435
Microbiology	0.057	34	56	119
S. pneumoniae	0.56	68	125	284
Influenza	0.36	68	125	284

Table C.3: Comparison of clinical characteristics between patients with sepsis due to either CAP or FP in the GAinS cohort. Numerical variables were tested with Mann-Whitney U tests. Categorical variables were tested with χ^2 tests. Correction for multiple testing were performed within numerical or categorical variables separately. *For patients with serial timepoint, clinical variables recorded here are those measured on the day of the first available sample included in MS2019. Abbreviations and definition of the clinical variables are the same as described for Table 3.6

	With community-acquired pneumonia	**With fecal peritonitis**	**FDR**
No. patients	769	407	
Age median (IQR), years	63 (51-74)	67 (57-77)	**6.4E-05**
Men, No./total (%)	441/769 (57)	197/407 (48)	**0.0068**
Pre-existing conditions, No./total (%)			
Heart/vascular diseases	315/769 (41)	170/406 (42)	0.85
Respiratory diseases	423/769 (55)	125/406 (31)	**2.0E-14**
Current or ex- smoker	305/742 (41)	76/393 (19)	**7.9E-13**
Malignancy or immune disease	95/769 (12)	115/406 (28)	**4.3E-11**
Diabetes	155/769 (20)	62/406 (15)	0.06
APACHE II score at day 1, median (IQR)	15 (11-19)	13 (9-18)	**0.0024**
SOFA scores*, median (IQR)			
Cardiovascular	1 (0-3.25)	1 (0-4)	**0.00087**
Respiratory	2 (2-2)	2 (2-2)	**7.7E-15**
Kidney	0 (0-1)	0 (0-2)	0.61

Table C.3 continued from previous page

	With community-acquired pneumonia	With fecal peritonitis	FDR
Liver	0 (0-0)	0 (0-0)	0.84
Hematological	0 (0-1)	0 (0-1)	0.65
Neurological	0 (0-0)	0 (0-0)	**0.002**
Total	5 (3-8)	6 (3-8)	0.72
Physiological variables*, median (IQR)			
Lowest mean arterial pressure, mmHg	66 (60-75)	66 (60-73)	0.89
Lowest systolic blood pressure, mmHg	98 (87-110)	98 (87-110)	0.61
Highest heart rate, beats/min	110 (95-125)	109 (97-123)	0.91
Lowest heart rate, beats/min	79 (69-90)	80 (70-90)	0.72
Arterial pH	7.39 (7.29-7.45)	7.36 (7.29-7.41)	**0.023**
Respiratory rate	30 (23-36)	22 (18-27)	**6.8E-36**
Partial pressure of oxygen (PaO2), kPa	8.5 (7.7-9.6)	10 (8.8025-11.6)	**3.3E-33**
Fraction of inspired oxygen (FiO2)	0.45 (0.35-0.6)	0.35 (0.28-0.4)	**2.5E-27**
PaO2/FiO2, kPa	19 (13.5-25.8)	30 (21.5-38.9)	**8.7E-40**
Partial pressure of CO2, kPa	5.6 (4.63-6.9)	5.2 (4.7-5.9)	**5.8E-05**
Lactate, mmol/L	1.6 (1.2-2.3)	1.9 (1.4-3.2)	**8.5E-05**
Bicarbonate, mmol/L	25 (21-29)	23 (20-26)	**7.7E-10**
Highest urea, mmol/L	8.7 (5.68-14.1)	9.85 (6-14.48)	0.21
Urine volumne, mL/24hrs	1785 (1100-2691)	1331 (826-1905)	**2.8E-10**
Highest creatinine, μmol/L	86 (60-141)	93 (62-144.5)	0.51
Lowest creatinine, μmol/L	80 (56-123.5)	84 (59-124)	0.45
Highest bilirubin, μmol/L	10 (6.5-18)	11 (7-18)	0.67
Alanine transaminase (AST), units/L	46 (28-85)	40 (20.5-57.5)	0.12
The international normalized ratio	1.2 (1-1.3)	1.2 (1.1-1.4)	**0.002**
Prothrombin time, seconds	14.5 (12.3-16.4)	16.5 (13.6-19.12)	**0.0001**
Aspartate transaminase (ALT), units/L	36 (21-67)	20 (14-31.25)	**1.1E-18**
AST/ALT ratio	1.4 (1.0-1.9)	1.5 (0.9-1.9)	0.89
Lowest platelets, $\times 10^3/\mu$L	202.5 (136-276)	207 (146.5-297.5)	0.19
Highest temperature, C	37.5 (36.9-38.2)	37.3 (36.8-37.8)	**0.00052**
Lowest temperature, C	36.2 (35.8-36.8)	36.1 (35.7-36.6)	0.11
Highest white cell count, $\times 10^3/\mu$L	13.1 (9.2-17.5)	12.4 (8.4-18.75)	0.41
Lowest white cell count, $\times 10^3/\mu$L	11.6 (7.8-16.25)	10.6 (6.55-16.15)	**0.025**
Haematocrit, %	35 (31.125-40)	31 (28-36)	**9.5E-09**
Lymphocyte count, $\times 10^3/\mu$L	0.9 (0.5-1.37)	0.8 (0.5-1.1)	**0.0023**
Monocyte count, $\times 10^3/\mu$L	0.64 (0.38-1)	0.5 (0.24-0.8)	**1.2E-06**
Polymorphonucleocyte count, $\times 10^3/\mu$L	10.1 (7-14.9)	10.3 (6.4-15.2)	0.74
Vasopressors or inotropes*, No./total (%)	290/764 (38)	194/401 (48)	**0.0014**
Duration of inotrope/vasopressor support, median(IQR), days	1 (0-4)	2 (0-4)	**0.013**
Shock*, No./total (%)	429/756 (57)	248/398 (62)	0.10
Mechanical ventilation/CPAP*, No./total (%)	520/764 (68)	228/401 (57)	**0.00048**
Duration of mechanical respiratory support, median(IQR), days	5 (1-11)	2 (1-6)	**4.1E-08**

Table C.3 continued from previous page

	With community-acquired pneumonia	With fecal peritonitis	FDR
ARDS, No./total (%) No ARDS Mild Moderate Severe	 502/769 (65) 10/769 (1.3) 88/769 (11) 169/769 (22)	 348/406 (86) 14/406 (3.4) 25/406 (6.1) 19/406 (4.7)	**9.1E-16**
Acute renal failure*, No./total (%)	162/764 (79)	85/401 (21)	1.00
Renal replacement therapy*, No./total (%)	76/764 (10)	47/401 (12)	0.43
Duration of renal support, median(IQR), days	0 (0-0)	0 (0-0)	0.74
Treated with activated protein C, No./total (%)	28/769 (3.6)	16/406 (3.9)	0.85
Corticosteroids, No./total (%)	187/769 (24)	88/406 (22)	0.41
28-day mortality, No./total (%)	150/762 (20)	48/404 (12)	**0.0017**
6-month mortality, No./total (%)	203/762 (73)	85/404 (79)	0.06

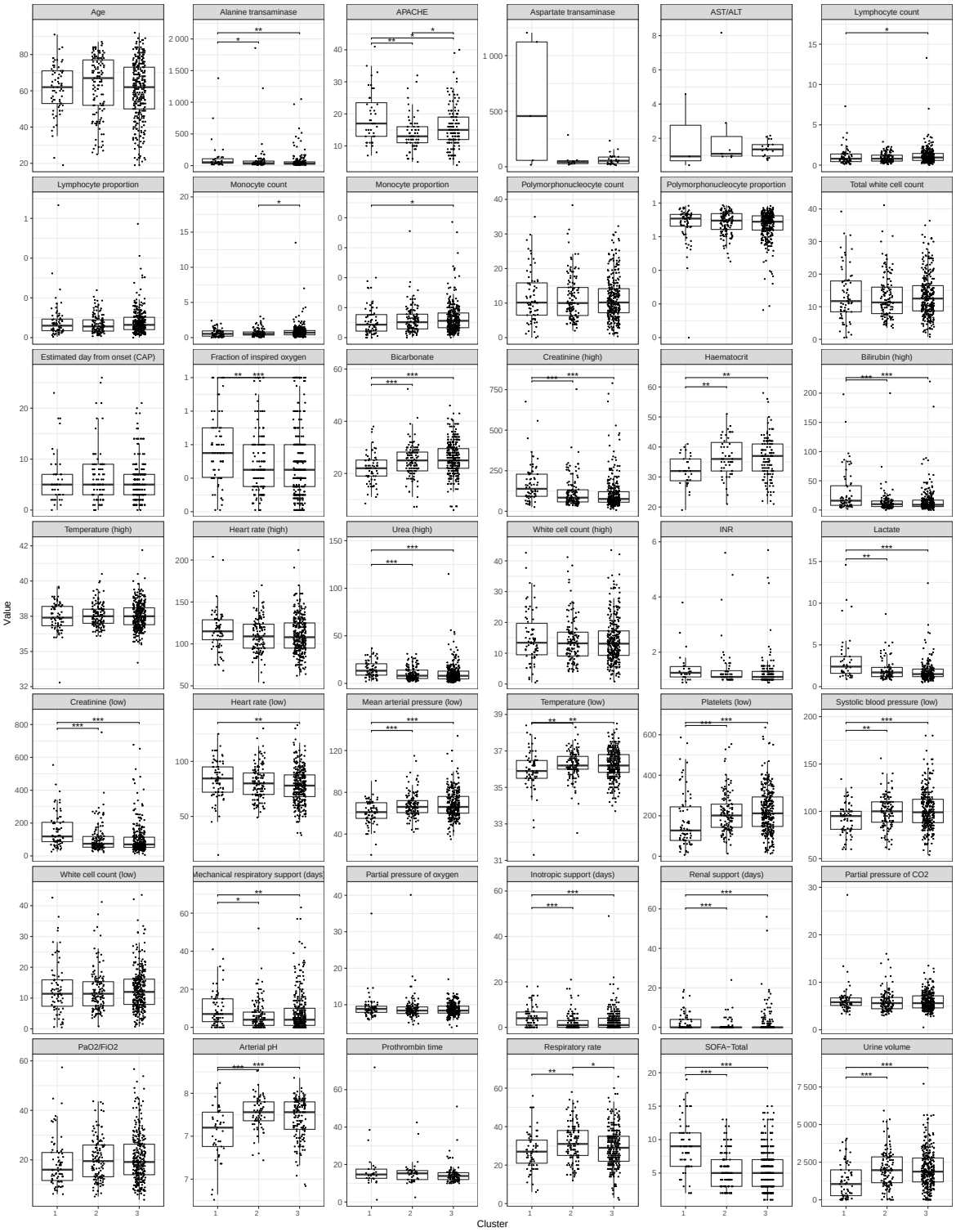


Figure C.2: Box plots of numerical clinical variables between the clusters in CAP patients. Discovery cohort cluster assignments (ConC) from mixed aetiology were used. Only samples from patients with sepsis due to CAP were compared and plotted. Asterisks indicate significance from Dunn's post-hoc tests (* FDR <0.05; ** FDR <0.01; *** FDR <0.001). For visualisation, two ConC3 sample points with INR = 14 and 99 were omitted. INR: international normalised ratio for prothrombin time. PaO_2/FiO_2: partial pressure of oxygen divided by fraction of inspired oxygen.

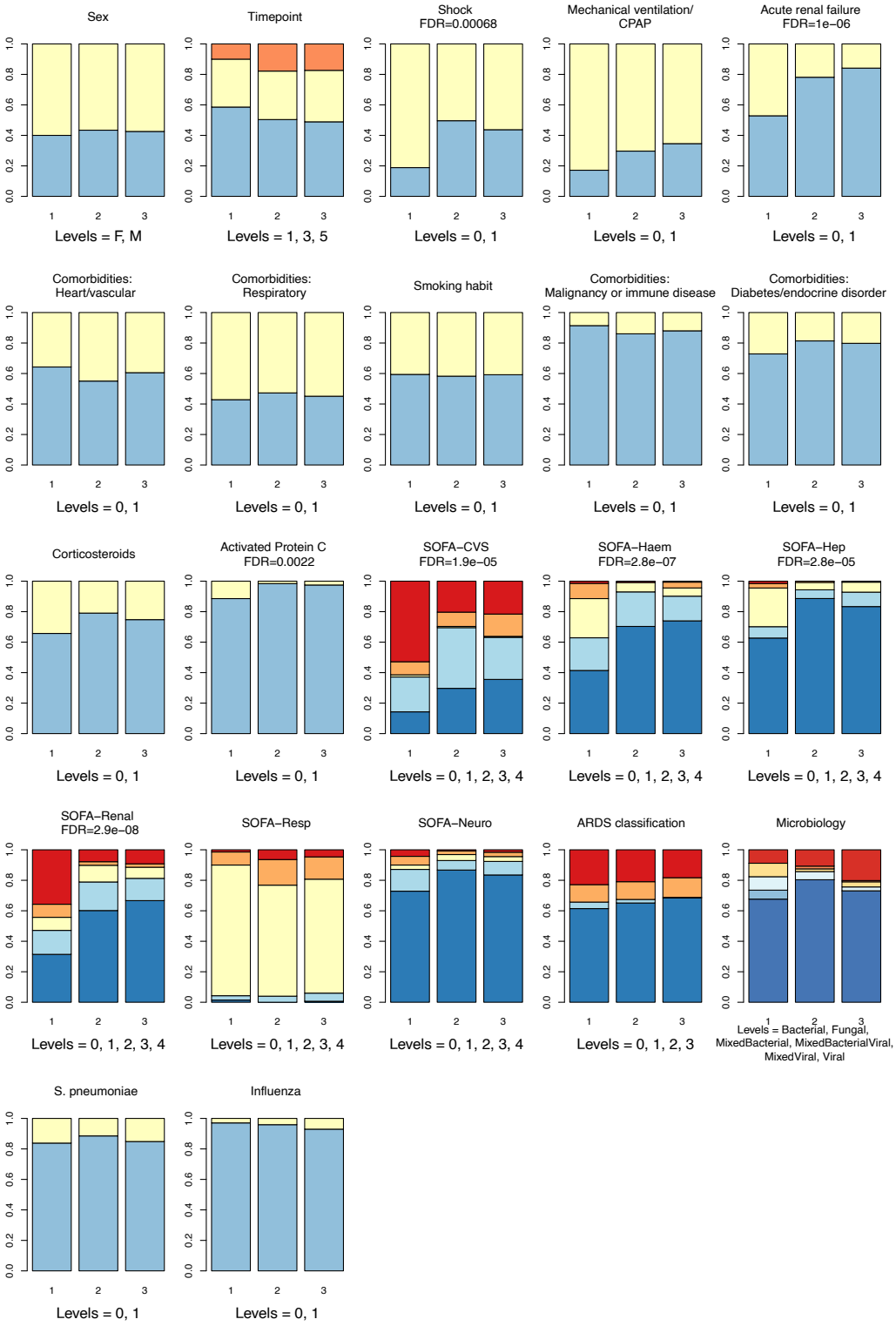
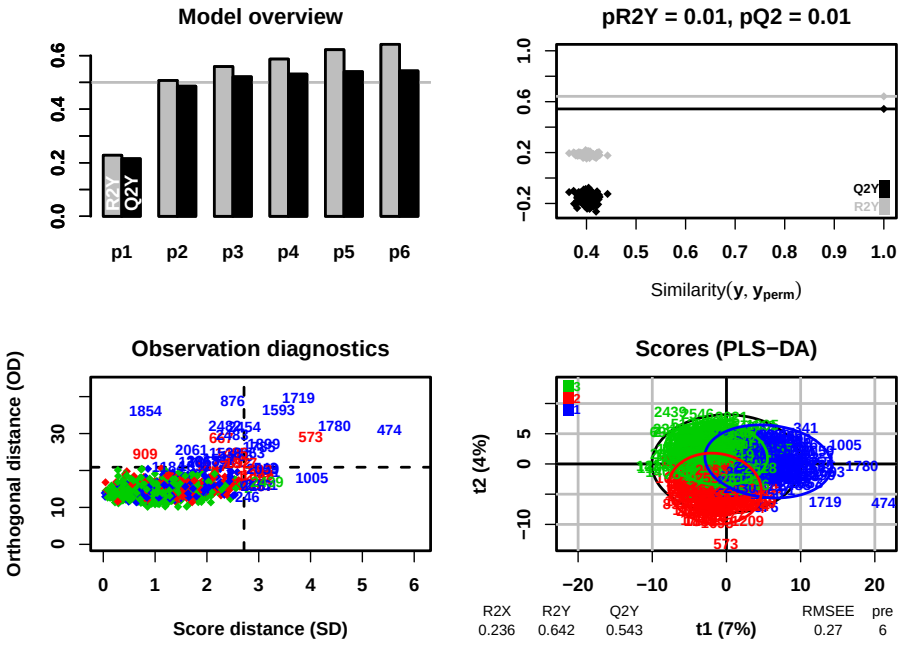
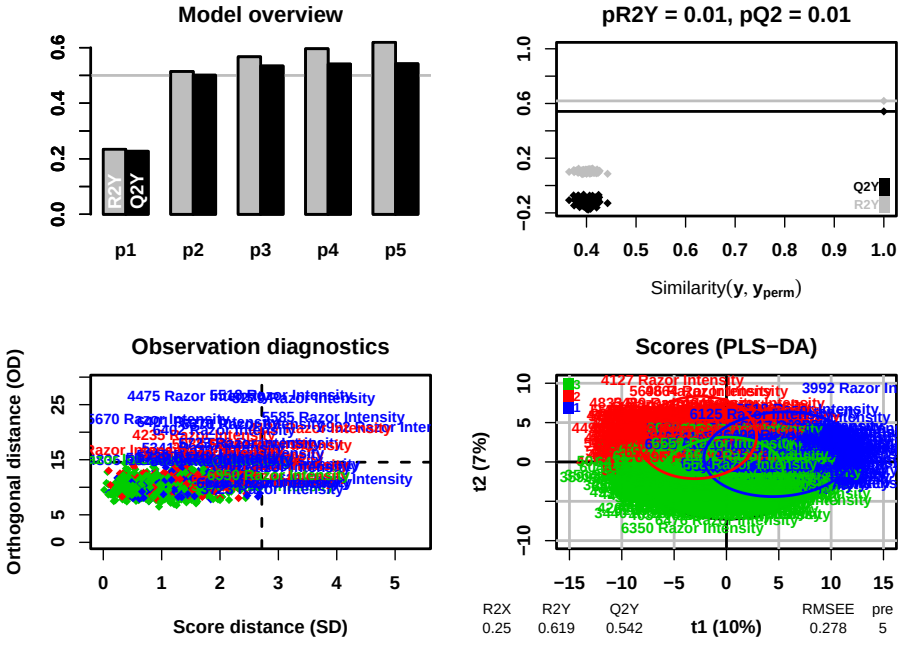


Figure C.3: Bar plots of categorical clinical variables between the clusters in CAP patients. Discovery cohort cluster assignments (ConC) from mixed aetiology were used. Only samples from patients with sepsis due to CAP were compared and plotted. FDR values labelled are from χ^2 tests, except for SOFA organ scores and ARDS classification for which the significance is derived from Kruskal-Wallis tests. For each variable, colours from the bottom to the top of the bars are in the order of the levels stated.

Figure C.4: PLS-DA model overview with 269 (a) or 138 (b) protein candidates input. Top left: inertia barplot suggests that 6 (a) and 5 (b) orthogonal components may be sufficient to capture most of the inertia; Top right: significance diagnostic: the R2Y and Q2Y of the model (horizontal lines) were compared with the corresponding values (dots) obtained after 100 random permutations of the y label (including one true label); Bottom left: outlier diagnostics; Bottom right: distribution of samples from the 3 classes on the first two components. R2X: fraction of X variation explained. R2Y: fraction of Y variation explained, within six-seventh of training set samples. Q2Y: fraction of Y variation explained when applying the model to one-seventh of training set samples as a further cross-validation test set. A good value for Q2Y is a value that is close to the R2Y, which means that the PLS model works independently of the specific training set data. RMSSE: square root of the mean error between the actual and predicted responses. pre: number of predictive components selected.



(a)



(b)

Table C.4: Cause of death for patients with the last available samples (among Day1/3/5) assigned to one of the three clusters. Percentages are the number of patients with the corresponding cause of death stated, out of the total number of recorded death within 6 months in the cluster. For each patient there could be more than one cause recorded. Cluster assignments in the validation cohort were predicted using the elastic net model with 269 candidate proteins, as described in Section 4.3.1.

Cause of death	Discovery cohort			Validation cohort		
	ConC1	ConC2	ConC3	ConC1	ConC2	ConC3
Intractable cardiovascular failure	18.8%	5.7%	4.9%	3.8%	10.5%	4.2%
Limitation of therapy	33.3%	37.7%	29.6%	46.2%	47.4%	31.3%
Failure to resolve organ system dysfunction	47.9%	43.4%	40.7%	42.3%	52.6%	35.4%
Unrelated cardiac/pulmonary event	0.0%	5.7%	2.5%	3.8%	5.3%	10.4%
Persistent or recurrent sepsis	20.8%	15.1%	13.6%	11.5%	21.1%	16.7%
Other unrelated cause	6.3%	17.0%	19.8%	11.5%	15.8%	14.6%
No. death in 6mth with a cause recorded	48	53	81	26	19	48

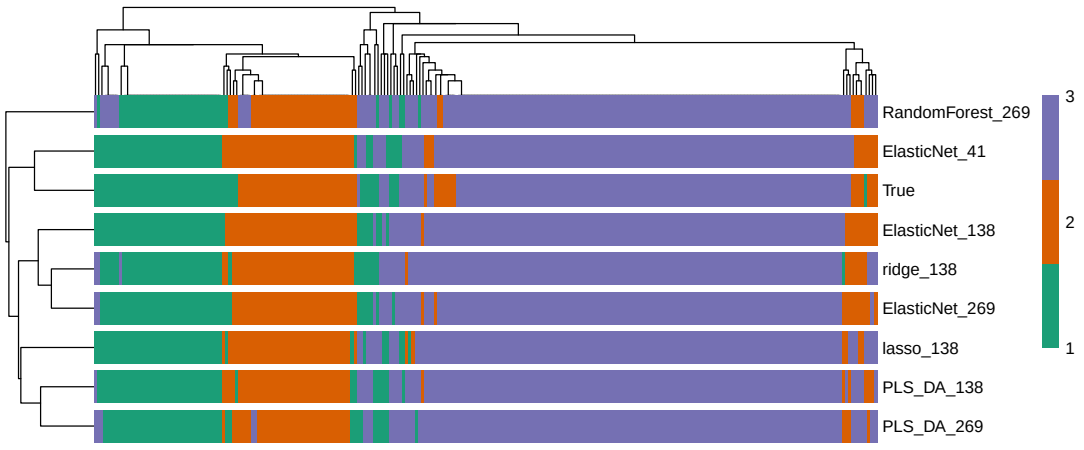


Figure C.5: Clustered heatmap of cluster memberships (1/2/3) predicted in the 8 models or from the true clustering result, within the 244 test set samples of the GAinS discovery cohort.

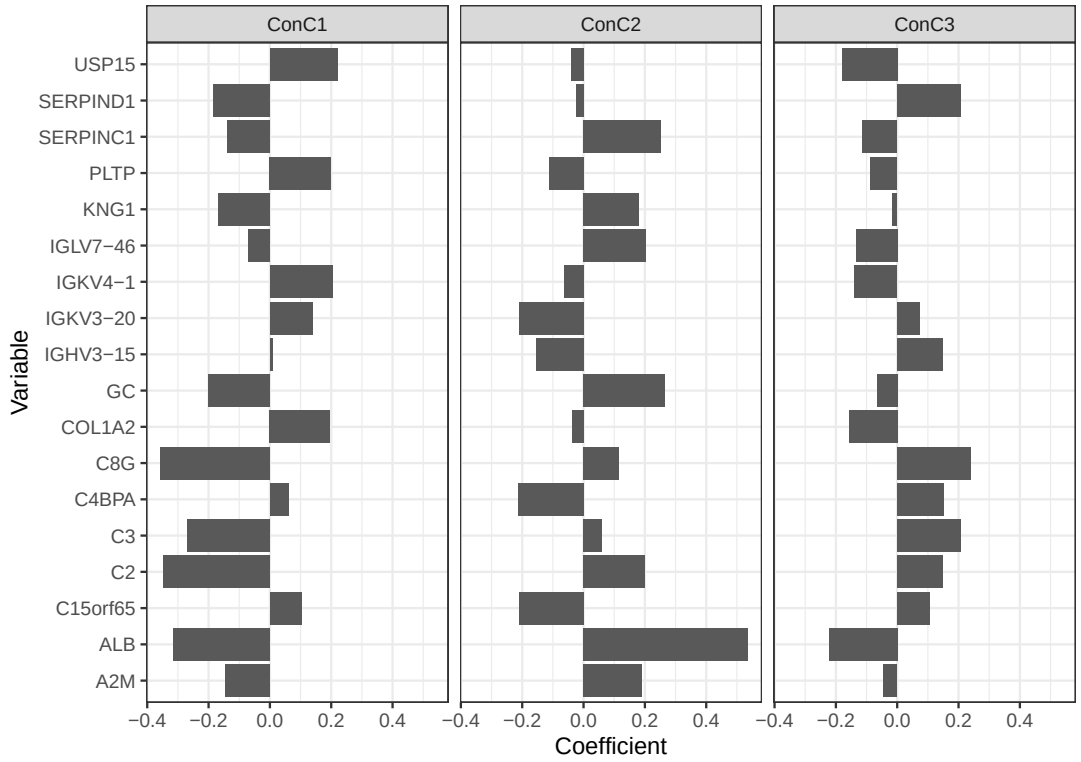


Figure C.6: Top protein contributions to the elastic net model with 138 proteins as input. Top 10 proteins with the largest absolute values of coefficients were included for each of the 3 models, consolidating to a list of 18 proteins. A positive coefficient indicates that a higher abundance of the corresponding protein predict the sample to be in the cluster; while a negative coefficient indicates that a lower abundance predict the samples to be in the cluster.

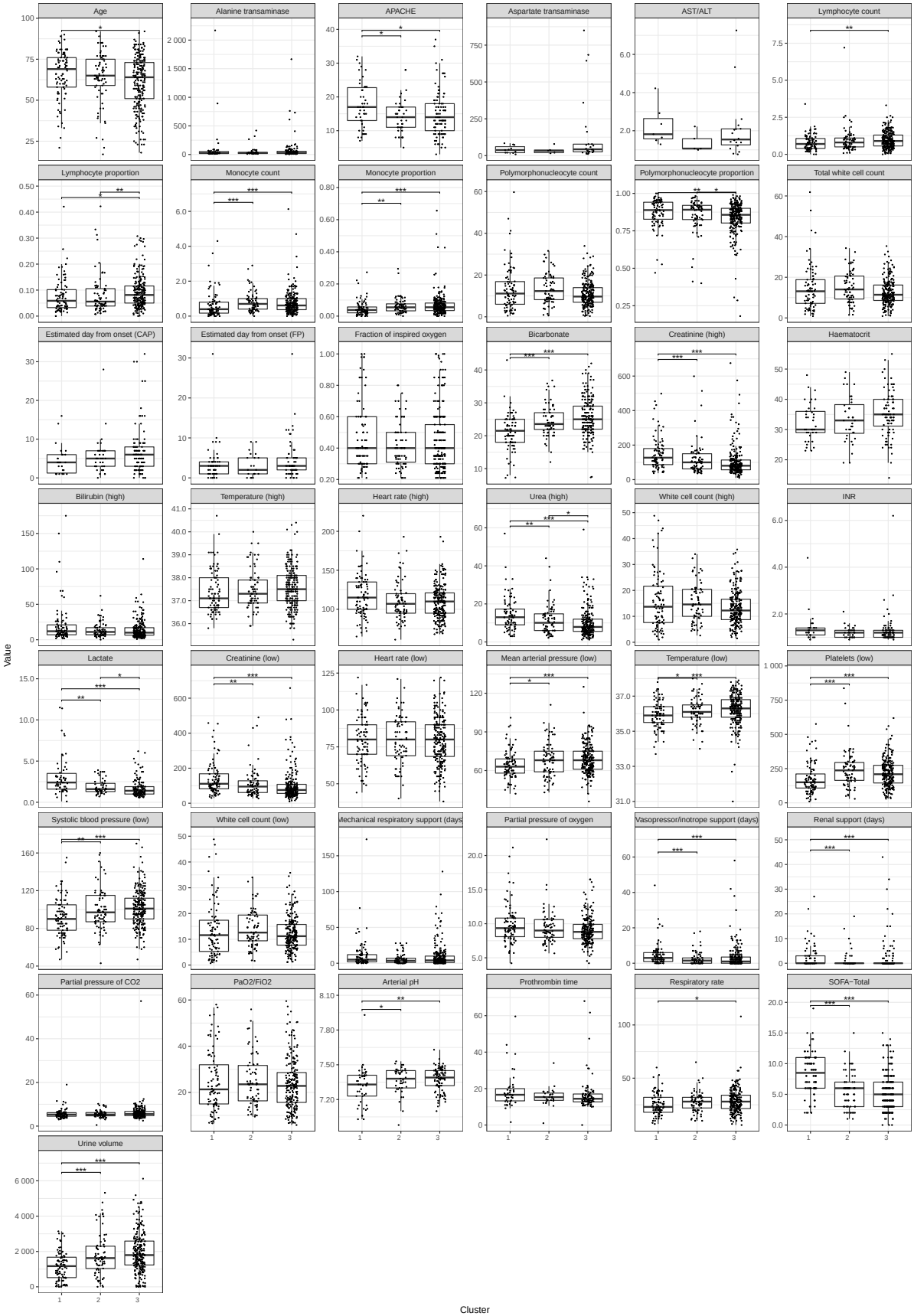


Figure C.7: Box plots of numerical clinical variables between the three validation cohort clusters. Asterisks indicate significance from Dunn's post-hoc tests (* FDR <0.05; ** FDR <0.01; *** FDR <0.001). For visualisation, one ConC1 sample point with INR =21.8 was omitted.

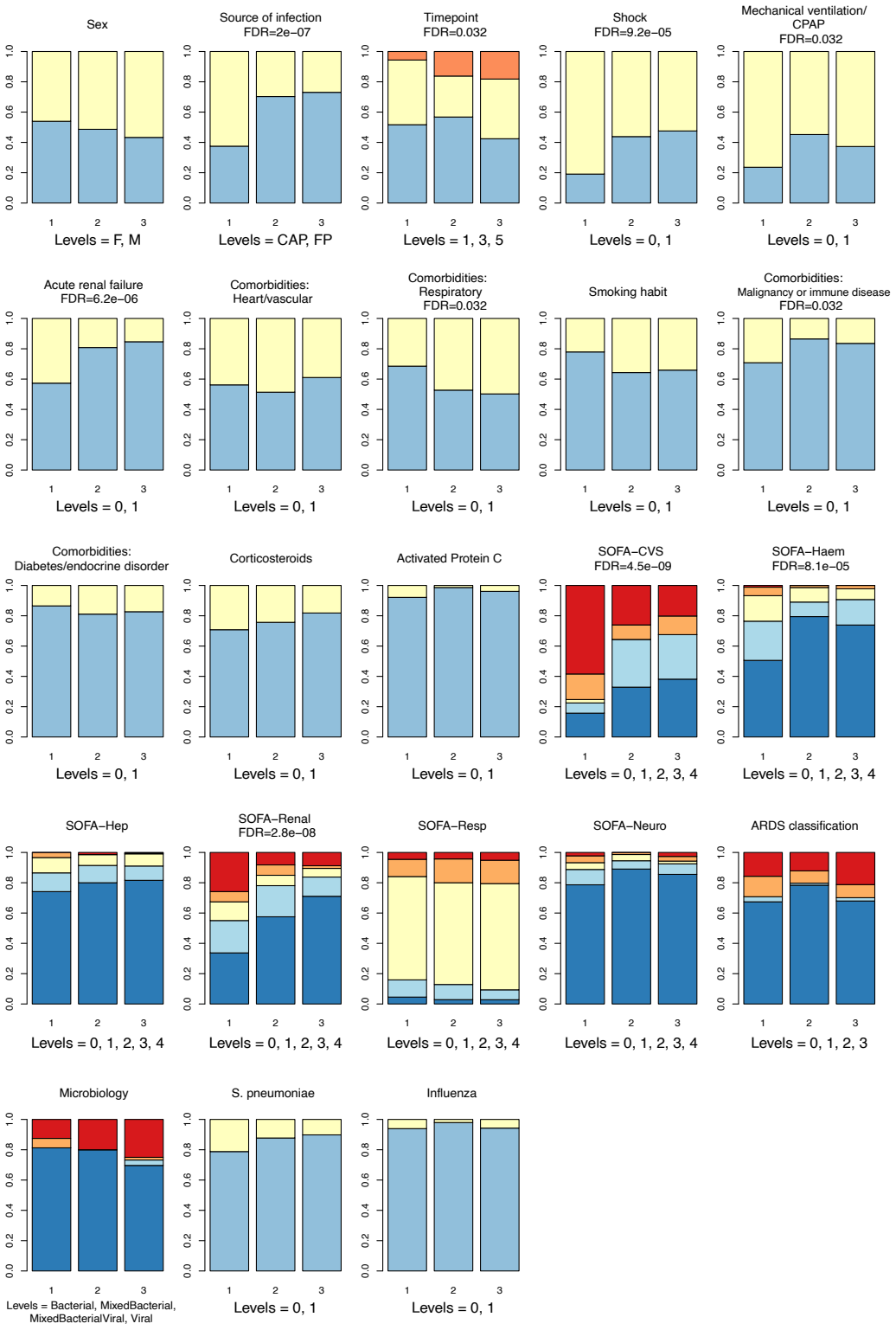


Figure C.8: Bar plots of categorical clinical variables between the validation cohort clusters. FDR values labelled are from χ^2 tests, except for SOFA organ scores and ARDS classification for which the significance is derived from Kruskal-Wallis tests. For each variable, colours from the bottom to the top of the bars are in the order of the levels stated. "Microbiology" was only available for part of the CAP patients.

Table C.5: Comparison of continuous clinical variables between the predicted clusters, in validation cohort. Only the first samples per patient were used for comparison. KW: Kruskal-Wallis test. Dunn's: Dunn's post-hoc test. "Dunn's FDR" shows the p values corrected within the 3 pairs tested for each variable, in the order of "Dunn's test pairs".

Clinical variable	KW FDR	Dunn's FDR	Dunn's test pairs	N(patients) with information		
				ConC1	ConC2	ConC3
Age	0.064	0.023	ConC1vs3	89	74	231
Estimated day from onset (FP)	0.3			53	21	61
Estimated day from onset (CAP)	0.16			26	37	108
Systolic blood pressure (low)	3.5E-04	8.6e-03, 2.7e-05	ConC1vs2, ConC1vs3	89	73	227
Mean arterial pressure (low)	0.0047	0.017, 0.00056	ConC1vs2, ConC1vs3	89	73	225
Heart rate (high)	0.11			89	73	227
Heart rate (low)	0.9			89	73	227
Vasopressor/Inotrope support (days)	9.2E-06	2.9e-05, 1.6e-06	ConC1vs2, ConC1vs3	89	74	231
Arterial pH	0.038	0.022, 0.0087	ConC1vs2, ConC1vs3	46	41	96
Respiratory rate	0.071	0.015	ConC1vs3	88	73	227
Partial pressure of oxygen	0.11			88	69	212
Fraction of inspired oxygen	0.58			89	73	226
PaO2/FiO2	0.82			88	69	212
Partial pressure of carbon dioxide	0.17			87	69	212
Mechanical respiratory support (days)	0.18			89	74	231
Lactate	4.0E-07	4.7e-03, 1.4e-08, 1.4e-02	ConC1vs2, ConC1vs3, ConC2vs3	58	49	139
Bicarbonate	1.5E-07	3.0e-04, 2.4e-09	ConC1vs2, ConC1vs3	85	68	214
Urea (high)	3.5E-06	6.9e-03, 1.4e-07, 2.8e-02	ConC1vs2, ConC1vs3, ConC2vs3	87	72	224
Urine volume	3.5E-07	2.1e-04, 7.5e-09	ConC1vs2, ConC1vs3	89	73	226
Creatinine (high)	3.6E-07	9.5e-04, 8.6e-09	ConC1vs2, ConC1vs3	89	73	226
Creatinine (low)	5.4E-06	3.7e-03, 2.2e-07	ConC1vs2, ConC1vs3	89	73	226
Renal support (days)	8.1E-05	6e-04, 8e-06	ConC1vs2, ConC1vs3	89	74	230
Aspartate transaminase	0.5			11	5	25
Alanine transaminase	0.9			54	46	129
AST/ALT	0.3			7	5	18
Prothrombin time	0.12			43	28	88
The international normalized ratio	0.15			28	33	95
Bilirubin (high)	0.36			89	69	211

Table C.5 continued from previous page

Clinical variable	KW FDR	Dunn's FDR	Dunn's test pairs	N(patients) with information		
				ConC1	ConC2	ConC3
Temperature (high)	0.16			89	73	227
Temperature (low)	0.001	0.041, 0.0001	ConC1vs2, ConC1vs3	89	73	227
Platelets (low)	6.8E-05	2.8e-05, 3.7e-05	ConC1vs2, ConC1vs3	89	73	225
White cell count (high)	0.13			89	73	226
White cell count (low)	0.24			89	73	226
Haematocrit	0.12			37	36	84
Lymphocytes (raw)	0.026	0.005	ConC1vs3	83	69	226
Monocytes (raw)	0.001	0.00045, 0.00034	ConC1vs2, ConC1vs3	82	69	225
Polymorphonucleocytes (raw)	0.12			83	69	226
Total white cells (raw)	0.12			82	69	225
Lymphocytes (proportion)	0.01	0.012, 0.0073	ConC1vs3, ConC2vs3	82	69	225
Monocytes (proportion)	0.0011	0.0018, 0.00018	ConC1vs2, ConC1vs3	82	69	225
Polymorphonucleocytes (proportion)	0.0073	0.0017, 0.048	ConC1vs3, ConC2vs3	82	69	225
ARDS classification	0.24			89	74	231
APACHE	0.066	0.029, 0.016	ConC1vs2, ConC1vs3	46	40	93
SOFA-Total	9.7E-11	5.2e-07, 6.4e-13	ConC1vs2, ConC1vs3	88	67	200
SOFA-CVS	4.5E-09	5.3e-06, 4.6e-11	ConC1vs2, ConC1vs3	89	73	228
SOFA-Haem	8.1E-05	5.5e-05, 3.4e-05	ConC1vs2, ConC1vs3	89	73	226
SOFA-Hep	0.36			89	70	212
SOFA-Renal	2.8E-08	3.1e-04, 3.1e-10, 3.9e-02	ConC1vs2, ConC1vs3, ConC2vs3	89	73	228
SOFA-Resp	0.35			88	70	214
SOFA-Neuro	0.21			89	73	228

Table C.6: Comparison of categorical clinical variables between predicted clusters, in validation cohort. Only the first samples per patient are used.

Clinical variable	χ^2 FDR	N(patients) with information		
		ConC1	ConC2	ConC3
Sex	0.29	89	74	231
Diagnosis	2.0E-07	88	74	230
Timepoint	0.032	89	74	231
Shock	9.2E-05	89	73	225
Mechanical ventilation/CPAP	0.032	89	73	228
Acute renal failure	6.2E-06	89	73	228
Comorbidities: Heart/vascular	0.38	89	74	231
Comorbidities: Respiratory	0.032	89	74	231
Comorbidities: Malignancy or immune disease	0.032	89	74	231
Comorbidities: Diabetes/endocrine disorders	0.64	89	74	231
Smoking habit	0.16	86	70	223
Corticosteroids	0.16	89	74	231
Activated Protein C	0.18	89	74	231
Microbiology	0.64	16	20	56
S. pneumoniae	0.29	33	49	157
Influenza	0.64	33	49	157

Table C.10: Summary of baseline characteristics and outcome data of 44 VANISH patients. All durations were counted within the shortest one among 28 days post-randomisation, or time to mortality, or length of hospital stay. [1]Recent surgery is defined as admitted to ICU following surgery. [2]Kidney failure is defined as having acute kidney injury stage 3; other organ failures defined as having a SOFA score of 3 or more. [3]28-day survivors as a proportion of patients with no kidney failure at baseline. [4]Other patients = those who had kidney failure, died, or both at any time. [5]Inotropes defined as dobutamine, epinephrine, milrinone, dopamine, dopexamine. Abbreviations: APACHE, Acute Physiology and Chronic Health Evaluation; BMI, body mass index; COPD, chronic obstructive pulmonary disease; IQR, interquartile range; PaO_2/FiO_2, arterial oxygen partial pressure to fractional inspired oxygen.

	Displayed as	Level	N
Baseline Characteristics			
Age	median (IQR)	64.5 (53.8-77.2)	44
Men	No./total (%)	32/44 (73)	44
Weight	median (IQR)	70 (60-85.2)	44
BMI	median (IQR)	23 (20.8-30.6)	44
Caucasian ethnicity	No./total (%)	34/44 (77)	44
Recent surgery[1]	No./total (%)	2/44 (5)	44
APACHE II score at baseline	median (IQR)	28 (24-32)	43
Pre-existing conditions			
- Ischemic heart disease	No./total (%)	5/44 (11)	44
- Severe COPD	No./total (%)	3/44 (7)	44

Table C.10 continued from previous page

	Displayed as	Level	N
- Cirrhosis	No./total (%)	5/44 (11)	44
- Cancer	No./total (%)	6/44 (14)	44
- immunocompromised	No./total (%)	4/44 (9)	44
- diabetes	No./total (%)	7/44 (16)	44
Organ failure[2]			
- Respiratory	No./total (%)	26/43 (60)	43
- Renal	No./total (%)	9/44 (20)	44
- Liver	No./total (%)	4/40 (10)	40
- Haematological	No./total (%)	4/43 (9)	43
- Neurological	No./total (%)	22/41 (54)	41
Physiological variables			
- Mean arterial pressure, mmHg	median (IQR)	66 (59.8-72.2)	44
- Heart rate, beats/min	median (IQR)	93 (79.8-110)	44
- Central venous pressure, mmHg	median (IQR)	13 (10-17.5)	26
- Lactate, mmol/L	median (IQR)	2.3 (1.5-3.8)	44
- PaO_2/FiO_2 (kPa)	median (IQR)	21.5 (15-31.9)	43
- Creatinine, μmol/L	median (IQR)	119 (80.2-193.2)	44
- Bilirubin, μmol/L	median (IQR)	13 (8.5-31.2)	40
- Platelets, $x10^3/\mu L$	median (IQR)	199 (114-255)	43
- Glasgow coma score	median (IQR)	7 (3-15)	41
Mechnical ventilation	No./total (%)	35/44 (80)	44
Renal replacement	No./total (%)	1/44 (2)	44
Volume of intravenous fluid in previous 4h, mL	median (IQR)	1101 (530.2-1596.2)	44
Patients receiving before randomization	No./total (%)	40/44 (91)	44
Noradrenaline dose at randomisation	median (IQR)	0.2 (0.1-0.3)	44
Outcome Data			
28-day survivors who never developed renal failure[3]	No./total (%)	18/35 (51)	35
Renal failure-free days in other patients[4]	median (IQR)	11.5 (0.2-24)	26
28-day mortality	No./total (%)	15/44 (34)	44
Hospital mortality	No./total (%)	15/44 (34)	44
ICU mortality	No./total (%)	11/44 (25)	44
Renal failure	No./total (%)	24/44 (55)	44
- survivors	No./total (%)	11/29 (38)	29
- nonsurvivors	No./total (%)	13/15 (87)	15
Duration of renal failure, days	median (IQR)	4 (2-7.5)	24
- survivors	median (IQR)	3 (2-12)	11
- nonsurvivors	median (IQR)	4 (2-7)	13
Weaned from vasopressors for >24h	No./total (%)	40/44 (91)	44
Time to shock reversal, hours	median (IQR)	56.8 (33.6-92.9)	40
Use of inotropes[5]	No./total (%)	5/44 (11)	44
Duration of mechanical ventilation, days	median (IQR)	6.5 (3-17)	38
Mean SOFA-total score over ICU stay	median (IQR)	4.8 (3.8-7.9)	44
Length of ICU stay, days	median (IQR)	9.5 (4.8-18.5)	44

Table C.10 continued from previous page

	Displayed as	Level	N
Length of hospital stay, days	median (IQR)	22.5 (14.2-46.5)	44

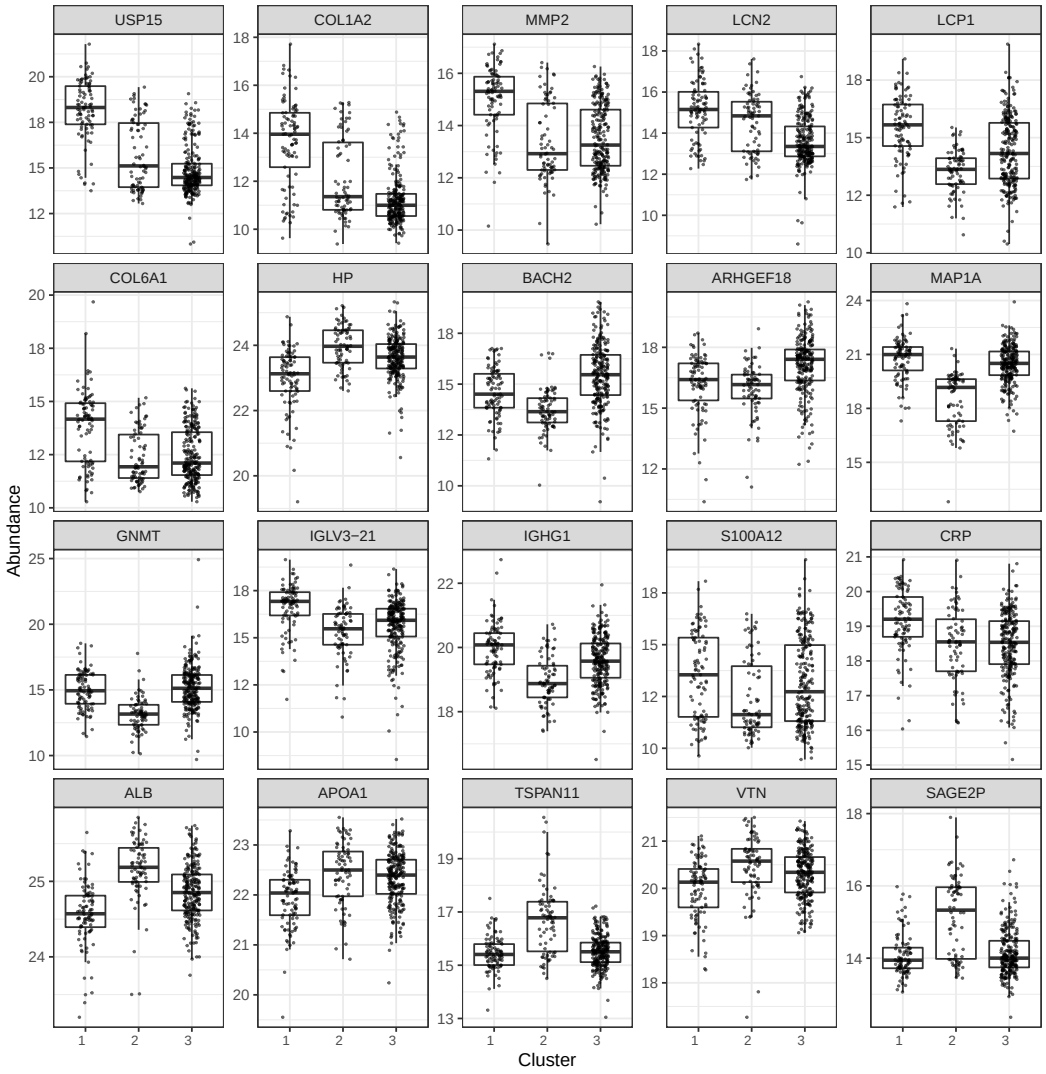
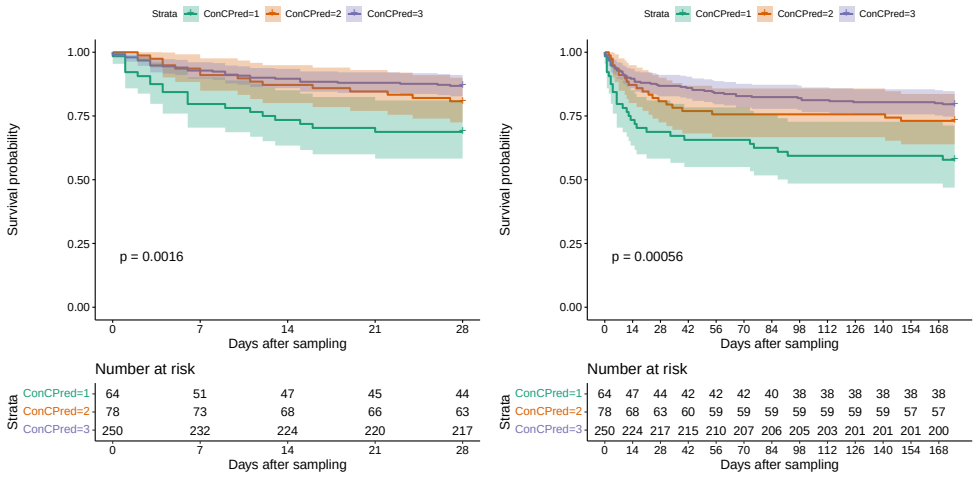
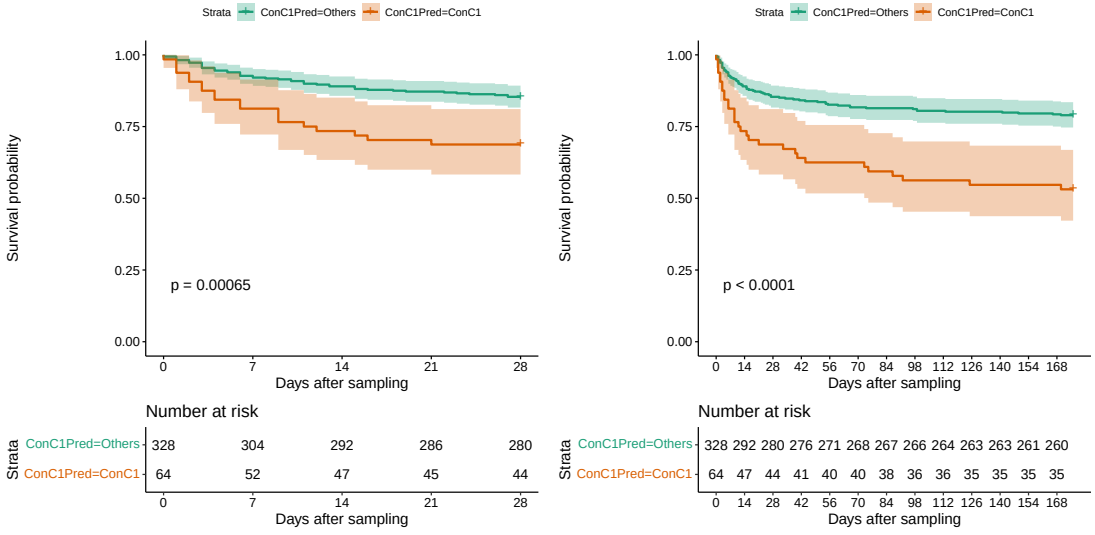


Figure C.9: Boxplots of protein abundance across the three predicted clusters, in validation cohort. Proteins described in the text are plotted.



(a) 8-protein 3-cluster model, 28-day mortality

(b) 8-protein 3-cluster model, 6-month mortality

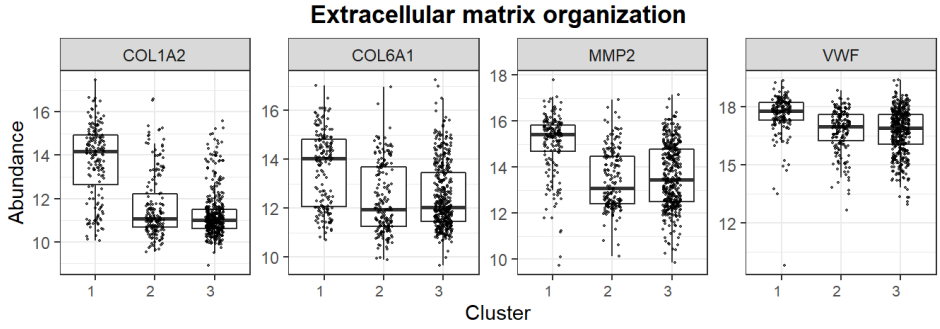


(c) 10-protein 2-cluster model, 28-day mortality

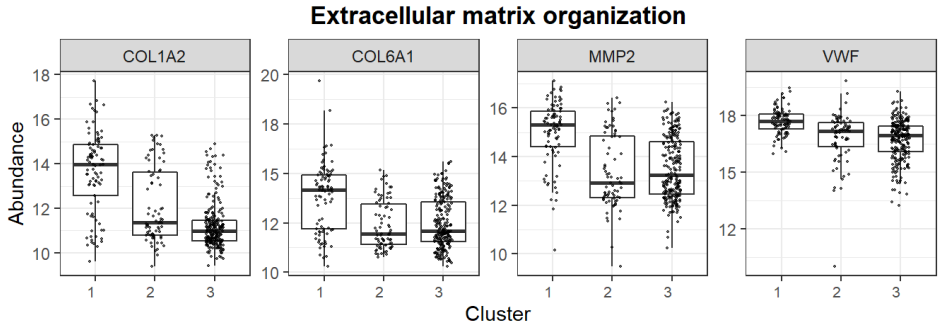
(d) 10-protein 2-cluster model, 6-month mortality

Figure C.10: Kaplan-Meier curves for the 3-cluster or 2-cluster minimal models. Patient stratification is evaluated in the validation cohort, using clusters predicted from the last available sample of each patient. Log-rank p values are labelled. ConCPred – predicted ConC cluster.

(a) In MS2019 discovery cohort



(b) In MS2019 validation cohort



(c) In MS192

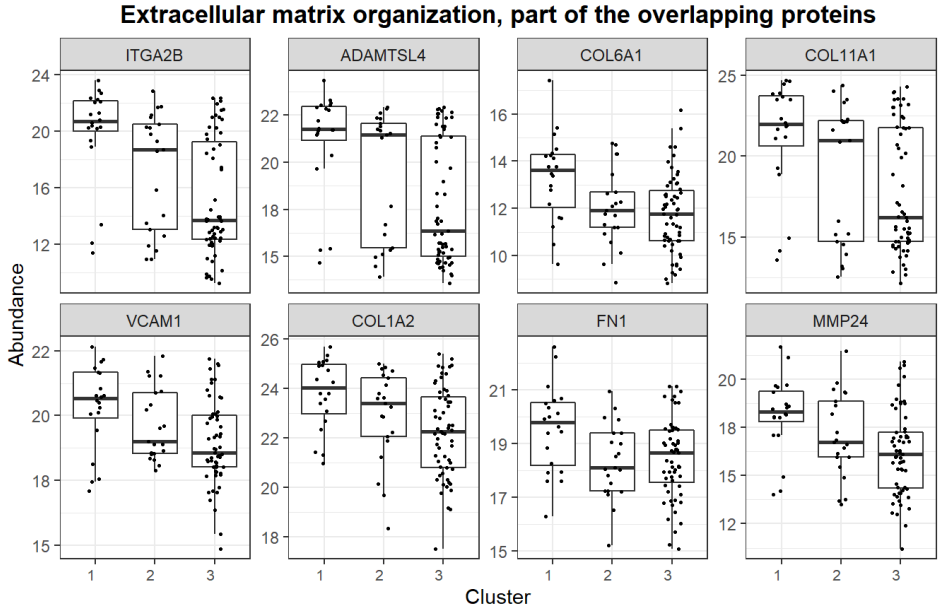
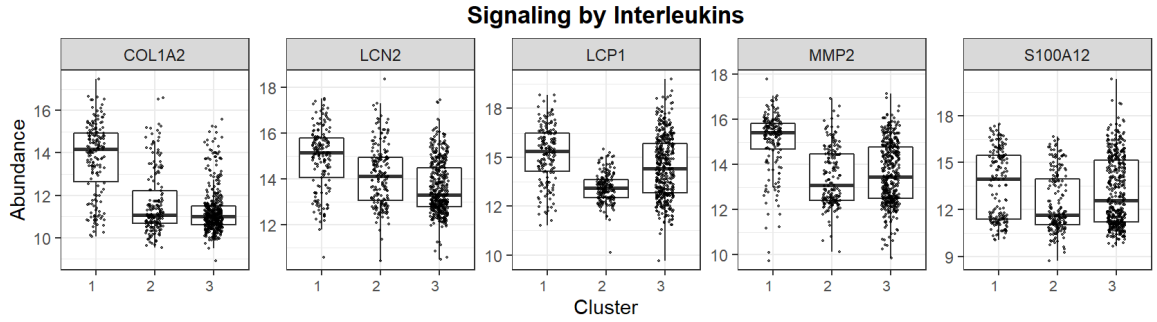
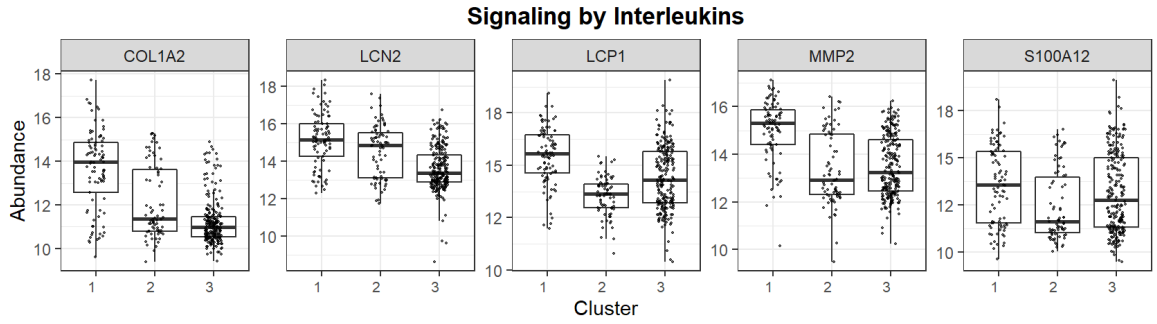


Figure C.11: Distribution of proteins annotated for **ECM organisation** cross the ConC clusters in each of the dataset. Only first samples of each patient were plotted.

(a) In MS2019 discovery cohort



(b) In MS2019 validation cohort



(c) In MS192

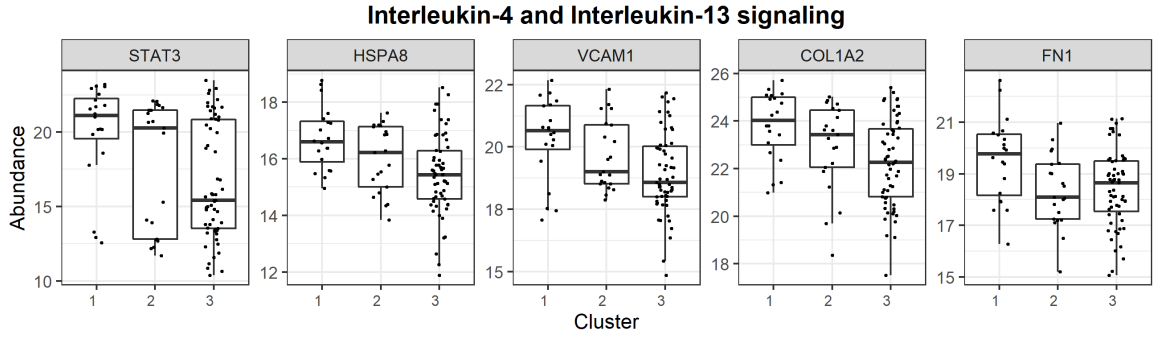


Figure C.12: Distribution of proteins annotated for **interleukin signalling** across the ConC clusters in each of the dataset. Only first samples of each patient were plotted.

Table C.7: List of the top 12 proteins removed in the MS192 dataset by the Pierce^{TM} Top 12 Abundant Protein Depletion Spin Columns.

From product information	Gene name	UniProt accession
α1-Acid Glycoprotein	ORM1, ORM2	P02763, P19652
α1-Antitrypsin	SERPINA1	P01009
α2-Macroglubulin	A2M	P01023
Albumin	ALB	P02768
Apolipoprotein A-I	APOA1	P02647
Apolipoprotein A-II	APOA2	P02652
Fibrinogen	FGA, FGB, FGG	P02671, P02675, P02679
Haptoglobin	HP	P00738
IgA	IGHA1, IGHA2	P01876, P01877
IgG	IGHG1, IGHG2, IGHG3, IGHG4	P01857, P01859, P01860, P01861
IgM	IGHM	P01871
Transferrin	LTFTF	P02787

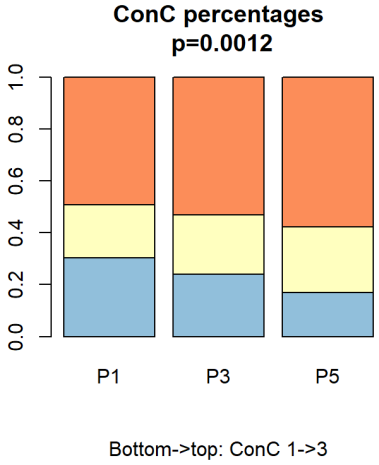


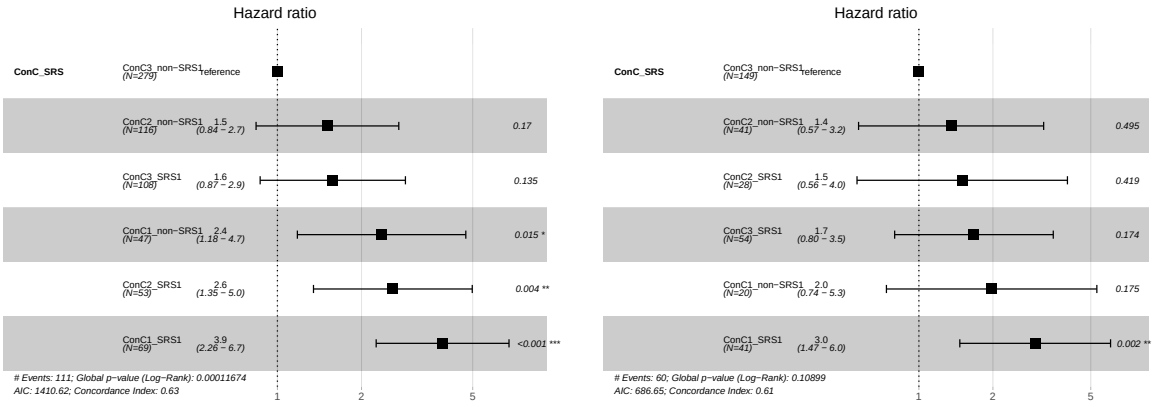
Figure C.13: Proportion of the three ConC clusters in plasma samples taken at day 1/3/5. GAinS discovery and validation cohorts were included. P values were given by a χ^2 tests.

Table C.8: Evidence for each entry in the summary graph Fig.4.21 of molecular characteristics of the clusters. Pathway enrichment refers to enrichment of proteins with either significantly higher or lower abundance in each contrast, using GOBP or Reactome annotations. Directions indicated were verified by visual examinations of the distributions across the clusters for the overlapping DA proteins. [1]This is a general term including phagocytosis, Fc-γ/ϵ receptor signalling, leukocyte migration, defense response to bacterium, and complement activation of classical pathway. The overlapping proteins annotated for these pathways as well as for "B cell signalling" were mainly constituted of immunoglobulins including the constant and variable regions. Abbreviations: disc – MS2019 discovery cohort; vali – MS2019 validation cohort; HV – healthy volunteer; DA - differentially abundant.

Term listed	Cluster	Direction	Evidence from		
			Dataset	Contrast	Approach
USP15, COL1A2	ConC1	high	disc, vali	ConC1vs3, ConC1vs2	Top DA proteins
MAP1A, GNMT	ConC2	low	disc, vali	ConC2vs3, ConC1vs2	Top DA proteins
TSPAN11	ConC2	high	disc, vali	ConC2vs3, ConC1vs2	Top DA proteins
Cytokines and chemokines	ConC1	high	Luminex	ConC1vs3	Individual DA proteins
	ConC3	low	Luminex	ConC1vs3	Individual DA proteins
Regulation of immune response[1]	ConC1	high	disc, vali, MS192	ConC1vs2, ConC1vs3	Pathway enrichment
Immunoglobulins	ConC1	high	disc, vali	ConC1vs2	Individual DA proteins
	ConC2	low	disc, vali	ConC1vs2	Individual DA proteins
B cell signalling	ConC1	high	disc, vali	ConC1vs2	Pathway enrichment
			disc	ConC1vsHV	Pathway enrichment
			Luminex	ConC1vs3	Individual protein functions
	ConC2	low	disc, vali	ConC1vs2	Pathway enrichment
Interleukin signalling	ConC1	high	disc, vali	ConC1vs2, ConC1vs3	Pathway enrichment
			MS192	ConC1vs3	Pathway enrichment
	ConC3	low	MS192	ConC1vs3	Pathway enrichment
ECM organisation	ConC1	high	disc, vali	ConC1vs2, ConC1vs3	Pathway enrichment
			MS192	ConC1vs3	Pathway enrichment
	ConC3	low	MS192	ConC1vs3	Pathway enrichment
Collagen fibril organisation	ConC1	high	MS192	ConC1vs3	Pathway enrichment
	ConC3	low	MS192	ConC1vs3	Pathway enrichment
Cell division	ConC1	high	MS192	ConC1vs3	Pathway enrichment
	ConC3	low	MS192	ConC1vs3	Pathway enrichment
Lipoprotein metabolic process	ConC1	low	disc, vali	ConC1vs2	Pathway enrichment
			MS192	ConC1vs3	Pathway enrichment
	ConC2	high	disc, vali	ConC1vs2	Pathway enrichment
	ConC3	high	MS192	ConC1vs3	Pathway enrichment
Transport of small molecules	ConC1	low	disc	ConC1vs2	Pathway enrichment
			MS192	ConC1vs3	Pathway enrichment
	ConC2	high	disc	ConC1vs2	Pathway enrichment
	ConC3	high	MS192	ConC1vs3	Pathway enrichment

Table C.9: Comparison of USP15 gene expression in leukocytes. A positive $\log_2$FC indicates higher in ConC1 or SRS1. $\log_2(1.5)=0.585$. $\log_2(1.2)=0.263$.

Groups compared	GE measured by	FDR	$\log_2$FC
ConC1 vs ConC2/3	Microarray	0.64	-0.043
	RNA-seq	0.14	0.093
SRS1 vs non-SRS1	Microarray	0.0011	0.18
	RNA-seq	6.5e-30	0.47



(a) In discovery cohort **(b)** In validation cohort

Figure C.14: Univariate Cox proportional hazard regression on 28-day mortality on combinations of ConC and SRS classifications. Model is built on 672 or 333 patients with both ConC and SRS assignments available in the discovery or validation cohort, respectively. Cluster assignments of the last available samples per patient were used. Hazards at 28-day post sampling were compared.

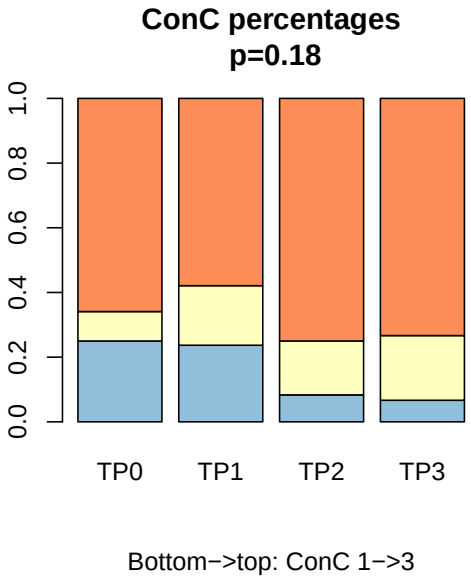


Figure C.15: Percentages of ConC membership in VANISH across the timepoints. Lengths of the colour bars from bottom to top represents the percentages within each timepoint of ConC1/2/3, respectively. The value is given by a χ^2 test.

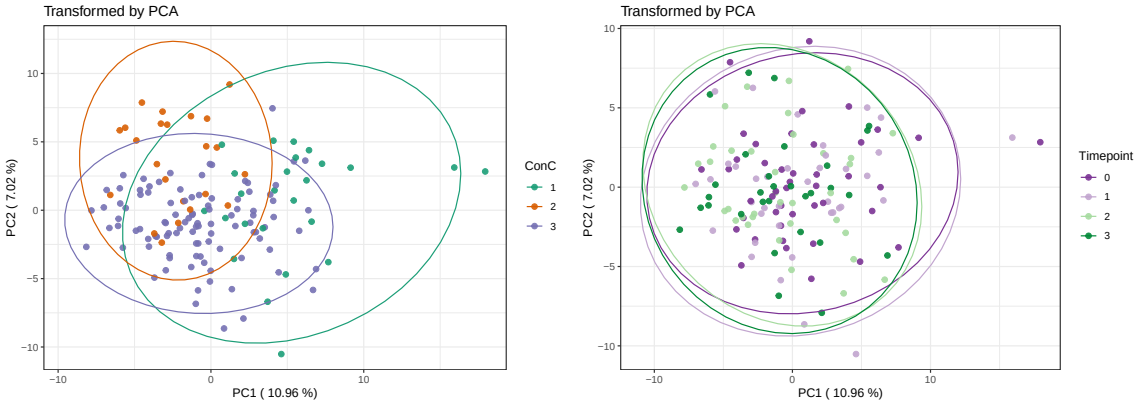


Figure C.16: Distribution of VANISH samples based on protein profiles, after PCA transformation. Separation between data ellipses can be observed between ConC clusters but not between the timepoints.

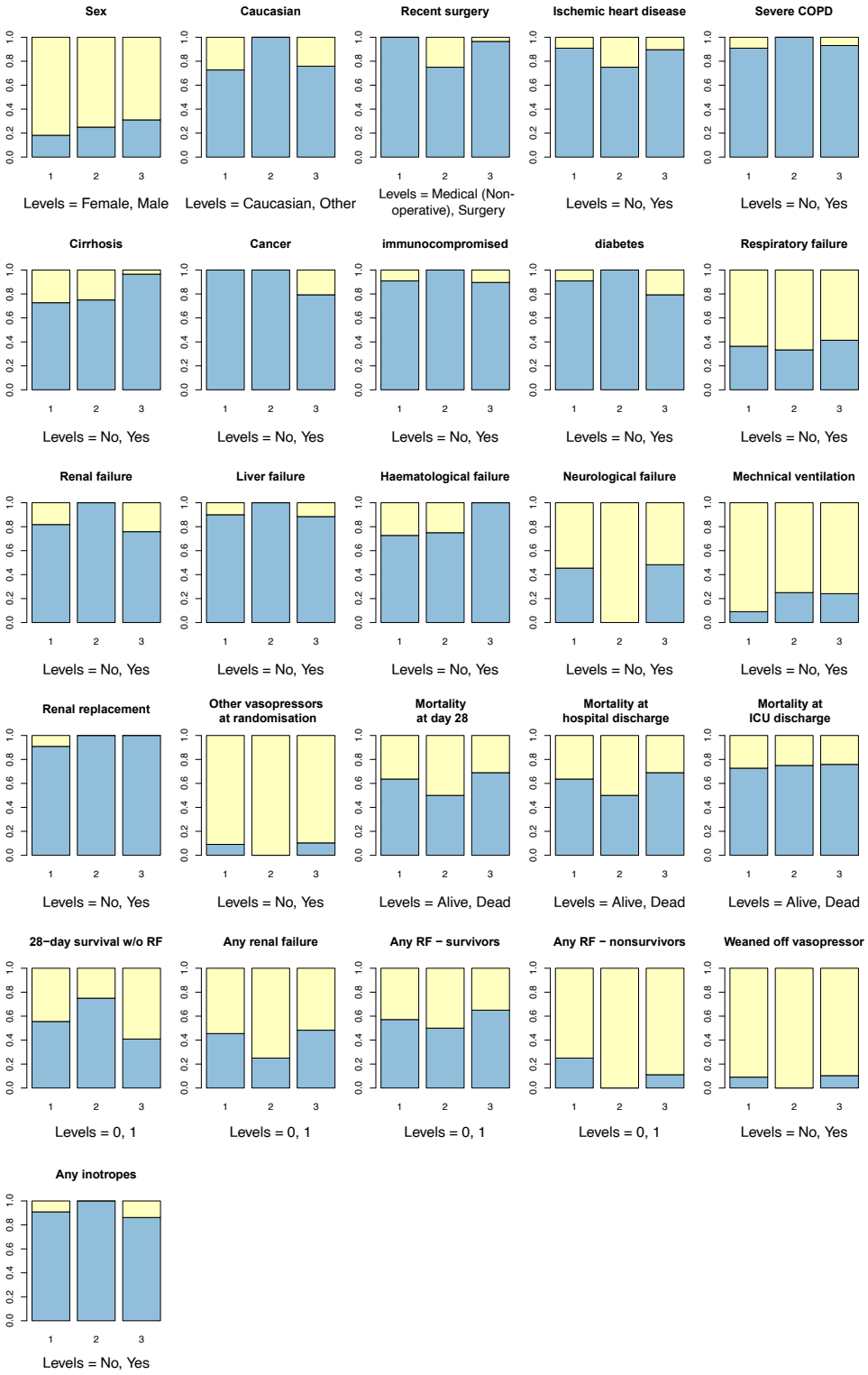


Figure C.17: Comparison of categorical clinical variables between the ConC clusters in VANISH, using baseline measurements (where applicable) and baseline cluster assignments. Length of the vertical bars represent the proportion of the corresponding level of the variable. Patient numbers were at most 11/4/29 for ConC1/2/3 respectively, subject to availability. For each variable, colours from the bottom to the top of the bars are in the order of the levels stated. Levels 0/1 represent the absence/presence of the event stated, respectively. The clinical variables should be interpreted the same as detailed in Table C.10. RF – renal failure.

D

Appendix to Chapter 5

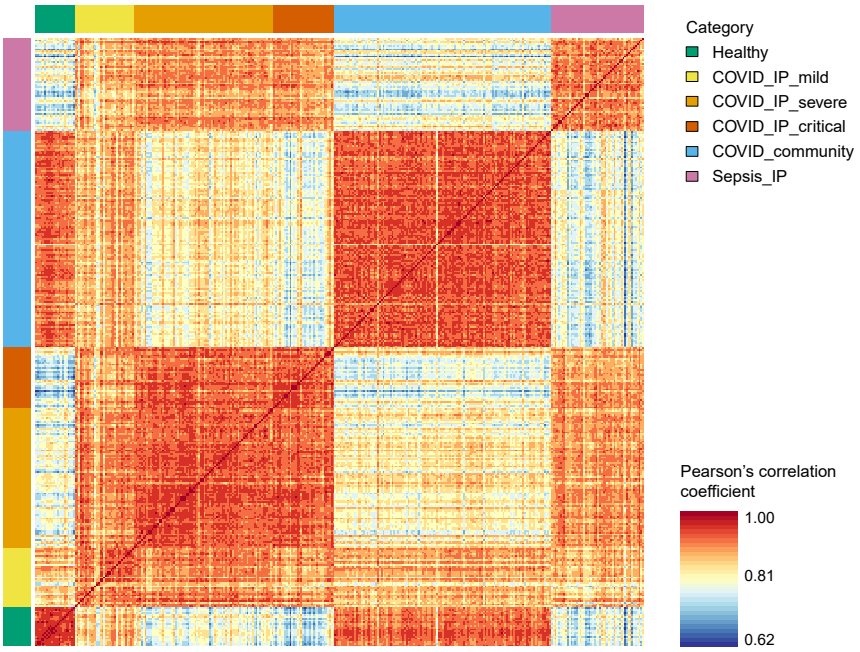


Figure D.1: Correlations between COMBAT samples. Sample are ordered by categories.

Table D.1: The panel of 51 cytokines measured by Luminex in COMBAT and the limits of detection (LODs).
The 47 cytokines with "TRUE" in "Keep" were used in my analysis. The group of "Healthy controls (serum)" was only included for quality control by the team in Luminex data generation but not in any downstream analysis.

Gene name	Uniprot accession	ULOD (pg/mL)	LLOD (pg/mL)	<LLOD	>ULOD	Keep	Percentage <LLOD or >ULOD in group									
							HV (plasma)	COVID mild	COVID severe	COVID critical	COVID critical SGUL (serum)	COVID HCW	Flu SGUL (serum)	Sepsis	Plate controls	Healthy controls (serum)
CCL18	P55774	216000	99.0	0.0%	8.0%	TRUE	0.0%	10.3%	4.5%	11.4%	16.7%	0.0%	9.1%	14.8%	0.0%	0.0%
LTF	P02788	4996500	2284.5	0.0%	2.0%	TRUE	0.0%	0.0%	1.5%	0.0%	9.5%	0.0%	4.5%	1.9%	0.0%	0.0%
LCN2	P80188	1568000	717.0	0.0%	0.0%	TRUE	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
MPO	P05164	1485500	679.0	0.0%	0.3%	TRUE	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.9%	0.0%	0.0%
CCL2	P13500	36350	16.6	0.0%	0.0%	TRUE	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
CCL3	P10147	112100	51.3	77.1%	0.0%	TRUE	85.3%	82.1%	74.6%	62.9%	73.8%	81.8%	77.3%	75.9%	83.3%	87.5%
CCL4	P13236	145650	66.6	24.1%	0.0%	TRUE	38.2%	33.3%	19.4%	22.9%	14.3%	36.4%	4.5%	18.5%	55.6%	12.5%
CCL17	Q92583	150100	68.6	67.0%	0.0%	TRUE	61.8%	64.1%	68.7%	82.9%	42.9%	81.8%	68.2%	66.7%	66.7%	87.5%
CXCL1	P09341	63550	29.1	62.8%	0.0%	TRUE	70.6%	56.4%	73.1%	80.0%	21.4%	77.3%	50.0%	66.7%	50.0%	87.5%
CXCL5	P42830	59100	27.0	31.5%	0.0%	TRUE	0.0%	51.3%	44.8%	65.7%	0.0%	68.2%	9.1%	35.2%	0.0%	6.3%
IFNB1	P01574	15000	6.9	96.0%	0.0%	TRUE	97.1%	92.3%	95.5%	94.3%	97.6%	95.5%	100.0%	98.1%	88.9%	100.0%
IFNG	P01579	53400	24.4	94.6%	0.0%	TRUE	100.0%	97.4%	94.0%	94.3%	95.2%	95.5%	95.5%	87.0%	94.4%	100.0%
IL1A	P01583	5800	2.7	79.9%	0.0%	TRUE	88.2%	74.4%	77.6%	85.7%	78.6%	77.3%	86.4%	74.1%	72.2%	100.0%
IL1B	P01584	19150	8.8	98.9%	0.0%	FALSE	97.1%	94.9%	100.0%	100.0%	100.0%	100.0%	100.0%	98.1%	100.0%	100.0%
IL3	P08700	110300	50.4	96.3%	0.0%	FALSE	94.1%	94.9%	97.0%	100.0%	95.2%	95.5%	100.0%	94.4%	94.4%	100.0%
IL6	P05231	5150	2.4	28.1%	1.4%	TRUE	94.1%	17.9%	4.5%	8.6%	0.0%	90.9%	4.5%	7.4%	94.4%	100.0%
IL10	P22301	4900	2.2	71.9%	0.0%	TRUE	94.1%	76.9%	67.2%	45.7%	69.0%	95.5%	50.0%	61.1%	100.0%	100.0%
IL13	P35225	436350	199.5	77.4%	0.0%	TRUE	91.2%	76.9%	82.1%	82.9%	59.5%	90.9%	54.5%	70.4%	88.9%	87.5%
IL17A	Q16552	17150	7.8	88.3%	0.0%	TRUE	94.1%	92.3%	95.5%	97.1%	88.1%	95.5%	77.3%	63.0%	94.4%	100.0%
LTA	P01374	8050	3.7	97.1%	0.0%	TRUE	97.1%	100.0%	97.0%	85.7%	100.0%	100.0%	100.0%	96.3%	100.0%	100.0%
CSF1	P09603	117100	53.5	48.1%	0.0%	TRUE	94.1%	25.6%	34.3%	17.1%	40.5%	100.0%	27.3%	33.3%	100.0%	100.0%
NGF	P01138	4600	2.1	31.2%	0.0%	TRUE	73.5%	28.2%	11.9%	8.6%	14.3%	77.3%	9.1%	14.8%	72.2%	100.0%
S100A9	P06702	29950	13.7	0.0%	4.3%	TRUE	0.0%	0.0%	3.0%	11.4%	7.1%	0.0%	13.6%	5.6%	0.0%	0.0%
CLEC11A	Q9Y240	393450	179.9	0.6%	0.0%	TRUE	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.9%	0.0%	6.3%
TFPI	P10646	376000	171.9	0.0%	0.0%	TRUE	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TGFA	P01135	8400	3.8	70.5%	0.0%	TRUE	97.1%	92.3%	86.6%	62.9%	4.8%	95.5%	36.4%	75.9%	100.0%	43.8%
THPO	P40225	443300	202.7	38.1%	0.0%	TRUE	67.6%	33.3%	28.4%	28.6%	16.7%	63.6%	22.7%	37.0%	77.8%	50.0%
TNF	P01375	9300	4.3	50.1%	0.0%	TRUE	97.1%	51.3%	44.8%	11.4%	21.4%	100.0%	18.2%	37.0%	100.0%	93.8%
CCL11	P51671	81300	37.2	47.6%	0.0%	TRUE	44.1%	51.3%	38.8%	48.6%	47.6%	63.6%	59.1%	57.4%	11.1%	50.0%
CCL19	Q99731	11050	5.1	0.0%	0.3%	TRUE	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.9%	0.0%	0.0%
CCL20	P78556	10850	5.0	28.4%	0.9%	TRUE	76.5%	33.3%	16.4%	0.0%	0.0%	68.2%	9.1%	7.4%	94.4%	87.5%
CD40LG	P29965	253350	115.9	22.1%	0.0%	TRUE	32.4%	25.6%	17.9%	8.6%	0.0%	31.8%	13.6%	38.9%	38.9%	18.8%
CD163	Q86VB7	6682500	3055.6	0.0%	0.9%	TRUE	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.5%	3.7%	0.0%	0.0%
F3	P13726	7450	3.4	0.0%	0.0%	TRUE	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
C5	P01031	919450	420.4	2.6%	0.0%	TRUE	0.0%	0.0%	0.0%	0.0%	0.0%	9.1%	0.0%	3.7%	27.8%	0.0%

Table D.1 continued from previous page

CXCL10	P02778	1250	0.6	0.0%	17.5%	TRUE	0.0%	5.1%	32.8%	57.1%	9.5%	0.0%	18.2%	16.7%	0.0%	0.0%
EGF	P01133	11450	5.3	55.9%	0.0%	TRUE	26.5%	84.6%	71.6%	74.3%	0.0%	59.1%	50.0%	79.6%	50.0%	18.8%
FGF2	P09038	2500	1.2	61.0%	0.0%	TRUE	44.1%	69.2%	62.7%	62.9%	33.3%	95.5%	72.7%	42.6%	100.0%	93.8%
CSF3	P09919	28950	13.3	56.4%	0.9%	TRUE	82.4%	69.2%	59.7%	48.6%	42.9%	86.4%	36.4%	25.9%	94.4%	75.0%
CSF2	P04141	15000	6.9	54.4%	0.0%	TRUE	100.0%	48.7%	26.9%	5.7%	66.7%	95.5%	36.4%	50.0%	100.0%	93.8%
GZMB	P10144	13650	6.3	43.6%	0.0%	TRUE	79.4%	48.7%	14.9%	8.6%	31.0%	77.3%	31.8%	40.7%	100.0%	100.0%
IFNA1	P01562	13250	6.1	92.8%	0.0%	TRUE	100.0%	87.2%	83.6%	88.6%	100.0%	100.0%	81.8%	98.1%	100.0%	100.0%
IL2	P60568	37200	17.0	96.0%	0.0%	TRUE	97.1%	100.0%	100.0%	91.4%	92.9%	100.0%	86.4%	92.6%	100.0%	100.0%
IL5	P05113	9650	4.4	86.2%	0.0%	TRUE	91.2%	87.2%	77.6%	80.0%	92.9%	86.4%	90.9%	85.2%	88.9%	100.0%
CXCL8	P10145	4650	2.2	22.1%	0.6%	TRUE	47.1%	35.9%	16.4%	2.9%	0.0%	95.5%	4.5%	3.7%	55.6%	18.8%
IL12A	P29459	179950	82.3	98.3%	0.0%	FALSE	100.0%	97.4%	98.5%	97.1%	95.2%	100.0%	100.0%	98.1%	100.0%	100.0%
IL15	P40933	7700	3.5	38.7%	0.0%	TRUE	88.2%	43.6%	31.3%	11.4%	14.3%	86.4%	9.1%	13.0%	77.8%	93.8%
IL23A	Q9NPF7	173900	79.5	69.9%	0.0%	TRUE	97.1%	64.1%	55.2%	37.1%	81.0%	90.9%	63.6%	68.5%	94.4%	87.5%
IL33	O95760	14050	6.4	96.0%	0.0%	FALSE	100.0%	97.4%	97.0%	97.1%	92.9%	95.5%	90.9%	92.6%	100.0%	100.0%
OSM	P13725	449250	205.4	79.7%	0.0%	TRUE	88.2%	79.5%	76.1%	68.6%	88.1%	77.3%	81.8%	77.8%	83.3%	81.3%
TREM1	Q9NP99	114350	52.3	6.6%	0.0%	TRUE	8.8%	5.1%	1.5%	0.0%	0.0%	31.8%	0.0%	3.7%	22.2%	25.0%

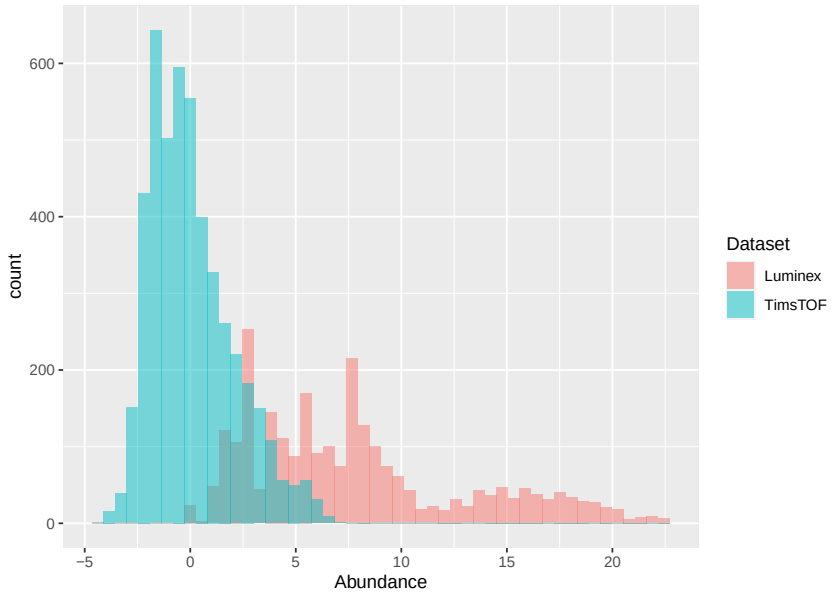


Figure D.2: TimsTOF and Luminex data for COVID-19 and flu serum samples, not quantile-normalised. Histogram of all protein values after pre-processing, measured by either the TimsTOF or Luminex platform. Luminex data are $\log_2$ transformed.

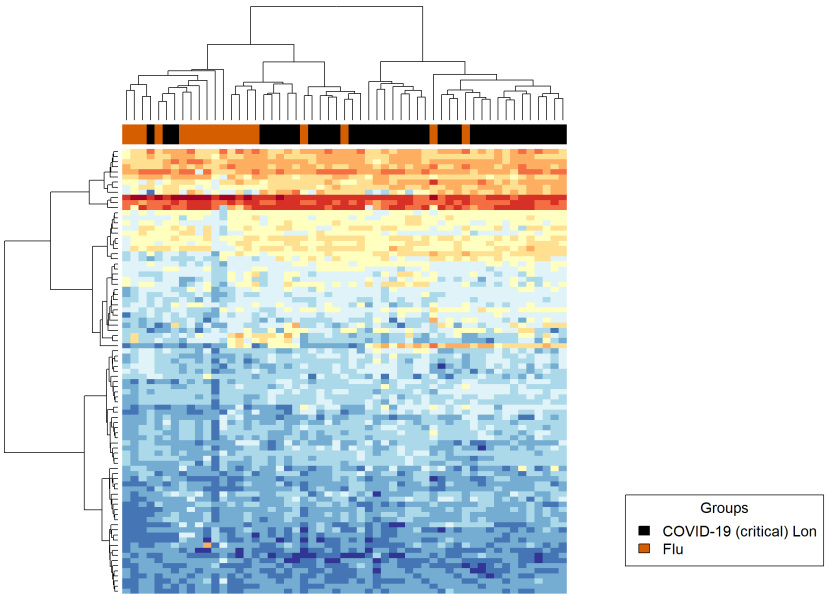


Figure D.3: Heatmap and hierarchical clustering in serum samples using TimsTOF data only.

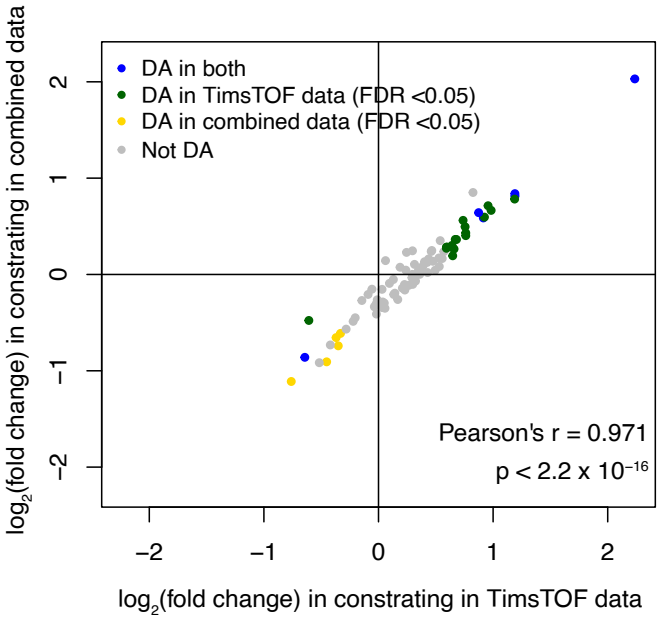


Figure D.4: Correlation of fold changes obtained in performing the COVID-flu contrast using only TimsTOF data or the TimsTOF-Luminex combined data. DA = differentially abundant.

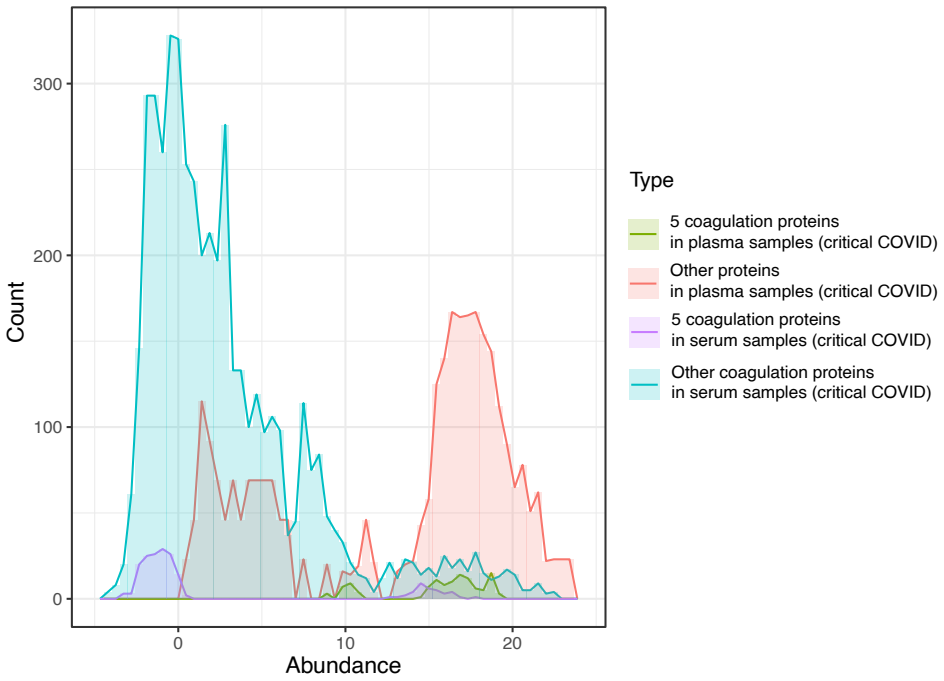


Figure D.5: Histograms of protein abundance in plasma and serum samples from critically ill COVID-19 patients, differentiating between five proteins involved in regulating coagulation (PROS1, SERPINF2, SERPINC1, PLG, and TFPI) and the other proteins in each dataset.

Bibliography

Abe, T., AlSarhan, M., Benakanakere, M. R., et al. (2015). "The B Cell–Stimulatory Cytokines BLyS and APRIL Are Elevated in Human Periodontitis and Are Required for B Cell–Dependent Bone Loss in Experimental Murine Periodontitis". *The Journal of Immunology* 195: pp. 1427–1435.

Ackermann, M., Verleden, S. E., Kuehnel, M., et al. (2020). "Pulmonary Vascular Endothelialitis, Thrombosis, and Angiogenesis in Covid-19". *New England Journal of Medicine* 383: pp. 120–128.

Aebersold, R. and Mann, M. (2016). "Mass-spectrometric exploration of proteome structure and function". *Nature 2016 537:7620* 537: pp. 347–355.

Albrechtsen, N. J. W., Geyer, P. E., Doll, S., et al. (2018). "Plasma Proteome Profiling Reveals Dynamics of Inflammatory and Lipid Homeostasis Markers after Roux-En-Y Gastric Bypass Surgery". *Cell Systems* 7: pp. 601–612.

Anderson, N. L. and Anderson, N. G. (2002). "The Human Plasma Proteome: History, Character, and Diagnostic Prospects". *Molecular and Cellular Proteomics* 1: pp. 845–867.

Anderson, N. L., Ptolemy, A. S., and Rifai, N. (2013). "The Riddle of Protein Diagnostics: Future Bleak or Bright?" *Clinical Chemistry* 59: pp. 194–197.

Angus, D. C., Marrie, T. J., Scott Obrosky, D., et al. (2002). "Severe community-acquired pneumonia: Use of intensive care services and evaluation of American and British Thoracic Society diagnostic criteria". *American Journal of Respiratory and Critical Care Medicine* 166: pp. 717–723.

Antcliffe, D., Jiménez, B., Veselkov, K., et al. (2017). "Metabolic Profiling in Patients with Pneumonia on Intensive Care". *EBioMedicine* 18: pp. 244–253.

Antcliffe, D. B., Burnham, K. L., Al-Beidh, F., et al. (2019). "Transcriptomic signatures in sepsis and a differential response to steroids from the VaNISH randomized trial". *American Journal of Respiratory and Critical Care Medicine* 199: pp. 980–986.

Antcliffe, D. B., Wolfer, A. M., O'Dea, K. P., et al. (2018). "Profiling inflammatory markers in patients with pneumonia on intensive care". *Scientific Reports* 8: pp. 1–9.

Aran, D., Hu, Z., and Butte, A. J. (2017). "xCell: Digitally portraying the tissue cellular heterogeneity landscape". *Genome Biology* 18: pp. 1–14.

Argelaguet, R., Arnol, D., Bredikhin, D., et al. (2020). "MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data". *Genome Biology* 21: pp. 1–17.

Arunachalam, P. S., Wimmers, F., Mok, C. K. P., et al. (2020). "Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans". *Science* 369: pp. 1210–1220.

Atarashi, K., Suda, W., Luo, C., et al. (2017). "Ectopic colonization of oral bacteria in the intestine drives T H 1 cell induction and inflammation". *Science (New York, N.Y.)* 358: pp. 359–365.

Atarashi, K., Tanoue, T., Ando, M., et al. (2015). "Th17 Cell Induction by Adhesion of Microbes to Intestinal Epithelial Cells". *Cell* 163: pp. 367–380.

Atarashi, K., Tanoue, T., Oshima, K., et al. (2013). "Treg induction by a rationally selected mixture of Clostridia strains from the human microbiota". *Nature* 500: pp. 232–236.

Baghela, A., Pena, O. M., Lee, A. H., et al. (2022). "Predicting sepsis severity at first clinical presentation: The role of endotypes and mechanistic signatures". *eBioMedicine* 75: p. 103776.

Bateman, A., Martin, M. J., Orchard, S., et al. (2021). "UniProt: the universal protein knowledgebase in 2021". *Nucleic Acids Research* 49: pp. D480–D489.

Beltrán-García, J., Osca-Verdegal, R., Pallardó, F. V., et al. (2020). "Sepsis and Coronavirus Disease 2019: Common Features and Anti-Inflammatory Therapeutic Approaches". *Critical Care Medicine* 48: pp. 1841–1844.

Bhatt, P. J., Shiau, S., Brunetti, L., et al. (2020). "Risk Factors and Outcomes of Hospitalized Patients With Severe Coronavirus Disease 2019 (COVID-19) and Secondary Bloodstream Infections: A Multicenter Case-Control Study". *Clinical Infectious Diseases* 2019: pp. 1–9.

Blaurock, N., Schmerler, D., Hünniger, K., et al. (2016). "C-terminal alpha-1 antitrypsin peptide: A new sepsis biomarker with immunomodulatory function". *Mediators of Inflammation* 2016.

Bone, R. C., Balk, R. A., Cerra, F. B., et al. (1992). "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis". *Chest* 101: pp. 1644–1655.

Brunson, J. C. (2020). "ggalluvial: Layered Grammar for Alluvial Plots". *Journal of Open Source Software* 5: p. 2017.

Burnap, S. A., Mayr, U., Shankar-Hari, M., et al. (2021). "A Proteomics-Based Assessment of Inflammation Signatures in Endotoxemia". *Mol Cell Proteomics* 20: p. 100021.

Burnham, K. L., Davenport, E. E., Radhakrishnan, J., et al. (2017). "Shared and Distinct Aspects of the Sepsis Transcriptomic Response to Fecal Peritonitis and Pneumonia". *American Journal of Respiratory and Critical Care Medicine* 196: pp. 328–339.

Cairns, D. A., Barrett, J. H., Billingham, L. J., et al. (2009). "Sample size determination in clinical proteomic profiling experiments using mass spectrometry for class comparison". *Proteomics* 9: pp. 74–86.

Calfee, C. S., Delucchi, K., Parsons, P. E., et al. (2014). "Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials". *The Lancet Respiratory Medicine* 2: pp. 611–620.

Calfee, C. S., Delucchi, K. L., Sinha, P., et al. (2018). "Acute respiratory distress syndrome subphenotypes and differential response to simvastatin: secondary analysis of a randomised controlled trial". *The Lancet Respiratory Medicine* 6: pp. 691–698.

Camicia, G., Pozner, R., and De Larrañaga, G. (2014). "Neutrophil extracellular traps in sepsis". *Shock (Augusta, Ga.)* 42: pp. 286–294.

Cano-Gamez, E. (2022). *SepstratifieR: Stratification of infectious disease patients by gene expression profile*. R package version 0.0.0.9000.

Cano-Gamez, E., Burnham, K. L., Goh, C., et al. (2022). "An immune dysfunction score for stratification of patients with acute infection based on whole blood gene expression". *medRxiv*: p. 2022.03.17.22272427.

Cevik, M., Kuppalli, K., Kindrachuk, J., et al. (2020). "Virology, transmission, and pathogenesis of SARS-CoV-2". *BMJ* 371.

Chakraborty, R. K. and Burns, B. (2021). "Systemic Inflammatory Response Syndrome". *StatPearls*.

Chang, K. C., Unsinger, J., Davis, C. G., et al. (2007). "Multiple triggers of cell death in sepsis: death receptor and mitochondrial-mediated apoptosis". *The FASEB Journal* 21: pp. 708–719.

Chawade, A., Alexandersson, E., and Levander, F. (2014). "Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets". *Journal of proteome research* 13: pp. 3114–3120.

Chella Krishnan, K., Kurt, Z., Barrere-Cain, R., et al. (2018). "Integration of Multi-omics Data from Mouse Diversity Panel Highlights Mitochondrial Dysfunction in Non-alcoholic Fatty Liver Disease". *Cell Systems* 6: 103–115.e7.

Cheng, S. C., Scicluna, B. P., Arts, R. J., et al. (2016). "Broad defects in the energy metabolism of leukocytes underlie immunoparalysis in sepsis". *Nature Immunology 2016 17:4* 17: pp. 406–413.

Chow, S.-C., Shao, J., Wang, H., et al. (2017). *Sample size calculations in clinical research*. chapman and hall/CRC.

Connors, J. M. and Levy, J. H. (2020). *COVID-19 and its implications for thrombosis and anticoagulation*. Tech. rep.

Coscia, F., Watters, K. M., Curtis, M., et al. (2016). "Integrative proteomic profiling of ovarian cancer cell lines reveals precursor cell associated proteins and functional status". *Nature Communications* 7: pp. 1–14.

COVID-19 Forecasting Team (2022). "Variation in the COVID-19 infection–fatality ratio by age, time, and geography during the pre-vaccine era: a systematic analysis". *The Lancet* 399: pp. 1469–1488.

COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium (2022). "A blood atlas of COVID-19 defines hallmarks of disease severity and specificity". *Cell* 185: 916–938.e58.

Cowburn, A. S., Sladek, K., Soja, J., et al. (1998). "Overexpression of leukotriene C4 synthase in bronchial biopsies from patients with aspirin-intolerant asthma". *The Journal of clinical investigation* 101: pp. 834–846.

Cuello, F., Shankar-Hari, M., Mayr, U., et al. (2014). "Redox State of Pentraxin 3 as a Novel Biomarker for Resolution of Inflammation and Survival in Sepsis". *Molecular and Cellular Proteomics* 13: pp. 2545–2557.

Dahlén, B., Nizankowska, E., Szczeklik, A., et al. (1998). "Benefits from adding the 5-lipoxygenase inhibitor zileuton to conventional therapy in aspirin-intolerant asthmatics". *American journal of respiratory and critical care medicine* 157: pp. 1187–1194.

Daix, T., Guerin, E., Tavernier, E., et al. (2018). "Multicentric Standardized Flow Cytometry Routine Assessment of Patients With Sepsis to Predict Clinical Worsening". *Chest* 154: pp. 617–627.

Davenport, E. E., Burnham, K. L., Radhakrishnan, J., et al. (2016). "Genomic landscape of the individual host response and outcomes in sepsis: A prospective cohort study". *The Lancet Respiratory Medicine* 4: pp. 259–271.

de Azambuja Rodrigues, P. M., Valente, R. H., Brunoro, G. V. F., et al. (2021). "Proteomics reveals disturbances in the immune response and energy metabolism of monocytes from patients with septic shock". *Scientific Reports 2021 11:1* 11: pp. 1–10.

De Coux, A., Tian, Y., Deleon-Pennell, K. Y., et al. (2015). "Plasma glycoproteomics reveals sepsis outcomes linked to distinct proteins in common pathways". *Critical Care Medicine* 43: pp. 2049–2058.

Demerle, K. M., Angus, D. C., Baillie, J. K., et al. (2021). "Sepsis Subclasses: A Framework for Development and Interpretation". *Critical Care Medicine* 49: pp. 748–759.

Demichev, V., Tober-Lau, P., Lemke, O., et al. (2021). "A time-resolved proteomic and prognostic map of COVID-19". *Cell Systems* 12: 780–794.e7.

Diao, B., Wang, C., Tan, Y., et al. (2020). "Reduction and Functional Exhaustion of T Cells in Patients With Coronavirus Disease 2019 (COVID-19)". *Frontiers in Immunology* 11: p. 827.

Dijk, D. van, Sharma, R., Nainys, J., et al. (2018). "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion". *Cell* 174: 716–729.e27.

Dolgachev, V., Panicker, S., Balijepalli, S., et al. (2018). "Electroporation-mediated delivery of FER gene enhances innate immune response and improves survival in a murine model of pneumonia". *Gene Therapy* 25: pp. 359–375.

Drewry, A., Samra, N., Skrupky, L., et al. (2014). "Persistent lymphopenia after diagnosis of sepsis predicts mortality". *Shock* 42: pp. 383–391.

Fabregat, A., Sidiropoulos, K., Viteri, G., et al. (2018). "Reactome diagram viewer: data structures and strategies to boost performance." *Bioinformatics (Oxford, England)* 34. Provided by Reactome. Citation Accessed on Fri Feb 11 2022: pp. 1208–1214.

Fang, H., Beckmann, G., Bountra, C., et al. (2019). "A genetics-led approach defines the drug target landscape of 30 immune-related traits". *Nature Genetics 2019 51:7* 51: pp. 1082–1091.

Fang, H., Knezevic, B., Burnham, K. L., et al. (2016). "XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits". *Genome Medicine* 8: p. 129.

Fang, H. and Knight, J. C. (2022). "Priority index: database of genetic targets in immune-mediated disease". *Nucleic Acids Research* 50: pp. D1358–D1367.

Ferguson, N. D., Fan, E., Camporota, L., et al. (2012). "The Berlin definition of ARDS: An expanded rationale, justification, and supplementary material". *Intensive Care Medicine* 38: pp. 1573–1582.

Filbin, M. R., Mehta, A., Schneider, A. M., et al. (2020). "Plasma proteomics reveals tissue-specific cell death and mediators of cell-cell interactions in severe COVID-19 patients". *bioRxiv*.

Fjell, C. D., Thair, S., Hsu, J. L., et al. (2013). "Cytokines and signaling molecules predict clinical outcomes in sepsis". *PLoS ONE* 8.

Fong, T. G., Chan, N. Y., Dillon, S. T., et al. (2021). "Identification of Plasma Proteome Signatures Associated With Surgery Using SOMAscan". *Annals of surgery* 273: pp. 732–742.

Fong, Y., Sebestyen, K., Yu, X., et al. (2013). "nCal: an R package for non-linear calibration". *Bioinformatics* 29: pp. 2653–2654.

Fresán, U., Guevara, M., Trobajo-Sanmartín, C., et al. (2021). "Hypertension and Related Comorbidities as Potential Risk Factors for COVID-19 Hospitalization and Severity: A Prospective Population-Based Cohort Study". *Journal of Clinical Medicine 2021, Vol. 10, Page 1194* 10: p. 1194.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". *Journal of statistical software* 33: p. 1.

Genga, K. R. and Russell, J. A. (2017). "Update of Sepsis in the Intensive Care Unit." *Journal of innate immunity* 9: pp. 441–455.

Georg, P., Astaburuaga-García, R., Bonaguro, L., et al. (2022). "Complement activation induces excessive T cell cytotoxicity in severe COVID-19". *Cell* 185: 493–512.e25.

Geyer, P. E., Holdt, L. M., Teupser, D., et al. (2017). "Revisiting biomarker discovery by plasma proteomics". *Molecular Systems Biology* 13: p. 942.

Geyer, P. E., Kulak, N. A., Pichler, G., et al. (2016a). "Plasma Proteome Profiling to Assess Human Health and Disease". *Cell Systems* 2: pp. 185–195.

Geyer, P. E., Voytik, E., Treit, P. V., et al. (2019). "Plasma Proteome Profiling to detect and avoid sample-related biases in biomarker studies". *EMBO Molecular Medicine* 11.

Geyer, P. E., Wewer Albrechtsen, N. J., Tyanova, S., et al. (2016b). "Proteomics reveals the effects of sustained weight loss on the human plasma proteome". *Molecular systems biology* 12: p. 901.

Ghoussaini, M., Mountjoy, E., Carmona, M., et al. (2021). "Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics". *Nucleic Acids Research* 49: pp. D1311–D1320.

Giamarellos-Bourboulis, E. J., Netea, M. G., Rovina, N., et al. (2020). "Complex Immune Dysregulation in COVID-19 Patients with Severe Respiratory Failure". *Cell Host and Microbe* 27: 992–1000.e3.

Gill, J. R., Sheng, Z. M., Ely, S. F., et al. (2010). "Pulmonary pathologic findings of fatal 2009 pandemic influenza A/H1N1 viral infections". *Archives of pathology and laboratory medicine* 134: pp. 235–243.

Goh, C., Golubchik, T., Ansari, M. A., et al. (2019). "Targeted metagenomic sequencing enhances the identification of pathogens associated with acute infection". *bioRxiv*: p. 716902.

Goh, C. and Knight, J. C. (2017). "Enhanced understanding of the host–pathogen interaction in sepsis: new opportunities for omic approaches". *The Lancet Respiratory Medicine* 5: pp. 212–223.

Gold, L., Ayers, D., Bertino, J., et al. (2010). "Aptamer-based multiplexed proteomic technology for biomarker discovery". *PloS one* 5.

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). "Coming of age: ten years of next-generation sequencing technologies". *Nature Reviews Genetics 2016 17:6* 17: pp. 333–351.

Gordon, A. C., Mouncey, P. R., Al-Beidh, F, et al. (2021). "The members of the writing committee (". *N Engl J Med* 384: pp. 1491–502.

Gordon, A. C., Mason, A. J., Thirunavukkarasu, N., et al. (2016). "Effect of Early Vasopressin vs Norepinephrine on Kidney Failure in Patients With Septic Shock: The VANISH Randomized Clinical Trial." *JAMA* 316: pp. 509–18.

Guinney, J., Dienstmann, R., Wang, X., et al. (2015). "The consensus molecular subtypes of colorectal cancer". *Nature Medicine 2015 21:11* 21: pp. 1350–1356.

Gustave, C.-A., Gossez, M., Demaret, J., et al. (2018). "Septic Shock Shapes B Cell Response toward an Exhausted-like/Immunoregulatory Profile in Patients". *Journal of immunology (Baltimore, Md. : 1950)* 200: pp. 2418–2425.

Gutmann, C., Takov, K., Burnap, S. A., et al. (2021). "SARS-CoV-2 RNAemia and proteomic trajectories inform prognostication in COVID-19 patients admitted to intensive care". *Nature Communications* 12.

Halstead, E. S., Umstead, T. M., Davies, M. L., et al. (2018). "GM-CSF overexpression after influenza a virus infection prevents mortality and moderates M1-like airway monocyte/macrophage polarization". *Respiratory research* 19: p. 3.

Harberts, E., Liang, T., Yoon, S. H., et al. (2020). "Toll-like Receptor 4-Independent Effects of Lipopolysaccharide Identified Using Longitudinal Serum Proteomics". *Journal of Proteome Research* 19: pp. 1258–1266.

Hasin, Y., Seldin, M., and Lusis, A. (2017). "Multi-omics approaches to disease". *Genome Biology 2017 18:1* 18: pp. 1–15.

Hastie, T., Tibshirani, R., Narasimhan, B., et al. (2021). *impute: impute: Imputation for microarray data*. R package version 1.66.0.

Hayashi, N., Yamaguchi, S., Rodenburg, F., et al. (2019). "Multiple biomarkers of sepsis identified by novel time-lapse proteomics of patient serum". *PLoS ONE* 14.

Herberg, J. A., Kaforou, M., Wright, V. J., et al. (2016). "Diagnostic Test Accuracy of a 2-Transcript Host RNA Signature for Discriminating Bacterial vs Viral Infection in Febrile Children". *JAMA* 316: p. 835.

Hohn, A., Iovino, I., Cirillo, F., et al. (2018). "Bioinformatical Analysis of Organ-Related (Heart, Brain, Liver, and Kidney) and Serum Proteomic Data to Identify Protein Regulation Patterns and Potential Sepsis Biomarkers".

Horby, P., Lim, W. S., Emberson, J. R., et al. (2021). "Dexamethasone in Hospitalized Patients with Covid-19". *New England Journal of Medicine* 384: pp. 693–704.

Hore, V., Viñuela, A., Buil, A., et al. (2016). "Tensor decomposition for multiple-tissue gene expression experiments". *Nature genetics* 48: pp. 1094–1100.

Hotchkiss, R. S., Monneret, G., and Payen, D. (2013). "Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy". *Nature Reviews Immunology 2013 13:12* 13: pp. 862–874.

Huard, B., McKee, T., Bosshard, C., et al. (2008). "APRIL secreted by neutrophils binds to heparan sulfate proteoglycans to create plasma cell niches in human mucosa". *The Journal of clinical investigation* 118: pp. 2887–2895.

Huber, W., Heydebreck, A. von, Sultmann, H., et al. (2002). "Variance stabilization applied to microarray data calibration and to the quantification of differential expression". *Bioinformatics* 18: S96–S104.

Huber-Lang, M., Lambris, J. D., and Ward, P. A. (2018). "Innate immune responses to trauma". *Nature Immunology 2018 19:4* 19: pp. 327–341.

Hung, M. S., Chen, I. C., Lin, P. Y., et al. (2016). "Epidermal growth factor receptor mutation enhances expression of vascular endothelial growth factor in lung cancer". *Oncology Letters* 12: pp. 4598–4604.

Ignjatovic, V., Geyer, P. E., Palaniappan, K. K., et al. (2019). "Mass Spectrometry-Based Plasma Proteomics: Considerations from Sample Collection to Achieving Translational Data". *Journal of proteome research* 18: p. 4085.

Ilies, M., Iuga, C. A., Loghin, F., et al. (2017). "Impact of blood sample collection methods on blood protein profiling studies". *Clinica Chimica Acta* 471: pp. 128–134.

Inforzato, A., Rivieccio, V., Morreale, A. P., et al. (2008). "Structural Characterization of PTX3 Disulfide Bond Network and Its Multimeric Status in Cumulus Matrix Organization". *Journal of Biological Chemistry* 283: pp. 10147–10161.

Introna, M, Alles, V., Castellano, M, et al. (1996). "Cloning of Mouse ptx3, a New Member of the Pentraxin Gene Family Expressed at Extrahepatic Sites". *Blood* 87: pp. 1862–1872.

Iqbal, A. J., Barrett, T. J., Taylor, L., et al. (2016). "Acute exposure to apolipoprotein a1 inhibits macrophage chemotaxis in vitro and monocyte recruitment in vivo". *eLife* 5.

Jackson, J. K., Gleave, M. E., Gleave, J., et al. (2005). "The inhibition of angiogenesis by antisense oligonucleotides to clusterin". *Angiogenesis 2005 8:3* 8: pp. 229–238.

James, G., Witten, D., Hastie, T., et al. (2013). *An introduction to statistical learning*. Vol. 112. Springer.

Janciauskiene, S., Wrenger, S., Immenschuh, S., et al. (2018). "The multifaceted effects of Alpha1-Antitrypsin on neutrophil functions". *Frontiers in Pharmacology* 9: p. 341.

Jankovska, E., Svitek, M., Holada, K., et al. (2019). "Affinity depletion versus relative protein enrichment: A side-by-side comparison of two major strategies for increasing human cerebrospinal fluid proteome coverage". *Clinical Proteomics* 16: pp. 1–10.

Jiang, L., Wang, M., Lin, S., et al. (2019). "A Quantitative Proteome Map of the Human Body": pp. 1–15.

Kalil, A. C., Patterson, T. F., Mehta, A. K., et al. (2021). "Baricitinib plus Remdesivir for Hospitalized Adults with Covid-19". *New England Journal of Medicine* 384: pp. 795–807.

Kalil, A. C. and Thomas, P. G. (2019). *Influenza virus-related critical illness: Pathophysiology and epidemiology*.

Kangelaris, K. N., Prakash, A., Liu, K. D., et al. (2015). "Increased expression of neutrophil-related genes in patients with early sepsis-induced ARDS". *American Journal of Physiology - Lung Cellular and Molecular Physiology* 308: pp. L1102–L1113.

Karapetis, C. S., Khambata-Ford, S., Jonker, D. J., et al. (2008). " K-ras Mutations and Benefit from Cetuximab in Advanced Colorectal Cancer ". *New England Journal of Medicine* 359: pp. 1757–1765.

Karp, N. A., Huber, W., Sadowski, P. G., et al. (2010). "Addressing accuracy and precision issues in iTRAQ quantitation." *Molecular and cellular proteomics : MCP* 9: pp. 1885–97.

Karpe, F., Vasan, S. K., Humphreys, S. M., et al. (2018). "Cohort Profile: The Oxford Biobank Why was the cohort set up?" *International Journal of Epidemiology*.

Kassambara, A., Kosinski, M., and Biecek, P. (2021). *survminer: Drawing Survival Curves using 'ggplot2'*.

Kaukonen, K.-M., Bailey, M., Pilcher, D., et al. (2015). "Systemic Inflammatory Response Syndrome Criteria in Defining Severe Sepsis". *New England Journal of Medicine* 372: pp. 1629–1638.

Kiehntopf, M., Schmerler, D., Brunkhorst, F. M., et al. (2011). "Mass Spectometry-Based Protein Patterns in the Diagnosis of Sepsis/Systemic Inflammatory Response Syndrome". *Shock* 36: pp. 560–569.

Kivelä, M., Arenas, A., Barthelemy, M., et al. (2014). "Multilayer networks". *Journal of Complex Networks* 2: pp. 203–271. arXiv: 1309 . 7233.

Kleinewietfeld, M., Manzel, A., Titze, J., et al. (2013). "Sodium chloride drives autoimmune disease by the induction of pathogenic TH17 cells". *Nature* 496: pp. 518–522.

Kong, A. T., Leprevost, F. V., Avtonomov, D. M., et al. (2017). "MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics". *Nature Methods 2017 14:5* 14: pp. 513–520.

Kosinski, T., Heilig, R., Bensaddek, D., et al. (2019). "Plasma proteomics goes high throughput-timsTOF Pro with PASEF and 4D feature alignment to quantify 500 plasma proteins in 11.5 min". *Bruker Daltonics* 03-2019, LCMS-151, 1867805.

Kuhn, M. (2021). *caret: Classification and Regression Training*.

Kumar, V. (2018). "T cells and their immunometabolism: A novel way to understanding sepsis immunopathogenesis and future therapeutics". *European Journal of Cell Biology* 97: pp. 379–392.

Kwok, A. J., Allcock, A., Ferreira, R. C., et al. (2022). "Identification of deleterious neutrophil states and altered granulopoiesis in sepsis". *medRxiv*: p. 2022.03.22.22272723.

Kwok, A. J., Mentzer, A., and Knight, J. C. (2020). "Host genetics and infectious disease". *Nature Reviews Genetics*.

Langfelder, P. and Horvath, S. (2008). "WGCNA: An R package for weighted correlation network analysis". *BMC Bioinformatics* 9: pp. 1–13.

Langley, R. J., Tsalik, E. L., Van Velkinburgh, J. C., et al. (2013). "Sepsis: An integrated clinico-metabolomic model improves prediction of death in sepsis". *Science Translational Medicine* 5.

Lee, G. W., Goodman, A. R., Lee, T. H., et al. (1994). "Relationship of TSG-14 protein to the pentraxin family of major acute phase proteins." *The Journal of Immunology* 153.

Leek, J. T., Johnson, W. E., Parker, H. S., et al. (2021). *sva: Surrogate Variable Analysis*. R package version 3.40.0.

Levy, M. M., Fink, M. P., Marshall, J. C., et al. (2003). "2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference". *Critical Care Medicine* 31: pp. 1250–1256.

Li, H., Liu, L., Zhang, D., et al. (2020). "SARS-CoV-2 and viral sepsis: observations and hypotheses". *The Lancet* 395: pp. 1517–1520.

Li, M., Ren, R., Yan, M., et al. (2022). "Identification of novel biomarkers for sepsis diagnosis via serum proteomic analysis using iTRAQ-2D-LC-MS/MS". *Journal of Clinical Laboratory Analysis* 36: e24142.

Liaw, A. and Wiener, M. (2002). "Classification and Regression by randomForest". *R News* 2: pp. 18–22.

Liston, A., Humblet-Baron, S., Duffy, D., et al. (2021). "Human immune diversity: from evolution to modernity". *Nature Immunology 2021*: pp. 1–11.

Lorenz, R. and Brauer, M. (1988). "Platelet factor 4 (PF4) in septicaemia". *Infection* 16: pp. 273–276.

Lötvall, J., Akdis, C. A., Bacharier, L. B., et al. (2011). "Asthma endotypes: A new approach to classification of disease entities within the asthma syndrome". *Journal of Allergy and Clinical Immunology* 127: pp. 355–360.

Macdonald-Dunlop, E., Klarić, L., Folkersen, L., et al. (2021). "Mapping genetic determinants of 184 circulating proteins in 26,494 individuals to connect proteins and diseases". *medRxiv* 10: p. 39.

MacIntyre, C. R., Chughtai, A. A., Barnes, M., et al. (2018). "The role of pneumonia and secondary bacterial infection in fatal and serious outcomes of pandemic influenza a(H1N1)pdm09". *BMC infectious diseases* 18.

Malmström, E., Kilsgård, O., Hauri, S., et al. (2016). "Large-scale inference of protein tissue origin in gram-positive sepsis plasma using quantitative targeted proteomics". *Nature Communications* 7: p. 10261.

Mantovani, A., Byrne, C. D., Zheng, M. H., et al. (2020). "Diabetes as a risk factor for greater COVID-19 severity and in-hospital death: A meta-analysis of observational studies". *Nutrition, Metabolism and Cardiovascular Diseases* 30: pp. 1236–1248.

Marshall, J. C. (2014). "Why have clinical trials in sepsis failed?" *Trends in Molecular Medicine* 20: pp. 195–203.

Maslove, D. M., Shapira, T., Tyryshkin, K., et al. (2019). "Validation of diagnostic gene sets to identify critically ill patients with sepsis". *Journal of Critical Care* 49: pp. 92–98.

Maslove, D. M., Tang, B., Shankar-Hari, M., et al. (2022). "Redefining critical illness". *Nature Medicine* 28: pp. 1141–1148.

Matsushita, M. (2010). "Ficolins: complement-activating lectins involved in innate immunity". *Journal of innate immunity* 2: pp. 24–32.

May, S. M., Abbott, T. E., Del Arroyo, A. G., et al. (2020). "MicroRNA signatures of perioperative myocardial injury after elective noncardiac surgery: a prospective observational mechanistic cohort study". *BJA: British Journal of Anaesthesia* 125: p. 661.

May, S. M., Reyes, A., Martir, G., et al. (2019). "Acquired loss of cardiac vagal activity is associated with myocardial injury in patients undergoing noncardiac surgery: prospective observational mechanistic cohort study". *British Journal of Anaesthesia* 123: pp. 758–767.

Mayr, F. B., Yende, S., and Angus, D. C. (2014). "Epidemiology of severe sepsis." *Virulence* 5: pp. 4–11.

Meier, F., Beck, S., Grassl, N., et al. (2015). "Parallel accumulation-serial fragmentation (PASEF): Multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device". *Journal of Proteome Research* 14: pp. 5378–5387.

Meijer, B., Gearry, R. B., and Day, A. S. (2012). "The role of S100A12 as a systemic marker of inflammation". *International journal of inflammation* 2012.

Melani, R. D., Gerbasi, V. R., Anderson, L. C., et al. (2022). "The Blood Proteoform Atlas: A reference map of proteoforms in human hematopoietic cells". *Science* 375: pp. 411–418.

Messner, C. B., Demichev, V., Bloomfield, N., et al. (2021). "Ultra-fast proteomics with Scanning SWATH". *Nature Biotechnology 2021 39:7* 39: pp. 846–854.

Messner, C. B., Demichev, V., Wendisch, D., et al. (2020). "Ultra-High-Throughput Clinical Proteomics Reveals Classifiers of COVID-19 Infection". *Cell Systems* 11: 11–24.e4.

Midwood, K. S. and Orend, G. (2009). "The role of tenascin-C in tissue injury and tumorigenesis". *Journal of cell communication and signaling* 3: pp. 287–310.

Mills, E. L., Kelly, B., Logan, A., et al. (2016). "Succinate Dehydrogenase Supports Metabolic Repurposing of Mitochondria to Drive Inflammatory Macrophages". *Cell* 167: 457–470.e13.

Mishra, J., Dent, C., Tarabishi, R., et al. (2005). "Neutrophil gelatinase-associated lipocalin (NGAL) as a biomarker for acute renal injury after cardiac surgery". *The Lancet* 365: pp. 1231–1238.

Misra, B. B., Langefeld, C., Olivier, M., et al. (2019). "Integrated omics: Tools, advances and future approaches". *Journal of Molecular Endocrinology* 62: R21–R45.

Mitchell, A. J., Gray, W. D., Hayek, S. S., et al. (2016). "Platelets confound the measurement of extracellular miRNA in archived plasma". *Scientific Reports 2016 6:1* 6: pp. 1–11.

Montasser, M. E., Van Hout, C. V., Miloscio, L., et al. (2021). " Genetic and functional evidence links a missense variant in B4GALT1 to lower LDL and fibrinogen ". *Science* 374: pp. 1221–1227.

Morris, J. H., Apeltsin, L., Newman, A. M., et al. (2011). "ClusterMaker: A multi-algorithm clustering plugin for Cytoscape". *BMC Bioinformatics* 12: pp. 1–14.

Murphy, K. and Weaver, C. (2016). *Janeway's immunobiology*. Garland science.

Nelson, M. R., Tipney, H., Painter, J. L., et al. (2015). "The support of human genetic evidence for approved drug indications". *Nature Genetics 2015 47:8* 47: pp. 856–860.

Neyton, L. P., Zheng, X., Skouras, C., et al. (2022). "Molecular Patterns in Acute Pancreatitis Reflect Generalizable Endotypes of the Host Response to Systemic Injury in Humans". *Annals of Surgery* 275: E453–E462.

Ng, P. C., Ang, I. L., Chiu, R. W. K., et al. (2010). "Host-response biomarkers for diagnosis of late-onset septicemia and necrotizing enterocolitis in preterm infants". *Journal of Clinical Investigation* 120: pp. 2989–3000.

Nie, X., Qian, L., Sun, R., et al. (2021). "Multi-organ proteomic landscape of COVID-19 autopsies". *Cell* 184: 775–791.e14.

Nieman, M. T. (2016). "Protease-activated receptors in hemostasis". *Blood* 128: pp. 169–177.

Niu, L., Geyer, P. E., Wewer Albrechtsen, N. J., et al. (2019). "Plasma proteome profiling discovers novel proteins associated with non-alcoholic fatty liver disease". *Molecular Systems Biology* 15: e8793.

Niu, L., Sulek, K., Vasilopoulou, C. G., et al. (2021). "Defining NASH from a multi-omics systems biology perspective". *Journal of Clinical Medicine* 10: p. 4673.

Niu, L., Thiele, M., Geyer, P. E., et al. (2020). "A paired liver biopsy and plasma proteomics study reveals circulating biomarkers for alcohol-related liver disease". *bioRxiv*: p. 2020.10.16.337592.

O'Callaghan, D. J., O'Dea, K. P., Scott, A. J., et al. (2015). "Monocyte tumor necrosis factor-α-converting enzyme catalytic activity and substrate shedding in sepsis and noninfectious systemic inflammation". *Critical Care Medicine* 43: pp. 1375–1385.

Omenn, G. S., States, D. J., Adamski, M., et al. (2005). "Overview of the HUPO Plasma Proteome Project: Results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database". *PROTEOMICS* 5: pp. 3226–3245.

Paich, H. A., Sheridan, P. A., Handy, J., et al. (2013). "Overweight and obese adult humans have a defective cellular immune response to pandemic H1N1 Influenza a virus". *Obesity* 21: pp. 2377–2386.

Papafilippou, L., Claxton, A., Dark, P., et al. (2020). "Protein corona fingerprinting to differentiate sepsis from non-infectious systemic inflammation". *Nanoscale* 12: pp. 10240–10253.

Parks, W. C., Wilson, C. L., and López-Boado, Y. S. (2004). "Matrix metalloproteinases as modulators of inflammation and innate immunity". *Nature Reviews Immunology 2004 4:8* 4: pp. 617–629.

Patel, P. B., Brett, S. J., O'Callaghan, D., et al. (2020). "Methylnaltrexone for the treatment of opioid-induced constipation and gastrointestinal stasis in intensive care patients. Results from the MOTION trial". *Intensive Care Medicine* 46: pp. 747–755.

Pierrakos, C., Velissaris, D., Bisdorff, M., et al. (2020). "Biomarkers of sepsis: time for a reappraisal". *Critical Care 2020 24:1* 24: pp. 1–15.

Pimienta, G., Heithoff, D. M., Rosa-Campos, A., et al. (2019). "Plasma Proteome Signature of Sepsis: a Functionally Connected Protein Network". *PROTEOMICS* 19: p. 1800389.

R Core Team (2015). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing.

Rautanen, A., Mills, T. C., Gordon, A. C., et al. (2015). "Genome-wide association study of survival from sepsis due to pneumonia: An observational cohort study". *The Lancet Respiratory Medicine* 3: pp. 53–60.

Reddy, K., Sinha, P., O'Kane, C. M., et al. (2020). "Subphenotypes in critical care: translation into clinical practice". *The Lancet Respiratory Medicine* 8: pp. 631–643.

Reich, M., Liefeld, T., Gould, J., et al. (2006). "GenePattern 2.0". *Nature genetics* 38: pp. 500–501.

Reinhart, K., Daniels, R., Kissoon, N., et al. (2017). "Recognizing Sepsis as a Global Health Priority — A WHO Resolution". *New England Journal of Medicine* 377: pp. 414–417.

Reslova, N., Michna, V., Kasny, M., et al. (2017). "xMAP technology: Applications in detection of pathogens". *Frontiers in Microbiology* 8: p. 55.

Reyes, M., Filbin, M. R., Bhattacharyya, R. P., et al. (2020). "An immune-cell signature of bacterial sepsis". *Nature Medicine 2020 26:3* 26: pp. 333–340.

Reyes, M., Filbin, M. R., Bhattacharyya, R. P., et al. (2021). "Plasma from patients with bacterial sepsis or severe COVID-19 induces suppressive myeloid cell production from hematopoietic progenitors in vitro". *Science Translational Medicine* 13: p. 9599.

Rieckmann, J. C., Geiger, R., Hornburg, D., et al. (2017). "Social network architecture of human immune cells unveiled by quantitative proteomics". *Nature Immunology* 18: pp. 583–593. arXiv: 0706.4396.

Ríos-Toro, J. J., Márquez-Coello, M., García-Álvarez, J. M., et al. (2017). "Soluble membrane receptors, interleukin 6, procalcitonin and C reactive protein as prognostic markers in patients with severe sepsis and septic shock". *PLOS ONE* 12: e0175254.

Ritchie, M. E., Phipson, B., Wu, D., et al. (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies". *Nucleic acids research* 43: e47.

Roh, J. S. and Sohn, D. H. (2018). "Damage-Associated Molecular Patterns in Inflammatory Diseases". *Immune Network* 18.

Rosengren, A. T., Nyman, T. A., and Lahesmaa, R. (2005). "Proteome profiling of interleukin-12 treated human T helper cells". *Proteomics* 5: pp. 3137–3141.

Rubio, I., Osuchowski, M. F., Shankar-Hari, M., et al. (2019). "Review Current gaps in sepsis immunology: new opportunities for translational research". *www.thelancet.com/infection*.

Rudd, K. E., Johnson, S. C., Agesa, K. M., et al. (2020). "Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study". *The Lancet* 395: pp. 200–211.

Sadrzadeh, S. M. and Bozorgmehr, J. (2004). "Haptoglobin phenotypes in health and disorders". *American journal of clinical pathology* 121 Suppl.

Scherag, A., Schöneweck, F., Kesselmeier, M., et al. (2016). "Genetic Factors of the Disease Course after Sepsis: A Genome-Wide Study for 28 Day Mortality". *EBioMedicine* 12: p. 239.

Schiess, R., Wollscheid, B., and Aebersold, R. (2009). "Targeted proteomic strategy for clinical biomarker discovery". *Molecular Oncology* 3: pp. 33–44.

Schwenk, J. M., Igel, U., Kato, B. S., et al. (2010). "Comparative protein profiling of serum and plasma using an antibody suspension bead array approach". *Proteomics* 10: pp. 532–540.

Scicluna, B. P., Klein Klouwenberg, P. M. C., Vught, L. A. van, et al. (2015). "A Molecular Biomarker to Diagnose Community-acquired Pneumonia on Intensive Care Unit Admission". *American Journal of Respiratory and Critical Care Medicine* 192: pp. 826–835.

Scicluna, B. P., Vught, L. A. van, Zwinderman, A. H., et al. (2017). "Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study". *The Lancet Respiratory Medicine* 5: pp. 816–826.

Seymour, C. W., Gomez, H., Chang, C. C. H., et al. (2017). "Precision medicine for all? Challenges and opportunities for a precision medicine approach to critical illness". *Critical Care* 21: pp. 1–11.

Seymour, C. W., Kennedy, J. N., Wang, S., et al. (2019). "Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis". *JAMA - Journal of the American Medical Association*. Vol. 321. American Medical Association: pp. 2003–2017.

Shannon, P., Markiel, A., Ozier, O., et al. (2003). "Cytoscape: a software environment for integrated models of biomolecular interaction networks". *Genome research* 13: pp. 2498–2504.

Shao, Z., Nishimura, T., Leung, L. L., et al. (2015). "Carboxypeptidase B2 deficiency reveals opposite effects of complement C3a and C5a in a murine polymicrobial sepsis model". *Journal of thrombosis and haemostasis : JTH* 13: pp. 1090–1102.

Shappell, C. N., Klompas, M., and Rhee, C. (2020). "Does Severe Acute Respiratory Syndrome Coronavirus 2 Cause Sepsis?" *Critical care medicine*: pp. 10–12.

Sharma, N. K., Ferreira, B. L., Tashima, A. K., et al. (2019). "Lipid metabolism impairment in patients with sepsis secondary to hospital acquired pneumonia, a proteomic analysis". *Clinical Proteomics* 16: pp. 1–13.

Sharma, N. K., Tashima, A. K., Brunialti, M. K. C., et al. (2017). "Proteomic study revealed cellular assembly and lipid metabolism dysregulation in sepsis secondary to community-acquired pneumonia". *Scientific Reports* 7: pp. 1–13.

Shen, B., Yi, X., Sun, Y., et al. (2020). "Proteomic and Metabolomic Characterization of COVID-19 Patient Sera". *Cell* 182: 59–72.e15.

Shubin, N. J., Navalkar, K., Sampson, D., et al. (2020). "Serum Protein Changes in Pediatric Sepsis Patients Identified With an Aptamer-Based Multiplexed Proteomic Approach". *Critical Care Medicine* 48: e48–e57.

Silasi-Mansat, R., Zhu, H., Popescu, N. I., et al. (2010). "Complement inhibition decreases the procoagulant response and confers organ protection in a baboon model of Escherichia coli sepsis". *Blood* 116: pp. 1002–1010.

Silver Spring (MD): Food and Drug Administration (US); Bethesda (MD): National Institutes of Health (US) (2016). "BEST (Biomarkers, EndpointS, and other Tools) Resource". *BEST (Biomarkers , EndpointS , and other Tools) Resource*.

Singer, M., Deutschman, C. S., Seymour, C. W., et al. (2016). "The third international consensus definitions for sepsis and septic shock (Sepsis-3)". *JAMA* 315: pp. 801–810. arXiv: 15334406.

Sinha, P., Furfaro, D., Cummings, M. J., et al. (2021). "Latent Class Analysis Reveals COVID-19-related ARDS Subgroups with Differential Responses to Corticosteroids".

Sinitcyn, P., Daniel Rudolph, J., and Cox, J. (2018). "Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data". *Annual Review of Biomedical Data Science* 1: annurev–biodatasci–080917–013516.

Sørensen, T. I., Nielsen, G. G., Andersen, P. K., et al. (1988). "Genetic and Environmental Influences on Premature Death in Adult Adoptees". *New England Journal of Medicine* 318: pp. 727–732.

Stanski, N. L. and Wong, H. R. (2019). "Prognostic and predictive enrichment in sepsis". *Nature Reviews Nephrology*.

Stein, M. M., Hrusch, C. L., Gozdz, J., et al. (2016). "Innate Immunity and Asthma Risk in Amish and Hutterite Farm Children". *The New England journal of medicine* 375: pp. 411–421.

Subramanian, A., Tamayo, P., Mootha, V. K., et al. (2005). "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles". *Proceedings of the National Academy of Sciences* 102: pp. 15545–15550.

Suhre, K., Arnold, M., Bhagwat, A. M., et al. (2017). "Connecting genetic risk to disease end points through the human blood plasma proteome". *Nature Communications 2017 8:1* 8: pp. 1–14.

Sun, B. B., Maranville, J. C., Peters, J. E., et al. (2018). "Genomic atlas of the human plasma proteome". *Nature* 558: pp. 73–79.

Swathi Raju, M., Jahnavi, V., Kamaraju, R. S., et al. (2016). "Continuous evaluation of changes in the serum proteome from early to late stages of sepsis caused by Klebsiella pneumoniae". *Molecular Medicine Reports* 13: pp. 4835–4844.

Sweeney, T. E., Azad, T. D., Donato, M., et al. (2018a). "Unsupervised analysis of transcriptomics in bacterial sepsis across multiple datasets reveals three robust clusters". *Critical Care Medicine* 46: pp. 915–925.

Sweeney, T. E., Liesenfeld, O., Wacker, J., et al. (2021). "Validation of Inflammopathic, Adaptive, and Coagulopathic Sepsis Endotypes in Coronavirus Disease 2019". *Critical Care Medicine*: E170–E178.

Sweeney, T. E., Perumal, T. M., Henao, R., et al. (2018b). "A community approach to mortality prediction in sepsis via gene expression analysis". *Nature Communications 2018 9:1* 9: pp. 1–10.

Szklarczyk, D., Gable, A. L., Lyon, D., et al. (2019). "STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets". *Nucleic Acids Research* 47: pp. D607–D613.

Thavarajah, T., Dos Santos, C. C., Slutsky, A. S., et al. (2020). "The plasma peptides of sepsis". *Clinical Proteomics* 17: pp. 1–18.

Theilgaard-Mönch, K., Jacobsen, L. C., Nielsen, M. J., et al. (2006). "Haptoglobin is synthesized during granulocyte differentiation, stored in specific granules, and released by neutrophils in response to activation". *Blood* 108: pp. 353–361.

Therneau, T. M. and Grambsch, P. M. (2000). *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Thévenot, E. A., Roux, A., Xu, Y., et al. (2015). "Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses". *Journal of proteome research* 14: pp. 3322–3335.

Thul, P. J., Akesson, L., Wiking, M., et al. (2017). "A subcellular map of the human proteome". *Science* 356.

Toledo, A. G., Golden, G., Campos, A. R., et al. (2019). "Proteomic atlas of organ vasculopathies triggered by Staphylococcus aureus sepsis". *Nature Communications* 10: pp. 1–13.

Tomic, A., Tomic, I., Rosenberg-Hasson, Y., et al. (2019). "SIMON, an Automated Machine Learning System, Reveals Immune Signatures of Influenza Vaccine Responses". *The Journal of Immunology* 203: pp. 749–759.

Tong, Y., Ku, X., Wu, C., et al. (2019). "Data-independent acquisition-based quantitative proteomic analysis reveals differences in host immune response of peripheral blood mononuclear cells to sepsis". *Scandinavian Journal of Immunology* 89: e12748.

Tridente, A., Clarke, G. M., Walden, A., et al. (2014). "Patients with faecal peritonitis admitted to European intensive care units: An epidemiological survey of the GenOSept cohort". *Intensive Care Medicine* 40: pp. 202–210.

Tu, C., Rudnick, P. A., Martinez, M. Y., et al. (2010). "Depletion of abundant plasma proteins and limitations of plasma proteomics". *Journal of Proteome Research* 9: pp. 4982–4991.

Tyanova, S., Temu, T., and Cox, J. (2016a). "The MaxQuant computational platform for mass spectrometry-based shotgun proteomics". *Nature Protocols* 11: pp. 2301–2319.

Tyanova, S., Temu, T., Sinitcyn, P., et al. (2016b). "The Perseus computational platform for comprehensive analysis of (prote)omics data". *Nature Methods* 13: pp. 731–740. arXiv: 1412.3197.

Uhlén, M., Fagerberg, L., Hallström, B. M., et al. (2015). "Tissue-based map of the human proteome". *Science* 347.

Uhlén, M., Karlsson, M. J., Hober, A., et al. (2019). "The human secretome". *Science Signaling* 12.

Välikangas, T., Suomi, T., and Elo, L. L. (2016). "A systematic evaluation of normalization methods in quantitative label-free proteomics". *Briefings in Bioinformatics* 19: bbw095.

Van Der Poll, T., Van De Veerdonk, F. L., Scicluna, B. P., et al. (2017). "The immunopathology of sepsis and potential therapeutic targets". *Nature Reviews Immunology* 17: pp. 407–420.

Van Vught, L. A., Scicluna, B. P., Wiewel, M. A., et al. (2017). "Association of gender with outcome and host response in critically ill sepsis patients". *Critical Care Medicine*.

Wabnitz, G. H., Köcher, T., Lohneis, P., et al. (2007). "Costimulation induced phosphorylation of L-plastin facilitates surface transport of the T cell activation molecules CD69 and CD25". *European journal of immunology* 37: pp. 649–662.

Walden, A. P., Clarke, G. M., McKechnie, S., et al. (2014). "Patients with community acquired pneumonia admitted to European intensive care units: An epidemiological survey of the GenOSept cohort". *Critical Care* 18: pp. 1–9.

Wang, B., Mezlini, A. M., Demir, F., et al. (2014). "Similarity network fusion for aggregating data types on a genomic scale". *Nature Methods 2014 11:3* 11: pp. 333–337.

Wang, Q., Li, S., Tang, X., et al. (2019). "Lipocalin 2 Protects Against Escherichia coli Infection by Modulating Neutrophil and Macrophage Function". *Frontiers in Immunology* 10: p. 2594.

Wang, S., Song, R., Wang, Z., et al. (2018). "S100A8/A9 in inflammation". *Frontiers in Immunology* 9: p. 1298.

Wilkerson, M. D. and Hayes, D. N. (2010). "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking". *Bioinformatics (Oxford, England)* 26: pp. 1572–1573.

Wishart, D. S., Feunang, Y. D., Guo, A. C., et al. (2018). "DrugBank 5.0: a major update to the DrugBank database for 2018". *Nucleic Acids Research* 46: pp. D1074–D1082.

Wong, H. R., Caldwell, J. T., Cvijanovich, N. Z., et al. (2019). "Prospective clinical testing and experimental validation of the Pediatric Sepsis Biomarker Risk Model". *Science Translational Medicine* 11: p. 9000.

Wong, H. R., Cvijanovich, N. Z., Anas, N., et al. (2015). "Developing a clinically feasible personalized medicine approach to pediatric septic shock". *American Journal of Respiratory and Critical Care Medicine* 191: pp. 309–315.

Wrona, E., Potemski, P., Sclafani, F., et al. (2021). "Leukemia Inhibitory Factor: A Potential Biomarker and Therapeutic Target in Pancreatic Cancer". *Archivum Immunologiae et Therapiae Experimentalis* 69: pp. 1–8.

Xochelli, A., Baliakas, P., Kavakiotis, I., et al. (2017). "Chronic Lymphocytic Leukemia with Mutated IGHV4-34 Receptors: Shared and Distinct Immunogenetic Features and Clinical Outcomes". *Clinical cancer research : an official journal of the American Association for Cancer Research* 23: pp. 5292–5301.

Yang, Z. Z., Grote, D. M., Ziesmer, S. C., et al. (2011). "Soluble IL-2RA facilitates IL-2-mediated immune responses and predicts reduced survival in follicular B-cell non-Hodgkin lymphoma". *Blood* 118: pp. 2809–2820.

Zador, Z., Landry, A., Cusimano, M. D., et al. (2019). "Multimorbidity states associated with higher mortality rates in organ dysfunction and sepsis: A data-driven analysis in critical care". *Critical Care* 23: pp. 1–11.

Zhang, J., Luo, Y., Wang, X., et al. (2019). "Global transcriptional regulation of STAT3- and MYC-mediated sepsis-induced ARDS". *Therapeutic Advances in Respiratory Disease* 13.

Zhang, J., Hu, Z. D., Song, J., et al. (2015). "Diagnostic Value of Presepsin for Sepsis: A Systematic Review and Meta-Analysis". *Medicine* 94: e2158.

Zhang, K., Zhang, S., Cui, W., et al. (2021). "Development and Validation of a Sepsis Mortality Risk Score for Sepsis-3 Patients in Intensive Care Unit". *Frontiers in Medicine* 7: p. 1142.

Zhong, B., Liu, X., Wang, X., et al. (2013). "Ubiquitin-specific protease 25 regulates TLR4-dependent innate immune responses through deubiquitination of the adaptor protein TRAF3". *Science signaling* 6.

Zhong, W., Edfors, F., Gummesson, A., et al. (2021). "Next generation plasma proteome profiling to monitor health and disease". *Nature Communications 2021 12:1* 12: pp. 1–12.

Zhou, Q., Cheng, C., Wei, Y., et al. (2020). "USP15 potentiates NF-κB activation by differentially stabilizing TAB2 and TAB3". *The FEBS journal* 287: pp. 3165–3183.

Zhu, Y. P., Wan, F. N., Shen, Y. J., et al. (2015). "Reactive stroma component COL6A1 is upregulated in castration-resistant prostate cancer and promotes tumor growth". *Oncotarget* 6: pp. 14488–14496.

Zierer, J., Pallister, T., Tsai, P. C., et al. (2016). "Exploring the molecular basis of age-related disease comorbidities using a multi-omics graphical model". *Scientific Reports 2016 6:1* 6: pp. 1–10.

Zou, Q., Jin, J., Hu, H., et al. (2014). "USP15 stabilizes MDM2 to mediate cancer-cell survival and inhibit antitumor T cell responses". *Nature immunology* 15: pp. 562–570.