

Using multi-omics to investigate the host response to COVID-19

Jack Gisby

Submitted to Imperial College London in part fulfilment of the requirements for the
degree of Doctor of Philosophy

June 2023

Centre for Inflammatory Disease

Department of Immunology and Inflammation, Faculty of Medicine

Imperial College London

Abstract

Many individuals with coronavirus disease 2019 (COVID-19) are asymptomatic or have only mild symptoms. However, a subset of patients experience severe disease, which can progress to respiratory distress and organ failure. End-stage kidney disease (ESKD) patients are at a particularly high risk of fatal COVID-19. Despite this, most ESKD patients could not fully self-isolate during the UK national lockdowns as they required regular haemodialysis treatments. As a result, there was a unique opportunity for longitudinal sampling of both outpatients and inpatients with COVID-19 in a group at high risk of severe disease.

New technologies, that permit the concurrent measurement of thousands of molecular traits, have the potential to grant us an improved understanding of the mechanisms underlying severe COVID-19 and could help prioritise new therapeutic targets. Blood transcriptomics and plasma proteomic profiling were applied to two cohorts of ESKD patients with COVID-19. Using these data, I characterised the temporal profiles of molecules following COVID-19 onset. For instance, LRRC15, a proposed receptor for SARS-CoV-2, fell in abundance in patients with severe disease but remained relatively stable in milder cases. Samples were also collected from patients following clinical recovery from COVID-19, which revealed that pathways related to clotting and inflammation remain dysregulated for months after infection by SARS-CoV-2.

These cohorts were collected as part of an observational study, so I could not discern between molecules that drive disease, versus those that are simply downstream effects of clinical deterioration. As a follow-up, Mendelian randomisation was performed. This involved the use of genetic variants as instruments to identify proteins that may cause severe COVID-19. While this approach did not conclusively identify proteins that mediate COVID-19 outcomes, it did reveal a mechanism by which differential splicing of the transmembrane region of the Fas cell death receptor influenced its abundance in plasma.

Acknowledgements

I would like to express my gratitude to my supervisor, Jimmy, for his guidance, support, and encouragement throughout my PhD journey. My research journey has been further enriched by discussions with Artemis Papadaki on the minutiae of the statistical methods used to analyse the datasets.

I would also like to thank all the patients and clinical staff at the Imperial College Healthcare NHS Trust, as well as those involved in sample processing and patient recruitment, for their invaluable contributions to my research.

I am deeply indebted to my partner, Maryam, who begrudgingly listened to me talk about my work. I am grateful to my parents, Ali and Jamie, for their support and for offering a safe haven during the national lockdowns. I would also like to thank my cat, Pebbles, who provided me with companionship and support during my research. Finally, I would like to give my heartfelt thanks to Croydon Central Domino's Pizza and Chat House Tandoori for providing much-needed sustenance while writing my thesis.

Statement of originality

I declare that the work presented in this thesis is my own. The contributions of others are described in the ***Contributions*** sections of Chapters 2, 3 and 4.

Copyright declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a [Creative Commons Attribution 4.0 International Licence](#) (CC BY).

Under this licence, you may copy and redistribute the material in any medium or format for both commercial and non-commercial purposes. You may also create and distribute modified versions of the work. This on the condition that you credit the author.

When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Contents

1	Introduction	15
1.1	Data-driven approaches to the life sciences	15
1.2	The layers of the central dogma	17
1.2.1	Genomics	17
1.2.2	Transcriptomics	22
1.2.3	Proteomics	29
1.3	Holistic approaches to analysing biological systems	35
1.3.1	Modelling molecular data	35
1.3.2	Integrating molecular measurements with genetic variants	45
1.4	The effect of COVID-19 on an at-risk population	50
1.4.1	COVID-19	50
1.4.2	End-stage kidney disease	56
1.4.3	Susceptibility of ESKD patients to severe COVID-19	58
1.5	Summary and objectives	59
2	Longitudinal proteomic profiling of dialysis patients with COVID-19	61
2.1	Chapter summary	61
2.2	Introduction	62
2.2.1	Longitudinal proteomic studies of COVID-19	62
2.2.2	The effect of ESKD and dialysis on plasma protein abundance	64
2.2.3	Contributions	66
2.3	Methods	67
2.3.1	Ethical approval	67
2.3.2	Main cohort	67
2.3.3	Validation cohort	70

2.3.4	Clinical severity scores	72
2.3.5	Proteomic assays	72
2.3.6	Normalisation and quality assessment and control.....	72
2.3.7	Missing values	75
2.3.8	Principal Components Analysis.....	75
2.3.9	Consensus clustering.....	75
2.3.10	Differential protein abundance analysis: ESKD versus healthy	75
2.3.11	Differential protein abundance analysis: COVID-19 positive versus negative.	76
2.3.12	Testing for associations between proteins and clinical severity	77
2.3.13	Testing for associations between proteins and clinical laboratory tests	77
2.3.14	Testing for associations between proteins and ethnicity.....	78
2.3.15	Alternative estimation of the FDR using the plug-in method	78
2.3.16	Empirical <i>p</i> -value calculation.....	79
2.3.17	Supervised learning	79
2.3.18	Survival analysis using joint modelling	80
2.3.19	Longitudinal analysis.....	80
2.3.20	Data and code availability	81
2.4	Results.....	82
2.4.1	Proteomic effects of COVID-19.....	85
2.4.2	Validation of results in an independent cohort.....	89
2.4.3	Proteins associated with COVID-19 severity.....	90
2.4.4	Potential confounders do not materially affect results	94
2.4.5	Benjamini-Hochberg ensures control of the false discovery rate	97
2.4.6	Comparing the effects of COVID-19 infection and severity.....	99
2.4.7	The effect of kidney failure and dialysis on the proteome.....	101
2.4.8	Comparison to COVID-19 patients without ESKD	103

2.4.9	Supervised learning	104
2.4.10	Proteins associated with risk of death	105
2.4.11	Proteins with distinct temporal profiles in severe COVID-19.....	109
2.4.12	Associations with clinical laboratory tests.....	113
2.4.13	Associations with ethnicity	115
2.5	Discussion	116
3	Evaluating the causal effects of proteins on severe COVID-19	126
3.1	Chapter summary	126
3.2	Introduction	127
3.2.1	Genetics of COVID-19	127
3.2.2	Contributions.....	128
3.3	Methods.....	129
3.3.1	Identifying COVID-19-associated variants.....	129
3.3.2	Selecting proteins robustly associated with COVID-19 severity	131
3.3.3	Genetic variants associated with circulating plasma proteins	131
3.3.4	Data harmonisation and clumping.....	132
3.3.5	Two-sample Mendelian Randomisation	132
3.3.6	MR restricted to <i>cis</i> -pQTLs	133
3.3.7	Pleiotropy robust methods and testing	133
3.3.8	Heterogeneity testing	134
3.3.9	Colocalisation testing	134
3.3.10	Replicating the FAS pQTL using alternative assays.....	135
3.3.11	The association of FAS and COVID-19 severity.....	135
3.3.12	Calculating splice QTLs	135
3.4	Results.....	137
3.4.1	Defining instruments and outcomes	137

3.4.2	Mendelian randomisation	138
3.4.3	Colocalisation	143
3.4.4	Investigating alternative outcomes	145
3.4.5	Assessing the FAS pQTL.....	147
3.4.6	Investigating the mechanism of the FAS pQTL	148
3.4.7	Assessing the COVID-19 locus.....	153
3.5	Discussion	155
4	Multi-omics of dialysis patients with COVID-19	158
4.1	Chapter summary	158
4.2	Introduction	159
4.2.1	The effect of ESKD and dialysis on the PBMC transcriptome	159
4.2.2	Contributions.....	160
4.3	Methods.....	161
4.3.1	Patient cohorts and ethical approval	161
4.3.2	Clinical severity scoring	164
4.3.3	PBMC collection protocol.....	164
4.3.4	Plasma collection.....	164
4.3.5	RNA-seq of PBMCs	165
4.3.6	Plasma proteomics	166
4.3.7	Comparison of ESKD patients on dialysis to healthy controls	167
4.3.8	Comparison of COVID-19 positive and negative patients.....	168
4.3.9	Molecular associations with COVID-19 severity	169
4.3.10	Gene set variation analysis	169
4.3.11	Robust rank aggregation.....	170
4.3.12	Modelling modular longitudinal trajectories	170
4.3.13	Additional WGCNA module annotation and association testing	172

4.3.14	Cell fraction imputation	172
4.3.15	Multi-omic longitudinal factor analysis with MEFISTO	172
4.3.16	Longitudinal modelling of cytokines and cytokine receptors	173
4.3.17	Supervised learning	173
4.3.18	Comparison of preinfection and recovery timepoints.....	174
4.3.19	Data and code availability	175
4.4	Results.....	176
4.4.1	The effect of COVID-19	178
4.4.2	Transcriptomic and proteomic changes associated with COVID-19 severity	183
4.4.3	Exploring the effects of including additional clinical covariates.....	187
4.4.4	The effect of ESKD and dialysis.....	190
4.4.5	Concordance with studies of COVID-19 for non-ESKD patients.....	197
4.4.6	Severe COVID-19 is associated with dynamic molecular trajectories.....	206
4.4.7	The temporal dynamics of cytokines in severe COVID-19.....	214
4.4.8	Plasma LRRRC15 predicts COVID-19 severity	220
4.4.9	Persistent deranged platelet and coagulation pathways in convalescence...	225
4.5	Discussion	228
5	Conclusions and future directions	238
6	Bibliography	242

List of figures

Figure 1.1 A broad overview of mRNA sequencing	26
Figure 2.1 Baseline characteristics of the Main cohort	69
Figure 2.2 Principal components analysis in relation to assay plate	74
Figure 2.3 Study design	83
Figure 2.4 Serial sampling of the Main cohort	84
Figure 2.5 Principal components analysis	86
Figure 2.6 Consensus clustering	87
Figure 2.7 Proteins dysregulated in COVID-19	88
Figure 2.8 Comparison to the Validation cohort	90
Figure 2.9 Principal components analysis in relation to clinical severity	91
Figure 2.10 Proteins dysregulated in severe COVID-19	92
Figure 2.11 Selected proteins strongly associated with COVID-19 severity	93
Figure 2.12 Sensitivity analyses adjusting for diabetes status and cause of ESKD	95
Figure 2.13 Sensitivity analysis adjusting for time since last haemodialysis	96
Figure 2.14 Permutation analysis to demonstrate successful FDR control	98
Figure 2.15 Comparison of proteins differentially expressed in COVID-19 with those associated with clinical severity	100
Figure 2.16 The effect of ESKD and dialysis on the plasma proteome	102
Figure 2.17 Comparison with the report of Filbin <i>et al.</i>	104
Figure 2.18 Prediction of severe COVID-19	106
Figure 2.19 Proteins associated with risk of death	107
Figure 2.20 Proteins associated with risk of death and their associations with clinical severity and laboratory measurements	108
Figure 2.21 Modelling of temporal protein trajectories	110
Figure 2.22 Temporal trajectories for proteins with a significant time x severity interaction	111
Figure 2.23 Raw data points for modelling of temporal protein trajectories	112
Figure 2.24 Associations of clinical laboratory markers with plasma proteins	114
Figure 3.1 Mendelian randomisation study design and data sources	130

Figure 3.2 Regional associations for the FAS <i>cis</i> -pQTL and COVID-19 outcomes	144
Figure 3.3 The effect of changing priors on FAS-COVID-19 colocalisation	145
Figure 3.4 Plasma FAS is associated with COVID-19 severity.....	146
Figure 3.5 Validation of MR estimates of the causal effect of soluble FAS protein on COVID-19 outcomes.....	148
Figure 3.6 Regional associations for FAS pQTL and eQTLs	150
Figure 3.7 Matching regional association patterns for the FAS pQTL and sQTL	151
Figure 3.8 A putative mechanism for the effect of FAS splicing	152
Figure 3.9 Regional associations for the FAS <i>cis</i> -pQTL and the 7 th release of the COVID-19 A2 outcome	154
Figure 4.1 Human Protein Atlas classification of proteins measured by the SomaScan v4.1 assay.....	167
Figure 4.2 Study design.....	177
Figure 4.3 Cohort summary.....	178
Figure 4.4 Principal components analysis coloured by COVID-19 status.....	180
Figure 4.5 Transcriptomic enrichment analysis comparing COVID-19 positive and negative samples.	180
Figure 4.6 Intra-individual comparison of pre-infection and acute COVID-19.	181
Figure 4.7 Proteomic enrichment analysis comparing COVID-19 positive and negative samples.	182
Figure 4.8 Association of the PBMC transcriptome and COVID-19 severity.....	184
Figure 4.9 Plasma proteomic associations with COVID-19 severity.....	186
Figure 4.10 The impact of clinical covariates on the transcriptome and proteome ...	188
Figure 4.11 The effect of including additional covariates in linear models	189
Figure 4.12 Overall abundance and expression for healthy controls and ESKD patients	194
Figure 4.13 The effect of ESKD and dialysis on the PMBC transcriptome.....	195
Figure 4.14 The effect of ESKD and dialysis on the plasma proteome	196
Figure 4.15 Comparison to the effect of COVID-19 in the COMBAT Consortium cohort	198

Figure 4.16 Comparison to the effect of COVID-19 severity in the COMBAT Consortium cohort	199
Figure 4.17 Overlap of significant genes and gene sets with data from the COMBAT study	200
Figure 4.18 Comparison of the effect of COVID-19 in the data of Filbin <i>et al.</i>	202
Figure 4.19 Comparison of the effect of COVID-19 severity in the data of Filbin <i>et al.</i>	203
Figure 4.20 Overlap of significant proteins and protein sets with data of Filbin <i>et al.</i>.....	204
Figure 4.21 Comparing the effect of COVID-19 in each cohort.....	205
Figure 4.22 Longitudinal profiles of transcriptomic modules.....	210
Figure 4.23 A multi-modal factor representing COVID-19 severity.....	211
Figure 4.24 Association of transcriptomic modules with imputed cell abundances .	212
Figure 4.25 Longitudinal profiles of protein modules and hub proteins.....	213
Figure 4.26 Cytokines with different temporal profiles depending on COVID-19 severity	215
Figure 4.27 Example cytokine profiles and underlying protein abundance.....	216
Figure 4.28 Comparison of SomaScan and Olink assays	218
Figure 4.29 The longitudinal profile of plasma EPOR protein.....	219
Figure 4.30 Supervised learning to predict COVID-19 severity from molecular features	221
Figure 4.31 Predicting COVID-19 severity using random forests	222
Figure 4.32 The profile of LRRC15 in plasma following COVID-19 infection	224
Figure 4.33 Persistent dysregulation of immune cell gene expression after recovery from COVID-19	226

List of tables

Table 1.1 Types of point mutations	19
Table 2.1 Characteristics of the Main cohort	68
Table 2.2 Characteristics of the controls	70
Table 2.3 Characteristics of the Validation cohort.....	71
Table 2.4 Missing values for clinical parameters.....	78
Table 3.1 Summary of Mendelian randomisation results	141
Table 3.2 Full <i>pan</i> -Mendelian randomisation results.....	142
Table 4.1 Characteristics of the Wave 1 cohort	162
Table 4.2 Characteristics of the Wave 2 cohort	163
Table 4.3 Characteristics of the controls	163
Table 4.4 Transcriptomic modules associated with disease trajectory	208
Table 4.5 Proteomic modules associated with disease trajectory	209
Table 4.6 Genes that do not return to baseline after recovery from COVID-19	227

1 Introduction

1.1 Data-driven approaches to the life sciences

Traditionally, the field of molecular biology has relied on a reductionist hypothesis-based approach¹. This has involved separating biological systems into their constituent parts, enabling the study of these components in isolation. This approach has yielded an understanding of the building blocks of biology and has provided a basis for understanding living organisms. However, living systems are more than the sum of their parts. In contrast, systems biology attempts to take a holistic viewpoint in which biological systems are studied as a whole; such an approach has been facilitated by the development of the “-omics” technologies, which allow for the concurrent measurement of entire categories of molecular features.

The first -omics field, genomics, began to be developed in the 1980s, enabling the advent of the Human Genome Project (HGP) in 1990 and its completion in 2003². The project was largely successful in its aims, having sequenced and assembled most of the human nuclear genome and permitted the identification of most protein-coding genes. The HGP gave way to studies that aimed to link genetic variation with observable traits. These included genome-wide association studies, which have implicated many genomic loci in the development of common diseases. However, many such loci fall outside of coding sequences, so elucidating the mechanisms by which they influence disease risk remains challenging^{2,3}. As a result, identifying the functional pathways that link genotype to phenotype has become a major focus of genomics research.

Entirely new -omics fields have emerged following the development of high-throughput technologies that enable the concurrent measurement of thousands of molecular traits at scale, such as microarrays and next-generation sequencing (NGS)⁴. These approaches have presented new opportunities for understanding disease heterogeneity and the discovery of biomarkers. While genomic information is effectively static, these other 'omic layers represent a dynamic system that more closely captures the state of the system's phenotype⁵. The most established tool for discovering pathways underlying disease is transcriptomics, which involves the measurement of RNA transcripts. More recently, technologies have been

developed that can measure the abundance of thousands of proteins. Proteins are more stable than RNA and most clinical tests measure proteins, so proteomics may have greater potential to identify clinically relevant biomarkers. Furthermore, proteins are the targets of most drugs, so studying proteins has the potential to identify therapeutic targets⁶. A major benefit of molecular data is that it can be integrated with genetic data to elucidate the functional pathways by which genetic variants mediate disease.

Advances in experimental methods have led to the generation of huge volumes of biological data⁷⁻¹⁰; the amount of data being generated is increasing at a far greater rate than our capacity to process and store it. It is expected that the volume of genomic data generated will continue to increase¹¹, largely due to its rapidly decreasing cost. As the cost of sequencing continues to decrease, it may be that experimental design, sample collection and downstream analyses become the limiting factors for these studies^{12,13}. The increasing proportion of time spent on data analysis and the complexity of the interactions between effector molecules is indicative of the challenge involved in deriving biological insights from these datasets. An individual -omics layer tends to allow for the generation of a list of differences associated with the disease and other traits of interest⁴. Alone, however, this information does not tell us about the aetiology of associations. There exists a compendium of methods to further investigate these datasets; these include dimensionality reduction or factorisation methods, clustering, and pathway enrichment analysis. Furthermore, more complex experimental designs may be employed, such as longitudinal studies in which measurements are taken in a repeated manner. Determination of individual temporal profiles can help us better understand the underlying factors involved in disease through the temporal ordering of molecular changes¹⁴. In study designs where multiple -omics modalities are measured, we can additionally integrate the data modalities to understand the interaction between molecules and generate novel insights.

1.2 The layers of the central dogma

The central dogma of molecular biology describes how information is transferred from DNA within biological systems¹⁵. Specifically, the information it contains can be transcribed to messenger RNA (mRNA) which can in turn be interpreted by ribosomes to synthesise proteins through the process of translation. The central dogma stipulates that sequence information does not generally flow in the reverse direction (i.e., from RNA or proteins to DNA) and that proteins cannot be synthesised directly from DNA. While there are exceptions to these rules, for instance in the case of reverse transcription, the central dogma provides a useful framework for interpreting the flow of information within cells. In this introduction, I provide a broad overview of the layers of the central dogma, including how they can be quantified and how these measurements can be applied to investigate biological systems.

1.2.1 Genomics

1.2.1.1 *Inheritance and variation*

According to the central dogma, the information required to synthesise proteins is stored by and flows from the genetic code, so it is essential that this information be passed to successive generations. The offspring of eukaryotic organisms are produced by sexual reproduction. This process begins with meiosis, during which the DNA of a diploid cell is replicated followed by two rounds of cell division. This results in the formation of four haploid gametes that contain only one of each autosome and one sex chromosome. Two key properties of meiosis allow for the shuffling of genetic material. Firstly, after DNA replication but before cell division, genetic recombination occurs for each pair of homologous autosomes. Transfer of information between non-sister chromatids leads to the generation of autosomes with novel combinations of alleles. Secondly, the homologous autosomes are distributed randomly among the gametes.

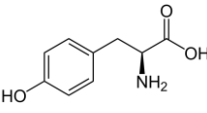
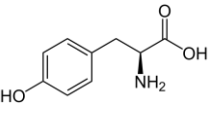
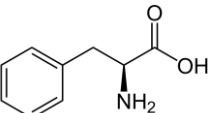
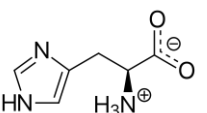
Differences in the genetic code between individuals are referred to as variation. These differences are, in part, explained by the shuffling of genetic material during meiosis. However, heritable genetic variants are also produced by random mutations. There are many types of mutations that arise through a variety of processes, including errors introduced during DNA replication and recombination. The simplest and most common form of variation are single nucleotide polymorphisms (SNPs), which are differences at a single base in the genome. For

instance, consider a base in the genome at which in most individuals the position is occupied by an A nucleotide, while a minority instead have a T allele. In this case, an individual could inherit the same allele from both parents (i.e., a genotype of AA or TT), or they could inherit different alleles from each parent (i.e., a genotype of AT). The most common allele (A) is referred to as the major allele, whilst the less frequent (T) is called the minor allele. For a variant to be considered an SNP, it must have a minor allele frequency (MAF) of at least 1%. Else, the variant is usually considered to be rare and is referred to as a single nucleotide variation (SNV) instead.

If a point mutation occurs within a protein coding sequence it has the potential to change the base composition of a protein. In coding sequences, each set of three bases is referred to as a codon, each of which encodes for a particular amino acid or a stop codon. Multiple codons can encode the same amino acid, meaning that not all base substitutions will affect the amino acid sequence (**Table 1.1**). For instance, both ACT and ACC encode for threonine. These types of mutations are referred to as synonymous, or silent. In contrast, non-synonymous mutations affect the amino acid sequence of a protein. These include missense mutations, which lead to the substitution of an amino acid. If such a mutation introduces an amino acid with similar physicochemical properties to the original (i.e., a conservative mutation), it may have little effect, depending on its position in the protein (**Table 1.1**). If the new amino acid has significantly different properties (i.e., a non-conservative mutation), such as an opposing polarity, then it is more likely to change the structure and function of the protein. Substitutions that introduce a stop codon, which includes TAA, TAG and TGA, are nonsense mutations. The introduction of a premature termination stop codon can lead to the production of a truncated protein. In such cases, a protein may not be produced at all due to the activation of nonsense-mediated decay, a pathway that degrades transcripts with abnormal terminating codons. Premature termination can also be averted at the stage of translation by suppressor tRNAs. Finally, in the case of read-through mutations, stop codons can be replaced by sense codons that encode an amino acid leading to aberrant extension of a protein.

Table 1.1 Types of point mutations

The effects of point mutations in coding sequences on the amino acid sequence of a protein.

Type of mutation	None	Silent	Missense		Nonsense
			Conservative	Non-conservative	
DNA Codon	TAC	TAT	TTC	CAC	TAA
Amino acid encoded	Tyr 	Tyr 	Phe 	His 	Stop

Outcomes of mutations can be described as beneficial, neutral, or deleterious. Deleterious mutations can disrupt a protein's ability to function. Such an outcome is particularly common in the case of nonsense mutations, which are often harmful. Neutral mutations have neither a significant deleterious nor beneficial effect, as is often the case for conservative missense mutations. Finally, while rare, mutations can have a beneficial effect. As an example, the SNP rs334 is located in the 6th codon of the β -globin gene. The most common allele is A and is part of a codon corresponding to glutamate, whereas the less common T allele leads to substitution by valine. This missense mutation leads to the production of defective haemoglobin capable of polymerisation. It manifests as sickle cell disease in individuals homozygous for the abnormal allele (TT). However, individuals that are heterozygous for rs334 (AT) generally do not suffer from symptoms and are more resistant to malaria compared to those that are heterozygous for the normal (AA) allele. Therefore, the T allele of rs334 has both deleterious and beneficial effects.

Mutations in coding sequences can directly affect the structure and function of proteins. However, most SNPs are found outside of coding sequences. While they will not lead to an amino acid substitution, these mutations can still elicit a phenotypic effect. For instance, some mutations can modulate the expression of genes. Examples of such mutations are those that change the base compositions of promoters, sequences at which RNA polymerase can bind to initiate transcription, and enhancers, to which transcription factors bind that modulate the expression of genes. Variants can also be present at splice sites, thereby controlling exon excision and potentially affecting the structure and function of synthesised proteins.

SNPs are not the only form of genetic variation. Small insertions or deletions of sequences (indels) are numerous in the genome. In-frame indels in coding sequences, whose lengths are divisible by three, lead to the insertion or deletion of an amino acid. Any other indel will lead to a frameshift mutation that leads to a change in the grouping of codons, thereby changing the entire downstream amino acid sequence. The change in reading frame will almost certainly be deleterious, leading to loss of function for the protein. Variants that affect larger genomic regions are classified as structural variants. Structural variants include translocations, the movement of large sequences to another site in the genome, and inversions, in which a locus is inverted with respect to its original position. Another class of structural variants are copy number variants, which are a result of larger indels and duplications that lead to differences in the number of repeats of a particular sequence in the genome.

Given that genetic material is shuffled during meiosis, one might expect that each variant in the genome is assorted independently within the population. However, this is not the case, and in fact alleles are not randomly distributed. The further away two loci are on a chromosome, the greater chance of a recombination event that separates them. In practice, loci that are close together are more likely to segregate together. At a population level, this concept is referred to as linkage disequilibrium (LD) and acts as a measure of the association between two loci. The LD between two variants is commonly quantified as the square of their correlation coefficient (r^2). Recombination itself is not random and occurs at certain sites preferentially. As a result, there are regions in the genome that have undergone little recombination in the population. These loci are referred to as haplotype blocks.

1.2.1.2 Genome-wide methods for identifying genetic variants

A common and cost-effective way to measure genetic variants is with an SNP array. These are DNA microarrays containing a variety of oligonucleotide probes that are each able to bind a particular sequence. These probes are sufficiently specific that they can be used to target a specific allele, allowing SNP arrays to be designed that can profile hundreds of thousands of SNPs in an individual's genome. However, given that estimates for the number of common SNPs in the human genome are in excess of 10 million, these data alone are insufficient for the exhaustive study of common genetic variants. Imputation can be used to ascertain unmeasured genotypes based on detailed reference data of human haplotype blocks. These references are commonly based on whole genome sequencing (WGS) data, which can

provide genotyping data without relying on reference haplotype maps. While WGS remains more expensive than SNP arrays, the cost of sequencing continues to drop¹⁶, making it an increasingly attractive option for studies of population genetics. For instance, the 1000 Genomes Project initially involved the application of WGS to 1,092 individuals and the UK Biobank project has generated WGS data for 200,000 participants to date.

1.2.1.3 Genome-wide association studies

Improvements in technological and analytical methods have allowed for large-scale genotyping, which has in turn given rise to Genome-Wide Association Studies (GWAS). These observational studies involve both the genome-wide measurement of variants (usually SNPs, but more recently have been extended to include indels) and the measurement of one or more traits of interest. Often, such traits represent a binary phenotype, such as the presence or absence of a disease. Therefore, GWAS are often case-control studies that compare the frequencies of alleles between cases (e.g., a group of individuals with a disease) and controls (e.g., healthy controls). Specifically, logistic regression may be used to identify alleles with significantly different frequencies in the cases, indicating that a particular variant, or genomic region, may influence the trait being studied. Association testing can also be applied to identify variants linked with quantitative measurements. Quantitative trait loci (QTLs) are regions of DNA that are associated with quantitative traits. In practice, the identification of QTLs (QTL mapping) can be performed in a similar fashion to the analysis of traditional GWAS by modelling the trait as an outcome of linear regression.

GWAS generally focus on relatively common variants. This is, in part, because many such studies are underpowered to detect associations with rarer variants. These studies often filter variants based on MAF, although the choice of threshold varies based on the number of participants that have been genotyped, the expected effect sizes and the population being studied¹⁷. GWAS are therefore predicated on the concept that common diseases are generally caused by common variants (referred to as the Common Disease-Common Variant hypothesis¹⁷). If a disease is relatively common in the population, then its causes must also be common. If a single or small number of variants were the cause of a common disease, they would have to be highly penetrant (i.e., have a large effect). However, such deleterious mutations would be subject to strong negative selection pressure, and so would be unlikely to become sufficiently prevalent in the population. Common variants, on the other hand,

generally have low penetrance and as such are under lower selection pressure, allowing them to propagate. Since each of these variants have only a small effect on the risk of disease, and because said disease is common in the population, there must be many of them. Of course, biology is rarely so simple, and so some cases of common diseases can be caused by rarer, more highly penetrant, variants. For instance, one of the highly penetrant BRCA1 mutations is present in less than 1% of the population¹⁸. While the BRCA1 and BRCA2 mutations are only present in a minority (<10%) of patients with breast cancer, the BRCA1 and BRCA2 mutations are respectively associated with an ~57% and ~49% increase in the risk of breast cancer¹⁹.

1.2.2 Transcriptomics

1.2.2.1 Gene expression

Genetic information is transferred to RNA by the process of transcription. Most RNA is non-coding; this includes ribosomal RNA, which is by far the most abundant type of transcript observed in human samples²⁰. However, studies of the transcriptome commonly focus on mRNA, which acts as an intermediary template for a functional protein. Transcription is a multi-stage process that can influence both the abundance of mRNA transcripts and the specific portions of a gene that are transcribed. In eukaryotes, transcripts formed by RNA polymerase are initially referred to as pre-mRNA and cannot be translated without further processing. This involves the addition of a 5' cap (i.e., the addition of a modified guanosine nucleotide to the 5' end) and a 3' poly(A) tail (i.e., the addition of a chain of adenine nucleotides to the 3' end). When a gene is copied it contains both exons (sequences that encode a portion of a protein) and introns (sequences that must be removed by splicing prior to the formation of a mature mRNA transcript). Additionally, many genes undergo alternative splicing, in which multiple mRNA isoforms can be produced by the exclusion of exons. Different mRNA isoforms can lead to the synthesis of proteins with different structures and functions.

The expression of genes as RNA transcripts is of interest because they give rise to functional proteins, and they represent one of the main channels by which genetic variants can modulate observable characteristics. There are many mechanisms by which the abundance of mature mRNA can be regulated. The rate of transcription itself can be modulated by controlling the activity of RNA polymerases. One way in which this control can be achieved is by modifying

the accessibility of a gene. In eukaryotic nuclei, DNA exists as a complex of DNA and proteins that is referred to as chromatin. DNA is wound around histones, which are packed tightly within the nucleus. Modifications to the chromatin structure through the modification and remodelling of histones can affect the physical accessibility of genes to RNA polymerase and transcription factors (TFs). TFs are proteins that bind to specific DNA sequences known as enhancers to activate or repress the transcription of a particular gene by RNA polymerase. The levels of TFs themselves can be differentially regulated in different cells or disease states. The binding of proteins to DNA can also be affected by methylation of cytosine nucleotides. DNA methylation can in turn affect chromatin remodelling and histone modification.

There are also several ways in which the abundance of mature mRNA can be controlled post-transcriptionally. The processing of mRNA can be controlled, for instance, at the capping and splicing stages. Furthermore, RNA is not a permanent fixture of cells and is degraded over time in the nucleus and cytoplasm²¹. The half-life of mRNA is approximately 10 hours in human cells²². However, this degradation can vary for different transcripts; for instance, the length of the 3' poly(A) tail attached to mRNA transcripts influences the stability of the molecule²¹. More rarely, mRNA sequences can also be edited post-transcriptionally²³. As a result, the observed abundance of mRNA within cells is a result of a complex set of regulatory mechanisms. Measuring gene expression, therefore, provides information on the sum of these processes, representing the overall transcriptional control of a particular gene.

1.2.2.2 A brief history of transcriptomics

Transcriptomics aims to study the genome-wide transcriptional state (i.e., the complete set of RNA or mRNA) of a system. Measuring the transcriptome in its entirety has historically been a non-trivial task. Early attempts at measuring genome-wide expression used low-throughput sequencing methods, such as Sanger sequencing; these approaches involved sequencing complementary DNA (cDNA) derived from mRNA transcripts^{24–26}. Additionally, highly targeted technologies were developed to quantify the transcripts of individual genes, such as northern blotting²⁷ and reverse transcriptase quantitative PCR²⁸. But these methods were unsuitable for quantitatively assessing the entire transcriptome.

While they are now less popular, microarray-based technologies have been extensively used to perform transcriptomic profiling. These function in a similar manner to the SNP arrays that

are used for genotyping. Microarrays are 2-dimensional grids of spots that each contain a specific DNA sequence, referred to as probes. These probes typically encode part of a gene or some other transcribed element of interest. Fluorescently labelled cDNA, produced by reverse transcription of mRNA, can be incubated with these probes. Subsequently, hybridisation between probes and their target sequences can be quantified and represents relative transcript abundance in the original sample. More complex microarrays have also been developed, such as tiling arrays, which contain a high density of overlapping probes, enabling the discovery and quantification of alternative splicing events²⁹. While microarrays are a cost-effective means of measuring gene expression, they can only measure sequences that are known *a priori*. Microarrays also have minimum and maximum detection limits. Non-specific binding leads to a background signal, limiting the sensitivity of these methods for lowly-expressed genes. Furthermore, microarrays suffer from truncation of fluorescence signal when expression is very high, so very high levels of expression are not captured.

1.2.2.3 RNA sequencing

The development and falling costs of high throughput sequencing technologies have enabled the increasing use of RNA sequencing (RNA-seq) to measure genome-wide expression. Unlike hybridisation-based methods, RNA-seq is truly untargeted in that it can measure the abundance of the entire set of transcripts in the sample without the need for the development of targeted probes. RNA-seq does not have an inherent limit for the quantification of transcripts, with an estimated dynamic range encompassing 5 orders of magnitude³⁰. Sequences, rather than abundance, are the output of an RNA-seq experiment, permitting more in-depth analysis of splicing junctions and the identification of sequence variants²⁹.

RNA must be extracted and processed before RNA-seq can be performed; these steps are referred to as sample and library preparation. The specific methods used for these steps differ but invariably include several key stages (**Figure 1.1**). These commonly include:

1. **RNA isolation.** RNA must first be extracted from a biological sample. Contaminating DNA must also be removed. Samples may additionally be assessed for both the integrity and amount of RNA.
2. **RNA enrichment/depletion.** The extracted RNA may then be enriched or depleted to ensure high representation of a specific class of RNA. In the experiments described in

this thesis, mRNA was enriched by selecting for RNA molecules with poly(A) tails. Haemoglobin RNA depletion was also performed in an attempt to remove RNA from erythrocytes.

3. **Fragmentation.** Most high-throughput sequencing methods can measure reads of approximately 100 base pairs (bp), which is shorter than typical mRNA transcripts. For this reason, the RNA is fragmented.
4. **Reverse transcription.** RNA is reverse transcribed to cDNA. DNA is more stable and so is a more suitable medium for storing sequence information. Furthermore, sequence amplification is required at subsequent stages, which is based on DNA polymerases.
5. **Adapter ligation.** Sequence adapters are ligated to each cDNA molecule, which are required for sequencing.
6. **Sequence amplification.** Amplification by polymerase chain reaction (PCR) can optionally be performed, which can be useful when there is only a meagre amount of RNA available in the original sample.
7. **Size selection.** Sequences can be selected that are of an appropriate size for sequencing.

Following the preparation of cDNA libraries, RNA sequencing can be performed. For the cohorts described in this thesis, RNA-seq was performed using Illumina dye sequencing. Sequencing takes place within a flow cell. The flow cell consists of lanes coated with oligonucleotides that are complementary to the adapter sequences that were previously ligated to the cDNA library. The first step in sequencing is to amplify the bound sequences through bridge amplification. This process uses PCR to create millions of clusters, each containing many cloned sequences, within the flow cell. Sequencing by synthesis can then take place in parallel for each cluster in the flow cell. This process begins with the attachment of a primer to the adapter of each sequence. Next, a single fluorescent nucleotide is added to each sequence that is complementary to the first base of the cDNA sequence. Each nucleotide (i.e., A, T, C and G) are bonded with a fluorophore tag of a different colour. This tag blocks the addition of further nucleotides, so it must be cleaved before sequencing can continue. Upon activation by a laser, the colour of the newly added nucleotide can be recorded by the sequencer for each cluster. At this point, the fluorophore tag may be removed, allowing the next base to be added. This process is repeated until a specific number of bases, or the entire

strand, has been sequenced. Sequencing can be performed for just one end of the cDNA sequences (single-end) or both (paired-end).

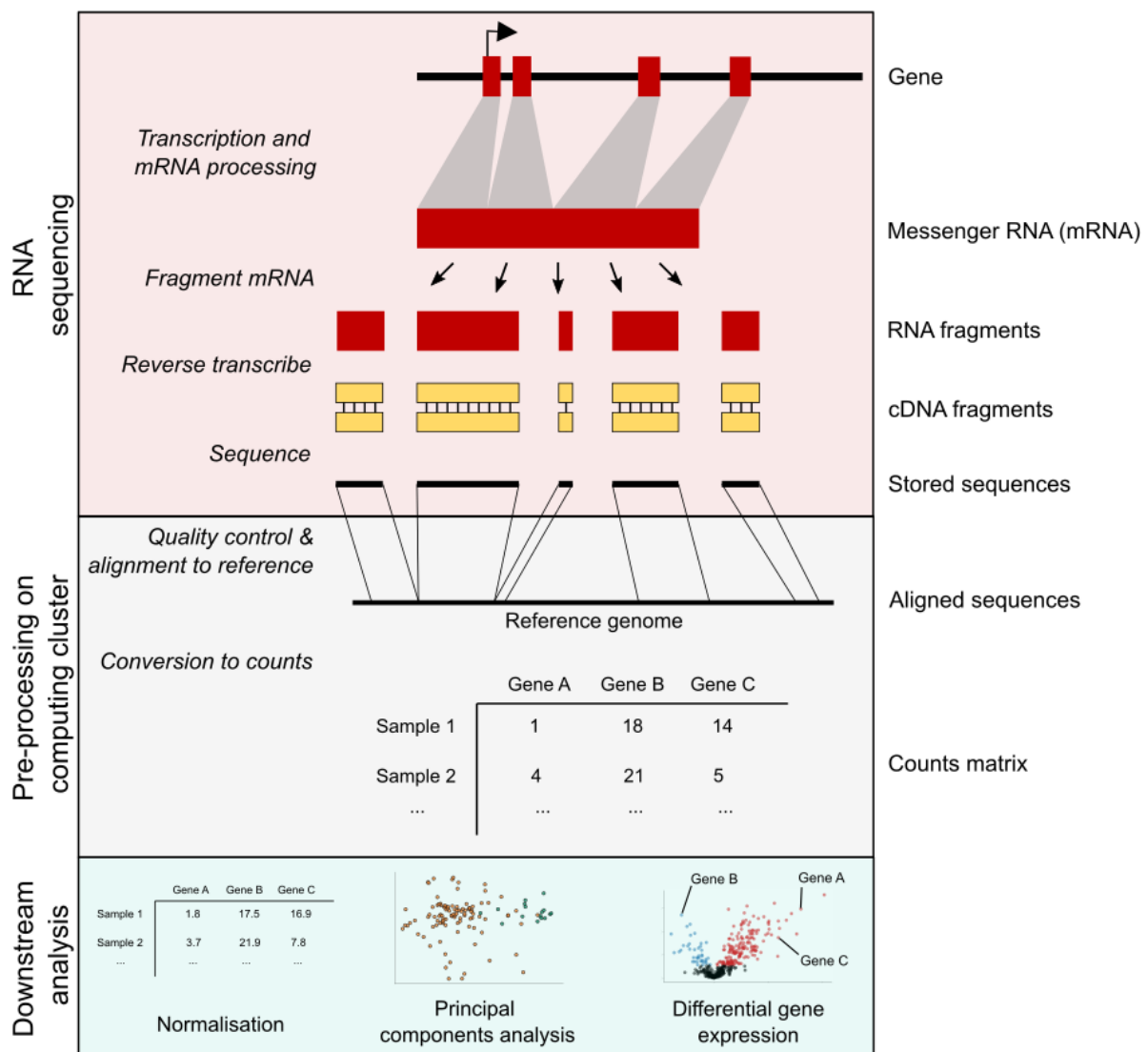


Figure 1.1 A broad overview of mRNA sequencing

An RNA sequencing experiment can be broadly divided into three main stages. Firstly, mRNA from a biological sample must be extracted and sequenced. Secondly, the sequences that are generated must be processed. This invariably involves quality control, alignment of sequences and quantification of the expression of genes. Sequence processing is computationally demanding, so these steps will usually be performed in parallel on a computing cluster. Finally, downstream analysis can be performed.

RNA-seq is commonly performed to quantify the abundance of transcripts in pooled cell populations from a particular tissue. Since it is performed for a heterogeneous mixture of many cells in bulk, it is commonly referred to as bulk RNA-seq. A more recently developed technique, single-cell RNA-seq, has become increasingly used to profile the transcriptomes of individual cells³¹. These data can additionally provide information on the presence, abundance and expression profiles of specific cell types.

1.2.2.4 Processing RNA sequences

High-throughput sequencing experiments generally store data as plain text in a file format known as FASTQ. Each sequence in a FASTQ file is described by four lines:

1. The first line begins with an “@” symbol and acts as a sequence identifier. This identifier may contain information about the sequencing run.
2. The second line contains the actual sequence, encoded as A, C, T, G and N, where N represents an unknown base.
3. The third line is a separator that does not contain any information.
4. The final line is a phred score encoded as ASCII characters. This line contains a character (i.e., a phred score) for each sequenced base.

Phred scores provide an estimation of the sequencing quality. From them, we can ascertain the probability of an incorrect base call. This probability (P) can be derived from the phred score (Q), as follows³²:

$$Q = -10 \log_{10} P$$

Therefore, a phred score of 10 indicates a 1/10 chance of an incorrect base call, while a phred score of 30 indicates an estimated 1/1,000 chance of an incorrect base call³². These scores are useful for identifying quality issues before downstream analysis. The FastQC program³³ is commonly used to perform quality control for these sequences. Among other things, this program provides a summary of the phred scores across all of the sequences in a FASTQ file. At this stage, quality issues may be identified for some sequences, particularly towards the end of the read. It is also common to find overrepresented sequences in the data, for instance where adapters have been sequenced. For this reason, RNA-seq pipelines commonly employ a trimming stage to trim poor-quality bases³⁴.

Following this, RNA sequences must be aligned (**Figure 1.1**). Usually, the reads are aligned to an existing reference genome for the organism being studied. In the absence of a reference genome or transcriptome, then *de novo* alignment is required. Since this thesis involves only human participants, a high-quality reference genome was available for alignment. The Splice Transcripts Alignment to a Reference (STAR)³⁵ application is commonly used for this purpose. As the name implies, STAR is a splice-aware aligner, meaning that it accounts for the fact that RNA transcripts will not align continuously in the genome since introns are not present in mature mRNA sequences and alternative splicing can lead to the skipping of exons. STAR is based on the concept of the “maximal mappable prefix” (MMP). An MMP is the longest sequence of a particular read that exactly matches one or more locations in the reference genome. In the first step, seed MMPs are found that match the first portion of the read. If the read does not fully map, for instance due to the presence of an intron, the MMP search is then repeated for the remaining portion of the read. In this way, STAR can extend the alignment through mismatches, using the previous MMP as an anchor. If the tail of a read is poor quality and does not have a good match in the genome, for instance due to the presence of an adapter or incorrect bases, STAR may ignore it. Each alignment STAR finds for a particular read is given a score, based on matches, mismatches, deletions and splice junction gaps³⁵. Finally, STAR will report the alignments that exceed a given threshold.

Alignment allows us to associate reads with known genes and is required to estimate the abundance of particular transcripts or genes. Since the positions of genes within the genome are known, automated software can be used to check their overlap with reads, thereby quantifying the number of reads that map to a given gene. Commonly used programs for this purpose include HTSeq-count³⁶ and featureCounts³⁷. By default, the HTSeq-count algorithm operates in union mode, meaning that a read is only counted if it overlaps a single gene, and is treated as ambiguous if the read simultaneously overlaps two or more genes. These programs output a sample by genes matrix of counts that can be used for downstream analysis. The Salmon program³⁸ represents an alternative approach and can be used to perform quantification without the need for an alignment step. Whereas alignment provides information on the precise base-to-base relationship between reads and the reference sequence, Salmon uses mapping to identify the origin of each read, but not the exact

alignment. As a result, Salmon can quantify the number of reads mapping to each transcript very quickly and accurately³⁸.

1.2.2.5 Applications

Transcriptomic profiling has revealed the dynamic nature of gene expression across different biological contexts. For instance, the various mechanisms that control gene expression give rise to unique transcriptional profiles in different tissues, which are essential for their functions. This is in part due to the action of tissue-specific transcription factors that control networks of co-expressed genes^{39,40}. The Genotype-Tissue expression project, which involved performing RNA sequencing (RNA-seq) for 54 tissues, has provided a wealth of information on the tissue-specific control of gene expression. These data revealed that the variation in expression observed between tissues (47%) is far higher than that between individuals (4%)⁴¹. Furthermore, while many genes display preferential expression in particular human tissues, only a few (~200) are exclusively expressed in a single tissue⁴¹. Single-cell RNA-seq (scRNA-seq) has since permitted more granular explorations of cell-specific expression profiles. Such studies have highlighted both shared and discordant transcriptional features of particular cell types between different tissues^{31,42}.

Changes in the transcriptome also occur in response to the environment⁴³. For instance, transcriptomics can be applied to identify genes involved in the bacterial stress response⁴⁴, or transcriptional changes in cancer cells in response to drug treatments⁴⁵. Furthermore, transcriptomics is often applied to investigate the effects of a treatment or disease in the tissues of human patients. Commonly, these studies will involve profiling cells extracted from blood samples. For instance, the bulk gene expression of blood leukocytes has been studied in the context of inflammation, such as in sepsis⁴⁶, systemic lupus erythematosus⁴⁷ and rheumatoid arthritis⁴⁸. These studies can identify both dysregulation of gene expression and changes in cell populations and, in the case of RNA-seq, reveal changes in splicing^{49,50}.

1.2.3 Proteomics

1.2.3.1 Protein abundance

While RNA can provide instructions for polypeptide synthesis, proteins themselves are the effector molecules that directly carry out most cellular functions. Following transcription, mature mRNA transcripts may proceed to the endoplasmic reticulum, where they can be

processed by ribosomes to initiate polypeptide synthesis. For some proteins, this process is sufficient to synthesise a functional protein; others comprise multiple polypeptide subunits or require further processing in the form of post-translational modifications. As discussed previously, the abundance of RNA transcripts is controlled by a multitude of processes. In turn, the cellular abundance of a particular protein is controlled to an extent by the abundance of the mRNA transcripts that encode it. However, studies based on various biological systems using a variety of experimental approaches have found that gene expression typically accounts for less than 50% of the variation in protein abundance⁵¹. The poor concordance observed in these studies could, in part, be explained by their limited sample sizes and measurement error inherent to the technological platforms used. However, it is clear from these data that a significant proportion of protein abundance is a result of processes downstream of mRNA.

The abundance of proteins is additionally controlled by translational regulation and degradation pathways⁵¹. A study of mammalian cells found that translation was the greatest predictor of protein abundance⁵² (as compared to transcription and protein degradation). In this study, protein degradation was less important, perhaps owing to the increased half-life of proteins relative to RNA (proteins having, on average, five times longer half-life⁵²). However, other studies point to the importance of protein stability in determining protein abundance⁵³. Furthermore, the abundance of proteins in extracellular fluids is influenced by additional mechanisms, including their rate of secretion and clearance. Overall, post-transcriptional mechanisms play an important role in regulating protein abundance. Therefore, their measurement provides data that is orthogonal to gene expression, motivating the direct study of proteins to understand the present state of a biological system.

Transcriptomics studies usually measure the abundance of transcripts within cells. For instance, populations of leukocytes or peripheral blood mononuclear cells can be isolated to understand their current transcriptional state. However, in the case of proteomics, studies commonly measure the abundance of proteins in extracellular fluids. For instance, instead of attempting to study protein abundance within cells in the blood, researchers will usually measure protein concentrations within plasma. When studying mediums such as plasma, not only is protein abundance affected by intracellular mechanisms of regulation but also the combined effects of secretion from multiple tissues. Most plasma proteins are synthesised in

the liver (including albumin, complement, fibrinogen and other coagulation factors). The remaining proteins found in plasma originate from a diverse range of tissues. These include immunoglobulins, which are synthesised by plasma cells. Some proteins are secreted into the blood in response to infection and inflammation, such as cytokines. These include Interleukin-6 (IL6) and Tumour necrosis factor alpha (TNF- α), which are secreted by immune cells in the blood. Many acute-phase reactants are produced by the liver, such as C-reactive protein (CRP). Some are released by other tissues, such as von Willebrand factor (vWF) and Plasminogen activator inhibitor, which are secreted by the vascular endothelium. Other proteins, that are canonically considered to be intracellular or membrane proteins, can also be found in the blood. While they are not actively secreted, these proteins may be released from cells due to tissue damage and cell death⁵⁴.

Many distinct processes contribute to the composition of proteins in plasma. On one hand, this complicates identifying the origin and function of particular proteins in the blood. On the other, proteins in plasma have the potential to reveal information about dysregulated processes in many other tissues. For instance, acute-phase reactants in plasma act as powerful diagnostic markers for detecting inflammation across the entire body.

1.2.3.2 Methods of measurement

While transcriptomics provides important information on the regulation of genes, most transcripts do not exert their effects directly. Given that there is limited correlation between mRNA and protein abundance, transcriptomics cannot provide a complete picture of the phenotypic state of a biological system. This motivates the study of proteins with proteomics. However, the wide-scale study of proteins remains challenging, and the proteomics field is considerably less mature than genomics and transcriptomics. High throughput and cost-effective methods that surveil the entire set of existing proteins are yet to be developed.

As was the case for studies of gene expression, protein abundance was originally studied in a highly targeted fashion⁵⁵. Early methods included electrophoresis-based techniques, such as SDS-PAGE, that allow for the separation of proteins by their molecular weight. Further characterisation of the separated proteins can be performed, either using antibody- or mass spectrometry-based methods. For instance, in a Western blot, electrophoresis is performed followed by immunostaining, allowing for the probing of separated proteins using antibodies.

ELISAs (enzyme-linked immunosorbent assays) are another popular method that involve quantifying the presence of proteins using enzyme-conjugated antibodies.

While these methods are still used routinely, their highly targeted nature make them unsuitable for proteomic profiling. One of the most popular high-throughput methods for proteomics is mass spectrometry (MS)⁵⁶. Prior to MS, peptides are usually fragmented in a shotgun approach using enzymes (e.g., trypsin) before liquid chromatography is used to separate them⁵⁷. MS can then be performed to estimate the mass of peptides in the mixture. Since the mass of a peptide alone is often insufficient for protein identification, tandem mass spectrometry is used. Collision-induced dissociation is performed to fragment the peptide, and a second round of MS (MS²) allows for the analysis of these fragments. The elution time, molecular weight and MS² spectrum of each peptide are commonly compared to known values from organism-specific databases. Since post-translational modifications change the mass of a protein, they can sometimes also be studied using MS-based methods⁵⁶. The intensity of peaks in MS allows for the relative quantification of identified proteins.

The other major category of proteomics technologies are the binding- or affinity-based methods. Traditional immunoassays measure protein abundance by measuring the binding of an antibody. This is a sensitive method of measuring protein abundance that is commonly used in diagnostic testing. However, in the case of proteomics, we are interested in measuring the abundance of many proteins simultaneously, prompting the development of multiplex assays. However, as the number of proteins measured increases, so too does the potential for cross-reactivity due to the non-specific binding of antibodies⁵⁸, making it difficult to develop multiplex assays for more than ~10 proteins. Proximal extension assays (PEA) were developed by Olink to overcome this limitation. The PEA technology involves the development of a pair of antibodies specific to each protein. Each pair of antibodies are attached to complementary oligonucleotides that have a unique sequence. As a result, DNA hybridisation should only occur if both antibodies bind to the same protein. Therefore, unlike in traditional immunoassays, cross-reactions are less likely to be detected. Hybridisation can be quantified by quantitative polymerase chain reaction (qPCR) or DNA sequencing. Olink Explore 3072, the most recent iteration of the Olink platform, allows for the measurement of approximately 3,000 proteins.

Another popular affinity-based method is based on modified oligonucleotide aptamers, rather than antibodies. Compared to antibodies, aptamers are more stable and less expensive. Currently, the multiplex SomaScan assays developed by SomaLogic employ over 7,000 aptamers (known as “SOMAmers”) to measure the abundance of over 6,000 unique proteins, approximately double that of the broadest Olink platform. These assays attempt to overcome the multiplexing cross-reactivity problem observed for antibody-based methods using binding kinetics, under the assumption that non-specific interactions will dissociate more rapidly⁵⁹. The bound SOMAmers can be quantified using a custom DNA microarray.

Comparison of the SomaScan and Olink assays reveal generally positive correlations between assays targeting the same protein, although only a minority of assays are strongly correlated (i.e., for less than 50% of proteins $r^2 \geq 0.7$)^{60–62}. Plasma protein abundance can be affected by genetic variants within or close to the gene encoding the protein. Identifying such an association, therefore, provides validation that the assay is truly measuring its target. A lower proportion of SOMAmers are associated with variants close to the gene encoding their targets compared to Olink reagents⁶², which may indicate higher specificity of Olink immunoassays. Advantages of the SomaScan platform include the fact it measures the proteome in greater breadth. The SomaScan assays may have lower intra-study variability compared to the Olink assays, as determined by measuring the coefficient of variation for technical replicates⁶².

These affinity-based assays are targeted methods, since they are designed to quantify a set of known proteins for which epitope-specific antibodies or aptamers have been developed. The proteins measured by these assays may be biased towards those that are known to be secreted, proteins of interest to potential customers and proteins for which antibodies are readily available. In contrast, MS-based shotgun proteomics is usually considered untargeted and can identify previously unknown proteins. Yet, paradoxically, affinity-based assays can provide a more comprehensive analysis of the proteome. Compared to affinity-based assays, MS methods suffer from poor sensitivity and a lower dynamic range. As a result, MS-based proteomics studies usually detect fewer proteins than affinity-based assays^{63,64}. Furthermore, since MS-based methods struggle to quantify proteins of low abundance, they usually fail to detect key inflammatory mediators such as cytokines in plasma. MS-based studies are better placed to detect the major constituents of the blood proteome, such as albumin, globulins, clotting factors and acute phase proteins⁶³. Additionally, some proteins will be detected in

some patients by MS but not others, leading to missing values; this is usually less of a problem for affinity-based assays⁶³. In summary, both affinity- and MS-based proteomics are biased towards the detection of different fractions of the proteome. However, affinity-based methods can detect important inflammatory mediators in plasma, making them practical for the study of inflammatory disease.

1.2.3.3 Applications

While the field of proteomics is not as well established as genomics and transcriptomics, proteomic profiling is becoming increasingly popular for investigating human disease. Most such studies measure protein abundance in plasma or serum. For instance, Olink and SomaScan assays have been used to characterise the response to inflammatory diseases, such as in common variable immunodeficiency⁶⁵, collagenous gastritis⁶⁶, systemic lupus erythematosus⁶⁷ and inflammatory bowel disease⁶⁸. Proteomics has also been employed to investigate the effect of drug treatments. The SomaScan assays were used to assess the effect of prednisone and infliximab treatment in patients with inflammatory bowel disease (IBD), which revealed downregulation of inflammatory cytokines and known markers of IBD progression⁶⁹. In another study, treatment of myocardial infarction patients with tocilizumab (which disrupts proinflammatory IL6 signaling by binding to IL6R) resulted in dysregulation of relevant inflammatory markers, including IL6 and IL6R⁷⁰.

1.3 Holistic approaches to analysing biological systems

1.3.1 Modelling molecular data

1.3.1.1 *Differential expression analysis*

Transcriptomic and proteomic profiling are commonly performed to identify molecular markers that have different average levels of abundance across two or more groups. For instance, in a case-control design, samples may be taken from a group of patients with a disease and healthy controls. The simplest approach is to compare each feature between the groups separately. This ubiquitous approach is referred to as differential expression analysis. In this thesis, I use the term “differential expression” when discussing transcriptomic features, whereas I use the term “differential abundance” for proteins, given that the abundance of proteins also reflects processes besides transcription (see **Section 1.2.3.1**). It should be acknowledged, however, that even the abundance of transcripts in a sample does not solely represent the process of expression (see **Section 1.2.2.1**).

Differential expression and abundance testing are usually performed using linear models, which are beholden to the following assumptions:

1. **Linearity**: The relationship between the independent variables and outcome must be linear and additive.
2. **Independence**: The residuals (error terms) should not be correlated.
3. **Multicollinearity**: The independent variables should not be highly correlated.
4. **Homoscedasticity**: The residuals must have constant variance (i.e., no mean-variance relationship).
5. **Normality**: The residuals must be normally distributed.

Linear models cannot capture the full complexity of gene expression and protein abundance, but they can provide a simplification of reality that can be useful for understanding biological processes. This general principle is described by statisticians using the popular aphorism “All models are wrong, but some are useful.” With this in mind, the regression assumptions should be considered with respect to their impact on their usefulness for understanding biological systems, rather than being followed dogmatically.

In this thesis, I present molecular analyses based on serial samples taken over time. This is at odds with the independence assumption. In a repeated measures study design, inter-individual variance is often greater than intra-individual⁷¹. For example, while the plasma abundance of a protein could vary greatly between individuals, we may observe very similar abundances in repeated samples taken from the same patient. Failure to model this correlation structure can lead to biased coefficients and standard errors, which generally leads to anti-conservative findings^{72,73}. So, in the case of repeated measures taken from the same individual, respecting the independence assumption is of great practical importance.

One option for considering this non-independence is to include the individual as a fixed effect; however, when the individual is confounded with a phenotype of interest, the parameters cannot be estimated because the linear model's design matrix is singular. In such cases, most standard workflows cannot perform valid hypothesis testing as they ignore the correlation structure between samples^{73,74}. To remedy this, researchers sometimes elect to discard their repeated measurements or take an average across each individual. This may be statistically valid but will result in a loss of information compared to fitting models to the entire dataset⁷³. The gold standard for modelling repeated measures data is with linear mixed models (LMMs). Standard linear models only permit the specification of fixed effects, which allow for the estimation of the mean effect of a covariate in the population. However, since it is an average, this effect is assumed to be the same for all individuals. LMMs additionally allow for the specification of random effects. These permit the estimation of individual-specific effects and consider the correlation between repeated measurements. Random effects may be used to estimate both individual-specific intercepts and slopes.

In the case of the Olink and SomaScan assays, whose measurements tend to be approximately normally distributed for each feature following log transformation, standard LMMs are often applied directly to their measurements⁷⁵. RNA-seq data, on the other hand, measures the number of transcripts in a sample and is therefore considered to be count data. These data require specialised normalisation and/or modelling procedures for valid differential expression analysis. Most studies analysing RNA-seq employ either the edgeR⁷⁶ or DESeq2⁷⁷ packages to compare groups of samples. These packages are convenient because data normalisation and modelling can be performed within a single software ecosystem. These methods have advantages for modelling RNA-seq data, because: i) they are both based on

generalised linear models with a negative binomial link function, which have been shown to be appropriate for transcript counts; and ii) information is shared across genes by estimating the variance parameters globally. Another method that performs well for RNA-seq data is limma-voom^{78,79}. The voom approach models these data as log2 transformed counts per million (logCPM) and incorporates their estimated mean-variance relationship to adjust for heteroscedasticity⁷⁸. However, none of these packages support the specification of random effects and so they are not appropriate for analysis of repeated measures designs.

One option is to fit LMMs directly to normalised counts data. Alternatively, there are a small number of specialist workflows designed for analysing RNA-seq experiments with non-independent samples. Some analysts have attempted to use the duplicateCorrelation functionality in the limma workflow to account for repeated measures⁷⁹. In such a model, the correlation between replicates is considered using a genome-wide dispersion estimate. Recently, however, an extension to the limma-voom framework, Dream, has been developed that estimates gene-level random effects⁷³. Using Dream has advantages over standard LMMs, as precision weights are used to model measurement error and the voom approach models the mean-variance relationship of the log counts⁷⁸. Alternatively, the glmmSeq package can be used for differential expression testing, which applies generalised LMMs (implemented in the glmer package) to model RNA-seq counts⁷⁴. The most popular packages for differential expression analysis are developed under the assumption that modelling counts directly with a negative binomial link function has an advantage over applying standard linear models to log transformed data^{76,77}. Other studies argue that weighted linear models can perform better in practice and that generalised linear mixed models with a negative binomial link may have limited power^{73,80–83}. Studies that have compared the performance of popular differential expression methods indicate that both approaches perform well in practice^{84,85}.

1.3.1.1.1 Multiple testing

For each molecule being tested, differential expression tests are typically applied to find differences between two or more groups. When considering a single statistical test, the *p*-value provides the probability of observing a difference at least as extreme as that actually observed, under the null hypothesis that there is no difference between the groups. Typically, researchers define a *p*-value cut-off point (referred to as alpha, or α) before performing an experiment. An alpha value of 0.05 is commonly chosen, in which case there is a 5% chance

of a false positive result (i.e., a type 1 error). In differential expression analysis, however, statistical tests are applied sequentially to a large set of molecules. If we were to compare 20,000 genes between two groups for which there were no true differences, we would observe p -values of less than 0.05 for approximately 1,000 genes. Therefore, p -values only provide appropriate control of the false-positive rate for a single test in isolation. This is referred to as the multiple testing problem.

It is essential to compensate for this phenomenon. Specifically, we wish to define a level of confidence for a group of tests. One way to do this is to control the Familywise Error rate (FWER). Whereas a p -value is the probability of a false-positive in a single test, the FWER allows us to control the probability of finding at least one false-positive in a group of tests. One of the most popular methods for controlling the FWER is the Bonferroni correction. Using this correction, we can define a new p -value cut-off by dividing the alpha by the number of tests performed (n). Using this formula (α / n), our original alpha of 0.05 would become 0.0000025 after Bonferroni correction for 20,000 genes. In GWAS, a p -value threshold of 5×10^{-8} has become the standard cut-off, which was calculated based on a Bonferroni correction for all independent common SNPs in the human genome. While the Bonferroni correction certainly provides sufficient control of the false positive rate, for many applications it is considered conservative because it leads to a high false negative rate (i.e., type 2 errors)⁸⁶.

In the context of differential expression analyses, researchers generally attempt to control the false discovery rate (FDR) instead. Whereas the Bonferroni correction controls the probability of at least one false positive, FDR-based methods attempt to control the proportion of discoveries that are false positives. For instance, if we were to identify 2,000 differentially expressed genes at an FDR threshold of 5%, there would be 100 false positives. Therefore, the FDR provides a balance that limits both false positive results and false negatives. One of the most popular methods for controlling FDR is the Benjamini-Hochberg (BH) approach⁸⁷. As for the Bonferroni correction, the p -value threshold can be modified to compensate for the multiple-testing problem. Alternatively, we can use the original p -value threshold (e.g., the commonly used α of 0.05) and adjust the p -values themselves instead⁸⁸. This procedure is outlined as follows:

1. Order the list of p -values from smallest to largest.

2. Multiply each p -value by the number of tests performed then divide by its ranking in the ordered list.
3. Replace any p -values that are greater than 1 with 1.
4. Compare each p -value to the subsequent value in the list. If the predecessor is greater, then set its value to the subsequent p -value.

1.3.1.2 *Survival analysis*

In differential expression and abundance analysis, expression or abundance are respectively modelled as an outcome of linear models. Conceptually, this allows us to understand the effect of predictor variables, such as the presence or severity of disease, on the molecule of interest. In this thesis, I also use LMMs to understand the effect of time from disease onset on gene expression and protein abundance. Survival analysis takes a different approach, where the dependent variable is the time to some event, such as the progression of disease or death. Such an analysis is usually carried out using Cox Proportional Hazards regression. These models are convenient because they consider the temporal aspect of a disease outcome and allow for the quantification of how predictors affect the risk of an event. Furthermore, these models are designed to consider censored outcomes. In the case of survival, the data are usually right censored, because it is rare that all of the patients will have died during the study period.

Many studies consider the impact of relatively stable covariates (i.e., time-invariant) on the outcome risk⁸⁹. For instance, variables such as age or sex might be considered as being fixed for the duration of the study. Alternatively, a study might measure molecular measurements for each patient at a particular time (e.g., at study enrolment) and use these as predictors in a survival model⁹⁰. This approach could demonstrate the prognostic value of a biomarker measured early in the disease course. However, these molecular factors are usually not stable throughout the disease course, and so their values at disease onset may differ greatly from measurements taken at the time of the event of interest. In studies where serial measurements are available, it is therefore of interest to make use of such data.

Joint models were designed to allow for the estimation of the effect of time-varying covariates, such as molecular markers, on survival^{91,92}. These models include two sub-models. The first is an LMM, which allows for the estimation of the dynamic profile of molecules over time.

LMMs are ideal for this purpose because the inclusion of random effects allows for the modelling of individual-specific profiles over time. They can also be used when sampling has been performed sporadically for each individual, as opposed to being taken at specific intervals. The second sub-model is a Cox model. The individual trajectories estimated using LMMs can be included in the survival sub-model, thus allowing for the estimation of the effect of time-varying covariates on time-to-outcome. The LMM component can estimate the uncertainty in the value of the covariate, which is explicitly acknowledged by the joint modelling procedure⁹². In summary, joint modelling combines LMMs with a Cox model, thereby allowing the estimation of the effect of a time-varying covariate on time-to-outcome. In this thesis, I use joint modelling to leverage longitudinal data for the identification of proteins that influence the risk of death during the course of COVID-19.

1.3.1.3 Predicting disease outcomes

Regression models (including linear models, logistic regression and Cox models) are ubiquitous for performing statistical inference. In the context of molecular data, these models are usually fit for a single molecule at a time, along with a small number of covariates. They can also be used for classification tasks, for instance to predict whether a person has a particular disease or is likely to face a particular outcome. However, the predictive performance of a single molecule is usually limited compared to a model that considers more predictors. These models are capable of being fit to larger numbers of predictors, however these models rarely generalise well to new data when the number of features greatly exceeds the number of samples (a problem that is commonly referred to as “ $p \gg n$ ”). This phenomenon can be understood using the concept of the bias-variance trade-off. Bias is the average difference between the model’s prediction and the true value, and so represents a systematic deviation from the true estimator. Models with high bias are considered to be simple and tend to underfit the data on which the model was fit (i.e., the “training” data). Variance is the sensitivity to small differences in these data. Models with high variance are generally more complex, and while they usually fit the training data very well, they may not perform well on unseen examples. Achieving a good balance between bias and variance (i.e., a model with appropriate complexity) leads to a model that can generalise well.

When fit to a high number of predictors, linear models may be overly complex, and for this reason they may overfit the training data and fail to generalise. To create a linear model with

good predictive performance, we must therefore reduce its complexity. Regularisation techniques can reduce the variance in regression models, at the cost of increased bias⁹³. The coefficients of standard linear models are estimated by minimising the sum of the squared differences between the actual values for the outcome and the predicted values (i.e., the residual sum of squares or RSS). Lasso and Ridge regression also attempt to minimise the RSS, but they each introduce an additional penalty term based on the model coefficients. Ridge regression introduces the L2 norm, which is equal to the square of the magnitude of the model's coefficients, whereas Lasso uses the L1 norm, which is equal to the absolute sum of the model's coefficients. In both Ridge and Lasso models, the coefficients will be shrunk towards zero; the key difference is that, in the former, the coefficients will never reach zero, whereas in the latter they will. Therefore, Lasso regression will effectively remove predictors from the model, thereby performing variable selection. In both models, the penalty term is usually multiplied by an additional parameter (λ), allowing the extent of regularisation to be controlled to ensure a generalisable solution. Elastic Net models are also commonly used, which simultaneously apply both the L1 and L2 penalties, thereby combining the features of both Lasso and Ridge.

These regularised regression techniques perform well for prediction tasks involving high-dimensional molecular data^{94,95}. A benefit of these, compared to more complex machine learning techniques, is that linear models are relatively easy to interpret; they are simply a linear combination of the input features. Lasso in particular is popular for molecular data, as it additionally performs feature selection. This allows researchers to identify small subsets of highly predictive features in their data. The features in molecular datasets are often, however, highly correlated with one another⁹⁶. In the case of two highly correlated features, lasso may retain one in the model, while shrinking the coefficient of the other to zero.

Since Lasso and Ridge are based on linear models, they are not generally able to model interactions and non-linear relationships between predictors. In contrast, various supervised learning methods can be used to model non-linear relationships, such as random forests. These models have also been successfully applied to regression and classification tasks based on molecular data^{97,98}. Random forests are based on decision trees. Each internal node of a decision tree represents a test based on the data, whilst branches from this node represent outcomes based on this test. The terminal (leaf) nodes represent the final decision,

which in a classification task will represent the predicted category, whereas in a regression task will be a prediction for the value of the outcome. The random forests algorithm involves growing a multitude of decision trees. At the prediction stage, each of these trees is allowed to vote on the final prediction. However, if an algorithm simply built a set of trees based on the full training set, one would likely observe the creation of highly correlated or even identical trees. For this reason, the random forests algorithm employs bagging to build each tree on a slightly different dataset, thereby leading to the creation of a variety of tree structures. Specifically, bagging involves taking a random sample of the training data (with replacement) before creating each tree. Additionally, a random subset of features is selected when choosing the feature at each node. This further increases the variation in the features selected by the trees.

1.3.1.4 Unsupervised approaches

Supervised learning algorithms, such as lasso regression and random forests, can be trained on labelled data to create a model that can predict the label of unseen data. This contrasts with unsupervised learning, which broadly refers to methods that capture patterns in data in the absence of known categories. Since these methods are not biased toward predetermined factors of interest, they can be particularly useful for exploratory analysis. For instance, principal components analysis (PCA) is a ubiquitous method for investigating the major sources of variation in -Omics datasets. PCA involves rotating high-dimensional data such that the first dimension captures the axis of maximal variance, the second is the orthogonal axis that captures the maximal remaining variance, and so on. These new dimensions are referred to as principal components (PCs) which are linear combinations of the original variables in the data.

Different populations can have systematic differences in allele frequencies, commonly due to ancestry, and these differences can be identified by PCA. If the groups being compared in GWAS have considerably different genetic backgrounds, this may confound associations between variants and traits. Therefore, PCs are commonly included as covariates to correct for population stratification in GWAS. PCA is also commonly used for the exploration of large molecular datasets. Such an analysis will reveal the major sources of variation in the data, which may be technical or biological in nature. When performing large-scale molecular profiling, samples are often processed in batches because these measurement methods can

often only process a limited number of samples at a time. However, this can lead to samples being processed under different conditions. Such conditions may include the operator, the day, the instruments and reagents used, and even the location in which the samples are processed. These factors can introduce sources of variation in the data, which are commonly referred to as batch effects. If these technical factors are large compared to other sources of variation in the data, then they will be revealed by PCA among the first components. This propensity to reveal and diagnose key technical factors makes it an essential step in the quality control of molecular data. PCA can also reveal important biological factors and may in some instances reveal groups of samples according to their molecular profiles. This may, for instance, be of interest for the identification of disease subtypes.

One drawback of PCA is that only two to three linear components can be effectively visualised at a time. This may not be a problem in studies where the first two components explain a significant proportion of the total variance, but for others PCA may fail to effectively visualise key patterns in the data. This limitation is particularly noticeable in studies that employ single-cell RNA-seq, where researchers generally attempt to visualise clusters of heterogeneous cell types in just two axes. Generally, there is a large amount of overlap between cell types in the first two PCs for single-cell expression data⁹⁹. For this reason, alternative dimensionality reduction methods such as t-distributed Stochastic Neighbourhood Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP) are commonly employed for visualisation. Unlike PCA, these methods allow for the modelling and visualisation of non-linear relationships in two dimensions. For many applications, UMAP has replaced t-SNE, partly owing to claims that the method better preserves distances between data points¹⁰⁰. However, criticisms have been levied against these methods for their failure to preserve either the local or global structure of the data or distances between observations¹⁰¹. Distortions introduced by these methods may limit the usefulness of these methods for the exploration and visualisation of high-dimensional data.

Studies are increasingly measuring multiple types of -Omics data with the hope that combining modalities will reveal insights beyond the sum of their parts; they also have the potential to reveal the interplay between different molecular factors. As a result, there is increasing interest in the development of methods for the integration and joint analysis of these data modalities. One such method is Multi-Omics Factor Analysis (MOFA)^{102,103}. More generally, factor

analysis is related to PCA in that it can be used to decompose high-dimensional data into a smaller number of variables. The two methods are generally computed differently, although their results often converge. Whereas PCA is merely a rotation of the existing data, factor analysis presumes the existence of a set of latent factors that underly the observed data. Conceptually, PCs are merely linear combinations of variables that explain a high amount of variance, whereas factor analysis aims to identify the source of variation in the data. MOFA extends these methods to identify factors that are shared across multiple data modalities¹⁰². These latent variables may be explored further by identifying molecular features that are aligned with the factor and by testing their association with relevant clinical parameters. For instance, MOFA was applied to infer a factor that predicted drug response in data from chronic lymphocytic leukaemia patients that was related to both a somatic mutation in the immunoglobulin heavy-chain variable region gene and downstream transcripts with known IGHV associations¹⁰².

Another key goal of unsupervised learning is clustering. It is commonly of interest to identify groups of samples based on molecular data. Such an approach could, for instance, reveal subgroups of patients with different underlying molecular processes. The identification of different biological mechanisms in disease could potentially inform a personalised medicine approach in which treatments are adapted to the patient. Two key choices to be made during such an analysis are the clustering method to be used and the number of clusters (K). The ideal clustering method may depend on the properties of the data being studied and the goal of the analysis. Selecting the correct value for K , however, remains challenging. One approach is consensus clustering, which employs a resampling-based approach to assess the stability of the identified clusters for different values of K ¹⁰⁴. Unfortunately, such methods are prejudiced towards higher values of K , since smaller clusters are inherently more stable, and may find evidence for the presence of clusters in unimodal data¹⁰⁵. To remedy this, John *et al.* developed a method that employs Monte Carlo to simulate reference data that preserves the covariance structure of the original dataset but for which there are no clusters (i.e., $K = 1$)¹⁰⁶. Cluster stability is measured using the proportion of ambiguous clustering (PAC) metric for both the reference and real data. For a series of values of K , the cluster stability is compared to that of the simulated null distribution. This allows for formal testing of the null hypothesis, that there is only a single cluster in the data ($K = 1$). Many studies attempt to cluster samples

based on molecular measurements to identify heterogeneous patient groups, however few appropriately consider the possibility that evidence of such structure is absent in their data. For instance, one study based on patients with Systemic Lupus Erythematosus used hierarchical clustering to stratify patients into seven groups based on their transcriptomes¹⁰⁷. However, subsequent investigation using the M3C method revealed the absence of significant clustering, and that selecting a K of 3 would have resulted in more stable clustering¹⁰⁶.

1.3.2 Integrating molecular measurements with genetic variants

1.3.2.1 Colocalising genetic traits

Many studies measuring large-scale molecular datasets are observational, in that the variable of interest (e.g., treatment or disease) is not controlled by the researchers. These studies are commonly used to investigate the effect of a treatment or disease compared to a control group; however, the assignment of individuals to these groups is not random, so there may be other differences between the groups being studied. For instance, some diseases, such as Systemic Lupus Erythematosus (SLE), are more common in females relative to males. A simple observational study might enrol a group of patients with SLE and compare them to a group of healthy individuals. However, because of this naïve recruitment strategy, there would likely be a greater proportion of females in the SLE group relative to the healthy group. Should the researchers perform transcriptomic profiling, it would be unsurprising if they found differential expression of genes on the Y chromosome between the groups. However, it would be incorrect to assume that the expression of these genes increases susceptibility to SLE. Rather, sex confounds the relationship between SLE susceptibility and the expression of genes on the Y chromosome. Some observational studies use matching to reduce the influence of confounders. For instance, in the case of SLE, controls could be selected to mirror the sex of the cases, which would alleviate confounding by sex. Alternatively, potential confounders can be statistically controlled for. However, we usually cannot perfectly characterise confounders, so there may be residual confounding due to measurement error. Furthermore, both matching and statistical adjustment require that all confounding variables are known to the researchers. We cannot verify that all confounders have been appropriately accounted for, so observational studies are usually limited to the identification of associations, as opposed to causal factors.

In contrast, randomised controlled trials (RCTs) involve performing an intervention. For instance, an RCT might involve the randomisation of enrolled patients to a placebo or a treatment arm. Randomisation allows for control over confounding variables. Additionally, reverse causation is avoided since the researchers perform the intervention. While these studies are ideal for identifying causal factors, RCTs are not always practical. For instance, an RCT comparing patients with SLE to healthy controls is not viable, as we would not be able to choose which individuals develop the disease.

While the transcriptome and proteome are subject to constant variability, genetic variants are fixed at conception. This property allows observational studies, such as GWAS, to avoid reverse causation. In this thesis I make use of Mendelian randomisation (MR), a method which uses genetic variants to make causal inferences; it considers the effect of a risk factor (the exposure) on another trait (the outcome). MR is built upon Mendel's laws of segregation, that alleles are inherited randomly from parents, and independent assortment, which states that alleles are inherited independently of each other (with the exception of Linkage Disequilibrium, see **Section 1.2.1.1**). MR can be considered akin to a "natural" RCT, in which the random segregation of alleles at birth is used to determine groupings. So, MR uses genetic variants as instrumental variables (IVs) for causal inference. Specifically, the causal effect of an exposure on some outcome can be estimated using a Wald ratio, which is based on the division of the association between the variant and outcome by the variant-exposure association. Additionally, if there are multiple instrumental variables, these ratios can be meta-analysed using inverse variance weighted (IVW) regression. These methods are, however, beholden to three key assumptions with respect to the IVs¹⁰⁸:

1. **Relevance assumption.** The IVs are strongly associated with the exposure.
2. **Independence assumption.** There are no unmeasured confounders between the IVs and the outcome.
3. **Exclusion restriction (ER).** The IVs only affect the outcome through the exposure (i.e., there is no horizontal pleiotropy).

In practice, the instrumental variables are often selected based on their association with the exposure, so the relevance assumption is relatively simple to assess. The independence assumption may be violated by systematic ancestry differences, although this can be mitigated

by adjusting for population stratification in the GWAS for the outcome. The ER assumption tends to garner more attention from MR practitioners, and several methods have been developed to assess its plausibility. An individual SNP can exhibit horizontal pleiotropy if it has a positive or negative effect on the outcome through a pathway other than the risk factor being modelled. In some cases, horizontal pleiotropy is balanced. This means that, although individual variants exhibit horizontal pleiotropy, their effects cancel out (i.e., their effects have a mean of zero). While strictly this violates the ER assumption, the estimates of standard MR approaches, such as IVW, will likely still be unbiased¹⁰⁹. More specifically, the fixed-effects IVW approach assumes the causal effect estimates are consistent across the IVs; however, random effects can be introduced to the IVW regression to account for balanced horizontal pleiotropy. In unbalanced (or “directional”) horizontal pleiotropy, the effects of the instruments on alternative pathways do not cancel out (i.e., their effects do not have a mean of zero). In this case, the results of both IVW approaches will be invalid. Unbalanced horizontal pleiotropy can, however, be accounted for using an intercept term, as is implemented by Egger regression. This method is still bound to the InSIDE assumption, however, which holds that the effects of variants on the exposure are not correlated with their effect on the alternative pathway(s)¹⁰⁹.

Linkage disequilibrium can also confound associations with variants, further complicating the interpretation of genetic associations. GWAS rarely identify associations for only a single variant at a particular locus. Instead, there are generally multiple variants identified at each region due to LD, which manifests visually as association peaks. It can be difficult to ascertain which variant at a particular locus causally affects the trait, as opposed to those variants that are simply correlated with the causal variant. This phenomenon also complicates the interpretation of MR studies, as it is difficult to establish whether the instrumental variable itself has a direct effect on the exposure and outcome, or if it is simply in LD with the relevant variant. Visual assessment of the regional association patterns at the loci for both the exposure and outcome might indicate a common association signal, but it is not a trivial task to rigorously evaluate whether these signals share the same underlying causal variant. Statistical methods have been developed for this purpose and are referred to as colocalisation tests. Many such methods assume that there is only a single variant in the region that is truly associated with

each of the measured traits^{110–112}. This is not always the case, and so recent efforts have focussed on considering multiple causal variants¹¹³.

1.3.2.2 *Investigating genetic variants using molecular quantitative trait loci*

GWAS have been successful in identifying links between genetic variants and traits, such as disease risk. However, the majority of variants are found outside of coding regions¹¹⁴, and so do not exert a direct effect on protein sequence. Consequently, a major challenge for GWAS has been in the interpretation of how variants exert their effect on biology. One way in which non-coding variants can affect a trait is by modulating the expression of genes. In fact, genetic variants can be tested for their association with the expression of a gene, as measured by transcriptomic technologies such as RNA-seq. Expression quantitative trait loci (eQTLs) have improved our understanding of transcriptional regulation and have become useful tools for describing the mechanisms by which variants exert their effect.

Colocalisation can be used to ascertain whether disease-causing variants also influence gene expression. If colocalisation is found, it may be interpreted as evidence that that gene is a causative factor. For example, colocalisation was applied to follow-up variants for 8 traits, including fasting glucose and fasting proinsulin, to eQTLs identified for 45 different tissues as part of the GTEx project^{115,116}. This procedure suggested causal roles for genes on fasting glucose and fasting proinsulin, such as *MADD*. *MADD* knockout mice were found to be hyperglycaemic¹¹⁷. So, colocalisation analysis provides causal evidence for the role of this gene in humans. Mendelian randomisation is also commonly applied to investigate links between eQTLs and disease. Specifically, eQTLs are used as instruments to assess the causative effect of gene expression (the exposure) on an outcome. For instance, Zhu *et al.*¹¹⁸ used MR to estimate the causal effect of 126 genes on 5 complex traits. These included a causal effect of *TRAF1* on rheumatoid arthritis (RA). The *TRAF1-C5* locus was among the first regions identified by GWAS of RA¹¹⁹. It was not clear which of these genes was functionally relevant, although *TRAF1* was believed to be the most likely candidate as the protein it encodes binds known mediators of RA¹²⁰. So, MR was likely able to prioritise the causal mediator at this locus.

While eQTL studies have been effective in identifying causal mediators, mRNA itself does not usually have a direct effect on disease. Its effect is generally mediated through proteins.

However, the correlation between transcript abundance and the concentration of the proteins they encode is limited (see **Section 1.2.3.1**). This motivates the use of protein abundance (i.e., pQTLs), rather than gene expression, to investigate the effects of genetic variants. Identifying proteins that cause disease also has the potential to identify drug-repurposing opportunities or novel targets. Genotyping and SomaScan proteomics performed for 3,301 participants resulted in the identification of 1,927 genotype-protein associations⁶. MR and colocalisation permitted the identification of proteins with putative causal effects on disease risk, including those that are the targets of existing drugs. For instance, a variant associated with higher plasma abundance of GP1BA, a platelet receptor for von Willebrand factor, was also associated with higher platelet count. A randomised trial of egaaptivon pegol, which inhibits vWF-GP1BA binding, indicated promise as an antithrombotic agent¹²¹. These data illustrate the potential for pQTLs to identify therapeutic targets.

1.4 The effect of COVID-19 on an at-risk population

1.4.1 COVID-19

Coronaviruses are a family of spherical RNA viruses that are named after the unique crown-like spikes on their surface. Some coronaviruses cause mild respiratory illness, including the common cold. Others can cause more severe respiratory illnesses, such as SARS (severe acute respiratory syndrome) and MERS (middle east respiratory syndrome). SARS-CoV-2 is a novel coronavirus strain that causes coronavirus disease 2019 (COVID-19), which was declared an ongoing pandemic by the World Health Organisation (WHO) as of March 2020¹²². SARS-CoV-2 typically incubates for 5 days, at which point symptoms may start to appear. A significant portion of individuals are asymptomatic or have only mild symptoms^{123,124}, including fever, fatigue and a dry cough¹²⁵. Some patients, however, develop severe disease characterised by marked inflammation and respiratory failure.

1.4.1.1 Clinical features

During the initial wave of SARS-CoV-2 infections (March 2020) in the US, most patients hospitalised with COVID-19 (75%) were over 50 years old, and the majority (54%) were male¹²⁶. 90% of these patients had at least one underlying condition, including hypertension (50%), obesity (48%), chronic lung disease (35%), diabetes mellitus (28%) and cardiovascular disease (28%)¹²⁶. These data provided early indications that older adults, males and those with particular underlying medical conditions may be disproportionately affected by COVID-19. Similar findings were made by an early (April-May 2020) study in the UK ($n = 20,133$)¹²⁷, which additionally reported that admitted patients were enriched for chronic kidney disease (16%). In this study, 41% of patients were discharged alive, while 26% died and 34% were continuing to receive care. As of 27/04/2023, there have been 224,448 deaths in the UK in which COVID-19 has been listed on the death certificate¹²⁸.

While most COVID-19 infections are mild, some patients will develop severe disease, involving dyspnoea, shortness of breath, and hypoxemia, an abnormally low level of oxygen in the blood¹²⁵. In some cases, the disease will progress to a critical state, commonly characterised by pneumonia, acute respiratory distress syndrome, septic shock, multi-organ failure and death¹²⁹. Other complications of severe COVID-19 include acute kidney, cardiac and liver injury, gastrointestinal symptoms, and coagulopathies¹³⁰.

1.4.1.2 Mechanism of infection

The receptor binding domain (RBD) of the SARS-CoV-2 spike glycoprotein (spike protein) binds to angiotensin-converting enzyme 2 (ACE2), a membrane-bound protein that acts as the host cellular receptor. To release its RNA into the cytoplasm, the virus must first fuse with the cell membrane of the host cell. This, however, requires that the spike protein is first activated by proteolytic cleavage mediated by host proteases¹³¹. There are two distinct pathways by which this can come about^{132,133}. In the first, endosomal entry, the ACE2/SARS-CoV-2 complex is engulfed by clathrin-mediated endocytosis. Given that the environment is acidic, cathepsins may cleave the spike protein, thereby activating it. The second, more common, pathway is not dependent on internalisation by endocytosis but requires the membrane protein TMPRSS2, which can cleave the spike protein at the cell surface. In either pathway, a fusion pore is formed, permitting the release of SARS-CoV-2 RNA into the cytoplasm. TMPRSS2-mediated cell entry is generally more efficient; however, the endosomal pathway allows for cell entry in its absence¹³³. Inside the cell, the viral RNA may be translated, after which the virus' genome may be replicated¹³². Virus-containing vesicles are assembled by the endoplasmic reticulum and Golgi. Finally, the contents of these vesicles are released by exocytosis.

SARS-CoV-2 is transmitted through respiratory droplets and aerosols, which when inhaled allow the virus to infect ACE2- and TMPRSS2-expressing ciliated airway cells in the upper respiratory tract¹²⁵ (i.e., the nasopharynx and trachea). The virus initially replicates locally in the upper airways, however at least in some cases the virus can migrate down the respiratory tract¹³⁴, a process which may be accelerated by the inhalation of viral particles. Alveolar-type 2 (AT2) epithelial cells, which express both ACE2 and TMPRSS2, are particularly susceptible to infection¹²⁵. SARS-CoV-2 may also spread to other tissues that express ACE2, such as myocardial cells and the cells of the proximal tubule in the kidney¹³⁵.

1.4.1.3 Pathogenesis of severe COVID-19

The initial response to infection by SARS-CoV-2 is facilitated by the innate immune system. Pattern recognition receptors (PRRs) allow for the early detection of viral particles. For instance, cytosolic pattern receptors (including toll-like receptors and retinoic acid-inducible gene 1-like receptors) can detect RNA viruses following cell entry¹³². These are expressed both by innate immune cells, such as macrophages, and airway epithelial cells. Detection of

the virus triggers the production and secretion of type 1 interferons (IFNs), which in turn activate antiviral defence mechanisms in neighbouring cells, which attempt to interfere with viral infection and encourage cell death. SARS-CoV-2 is a relatively weak IFN-1 inducer, in part owing to its ability to block IFN activation by PPRs. The SARS-CoV-2 ORF9b protein has been shown to suppress types 1 and 3 IFN production by interacting with RNA-sensing pathways¹³⁶, while ORF6, ORF8 and nucleocapsid proteins have been shown to antagonise IFN1 pathways¹³⁷. An insufficient or delayed IFN response fails to contain SARS-CoV-2, leading to viral spread and an excessive inflammatory response¹²⁵.

Following the initial stages of COVID-19, involving the innate immune response and damage caused directly by SARS-CoV-2, some patients exhibit a disproportionate inflammatory response. This is characterised by high levels of proinflammatory cytokines, including interleukin-6 (IL6), CCL2 and CXCL10, and other inflammatory markers, including D-dimer and C-reactive protein¹²⁵. This hyperinflammatory state is brought about by both innate and adaptive factors. SARS-CoV-2 specific antigens are presented by antigen-presenting cells, including macrophages and dendritic cells, leading to activation of T and B cells. The infiltration of the lungs by immune cells, such as macrophages and neutrophils, and the excessive production of cytokines by activated cells, including monocytes and helper (CD4+) T cells, can result in excessive local and systemic inflammation. These observations led to comparisons between COVID-19 and other disorders, such as cytokine storm syndrome (CSS). The comparison of COVID-19 to CSS has, however, come under criticism, in part because in COVID-19 the levels of inflammatory mediators such as IL6 generally do not reach the levels observed in these other syndromes^{138,139}. Nevertheless, both severe COVID-19 and CSS are characterised by the excessive release of proinflammatory cytokines resulting in the injury of host cells and the development of acute lung injury.

Endothelial and alveolar epithelial cell damage can disrupt the barrier between the alveoli and vasculature, leading to pulmonary oedema, the build-up of fluid from the circulation in extravascular spaces of the lungs. This can manifest clinically as dyspnoea and hypoxaemia, and can progress to acute respiratory distress syndrome (ARDS)¹⁴⁰, a form of respiratory failure involving widespread pulmonary inflammation. ARDS is commonly observed in patients with severe COVID-19 and is also known to develop following infection by SARS-CoV and MERS-CoV¹⁴¹. Inflammatory cytokines and other stimuli, such as hypoxia, can contribute to

alveolar epithelial and endothelial damage, leading to a “leaky” state^{125,140}. AT2 cells are directly targeted by SARS-CoV-2, leading to diffuse alveolar damage and impaired respiratory function. This allows circulating immune cells, inflammatory mediators, and damage-associated molecular patterns to infiltrate the lungs, leading to further epithelial disruption. Damage to the endothelium attracts circulating immune cells, which can in turn release inflammatory mediators and encourage endothelial activation, which further contributes to a proinflammatory state and triggers endothelial cell death in COVID-19 patients¹⁴². There is also evidence that endothelial cells, which express ACE2, could be directly targeted by SARS-CoV-2¹⁴².

Patients with COVID-19 have an elevated risk of thrombotic events, including pulmonary embolism and deep vein thrombosis^{143,144}. While healthy lung vasculature acts to prevent clotting, activated endothelial cells encourage a prothrombotic state. Markers of endothelial death and activation are released in patients with severe COVID-19¹⁴⁵, and these patients have elevated circulating endothelial cells indicating extensive disruption¹⁴⁶. D-dimer, a product of fibrin degradation, and von Willebrand factor (vWF), which mediates the adhesion of platelets and the endothelium, have been recognised as prognostic markers of COVID-19 severity^{147,148}. These observations highlight coagulation dysfunction as a feature of COVID-19 and its link with endothelial activation. Endothelial disruption is extensive in COVID-19, and this is likely a major contributor towards COVID-19-related thrombosis.

Neutrophils have been observed to accumulate in the alveolar space in both patients with ARDS and COVID-19^{140,149}. In the case of COVID-19, neutrophils are recruited early in the disease course due to viral-mediated injury, encouraged by the release of neutrophil chemotactic factors. Neutrophils also accumulate intravascularly in patients with severe COVID-19¹⁴⁹. These cells form neutrophil extracellular traps (NETs), webs of chromatin and granule proteins that aid in killing pathogens extracellularly. NETs are associated with severe COVID-19, and a higher proportion of neutrophils in these patients are of an active low-density form that are more likely to form NETs¹⁴⁹. While NETs have antiviral roles, they can also damage host cells, further contributing to the destruction of the alveolar epithelium and vascular endothelium. NETs can activate the vascular endothelium, leading to the induction of vWF and promoting coagulation¹⁵⁰. They additionally act as a scaffold for platelet adhesion

that can contribute to thrombosis. Components of the NETs themselves, such as histones, may directly interact with clotting factors¹⁵¹.

1.4.1.4 Long-term consequences of COVID-19

During the course of SARS-CoV-2 infection, there is an elevated risk of pulmonary, kidney, cardiac, liver and neurological complications^{129,130}. Additionally, it is becoming increasingly recognised that the elevated risk of some conditions extends beyond the acute phase of the illness. A study of individuals with COVID-19 and at least one thromboembolic event found that the relative risk of pulmonary embolism (PE) and deep vein thromboembolism (DVT) were elevated between 29-56 days following COVID-19 diagnosis¹⁴³. Another study found that the relative risk of DVT was elevated for 70 days following COVID-19, PE for 110 days, and bleeding for 60 days¹⁴⁴. There are also other cardiovascular manifestations following the resolution of acute COVID-19, including increased risk of myocarditis, pericarditis, heart failure, dysrhythmias, and heart disease.

Not only does COVID-19 confer an increased risk of long-term complications, but symptoms do not always fully resolve following the resolution of infection. Persisting symptoms are referred to as “long COVID” or “post-acute sequelae of COVID-19”, a condition which has been estimated to affect over 65 million individuals worldwide¹⁵². There are multiple different definitions and diagnostic criteria for long COVID¹⁵³, making it challenging to study and treat. The definitions loosely converge on the principle that long COVID involves ongoing symptoms that cannot be explained by other conditions, from approximately 8 weeks following COVID-19 onset. Common symptoms (occurring in at least 20% of cases) include fatigue, headache, attention disorder, hair loss, dyspnoea, loss of taste/smell and breathlessness after exercising¹⁵⁴. Long COVID often leads to persistent fatigue, and some patients fulfil the diagnostic criteria for myalgic encephalomyelitis/chronic fatigue syndrome¹⁵⁵. Markers of coagulation (e.g., D-dimer), cardiovascular damage (e.g., NT-pro-BNP) and inflammatory markers (e.g., CRP) have been observed in patients with long COVID^{154,155}, pointing to underlying molecular changes in these individuals.

1.4.1.5 Treatment and vaccination

A variety of treatments for COVID-19 have been investigated for COVID-19. Some target viral replication directly, whereas others attempt to modulate the response to infection¹⁵⁶. Antiviral

therapies include remdesivir, which was found to reduce COVID-19-related hospitalisation or death by 87% in the PINETREE randomised trial¹⁵⁷. Remdesivir inhibits SARS-CoV-2 RNA-dependent RNA polymerase, thereby reducing viral replication, including in airway epithelial cells. Also demonstrated to be effective at reducing the risk of hospitalisation or death is molnupiravir, which causes deleterious mutations to accumulate in the SARS-CoV-2 genome. Nirmatrelvir–ritonavir consists of nirmatrelvir, which targets the SARS-CoV-2 3-chymotrypsin-like cysteine protease enzyme, which is required by the virus for replication, and ritonavir, which amplifies the action of nirmatrelvir. Combined, these two treatments reduced the risk of severe COVID-19 by 89% in the EPIC-HR randomised trial¹⁵⁸. Anti-inflammatory therapies have also been demonstrated to be effective in the treatment of COVID-19. These include dexamethasone, the first treatment shown to reduce mortality from COVID-19 in a randomised trial¹⁵⁹. Dexamethasone is a corticosteroid that restrains inflammatory processes and was already in use for the treatment of inflammation. Also effective in reducing mortality is tocilizumab, which is a monoclonal antibody that specifically blocks the interleukin-6 receptor¹⁶⁰. Like dexamethasone, it dampens inflammatory pathways and is already applied in the treatment of inflammatory and autoimmune disorders, including rheumatoid arthritis and cytokine release syndrome. However, it has a more targeted effect in reducing interleukin-6 (IL6) signaling, the abundance of which is highly associated with COVID-19 severity¹⁶¹. It reduced mortality in randomised trials of hospitalised¹⁶⁰ and critically ill¹⁶² patients.

While these treatments are generally only used in the treatment of severe COVID-19, vaccines have been developed and widely disseminated. In the UK, vaccine effectiveness (i.e., the reduction in risk as a result of receiving a vaccine) against COVID-19-related hospitalisation was 52.2% for one dose and 77.6% for three doses¹⁶³. Furthermore, vaccine effectiveness against COVID-19 mortality was 58.7% for one dose and 93.2% for three doses. Since COVID-19 vaccination reduces the risk of severe COVID-19, and the incidence of long COVID is substantially elevated in hospitalised patients, it was hypothesised that vaccination prior to SARS-CoV-2 infection may reduce the risk of developing long COVID. A review investigating this hypothesis identified six relevant studies, all of which found that vaccines reduce the risk of developing long COVID in individuals with mild to moderate COVID-19¹⁶⁴. However, there have been inconsistent findings on the effect of vaccination for patients that have already developed long COVID.

1.4.2 End-stage kidney disease

The kidneys have an essential role in maintaining homeostasis by filtering waste and excess fluids from the blood. This filtration is carried out by microscopic structures known as nephrons, of which each kidney contains approximately 1 million. Within each nephron is a glomerulus, a tuft of capillaries that are surrounded by the Bowman's capsule. This structure permits small molecules in the blood, including water, glucose, and waste products, to move into the Bowman's capsule. The filtered fluid proceeds through tubules in the nephron, which permit water and other useful substances, such as glucose and amino acids, to be reabsorbed into the blood. Other substances, including waste products like urea and creatine, are excreted from the body as urine. If these substances are not removed, they accumulate in the blood and can cause damage. Such a build-up is referred to as uraemia, which is associated with a range of symptoms including fatigue, nausea, vomiting and muscle atrophy.

The amount of filtration required by the kidneys can change, for instance due to the intake of both food and fluids¹⁶⁵. The kidneys possess a functional reserve that allows them to adapt to these changes by increasing their glomerular filtration rate (GFR). This renal reserve allows the kidneys to adapt to short-term changes in filtration demands. However, more long-lived changes, such as changes in body mass during normal growth, obesity and pregnancy, require structural adaptations of the nephrons. While the kidneys cannot generate additional nephrons, existing nephrons can grow to increase their filtration rate. Higher renal load and glomerular hypertension can upregulate the epidermal growth factor receptor (EGFR) and its ligand transforming growth factor alpha (TGFA), which in turn leads to hypertrophy of the nephrons, thereby increasing the glomerular surface area available for filtration^{166,167}.

While renal hypertrophy can increase GFR, excessive growth can lead to barrier dysfunction in the glomerulus¹⁶⁵. Podocytes, epithelial cells that line the Bowman's capsule and permit filtration into the nephrons, are also required to grow in response to glomerular hypertrophy. However, excessive hypertrophy can lead to a sparsity of these cells, in part because sheer stress causes podocytes to detach from the epithelial lining. This now dysfunctional barrier permits the movement of proteins from the blood, leading to proteinuria, the presence of excessive levels of proteins in the urine. Renal tubules also undergo hypertrophy in response to renal hyperfiltration, and these increased filtration demands promote oxidative stress. A combination of proteinuria and increased cell stress can contribute to tubular cell loss and the

activation of inflammatory pathways. Cellular damage and inflammation trigger the induction of fibrotic pathways, including those mediated by transforming growth factor-beta (TGFA)¹⁶⁶. These pathways trigger the infiltration of the kidneys by immune cells and fibroblasts, which further contribute to inflammation, promote the deposition of extracellular matrix (ECM) and lead to scar formation in the kidneys. In summary, while there are mechanisms by which the kidneys can increase their filtration rate, excessive nephron hypertrophy leads to kidney damage and disruption. As a result, the demands on the remaining become even greater, in turn leading to further injury.

The mechanisms by which the kidney can adapt to changes in load and nephron loss are not always sufficient to maintain an adequate filtration rate. Chronic kidney disease (CKD) is defined by loss of kidney function, as measured by markers of proteinuria and GFR, or structural abnormalities that persist for at least three months. An estimated GFR (eGFR) of less than 90ml/min and an albumin:creatinine ratio (ACR) of less than 3mg/mmol are considered abnormal. A study of 68 global populations estimated the prevalence of CKD to be 13.4% (95% confidence interval [CI] of 11.7-15.1%)¹⁶⁸. Renal injury is irreversible, and kidney function can continue to decline in CKD patients leading to progression of the disease. End-stage kidney disease (ESKD) is the final stage of CKD, which is marked by almost complete renal failure (eGFR < 15ml/min), at which point the kidneys can no longer function on their own. While CKD incidence is higher in older people (>65 years) and women, younger people (≤65 years) and men are at greater risk of progressing to ESKD¹⁶⁸. ESKD is fatal without kidney replacement therapy, which involves either kidney transplantation or maintenance dialysis.

A variety of factors can contribute to the development of kidney disease. Major causes include hypertension and diabetes. High blood pressure and raised blood sugar can directly damage the glomeruli. Inflammatory conditions, such as systemic lupus erythematosus, can also cause damage to renal blood vessels leading to glomerulonephritis. Additionally, genetic factors can predispose individuals to CKD¹⁶⁹. For instance, mutations in the uromodulin (*UMOD*) gene can influence the development of kidney disease. *UMOD*, the most abundant protein in urine, is expressed in renal tubular cells where it may protect against inflammation.

1.4.3 Susceptibility of ESKD patients to severe COVID-19

Patients with ESKD have an elevated risk of severe COVID-19. A group of 59 ESKD patients admitted to a New York City medical centre with COVID-19 had poor outcomes relative to the general population, having a 31% overall mortality and 75% mortality for those that required mechanical ventilation¹⁷⁰. The susceptibility of ESKD patients to severe COVID-19 can be partially explained by the fact that these patients are enriched for other risk factors. Of the patients admitted to the New York medical centre, 98% suffered from hypertension, 69% had diabetes, 46% had coronary artery disease and 17% had pulmonary disease¹⁷⁰. However, even after considering common comorbidities, impaired kidney function emerges as a key risk factor for COVID-19-related mortality. An analysis of 17,278,392 NHS primary care records in England found that an eGFR of less than 30ml/min conferred a hazard ratio of 2.52 (95% CI of 2.33-2.72) for COVID-19-related death, after adjustment for related conditions including BMI, hypertension and diabetes¹⁷¹.

ESKD patients suffer from immune dysfunction, putting them at a higher risk of death from infections¹⁷². Individuals with ESKD simultaneously suffer from chronic low-grade inflammation but are also considered to be immunosuppressed. Renal damage and uraemia stimulate the release of proinflammatory cytokines, which in turn leads to chronic activation of the immune system and systemic inflammation. Despite this, ESKD patients suffer from deficiency of both innate and adaptive immunity and display an impaired ability to respond to pathogens. The populations of various immune cells are depleted in these individuals¹⁷³: neutrophils, monocytes and macrophages are depleted, leading to dysfunctional phagocytosis; dendritic cells are impaired, leading to a weakened response to infection; and populations of T and B cells can also be reduced, impeding the ability of these patients to launch an appropriate adaptive immune response. ESKD patients also display a weak immune response to vaccines, including those for influenza and SARS-CoV-2¹⁷³⁻¹⁷⁵, making them less effective in this population. Additionally, ESKD patients are usually required to attend regular haemodialysis sessions, putting them at increased risk of exposure to viral infections. In summary, there are numerous factors leading to both an increased risk of infection and progression to severe disease in ESKD patients. Therefore, these patients represent a vulnerable population that are important to study in the context of COVID-19.

1.5 Summary and objectives

In this dissertation, I characterised the molecular effects of severe COVID-19 using genomic, transcriptomic, and proteomic data. This work was carried out under the hypothesis that large-scale molecular investigations of COVID-19 have the potential to reveal insights and identify biomarkers that would otherwise be missed by studies targeting smaller groups of molecules. In this work, I explore markers of severe COVID-19 that are not typically measured by studies employing targeted molecular assays, which tend to focus on genes and proteins that are already known to be involved in the disease process, such as cytokines.

The primary cohorts studied in this dissertation involved ESKD patients with COVID-19, all of whom were undergoing regular dialysis treatments. I hypothesised that, by using ESKD patients as controls, the molecular signatures identified in this work will be generalisable to otherwise healthy patients with COVID-19. I tested this hypothesis by comparing my results with those based on cohorts of patients without kidney disease. I also explored the extent to which ESKD and dialysis affects molecules in the blood by comparing these patients to healthy volunteers with normal renal function. These data allowed me to consider the contribution of these factors to my main findings and provided a molecular characterisation of a group of patients that have not been well-studied using -omics technologies.

The specific aims that I attempt to satisfy throughout this thesis are as follows:

1. Identify PBMC transcripts and plasma proteins that are dysregulated in COVID-19. To this end, I will:
 - i. Identify molecules that are dysregulated in ESKD patients with COVID-19 relative to non-infected ESKD controls.
 - ii. Identify molecules that are associated with COVID-19 severity.
 - iii. Attempt to understand the extent to which the response of ESKD patients to SARS-CoV-2 infection reflects that of otherwise healthy patients.
 - iv. Characterise the temporal dynamics of molecules following COVID-19 onset.
2. Investigate whether proteins have a causal role in the susceptibility to severe COVID-19 by incorporating genomic and proteomic data.
3. Identify PBMC transcripts and plasma proteins that are dysregulated in ESKD patients on dialysis, relative to otherwise healthy patients.

In Chapter 2, I begin to realise the first aim through the investigation of 436 plasma proteins in a cohort of ESKD patients with COVID-19. One outcome of this chapter is the identification of proteins associated with COVID-19 severity. Chapter 3 attempts to extend Chapter 2 and fulfil the second aim by using genetic data to assess the causal effect of these proteins on susceptibility to severe COVID-19. Chapter 4 also extends Chapter 2 and further satisfies the first aim by investigating 6,323 plasma proteins and the PBMC transcriptome in two cohorts of ESKD patients with COVID-19. The third aim was accomplished by the comparison of non-infected ESKD controls to a healthy cohort in Chapters 2 and 4.

2 Longitudinal proteomic profiling of dialysis patients with COVID-19

2.1 Chapter summary

Whilst a significant portion of individuals infected by SARS-CoV-2 are asymptomatic or have only mild symptoms, a subset of patients develop severe disease, involving marked inflammation and thrombosis. Characterising the mechanisms of severe illness will grant us an improved understanding of COVID-19 and may reveal therapeutic targets. End-stage kidney disease (ESKD) patients are at a particularly high risk of severe COVID-19. Despite this, ESKD patients cannot shield as they must continue to receive haemodialysis treatment. These patients provided a unique opportunity for longitudinal sampling from early symptoms through hospitalisation. Serial blood sampling was performed for 55 ESKD patients infected by SARS-CoV-2 in April and May of 2020 ($n = 256$ serial samples). Also recruited were 51 non-infected ESKD controls and 35 healthy controls. A further 46 COVID-19 patients were enrolled to form a validation cohort. The levels of 436 circulating proteins were measured using Olink immunoassays. These proteins were selected for their roles in inflammatory and cardiovascular processes, allowing for the characterisation of key mediators of disease.

I compared COVID-19 cases and controls, which revealed 221 differentially abundant proteins. These results were consistent with both a validation cohort of ESKD patients and external datasets that included patients with normal renal function. 203 proteins were associated with COVID-19 severity, which included IL6, a proinflammatory cytokine. I applied supervised learning to predict COVID-19 severity and extract key biomarkers, which included KRT19, a marker of epithelial injury. The availability of dense serial samples permitted the modelling of temporal protein profiles following COVID-19 onset. Thirty-two proteins had distinct temporal profiles in severe disease compared to mild disease, including ACE2, the receptor for SARS-CoV-2. Through survival analysis with joint models, I found 69 markers associated with COVID-19-related death. This study characterised the dynamic profiles of proteins in severe COVID-19 and provides insight into the processes dysregulated in a group of at-risk patients. These findings were published in *eLife*¹⁷⁶.

in this study, I also compared ESKD patients on dialysis to healthy controls, finding marked dysregulation of the proteome. The most dysregulated proteins had clear links to renal injury, such as fibroblast growth factor 23, a hormone that stimulates phosphate reabsorption by the kidney, and Cystatin C, a marker of renal function.

2.2 Introduction

2.2.1 Longitudinal proteomic studies of COVID-19

A multitude of studies have investigated the COVID-19 plasma proteome, many of which were cross-sectional and simply considered proteomic differences between COVID-19 cases and controls, or identified proteins correlated with disease severity. These studies are observational and are thus limited by the fact they cannot discern between proteins that drive disease and those that are simply downstream markers of the response to COVID-19. Longitudinal studies, which take multiple molecular measurements over time from the same individuals, are also observational and so causality cannot be strictly evaluated. However, these studies may be better at prioritising causal factors as they allow for the identification of early markers of clinical deterioration through the temporal ordering of molecular changes. In other words, if a protein changes in abundance before clinical deterioration it may be more likely to drive disease.

However, longitudinal studies of individuals infected with SARS-CoV-2 were logistically challenging during the initial stages of the COVID-19 pandemic, as lockdowns made obtaining samples from non-hospitalised patients challenging. Serial measurements could be taken from hospitalised patients, but such studies fail to capture the full spectrum of COVID-19 severity. Patients undergoing haemodialysis, on the other hand, were unable to shield because they had to attend regular haemodialysis appointments. In the study described in this chapter, this permitted the longitudinal proteomic profiling of both hospitalised and non-hospitalised patients with COVID-19.

A small proportion of proteomics studies of COVID-19 included serial measurements^{177–182}. Here, I will focus on such investigations that were performed prior to or concurrently with the present study. One such study, carried out by Lucas *et al.*¹⁷⁷, used array-based technologies to measure the abundance of cytokines and chemokines in the serum of COVID-19 patients

admitted to Yale New Haven Hospital. Longitudinal protein measurements were analysed in two ways. The first method involved grouping timepoints into 5-day bins, followed by calculation and visual inspection of error bars (calculated using the standard error of the mean). The second method involved grouping patients into “mild” and “severe” disease severity categories, before performing linear regression for each group. While these approaches may be sufficient for identifying general trends, they are not ideal because they do not statistically account for the correlation between repeated measurements taken from the same individuals. Furthermore, regressions performed separately for two groups are not sufficient to identify severity-dependent trajectories. In another study by Haljasmägi *et al.*¹⁷⁸, plasma proteins were profiled for hospitalised COVID-19 patients using both Olink and other assays. This study employed a similar modelling method, in which the authors applied loess and linear regression separately to the intensive care unit (ICU) and non-ICU patient groups. Loess is a regression method that fits a smooth non-linear curve by fitting standard linear models to local subsets of data. While linear regression and loess can visualise general trends in the data, they do not take into account correlation between repeated measurements, so they are not suitable for testing for the presence of severity-dependent profiles.

Demichev *et al.*¹⁸¹ modelled the longitudinal trends of plasma proteins for COVID-19 patients admitted to the Charité University Hospital. However, this modelling approach only considered two samples for each patient: the first and last sampling time points at the peak of the disease. This approach allowed the researchers to identify proteins that have dynamic profiles contemporaneous with clinical changes, which is certainly of interest. However, it did not allow for the modelling of protein profiles across the entirety of the disease course. The authors did attempt to do so with loess, which was performed separately for patients that did and did not require invasive mechanical ventilation, but as discussed previously this approach is not optimal for modelling repeated measurements and identifying group-dependent trajectories. Shu *et al.*¹⁷⁹ used mass spectrometry to perform plasma proteomic profiling for patients with COVID-19 at multiple timepoints. However, comparisons were mostly performed between COVID-19 cases and healthy controls. The lack of longitudinal modelling is likely explained by the fact that there were very few patients in each cohort. Rodriguez *et al.*¹⁸⁰ performed Olink plasma proteomics for 17 patients hospitalised with COVID-19. Repeated measurements were obtained for some patients, allowing for the visualisation of their individual protein profiles

following admission. However, as the authors explain in their discussion, the study did not have sufficient statistical power to model severity-specific protein profiles. A study by Thwaites *et al.*¹⁶⁰ involved the plasma proteomic profiling of 510 COVID-19 patients. Linear modelling was applied separately to patients with severe and moderate disease, similarly to some of the aforementioned studies.

In my opinion, the study that most successfully modelled the longitudinal trajectories of plasma proteins following COVID-19 onset was that of Filbin *et al.*⁷⁵. This study involved 384 patients admitted with acute respiratory distress to the Massachusetts General Hospital. Samples were collected on days 0, 3, 7 and 28. The authors employed linear mixed models to identify plasma proteins that had significantly different longitudinal profiles depending on disease severity. Specifically, they modelled the interaction between time and severity. Proteins with severity-dependent trajectories included IL6 and ACE2.

In summary, while several studies collected serial measurements of plasma proteins, very few were able to effectively model their temporal profiles. In the present study, sufficient serial samples were collected for the robust modelling of protein trajectories following COVID-19 onset. I used a similar approach to that employed by Filbin *et al.*⁷⁵. A key difference, however, is that while Filbin *et al.* limited their analysis to fixed timepoints in the first week of COVID-19, I was able to model the temporal dynamics of proteins for three weeks following COVID-19 onset.

2.2.2 The effect of ESKD and dialysis on plasma protein abundance

The primary goal of the present study was to identify proteins associated with COVID-19. However, given that the study consists of ESKD patients on dialysis, it was unclear the extent to which its findings were applicable to patients with normal kidney function. There are a dearth of studies applying broad proteomic profiling to ESKD patients, and so the molecular effects of very severe kidney disease and dialysis are not clear. Olink assays were additionally applied to a group of healthy controls, allowing for the characterisation of the proteome of ESKD patients on dialysis and a better understanding of how these patients may be affected by COVID-19.

Many studies have investigated proteomic associations with estimated glomerular filtration rate (eGFR). However, most of these studies have been carried out in cohorts with normal

renal function or early-stage CKD. One such study measured 1,054 circulating proteins using the SomaScan platform in 938 patients with stable coronary heart disease¹⁸³, 258 of which had CKD while 679 did not. 709 proteins were associated with eGFR (although multiple testing correction was not performed) and 21 proteins were prognostic of cardiovascular risk in patients with CKD. Among these were tumour necrosis factor receptor superfamily members (TNFRSF1A, EDA2R, TNFRSF9), fibroblast growth factor 7 (FGF7), beta-2-microglobulin (B2M) and the epidermal growth factor receptor (EGFR), a protein expressed in the kidneys in response to kidney damage¹⁸⁴. Another study, also based on the SomaScan platform, considered 993 plasma proteins measured for 2,882 individuals from four cohorts¹⁸⁵; only 196 participants, however, had CKD. They found 57 plasma proteins to be robustly associated with eGFR, including cystatin C (CST3), B2M, insulin growth factor binding protein 6 (IGFBP-6) and tumour necrosis factor receptor superfamily member 19L (RELTL). Follow-up with Mendelian randomisation (MR) analysis revealed a potential causal effect of melanoma inhibitory activity protein (MIA), cystatin M (CST6) and carbonic anhydrase III (CA3) on eGFR. A study that applied proteomic profiling of 80 plasma proteins in older adults identified 28 to be significantly associated with eGFR, including cathepsins D and L (CTSD, CTSL), growth/differentiation factor 15 (GDF15), fatty acid binding protein 4 (FABP4) and fibroblast growth factor 23 (FGF23)¹⁸⁶. Among the 1,047 participants of this study, 409 had normal kidney function and none had ESKD. A study carried out for 389 individuals (including 15 participants with ESKD) in Sweden measured 2,893 plasma proteins using SomaScan assays¹⁸⁷. The authors identified proteins correlated with eGFR, including insulin-like growth factor-binding protein 6 (IGFBP6), CST3 and B2M.

Some proteomic studies instead focussed on early markers that are prognostic of adverse kidney outcomes. Niewczas *et al.*¹⁸⁸ examined 194 plasma proteins in three cohorts of patients with diabetes. 17 proteins were identified that were associated with a 10-year risk of developing ESKD, six of which were TNFRSF members. Another study considered the association of 4,877 plasma proteins with the development of adverse kidney outcomes in 9,406 adults¹⁸⁹. Thirteen proteins were significantly associated with a 50% eGFR decline in both follow-up periods, including follistatin-related protein 3 (FSTK3) and TNFRSFs.

While these studies have been effective in identifying markers of renal function and predictors of decline, many of them have been carried out in patients with normal function and few have

included patients with severe CKD or ESKD. One study measured 614 plasma proteins in 42 patients with chronic kidney disease (CKD), allowing for the comparison of early- and late-stage CKD¹⁹⁰. 60 proteins varied significantly between the groups, including CST3, angiopoietin-2 (ANGPT2), matrix metalloproteinase 7 (MMP7), FSTL3 and B2M. Romanova *et al.*¹⁹¹ compared 26 adults with ESKD to 10 healthy controls, finding dysregulation of 82 proteins, including cytokines, apolipoproteins (APOA1, APOA4, APOE) and Heat Shock Protein 90 Beta Family Member 2 (HSP90B2). Proteomic analysis of 14 moderate (stage 2-3) CKD patients and 15 ESKD patients on dialysis revealed 39 dysregulated proteins in ESKD compared to the CKD patients. These included IGFBP6, B2M, haemoglobin subunit beta (HBB), coagulation factor XII (F12) and complement-related proteins (e.g., CFD, C1S). Some studies have specifically attempted to investigate the effect of dialysis on the plasma proteome. It has been shown that there are proteomic differences between patients undergoing high-flux haemodialysis vs online haemodiafiltration¹⁹² and between high- and low-flux haemodialysis, which included immune-related proteins and complement¹⁹³. In summary, there have been few well-powered studies assessing the effects of ESKD and dialysis on the plasma proteome.

2.2.3 Contributions

Most of the results described in this chapter were published in *eLife*¹⁷⁶. This was, of course, a collaborative project and full details of individual contributions are provided in the publication. Candice L Clarke, Nicholas Medjeral-Thomas, Talat H Malik, Paige M Mortimer, Norzawani B Buang, Marie Pereira, Frederic Toulza, Ester Fagnano, Marie-Anne Mawhin, Emma E Dutton, Lunnathaya Tapeng and Marina Botto processed the samples. Candice L Clarke, Nicholas Medjeral-Thomas, Shanice Lewis, Eleanor Sandhu, Stephen P McAdoo, Maria F Predecki, Michelle Willicombe, David C Thomas and James E Peters had roles in data curation, patient recruitment, sample collection and clinical phenotyping. Artemis Papadaki contributed to the statistical analysis. Brian Tom, Jessica Barrett, Arianne C Richard and Paul DW Kirk provided advice on statistical approaches. The study was conceived and designed by Matthew C Pickering, Marina Botto, Michelle Willicombe, David C Thomas and James E Peters. Many authors also played a significant role in the interpretation of the data. I would also like to thank the patients who volunteered for this study and the staff at Imperial College Healthcare NHS Trust.

I performed the analyses described in this chapter and created the figures and tables. Some of the text, figures and tables in this chapter are repurposed from the *eLife* manuscript¹⁷⁶, which was published under the CC BY 4.0 license.

The CC BY 4.0 license is available at <https://creativecommons.org/licenses/by/4.0/>.

2.3 Methods

2.3.1 Ethical approval

All participants (patients and controls) were recruited from the Imperial College Renal and Transplant Centre and its satellite dialysis units in London, and provided written informed consent before participation. Study ethics were reviewed by the UK National Health Service (NHS) Health Research Authority (HRA) and Health and Care Research Wales (HCRW) Research Ethics Committee (reference 20/WA/0123: The impact of COVID-19 on patients with renal disease and immunosuppressed patients). Ethical approval was given.

2.3.2 Main cohort

55 COVID-19 positive haemodialysis patients were recruited, either as outpatients or as inpatients (**Table 2.1**). All patients were receiving in-centre outpatient haemodialysis before COVID-19 diagnosis. COVID-19 was confirmed in all cases with positive nasal PCR for the SARS-CoV-2 virus. Patients were recruited during the first UK national lockdown, with recruitment from 8th April – 30th May 2020. Blood was collected in EDTA tubes and centrifuged to obtain plasma, before being stored at –80°C. Sample processing was performed within 4 hours of venepuncture. Where feasible, serial blood samples were taken; in total, 259 samples were obtained, although 3 subsequently failed quality control (see **Section 2.3.6**). The median number of serial samples was 5 (range 1-10). In addition, 51 COVID-19 negative haemodialysis controls were recruited. COVID-19 negative haemodialysis controls were selected to mirror the cases in terms of demographic features (age, sex, ethnicity) (**Figure 2.1, Table 2.1**). A set of 35 healthy volunteers were also recruited who did not have known kidney problems (**Table 2.2**). These samples were taken from staff at Imperial College London and the National Health Service. Neither the ESKD controls nor the healthy controls displayed clinical features of any other infection.

Table 2.1 Characteristics of the Main cohort

Patient subsets were defined according to their peak clinical severity over the course of the illness. 'South Asian' represents individuals with Indian, Pakistani, or Bangladeshi ancestry.

Q1 = lower quartile. Q3 = upper quartile. GN = glomerulonephritis. HTN = hypertension. N/A = not applicable.

*One patient had type 1 diabetes, while the remainder had type 2.

	COVID-19 positive ESKD patients (n = 55)			ESKD controls (n = 51)
	Overall	Patients with peak severity of mild or moderate (n = 28)	Patients with peak severity of severe or critical (n = 27)	
Age Median (Q1-Q3)	72.2 62.5-77.3	73.4 65.5-76.4	68.5 61.8-78.8	70.1 62.2-75.1
Sex M F	39 (70.9%) 16 (29.1%)	18 (64.3%) 10 (35.7%)	21 (77.8%) 6 (22.2%)	36 (70.6%) 15 (29.4%)
Ethnicity White Black South Asian Asian (other) Other	16 (29.1%) 8 (14.5%) 18 (32.7%) 4 (7.3%) 9 (16.4%)	5 (17.9%) 5 (17.9%) 10 (35.7%) 1 (3.6%) 7 (25.0%)	11 (40.7%) 3 (11.1%) 8 (29.6%) 3 (11.1%) 2 (7.4%)	13 (25.5%) 8 (15.7%) 20 (39.2%) 3 (5.9%) 7 (13.7%)
Diabetes	34 (61.8%) *	16 (57.1%)	18 (66.7%)	24 (47.1%) *
Current Smoker	1 (1.8%)	1 (3.6%)	0 (0%)	0 (0%)
ESKD Cause DN Genetic GN HTN/Vascular Other Unknown	29 (52.7%) 1 (1.8%) 3 (5.5%) 5 (9.1%) 8 (14.5%) 9 (16.4%)	14 (50.0%) 1 (3.6%) 1 (3.6%) 3 (5.5%) 5 (17.9%) 4 (14.3%)	15 (55.6%) 0 (0%) 2 (7.4%) 2 (7.4%) 3 (11.1%) 5 (18.5%)	20 (39.2%) 1 (2.0%) 9 (17.6%) 7 (13.7%) 4 (7.8%) 10 (19.6%)
Hospitalisation due to COVID-19†	33 (60%)	6 (21.4%)	27 (100%)	N/A
Fatal COVID-19	9 (16.3%)	0 (0%)	9 (33.3%)	N/A

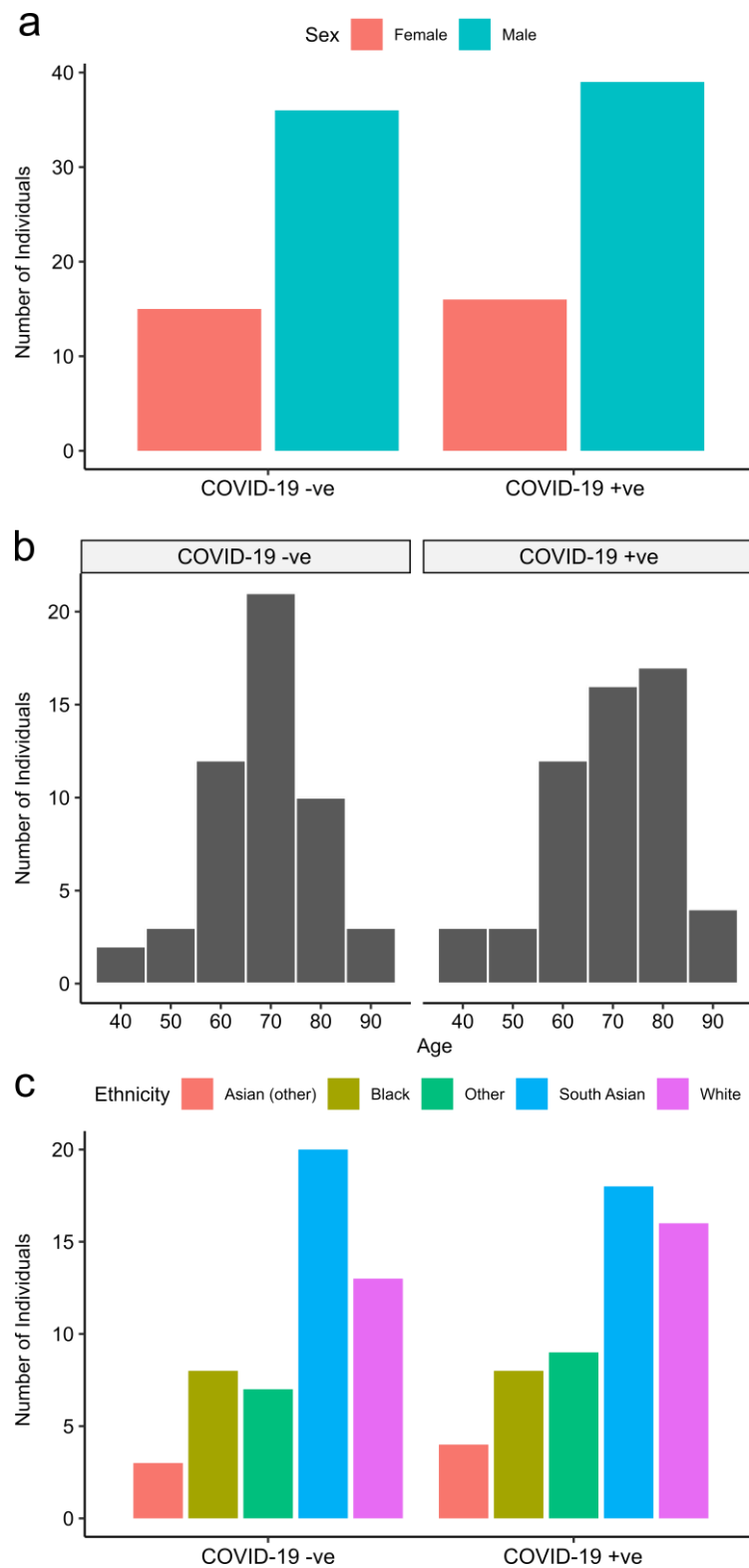


Figure 2.1 Baseline characteristics of the Main cohort

The number of COVID-19 positive and negative patients (plasma) stratified by: **a**) sex, **b**) age, and **c**) ethnicity.

Table 2.2 Characteristics of the controls

'South Asian' represents individuals with Indian, Pakistani, or Bangladeshi ancestry.

Q1 = lower quartile. Q3 = upper quartile.

	ESKD controls (n = 51)	Healthy controls (n = 35)
Age		
Median	70.1	46.6
(Q1-Q3)	62.2-75.1	35.8-54.2
Sex		
M	36 (70.6%)	18 (51.4%)
F	15 (29.4%)	17 (48.6%)
Ethnicity		
White	13 (25.5%)	8 (22.9%)
Black	8 (15.7%)	8 (22.9%)
South Asian	20 (39.2%)	9 (25.7%)
Asian (other)	3 (5.9%)	8 (22.9%)
Other	7 (13.7%)	2 (5.7%)

2.3.3 Validation cohort

A separate set of 46 COVID-19 positive haemodialysis patients were also recruited (**Table 2.3**). These patients were recruited from the same centre, but slightly earlier than the Main cohort (recruitment commenced on 30th March 2020). For these patients, blood was collected in serum tubes and centrifuged to obtain serum. At this time, there was very limited access to laboratory facilities and so plasma was not collected from these patients. 5 were outpatients and 41 were inpatients, reflecting the fact that UK policy was weighted towards inpatient testing at the time these patients were recruited. 33 of 46 patients (71.7%) had severe or critical disease (by peak WHO severity), and 9 (19.6%) patients died. For 40 patients, only one sample from a single timepoint was collected, and for 6 patients, 2 samples were collected. Serum samples from 11 non-infected haemodialysis patients were used to provide controls for the Validation cohort (collected at the same time as plasma from a subset of the ESKD control group recruited as part of the Main cohort).

Table 2.3 Characteristics of the Validation cohort

Patient subsets were defined according to their peak clinical severity over the course of the illness. 'South Asian' represents individuals with Indian, Pakistani, or Bangladeshi ancestry.

Q1 = lower quartile. Q3 = upper quartile. GN = glomerulonephritis. HTN = hypertension. N/A = not applicable.

	COVID-19 positive ESKD patients (n = 46)	COVID-19 negative ESKD controls (n = 11)
Age		
Median (Q1-Q3)	64.3 60.3-73.0	71.6 (61.7-73.9)
Sex		
M	32 (69.6%)	8 (72.3%)
F	14 (30.4%)	3 (27.3%)
Ethnicity		
White	11 (23.9%)	3 (27.3%)
Black	8 (17.4%)	3 (27.3%)
South Asian	12 (26.1%)	3 (27.3%)
Asian (other)	7 (15.2%)	0 (0%)
Other	8 (17.4%)	2 (18.2%)
Diabetes	29 (63.0%)	6 (54.5%)
Current Smoker	2 (4.3%)	0 (%)
ESKD Cause		
DN	19 (41.3%)	5 (45.5%)
Genetic	1 (2.2%)	0 (0%)
GN	7 (15.2%)	1 (9.1%)
HTN/Vascular	3 (6.5%)	1 (9.1%)
Other	3 (6.5%)	2 (18.2%)
Unknown	13 (28.3%)	2 (18.2%)
Hospitalisation due to COVID-19	41 (89.1%)	N/A
Severe or critical COVID-19	33 (71.7%)	N/A
Fatal COVID-19	9 (19.6%)	N/A

2.3.4 Clinical severity scores

Severity scoring was performed based on WHO classifications (WHO clinical management of COVID-19: Interim guidance 27 May 2020) adapted for clinical data available from electronic medical records. 'Mild' was defined as COVID-19 symptoms but no evidence of pneumonia and no hypoxia. 'Moderate' was defined as symptoms of pneumonia or hypoxia with oxygen saturation (SaO₂) greater than 92% on air, or an oxygen requirement no greater than 4L/min. 'Severe' was defined as SaO₂ less than 92% on air, or respiratory rate more than 30 per minute, or oxygen requirement more than 4L/min. 'Critical' was defined as organ dysfunction or shock or the need for high dependency or intensive care support (i.e., the need for non-invasive ventilation or intubation). Severity scores were charted throughout a patient's illness. I defined the overall severity/clinical course for each patient as the peak severity score that occurred during the patient's illness.

2.3.5 Proteomic assays

Plasma and serum proteomic measurements were performed using Olink proximity extension immunoassays (<https://www.olink.com/products/>). Five 92-protein multiplex Olink panels were run ("inflammation", "immune response", "cardiometabolic", "cardiovascular 2" and "cardiovascular 3"), resulting in 460 measurements per sample. Since a small number of proteins were measured on more than one panel, a total of 436 unique proteins were measured. In this chapter, I report these proteins by the symbols of the genes that encode them. The Olink assays were run at UMC Utrecht using 88 samples/plate, all of which were run in a single batch. Samples were randomly allocated to plates and wells given two constraints. Firstly, it was ensured each plate had a balance of cases and controls. Secondly, once a patient was allocated to a plate, all further serial samples from that patient were assigned to the same plate.

2.3.6 Normalisation and quality assessment and control

The data were normalised using standard Olink workflows to produce relative protein abundance on a log₂ scale ("NPX"). I performed quality assessment by a) examination of Olink internal controls, and b) inspection of boxplots, relative log expression (RLE) plots¹⁹⁴ and principal components analysis (PCA). Following these steps, 3 poor quality samples were removed. Two of these samples exhibited very low median NPX across all proteins. The third displayed high variability in RLE plots and was flagged with a warning based on Olink's internal

controls. In addition, 5 samples failed quality control on a single proteomic panel only, with the remaining panels passing QC. For these samples, I set proteins on the panel that failed QC to missing, and the data for the remaining proteins were retained.

For approximately 10% of the protein assays, more than 25% of their measurements fell below the calculated limited of detection (LOD). According to Olink guidance, I did not remove these assays or remove individual measurements below LOD. This approach has been reported by Olink to increase statistical power while avoiding truncation of protein distributions and increases in false positive results. Visual examination of these assays confirmed that their measurements were approximately normally distributed.

13 proteins were assayed more than once due to their inclusion in multiple Olink panels. For plasma, the median correlation between the assays was 0.986 with an interquartile range (IQR) of 0.974-0.993 and a range of 0.925 to 0.998. For serum, the median correlation between the assays was 0.991 with an interquartile range of 0.952-0.995 and a range of 0.737-0.999. I removed duplicate assays at random prior to subsequent analysis.

Principal components analysis revealed no substantial impact of plate effects (**Figure 2.2**).

For 11 ESKD controls, there were contemporaneous plasma and serum samples. To assess the comparability of these two matrices, I calculated Pearson's correlation coefficient between the assays for each protein. 344/436 (78.9%) of proteins had a Pearson's $r > 0.5$. The proteins with the lowest estimated Pearson correlation coefficient were AZU1, STK4 and TANK. However, this comparison had a relatively small sample size ($n = 11$) and these samples were from control patients without infection, so caution should be made in extrapolating these findings to the context of active infection, where protein dynamic ranges may be different.

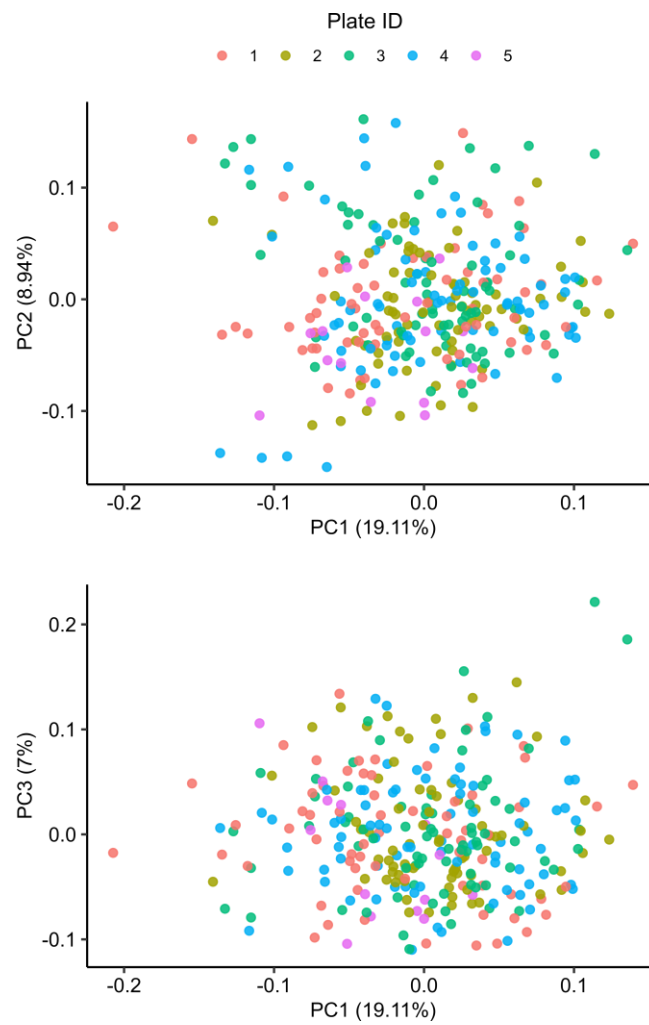


Figure 2.2 Principal components analysis in relation to assay plate

Principal components analysis of the Main cohort. Samples are coloured by plate.

2.3.7 Missing values

Following QC, 0.22% of data points were missing for the Main cohort and 0.35% for the Validation cohort. For analyses that required no missing values (PCA and supervised learning), I imputed missing values as follows. The dataset was first scaled and centred, and missing values were imputed using caret's k-nearest neighbours (kNN) method¹⁹⁵. The 5 closest samples (by Euclidean distance) were used to estimate each missing value.

2.3.8 Principal Components Analysis

I used singular value decomposition to perform principal components analysis (PCA). I first performed PCA to compare the ESKD controls of the Main cohort to samples from healthy individuals (**Figure 2.16a**). I also applied PCA to proteomic data from the entire Main cohort (which included plasma samples from both COVID-19-infected and non-infected ESKD patients) (**Figure 2.5**). I then used the loadings from the Main cohort together with the proteomic data from the Validation cohort to calculate principal components scores (**Figure 2.5**). This enabled projection of the Validation cohort data into the PC space of the Main cohort.

2.3.9 Consensus clustering

I performed consensus clustering using the M3C package¹⁰⁶ to identify the optimal number of clusters. M3C uses a Monte Carlo approach to generate a reference dataset, in which there are no clusters, but the covariance structure of the original data is conserved. Clustering is performed in both the original and reference data. This allows for the evaluation of whether the clusters in the original data are more stable than in the reference data. Therefore, M3C allows us to test the null hypothesis, that there are no clusters in the data. Among the metrics calculated by M3C is RCSI, the relative cluster stability index, which indicates the stability of the clusters in the real data relative to the mean stability of clusters in the reference data. M3C also produces *p*-values for the evaluation of the null hypothesis, that there are no clusters in the data. M3C clustering was performed using default parameters (iters = 25, repsreal = 100, repsref = 100, clusteralg = "pam").

2.3.10 Differential protein abundance analysis: ESKD versus healthy

I used linear models to compare 55 samples from patients with ESKD to a group of 35 healthy patients without kidney failure (**Table 2.2**). Linear models were fit using R's lm function and

type 2 F tests were applied with the `car::Anova` function¹⁹⁶. Age, sex and ethnicity were included as covariates. The regression model in Wilkinson-style notation was:

$$\text{NPX} \sim \text{eskd_status} + \text{sex} + \text{age} + \text{ethnicity}$$

2.3.11 Differential protein abundance analysis: COVID-19 positive versus negative

I performed differential protein abundance analyses between COVID-19 positive and negative samples using linear mixed models (LMMs), to account for the use of serial samples from the same individuals (for a more detailed explanation, see **Section 1.3.1.1**). This analysis compared 256 samples from 55 COVID-19 patients with 51 non-infected patients (1 sample per non-infected patient). LMMs were fit using R's `lmerTest::lmer` function^{197,198}. Age, sex and ethnicity were included as covariates. I used a random intercept term to estimate the variability between individuals in the study and account for repeated measures. The regression model was:

$$\text{NPX} \sim \text{covid_status} + \text{sex} + \text{age} + \text{ethnicity} + (1 \mid \text{individual})$$

Where NPX represents protein abundance and `covid_status` was a categorical variable (infected/non-infected). Sex and ethnicity were also categorical variables. Age was a quantitative variable. I calculated *p*-values using a type 3 F test in conjunction with Satterthwaite's method for estimating the degrees of freedom for fixed effects¹⁹⁷. The regression model was fit for each of the 436 proteins individually. Multiple testing correction was performed using the Benjamini-Hochberg method and 5% FDR was used for the significance threshold.

I applied the same approach to the Validation cohort. This analysis comprised 52 serum samples from 46 COVID-19 positive patients versus 11 samples from non-infected patients samples (1 sample per non-infected patient).

As sensitivity analyses, I repeated the differential abundance analyses between cases and controls for the Main cohort adjusting for additional covariates and comparing this to the basic model (i.e., using age, sex, and ethnicity alone). This was performed for each of the following parameters: diabetes status, cause of ESKD, and time to last haemodialysis.

2.3.12 Testing for associations between proteins and clinical severity

For testing the association of plasma proteins with the 4-level WHO severity rating (mild, moderate, severe and critical) within COVID-19 positive cases from the Main cohort ($n = 256$ samples from 55 patients), I employed a similar linear mixed modelling approach to the COVID-19 positive versus negative differential abundance analysis; however, for this analysis, I replaced the covid_status term with a severity variable encoded using orthogonal polynomial contrasts to account for ordinal nature of severity levels. As before, age, sex and ethnicity were included as covariates. As a sensitivity analysis, I repeated the modelling with time to last haemodialysis (days) as an additional covariate.

2.3.13 Testing for associations between proteins and clinical laboratory tests

I also employed the linear mixed modelling strategy to test the association of temporal clinical laboratory variables and protein levels, with the value of the clinical variable (as a quantitative trait) used in place of covid_status. Only COVID-19 positive patients were included in this analysis. Contemporaneous lab measurements were not available for all samples. This varied according to the clinical lab parameter (**Table 2.4**). I calculated correlations between clinical laboratory variables and protein levels using the R package rmcrr, which determines the overall within-individual relationship among paired measures that have been taken on two or more occasions¹⁹⁹. The rmcrr method adjusts for non-independence between measurements from the same individual with the analysis of covariance (ANCOVA) approach¹⁹⁹; the method models the association for each individual using parallel regression lines that each have separate intercepts. Therefore, rmcrr is able to model the intra-individual variation between measurements, improving statistical power without relying on an assumption that samples are independent.

Table 2.4 Missing values for clinical parameters

The proportion of proteomic timepoints for which there were no data are indicated for each clinical measurement.

Parameter	Proportion missing
White cell count	34%
Neutrophils	34%
Monocytes	34%
Lymphocytes	34%
C-reactive protein	36%
Ferritin	64%
Troponin	65%
D-dimer	70%

2.3.14 Testing for associations between proteins and ethnicity

I performed testing of protein levels and ethnicity separately in COVID-19 negative ESKD patients and COVID-19 positive ESKD patients. These analyses were limited to individuals who were White, South Asian (Indian, Pakistani, or Bangladeshi ancestry) or Black as there were too few individuals from other ethnic groups for meaningful interpretation. For COVID-19 negative patients (1 sample per patient), I performed linear regression for each protein with ethnicity as the predictor variable, and age and sex as covariates. For COVID-19 positive patients I used LMMs to account for serial samples from the same individual, again with age, and sex as covariates.

2.3.15 Alternative estimation of the FDR using the plug-in method

To provide additional support that the Benjamini-Hochberg procedure was providing adequate control of the FDR, I also used the plug-in method of Tibshirani²⁰⁰ as an alternative method to estimate the FDR, as described below.

1. R was defined as the number of associations declared significant in the real data.
2. C was defined as the test statistic used as the significance threshold used in the real data (i.e., that corresponding to an adjusted p -value of 0.05).

3. The expected number of proteins found to be significant under the null hypothesis, that no proteins are differentially abundant between COVID-19 positive versus negative patients (i.e., false positives), was estimated using a permutation strategy. I randomly permuted each individual's COVID-19 status label 100,000 times and, in each case, repeated the differential abundance analysis on the permuted data. The number of false positives ($\hat{V}$) was then estimated as the number of associations with test statistic $> C$ in 100,000 permutations of the data, divided by the number of permutations.

4. The estimated FDR was then calculated as $\hat{V}/R$.

I implemented a similar approach to test for the association of proteins with severity scores within cases. While the case versus control labels were the same for each individual (except for sample C105, for which we had obtained both a COVID-19 positive and negative sample), each individual could have different disease severity at each timepoint during acute infection. Therefore, labels were permuted at the individual level for the case versus control analysis, whereas permutation of labels was done at the sample level for the severity analysis.

2.3.16 Empirical p -value calculation

As a complementary analysis, based on the approach of Filbin *et al.*²⁰¹, I estimated the empirical P -value for the likelihood of observing as many significant proteins as were identified in the real data if the null hypothesis of no differentially abundant proteins in cases versus controls were true. I again used 100,000 permutations of the case-control labels to estimate the null distribution. I performed Benjamini-Hochberg adjustment on the nominal p -values of each permutation and counted the number of significant proteins (adjusted p -value < 0.05) in each permutation.

2.3.17 Supervised learning

I fit random forest models using R's randomForest and caret packages^{195,202}. Data was centred, scaled, and imputed as in **2.3.7 Missing values** with the caveat that, during cross-validation, the pre-processing procedure was first applied on the resampled (training) data before the same method was applied without re-calculation to the holdout (test) set. To estimate model accuracy, I used 4-fold cross-validation. The cross-validation procedure was repeated 100 times. The model's parameters were kept constant at 500 trees and the mtry

value (number of proteins randomly sampled as candidates at each node) was set to the square root of the number of features. After parameter estimation, I fit a final model trained using the entirety of the dataset. This model was used for subsequent feature extraction. Random forest feature extraction was carried out using the R `randomForestExplainer` package²⁰³. I made use of the following importance measures: accuracy decrease (the average decrease in prediction accuracy upon swapping out a feature), number of trees (the number of trees with a node corresponding to a feature) and mean minimal depth (the average depth at which a node corresponding to a feature occurs). Three models were generated with different input features: i) proteomic data alone; ii) clinical parameters alone; iii) proteomic data and clinical parameters. Clinical parameters included sex, age, ethnicity, cause of ESKD, comorbidities, smoking status, radiological evidence of pulmonary infiltrates, and clinical laboratory tests.

2.3.18 Survival analysis using joint modelling

Following scaling and centering, I fit LMMs for each protein to capture the temporal trajectories of each individual. A polynomial spline of degree 2 was used to model protein concentration with respect to time (from symptom onset, measured in days). The curve was fit for samples that were taken between 1 and 28 days from first symptoms, inclusive. Both random intercepts and random slopes were estimated for each individual, as per the following Wilkinson-style formula:

$$\text{NPX} \sim \text{time} + (\text{time} \mid \text{individual})$$

These were joined to a Cox regression model using the `jointModel` package⁹¹ in order to estimate the association of each protein with risk for death. For further discussion of the joint modelling approach, see **Section 1.3.1.2**. A Wald test was used to calculate *p*-values for the association between the linear mixed model and Cox regression. Benjamini-Hochberg adjustment was applied, with an adjusted *p*-value of 0.05 used as the significance threshold.

2.3.19 Longitudinal analysis

I again fit LMMs to explicitly model the temporal profile of each protein. In this instance, however, I set up the model to test for each protein a) whether the protein significantly changed over time following symptom onset and b) whether the protein changes over time differently in individuals with a mild versus severe disease course. The latter was performed statistically

by testing for an interaction effect between time and clinical course. For this analysis, I binarised patients into severe or non-severe clinical courses according to the peak WHO severity disease of their illness. Patients with a peak WHO score of mild or moderate were considered non-severe and those with a peak score of severe or critical were considered severe.

I then used R's `bs` function to fit a polynomial spline of degree 2 to model protein concentration with respect to time (from symptom onset, measured in days)²⁰⁴. The model was fit for samples that were taken between 1 and 21 days from symptom onset, inclusive. Random slopes were estimated with respect to time, in addition to random intercepts, to account for each individual's unique disease course. For each protein, I fit the following model (Wilkinson-style notation):

$$\text{NPX} \sim \text{time} * \text{severity} + \text{sex} + \text{age} + \text{ethnicity} + (\text{time} | \text{individual})$$

To identify proteins that changed significantly over time, I examined the *p*-values for the main effect of time. To identify proteins with distinct temporal profiles between severe and non-severe cases, we examined the *p*-values for the time x severity interaction term. For each of these two research questions, *p*-values were adjusted for the multiple proteins tested using the Benjamini-Hochberg method and 5% FDR was used as the significance threshold.

2.3.20 Data and code availability

Underlying source data for all analyses (individual-level proteomic and clinical phenotyping data) are available without restriction from the Dryad Digital Repository:

<https://doi.org/10.5061/dryad.6t1g1jwxj>

Code is available from GitHub:

https://github.com/jackgisby/longitudinal_olink_proteomics

Which has also been copy archived:

[swh:1:rev:32f08137859d44707ec4f086eed9af9b9ee91a87](https://www.swh.io/rev/32f08137859d44707ec4f086eed9af9b9ee91a87)

2.4 Results

In this chapter, I describe the proteomic profiling of ESKD patients with COVID-19, recruited during the first wave of COVID-19 in April and May of 2020 (**Figure 2.3**). The Main cohort included 55 ESKD patients with COVID-19 (**Table 2.1**). Blood samples were taken as soon as feasible following COVID-19 diagnosis and serial sampling was performed for 51/55 patients. At the time of the initial sample, 30 patients were outpatients attending haemodialysis sessions and 25 were hospitalised inpatients, although 8 patients who were recruited as outpatients were subsequently admitted to hospital with COVID-19 over the course of the study (**Figure 2.3, Figure 2.4**). Nine patients died. Another 51 non-infected haemodialysis patients were recruited as controls, mirroring the age, sex and ethnicity distribution of the COVID-19 cases (**Table 2.2**). The Olink proteomics platform was applied to measure 436 proteins in 256 plasma samples from the COVID-19 patients and the 51 control samples. Proteomic profiling was also simultaneously performed for a group of healthy patients (**Table 2.2**), allowing for the investigation of proteins dysregulated in ESKD patients undergoing dialysis.

In addition, Olink proteomic assays were performed for 52 serum samples from a separate set of 46 COVID-19-positive ESKD patients (hereafter referred to as the Validation cohort) and 11 serum samples from ESKD COVID-19-negative controls (a subset of the Main cohort controls) (**Table 2.3**). For most patients in the Validation cohort, only a single timepoint was available. A higher proportion of these patients (41/46, 89%) were hospitalised and had severe disease (**Table 2.3**) than in the Main cohort (**Table 2.1**). Unless otherwise stated, results in this chapter were derived from the Main cohort.

The proteins measured consisted of five multiplex ‘panels’ focussed on proteins relevant to immuno-inflammation, cardiovascular, and cardiometabolic disease. I performed enrichment analysis using string-db²⁰⁵ and the Reactome database¹⁶⁰, finding an overrepresentation of immune-related proteins on these panels. Among the most significantly enriched terms were “Immune System”, “Signaling by Interleukins”, “Extracellular matrix organisation”, “Cell surface interactions at the vascular wall”, “Neutrophil degranulation”, “Chemokine receptors bind chemokines” and “Formation of Fibrin Clot”. It is important to keep in mind the targeted nature of these panels when interpreting the results presented in this chapter.

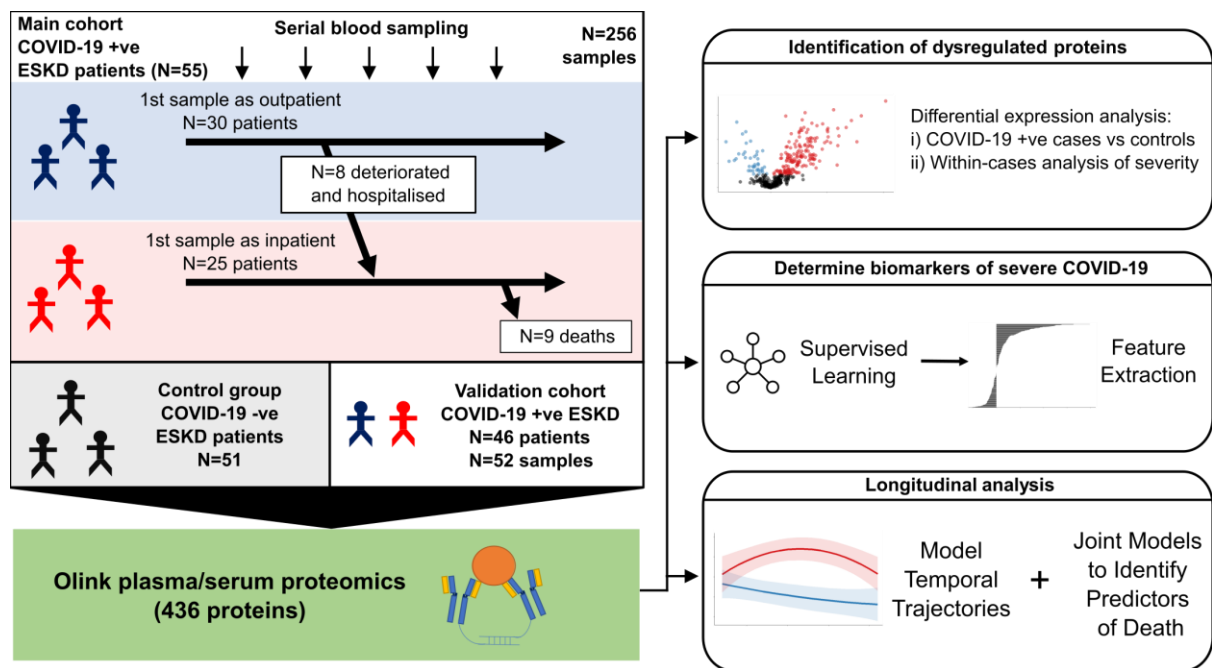


Figure 2.3 Study design

Schematic representing a summary of the patient cohorts, sampling, and the major analyses. Blue and red stick figures represent outpatients and hospitalised patients, respectively.

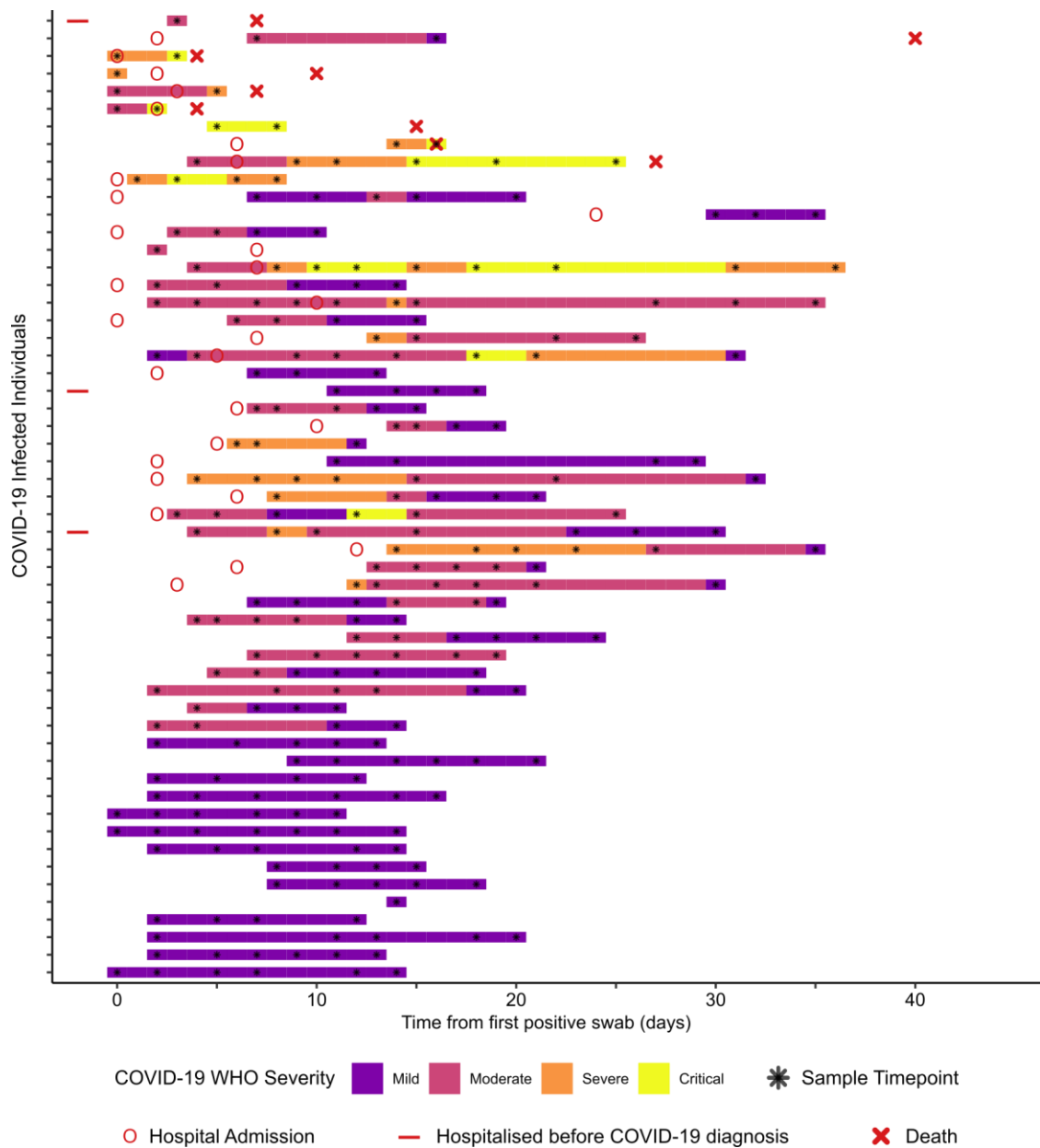


Figure 2.4 Serial sampling of the Main cohort

Timing of serial blood sampling in relation to the clinical course of COVID-19. Black asterisks indicate when samples were obtained. Three patients were already hospitalised prior to COVID-19 diagnosis (indicated by red bars).

2.4.1 Proteomic effects of COVID-19

PCA of proteomic data from the Main cohort demonstrated differences between samples from COVID-19-positive cases and controls (**Figure 2.5**); the two groups did not, however, separate into visually distinct clusters. I performed a formal clustering analysis using M3C¹⁰⁶, which employs a Monte Carlo-based method to ascertain whether clusters are more stable than would be expected under a null hypothesis, that there are no clusters in the data. This analysis revealed that the optimal number of clusters (K) was 3, but that these clusters were not significantly more stable than those identified in a reference dataset, in which there were no clusters (**Figure 2.6a**). So, a purely unsupervised approach was not able to clearly distinguish between the plasma proteome of ESKD patients with and without COVID-19.

To dissect the effects of COVID-19 on the plasma proteome, I performed a differential abundance analysis in the Main cohort between COVID-19 cases ($n = 256$ samples passing quality control [QC] from 55 patients) and non-infected ESKD controls ($n = 51$) using linear mixed models, which account for serial samples from the same individual. This revealed the upregulation of 181 proteins in COVID-19 and downregulation of 40 (**Figure 2.7**). In this chapter, I refer to proteins by the names of the genes that encode them. The most strongly upregulated proteins (in terms of fold change) were DDX58, CCL7, IL6, CXCL11, KRT19, and CXCL10, and the most strongly downregulated were SERPINA5, CCL16, FABP2, PON3, ITGA11, and MMP12 (**Figure 2.7**). The most highly upregulated proteins predominantly consisted of cytokines and proteins involved in the interferon response, whereas there was downregulation of proteins involved in lipid metabolism and those that interact with the extracellular matrix.

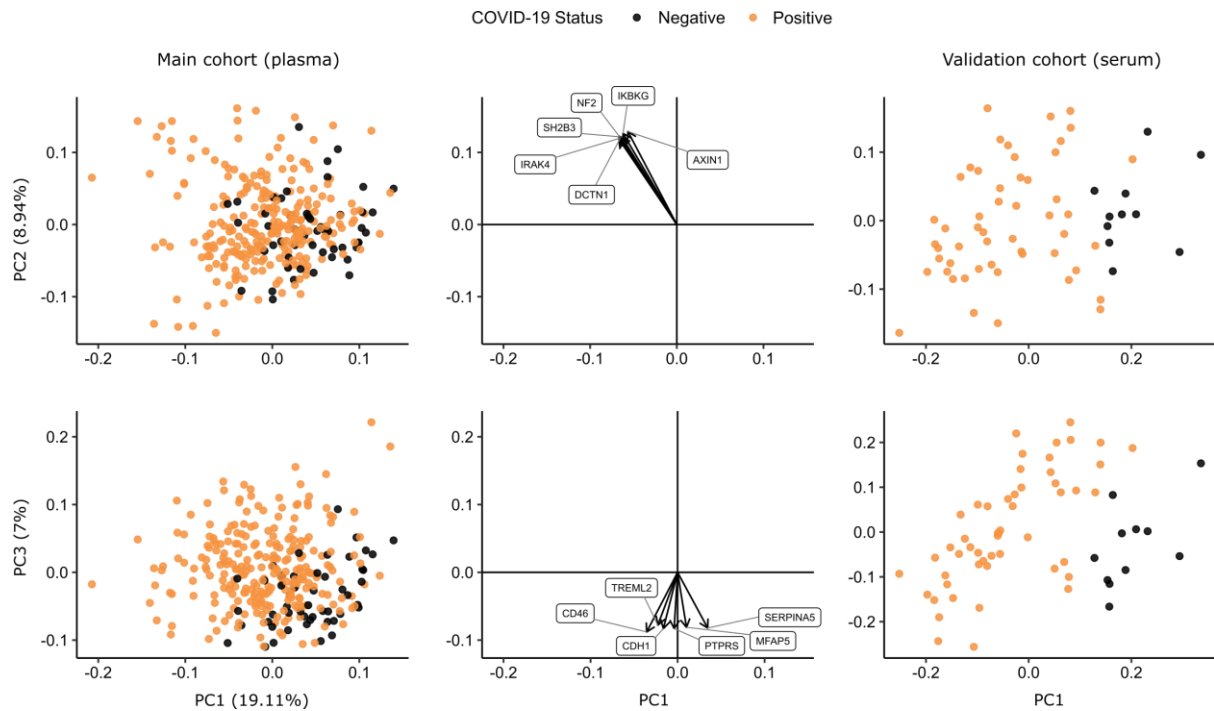


Figure 2.5 Principal components analysis

PC = principal component. Each point represents a sample. Colouring indicates COVID-19 status. **left)** Main cohort. Due to serial sampling, there are multiple samples for most patients. The proportion of variance explained in the Main cohort by each PC is shown in parentheses on the axis labels. **centre)** The directions and relative sizes of the 6 largest PC loadings are plotted as arrows. **right)** The Validation cohort. Samples are projected into the PCA coordinates calculated for the Main cohort.

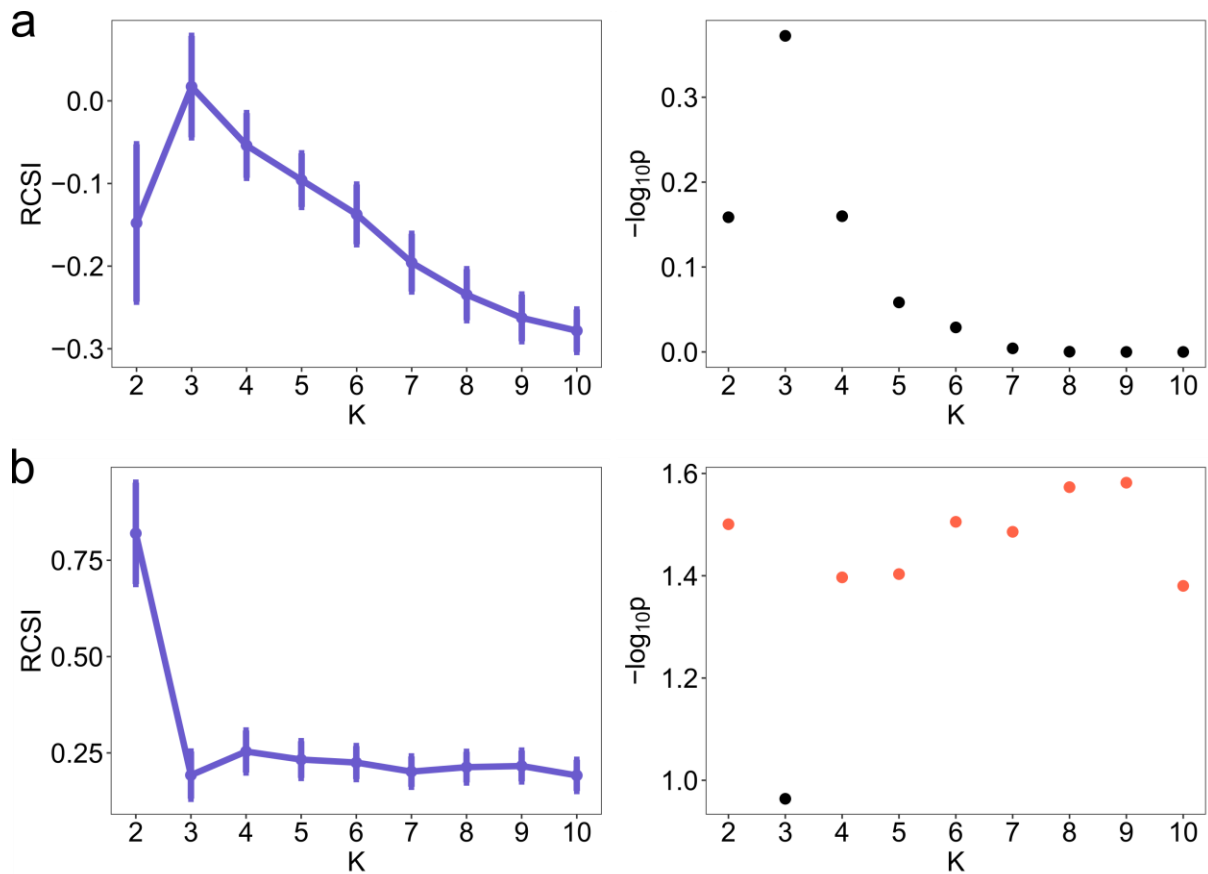


Figure 2.6 Consensus clustering

Visualisation of consensus clustering statistics as the number of clusters (K) varies, performed for the Main cohort (a) and Validation cohort (b). Consensus clustering performed by M3C involves comparing the observed clusters to those identified in a reference dataset generated by a Monte Carlo procedure, in which there are no clusters. The relative cluster stability index (RCSI; left) indicates the stability of the observed clusters relative to those in the reference data. Higher RCSI values indicate more stable clustering. P -values (right) are shown for the evaluation of the null hypothesis, that there are no clusters in the data.

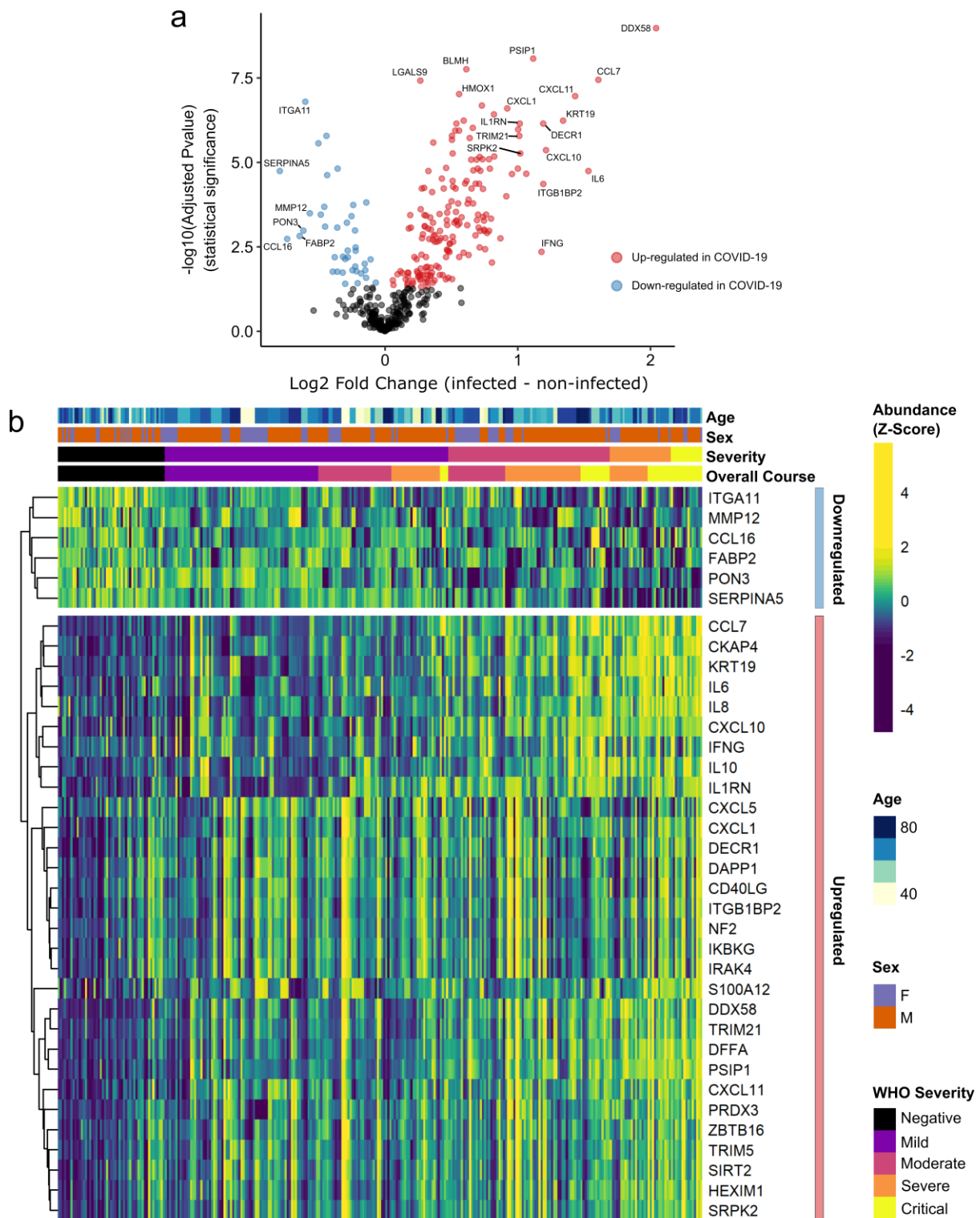


Figure 2.7 Proteins dysregulated in COVID-19

a) Proteins up-regulated (red) or down-regulated (blue) in COVID-19 positive patients versus negative ESKD patients. $n = 256$ plasma samples from 55 COVID-19 positive patients, versus $n = 51$ ESKD controls (1 sample per control patient). P-values from linear mixed models after Benjamini-Hochberg adjustment; significance threshold = 5% FDR; dark-grey = non-significant. **b)** Heatmap showing protein levels for selected proteins with the largest fold changes in differential abundance analysis (Main cohort). Each column represents a sample ($n = 256$ COVID-19 samples and 51 non-infected samples). Each row represents a protein. Proteins are annotated using the symbol of their encoding gene. For legibility, not all significantly associated proteins are shown; the heatmap is limited to the 17% most up- or down-regulated proteins (by fold change) of those with a significant association. Proteins are

ordered by hierarchical clustering. Samples are ordered by WHO severity at the time of blood sample ('Severity'). 'Overall course' indicates the peak WHO severity over the course of the illness.

2.4.2 Validation of results in an independent cohort

While the Validation cohort was smaller than the Main cohort ($n = 52$ serum samples from 46 patients with COVID-19) and most patients were only sampled once, it provided an excellent opportunity for validation. I first projected the data from the Validation cohort into the PC space of the Main cohort to examine how well the separation of cases and controls replicated (see **Section 2.3.8**). Surprisingly, the separation of infected and non-infected patients was clearer than in the Main cohort (**Figure 2.5**). This may be because there were a higher proportion of hospitalised patients (41 of 46 patients) in the Validation cohort (**Table 2.1, Table 2.3**). Clustering with M3C¹⁰⁶ identified 2 clusters ($K = 2$) in these data (**Figure 2.6b**). All the non-infected patients were in cluster 1, along with 20 infected patients. Of the infected patients in cluster 1, 40% had a mild/moderate disease course, whereas only 22% of the patients in cluster 2 had a mild/moderate clinical course (defined as the peak WHO severity during the illness). This further highlights differences between the cohorts, since patients did not separate into discrete clusters in the Main cohort (**Figure 2.6a**).

Next, I directly compared COVID-19 cases and controls in the Validation cohort, finding 201 differentially abundant proteins (5% FDR). Of the 221 proteins that were dysregulated in the Main cohort, 150 (69.7%) were also identified in the Validation cohort at 5% FDR (**Figure 2.8a**). Effect sizes in each dataset were highly correlated ($r = 0.80$; **Figure 2.8b**), demonstrating that these findings are highly reproducible despite differences in sample size, disease severity and matrix (while plasma samples were collected for the Main cohort, only serum was available for the Validation cohort).

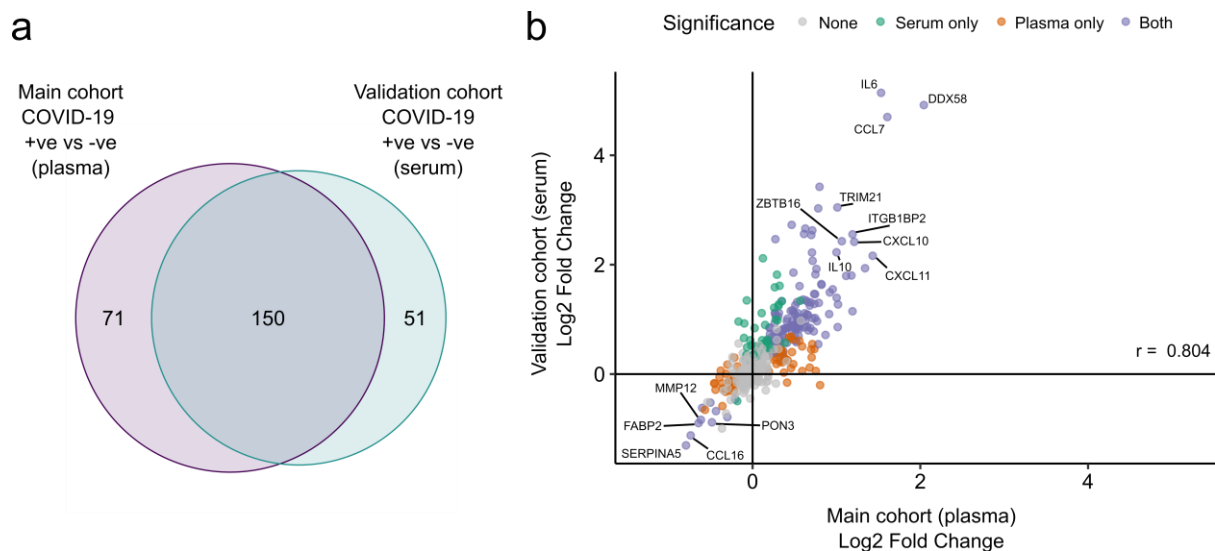


Figure 2.8 Comparison to the Validation cohort

Differential abundance analyses were performed using linear mixed models. Main cohort analysis (plasma samples): 256 samples from 55 COVID-19 patients versus 51 non-infected patient samples (single time-point). Validation cohort (serum samples): 52 samples from 55 COVID-19 patients and 11 non-infected patient samples (single time-point). **a)** Overlap between the significant associations in the differential abundance analysis between ESKD patients with and without COVID-19 in each cohort. 5% FDR was used as the significance threshold in both analyses. **b)** Comparison of estimated effect sizes for all 436 proteins in the differential abundance analyses (COVID-19 positive versus negative) in each cohort. Each point represents a protein. Pearson's r is shown.

2.4.3 Proteins associated with COVID-19 severity

The scores plots were reproduced for the first two PCs and recoloured by clinical severity at the time of sampling (defined by WHO severity scores, graded as mild, moderate, severe, or critical) (**Figure 2.9a**). I observed previously that there was no clear separation between the samples, and within-cases consensus clustering failed to identify significant structure in the data. Despite this, there was a noticeable gradient of COVID-19 severity, which was best captured by principal components 1 and 3 (**Figure 2.9a**). To determine proteins associated with COVID-19 severity, I tested for associations between proteins and WHO severity score at the time of blood sampling, using linear mixed models with severity encoded as an ordinal predictor (see **Section 2.3.12**). 203 proteins were associated with COVID-19 severity (**Figure 2.10**), the majority of which were positively associated with severity, with only 42 downregulated in severe disease. The most highly upregulated protein in severe disease was interleukin 6 (IL6), mirroring previous observations^{162,206,207}. There was also upregulation of proteins involved in the interferon response to viral infection (e.g., IFNG, DDX58), monocyte chemotaxis (e.g., CCL2, CCL7, CXCL10), neutrophil activation and degranulation (e.g.,

PRTN3, MPO), and epithelial injury (e.g., KRT19, AREG, PSIP1, GRN) (**Figure 2.11**). The most downregulated protein was serpin family A member 5 (SERPINA5), which plays a role in thrombosis through the inhibition of coagulation factors. Also highly downregulated were leptin (LEP), fetuin B (FETUB) and serpin family A member 12 (SERPINA12) (**Figure 2.10**).

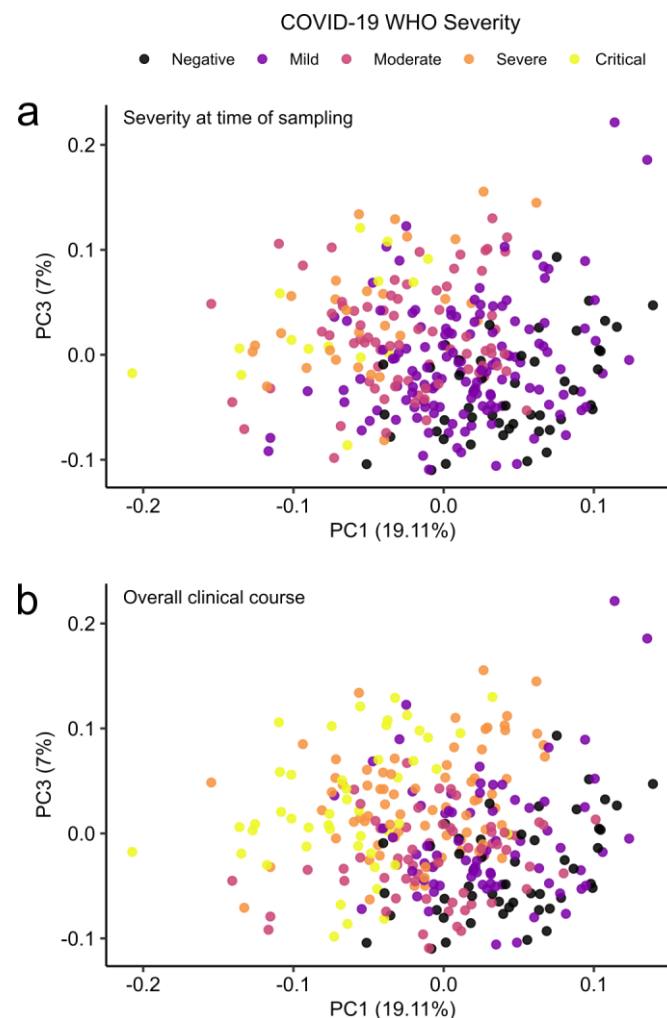


Figure 2.9 Principal components analysis in relation to clinical severity

a) Colouring indicates WHO severity at the time of sampling. **b)** Colouring indicates the overall clinical course (i.e., peak WHO severity) for the patient from which that sample was taken.

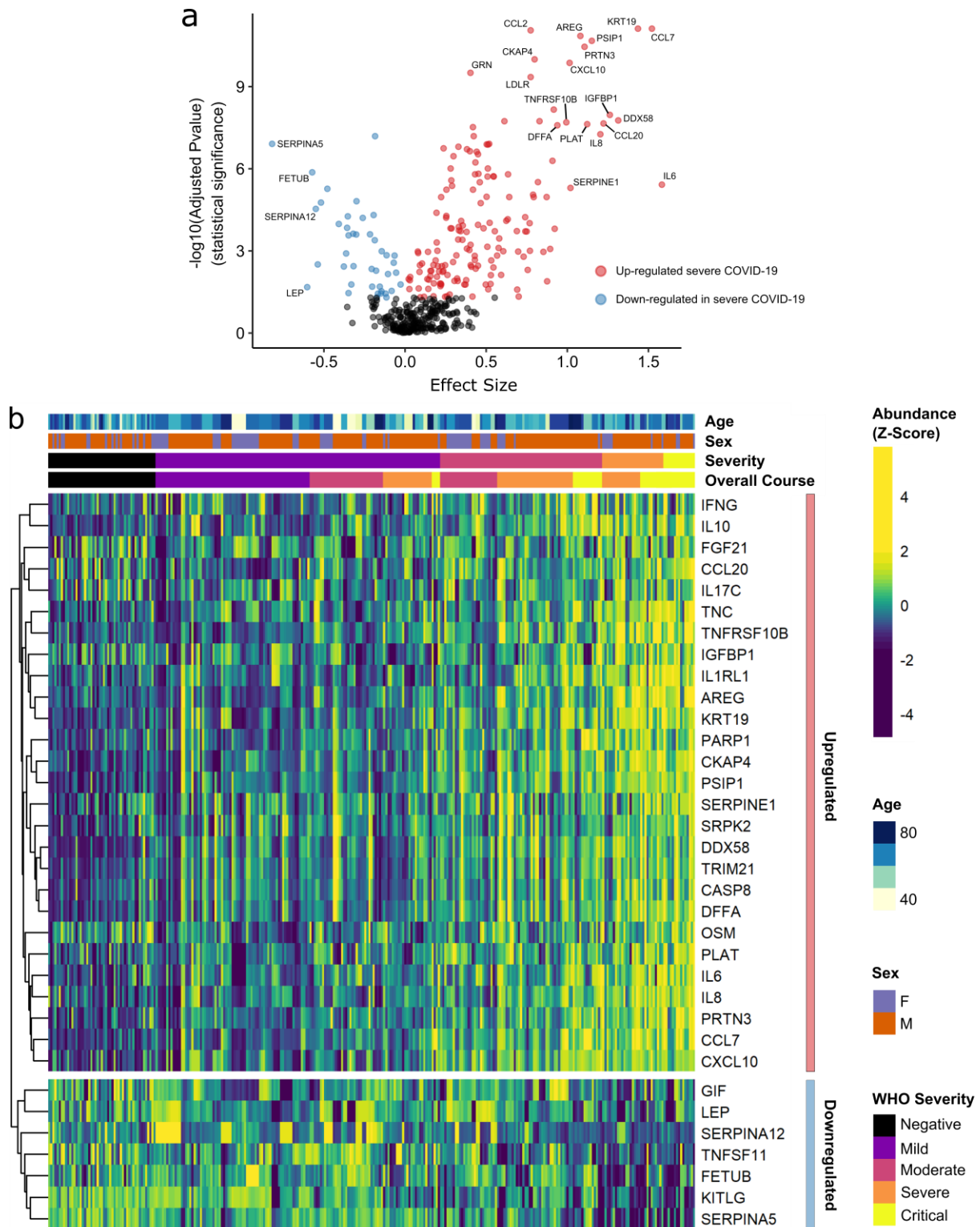


Figure 2.10 Proteins dysregulated in severe COVID-19

a) Proteins associated with disease severity. Associations of protein levels against WHO severity score at the time of sampling. Linear gradient indicates the effect size. A positive effect size (red) indicates that an increase in protein level is associated with increasing disease severity and a negative gradient (blue) the opposite. $n = 256$ plasma samples from 55 COVID-19 positive patients. For **a)** and **b)**: P-values from linear mixed models after Benjamini-Hochberg adjustment; significance threshold = 5% FDR; dark-grey = non-significant. **b)** Heatmap showing protein levels for selected proteins with strong associations with severity. Each column represents a sample ($n = 256$ COVID-19 samples and 51 non-infected samples). Each row represents a protein. Proteins are annotated using the symbol of their encoding gene. For legibility, not all significantly associated proteins are shown; the heatmap is

limited to the 17% most up- or down-regulated proteins (by effect size) of those with a significant association. Proteins are ordered by hierarchical clustering. Samples are ordered by WHO severity at the time of blood sample ('Severity'). 'Overall course' indicates the peak WHO severity over the course of the illness.

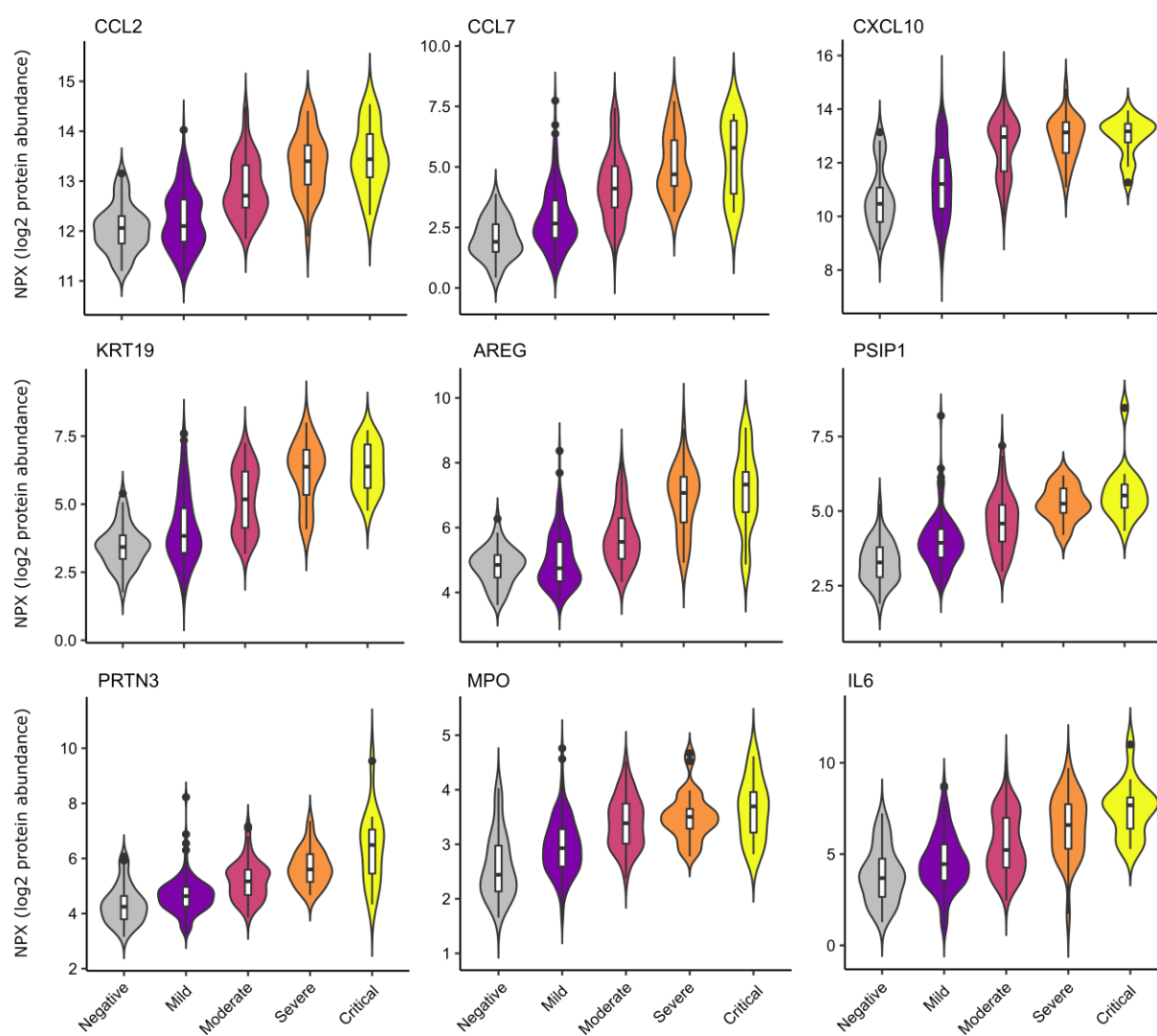


Figure 2.11 Selected proteins strongly associated with COVID-19 severity

Violin plots showing the distribution of plasma protein levels according to COVID-19 status at the time of blood draw. Boxplots indicate median and interquartile range. $n = 256$ samples from 55 COVID-19 patients and 51 samples from non-infected patients. WHO severity indicates the clinical severity score of the patient at the time the sample was taken. Mild $n = 135$ samples; moderate $n = 77$ samples; severe $n = 29$ samples; critical $n = 15$ samples. Upper panel: monocyte chemokines. Middle panel: markers of epithelial injury. Lower panel: 2 neutrophil proteases, and IL6.

2.4.4 Potential confounders do not materially affect results

Although the COVID-19-negative controls were well matched in terms of age, sex, and ethnicity (**Table 2.1**), it was not feasible to find control patients with perfectly matched comorbidities in the context of the ongoing healthcare emergency at the time of patient recruitment. There was a higher prevalence of diabetes in the COVID-19 cases compared to the controls (61.8% versus 47.1%, respectively; **Table 2.1**). To evaluate whether differing rates of diabetes impacted the proteins identified as differentially abundant between cases and controls, I performed a sensitivity analysis in which diabetes was added as an additional covariate in the linear mixed model. This did not materially affect the findings; estimated fold changes and $-\log_{10}$ -transformed p -values from models with and without the inclusion of diabetes were highly correlated (Pearson's $r > 0.99$, and $r = 0.96$, respectively; **Figure 2.12a-b**). There were also differences in the underlying cause of ESKD in cases compared to controls (**Table 2.1**). I therefore performed a further sensitivity analysis adjusting for the underlying cause of renal failure. Similarly, this did not make any meaningful difference to the results (Pearson $r > 0.99$ for estimated fold change, and $r = 0.95$ for $-\log_{10} p$ -values; **Figure 2.12c-d**).

All samples were taken prior to haemodialysis to minimise the impact of this treatment on the proteome. Furthermore, for most samples (86.6%), the most recent haemodialysis session was between 48 and 72 hours prior to the blood draw. This consistency should reduce the impact of haemodialysis on the plasma proteome. Nevertheless, I also considered the possibility that the timing of haemodialysis could affect the differential abundance analyses. I evaluated the effect of this factor through a sensitivity analysis, in which time from last haemodialysis was included as a covariate. The observed proteomic associations with COVID-19 were not materially affected by this; $-\log_{10} p$ -values and estimated effect sizes were highly correlated with those obtained prior to the inclusion of this variable in the regression models (Pearson $r > 0.99$ for both the effect size estimates and $-\log_{10} p$ -values; **Figure 2.13a-b**). Timing from last haemodialysis could also affect proteomic associations with severity, particularly if clinical deterioration affected the regularity of these treatments. However, I found that adjusting for time from last dialysis did not substantially affect estimated associations between protein abundance and contemporaneous COVID-19 severity (Pearson $r > 0.99$ for both the effect size estimates and $-\log_{10} p$ -values; **Figure 2.13c-d**).

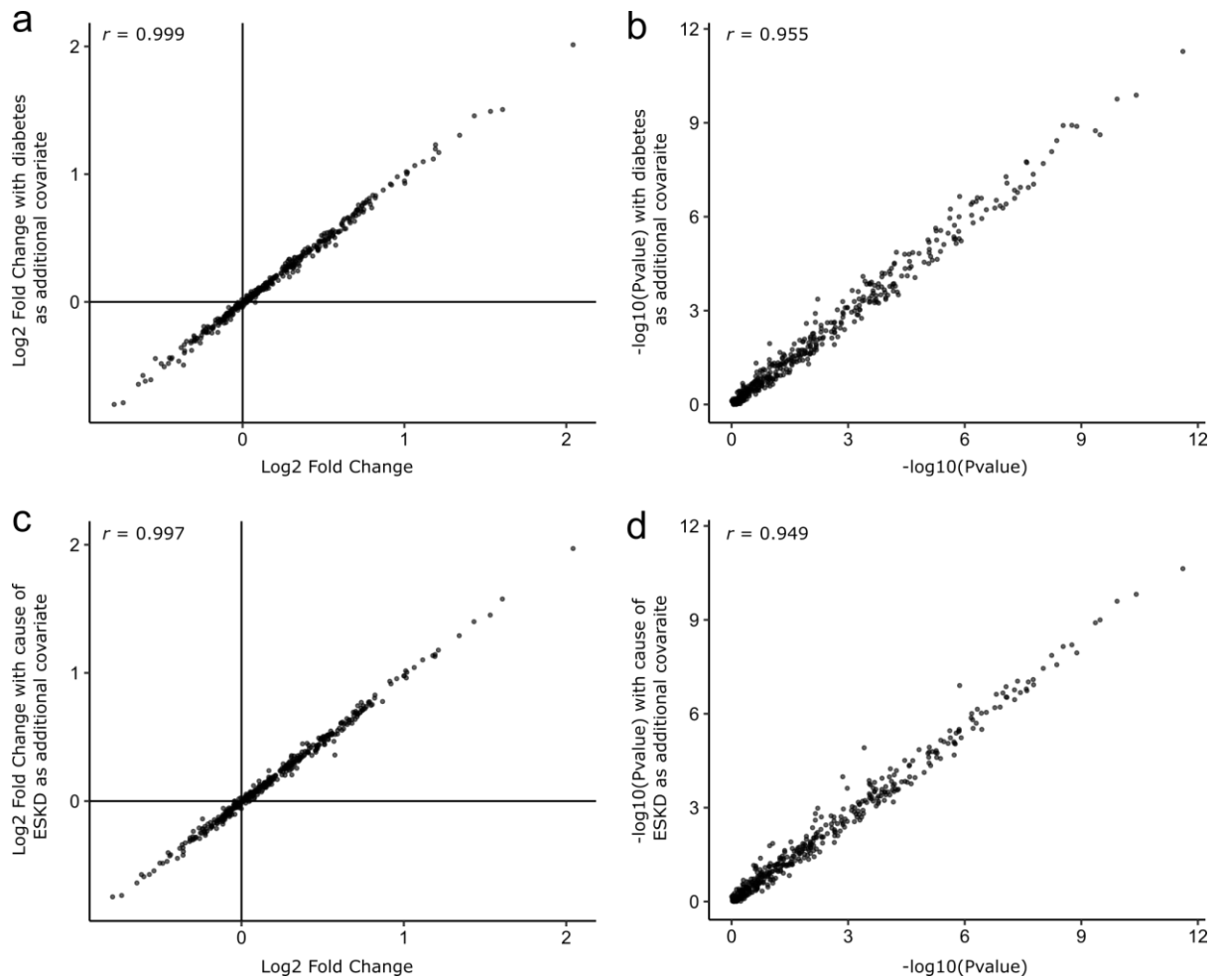


Figure 2.12 Sensitivity analyses adjusting for diabetes status and cause of ESKD

As sensitivity analyses, the COVID-19 positive versus negative differential abundance regressions were repeated adding diabetes status (**a-b**) and cause of ESKD (**c-d**) as additional covariates. The basic model included age, sex and ethnicity as covariates. Each point represents a protein. A comparison of -log10 p-values and effect sizes are shown for all 436 proteins. r indicates Pearson's correlation coefficient.

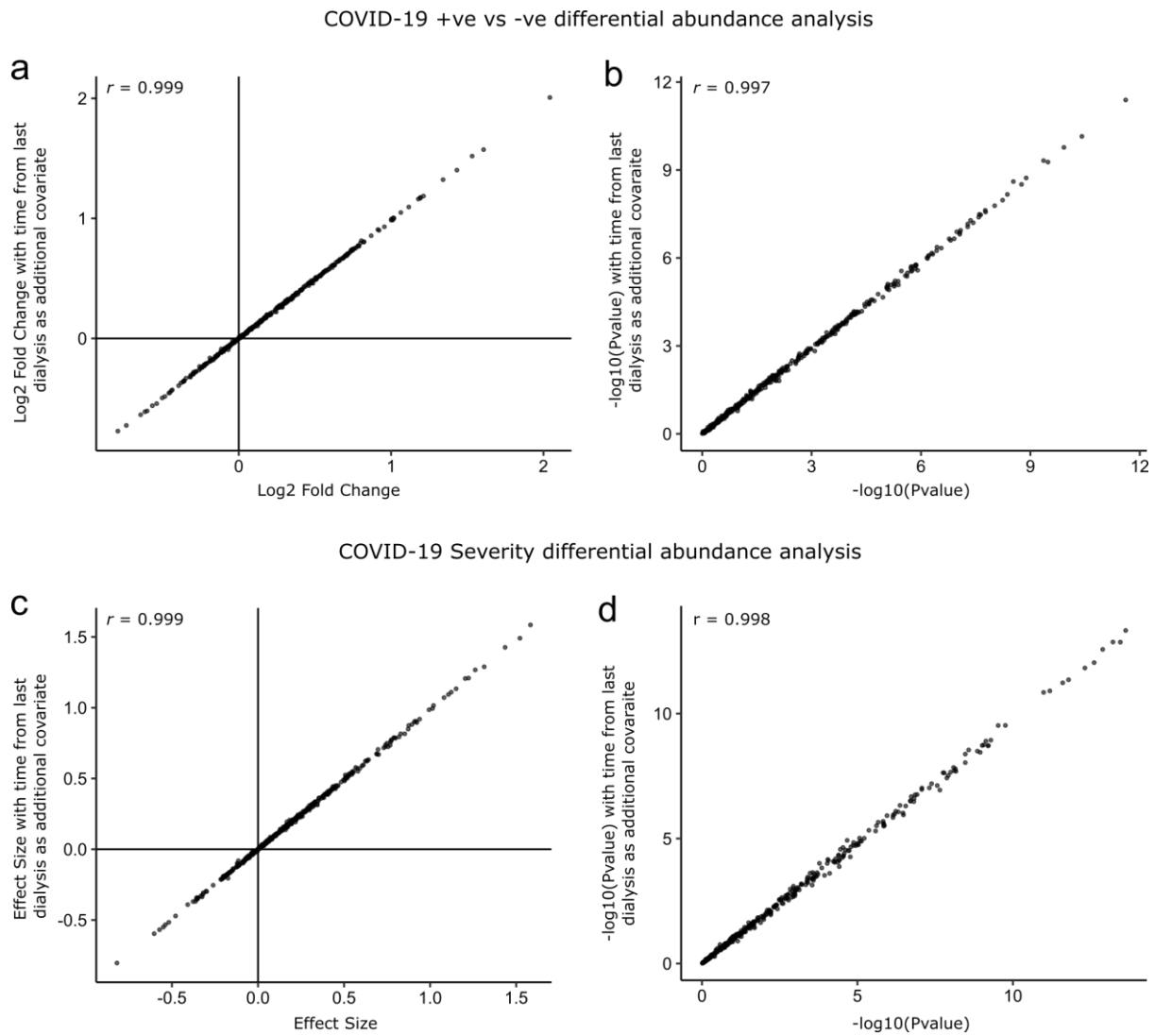


Figure 2.13 Sensitivity analysis adjusting for time since last haemodialysis

Comparison of results obtained with and without adding time since last haemodialysis as an additional covariate to the regression models. **a-b)** COVID-19 positive versus negative differential expression analysis. **c-d)** Severity analysis. Each point represents a protein. r indicates Pearson's correlation coefficient.

2.4.5 Benjamini-Hochberg ensures control of the false discovery rate

A high proportion of the measured proteins were significantly differentially abundant in the comparison of COVID-19 cases and controls (**Figure 2.7**). Given the highly targeted nature of the Olink panels (enriched for immune and inflammation-related proteins; **Section 2.4**), this was not surprising. Nevertheless, to ensure that the Benjamini–Hochberg adjustment of p -values was correctly controlling the false discovery rate (FDR) at the 5% level, I performed two further analyses.

The first involved estimating the FDR using an alternative method (the plug-in procedure²⁰⁰; **Section 2.3.15**). The plug-in method estimated an FDR for the comparison of COVID-19 cases and controls of 6.2%. The same procedure implemented for the analysis of contemporaneous severity estimated control of the FDR at 5.7%. This indicated that the Benjamini-Hochberg adjustment was appropriately controlling the FDR.

Second, I used permutations to estimate the distribution of the number of proteins expected to be declared significant under the null hypothesis of no association between any proteins and COVID-19 (see **Section 2.3.16**). The distribution of the number of proteins declared significant is shown in **Figure 2.14**; on no occasion in 100,000 permutations did we observe more proteins declared significant than in the real data. Thus, the empirical p -value (i.e., the proportion of permutations for which there were more than or equal to the number of associations seen for the real data) is less than $1/100,000$ ($p \leq 1 \times 10^{-5}$). This indicated that the probability of observing the number of differentially abundant proteins I identified was highly unlikely under the null. I also applied this method to the association testing of proteins with severity scores within cases (**Figure 2.14**). Again, on no occasion in 100,000 permutations were more proteins declared significant than in the real data (empirical p -value $< 1 \times 10^{-5}$).

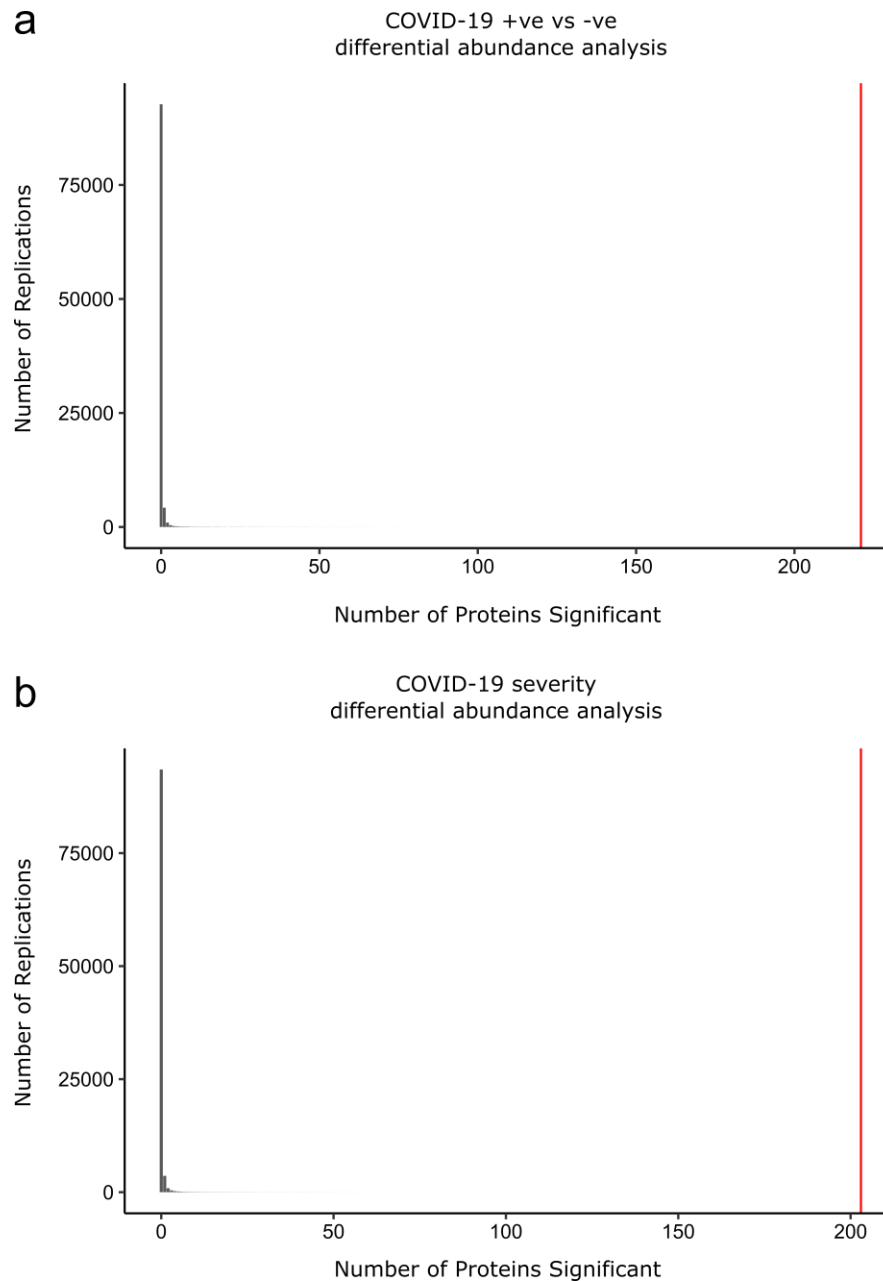


Figure 2.14 Permutation analysis to demonstrate successful FDR control

Histogram showing the distribution of the number of associations declared significant (FDR 5%) after random permutation of class labels (100,000 replications). The vertical red line denotes the number of proteins that were significant in the analysis with the true sample labels. **a)** the COVID-19 positive versus negative differential abundance analysis. **b)** the COVID-19 severity differential abundance analysis.

2.4.6 Comparing the effects of COVID-19 infection and severity

I next considered the similarities and differences between the set of proteins associated with COVID-19 severity and those that were differentially abundant between COVID-19 cases and controls. Most severity-associated proteins were also identified as being dysregulated in the COVID-19-positive versus -negative analysis (140/203; 69%; **Figure 2.15a**). Log fold changes for proteins in COVID-19 versus non-infected patients were also correlated with effect sizes in the severity analysis, such that the proteins most upregulated in cases versus controls also tended to show the greatest upregulation in severe disease (**Figure 2.15b**). However, there were some exceptions that were strongly associated with severity but did not have significantly different levels between infected and non-infected patients. These included CCL20, IL17C and oncostatin M (OSM) (**Figure 2.15c**).

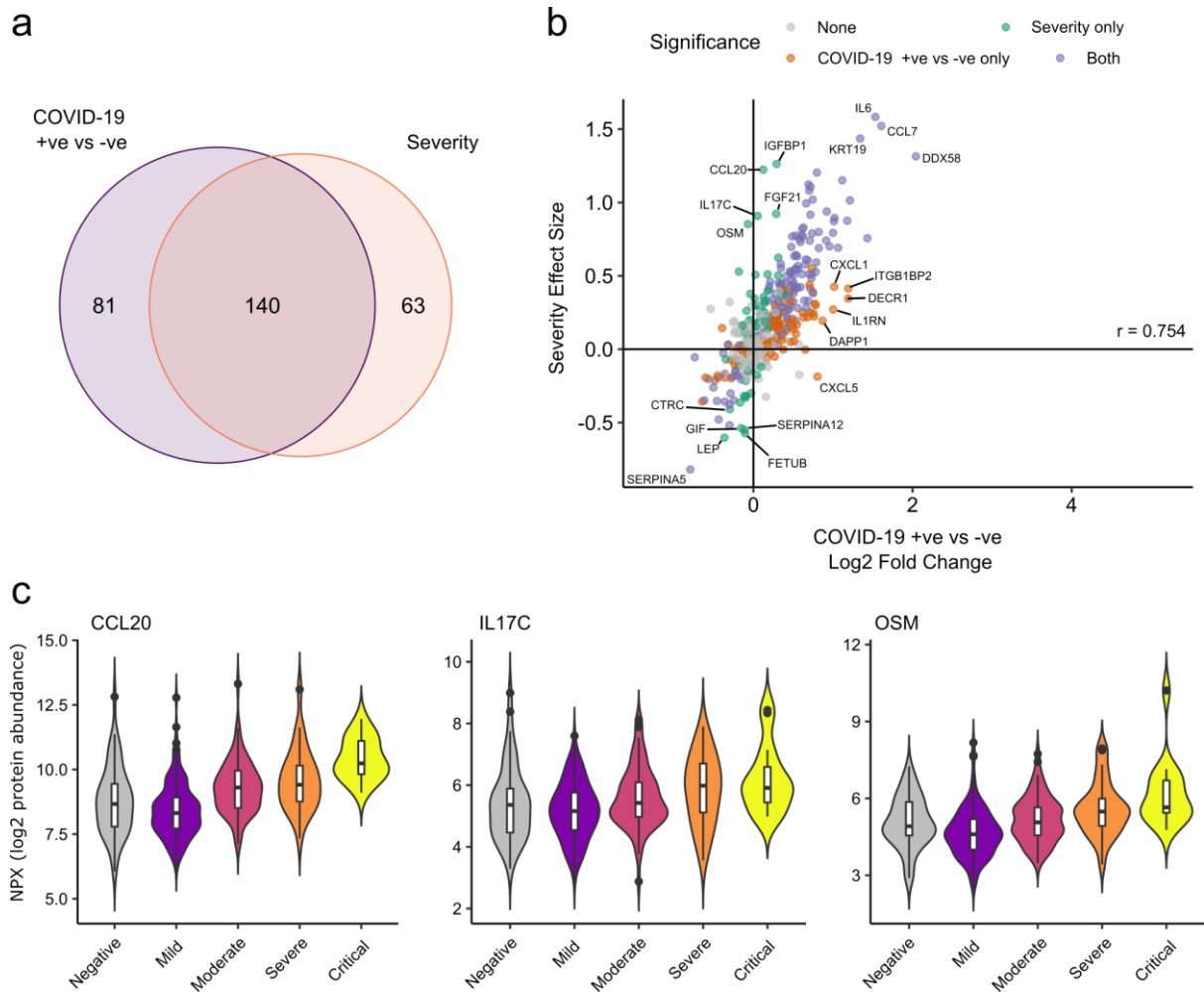


Figure 2.15 Comparison of proteins differentially expressed in COVID-19 with those associated with clinical severity

a) Overlap between the proteins significantly differentially expressed in COVID-19 ($n=256$ COVID-19 samples and 51 non-infected samples) versus those associated with severity (within cases analysis, $n = 256$ samples) (Main cohort). 5% FDR was used as the significance cut-off in both analyses. **b)** Comparison of effect sizes for each protein in the COVID-19 positive versus negative analysis (x-axis) and severity analysis (y-axis). Each point represents a protein. Pearson's r is shown. **c)** Examples of proteins specifically associated with severity but not significantly differentially abundant in the comparison of all cases versus controls. Violin plots showing the distribution of plasma protein levels according to COVID-19 status at the time of blood draw. Boxplots indicate median and interquartile range. $n = 256$ samples from 55 COVID-19 patients and 51 samples from non-infected patients. WHO severity indicates the clinical severity score of the patient at the time the sample was taken. Mild $n = 135$ samples; moderate $n = 77$ samples; severe $n = 29$ samples; critical $n = 15$ samples.

2.4.7 The effect of kidney failure and dialysis on the proteome

While the main aim of the study described in this chapter was to investigate COVID-19, the cohorts studied consisted of ESKD patients on dialysis. These patients are enriched for other diseases^{208,209} (e.g., diabetes and cardiovascular disease) and exhibit a chronic pro-inflammatory state²¹⁰. Furthermore, both dialysis and renal impairment have been shown to affect the abundance of proteins in plasma^{192,193,211–213} (see **Section 2.2.2**). I therefore sought to characterise the differences between ESKD patients on dialysis (with no known current infections) ($n = 51$) and healthy individuals ($n = 35$). PCA revealed that the groups separated into two distinct clusters along the first principal component (PC1) (**Figure 2.16a**). I also attempted to recover these clusters using unsupervised clustering analysis. Consensus clustering with M3C confirmed that there were two clusters in the data ($K = 2$, p -value = 0.005; **Figure 2.16b**) and these perfectly aligned with kidney status.

I next performed a differential abundance analysis between these groups to identify the specific proteins that were dysregulated (see **Section 2.3.10**). Most of the proteins (299/436) measured by the Olink assays were significantly ($FDR < 5\%$) upregulated in ESKD patients undergoing dialysis, several of which had direct links to kidney disease (**Figure 2.16c**). The most upregulated protein was fibroblast growth factor 23 (FGF23), a hormone that suppresses phosphate reabsorption in the kidneys²¹⁴. There was also upregulation of cytokines (e.g., SPP1, CXCL9, CCL15, CX3CL1, IL12B, IL17C, CCL3, CCL14), proteins related to adhesion (e.g., CXADR, CEACAM8, MEPE, COL18A1, VEGFA, F11R, LGALS4) and other immune-related markers (e.g., NCR1, PRSS2, PLAUR, LILRB4, TREM1, DEFA1, PGLYRP1), pointing to a pro-inflammatory state. Other highly upregulated proteins included tumour necrosis factor receptor superfamily members (TNFRSF11A, TNFRSF1A, TNFRSF14, TNFRSF9, TNFRSF1B, TNFRSF10B, LTBR, TNFRSF10A, EDAR), insulin-like growth factor binding proteins (IGFBP1, IGFBP6, IGFBP2, IGFBP7), other growth factors (FGF21, EGFR, GDF15, GH1, HGF), cystatins (CSTB, CST3, CST5), C-type lectin binding domain-containing proteins (CLEC7A, CLEC4D), fatty acid binding proteins (FABP4, FABP6) and myoglobin (MB). The most downregulated protein was uromodulin (UMOD), which is the most abundant protein in urine and may have a protective effect against kidney injury by attenuating inflammation²¹⁵ (**Figure 2.16c**). There were seven other proteins downregulated in the ESKD patients: FCN2, CNDP1, low-density lipoprotein receptor (LDLR), ADAMTS13, epidermal growth factor

receptor (EGFR), selectin P ligand (SELPLG), and cartilage acidic protein 1 (CRTAC1). Collectively, these results indicate that the proteome of ESKD patients on dialysis is distinct from healthy patients, and proteins related to kidney function are particularly affected.

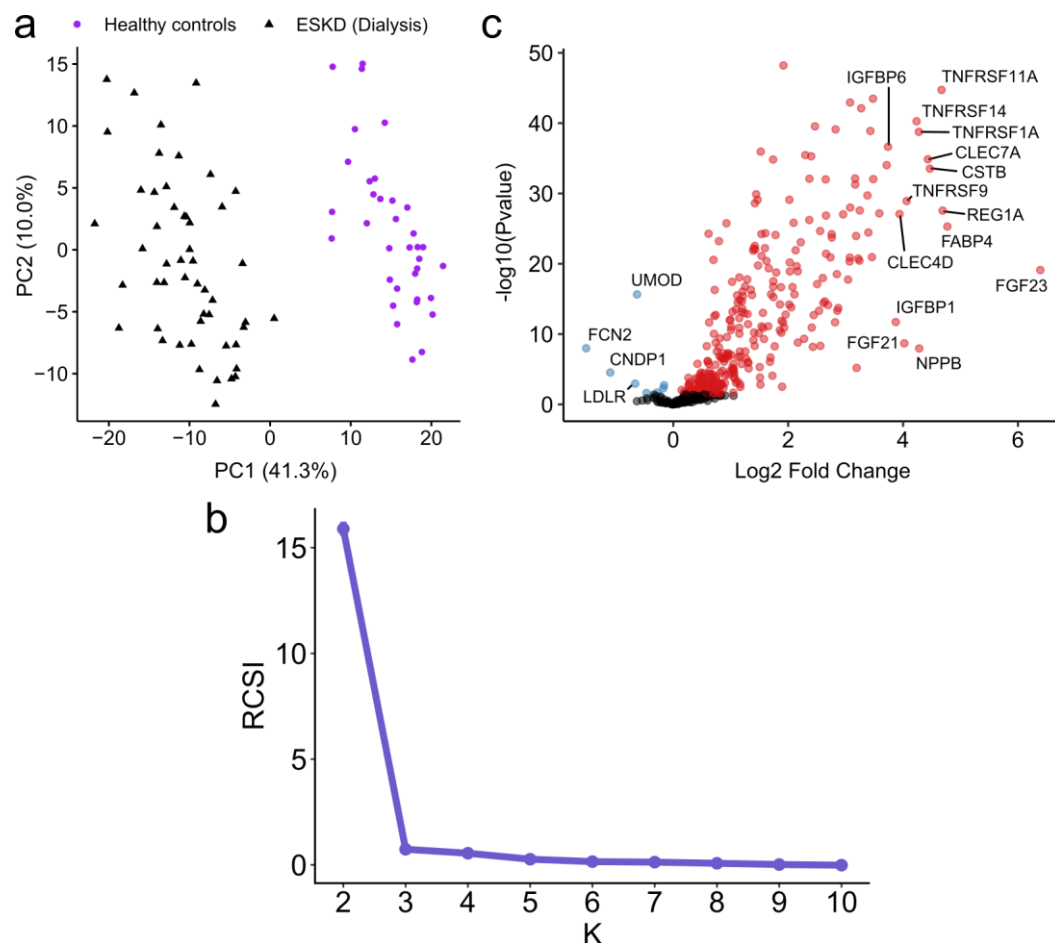


Figure 2.16 The effect of ESKD and dialysis on the plasma proteome

Comparison of non-infected ESKD patients on dialysis and healthy controls. **a)** Principal components analysis for ESKD patients on dialysis and healthy controls. Each point represents a sample. The proportion of variance explained by each principal component (PC) is shown in parentheses on the axis labels. **b)** Relative cluster stability indexes (RCSI) for values of K (i.e., cluster number). Higher RCSI scores indicate more stable clustering. **c)** Proteins up-regulated (red) or down-regulated (blue) in 51 ESKD patients on dialysis relative to 35 healthy individuals (1 sample per patient). Significance was controlled at 5% FDR by Benjamini-Hochberg adjustment; dark-grey = non-significant.

2.4.8 Comparison to COVID-19 patients without ESKD

ESKD patients on dialysis are more likely to have severe COVID-19 (see **Section 1.4.3**) and their plasma proteomes are unique (see **Section 2.2.2**). Therefore, the proteomic response to infection by SARS-CoV-2 may be different in this group compared to healthy individuals. For this reason, I attempted to determine whether the findings described in this chapter were consistent with proteomic studies of patients that were not selected based on kidney status^{75,177,216,217}. Of the 221 proteins that were differentially abundant in the analysis of COVID-19-positive versus -negative ESKD patients, 116 associations had been previously reported in these four studies. Of the 203 proteins associated with severity, 165 had previously been reported. This indicates that the effects of COVID-19 on the proteome are broadly consistent regardless of kidney function.

I focussed in more detail on the study by Filbin *et al.*⁷⁵ due to their cohort's large sample size, the availability of serial measurements, the use of similar modelling strategies and the use of Olink immunoassays. This study comprised 384 patients with acute respiratory distress (306 COVID-19 positive and 79 COVID-19 negative) and measured 1,472 proteins using the Olink Explore platform; this included 417 of the 436 proteins measured in the ESKD cohort. Of the 221 proteins identified by the comparison of COVID-19 cases and controls in the Main cohort, 210 were measured in their study. Of these, 100 (47.6%) were significant in their differential abundance analysis of COVID-19-positive versus COVID-19-negative respiratory distress patients. In addition, there was a strong correlation (Pearson's $r = 0.69$) between the estimated log-fold changes in the ESKD and non-ESKD cohorts (**Figure 2.17**). Of the 203 proteins associated with severity in the ESKD cohort, 192 were measured in the study of Filbin *et al.* Of these, 157 proteins were significantly associated with severity, so there was a concordance of 81.8%. Therefore, despite differences in study design and clinical populations, there were notable similarities between the ESKD cohort and the cohort of Filbin *et al.*

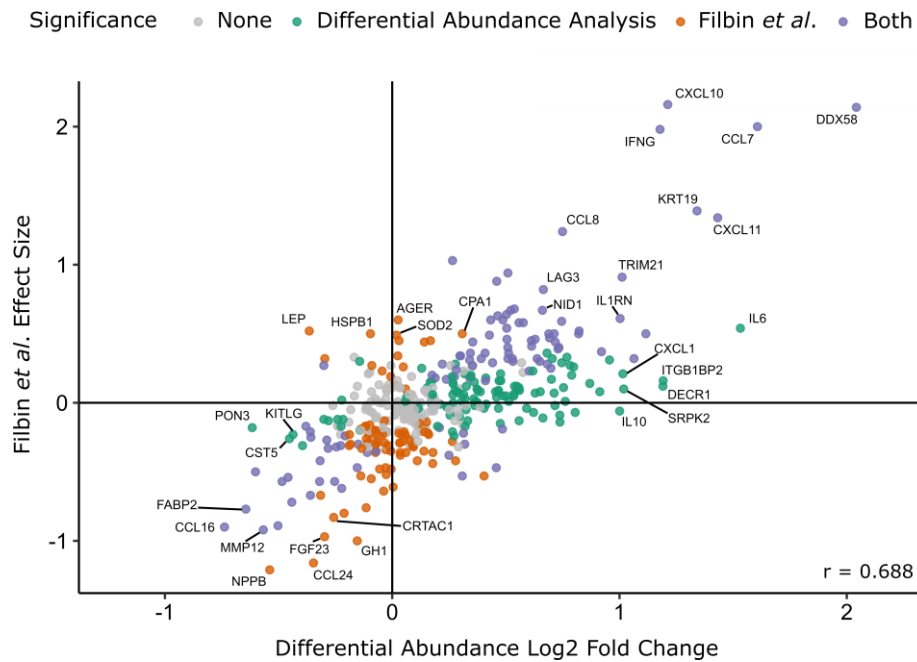


Figure 2.17 Comparison with the report of Filbin *et al.*

Comparison of log2 fold change for COVID-19 positive versus negative ESKD patients in our study versus COVID-19 positive versus negative respiratory distress patients in the report by Filbin *et al.*⁷⁵. Colours indicate whether a protein was significantly differentially abundant in each study. Pearson's r is shown.

2.4.9 Supervised learning

Principal components analysis revealed that some samples from patients who had mild or moderate disease at the time of sampling clustered with samples from patients with severe disease (**Figure 2.9a**). Examination of the same PCA plot this time labelling samples according to the patient's overall clinical course (measured by peak WHO severity score over the duration of the illness) (**Figure 2.9b**), revealed that these samples came from individuals who subsequently developed severe or critical disease. This suggests that there are molecular changes that predate clinical deterioration. To evaluate this further, I used supervised learning approaches to ascertain whether the proteomic signature of the first blood sample for each patient in the Main cohort could identify whether the patient either had clinically severe COVID-19 at the time of sampling or would develop severe disease in the future. Whereas differential expression analyses consider each protein marker separately, there exist supervised learning techniques that allow for the examination of all proteins concurrently; these methods can also capture non-linear relationships between proteins in the dataset. Using random forests, I trained a classifier on the first sample for each COVID-19 patient to predict the overall clinical

course, defined by peak WHO severity. For this analysis, I binarised clinical courses into either WHO mild/moderate or severe/critical.

The random forests method achieved 71% accuracy in predicting peak severity. By contrast, a model that used only clinically available predictors (demographics, comorbidities and clinical laboratory results), achieved only 66% accuracy. Combining both proteomic and clinical predictors into a single model did not improve accuracy (71%) compared to using proteomic predictors alone, suggesting that the information contained in the clinical predictors is captured at the proteomic level. While broad proteomic profiling could not be applied in clinical practice for risk stratification, the features selected by the classifier highlight proteins of potential biological importance. I therefore interrogated the model to identify key proteins by calculating feature importance metrics (see **Section 2.3.17**). The most important proteins for predicting the presence of current or future severe disease were IL18BP, CTSD (Cathepsin D), GDF15, KRT19, TNFSF11 and IL1RL1 (ST2) (**Figure 2.18**).

2.4.10 Proteins associated with risk of death

In the Main cohort, 9/55 patients died. I therefore sought to identify proteins associated with the risk of death. However, standard survival analysis techniques are insufficient to incorporate the serial samples available for this cohort (see **Section 1.3.1.2**). To leverage the dynamic nature of repeated protein measurements for the prediction of death, I utilised joint models, which combine linear mixed models and Cox models^{91,218} (see **Section 2.3.18**). This analysis identified 44 proteins for which increased concentration was associated with increased risk of death (**Figure 2.19, Figure 2.20**), including cystatin C (CST3), IL22RA1, azurocidin 1 (AZU1), CCL28 and spondin 1 (SPOND1). There were also 25 proteins for which increased concentration was associated with a reduced risk of death, including CD84, TNFSF12, TANK, PRKCQ and adrenomedullin (ADM).

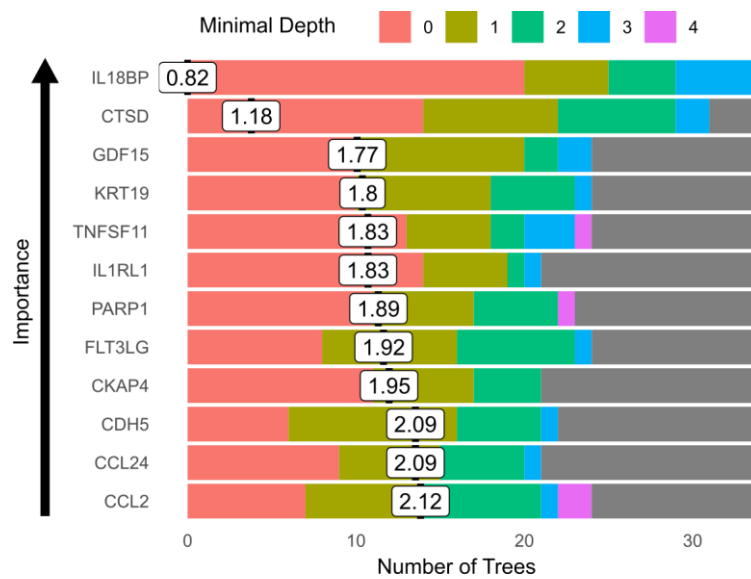


Figure 2.18 Prediction of severe COVID-19

The 12 most important proteins for predicting overall clinical course (defined by peak COVID-19 WHO severity) using Random Forests supervised learning. If a variable is important for prediction, it is likely to appear in many decision trees (number of trees) and be close to the root node (i.e., have a low minimal depth). The mean minimal depth across all trees (white box) was used as the primary feature selection metric.

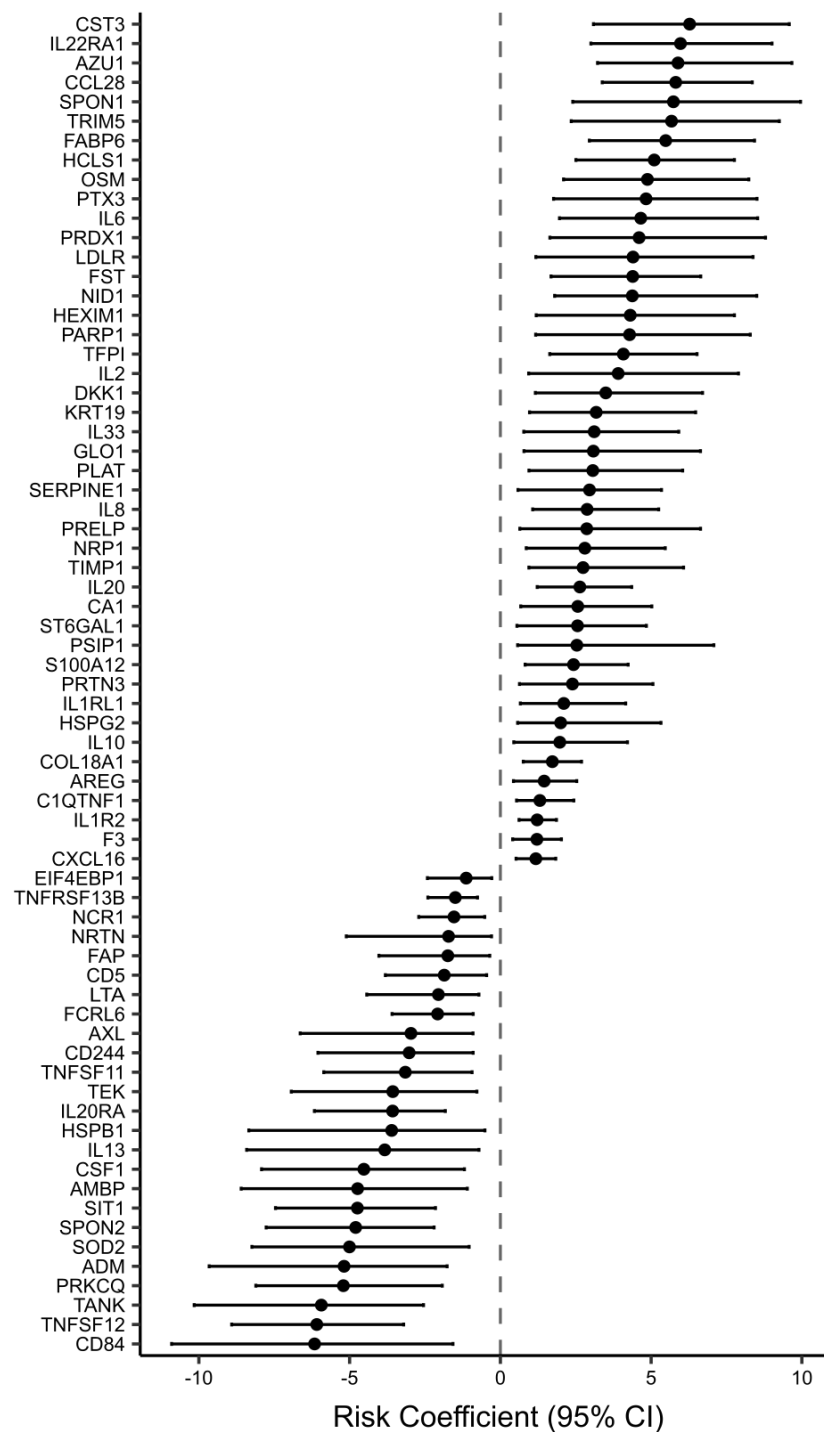


Figure 2.19 Proteins associated with risk of death

Proteins that are significant predictors of death (Benjamini-Hochberg adjusted $p < 0.05$). $n = 256$ samples from 55 COVID-19 positive patients, of whom 9 died. Risk coefficient estimates are from a joint model. Bars indicate 95% confidence intervals. For proteins with a positive risk coefficient, a higher concentration corresponds to a high risk of death, and vice versa for proteins with negative coefficients.

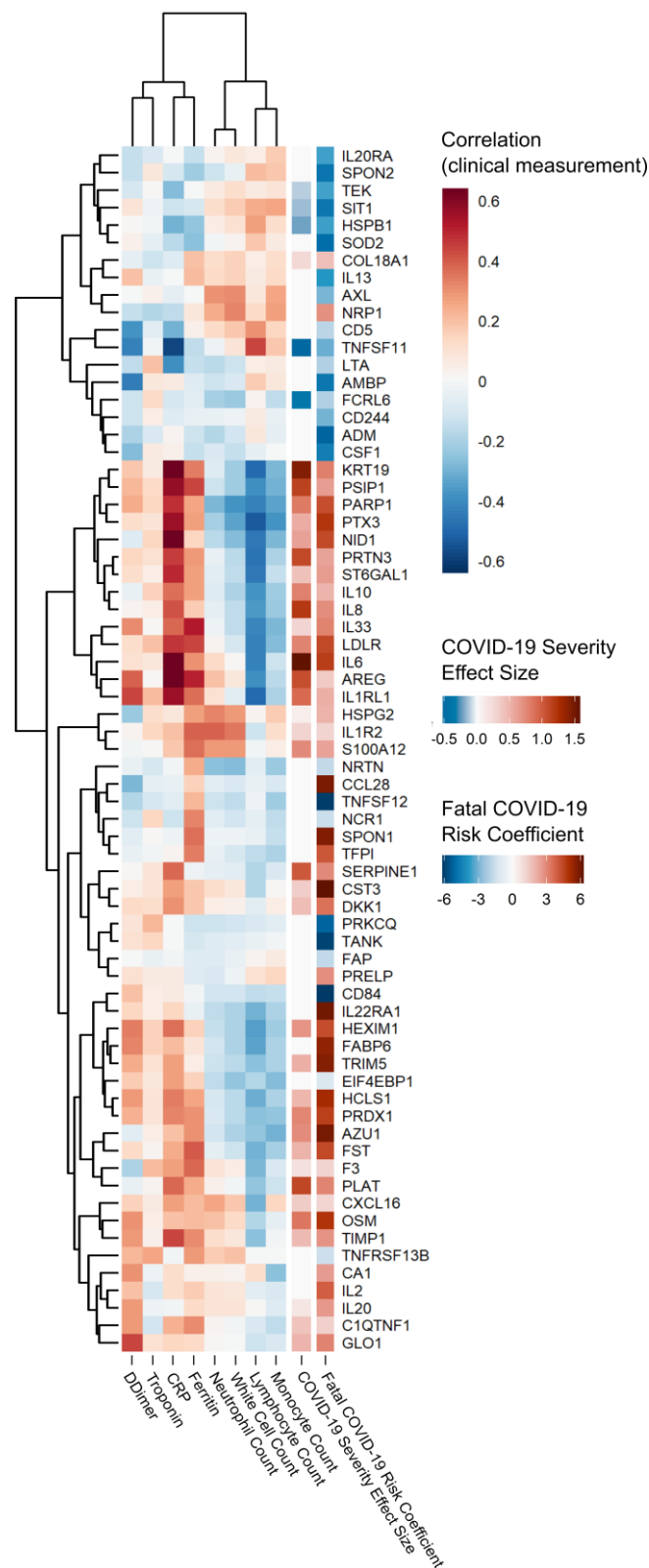


Figure 2.20 Proteins associated with risk of death and their associations with clinical severity and laboratory measurements

Proteins significantly associated with risk of death (5% FDR) according to joint models are shown. The estimated linear effect sizes from LMMs testing for associations with severity are also shown. Correlations between protein levels and contemporaneous clinical laboratory marker values were calculated using `rmcorr`¹⁹⁹. The rows and columns of the clinical marker correlation matrix are ordered by hierarchical clustering.

2.4.11 Proteins with distinct temporal profiles in severe COVID-19

The immune response to infection is dynamic, and therefore snapshot measurements provide only partial insights. Leveraging the dense serial sampling in our dataset (**Figure 2.4**), I modelled the temporal trajectory of each protein and asked whether or not protein trajectories differed in patients with a severe/critical versus mild/moderate overall clinical course. This was achieved using linear mixed models that included a term for the time from first symptoms (in days), another for the overall clinical course (i.e., the peak severity during the course of the illness) and a term for the interaction between time and severity (see **Section 2.3.19**).

178 proteins were significantly associated with time from first symptoms (5% FDR), demonstrating the temporal variability in plasma proteins across the disease course. Moreover, I identified 32 proteins for which there was a significant interaction between time and severity, i.e., proteins displaying differential temporal trajectories between mild/moderate and severe/critical infections (**Figure 2.21, Figure 2.22, Figure 2.23**). The proteins with the most significant interaction effects were the integrins ITGA11 and ITGB6, intracellular adhesion molecule 1 (ICAM1), TNFRSF10B and the urokinase plasminogen activator receptor (PLAUR). The abundance of these proteins rose in the more severe patients but remained stable following COVID-19 onset in milder cases (**Figure 2.21, Figure 2.23**). ACE2, the receptor for SARS-CoV-2, also exhibited this pattern. The profile of ITGA11 contrasted these proteins, as its abundance fell over time in the more severe patients but remained relatively flat in the milder group.

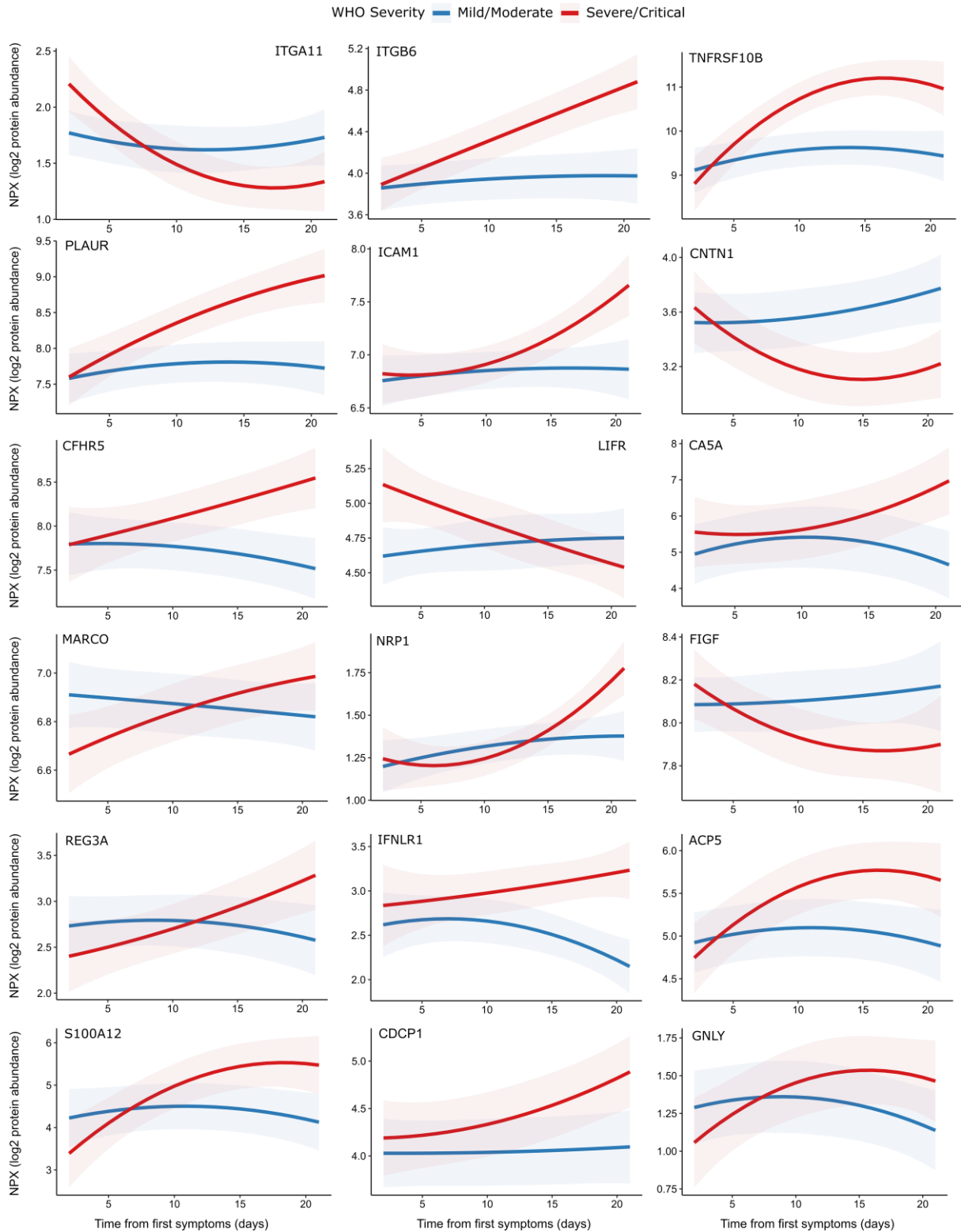


Figure 2.21 Modelling of temporal protein trajectories

The top 18 proteins displaying the most significantly (5% FDR) different longitudinal trajectories between patients with a mild or moderate ($n = 28$) versus severe or critical ($n = 27$) overall clinical course (defined by peak WHO severity). Means and 95% confidence intervals for each group, predicted using linear mixed models (see **Section 2.3.19**), are plotted. The remainder of the significant proteins are shown in **Figure 2.22**. Individual data points are shown in **Figure 2.23**.

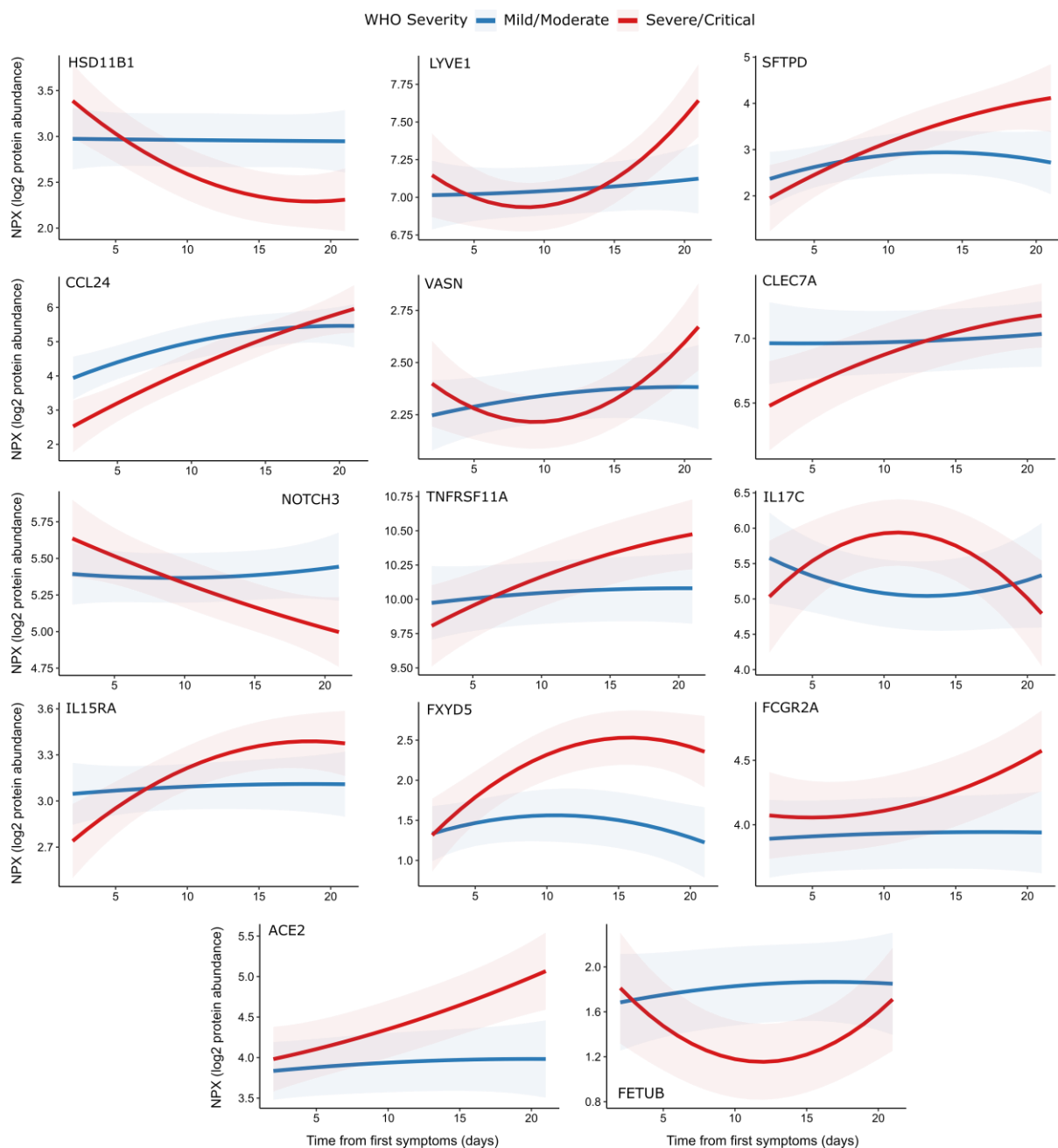


Figure 2.22 Temporal trajectories for proteins with a significant time x severity interaction

Proteins significant at 5% FDR but not shown in **Figure 2.21** are displayed here.

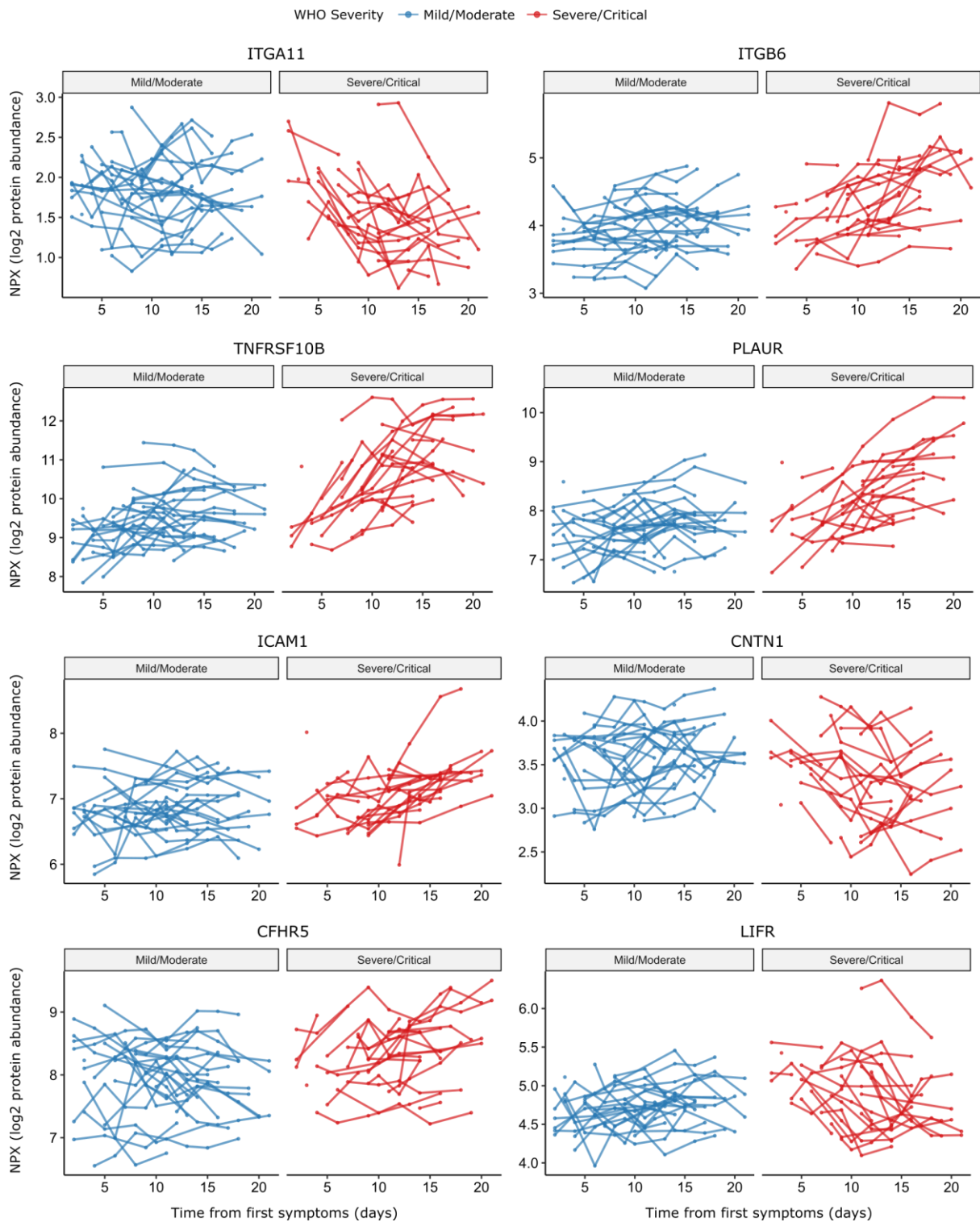


Figure 2.23 Raw data points for modelling of temporal protein trajectories

The 8 most significant proteins from **Figure 2.21** are displayed.

2.4.12 Associations with clinical laboratory tests

A number of routine clinical laboratory tests have well-characterised associations with COVID-19 (e.g., elevated inflammatory markers, d-dimer and reduced lymphocyte count)²¹⁹. I therefore compared our proteomic data from COVID-19 patients at each timepoint to contemporaneous clinical laboratory measurements using linear mixed models (see **Section 2.3.13**). There were associations between plasma proteins and all clinical measurements except troponin (**Figure 2.24**). The proportion of missing values varied for each clinical measurement (**Table 2.4**), which in part explains why some parameters were associated with a greater number of proteins. There was an overlap between the proteins associated with COVID-19 severity, lymphocyte count and C-reactive protein (CRP) (**Figure 2.10, Figure 2.24**). For instance, IL6, KRT19, IFNG and CXCL10 were strongly associated with raised CRP and ferritin and reduced lymphocyte counts. This echoes previous findings that severe COVID-19 is characterised by an enhanced inflammatory state and lymphopenia. Monocyte count was strongly associated with levels of CCL2 and CCL7, monocyte-chemotactic proteins; this may reflect the secretion of these proteins at the site of infection, thereby drawing monocytes out of the blood. Monocyte count was also associated with the monocyte chemokines CCL7 and CCL17, and the monocyte chemoattractant CCL16. Elevated neutrophil count was highly associated with Oncostatin-M (OSM), which regulates IL6, GCSF and GMCSF production, and with the proteases MMP9 and defensin.

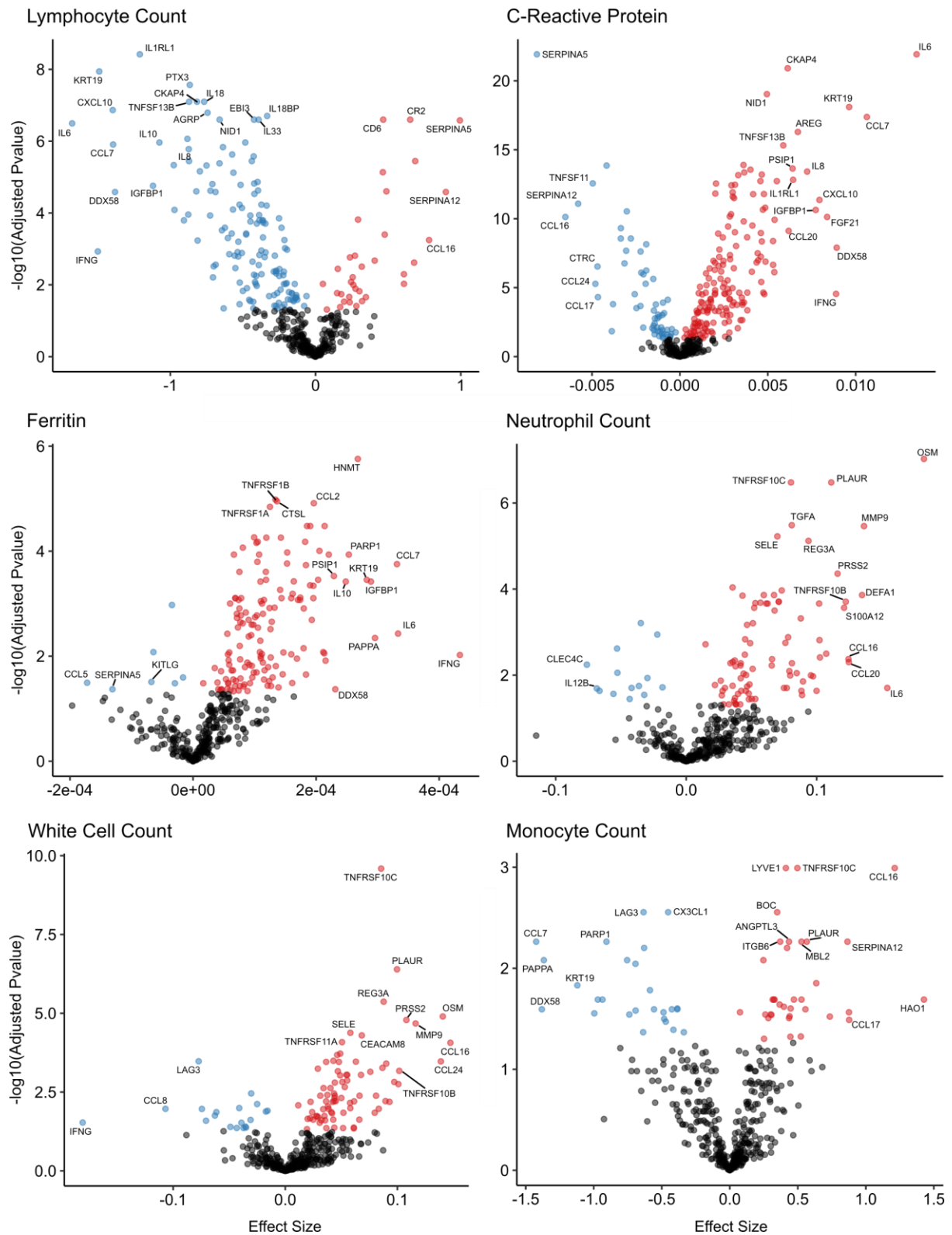


Figure 2.24 Associations of clinical laboratory markers with plasma proteins

Proteins that are positively (red) or negatively (blue) associated with clinical laboratory parameters (5% FDR). p -values from differential abundance analysis using linear mixed models after Benjamini-Hochberg adjustment. Dark-grey = non-significant. Only two significant associations were found for d-dimer (PIGR and EPCAM, both down-regulated).

2.4.13 Associations with ethnicity

In the UK, individuals from ethnic minorities are at higher risk of severe disease and death from COVID-19¹⁷¹. I therefore examined whether any of the proteins we measured exhibited differences between ethnicities (i.e., White, South Asian and Black), analysing COVID-19 positive cases and controls separately (see **Section 0**). In COVID-19 negative ESKD patients, no proteins were significantly associated with ethnicity in a multivariable model adjusting for age and sex. In COVID-19 positive ESKD patients, there was the potential for protein associations with ethnicity to be confounded by disease severity. To account for this, I included severity as a covariate, in addition to age and sex. A single protein, lymphocyte antigen 75 (LY75), was associated with ethnicity (nominal $p = 0.0001$, Benjamini-Hochberg adjusted $p = 0.04$). This protein was elevated in white patients. Using the same within-cases analysis strategy in the Validation cohort, I found no proteins significantly associated with ethnicity after multiple testing correction, although the nominal p -value for LY75 was 0.025.

2.5 Discussion

In this chapter, I described the analysis of 436 plasma proteins in ESKD patients with and without COVID-19. Since these proteins were selected based on their roles in inflammation or cardiovascular processes, this investigation focussed on key proteins that are likely to be functionally relevant in COVID-19. However, this bias also limited the scope of the current study to a small portion of the proteome compared to more wide-angled immunoassays and mass spectrometry-based approaches. Furthermore, the overrepresentation of immune- and cardiovascular-related proteins prevented the use of pathway analyses.

Comparison of COVID-19 cases and controls revealed upregulated markers of the interferon response to viral infection, including interferon-gamma (IFNG) and RNA sensor RIG-I (DDX58). DDX58 recognises cytoplasmic viral RNA, leading to the production of type 1 interferons and pro-inflammatory cytokines. However, SARS-CoV-2 N protein can bind to DDX58, which may disrupt its ability to discriminate between self- and non-self-RNA²²⁰. There was also upregulation of numerous inflammatory cytokines, among which was C-X-C motif chemokine ligand 10. CXCL10 is induced by IFNG to regulate inflammation, including the recruitment of effector T cells, natural killer cells and monocytes to the site of infection; it, along with IL6, have been noted as a key marker of severe COVID-19²²¹. There was also upregulation of intracellular epithelial proteins, including KRT19, a filament protein, and PSIP1, which is associated with stress-induced apoptosis.

ESKD patients with COVID-19 were matched with non-infected controls based on age, sex and ethnicity (**Figure 2.1**). However, this does not negate the possibility of measured or unmeasured confounders between the cases and controls. For instance, there were differences in the prevalence of diabetes and the underlying causes of ESKD between COVID-19-positive cases and controls (**Table 2.1**), although adjusting for these variables did not materially affect results (**Figure 2.12**). This is not to say, however, that the current study is free from selection bias, leading to unmeasured differences between the cohorts. Additionally, the healthy controls recruited as part of this study were not so well matched with the non-infected ESKD patients. Of note, the ESKD patients were substantially older than the healthy controls (ESKD median age = 70, healthy median age = 47) (**Table 2.2**). There was also a higher proportion of males in the ESKD group (71%) compared to the healthy controls (51%).

Furthermore, blood was drawn during the course of haemodialysis for ESKD patients, which could represent a source of technical variation. These differences represent clear limitations for the comparisons between ESKD patients on dialysis and healthy controls carried out in this chapter, as these results could be confounded with biological and technical factors.

The comparison of COVID-19 positive and negative ESKD patients revealed 221 proteins that were differentially abundant in COVID-19, over half the number that were measured (50.7%). Permutation analyses indicated that the false discovery rate was appropriately controlled at a level of 5%. Analysis of a second cohort of COVID-19 positive patients revealed consistent results, providing validation of these findings. However, the controls for the Validation cohort were a subset of the controls sampled as part of the Main cohort. Therefore, this comparison is limited by the fact the Validation cohort was not completely independent.

COVID-19 patients did not form distinct groups according to their severity levels in PCA (**Figure 2.9**) and consensus clustering indicated there was no significant structure in the data. However, there was a gradient of severity from mild to critical disease. Analysis of COVID-19 severity at the time of sampling with linear mixed models revealed 203 associated proteins (**Figure 2.10**). As for the comparison of cases and controls, many cytokines and chemokines were associated with COVID-19 severity. Interleukin-6 (IL6) was the most upregulated protein with respect to contemporaneous COVID-19 severity (**Figure 2.11**), which has also been reported as being upregulated in severe COVID-19 by previous studies^{75,180}. IL6 is produced in response to infections and tissue injury¹⁶²; it induces the synthesis of acute phase proteins (such as C-reactive protein and serum amyloid A), leading to the stimulation of inflammatory pathways. The IL6 receptor antagonist tocilizumab has been shown to improve survival in critically ill COVID-19 patients¹⁶², indicating that excessive IL6-IL6R signaling contributes to severe disease. The prioritisation of IL6 by differential abundance testing demonstrates the value of plasma proteins in the discovery of candidate drug targets. While cytokines and chemokines are key mediators of the response to COVID-19, it should be noted that the Olink panels used in this study were highly enriched for the Reactome term “Cytokine Signaling in the Immune system” ($p\text{-value} = 5.3 \times 10^{-54}$), with 107/436 proteins being listed under this pathway. So, it is unsurprising that these proteins make up a high proportion of the results.

Most of the severity-associated proteins were also differentially abundant in the comparison of COVID-19 infected versus non-infected patients (69%) (**Figure 2.15a**), and for many proteins the effect size of severity was proportional to the fold change between cases and controls (**Figure 2.15b**). This suggests that, for many proteins, the difference between severe and mild COVID-19 and non-infected patients is similar, but more substantial in more severe patients (i.e., severe patients display a similar, but exaggerated, immune response compared to milder patients, as opposed to severe disease being mediated by an orthogonal set of proteins and pathways). This is consistent with the fact that the proteomic differences between controls and patients with mild COVID-19 in PC space occur on the same axes that separate controls and severe patients (**Figure 2.9**). There were, however, exceptions where proteins associated with severity were not differentially abundant between cases and controls (e.g., OSM, IL17C and CCL20). These proteins may reflect processes that are specific to severe disease, which could indicate their promise as therapeutic targets.

ESKD patients are at increased risk of severe COVID-19^{208,209} and renal impairment itself is an independent risk factor for fatal COVID-19¹⁷¹. Renal function has been linked to changes in the proteome^{211–213}, and haemodialysis may further alter plasma protein abundance^{192,193} (see **Section 2.2.2**). I further investigated the effects of ESKD and dialysis on the proteome by comparing these patients with healthy controls. This revealed marked dysregulation of the proteome in ESKD, and unsupervised consensus clustering identified groupings that perfectly aligned with kidney status. The protein most highly upregulated in ESKD patients was FGF23, a hormone that controls the reabsorption of phosphate by the kidney; its abundance has previously been shown to predict the progression of renal disease²¹⁴. The most downregulated was Uromodulin, the most abundant protein in urine. The clear links between these proteins and kidney function suggest that the differential abundance analysis reflects at least in part associations with kidney failure, as opposed to other biological or technical factors. Another highly upregulated protein with well-known links to kidney disease was Cystatin C (CST3), which is used as a marker of glomerular filtration²²². CST3 is filtered from plasma by the glomerular capillaries, and so its abundance in plasma closely reflects kidney function. Also upregulated was Cystatin B (CSTB), an intracellular protein used as a urinary marker of proximal tubule injury which may be released into urine and serum upon epithelial cell death^{223,224}.

Decline in kidney function is associated with inflammation. This was illustrated by the upregulation of a wide array of immune-related markers in the ESKD patients relative to the healthy controls, including cytokines and adhesion molecules. There was also upregulation of many Tumour Necrosis Factor Receptor Superfamily (TNFRSF) members and Insulin-like Growth Factor-Binding Proteins (IGFBP), which may reflect elevated inflammation and renal cell death. Among these was IGFBP1, which has been linked to glucose tolerance, blood pressure and cardiovascular disease²²⁵. Regenerating Family Member 1 Alpha (REG1A) was highly upregulated, a protein secreted by the pancreas that is associated with islet cell regeneration; its abundance is elevated in patients with diabetes²²⁶. Also upregulated in ESKD was Pro-Brain Natriuretic Peptide (NPPB), which has key roles in cardiovascular homeostasis, and high plasma abundance can be indicative of myocardial injury²²⁷. NPPB is also known to be elevated in renal injury and may be stimulated by reduced glomerular cardiac function²²⁷. Fatty acid binding proteins (FABP) have also been shown to be raised not only in the plasma of diabetic patients²²⁸ but also in obesity and atherosclerosis²²⁹. In the present study, FABP4 and FABP6 were elevated in the plasma of patients with ESKD. Furthermore, there was downregulation of the Low-Density Lipoprotein Receptor (LDLR) in the ESKD patients, which mirrors previous findings in patients with chronic kidney disease²³⁰. Reduced levels of LDLR may contribute to dyslipidaemia, which could, in turn, increase the risk of cardiovascular disease²³⁰. Also downregulated was ADAMTS13, which cleaves von Willebrand Factor, and SLPLG, which is involved in the adhesion of leukocytes to activated platelets. All considered, there were a range of proteins dysregulated in the ESKD patients with links to obesity, diabetes, cardiovascular disease, and thrombosis, all of which are comorbid with kidney disease^{208,209}. On one hand, these proteins could reflect the direct effect of kidney failure, however they could also simply reflect a higher rate of these disorders in the ESKD cohort.

While ESKD and dialysis have a sizeable effect on the proteome, both the COVID-19 cases and controls had ESKD and were undergoing haemodialysis. Therefore, analysis by COVID-19 status should not reflect associations of proteins with renal function. However, it does not necessarily follow that these results can be generalised to patients with normal renal function. As discussed previously (see **Section 1.4.3**), ESKD patients have a reduced ability to respond to viruses and suffer from chronic low-grade inflammation²¹⁰. Therefore, their immune response to COVID-19 may not reflect that of otherwise healthy patients. For this reason, I

compared the proteins differentially abundant in COVID-19 patients with ESKD to another study, carried out by Filbin *et al.*⁷⁵. Despite differences in clinical populations (i.e., ESKD vs. non-ESKD) and control groups (i.e., non-infected vs. active non-COVID-19 respiratory infection), the effect sizes in both studies were highly consistent (**Figure 2.17**). The associations with COVID-19 severity observed in the Main cohort were also highly concordant with the findings of Filbin *et al.*⁷⁵, with 81.8% of the proteins associated with severe COVID-19 in ESKD patients being reported in their data. These data suggest that the inflammatory response to COVID-19, as reflected in the plasma proteome, is similar regardless of kidney disease. While they require validation in cohorts of patients with normal renal function, the findings presented in this chapter may be applicable to patients with normal kidney function.

Despite a relatively small sample size (27 severe/critical and 28 mild/moderate individuals), supervised learning with random forests was able to predict peak severity based on each individual's first sample with an accuracy of 71%. Predictive performance was not improved upon including clinical predictors in the model, but a model based only on these predictors was less accurate (66% versus 71%), indicating the potential prognostic value of plasma proteins for future clinical deterioration. However, it should be noted that some patients had reached their peak clinical severity at the time their first blood sample was taken. Furthermore, this model was not tested in an external validation cohort, so its generalisability may be limited. The purpose of this analysis, however, was not to create a clinical risk prediction model, but to identify features of potential biological importance. The pro-inflammatory cytokine interleukin-18-binding protein (IL18BP) had the lowest mean minimal depth of all proteins. IL18BP functions as a decoy receptor that inhibits the binding of IL18 to its receptor, thus downregulating interferon gamma (IFNG). IL18BP is also itself induced by IFNG in a negative feedback loop²³¹. Both IFNG and IL18BP were upregulated in severe COVID-19. The importance of IL18BP for predicting COVID-19 severity could therefore be a result of its association with the IFNG-dependent response to viral infection. KRT19, a cytokeratin expressed in epithelial cells, was both highly associated with severe COVID-19 and emerged as a key predictor according to random forests. This protein is likely a marker of lung epithelial damage²³², as these cells are targeted by SARS-CoV-2. Additionally, proteins involved in the repair of epithelial tissues (including PSIP1, AREG and GRN) were positively associated with COVID-19 severity. The second most predictive protein, the lysosomal peptidase cathepsin D

(CTSD), may also be required by involved in epithelial cell renewal²³³. Another predictor, PARP1, has roles in immune cell recruitment and transcription of proinflammatory cytokines, so may contribute to a proinflammatory state. PARP1 has also been linked to cardiovascular and inflammatory disorders²³⁴. Growth differentiation factor 15 (GDF15) has also been linked to aberrant inflammatory and vascular processes and is produced following kidney and lung injury²³⁵. In these predictors, there was a disproportionate representation of proteins related to epithelial and endothelial cell damage, likely reflecting the destruction of these tissues in severe COVID-19.

While many studies have considered the impact of COVID-19 on the plasma proteome, they have largely ignored temporal information (see **Section 2.2.1**). Most studies perform differential abundance analyses, similarly to those carried out in **Section 2.4.3**. This simplistic approach ignores the fact that the response to SARS-CoV-2 infection is not binary, but a dynamic process. Clinical deterioration does not typically occur until 7-10 days following the first appearance of symptoms, indicating distinct phases of infection. This suggests that temporal information could improve our understanding of COVID-19 and inform the optimal timing of therapeutic interventions. The ESKD patients in this study were required to attend regular haemodialysis sessions, enabling the collection of serial samples and allowing for the longitudinal modelling of proteins following COVID-19 symptom onset. Another advantage of studying patients on dialysis was that they allowed the collection of samples in both the outpatient and inpatient settings. Most people with mild cases of COVID-19 were advised to stay at home and continue to isolate. Patients undergoing haemodialysis, on the other hand, had to continue to attend regular haemodialysis appointments, allowing for the inclusion of samples with mild and moderate disease. This presented the unique opportunity to compare the disease course of patients with severe disease courses to those with a more indolent course. Using these data, I was able to identify proteins that had different temporal trajectories conditional on COVID-19 severity.

Proteins could be roughly categorised into those that rose in abundance following symptom onset in more severe cases, and those that fell. ACE2, the receptor for SARS-CoV-2, exhibited a rising profile in severe COVID-19 but its abundance was relatively stable in milder patients. Baseline plasma ACE2 abundance was also positively correlated with severity in the data of Filbin *et al.*^{75,236}, and patients with kidney disease had higher ACE2 levels²³⁶. These data

appear to contrast findings in mice that indicate that *ACE2* expression is protective against severe acute lung injury²³⁷. The Olink immunoassays are measuring *ACE2* levels in plasma, and so likely do not mirror the abundance of membrane-bound *ACE2* in the lungs and vascular endothelium. Increased circulating *ACE2* could indicate greater tissue damage in severe disease. Circulating S100A12 (ENRAGE) and ICAM1 rose following the onset of severe COVID-19. S100A12 binds to AGER (RAGE) to upregulate pro-inflammatory cytokines and cell adhesion molecules, including ICAM1. S100A12 is secreted by leukocytes (neutrophils, monocytes and macrophages) and epithelial cells in the lungs. Other studies of plasma proteins and the expression profiles of cells in the blood have implicated S100A12^{182,217,238}. Additionally, a study of acute respiratory distress syndrome also found that the abundance of S100A12 and soluble RAGE was increased in bronchoalveolar lavage fluid and endothelial cells²³⁹. S100A12 contributes to a proinflammatory state, including activation of pulmonary tissues and accumulation of neutrophils, disrupting the integrity of the vascular endothelium. These processes are also reflected in the upregulation of ICAM1, which is expressed by endothelial cells. The integrins ITGB6 and ITGA11 displayed inverse profiles, with ITGB6 rising in more severe disease and ITGA11 falling. Both proteins are involved in wound healing and fibrogenesis, and their dysregulation may reflect lung injury in severe disease. ITGA11 binds to type 1 collagens that accumulate in fibrotic tissues^{240,241}; its decreased abundance in plasma may therefore reflect increased binding during fibrosis in severe COVID-19. Many of the proteins associated with COVID-19 severity, identified as key predictors by random forests and found to have dynamic profiles in severe disease have links to the destruction of pulmonary and vascular tissues. The impaired vascular endothelium is characteristic of patients with COVID-19 and contributes to the prothrombotic state observed in these individuals^{143,144,242}.

After considering the differences between patients with mild and more severe disease courses, I next considered proteins that were specifically dysregulated in patients who died. Many studies investigating this question have considered the extent to which proteins measured at baseline (i.e., immediately following the diagnosis of COVID-19) are associated with risk of death using Cox regression^{243,244}. However, in the present study there were multiple measurements available for each individual; incorporating this information into survival analysis may be more informative than using baseline measurements alone²⁴⁵. Linear mixed

models, however, are not appropriate for modelling time-to-event data, and standard Cox regression models cannot model time-varying covariates. Joint models, on the other hand, can incorporate an LMM submodel (for estimating the change in protein abundance over time) with a survival model (for estimating the association with time to death). Such a model can estimate the contemporaneous association of protein abundance with time of death²⁴⁵.

Joint survival models revealed that the temporal abundance of 44 proteins was associated with an increased risk of death (**Figure 2.19**). These included CST3, which was also predictive of clinical deterioration in the study of Demichev *et al.*¹⁸¹. CST3 is also used as a marker of kidney function. IL22RA1 was also highly associated with the risk of death; it forms a heterodimer with IL10RB to act as a receptor for IL22, which is linked to lung epithelial cell survival and induces pro-inflammatory cytokines, including S100 proteins. The expression of the IL22 receptor is upregulated in bronchial epithelial cells in response to influenza infection and plays an antimicrobial role²⁴⁶. AZU1 and PRTN3 were also associated with the risk of death, and the genes encoding these proteins are adjacent to each other on chromosome 19; both are components of azurophil granules, specialised neutrophil lysozymes with anti-microbial properties that fuse with phagocytic vacuoles. Myeloperoxidase (MPO), another protein found in azurophilic granules, was associated with severe COVID-19, but not risk of death. It should be noted, however, that the proteins measured in the current study were themselves enriched for “Neutrophil degranulation” ($p\text{-value} = 1.8 \times 10^{-13}$), which included 45/436 proteins. It was therefore inevitable that this study would uncover markers related to neutrophils.

Among the 25 proteins associated with reduced risk of death (**Figure 2.19**) was adrenomedullin (ADM), which has a protective effect in sepsis as it can counteract excessive production of proinflammatory cytokines and protect against endothelial damage^{247,248}. Considering this, and the apparent protective effect of Adrenomedullin in the present study, it is plausible that treatment with ADM could also be beneficial for patients with severe COVID-19. Some proteins associated with reduced risk of death had both pro- and anti-inflammatory roles. While TNFSF12 can induce the production of inflammatory cytokines, it can also inhibit the innate immune response²⁴⁹ and promote tissue repair²⁵⁰. While some reports suggest IL13 could drive COVID-19 severity²⁵¹, it is also known to inhibit the production of inflammatory cytokines²⁵². Many studies that consider many molecular measurements simultaneously,

including the one described in this chapter, perform simple comparisons between groups that describe biomarkers as up- or down-regulated. While this approach can reveal important markers, it encourages a simplistic view of their function and, in the context of COVID-19, means that many proteins are considered to either drive or constraint the immune response. However, many proteins have multiple functions and can modulate inflammation in different ways depending on context.

Most of the participants in the present study were of non-white ethnicity. These patients are at higher risk of severe COVID-19, making them an important group to study¹⁷¹. These data additionally presented an opportunity to consider molecular features that distinguish ethnicity in patients with COVID-19. A single protein, lymphocyte antigen 75 was significantly associated with ethnicity in the COVID-19 patients of the Main cohort, but there was not clear evidence that it confers a differential risk of COVID-19 outcomes. Furthermore, given that these data are observational, this association could be influenced by unmeasured confounders or failure to completely adjust for COVID-19 severity. While these analyses failed to identify substantial ethnicity-related variation in the measured proteins, an important caveat is that there were relatively modest numbers of individuals from each broad ethnic group, so statistical power was limited. Larger multi-ethnic studies are needed to adequately consider ethnicity-related differences in the response to COVID-19.

In summary, the data presented in this chapter provide insight into the dysregulation of the plasma proteome during COVID-19 infection and in severe disease. The collection of serial samples permitted the characterisation of the dynamic profiles of plasma proteins following COVID-19 infection. These results highlight markers of inflammation and neutrophil activation and point to the destruction of the lung epithelium and vascular endothelium. These findings appear to be largely applicable to patients with normal kidney function and could inform the selection of novel therapeutic targets.

Key findings:

- Identification of dysregulated proteins in severe COVID-19, including markers that point to the disruption of the lung epithelium and vascular endothelium.
- An in-depth characterisation of the temporal profiles of proteins following COVID-19 symptom onset in both outpatients and inpatients.

- Identification of proteins that predict severe disease and markers associated with survival.

Future directions:

- There is a need to differentiate between proteins that drive severe COVID-19, versus those that are simply downstream markers of inflammation. The present study was followed up by a Mendelian Randomisation study that aimed to use genetic variants to identify causal mediators of severe disease (Chapter 3).
- The present study focussed on key markers of inflammation but may have missed relevant targets that were not measured. It would be beneficial to consider less commonly studied proteins and integrate other -omics modalities (Chapter 4).

3 Evaluating the causal effects of proteins on severe COVID-19

3.1 Chapter summary

Numerous studies have employed proteomic profiling to identify markers of severe COVID-19. However, most of these studies are observational, so they cannot distinguish between proteins that are downstream markers of clinical deterioration, versus those that drive the development of severe disease. Mendelian randomisation (MR), on the other hand, uses genetic variants as instrumental variables to identify risk factors that have a causal effect on an outcome. In the present study, we followed up 157 proteins known to be robustly associated with COVID-19 severity using MR. This investigation was facilitated by protein quantitative trait loci (pQTL) provided by the SCALLOP consortium and meta-analyses of COVID-19 genome-wide association studies performed by the Host Genetics Initiative. Using MR, we identified five proteins with the potential to be causal mediators of severe disease, but ultimately none of these proved to be robust.

MR assumes that genetic variants only affect the outcome through the risk factor. We assessed the plausibility of this assumption using sensitivity analyses. These methods failed for LGALS9, TNFRSF10A and CCL2, reducing confidence in these proteins as plausible causal mediators. Most sensitivity analyses indicated that EPHB4 was a strong candidate, however the effect of this protein on severe COVID-19 was predominantly mediated by *trans*-pQTLs. Since *trans*-pQTLs do not have a direct link to the protein whose abundance is being modulated, there is a greater chance of horizontal pleiotropy. The strongest evidence of a causal effect was for the Fas cell death receptor (FAS). However, these findings did not replicate in a larger meta-analysis of the effects of variants on very severe COVID-19. Therefore, we did not identify evidence for a causal effect of any of the proteins studied on the risk of COVID-19 outcomes. Further investigation revealed that the FAS pQTL was not mediated through gene expression. Instead, colocalisation analysis indicated that this pQTL arises from differential splicing of *FAS* exon 6, which encodes its trans-membrane domain. It is likely that the FAS protein assay measures soluble FAS, which acts as a decoy receptor for the FAS ligand (FAS), thereby reducing signaling by membrane-bound FAS.

3.2 Introduction

Many studies have applied molecular profiling techniques to measure the abundance of circulating plasma proteins in COVID-19^{177–182}. One such study was described in **Chapter 2** of this thesis¹⁷⁶. Another study by Filbin *et al.*⁷⁵ also applied Olink immunoassays to individuals infected by SARS-CoV-2. These observational studies assessed associations between plasma proteins and COVID-19 severity. This chapter aims to evaluate whether these COVID-19-associated proteins play a causal role in mediating severe COVID-19, primarily using two-sample Mendelian randomisation.

3.2.1 Genetics of COVID-19

Numerous studies have performed genome-wide association studies (GWAS) for COVID-19 outcomes. One of the largest of these studies involved the Genetics Of Mortality In Critical Care (GenOMICC) cohort, which involved 2,244 patients critically ill with COVID-19 and ancestry-matched controls from UK Biobank²⁵⁴. In this study, fifteen independent loci reached genome-wide significance (p -value $< 5 \times 10^{-8}$). The most significant variant was situated on chromosome 3, although this locus was very broad and encapsulated several genes, including *SLC6A20*, *LZTFL1*, *CXCR6*, *CCR2* and *CCR3*. Other loci included those nearby the *TYK2*, *DPP9* and *IFNAR2* genes and one locus was situated in the *OAS* gene cluster (*OAS1*, 2 and 3). Transcriptome-wide association analysis (TWAS) was performed, which involved predicting gene expression from genotype and identifying transcripts associated with COVID-19 status. This approach prioritised the genes *CXCR6*, *CCR3*, *CCR2*, *MAT2B* and *OAS3*. Mendelian randomisation (MR) was performed using GTEx eQTLs¹¹⁵ as instruments, which revealed a potential causal role for *IFNAR2* and *TYK2*. The COVID-19 Host Genetics Initiative performed meta-analyses of 46 distinct studies, which combined genetic data from 49,562 COVID-19 cases, 6,179 of which were critically ill. After adjustment for both multiple testing and the number of traits measured, thirteen significant loci were identified, including the aforementioned associations identified by GenOMICC. While these studies implicated multiple loci, the molecular mechanisms by which these variants exert their effect on COVID-19 outcomes were not immediately clear. The study described in this chapter aimed to elucidate these mechanisms under the hypothesis that they were mediated by plasma proteins.

3.2.2 Contributions

Most of the results described in this chapter were made available as part of a *medRxiv* preprint²⁵³. This was a highly collaborative project that involved both analytical contributions and sharing of data from many sources. I developed the computational pipeline that performed LD clumping and harmonisation, Mendelian randomisation (MR), and sensitivity analyses, although the overarching design of the pipeline was discussed with many collaborators. Lucija Klaric, Artemis Papadaki and I used this pipeline to perform proteome-wide MR analyses to identify factors that may cause COVID-19. I performed the colocalisation analyses and Lucija Klaric performed the conditional analyses. Alex Tokolyi performed the splicing analysis described in **Section 3.3.12**, which was based on data from the INTERVAL cohort. The project was facilitated by analytical work and data provided by members of the Systematic and Combined AnaLysis of Olink Proteins (SCALLOP) consortium. Cohort-specific acknowledgements are provided in Supplementary Table 2 of the preprint²⁵³. The SCALLOP pQTL data were meta-analysed by Erin Macdonald-Dunlop, Lucija Klaric, Jing Hua Zhao and Marisa D. Muckian. The project also relied on the UK Biobank resource, approved under application 19655, and data from numerous studies of the genetics of COVID-19, which were meta-analysed by the COVID-19 host genetics initiative.

In summary, I developed the main computational pipeline used to generate MR results, perform sensitivity analyses and apply colocalisation tests. Some of the text, figures and tables in this chapter are repurposed from the *medRxiv* preprint²⁵³. I created the figures and tables presented in this chapter, except for **Figure 3.8a**, which was created by Alex Tokolyi, **Figure 3.8b**, which was based on the design of James E Peters, and **Figure 3.5** and **Figure 3.4a**, which were created by Lucija Klaric.

3.3 Methods

3.3.1 Identifying COVID-19-associated variants

Associations of genetic variants with COVID-19 outcomes were based on the meta-analyses performed by the COVID-19 Host Genetics Initiative (HGI) (<https://www.covid19hg.org/>). Specifically, we selected meta-analyses that HGI calculated based on individuals with European ancestry, to match the design of the pQTL studies. The analyses described in this chapter are based on the 5th HGI release (release date 7th January 2021). Data from 23-and-me were excluded. Data were obtained for three COVID-19 outcomes (**Figure 3.1**):

- **A2**: Very severe respiratory confirmed COVID-19 (compared to population)
- **B2**: Hospitalised COVID-19 (compared to population)
- **C2**: COVID-19 (compared to population)

The data were obtained from the following sources:

https://storage.googleapis.com/covid19-hg-public/20201215/results/20210107/COVID19_HGI_A2_ALL_eur_leave_ukbb_23andme_20210107.b37.txt.gz

https://storage.googleapis.com/covid19-hg-public/20201215/results/20210107/COVID19_HGI_B2_ALL_eur_leave_ukbb_23andme_20210107.b37.txt.gz

https://storage.googleapis.com/covid19-hg-public/20201215/results/20210107/COVID19_HGI_C2_ALL_eur_leave_ukbb_23andme_20210107.b37.txt.gz

Data from the 7th HGI release (release date 8th April 2022) for the A2 outcome were additionally obtained from:

https://storage.googleapis.com/covid19-hg-public/freeze_7/results/20220403/main/sumstats/COVID19_HGI_A2_ALL_leave_23andme_20220403_GRCh37.tsv.gz

We removed variants with a p -value for heterogeneity of less than 0.001 before further analysis (see **Section 3.3.8 Heterogeneity testing**).

The study described in this chapter aimed to identify which of the severity-associated proteins were causally associated with COVID-19. For this reason, A2, which represents the variants associated with very severe COVID-19, was treated as the primary outcome in this study.

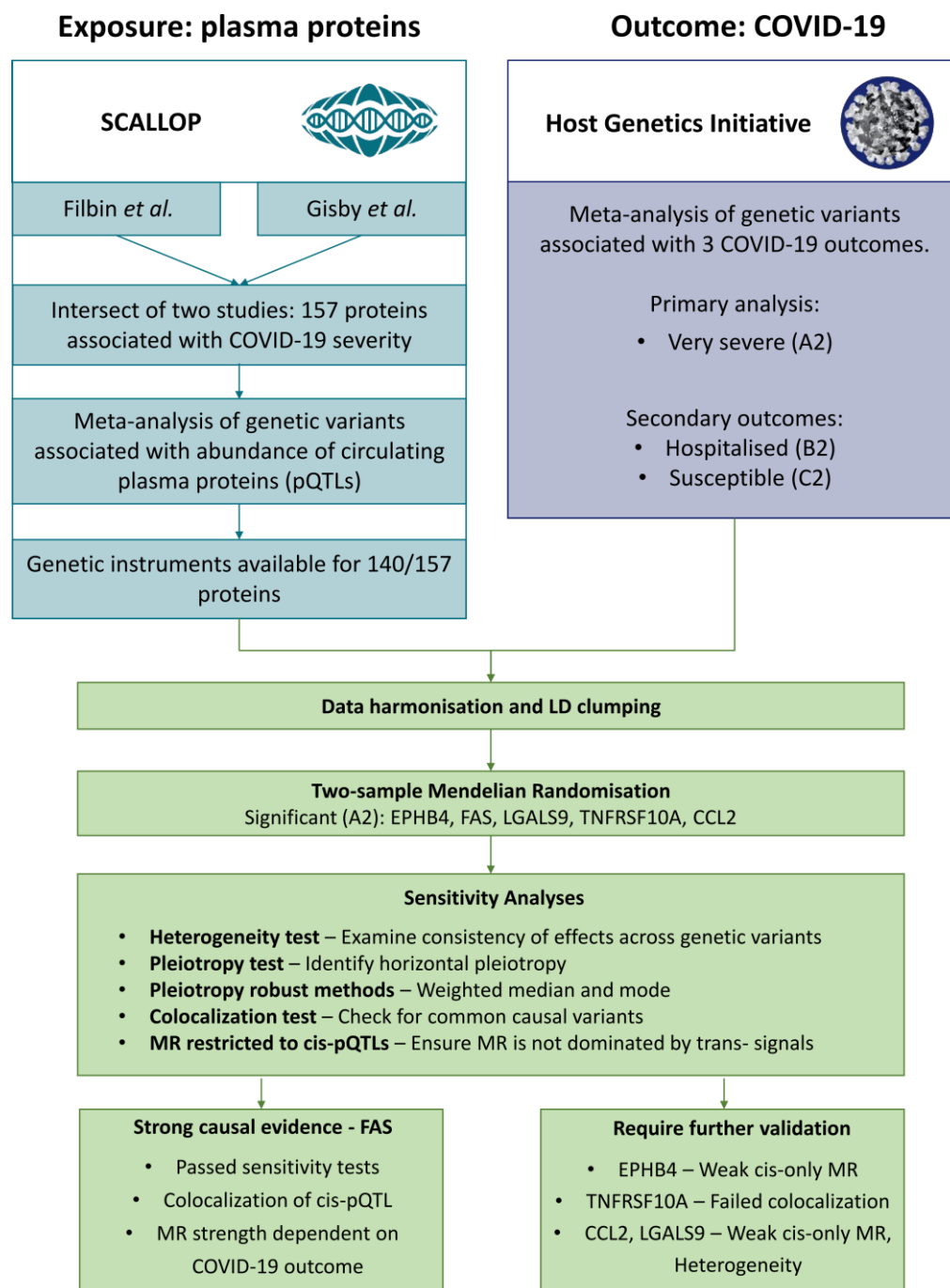


Figure 3.1 Mendelian randomisation study design and data sources

Severity-associated protein biomarkers were identified from the studies by Filbin *et al.* and Gisby *et al.*

3.3.2 Selecting proteins robustly associated with COVID-19 severity

Numerous studies have applied Olink immunoassays to identify plasma proteins that are associated with COVID-19 severity^{75,176}. However, these studies are observational, so they cannot determine which of these associations are causal in nature. We selected two studies to define a list of proteins robustly associated with COVID-19 severity. The first study was published by Filbin *et al.*⁷⁵ in 2021, consisting of hospitalised patients with COVID-19. The second was our study, Gisby *et al.*¹⁷⁶, as described in **Chapter 2**. We selected plasma proteins that were significantly associated (5% FDR) with COVID-19 severity in these two studies. This resulted in 157 proteins (**Figure 3.1**).

3.3.3 Genetic variants associated with circulating plasma proteins

Genetic instruments are required to perform Mendelian randomisation. Specifically, to assess the role of proteins as causal mediators, we needed to identify protein quantitative trait loci (pQTLs) for the 157 proteins identified by observational studies. To this end, we accessed data from the SCALLOP consortium²⁵⁵. These data were sourced from multiple cohorts, each of which performed proteomics using one or more of five Olink immunoassay panels (“Inflammation”, “Cardiovascular2”, “Cardiovascular3”, “Cardiometabolic” and “Immune Response”). The details of the cohorts and cohort-specific ethical approval are provided in Supplementary Table 2 of the preprint associated with this work²⁵³:

<https://www.medrxiv.org/content/10.1101/2021.04.01.21254789v1.supplementary-material>

Prior to the study presented in this chapter, each SCALLOP cohort performed pQTL mapping. This involved first applying a rank-based inverse-normal transformation to the values of each protein. Then, genome-wide association analysis was performed for each protein and variant using an additive regression model, with protein abundance as an outcome variable and genotype as an explanatory variable. The models were adjusted for age and sex in addition to cohort-specific covariates. In each cohort, population structure was accounted for either by including principal components in the model or by considering relatedness. Following this, quality control was applied to the summary statistics for each cohort using EasyQC²⁵⁶, using the protocol outlined by Winkler *et al.*²⁵⁶.

Finally, these cohorts were meta-analysed to provide a single set of robust instruments for MR. Specifically, the meta-analyses for the cardiometabolic and immune response panels

were carried out by Lucija Klaric and Marisa D. Muckian. Meta-analysis of the cardiovascular and inflammation panels was carried out by Erin Macdonald-Dunlop and Jing Hua Zhao, respectively. These analyses were performed using METAL, with the inverse-variance fixed effect method and “STDERR” option²⁵⁷, followed by correction for genomic control. Resulting summary statistics were retained for variants with a minor allele frequency (MAF) of greater than 0.01 and heterogeneity (I^2) below 75%.

3.3.4 Data harmonisation and clumping

For each set of pQTL summary statistics, we selected variants that achieved genome-wide significance (p -value $< 5 \times 10^{-8}$). We further filtered these data such that only variants present in the COVID-19 GWAS summary statistics remained. We then applied the R TwoSampleMR::harmonise_data function²⁵⁸ to harmonise the datasets such that the effects of the exposure and outcome were relative to the same allele.

In many cases, the remaining genetic variants were non-independent due to linkage disequilibrium (LD). Some software packages, such as those implemented in the MendelianRandomization package²⁵⁹, accept as input an LD correlation matrix, thus taking into account the correlation between variants. However, other MR methods, such as those implemented in the R TwoSampleMR package²⁵⁸, assume that genetic instruments are independent; using such methods with correlated variants could lead to “double-counting” of effects, potentially leading to their overestimation. To ensure the genetic instruments for each protein were largely independent, we performed pruning using Plink 1.9²⁶⁰, with the parameters `clump_r2 = 0.01` and `clump_kb = 10,000`. We carried this out using an LD reference panel based on 10,000 randomly selected and unrelated individuals of British ancestry (as determined using genomic data) from UK Biobank. Finally, we removed variants that were duplicated by either position or ID using Plink 2.0 (with the parameter “--rm-dup exclude-all”) and excluded variants in the major histocompatibility (MHC) region (hg19 genome position chr6:26,000,000-34,000,000).

3.3.5 Two-sample Mendelian Randomisation

We evaluated the causal effect of each protein on each outcome using two-sample Mendelian Randomisation (MR) with the TwoSampleMR package²⁵⁸. In cases where there were multiple uncorrelated variants available as genetic instruments, MR was carried out using inverse

variance weighted (IVW) fixed effects (FE) regression, implemented within the function `TwoSampleMR::mr_ivw_fe`. Where only a single variant was available as a genetic instrument, we instead applied a Wald Ratio (WR) test using the function `TwoSampleMR::mr_wald_ratio`. We applied the Benjamini-Hochberg method to adjust the resulting *p*-values and defined the significance threshold as 5% FDR.

3.3.6 MR restricted to *cis*-pQTLs

I refer to the primary MR as the *pan*-pQTL analysis, because both *trans*- and *cis*-pQTLs were used as genetic instruments. The benefit of this approach is that the number of available independent genetic instruments are maximised, increasing power and allowing for evaluation of the consistency of effects across a greater number of independent variants. However, *cis* instruments, that are situated in close proximity to the gene encoding the protein whose abundance is being modulated, may be less likely to be affected by horizontal pleiotropy²⁶¹. For instance, some *trans* loci are highly pleiotropic, being simultaneously associated with the abundance of many proteins²⁶². As a sensitivity analysis, we repeated MR analysis for each protein and restricted the genetic instruments to *cis*-pQTLs only. Following an additional filtering step to select variants within 0.5Mb of the gene that encodes the protein, we performed LD pruning and MR as described previously.

3.3.7 Pleiotropy robust methods and testing

The presence of horizontal pleiotropy can violate the exclusion-restriction (ER) MR assumption (see **Section 1.3.2.1**). In particular, the IVW (FE) MR method assumes that the estimate of the causal effect is consistent across the variants (i.e., there is no balanced horizontal pleiotropy) and that the intercept for the regression is zero, which may not be the case in the presence of unbalanced horizontal pleiotropy and could lead to false-positive results. For this reason, we performed sensitivity analyses that are more robust to the presence of horizontal pleiotropy, including the weighted median (using `TwoSampleMR::mr_weighted_median`) and weighted mode (using `TwoSampleMR::mr_weighted_mode`) methods. The Maximum likelihood method (using `TwoSampleMR::mr_two_sample_ml`) was also applied, which allows for uncertainty in the effect size of the genetic associations with the exposure and for genetic associations with the exposure and outcome for each variant to be correlated²⁶³. The MR Egger regression method was also applied (using `TwoSampleMR::mr_egger_regression`), which is similar to the IVW

method with a key difference being that the intercept is unconstrained, allowing for the detection and correction of bias due to unbalanced horizontal pleiotropy²⁶⁴.

3.3.8 Heterogeneity testing

If the causal estimates are not consistent across the IVs, this could indicate the presence of horizontal pleiotropy and a departure from the core MR assumptions. We calculated heterogeneity statistics for each MR analysis using the `TwoSampleMR::mr_heterogeneity` function²⁵⁸.

3.3.9 Colocalisation testing

The most significant variant at a given genomic locus is not necessarily mechanistically relevant; it may simply be correlated with the relevant variant. As such, even if a particular variant is both significantly associated with protein abundance and COVID-19 outcomes, there may be different underlying variants mediating each of the associations. This motivated the use of colocalisation analysis as a sensitivity analysis to differentiate between causal relationships and confounding by linkage disequilibrium. Colocalisation tests whether regional genetic association signals for different traits come from distinct or the same shared causal variant. The Bayesian colocalisation method implemented by the `coloc.abf` function in the `coloc` R package estimates posterior probabilities (PP) for five hypotheses: no association with either trait (H0), association with only one trait (H1 or H2), association with both traits but under the effect of distinct underlying causal variants (H3), or association with both traits under the effect of a shared causal variant (H4), also referred to as colocalisation.

For each protein with a significant causal estimate according to MR, we applied colocalisation to compare the HGI A2 COVID-19 GWAS and the relevant *cis*-pQTL (i.e., the variants within 0.5Mb of the gene encoding each protein). A PP of greater than 0.8 was considered to be strong evidence in favour of a hypothesis. It is also important to consider, however, that these results are contingent on the specification of three priors, p_1 , p_2 and p_{12} . The first, p_1 , is the prior probability that a randomly selected SNP in a given region is causally associated with the first trait (similarly, p_2 represents the prior for the second trait). The third, p_{12} , is the prior probability that a randomly selected SNP in the region is associated with both traits. In the analyses described in this thesis, default priors ($p_1 = 1 \times 10^{-4}$, $p_1 = p_2$, $p_{12} = 1 \times 10^{-5}$) were applied. To assess the effects of this prior on the colocalisation of the FAS pQTL and the A2 outcome,

we carried out a sensitivity analysis involving the estimation of the competing hypotheses as p_{12} varies (**Figure 3.3**).

Colocalisation testing assumes that there is only a single causal variant at a given locus for each trait. However, a single genomic region may contain multiple variants that affect a single trait. Therefore, we additionally applied the Sum of Single Effects method¹¹¹ using `coloc::coloc.susie`, an extension to the standard `coloc::coloc.abf` R function.

We also applied `coloc` to investigate whether expression QTLs (eQTLs) from various cell types and studies would colocalise with the FAS *cis*-pQTL. These included eQTLs from multiple tissue types from GTEx v7, whole blood data from eQTLGen, and five sorted leukocyte subsets from Peters *et al.*²⁶⁵.

3.3.10 Replicating the FAS pQTL using alternative assays

In the primary analysis, we employed genetic instruments obtained from a meta-analysis of data from SCALLOP cohorts that used Olink immunoassays. As a secondary analysis, we validated these findings using an additional set of meta-analyses derived from proteomic profiling by SomaScan assays based on SomaLogic's aptamer-based platform²⁶². We ran the original MR pipeline in an identical manner to the primary analysis, except for the difference in protein-variant associations used as input.

3.3.11 The association of FAS and COVID-19 severity

Data from the study described in **Chapter 2**, also available from the associated publication¹⁷⁶, were used to visualise the association of circulating FAS abundance with contemporaneous COVID-19 severity. We applied association analysis with linear mixed models, as described previously in sections **1.3.1.1** and **2.3.12**, to test the association of plasma FAS abundance with COVID-19 status (comparing cases and controls) and severity at the time of blood sampling (using a 4-level ordinal scale for clinical severity), respectively.

3.3.12 Calculating splice QTLs

We investigated whether the FAS pQTL would colocalise with a splice QTL (sQTL). To this end, we accessed RNA sequencing data from the INTERVAL cohort. These data were generated as follows. For 5,000 individuals of this cohort, 3 ml of whole blood were collected in Tempus Blood RNA Tubes (ThermoFisher Scientific), following the manufacturer's

instructions. RNA extraction was performed by QIAGEN Genomic Services using an in-house developed protocol based on QIAGEN's proprietary silica technology. The quality of the extracted RNA was assessed using spectrophotometric analysis. RNA Integrity Number (RIN) values were determined using a TapeStation 4200 system (Agilent), following the manufacturer's protocol. Messenger RNA (mRNA) was isolated using a NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB). Globin depletion was performed using a KAPA RiboErase Globin Kit (Roche). RNA library preparation was done using a NEBNext Ultra II DNA Library Prep Kit for Illumina (NEB) on a Bravo WS automation system (Agilent). Libraries were pooled to 96-plex in equimolar amounts, quantified using a High Sensitivity DNA Kit on a 2100 Bioanalyzer (Agilent), and then normalised to 2.8 nM prior to sequencing. Samples were sequenced using 75 bp paired-end sequencing reads (reverse stranded) on a NovaSeq 6000 system (S4 flow cell, Xp workflow; Illumina). The sequence data quality was assessed using FastQC v0.11.8³³. Reads were aligned to the GRCh38 human reference genome (Ensembl GTF annotation v99) using STAR v2.7.3.a³⁵.

Alex Tokolyi identified genetic variants associated with differential splicing of *FAS* exon 6 using the following methodology. Transcript splice junctions were extracted with the regtools "junction extract" tool, and introns were clustered and excision ratios calculated and normalised for downstream splice QTL (sQTL) analysis according to the leafcutter²⁶⁶ pipeline (build #aa12b1e) with default parameters. *Cis*-sQTLs were calculated on the resulting leafcutter ratios for the excision event of *FAS* exon 6 (ENSE00003500194) and flanking introns, and genotypes with a minor allele frequency >0.01 in the region ± 200 kb of the *FAS* gene using tensorQTL 1.0.5²⁶⁷. Blood cell type proportions, sex, and 10 genomic and 10 splicing principal components were added as covariates to the linear model.

Finally, we performed colocalisation analysis based on summary statistics for the *FAS* *cis*-sQTL and *cis*-pQTL using both coloc::coloc.susie and coloc::coloc.abf, as described in **Section 3.3.9**.

3.4 Results

3.4.1 Defining instruments and outcomes

The study described in **Chapter 2** revealed proteins associated with contemporaneous COVID-19 severity. However, these data were obtained from patients with end-stage kidney disease (ESKD), and so the generalisability of these results to patients with normal kidney function was unclear. For this reason, we used results from a second study, carried out by Filbin *et al.* in the same year⁷⁵, to define a set of proteins robustly associated with COVID-19 severity. Both studies applied Olink proteomic profiling to individuals infected by SARS-CoV-2. Specifically, a set of 157 proteins was defined based on significance (5% FDR) in both studies (**Figure 3.1**). In this chapter, I describe the use of Mendelian randomisation (MR) and colocalisation to assess whether these proteins have causal effects on COVID-19 outcomes.

The data provided by these studies are insufficient for calculating pQTLs, as they did not perform genotyping. Furthermore, detecting associations with genetic variants generally requires larger sample sizes, in part due to the high multiple testing burden incurred from testing millions of SNPs. Collectively, the members of the SCALLOP²⁵⁵ consortium have carried out both Olink proteomic profiling and genotyping in individuals of European heritage. These studies applied 92-plex panels of Olink immunoassays. Since different panels were run by different SCALLOP contributors, the combined sample size for each set of proteins ranged from 3,658 and 26,494. We accessed these data for use as genetic instruments in MR (see **Section 3.3.2**).

We filtered the pQTL summary statistics based on genome-wide significance and removed variants in the MHC region (see **Section 3.3.4**), after which there were pQTLs were available for 140/157 proteins. For 30 proteins, only *cis*-pQTLs were available (i.e., local-acting variants situated within 0.5 Mb of the gene encoding the protein). 17 proteins had only *trans*-pQTLs (i.e., distal regulators of protein abundance). The remaining 93 proteins had both significant *cis*- and *trans*-pQTLs. We performed further filtering using pruning based on linkage disequilibrium, to remove variants with a correlation (r^2) of greater than 1% (**Figure 3.1**) (see **Section 3.3.4**).

For the outcome data, we obtained GWAS summary statistics from release 5 of the COVID-19 Host Genetics Initiative meta-analysis²⁶⁸. Summary statistics were available for multiple

outcomes, which were each compared to the general population. The present investigation primarily focussed on the very severe respiratory COVID-19 outcome (which I refer to as A2 or severe COVID-19), which is defined as hospitalised patients who required respiratory support and/or died. In some secondary analyses, we additionally used statistics based on the B2 (hospitalised COVID-19) and C2 (susceptible to COVID-19) outcomes.

3.4.2 Mendelian randomisation

We performed Mendelian randomisation testing for 140 proteins using the inverse-variance weighted (VW) fixed effect (FE) method. Five proteins had significant effects on critical COVID-19 (A2) after multiple testing correction (5% FDR) (**Figure 3.1, Table 3.1**). These included EPH receptor B4 (EPHB4), C-C motif chemokine ligand 2 (CCL2), galectin 9 (LGALS9), Tumour Necrosis Factor Receptor Superfamily Member 10A (TNFRSF10A) and the Fas cell surface death receptor (FAS).

Mendelian randomisation makes three key assumptions regarding the instrumental variables (IVs)¹⁰⁸ (see **Section 1.3.2.1**). When interpreting the findings of MR, it is therefore important to evaluate whether there is reasonable evidence that these assumptions are met. The relevance assumption, that the genetic variants associate with the risk factor of interest, is easily demonstrated empirically because the pQTLs used as instruments in the current study achieved genome-wide significance. The independence assumption, that there are no unmeasured confounders of the associations between genetic variants and outcome, is more difficult to assess. If ancestry, for instance, were associated with both a genetic variant and susceptibility to severe COVID-19, then this assumption would be violated. The pQTL and outcome data were restricted to individuals of European ancestry and population structure was accounted for in the associations of genetic variants with protein abundance and the outcome. This reduces the likelihood that there are unmeasured confounders between genetic variants and COVID-19 severity. Finally, the exclusion restriction (ER) assumption asserts that genetic variants must affect the outcome only through their effect on the risk factor of interest (i.e., there is no horizontal pleiotropy). In the present study, we applied statistical methods that can assess whether MR results are robust to horizontal pleiotropy as sensitivity analyses, including the weighted mode, weighted median, maximum likelihood, and Egger methods (**Table 3.1, Table 3.2, Section 3.3.7**). We also tested for heterogeneity to ascertain whether the effect estimates of multiple genetic variants were consistent (**Table 3.1, Section 3.3.8**).

While the primary MR included both *trans*- and *cis*-variants, we repeated the analysis using only the *cis*-acting pQTLs, which have a direct link to the gene encoding the protein (**Table 3.1**).

For TNFRSF10A, Cochran's Q value (an estimate of heterogeneity) was not significant (p -value < 0.05) (**Table 3.1**), indicating that there was no substantial heterogeneity between the instrumental variables for the effect of its expression on COVID-19 outcomes. Except for Egger, the MR methods suggested a protective effect of TNFRSF10A on COVID-19 severity (**Table 3.2**). On the other hand, the MR estimate for Egger regression was close to 0. The Egger intercept was significant (p -value < 0.05) (**Table 3.1**), indicating that unbalanced horizontal pleiotropy could be driving the apparent protective effect of TNFRSF10A. The majority of TNFRSF10A pQTLs were *cis*-acting, so re-analysis without *trans*-pQTLs did not lead to meaningful changes in regression results (**Table 3.1**).

In contrast, for CCL2 there was significant heterogeneity (Cochran's Q > 50 and p -value < 0.05) (**Table 3.1**). All MR methods returned positive estimates for the effect of CCL2 on the A2 outcome, although the weighted mode (p -value = 0.24), IVW random effects (RE) (p -value = 0.24) and Egger (p -value = 0.68) methods were non-significant (**Table 3.2**). Additionally, an MR analysis of *cis*-pQTLs for CCL2 was non-significant (p -value = 0.50) (**Table 3.1**). The effect of CCL2 on susceptibility to severe COVID-19 was largely based on *trans*-pQTLs; specifically, the results were largely driven by a highly significant pQTL near the 3p21.31 locus (in proximity to the CCR2 gene, which encodes a receptor for CCL2). These results cast doubt on the reliability of CCL2 as a causal mediator of COVID-19.

There was also significant heterogeneity for LGALS9 (**Table 3.1**). Furthermore, while the IVW and maximum likelihood methods indicated reduced LGALS9 was associated with severe COVID-19, the weighted median, weighted mode and Egger methods did not find a significant MR estimate (**Table 3.2**). MR carried out for only LGALS9 *cis*-pQTLs was also non-significant (**Table 3.1**). Collectively, these sensitivity analyses indicate that the initial MR for this protein may be unreliable.

There was no heterogeneity revealed for EPHB4 (**Table 3.1**). Furthermore, all MR methods indicated that reduced EPHB4 abundance is causally associated with severe COVID-19 (**Table 3.2**), although the Egger estimate was non-significant (p -value = 0.24). Despite this,

the Egger intercept for EPHB4 was also non-significant (p -value = 0.90) and the Egger estimate was similar to that of other methods (**Table 3.1**, **Table 3.2**). So, Egger regression does not appear to contradict the original MR findings. Only 1/5 of the instrumental variables used for the Egger MR were within 0.5 Mb of the gene encoding EPHB4. Unsurprisingly, the *cis*-pQTL MR for EPHB4 did not reach significance (p -value = 0.13); however, the MR estimate for this analysis (OR = 0.64; 95% confidence interval = 0.36-1.15) was similar to that of the original IVW FE method (OR = 0.56; 95% confidence interval = 0.44-0.71).

Significant heterogeneity was not found for FAS (p -value = 0.15), although Cochran's Q was relatively high ($Q = 29.1$). Every MR method found a significant effect of FAS abundance on COVID-19 severity (**Table 3.2**). Additionally, repeating MR for the FAS *cis*-pQTLs found similar results to the original *pan*-MR (**Table 3.1**).

Table 3.1 Summary of Mendelian randomisation results

Protein	EPHB4	CCL2	LGALS9	TNFRSF10A	FAS
# IV	5	7	26	28	7
<i>pan</i> -MR <i>p</i> -value	1.28×10 ⁻⁶	2.43×10 ⁻⁶	6.38×10 ⁻⁴	1.71×10 ⁻³	1.36×10 ⁻³
Adjusted <i>pan</i> -MR <i>p</i> -value	1.69×10 ⁻⁴	1.69×10 ⁻⁴	2.96×10 ⁻²	4.76×10 ⁻²	4.74×10 ⁻²
<i>pan</i> -MR OR (95% CI)	0.56 (0.44-0.71)	1.54 (1.29-1.84)	0.79 (0.69-0.91)	0.81 (0.71-0.92)	1.40 (1.14-1.72)
<i>cis</i> -MR <i>p</i> -value	0.13	0.50	0.16	1.91×10 ⁻³	6.70×10 ⁻⁴
<i>cis</i> -MR OR (95% CI)	0.64 (0.36-1.15)	0.71 (0.26-1.92)	0.90 (0.78-1.04)	0.81 (0.71-0.92)	1.44 (1.17-1.78)
Cochran's Q (<i>p</i> -value)	4.3 (0.36)	97.0 (1.1×10 ⁻¹⁸)	54.7 (5.4×10 ⁻⁴)	9.4 (0.36)	29.1 (0.15)
Egger intercept <i>p</i> -value	0.90	0.98	0.05	0.03	0.05

A summary of the main *pan*-MR results and the *cis*-MR, heterogeneity and Egger sensitivity tests.

IV – number of instruments used; MR *p*-value – Inverse-variance fixed effect (IVW-FE) *p*-value; Adjusted MR *p*-value – Benjamini-Hochberg adjusted MR *p*-value; OR – odds ratio; *pan* – analysis using all variants; *cis* – analysis using only *cis* variants; Cochran's Q – inverse variance weighted heterogeneity test.

Proteins are annotated using the symbols of the encoding gene. EPHB4 – Ephrin type-B receptor 4; CCL2 – C-C-motif chemokine 2; LGALS9 – Galectin-9; TNFRSF10A – Tumour necrosis factor receptor superfamily member 10A; FAS – FAS cell death receptor (also known as Tumour necrosis factor receptor superfamily member 6).

Table 3.2 Full *pan*-Mendelian randomisation results

Protein	MR Method	Beta	SE	<i>p</i>-value	Significance
FAS	MR Egger	0.76	0.20	1.23×10^{-02}	Significant
	IVW (random effects)	0.34	0.13	1.06×10^{-02}	Significant
	IVW (fixed effects)	0.34	0.10	1.36×10^{-03}	Significant
	Weighted median	0.42	0.12	3.44×10^{-04}	Significant
	Weighted mode	0.43	0.13	1.56×10^{-02}	Significant
	Maximum likelihood	0.34	0.11	1.38×10^{-03}	Significant
CCL2	MR Egger	0.46	1.03	6.77×10^{-01}	Non-Significant
	IVW (random effects)	0.43	0.37	2.41×10^{-01}	Non-Significant
	IVW (fixed effects)	0.43	0.09	2.43×10^{-06}	Significant
	Weighted median	0.26	0.13	4.65×10^{-02}	Significant
	Weighted mode	0.16	0.12	2.38×10^{-01}	Non-Significant
	Maximum likelihood	0.50	0.10	5.00×10^{-07}	Significant
EPHB4	MR Egger	-0.54	0.37	2.40×10^{-01}	Non-Significant
	IVW (random effects)	-0.59	0.13	3.06×10^{-06}	Significant
	IVW (fixed effects)	-0.59	0.12	1.28×10^{-06}	Significant
	Weighted median	-0.58	0.14	2.92×10^{-05}	Significant
	Weighted mode	-0.62	0.15	1.51×10^{-02}	Significant
	Maximum likelihood	-0.59	0.12	1.79×10^{-06}	Significant
LGALS9	MR Egger	0.01	0.15	9.52×10^{-01}	Non-Significant
	IVW (random effects)	-0.23	0.10	2.09×10^{-02}	Significant
	IVW (fixed effects)	-0.23	0.07	6.38×10^{-04}	Significant
	Weighted median	0.00	0.10	9.94×10^{-01}	Non-Significant
	Weighted mode	-0.03	0.10	7.98×10^{-01}	Non-Significant
	Maximum likelihood	-0.23	0.07	8.65×10^{-04}	Significant
TNFRSF10A	MR Egger	0.02	0.12	8.44×10^{-01}	Non-Significant
	IVW (random effects)	-0.21	0.07	2.53×10^{-03}	Significant
	IVW (fixed effects)	-0.21	0.07	1.71×10^{-03}	Significant
	Weighted median	-0.15	0.09	1.09×10^{-01}	Non-Significant
	Weighted mode	-0.15	0.09	1.07×10^{-01}	Non-Significant
	Maximum likelihood	-0.21	0.07	2.00×10^{-03}	Significant

Mendelian randomisation for all methods performed for all variants (i.e., the *pan*-MR). The five proteins were selected based on their adjusted *p*-value for the IVW (fixed effects) method (**Table 3.1**). The *p*-value and significance columns are based on unadjusted *p*-values. Positive betas indicate higher protein abundance is associated with a higher risk of severe COVID-19, whereas negative betas indicate that higher protein abundance is associated with a lower risk of severe COVID-19.

3.4.3 Colocalisation

We applied colocalisation analysis¹¹⁰ to assess whether there was confounding by linkage disequilibrium or whether the pQTL and COVID-19 associations were driven by the same underlying causal variant. We applied this approach to the variants in the *cis* region of the gene encoding each protein (i.e., those variants within 0.5Mb). The only *cis*-pQTL that colocalised with the regional association signal for COVID-19 severity (A2) was for *FAS*, with a posterior probability (PP) for colocalisation (H_4) of 0.95 (**Figure 3.2**). It is important to consider, however, that this result is contingent on the specification of prior probabilities, that encode the probability of a randomly selected variant in a given region being causally associated with one or both traits' outcome (see **Section 3.3.9**). The authors of the coloc package have pointed out that the default values for these priors may not be suitable in all instances and that their specifications can substantially alter inferences¹¹³. Specifically, misspecification of the p_{12} prior can affect the PP of H_4 , and its default (1×10^{-5}) may be too liberal. A sensitivity analysis performed to investigate the effect of changing p_{12} on PPs of the five hypotheses demonstrated that colocalisation (H_4) was most probable (i.e., $PP > 0.5$) when the p_{12} prior was as low as 1×10^{-6} (**Figure 3.3**). Simulation studies indicate that specifying p_{12} at 5×10^{-6} is generally robust¹¹³, suggesting that the original conclusion of colocalisation was accurate.

These colocalisation analyses assume that there is at most a single causal variant underlying each trait at a given locus. While such an assumption is convenient, it does not necessarily reflect reality¹¹³. For this reason, we performed a second colocalisation method using the Sum of Single Effects method¹¹¹, which explores the possibility of multiple causal variants. This method also confirmed colocalisation between the *FAS* pQTL and the A2 outcome.

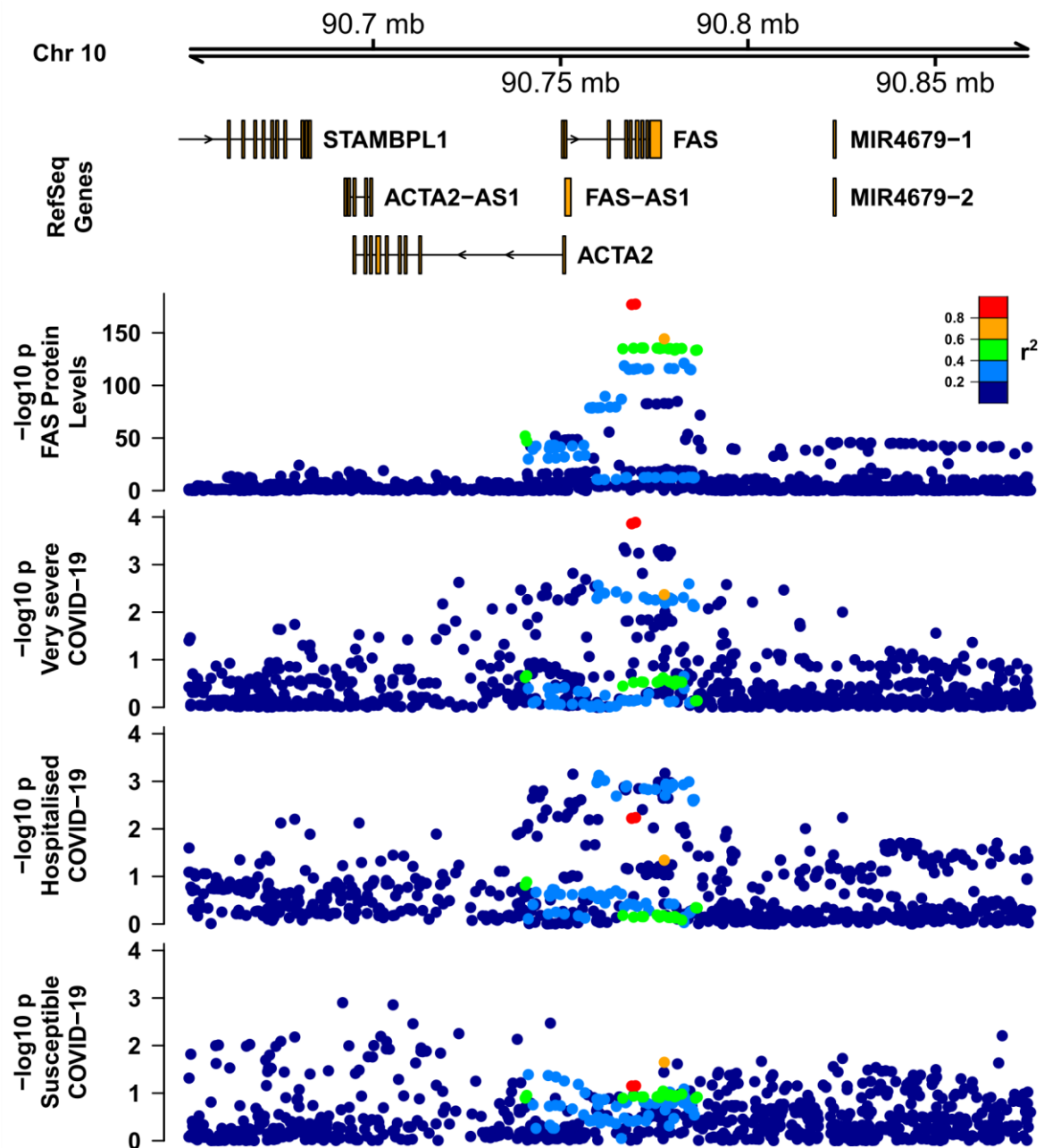


Figure 3.2 Regional associations for the FAS *cis*-pQTL and COVID-19 outcomes

Regional association plot (hg19 genome build) showing the *cis*-pQTL for FAS (plasma) and the association of variants with COVID-19 outcomes. The posterior probability of a shared causal variant (PP H_4) between FAS protein levels and very severe COVID-19 = 0.95. Very severe COVID-19 = A2; hospitalised COVID-19 = B2; susceptible COVID-19 = C2.

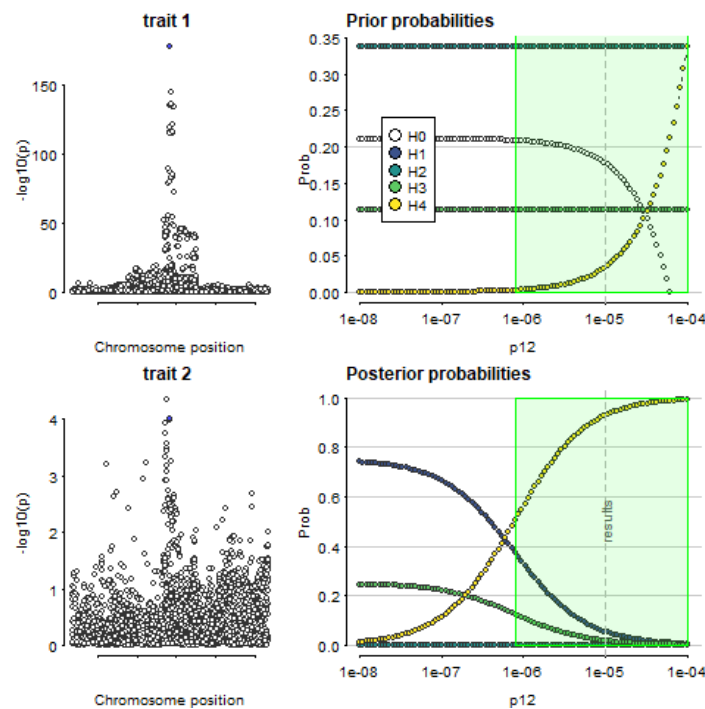


Figure 3.3 The effect of changing priors on FAS-COVID-19 colocalisation

The prior and posterior probabilities are plotted as a function of p_{12} for five competing hypotheses for the colocalisation of the FAS pQTL (trait 1; above) and severe COVID-19 (trait 2; below). The grey dotted line indicates the default value for the p_{12} prior. The green box indicates the values of p_{12} for which colocalisation (H_4) is the most probable hypothesis.

3.4.4 Investigating alternative outcomes

The primary COVID-19 outcome simultaneously represents both susceptibility to COVID-19 and risk of very severe COVID-19, as patients were compared to the general population (controls). Other outcomes included hospitalised COVID-19 compared to controls (B2) and all COVID-19 compared to controls (C2). We repeated the FAS MR for these outcomes. FAS did not significantly affect susceptibility to COVID-19 (C2) or hospitalisation with COVID-19 (B2) (**Figure 3.4a**). There appeared to be a dose-response relationship between the severity of the COVID-19 outcomes and the estimated effect of FAS. These findings suggested that levels of circulating FAS affect the risk of developing severe disease, but not susceptibility to COVID-19. This hypothesis is consistent with observational data from Chapter 2¹⁷⁶. In these data, FAS was positively associated with COVID-19 severity (adjusted p -value = 0.019), whereas there was no significant difference between its abundance in non-infected versus infected patients (adjusted p -value = 0.28) (**Figure 3.4b**).

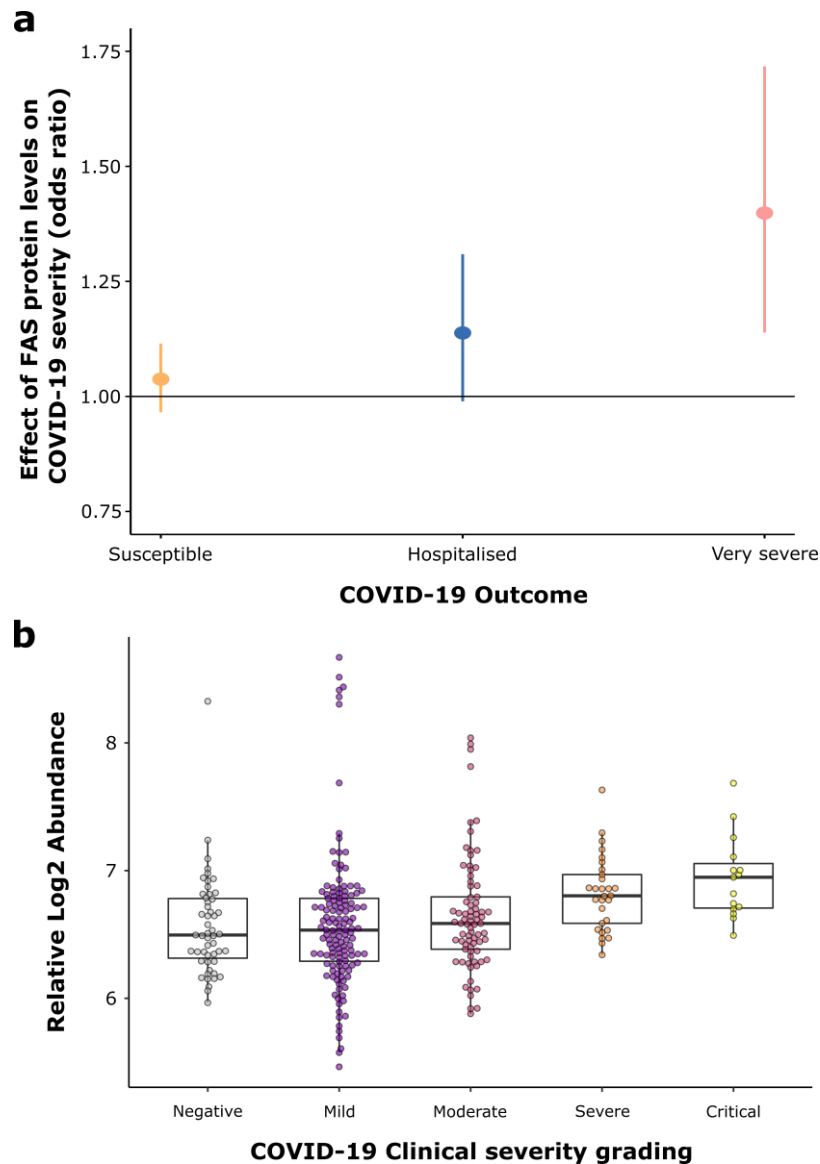


Figure 3.4 Plasma FAS is associated with COVID-19 severity

a) MR estimates of the causal effect of circulating FAS protein on different COVID-19 outcomes: susceptibility to infection (A2), hospitalisation (B2) and very severe disease (C2).

b) Soluble FAS protein levels in COVID-19 patients, stratified by clinical severity, and non-infected controls (data from Gisby *et al.*¹⁷⁶). Boxplots showing the distribution of plasma protein levels according to COVID-19 status at the time of blood draw. Boxplots indicate median and interquartile range. $n = 256$ samples from 55 COVID-19 patients and 51 samples from non-infected patients. 'COVID-19 status' indicates the clinical severity score of the patient at the time the sample was taken. Mild $n = 135$ samples; moderate $n = 77$ samples; severe $n = 29$ samples; critical $n = 15$ samples.

3.4.5 Assessing the FAS pQTL

The relevance assumption of MR requires the instrumental variables to be robustly associated with the exposure. For this reason, it is vital that the Olink immunoassays truly represent relative FAS abundance levels. In some cases, immunoassays lack specificity as a result of binding to a protein other than the intended target. Relative protein abundance can alternatively be measured using aptamer-based SomaScan assays. We repeated MR using a SomaScan-calculated pQTL⁶, which revealed similar results to the original findings (MR OR = 1.35, 95% CI = 1.14-1.60, p -value = 4.6×10^{-4}) (**Figure 3.5**). In some instances, the Olink and SomaScan assays are poorly correlated, or even anticorrelated, for the same protein target²⁶⁹. This result provides additional evidence that the pQTL is truly associated with FAS abundance, as opposed to an association driven by off-target binding.

While the statistical analyses thus far pointed to FAS as a mediator of COVID-19, we considered that the variant underlying the FAS pQTL could mediate COVID-19 through an alternative mechanism (e.g., the variant could also modulate the abundance of a second protein, which is truly causal for COVID-19). The PhenoScanner database, which contains over 65 billion associations for 150 million genetic variants, was used to search for other associations at this locus. Specifically, we searched the database for variants that were used as instruments in the FAS *cis*-MR and those that were in LD ($r^2 > 0.6$) with them. For the most significant FAS pQTL, rs7911226, there were no significant (p -value $< 1 \times 10^{-5}$) associations. For the secondary variant, rs28362318, there were associations with “Cause of death: subarachnoid haemorrhage from intracranial artery, unspecified” (p -value = 3.41×10^{-7}) and “Treatment with nebivolol” (p -value = 9.51×10^{-6}). Subarachnoid haemorrhage is bleeding into the subarachnoid space in the brain. High blood pressure increases the risk of spontaneous subarachnoid haemorrhage and antihypertensives are employed in its treatment²⁷⁰. Nebivolol is a beta blocker used to treat high blood pressure and heart failure. Overall, the FAS pQTL was mostly independent of other traits according to PhenoScanner.

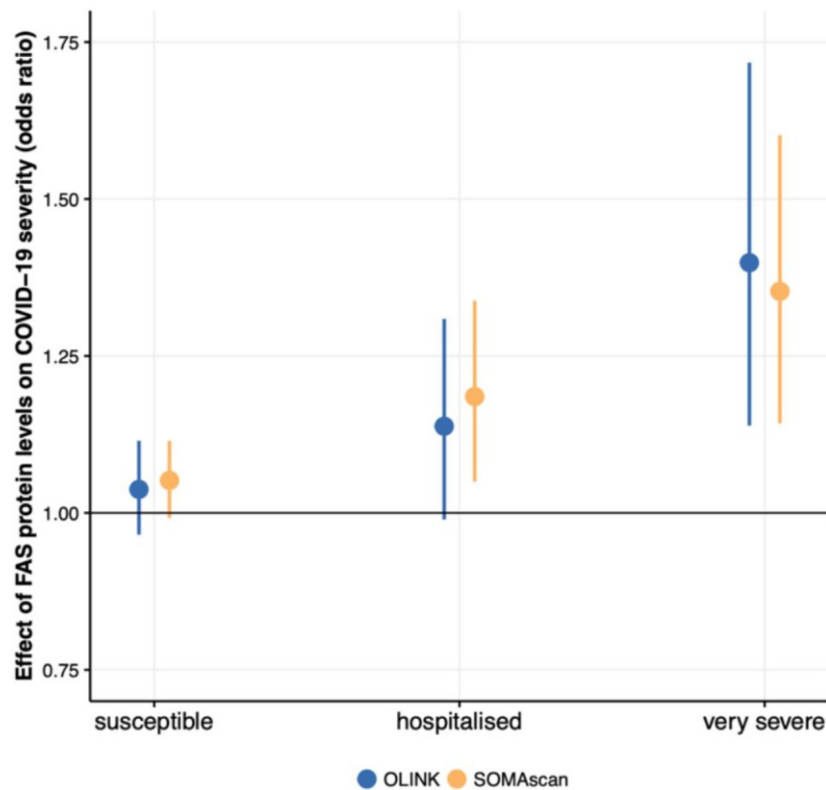


Figure 3.5 Validation of MR estimates of the causal effect of soluble FAS protein on COVID-19 outcomes

Genetic instruments were obtained from another pQTL study of FAS levels measured with a different proteomic platform (SOMAscan⁶). 'Olink' indicates the MR estimate obtained in the primary analysis.

3.4.6 Investigating the mechanism of the FAS pQTL

The sentinel *cis*-pQTL variant was rs982764. This single nucleotide polymorphism (SNP) is located between exons 4 and 5 of the *FAS* gene. The T allele of this variant is associated with greater levels of plasma FAS abundance. Since the variant is situated in an intron and is not in LD ($r^2 > 0.2$) with any non-synonymous coding variants, it was expected that this variant would also be associated with *FAS* gene expression (i.e., an eQTL). We examined eQTL data from multiple tissues in the GTEx Catalogue²⁷¹, whole blood data from eQTLGen²⁷² and sorted immune cell subsets²⁶⁵. There was evidence for a *FAS* eQTL in these three datasets, including evidence in the lung, blood, adipose tissue and artery tissues from GTEx and monocytes from Peters *et al.*²⁶⁵. However, the most significant variant for the eQTL was not rs982764 (**Figure 3.6**). We performed colocalisation analysis, which revealed that the eQTLGen blood eQTL and GTEx lung eQTL were mediated by different underlying causal variants to the FAS pQTL (PP of distinct causal variants [H_3] of 1.00 for both datasets). However, both the pQTL and eQTL displayed highly significant and broad peaks in Manhattan plots (**Figure 3.6**), suggesting

that they could be mediated by multiple underlying causal variants. Standard colocalisation analysis could fail to appropriately compare these loci, whereas the Sum of Single Effects¹¹¹ (SuSiE) can account for conditionally independent signals. However, re-analysis with SuSiE confirmed that the eQTL and pQTL signals did not colocalise (PP of distinct causal variants $[H_3] = 1.00$ for every set of distinct variants). Therefore, the effect of this variant on FAS plasma protein abundance could be explained by neither changes in protein sequence nor modulation of gene expression.

The *FAS* gene is made up of nine exons, which combine to generate several protein-coding and non-coding isoforms²⁷³ (**Figure 3.7**). Isoform 1 has an extracellular domain, encoded by exons 1 to 5, a transmembrane domain, encoded by exon 6, and a cytoplasmic domain, encoded by exons 7 and 9. This isoform is a cell surface receptor that binds to the cytokine FAS ligand (FASL). An intracellular signaling cascade is triggered upon binding of FASL to FAS, leading to apoptosis. Another FAS protein isoform is formed when alternative splicing leads to the skipping of exon 6²⁷⁴ (transcript isoform *FAS* Δ Ex6; **Figure 3.7**). This form of FAS, which lacks the transmembrane domain, is soluble and cannot activate apoptotic pathways. Instead, soluble FAS (sFAS) acts as a decoy receptor for FASL and opposes the action of membrane-bound FAS (mFAS) by reducing FAS-FASL signaling²⁷⁵.

It is possible that mFAS could be released into plasma due to cell damage and death. This is particularly plausible in severe COVID-19, due to increased cell turnover and viral hijacking of host cellular machinery²⁷⁶. However, given that the pQTL data were measured in otherwise healthy individuals, it is more likely that the Olink and SomaScan assays reflect the relative abundance of sFAS. This prompted an investigation into whether the pQTL for circulating FAS abundance reflected differential splicing of exon 6. We obtained splice QTL (sQTL) data from the INTERVAL cohort. Specifically, summary statistics were obtained for the associations between variants and a shift towards transcript isoforms lacking exon 6 in whole-blood RNA sequencing data (see **Section 3.3.12**). These data indicated that the most significant FAS *cis*-pQTL was associated with the skipping of exon 6 (p -value = 2.1×10^{-22}) (**Figure 3.8a**). Unlike the eQTL, the sQTL displayed a similar association pattern to that of the pQTL (**Figure 3.7**). We performed colocalisation analysis, which indicated that the sQTL and pQTL were likely mediated by the same underlying causal variants (PP for shared causal variants $[H_4]$ of 0.998). We also confirmed this result by reanalysis with SuSiE. The data presented here uncover a

putative mechanism for the mediation of FAS protein abundance in plasma through alternative splicing (**Figure 3.8b**).

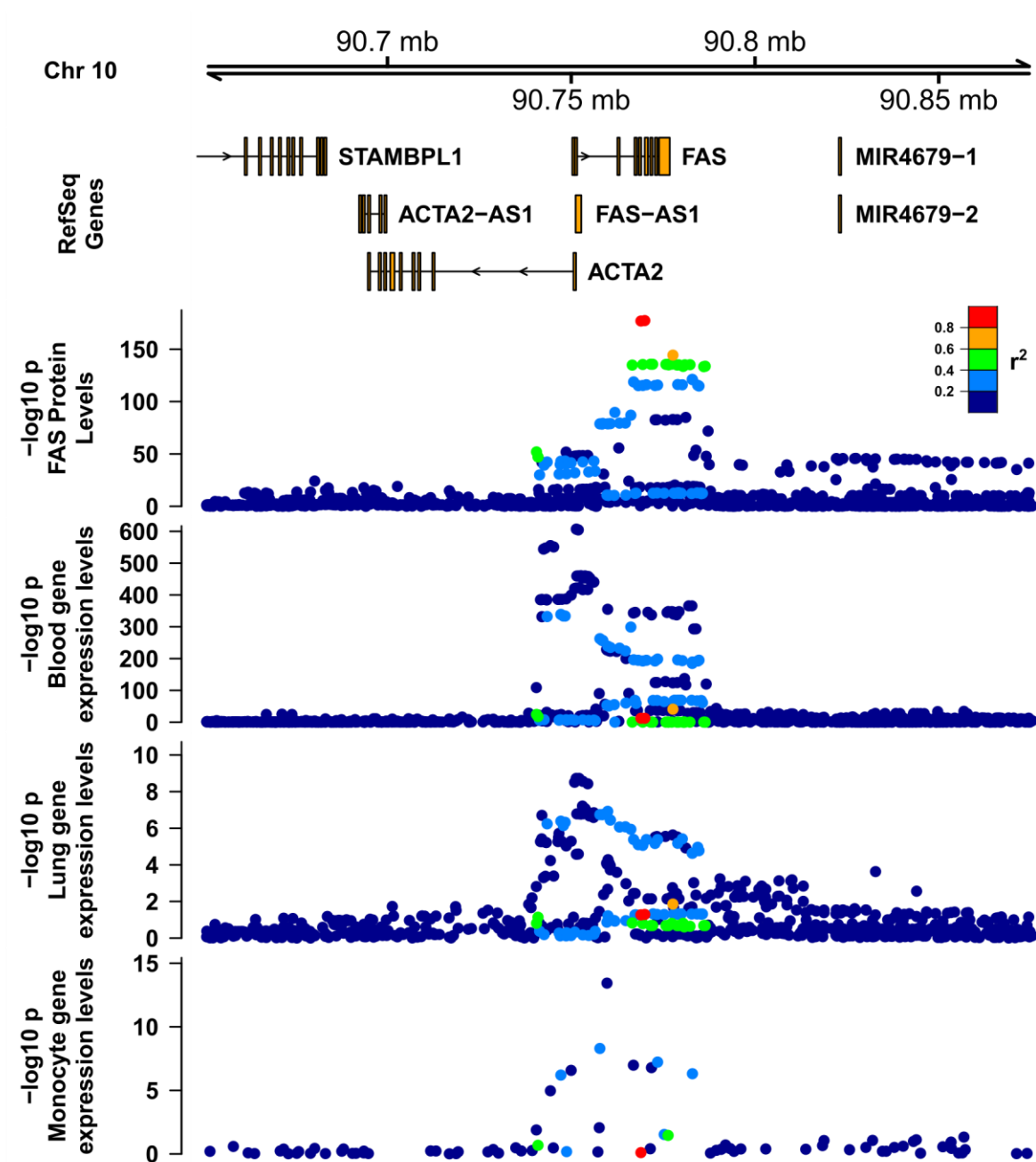


Figure 3.6 Regional associations for FAS pQTL and eQTLs

Tracks from top to bottom: FAS plasma pQTL (SCALLOP meta-analysis), FAS eQTL in whole blood (eQTLGen), lung (GTEx v7) and blood monocytes (Peters *et al.*²⁶⁵). Colour indicates linkage disequilibrium (r^2) with the sentinel pQTL variant.

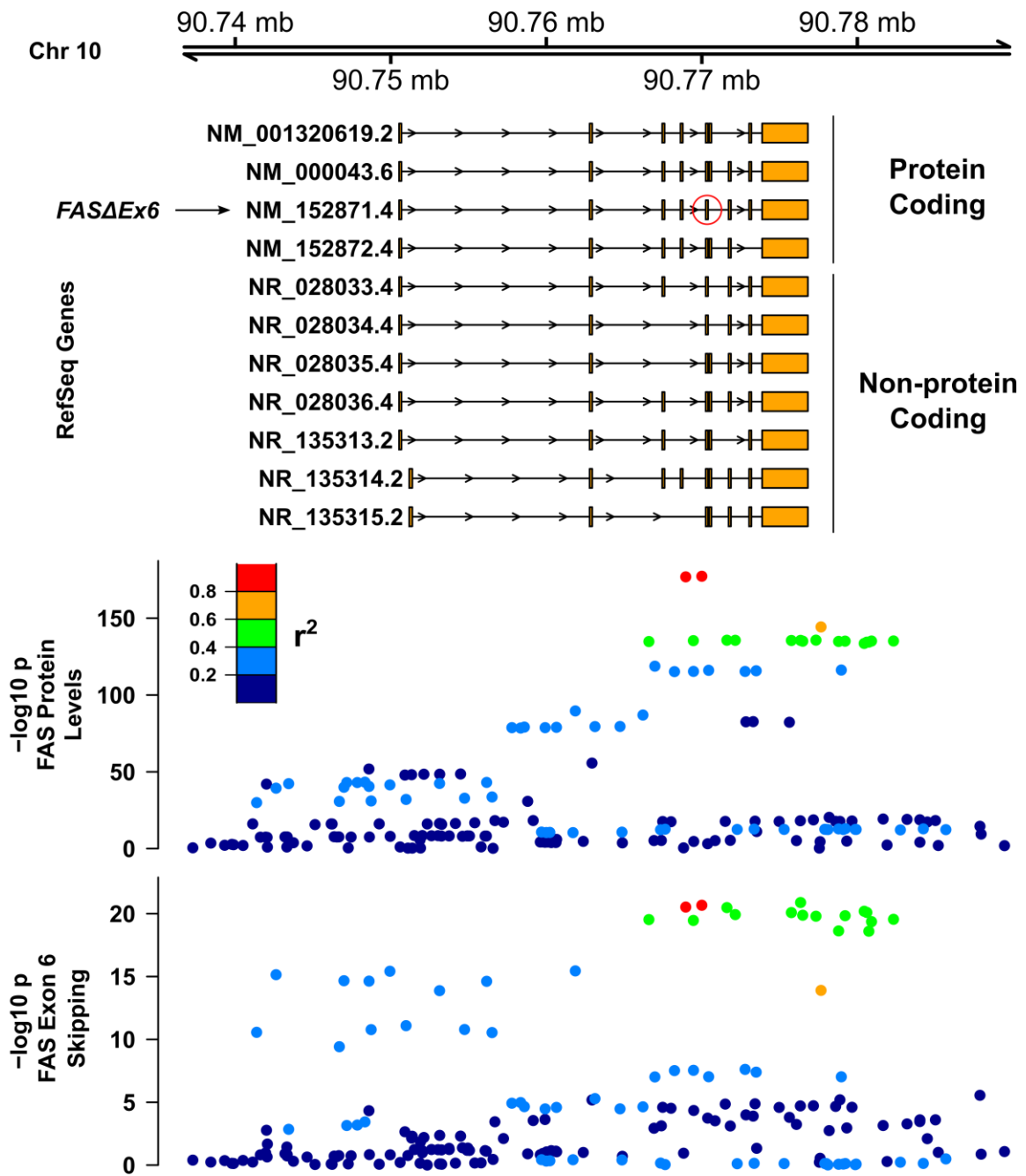


Figure 3.7 Matching regional association patterns for the FAS pQTL and sQTL

Regional association plot showing (from top to bottom): transcript isoforms, the soluble FAS cis-pQTL, and the associations with FAS exon 6 splicing. FAS Δ Ex6 indicates the transcript isoform lacking exon 6 (red circle). The posterior probability of a shared causal variant (PP H4) between FAS protein levels and exon 6 splicing = 0.99. Colour indicates linkage disequilibrium (r^2) with the sentinel pQTL variant.

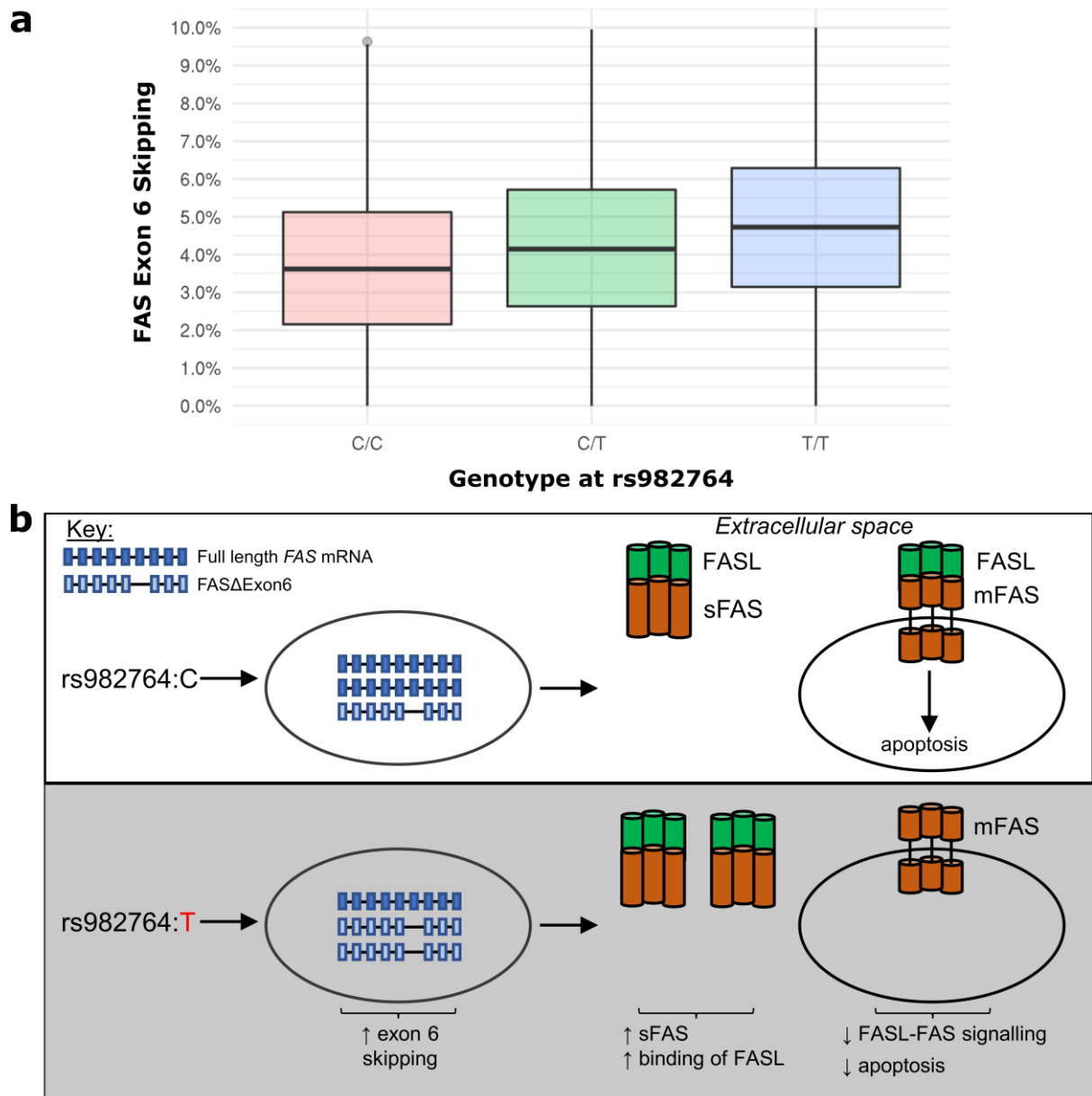


Figure 3.8 A putative mechanism for the effect of FAS splicing

a) Boxplot showing exon 6 splice quantitative trait locus (sQTL). Number of individuals by genotype at rs982764: 443 (CC), 1992 (CT), 2329 (TT). In 14 individuals genotype could not be reliably ascertained. Y-axis represents % of transcripts with exon 6 skipping. Splicing was associated with this genotype according to linear models (p -value = 4×10^{-22}).

b) Proposed model by which genetic variation in *FAS* could contribute to increase soluble FAS. A non-coding variant acts as a splice QTL. The sentinel *FAS* pQTL (rs982764:T) is associated with an increased proportion of transcripts lacking exon 6 resulting in higher levels of soluble FAS (sFAS). sFAS acts as a decoy receptor for FASL (FASL), blocking FASL binding to membrane-bound FAS (mFAS) on the cell surface and thus reducing apoptosis.

3.4.7 Assessing the COVID-19 locus

While the MR for FAS was significant at 5% FDR (adjusted p -value = 0.047), none of the variants considered in this analysis achieved genome-wide significance (i.e., 5×10^{-8}) with respect to the very severe COVID-19 (A2) outcome (**Figure 3.2**). Note that this is not strictly a requirement for MR analysis, and colocalisation was found between the FAS pQTL and the A2 outcome (PP = 0.95). However, it does call into question the reliability of the associations between these variants and COVID-19 severity. It was therefore of interest to assess whether these associations replicated in other cohorts. Since the preprint associated with this chapter was made available²⁵³, the 7th round of the HGI meta-analysis was released. Whereas release 5 of the A2 outcome used so far in this chapter included 5,582 cases, release 7 included 18,152 patients with very severe COVID-19. The COVID-19-associated locus in the *FAS* gene region observed in release 5 (**Figure 3.2**) is not visible in release 7 (**Figure 3.9**). Colocalisation analysis was performed for the FAS pQTL and the release 7 A2 outcome. The PP of H_4 was only 0.07, indicating that colocalisation was unlikely. The most probable outcome was H_1 (PP = 0.73), indicating that there is a variant in the region associated with levels of circulating FAS, but none associated with severe COVID-19. Re-analysis with SuSiE also indicated that colocalisation was unlikely. Therefore, the COVID-19-associated locus in the *FAS* gene region did not replicate in this new release. This indicates that the original MR finding, that genetic predisposition to higher circulating sFAS increases the risk of very severe COVID-19, is not reliable.

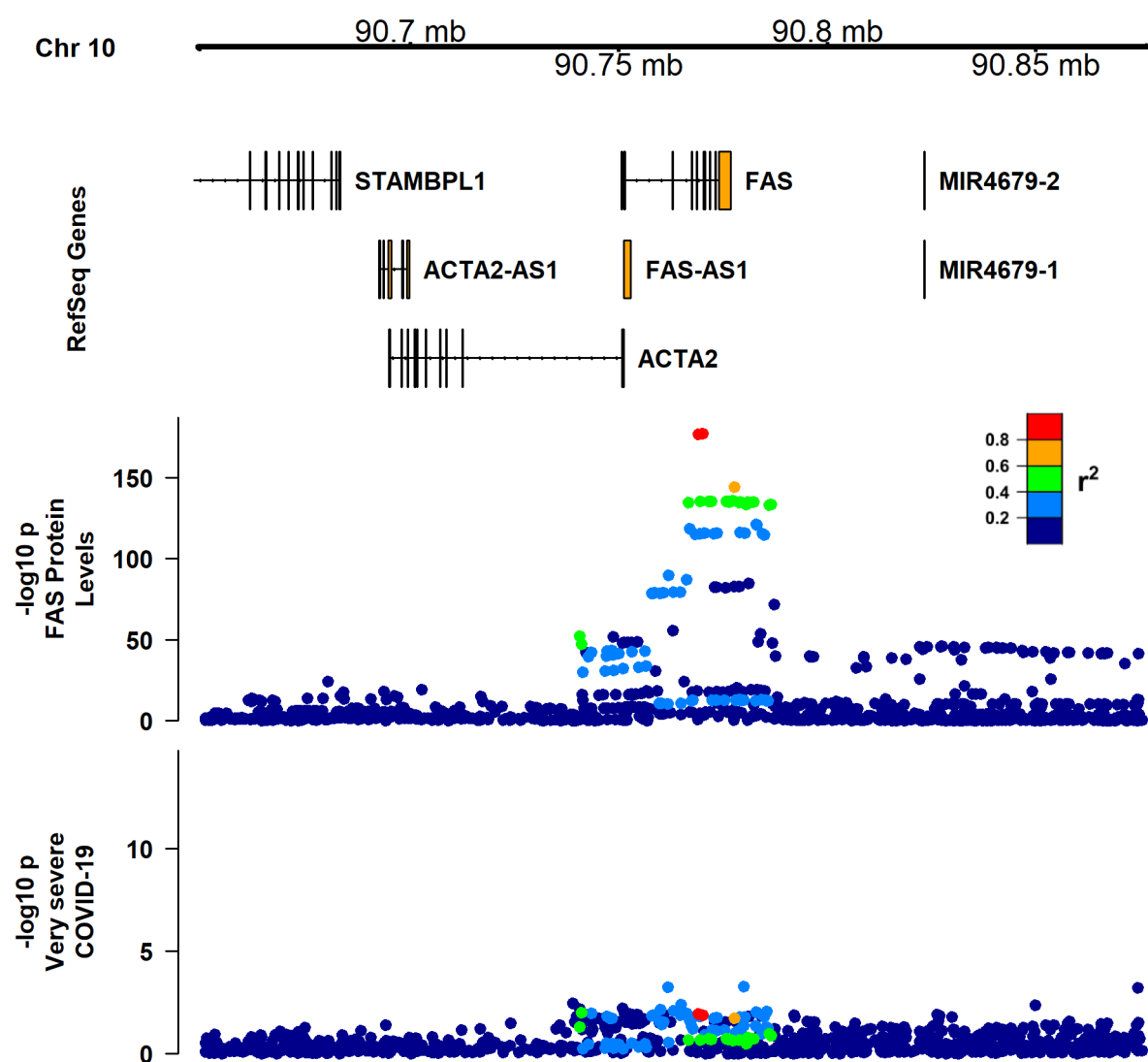


Figure 3.9 Regional associations for the FAS *cis*-pQTL and the 7th release of the COVID-19 A2 outcome

Regional association plot (hg19 genome build) showing the *cis*-pQTL for FAS (plasma) and the association of variants with the COVID-19 A2 outcome for HGI release 7. Colour indicates linkage disequilibrium (r^2) with the sentinel pQTL variant.

3.5 Discussion

A subset of patients infected with the SARS-CoV-2 virus develop severe disease characterised by marked inflammation, immunopathology, and thrombosis^{123,124}. While vaccination programmes have been effective, at the time this work was commenced there remained a need for therapies for those with severe COVID-19. Proteins are the targets of most drugs, making proteomic profiling an ideal tool for identifying potential therapeutic targets. In this study, we attempted to use Mendelian randomisation (MR) to assess whether COVID-19 severity-associated proteins play a causal role in severe disease. The strongest evidence for a causal association with severe COVID-19 was observed for the FAS cell death receptor. However, this finding did not replicate in a more recent release of the HGI meta-analysis.

As an additional outcome of this work, we identified a mechanism by which a genetic variant appears to increase the abundance of soluble FAS (sFAS) in plasma through differential splicing of its transmembrane domain (**Figure 3.8**). While the FAS *cis*-pQTL was mediated by different underlying variants to its eQTL, it did colocalise with a QTL for the differential splicing of *FAS* exon 6. Since the sixth exon of *FAS* encodes for its transmembrane domain, mRNA transcripts missing it encode for a soluble protein that can no longer trigger the intracellular cascades required for apoptosis²⁷⁵; furthermore, sFAS acts as a decoy receptor for FASL, inhibiting mFAS-FASL signaling. Therefore, genetic predisposition to produce a higher proportion of sFAS could impair FAS-mediated apoptosis. However, while it is likely that the Olink and SomaScan assays reflect the abundance of sFAS, we were not able to verify which isoform was measured. Furthermore, it is possible that differential splicing of exon 6 could modify the epitope to which the Olink and SomaScan assays bind, leading to a change in apparent protein abundance.

The sentinel FAS pQTL was not in LD with variants associated with any traits according to PhenoScanner. The variant rs28362318, however, which was included as an instrument in the FAS *cis*-MR, was associated with subarachnoid haemorrhage and treatment with antihypertensives. Activation of apoptosis by FASL has also been associated with diabetes and cardiovascular events²⁷⁷, which may explain why this variant is associated with traits related to the cardiovascular system and blood pressure.

Zhou *et al.* used MR to investigate the effect of 1,425 circulating proteins (as measured by SomaScan assays) on COVID-19 severity and susceptibility²⁷⁸. This approach identified OAS1, ABO and IL10RB as potential causal mediators, although only OAS1 had no observed pleiotropic effect and colocalised across all three COVID-19 outcomes (very severe, hospitalised and susceptible). While these authors also identified a significant association between FAS and susceptibility to very severe COVID-19 (A2), their study design prioritised proteins that were associated with at least two COVID-19 outcomes. The results presented in this thesis demonstrate that FAS was only significantly associated with the A2 outcome as of HGI release 5 (**Figure 3.4, Figure 3.9**). Since the approach used in this chapter only involved proteins found to be associated with severe COVID-19 in observational data, we did not identify OAS1, IL10RB or ABO.

Since the work described in the current chapter was released on medRxiv²⁵³, another study reported colocalisation of the FAS *cis*-pQTL using the same outcome (A2 HGI release 4²⁷⁹). The authors also noted that FAS plasma protein abundance likely reflects sFAS, and that increased abundance of sFAS could reduce mFAS-FASL signaling. They did not, however, investigate the role of splicing in the abundance of sFAS. They also did not attempt to replicate the observed associations with COVID-19 in the *FAS* gene region, either in external studies or in more recent releases of HGI.

Key findings:

- Through Mendelian randomisation analysis, proteins were identified with potential causal roles in disease susceptibility and the development of severe COVID-19. However, follow-up analyses cast doubt on the reliability of these findings.
- A mechanism was identified by which a genetic variant appears to affect apparent plasma abundance of FAS by modulating splicing. The splice variant is associated with differential splicing of FAS exon 6, that encodes the protein's transmembrane domain, which could change the relative expression of the soluble FAS isoform.

Future directions:

- Similar studies could be carried out in the broader proteome to assess whether other proteins could have a causal role in disease susceptibility and the development of severe COVID-19.

- Future MR studies could leverage larger GWAS meta-analyses of COVID-19 outcomes, thereby improving statistical power to detect potential causal relationships.

4 Multi-omics of dialysis patients with COVID-19

4.1 Chapter summary

In this chapter, I investigate the blood multi-ome of ESKD patients with COVID-19 using plasma proteomic data from SomaScan assays and transcriptomic data from RNA-sequencing of peripheral blood mononuclear cells. These approaches were applied to the cohort described previously in Chapter 2, which consisted of ESKD patients with COVID-19 sampled serially during the initial wave of infections in March and April of 2020. Additionally, 17 of the ESKD controls sampled in 2020 were later infected by SARS-CoV-2 in 2021. In this chapter, I additionally describe the analysis of these patients, for whom samples were collected prior to infection, serially during the acute COVID-19 episode and following recovery from COVID-19.

I observed transcriptomic and proteomic associations with COVID-19 infection and severity, which converged on unifying themes, such as the dysregulation of histone pathways and the interferon response. Using longitudinal modelling, I identified modules of genes with dynamic profiles in severe disease. These appeared to reflect the activity of particular cell types in the PBMC fraction, including low-density granulocytes and T lymphocytes. Supervised learning indicated that *LRRC15*, a SARS-CoV-2 spike receptor, was particularly predictive of current clinical severity. I compared the transcriptome of patients sampled before infection and following recovery from COVID-19, finding prolonged elevation of genes involved in vascular, clotting and inflammatory pathways. These findings were published in *Nature Communications*²⁷⁶.

Additionally, in this chapter I considered differences between ESKD patients on dialysis and healthy controls, which revealed transcriptional dysregulation of pathways activated by erythropoietin. I also found marked changes in the ESKD proteome, including pathways relating to ephrin, interleukin and cytokine signaling and dysregulation of triglyceride metabolism.

4.2 Introduction

In Chapter 2, I described the application of Olink proteomics to measure 436 cardiovascular- and immune-related proteins in a cohort of patients sampled during the first wave of COVID-19. These data allowed for the investigation of key proteins following symptom onset, but the Olink panels were highly targeted and measured only a small portion of the proteome. Most studies of circulating proteins in COVID-19 have used the Olink immunoassay technology^{75,176,180}, or mass spectrometry^{181,280}. The broadest Olink assay system available at the time of the present study, used in the study of Filbin *et al.*⁷⁵, measures 1,472 proteins, while mass spectrometry is generally limited to reliable detection of fewer than 1,000 plasma proteins and lacks sensitivity for low abundance proteins. A small number of studies have employed the aptamer-based SomaScan v4 platform, which measures 4,665 unique proteins^{75,243,281–283}. The work I present in this chapter is based on the SomaScan v4.1 panel, which measured 6,323 unique proteins. This was paired with bulk RNA sequencing (RNA-seq) to permit broader hypothesis-free analyses of COVID-19. In addition to the cohort described in Chapter 2, this chapter was strengthened by data from the second wave of infections, which allowed me to compare samples taken before, during and after COVID-19 from the same individuals. While the members of the Wave 1 cohort were infected early in the SARS-CoV-2 pandemic, individuals in the Wave 2 cohort were likely infected with the Alpha variant²⁸⁴, by which time the vaccine rollout had begun, and hospitals were better prepared to deal with severe COVID-19.

4.2.1 The effect of ESKD and dialysis on the PBMC transcriptome

An additional aim of the present study was to investigate the effect of ESKD and dialysis on the blood multi-ome. Previously, I described studies of the effect of renal function and dialysis on the plasma proteome (see **Section 2.2.2**). An even more limited set of studies have investigated the PBMC transcriptome in similar contexts. PBMC transcriptomics has been performed for kidney allograft recipients^{285,286}, however these are very different circumstances compared to ESKD patients undergoing dialysis. One study performed both bulk and single-cell RNA-seq for ten healthy controls and ten ESKD patients on haemodialysis²⁸⁷. These data revealed dysregulation of genes related to the renin-angiotensin system, TNF and cytokine signaling, and the PI3K, MAPK, and NFκB pathways. Prior to this, Luo *et al.*²⁸⁸ performed a similar study, also involving the application of single-cell RNA-seq (scRNA-seq) to 10 healthy

volunteers and 10 ESKD patients. This analysis revealed transcriptional dysregulation of similar pathways, in addition to STAT5 signaling and fatty acid metabolism.

4.2.2 Contributions

Most of the results described in this chapter were published in *Nature Communications*²⁷⁶. This was, of course, a collaborative project and full details of individual contributions are provided in the publication. The contributions set out in **Section 2.2.3** also apply to the study presented in this chapter, since the same cohort was studied here. Additionally, Norzawani B Buang contributed to the transcriptomic data analysis and processing, including the initial processing of the RNA-seq reads, application of Gene Set Variation Analysis (GSVA) to identify enriched pathways and relative cell abundance imputation with CIBERSORTx.

I performed the analyses described in this chapter and created the figures and tables. Some of the text, figures and tables in this chapter are repurposed from the *Nature Communications* manuscript²⁷⁶, which was published under the CC BY 4.0 license.

The CC BY 4.0 license is available at <https://creativecommons.org/licenses/by/4.0/>.

4.3 Methods

4.3.1 Patient cohorts and ethical approval

All participants were recruited from the Imperial College Renal and Transplant Centre and its satellite dialysis units in London (United Kingdom) and provided written informed consent before participation. Study ethics were reviewed by the UK National Health Service (NHS) Health Research Authority (HRA) and Health and Care Research Wales (HCRW) Research Ethics Committee (reference 20/WA/0123: The impact of COVID-19 on patients with renal disease and immunosuppressed patients). Ethical approval was given. Study volunteers provided informed consent and did not receive financial or other compensation for participating in the study.

Two cohorts of ESKD patients with COVID-19 were recruited, all of whom were receiving outpatient haemodialysis before acquiring COVID-19. To minimise variation related to the timing of dialysis, blood samples were taken prior to haemodialysis treatment. The first cohort were recruited during the initial phase of the COVID-19 pandemic (April-May 2020). In **Chapter 2**, this cohort was referred to as the “Main cohort”. In this chapter, I refer to these patients as the “Wave 1 cohort”. Blood samples were taken from 53 patients with COVID-19 (**Table 4.1**). Serial blood sampling was carried out where feasible. 59 non-infected haemodialysis patients were contemporaneously recruited to provide a control group (**Table 4.3**). Samples were also obtained for a group of 41 healthy volunteers, who did not have kidney disease and did not have clinically apparent infections (**Table 4.3**). These samples were taken from staff at Imperial College London and the National Health Service.

The Wave 2 cohort consisted of 17 ESKD patients with COVID-19 infected in January-March 2021 (**Table 4.2**). These 17 individuals had all been recruited as part of the COVID-19 negative control group during Wave 1, and so a pre-infection sample collected in April/May 2020 (8-9 months preceding infection) was also available. Serial samples were acquired approximately every 2-3 days following COVID-19 diagnosis. Samples were taken for 12 of these patients approximately two months following COVID-19 onset (range 41-55 days from the initial sample). Convalescent samples were unavailable for four patients who died and for one patient due to logistical difficulties in sample collection.

Table 4.1 Characteristics of the Wave 1 cohort

DN = diabetic nephropathy. GN = glomerulonephritis. HTN = hypertension. IQR = inter-quartile range. Subsets were defined according to peak WHO severity over the course of the illness.

*One patient had type 1 diabetes while the remainder had type 2.

†3 patients were hospitalised prior to COVID-19 diagnosis. 8 patients diagnosed with COVID-19 as outpatients subsequently deteriorated were hospitalised.

	Overall (n = 53)	Peak severity mild or moderate (n = 28)	Peak severity severe or critical (n = 25)
Age			
Median	72.0	73.0	68.0
(IQR)	62.0-76.0	64.8-76.2	62.0-76.0
Sex			
M	37 (69.8%)	18 (64.3%)	19 (76.0%)
F	16 (30.2%)	10 (35.7%)	6 (24.0%)
Ethnicity			
Asian	21 (39.6%)	11 (39.3%)	10 (40.0%)
White	17 (32.1%)	6 (21.4%)	11 (44.0%)
Black	8 (15.1%)	5 (17.9%)	3 (12.0%)
Other	7 (13.2%)	6 (21.4%)	1 (4.0%)
Diabetes	32* (60.4%)	16* (57.1%)	16 (64.0%)
Current smoker	1 (1.9%)	1 (3.6%)	0 (0.0%)
ESKD cause			
DN	27 (50.9%)	14 (50.0%)	13 (52.0%)
HTN/vascular	5 (9.4%)	3 (10.7%)	2 (8.0%)
GN/autoimmune	3 (5.7%)	1 (3.6%)	2 (8.0%)
Genetic	1 (1.9%)	1 (3.6%)	0 (0.0%)
Other/unknown	17 (32.1%)	9 (32.1%)	9 (36.0%)
Hospitalisation due to COVID-19†	32 (60.4%)	7 (25.0%)	25 (100%)
Fatal COVID-19	9 (17.0%)	0 (0.0%)	9 (36.0%)

Table 4.2 Characteristics of the Wave 2 cohort

DN = diabetic nephropathy. GN = glomerulonephritis. HTN = hypertension. IQR = inter-quartile range. Subsets were defined according to peak WHO severity over the course of the illness.

	Overall (n = 17)	Peak severity mild or moderate (n = 9)	Peak severity severe or critical (n = 8)
Age			
Median	72.0	72.0	70.5
(IQR)	61.0-77.0	60.0-75.0	64.0-80.5
Sex			
M	10 (58.8%)	4 (44.4%)	6 (75.0%)
F	7 (41.2%)	5 (55.6%)	2 (25.0%)
Ethnicity			
Asian	11 (64.7%)	6 (66.7%)	5 (62.5%)
White	4 (23.5%)	2 (22.2%)	2 (25.0%)
Black	1 (5.9%)	0 (0.0%)	1 (12.5%)
Other	1 (5.9%)	1 (11.1%)	0 (0.0%)
Diabetes	11 (64.7%)	6 (66.7%)	5 (62.5%)
Current smoker	0 (0.0%)	0 (0.0%)	0 (0.0%)
ESKD cause			
DN	10 (58.8%)	6 (66.7%)	4 (50.0%)
HTN/vascular	0 (0.0%)	0 (0.0%)	0 (0.0%)
GN/autoimmune	1 (5.9%)	0 (0.0%)	1 (12.5%)
Genetic	0 (0.0%)	0 (0.0%)	0 (0.0%)
Other/unknown	6 (35.3%)	3 (33.3%)	3 (37.5%)
Hospitalisation due to COVID-19	9 (52.9%)	1 (11.1%)	8 (100%)
Fatal COVID-19	4 (23.5%)	0 (0.0%)	4 (50.0%)

Table 4.3 Characteristics of the controls

IQR = inter-quartile range.

	ESKD controls (n = 59)	Healthy controls (n = 41)
Age		
Median	69.0	47.0
(IQR)	61.5-75.0	38.0-55.0
Sex		
M	42 (71.2%)	21 (51.2%)
F	17 (28.8%)	20 (48.8%)
Ethnicity		
Asian	30 (50.9%)	17 (41.5%)
White	14 (23.7%)	13 (31.7%)
Black	8 (13.6%)	8 (19.5%)
Other	7 (11.9%)	3 (7.3%)

4.3.2 Clinical severity scoring

Disease severity was assessed using a four-level ordinal score, categorised as mild, moderate, severe, or critical, based on the WHO clinical management of COVID-19: Interim guidance 27 May 2020. Mild was defined as COVID-19 symptoms but no evidence of pneumonia and no hypoxia. Moderate was defined as symptoms of pneumonia or hypoxia with oxygen saturation (SaO₂) greater than 92% on air, or an oxygen requirement no greater than 4 L/min. Severe was defined as SaO₂ less than 92% on air, or respiratory rate more than 30 per minute, or oxygen requirement more than 4 L/min. Critical was defined as organ dysfunction or shock or the need for high dependency or intensive care support (i.e., the need for non-invasive ventilation or intubation). Disease severity scores were recorded throughout the illness, such that samples from the same individual could have different severity scores according to the temporal evolution of the disease. The overall clinical course was defined for each patient as the peak severity score that occurred during the patient's illness. Different downstream analyses utilise either the severity at the time of sample (i.e., the sample-level severity) or the overall clinical course (i.e., the patient-level severity), as described in the relevant sections below.

4.3.3 PBMC collection protocol

Peripheral blood mononuclear cells (PBMCs) were obtained by density gradient centrifugation using Lymphoprep (STEMCELL Technologies, Canada). Approximately 20 ml of blood was diluted 1x with phosphate-buffered saline (PBS) with the addition of 2% FBS and layered on top of 15 ml of Lymphoprep solution. The samples were then centrifuged at 800 g for 20 minutes at room temperature without break. PBMCs were collected from the interface and washed twice with PBS/2%FBS. 2 million PBMCs were centrifuged down to form a pellet and resuspended in 350 μ l RLT buffer + 1% β -Mercaptoethanol (from Qiagen RNeasy kit) for RNA extraction. The remaining PBMCs were cryopreserved in 1 ml freezing medium (FBS 10% DMSO) and stored at -80°C .

4.3.4 Plasma collection

5 ml of blood was collected in EDTA tubes and centrifuged at 1000 x g for 15 mins. Plasma was extracted and frozen at -80°C .

4.3.5 RNA-seq of PBMCs

RNA extraction and sequencing were done at GENEWIZ facilities (Leipzig, Germany). Total RNA was extracted using RNeasy Mini kits (Qiagen) as per the manufacturer's instructions, with an additional purification step by on-column DNase treatment using the RNase-free DNase Kit (Qiagen) to remove any genomic DNA. Total RNA quality and concentration were analysed using Agilent TapeStation (Agilent Tech Inc.). Samples with RIN values ≥ 6.0 and ≥ 100 ng of total RNA were used to generate RNA-seq libraries. RNA-seq libraries were made using NEBnext ultra II RNA directional kit per the manufacturer's instruction. Poly-A RNA was purified using poly-T oligo-attached magnetic beads followed by hemoglobin mRNA depletion using QIAseq FastSelect Globin Kit to remove potential contaminating RNA from red blood cells. Then, the first and second cDNA strand synthesis was performed. Next, cDNA 3' ends were adenylated and adapters ligated followed by library amplification. The libraries were size selected using AMPure XP Beads (Beckman Coulter), purified and their quality was checked using a short sequencing run on MiSeq Nano. Samples were randomized to avoid confounding of batch effects with clinical status and multiplexed libraries were run on 29 lanes of the Illumina HiSeq platform to generate approximately 30 million x 150bp paired-end reads per sample.

I performed initial quality control and alignment using the nf-core RNA-seq v3.2 pipeline²⁸⁹ based on nextflow²⁹⁰, a workflow management system. I then applied FastQC³³ to evaluate and merge paired reads before trimming adapters using Trimalore³⁴. I aligned reads to GRCh38 using STAR³⁵. Finally, htseq-count³⁶ was used to generate a counts matrix.

For the Wave 1 cohort, transcriptomic data were available for 179 samples from 51 COVID-19 positive ESKD patients (median 3 samples per patient, range 1-8), plus 55 non-infected ESKD patient samples. For the Wave 2 cohort (17 patients), transcriptomic data were available for 90 samples collected during acute COVID-19 infection (median of 6 samples per patient, range 3-7), plus 17 pre-infection samples and 12 convalescent samples.

Prior to further analysis, I removed genes with insufficient counts using edgeR's filterByExpr function⁷⁶. For differential expression analyses, the group argument was set to the main group of interest. For all analyses, gene expression was TMM normalised²⁹¹, converted to counts per million (CPM) and log-transformed. I primarily used ENSEMBL identifiers²⁹², however for

plots I report the HGNC gene ID²⁹³ where available. For analyses that considered multiple genes simultaneously (PCA, M3C, WGCNA, MEFISTO, supervised learning), I additionally removed genes with low variance (unless otherwise stated, I removed 33% of genes with the lowest maximum absolute deviation)¹⁰⁶, and centered and scaled the data.

4.3.6 Plasma proteomics

Proteomics was performed on EDTA plasma samples using the aptamer-based SomaScan platform (SomaLogic, Boulder, Colorado, USA). These data were normalised by the standard procedures applied to SomaScan assays. The final normalisation step, referred to as “Median Signal Normalization to a Reference”, scales the data for each sample against a reference cohort, under the assumption that each sample contains a similar protein content. While this may be a sensible assumption when comparing COVID-19 cases and controls, there were systematic differences between ESKD patients on dialysis and healthy participants (**Figure 4.12b**). Additionally, Olink assays previously displayed upregulation of the majority of proteins in ESKD patients relative to healthy controls (**Figure 2.16**). Therefore, for analyses comparing these groups, this normalisation step was not applied.

The SomaScan v4.1 assay contains 7,288 modified aptamers (Somamers) that target human proteins. Since more than one aptamer may target the same protein, these 7,288 aptamers map to 6,347 unique proteins. I removed 48 Somamers due to QC failure, so the final dataset contained 7,240 Somamers representing 6,323 unique proteins. I annotated these proteins using the Human Protein Atlas⁵⁴. 4,980 proteins were labelled as intracellular, 1,586 were annotated as membrane proteins and 1,160 as secreted (**Figure 4.1a**). Note that many proteins were labelled as both intracellular and as membrane or secreted (**Figure 4.1b**).

I report proteins by their corresponding HGNC gene ID²⁹³, which provides a more standardised nomenclature compared to protein names and allows direct comparison with the transcriptomic data.

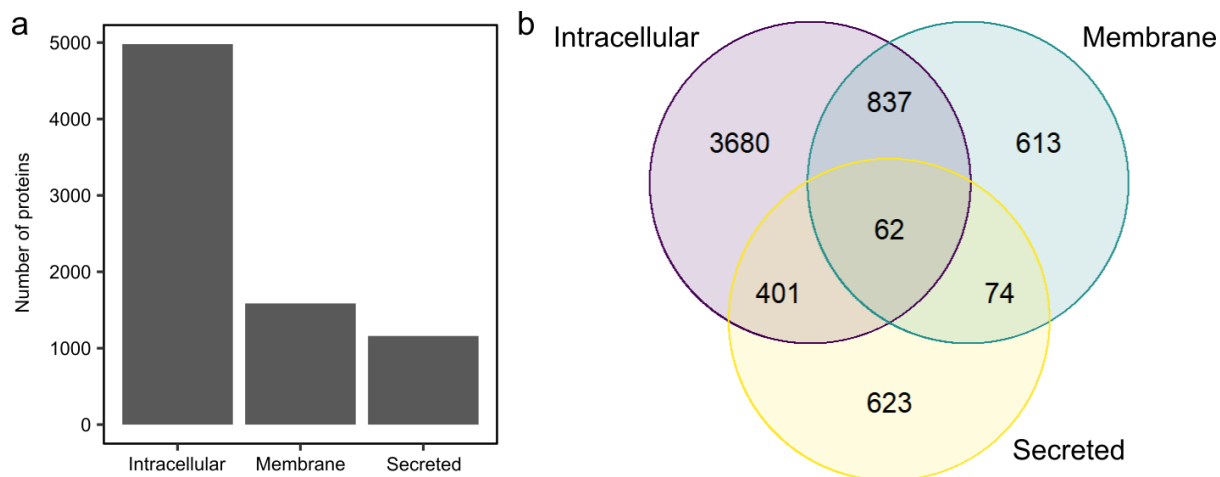


Figure 4.1 Human Protein Atlas classification of proteins measured by the SomaScan v4.1 assay

a) The number of unique proteins measured that were labelled as intracellular, membrane and secreted. **b)** Venn diagram illustrating the overlap between the annotations.

Where multiple Somamers related to the same protein, I retained these Somamers for univariate analyses such as differential abundance analyses. However, for analyses that considered multiple proteins simultaneously (PCA, M3C, WGCNA, MEFISTO, supervised learning), I selected one Somamer at random to represent each protein. The abundance values for each Somamer were inverse-rank normalised prior to downstream analyses.

For the Wave 1 cohort, proteomic data were available for 86 samples from 37 COVID-19 positive ESKD patients (median 3 samples per patient, range 1-3), plus 53 non-infected ESKD patients. For the Wave 2 cohort ($n = 17$ patients), proteomic data were available for 102 samples collected serially during acute COVID-19 infection (median of 6 samples per patient, range 5-7) and 16 pre-infection samples (for one patient, a pre-infection plasma sample was unavailable). Additionally, I excluded one COVID-19 positive sample in the Wave 2 cohort that failed QC from downstream analyses.

4.3.7 Comparison of ESKD patients on dialysis to healthy controls

I used linear models to compare patients with ESKD to a group of healthy patients with normal kidney function. Age, sex and ethnicity were included as covariates. The regression model in Wilkinson-style notation was:

$$E \sim \text{eskd_status} + \text{sex} + \text{age} + \text{ethnicity}$$

where E represents expression (gene or protein, depending on the data type being analysed) and eskd_status was a categorical variable (ESKD on dialysis or healthy).

For the transcriptomic data, I applied gene-wise negative binomial generalised linear models with quasi-likelihood tests⁷⁶ (i.e., using the estimateDisp, glmQLFit and glmQLFTest functions from the edgeR package). For the protein data, I fit linear models using R's lm function and applied type 2 F tests with the car::Anova function¹⁹⁶. I performed multiple testing correction using the Benjamini-Hochberg method and set the significance threshold as 1% FDR.

4.3.8 Comparison of COVID-19 positive and negative patients

I compared COVID-19 positive and negative patients using linear mixed models (LMMs), which account for serial samples from the same individual¹⁹⁸. Age, sex and ethnicity were included as covariates. I used a random intercept term to estimate the variability between individuals in the study and thus account for repeated measures. I performed differential expression analyses for the transcriptomic data and the proteomic data. The regression model for these analyses in Wilkinson-style notation was:

$$E \sim \text{covid_status} + \text{sex} + \text{age} + \text{ethnicity} + (1 \mid \text{individual})$$

where E represents expression (gene or protein, depending on the data type being analysed) and covid_status was a categorical variable (COVID-19 positive or negative).

For differential abundance testing of proteins, I applied LMMs using the lmerTest R package¹⁹⁷. Differential gene expression analysis was performed using the same model formula, applied using the differential expression for repeated measures (dream) pipeline⁷³ in the variancePartition R package⁷¹. Regardless of data type, I fit LMMs using restricted maximum likelihood (REML) and calculated *p*-values using a type 3 F-test, in conjunction with Satterthwaite's method for estimating the degrees of freedom for fixed effects¹⁹⁷. Multiple testing correction was performed using the Benjamini-Hochberg method and a 1% FDR was used for the significance threshold.

The Wave 1 cohort was analysed separately from the Wave 2 cohort. For Wave 1, I compared samples from COVID-19 positive ESKD patients to COVID-19 negative ESKD patients. For

Wave 2, I compared samples from COVID-19 positive ESKD patients to samples from these same patients taken approximately 8 months prior to infection.

When reporting the number of differentially expressed proteins in the text I refer to the number of unique proteins rather than the number of significant Somamers.

4.3.9 Molecular associations with COVID-19 severity

I performed a within-cases analysis, testing for the association of gene expression with COVID-19 severity at the time of sampling. I used the four-level WHO severity rating (mild, moderate, severe, critical), which could vary between samples from the same individual reflecting the clinical status at the time the sample was taken. I again used LMMs to account for samples from the same individual. The regression model was as follows:

$$E \sim \text{covid_severity} + \text{sex} + \text{age} + \text{ethnicity} + (1 \mid \text{individual})$$

The covid_severity variable represents severity at the time of the sample and was encoded using orthogonal polynomial contrasts to account for the ordinal nature of severity levels.

COVID-19 positive samples from the Wave 1 cohort were analysed separately from those from the Wave 2 cohort.

The same approach was used for the proteomics data.

4.3.10 Gene set variation analysis

To identify pathways that were up- or down-regulated in COVID-19 positive versus negative samples, I applied gene set variation analysis (GSVA)²⁹⁴. To define gene sets, I used the MSigDb C2 canonical pathways²⁹⁵. Sets with less than ten genes were discarded. I additionally included a gene set for the peripheral immune response defined for patients with severe COVID-19²⁹⁶ and a set of type 1 interferons active in patients with systemic lupus erythematosus (SLE)²⁹⁷. After the reduction of genes into gene sets, I performed testing for dysregulated pathways using the same linear mixed modelling approach as for the differential gene expression analyses. *p*-values were adjusted by Benjamini-Hochberg, with a significance threshold of 1% FDR. This same approach was applied to the proteomics data.

To assess the key molecules underpinning enriched pathways, I examined the genes and proteins that comprise these pathway terms and identified which of these featured most prominently in the differential gene expression and protein abundance analyses.

I repeated this procedure for i) testing of association of pathways with severity at the time of sample using the 4-level ordinal score and ii) testing for differences between ESKD patients on dialysis and healthy participants. For the proteomic comparison of ESKD patients and healthy controls, I applied the “plage” method implemented in the GSVA package. The standard GSVA method ranks genes on a per-sample basis²⁹⁴, whereas the PLAGE method does not. Sample-based normalisation may not be appropriate for the comparison of ESKD patients and healthy controls, as discussed in **Section 4.3.6**.

4.3.11 Robust rank aggregation

The Wave 1 and Wave 2 cohorts were analysed separately for both the differential expression analyses between COVID-19 positive and negative samples and for the within-cases severity analyses. To identify the associations that were most consistent between the Wave 1 and Wave 2 cohorts, for each analysis, I integrated the *p*-values for each cohort using robust rank aggregation (RRA)²⁹⁸. This method identifies features that are ranked higher than expected across multiple lists. RRA generates a significance score analogous to a *p*-value; I -log10-transformed these values such that a larger score indicates more consistent associations between the Wave 1 cohort and the Wave 2 cohort. I applied RRA to the results of the transcriptomic, proteomic and GSVA analyses comparing COVID-19 positive versus negative samples from Wave 1 and Wave 2. Similarly, I applied it to the analyses testing for the association of molecular features with COVID-19 severity at the time of sampling.

4.3.12 Modelling modular longitudinal trajectories

I examined the temporal trajectories of the transcriptome following infection, by explicitly modelling molecular markers with respect to time following COVID-19 symptom onset. I used a two-step approach.

Step 1. To aid biological interpretation, I first applied a dimension reduction strategy using weighted gene correlation network analysis (WGCNA)⁹⁶ to identify modules of correlated molecular features. For this analysis, I combined samples from the Wave 1 and Wave 2 cohorts. Additionally, since the goal was to perform longitudinal analysis, I only selected

patients who had been sampled at least three times prior to 21 days following COVID-19 symptom onset. The default implementation of WGCNA is not designed for use with non-independent samples²⁹⁹, so I modified the analysis pipeline by generating a correlation matrix using a repeated measures correlation metric (rmcorr) that considers non-independence between samples¹⁹⁹. I used WGCNA's pickSoftThreshold.fromSimilarity function to pick the minimum soft-thresholding power that satisfied the minimum scale-free topology fitting index ($r^2 > 0.85$) and was associated with a maximum mean connectivity of 100. I subsequently defined signed adjacency and topological overlap matrices before applying average-linkage hierarchical clustering. I cut this tree with a hybrid dynamic tree-cutting algorithm, with the parameters deepSplit = 4 and minClusterSize = 30³⁰⁰. Finally, I defined eigengenes for each module and merged those with a distance of less than 0.25. The eigengenes provide a numerical representation for each module of co-expressed genes. I used the same approach to analyse the proteomic data.

Step 2. To examine the trajectory of each module over time, I fit LMMs with time from symptom onset as an independent variable and the eigengene (or eigenprotein, in the case of proteomic modules) as the dependent variable. Time was defined for each sample as time from the first onset of symptoms; where the date of first symptoms was not available, I instead used the date of the first positive swab. Samples that were taken more than 21 days from each individual's baseline date were excluded. I used R's bs function to fit a polynomial spline of degree two to model the expression of modules with respect to time from baseline²⁰⁴. To test whether modules displayed different temporal patterns according to the overall clinical course of COVID-19 (defined as a binary variable indicating whether the peak WHO severity score was mild/moderate or severe/critical), I included clinical course as a covariate in the model and an interaction term between time from symptom onset and clinical course (TxCC).

The regression model used is displayed using Wilkinson-style notation below.

```
eigen-expression ~ clinical_course * time + sex + age + ethnicity + wave + (1 | individual)
```

I extracted the *p*-values for the TxCC term in this model and applied Benjamini-Hochberg adjustment, using 5% FDR as the significance threshold. A significant interaction effect for the TxCC term indicates that the module has a different temporal profile in mild/moderate versus severe/critical disease.

4.3.13 Additional WGCNA module annotation and association testing

To better understand the biological information reflected in the transcriptomic and proteomic modules, I further characterised them through a multi-pronged analytical strategy. I tested the association of eigen-genes and eigen-proteins with other variables. First, I tested for the association of the modules with WHO severity at the time of the sample using the LMM approach described above in **Section 4.3.9**, i.e.:

$$\text{eigen-expression} \sim \text{covid_severity} + \text{sex} + \text{age} + \text{ethnicity} + \text{wave} + (1 \mid \text{individual})$$

Second, since PBMCs represent a mixed population of immune cells, I investigated whether disease trajectory-associated transcriptomic modules might reflect a shift in cell type proportions. To this end, I applied CIBERSORTx, a computational algorithm to impute immune cell fractions from RNA-seq data³⁰¹ (see **Section 4.3.14**). I then tested for correlations between these imputed immune cell proportions and module eigengenes using LMM:

$$\text{eigen-expression} \sim \text{cell_fraction} + \text{sex} + \text{age} + \text{ethnicity} + \text{wave} + (1 \mid \text{individual})$$

Both these models included an additional fixed effect (wave) to reflect the cohort.

Third, I performed pathway enrichment analysis on the modules using the R package clusterProfiler's enricher function³⁰². Gene sets were defined using MSigDB C2 canonical pathways²⁹⁵.

Lastly, to understand the relationship between the transcriptomic and proteomic modules, I performed correlation analysis using LMMs.

5% FDR was used for statistical significance for these analyses.

4.3.14 Cell fraction imputation

I used CIBERSORTx³⁰¹ to impute cell fractions from the normalised bulk RNA-seq dataset. The program was run with default parameters. I inferred the cell fractions of 22 immune cell types in the isolated PBMCs of each sample using the LM22 signature matrix file³⁰³.

4.3.15 Multi-omic longitudinal factor analysis with MEFISTO

MEFISTO³⁰⁴ is an extension of Multi-Omics Factor Analysis (MOFA) that can exploit temporal relationships between samples to find factors that change over time (from baseline). I used this method to find joint factors of variation in the transcriptomic and proteomic datasets. For

the MEFISTO analysis, I used the same set of samples as in the network analysis and applied the same pre-processing steps to the data (see **Section 4.3.12**). Additionally, I removed genes with the lowest maximum absolute deviation¹⁰⁶ such that the number of genes retained were equal to the number of unique proteins measured (6,323) to avoid imbalanced numbers of features between the transcriptomic and proteomic data, which can impact the MEFISTO algorithm. Using the slow convergence criterion, MEFISTO identified 8 factors that had a minimal variance explained of 1% in at least one data modality.

I then applied the longitudinal model described earlier to test for an interaction effect between time from first symptoms and clinical course (TxCC), with a latent factor identified by MEFISTO as the dependent variable. The regression model used is displayed using Wilkinson-style notation below:

$$\text{latent_factor} \sim \text{clinical_course} * \text{time} + \text{sex} + \text{age} + \text{ethnicity} + \text{wave} + (1 \mid \text{individual})$$

4.3.16 Longitudinal modelling of cytokines and cytokine receptors

I modelled the temporal profiles of 232 plasma proteins that fell within the KEGG pathway “Cytokine-cytokine receptor interaction”. As for the longitudinal analyses described earlier, I used LMMs with a time x clinical course interaction (TxCC) term.

$$P \sim \text{clinical_course} * \text{time} + \text{sex} + \text{age} + \text{ethnicity} + \text{wave} + (1 \mid \text{individual})$$

P-values for the time x clinical course interaction were extracted and adjusted for multiple testing with the Benjamini-Hochberg procedure, with a significance threshold of 5% FDR.

4.3.17 Supervised learning

The goal of this analysis was to predict clinical severity from the molecular features (transcriptomic, proteomic or both). I performed supervised learning using the R caret framework¹⁹⁵, which uses the randomForest package to fit random forest models and glmnet³⁰⁵ to fit lasso models. For this analysis, I only included samples on which both transcriptomics and proteomics had been performed. I then selected the earliest sample for each individual at which they had reached their peak COVID-19 WHO severity score, so that there was one sample per patient. I then categorised the clinical severity score corresponding to each sample into a binary variable such that patients with a WHO severity score of mild or moderate were considered mild/moderate and those with a WHO score of severe or critical

were considered severe/critical. In total, there were $n = 37$ mild/moderate samples and $n = 14$ severe/critical samples included.

I trained models using Monte Carlo cross-validation for i) the plasma proteomic data alone (6,323 features), ii) the PBMC RNA-seq data alone (12,225 features) and iii) the combined proteomic and RNA-seq datasets. The first step in this training process was to create 200 random partitions of the data, such that 80% of the data was used to train the model in each resample and 20% was retained as a validation set. In each resample, I calculated the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. I then calculated confidence intervals for the 200 AUC-ROC values generated for each model and feature type.

The random forest model's parameters were kept constant at 500 trees and the mtry value (number of proteins randomly sampled as candidates at each node) was calculated as the square root of the number of features. After cross-validation, I fit a final random forest model using the entirety of the dataset. I extracted important features from this model using the randomForestExplainer R package²⁰³, based on the accuracy decrease metric (the average decrease in prediction accuracy upon swapping out a feature). For the lasso model, the lambda value that maximised the mean AUC-ROC during cross-validation was selected. I recorded the features selected by the lasso model in each data resample. Feature importance was subsequently defined as the number of models in which each feature had a non-zero coefficient. The feature importance metrics from both models were scaled by dividing their values by the maximum value, such that the most important feature has an importance metric of 1.

4.3.18 Comparison of preinfection and recovery timepoints

For the 12 individuals in the Wave 2 cohort for whom a convalescent sample had been collected (approximately 2 months post-infection; range 41-55 days from the initial sample), I performed a differential gene expression analysis comparing these samples to the paired pre-infection samples using LMMs, implemented within the R dream package⁷³. Age, sex and ethnicity were included as covariates and a random intercept term was used to account for the paired nature of the samples. Statistical significance was defined as 1% FDR. To identify enriched pathways in the list of differentially expressed genes, I performed overrepresentation

analyses using the same approach as described above for annotating the WGCNA modules (see **Section 4.3.13**).

4.3.19 Data and code availability

The individual-level transcriptomics (counts), proteomics and flow cytometry data are available without restriction from Zenodo (<https://doi.org/10.5281/zenodo.6497251>). The raw RNA-seq reads are under restricted access to comply with UK GDPR legislation and have been deposited in the European Phenome-Genome archive (EGA) under study accession EGAS00001006778 (<https://ega-archive.org/studies/EGAS00001006778>); requests for access can be made to the Data Access Committee.

An archived GitHub repository containing custom analysis code is available from Zenodo (<https://doi.org/10.5281/zenodo.7333789>).

In this study, I utilised the whole blood bulk RNA-seq generated by the COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium study³⁰⁶, which is available from Zenodo (<https://doi.org/10.5281/zenodo.6120249>). I also used the SomaScan proteomics data of Filbin *et al.*⁷⁵, accessed through Mendeley Data (<https://doi.org/10.17632/nf853r8xsj.2>).

4.4 Results

In this chapter, I investigate data generated for two cohorts of ESKD patients on haemodialysis presenting with COVID-19 (**Figure 4.2**). The first cohort studied was the “Main cohort” described in **Chapter 2**. In this chapter, I refer to this cohort as the “Wave 1” cohort, as it consisted of 53 patients recruited during the initial phase of the COVID-19 pandemic (April-May 2020) (**Table 4.1**). Serial samples were acquired for these patients, although this sampling was not performed consistently for all individuals given the immense pressure on hospital services and the national lockdown that was in effect at the time (**Figure 4.3**). Disease severity was assessed using a WHO four-level ordinal score, with each patient categorised at each timepoint into mild, moderate, severe, or critical. Of the 53 patients, 25 had a peak illness severity score of severe or critical (hereafter severe/critical) and 28 mild or moderate (hereafter mild/moderate). 59 non-infected haemodialysis patients were contemporaneously recruited to provide a control group. Also recruited were 41 healthy participants, who had neither kidney disease nor known infections (**Table 4.3**).

In this chapter, I additionally analyse a cohort of 17 ESKD patients with COVID-19 who were infected during the resurgence of cases in January-March 2021 (**Figure 4.2, Table 4.2**). I refer to these patients as the “Wave 2” cohort. All had been initially recruited as part of the COVID-19 negative control group during Wave 1, meaning that a pre-infection sample was available, collected 8-9 months earlier. Serial samples were systematically acquired for all patients in the Wave 2 cohort at regular intervals (every 2-3 days over the course of the acute illness) (**Figure 4.3**). 9 patients had a peak illness severity of severe/critical (of whom 4 died), and 8 mild/moderate. For 12 of these patients, convalescent samples were acquired approximately two months following infection.

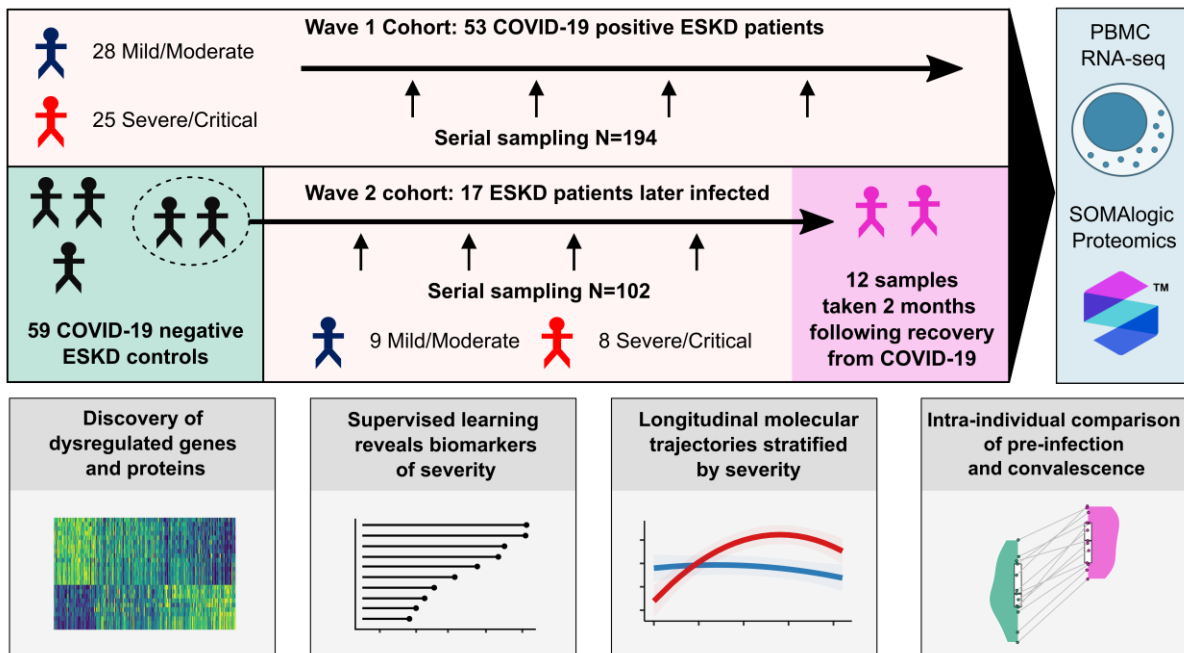


Figure 4.2 Study design

Graphical summary of the patient cohorts, sampling, and major analyses. Wave 1 patients were recruited in Spring 2020. 17 of the COVID-19 negative ESKD patients recruited as controls in Wave 1 were recruited again as COVID-19 positive cases in Wave 2 (2021). For 12/13 survivors in Wave 2, a convalescent sample was obtained approximately 2 months following recovery. Thus, for Wave 2, there were paired pre-infection, acute infection and post-infection samples from the same individuals.

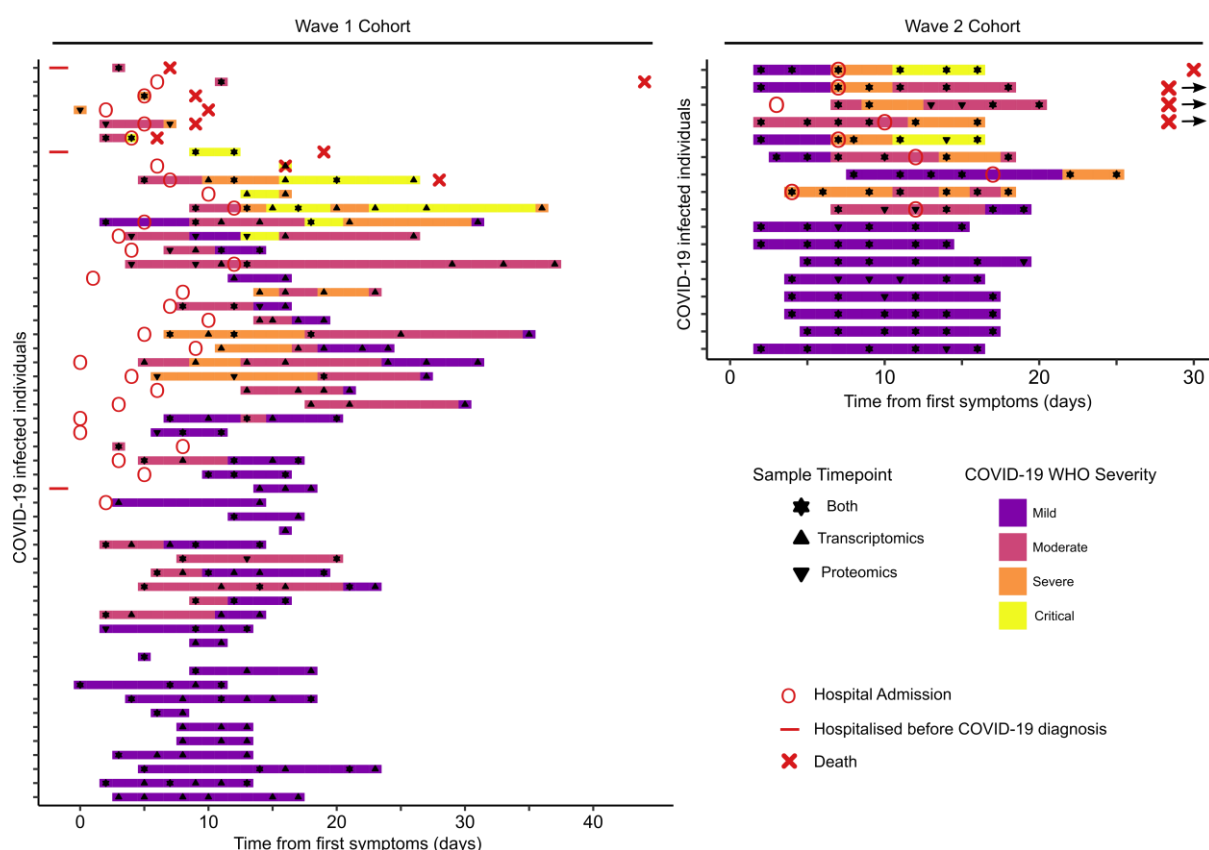


Figure 4.3 Cohort summary

For each cohort, the timing of the serial blood sampling is shown by triangles and the temporal COVID-19 severity by coloured bars. Three patients were hospitalised prior to COVID-19 diagnosis in the Wave 1 cohort. Three of the four patients in the Wave 2 cohort with fatal outcomes died >30 days from first positive swab.

4.4.1 The effect of COVID-19

RNA sequencing (RNA-seq) was performed for peripheral blood mononuclear cells (PBMCs). Principal components analysis (PCA) revealed a clear effect of COVID-19 in both Wave 1 (COVID-19 positive and negative patient samples) and Wave 2 (pre-infection and subsequent COVID-19 positive samples from the same individuals) (**Figure 4.4**). In the Wave 1 cohort, I performed differential gene expression analysis between COVID-19 positive ($n = 179$ samples from 51 patients) and negative samples ($n = 55$) using linear mixed models (LMMs) to account for repeated samples from the same individuals. This approach identified 3,026 significantly up-regulated and 3,329 down-regulated genes (at a 1% false discovery rate [FDR]). For the Wave 2 cohort, I compared COVID-19 positive samples ($n = 90$ samples from 17 individuals) with pre-infection samples from these same individuals. This analysis revealed 2,871 up-regulated and 3,325 down-regulated genes (1% FDR, LMM). In total, 3,468 genes were

significantly differentially expressed (1% FDR) in both the Wave 1 and the Wave 2 cohorts. However, using a hard threshold (i.e., based on p -values) to identify the intersection of significant features can underestimate the commonality between two datasets³⁰⁷. To provide a measure of consistency that is not dependent on the significance threshold, I compared the estimated effect size (log2 fold change) between COVID-19 positive and negative samples for each gene in the Wave 1 and Wave 2 cohorts. These were highly concordant (Pearson's $r = 0.80$), despite differences in the prevalent SARS-CoV-2 variant and developments in medical management (8 of 17 patients in the Wave 2 cohort received glucocorticoids).

The large numbers of differentially expressed genes in these cohorts indicated widespread transcriptomic changes in COVID-19. However, it is difficult to interpret long lists of differentially expressed features meaningfully and systematically. To gain insight into the biological pathways underlying these changes, I used Gene Set Variation Analysis (GSVA)²⁹⁴ to compare COVID-19 positive and negative ESKD samples. Still, this analysis resulted in the identification of a large number of significantly enriched pathways for both cohorts. To remedy this, I used robust rank aggregation (RRA) to identify genes and pathways that were consistently differentially expressed in both cohorts. When interpreting these data, I focussed on the pathways with the highest RRA scores (**Figure 4.5**) and considered genes that were both members of these pathways and were significantly differentially expressed in both cohorts (some of which are visualised in **Figure 4.6a**). Enriched pathways included those related to cell cycle (e.g., “Polo-like kinase mediated events”, which are involved in the cellular response to DNA damage) and host defence (e.g., “Complement cascade”, “Fc-gamma receptor-dependent phagocytosis”, and “Parasite infection”) (**Figure 4.5**). This analysis also highlighted leukocyte-endothelial interactions (“Cell surface interactions at the vascular wall”, which included *SELL* and *CEACAM-1*, -3, -6 and -8 genes) (**Figure 4.5**). Examples of marked changes in gene expression between the pre-infection and first acute infection sample in the Wave 2 cohort included components of the “Immunoregulatory interactions between a lymphoid and a non-lymphoid cell pathway” term (e.g., *SIGLEC1*, *SIGLEC9*, *SELL*, all increased) and “Development and heterogeneity of the ILC family” (e.g., *IFNG*, *GATA3*, *RORA*, all decreased) (**Figure 4.6a**).

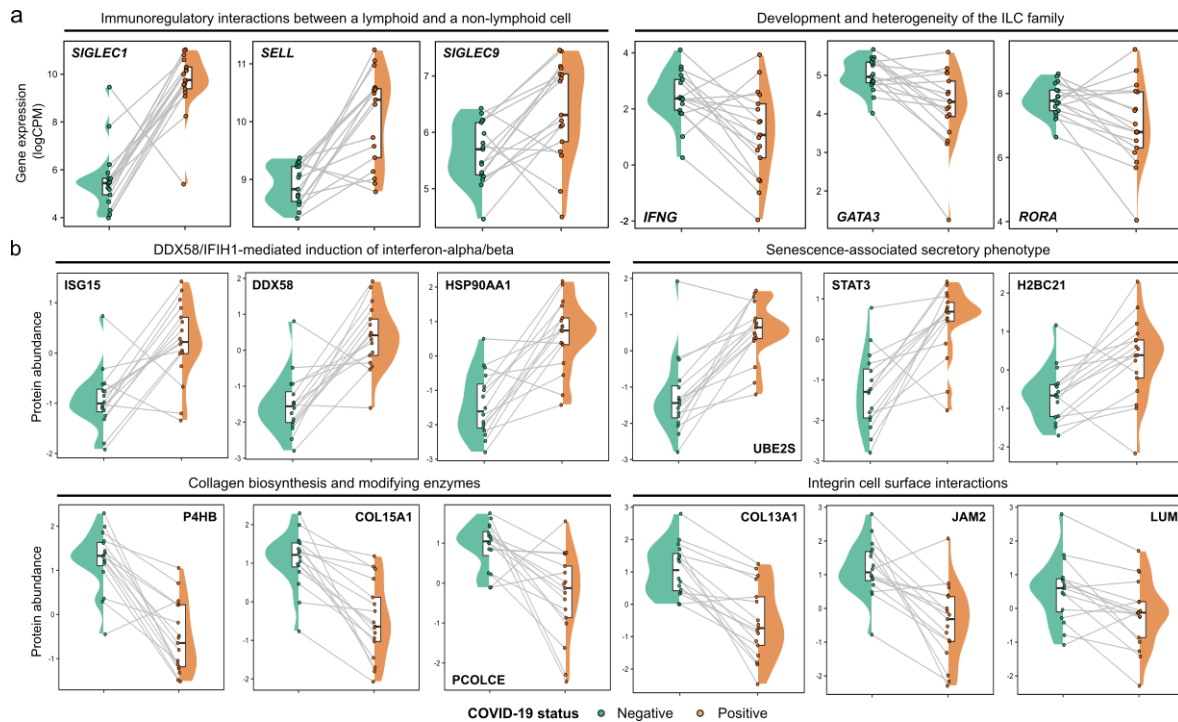


Figure 4.6 Intra-individual comparison of pre-infection and acute COVID-19.

Paired violin plots showing intra-individual comparisons of pre-infection and most severe sample (Wave 2 cohort) during COVID-19 for selected genes and proteins. Grey lines link each individual's pre-infection and infection samples; these samples are represented by points and coloured by COVID-19 status. Shaded areas indicate kernel density estimates. For boxplots, centre = median, upper bound = upper quartile, lower bound = lower quartile. All genes and proteins shown were significantly associated with infection (1% FDR) in both cohorts. Genes and proteins are grouped by their membership to pathways that were significantly enriched (1% FDR) in GSVA. **a**) Paired violin plots displaying gene expression ($n = 34$ samples from 17 individuals). **b**) Paired violin plots showing protein abundance ($n = 32$ samples from 16 individuals).

I next assessed the circulating proteome, based on 6,323 unique proteins measured with the SomaScan platform. PCA again indicated clear differences between COVID-19 positive and negative samples (**Figure 4.4**). I employed differential abundance analysis using LMMs to compare COVID-19 positive and negative samples. In the Wave 1 cohort (86 samples from 37 COVID-19 positive ESKD patients versus 53 non-infected ESKD patient samples), there were 1,273 proteins significantly (1% FDR) differentially abundant between COVID-19 positive and negative samples. In Wave 2, a comparison of COVID-19 positive samples ($n = 102$ samples from 17 patients) with pre-infection samples from the same individuals identified 5,265 differentially abundant proteins. The effect sizes were generally concordant between the cohorts (Pearson's $r = 0.57$), although the correlation was not as strong as was observed

for the expression of genes by PBMCs. There were 730 proteins significantly differentially abundant in both the Wave 1 and the Wave 2 cohorts.

I performed enrichment analysis based on the protein data and considered the pathways that were consistently dysregulated in both cohorts according to RRA. These pathways included “DDX58 IFIH1 mediated induction of interferon-alpha/beta”, “Wilk *et al.*, 2021 IFN module”²⁹⁶, “Host-pathogen interaction of human coronaviruses interferon induction” and “SARS-CoV-2 innate immunity evasion and cell-specific immune response”, reflecting host anti-viral responses (**Figure 4.7**). Furthermore, the prioritisation of pathways specifically related to SARS-CoV-2 and coronaviruses provides a validation of the analysis. Highly up-regulated proteins within these pathways included STAT1; DDX58 and ISG15, both crucial to the IFN-mediated antiviral response in COVID-19³⁰⁸; IFITM3, which is up-regulated in lung epithelial cells during early SARS-CoV-2 infection³⁰⁹; and the chemokines CXCL11, CXCL1, CXCL6, CXCL5 and CXCL10. Another significantly up-regulated pathway was “Senescence-associated secretory phenotype”, which included up-regulated ubiquitin-conjugating enzymes (UBE2S, UBE2E1), histones (H2BC21, H2BU1) and STAT3 (**Figure 4.6b**). Down-regulated pathways included “Integrin cell surface interactions” and “Collagen biosynthesis and modifying enzymes” which contained collagen proteins (e.g., COL11A2, COL13A1, COL15A1) and related enzymes (e.g., P4HB, PCOLCE) (**Figure 4.6b**).

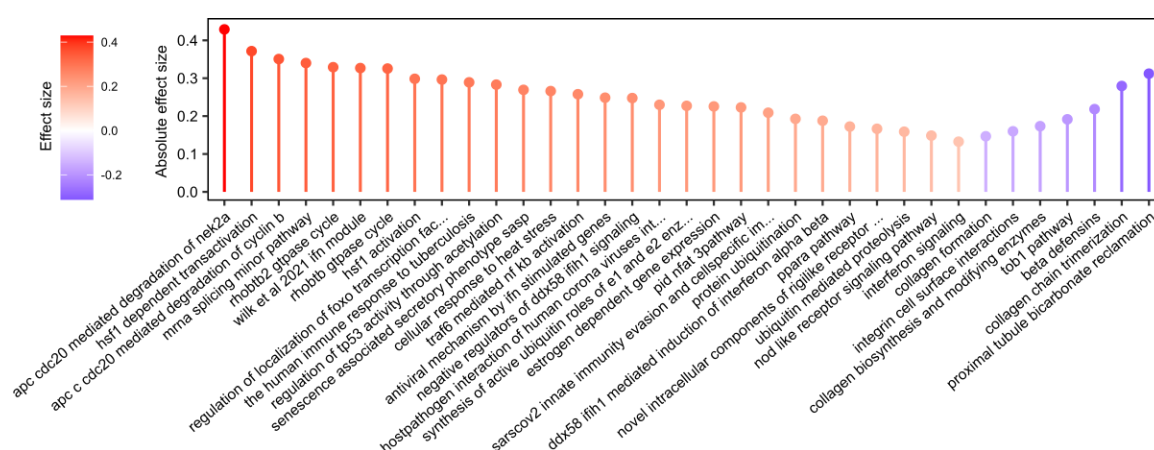


Figure 4.7 Proteomic enrichment analysis comparing COVID-19 positive and negative samples.

The 30 protein pathway enrichment terms with the greatest RRA scores (indicating consistent dysregulation in both the Wave 1 and Wave 2 proteomic datasets), ordered by effect size. All pathway terms shown were significantly enriched in the individual cohort analyses (1% FDR). Red = up-regulated in COVID-19 positive versus negative; blue = down-regulated.

4.4.2 Transcriptomic and proteomic changes associated with COVID-19 severity

In both cohorts, the PCA of the PBMC transcriptomics revealed differences according to both severity at the time of sampling and overall clinical course (defined by peak severity score) (**Figure 4.8a**). There was a gradient of severity reflected in the molecular data. I next assessed which molecular features were associated with severity at the time of blood sampling, encoded as an ordinal variable. This approach identified 3,522 transcripts that were significantly associated with contemporaneous severity in the Wave 1 cohort and 657 in the Wave 2 cohort (1% FDR). 363 genes were significantly associated in both cohorts.

I subsequently applied GSVA to identify pathways and used RRA to combine results from each cohort. The up-regulated transcriptomic pathways in more severe disease included those involved in oxidative stress (“Glutathione metabolism”, “Detoxification of reactive oxygen species”), “Transcriptional regulation of granulopoiesis”, pathways containing numerous histone-encoding genes (“HDACs deacetylate histones”, “Diseases of programmed cell death”, “RHO GTPases activate PKNs”) and “Complement and coagulation cascades” (**Figure 4.8b, c**). Down-regulated pathway terms included “TCRA pathway”, “Pathogenesis of SARS-CoV-2 mediated by nsp9-nsp10 complex”, “TP53 activity”, and “PD1 signaling”, indicating a lower abundance of molecules associated with T cell activation in more severe COVID-19 (**Figure 4.8b, c**).

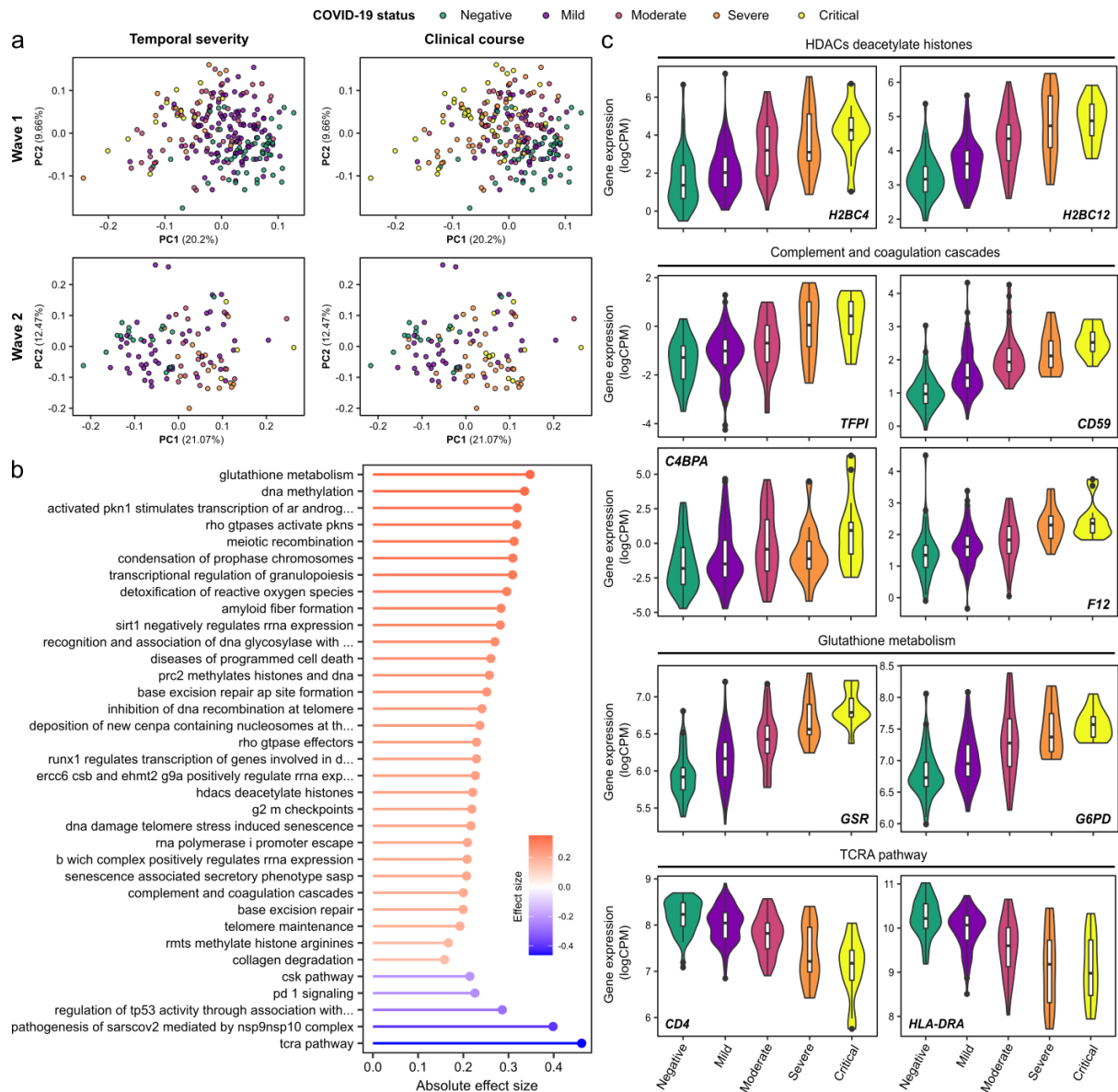


Figure 4.8 Association of the PBMC transcriptome and COVID-19 severity

a) PCA of the PBMC transcriptome. Each point represents a sample and is coloured by contemporaneous COVID-19 WHO severity (left) and overall clinical course (right). The variance explained by PC1 and 2 is indicated in parentheses. **b)** The 30 GSEA transcriptomic pathway enrichment terms with the greatest RRA scores. All were significantly enriched in both Wave 1 and 2 cohorts (1% FDR). Terms are ordered and coloured by their effect size. Red = up-regulated in more severe COVID-19; blue = down-regulated. **c)** Violin plots show gene expression values (Wave 1 cohort; $n = 179$ samples from 51 COVID-19 positive ESKD patients and 55 samples from COVID-19 negative ESKD patients) stratified by COVID-19 status and severity (at time of sample) for selected genes. Shaded areas indicate kernel density estimates. For boxplots, centre = median, upper bound = upper quartile (UQ), lower bound = lower quartile (LQ), upper whisker = largest value at most $1.5 \times \text{IQR}$ (inter-quartile range) from the UQ, lower whisker = smallest value at most $1.5 \times \text{IQR}$ from the LQ, points = samples outside of the range of the whiskers. All genes shown were significantly associated (1% FDR) with severity in both the Wave 1 and 2 cohorts. Genes are grouped by their membership to pathways that were significantly enriched (1% FDR) in GSEA.

PCA of the proteomic data revealed differences according to clinical severity (**Figure 4.9a**). I performed differential abundance analysis with LMMs, which identified 148 and 1,625 proteins associated (1% FDR) with disease severity in the Wave 1 (86 COVID-19 positive samples) and Wave 2 (102 COVID-19 positive samples) datasets, respectively. 98 proteins were associated with severity in both datasets. I next applied pathway analysis, revealing a total of 15 severity-associated pathway terms that reached statistical significance (1% FDR) in both cohorts (**Figure 4.9b**). Pathways upregulated in more severe disease were “HDACs deacetylate histones”, which included various histones (**Figure 4.9c**). There was also upregulation of pathways related to transcriptional regulation (e.g., “mRNA splicing minor pathway”, “Spliceosome”, “RNA polymerase II transcription termination”, “Processing of capped intron-containing pre mRNA”); differentially abundant members of these pathways included serine and arginine-rich splicing factors (SRSF6 and SRSF7) and those related to transcriptional regulation (e.g., SNRNP25, PUF60, YBX1) (**Figure 4.9c**). The most down-regulated pathways included “PD-1 signaling” and “T-cell receptor and costimulatory signaling” (**Figure 4.9b, c**), echoing the results of the transcriptomic pathway analysis (**Figure 4.8b, c**).

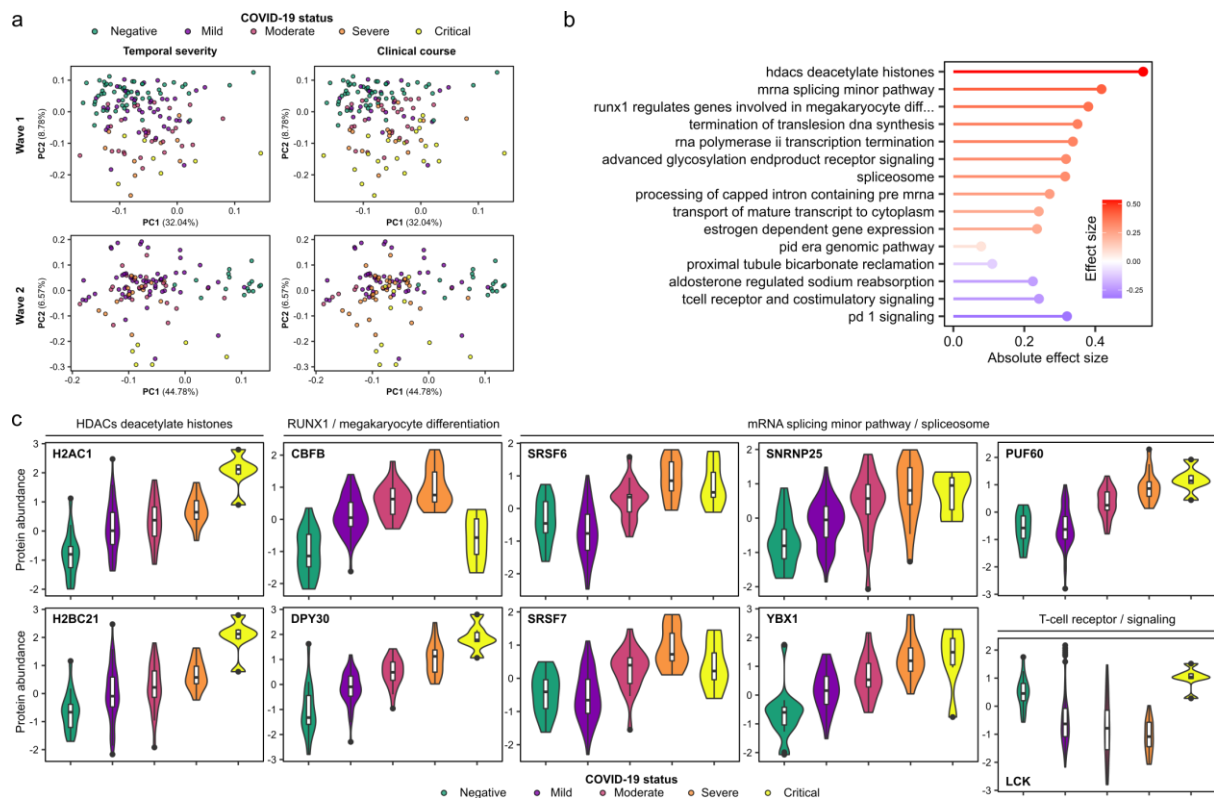


Figure 4.9 Plasma proteomic associations with COVID-19 severity

a) PCA of the proteome. Each point represents a sample and is coloured by contemporaneous COVID-19 severity (left) and the overall clinical course of the patient (right). **b)** The GSEA protein pathway enrichment terms significantly associated (1% FDR) with contemporaneous severity in both the Wave 1 and Wave 2 proteomic datasets. All pathway terms shown were significantly enriched in the individual cohort analyses (1% FDR). Terms are ordered and coloured by their effect size. Terms up-regulated in more severe COVID-19 are coloured red while down-regulated terms are blue. **c)** Violin plots show plasma protein abundance (Wave 2 cohort; 102 samples from 17 individuals and 16 pre-infection samples) stratified by COVID-19 status and severity (at the time of sample) for selected proteins. All proteins shown were significantly associated (1% FDR) with contemporaneous severity in the Wave 2 cohort. Shaded areas indicate kernel density estimates. For boxplots, centre = median, upper bound = upper quartile (UQ), lower bound = lower quartile (LQ), upper whisker = largest value at most 1.5 * IQR (inter-quartile range) from the UQ, lower whisker = smallest value at most 1.5 * IQR from the LQ, points = samples outside of the range of the whiskers.

4.4.3 Exploring the effects of including additional clinical covariates

I considered whether clinical factors related to ESKD might affect the PBMC transcriptome or the plasma proteome, and potentially confound LMM-based analyses of COVID-19 status or severity. I first evaluated whether the cause of ESKD, diabetes or duration of time since first commencing haemodialysis had any impact on PCA plots for the transcriptome or the proteome. These covariates did not have any visible effect on the PCA plots, in either COVID-19 negative or COVID-19 positive ESKD patients (**Figure 4.10**). Nevertheless, to ensure that the results were not confounded by the effect of ESKD-related variables, I next performed sensitivity analyses comparing the primary statistical models to a model with additional covariates (adjusting for the cause of ESKD, diabetes status and duration of time spent on dialysis). I performed these sensitivity analyses for i) the Wave 1 differential expression and abundance analyses comparing COVID-19 positive and negative ESKD patients, and ii) the transcriptomic and proteomic analyses testing for molecular associations with clinical severity. I found that adjustment for these covariates had very little effect on the comparisons of interest (**Figure 4.11**). The Wave 2 differential gene expression and protein abundance models compared COVID-19 positive samples to pre-infection samples from the same individuals, so inter-individual variation is inherently controlled for.

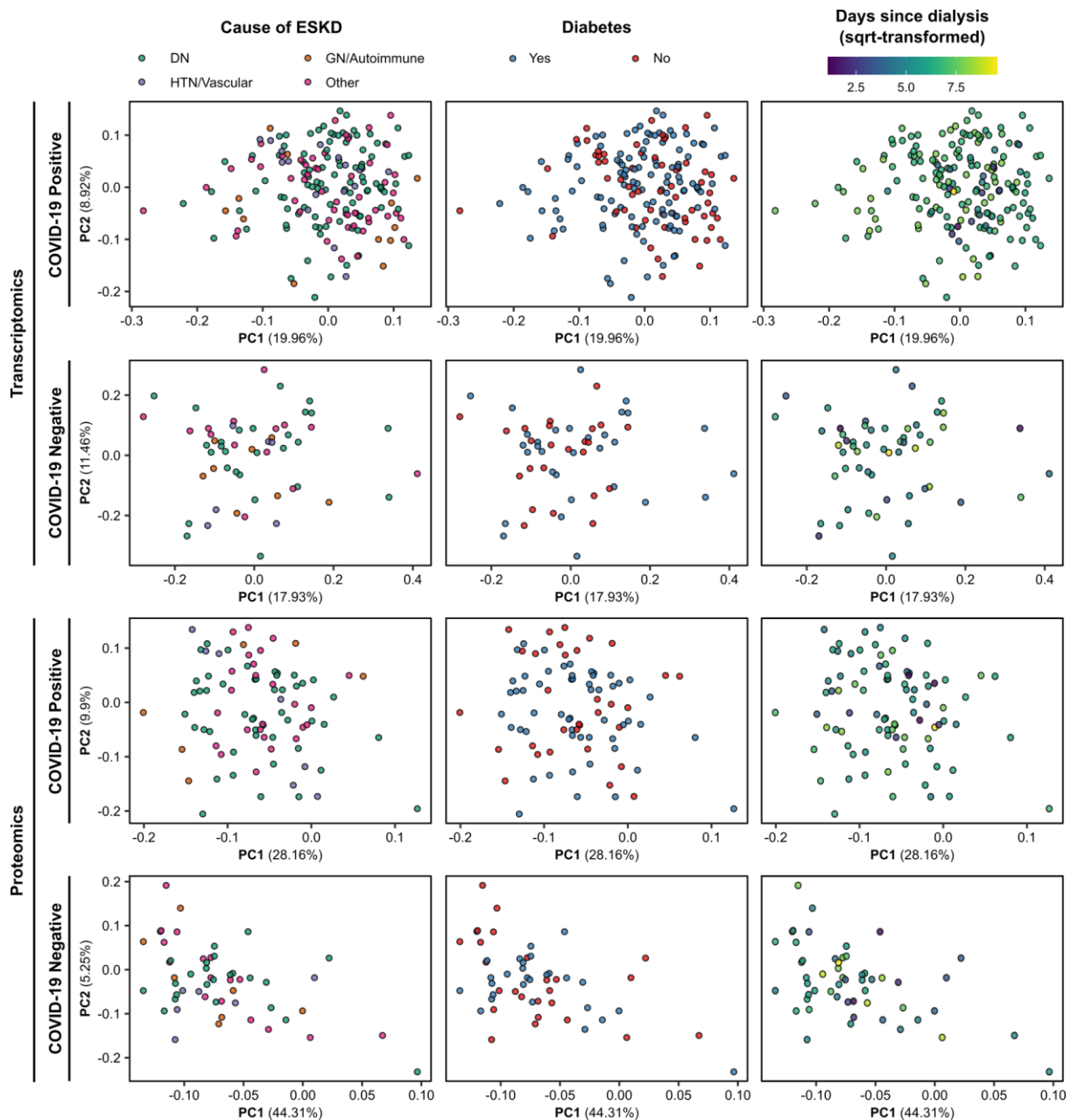


Figure 4.10 The impact of clinical covariates on the transcriptome and proteome

Principal components analysis (PCA) to visualise the impact of clinical covariates on PBMC transcriptome and plasma proteome. PCA plots of the transcriptome and proteome in the Wave 1 cohort, stratified by COVID-19 positive and negative samples. Samples are coloured according to clinical covariates. Left panel: cause of ESKD (DN = diabetic nephropathy, GN = glomerulonephritis, HTN = hypertension). Middle panel: diabetes status (all but one of the diabetic patients had type 2 diabetes). Right panel: time since the first initiation of haemodialysis (measured as the square root of days since the patient first began haemodialysis). The variance explained by PC1 and 2 is indicated in parentheses.

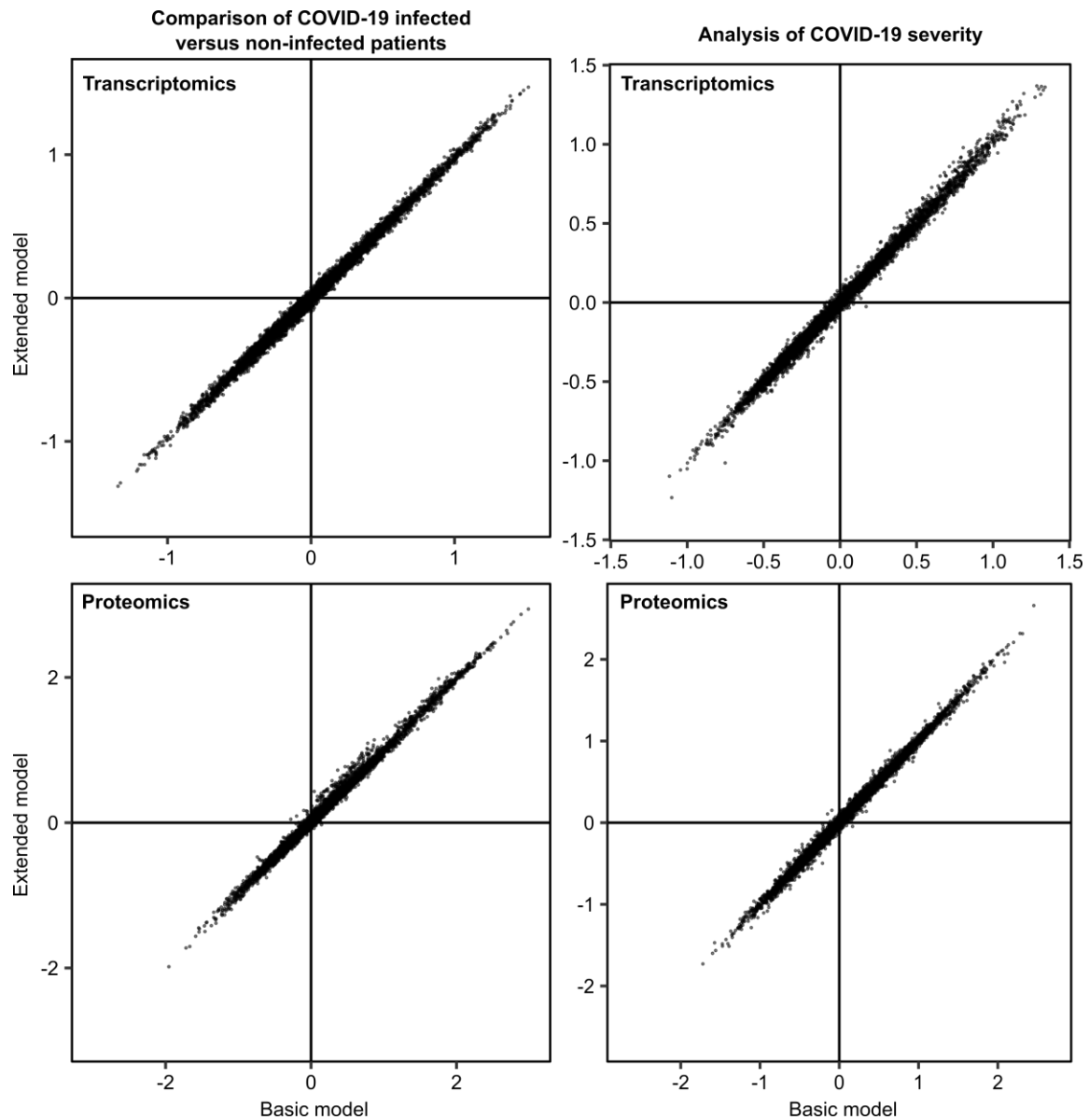


Figure 4.11 The effect of including additional covariates in linear models

The effect of additional covariates on the estimates for each protein and gene analysed in the Wave 1 cohort. Each point represents a molecular feature (genes in upper panels, proteins in lower panels). The x-axis for each panel represents the LMM model estimates (betas) for the molecular feature from the model adjusted for age, sex and ethnicity. The y-axis represents the betas from an extended model with additional covariates representing the cause of ESKD (diabetic nephropathy, glomerulonephritis/autoimmune, hypertension/vascular, or other/unknown), diabetes status, and time since dialysis (measured as the square root of days since the patient began dialysis).

4.4.4 The effect of ESKD and dialysis

Kidney failure and dialysis treatments affect both the plasma proteome (see **Section 2.2.2**) and the PBMC transcriptome (see **Section 4.2.1**). Additionally, I observed marked differences between the proteome (as measured by Olink assays targeting 436 unique proteins) of healthy individuals and the ESKD patients sampled as controls in the Wave 1 cohort (see **Section 2.4.7**). To investigate these differences further, I compared the PBMC transcriptome and the broader plasma proteome (as measured by the SomaScan assays) between ESKD patients on dialysis (with no known current infections) and healthy individuals (**Table 2.2**).

As was the case for the Olink data (**Figure 4.12a**), there was a greater average protein abundance across all SomaScan assays for the ESKD patients (**Figure 4.12b**). This difference was not as stark for the SomaScan compared to that observed in the Olink data, however. This may be because Olink assays are enriched for proteins relating to inflammation and vascular processes, whilst the SomaScan assays cover a much broader range of the proteome. For the COVID-19-related analyses described in this chapter, the SomaScan data were normalised such that each sample is scaled based on a reference cohort. As a result, each sample is constrained to have similar median protein abundance, under the assumption that each sample contains a similar protein content. In this case, considering that there is a different total protein content in ESKD patients on dialysis relative to healthy controls, the assumption of this normalisation step does not hold. Therefore, for the comparison of these groups only, this final normalisation step was not applied.

In contrast, there were fewer total transcripts observed on average for the ESKD patients (**Figure 4.12c**). This is indicative of a systematic difference in sequencing depth between the groups of participants, which could reflect factors that are technical in nature (e.g., differences in the way blood was taken), or those that are biological (e.g., higher cell death in ESKD patients). Most methods for normalising RNA-seq data assume that most genes are not differentially expressed. For instance, the trimmed mean of M values (TMM) method corrects for differences in sequencing depth based on a set of genes that are presumed to be consistent across samples. However, in the present study, it is not clear whether such an assumption holds. In the absence of methods that can correct for sequencing depth without similar assumptions, I performed TMM normalisation²⁹¹ prior to further comparisons between healthy participants and ESKD patients, as has been done for all the RNA-seq-based analyses

described in this chapter. The average expression of genes following normalisation was more balanced between the groups of participants (**Figure 4.12d**). However, when interpreting these data, it is important to consider that the normalisation process may have removed relevant biological signals or introduced spurious factors.

There were clear differences between the patient groups in the scores plot based on the PCA of the PBMC transcriptome, although they could not be completely separated by the first two PCs (**Figure 4.13a**). As was observed for the Olink proteome (**Figure 2.16a**), PCA calculated for the SomaScan proteome revealed complete separation between samples from healthy individuals and ESKD patients on dialysis (**Figure 4.14a**).

I next attempted to identify specific proteins and genes that were dysregulated in ESKD patients on dialysis. First, I applied differential expression analysis to identify 4,137 genes that were upregulated and 3,853 that were downregulated in these patients ($\text{FDR} < 1\%$), when compared to healthy controls (55 ESKD patients vs. 41 healthy controls; **Figure 4.13b**). I observed extremely high log-fold changes ($\log\text{FC}$) for some genes (**Figure 4.13b**), such as *IGFBP*, which had an estimated $\log\text{FC}$ of -11.2 . While this gene was expressed (i.e., one or more reads mapped to the gene) in only 16% of the ESKD patients (median mapped reads = 0), it was expressed in 95% of the healthy individuals (median mapped reads = 179). The most down-regulated genes (i.e., those with a $\log\text{FC} < -5$) had common roles related to cellular adhesion and the extracellular matrix (ECM).

I applied GSVA to identify 1,466 upregulated and 113 downregulated pathways of genes ($\text{FDR} < 1\%$; **Figure 4.13c**). I observed a great deal of overlap between many of the most upregulated pathways. Many of these pathways (e.g., “Proteasome”, “Cellular response to hypoxia” and “Hedgehog ligand biogenesis”, “IRAK1 recruits IKK complex”, “Interleukin 1 signaling”, “Prolonged ERK activation events”, “RAS pathway”, “NFKB is activated and signals survival” and “ECM pathway”) included genes related to the proteasome and IL1 signaling. These collectively included significantly upregulated proteasome genes (e.g., *PSMA1*, *PSMB1*, *PSMC1*, *PSMD1*, *PSME1*, *PSMF1*), ubiquitin (e.g., *UBB*, *UBC*) and related enzymes (e.g., *SEL1L*, *VCP*, *P4HB*, *ADAM*, *PELI1*, *PELI2*), mitogen-activated protein kinases (e.g., *MAPK1*, *MAP2K1*), proteins related to nuclear factor kappa-B (e.g., *NFKB1*, *NFKBIB*, *RELA*) and inhibitor of kappa-B kinase complex (*IRAK1*, *UBE2N*, *TRAF6*, *CHUK*, *IKBKG*), and other

proteins related to interleukin-1 signaling (e.g., *IL1B*, *IL1RAP*, *IL1R2*, *IRAK3*, *TOLLIP*). These pathways may indicate inflammatory signaling by PBMCs in the ESKD patients. There were other highly upregulated immune-related pathways, such as “Regulation of IFNG signaling”, “Signaling by CSF3 G CSF”, “IL10 anti-inflammatory signaling pathway” and “Class 1 MHC mediated antigen processing presentation”. A wide variety of genes within these pathways were significantly upregulated, including but not limited to those encoding ubiquitin- and proteasome-related proteins, signal transducer and activator of transcription-related proteins (e.g., *STAT1*, *STAT3*, *STAT5A*, *PIAS1*), protein tyrosine phosphatase non-receptors (e.g., *PTPN2*, *PTPN6*), tyrosine kinases (e.g., *HCK*, *TYK2*, *SYK*), major histocompatibility complex proteins (e.g., *HLA-A*, *B*, *C*, *E*, *G*), calcium-binding proteins (e.g., *S100A8*, *S100A12*), *IL10*, *JAK2* and *IFNGR2*.

While these results imply inflammatory dysregulation, many of the aforementioned genes and pathways are more specifically involved in the erythropoietin (EPO) receptor signaling cascade. EPOR binds to erythropoietin (EPO) and is involved in the proliferation and survival of erythroid cells. EPO is produced by interstitial fibroblasts in the kidney, so as kidney disease progresses EPO production falls, and consequently patients require exogenous administration of recombinant EPO. Upon binding of EPO, EPOR dimerises, leading to the induction of intracellular signaling pathways in conjunction with the JAK2 kinase. The most well-recognised of these pathways is mediated by STAT5, however additional signaling cascades may be triggered, including the PI3K/Akt and Ras/MAPK pathways, the STAT1 and STAT3 transcription factors, and the LYN tyrosine kinase. Pathways relating to erythropoietin were enriched in the ESKD patients, including “erythropoietin activates phosphoinositide 3 kinase”, “erythropoietin activates RAS” and “signaling by erythropoietin”. Upregulated genes in these pathways included *EPOR*, *JAK2*, *LYN*, phosphoinositide 3-kinases (*PIK3CG*, *PIK3CA*, *PIK3R5*, *PIK3R1*), *GAB1*, *IRS2*, *RAPGEF1*, *KRAS*, *NRAS*, *STAT5A* and *SHC1*. Collectively, these findings imply transcriptomic changes in PBMCs associated with EPO-EPOR signaling.

The pathways “Response of EIF2AK1 HRI to heme deficiency” and “Erythrocytes take up carbon dioxide and release oxygen” were upregulated in ESKD patients; they contained genes involved in haematopoiesis, stress and haem deficiency (e.g., *EIF2AK1*, *CEBP*, *ATF4*, *CA1*, *RHAG*, *HBA2*, *HBB*, *AQP1*). Some of these, such as Rh-Associated Glycoprotein (*RHAG*) which was the 7th most upregulated gene, are erythrocyte-specific.

I observed that many of the most downregulated pathways were related to collagens, the ECM and integrins (e.g., “Collagens”, “miRNA targets in ECM and membrane receptors”, “Basement membranes”, “Matrisome” and “ECM receptor interaction”). These pathways included collagens (e.g., *COL3A1*, *COL1A2*, *COL12A1* and *COL1A1*) and related enzymes (e.g., *PCOLCE*, *LOX*, *DDR1*, *PXDN*), laminins (e.g., *LAMA2*, *LAMA4*, *LAMB1*), integrins (e.g., *ITGA6*, *ITGA11*), insulin-like growth factor-binding proteins (e.g., *CCN1*, *CCN2*, *IGFBP2*, *IGFBP5*), fibulins (e.g., *FBLN1*, *FBLN5*), platelet-derived growth factor subunit B (*PDGFB*) and fibroblast growth factor 2 (*FGF2*). Transcriptional dysregulation of ECM components could indicate disruption of vascular adhesion by PBMCs.

I next applied differential abundance analysis to identify dysregulated proteins (between 53 ESKD patients on dialysis and 41 healthy controls), finding 2,222 upregulated and 80 downregulated (**Figure 4.14b**). The most upregulated proteins included those previously found to be associated with kidney function, such as B2M, TFF3, FABP4 and CST6. Relatively few genes were downregulated, and these included endosomal transmembrane epsin interactor 1 (*ENTREP1*), glutathione peroxidase 3 (*GPX3*), haptoglobin (HP) and platelet factor 4 (PF4). I applied enrichment analysis to find 571 upregulated and 9 downregulated pathways (**Figure 4.14c**). The most upregulated pathway related to ephrin signaling, which included ephrins (e.g., *EFNA1*, *EFNB1*) and ephrin receptors (e.g., *EPHA1*, *EPHB1*, *EPHB4*). The third most upregulated term was “WNT signaling pathway”, which overlapped the pathway “WNT signaling in kidney disease”. Some of the upregulated pathways represented inflammatory processes, such as “MMP cytokine connection” and “Other interleukin signaling”, which collectively involved colony-stimulating factors (*CSF3*, *CSF1R*, *CSF1*), interleukins (*IL34*, *IL16*, *IL10RB*, *IL2RB*, *IL15RA*, *IL15*) and tumour necrosis factor receptor superfamily members (*TNFRSF1A*, *TNFRSF8*, *TNFRSF1B*). “Ghrelin pathway” was also highly upregulated; most of the significantly upregulated members of this pathway were insulin-like growth factor binding proteins (*IGFBP6*, *IGFBP2*, *IGFBP4*, *IGFBP1*).

Some upregulated terms, including “metabolism of fat-soluble vitamins” and “triglyceride metabolism”, consisted of various apolipoproteins (e.g., *APOA1*, *POC2*), glypicans (e.g., *GPC1*, *GPC4*), heparan sulfate proteoglycan (*HSPG2*), fatty acid binding proteins (*FABP1*, 3 and 4), lipoprotein lipase (*LPL*), LDL receptor-related proteins (*LRP10*, *LRP12*). The upregulation of FABPs was also observed in the Olink data (see **Section 2.4.7**); they are

implicated in diabetes²²⁸, obesity and atherosclerosis²²⁹. In contrast, APOM and APOB (the ligand for the low-density lipoprotein receptor) were downregulated. These results point to dysregulation in lipid metabolism, which could contribute to dyslipidaemia and cardiovascular disease²³⁰. Some of the proteins in these pathways also interact with the ECM. Furthermore, there was upregulation of “cell junction organisation”, “integrin A4B1 pathway”, “non-integrin membrane ECM interactions”, “collagens” and “basement membranes. These collectively included collagens (e.g., COL15A1, COL6A2), laminins and related proteins (e.g., LAMC2, NET1), adhesion molecules (e.g., ITGB1, JAM2, VCAM1, SPP1, CADM3, NECTIN2), cadherins (e.g., CDH1, CDH11), nectin cell adhesion molecules (e.g., NECTIN1, NECTIN2), integrins (e.g., ITGA5, ITGB1), urokinase plasminogen activator (PLAUR), vascular cell adhesion protein 1 (VCAM1) and vascular endothelial growth factor A (VEGFA). These findings may reflect the deposition of ECM as a result of renal fibrosis³¹⁰ and vascular damage as a result of haemodialysis.

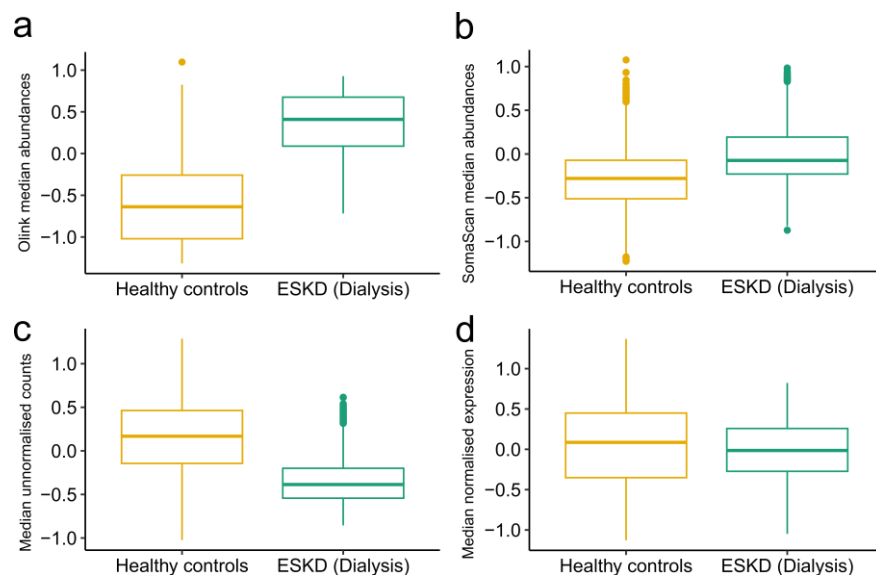


Figure 4.12 Overall abundance and expression for healthy controls and ESKD patients

Each feature (i.e., gene or protein) was first scaled and centred. Medians were then calculated for healthy controls and ESKD patients on dialysis for each feature. These feature-wise medians were then summarised as boxplots. The original features were **a**) Olink normalised protein expression (NPX) **b**) SomaScan log-transformed relative fluorescence units (RFU) **c**) raw RNA-seq counts and **d**) TMM-normalised log-transformed counts per million (logCPM).

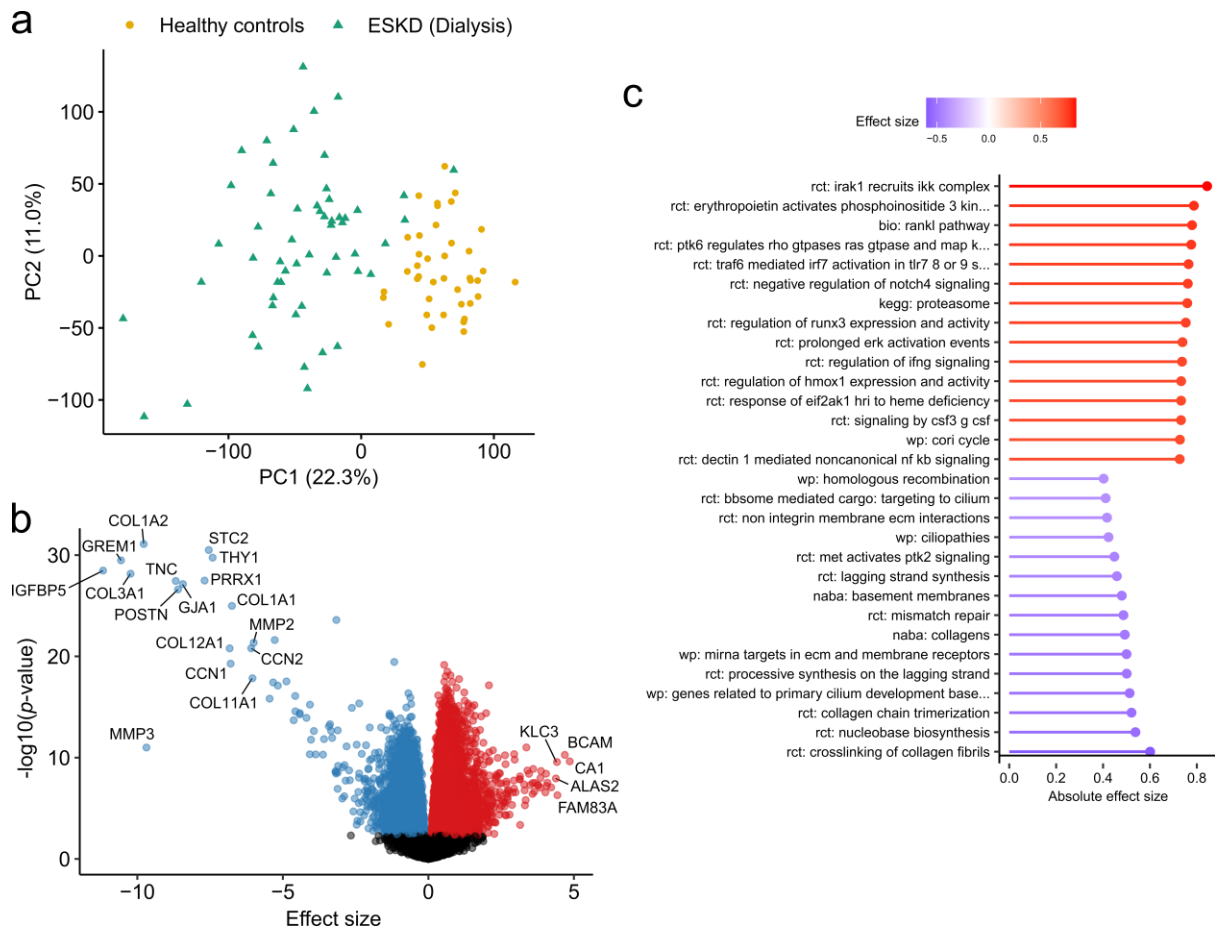


Figure 4.13 The effect of ESKD and dialysis on the PMBC transcriptome

Comparison of non-infected ESKD patients on dialysis and healthy controls. **a**) PCA of the PBMC transcriptome. Each point represents a sample and is coloured by kidney status. The variance explained by PC1 and 2 is indicated in parentheses. **b**) Proteins upregulated (red) or downregulated (blue) in 55 ESKD patients on dialysis relative to 41 healthy individuals (1 sample per patient). Significance was controlled at 1% FDR by Benjamini-Hochberg adjustment; dark-grey = non-significant. **c**) The 15 most upregulated and 15 most downregulated transcriptomic pathway enrichment terms. All terms were significantly enriched at 1% FDR. Terms are ordered and coloured by their effect size. Red = upregulated in ESKD patients on dialysis; blue = downregulated.

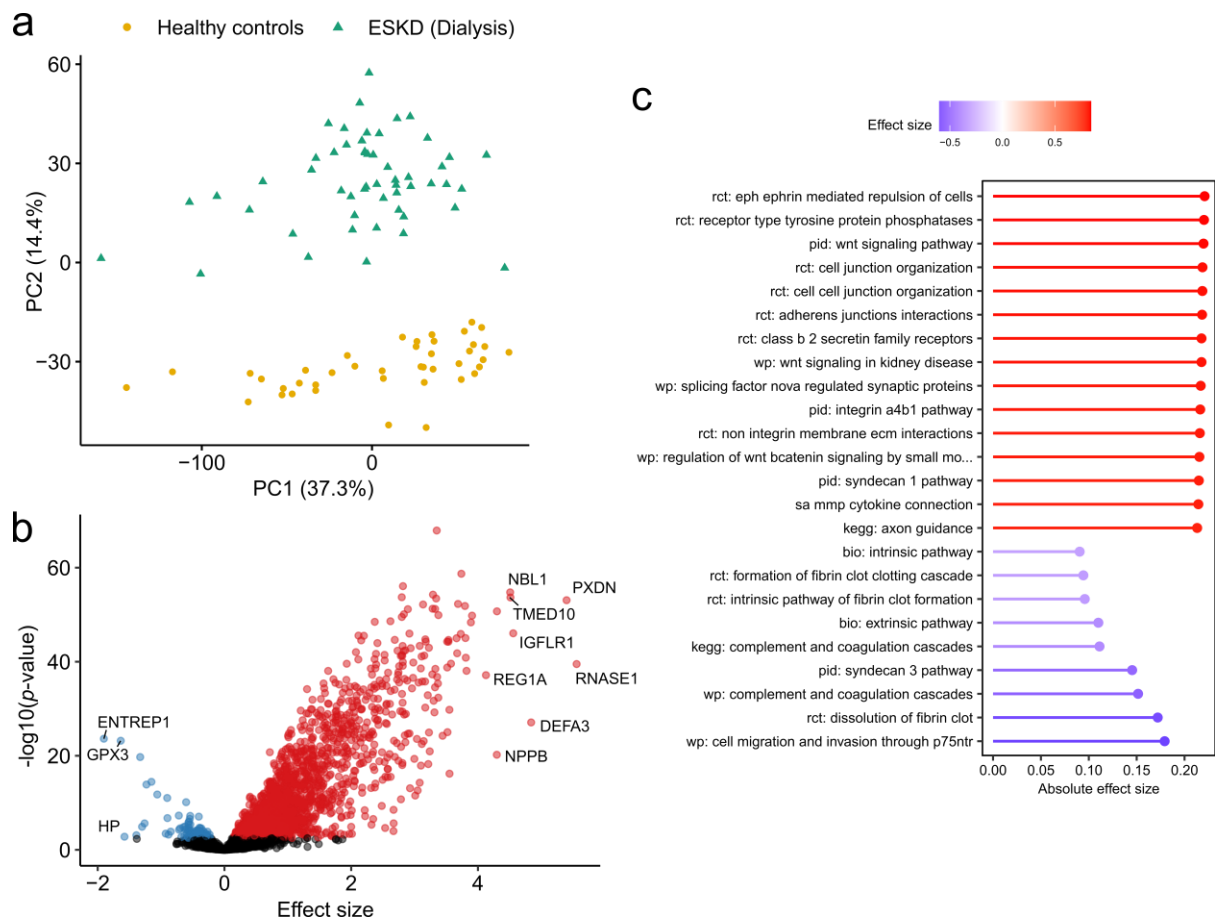


Figure 4.14 The effect of ESKD and dialysis on the plasma proteome

Comparison of non-infected ESKD patients on dialysis and healthy controls. **a)** PCA of the plasma proteome. Each point represents a sample and is coloured by kidney status. The variance explained by PC1 and 2 is indicated in parentheses. **b)** Proteins upregulated (red) or downregulated (blue) in 53 ESKD patients on dialysis relative to 41 healthy individuals (1 sample per patient). Significance was controlled at 1% FDR by Benjamini-Hochberg adjustment; dark-grey = non-significant. **c)** The 15 most upregulated and 15 most downregulated proteomic pathway enrichment terms. All terms were significantly enriched at 1% FDR. Terms are ordered and coloured by their effect size. Red = upregulated in ESKD patients on dialysis; blue = downregulated.

4.4.5 Concordance with studies of COVID-19 for non-ESKD patients

There were marked differences observed in the proteome and transcriptome of ESKD patients on dialysis compared to healthy patients. Therefore, I next sought to investigate whether the COVID-19-related signatures that I observed in this chapter were specific to ESKD patients, or mirrored those identified in other non-ESKD patient cohorts. To this end, I compared the RNA-seq results to the COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium study³⁰⁶, which is also from a United Kingdom healthcare setting. The COMBAT study performed whole blood bulk RNA-seq from 77 individuals with COVID-19 and 10 healthy controls. I re-analysed these data so that the analytical approach was as similar as possible to that used in the present study.

For the differential gene expression analysis between COVID-19 positive and negative samples, I found a high degree of concordance between the effect sizes in the ESKD cohort and the COMBAT study (COMBAT versus Wave 1 $r = 0.68$; COMBAT versus Wave 2 $r = 0.68$) (**Figure 4.15, Figure 4.17**), despite the use of whole blood in the COMBAT study as opposed to the ESKD cohort, which is based on PBMCs. The pathway-level GSVA analysis was similarly consistent (COMBAT versus Wave 1 $r = 0.54$; COMBAT versus Wave 2 $r = 0.66$) (**Figure 4.15, Figure 4.17**). The results for the association of gene expression with COVID-19 severity in the ESKD cohort were also generally consistent with the COMBAT data. For gene-level analysis, Pearson's r was 0.56 and 0.59 for Wave 1 and Wave 2, respectively (**Figure 4.16, Figure 4.17**). For GSVA gene sets, r was 0.44 and 0.37 for Wave 1 and 2, respectively (**Figure 4.16, Figure 4.17**). Given that the COMBAT RNA-seq data were generated on whole blood (versus PBMC in the ESKD cohorts) and considering potential sources of non-biological variation between studies, these levels of inter-study concordance suggest that the transcriptomic signatures related to both COVID-19 (versus negative controls) and to COVID-19 severity are similar in the ESKD and non-ESKD patients.

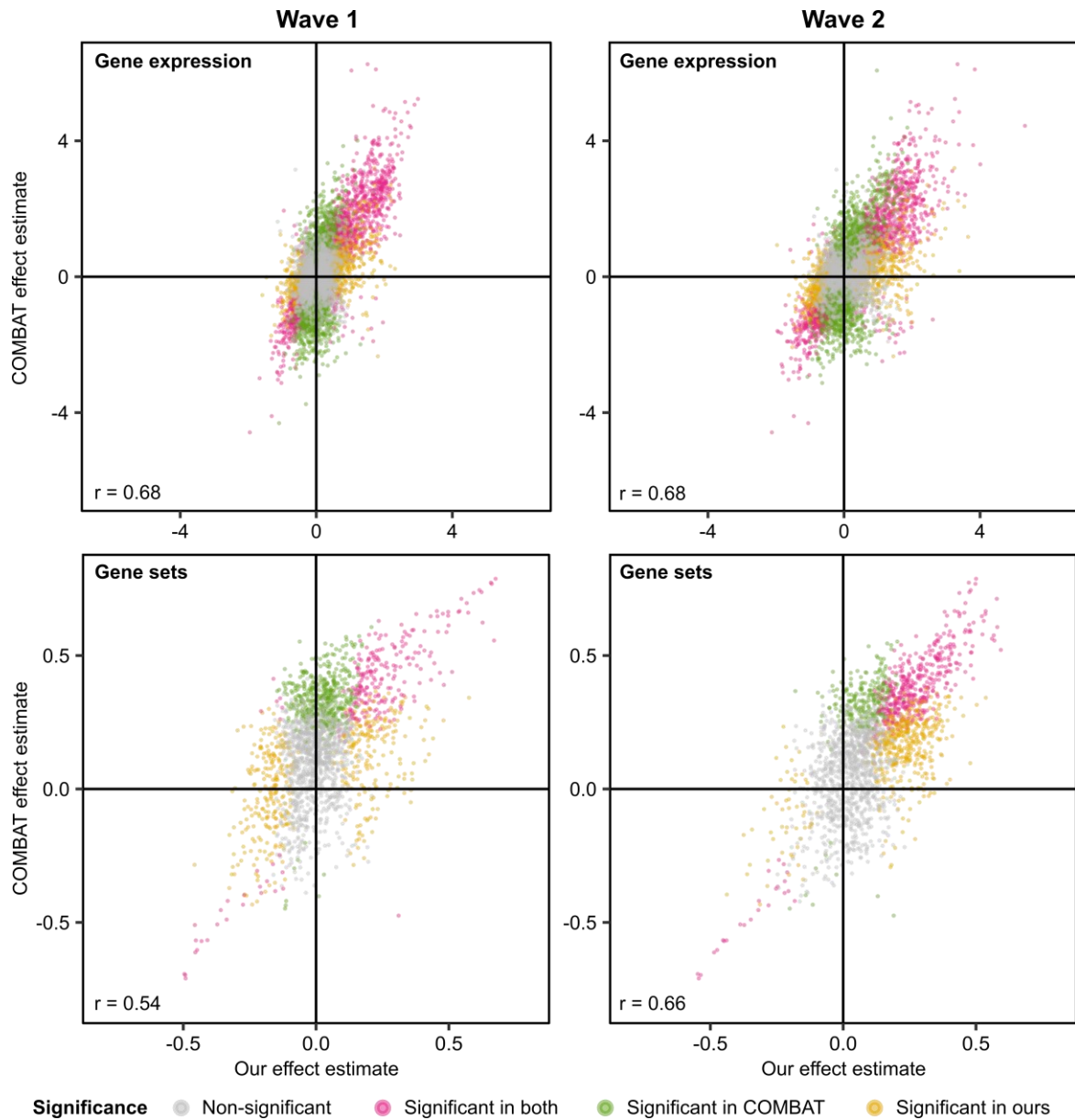


Figure 4.15 Comparison to the effect of COVID-19 in the COMBAT Consortium cohort

The effect estimates for the differential expression of genes between COVID-19 positive vs. negative samples were compared between the cohorts in this study (ESKD patients; “Our effect estimate”) and a re-analysis of the data published by the COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium. Colour represents statistical significance (FDR < 0.01) in each study. Pearson’s correlation (r) is displayed for each comparison.

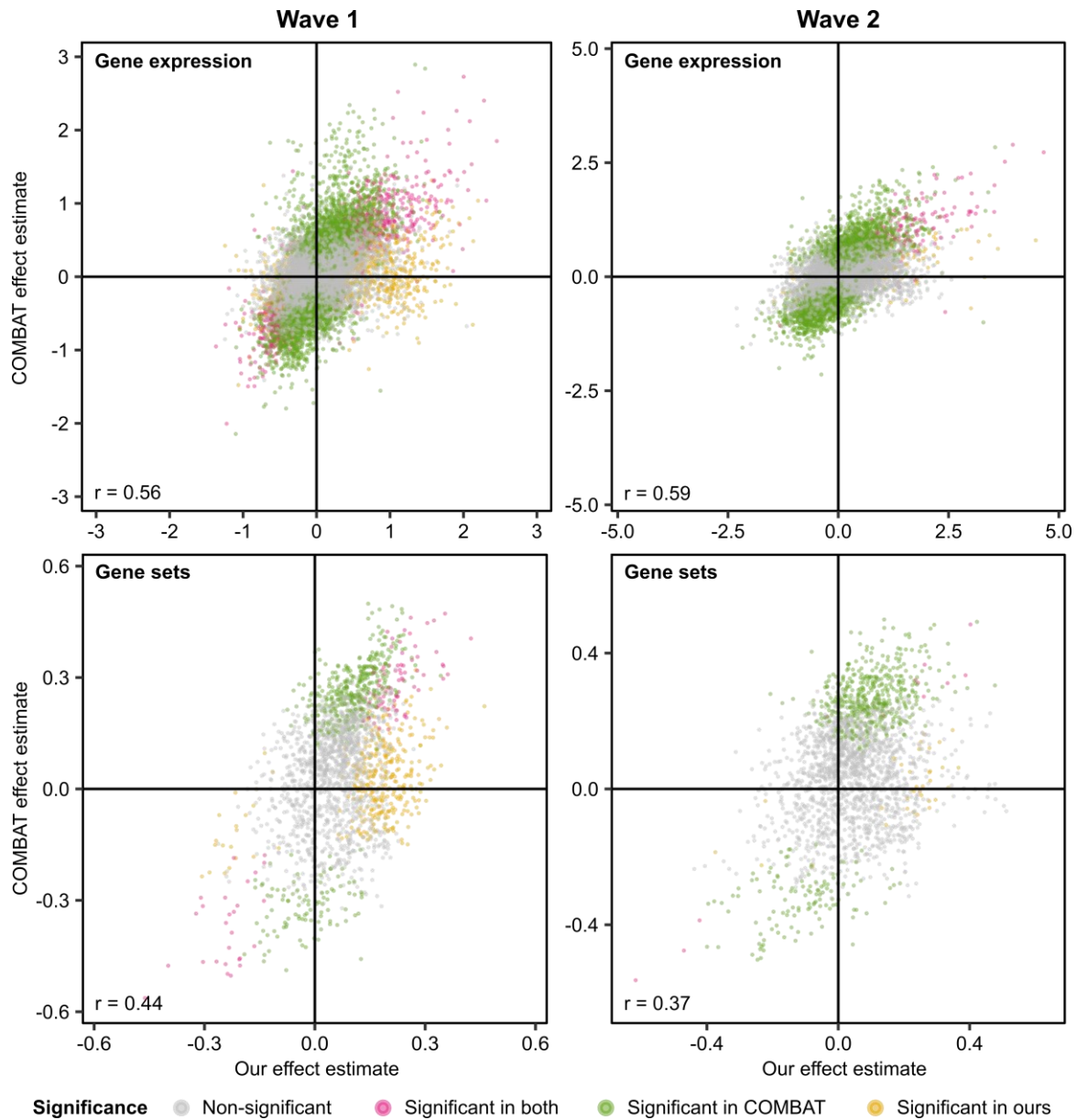


Figure 4.16 Comparison to the effect of COVID-19 severity in the COMBAT Consortium cohort

The effect estimates for the association of genes contemporaneous COVID-19 severity were compared between the cohorts in this study (“Our effect estimate”) and a re-analysis of the data published by the COvid-19 Multi-omics Blood ATlas (COMBAT) Consortium. Colour represents significance ($FDR < 0.01$) in each study. Pearson’s correlation (r) is displayed for each comparison.

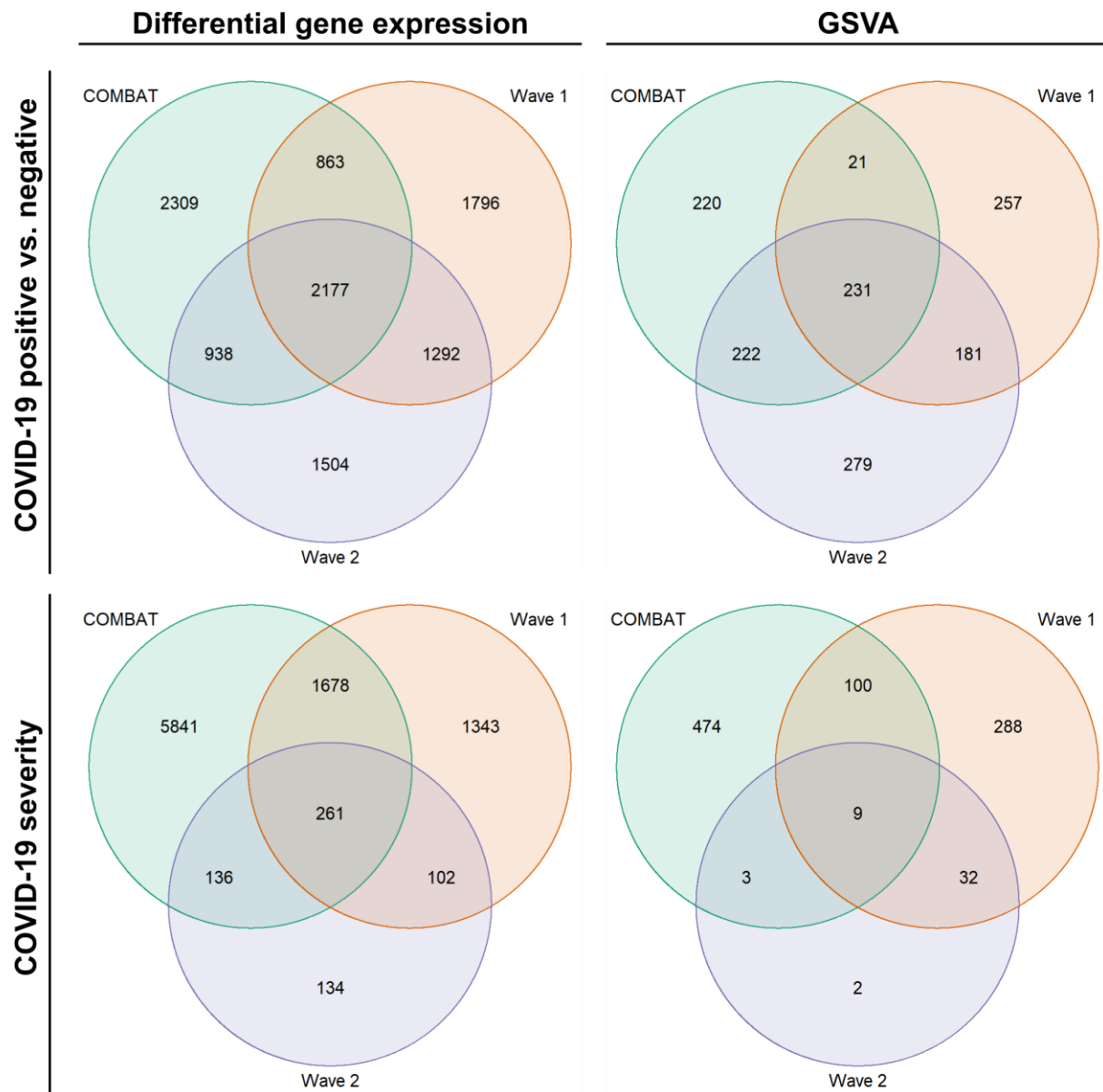


Figure 4.17 Overlap of significant genes and gene sets with data from the COMBAT study

Genes and GSVA gene sets with Benjamini-Hochberg adjusted p -values of less than 0.01 were considered significant in each cohort. The upper panels compare COVID-19 positive versus negative samples. The lower panels compare associations with COVID-19 severity. Statistical testing was performed using LMMs.

I then compared the plasma proteomic data to those of Filbin *et al.*⁷⁵, who also used the SomaScan proteomic platform and collected samples at multiple timepoints during acute COVID-19. The effect correlations between these studies were lower than observed for the cross-study transcriptomic comparisons. The effect estimates for the COVID-19 positive versus negative analyses in the ESKD cohorts and the Filbin data were moderately correlated for the Wave 1 analysis ($r = 0.45$), but less so for Wave 2 ($r = 0.28$) (**Figure 4.18, Figure 4.20**). There was a similar pattern for the GSVA protein pathway analysis (Wave 1 $r = 0.46$; Wave 2 $r = 0.22$) (**Figure 4.18, Figure 4.20**). For the analysis of proteins associated with severity, the correlation of effect estimates with the Filbin data was low for the Wave 1 analysis ($r = 0.20$) but stronger for the Wave 2 analysis ($r = 0.55$). Once again, I observed a similar pattern for the pathway analyses (Wave 1 $r = 0.22$; Wave 2 $r = 0.39$) (**Figure 4.19, Figure 4.20**).

Given the relatively poor concordance between the SomaScan proteomics results for the ESKD cohorts and the data of Filbin *et al.*, I considered whether these differences were due to an ESKD-specific response to COVID-19. Previously (in **Chapter 2**), I described a differential abundance analysis for the comparison proteins, as measured by Olink immunoassays, between COVID-19 cases and controls. The effect sizes for this analysis were correlated ($r = 0.69$) with the effect sizes estimated for a similar comparison performed for the Olink data of Filbin *et al.* (see **Section 2.4.8**). Yet, when such a comparison was made for the same cohorts using the SomaScan assays, there appeared to be greater discordance ($r = 0.45$) (**Figure 4.18**). Therefore, the discordance with Filbin *et al.* appears to be limited to the SomaScan assays. This may suggest that the low correlation between the findings in the ESKD cohorts and the data of Filbin *et al.* are not primarily driven by ESKD. On the other hand, kidney failure may have a greater impact on the proteins measured by the SomaScan assays. The effect sizes estimated for the comparison of cases and controls were also relatively weakly correlated between the Wave 1 and Wave 2 cohorts for the proteome ($r = 0.57$) compared to the transcriptome ($r = 0.80$) (**Figure 4.21**). This lends support to the hypothesis that factors other than ESKD explain the poor concordance between the proteomic findings and the data of Filbin *et al.*

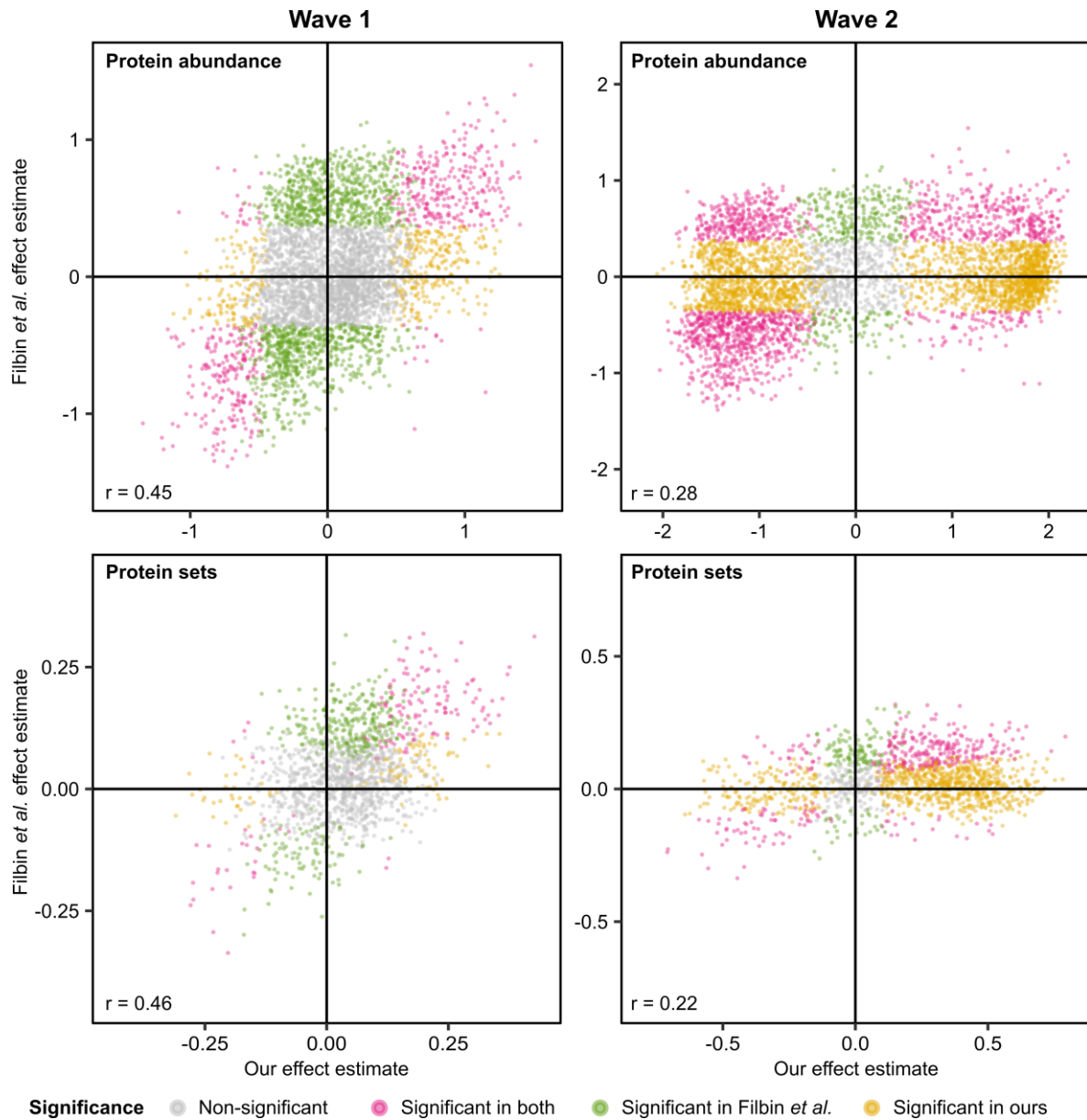


Figure 4.18 Comparison of the effect of COVID-19 in the data of Filbin *et al.*

The effect estimates for the differential abundance of proteins between COVID-19 positive vs. negative patients were compared between the cohorts in this study ("Our effect estimate") and a re-analysis of the data published by Filbin *et al.* Colour represents significance (FDR < 0.01) in each study. Pearson's correlation (r) is displayed for each comparison.

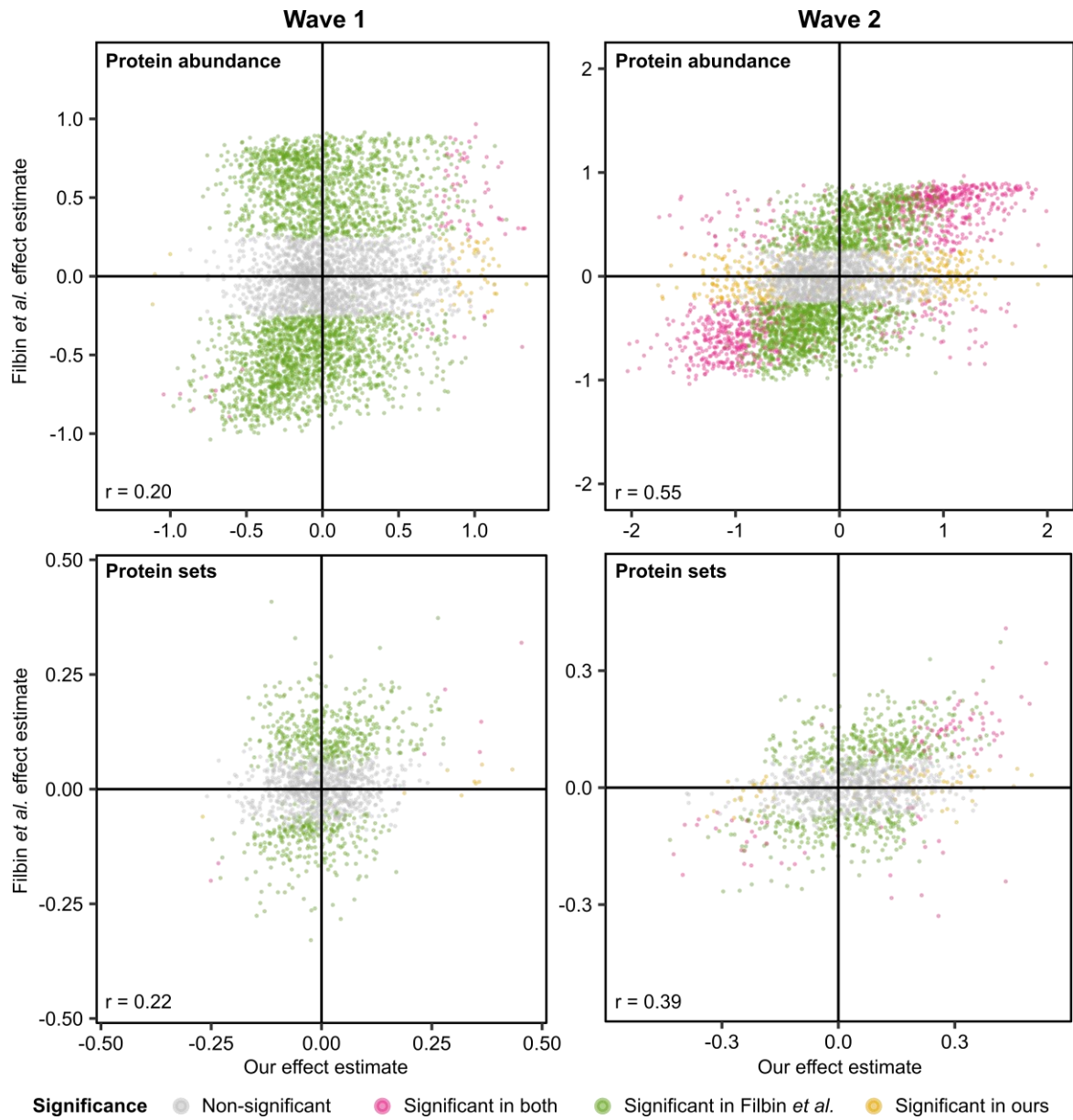


Figure 4.19 Comparison of the effect of COVID-19 severity in the data of Filbin *et al.*

The effect estimates for the association of proteins with COVID-19 severity were compared between the cohorts in this study ("Our effect estimate") and a re-analysis of the data published by Filbin *et al.* Colour represents significance (FDR < 0.01) in each study. Pearson's correlation (r) is displayed for each comparison.

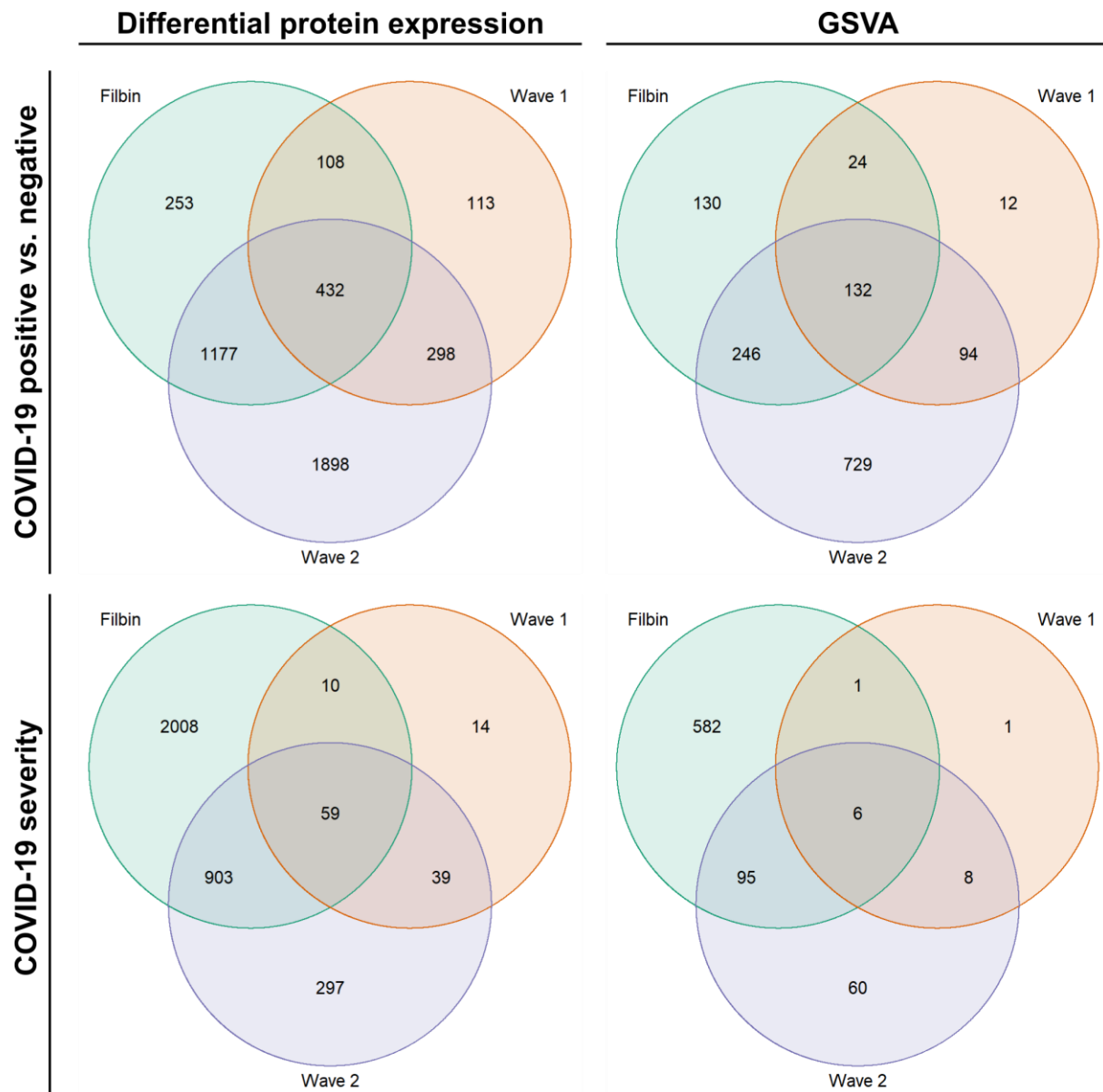


Figure 4.20 Overlap of significant proteins and protein sets with data of Filbin *et al.*

Proteins and GSVA protein sets with Benjamini-Hochberg adjusted p -values of less than 0.01 were considered significant in each cohort. The upper panels compare COVID-19 positive versus negative samples. The lower panels compare associations with COVID-19 severity. Statistical testing was performed using LMMs.

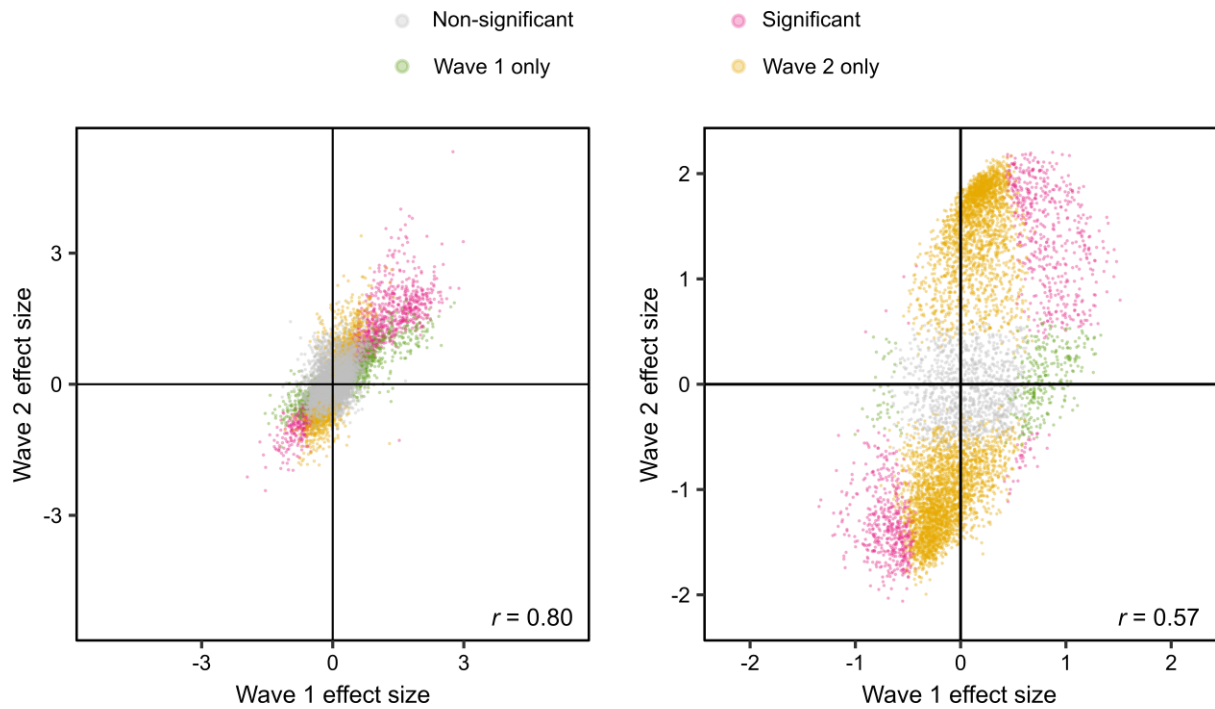


Figure 4.21 Comparing the effect of COVID-19 in each cohort

Comparison of effect sizes (coefficients from LMM) for the Wave 1 and Wave 2 COVID-19 positive versus negative differential expression and abundance analyses. Each point is a gene or protein, coloured according to their significance in the Wave 1 and 2 analyses. r = Pearson's correlation coefficient. **Left**) Genes. **Right**) proteins.

4.4.6 Severe COVID-19 is associated with dynamic molecular trajectories

I next examined the temporal trajectories of the transcriptome and the proteome during COVID-19 by explicitly modelling molecular profiles with respect to time following symptom onset (see **Section 4.3.12**). To aid biological interpretation, I first applied a dimension reduction strategy using weighted gene correlation network analysis (WGCNA)⁹⁶. WGCNA identified 23 modules of co-expressed genes (which are denoted with the prefix t), and 12 proteomic modules (denoted with p). Longitudinal modelling revealed 8 transcriptomic and 5 proteomic modules with significantly (5% FDR) different temporal patterns in patients with mild/moderate versus severe/critical disease (LMM time x clinical course (TxCC) interaction; **Section 4.3.12; Table 4.4, Table 4.5**). Typically, the modules displayed a flat temporal profile in mild/moderate COVID-19, whereas there was a dynamic profile in severe/critical disease (**Figure 4.22a**). Some modules rose with time in severe/critical patients (e.g., tB, tL, p9 and p12), whilst others dropped (e.g., tC, tP, tI, p7). Examples of individual genes from module tB exhibiting this behaviour include *MMP9*, *ORM1* and *LRRN1* (**Figure 4.22b**).

I identified significant associations between modules, with the transcriptomic and proteomic modules clustering into larger positively or negatively correlated groupings (**Figure 4.22c**). The inter-modular associations appeared to strongly reflect the association with COVID-19 severity at time of sampling (**Table 4.4, Table 4.5**), implying that this is a strong underlying factor in the -omics data. Consistent with this, integrated analysis of the transcriptomic and proteomic datasets using MEFISTO³¹¹ revealed a single factor that had a significantly different trajectory in severe/critical versus mild/moderate disease ($p = 5.4 \times 10^{-12}$, LMM TxCC) (**Figure 4.23**).

I characterised the modules by pathway analysis (**Figure 4.22a, Table 4.4, Table 4.5**). I also investigated whether disease trajectory-associated transcriptomic modules might reflect a shift in cell-type proportions, estimated using the CIBERSORTx algorithm (see **Section 4.3.13; Figure 4.24**). The severity-associated modules tB and tJ were both strongly positively associated with imputed neutrophil abundances, and negatively associated with lymphocyte subsets. The presence of a neutrophilic gene signature in the PBMC preparations may indicate the presence of low-density granulocytes. Consistent with this, hub genes in Module tB (including *TECPR2*, *CSF3R*, *STX3*; **Figure 4.22b, Table 4.4**) are associated with granulocytes and autophagy, and pathway analysis of the module genes revealed enrichment

for pathways including “Neutrophil degranulation” and “ROS and RNS production in phagocytes. Module tB also contains genes encoding calcium-binding proteins (e.g., *S100A6*, *S100A9*, *S100A11*, *S100A12*) that play important roles in regulating inflammatory pathways³¹², as well as histones (e.g., *H2BC4*), integrins (e.g., *ITGA1*, *ITGAM*, *ITGB4*, *ITGAX*, *ITGAD*), adhesion molecules (e.g., *CEACAM1*, *CEACAM3*, *CEACAM4*, *ICAM3*), oncostatin M (*OSM*) and colony-stimulating factor 1 (*CSF1*). The tL module, which also displayed a rising trajectory in worse disease, was strongly positively associated with imputed plasma cell proportion (**Figure 4.22a**, **Table 4.4**, **Figure 4.24**) and many of its members encoded immunoglobulins. In contrast to tB, tJ and tL, the other transcriptomic modules (tP, tC, tF, tI, tN) all displayed a decreasing trajectory in patients with worse disease (**Figure 4.22a**). These transcriptomic modules tended to be positively associated with imputed lymphocyte subset proportions (**Figure 4.24**). This pattern is consistent with the observation that lymphopenia is a key feature of severe COVID-19^{313,314}.

Three severity-associated proteomic modules were positively associated with both COVID-19 severity and transcriptomic modules tB, tJ and tL; these were p8 and p9 (both enriched for pathways related to RNA splicing), and p12 (significantly enriched for the pathway HDACs deacetylate histones) (**Figure 4.25**, **Table 4.5**). The latter is consistent with previous analyses, in that a histone pathway signature was prominently associated with COVID-19 severity in both the RNA-seq (**Figure 4.8b, c**) and plasma proteomic data (**Figure 4.9b, c**). The remaining proteomic modules, p1 and p7, did not have clear functional enrichments, although they both contained immune-related proteins.

Table 4.4 Transcriptomic modules associated with disease trajectory

Transcriptomic modules associated with disease trajectory (i.e., with a significant TxCC interaction) are tabulated and assigned names that are representative of their members. Size = the number of genes assigned to the module. Primary cell type indicates the CIBERSORTx cell type with the greatest positive correlation with the module's eigengene. Severity association indicates whether the module's eigengene is positively (↑) or negatively (↓) associated with contemporaneous WHO severity (5% FDR, LMM).

* Module tJ was positively correlated with severity, but this association was not significant at 5% FDR.

R = Reactome. WP = WikiPathways. PID = Pathway Interaction Database.

Module code	Module name	Size	Selected enrichments	Primary cell type	Hub genes	Severity association
tB	Granulocyte cell-like module 1	1,309	R: Neutrophil degranulation R: Innate immune system R: ROS and RNS production in phagocytes R: Oxidative stress-induced senescence	Neutrophils	<i>TECPR2, CSF3R, STX3, MMP25, BASP1, MBOAT7, NCF4, GLT1D1, RNF24, DHX34</i>	↑
tJ	Granulocyte cell-like module 2	169	R: Neutrophil degranulation R: Innate immune system R: Antimicrobial peptides NABA: Matrisome	Macrophages (M0)	<i>CEACAM8, BPI, CD24, CEACAM6, ABCA13, DEFA4, LTF, AZU1, ELANE, LCN2</i>	↑*
tL	Plasma cell module	655	R: Cell cycle R: MHC class II antigen presentation WP: DNA damage response R: Factors involved in megakaryocyte development and platelet production	Plasma cells	<i>RRM2, FOXM1, MKI67, TPX2, BUB1, KIFC1, TK1, MZB1, CDK1, CCNA2</i>	↑
tP	Nuclear and cell cycle module	464	None significant	B cells naive	<i>SNRPA1, CEP95, THUMPD2, CENPC, ZNF326, NDUFAF5, ILKAP, SUPV3L1, USP36, GOLT1B</i>	↓
tC	T-cell activity module 1	545	KEGG: Natural killer cell-mediated cytotoxicity R: Immunoregulatory interactions between a lymphoid and a non-lymphoid cell PID: CD8 TCR downstream pathway	NK cells resting	<i>SAMD3, ADGRG1, ZAP70, PYHIN1, FCRL6, PRSS23, FGFBP2, NFATC2, PTCH1, LLGL2</i>	↓
tF	T-cell activity module 2	761	WP: T-cell receptor and co-stimulatory signaling	T cells CD4 memory resting	<i>PLCG1, LCK, UBASH3A, ABCD2, LINC00649, PRKCQ-AS1, TC2N, LINC01550, SEPTIN1, NLRC3</i>	↓
tI	Monocyte module	299	R: PD1 signaling KEGG: Viral myocarditis	Monocytes	<i>PLXNB2, NAAA, CSF1R, PSAP, CARD9, SLC7A7, PEA15, ARHGEF10L,</i>	↓

			KEGG: Antigen processing and presentation KEGG: Cell adhesion molecules R: Interferon-gamma signaling		<i>ZNF385A, ARRB1</i>	
tN	Allergy-related module	54	WP: IL-3 signaling pathway KEGG: Asthma	T cells CD4 memory resting	<i>HDC, LINC02458, CPA3, GATA2, AKAP12, MS4A2, ENPP3, FCER1A, TRIM51EP, SLC45A3</i>	↓

Table 4.5 Proteomic modules associated with disease trajectory

Proteomic modules associated with disease trajectory (i.e., with a significant TxCC interaction) are tabulated and assigned names that are representative of their members. Size = the number of proteins assigned to the module. Severity association indicates whether the module's eigenprotein is positively (↑) or negatively (↓) associated with contemporaneous WHO severity (5% FDR, LMM).

R = Reactome. WP = WikiPathways. PID = Pathway Interaction Database.

Module code	Module name	Size	Selected enrichments	Hub genes	Severity association
p12	Histone-associated module	40	R: HDACs deacetylate histones KEGG: Systemic lupus erythematosus R: Chromatin modifying enzymes R: HCMV late events	H2AC1, H2AW, H2BU1, H2BC21, H2BC12, H2AZ1, CELF2, H2AC11, EEF1B2, MMP17	↑
p9	Splicing and nuclear module 1	241	PID: FRA pathway WP: Striated muscle contraction R: Complement cascade KEGG: Spliceosome	CLSTN3, KHSRP, OIT3, NELFA, EWSR1, CETP, PUF60, SRSF6, NXT1, ARHGAP36	↑
p8	Splicing and nuclear module 2	107	R: Processing of capped intron-containing pre-mRNA R: mRNA splicing KEGG: Spliceosome	CILP, FUBP1, TSSC4, ALKAL2, STMN3, STMN1, STMN2, UBE2Z, MAFG, LEMD1	↑
p1	Mixed immune module 1	281	WP: Development of pulmonary dendritic cells and macrophage subsets	MRPL52, PLA2G2C, ITGAL, NLRP4, GDF3, VSTM4, ITGB1, SVBP, ANXA8, UBD	↑
p7	Mixed immune module 2	66	None significant	DEFB135, SPINK14, VAT1L, NT5E, PAXIP1, ZHX3, ODC1, PLEKHM2, DEFB112, F3	↓

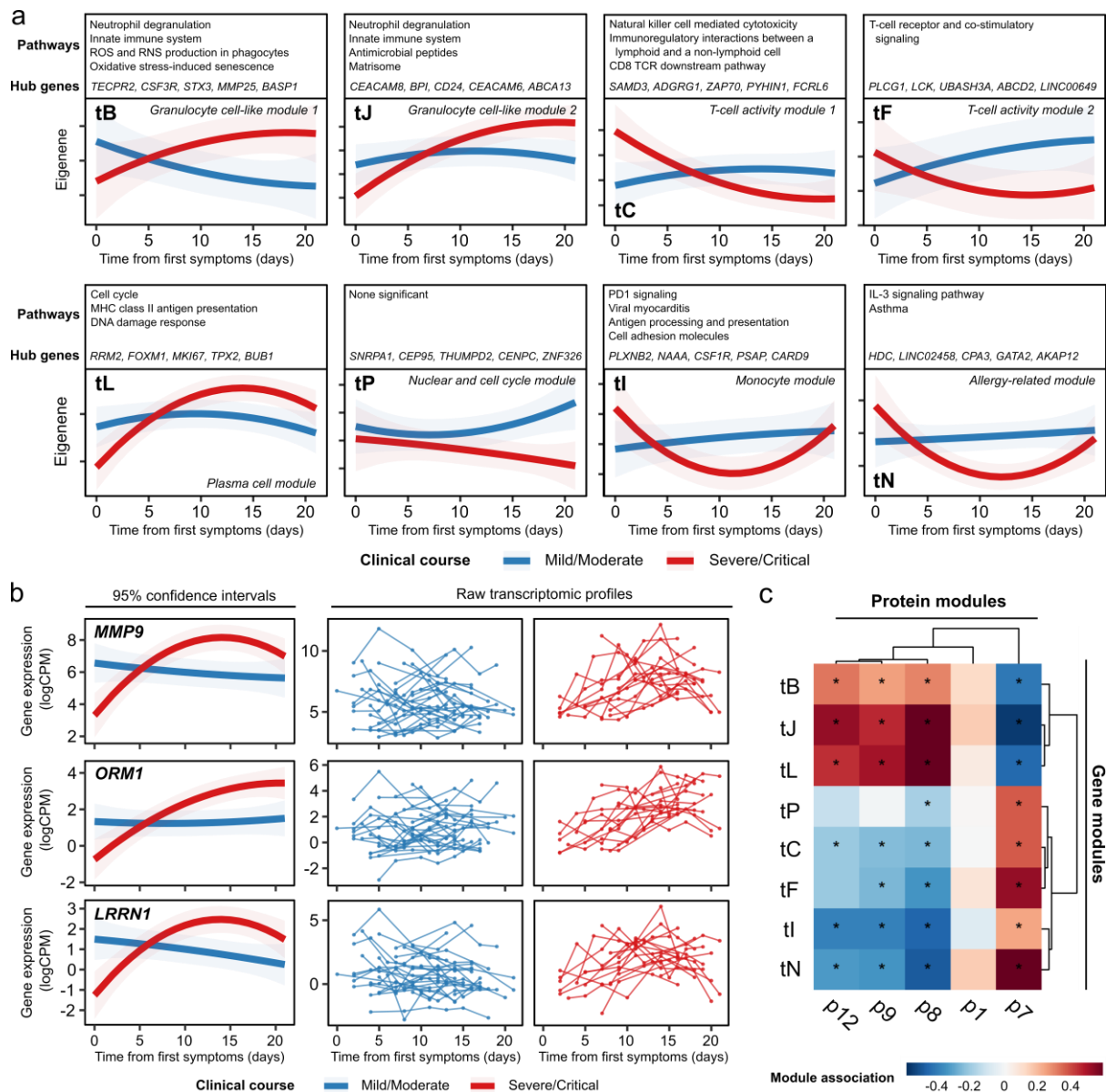


Figure 4.22 Longitudinal profiles of transcriptomic modules

a) The longitudinal profiles of significant (TxCC, LMM, FDR < 0.05) gene modules, stratified by clinical course. Lines represent estimated marginal means and shaded areas represent their 95% confidence intervals. **b)** Modelled longitudinal profiles of the three genes within module B with the most significant TxCC interaction effects (LMM). Left: lines represent estimated marginal means and shaded areas represent their 95% confidence intervals. Right: individual-level data ($n = 169$ samples from 40 individuals). **c)** Heatmap displaying associations (LMM) between transcriptomic and proteomic modules (right). Red = positive correlation, blue = negative correlation. Significant associations (5% FDR) are represented by an asterisk.

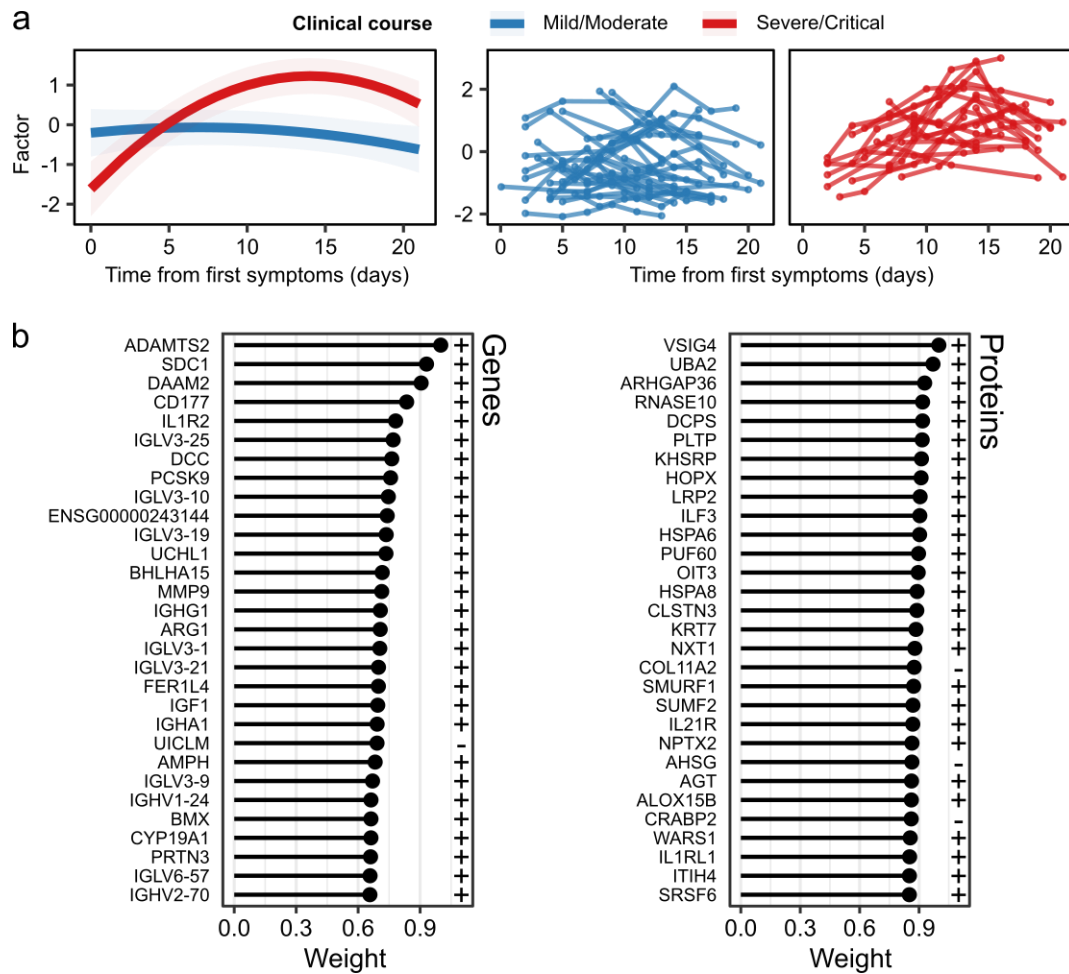


Figure 4.23 A multi-modal factor representing COVID-19 severity

a) The longitudinal profile of a factor identified by MEFISTO with a significant (5% FDR; LMM) TxCC interaction. Lines represent estimated marginal means and shaded areas represent their 95% confidence intervals (left) and the trajectory of the factor for each individual (right). **b)** The weights of individual genes and proteins with respect to the factor identified by MEFISTO. A plus sign indicates molecules positively associated with the factor while molecules negatively associated with the factor have a minus sign.

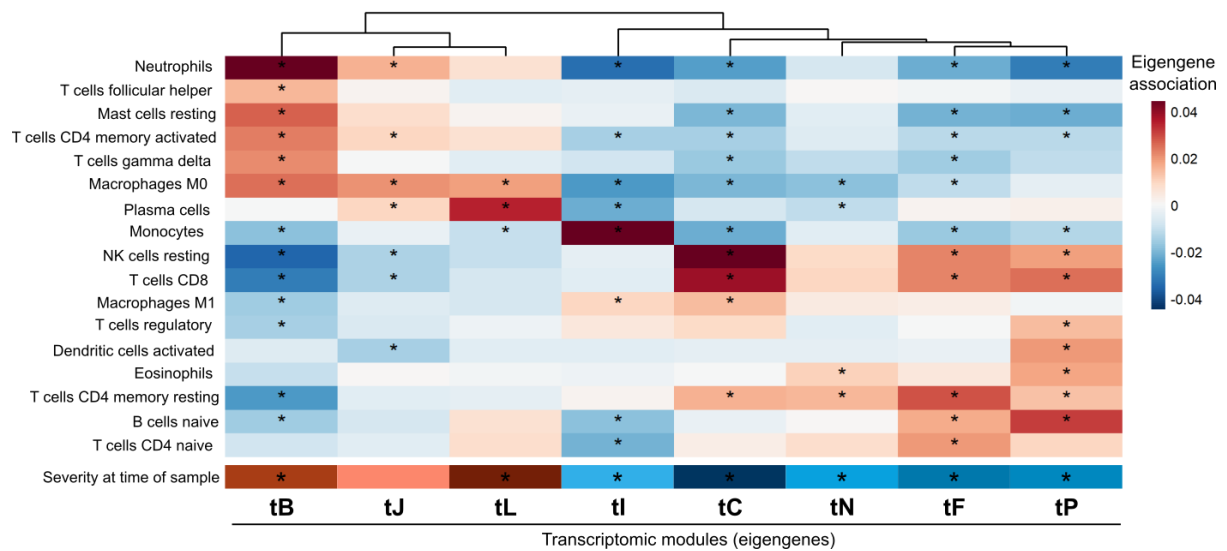


Figure 4.24 Association of transcriptomic modules with imputed cell abundances

LMM association of modules (columns) with cell abundances (rows) imputed using CIBERSORTx. All modules displayed have significant TxCC interaction effects (5% FDR, LMM). The bottom row shows the LMM association with the severity score at time of sampling. Asterisks represent significant (5% FDR) associations.

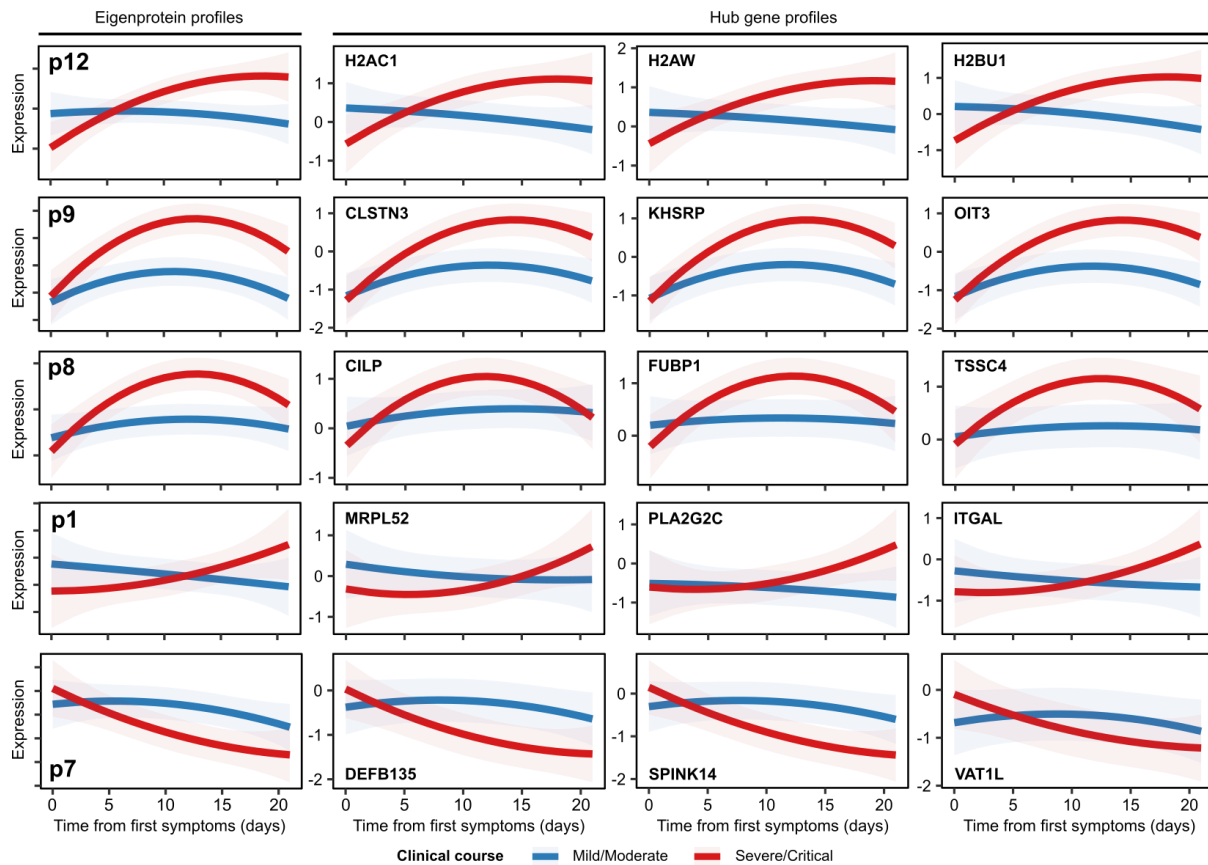


Figure 4.25 Longitudinal profiles of protein modules and hub proteins

Modelled longitudinal profiles of protein modules with significant (5% FDR; LMM) TxCC interactions (left) and the profiles of their top three most central hub proteins (right). Lines represent LMM estimated marginal means and shaded areas represent their 95% confidence intervals. Red = patients with a severe/critical clinical course; blue = mild/moderate clinical course.

4.4.7 The temporal dynamics of cytokines in severe COVID-19

Many of the plasma proteins identified in the differential abundance and modular analyses are primarily known to have intracellular roles. For instance, the protein modules p8 and p9 were enriched for pathways related to splicing and transcriptional regulation (**Table 4.5**). The Human Protein Atlas annotations for the 6,323 unique proteins measured by the SomaScan panel demonstrated that, while there were almost 5,000 proteins annotated as being intracellular, only around 1000 were annotated as secreted (**Figure 4.1a**). The elevation of these proteins in severe disease could simply be a readout of increased cell turnover, death, stress, and viral hijacking of host cellular machinery. This motivated a more focused analysis of proteins whose primary biological role is to act extra-cellularly (e.g., cytokines, chemokines, growth factors and their receptors). These classes of proteins are important therapeutic targets in inflammatory diseases³¹⁵. Specifically, I modelled the temporal profiles of 232 proteins that made up the KEGG pathway “Cytokine-cytokine receptor interactions”. Fifty proteins had significantly different profiles in patients with a severe/critical clinical course versus those with mild/moderate ones (TxCC interaction effect, 5% FDR). Proteins exhibited distinct patterns of divergence between severe/critical and mild/moderate disease over time (**Figure 4.26**). Some (e.g., IL1b, IL6, IL15RA, CCL2) showed a relatively stable temporal profile in mild/moderate patients but rising trajectories in severe/critical patients (**Figure 4.27b**). Others (e.g., CCL15, TNFSF13B (BAFF), PDGFRB, EDAR, IFNA10, IFNA13, IFNA16, IFNE and IFNL3) were elevated early in the disease course and decreased over time, but displayed more marked initial elevations in severe/critical patients (**Figure 4.27c**). CD40LG, TNFSF10 (TRAIL) and IL11 were reduced in the severe/critical versus the mild/moderate group at early timepoints but increased in severe/critical patients later (**Figure 4.27d**). Conversely, leptin, INHBA (inhibin A), and CCL22 were initially higher in severe/critical than mild/moderate patients but with the reverse pattern later on (**Figure 4.27e**). These data demonstrate the dynamic nature of soluble proteins in severe COVID-19. Studies without dense serial sampling are limited to a single snapshot, which may be an oversimplification of the response to viral infection.

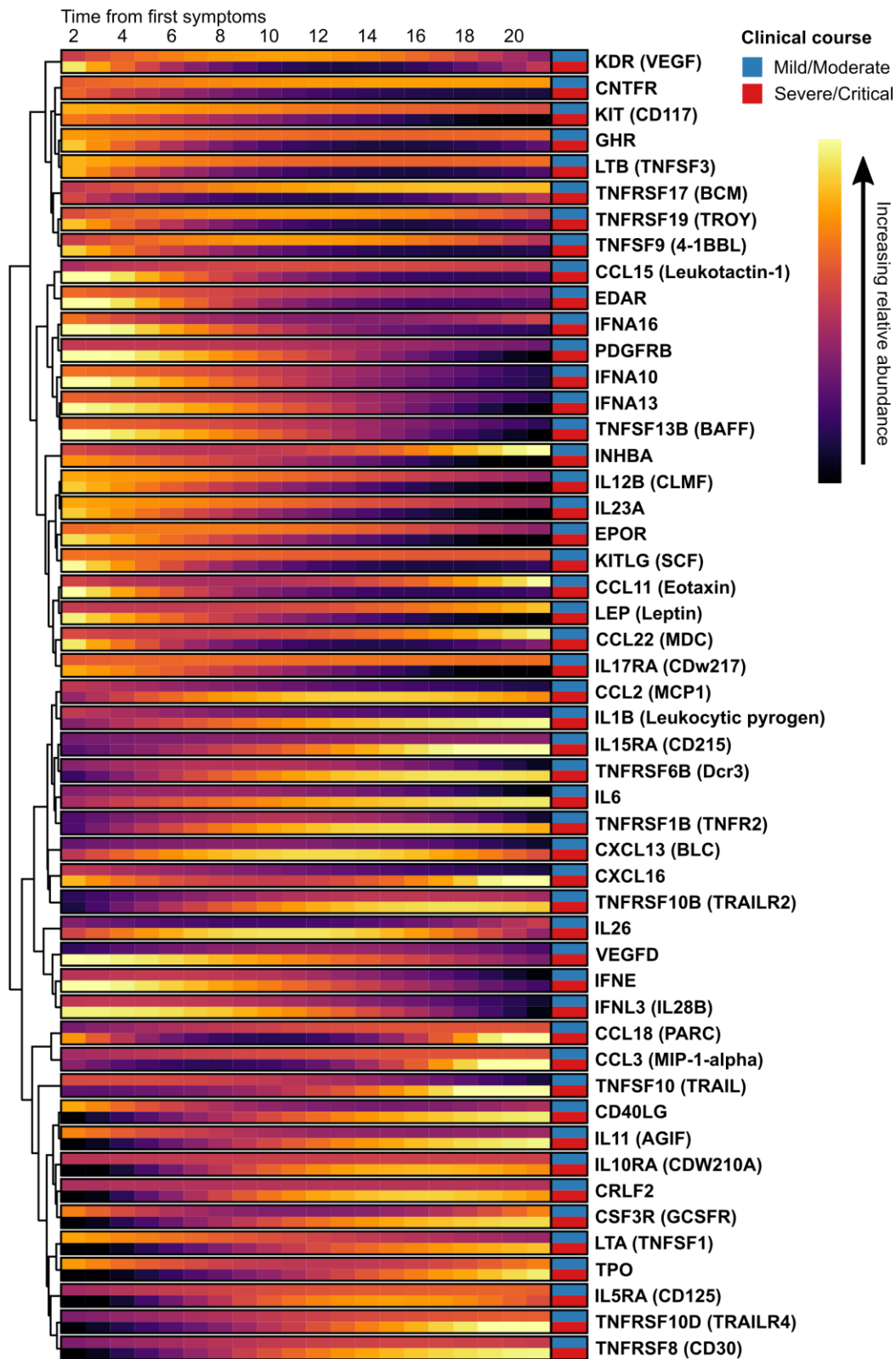


Figure 4.26 Cytokines with different temporal profiles depending on COVID-19 severity

Dynamic temporal changes in circulating cytokines and receptors vary between severe and mild COVID-19. The heatmap displays proteins with a significantly different temporal profile in mild vs severe disease (TxCC, LMM, FDR < 0.05). The colour indicates LMM estimated marginal means over time, stratified by patient group ($n = 169$ samples from 40 individuals). In other words, the profile of each protein is visualised over time separately for the mild/moderate and severe/critical groups. Proteins are clustered based on the temporal profile of the discordance

between mild/moderate and severe/critical disease. Proteins are annotated using gene symbols, with alternative common protein identifiers in parentheses.

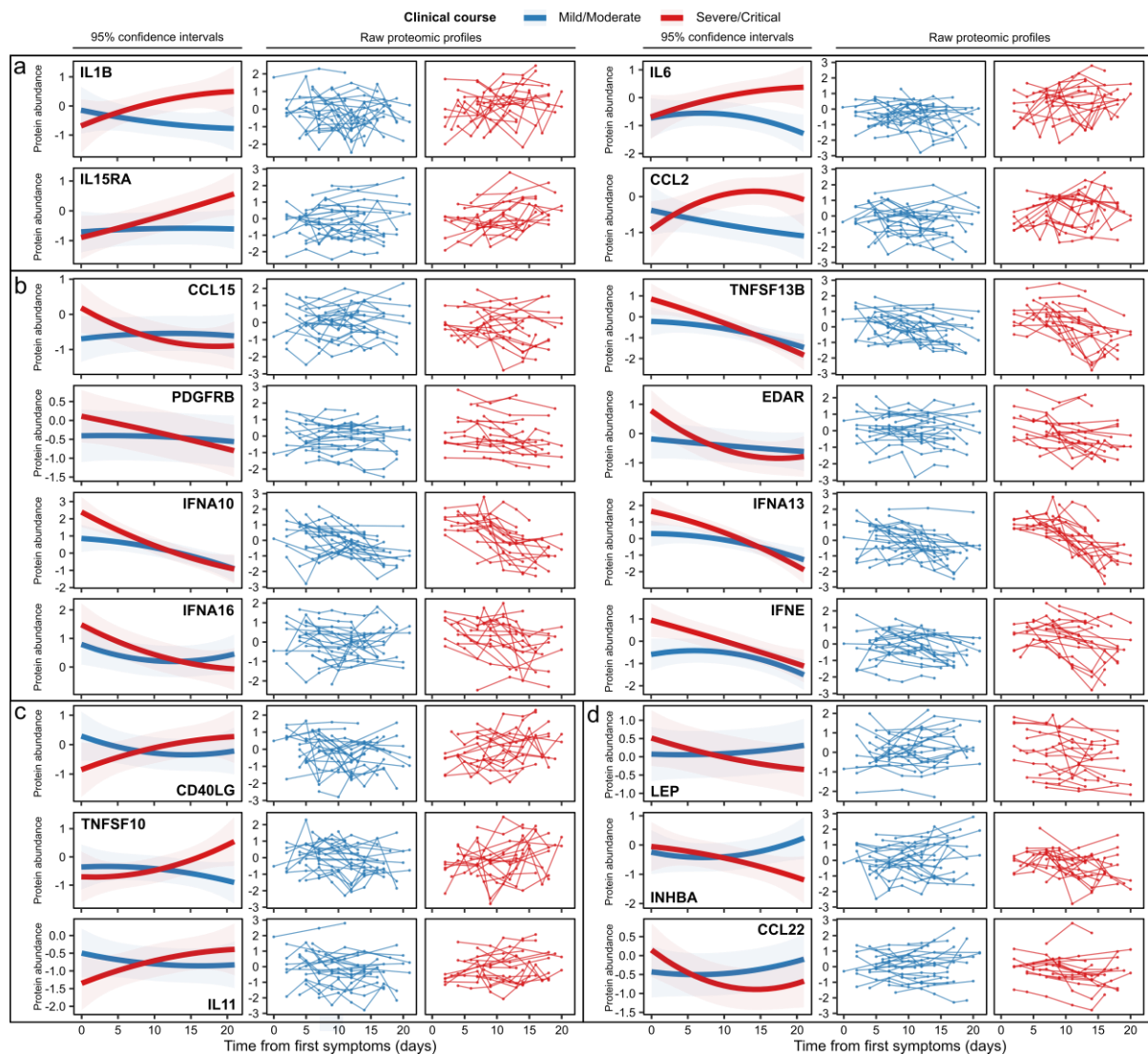


Figure 4.27 Example cytokine profiles and underlying protein abundance

Examples of proteins with differing patterns of discordance over time in severe/critical versus mild/moderate patients (TxCC, LMM, FDR <0.05). Lines represent estimated marginal means and shaded areas represent their 95% confidence intervals. The raw proteomic profiles display protein abundance for each individual ($n = 169$ samples from 40 individuals).

Another proteomic study of COVID-19, carried out by Filbin *et al.*, applied a very similar mixed modelling approach to investigate severity-specific profiles using Olink immunoassays. A subtle difference is that they encoded time as a categorical variable (0, 3 and 7 days) while I treated it as a continuous variable. Of the 50 cytokines exhibiting a significant time x clinical course interaction effect in the ESKD cohort, 32 were measured on the Olink panel used by Filbin *et al.* Of these 32 proteins, they found that 28 had a significant (FDR < 0.05) time x severity interaction term. There is limited concordance between the Olink and SomaScan assays (as illustrated by **Figure 4.28**), so the four discordant proteins may be explained by the limited concordance between the assay platforms.

Filbin *et al.* also performed proteomics using version four of the SomaScan platform v4 (as opposed to v4.1, which was used in the present study), which measured 46/50 of the cytokines significant in the analysis of the ESKD cohort. I inverse-rank normalised their SomaScan data for these cytokines and applied LMMs with an interaction term for time and severity. 33/46 of the cytokines replicated (at unadjusted $p < 0.05$) in this analysis. These results show that there is a relatively high degree of concordance in terms of plasma cytokines and receptors that have a different longitudinal profile between mild and severe patients in the ESKD cohort and a non-ESKD patient group. This suggests that severe COVID-19 is characterised by common dynamic cytokine changes across patient groups.

One exception to this was EPOR (erythropoietin receptor). In the ESKD patients, EPOR exhibited a significantly different temporal profile in severe/critical versus mild/moderate disease ($p = 1.7 \times 10^{-8}$, LMM time x disease course interaction). In contrast, the temporal profile was not significantly different among differing disease severity strata in the data of Filbin *et al.* ($p = 0.47$, LMM, TxCC interaction) (**Figure 4.29a**). As discussed previously (**Section 2.4.7**), in ESKD there is a loss of EPO production by the kidney. Visualisation of the ESKD patient data revealed a decline in EPOR abundance over time in severe COVID-19 and a more stable profile in mild/moderate disease (**Figure 4.29b**); these contrasting profiles could reflect the changes in erythropoietin responsiveness that accompany critical illness and sepsis in ESKD patients^{316,317}.

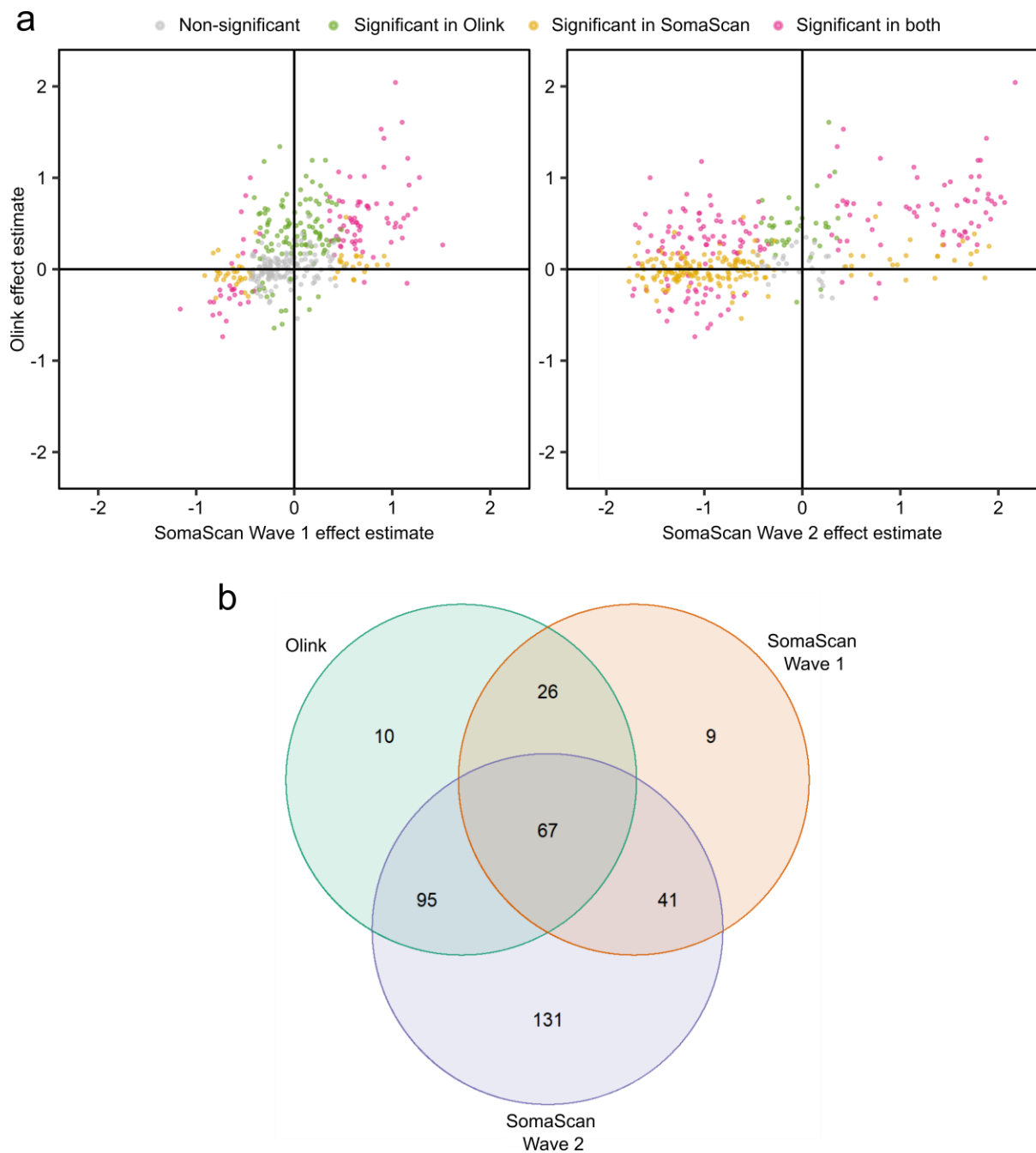


Figure 4.28 Comparison of SomaScan and Olink assays

The SomaScan data for both the Wave 1 and Wave 2 cohorts were compared to the Olink results for the Wave 1 cohort (i.e., the results described in **Section 2.4.1 Proteomic effects of COVID-19**). **a**) The effect estimates for the differential abundance of proteins between COVID-19 positive vs. negative patients. Colour represents significance ($FDR < 0.05$) in each study. **b**) The overlap of significant proteins for the comparison of COVID-19-infected and non-infected patients.

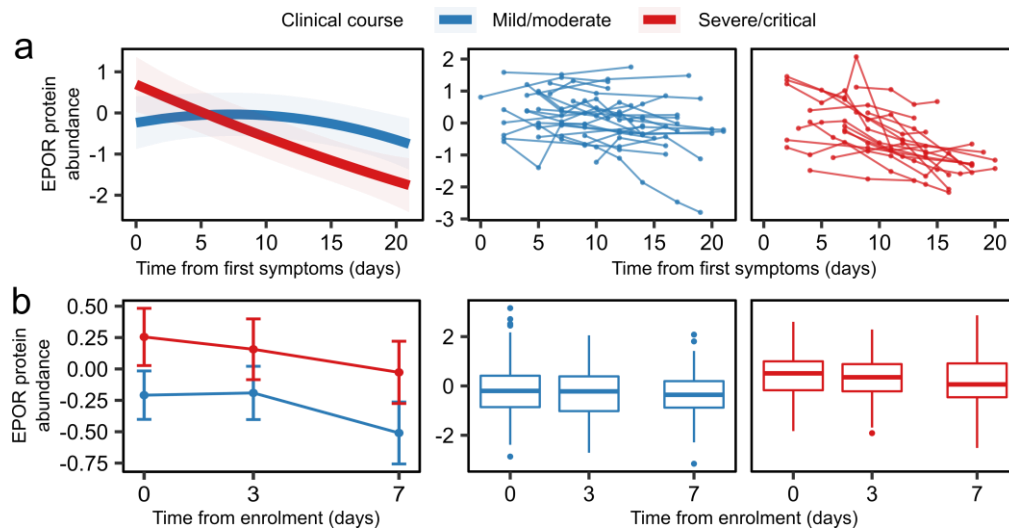


Figure 4.29 The longitudinal profile of plasma EPOR protein

a) The temporal profile of EPOR in the ESKD cohorts, stratified by severity of the patient's overall clinical course. Left: lines represent LMM estimated marginal means and shaded areas represent their 95% confidence intervals. Right: raw data for each individual ($n = 169$ samples from 40 individuals). **b)** The profile of EPOR in the data from Filbin et al. (non-ESKD cohort; $n = 655$ samples from 306 individuals). Left: points represent LMM estimated marginal means and error bars indicate their 95% confidence intervals. Right: box and whiskers plots showing the underlying protein data, where centre = median, upper bound = upper quartile (UQ), lower bound = lower quartile (LQ), upper whisker = largest value at most $1.5 \times \text{IQR}$ (inter-quartile range) from the UQ, lower whisker = smallest value at most $1.5 \times \text{IQR}$ from the LQ, points = samples outside the range of the whiskers.

4.4.8 Plasma LRRC15 predicts COVID-19 severity

I next investigated whether clinical severity could be inferred from the transcriptomic and/or proteomic data and which had better predictive performance. For each COVID-19 patient, I selected the first sample at the patient's peak severity score so that there was one sample per patient. To predict COVID-19 severity at the time of sampling, I employed two supervised learning methods: lasso and random forests. I applied these separately for i) the plasma proteomic data; ii) the PBMC transcriptomic data; and iii) the combination of both (the multi-omic data). For this analysis, I combined the COVID-19 cases from both cohorts. The area under the curve (AUC) was estimated using Monte Carlo cross-validation (see **Section 4.3.17**).

The proteomic-based models outperformed the transcriptome-based ones, with non-overlapping 95% confidence intervals (**Figure 4.30a, Figure 4.31a**). The lasso model generated on the proteome had an estimated AUC of 0.93 (versus 0.86 for the transcriptome). The random forests model generated on the proteome had an AUC of 0.88 (versus 0.83 for the transcriptome). The models based on the proteome alone also had greater predictive performance than those trained on the multi-omic data, although the confidence intervals for the AUC estimates overlapped (**Figure 4.30a, Figure 4.31a**).

I next examined the supervised learning models to identify the most important biomarkers that differentiate severe/critical disease (**Section 4.3.17, Figure 4.30b, Figure 4.31b**). Although only a minority of the input features to the multi-omic model were proteins (34%; 6,323/18,548), proteins made up the majority of the top 15 most important predictors (10/15 for lasso and 9/15 for random forests). This, and the finding that the plasma proteome was a superior classifier of severity than the PBMC transcriptome, highlights that plasma proteins provide a valuable read-out of the pathophysiological processes in severe COVID-19.

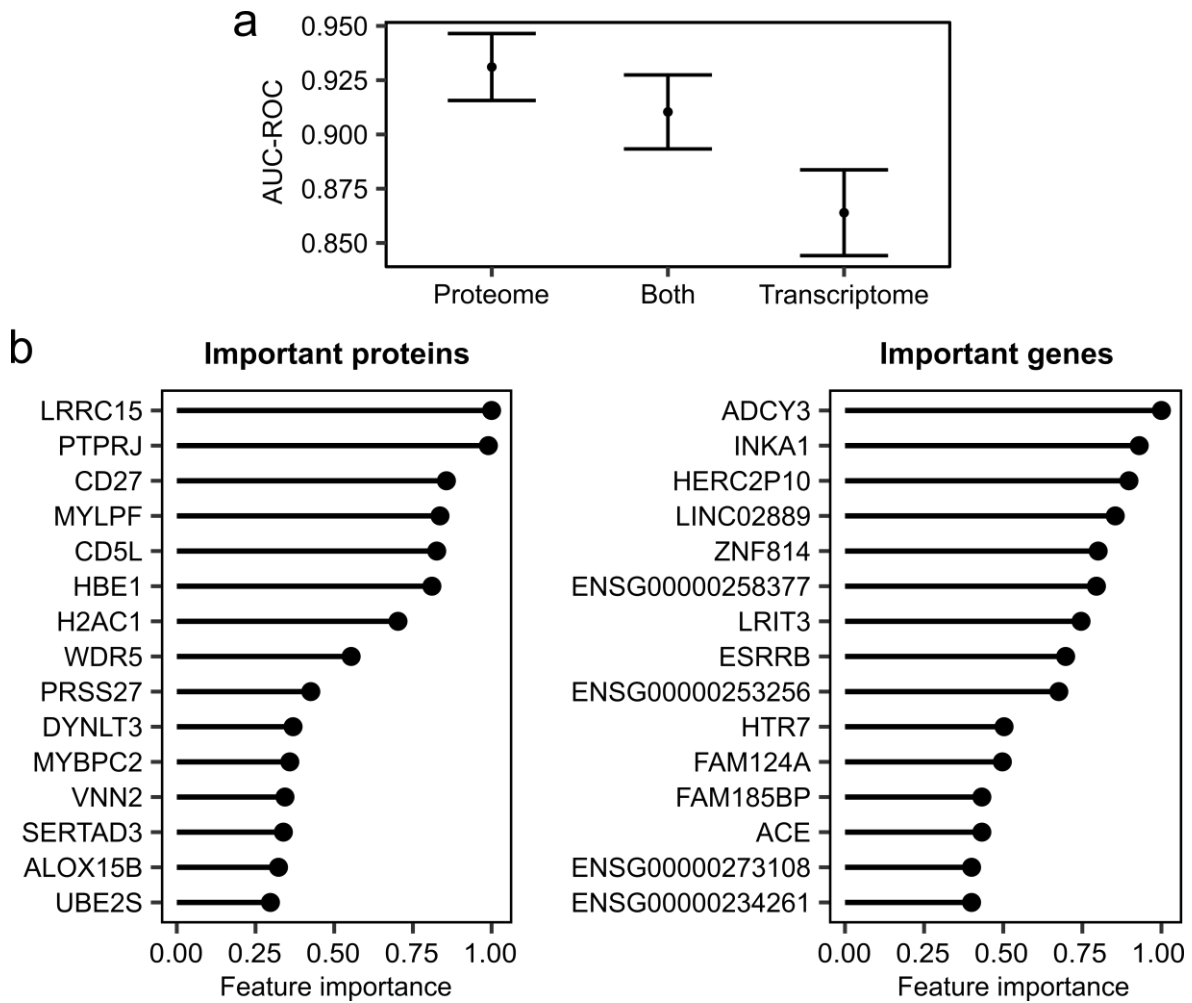


Figure 4.30 Supervised learning to predict COVID-19 severity from molecular features

a) Point estimates (mean) and 95% confidence intervals of the area under the receiver operating characteristic curve (AUC-ROC) for predicting COVID-19 severity (from 200 cross-validation resamples of 51 independent samples) using lasso regression. Both = supervised learning on the combined proteomic and transcriptomic data.

b) Important proteins (left) and genes (right) for the lasso model. Feature importance is scaled between 0 and 1, where 1 represents the most important feature.

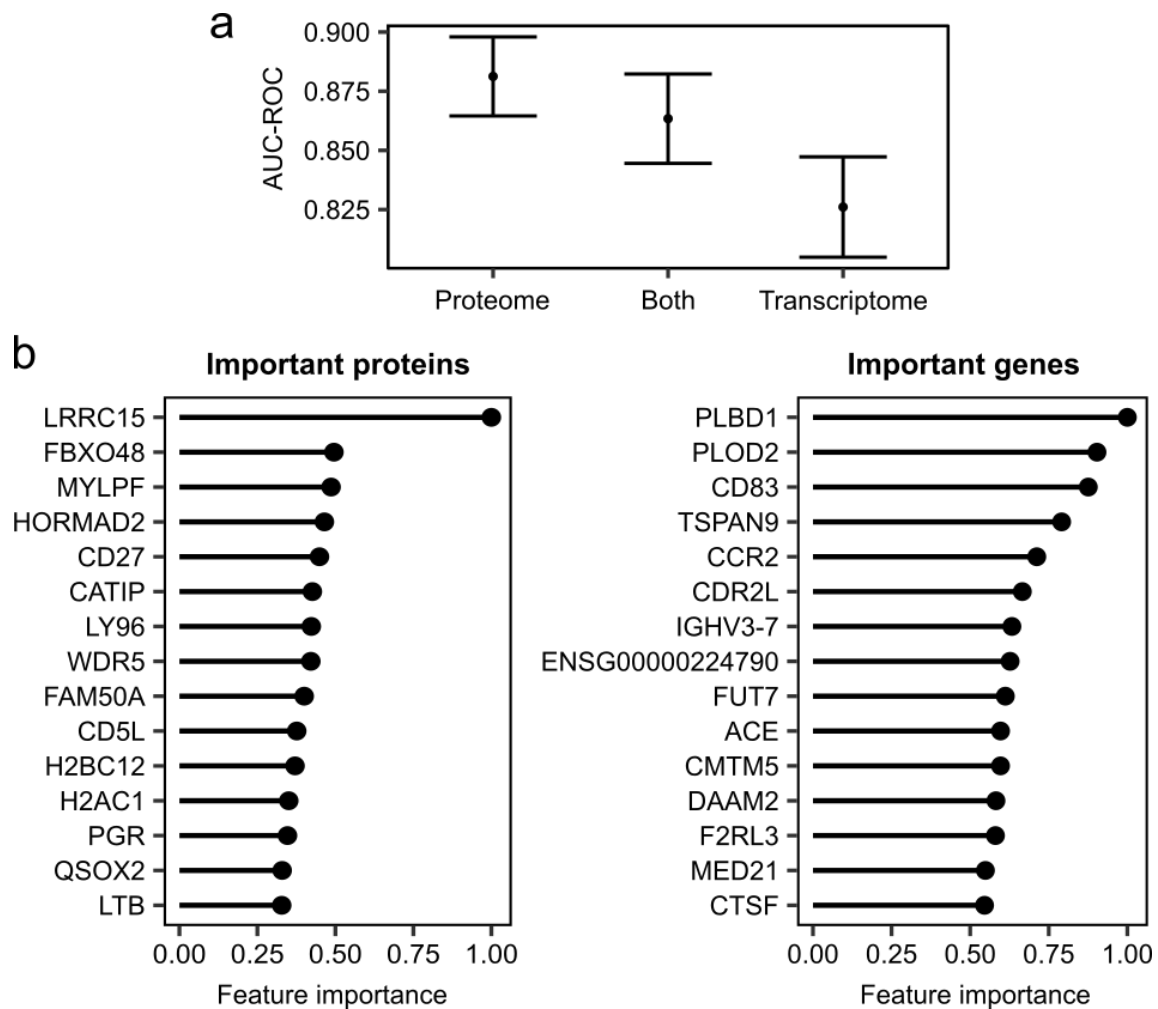


Figure 4.31 Predicting COVID-19 severity using random forests

a) Point estimates (mean) and 95% confidence intervals of the area under the curve from receiver operating characteristic curves (AUC-ROC) for predicting COVID-19 severity (from 51 independent samples, i.e., one sample per patient) using random forests. Both = supervised learning on the combined proteomic and transcriptomic data.

b) Important proteins (left) and genes (right) for the random forests model. Feature importance is scaled between 0 and 1, where 1 represents the most important feature.

Both lasso and random forests identified plasma leucine-rich repeat containing 15 (LRRC15) protein levels as the most important biomarker of COVID-19 severity. Interestingly, around the time this analysis was carried out, this protein had been recently identified by three independent studies as a receptor for SARS-COV-2^{318–320}. I examined LRRC15's longitudinal trajectory over the course of COVID-19 infection, finding that it displayed a different temporal profile dependent on the disease course ($p = 6.5 \times 10^{-8}$, TxCC interaction, LMM). The concentration was stable in most individuals with mild/moderate COVID-19 (**Figure 4.32**), whereas it decreased over time in severe/critical patients. Thus, a snapshot level of LRRC15 and its dynamic profile over time can convey information on the current clinical state of the patient and the overall course of the disease, respectively.

Previously, I found that the dynamic profiles of some cytokines did not replicate in studies of COVID-19 that did not involve ESKD patients (e.g., EPOR) (**Figure 4.29**). Therefore, I investigated whether LRRC15 is a biomarker of COVID-19 severity specifically in ESKD or applies to more general patient groups. I identified two external studies in non-ESKD patients that measured plasma LRRC15 using the SomaScan platform. Su *et al.*²⁴³ performed proteomic profiling in two general population cohorts. In this study, LRRC15 levels were significantly associated with their COVID-19 outcome (Benjamini-Hochberg adjusted p -value = 10^{-10}), with lower levels associated with severe disease.

The study of Filbin *et al.*⁷⁵ found that individuals with more severe COVID-19 had a different temporal profile of plasma LRRC15 (again measured using SomaScan) compared to less severe cases. I re-analysed their data to model and visualise the interaction between time and severity for LRRC15. This revealed that, as observed in the ESKD cohorts, LRRC15 abundance fell over time in more severe patients relative to milder cases (**Figure 4.32**). These findings suggest that LRRC15 is also a biomarker of severe COVID-19 in more general patient cohorts.

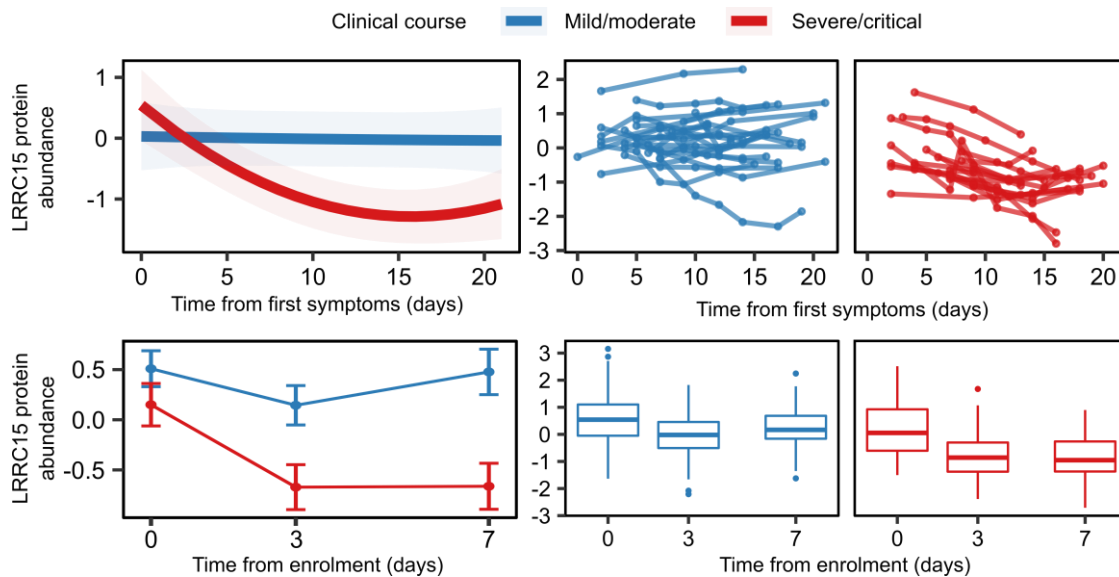


Figure 4.32 The profile of LRRC15 in plasma following COVID-19 infection

Above) The profile of LRRC15 plasma protein concentration over time, stratified by severity of the patient's overall clinical course ($n = 169$ samples from 40 individuals). Left: lines represent estimated LMM marginal means and shaded areas represent their 95% confidence intervals. Right: raw data for each individual.

Below) The longitudinal profile of LRRC15 in the data published by Filbin *et al.* ($n = 655$ samples from 306 individuals). Left: points represent LMM estimated marginal means and error bars indicate their 95% confidence intervals. Right: box and whiskers plots showing the underlying protein data, where centre = median, upper bound = upper quartile (UQ), lower bound = lower quartile (LQ), upper whisker = largest value at most $1.5 \times \text{IQR}$ (inter-quartile range) from the UQ, lower whisker = smallest value at most $1.5 \times \text{IQR}$ from the LQ, points = samples outside of the range of the whiskers.

In the supervised learning analysis, I elected to train models using cross-validation performed on the combined data from the Wave 1 and Wave 2 cohorts. This approach maximised the training data available for modelling and allowed for the estimation of confidence intervals for the performance metrics, which was an important consideration given the relatively modest sample size. However, this also meant that I could not test the models on a completely independent cohort. Therefore, to ensure that our models of COVID-19 severity would perform well given new data, I performed a secondary analysis wherein a second set of random forests and lasso models were trained on the transcriptomic and proteomic datasets using cross-validation in the Wave 1 cohort alone. I was then able to test the models in the Wave 2 cohort to estimate the AUC and evaluate the ability of these models to extrapolate to truly unseen data. Both the proteomic (lasso AUC = 0.84, random forests AUC = 0.77) and transcriptomic (lasso AUC = 0.70, random forests AUC = 0.96) models had predictive value in the unseen Wave 2 cohort, indicating that these supervised learning models can generalise to completely unseen data. There was, however, greater variability between the performance metrics of the

four models than those estimated in the primary analysis using cross-validation. However, it is important to stress that these AUC estimates are generated from testing a single time on the Wave 2 cohort and thus are unlikely to be as accurate as the estimates from the primary analysis (cross-validation on the combined Wave 1 and 2 data) which were averaged over 200 cross-validation iterations.

4.4.9 Persistent deranged platelet and coagulation pathways in convalescence

For 12 of the 17 patients in the Wave 2 cohort, a sample was obtained after clinical recovery approximately two months following the acute infection. PCA analysis of the PBMC transcriptome showed that, while pre-COVID-19 and convalescent samples appeared more similar than samples taken during COVID-19, there were differences between the convalescent samples and their pre-infection counterparts (**Figure 4.33a**), indicating that they have not fully returned to baseline. I compared the convalescent samples to their paired pre-COVID-19 samples, finding 25 significantly differentially expressed genes (1% FDR), of which 24 were up-regulated post-COVID-19 (**Figure 4.33b, Table 4.6**).

Up-regulated clotting-related genes included *PF4* (encoding platelet factor 4) and the related gene *PF4V1* (platelet factor 4 variant 1). Of note, these genes are located in the same genomic region on chromosome 4, along with the chemokine *CXCL5*, which was also significantly up-regulated. Another nearby gene, Pro-Platelet Basic Protein (*PPBP* or *CXCL7*), had a higher abundance in convalescent samples, although it did not quite reach significance at 1% FDR (nominal $p = 3.3 \times 10^{-5}$, adjusted $p = 0.016$). This gene encodes a neutrophil chemoattractant released by activated platelets. The up-regulation of these neighbouring genes suggests they are influenced by a shared genomic regulatory element. Other genes upregulated in convalescence included *ESAM* (Endothelial Cell Adhesion Molecule), which has roles in thrombosis and platelet aggregation, and *CAVIN2*, which is required for endothelial cell maintenance and proliferation. Additionally, there were upregulated genes involved in platelet activation (e.g., *ALOX12*, *FSTL1*, *PTGS1*), cell adhesion (e.g., *ITGB5*), and calcium binding and muscle contraction (e.g., *CALD1*, *PVALB*).

Overrepresentation analysis of the 25 differentially expressed genes revealed significant enrichment of terms including “Platelet activation, signaling and aggregation”, “Formation of fibrin clot/clotting cascade”, “Chemokine signaling pathway”, “SARS-CoV-2 innate immunity

evasion and cell-specific immune response” and “Smooth muscle contraction” (**Figure 4.33c**). These data suggest persistent activation of abnormal processes for a considerable time after clinical recovery. In particular, they implicate the vascular and clotting systems, which may have implications for the long-term risk of thrombosis.

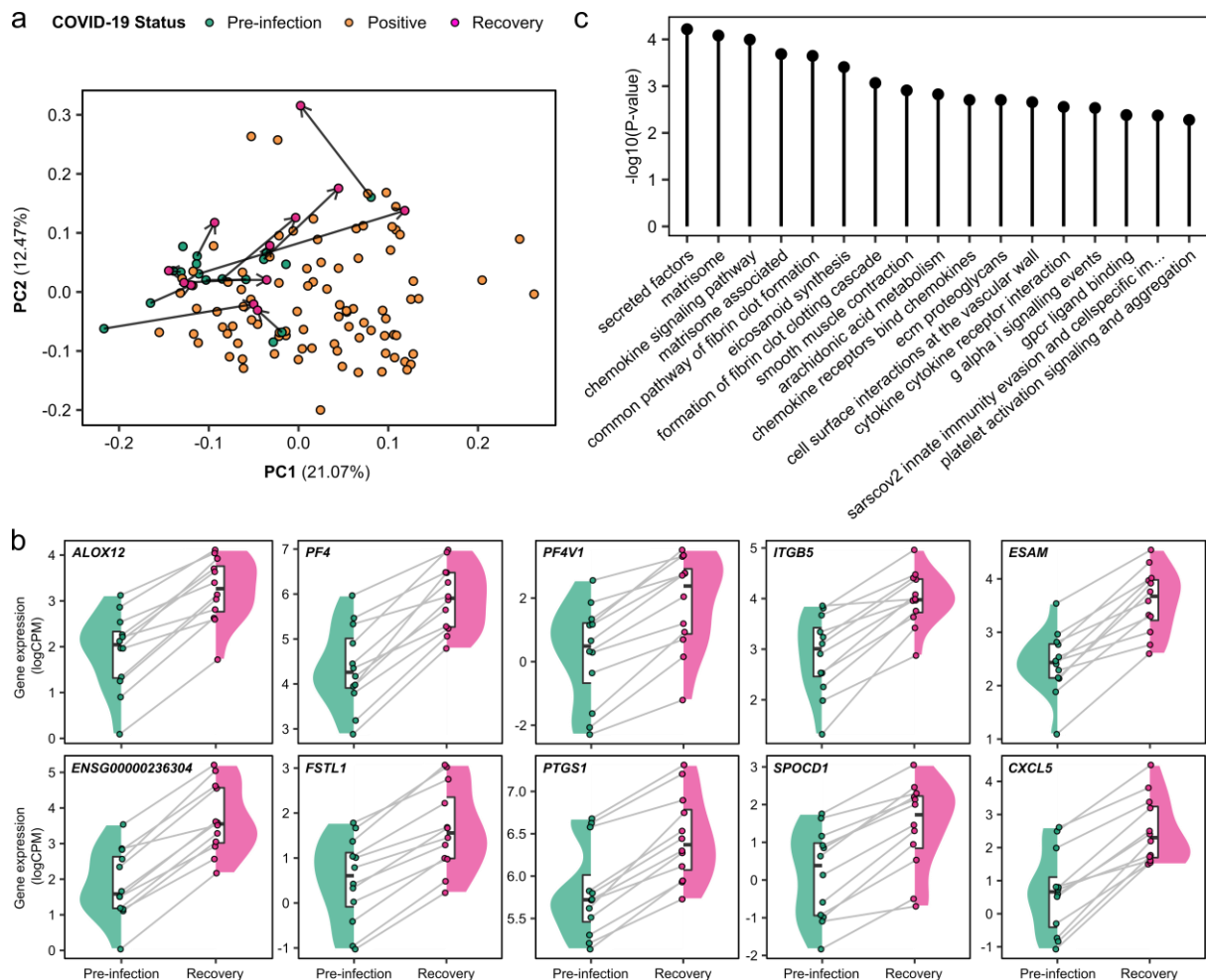


Figure 4.33 Persistent dysregulation of immune cell gene expression after recovery from COVID-19

a PCA of the Wave 2 PBMC transcriptomic data, including pre-infection, infection and recovery samples (taken 2 months after the acute illness). Each point represents a sample. Arrows link recovery samples to the pre-infection sample from the same individual.

b Paired violin plots for differentially expressed genes in recovery versus pre-infection samples ($n = 24$ samples from 12 individuals). Grey lines link each individual's pre-infection sample to their recovery sample; these samples are represented by points. Shaded areas indicate kernel density estimates. For boxplots, centre = median, upper bound = upper quartile, lower bound = lower quartile.

c All significantly enriched (5% FDR) pathway terms for the differentially expressed genes in recovery versus pre-infection samples. Overrepresentation testing was performed for each gene set with a one-sided Fisher's exact test.

Table 4.6 Genes that do not return to baseline after recovery from COVID-19

Genes that are significantly differentially expressed (1% FDR, LMM) in recovery versus pre-infection samples ($n = 24$ samples from 12 patients). The estimate represents the estimated difference in logCPM (counts per million).

Gene ID	Gene name	Estimate (recovery - pre-infection)	P-value	Adjusted P-value
ENSG00000236304	lncRNA	1.84	5.37×10^{-08}	9.20×10^{-04}
FSTL1	Follistatin Like 1	1.17	1.60×10^{-07}	1.37×10^{-03}
PTGS1	Prostaglandin-Endoperoxide Synthase 1	0.64	3.69×10^{-07}	2.11×10^{-03}
SPOCD1	SPOC Domain Containing 1	1.31	6.41×10^{-07}	2.53×10^{-03}
CXCL5	C-X-C Motif Chemokine Ligand 5	1.84	7.39×10^{-07}	2.53×10^{-03}
ALOX12	Arachidonate 12-Lipoxygenase, 12S Type	1.33	1.24×10^{-06}	3.53×10^{-03}
PF4	Platelet Factor 4	1.53	1.71×10^{-06}	3.91×10^{-03}
ESAM	Endothelial Cell Adhesion Molecule	1.18	1.83×10^{-06}	3.91×10^{-03}
MT-RNR1	Mitochondrially Encoded 12S rRNA	1.00	2.61×10^{-06}	4.13×10^{-03}
MMD	Monocyte To Macrophage Differentiation Associated	0.77	2.80×10^{-06}	4.13×10^{-03}
ENSG00000240093	lncRNA	-0.71	2.82×10^{-06}	4.13×10^{-03}
MTURN	Maturin, Neural Progenitor Differentiation Regulator Homolog	0.54	3.11×10^{-06}	4.13×10^{-03}
GNG11	G Protein Subunit Gamma 11	1.53	3.13×10^{-06}	4.13×10^{-03}
CAVIN2	Caveolae Associated Protein 2	1.15	4.51×10^{-06}	5.52×10^{-03}
DOK6	Docking Protein 6	1.59	5.00×10^{-06}	5.71×10^{-03}
LINC00989	lncRNA	1.25	5.76×10^{-06}	5.81×10^{-03}
SPARC	Secreted Protein Acidic And Cysteine Rich	1.65	6.07×10^{-06}	5.81×10^{-03}
PF4V1	Platelet Factor 4 Variant 1	1.60	6.11×10^{-06}	5.81×10^{-03}
ABLIM3	Actin Binding LIM Protein Family Member 3	1.41	6.88×10^{-06}	6.07×10^{-03}
MFAP3L	Microfibril Associated Protein 3 Like	0.74	7.09×10^{-06}	6.07×10^{-03}
CALD1	Caldesmon 1	1.78	7.79×10^{-06}	6.23×10^{-03}
ITGB5	Integrin Subunit Beta 5	1.10	8.00×10^{-06}	6.23×10^{-03}
LINC01750	lncRNA	2.00	1.13×10^{-05}	8.22×10^{-03}
PCSK6	Proprotein Convertase Subtilisin/Kexin Type 6	1.22	1.15×10^{-05}	8.22×10^{-03}
PVALB	Parvalbumin	2.04	1.46×10^{-05}	9.97×10^{-03}

4.5 Discussion

In this chapter, I described the analysis of transcriptomic and proteomic data generated for ESKD patients with and without COVID-19. This work extends that described in Chapter 2 in two main ways. Firstly, an additional cohort was considered in this chapter consisting of patients infected with SARS-CoV-2 in early 2021. Secondly, bulk RNA-seq of PBMCs was performed in conjunction with SomaScan plasma proteomics. The SomaScan assays permitted the measurement of far more proteins (6,323) than those measured by the Olink panels employed previously (436). Therefore, in this chapter, I was able to perform a broader analysis of the COVID-19 multi-ome.

I compared the estimated effect of COVID-19 for each protein and gene between the Wave 1 and Wave 2 cohorts (**Figure 4.21**). There were high levels of consistency between the cohorts, providing some validation of these findings. However, this comparison was limited by the fact that these cohorts are not truly independent because the non-infected controls of the Wave 2 cohort were a subset of the Wave 1 cohort controls. So, it is still important to consider whether these results replicate in fully independent external cohorts. For the comparison of COVID-19 cases and controls and the investigation of COVID-19 severity, I used Robust Rank Aggregation (RRA) to identify molecules and pathways that were consistently dysregulated in both cohorts. This allowed for the prioritisation of more robust features that may be more likely to replicate in other cohorts.

There were several common themes between the transcripts and proteins associated with severe COVID-19, including enrichment for pathways related to histones, the interferon response, granulopoiesis, clotting, TCR activation, and the cell cycle. For instance, histone-encoding genes and plasma levels of histone proteins were both upregulated in more severe COVID-19. Upregulation of cell cycle processes and genes encoding histones could reflect the increased proliferation of immune cells, as sufficient histones are required for DNA replication during mitosis³²¹. Histones are predominantly intracellular proteins, so the differential abundance of these proteins in plasma could reflect higher levels of cell damage and death. Histones have also been shown to act as damage-associated molecular patterns (DAMPs), that can directly ligate to toll-like receptors, leading to an inflammatory response³²². More specifically, they can direct damage to epithelial and endothelial cells³²³. Additionally,

histones are constituents of neutrophil extracellular traps (NETs). NETs are extracellular structures produced by neutrophils to trap pathogens. NETs can contribute to tissue injury in COVID-19 and may increase the risk of thrombotic events³²⁴.

Using weighted gene correlation network analysis, I identified modules of coexpressed genes with dynamic profiles in severe COVID-19, many of which appeared to reflect the activity of particular cell types in the PBMC fraction. For instance, several members of module tB encoded proteins related to neutrophils, including components of the phagocyte NADPH oxidase. This gene signature may reflect the presence of low-density granulocytes in the PBMC fraction; these cells have been observed in PBMCs, and they are more abundant in patients with sepsis³²⁵. The presence of low-density granulocytes may indicate the prolonged activation of neutrophils. Furthermore, low-density granulocytes themselves have been found to be more abundant in patients with severe COVID-19 and may be predisposed to NET formation¹⁴⁹. In line with this, the module contained a number of histone-encoding genes. Other modules that appeared to reflect cell types in the PBMC fraction included those related to T cell activity, whose expression fell in more severe disease, likely reflecting lymphopenia which is a well-known feature of severe COVID-19²¹⁹. The modular analysis also highlights a challenge in interpreting bulk RNA-seq data: transcriptomic signatures may indicate changes in gene expression, but they may also reflect changes in the distribution of cell types within the PBMC fraction. In the publication associated with this chapter, flow cytometry data was additionally performed in an attempt to identify such shifts²⁷⁶.

Pathways related to the control of transcription and translation were upregulated in more severe COVID-19 cases, perhaps reflecting dysregulation of splicing, translation and nucleic acid metabolism in SARS-CoV-2 infected cells^{326,327}. The dysregulation of these canonically intracellular proteins likely points to tissue damage and increased cell turnover. Cytokines, on the other hand, are soluble proteins that play a major role in the pathogenesis of inflammatory diseases and are important targets of existing drugs³¹⁵. Longitudinal analysis of cytokines and their receptors revealed divergent temporal trajectories between disease severity strata, manifesting in several patterns (**Figure 4.26**). For example, IL6 has similar plasma abundance at symptom onset regardless of disease severity. However, it increased in abundance over time in severe and critical patients, whereas it fell in patients with a milder clinical course (**Figure 4.26**). IL6 contributes to the excessive pro-inflammatory state observed in patients

with severe COVID-19 and tocilizumab, which inhibits IL6-IL6R signaling, has been shown to improve survival in critically ill patients¹⁶⁰. CCL22 was more abundant in severe/critical patients near the onset of disease, but reduced in concentration over time (**Figure 4.26**); in contrast, in mild/moderate patients, its abundance in plasma increased across the disease course. CCL22 is involved in the chemotaxis of a variety of immune cells (including dendritic cells, monocytes, activated T cells and natural killer cells). Mice deficient for CCL22 were found to be susceptible to inflammatory disease³²⁸, suggesting a role in curbing the overactivation of inflammatory processes. Its absence in the later stages of severe COVID-19 could contribute to a prolonged inflammatory response in these patients. The type 1 interferons IFNA16, IFNA10 and IFNA13 (encoded by neighbouring genes on 9p21.3) had similar profiles, in which their abundance was high at symptom onset and reduced over the course of the disease (**Figure 4.26**); this pattern was more extreme in severe and critical patients relative to mild and moderate patients. In mice infected with SARS-CoV, a delayed type 1 interferon response leads to elevated levels of inflammatory cytokines and more severe disease³²⁹. Since the proteins in the current study were measured following symptom onset, the initial response to infection is missed. Therefore, the initially higher levels of type 1 interferons observed in **Figure 4.26** could reflect a delayed response in more severe COVID-19. The differential interferon response following SARS-CoV-2 infection could also be related to viral load.

When compared to healthy controls, there was extensive dysregulation of both transcripts and proteins in ESKD patients, and these perturbations included inflammatory mediators. In the transcriptomic data, pathway analysis converged on pathways downstream of erythropoietin receptor (EPOR) signaling. Both *EPOR* and its downstream signalling pathways were transcriptionally upregulated in the PBMCs of the ESKD patients. In addition to its well-known effects on erythropoiesis, EPO binds to EPOR on the surface of PBMCs to stimulate intracellular signaling pathways (including STAT5, JAK2, ERK and MAPK), which may, in turn, mediate inflammation³³⁰. For instance, treatment of leukocytes with EPO modulated the production of cytokines, with potential anti-inflammatory effects³³¹.

There was also a greater abundance of pathways related to haem signaling in the PBMCs of ESKD patients on dialysis. These included genes encoding haemoglobin subunits and erythrocyte-specific genes, such as rh-associated glycoprotein. A small proportion of PBMCs are circulating stem cells, including erythroid progenitors, which could explain the presence of

these genes in the present study³³²; these progenitors are elevated in the PBMC fraction of patients with heightened erythropoiesis. The upregulation of erythrocyte-specific genes in ESKD could therefore indicate the expansion of circulating erythrocyte precursors. If this is the case, then it is plausible that these cells are the source of EPO-EPOR disruption observed in ESKD patients. However, haemoglobin mRNA depletion was performed as part of the RNA-seq protocol, which should have removed transcripts from contaminating erythrocytes. Differential expression of haemoglobin subunits by PBMCs can represent dysregulation of oxidative stress as a result of inflammation^{333,334}, providing an alternative explanation for this gene signature. However, these findings should be interpreted in light of the fact that there were systematic differences in average gene expression between ESKD patients and healthy participants. The normalisation method I applied can correct for these differences given that most of the genes measured are not differentially expressed. However, it is difficult to assess whether this is truly the case in the present study. Therefore, the differential expression of these genes could be an artefact of TMM normalisation applied under erroneous assumptions.

Previously, bulk and single-cell RNA sequencing of PBMCs was performed for patients undergoing haemodialysis and healthy controls^{287,288}. The authors carried out enrichment analyses in various cell subsets, which uncovered pathways similar to those observed in the present study, including fatty acid and haem metabolism, interferon-gamma, tumour necrosis factors and a multitude of signaling pathways (including JAK-STAT, NFkB, IFNG and PI3K) in various cell subsets. There have been numerous proteomic studies of kidney disease (see **Section 2.2.2**), although few of these have specifically investigated the effects of ESKD and dialysis on the plasma proteome. Many of the markers implicated in these studies were also found to be dysregulated in the analysis described in this chapter. For instance, one such study highlighted the prognostic value of tumour necrosis factor receptor superfamily members (TNFRSFs) in predicting the development of ESKD¹⁸⁸. Another investigation of chronic kidney disease patients also found upregulation of cystatin C, FGF23 and TNFRSFs using SomaScan assays³³⁵.

While ESKD and dialysis have pronounced effects on both the transcriptome and proteome, both the COVID-19 cases and controls sampled in this study consisted of ESKD patients undergoing haemodialysis. So, the results of these comparisons should reflect the effect of COVID-19, rather than kidney disease. However, it is difficult to ascertain whether these

patients respond to COVID-19 in the same way as individuals with normal renal function. For this reason, I performed a formal comparison of these results to other transcriptomic and proteomic datasets. The transcriptomic findings were remarkably similar to those of the COMBAT study³⁰⁶, despite PBMCs being measured for the ESKD cohort, as opposed to whole blood in the COMBAT study. However, there was lower concordance between the ESKD cohorts and the results of Filbin *et al.*⁷⁵. It should be noted that comparing studies in this manner has inherent limitations, as it is impossible to determine the source of such discordance. Differences between the studies could be due to biological differences, such as those caused by ESKD or dialysis, or they could be explained by differences in study design, statistical power, assay platforms and other sources of non-biological variation. For instance, whilst the ESKD controls did not have known infections, those of Filbin *et al.* presented with acute respiratory distress. Furthermore, Filbin *et al.* used an earlier version of the SomaScan platform, so there could be differences in the protein affinities of the assays. Alternatively, it is plausible that the differences in circulating proteins between the studies could be due to kidney disease, as circulating proteins are affected by renal impairment^{211–213,336}. However, it should be noted that the results based on the Olink proteomic assays performed for the Wave 1 cohort were relatively concordant with the data of Filbin *et al.* (see **Section 2.4.8**). So, the observed discordance appears to be limited to the SomaScan data gathered by these studies. Furthermore, there was limited correlation between the SomaScan-based models based on the Wave 1 and Wave 2 cohorts. Therefore, these differences might represent the technical limitations of the SomaScan platform, as opposed to biological differences related to kidney function. Overall, comparisons revealed that the results presented in this chapter were generally similar to those identified in cohorts of COVID-19 patients without ESKD, albeit less so for SomaScan plasma proteomics than for RNA-seq and Olink proteomics. Regardless, caution should be taken in attempting to generalise the findings presented in this chapter to other patient populations.

Differential expression and abundance analyses between groups allow for the estimation of a single effect size. This allows for the simple comparison of estimates between studies, permitting the use of correlation to estimate concordance. However, it is far more challenging to ascertain whether temporal protein trajectories replicate in other studies. Regardless, I found that the majority of cytokines with severity-dependant temporal profiles were also

identified by a similar modelling strategy employed by Filbin *et al.*⁷⁵. However, while the temporal dynamics of EPOR were not dependent on severity in the data of Filbin *et al.*, its abundance fell more rapidly in ESKD patients with severe COVID-19. While anaemia is treated in kidney disease patients using erythropoiesis-stimulating agents (ESA), inflammation can cause ESA hyporesponsiveness^{316,317}. Higher levels of inflammation in severe COVID-19, and therefore reduced response to ESA, could lead to increased expression of EPOR to combat anaemia, explaining its unique profile in the ESKD cohort. Alternatively, hospital admission could disrupt ESA treatments. Finally, other differences between the studies could explain the disparate profiles of EPOR.

Supervised learning of the plasma proteome appeared to better capture disease severity than the PBMC transcriptome. Furthermore, proteins had greater importance than genes according to algorithms trained on both sets of features. These findings suggest a greater impact of COVID-19 on the circulating proteome of ESKD patients than the immune cell transcriptome. While the transcriptomic data gathered in this study only contained information on the state of PBMCs, plasma proteins could reflect pathogenesis and tissue injury in a greater number of tissue types and organs. However, the possibility that technical factors account for these differences cannot be excluded.

The supervised learning approach did not leverage serial samples, so only one sample per patient was used during the training process. Furthermore, these models were trained to predict current severity, as opposed to future clinical deterioration. Therefore, these models likely have limited prognostic value for determining patients that will develop severe COVID-19. Larger studies of COVID-19 that include patients with normal renal function are better placed to develop such models, as has been done by Filbin *et al.*⁷⁵ and for other large molecular datasets^{243,244}. While there were limitations to the supervised learning approach employed in this chapter, its main goal was not prediction itself. Rather, I aimed to extract key biomarkers. To a large extent, the predictive power of the proteome could be explained by leucine-rich repeat-containing protein 15 (LRRC15). Plasma abundance of LRRC15 alone was highly predictive of severity (AUC = 0.94). This protein is not measured by Olink panels, which may explain why this protein has not been highlighted by previous studies in the context of COVID-19. I found that the abundance of LRRC15 was similar regardless of severity at symptom onset (**Figure 4.32**); its levels remained stable in the plasma of mild and moderate

patients but decreased over time in severe and critical illness. Therefore, it is possible that cross-sectional studies measuring protein abundance at disease onset may not find that it is differentially abundant. Despite this, data mining of previous studies revealed similar findings in two cohorts that included patients with normal renal function^{75,243}.

LRRC15 is a transmembrane protein that interacts with the extracellular matrix and has roles in cell adhesion and migration. Much of the literature relating to LRRC15 describes its aberrant expression in a variety of tumours, including those of the breast and prostate^{337–339}. However, there is also evidence that LRRC15 has roles in mediating inflammation. LRRC15 is highly differentially expressed in the synovial tissues of patients with rheumatoid arthritis (RA)^{340,341}. Knockdown of LRRC15 *in vitro* led to a reduction in signaling by proinflammatory cytokines and suppression of abnormal proliferation³⁴¹. LRRC15 has also been shown to mediate viral infection. LRRC15 impedes adenovirus infection by redistributing its host receptor away from the cell surface³³⁹. More specifically, it has recently been shown to interact with SARS-CoV-2. Three independent studies employed systematic screens for proteins that interact with the SARS-CoV-2 spike protein, and they collectively converged upon LRRC15^{318,320,342}.

Shilts *et al.*³¹⁸ employed two independent large-scale screening methods to identify human proteins that might interact with the SARS-CoV-2 spike protein. The first method involved arraying a library of over 2,000 human cDNAs encoding human membrane proteins before transfecting them into HEK293 cells. Each of the transfected cells were stained with fluorescent spike protein and binding was measured by flow cytometry. The second screen employed a pooled genome-wide CRISPR activation library to identify genes which, when upregulated, induce the binding of spike protein. Both screens identified LRRC15 among the top-ranked candidates. They went on to show that LRRC15, like ACE2, binds the C-terminal domain of spike protein. Furthermore, they found that the expression of LRRC15 was insufficient to permit viral entry. Song *et al.*³⁴² performed a CRISPRa screen on the A375 melanoma cell line, also identifying LRRC15 and ACE2 among the highest confidence SARS-CoV-2 receptors. The authors found that co-culturing HeLa-sgLRRC15 cells with HeLa-ACE2 cells reduced pseudovirus entry. They also re-analysed lung single-cell RNA-seq datasets based on patients with COVID-19^{343,344}, finding that LRRC15 was most highly expressed in fibroblasts, and was particularly expressed by pathological fibroblasts that were abundant in the lungs of COVID-19 patients. Finally, Loo *et al.*³²⁰, also employed a CRISPRa screen to

identify proteins that could bind to spike protein based on HEK293T cells. Once again, the screen identified ACE2 and LRRC15 as the top binding partners for the SARS-CoV-2 spike protein. In this study, fibroblasts expressing LRRC15 were co-cultured with HEK293T cells expressing ACE2 and TMPRSS2; they found that co-culturing LRRC15+ fibroblasts reduced viral entry by pseudovirus and infection by authentic virus. Both Song *et al.*³⁴² and Loo *et al.*³²⁰ found that LRRC15 can act *in trans* to inhibit SARS-CoV-2 cell entry to ACE2+ cells. It may be that LRRC15 acts as a decoy receptor by binding virions that cannot then enter cells via ACE2. The data presented in this chapter provide *in vivo* human evidence to suggest that LRRC15 may be important in the host response to SARS-CoV-2 and are consistent with a model in which a failure to up-regulate LRRC15 increases the risk of severe COVID-19 because of the lack of a receptor that inhibits its entry to cells. Alternatively, Loo *et al.*³²⁰ posits that depletion of lung LRRC15/SARS-CoV-2 complexes could lead to decreased plasma abundance of LRRC15. While these studies indicate that LRRC15 could play a role in mediating severe COVID-19, the mechanism by which LRRC15 could act to impede SARS-CoV-2 infection *in vivo* remains unclear.

In the Wave 2 cohort described in this chapter, samples were taken prior to SARS-CoV-2 infection and approximately two months following symptom onset. Comparison of these timepoints revealed ongoing upregulation of the expression of genes related to the complement, platelet and coagulation pathways. Among the genes upregulated in convalescence was platelet factor 4 (*PF4* or *CXCL4*), which encodes a protein secreted by activated platelets. PF4 binds the anticoagulant heparin, thereby promoting platelet aggregation and coagulation. In mice infected with influenza (H1N1), PF4 was involved in the recruitment of neutrophils, the modulation of inflammation in the lungs and displayed essential anti-viral properties³²⁹. The genes *PF4*, *PF4V1*, *CXCL5* and *PPBP*, which are neighbours on chromosome 4 (4q13.3), may be coregulated in convalescence. During acute SARS-CoV-2 infection in mice, *CXCL5* is expressed by lung epithelial cells to recruit neutrophils and induce inflammation, although it does not appear to modulate viral clearance³⁴⁵. *CXCL5* released by platelets and eosinophils may also contribute to the remodelling of connective tissues³⁴⁶ and limit the formation of atherosclerotic plaques³⁴⁷. Pro-platelet basic protein (*PPBP* or *CXCL7*) encodes a neutrophil chemoattractant released by activated platelets. Other genes upregulated in convalescence included *ALOX12*, which is highly expressed in platelets and

has roles in vascular processes³⁴⁸. Mutations within *ALOX12* are associated with thrombocytopenia³⁴⁹ and cardiovascular events^{350,351}, including hypertension and myocardial infarction. *ALOX12* has also been implicated in clotting, platelet aggregation, atherosclerosis and vascular injury^{352–355}. Also upregulated was *FSTL1*, which promotes the survival of endothelial cells³⁵⁶ and is associated with platelet count³⁵⁷, and *PTGS1*, which is involved in vasoconstriction and modulation of platelet aggregation³⁵⁸.

The genes that remained dysregulated following recovery from COVID-19 were overwhelmingly associated with vascular processes and platelet activation and aggregation. There is a large volume of literature describing the elevated risk of thrombotic events during COVID-19^{148,242,359}, however there is also an increasing body of evidence indicating that the risk of thrombo-embolism is increased beyond the acute phase of infection. One such study was carried out based on 1,449 individuals with positive SARS-CoV-2 tests and at least one thromboembolic event¹⁴³. Incident rate ratios indicated that the risk of pulmonary embolism was 3.5-fold higher between 28 and 56 days following COVID-19 diagnosis. Another study, based on 1,057,174 patients, found that the relative risk of pulmonary embolism was significantly increased for 110 days following COVID-19 onset, and the risk of deep vein thrombosis was also elevated for 70 days¹⁴⁴. The data presented in this chapter provide evidence of a molecular basis for the persistent risk of thromboembolic events observed following clinical recovery from COVID-19.

In summary, the analyses presented in this chapter identified many genes and proteins associated with COVID-19 infection and severity. Modelling temporal molecular changes during acute infection revealed activation of neutrophils and dynamic profiles of cytokines in severe disease. The plasma abundance of LRRC15, an alternative receptor for SARS-CoV-2, was able to discriminate between severe and non-severe patients. Furthermore, a comparison of samples taken before and after infection revealed that the expression of genes related to clotting do not return to normal levels following COVID-19.

Key findings:

- Transcriptomics and broad proteomic panels permitted the characterisation of temporal molecular changes following COVID-19 symptom onset.
- LRRC15, a coreceptor for SARS-CoV-2, was identified as a key predictor of severe COVID-19.

- Identification of a clotting signature that is elevated for months following COVID-19 onset.

Future directions:

- Detected transcriptomic expression patterns appear to largely reflect changes in cell abundances. These changes could be further dissected with single-cell RNA-seq.
- The measurement of LRRC15 in the lungs of COVID-19 patients could help elucidate its role in SARS-CoV-2 infection.

5 Conclusions and future directions

Modern technologies allow for the profiling of molecular signatures on an unprecedented scale, thereby presenting opportunities for the hypothesis-free systems-level characterisation of disease. In this thesis, I investigated the blood proteome and transcriptome in end-stage kidney disease (ESKD) patients with COVID-19. These patients were not able to shield during the lockdowns because they had to attend regular haemodialysis appointments, affording the opportunity to collect serial samples from both outpatients and inpatients throughout the course of COVID-19. These data permitted the characterisation of the temporal molecular response to COVID-19 in an at-risk group of patients.

I identified modules of genes and proteins with dynamic temporal profiles. Many of the gene modules appeared to reflect the activity of particular cell types in the PBMC fraction. A limitation of bulk RNA sequencing (RNA-seq) is that it cannot discern between changes in gene expression and cell type composition. Cell abundance imputation was performed to dissect changes in cell abundances, but this method does not provide definitive information on the relationship between observed transcripts and the cells they were originally expressed by. This motivates the use of single-cell RNA-seq to further investigate the transcriptomes of the patients in these cohorts. This approach would allow for a more detailed characterisation of the PBMC transcriptome. Such data has, in fact, already been generated for both the Wave 1 and Wave 2 cohorts; as of the completion of this thesis, these data are being analysed and promise to explore the pathways revealed by bulk transcriptomics.

I employed supervised learning to identify COVID-19 severity from molecular measurements. To do this, however, I included only one sample from each patient. Therefore, the sample sizes for these analyses were relatively small, limiting the data available for prediction. Furthermore, samples were not available at baseline for many patients. Therefore, these models were not necessarily predicting future severity, but instead the severity at the time of sampling. This limited the utility of these models for predicting future deterioration. It may be of interest for studies with larger numbers of participants to design models that are better placed to do this. Regardless, these models were able to identify key biomarkers. One particularly predictive protein was LRRC15, which fell in plasma abundance following COVID-19 onset in more severe patients but remained relatively stable in those with milder disease

courses. There is evidence from *in vitro* studies that this protein may act as an inhibitory receptor for SARS-CoV-2 replication and infection. While LRRC15 could be an important mediator of COVID-19, it remains unclear how it could suppress SARS-CoV-2 infection *in vivo* and how this relates to its abundance in plasma. It would be informative to understand the dynamics of its abundance in the lungs during COVID-19. While there have been proteomic studies of bronchoalveolar lavage fluid from COVID-19 patients, none have measured LRRC15, so this may be an avenue for future studies to explore.

In Chapter 2, I reported proteins with links to severe COVID-19. However, these data were observational, and so they were limited to identifying associations, rather than causal mediators of disease. In an attempt to overcome this limitation, Mendelian randomisation (MR) was performed to follow up proteins that were associated with COVID-19 severity in observational studies. While initial data suggested a role for FAS in severe COVID-19, this finding was not supported by a larger meta-analysis of COVID-19 patients. Currently, many MR studies that use molecular data do not attempt to replicate their discoveries in independent cohorts. The results of such studies are used to infer causal relationships and may be used as evidence to prioritise therapeutic targets. It is therefore vital that MR practitioners consider the robustness of their results, which may be bolstered by replication. This work also demonstrated how variants associated with protein abundance are not necessarily mediated through gene expression. The FAS pQTL likely reflected changes in the plasma abundance of the soluble form of FAS mediated through differential splicing of the exon encoding its transmembrane portion. It is important for studies attempting to discern the molecular mechanism through which genetic variants act to consider splicing, in addition to transcript and protein abundances. Additionally, it should be noted that multiplex binding assays can suffer from off-target binding leading to erroneous signals. Studies that measure proteins using these technologies should ideally attempt to verify the true target of the assay using alternative measurement systems.

ESKD patients are at higher risk of fatal COVID-19. Furthermore, the analyses presented in this thesis demonstrate distinct differences between ESKD patients on dialysis and healthy individuals. Therefore, there is no guarantee that the molecular response of these patients to COVID-19 reflects those of the general population. In general, comparisons to studies of general patient cohorts suggested that ESKD patients respond to COVID-19 in a similar way

to patients with normal kidney function. However, these simplistic comparisons were not sufficient to robustly compare the COVID-19 responses of individuals with normal renal function and those of ESKD patients. To investigate the ESKD-specific response to infection, COVID-19 patients with normal renal function would need to be sampled to provide a comparator group. However, there would have been logistical challenges for collecting such measurements in 2020 and 2021 when the Wave 1 and Wave 2 cohorts were respectively sampled. Sampling non-hospitalised patients was difficult during the lockdowns. If future studies attempt to investigate ESKD patients with COVID-19, it would be worthwhile to simultaneously consider patients with normal renal function. Such a study could characterise the ESKD-specific molecular response to COVID-19 by modelling the interaction between kidney status and COVID-19 outcomes.

Some ESKD patients were sampled both as non-infected controls in 2020 and during their acute COVID-19 episode in 2021. These data permitted an intra-individual comparison of samples taken from patients prior to and during COVID-19. Additionally, many of these patients were sampled following clinical recovery from COVID-19. This approach revealed that transcriptomic pathways related to clotting and inflammation did not return to baseline after recovery, which may help explain the elevated risk of thrombotic events following COVID-19. However, since these data included only 12 ESKD patients, it is unclear to what extent these findings apply to the general population. Furthermore, these samples were taken shortly after clinical recovery, so I could not ascertain how long these changes persisted. Larger systems-level studies should be carried out to robustly assess the long-term effects of COVID-19 in patients with normal renal function.

Many studies have considered molecular associations with eGFR and risk factors for developing kidney disease. However, few studies have directly studied the broad plasma proteome or PBMC transcriptome in ESKD patients on dialysis. In this thesis, I was able to characterise the proteomic effects of ESKD and dialysis by comparing these patients to healthy controls. ESKD and dialysis had pronounced effects on the blood multi-ome. However, the extent of the molecular dysregulation in ESKD compared to healthy individuals made interpretation of the data challenging. How can we prioritise relevant pathways if virtually everything is dysfunctional? Furthermore, how can we ascertain which effects are a result of kidney disease, as opposed to those that are the result of haemodialysis? Future studies

should attempt to disentangle these factors and identify pathways that are likely to drive disease. An ideal approach would be to collect serial samples from chronic kidney disease patients. This would permit molecular profiling to be performed as kidney function declines, thus revealing the sequence of changes in the blood multi-ome as CKD worsens. Following patients before and after the commencement of dialysis could additionally characterise both the short- and long-term molecular effects of this treatment.

In this thesis, I described the dynamics of plasma proteins following COVID-19 onset, facilitated by wide-angled proteomics assays. While the value of these technologies is clear, they are constrained by the need for reagents specific to each protein. Cost-effective technologies that measure the entire proteome would be able to facilitate truly hypothesis-free analysis. Furthermore, whereas RNA-seq can identify splicing variants, multiplex binding assays cannot differentiate protein isoforms or reveal post-translational modifications. Such capabilities would have been beneficial, for instance, in determining whether the FAS plasma pQTL represented its soluble or membrane-bound isoform. In this instance, the isoform could have been determined through targeted follow-up experiments, however it would be beneficial to obtain this information for the entire set of proteins measured. The further development of high-throughput proteomic technologies, such as the improvement of mass spectrometry-based approaches, may allow us to explore the entire proteome and clearly elucidate the identities of the proteins that have been measured.

6 Bibliography

1. Regenmortel, M. H. V. van. Reductionism and complexity in molecular biology. *EMBO Rep* **5**, 1016–1020 (2004).
2. Moraes, F. & Góes, A. A decade of human genome project conclusion: Scientific diffusion about our genome knowledge. *Biochemistry and Molecular Biology Education* **44**, 215–223 (2016).
3. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **412**, 565–566 (2001).
4. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol* **18**, 1–15 (2017).
5. Franklin, S. & Vondriska, T. M. Genomes, Proteomes, and the Central Dogma. *Circ Cardiovasc Genet* **4**, 576–576 (2011).
6. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
7. Dolinski, K. & Troyanskaya, O. G. Implications of Big Data for cell biology. *Mol Biol Cell* **26**, 2575–2578 (2015).
8. Berger, B., Daniels, N. M. & Yu, Y. W. Computational biology in the 21st century. *Commun ACM* **59**, 72–80 (2016).
9. Leonelli, S. The challenges of big data biology. *Elife* **8**, e47381 (2019).
10. Navarro, F. C. P. *et al.* Genomics and data science: an application within an umbrella. *Genome Biol* **20**, 109 (2019).
11. Stephens, Z. D. *et al.* Big data: Astronomical or genetical? *PLoS Biol* **13**, 1–11 (2015).
12. Sboner, A., Mu, X., Greenbaum, D., Auerbach, R. K. & Gerstein, M. B. The real cost of sequencing: higher than you think! *Genome Biol* **12**, 125 (2011).

13. Schwarze, K. *et al.* The complete costs of genome sequencing: a microcosting study in cancer and rare diseases from a single center in the United Kingdom. *Genetics in Medicine* **22**, 85–94 (2020).
14. Sperisen, P., Cominetti, O. & Martin, F. P. J. Longitudinal omics modeling and integration in clinical metabonomics research: Challenges in childhood metabolic health research. *Front Mol Biosci* **2**, 1–11 (2015).
15. Crick, F. Central Dogma of Molecular Biology. *Nature* **227**, 561–563 (1970).
16. Wetterstrand, K. A. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). www.genome.gov/sequencingcostsdata.
17. Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* **6**, 109–118 (2005).
18. Berry, D. A., Parmigiani, G., Sanchez, J., Schildkraut, J. & Winer, E. Probability of carrying a mutation of breast-ovarian cancer gene BRCA1 based on family history. *J Natl Cancer Inst* **89**, 227–38 (1997).
19. Chen, S. & Parmigiani, G. Meta-analysis of BRCA1 and BRCA2 penetrance. *J Clin Oncol* **25**, 1329–33 (2007).
20. Zhao, S., Zhang, Y., Gamini, R., Zhang, B. & von Schack, D. Evaluation of two main RNA-seq approaches for gene quantification in clinical RNA sequencing: polyA+ selection versus rRNA depletion. *Sci Rep* **8**, 4781 (2018).
21. Chen, C.-Y. A., Ezzeddine, N. & Shyu, A.-B. Messenger RNA half-life measurements in mammalian cells. *Methods Enzymol* **448**, 335–57 (2008).
22. Yang, E. *et al.* Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res* **13**, 1863–72 (2003).
23. Kleinman, C. L., Adoue, V. & Majewski, J. RNA editing of protein sequences: a rare event in human transcriptomes. *RNA* **18**, 1586–96 (2012).

24. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**, 1413–5 (2008).
25. Marra, M. A., Hillier, L. & Waterston, R. H. Expressed sequence tags--ESTablishing bridges between genomes. *Trends Genet* **14**, 4–7 (1998).
26. Boguski, M. S., Tolstoshev, C. M. & Bassett, D. E. Gene discovery in dbEST. *Science* **265**, 1993–4 (1994).
27. Alwine, J. C., Kemp, D. J. & Stark, G. R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc Natl Acad Sci U S A* **74**, 5350–4 (1977).
28. Becker-André, M. & Hahlbrock, K. Absolute mRNA quantification using the polymerase chain reaction (PCR). A novel approach by a PCR aided transcript titration assay (PATTY). *Nucleic Acids Res* **17**, 9437–46 (1989).
29. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57–63 (2009).
30. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5**, 621–628 (2008).
31. Karlsson, M. *et al.* A single-cell type transcriptomics map of human tissues. *Sci Adv* **7**, eabh2169 (2021).
32. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186–94 (1998).
33. Andrews, S. *et al.* FastQC: a quality control tool for high throughput sequence data. Preprint at <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (2012).
34. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal; Vol 17, No 1: Next Generation Sequencing Data Analysis* DO - 10.14806/ej.17.1.200 (2011).

35. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
36. Anders, S., Pyl, P. T. & Huber, W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
37. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
38. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417–419 (2017).
39. Pierson, E., Consortium, the Gte., Koller, D., Battle, A. & Mostafavi, S. Sharing and Specificity of Co-expression Networks across 35 Human Tissues. *PLoS Comput Biol* **11**, e1004220- (2015).
40. Huang, J., Zheng, J., Yuan, H. & McGinnis, K. Distinct tissue-specific transcriptional regulation revealed by gene regulatory networks in maize. *BMC Plant Biol* **18**, 111 (2018).
41. Melé, M. *et al.* The human transcriptome across tissues and individuals. *Science* (1979) **348**, 660–665 (2015).
42. Eraslan, G. *et al.* Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* (1979) **376**, eabl4290 (2023).
43. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLoS Comput Biol* **13**, e1005457- (2017).
44. Avican, K. *et al.* RNA atlas of human bacterial pathogens uncovers stress dynamics linked to infection. *Nat Commun* **12**, 3282 (2021).
45. Rambow, F. *et al.* Toward Minimal Residual Disease-Directed Therapy in Melanoma. *Cell* **174**, 843-855.e19 (2018).
46. Davenport, E. E. *et al.* Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study. *Lancet Respir Med* **4**, 259–71 (2016).

47. Mistry, P. *et al.* Transcriptomic, epigenetic, and functional analyses implicate neutrophil diversity in the pathogenesis of systemic lupus erythematosus. *Proceedings of the National Academy of Sciences* **116**, 25222–25228 (2019).
48. Lewis, M. J. *et al.* Molecular Portraits of Early Rheumatoid Arthritis Identify Clinical and Treatment Response Phenotypes. *Cell Rep* **28**, 2455–2470.e5 (2019).
49. Halperin, R. F. *et al.* Improved methods for RNAseq-based alternative splicing analysis. *Sci Rep* **11**, 10740 (2021).
50. Panousis, N. I. *et al.* Combined genetic and transcriptome analysis of patients with SLE: distinct, targetable signatures for susceptibility and severity. *Ann Rheum Dis* **78**, 1079 (2019).
51. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat Rev Genet* **13**, 227–232 (2012).
52. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
53. Vogel, C. *et al.* Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Mol Syst Biol* **6**, 400 (2010).
54. Uhlen, M. *et al.* Tissue-based map of the human proteome. *Science (1979)* **347**, 1260419–1260419 (2015).
55. Aslam, B., Basit, M., Nisar, M. A., Khurshid, M. & Rasool, M. H. Proteomics: Technologies and Their Applications. *J Chromatogr Sci* **55**, 182–196 (2017).
56. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).
57. Bekker-Jensen, D. B. *et al.* An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst* **4**, 587–599.e4 (2017).
58. Juncker, D., Bergeron, S., Laforte, V. & Li, H. Cross-reactivity in antibody microarrays and multiplexed sandwich assays: shedding light on the dark side of multiplexing. *Curr Opin Chem Biol* **18**, 29–37 (2014).

59. Gold, L. *et al.* Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5**, e15004 (2010).
60. Raffield, L. M. *et al.* Comparison of Proteomic Assessment Methods in Multiple Cohort Studies. *Proteomics* **20**, e1900278 (2020).
61. Pietzner, M. *et al.* Synergistic insights into human health from aptamer- and antibody-based proteomic profiling. *Nat Commun* **12**, 6822 (2021).
62. Katz, D. H. *et al.* Proteomic profiling platforms head to head: Leveraging genetics and clinical traits to compare aptamer- and antibody-based methods. *Sci Adv* **8**, eabm5164 (2023).
63. Billing, A. M. *et al.* Complementarity of SOMAscan to LC-MS/MS and RNA-seq for quantitative profiling of human embryonic and mesenchymal stem cells. *J Proteomics* **150**, 86–97 (2017).
64. Petrer, A. *et al.* Multiplatform Approach for Plasma Proteomics: Complementarity of Olink Proximity Extension Assay Technology to Mass Spectrometry-Based Protein Profiling. *J Proteome Res* **20**, 751–762 (2021).
65. Berbers, R.-M. *et al.* Targeted Proteomics Reveals Inflammatory Pathways that Classify Immune Dysregulation in Common Variable Immunodeficiency. *J Clin Immunol* **41**, 362–373 (2021).
66. Curci, D., Dillon, S. T., Gu, X., Winter, H. & Libermann, T. A. Proteome-Wide Analysis Using SOMAscan Identifies and Validates Epidermal Growth Factor as a Disease Marker of Collagenous Gastritis. *Gastro Hep Advances* **1**, 689–702 (2022).
67. Petrackova, A. *et al.* Serum protein pattern associated with organ damage and lupus nephritis in systemic lupus erythematosus revealed by PEA immunoassay. *Clin Proteomics* **14**, 32 (2017).
68. Kalla, R. *et al.* Serum proteomic profiling at diagnosis predicts clinical course, and need for intensification of treatment in inflammatory bowel disease. *J Crohns Colitis* **15**, 699–708 (2021).

69. Heier, C. R. *et al.* Identification of Pathway-Specific Serum Biomarkers of Response to Glucocorticoid and Infliximab Treatment in Children with Inflammatory Bowel Disease. *Clin Transl Gastroenterol* **7**, e192 (2016).
70. George, M. J. *et al.* Novel Insights Into the Effects of Interleukin 6 Antagonism in Non-ST-Segment-Elevation Myocardial Infarction Employing the SOMAscan Proteomics Platform. *J Am Heart Assoc* **9**, e015628 (2020).
71. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, 483 (2016).
72. Jostins, L., Pickrell, J. K., MacArthur, D. G. & Barrett, J. C. Misuse of hierarchical linear models overstates the significance of a reported association between OXTR and prosociality. *Proc Natl Acad Sci U S A* **109**, E1048 (2012).
73. Hoffman, G. E. & Roussos, P. Dream: powerful differential expression analysis for repeated measures designs. *Bioinformatics* **37**, 192–201 (2021).
74. Rivellese, F. *et al.* Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial. *Nat Med* **28**, 1256–1268 (2022).
75. Filbin, M. R. *et al.* Longitudinal proteomic analysis of severe COVID-19 reveals survival-associated signatures, tissue-specific cell death, and cell-cell interactions. *Cell Rep Med* **2**, 100287 (2021).
76. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
77. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
78. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29 (2014).
79. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).

80. Hawinkel, S., Rayner, J. C. W., Bijmans, L. & Thas, O. Sequence count data are poorly fit by the negative binomial distribution. *PLoS One* **15**, e0224909 (2020).
81. Yu, L., Fernandez, S. & Brock, G. Power analysis for RNA-Seq differential expression studies using generalized linear mixed effects models. *BMC Bioinformatics* **21**, 198 (2020).
82. Reeb, P. D. & Steibel, J. P. Evaluating statistical analysis models for RNA sequencing experiments. *Front Genet* **4**, 178 (2013).
83. Rocke, D. M. *et al.* Excess False Positive Rates in Methods for Differential Gene Expression Analysis using RNA-Seq Data. *bioRxiv* 020784 (2015) doi:10.1101/020784.
84. Assefa, A. T. *et al.* Differential gene expression analysis tools exhibit substandard performance for long non-coding RNA-sequencing data. *Genome Biol* **19**, 96 (2018).
85. Baik, B., Yoon, S. & Nam, D. Benchmarking RNA-seq differential expression analysis methods using spike-in and simulation data. *PLoS One* **15**, e0232271- (2020).
86. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics* **24**, 1202–1205 (2016).
87. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300 (1995).
88. Yekutieli, D. & Benjamini, Y. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *J Stat Plan Inference* **82**, 171–196 (1999).
89. Ibrahim, J. G., Chu, H. & Chen, L. M. Basic concepts and methods for joint models of longitudinal and survival data. *J Clin Oncol* **28**, 2796–801 (2010).
90. Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E. & Groothuis-Oudshoorn, C. G. M. Time-varying covariates and coefficients in Cox regression models. *Ann Transl Med* **6**, 121 (2018).

91. Rizopoulos, D. JM: An R package for the joint modelling of longitudinal and time-to-event data. *J Stat Softw* **35**, 1–33 (2010).
92. Rizopoulos, D. The R Package Jmbayes for Fitting Joint Models for Longitudinal and Time-to-Event Data Using MCMC. *J Stat Softw* **72**, 1–46 (2016).
93. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58**, 267–288 (1996).
94. Ghosh, D. & Chinnaiyan, A. M. Classification and selection of biomarkers in genomic data using LASSO. *J Biomed Biotechnol* **2005**, 147–54 (2005).
95. Frost, H. R. & Amos, C. I. Gene set selection via LASSO penalized regression (SLPR). *Nucleic Acids Res* **45**, e114 (2017).
96. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, (2008).
97. Chen, X. & Ishwaran, H. Random forests for genomic data analysis. *Genomics* **99**, 323–9 (2012).
98. Díaz-Uriarte, R. & Alvarez de Andrés, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **7**, 3 (2006).
99. Xiang, R. *et al.* A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data. *Front Genet* **12**, 646936 (2021).
100. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv* (2018).
101. Chari, T. & Pachter, L. The Specious Art of Single-Cell Genomics. *bioRxiv* 2021.08.25.457696 (2022) doi:10.1101/2021.08.25.457696.
102. Argelaguet, R. *et al.* Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol* **14**, e8124 (2018).
103. Argelaguet, R. *et al.* MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol* **21**, 111 (2020).

104. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data. *Mach Learn* **52**, 91–118 (2003).
105. Şenbabaoğlu, Y., Michailidis, G. & Li, J. Z. Critical limitations of consensus clustering in class discovery. *Sci Rep* **4**, 6207 (2014).
106. John, C. R. *et al.* M3C: Monte Carlo reference-based consensus clustering. *Sci Rep* **10**, 1–14 (2020).
107. Banchereau, R. *et al.* Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* **165**, 551–65 (2016).
108. Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies: A guide, glossary, and checklist for clinicians. *BMJ (Online)* **362**, (2018).
109. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum Mol Genet* **27**, R195–R208 (2018).
110. Giambartolomei, C. *et al.* Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. *PLoS Genet* **10**, e1004383- (2014).
111. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet* **17**, e1009440- (2021).
112. Fortune, M. D. *et al.* Statistical colocalization of genetic risk variants for related autoimmune diseases in the context of common controls. *Nat Genet* **47**, 839–846 (2015).
113. Wallace, C. Eliciting priors and relaxing the single causal variant assumption in colocalisation analyses. *bioRxiv* 1–20 (2019) doi:10.1101/838946.
114. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* **106**, 9362–9367 (2009).
115. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–60 (2015).

116. Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet* **99**, 1245–1260 (2016).
117. Li, L. *et al.* IG20/MADD plays a critical role in glucose-induced insulin secretion. *Diabetes* **63**, 1612–23 (2014).
118. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481–487 (2016).
119. Plenge, R. M. *et al.* TRAF1–C5 as a Risk Locus for Rheumatoid Arthritis — A Genomewide Study. *New England Journal of Medicine* **357**, 1199–1209 (2007).
120. Tsitsikov, E. N. *et al.* TRAF1 Is a Negative Regulator of TNF Signaling: Enhanced TNF Signaling in TRAF1-Deficient Mice. *Immunity* **15**, 647–657 (2001).
121. Gilbert, J. C. *et al.* First-in-Human Evaluation of Anti–von Willebrand Factor Therapeutic Aptamer ARC1779 in Healthy Volunteers. *Circulation* **116**, 2678–2686 (2007).
122. Wang, L., Wang, Y., Ye, D. & Liu, Q. Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence. *Int J Antimicrob Agents* **55**, 105948 (2020).
123. Lavezzo, E. *et al.* Suppression of COVID-19 outbreak in the municipality of Vo, Italy. *medRxiv* 2020.04.17.20053157 (2020) doi:10.1101/2020.04.17.20053157.
124. Long, Q. *et al.* Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections. *Nat Med* (2020) doi:10.1038/s41591-020-0965-6.
125. Lamers, M. M. & Haagmans, B. L. SARS-CoV-2 pathogenesis. *Nat Rev Microbiol* **20**, 270–284 (2022).
126. Garg, S. *et al.* Hospitalization Rates and Characteristics of Patients Hospitalized with Laboratory-Confirmed Coronavirus Disease 2019 — COVID-NET, 14 States, March 1–30, 2020. *MMWR Morb Mortal Wkly Rep* **69**, 458–464 (2020).
127. Docherty, A. B. *et al.* Features of 20 133 UK patients in hospital with covid-19 using the ISARIC WHO Clinical Characterisation Protocol: prospective observational cohort study. *BMJ* **369**, m1985 (2020).
128. Deaths in United Kingdom. <https://coronavirus.data.gov.uk/details/deaths>.

129. Pascarella, G. *et al.* COVID-19 diagnosis and management: a comprehensive review. *J Intern Med* **288**, 192–206 (2020).
130. Li, X. *et al.* Clinical determinants of the severity of COVID-19: A systematic review and meta-analysis. *PLoS One* **16**, e0250602- (2021).
131. Shang, J. *et al.* Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences* **117**, 11727–11734 (2020).
132. Alipoor, S. D. *et al.* COVID-19: Molecular and Cellular Response. *Front Cell Infect Microbiol* **11**, (2021).
133. Jackson, C. B., Farzan, M., Chen, B. & Choe, H. Mechanisms of SARS-CoV-2 entry into cells. *Nat Rev Mol Cell Biol* **23**, 3–20 (2022).
134. Mason, R. J. Pathogenesis of COVID-19 from a cell biology perspective. *European Respiratory Journal* **55**, 2000607 (2020).
135. Xu, H. *et al.* High expression of ACE2 receptor of 2019-nCoV on the epithelial cells of oral mucosa. *Int J Oral Sci* **12**, 8 (2020).
136. Han, L. *et al.* SARS-CoV-2 ORF9b antagonizes type I and III interferons by targeting multiple components of the RIG-I/MDA-5–MAVS, TLR3–TRIF, and cGAS–STING signaling pathways. *J Med Virol* **93**, 5376–5389 (2021).
137. Li, J.-Y. *et al.* The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. *Virus Res* **286**, 198074 (2020).
138. Leisman, D. E. *et al.* Cytokine elevation in severe and critical COVID-19: a rapid systematic review, meta-analysis, and comparison with other inflammatory syndromes. *Lancet Respir Med* **8**, 1233–1244 (2020).
139. Chen, L. Y. C., Hoiland, R. L., Stukas, S., Wellington, C. L. & Sekhon, M. S. Confronting the controversy: interleukin-6 and the COVID-19 cytokine storm syndrome. *European Respiratory Journal* **56**, 2003006 (2020).
140. Meyer, N. J., Gattinoni, L. & Calfee, C. S. Acute respiratory distress syndrome. *The Lancet* **398**, 622–637 (2021).

141. Li, X., Geng, M., Peng, Y., Meng, L. & Lu, S. Molecular immune pathogenesis and diagnosis of COVID-19. *J Pharm Anal* **10**, 102–108 (2020).
142. Varga, Z. *et al.* Endothelial cell infection and endotheliitis in COVID-19. *The Lancet* **395**, 1417–1418 (2020).
143. Ho, F. K. *et al.* Thromboembolic Risk in Hospitalized and Nonhospitalized COVID-19 Patients: A Self-Controlled Case Series Analysis of a Nationwide Cohort. *Mayo Clin Proc* **96**, 2587–2597 (2021).
144. Katsoularis, I. *et al.* Risks of deep vein thrombosis, pulmonary embolism, and bleeding after covid-19: nationwide self-controlled cases series and matched cohort study. *BMJ* **377**, (2022).
145. Goshua, G. *et al.* Endotheliopathy in COVID-19-associated coagulopathy: evidence from a single-centre, cross-sectional study. *Lancet Haematol* **7**, e575–e582 (2020).
146. Guervilly, C. *et al.* Circulating Endothelial Cells as a Marker of Endothelial Injury in Severe COVID -19. *J Infect Dis* **222**, 1789–1793 (2020).
147. Bonaventura, A. *et al.* Endothelial dysfunction and immunothrombosis as key pathogenic mechanisms in COVID-19. *Nat Rev Immunol* **21**, 319–329 (2021).
148. Berger, J. S. *et al.* Prevalence and Outcomes of D-Dimer Elevation in Hospitalized Patients With COVID-19. *Arterioscler Thromb Vasc Biol* (2020) doi:10.1161/atvbaha.120.314872.
149. Leppkes, M. *et al.* Vascular occlusion by neutrophil extracellular traps in COVID-19. *EBioMedicine* **58**, (2020).
150. Michels, A. *et al.* Histones link inflammation and thrombosis through the induction of Weibel–Palade body exocytosis. *Journal of Thrombosis and Haemostasis* **14**, 2274–2286 (2016).
151. Fuchs, T. A. *et al.* Extracellular DNA traps promote thrombosis. *Proceedings of the National Academy of Sciences* **107**, 15880–15885 (2010).

152. Ballering, A. V, van Zon, S. K. R., olde Hartman, T. C. & Rosmalen, J. G. M. Persistence of somatic symptoms after COVID-19 in the Netherlands: an observational cohort study. *The Lancet* **400**, 452–461 (2022).
153. Davis, H. E., McCorkell, L., Vogel, J. M. & Topol, E. J. Long COVID: major findings, mechanisms and recommendations. *Nat Rev Microbiol* **21**, 133–146 (2023).
154. Lopez-Leon, S. *et al.* More than 50 long-term effects of COVID-19: a systematic review and meta-analysis. *Sci Rep* **11**, 16144 (2021).
155. Kedor, C. *et al.* A prospective observational study of post-COVID-19 chronic fatigue syndrome following the first pandemic wave in Germany and biomarkers associated with symptom severity. *Nat Commun* **13**, 5104 (2022).
156. Murakami, N. *et al.* Therapeutic advances in COVID-19. *Nat Rev Nephrol* **19**, 38–52 (2023).
157. Gottlieb, R. L. *et al.* Early Remdesivir to Prevent Progression to Severe Covid-19 in Outpatients. *New England Journal of Medicine* **386**, 305–315 (2021).
158. Hammond, J. *et al.* Oral Nirmatrelvir for High-Risk, Nonhospitalized Adults with Covid-19. *New England Journal of Medicine* **386**, 1397–1408 (2022).
159. Horby, P. *et al.* Dexamethasone in Hospitalized Patients with Covid-19. *N Engl J Med* **384**, 693–704 (2021).
160. Abani, O. *et al.* Tocilizumab in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. *The Lancet* **397**, 1637–1645 (2021).
161. Coomes, E. A. & Haghbayan, H. Interleukin-6 in Covid-19: A systematic review and meta-analysis. *Rev Med Virol* **30**, e2141 (2020).
162. Gordon, A. C. *et al.* Interleukin-6 Receptor Antagonists in Critically Ill Patients with Covid-19. *N Engl J Med* **384**, 1491–1502 (2021).
163. Office for National Statistics (ONS). COVID-19 vaccine effectiveness estimated using Census 2021 variables, England: 31 March 2021 to 20 March 2022.

<https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/deaths/bulletins/covid19vaccineeffectivenessestimatedusingcensus2021variablesengland/31march2021to20march2022> (2023).

164. Notarte, K. I. *et al.* Impact of COVID-19 vaccination on the risk of developing long-COVID and on existing long-COVID symptoms: A systematic review. *EClinicalMedicine* **53**, (2022).
165. Romagnani, P. *et al.* Chronic kidney disease. *Nat Rev Dis Primers* **3**, 17088 (2017).
166. Laouari, D. *et al.* TGF- α mediates genetic susceptibility to chronic kidney disease. *J Am Soc Nephrol* **22**, 327–35 (2011).
167. Helal, I., Fick-Brosnahan, G. M., Reed-Gitomer, B. & Schrier, R. W. Glomerular hyperfiltration: definitions, mechanisms and clinical implications. *Nat Rev Nephrol* **8**, 293–300 (2012).
168. Hill, N. R. *et al.* Global Prevalence of Chronic Kidney Disease – A Systematic Review and Meta-Analysis. *PLoS One* **11**, e0158765- (2016).
169. Köttgen, A. *et al.* Multiple loci associated with indices of renal function and chronic kidney disease. *Nat Genet* **41**, 712–717 (2009).
170. Valeri, A. M. *et al.* Presentation and Outcomes of Patients with ESKD and COVID-19. *J Am Soc Nephrol* **31**, 1409–1415 (2020).
171. Williamson, E. J. *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430–436 (2020).
172. Kato, S. *et al.* Aspects of immune dysfunction in end-stage renal disease. *Clin J Am Soc Nephrol* **3**, 1526–33 (2008).
173. Mastalerz-Migas, A., Gwiazda, E. & Brydak, L. B. Effectiveness of influenza vaccine in patients on hemodialysis--a review. *Med Sci Monit* **19**, 1013–8 (2013).
174. Chen, J.-J. *et al.* Immunogenicity Rates After SARS-CoV-2 Vaccination in People With End-stage Kidney Disease: A Systematic Review and Meta-analysis. *JAMA Netw Open* **4**, e2131749 (2021).

175. Anand, S. *et al.* Antibody Response to COVID-19 Vaccination in Patients Receiving Dialysis. *J Am Soc Nephrol* **32**, 2435–2438 (2021).
176. Gisby, J. *et al.* Longitudinal proteomic profiling of dialysis patients with COVID-19 reveals markers of severity and predictors of death. *Elife* **10**, 2020.11.05.20223289 (2021).
177. Lucas, C. *et al.* Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* **584**, 463–469 (2020).
178. Haljasmägi, L. *et al.* Longitudinal proteomic profiling reveals increased early inflammation and sustained apoptosis proteins in severe COVID-19. *Sci Rep* **10**, 20533 (2020).
179. Shu, T. *et al.* Plasma Proteomics Identify Biomarkers and Pathogenesis of COVID-19. *Immunity* **53**, 1108-1122.e5 (2020).
180. Rodriguez, L. *et al.* Systems-Level Immunomonitoring from Acute to Recovery Phase of Severe COVID-19. *Cell Rep Med* **1**, 100078 (2020).
181. Demichev, V. *et al.* A time-resolved proteomic and prognostic map of COVID-19. *Cell Syst* **12**, 780-794.e7 (2021).
182. Thwaites, R. S. *et al.* Inflammatory profiles across the spectrum of disease reveal a distinct role for GM-CSF in severe COVID-19. *Sci Immunol* **6**, eabg9873 (2021).
183. Yang, J. *et al.* Impact of Kidney Function on the Blood Proteome and on Protein Cardiovascular Risk Biomarkers in Patients With Stable Coronary Heart Disease. *J Am Heart Assoc* **9**, e016463 (2020).
184. Tang, J., Liu, N. & Zhuang, S. Role of epidermal growth factor receptor in acute and chronic kidney injury. *Kidney Int* **83**, 804–810 (2013).
185. Matías-García, P. R. *et al.* Plasma Proteomics of Renal Function: A Transethnic Meta-Analysis and Mendelian Randomization Study. *J Am Soc Nephrol* **32**, 1747–1763 (2021).

186. Carlsson, A. C. *et al.* Use of Proteomics To Investigate Kidney Function Decline over 5 Years. *Clin J Am Soc Nephrol* **12**, 1226–1235 (2017).
187. Christensson, A. *et al.* The Impact of the Glomerular Filtration Rate on the Human Plasma Proteome. *Proteomics Clin Appl* **12**, e1700067 (2018).
188. Niewczasz, M. A. *et al.* A signature of circulating inflammatory proteins and development of end-stage renal disease in diabetes. *Nat Med* **25**, 805–813 (2019).
189. Grams, M. E. *et al.* Proteins Associated with Risk of Kidney Function Decline in the General Population. *Journal of the American Society of Nephrology* **32**, (2021).
190. Gold, L. *et al.* Aptamer-based multiplexed proteomic technology for biomarker discovery. *PLoS One* **5**, e15004 (2010).
191. Romanova, Y. *et al.* Proteomic Analysis of Human Serum from Patients with Chronic Kidney Disease. *Biomolecules* **10**, (2020).
192. Martínez-Alonso, E. *et al.* Proteomic analysis of plasma proteins of high-flux haemodialysis and on-line haemodiafiltration patients reveals differences in transthyretin levels related with anaemia. *Sci Rep* **10**, 16029 (2020).
193. Han, S. *et al.* Proteomics investigation of the changes in serum proteins after high- and low-flux hemodialysis. *Ren Fail* **40**, 506–513 (2018).
194. Gandolfo, L. C. & Speed, T. P. RLE plots: Visualizing unwanted variation in high dimensional data. *PLoS One* **13**, 1–9 (2018).
195. Kuhn, M. Building Predictive Models in R Using the caret Package. *J Stat Softw* **28**, 1–26 (2008).
196. John Fox & Sanford Weisberg. *An R Companion to Applied Regression*. (Sage, 2019).
197. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest Package: Tests in Linear Mixed Effects Models. *J Stat Softw* **82**, (2017).
198. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* **67**, (2015).

199. Bakdash, J. Z. & Marusich, L. R. Repeated measures correlation. *Front Psychol* **8**, 1–13 (2017).
200. Hastie T, Tibshirani R, F. J. High-Dimensional Problems: $p \gg N$. in *The Elements of Statistical Learning* 649–698 (Springer Series in Statistics, 2001).
201. Filbin, M. R. *et al.* Plasma proteomics reveals tissue-specific cell death and mediators of cell-cell interactions in severe COVID-19 patients. *bioRxiv* 2020.11.02.365536 (2020) doi:10.1101/2020.11.02.365536.
202. Leo, B. Random forests. *Mach Learn* **45**, 5–32 (2001).
203. Paluzynska, A., Biecek, P. & Jiang, Y. randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance. Preprint at <https://cran.r-project.org/package=randomForestExplainer> (2020).
204. Perperoglou, A., Sauerbrei, W., Abrahamowicz, M. & Schmid, M. A review of spline function procedures in R. *BMC Med Res Methodol* **19**, 1–16 (2019).
205. Szklarczyk, D. *et al.* STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607–D613 (2019).
206. Wu, C. *et al.* Risk Factors Associated With Acute Respiratory Distress Syndrome and Death in Patients With Coronavirus Disease 2019 Pneumonia in Wuhan, China. *JAMA Intern Med* **180**, 934–943 (2020).
207. Li, X. *et al.* Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *J Allergy Clin Immunol* **146**, 110–118 (2020).
208. Fraser, S. D. S. *et al.* The burden of comorbidity in people with chronic kidney disease stage 3: a cohort study. *BMC Nephrol* **16**, 193 (2015).
209. Wattanakit, K. & Cushman, M. Chronic kidney disease and venous thromboembolism: epidemiology and mechanisms. *Curr Opin Pulm Med* **15**, 408–12 (2009).
210. Carrero, J. J. & Stenvinkel, P. Inflammation in end-stage renal disease--what have we learned in 10 years? *Semin Dial* **23**, 498–509 (2010).

211. Naseeb, U. *et al.* Complementary LC-MS/MS Proteomic Analysis of Uremic Plasma Proteins. *J Coll Physicians Surg Pak* **25**, 606–609 (2015).
212. Christensson, A. *et al.* The Impact of the Glomerular Filtration Rate on the Human Plasma Proteome. *Proteomics Clin Appl* **12**, e1700067 (2018).
213. Lind, L. *et al.* Longitudinal effects of aging on plasma proteins levels in older adults - associations with kidney function and hemoglobin levels. *PLoS One* **14**, e0212060 (2019).
214. Fliser, D. *et al.* Fibroblast Growth Factor 23 (FGF23) Predicts Progression of Chronic Kidney Disease: The Mild to Moderate Kidney Disease (MMKD) Study. *Journal of the American Society of Nephrology* **18**, (2007).
215. El-Achkar, T. M. *et al.* Tamm-Horsfall protein protects the kidney from ischemic injury by decreasing inflammation and altering TLR4 expression. *Am J Physiol Renal Physiol* **295**, F534-44 (2008).
216. Shen, B. *et al.* Proteomic and Metabolomic Characterization of COVID-19 Patient Sera. *Cell* **182**, 59-72.e15 (2020).
217. Arunachalam, P. S. *et al.* Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science (1979)* **6261**, 1–18 (2020).
218. Ibrahim, J. G., Chu, H. & Chen, L. M. Basic concepts and methods for joint models of longitudinal and survival data. *J Clin Oncol* **28**, 2796–2801 (2010).
219. Guan, W. *et al.* Clinical characteristics of coronavirus disease 2019 in China. *New England Journal of Medicine* **382**, 1708–1720 (2020).
220. Chen, K. *et al.* SARS-CoV-2 Nucleocapsid Protein Interacts with RIG-I and Represses RIG-Mediated IFN- β Production. *Viruses* **13**, (2021).
221. Blot, M. *et al.* CXCL10 could drive longer duration of mechanical ventilation during COVID-19 ARDS. *Crit Care* **24**, 632 (2020).
222. Peralta, C. A. *et al.* Cystatin C identifies chronic kidney disease patients at higher risk for complications. *J Am Soc Nephrol* **22**, 147–55 (2011).

223. Gordin, E. *et al.* Urinary clusterin and cystatin B as biomarkers of tubular injury in dogs following envenomation by the European adder. *Res Vet Sci* **134**, 12–18 (2021).
224. Yerramilli, M., Farace, G., Quinn, J. & Yerramilli, M. Kidney Disease and the Nexus of Chronic Kidney Disease and Acute Kidney Injury: The Role of Novel Biomarkers as Early and Accurate Diagnostics. *Vet Clin North Am Small Anim Pract* **46**, 961–93 (2016).
225. Rajwani, A. *et al.* Increasing circulating IGFBP1 levels improves insulin sensitivity, promotes nitric oxide production, lowers blood pressure, and protects against atherosclerosis. *Diabetes* **61**, 915–24 (2012).
226. Chen, Z., Downing, S. & Tzanakakis, E. S. Four Decades After the Discovery of Regenerating Islet-Derived (Reg) Proteins: Current Understanding and Challenges. *Front Cell Dev Biol* **7**, (2019).
227. Srisawasdi, P., Vanavanan, S., Charoenpanichkit, C. & Kroll, M. H. The Effect of Renal Dysfunction on BNP, NT-proBNP, and Their Ratio. *Am J Clin Pathol* **133**, 14–23 (2010).
228. Rodríguez-Calvo, R. *et al.* Fatty acid binding protein 4 (FABP4) as a potential biomarker reflecting myocardial lipid storage in type 2 diabetes. *Metabolism* **96**, 12–21 (2019).
229. Zhang, Y. *et al.* High expression of FABP4 and FABP6 in patients with colorectal cancer. *World J Surg Oncol* **17**, 171 (2019).
230. Chan, D. T. *et al.* Chronic kidney disease delays VLDL-apoB-100 particle catabolism: potential role of apolipoprotein C-III. *J Lipid Res* **50**, 2524–31 (2009).
231. Harel, M., Girard-Guyonvarc'h, C., Rodriguez, E., Palmer, G. & Gabay, C. Production of IL-18 Binding Protein by Radiosensitive and Radioresistant Cells in CpG-Induced Macrophage Activation Syndrome. *The Journal of Immunology* **205**, 1167–1175 (2020).
232. Saha, S. K. *et al.* KRT19 directly interacts with β -catenin/RAC1 complex to regulate NUMB-dependent NOTCH signaling pathway and breast cancer properties. *Oncogene* **36**, 332–349 (2017).

233. Saftig, P. *et al.* Mice deficient for the lysosomal proteinase cathepsin D exhibit progressive atrophy of the intestinal mucosa and profound destruction of lymphoid cells. *EMBO J* **14**, 3599–3608 (1995).
234. Henning, R. J., Bourgeois, M. & Harbison, R. D. Poly(ADP-ribose) Polymerase (PARP) and PARP Inhibitors: Mechanisms of Action and Role in Cardiovascular Disorders. *Cardiovasc Toxicol* **18**, 493–506 (2018).
235. Zimmers, T. A. *et al.* Growth differentiation factor-15/macrophage inhibitory cytokine-1 induction after kidney and lung injury. *Shock* **23**, (2005).
236. Kragstrup, T. W. *et al.* Plasma ACE2 predicts outcome of COVID-19 in hospitalized patients. *PLoS One* **16**, e0252799- (2021).
237. Imai, Y. *et al.* Angiotensin-converting enzyme 2 protects from severe acute lung failure. *Nature* **436**, 112–116 (2005).
238. Schulte-Schrepping, J. *et al.* Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. *Cell* **182**, 1419-1440.e23 (2020).
239. Wittkowski, H. *et al.* Neutrophil-derived S100A12 in acute lung injury and respiratory distress syndrome. *Crit Care Med* **35**, (2007).
240. Bansal, R. *et al.* Integrin alpha 11 in the regulation of the myofibroblast phenotype: implications for fibrotic diseases. *Exp Mol Med* **49**, e396–e396 (2017).
241. D'Agnillo, F. *et al.* Lung epithelial and endothelial damage, loss of tissue repair, inhibition of fibrinolysis, and cellular senescence in fatal COVID-19. *Sci Transl Med* **13**, eabj7790 (2023).
242. Gorog, D. A. *et al.* Current and novel biomarkers of thrombotic risk in COVID-19: a Consensus Statement from the International COVID-19 Thrombosis Biomarkers Colloquium. *Nat Rev Cardiol* 1–21 (2022) doi:10.1038/s41569-021-00665-7.
243. Su, C.-Y. *et al.* Circulating proteins to predict adverse COVID-19 outcomes. *medRxiv* 2021.10.04.21264015 (2021) doi:10.1101/2021.10.04.21264015.

244. Duijvelaar, E., Gisby, J., Peters, J., Bogaard, H. J. & Aman, J. Longitudinal plasma proteomics reveals alveolar-capillary barrier disruption in critically ill COVID-19 patients. *Res Sq* (2023) doi:10.21203/rs.3.rs-2872132/v1.
245. Long, J. D. & Mills, J. A. Joint modeling of multivariate longitudinal data and survival data in several observational studies of Huntington's disease. *BMC Med Res Methodol* **18**, 138 (2018).
246. Guillon, A. *et al.* Neutrophil proteases alter the interleukin-22-receptor-dependent lung antimicrobial defence. *European Respiratory Journal* **46**, 771 (2015).
247. Gonzalez-Rey, E., Chorny, A., Varela, N., Robledo, G. & Delgado, M. Urocortin and Adrenomedullin Prevent Lethal Endotoxemia by Down-Regulating the Inflammatory Response. *Am J Pathol* **168**, 1921–1930 (2006).
248. Elke, G. *et al.* The use of mid-regional proadrenomedullin to identify disease severity and treatment response to sepsis - a secondary analysis of a large randomised controlled trial. *Crit Care* **22**, 79 (2018).
249. Maecker, H. *et al.* TWEAK attenuates the transition from innate to adaptive immunity. *Cell* **123**, 931–944 (2005).
250. Burkly, L. C., Michaelson, J. S., Hahm, K., Jakubowski, A. & Zheng, T. S. TWEAKing tissue remodeling by a multifunctional cytokine: Role of TWEAK/Fn14 pathway in health and disease. *Cytokine* **40**, 1–16 (2007).
251. Donlan, A. N. *et al.* IL-13 is a driver of COVID-19 severity. *JCI Insight* **6**, (2021).
252. Minty, A. *et al.* Interleukin-13 is a new human lymphokine regulating inflammatory and immune responses. *Nature* **362**, 248–250 (1993).
253. Klaric, L. *et al.* Mendelian randomisation identifies alternative splicing of the FAS death receptor as a mediator of severe COVID-19. *medRxiv* (2021) doi:10.1101/2021.04.01.21254789.
254. Pairo-Castineira, E. *et al.* Genetic mechanisms of critical illness in COVID-19. *Nature* **591**, 92–98 (2021).

255. Folkersen, L. *et al.* Genomic and drug target evaluation of 90 cardiovascular proteins in 30,931 individuals. *Nat Metab* **2**, 1135–1148 (2020).
256. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192–1212 (2014).
257. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
258. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, e34408 (2018).
259. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol* **46**, 1734–1739 (2017).
260. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-015-0047–8 (2015).
261. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet* **23**, R89–R98 (2014).
262. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
263. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data. *Genet Epidemiol* **37**, 658–665 (2013).
264. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol* **44**, 512–525 (2015).
265. Peters, J. E. *et al.* Insight into Genotype-Phenotype Associations through eQTL Mapping in Multiple Cell Types in Health and Immune-Mediated Disease. *PLoS Genet* **12**, e1005908 (2016).

266. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat Genet* **50**, 151–158 (2018).
267. Taylor-Weiner, A. *et al.* Scaling computational genomics to millions of individuals with GPUs. *Genome Biol* **20**, 228 (2019).
268. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur J Hum Genet* **28**, 715–718 (2020).
269. Raffield, L. M. *et al.* Comparison of Proteomic Assessment Methods in Multiple Cohort Studies. *Proteomics* **20**, 1900278 (2020).
270. Abraham, M. K. & Chang, W.-T. W. Subarachnoid Hemorrhage. *Emerg Med Clin North Am* **34**, 901–916 (2016).
271. null, null *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (1979)* **348**, 648–660 (2015).
272. Vösa, U. *et al.* Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet* **53**, 1300–1310 (2021).
273. Baeza-Centurion, P., Miñana, B., Schmiedel, J. M., Valcárcel, J. & Lehner, B. Combinatorial Genetics Reveals a Scaling Law for the Effects of Mutations on Splicing. *Cell* **176**, 549-563.e23 (2019).
274. Paronetto, M. P., Passacantilli, I. & Sette, C. Alternative splicing and cell survival: from tissue homeostasis to disease. *Cell Death Differ* **23**, 1919–1929 (2016).
275. Cascino, I., Fiucci, G., Papoff, G. & Ruberti, G. Three functional soluble forms of the human apoptosis-inducing Fas molecule are produced by alternative splicing. *J Immunol* **154**, 2706–13 (1995).
276. Gisby, J. S. *et al.* Multi-omics identify falling LRRC15 as a COVID-19 severity marker and persistent pro-thrombotic signals in convalescence. *Nat Commun* **13**, 7775 (2022).

277. Mattisson, I. Y. *et al.* Elevated Markers of Death Receptor-Activated Apoptosis are Associated with Increased Risk for Development of Diabetes and Cardiovascular Disease. *EBioMedicine* **26**, 187–197 (2017).
278. Zhou, S. *et al.* A Neanderthal OAS1 isoform protects individuals of European ancestry against COVID-19 susceptibility and severity. *Nat Med* **27**, 659–667 (2021).
279. Anisul, M. *et al.* A proteome-wide genetic investigation identifies several SARS-CoV-2-exploited host targets of clinical relevance. *Elife* **10**, e69719 (2021).
280. Gutmann, C. *et al.* SARS-CoV-2 RNAemia and proteomic trajectories inform prognostication in COVID-19 patients admitted to intensive care. *Nat Commun* **12**, 3406 (2021).
281. Galbraith, M. D. *et al.* Seroconversion stages COVID19 into distinct pathophysiological states. *Elife* **10**, e65508 (2021).
282. Paranjpe, I. *et al.* Proteomic Characterization of Acute Kidney Injury in Patients Hospitalized with SARS-CoV2 Infection. *medRxiv* 2021.12.09.21267548 (2021) doi:10.1101/2021.12.09.21267548.
283. Sullivan, K. D. *et al.* The COVIDome Explorer researcher portal. *Cell Rep* **36**, (2021).
284. Whitaker, M. *et al.* Variant-specific symptoms of COVID-19 in a study of 1,542,510 adults in England. *Nat Commun* **13**, 6856 (2022).
285. Zhuang, Q. *et al.* Single-Cell Transcriptomic Analysis of Peripheral Blood Reveals a Novel B-Cell Subset in Renal Allograft Recipients With Accommodation. *Front Pharmacol* **12**, 706580 (2021).
286. Dorr, C. *et al.* Differentially Expressed Gene Transcripts Using RNA Sequencing from the Blood of Immunosuppressed Kidney Allograft Recipients. *PLoS One* **10**, e0125045- (2015).
287. Wu, H. *et al.* Single-Cell RNA and ATAC Sequencing Reveal Hemodialysis-Related Immune Dysregulation of Circulating Immune Cell Subpopulations. *Front Immunol* **13**, 878226 (2022).

288. Luo, T. *et al.* A single-cell map for the transcriptomic signatures of peripheral blood mononuclear cells in end-stage renal disease. *Nephrology Dialysis Transplantation* **36**, 599–608 (2021).
289. Ewels, P. A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* **38**, 276–278 (2020).
290. Di Tommaso, P. *et al.* Nextflow enables reproducible computational workflows. *Nat Biotechnol* **35**, 316–319 (2017).
291. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25 (2010).
292. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res* **49**, D884–D891 (2021).
293. Tweedie, S. *et al.* Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Res* **49**, D939–D946 (2021).
294. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* **14**, 7 (2013).
295. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).
296. Wilk, A. J. *et al.* A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat Med* **26**, 1070–1076 (2020).
297. Buang, N. *et al.* Type I interferons affect the metabolic fitness of CD8⁺ T cells from patients with systemic lupus erythematosus. *Nat Commun* **12**, 1980 (2021).
298. Kolde, R., Laur, S., Adler, P. & Vilo, J. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics* **28**, 573–580 (2012).
299. Li, J. *et al.* Application of Weighted Gene Co-expression Network Analysis for Data from Paired Design. *Sci Rep* **8**, 622 (2018).
300. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).

301. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat Biotechnol* **37**, 773–782 (2019).
302. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation* **2**, 100141 (2021).
303. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat Methods* **12**, 453–457 (2015).
304. Velten, B. *et al.* Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nat Methods* **19**, (2022).
305. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw* **33**, 1–22 (2010).
306. Ahern, D. J. *et al.* A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* **185**, 916-938.e58 (2022).
307. Blangiardo, M. & Richardson, S. Statistical tools for synthesizing lists of differentially expressed features in related experiments. *Genome Biol* **8**, R54 (2007).
308. Liu, G. *et al.* ISG15-dependent activation of the sensor MDA5 is antagonized by the SARS-CoV-2 papain-like protease to evade host innate immunity. *Nat Microbiol* **6**, 467–478 (2021).
309. Hachim, M. Y. *et al.* Interferon-Induced Transmembrane Protein (IFITM3) Is Upregulated Explicitly in SARS-CoV-2 Infected Lung Epithelial Cells. *Front Immunol* **11**, 1372 (2020).
310. Genovese, F., Manresa, A. A., Leeming, D. J., Karsdal, M. A. & Boor, P. The extracellular matrix in the kidney: a source of novel non-invasive biomarkers of kidney fibrosis? *Fibrogenesis Tissue Repair* **7**, 4 (2014).
311. Velten, B., Braunger, J. M., Arnol, D., Argelaguet, R. & Stegle, O. Identifying temporal and spatial patterns of variation from multi-modal data using MEFISTO. *bioRxiv* 2020.11.03.366674 (2020) doi:10.1101/2020.11.03.366674.

312. Bagheri-Hosseinabadi, Z., Abbasi, M., Kahnouji, M., Ghorbani, Z. & Abbasifard, M. The prognostic value of S100A calcium binding protein family members in predicting severe forms of COVID-19. *Inflammation Research* **71**, 369–376 (2022).
313. Mathew, D. *et al.* Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science* **369**, (2020).
314. Su, Y. *et al.* Multi-Omics Resolves a Sharp Disease-State Shift between Mild and Moderate COVID-19. *Cell* **183**, 1479–1495.e20 (2020).
315. Schett, G., McInnes, I. B. & Neurath, M. F. Reframing Immune-Mediated Inflammatory Diseases through Signature Cytokine Hubs. *N Engl J Med* **385**, 628–639 (2021).
316. Karaboyas, A. *et al.* Inflammation and Erythropoiesis-Stimulating Agent Response in Hemodialysis Patients: A Self-matched Longitudinal Study of Anemia Management in the Dialysis Outcomes and Practice Patterns Study (DOPPS). *Kidney Med* **2**, 286–296 (2020).
317. Shah, H. H., Uppal, N. N. & Fishbane, S. Inflammation and Erythropoiesis-Stimulating Agent Hyporesponsiveness: A Critical Connection. *Kidney Med* **2**, 245–247 (2020).
318. Shilts, J. *et al.* LRRC15 mediates an accessory interaction with the SARS-CoV-2 spike protein. *bioRxiv* 2021.09.25.461776 (2021) doi:10.1101/2021.09.25.461776.
319. Song, J. *et al.* LRRC15 is an inhibitory receptor blocking SARS-CoV-2 spike-mediated entry in trans. *bioRxiv: the preprint server for biology* Preprint at <https://doi.org/10.1101/2021.11.23.469714> (2021).
320. Loo, L. *et al.* LRRC15 suppresses SARS-CoV-2 infection and controls collagen production. *bioRxiv* (2021) doi:10.1101/2021.11.09.467981.
321. Mei, Q. *et al.* Regulation of DNA replication-coupled histone gene expression. *Oncotarget* **8**, 95005–95022 (2017).
322. Huang, H. *et al.* Endogenous histones function as alarmins in sterile inflammatory liver injury through Toll-like receptor 9 in mice. *Hepatology* **54**, 999–1008 (2011).

323. Silk, E., Zhao, H., Weng, H. & Ma, D. The role of extracellular histone in organ injury. *Cell Death Dis* **8**, e2812 (2017).
324. Zuo, Y. *et al.* Neutrophil extracellular traps and thrombosis in COVID-19. *J Thromb Thrombolysis* **51**, 446–453 (2021).
325. Schenz, J., Obermaier, M., Uhle, S., Weigand, M. A. & Uhle, F. Low-Density Granulocyte Contamination From Peripheral Blood Mononuclear Cells of Patients With Sepsis and How to Remove It – A Technical Report. *Front Immunol* **12**, (2021).
326. Bojkova, D. *et al.* Proteomics of SARS-CoV-2-infected host cells reveals therapy targets. *Nature* **583**, 469–472 (2020).
327. Finkel, Y. *et al.* SARS-CoV-2 uses a multipronged strategy to impede host protein synthesis. *Nature* **594**, 240–245 (2021).
328. Rapp, M. *et al.* CCL22 controls immunity by promoting regulatory T cell communication with dendritic cells in lymph nodes. *Journal of Experimental Medicine* **216**, 1170–1181 (2019).
329. Channappanavar, R. *et al.* Dysregulated Type I Interferon and Inflammatory Monocyte-Macrophage Responses Cause Lethal Pneumonia in SARS-CoV-Infected Mice. *Cell Host Microbe* **19**, 181–193 (2016).
330. Felaco, P. *et al.* Erythropoietin induces miRNA-210 by JAK2/STAT5 signaling in PBMCs of End-stage Renal Disease patients. *FEBS J* **287**, 5167–5182 (2020).
331. Pesce, M. *et al.* Effect of erythropoietin on primed leucocyte expression profile. *Open Biol* **4**, 140026 (2014).
332. Zhang, X. *et al.* Peripheral blood mononuclear cells show prominent gene expression by erythroid progenitors in diseases characterized by heightened erythropoiesis. *Br J Haematol* **190**, e42–e45 (2020).
333. Brunyanszki, A. *et al.* Upregulation and Mitochondrial Sequestration of Hemoglobin Occur in Circulating Leukocytes during Critical Illness, Conferring a Cytoprotective Phenotype. *Mol Med* **21**, 666–675 (2015).

334. Derakhshani, A. *et al.* The Role of Hemoglobin Subunit Delta in the Immunopathy of Multiple Sclerosis: Mitochondria Matters. *Front Immunol* **12**, (2021).
335. Dubin, R. F. *et al.* Analytical and Biological Variability of a Commercial Modified Aptamer Assay in Plasma Samples of Patients with Chronic Kidney Disease. *J Appl Lab Med* **8**, 491–503 (2023).
336. Yang, J. *et al.* Impact of Kidney Function on the Blood Proteome and on Protein Cardiovascular Risk Biomarkers in Patients With Stable Coronary Heart Disease. *J Am Heart Assoc* **9**, e016463 (2020).
337. Satoh, K., Hata, M. & Yokota, H. High lib mRNA expression in breast carcinomas. *DNA Res* **11**, 199–203 (2004).
338. Stanbrough, M. *et al.* Increased expression of genes converting adrenal androgens to testosterone in androgen-independent prostate cancer. *Cancer Res* **66**, 2815–25 (2006).
339. O'Prey, J., Wilkinson, S. & Ryan, K. M. Tumor antigen LRRC15 impedes adenoviral infection: implications for virus-based cancer therapy. *J Virol* **82**, 5933–9 (2008).
340. Chen, Y.-J., Chang, W.-A., Hsu, Y.-L., Chen, C.-H. & Kuo, P.-L. Deduction of Novel Genes Potentially Involved in Osteoblasts of Rheumatoid Arthritis Using Next-Generation Sequencing and Bioinformatic Approaches. *Int J Mol Sci* **18**, (2017).
341. Ding, H. *et al.* RUNX1 Ameliorates Rheumatoid Arthritis Progression through Epigenetic Inhibition of LRRC15. *Mol Cells* **46**, 231–244 (2023).
342. Song, J. *et al.* LRRC15 inhibits SARS-CoV-2 cellular entry in trans. *PLoS Biol* **20**, e3001805- (2022).
343. Madissoon, E. *et al.* scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol* **21**, 1 (2019).
344. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single-cell RNA sequencing. *Nature* **587**, 619–625 (2020).

345. Liang, Y. *et al.* Role of neutrophil chemoattractant CXCL5 in SARS-CoV-2 infection-induced lung inflammatory innate immune response in an in vivo hACE2 transfection mouse model. *Zool Res* **41**, 621–631 (2020).
346. Persson, T. *et al.* Expression of the neutrophil-activating CXC chemokine ENA-78/CXCL5 by human eosinophils. *Clinical & Experimental Allergy* **33**, 531–537 (2003).
347. Rousselle, A. *et al.* CXCL5 limits macrophage foam cell formation in atherosclerosis. *J Clin Invest* **123**, 1343–1347 (2013).
348. Zheng, Z. *et al.* The biological role of arachidonic acid 12-lipoxygenase (ALOX12) in various human diseases. *Biomedicine & Pharmacotherapy* **129**, 110354 (2020).
349. Kaur, G., Jalagadugula, G., Mao, G. & Rao, A. K. RUNX1/core binding factor A2 regulates platelet 12-lipoxygenase gene (ALOX12): studies in human RUNX1 haplodeficiency. *Blood* **115**, 3128–3135 (2010).
350. Quintana, L. F., Guzmán, B., Collado, S., Clària, J. & Poch, E. A coding polymorphism in the 12-lipoxygenase gene is associated to essential hypertension and urinary 12(S)-HETE. *Kidney Int* **69**, 526–530 (2006).
351. Roumeliotis, A. K. *et al.* Association of ALOX12 gene polymorphism with all-cause and cardiovascular mortality in diabetic nephropathy. *Int Urol Nephrol* **50**, 321–329 (2018).
352. Limor, R. *et al.* Aldosterone up-regulates 12- and 15-lipoxygenase expression and LDL oxidation in human vascular smooth muscle cells. *J Cell Biochem* **108**, 1203–1210 (2009).
353. Yeung, J. *et al.* 12(S)-HETrE, a 12-Lipoxygenase Oxylipin of Dihomo- γ -Linolenic Acid, Inhibits Thrombosis via Gas Signaling in Platelets. *Arterioscler Thromb Vasc Biol* **36**, 2068–2077 (2016).
354. Allen-Redpath, K. *et al.* Phospholipid membranes drive abdominal aortic aneurysm development through stimulating coagulation factor activity. *Proceedings of the National Academy of Sciences* **116**, 8038–8047 (2019).

355. Zawada, A. M. *et al.* SuperTAG Methylation-specific Digital Karyotyping Reveals Uremia-induced Epigenetic Dysregulation of Atherosclerosis-Related Genes. *Circ Cardiovasc Genet* **5**, 611–620 (2012).
356. Ouchi, N. *et al.* Follistatin-like 1, a secreted muscle protein, promotes endothelial cell function and revascularization in ischemic tissue through a nitric-oxide synthase-dependent mechanism. *J Biol Chem* **283**, 32802–11 (2008).
357. Wilson, D. C. *et al.* Follistatin-like protein 1 is a mesenchyme-derived inflammatory protein and may represent a biomarker for systemic-onset juvenile rheumatoid arthritis. *Arthritis Rheum* **62**, 2510–2516 (2010).
358. Mitchell, J. A. *et al.* Endothelial cyclooxygenase-1 paradoxically drives local vasoconstriction and atherogenesis despite underpinning prostacyclin generation. *Sci Adv* **7**, eabf6054 (2023).
359. Wichmann, D. *et al.* Autopsy Findings and Venous Thromboembolism in Patients With COVID-19: A Prospective Cohort Study. *Ann Intern Med* **173**, 268–277 (2020).