
Functional Genomics of the Sepsis Response

Katie L. Burnham



Kellogg College

UNIVERSITY OF OXFORD

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

TRINITY TERM, 2017

Functional Genomics of the Sepsis Response

Katie Burnham, Kellogg College, Trinity Term 2017

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

Sepsis is defined as a dysregulated immune response to infection causing organ dysfunction, and is a major area of unmet clinical need. Although conventionally considered a unified disease with a common pathway to organ failure and death, substantial clinical and molecular heterogeneity is seen, which has limited efforts to understand pathophysiology and improve therapeutic strategies. Sepsis is associated with global changes in gene expression, and genetic variants are known to affect the response to infection. This thesis therefore uses an integrated functional genomics approach to investigate disease mechanisms and variation in the sepsis response.

Data are presented for 551 patients admitted to intensive care with sepsis due to community acquired pneumonia (CAP) or faecal peritonitis (FP). The sepsis response is explored using genome-wide gene expression and proteomics data, and molecular quantitative trait loci (QTL) are mapped in the context of disease. Comparisons with cardiac surgery patients are performed to identify shared and specific aspects of the host response.

The host transcriptomic response was largely shared across sources of sepsis, although some specificity relating to viral infection and interferon signalling was observed and validated in prospectively recruited patients. Expression-based sepsis response signature (SRS) subgroups previously described in CAP were validated, and were additionally observed in FP. SRS1 is associated with higher early mortality, and shows enrichment of pathways relating to T cell exhaustion, cell death, and endotoxin tolerance. Differences between SRS groups were also observed in the FP plasma proteome. Serial sampling enabled the investigation of temporal changes in gene expression and protein abundance within patients. Lastly, disease-relevant expression QTL were identified, and interactions with source of sepsis and SRS determined, highlighting the potential impact of regulatory variation on the sepsis response.

This thesis demonstrates the benefit of an integrative functional genomics approach to explore heterogeneity in sepsis, and highlights opportunities for patient stratification and personalised medicine.

ACKNOWLEDGEMENTS

First and foremost, I want to thank my two supervisors, Julian Knight and Emma Davenport. As my group leader, Julian has directed the project, guiding my work and providing me with a wide variety of opportunities throughout my DPhil. Emma has advised and helped with all aspects of the PhD experience, big and small, and continued doing so even after her moving to Boston required 7am Skype calls to talk through problems. She has been a pleasure to work with, and I look forward to continuing with our joint efforts.

All members of the J Knight lab deserve my gratitude for help given at various stages. In particular, I would like to thank Kat Doherty and Peter Humburg for advice and assistance early on in my PhD, and Cyndi Goh, Alex Mentzer, and Micky Carter for more recent discussion and help (and constant provision of brownies). I am also very grateful to the GAinS Investigators, and in particular Charles Hinds, for the opportunity to do this PhD, and for involving me in the wider efforts of the collaboration.

Lastly, I could not have succeeded without the support of my family and friends. My parents have always encouraged me in whatever I have chosen to do, and I fully expect my dad to take up the challenge of reading this thesis over the next ten years or so. The lab girls - Alicia and Anna (otherwise known as A²), Evelyn, and Sara - have been there throughout. The residents of 39 St Mary's Road, past and present - Liz, Nonie, and Sarah - may not have dragged Charlie to Candy Mountain, but never failed to cheer me up at the end of the day. Similarly, Sunday Fud - Cat, Jon, and Martin - has been really a lovely weekly tradition filled with laughter (mostly at Jon). Finally, all the X staircase group have been supportive throughout my DPhil, but my fellow PhD students - Ben and Talitha - made the write-up experience in particular so much more enjoyable, with a lovely week in Cambridge somehow resulting in my third results chapter.

DECLARATIONS

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

The Genomic Advances in Sepsis (GAinS) study began recruiting patients in 2005 from 34 intensive care units across the UK. From October 2013, I have been responsible for liaising with research nurses, providing supplies, maintaining sample records, and collecting and processing samples.

I processed 207 of the RNA samples presented in this thesis, with additional samples previously processed by Jayachandran Radhakrishnan, Narelle Magueri, Ashley Thorpe, and Emma Davenport. Microarray gene expression data were generated by the WTCHG Core Genomics facility. The Radhakrishnan 2010 data had been generated by Jayachandran Radhakrishnan, and the Davenport 2011 data by Emma Davenport. The cardiac surgery patients were recruited by Eduardo Svoren at Bart's and the London NHS Trust, who also processed the samples and performed his own analyses, which I have replicated for comparison to sepsis. Helen Lockstone, Peter Humburg, and Emma Davenport gave advice for microarray data processing and analysis. Some previous analyses referenced in this thesis had been performed by Emma Davenport: the derivation of a survival prediction model, the identification of two sepsis response subgroups, and the mapping of eQTL in the Davenport 2011 cohort (described in Davenport (2014)).

I carried out plasma sample processing for proteomics at the Target Discovery Institute together with Marie-Laetitia Thézenas, who generated, aligned, and mapped the mass spectrometry data. Philip Charles gave advice for data processing and analysis.

I processed 185 of the DNA samples presented in this thesis with the help of Cyndi Goh, with additional samples previously processed by Jayachandran Radhakrishnan, Emma Davenport, Anna Rautanen, and Tara Mills. Genotyping data for 275 CAP patients had been generated and processed by Emma Davenport. Genotyping data for 655 patients were generated at the Wellcome Trust Sanger Institute. Emma Davenport provided specific components of the genotyping QC and eQTL mapping pipelines, and Alexander Mentzer advised on imputation. The results of the GWAS for survival in sepsis were provided by Anna Rautanen and Alexander Mentzer.

SUBMITTED ABSTRACTS

Gene expression profiling of patients with severe sepsis due to community acquired pneumonia and faecal peritonitis reveals distinct gene signatures

Poster: Intensive Care Society State of the Art Meeting - Winner of the Research Poster Prize (2014)

KL Burnham, J Radhakrishnan, EE Davenport, E Svoren, P Humburg, AC Gordon, P Hutton, C Garrard, CJ Hinds, JC Knight, and the GAIN investigators

Distinct gene expression signatures revealed in severe sepsis patients with differing infectious aetiology

Poster: Quantitative Genomics (2015)

KL Burnham, J Radhakrishnan, EE Davenport, E Svoren, P Humburg, AC Gordon, P Hutton, C Garrard, CJ Hinds, JC Knight, and the GAIN investigators

Heterogeneity in the individual transcriptomic response to severe sepsis

Concurrent Oral Session: International Congress of Human Genetics (2016)

KL Burnham, EE Davenport, J Radhakrishnan, P Humburg, P Hutton, C Garrard, CJ Hinds, JC Knight

Inter-individual variation in the host transcriptomic response to sepsis

Oral Presentation: Quantitative Genomics (2016)

KL Burnham, EE Davenport, J Radhakrishnan, P Humburg, P Hutton, C Garrard, CJ Hinds, JC Knight

Shared and specific features of the individual transcriptomic response in sepsis due to faecal peritonitis

Poster: American Society of Human Genetics (2016)

KL Burnham, EE Davenport, J Radhakrishnan, E Svoren-Jabalera, P Humburg, P Hutton, C Garrard, AC Gordon, AVS Hill, CJ Hinds, JC Knight

ASSOCIATED PUBLICATIONS

Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study

The Lancet Respiratory Medicine (2016)

EE Davenport, **KL Burnham**, J Radhakrishnan, P Humburg, P Hutton, TC Mills, A Rautanen, AC Gordon, C Garrard, AVS Hill, CJ Hinds, JC Knight

XGR software for enhanced exploration and interpretation of genomic summary data, illustrated by application to immunological traits

Genome Medicine (2016)

H Fang, B Knezevic, **KL Burnham**, JC Knight

Shared and distinct aspects of the sepsis transcriptomic response to fecal peritonitis and pneumonia

The American Journal of Respiratory and Critical Care Medicine (2017)

KL Burnham, EE Davenport, J Radhakrishnan, P Humburg, AC Gordon, P Hutton, E Svoren-Jabalera, C Garrard, AVS Hill, CJ Hinds, JC Knight

Classification of patients with sepsis according to blood genomic endotypes: a prospective cohort study

The Lancet Respiratory Medicine (2017)

B Scicluna, LA van Vught, AH Zwinderman, MA Wiewal, EE Davenport, **KL Burnham**, P Nuernberg, MJ Schultz, J Horn, OL Cremer, MJ Bonten, CJ Hinds, HR Wong, JC Knight, T van der Poll on behalf of the MARS Consortium

A community approach to mortality prediction in sepsis via gene expression analysis

Nature Communications (2018)

TE Sweeney, TM Perumal, R Henao, M Nichols, JA Howrylak, A Choi, JF Bermejo-Martin, R Almansa, E Tamayo, EE Davenport, **KL Burnham**, CJ Hinds, JC Knight, CW Woods, SF Kingsmore, GS Ginsburg, HR Wong, G Parnell, B Tang, LL Moldawer, FE Moore, L Omberg, P Khatrri, EL Tsalik, LM Mangravite, RJ Langley

CONTENTS

Abstract	i
Acknowledgements	ii
Declarations	iii
Submitted Abstracts	iv
Associated Publications	v
Contents	vi
List of Figures	x
List of Tables	xii
Abbreviations	xiii
1 Introduction	1
1.1 Genetics and gene expression	2
1.2 Functional genomics in complex disease	7
1.3 Immunity and inflammation	18
1.4 Sepsis and the systemic inflammatory response	22
1.5 Specific aims and objectives	33
2 Materials and Methods	37
2.1 Genomic Advances in Sepsis	37
2.2 Transcriptomics	40
2.3 Mass spectrometry	43
2.4 Genotyping	45
2.5 QTL mapping	47
3 The Transcriptomic Response to Sepsis and Related Conditions	49
3.1 Introduction	49
3.2 Results: sepsis	57
3.3 Results: cardiac surgery	86
3.4 Results: comparison across SIRS sources	91
3.5 Discussion	100
3.6 Conclusions	106
4 Validation and Characterisation of Sepsis Response Signatures	107
4.1 Introduction	107
4.2 Results	117
4.3 Discussion	151
4.4 Conclusions	158

5	The Plasma Proteome in Sepsis due to Faecal Peritonitis	159
5.1	Introduction	159
5.2	Results	163
5.3	Discussion	187
5.4	Conclusions	193
6	Regulatory Genetic Variation in the Systemic Inflammatory Response	194
6.1	Introduction	194
6.2	Results	203
6.3	Discussion	240
6.4	Conclusions	246
7	General Discussion	247
7.1	The SIRS response	247
7.2	Implications of disease heterogeneity	249
7.3	Integration of multi-omic data	251
7.4	Limitations and future work	252
7.5	Conclusion	258
	Appendices	259
A	The transcriptomic response to sepsis and related conditions	259
B	Validation and characterisation of sepsis response signatures	280
C	The plasma proteome in sepsis due to faecal peritonitis	288
D	Regulatory genetic variation in the systemic inflammatory response	293
	Bibliography	299

LIST OF FIGURES

2.1	GAinS recruitment	38
3.1	Overview of GAinS gene expression cohorts	58
3.2	Overview of gene expression contrasts	59
3.3	Radhakrishnan 2010 cohort PCA	61
3.4	Davenport 2011 cohort PCA	63
3.5	Burnham 2014 cohort PCA	65
3.6	Burnham 2016 cohort PCA	67
3.7	PCA of the Radhakrishnan 2010 cohort by diagnosis	70
3.8	Correlation between gene expression PCs and clinical covariates	70
3.9	CAP vs FP volcano plot	72
3.10	CAP vs FP permutation histogram	72
3.11	Effect of cell count on differential expression	73
3.12	Enriched network in CAP vs FP	74
3.13	Viral gene expression signature volcano plot	76
3.14	CAP vs FP pathway enrichment replication	77
3.15	Boxplots of diagnostic model genes	78
3.16	ROC curves for gene expression-based diagnostic models	79
3.17	Dendrogram of serial samples	81
3.18	Day 1 vs day 5 volcano plot	82
3.19	Time course profiles: example genes	83
3.20	Time course profiles from onset: example genes	84
3.21	Sepsis time course profiles: example genes	85
3.22	Svoren 2011 cardiac surgery cohort PCA	87
3.23	Surgery response volcano plot	88
3.24	Surgery response: exemplar differentially expressed genes	89
3.25	Surgery response heatmap	90
3.26	Global surgery response	91
3.27	Sepsis response volcano plot	92
3.28	Sepsis response: comparison of CAP and FP	93
3.29	The CAP sepsis response: cohort comparison	95
3.30	Comparison of the septic and surgery responses	96
3.31	Comparison of sepsis and post-surgical samples	97
3.32	Comparison of sepsis and surgery response	98
3.33	Comparison of sepsis and traumatic injury responses	99
4.1	Davenport 2011 survival analysis: previous work	113
4.2	Kaplan-Meier plots for SRS subgroups: previous work	114
4.3	Subgroups in CAP: previous work	115
4.4	Survival model validation: ROC curve	118
4.5	Kaplan-Meier Plots for SRS validation	119
4.6	Validation of the CAP SRS subgroups	120
4.7	Validation of the CAP SRS pathway enrichment	121
4.8	Comparison of SRS model prediction and unsupervised clustering	122
4.9	Permutation of SRS groups	123

4.10	Key mediators of endotoxin tolerance	125
4.11	Defining endotoxin tolerance gene signatures	127
4.12	Use of ComBat to combine FP batches	130
4.13	Unsupervised clustering in FP	132
4.14	Correlation of differential gene expression in SRS for CAP and FP	133
4.15	Interaction between SRS and source of sepsis	134
4.16	ROC curves for SRS models in FP and CAP	135
4.17	PCA of FP samples with SRS assigned by CAP model	136
4.18	Boxplots of SRS_FP model genes	137
4.19	Comparison of SRS with paediatric endotypes	138
4.20	Gene set enrichment of SRS in paediatric endotypes	139
4.21	Timeline of SRS_FP group membership	140
4.22	FP model genes in serial samples	141
4.23	PCA to show evolution of SRS over time	142
4.24	SRS in all GAINs gene expression data	143
4.25	Evolution of SRS group membership	144
4.26	Kaplan-Meier plot based on SRS dynamics.	145
4.27	SRS1->2 vs SRS1 volcano plot	146
4.28	Boxplot of PTRF expression	147
4.29	Expression of candidate housekeeper genes	148
4.30	Output of the CAP SRS model with normalisation	149
4.31	Output of an SRS model for all-cause sepsis	151
5.1	Heatmap of protein detection across samples	164
5.2	Log transformation of raw data	166
5.3	Boxplots of Top12 proteins	167
5.4	Median centered data	167
5.5	Missingness in mass spectrometry data by phenotype	169
5.6	Heatmap of processed proteomics data	170
5.7	Subcellular localisation of detected proteins	171
5.8	Detection frequency according to subcellular location	171
5.9	Raw abundance according to subcellular location	172
5.10	Pathway enrichment: Gene Ontology Biological Processes	173
5.11	Pathway enrichment: Molecular Signatures Database	174
5.12	Principal component analysis of the proteomic data	175
5.13	Correlation between proteins and PC1	176
5.14	Correlation between samples	177
5.15	Unsupervised clustering of mass spectrometry data	178
5.16	Kaplan-Meier plot for proteomics clusters	179
5.17	Differential protein abundance between SRS groups	180
5.18	Permutation analysis for SRS differential abundance	180
5.19	Volcano plots for proteomics clusters	182
5.20	Boxplots of SRS biomarker proteins	183
5.21	ROC curve for protein-based SRS model	184
5.22	SPTN1 abundance over time	185
5.23	Comparison of protein abundance and gene expression	186
5.24	<i>FKBP1A</i> gene expression and protein abundance by SRS	187

6.1	Distance of eSNPs from transcriptional start sites	204
6.2	Epigenetic context of sepsis eQTL	205
6.3	Examples of SRS-specific eQTL in CAP	207
6.4	Sharing of SNPs between genotyping arrays	209
6.5	Filtering of imputed genotyping data	210
6.6	Comparison of genotyping and imputation coverage	210
6.7	GAInS eQTL genotyping data PCA	211
6.8	Gene expression principal components in sepsis eQTL	212
6.9	Manhattan plot of sepsis eQTL	214
6.10	Volcano plot of sepsis cis-eQTL	215
6.11	Replication of <i>IRF5</i> eQTL in sepsis	216
6.12	<i>CLEC12A</i> eQTL in sepsis	217
6.13	<i>PPP2R3C</i> eQTL in sepsis	218
6.14	Enrichment of Experimental Factor Ontology terms in eSNPs	219
6.15	Interactions of sepsis eQTL with source of sepsis	222
6.16	Permutation for diagnosis-eQTL interactions	223
6.17	Interaction of <i>OAS1</i> eQTL with source of sepsis	223
6.18	Interactions of sepsis eQTL with SRS group	225
6.19	Permutation for SRS-eQTL interactions	226
6.20	Examples of SRS-eQTL interactions	227
6.21	Gene expression PCs in surgery eQTL	228
6.22	Manhattan plot of eQTL in cardiac surgery patients	230
6.23	Volcano plot of surgery eQTL	231
6.24	Examples of eQTL in surgery patients	231
6.25	Interactions of surgery eQTL with status	233
6.26	Examples of eQTL interacting with operative status	234
6.27	Associations between SNPs and protein abundance	235
6.28	Correlation of eQTL Z scores in sepsis and surgery	237
6.29	Correlation of eQTL Z scores in sepsis and healthy volunteers . . .	238
A.1	Correlation between gene expression and clinical covariates	261
A.2	Histogram of viral vs non-viral permutation results	266
A.3	Day of sampling histogram	269
B.1	Effect of cell count on differential expression	282
B.2	Effect of cell count on differential expression	286
C.1	Mass spectrometry data sample medians	288
C.2	Identification of outlying samples	289
C.3	P value histogram for differential abundance analysis	289
D.1	Permutation of surgery interactions	298
D.2	The effect of PCs in pQTL mapping	298

LIST OF TABLES

2.1	GAinS samples	39
2.2	Clinical manifestations of SIRS	39
3.1	Radhakrishnan 2010 cohort demographics	60
3.2	Davenport 2011 cohort demographics	62
3.3	Burnham 2014 cohort demographics	64
3.4	Burnham 2016 cohort demographics	66
3.5	Covariates comparison	69
4.1	Endotoxin tolerance gene signatures	128
5.1	Proteomics cohort demographics	163
6.1	eQTL in sepsis patients	213
6.2	Overlap of eQTL with sepsis GWAS associations	220
6.3	eQTL interactions with source of sepsis	222
6.4	eQTL interactions with SRS group	224
6.5	eQTL in cardiac surgery patients	229
6.6	Interaction eQTL in cardiac surgery patients	232
6.7	Association of SNPs with protein abundance.	234
A.1	Comparison of clinical covariates between CAP and FP	259
A.2	Comparison of clinical covariates between CAP and FP	259
A.3	Comparison of clinical covariates between CAP and FP	260
A.4	Correlation of covariates with principal components	261
A.5	Differentially expressed genes by source of infection	263
A.6	Enriched pathways for CAP vs FP	264
A.7	Differentially expressed genes for viral infections	265
A.8	Differentially expressed genes by source of infection (validation)	266
A.9	Enriched pathways for CAP vs FP (validation)	268
A.10	Genes differentially expressed between day 1 and day 5 samples	269
A.11	Genes differentially expressed over time	270
A.12	Genes differentially expressed from time of onset of FP	272
A.13	Enriched pathways for days from onset (FP)	273
A.14	Genes differentially expressed over time (validation)	274
A.15	Genes differentially expressed in surgery	275
A.16	Genes differentially expressed in sepsis	276
A.17	Enriched pathways for the sepsis response	277
A.18	Genes differentially expressed between sepsis and surgery	278
A.19	Enriched pathways for sepsis vs surgery	279
A.20	Enriched pathways for the trauma response	279
B.1	Genes differentially expressed in SRS1 vs SRS	280
B.2	Enriched pathways for SRS1 vs SRS2	281
B.3	In vitro studies of the transcriptomic endotoxin tolerance response.	283
B.4	Genes differentially expressed in SRS1_FP vs SRS2_FP	285

B.5	Enriched pathways for SRS1_FP vs SRS2_FP	287
C.1	Proteins differentially abundant between SRS groups	290
C.2	Proteins differentially abundant between proteomic clusters	290
C.3	Enriched pathways for the proteomic clusters	291
C.4	Proteins differentially abundant over time	292
D.1	eQTL in sepsis patients	293
D.2	eQTL interactions with source of sepsis	294
D.3	eQTL interactions with SRS group	295
D.4	eQTL in cardiac surgery patients	296

ABBREVIATIONS

ACN	Acetonitrile
APACHE	Acute physiology and chronic health evaluation
ARDS	Acute respiratory distress syndrome
AS	Ankylosing spondylitis
AUC	Area under the curve
CAP	Community acquired pneumonia
DAMP	Damage-associated molecular pattern
DNA	Deoxyribonucleic acid
DE	Differentially expressed
eCRF	Electronic case report form
EFO	Experimental factor ontology
eIF	Eukaryotic initiation factor
ENCODE	Encyclopedia of DNA elements
eQTL	Expression quantitative trait locus
ET	Endotoxin tolerance
EWAS	Epigenome-wide association study
FC	Fold change
FDR	False discovery rate
FP	Faecal peritonitis
GAinS	Genomic Advances in Sepsis
GCTA	Genome-wide complex trait analysis
GenOSept	Genetics of Sepsis and Septic Shock
GEO	Gene expression omnibus
GO	Gene ontology
GOBP	Gene ontology biological process
GTE_x	Genotype-tissue expression
GWAS	Genome-wide association study
HRC	Haplotype reference consortium
HWE	Hardy-Weinberg equilibrium
IBD	Inflammatory bowel disease
IBS	Identity by state
ICU	Intensive care unit
IL	Interleukin
IPA	Ingenuity pathway analysis
LD	Linkage disequilibrium
LPS	Lipopolysaccharide
MAD	Mean absolute variation
MAF	Minor allele frequency
MAR	Missing at random
MARS	Molecular diagnosis and risk stratification of sepsis
MDSC	Myeloid derived suppressor cell
MHC	Major histocompatibility complex
MNAR	Missing not at random
MODS	Multiple organ dysfunction syndrome

MR	Misclassification rate
MS	Mass spectrometry
NET	Neutrophil extracellular traps
NFAT	Nuclear factor of activated T cell
NK	Natural killer
PAMP	Pathogen-associated molecular pattern
PBMC	Peripheral blood mononuclear cells
PBWT	Positional Burrows-Wheeler transform
PC	Principal component
PCA	Principal component analysis
PCT	Procalcitonin
PLS	Partial least squares
PPD	Plasma Proteome Database
pQTL	Protein quantitative trait locus
PRR	Pattern recognition receptor
QC	Quality control
qPCR	Quantitative polymerase chain reaction
qSOFA	Quick sequential organ failure assessment
QTL	Quantitative trait loci
RMA	Robust multi-array average
ROC	Receiver operator characteristic
SD	Standard deviation
SIRS	Systemic inflammatory response syndrome
SNP	Single nucleotide polymorphism
SOFA	Sequential organ failure assessment
SRS	Sepsis response signature
TF	Transcription factor
TFA	Trifluoroacetic acid
TLR	Toll like receptor
TNF	Tumour necrosis factor
TSS	Transcriptional start site
UK	United Kingdom
VEP	Variant effect predictor
WTCHG	Wellcome Trust Centre for Human Genetics

INTRODUCTION

This chapter presents the aims of this thesis, and provides an overview of pre-existing knowledge relevant to these goals; in particular, functional genomics approaches are described, and the literature describing their application in sepsis is reviewed

1.1 Genetics and gene expression	2
1.2 Functional genomics in complex disease	7
1.3 Immunity and inflammation	18
1.4 Sepsis and the systemic inflammatory response	22
1.5 Specific aims and objectives	33

The overall objective of this thesis is to investigate the nature and extent of variation in the individual host response to sepsis using a functional genomics approach. Sepsis is a heterogeneous clinical syndrome defined as a dysregulated host response to infection causing organ dysfunction. Applications of a genome-wide association approach have had limited success, suggesting that a functional genomics approach was required for further progression in understanding the disease. This work is conducted as part of the UK Genomic Advances in Sepsis study (GAinS, www.ukccg-gains.org).

In this thesis, I will investigate the molecular response to sepsis and related conditions using samples from patients recruited to the GAinS study. I will integrate data from complementary -omics approaches, including genotyping, genome-wide gene expression in peripheral blood leukocytes, and mass spectrometry data from plasma samples, with a view to gaining insight into

pathogenesis and identifying novel opportunities for therapeutics.

In this chapter, I will provide an overview of topics relevant to these aims in order to put the work described in this thesis in the context of current knowledge. Firstly, I will introduce the mechanisms by which genetic information is translated into biological function, and the methods that can be used to study these processes. I will then describe the host response to infection, and its dysregulation in the context of sepsis, and will summarise the applications of functional genomics approaches in sepsis described in the literature. Finally, I will detail my specific aims.

1.1 Genetics and gene expression

The human genome was first sequenced completely by the Human Genome Project Consortium in 2003 (Collins et al. 2003). This original reference sequence provides an overview of the genome based on the donors recruited to the project; however, a great deal of genetic diversity is observed across individuals. This is in the form of private mutations unique to a particular person, and variants that are shared across different individuals.

The sharing of specific combinations of contiguous variants, or haplotypes, can be informative about relatedness and population history. Members of the same population generally have more genetic variants in common than individuals from different populations, due to their shared ancestry. Additionally, variants that are located close together in the genome are frequently transmitted together, as there is only a short stretch of DNA between them upon which recombination can act during meiosis. Variants that are associated in this way are said to be genetically linked or in linkage disequilibrium (LD). Those further apart and therefore broken up by recombination are inherited independently.

1.1.1 Types of genetic variation

A polymorphism is a genetic locus that exists in multiple forms, or alleles, at appreciable frequencies in a population (e.g. minor allele $\geq 1\%$). Polymorphisms include single nucleotide polymorphisms (SNPs), insertions and deletions of single bases (indels), and larger structural changes such as copy number variants, genomic rearrangements, and chromosomal duplications. Their effects on phenotype range from being undetectable, which is the case for the vast majority of polymorphisms, to extremely severe, such as those causing Mendelian disease or lethality.

This variation has consequently been characterised by several consortia. For instance, the 1000 Genomes Project aimed to create a catalogue of human genetic variation in 26 populations (The 1000 Genomes Project Consortium 2012). More recently, genetic sequence or genotyping data have become available for large numbers of individuals from specific populations, which have been combined into large publicly available datasets such as the Haplotype Reference Consortium (HRC) panel (McCarthy et al. 2016). These data can be extremely informative for making inferences about population structure and history. Additionally, by identifying those genetic variants that contribute to disease risk, we may improve diagnosis, treatment, and prevention of disease.

1.1.2 Consequences of genetic variation

The functional consequences of a genetic variant depend on its location in the genome, and the type of change involved. For example, a base substitution in a protein-coding region of a gene (exon) that is translated into a change in the amino acid sequence of the protein can have an effect on the protein structure and function. This is the case in Mendelian diseases such as cystic fibrosis and sickle cell anaemia, and the effects of such variants are relatively straightforward

to predict. However, in most cases natural selection rapidly removes coding variants with strongly deleterious effects from the population. Moreover, only around 2% of the human genome encodes proteins. Disease-associated genetic variants therefore tend to have minor effects in isolation, and are thought to act largely through control of gene expression (Maurano et al. 2012).

1.1.3 Gene expression and its regulation

Gene expression refers to the synthesis of a functional product from a gene. The DNA sequence of a gene is transcribed into RNA, which in most cases is subsequently translated into protein; in the case of non-coding genes the functional product is an RNA molecule. As each cell contains the same genetic information, functional specialisation stems from the restriction of expression to particular subsets of genes. Gene expression is subject to synergistic and dynamic regulation, so the quantity of protein produced from each gene and thus its activity may be modulated over time. Observed phenotypes can thereby be a result of changes in the regulation of gene expression.

Gene expression is controlled through transcriptional initiation and elongation, mRNA processing and stability, and protein translation and degradation. The gene product may also be varied by the inclusion of different exons through alternative splicing. The majority of regulation, however, is thought to be at the level of transcriptional initiation.

1.1.4 Transcriptional regulation

RNA synthesis is initiated by the transcriptional machinery from the core promoter, which is adjacent to the transcriptional start site (TSS) of the gene. This process is controlled by the binding of transcription factors (TFs) to their cognate binding sites in the promoter and associated enhancers. Both the identity of these

TFs and the affinity with which they bind can affect transcriptional initiation positively or negatively; therefore, genetic variants in the TF binding site or the sequence encoding the TF itself can lead to differences in transcript quantities. The expression of TFs is also subject to regulation, providing a mechanism by which the expression of related sets of genes controlled by the same TF may be modulated in a coordinated fashion. Furthermore, the activity of particular enhancers can vary according to the cell type and state.

The identification of promoters, enhancers, and TF binding sites is consequently vital to understanding transcriptional regulation, and has been a major challenge in functional genomics. The Encyclopaedia of DNA Elements (ENCODE) and the FANTOM5 promoter and enhancer atlases address this need, providing publicly available data describing regulatory DNA elements in a variety of contexts (ENCODE Project Consortium 2012; Lizio et al. 2015).

Chromatin conformation and accessibility

Whilst the core promoter is close to the TSS, enhancers and other regulatory elements such as insulators and silencers are more distal. In order to affect transcription, these factors must interact with the transcriptional machinery at the promoter. This can involve very long range interactions, with contact achieved through the looping of intermediate DNA. Enhancers usually act in a modular fashion, meaning that multiple elements contribute to transcriptional control through the formation of complex regulatory networks. The activation of different combinations of promoter and enhancer elements is the main mechanism producing spatially or temporally specific expression patterns.

Gene expression may also be modulated by DNA packaging and organisation. Through association with histone proteins, eukaryotic DNA is assembled into nucleosomes separated by linker DNA. Nucleosome positioning is dynamic, and can affect the accessibility of binding sites and regulatory elements.

Transcriptional activation is associated with the repositioning of nucleosomes, and active regulatory elements tend to be found in open regions of the genome. Chromatin accessibility can be altered by TF binding, and is also associated with the presence or absence of specific epigenetic modifications.

Epigenetic modifications

Epigenetics refers to any modifications aside from changes to the underlying DNA sequence that affect gene expression. These modifications are reversible, and can also be heritable. They provide a mechanism by which gene expression can be modulated appropriately in response to environmental stimuli such as pathogens or chemicals. The profiling of epigenetic markers in different contexts, for instance by the Blueprint and Roadmap consortia (Adams et al. 2012; Bernstein et al. 2010), generates another layer of information from which regulatory mechanisms may be inferred.

Chromatin remodelling is classed as an epigenetic modification, but may also be a consequence of upstream epigenetic signals such as the acetylation or methylation of histones. Each histone may be chemically modified at a number of different sites, meaning that the set of possible changes is extremely large. Whilst the resulting histone code is not yet understood, some general observations have been made; for example, histone acetylation usually results in transcriptional activation. DNA can also be directly methylated at CpG sites, which is associated with repression of gene expression.

1.1.5 Regulation of translation and protein abundance

Although the bulk of control is at the level of transcriptional initiation, additional post-transcriptional regulation can occur. Multiple different steps are necessary for the production of protein from mRNA, and each of these can be modulated

to affect the quantity of the final gene product in a rapid and flexible manner.

Firstly, transcripts undergo processing (polyadenylation, capping, and splicing) prior to export. Splicing in particular presents a mechanism by which a variety of protein products may be created according to the requirements of the cell at the time. The processed mRNA is subsequently exported into the cytoplasm, at which stage quality control is performed. Transcript decay may then occur at different rates, for example due to the action of miRNAs and siRNAs.

Translation begins with the assembly of a pre-initiation complex involving eukaryotic initiation factors (eIFs), tRNA, and ribosomal subunits. This complex binds to the mRNA at the capped 5' end, and then scans along the transcript until it reaches the start codon. An association forms between this site and the initiator tRNA, resulting in a stable initiation complex with a complete ribosome that can subsequently move along the mRNA and build a protein.

Translational initiation is thought to be the rate limiting step in the production of protein from mRNA, and as such is the target for most translational regulation. Modulation of this process can be global; for instance, inhibiting or modifying elements of the translational machinery can repress the translation of all proteins. Alternatively, it can be targeted to specific mRNAs, such as through the binding of regulatory proteins to cis-regulatory elements in the 3' untranslated region. Subsequently, post-translational modifications and protein degradation are used to regulate protein activity.

1.2 Functional genomics in complex disease

Functional genomics refers to the use of high throughput technologies to understand the molecular basis of biological function. Specific aims can include identifying genetic variation and elucidating its consequences, assigning function to genes, and investigating how interactions between distant genomic

regions can organise genes into networks that drive a particular phenotype.

Complex diseases are influenced by a combination of environmental and heritable factors. Deducing the mechanisms involved in disease susceptibility, mortality, and morbidity is challenging, but a functional genomics approach can be particularly powerful (Steinmetz and Davis 2004). The many steps involved in the translation of DNA sequence into an observable phenotype can necessitate the integration of complementary data types to advance this goal. Those methods most relevant to this thesis are described in detail below.

1.2.1 Genotyping and genome-wide association studies

Identifying genetic variants is a common requirement in functional genomics studies. Whole genome sequencing provides a means of profiling all types of DNA variants, ranging from common SNPs to chromosomal rearrangements present only in one sample. Whilst this has the advantage of being unbiased, it is a relatively expensive approach. In rare diseases, where single low frequency variants affecting the protein sequence are most likely to be causal, whole exome sequencing may instead be used. This targeted sequencing of the transcribed regions of the genome greatly reduces costs, and successfully identifies a putative causal variant in around one third of cases (Taylor et al. 2015).

Common complex diseases are instead thought to be driven by multiple common variants of relatively low penetrance. This is referred to as the common variant-common disease hypothesis (Schork et al. 2009). Patients may have different combinations of risk alleles, and healthy individuals may carry known risk variants without developing the disease. Genome-wide association studies (GWAS) aim to identify disease-associated variants by comparing the frequency of each variant between patients and controls. Large numbers of case and control individuals must therefore be included to identify variants associated with the

disease, making whole genome sequencing impractical. Instead, genotyping arrays have generally been used in GWAS to interrogate a set of known SNPs.

Illumina's Infinium genotyping platform consists of an array of silica beads, each coated with thousands of copies of an oligonucleotide targeting a specific genomic region. Importantly, these sequences stop one base before the locus of interest. Fragments of sample DNA bind to complementary probes on the array, which DNA polymerase then extends by one base according to the sample DNA sequence. This causes a fluorescent signal to be emitted, its colour indicating which base has been incorporated. The genotype is determined from the intensity values for each base; for example, a 50:50 ratio indicates heterozygosity.

Thus far, GWAS have identified around 10,000 genetic variants associated with a range of diseases (MacArthur et al. 2017). In general terms, it has been confirmed that many polymorphisms with small individual effects contribute to each complex trait, necessitating the use of very large cohorts to discover associations (Visscher et al. 2017). For example, by using a sample size of 35,000 cases, 108 loci have been associated with schizophrenia (Ripke et al. 2014). An example of GWAS results being translated into new therapeutics is found in osteoporosis, where both Mendelian mutations and GWAS variants affecting the *SOST* gene have led to the development of the sclerostin inhibitor Romosozumab (Balemans et al. 2001; Brunkow et al. 2001; Stykarsdottir et al. 2009; Cosman et al. 2016).

Genotyping arrays usually profile known common SNPs throughout the genome (minor allele frequency (MAF) $> 1 - 5\%$). Associations between alleles due to LD reduce the number of sites that need to be assayed for GWAS, as they allow genomic variation to be captured using a relatively small number of tagging SNPs. However, it also means that the most associated variant is not necessarily causal, as it is more likely to instead be in LD with the causal SNP. Understanding the contribution of a locus to disease is made more difficult by the fact that the

vast majority of GWAS-associated SNPs are found in non-coding regions, and so do not have a clear impact on a protein. It is thought instead that they mediate disease risk through regulation of gene expression (Maurano et al. 2012).

1.2.2 Gene expression profiling

Gene expression is under precise spatial and temporal control, and disruption of normal regulatory networks can have severe consequences on phenotype. Exploration of gene expression in a disease context can therefore increase understanding of pathogenic mechanisms and suggest new therapeutic targets. Knowledge of a particular disease and the pathways involved may suggest a set of candidate genes whose expression may be assayed. This can provide information about the disease state across individuals; however, it is limited by the choice of genes, and does not permit the discovery of new biology. Instead, measuring the complete set of transcribed genes, or transcriptome, is an unbiased approach to investigating gene expression in disease.

Genome-wide gene expression can be measured using microarrays or RNA-sequencing. Similar to genotyping arrays, Illumina microarrays comprise specific probe sequences on beads, each targeting a particular transcribed sequence. In general, two or three probes are used per gene, and every probe sequence is present multiple times across the array to ensure reliable measurements. RNA is extracted from a sample and reverse transcribed into cRNA, which is then fluorescently labelled and incubated on the array. The targeted sequences bind their complement probe and emit fluorescence, with the amount of fluorescence detected for each probe being proportional to the original transcript quantity. In contrast, RNA-seq produces sequence reads for all transcripts present in a sample. These are then aligned to the reference genome, or assembled *de novo* to quantify expression based on the number of reads mapped to each gene .

The relative benefits of the two approaches are comparable to genotyping, although the cost of RNA-seq has decreased sufficiently for it to have largely replaced microarray technology in transcriptomics. Gene expression microarrays interrogate a defined set of transcripts, unlike in RNA-seq where all transcripts present are measured. Microarray probe design requires knowledge of the gene sequence, and it is not possible to detect novel transcripts. Measurements can be affected by alternative splicing or the presence of a SNP in a probe sequence, whereas RNA-seq permits the quantification of different transcript isoforms and allele-specific expression. Finally, microarray experiments frequently exhibit large batch effects relating to sample preparation and hybridisation. Whilst systematic variation is also observed in RNA-seq, it is more challenging to compare results between microarray experiments in a quantitative manner.

1.2.3 Differential gene expression analysis

A common goal in transcriptomics is to identify genes that are expressed at different levels in the disease state compared to healthy individuals. Differential expression may be a cause or consequence of disease, and can highlight genes and biological functions that are particularly relevant to pathogenesis. This can then enhance understanding of the development and progression of disease, and provide new targets for drug development. For example, differential expression analysis across human cancer highlighted common themes in metabolism, cell growth, and enhanced immune activity (Choi et al. 2005). This type of analysis requires the profiling of cases and controls in parallel to prevent systematic or batch effects leading to differences in measured expression.

Alternatively, subgroups of patients may be compared, for example to investigate the molecular response to a particular treatment, and to identify diagnostic or prognostic biomarkers. For example, a transcriptomic approach was used to identify signatures associated with allograft rejection in heart transplants

(Horwitz et al. 2004). These have been developed into a commercially available qPCR panel that is used to monitor patients (Deng et al. 2006). Furthermore, gene expression signatures may resolve clinically relevant disease subtypes, as in pulmonary disease where they have been used to define patient subgroups and to target treatment (Campbell et al. 2011).

Restricted access to disease relevant tissues, such as the brain in multiple sclerosis, can be a limiting factor in such studies. Peripheral blood gene expression profiling is appealing due to ease of sampling, and is pertinent to a range of conditions, particularly autoimmune disease, cancer, and infectious disease. Immune cells migrate to their sites of action via the blood, which interacts with every tissue type. Gene expression in peripheral blood cells can therefore reflect pathological changes in different parts of the body. Interpretation can be challenging due to the complex nature of the signalling networks and the mixed cell populations involved, but such data has provided insights in a number of disease contexts. For example, in haematologic cancers, blood gene expression signatures have been used to define disease subtypes associated with response to treatment and prognosis (Ebert and Golub 2004). Other malignancies may be diagnosed through detection of specific transcripts produced by tumour cells circulating in the blood (Smirnov et al. 2005). In systemic lupus erythematosus, an autoimmune disease affecting multiple organ systems, dysregulation of interferon signalling has been described using peripheral blood transcriptomics (Baechler et al. 2003). This is particularly pronounced in more severely affected patients, and has resulted in clinical trials of IFN blockade as a treatment for lupus (e.g. NCT02446899) (Oon et al. 2016).

1.2.4 Proteins and proteomics

As in gene expression studies, the comparison of protein quantities between diseased and healthy individuals can be informative. Candidate proteins may

be assayed using targeted techniques, or the whole proteome can be measured with high throughput methods. Mass spectrometry identifies input peptides based on their specific masses and fragmentation spectra. In clinical research, this is commonly used to describe all proteins present in a sample, together with their relative abundances. Differentially present or abundant proteins can be used as diagnostic biomarkers, and may provide opportunities for novel therapeutic interventions. In ankylosing spondylitis (AS), 2 proteins that differed in abundance between patients and controls were identified that exhibited high diagnostic value for AS (Fischer et al. 2012). Additionally, measurement of the HER2 protein is used to target treatment with trastuzumab in breast cancer (Banerjee and Smith 2010).

1.2.5 Annotation and enrichment analysis

Biological interpretation of the results of -omics studies presents a great challenge. Frequently, these are lists of genes or SNPs that might be functionally related. Knowledge discovery from such outputs relies on the availability of annotation data, and its application to the features of interest. For example, a set of genes involved in a particular canonical pathway may be defined. If a list of differentially expressed genes contains more genes from this pathway than would be expected by chance, it suggests that the pathway is relevant to the disease.

Hierarchical structures or ontologies can be used to store information on gene functions and the relationships between annotations, such as those maintained by the Gene Ontology (GO) Consortium. Annotation of sequence variants and regulatory regions is less comprehensive, but notably the GWAS catalogue has been mapped to the Experimental Factor Ontology (EFO). Other knowledge domains are represented by data from consortia such as ENCODE, FANTOM5, BLUEPRINT, and Roadmap, which are used in online tools such as the Variant Effect Predictor (VEP) and RegulomeDB to explore SNP function. Identification

of gene-SNP networks may be particularly useful in the future, but more experimental data (e.g. chromatin confirmation capture) are required to realise the full potential of such methods.

A number of tools aid biological interpretation of -omic data. For example, XGR is an R package and web-app that leverages multiple data sources to integrate available knowledge and reveal underlying biological themes in a systematic way (Fang et al. 2016). Specific features of XGR include the specification of the background for enrichment testing, the consideration of ontological structures where relevant, and the analysis of both gene and SNP data.

1.2.6 Mapping molecular quantitative traits

Elucidating the target of regulatory variants is a key challenge in functional genetics. Whilst they frequently regulate nearby genes, longer range interactions are also observed, and moreover, gene rich regions present a variety of plausible targets. This problem may be addressed systematically through expression quantitative trait loci (eQTL) mapping. eQTL describe statistical associations between a genetic variant (eSNP), and the amount of a particular gene transcript (expression associated eGene), which can be considered as a quantitative trait in the same way as phenotypes such as height. Such links between a SNP and transcript are thought to indicate instances of transcriptomic regulation, and so can be informative about gene expression in general as well as generating hypotheses about the impact of particular variants.

eQTL mapping involves testing all pairs of genetic variants and transcripts for significant associations, although separate analyses are often performed for local (eSNP within ~1Mb of the eGene) and distal interactions. These position-based groupings are frequently referred to as *cis* and *trans*-eQTL, relating to the presumed mode of action of putative regulatory variants. eSNPs close to their

eGene are generally located in *cis*-regulatory elements, such as transcription factor binding sites in the gene promoter. Variants located elsewhere in the genome are more likely to act in *trans*, *i.e.* through a transcription factor or functional RNA that regulates the eGene. *Trans*-eQTL can thus involve central regulators of sets of related genes (Rockman and Kruglyak 2006; Albert and Kruglyak 2015).

Gene expression, chromatin conformation, and epigenetic modifications are all known to vary according to cell type and developmental stage. The action of regulatory variants and the eQTL detected may consequently be context specific. It is therefore important to consider the most appropriate conditions under which to assay gene expression. For example, investigation of different leukocyte populations may be informative for autoimmune diseases. A number of studies in large numbers of healthy individuals describe eQTL in a variety of tissue types, providing resources that may be queried (Westra et al. 2013; Ardlie et al. 2015; Joehanes et al. 2017).

Moreover, the evidence suggesting that immune stimuli can have a profound effect on the regulatory landscape advocates for the use of primary cells from patients in disease research (Fairfax et al. 2014; Lee et al. 2014). Such considerations can limit the number of samples assayed, but can be instrumental in the identification of disease relevant variants. For example, eQTL specific to the response to *Mycobacterium tuberculosis* found SNPs that played a role in anti-mycobacterial immunity (Barreiro et al. 2012). Similarly, associations between genetic variants and the gene expression response to rhinovirus infection highlighted loci that could drive inter-individual variation (Çaliskan et al. 2015).

A similar approach may be applied to identify variants with an effect on protein abundance, or indeed other quantitative molecular traits such as metabolites and epigenetic modifications. Protein QTL (pQTL) have been identified for candidate proteins and across the whole proteome, in healthy individuals and in the context

of diseases including Alzheimer's and inflammatory bowel disease (IBD) (Melzer et al. 2008; Suhre et al. 2017; Kim et al. 2013; Di Narzo et al. 2017). Only the most recent studies achieve a similar scale to eQTL datasets in terms of the number of individuals and features assayed; for example, Sun et al. (2017) test 2994 proteins across 3301, finding 1927 pQTL associations. As technology continues to advance, this area of research is expected to expand correspondingly.

1.2.7 Epigenomics and functional characterisation

High throughput approaches such as those described above generate many hypotheses about disease relevant genes, but do not demonstrate mechanisms of action. This is particularly important for clinical translation and drug development. For example, it is assumed that changes in transcription lead to changes in protein levels that could then be targeted by treatment, but this is not always the case. Similarly, the overlap of a GWAS association with a known promoter can identify putative target genes, but the causal link is not proven. Allele-specific expression or gene editing with the CRISPR-Cas9 system may be used in conjunction with disease specific assays in relevant cell lines or primary cells to functionally validate such hypotheses. However, in order for these lower throughput assays to be cost effective, genes and variants must be first prioritised.

In both GWAS and QTL mapping, the peak association is frequently a marker for the causal variant. Fine mapping approaches and integration with epigenomic data can help resolve these associations and direct further functional work. This could include combining eQTL and GWAS data to assign targets to regulatory risk variants, or eQTL with proteomic data to demonstrate a consequence on protein abundance. Moreover, overlapping SNP data with epigenetic annotations can prioritise likely regulatory variants, and improve understanding of transcriptional control and disease progression at the molecular level (Rakyan et al. 2011). Whilst annotations defined

in healthy individuals are informative, the complexity and specificity of regulatory landscape advocates for the generation of context specific epigenetic maps. These could describe DNA methylation through arrays or bisulphite sequencing, transcription factor binding or histone modifications through chromatin immunoprecipitation, and long range chromatin interactions using chromatin conformation capture.

1.2.8 Missing heritability

As described above, GWAS studies have been successful in identifying risk variants for a wide range of diseases. However, compared to heritability estimates from family studies, these results explain only a small fraction of the variance attributed to genetic factors. This discrepancy is referred to as missing heritability, and has several possible explanations.

Firstly, missing heritability may be due to highly polygenic genetic effects, with the individual contribution of each SNP being so small that hundreds of thousands of cases would be required for detection. Variation in the trait under examination may reflect a heterogeneous genetic aetiology, meaning that specific risk variants may only be detected when homogeneous subgroups of patients are tested. Limitations in study design could have prevented the detection of associations; measurement error or misdiagnosis, inappropriate selection of control groups, and latent cases within the control population could all confound the results of GWAS. Moreover, the set of variants tested usually do not include rare SNPs, copy number variants, and structural variants, all of which could contribute to pathogenesis. Gene-gene and gene-environment interactions may play a large role, and are not considered in the usual study design. Lastly, genetic effects may in fact be epigenetic effects, and a growing number of epigenome-wide association studies (EWAS) are investigating this possibility.

Whilst each of these conjectures may apply in certain contexts, Genome-wide Complex Trait Analysis (GCTA) has demonstrated that the inclusion of all common SNPs in a linear model explained a substantially greater proportion of overall trait variance than known GWAS associations. This suggests that genetic heritability is in fact the result of a large number of common SNPs acting additively, but with such small individual effect sizes that the majority have not been detected (Yang et al. 2010).

1.3 Immunity and inflammation

The immune system is a network of molecules, cells, tissues, and organs that defends against infectious disease. If functioning appropriately, the immune system detects and destroys pathogens and other foreign molecules, whilst recognising healthy host tissue as not requiring a response. However, immune dysfunction plays a role in a number of diseases, ranging from immunodeficiencies such as severe combined immunodeficiency, to autoimmune diseases and allergies where a hyperactive immune system attacks normal cells. The human immune response consists of many layers of defence, divided into two interacting branches: the innate and adaptive immune responses.

1.3.1 The innate immune response

The innate immune system presents an immediate, acute, and non-specific response to infection. It comprises surface barriers to pathogens, such as the skin and secreted mucus, cell-mediated responses, and humoral components. It also activates aspects of the adaptive immune system (Turvey and Broide 2010).

Inflammation occurs in response to injury from infection, trauma, toxins, burns, or other tissue damage. It is an short-term response that both combats the

stimulus and repairs the damage caused. It is initiated by the recognition of pathogen- or damage-associated molecular patterns (PAMPs and DAMPs) by pattern recognition receptors (PRRs) on resident immune cells. These cells become activated and release inflammatory mediators, such as cytokines that trigger vasodilation, swelling, and fever; interferons with anti-viral effects; and chemokines, which attract leukocytes to the site of infection or damage. These cells can then migrate from the blood vessel into the affected tissue, where they become activated and kill invading pathogens, remove dead cells, and initiate repair. This involves phagocytosis of pathogens and cell debris, the destruction of infected host cells, and the release of additional signalling molecules that help to activate the adaptive immune response.

The complement system consists of a number of circulating precursor proteins that become activated during infection through a biochemical cascade. It propagates and enhances the innate immune response through lysis and opsonisation of pathogens, stimulation of inflammation, and attraction and activation of leukocytes.

1.3.2 The adaptive immune response

In contrast to the general defence mechanisms of innate immunity, the adaptive immune system develops pathogen-specific responses. Immunological memory is established, such that subsequent exposure produces an enhanced reaction, granting long-term protection. Cross-talk between the innate and adaptive immune systems is extensive, and vital for an effective and appropriate response to infection (Bonilla and Oettgen 2010).

The adaptive response is initiated by the binding of antigens to their cognate receptors. A huge diversity of T and B cell receptors is generated through somatic gene recombination, in order that millions of different antigens may

be recognised specifically and trigger a targeted response. Following antigen recognition, B cells become activated and secrete immunoglobulins. These antibodies circulate and bind specifically to antigens to inactivate them or mark them for destruction by the innate immune system. T cells recognise antigens presented by major histocompatibility complex (MHC) molecules on host cells. They then respond by killing infected cells, releasing signalling molecules to enhance the activity of other cells, or regulating the immune responses. Some lymphocytes differentiate into memory cells that can respond more rapidly upon subsequent infection by the same pathogen (Pennock et al. 2013).

1.3.3 Cells of the immune system

The cells of the immune system are derived from pluripotent haematopoietic stem cells via lymphoid progenitor cells or myeloid stem cells. They include a number of different cell types, each of which are adapted to particular roles but act in a synergistic and interdependent manner.

Phagocytosis is carried out by neutrophils, dendritic cells, monocytes, and macrophages. Neutrophils can additionally attack pathogens through the release of antimicrobial proteins and neutrophil extracellular traps (NETs), whilst monocytes present antigens to T cells and secrete pro- and anti-inflammatory signals. Other cells involved in innate immunity are mast cells, basophils, and eosinophils, which have secretory and regulatory roles, and natural killer (NK) cells, which destroy infected host cells (Turvey and Broide 2010).

As described above, the adaptive immune system depends upon antigen presenting cells (particularly macrophages, B cells, and dendritic cells) interacting with T cells. There are several T cell subsets, with cytotoxic T cells killing infected cells, helper T cells stimulating other immune cells, and regulatory T cells modulating the immune response (Pennock et al. 2013; Bonilla

and Oettgen 2010).

1.3.4 Genetics and infectious disease

Infectious diseases are caused by the invasion of the body by pathogenic microorganisms such as bacteria, viruses, parasites, or fungi. Whilst chronic autoimmune conditions are known to have a partially genetic aetiology, infectious diseases are generally considered to be entirely environmental diseases as they are caused by an external microbial agent. However, there is substantial evidence supporting a heritable component to the risk of death from infection.

Firstly, only a subset of individuals exposed to an infectious agent develop disease. This has been demonstrated for instance for tuberculosis and leprosy through twin studies (Comstock 1978; Misch et al. 2010). Additionally, adopted children have a greatly increased risk of dying from infection if one of their biological parents had died due to infectious disease, whereas no such effect was seen with adoptive parents (Sørensen et al. 1988). The importance of genetic variation in the MHC region to the immune response has been highlighted by a number of studies, with the first GWAS performed in the context of an infectious disease (HIV-1) identifying HLA SNPs that were strongly associated with viral load (Fellay et al. 2007).

The genetics of susceptibility to and outcome from infectious disease may be studied in a number of ways (Chapman and Hill 2012). GWAS have been carried out and significantly associated loci identified in infections such as malaria (Jallow et al. 2009), tuberculosis (Thye et al. 2010), and dengue (Khor et al. 2011). Moreover, insights may be gained through the investigation of rare monogenic and selective immunodeficiencies. For example, susceptibility to mycobacteria may be greatly increased by mutations that impair the interferon gamma response (van de Vosse and Dissel 2009). Lastly, regulation of gene

expression is central to the immune response. A functional genomics approach may be used to investigate both disease-associated polymorphisms and the dynamics of gene expression following infection to improve understanding of mechanisms of disease and the host immune response (Knight 2013).

1.4 Sepsis and the systemic inflammatory response

As mentioned above, the immune response can become dysregulated with severe consequences. The first consensus definitions published for sepsis described it as a systemic inflammatory response caused by an infection (Bone et al. 1992), with severe sepsis referring to a subgroup of patients with associated organ failure. Systemic inflammatory response syndrome (SIRS) refers more generally to dysregulation of inflammation, and can also occur following sterile insults such as burns, surgery, and traumatic injury (Bosmann and Ward 2013). More recent guidelines state that sepsis is life-threatening organ dysfunction caused by a dysregulated host response to infection (Singer et al. (2016); Sepsis-3).

Whilst the initial cause of sepsis is an infectious agent, it is the inappropriate host response to this pathogen that causes the symptoms observed, which may persist after the infection has been resolved. It was formerly thought that mortality and morbidity in sepsis were due to excessive inflammation, but anti-inflammatory drugs have failed to improve survival rates, and patients can in fact display immunosuppression even up to five years post-recovery (Quartin et al. 1997; Deans et al. 2005). This has prompted reinterpretation of the disease as more general and complex dysregulation of the immune response, with both pro- and anti-inflammatory processes involved in mortality (Wiersinga 2011; Hotchkiss et al. 2013).

1.4.1 Dysregulation of the inflammatory response

The innate immune system is activated strongly in sepsis. The consequential damage to the host results in the release of DAMPs, which in turn enhance and prolong the response. NF- κ B signalling and pro-inflammatory cytokines including tumour necrosis factor (TNF), interleukin-1 β (IL-1 β), IL-12, and IL-18 have been implicated in pathogenesis (van der Poll et al. 2017). Dysregulation of the complement and coagulation systems are also particularly prevalent (Wiersinga et al. 2014).

The complement cascade enhances the inflammatory response, and its dysregulation contributes to the tissue damage and organ dysfunction seen in sepsis. Similarly, excessive coagulation can lead to microvascular thrombosis and haemorrhage. This is compounded by the rapid release of NETs by neutrophils (NETosis), which can lead to the aggregation of platelets and thrombosis. In the controlled immune response, platelets maintain vascular integrity and prevent haemorrhage; however, platelet activation has been associated with organ dysfunction in sepsis. Disruption of the endothelial barrier is common, causing tissue oedema and hypotension. Interdependence and regulatory feedback between these functions and processes mean that their contribution to pathogenesis is both complex and dynamic, and varies extensively between patients (van der Poll et al. 2017).

1.4.2 Immunosuppression

Features of immunosuppression commonly observed in sepsis include lymphocyte exhaustion and apoptosis. T cells exposed to a high antigen load release reduced amounts of cytokines, and the proportion of immunosuppressive regulatory T cells is elevated (van der Poll et al. 2017). Antigen-presenting cells similarly express less HLA-DR, produce less cytokines, and are more subject to

apoptosis (Monneret et al. 2006). This immunosuppression has been related to epigenetic reprogramming (Carson et al. 2011).

Endotoxin tolerance is the phenomenon whereby cells that have previously been exposed to lipopolysaccharide (LPS), a component of bacterial cell walls, display a reduced response to a repeated stimulus (Beeson 1946). This is thought to be an adaptive mechanism limiting damage from excessive or inappropriate inflammation, and involves the precisely controlled regulation of several different gene sets through epigenetic remodelling. For example, pro-inflammatory cytokines including $\text{TNF}\alpha$ are repressed in tolerant cells, whereas antimicrobial genes are induced. It has been proposed that endotoxin tolerance plays a role in sepsis, and may be an important mechanism leading to the sustained immunoparalysis seen (Cavaillon and Adib-Conquy 2006). Inter-individual variation in the degree of tolerance developed may also be relevant to clinical heterogeneity (Fairfax et al. 2011).

1.4.3 Metabolic derangement

Metabolism and immunity are closely connected, and the immune dysregulation in sepsis corresponds with metabolic imbalances. Specifically, the Warburg effect is observed, where inflammatory and cancer cells shift from oxidative phosphorylation to aerobic glycolysis for more rapid production of ATP and intermediate molecules (Cheng et al. 2014). Metabolic effects in sepsis are complex, and relatively little research has examined the impact on pathogenesis and treatment.

1.4.4 Infectious aetiology of sepsis

Sepsis can be caused by any type of infection, with the prevailing opinion being that pathogenesis progresses via a common pathway regardless of infectious

aetiology. However, the causative organism affects treatment choice, and can be related to outcome; for example, a meta-analysis of bloodstream infection found that gram-negative bacteraemia had a higher mortality rate than gram-positive bacteraemia (Cohen et al. 2004). Identification of the cause of sepsis is therefore of great interest (Goh and Knight 2017).

Respiratory tract infections are the most common cause of sepsis, followed by abdominal and genitourinary infections. Historically, gram-negative bacteria were predominant, but the incidence of gram-positive infections have increased recently, potentially due to the preponderance of hospital-acquired infections (Mayr et al. 2014). The European Prevalence of Infection in Intensive Care study reported 62.2% gram-negative organisms amongst the approximately two thirds of severe sepsis patients with a positive blood culture. The predominant specific organisms were *Staphylococcus aureus* (20.5%), *Pseudomonas* species (19.9%), Enterobacteriaceae (16.0%), and fungi (19%) (Vincent et al. 2009). However, a recent epidemiological study in the USA estimated that 47% of sepsis cases in hospitals were culture negative and so without a confirmed aetiology (Gupta et al. 2016). Patients are not tested routinely for viral infections, but it has been suggested that a significant proportion of sepsis patients have a viral aetiology or a concomitant viral infection that may affect their outcome.

The known risk factors for sepsis largely reflect the risk of infection, and include age, male gender, black race, and socioeconomic status. Pre-existing chronic diseases are frequently exacerbated by acute infection, and patients with comorbidities have poorer long-term outcomes (Mayr et al. 2014).

1.4.5 Treatment and impact

Current treatments for sepsis focus on source control and elimination of the causative infection through intravenous antibiotics, and supportive care, often

in an intensive care unit (ICU) environment. Fluids and, in the case of shock, vasopressors are used to maintain blood pressure, and organ support such as dialysis and ventilation is provided as appropriate. Corticosteroids may be given to control inflammation, but their effects are not well understood. Activated protein C has been used as an anticoagulant and anti-inflammatory agent; however, a review of five clinical trials found that it was not consistently beneficial and the drug has been withdrawn (Martí-Carvajal et al. 2012).

Sepsis is a major cause of admission to intensive care and leads to an estimated 37,000 deaths per year in the UK alone (Daniels 2011). New insights into the pathophysiology and the development of targeted treatments are urgently required. A large number of anti-inflammatory agents have been tested, but due to inconsistent effects across trials, none are approved for use. Immune stimulants have been suggested as protection against secondary infections, but as these only develop in a minority of patients their potential benefit is limited (van der Poll et al. 2017). Clearer understanding of heterogeneity in the sepsis response is therefore required (Marshall 2014; Cohen et al. 2015).

1.4.6 Functional genomics in sepsis

The complexity of the sepsis response and the importance of gene expression regulation in the immune system suggest that functional genomics approaches may be extremely valuable for improving understanding of pathogenesis and investigating inter-individual variation. Previous work in the field that examines the genome-wide host response is reviewed briefly below.

Genome-wide association studies

Two GWAS for survival in adult sepsis have been published, which identified independent risk loci. Rautanen et al. (2015) found an intronic SNP in the *FER*

gene was associated with 28-day survival in a cohort restricted to patients with sepsis due to pneumonia. In a mouse model of lung contusion and bacterial pneumonia, delivery of the *FER* gene by electroporation improved survival rates through increased immune cell recruitment (Dolgachev et al. 2016). A second GWAS found an association between a SNP in the *VPS13A* gene and survival in all-cause sepsis (Scherag et al. 2016). A GWAS for sepsis in premature infants did not identify any significant associations (Srinivasan et al. 2017).

An alternative pharmacogenomics approach identified a number of SNPs associated with response to treatment (Man et al. 2013). This suggests that specific therapies are beneficial only to subgroups of patients with certain genotypes, and highlights an opportunity for personalised medicine in sepsis.

Genome-wide gene expression

The majority of published work on the host sepsis response has been in the field of transcriptomics. Sweeney and Khatri (2017) recently performed a systematic search for publicly available gene expression data in sepsis, and identified 39 data sets comprising 3,241 samples from 2,604 patients. This included only studies with a non-septic comparator group, and additional cohorts exist comprising just sepsis patients. Many of these studies were limited by small sample sizes (<100), and reviews suggest that relatively low levels of replication are partially due to the use of underpowered, heterogeneous cohorts (Tang et al. 2010; Wong et al. 2012a). Differences between adult and paediatric sepsis are also apparent from these studies. More recent publications have involved much larger cohorts, and together provide valuable insights into several areas of interest.

Diagnosis of sepsis: Early detection of sepsis is critical to ensure appropriate treatment, but blood cultures are slow and have low sensitivity. The use of host gene expression biomarkers are therefore an appealing alternative by which to

discriminate sepsis from non-infectious systemic inflammation. A number of studies have identified different gene sets that are able to separate sepsis from SIRS patients and/or healthy controls (e.g. Prucha et al. (2004), Johnson et al. (2007a), Sutherland et al. (2011), and Scicluna et al. (2015). More recently, combined analyses on publicly available data have allowed the derivation of more robust diagnostic signatures and permitted the comparison of several different predictors (Sweeney et al. 2015; Sweeney and Khatri 2017).

Identification of infection type: Biomarkers for infectious aetiology would be extremely useful for informing clinical management. Whilst Tang et al. (2009) did not find gene expression differences between Gram positive and Gram negative sepsis, identification of viral pneumonia may be more achievable and clinically useful (Tsalik et al. 2016). For example, Parnell et al. (2012) described a gene set identifying patients with influenza, and Sweeney et al. (2016) also derived a model to discriminate between bacterial and viral infection in order to inform treatment choice.

Organ failure: Similarly, gene expression signatures have been described for acute respiratory distress syndrome (ARDS) within sepsis patient cohorts (Howrylak et al. 2009; Dolinay et al. 2012). Studies of specific organ failure increase understanding of disease mechanisms, for example indicating the importance of the neutrophil response in ARDS (Kangelaris et al. 2015).

Mortality: Comparison of survivors and non-survivors can give insight into the factors that lead to poor outcomes, suggest therapeutic targets, and identify patients requiring additional support. Features such as early suppression of the adaptive immune response, disruption of the cell cycle, and upregulation of pro-inflammatory cytokines have been robustly associated with mortality (Sweeney and Wong 2016). However, relatively few genes may be differentially expressed

between eventual survivors and non-survivors at onset (Scicluna et al. 2015; Severino et al. 2014), and time of sampling is an important confounder given changes in gene expression over the course of disease progression (Cazalis et al. 2014; Kwan et al. 2013).

Several survival prediction models have been proposed (Pachot et al. 2006; Bermejo-Martin et al. 2010; Kwan et al. 2013; Tsalik et al. 2014; Cazalis et al. 2014), but the majority have not been validated or perform less well in prospectively recruited cohorts. However, the application of several different models to a combined cohort of publicly available data has demonstrated the potential of such an approach (Sweeney et al. 2016).

Subcategorisation of patients: This approach addresses the heterogeneity in the sepsis response, and aims to identify patient subgroups with clinically important differences in their disease response. Maslove et al. (2012) used unsupervised clustering to identify two subgroups, one of which had increased inflammatory and Toll-like receptor signalling. Differences in the expression of genes targeted by drugs such as vasopressin and hydrocortisone were also described. Parnell et al. (2013) instead described a set of genes that were downregulated in sepsis and correspond to pathways relevant to immunosuppression in sepsis. This gene signature was intended as a means of assaying the immune response state of patients, but has not been validated.

Paediatric sepsis: Wong and colleagues have assembled a large patient cohort and published extensively on paediatric sepsis (≤ 10 years old) (Wong 2013). Exploration of gene expression patterns identified early repression of the adaptive immune response (Cvijanovich et al. 2008), with related differences across developmental age groups (Wynn et al. 2011). This highlights the importance of patient age as a potential confounding factor for subsequent analyses. Novel therapeutic targets have been identified and

experimentally validated, namely inhibition of matrix metalloproteinase-8 activity and restoration of zinc homeostasis (Solan et al. 2012; Wong 2007).

Three paediatric septic shock endotypes (patient subgroups) with clinically relevant differences have been described (Wong et al. 2009a). Specifically, a subgroup with repression of T cell, B cell, and glucocorticoid receptor signalling was found to have greater illness severity and mortality rates. The subgroups also showed differences in their response to corticosteroids. A gene signature was developed to resolve these endotypes in new patients and the findings were replicated (Wong et al. 2010b; Wong et al. 2011). A panel of biomarkers has also been identified and tested for mortality prediction (Kaplan and Wong 2011; Wong et al. 2012a). This work in particular has involved the identification of candidate biomarkers through gene expression profiling, followed by measurement of protein abundance.

Proteomics

Proteomics in sepsis has frequently focussed on rapid pathogen identification. However, the host response has also been investigated, with the majority of studies using easily accessible sample types such as plasma and urine.

Diagnosis: Comparison of pooled sepsis and SIRS plasma samples identified proteins involved in the complement and coagulation pathways, with the potential to discriminate between sepsis and SIRS prior to clinical diagnosis (Shen et al. 2006; Lissauer et al. 2007). Similarly, comparison of plasma protein abundance in 10 adults with sepsis due to *Acinetobacter baumannii* to paired healthy donors identified differentially abundant lipoproteins and coagulation-related proteins (Soares et al. 2009). Profiling of the platelet proteome of six sepsis patients and paired healthy controls found five differentially abundant proteins including IL-1 β (Liu et al. 2014).

As proteomics methodologies have advanced, subsequent studies have used much larger sample sizes to address this same aim. For example 77 neonatal sepsis patients were compared to non-septic neonates on the basis of the plasma proteome. This identified a set of potential diagnostic biomarkers that could inform treatment, and that were prospectively validated (Ng et al. 2010). Additionally, Kiehntopf et al. (2011) compared the plasma proteome of 166 patients with SIRS of infectious and non-infectious aetiology and developed a classification algorithm that was validated in prospectively recruited patients. The proteins involved were implicated in infection-related protease activity.

Prognosis: Kalenka et al. (2006) identified six serum proteins involved in inflammation and cytoprotective signalling that were differentially abundant between 28 day survivors and non-survivors (n=18). Another prognostic study measured plasma glycopeptides in ten survivors and ten non-survivors. Proteins that were only detected in survivors were enriched for the complement cascades, coagulation cascade, cell adhesion molecules, extracellular matrix-receptor interaction, focal adhesion, and prion diseases (DeCoux et al. 2015). Differences in specific functions within the coagulation and complement cascades were also described. Lastly, Raju et al. (2016) generated longitudinal serum protein profiles for sepsis patients from diagnosis until recovery/mortality. Six proteins showed different patterns of expression between survivors (n=6) and non-survivors (n=6).

In a more integrated approach, Langley et al. (2013) compared the plasma proteomes and metabolomes of critically ill patients with and without sepsis due to CAP. Significant differences were found between survivors and non-survivors, and sepsis and SIRS patients. Mortality was associated with gluconeogenesis, and impaired fatty acid transport and β -oxidation. A survival prediction model based on metabolites and clinical features was described and validated.

Age and risk of sepsis: To explore the relationship between age and risk of sepsis, Cao and Robinson (2014) investigated differences in the plasma proteome of 39 CAP patients in two age brackets who did or did not develop sepsis. They identified 58 proteins with statistically significant differences in abundance in sepsis as a function of age, which related to the acute phase response, coagulation signalling, lipid metabolism, and reactive oxygen species production.

1.4.7 Response to surgery

Surgery induces a systemic host response involving the immune system and biological mechanisms associated with tissue repair and homeostasis. Considerable variation is seen across patients, with a dysfunctional host response associated with poor outcomes; for example, multiple organ dysfunction can develop, which has a mortality rate of up to 40%. The causes of an adverse response are not well understood, but comorbidities, surgical complications, and genetic predispositions are thought to be contributing factors. Investigation of the host response to surgery could improve understanding of the inflammatory response, its dysregulation, and its relationship to aetiology. For example, Liangos et al. (2010) studied the transcriptomic response of ten patients and identified differential regulation relating to ischaemia reperfusion and the inflammatory response.

The similarities between the surgical and sepsis responses, and the increased risk of an infection following an operation, make the transcriptomic response to surgery highly relevant to the study of sepsis. Elective surgery is particularly useful in this context, due to the availability of pre-insult samples, clear start and end points for the injury, and similarities in patient demographics.

1.4.8 Traumatic injury

Similarly, the inflammatory response to injury can become dysregulated. This may lead to multiple organ dysfunction and mortality. The innate immune system is activated by DAMPs, with neutrophils, monocytes, and NK cells playing a central role in the injury response. Necrosis from a burn or trauma additionally triggers the coagulation and fibrinolysis cascades. As noted for sepsis, immunosuppression in the form of T cell exhaustion is also described, increasing susceptibility to infection (Hotchkiss and Karl 2003).

The “Inflammation and the Host Response to Injury” research programme has published extensively on systems biology analysis of trauma. Notably, genome-wide gene expression data were generated for serial samples from 167 trauma patients in order to investigate the transcriptomic response to severe blunt trauma. Xiao et al. (2011) describe a “genomic storm”, with differential expression of >80% of the leukocyte transcriptome in the 28 days following injury, and specifically activation of the innate and suppression of the adaptive response. This was found to be highly comparable across patients, and to the burn response.

1.5 Specific aims and objectives

The aims of this thesis are to:

- **Investigate the host transcriptome in sepsis of different sources (Ch. 3)**

The two most common causes of sepsis in the UK are community acquired pneumonia (CAP) and faecal peritonitis (FP). The host transcriptome in sepsis due to FP has not yet been described, and the clinical differences observed between CAP and FP suggest that variation may also be observed in the molecular response to disease. I aim to:

1. describe the host transcriptome in sepsis due to FP and CAP for peripheral blood leukocytes
2. directly compare gene expression between CAP and FP patients
3. identify temporal changes in gene expression in CAP and FP

- **Compare the sepsis transcriptomic response to related conditions (Ch. 3)**

The sepsis response is related to the systemic inflammatory response to insults such as surgery and trauma. Comparison of the infectious and sterile responses might aid interpretation of the biological processes involved. I will therefore:

1. define the transcriptomic response in:
 - (a) sepsis due to CAP and FP, in contrast to non-septic samples
 - (b) traumatic injury, in contrast to healthy volunteers
 - (c) cardiac surgery, using pre- and post-operative samples
2. compare these responses across conditions

- **Investigate heterogeneity in the individual sepsis response (Ch. 4)**

Sepsis is a highly complex and heterogeneous syndrome. Previous work has described distinct gene expression signatures in CAP patients enrolled in GAINs (Davenport (2014), reviewed in chapter 4). Notably, these identified a subgroup of patients with relative immunosuppression and higher mortality. Here, I aim to validate and expand upon these findings, particularly focussing on whether these expression signatures could be useful for prognostication and patient stratification in the clinic. My specific aims are to:

1. validate the sepsis response signatures (SRS) in a new cohort of CAP patients
2. investigate inter-individual variation in FP and compare to CAP

3. investigate the biological and clinical relevance of inter-individual variation in the sepsis response
4. use serial samples to investigate whether SRS group membership varies over time

- **Investigate the proteome of sepsis patients (Ch. 5)**

Gene expression analyses are generally based on the hypothesis that changes in gene expression result in changes in protein abundance, which in turn have an effect on phenotype. Investigation of the proteome can complement transcriptomic work, adding information about biological activity and may be more useful for the identification of therapeutic targets. I will therefore carry out a pilot mass spectrometry experiment to:

1. describe the proteome of patients with sepsis due to FP
2. investigate differentially abundant proteins in the context of:
 - (a) sepsis response subgroups
 - (b) survival
 - (c) time from disease onset
3. compare the sepsis plasma proteome to the transcriptome

- **Investigate the impact of genetic variation on the individual immune response (Chapter 6)**

Control of the immune response is largely at the level of gene expression, and this has been found to be significantly modulated by regulatory variation. It is therefore important to improve our understanding of the mechanisms involved and how specific variants can have an effect on phenotype via gene expression. One approach to this is to map gene expression as a quantitative trait (eQTL mapping). In this chapter, I will explore the impact of genetic variation in the context of systemic inflammation through:

1. mapping QTL for
 - (a) gene expression in sepsis patients, identifying interactions with SRS group and source of sepsis
 - (b) gene expression in cardiac surgery patients
 - (c) plasma protein abundance in sepsis due to FP
2. comparing QTL across conditions and integrating other relevant data from the literature

MATERIALS AND METHODS

This chapter describes patient recruitment, sample processing, data acquisition, processing, and analysis for transcriptomics, mass spectrometry, and genotyping

2.1	Genomic Advances in Sepsis	37
2.2	Transcriptomics	40
2.3	Mass spectrometry	43
2.4	Genotyping	45
2.5	QTL mapping	47

2.1 Genomic Advances in Sepsis

The Genomic Advances in Sepsis (GAinS) study (www.ukccg-gains.org) was begun in 2005 by the UK Critical Care Genomics group to perform genomics research in sepsis based on a large, homogeneous, and well-characterised patient cohort. A resource of samples and phenotypic information from over 1100 patients admitted to 34 ICUs with severe sepsis due to community acquired pneumonia (CAP) or faecal peritonitis (FP) has since been assembled, with recruitment still ongoing at 7 sites across the UK (Table 2.1, Fig. 2.1).

Functional genomics data have been generated for several cohorts of sepsis patients recruited to GAinS. These datasets form the basis for this thesis, and those described here include two genotyping datasets, four microarray gene expression datasets, and one mass spectrometry proteomics dataset.

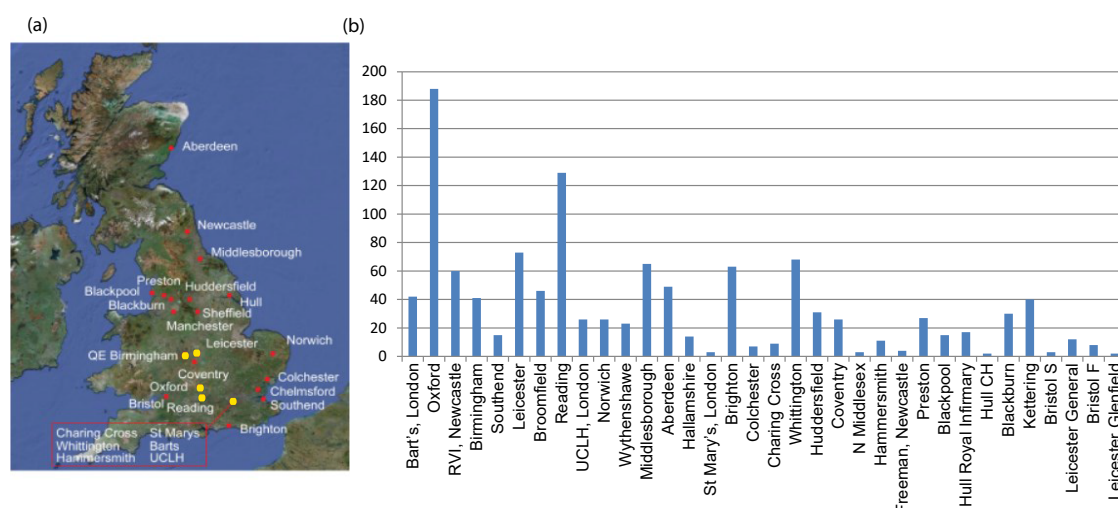


Figure 2.1: Overview of GAINs recruitment.

Functional genomics samples have been collected from 1178 patients recruited to the Genomic Advances in Sepsis (GAINs) study. (a) These patients were admitted to 34 intensive care units in 22 cities across the UK, shown on the map (adapted from Davenport 2014). Currently, seven of these sites are actively recruiting, indicated in yellow. (b) The total number of patients recruited at each site is shown as a bar chart.

2.1.1 GAINs patient recruitment and exclusion criteria

Sepsis patients were recruited through the GAINs study from 34 participating ICUs between 2005 and 2016. Ethics approval was granted nationally and for individual participating centres, and informed consent was obtained from the patient or from their legal representative. Adult patients (≥ 18 years of age) with sepsis due to community acquired pneumonia (CAP) or faecal peritonitis (FP) were enrolled by research nurses. Sepsis was diagnosed according to ACCP/SCCM guidelines (Sepsis-2; Bone et al. 1992), detailed in Table 2.2, and classed as severe when associated with organ dysfunction, hypoperfusion abnormality or sepsis-induced hypotension. All patients were classed as having severe sepsis during ICU admission using these guidelines. CAP was defined as febrile illness associated with cough, sputum production, breathlessness, leukocytosis and radiological features of pneumonia acquired in the community or within two days of admission to hospital. FP was defined as inflammation of the peritoneal membrane secondary to faecal contamination, diagnosed by

Number of patients	RNA		DNA		RNA + DNA	
	CAP	FP	CAP	FP	CAP	FP
Not yet processed	158	72	282	146	340	168
Processed	598	315	441	243	349	206
Profiled	410	139	592*	313*	350*	103
Total recruited	756	387	723	389	689	374

Table 2.1: Summary of available GAINs samples

The number of patients recruited to the GAINs study are given by cause of sepsis (CAP or FP) and stage of functional genomics profiling. DNA and RNA are extracted from at least one serial blood samples taken within the first five days following ICU admission.

* DNA samples from the GenOSept study were included in genotyping assays.

Clinical manifestations of SIRS

Temperature $>38^{\circ}\text{C}$ or $<36^{\circ}\text{C}$

Heart rate >90 beats/minute

Respiratory rate >20 breaths/minute (tachypnea) or $\text{PaCO}_2 < 32$ mm Hg (hyperventilation)

Alteration in white blood cell count: $>12,000/\text{cu mm}$, $<4,000/\text{cu mm}$, or $>10\%$ percent immature neutrophils

Table 2.2: Clinical manifestations of SIRS (Bone et al. 1992)

SIRS is diagnosed in patients with at least two of these clinical manifestations. Sepsis is diagnosed in patients with SIRS in the presence of an infection.

laparotomy. Exclusion criteria are immunocompromise, admission for palliative care only, and pregnancy.

2.1.2 Cardiac surgery patient recruitment and exclusion criteria

Patients scheduled for elective cardiac surgery were recruited by Dr Eduardo Svoren. Ethical approval was granted by the East London and the City Ethics Committee. Patients presenting for the first time for coronary artery bypass grafting, valve replacement, or repair requiring cardiopulmonary bypass were enrolled at St Bartholomew's Hospital following written consent. Patients were excluded if they were immunocompromised (HIV positive, neutropenic, or

regular therapy with immunosuppressants including steroids), undergoing an emergency operation, had malignancy, or could not give informed consent.

2.2 Transcriptomics

2.2.1 Sample collection

Serial samples for RNA extraction were taken from septic patients on the first, third, and/or fifth day after ICU admission. For surgical patients, samples were obtained before anaesthesia, and on the first day following surgery. For both cohorts, 5ml of blood were collected into Vacuette EDTA tubes, and the leukocyte population isolated at the bedside using the LeukoLOCK filter system (Life Technologies). The samples were stabilised using *RNAlater* and stored at -80°C . Plasma samples were also taken at each time point, along with a buffy coat and whole blood sample upon recruitment for DNA extraction.

Comprehensive anonymised clinical information was recorded for each sepsis patient using an electronic Case Report Form (eCRF), including differential white cell counts, mortality, and disease severity scores. The complete list of measurements recorded is given in the Appendix (Tables A.1, A.2, A.3).

2.2.2 RNA extraction

Purified RNA was extracted from the leukocyte samples using the Total RNA Isolation Protocol (Life Technologies). The contents of the filters were collected following lysis using a guanidine thiocyanate-based solution. Cellular proteins and DNA were degraded using Proteinase K and DNase I respectively, and RNA purified using magnetic bead technology. RNA quality and concentration was assessed using spectrophotometry (NanoDrop) and on-chip gel electrophoresis

for a subset of samples (BioAnalyzer 2100, Agilent).

2.2.3 Microarray analysis

Data had previously been generated at the Wellcome Trust Sanger Institute for the Radhakrishnan 2010 (127 CAP, 94 FP, 30 surgery samples) and Davenport 2011 (265 CAP, 126 surgery samples) cohorts (Radhakrishnan 2012; Davenport 2014; Svoren 2015). Genome-wide gene expression was quantified using the Illumina Human-HT-12v4 BeadChip comprising 47,231 probes at the Wellcome Trust Centre for Human Genetics (WTCHG), Oxford (Core Genomics) for a further two cohorts. The Burnham 2014 cohort consisted of 168 sepsis samples, using the first available time point with high quality RNA for each patient. For validation of survival comparisons, all available CAP non-survivors (n=57) were included with 57 age- and sex-matched CAP survivors. Additionally, 54 of these CAP patients were used together with 54 age and sex-matched FP samples to validate diagnostic comparisons. The Burnham 2016 cohort comprised all patients with good quality RNA from all time points available, for whom no gene expression data had previously been generated (72 FP, 72 CAP samples). Patients were allocated randomly to the microarray chips to avoid biases in expression estimates, with samples from the same patient placed on the same microarray chip to allow intra-individual comparison, and raw data were prepared using Illumina's Genomestudio (WTCHG).

2.2.4 Acquisition of publicly available data

Publicly available microarray gene expression data were accessed through NCBI GEO database using GEOquery (Davis and Meltzer 2007). These included data on traumatic injury (from The "Inflammation and the Host Response to Injury" research programme, Xiao et al. (2011)), paediatric sepsis (Wong et al.

2009a), and models of endotoxin tolerance (Fresno et al. 2009; Pena et al. 2011) (accession: GSE36809; GSE26440; GSE15219; GSE22248) .

2.2.5 Data processing and analysis

Data processing and analysis was carried out using R, unless otherwise stated. Demographic data were compared between groups by χ^2 test for count data, t-test for continuous data, and log rank test for survival.

All four gene expression datasets were processed in parallel using the same methods. Background intensities were subtracted and probes with a detection p-value <0.05 in at least 5% of samples retained for downstream analysis. The raw data were transformed and normalised using the Variance Stabilisation and Normalisation R package (Huber et al. 2002). In order to test the diagnostic, survival, and SRS models, validation data were normalised against the discovery data. Subsequently, datasets were combined using ComBat to correct for batch effects (Johnson et al. 2007b). Quality control (QC) checks including principal component analysis (PCA) were used to identify batch effects and outlying samples.

Differential gene expression was analysed using the R package limma (Smyth 2004), which fits a generalised linear model to the expression of each gene and employs an empirical Bayes approach to take overall variance in the dataset into account. Genes with an FDR <0.05 and a fold change ≥ 1.5 were considered to be differentially expressed. The rotation gene set test (ROAST) (Wu et al. 2010) was used to test for overlap in gene lists from different contrasts. Power calculations were performed using the R package sizepower (Qiu W and Whitmore 2006). Pathway enrichment analysis was carried out with Ingenuity Pathway Analysis (QIAGEN, www.qiagen.com/ingenuity, QIAGEN Redwood City), and with XGR (Fang et al. 2016) using the “lea” algorithm to account for the hierarchical

structure of the ontology and taking all genes tested for differential expression as the background.

Variables were selected for use in logistic regression prediction models using GeneRave (Kiiveri 2008). This package uses a sparse regression method to select a minimal number of explanatory variables from a large number of potential predictors. ROC curves were plotted with the ROCR package (Lengauer et al. 2005) to assess the models, using leave-one-out cross validation in the discovery cohort prior to testing on new data.

The FactoMineR package (Lê et al. 2008) was used for unsupervised clustering based on the 10% most variable probes. With this approach, agglomerative hierarchical clustering with Ward's method and Euclidean distance is used to initiate the appropriate number of clusters, followed by *k*-means clustering to consolidate group membership. The within-group sum of squares is calculated for different numbers of groups to find the point where further dividing the data no longer significantly decreases the within-group dispersion.

Partial least squares regression (PLS) was carried out to assess the global response to cardiac surgery using the operative status as a binary response variable and all gene expression probes that passed QC as explanatory variables (Wehrens and Mevik 2007). The supervised PLS approach identifies variance components that differentiate treatment groups, unlike PCA, which is unsupervised (Humburg et al. 2016).

2.3 Mass spectrometry

2.3.1 Sample processing

Proteins were identified and quantified in 60 samples from 55 patients. Sample preparation from plasma was carried out with Marie-Laetitia Thézenas (Target

Discovery Institute, University of Oxford). This involved delipidation, depletion of abundant proteins, protein precipitation, digestion, desalting, and drying. An initial optimisation experiment was carried out comparing two protein precipitation methods, involving either a chloroform/methanol method (Wessel and Flügge 1984) or a trichloroacetic acid/deoxycholate protocol (Bensadoun and Weinstein 1976). The latter protocol was found to yield more protein, and so was used for the main experiment.

Plasma samples were delipidated by centrifugation, followed by aspiration of 50 μ l from the plasma layer. Protein quantification was performed prior to processing using the BCA assay (Walker 1994). The 12 most abundant proteins were depleted using Top12 columns (Thermo, manufacturer's protocol). Briefly, the samples were incubated in the Top12 cartridges for one hour at room temperature, then the non-abundant proteins were recovered by centrifugation. Protein was precipitated and the pellet washed by centrifugation with acetone and solubilised in urea. The protein was then digested using the Smart digest kit (Thermo; manufacturer's protocol). The samples were desalted using acetonitrile (ACN), trifluoroacetic acid (TFA), and formic acid, and the elute was then vacuum dried and resuspended in 20 μ l 0.1% TFA and 1% ACN buffer. Peptide quantification was carried out by colorimetric assay (Thermo). Concentrations were normalised across the samples and 2 μ l injected into the Q-Exactive with a 1 hour gradient. A pool of all the samples was also created for data alignment.

2.3.2 Data processing and analysis

The data were generated using nano-liquid chromatography electrospray ionisation tandem mass spectrometry. Primary analysis was performed with Progenesis LC-MS software using default settings. All chromatograms were aligned based on retention score and mass to charge ratio (m/z) using data from pooled samples as a reference (Marie-Laetitia Thézenas). Species identification

was performed with PEAKS (Zhang et al. 2012) with a mass tolerance of 10ppm and allowing one missed cleavage. Carbamidomethylation of cysteine was a fixed modification, and oxidation of methionine and deamidation of asparagine and glutamine were used as variable modifications. The false discovery rate was controlled at 1% using a decoy search to generate the background, and an ion score cut-off of 20 was used. Protein abundance was then calculated in Progenesis LC-MS using the area under the curve for non-degenerate peptides.

The raw data were $\log_2$ transformed such that they followed a normal distribution. Samples were ranked by their standard deviation, and the bottom 90% were used to calculate the median and median absolute variation (MAD) for each feature. Median centering normalisation was then carried out, and features missing in 90% of samples or more removed. Remaining missing data points were imputed by random draws from a distribution with a mean and standard deviation estimated from the values available for that feature (Philip Charles, personal communication). Samples were contrasted using limma (Smyth 2004) to identify differentially abundant proteins. Multiple testing was accounted for using q values (Dabney et al. 2010).

2.4 Genotyping

2.4.1 DNA extraction

DNA had previously been extracted from buffy coat or whole blood samples using the Qiagen DNA extraction protocol, the automated Maxwell 16 Blood purification kit (Promega), or the QIAamp Blood Midi kit protocol (Qiagen) by Jay Radhakrishnan, Emma Davenport, Tara Mills, Anna Rautanen, and Narelle Maugeri. DNA from a further 185 samples was extracted using the Maxwell machine, which uses paramagnetic particles to purify DNA following cell lysis.

The DNA was quantified by spectrophotometry (NanoDrop) and the Quant-iT PicoGreen kit (Invitrogen).

2.4.2 Genotyping and data processing

Genotyping data had been previously generated for 275 CAP and 63 surgery patients and 730,525 SNPs using the Illumina HumanOmniExpress BeadChip (Davenport 2014). A further 655 patients were genotyped at the Wellcome Trust Sanger Institute using the Infinium CoreExome BeadChip (Illumina; 551,839 SNPs) and the Illuminus genotype calling algorithm. Genotyping QC was performed in PLINK 1.9 (Purcell and Chang 2015; Chang et al. 2015) according to the methods described in Anderson et al. (2010). Samples were excluded on the basis of discordant sex information, proportion of missing genotypes > 0.02 , heterozygosity rate, and identity by descent. SNPs were excluded if they had a missing data proportion > 0.05 , minor allele frequency (MAF) < 0.01 , and Hardy-Weinburg equilibrium (HWE) $P < 1 \times 10^{-5}$.

2.4.3 Imputation

The two sepsis genotyping datasets were imputed independently against the Haplotype Reference Consortium (HRC) release 1.1 panel using the Sanger Imputation Service (McCarthy et al. 2016). Pre-imputation checks were carried out using the HRC Imputation Tool (Will Rayner, WTCHG; www.well.ox.ac.uk/~wrayner/tools), which checks and corrects strand, alleles, position, Ref/Alt assignments, and frequency differences against the HRC panel. Genotypes were phased using Eagle2 (v2.0.5) (Loh et al. 2016) and imputed using PBWT (Durbin 2014). SNPs with an info score < 0.9 were removed, the two datasets were combined, and SNPs with a MAF < 0.05 or HWE $P < 0.001$ were removed. The final data comprised 910 individuals and 4,484,827 SNPs.

2.5 QTL mapping

Both gene expression and genotyping data were available for 468 sepsis patients, with multiple gene expression time points giving a total of 592 samples. The gene expression data were reduced to autosomal uniquely mapping probes that did not overlap a reported SNP, and eQTL were defined as cis when the distance between the SNP chromosomal position and transcriptional start site was $\leq 250\text{kb}$. Cis-eQTL mapping was performed using the methods described in Davenport et al. 2017. A linear mixed model was used to determine SNP-probe associations, including as covariates a random effect for patient, the first 30 principal components of the gene expression data, and the first six principal components of the genotyping data:

$$E_{i,j} = \theta + \beta_{\text{geno}} \cdot g_j + (\kappa_j | i) + \sum_{l=1}^{30} \phi_l \cdot pc_{i,l} + \sum_{m=1}^6 \gamma_m \cdot pc_{j,m}$$

$E_{i,j}$	gene expression for the i^{th} sample from the j^{th} subject
θ	intercept
β_{geno}	effect of genotype
g_j	genotype of the j^{th} subject
$(\kappa_j i)$	random effect for the i^{th} sample from the j^{th} subject
$pc_{i,l}$	principal component l of gene expression for sample i
$pc_{j,m}$	principal component m of genotyping for subject j

A Bonferroni corrected p-value cut off (5.2×10^{-9}) was used to determine the significance of baseline eQTL.

These baseline eQTL were tested for interactions with clinical features if there were at least two minor allele homozygotes, and significance was assessed by a likelihood ratio test:

$$E_{i,j} = \theta + \beta_{geno} \cdot g_j + (\kappa_j | i) + \sum_{l=1}^{30} \phi_l \cdot pc_{i,l} + \sum_{m=1}^6 \gamma_m \cdot pc_{j,m} + \beta_{clin} \cdot d_i + \beta_x \cdot d_i \cdot g_j$$

β_{clin}	effect of clinical feature
d_i	value of clinical feature
β_x	effect of interaction between genotype and clinical feature

Similarly, eQTL were mapped using gene expression and genotyping data for 53 cardiac surgery patients, each with two timepoints. The gene expression data were restricted to the genes differentially expressed between pre- and post-surgery samples. The first 15 principal components of the gene expression data were included in the model, and a corrected p-value threshold of 5.5×10^{-7} was used. Interactions with surgical status were then identified for significant eQTL. Associations for SNP-gene pairs tested across data sets were compared on the basis of Z scores (beta value/standard error).

eSNPs were annotated with chromatin features for LPS-stimulated monocytes generated by the Blueprint Consortium, including histone marks (H3K27ac, H3K4me1 and H3K4me3) and DNase I hypersensitivity (Saeed et al. 2014). The proportion of eSNPs overlapping each mark was compared to the proportion for SNPs within the cis window of the genes tested by Fisher's Exact test, and the distance to the nearest mark by Mann-Whitney test. The impact of specific SNPs was explored using the Variant Effect Predictor (Ensembl, McLaren et al. (2016)).

THE TRANSCRIPTOMIC RESPONSE TO SEPSIS AND RELATED CONDITIONS

This chapter explores the gene expression changes over time in sepsis due to community acquired pneumonia or faecal peritonitis, and compares this to the sterile inflammatory response following cardiac surgery or traumatic injury

3.1	Introduction	49
3.2	Results: sepsis	57
3.3	Results: cardiac surgery	86
3.4	Results: comparison across SIRS sources	91
3.5	Discussion	100
3.6	Conclusions	106

3.1 Introduction

Sepsis is a clinical syndrome with no diagnostic test, no currently approved targeted treatments, and enormous variation in clinical manifestation. It is the most common reason for ICU admittance, and mortality rates are estimated to be more than 10%, consistently rising to 40% in septic shock. Clinical management of sepsis therefore presents a great challenge. In this chapter, I aim to leverage transcriptomic data to explore the dysregulated host response that causes sepsis and thus gain insights into pathogenesis in sepsis and related conditions.

3.1.1 Difficulties in defining and diagnosing sepsis

Sepsis is currently defined as “life-threatening organ dysfunction caused by a dysregulated host response to infection”. This definition comes from The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) published in February 2016, which advocated moving away from the Sepsis-2 “SIRS” criteria as indicators for sepsis. According to these new guidelines, sepsis is indicated by the presence of clinical signals such as an altered mental state, a respiratory rate of 22 or more per minute, or a systolic blood pressure of 100mmHG or less, in patients with a suspected infection.

These simple and general indicators are suggestive of immune dysregulation or inflammation, but are not necessarily directly related to the molecular dysregulation that drives the disease. Hence, treatment strategies that focus on managing the symptoms of sepsis are unlikely to effectively address the underlying causative mechanisms of sepsis pathogenesis. The latter are poorly understood, and furthermore may be modulated by any combination of host-specific and external factors, including but not limited to the causative pathogen, comorbidities, age, sex, and genetic background. Consequently, there is wide variation in clinical presentation and response to treatment, compounding difficulties in diagnosis and disease management.

3.1.2 Transcriptional regulation of the inflammatory response

The inflammatory response to infection is extremely complex, and varies both according to the pathogen and across individuals. It is largely controlled at the level of gene expression, allowing rapid and highly regulated induction of the appropriate factors as and when required. A number of studies have investigated the healthy immune response by examining these transcriptional changes *in vitro* and *in vivo*. Differential expression of multiple coordinately regulated sets of

genes is seen soon after stimuli such as lipopolysaccharide (LPS: a component of gram-negative bacterial cell walls), causing in turn a cascade of further changes in transcript levels.

The sepsis response is evidently similarly complex and multi-faceted, and can also be investigated at the molecular level through genome-wide gene expression profiling. Comparison of patients to healthy individuals or those with a more appropriate immune response is useful in identifying disease-specific features at the molecular level. These can be used to develop diagnostic and prognostic tools, and may grant insights into disease mechanisms and thereby aid drug development.

3.1.3 The tissue of interest

As described above, dysregulation of the immune response is the unifying feature in sepsis. Patient leukocytes are therefore of great interest when studying the sepsis response, being both highly relevant to pathogenesis and relatively accessible in comparison to taking biopsies from the infection site. As such, the total blood leukocyte population was used for functional genomics profiling in the GAinS study.

However, “leukocytes” encompasses a variety of cell types, each with specific roles and programmes of gene regulation. Hence, they should ideally be studied individually, but there are important logistical considerations to cell-specific data generation. For example, cell separation using clinical samples is a time-sensitive process, requiring equipment not available in many of the hospitals involved in the study, and is costly in terms of reagents, time, and effort. Patient recruitment to the GAinS study was instead maximised through use of the LeukoLOCK filters, which retain and preserve leukocytes from a whole blood sample. Importantly, use of the LeukoLOCK system excludes the vast majority of globin mRNA

present in whole blood, which has been shown to reduce the sensitivity of microarray gene expression data, without the need for cell separation or specialist equipment. In addition, processing is extremely rapid and can be done at the bedside, preserving the gene expression profile at the time of collection.

As mentioned above, a notable limitation of this approach is that it precludes the possibility of investigating cell type specific effects. Moreover, variation in cell proportions between individuals could have a confounding effect on the analyses. If a gene expressed exclusively in neutrophils is perceived to be expressed at higher levels in one patient on the basis of whole blood profiling, this might be a true upregulation, or it might be because that patient has a higher neutrophil proportion. This should therefore be taken into account in downstream analysis and interpretation, through deconvolution of cell type-specific effects where possible. With this in mind, differential cell counts were recorded at the time of sampling.

In the particular case of sepsis, a comparison between whole blood and cell specific gene expression has been carried out by Wong et al. (2010a). This demonstrated strong correlation between gene expression patterns in whole blood and in leukocyte subtypes, suggesting that profiling whole blood is a reasonable approach to use to increase sample size significantly while not greatly obscuring signals in the resulting data.

3.1.4 Causes of sepsis

By definition, sepsis may be caused by any infection; however, a positive blood culture is only seen in around 40% of cases in UK hospitals. The causative infection is therefore unknown in a large number of patients, and patients may often be classified by infection site instead.

The GAINs study recruits patients with sepsis resulting from community

acquired pneumonia (CAP), the most common and therefore most studied cause of sepsis, and faecal peritonitis (FP). When known, the causative pathogen is also recorded, but this information is limited to fewer than half of the CAP patients in the GAINs cohort. CAP is defined as pneumonia acquired in the community or within two days of admission to hospital (Musher and Thorner 2014). FP is defined as inflammation of the peritoneal membrane secondary to faecal contamination; a heterogeneous bacterial population is therefore involved (Weledji and Ngowe 2013; Paul et al. 2015). It may be caused by diverticular disease, breakdown of surgical anastomosis, traumatic injury, and many other minor causes. FP is another common cause of sepsis in the UK, in which the transcriptional sepsis response has not yet been described. These diseases are straightforward to diagnose clinically, giving confidence in patient categorisation.

Although sepsis of all aetiologies has conventionally been considered as a single unified disease, it has been suggested that the initial infection could have an impact on pathogenesis (Goh and Knight 2017). A number of studies have investigated the clinical equivalence of sepsis resulting from different infections, for example comparing patients infected with gram-positive and gram-negative bacteria (Gogos et al. 2010). However, no detailed analyses of the effects on the sepsis response have been performed and the impact of causative infection on the sepsis transcriptomic response remains unclear. Sparse microbiological data limit comparisons across all aetiologies, yet information on the source of infection could also be used to compare disease subgroups based on infection site. CAP and FP have different clinical presentations and time courses, and are treated using different strategies; it is therefore plausible that differences in molecular signatures might also be observed.

In this chapter, I investigate the transcriptomic response in sepsis due to CAP and FP patients, describing the FP sepsis response for the first time and comparing

the two subgroups directly. By focussing on these dissimilar infectious origins, we would expect any differences in the sepsis response associated with infection site to be correspondingly pronounced.

3.1.5 Gene expression dynamics in the sepsis response

The sepsis response has conventionally been described as an early hyper-inflammatory host response driving pathogenesis, followed by compensatory immunosuppression that increases susceptibility to secondary infections. This implies that a clear progression from one state to the other is observed, which would allow appropriate treatments to be administered for each stage. More recently it has been recognised that these two phases may coincide and interact, and are both extremely important to disease pathogenesis and outcome (Hotchkiss et al. 2013; Wiersinga et al. 2014). The sepsis response state of an individual is additionally likely to vary over the course of disease progression. The significant heterogeneity seen across patients, in terms of disease severity and response to treatment, may correspond to differences in this host response at the time of observation.

In CAP, it is very difficult to determine the time of disease onset. Given the nature of the illness, patients may be admitted to ICU at very different stages in their disease history, and so the time of RNA sampling relative to admission is unlikely to correspond well to a disease time course across patients. A date of diagnosis has been recorded for each patient, which may improve resolution here; however, there could still be a great deal of variation between patients in terms of rate at which their sepsis response develops. The availability of serial samples from the same patient over the first five days in intensive care enables the comparison of gene expression over the course of disease progression without knowledge of the number of days since disease onset for each patient.

Serial samples are also available for the FP patients, allowing a comparable analysis to be performed. Moreover, there is usually a defined trigger of disease such as surgery, giving a known time of disease onset. This allows all samples to be placed on a general time line according to the number of days since the causative insult, empowering further analysis of gene expression dynamics in FP.

3.1.6 The dysregulated immune response in SIRS

Infection, tissue stress, and injury all trigger the inflammatory response, which may vary according to the stimulus type. As in sepsis, the host response to sterile injuries or surgery can become dysregulated, leading to systemic inflammation and organ damage in a minority of cases. Patients are diagnosed with SIRS (Systemic Inflammatory Response Syndrome) on the basis of the same clinical criteria as sepsis but in the absence of infection. Gene expression in SIRS following trauma and burns has been studied in particular by the “Inflammation and the Host Response to Injury” research programme (Xiao et al. 2011). They describe a “genomic storm” of differential expression and, as in sepsis, there is evidence that features of pro-inflammation and immunosuppression occur concurrently.

Comparisons between SIRS patients of differing causes could be informative for understanding the underlying causes of mortality and morbidity. Elective surgery in particular can also serve as a model for the host response to traumatic injury and sepsis, given the availability of pre-insult samples, clear start and end points for the injury, and similar patient demographics to sepsis cohorts. Analysis of global gene expression before and after surgery could help to resolve heterogeneity in the host response and identify the patients who are at greater risk of adverse peri-operative events, allowing targeting of interventions to these patients to improve outcomes.

In this chapter, I define the host response to cardiac surgery using paired pre- and post-operative gene expression data. Microarray data generated by Dr Eduardo Svoren are available for circulating peripheral blood leukocytes from a group of low risk patients undergoing elective cardiac surgery and cardiopulmonary bypass. These data were generated together with a cohort of GAinS sepsis patients, allowing definition of shared and specific features of the systemic inflammatory response in the presence and absence of an infectious component. Publicly available expression data from the “Inflammation and the Host Response to Injury” research programme enables further comparisons to the trauma response.

3.1.7 Aims

The overall aim of this chapter is to characterise the transcriptomic response to sepsis caused by CAP or FP, and investigate how the systemic inflammatory response is influenced by its cause. This will include comparison of the two sepsis subgroups directly, and contrasting sepsis with the SIRS response defined using in-house and publicly available datasets. Specifically, I will test the following hypotheses:

1. that a subset of genes are differentially expressed between patients with sepsis due to CAP and FP
2. that there are changes in gene expression over time in sepsis
3. that there are widespread transcriptomic changes following cardiac surgery
4. that there are differences in gene expression across the responses to sepsis, surgery, and blunt trauma

3.2 Results: sepsis

3.2.1 Description of sepsis cohorts

In this chapter, I analyse sepsis microarray gene expression data processed in several different batches over a six year period (Fig. 3.1). These consist of samples from patients enrolled in the GAINs study, recruited from 34 UK intensive care units (ICUs) between 2005 and 2016. Serial samples were obtained on the first, third, and/or fifth day of ICU stay. The total blood leukocyte population was isolated from whole blood for RNA extraction using the LeukoLOCK system (Invitrogen), and extensive clinical data were recorded using an electronic case report form (eCRF).

Two of these datasets were produced by former DPhil students in the J Knight lab (Radhakrishnan 2010: Dr Jayachandran Radhakrishnan; and Davenport 2011: Dr Emma Davenport), whilst I generated the later two as validation cohorts as part of this thesis (Burnham 2014 and Burnham 2016). Samples from cardiac surgery patients taken before and after their operation using the GAINs protocol were also included in the earlier two cohorts. These samples were obtained and processed by Dr Eduardo Svoren, and are described in greater detail below. These four cohorts are used in this chapter to compare gene expression between various clinical patient subgroups, as illustrated in Fig. 3.2.

Genome-wide expression data were generated for all cohorts using the Illumina Human-HT-12 v4 Expression BeadChip (48,000 probes), and I applied the same data processing pipeline to each dataset in parallel, enabling combination of the data and validation of results. An description of these four datasets and a summary of the quality control (QC) is given below.

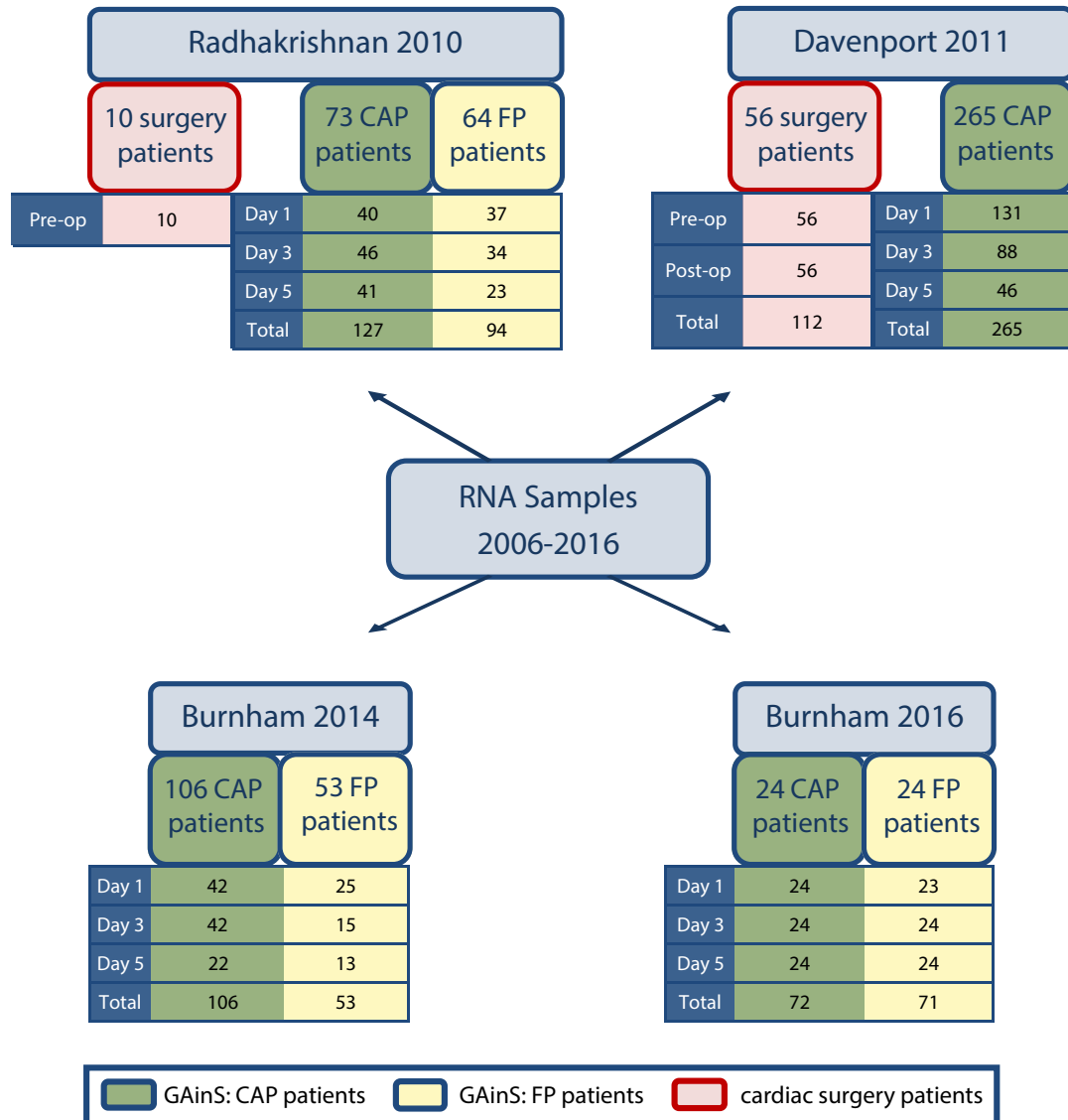


Figure 3.1: Overview of the GAINs gene expression datasets analysed in this thesis.

The number of samples for which gene expression data are available following quality control is given for each of the four cohorts analysed in this thesis. This is subdivided according to the cause of sepsis (CAP (*green*) or FP (*yellow*)), given the focus on source of infection in this chapter. Cohorts are named according to the person who generated the data and the year this was done. Samples from elective cardiac surgery patients (*red*), taken before and after their operation by Eduardo Svoren, are also included in the Radhakrishnan 2010 and Davenport 2011 cohorts.

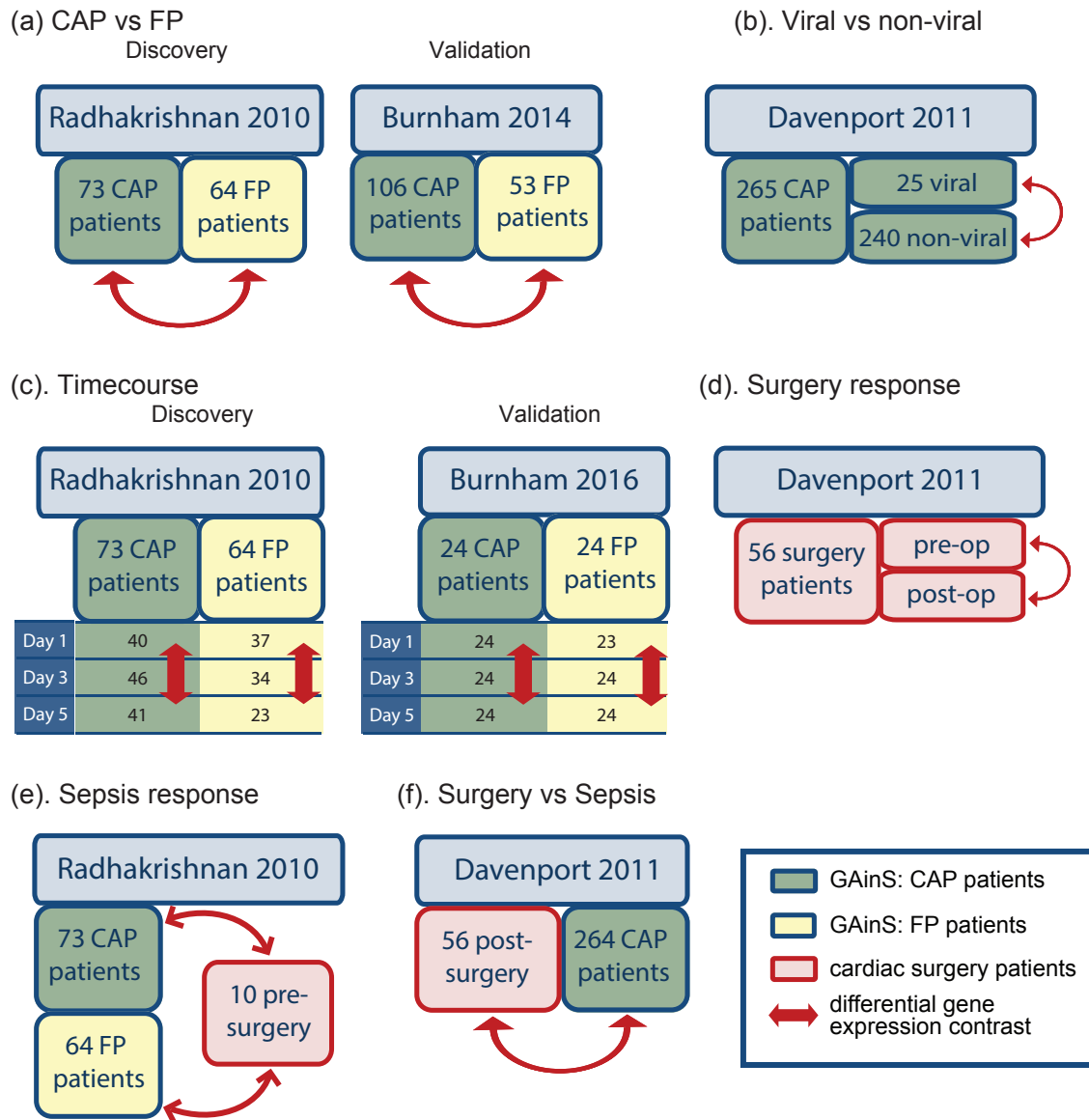


Figure 3.2: Overview of the gene expression contrasts performed in this chapter.

The cohort used and the number of samples contrasted are shown for each gene expression comparison. These include (a) CAP vs FP, (b) viral vs non-viral, (c) changes within patients over time, (d) the surgery response, (e) the sepsis response, and (f) post-surgery vs sepsis.

Radhakrishnan 2010

The Radhakrishnan 2010 dataset has been previously described in Radhakrishnan 2012. Data had been generated for all good quality samples from patients recruited to GAINs prior to January 2010. These included some serial samples taken within five days of ICU admission for a subset of the patients, giving a total of 234 samples from sepsis patients. Thirty samples from ten cardiac surgery patients were additionally processed in parallel; I retained only the pre-operative cardiac samples to represent a non-septic population with similar ages and comorbidities as the GAINs cohort. Seven replicate samples that had been included for technical quality control were removed prior to data normalisation.

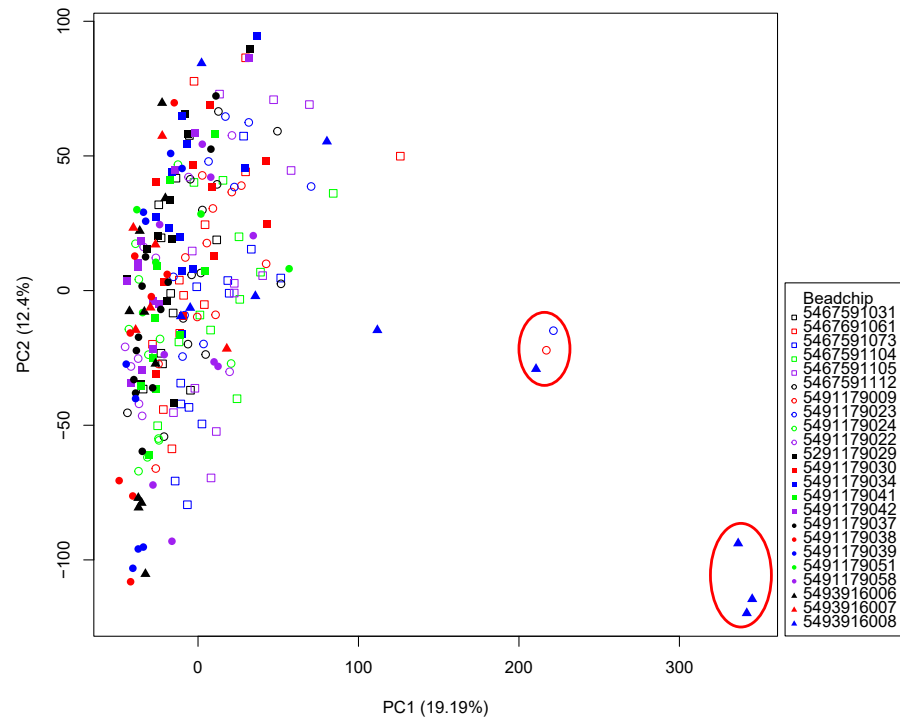
Six outlying sepsis samples were identified through Principal Component Analysis (PCA) and other QC measures (Fig. 3.3), and probes that did not have a detection p-value <0.05 in at least 5% of samples were removed (20,072 probes). Following normalisation and QC, expression data were available for 27,159 probes in 231 samples (127 CAP, 94 FP, 10 pre-surgery) from 147 patients (73 CAP, 64 FP, 10 pre-surgery).

	Overall	CAP	FP
Patients	137	73	64
Samples (D1, D3, D5)	221 (77, 80, 64)	127 (40, 46, 41)	64 (37, 34, 23)
Mean Age ($\pm$ SD)	62.7 ($\pm$ 16.6)	60.6 ($\pm$ 15.6)	65.1 ($\pm$ 17.5)
% Male	47.4%	50.6%	43.7%
Mortality rate	18.2%	20.5%	15.6%

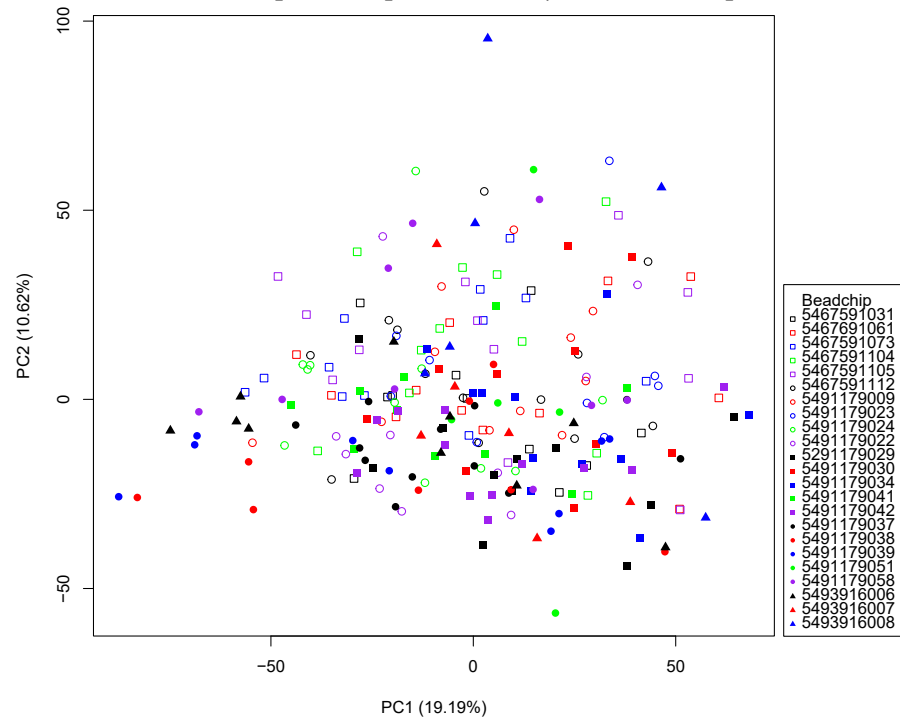
Table 3.1: Radhakrishnan 2010 Cohort Demographics

Demographic and clinical information are summarised for the whole cohort, and by source of infection. Serial samples were taken on the first (D1), third (D3), and fifth (D5) day of ICU admission, and mortality status at 28 days following admission recorded.

The final cohort had an overall mortality rate of 18.2%, which is consistent with that observed across the entire GAINs study. The mean age was 62.7, and 47.4% of patients were male. The samples included are relatively evenly spread across the three timepoints (Day 1 77, Day 3 80, Day 5 64). The CAP and FP subgroups are



(a) Principal Component Analysis of all samples



(b) Principal Component Analysis of 231 samples following outlier removal

Figure 3.3: Principal component analysis: Radhakrishnan 2010 cohort.

The first two principal components of the gene expression data are plotted, with the amount of variance explained by each noted; (a) when all samples are included in the analysis, and (b) after the removal of six outliers. Samples are coloured according to bead chip, and outlying samples are circled in red.

similarly representative of the whole study, with the FP patients having a lower mortality rate (15.6% vs 20.5%), a higher mean age (65.1 vs 60.6), and a lower proportion of males (43.7% vs 50.6%). A comprehensive comparison of clinical covariates between CAP and FP is given below. Only summary demographic information is available for the cardiac surgery cohort, which demonstrates that the mean age of the surgery cohort (66.4) is very comparable to the sepsis cohort, but that it has a higher proportion of males (73.9%).

Davenport 2011

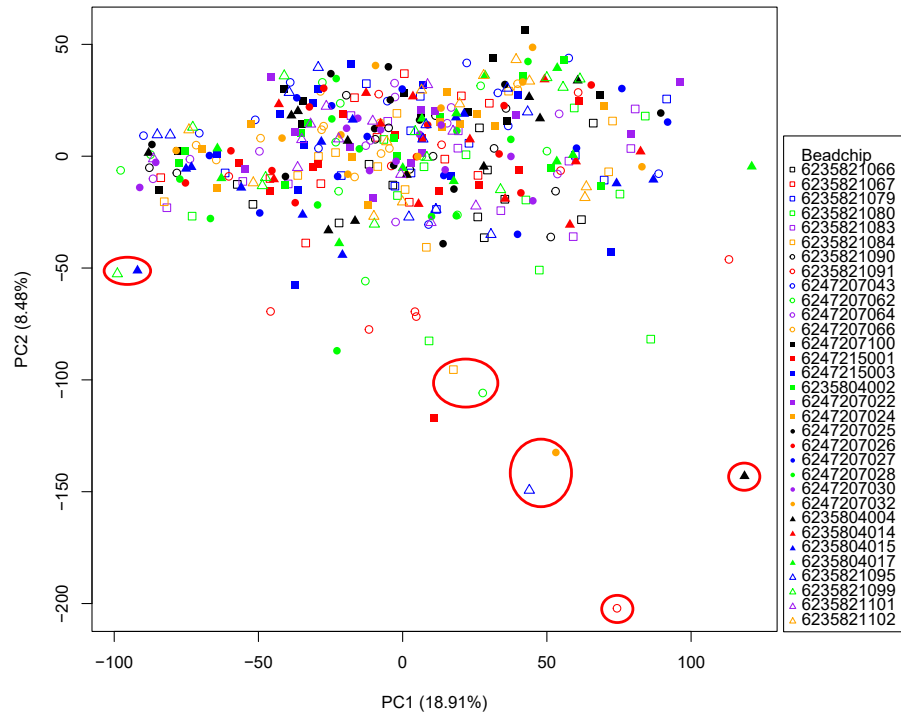
Also analysed in this chapter is the Davenport 2011 dataset, published as the discovery cohort of 265 CAP patients in Davenport et al. 2016. Data had been generated for the first available good quality sample for all CAP patients of European ancestry recruited to GAINs at the time, and were processed using identical quality control and processing as described above, resulting in the published dataset of 26,185 probes (Table 3.2). The cohort has an average age of 62, with a slightly higher proportion of males than observed in the Radhakrishnan 2010 cohort. The mortality rate at 28 days is 21%, with 27% of patients having a known bacterial infection and 9% a viral infection.

	CAP
Patients	265
Samples (D1, D3, D5)	265 (131, 88, 46)
Mean Age ($\pm$SD)	62 ($\pm$ 16)
% Male	55%
Mortality rate	21%

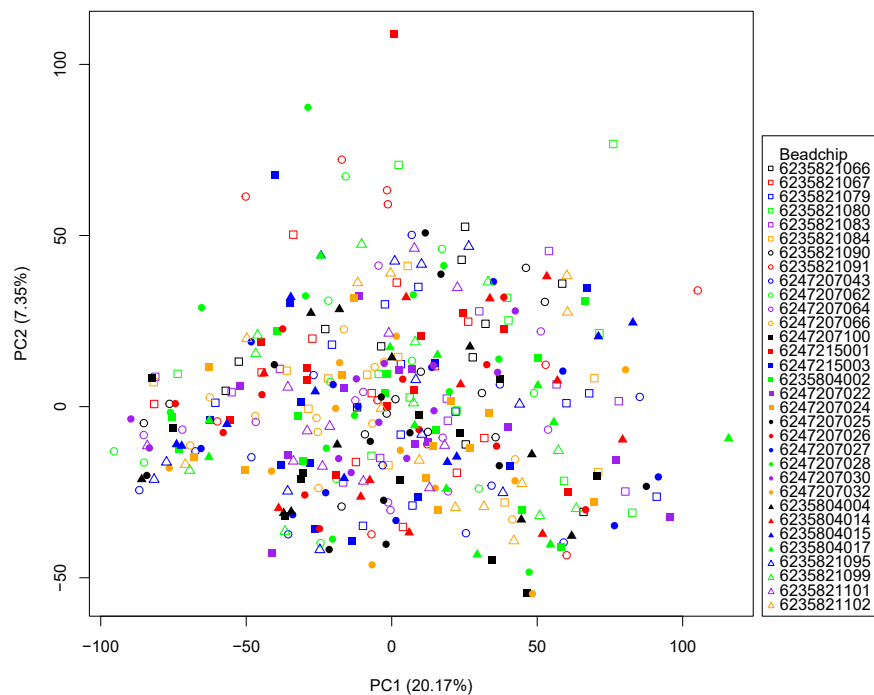
Table 3.2: Davenport 2011 Cohort Demographics

Demographic and clinical information are summarised for the whole cohort. Serial samples were taken on the first (D1), third (D3), and fifth (D5) day of ICU admission, and mortality status at 28 days following admission recorded.

However, these microarray data had been generated in combination with a larger cardiac surgery cohort (described in more detail below) to enable direct



(a) Principal Component Analysis of all samples



(b) Principal Component Analysis of 376 samples following outlier removal

Figure 3.4: Principal component analysis: Davenport 2011.

The first two principal components of the gene expression data are plotted, with the amount of variance explained by each noted; (a) when all samples are included in the analysis, and (b) after the removal of eight outliers circled in *red*. Samples are coloured according to bead chip.

comparison of the sepsis and surgery responses. I therefore renormalised the Davenport 2011 dataset as follows. Raw microarray data for pre- and post-surgical samples from 63 patients and 306 sepsis patients were imported into R, together with associated meta data. During the sample processing, four microarray chips had suffered from a hybridisation failure and so samples from these chips were excluded (34 sepsis, 14 surgery). Probes were filtered such that probes with a detection p-value > 0.05 in more than 95% of samples were excluded. The data were then normalised using the R package *vs*n and quality control plots generated. Eight outlying samples were identified and removed, and QC plots regenerated to check the data. The final dataset consists of 264 sepsis samples, and 56 pairs of cardiac samples (Fig. 3.4).

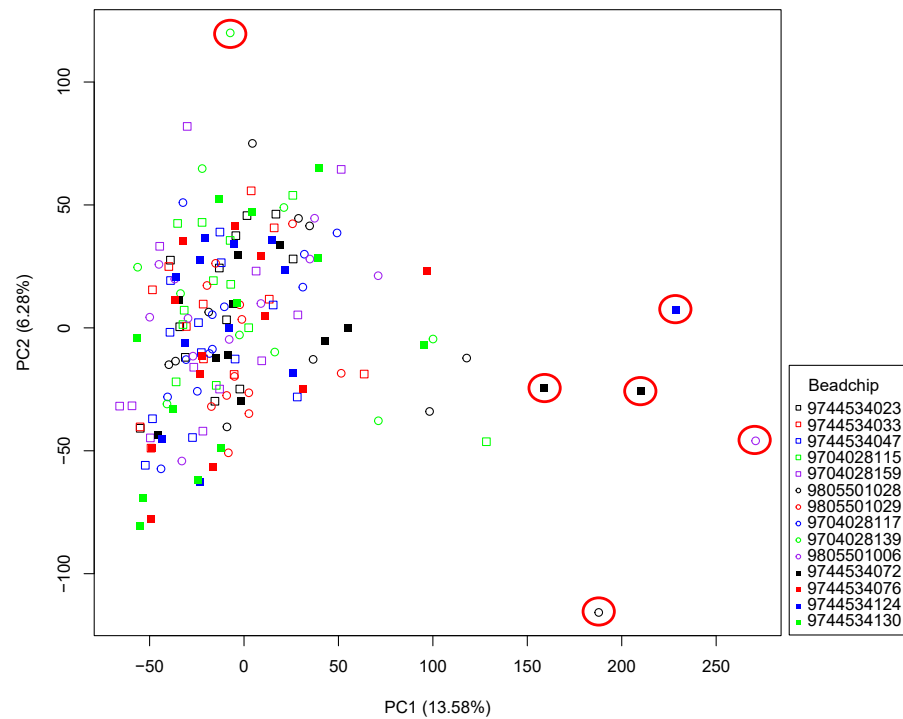
Burnham 2014

The Burnham 2014 cohort was designed as a validation cohort for the Radhakrishnan 2010 and Davenport 2011 cohorts, and therefore comprises the first available good quality sample for selected patients not included in either discovery cohort. First, I selected 57 CAP non-survivors and 57 age- and sex-matched CAP survivors for the validation of observations made in the Davenport 2011 data. The mortality rate is therefore 50% for the CAP samples pre-QC, and there is a higher mean age and proportion of males than in previous cohorts. A

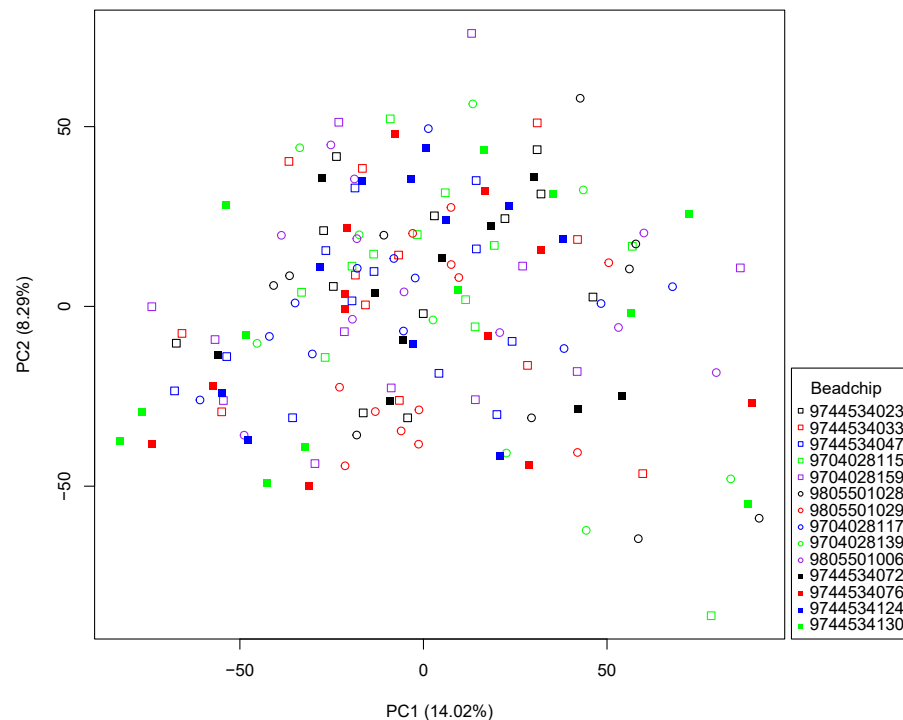
	Overall	CAP	FP
Patients	159	106	53
Samples (D1, D3, D5)	159 (67, 57, 35)	106 (42, 42, 22)	53 (25, 15, 13)
Mean Age ($\pm$ SD)	68.7 ($\pm$ 13.8)	69.1 ($\pm$ 14.4)	68 ($\pm$ 12.7)
% Male	69.2%	74.5%	58.5%
Mortality rate	37.1%	49.1%	13.2%

Table 3.3: Burnham 2014 Cohort Demographics

Demographic and clinical information are summarised for the whole cohort, and by source of infection. Serial samples were taken on the first (D1), third (D3), and fifth (D5) day of ICU admission, and mortality status at 28 days following admission recorded.



(a) Principal Component Analysis of all samples



(b) Principal Component Analysis of 159 samples following outlier removal

Figure 3.5: Principal component analysis: Burnham 2014.

The first two principal components of the gene expression data are plotted, with the amount of variance explained by each noted; (a) when all samples are included in the analysis, and (b) after the removal of six outliers circled in *red*. Samples are coloured according to bead chip.

further 54 FP patients age- and sex-matched to 54 of the CAP patients (chosen so that mortality was representative of the entire study population) were included, giving a total of 168 patient samples prior to quality control.

Following QC and data normalisation, data were available for 106 CAP and 53 FP patients (Table 3.3, Fig. 3.5). Due to the use of this cohort for validation, I used the vsn R package to normalise the CAP data against the Davenport 2011 cohort, and the whole dataset against the Radhakrishnan 2010 cohort; therefore, the data were restricted to the probes retained in each of these cohorts.

Burnham 2016

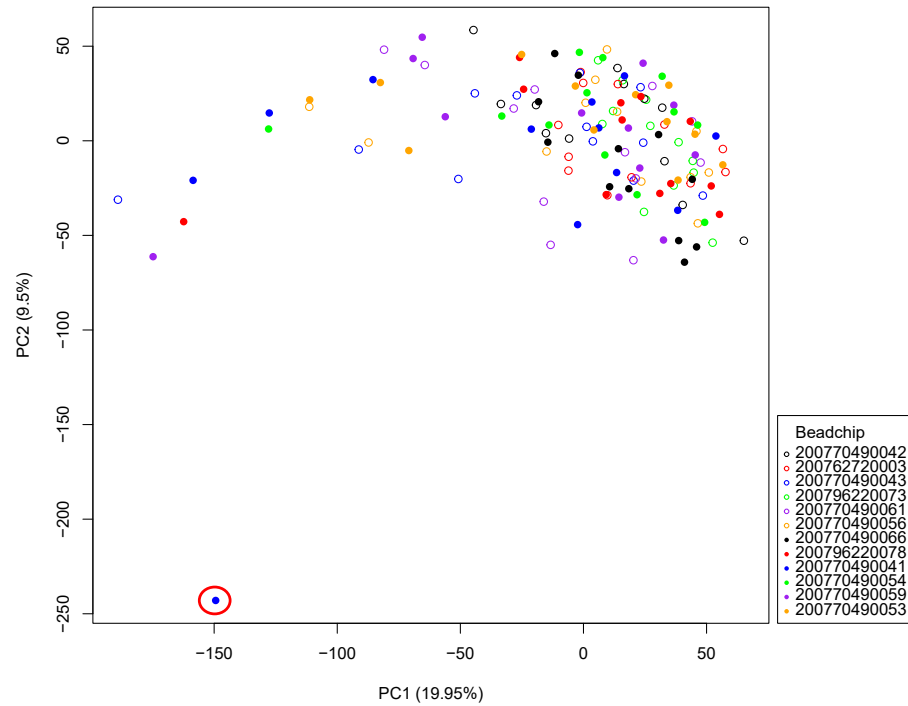
Finally, the Burnham 2016 cohort was designed for time course profiling. I selected patients with three timepoints (Day 1, 3, and 5 of ICU stay) from subjects with no gene expression data already existing. A total of 24 CAP and 24 FP patients were included, and one sample was identified as an outlier in QC as described above (Table 3.4, Fig. 3.6). All three samples from this patient were removed for time course analysis.

This cohort had a higher proportion of males (63%) and mean age (68.4) than the Radhakrishnan 2010 and Davenport 2011 cohorts, but a similar mortality rate at 28 days (21.3%). This is notable, given the requirement for a sample to be

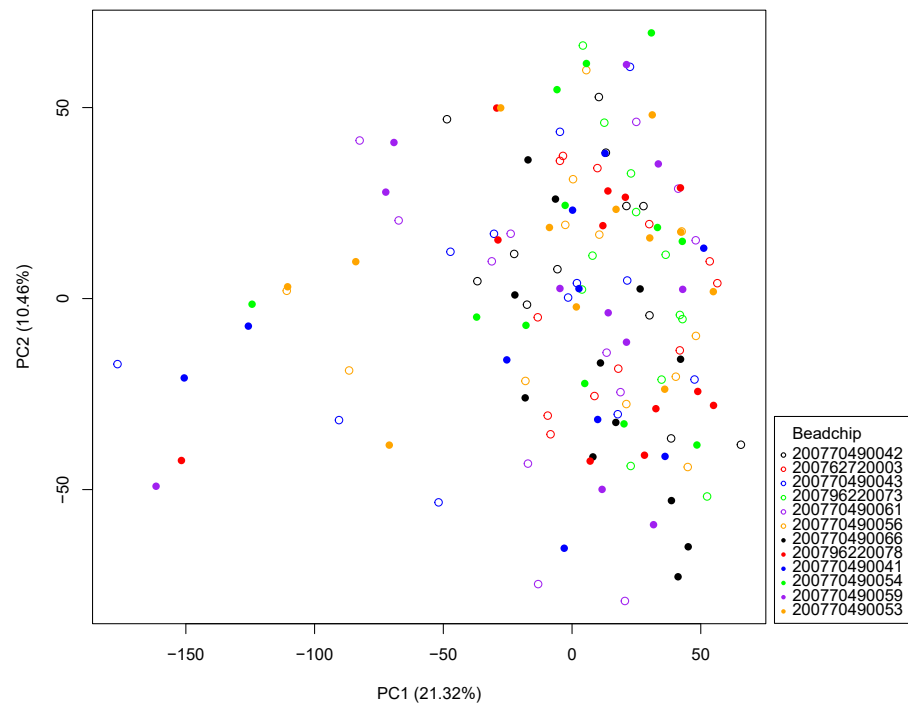
	Overall	CAP	FP
Patients	47	24	23
Samples (D1, D3, D5)	141 (47, 47, 47)	72 (24, 24, 24)	71 (23, 23, 23)
Mean Age ($\pm$SD)	68.4 ($\pm$ 13)	68.6 ($\pm$ 13.5)	68.2 ($\pm$ 12.5)
% Male	63%	68.5%	57.4%
Mortality rate	21.3%	20.4%	22.2%

Table 3.4: Burnham 2016 Cohort Demographics

Demographic and clinical information are summarised for each whole cohort, and by aetiology. Serial samples were taken for every patient on the first (D1), third (D3), and fifth (D5) day of ICU admission, and mortality status at 28 days following admission recorded.



(a) Principal Component Analysis of all samples



(b) Principal Component Analysis of 143 samples following outlier removal

Figure 3.6: Principal component analysis: Burnham 2016.

The first two principal components of the gene expression data are plotted, with the amount of variance explained by each noted (a) when all samples are included in the analysis, and (b) after the removal of one outlier circled in *red*. Samples are coloured according to bead chip.

available on the fifth day in ICU for inclusion in the cohort (and indeed for the first and third day). This could potentially bias the cohort towards patients who survive longer, and would certainly exclude any patient who died within the first four days in ICU. This should be considered in downstream analyses.

3.2.2 Comparison of clinical covariates in CAP and FP

Initial analysis focussed on whether or not the sepsis response was specific to the source of infection. This made use of the Radhakrishnan 2010 cohort as a discovery cohort and the Burnham 2014 cohort for validation, as these included patients with sepsis due either CAP or FP and so permitted direct comparison of these two common causes of sepsis (Fig. 3.1, 3.2a).

First, the two disease subgroups in the Radhakrishnan 2010 cohort were compared on the basis of all available demographic and phenotypic information, and a multiple testing correction applied (Table A.1, A.2, A.3). This demonstrated that there were a number of differences between the groups, some of which were expected from the nature of the causative infection type whilst others were less predictable. For example, FP patients had more instances of GI-related comorbidities while CAP patients had more respiratory comorbidities, higher respiratory rate, and oxygenation requirements. A higher lymphocyte count was also observed in the CAP subgroup. Although mortality did not differ significantly between the groups, CAP patients had higher APACHE II and SOFA scores on admission. Results for selected measures of particular interest are given in Table 3.5.

3.2.3 Correlation of covariates with gene expression

In order to investigate global variation in gene expression, I carried out principal component analysis on the top 10% most variably expressed transcripts (n=2716)

Covariate	CAP (n=73)	FP (n=64)	Adjusted p value
Age (years)	60.6 (15.6)	65.1 (17.5)	NS
Male sex	37 (50.6%)	28 (43.7%)	NS
APACHE II score on day 1 in ICU	18 (6.2)	15 (5.9)	0.005
SOFA score on day of sampling	6.8 (3.5)	5.4 (4.0)	0.03
14 day mortality	8 (11.0%)	6 (9.4%)	NS
28 day mortality	13 (17.8)	8 (12.5%)	NS
6 month mortality	17 (23.3%)	11 (17.2%)	NS
Mechanical Ventilation	51 (69.9%)	34 (53.1%)	0.05
Vasopressors/inotropes:			NS
- No dose	37 (50.7%)	35 (54.7%)	
- Low dose	1 (1.4%)	1 (1.6%)	
- Medium dose	10 (13.7%)	5 (7.8%)	
- High dose	24 (32.9%)	23 (35.9%)	
Mean Temperature (low)	36.4 (0.8)	36.1 (0.7)	0.018
Mean Haematocrit %	35.4 (6.9)	30 (5.9)	4.4x10 ⁻⁶
Proportion of leukocytes:			
Lymphocytes	9.17%	6.52%	0.008
Polynucleocytes	84.10%	87.98%	0.029
Mononucleocytes	6.73%	5.50%	NS
Mean Bicarbonate	25.0 (6.8)	22.1 (5.1)	0.008

Table 3.5: Comparison of clinical covariates between CAP and FP.

Demographic and clinical information are summarised for each cause of sepsis, presented as count (percentage) or mean (standard deviation). The CAP and FP subgroups were compared using the chi-square test for count data, the Student's t-test for continuous data, and the log-rank test for mortality.

across all samples. This revealed that the non-septic pre-surgical samples cluster together, and are clearly distinct from septic patients (Fig. 3.7). Within the sepsis samples, CAP and FP patients are not separated, but some association with principal component (PC) 1 is observed, with FP samples on average being more removed from the pre-surgical samples. This suggests that there could be small but detectable differences between the CAP and FP sepsis responses.

The first six PCs of the sepsis gene expression data (excluding the non-septic samples) describe >50% of the observed variance. These PCs were tested for correlation with a comprehensive set of clinical covariates, which are potential sources or outcomes of variation in gene expression (Fig. 3.8, Table A.4). For PC1, the strongest correlations were with source of infection ($r=-0.417$,

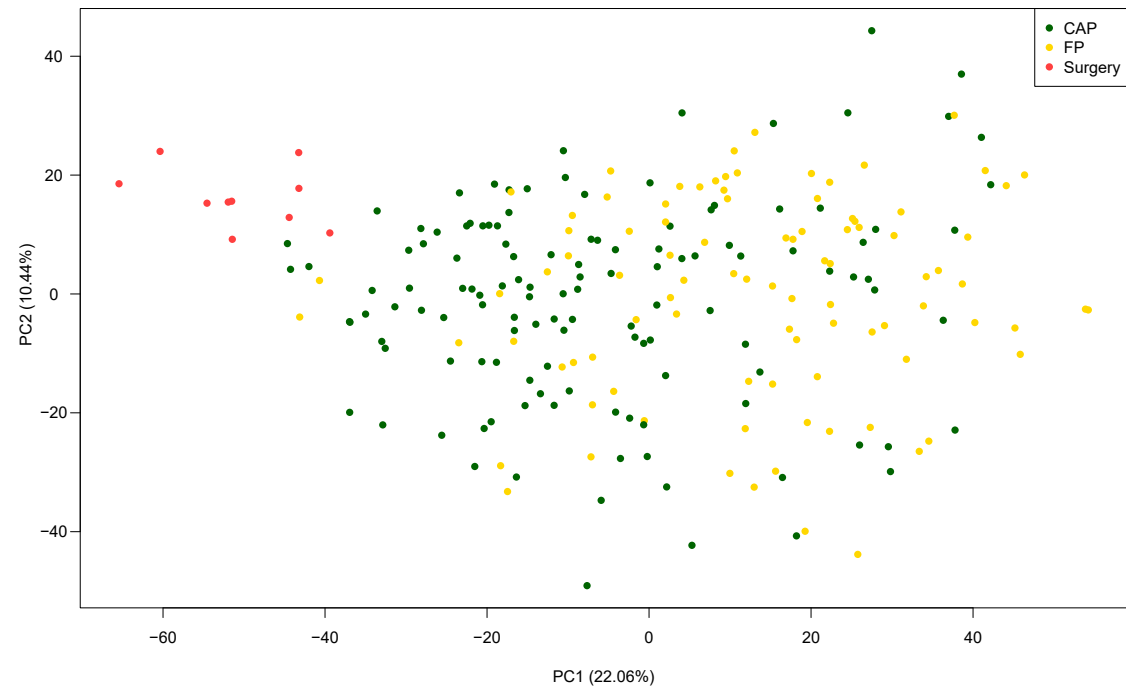


Figure 3.7: Principal Component Analysis (PCA) of the Radhakrishnan 2010 gene expression data.
The first two principal components of the top 2,716 (10%) most variable transcripts are plotted. Each point represents a sample, coloured according to the diagnosis of the patient (CAP *green*, FP *yellow*, surgery *red*). The proportion of variation explained by each principal component is given.

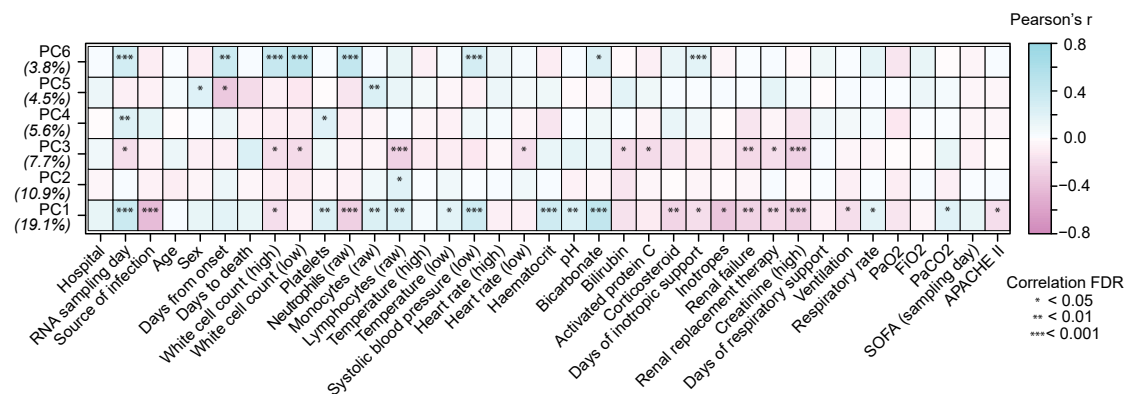


Figure 3.8: Heatmap showing correlation between principal components of gene expression data and clinical covariates.
Each of the first six principal components (50% of variance) of the Radhakrishnan 2010 sepsis gene expression data were tested for correlation with all available clinical covariates. Positive correlations are indicated in *blue* and negative correlations in *pink*, as measured by Pearson's *r*. Significant correlations are indicated by asterisks (*).

FDR=6.12x10⁻⁹), bicarbonate ($r=0.35$, FDR=2.2x10⁻⁸), and neutrophil count ($r=-0.29$, FDR=2.0x10⁻³). As noted above, this suggests that source of infection is associated with global differences in the transcriptomic response to sepsis, albeit weakly. The analysis was repeated on the CAP and FP samples separately, revealing some differences in the associations observed; for example, correlations were detected with neutrophil count, bicarbonate, inotrope use, and renal failure in CAP, and hypotension in FP (Fig. A.1).

3.2.4 Effect of cause of sepsis on gene expression

As described above, principal component analysis of sepsis patients revealed limited segregation according to source of infection. This suggests that CAP and FP do not result in completely distinct responses at the transcriptomic level. However, differences in the regulation of a small number of key genes could have an important effect on the patient response.

Genome-wide gene expression was contrasted between CAP and FP by linear regression (limma (Smyth 2004)) using the first available samples for each patient ($n_{FP}=64$, $n_{CAP}=73$; Fig. 3.2a). Genes were deemed significantly differentially expressed with a fold change ≥ 1.5 and an FDR < 0.05 . This identified 310 probes (263 genes) differentially expressed between the two diagnoses (Fig. 3.9, Table A.5). These included several genes of interest in the context of sepsis, in particular a set of interferon-induced antiviral genes (*IFIT1*, *IFIT2*, and *IFIT3*) upregulated in CAP; *BTN3A3*, an MHC class I gene; *IFIH1*, an innate immune receptor and cytoplasmic sensor of viral nucleic acids; *OAS2*, an interferon-induced antiviral enzyme; and *XAF1*, which counteracts an inhibitor of apoptosis.

Permutation analysis was used to determine if this relatively small number of differentially expressed genes was more than would be expected by chance.

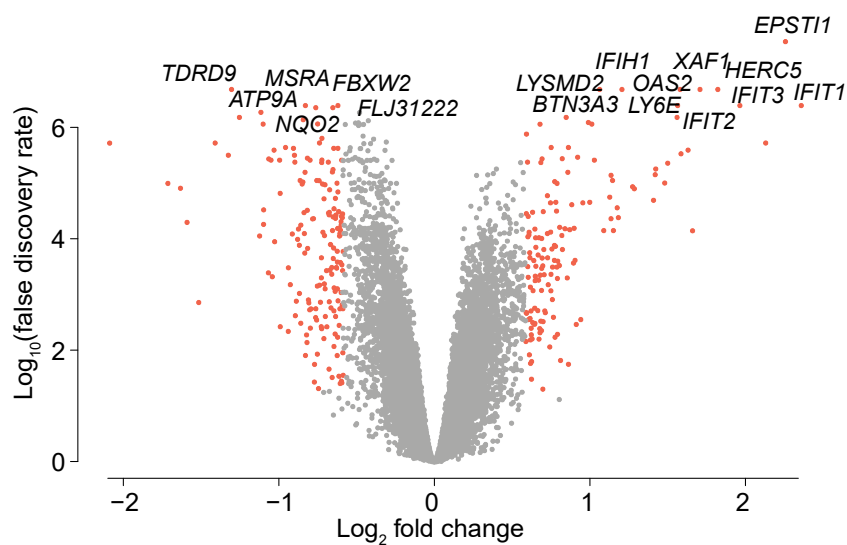


Figure 3.9: Comparison of gene expression between sources of infection.
A volcano plot showing differences in gene expression between patients with sepsis due to CAP and those with sepsis due to FP. Significance of differential expression is plotted against the log fold change, with a positive fold change indicating higher expression in CAP. Probes with FDR <0.05 and fold change ≥1.5 are plotted in *red*, and the most significant genes are labelled.

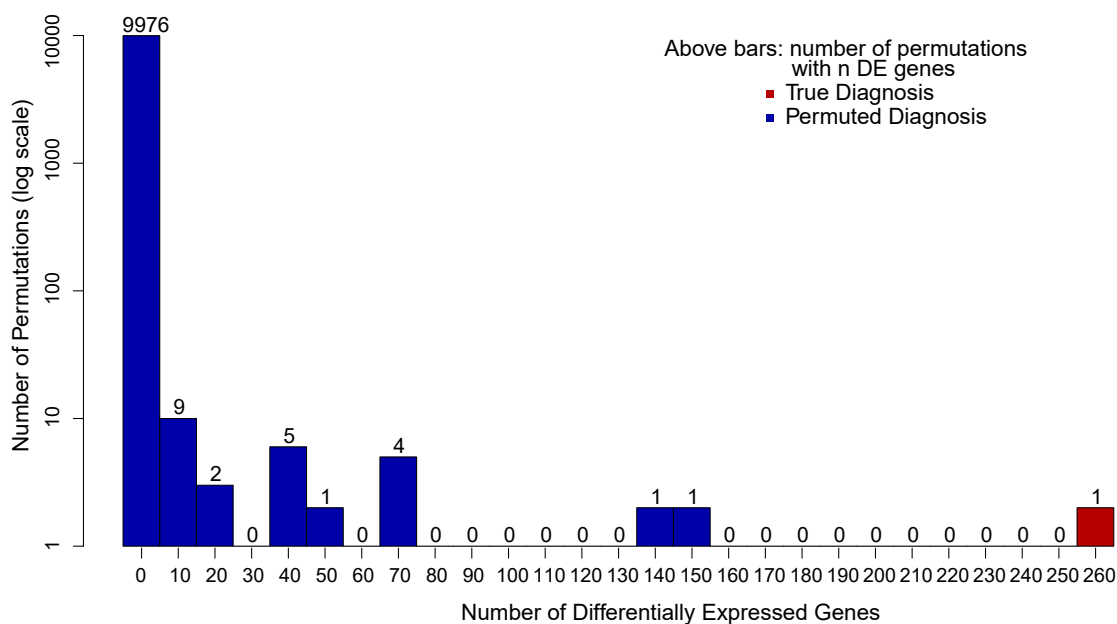


Figure 3.10: Histogram of CAP vs FP permutation results.
Diagnostic labels were permuted 9,999 times and differential gene expression analysis carried out. The number of significantly differentially expressed genes across all permutations is plotted, demonstrating that the true observation (263 genes) is significantly more than expected by chance (p <0.0001).

The diagnostic sample labels were permuted 9,999 times, and the differential expression contrast repeated. None of the permutations resulted in a greater number of differentially expressed genes than was observed for the true groups (Fig. 3.10).

3.2.5 Effect of cell count

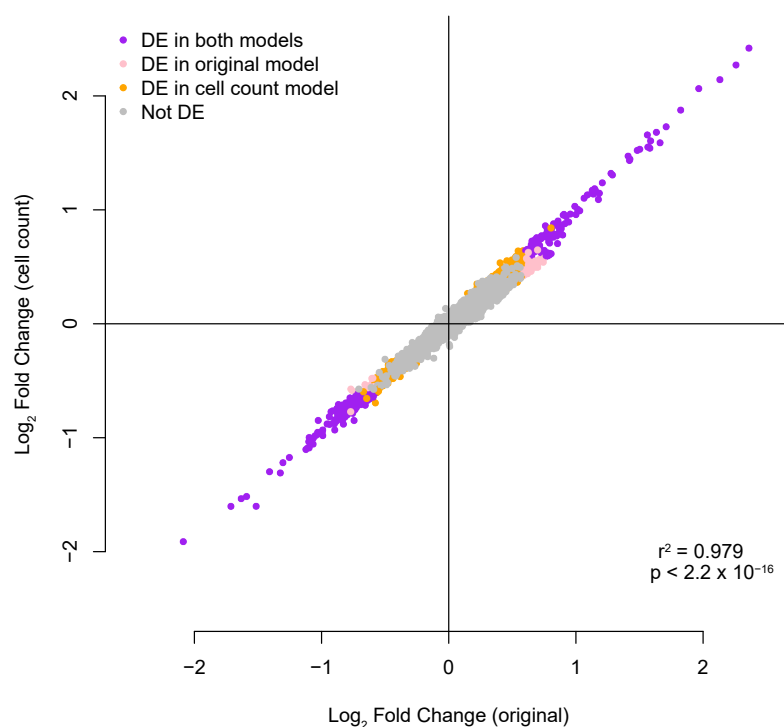


Figure 3.11: Effect of differential cell count on differential expression analysis.

The log fold change for differential expression of all genes between CAP and FP accounting for relative mononucleocyte, polynucleocyte, and lymphocyte counts is plotted against the fold changes for the same genes in the original model. Purple colouring indicates significant differential expression in both models; pink in the original model only; orange in the cell count model only; grey indicates the gene is not significantly differentially expressed between CAP and FP in either model.

In the comparison of clinical covariates, significant differences in total and differential cell counts were observed between the CAP and FP groups. Cell count could influence the detected gene expression, and thereby the differential gene expression analysis. In order to test if the analysis described above was robust to variation in cell count, it was repeated accounting for differences in relative mononucleocyte, polynucleocyte, and lymphocyte quantities in the regression

response was also identified ($p=1\times 10^{-39}$; Fig. 3.12). Upstream regulators predicted from the differentially expressed genes included TNF and LPS, which are consistent with IFN signalling and the pro-inflammatory response.

The R package XGR (Fang et al. 2016) was then used to test for enrichment of different sets of annotations, whilst correcting for the background of all genes retained after filtering in the Radhakrishnan 2010 dataset. Consistent with the analysis using IPA, significant enrichment was observed for the Gene Ontology biological processes “defence response to virus” ($\text{FDR}=5.2\times 10^{-6}$), “type I interferon signalling pathway” ($\text{FDR}=6.4\times 10^{-5}$, and cellular response to LPS” ($\text{FDR}=0.016$).

3.2.7 Viral response signatures

Given the prominence of a viral response gene signature in the contrast between CAP and FP, the impact of sepsis aetiology (bacterial vs viral) was investigated. Microbiological data is sparse in the GAINs study as a whole, with only 40% of CAP patients having a positive blood culture and thus a confirmed bacterial aetiology. Additionally, viral PCR was performed according to hospital procedures, generally only during influenza outbreaks, thus limiting virology data. In the Radhakrishnan 2010 cohort, only 6 of the 73 CAP patients had a confirmed viral infection; however, it is likely that additional patients had an undiagnosed viral infection.

The Davenport 2011 cohort was enriched for patients with an influenza infection following the H1N1 epidemic; this larger dataset was therefore used to contrast gene expression of CAP patients with and without a confirmed viral infection (25 vs 240 patients; Fig. 3.2b). This identified 39 differentially expressed genes ($\text{FDR}<0.05$, $\text{FC}\geq 1.5$; Fig. 3.13, Table A.7), which permutation analysis demonstrated was more than would be expected by chance ($p=0.0056$, Fig. A.2). Notably, this

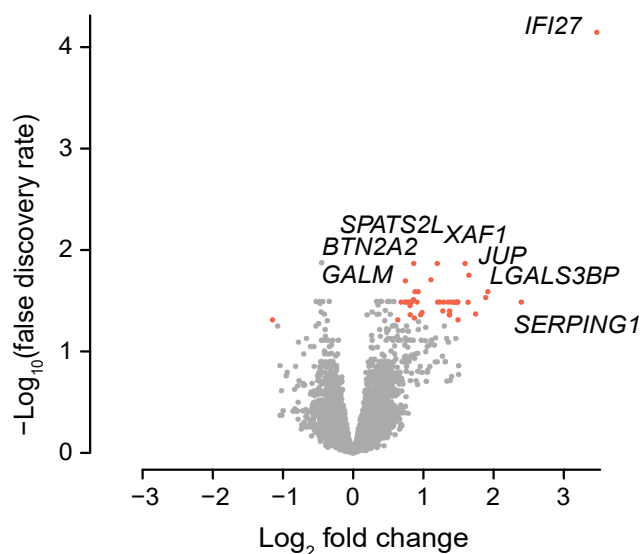


Figure 3.13: Comparison of gene expression between viral and non-viral infections.

A volcano plot showing differences in gene expression between patients with a confirmed viral infection and all other patients. Significance of differential expression is plotted against the log fold change, with a positive fold change indicating higher expression in viral infections. Probes with $FDR < 0.05$ and fold change ≥ 1.5 are plotted in red, and the most significant genes are labelled.

gene set included *LAMP3*, a dendritic cell glycoprotein induced upon influenza A infection (Zhou et al. 2011); and *IFI27*, induced by interferon α and previously described as a marker of influenza infection (Ioannidis et al. 2012). Using a gene set enrichment test (ROAST, from the limma package), this viral signature was significantly enriched in the genes differentially expressed between CAP and FP (24/39 genes differentially expressed; $p < 1 \times 10^{-4}$).

3.2.8 Validation of results

Samples were selected for the Burnham 2014 cohort to validate the differences in gene expression observed between CAP and FP. Power calculations based on the Radhakrishnan 2010 data demonstrated that a sample size of 53 in each group would give more than 95% power to detect differentially expressed genes with a fold change ≥ 1.5 (Fig. 3.2a). Source of infection, cell count, and day of sampling were again strongly associated with the first six gene expression

PCs. The differential expression analysis demonstrated high concordance of differentially expressed genes (Table A.8), with the fold changes in the 2010 data being highly correlated with the corresponding values in the 2014 cohort (Pearsons $r=0.83$, $p<2.2\times10^{-16}$). Pathway analysis with IPA and XGR also showed overlap in enriched pathways and functions, with just under half being replicated (FDR<0.05; Fig. 3.14, Table A.9).

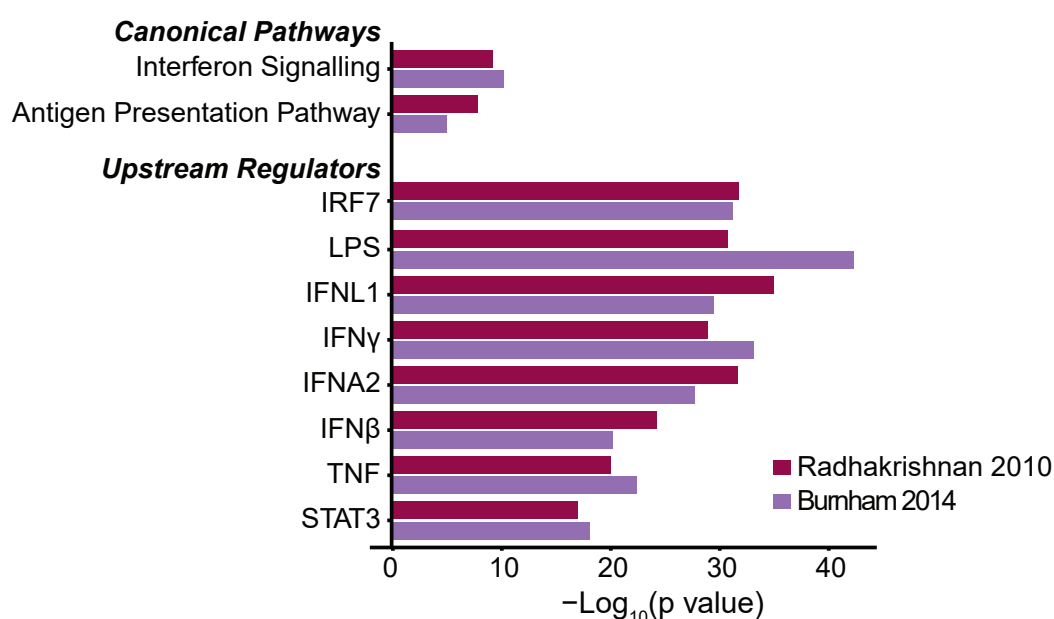


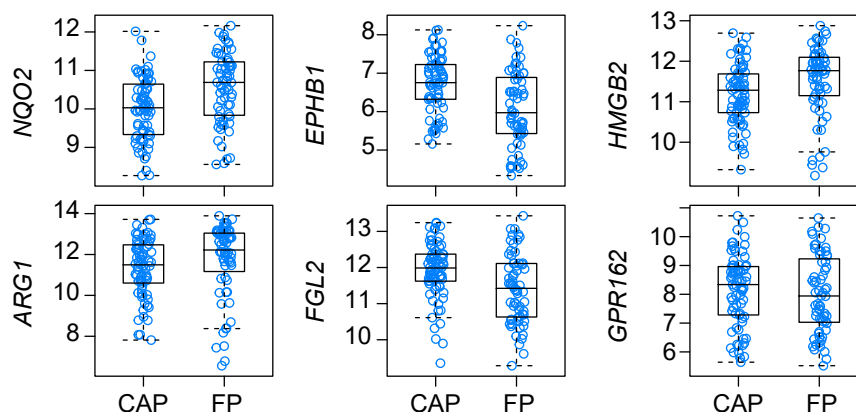
Figure 3.14: Replication of pathway enrichment analysis for source of infection. Canonical pathways and upstream regulators identified through pathway analysis of genes differentially regulated between CAP and FP are plotted by the significance of enrichment in the Radhakrishnan 2010 and Burnham 2014 datasets.

3.2.9 Diagnostic model

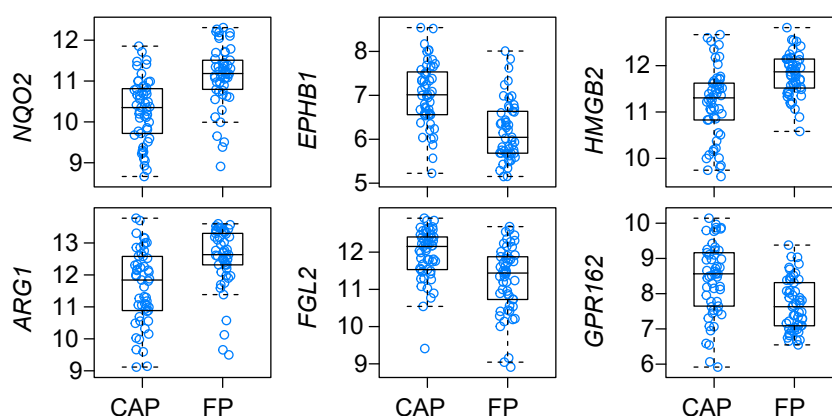
Differentiation between CAP and FP does not present a diagnostic challenge in the clinic. However, a gene expression-based model that successfully distinguishes between sepsis caused by pneumonia and peritonitis would demonstrate that the differential gene expression signatures are significant and robust. A sparse regression method (GeneRave (Kiiveri 2008)) was therefore employed to design a diagnostic model based on gene expression. Predictors were selected from the probes differentially expressed between CAP and FP with

moderate to high expression (>6.5) in at least 64 samples (the number of samples in the smaller FP subgroup), and that mapped to known genes.

This identified a gene set comprising *EPHB1*, *NQO2*, *ARG1*, *HMGB2*, *FGL2*, and *GPR162* that was predictive of source of infection (FP or CAP) for the Radhakrishnan 2010 cohort with reasonably high accuracy, demonstrating the magnitude and reproducibility of the differences observed between CAP and FP (Fig. 3.15a). The model was tested on the discovery cohort by leave-one-out validation (misclassification rate 29.9%, AUC 0.76; Fig. 3.16). In the Burnham 2014 validation cohort, the test error was 28.3% (AUC 0.80) (Fig. 3.15b).



(a) Radhakrishnan 2010 (discovery) cohort



(b) Burnham 2014 (validation) cohort

Figure 3.15: Boxplots of diagnostic model genes.

The $\log_2$ expression of six genes, *EPHB1*, *NQO2*, *ARG1*, *HMGB2*, *FGL2*, and *GPR162*, is plotted for CAP and FP patients in (a) the Radhakrishnan 2010 cohort, and (b) the Burnham 2014 cohort.

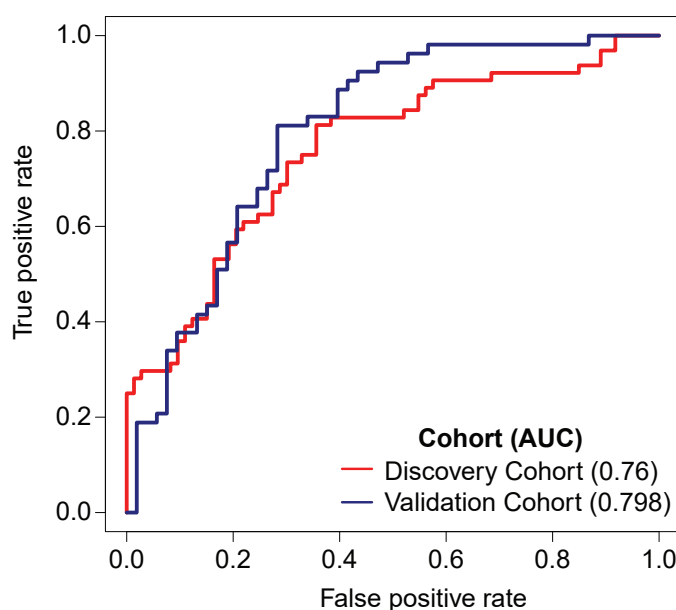


Figure 3.16: ROC curves for gene expression-based diagnostic models.

A model was designed to identify sepsis cause based on the expression of six genes (*EPHB1*, *NQO2*, *ARG1*, *HMGB2*, *FGL2*, and *GPR162*). The performance of the model was tested by leave-one-out validation in the discovery cohort and on the entire validation cohort. The ROC curve is plotted for each and the area under the curve (AUC) given.

Several of the genes selected for the diagnostic model are of biological interest. Firstly, *EPHB1* encodes the ephrin receptor B1, and is upregulated in CAP relative to FP. It plays a general role in targeted cell migration and adhesion, contributing to the development of immune networks including T cell development and migration. It has been shown to be essential in T-cell interactions during their activation, and has been implicated in inflammatory pain; for example, it is associated with the enhancement of lymphocyte migration and cytokine production in rheumatoid arthritis (Kitamura et al. 2008).

ARG1, the gene for arginase 1, has higher expression in FP than CAP. Arginase 1 is a cytosolic protein that causes depletion of arginine, a metabolite essential for normal T cell function. It has been previously implicated in sepsis, with reports that *ARG1* deficiency blocks endotoxaemia and that septic myeloid-derived suppressor cells (MDSCs) have increased *ARG1* expression. MDSCs are induced upon physical injury or surgery, leading to decreased intracellular arginine in T cells. This inhibits T cell proliferation, memory, and cytotoxicity, and could

lead to increased susceptibility to infection following surgery or traumatic injury. Given that all FP patients have undergone surgery, and the illness of some results from traumatic injury, this is a particularly interesting association (Zhu et al. 2014; Mathias et al. 2017).

HMGB2 is a member of the non-histone chromosomal high mobility group protein family. These proteins generally facilitate cooperative interactions between proteins by promoting DNA flexibility. This particular gene has a known role in inflammation, acting as an extracellular damage-associated molecular pattern (DAMP) secreted by activated cells or passively released by damaged cells, leading to the induction of inflammatory cytokines through TLR4 signalling (Yang et al. 2016; Zhou et al. 2016). As such, it is a critical mediator of lethal infection and injury, with a central role in the host inflammatory response to infection or sterile insults. It is downregulated in CAP vs FP.

Finally, *FGL2* is a fibrinogen-like secreted protein upregulated in CAP relative to FP. It is expressed in endothelial and T cells, and is induced by viral proteins and cytokines responding to viral inflammation. It has an immunosuppressive function in infection; as an effector of regulatory T cells, it suppresses T cell proliferation, dendritic cell maturation, and Th1 cytokine production. It can also enhance Th2 cytokines, and downregulate antigen presentation. *FGL2* has been reported to induce activated lung-infiltrating macrophages in COPD, and its activity in macrophages has been associated with outcome and disease severity in viral hepatitis. It is expressed at high levels in the lungs of SARS patients, reflecting its physiological role in immune coagulation (Hu et al. 2016).

3.2.10 Time from onset of disease

Time from onset of acute illness is likely to be an important influence on variation in gene expression, both between and within patients. However, this

is not straightforward to investigate. It is not expected that the first day in ICU for one patient corresponds well with day one for another patient, even with the same source of sepsis, due to the combined effect of host, microbial, and environmental factors that influence the host response over time. It is therefore more appropriate to compare serial samples taken over the course of ICU admission from the same patient to explore variation in gene expression within individuals over time. This approach also allows the investigation of differences in expression dynamics between causes of sepsis; given the difference in disease course between pneumonia and peritonitis, expression dynamics might be specific to a diagnostic subgroup.

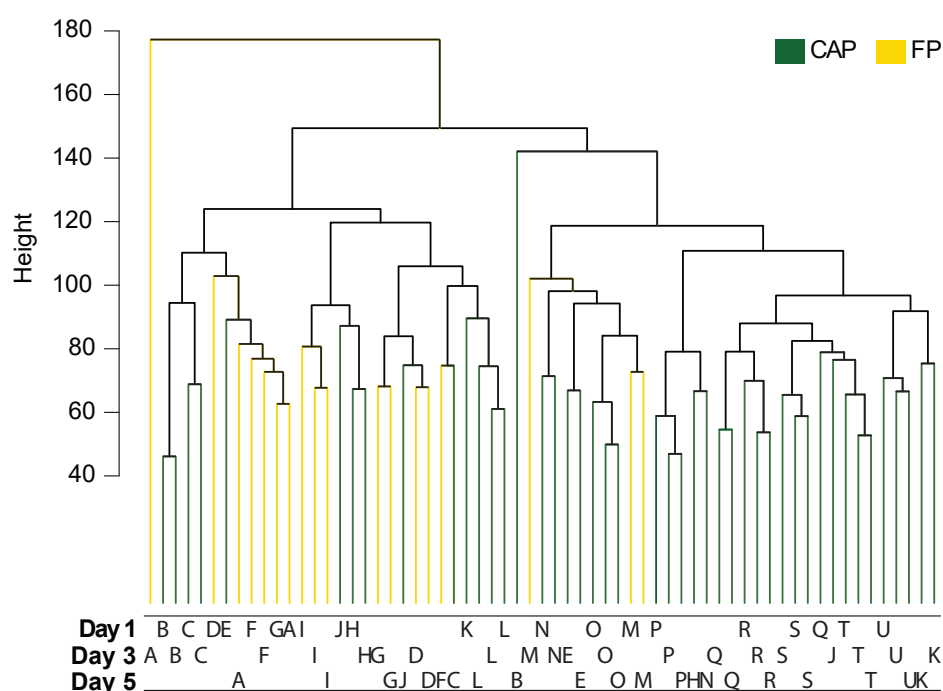


Figure 3.17: Dendrogram of serial samples for patients with FP and CAP.

Hierarchical clustering demonstrates the close relationship between samples from the same patient (each letter represents a different patient), and the same diagnostic subgroups (green indicates CAP, yellow FP). Some clustering of repeated samples from the same patient is observed, e.g. patients “I” and “O”; other patients have samples dispersed widely across the dendrogram (patients “K” and “F”).

The Radhakrishnan 2010 dataset profiles 221 samples from 137 patients; 63 patients have at least two samples. There was no difference in the proportion of samples taken on day 1, 3, and 5 between FP and CAP samples (Fig. A.3,

chi-square $P=0.350$). Hierarchical clustering was carried out on samples from patients with all three time points analysed, and did not show grouping of samples by day of RNA sampling (Fig. 3.17). Some clustering of repeated samples from the same patient was observed, whilst other patients had samples dispersed widely across the dendrogram, suggesting they had global changes in gene expression over time. For 48% of the samples from a serially sampled individual ($n=147$), the most similar sample identified by Pearson correlation was another sample from the same patient.

3.2.11 Gene expression dynamics within patients

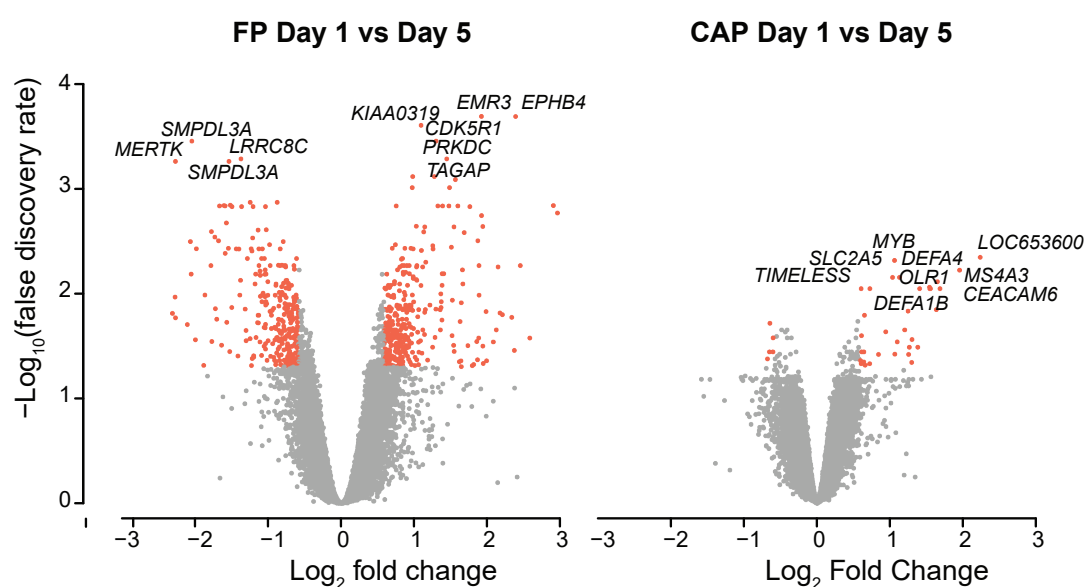


Figure 3.18: Differential expression within patients between day 1 and day 5.

A volcano plot showing differences in gene expression within patients with sepsis due to CAP and those with sepsis due to FP. Significance of differential expression is plotted against the log fold change, with a positive fold change indicating higher expression in day 5. Probes with $FDR < 0.05$ and fold change ≥ 1.5 are plotted in red, and the most significant genes are labelled.

A repeated measures regression model was fitted using day 1 and 5 samples from the same patients to identify genes changing expression over time (Fig. 3.2c). In FP ($n=9$), 714 genes were significantly differentially expressed within patients over time, compared to only 80 genes in CAP ($n=17$) (Fig. 3.18, Table A.10).

The R package timecourse was used to further explore the hypothesis that temporal changes in gene expression could depend on the source of infection. This analysis was restricted to patients with samples available at all three time points ($n_{CAP}=15$, $n_{FP}=6$). Again, temporal changes in gene expression were found to be more pronounced in FP patients, with the specific genes highly prioritised by this approach also differing between CAP and FP (Table A.11). For example, genes showing large temporal changes in FP included the kinase activator *CDK5R1* and the complement component *C1QB*, whilst in CAP, the cell cycle gene *NIT2* and the haematopoiesis regulatory transcription factor *MYB* were highly prioritised (Fig. 3.19).

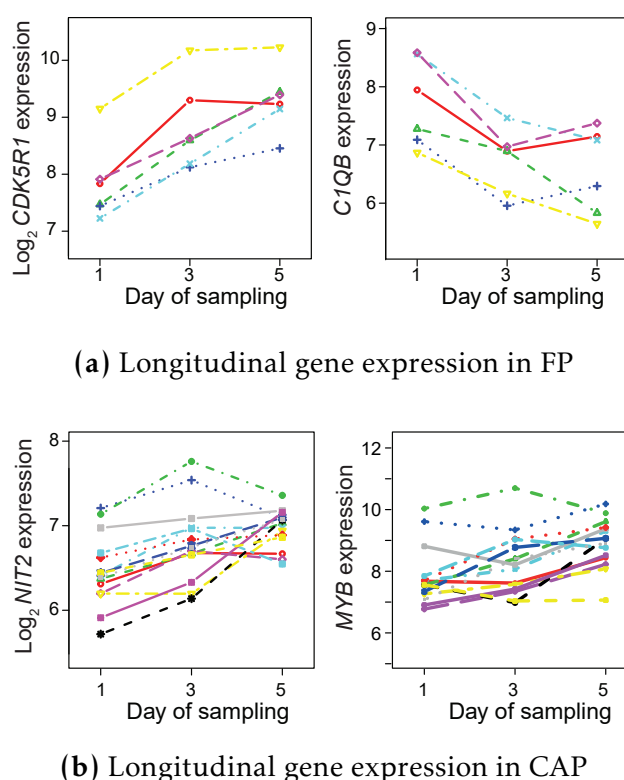


Figure 3.19: Gene expression within patients over time.

Gene expression is plotted over time in (a) FP patients and (b) CAP patients, for selected genes showing large changes over time. Each point represents a sample, with serial samples from the same patient connected by colour-coded lines.

3.2.12 Changes from time of onset in FP

For the FP patients, the time of the acute event had been recorded, and so it was possible to calculate the time since onset of disease with a reasonable degree of accuracy. Time course analysis was carried out in limma based on number of days from disease onset for the 2010 cohort. This identified 140 genes showing significant changes in expression over time (FDR <0.05, Table A.12). These included *AAK1*, which is involved in the regulation of endocytosis and thereby implicated in bacterial invasion, and *SNN*, a marker of mitochondrial dysfunction, which has been associated with multi-organ failure (Fig. 3.20).

Pathway analysis showed that this set of genes was significantly enriched for phagosome formation, NK cell signalling, and IL-3 signalling, as well as the biological functions cellular degranulation, chemotaxis, leukocyte activation, macrophage adhesion, phagocytosis, and bacterial infection (Table A.13).

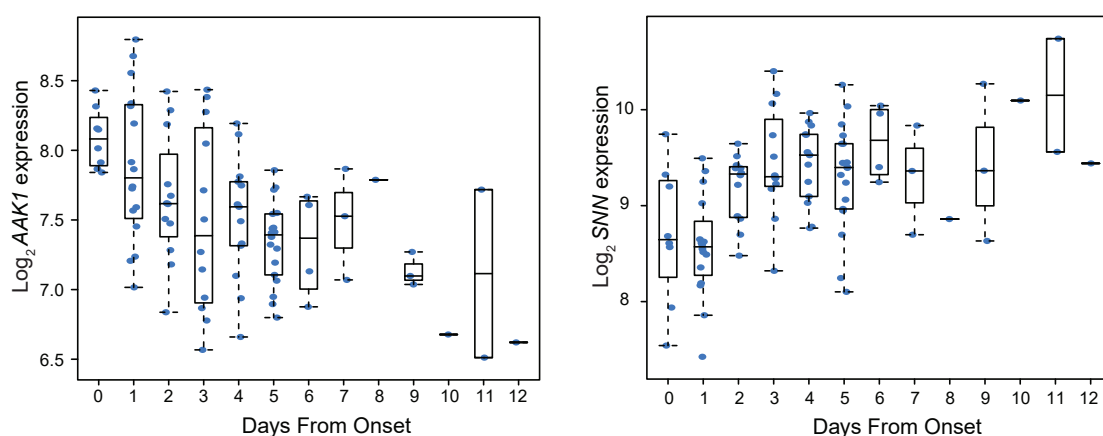


Figure 3.20: Examples of genes changing expression over time from disease onset. Expression of *AAK1* and *SNN* is plotted for each FP sample in the Radhakrishnan 2010 cohort against the number of days since disease onset.

3.2.13 Extension of gene expression dynamics

These results were validated in the Burnham 2016 dataset, which was designed to further explore longitudinal gene expression in CAP and FP, and comprised

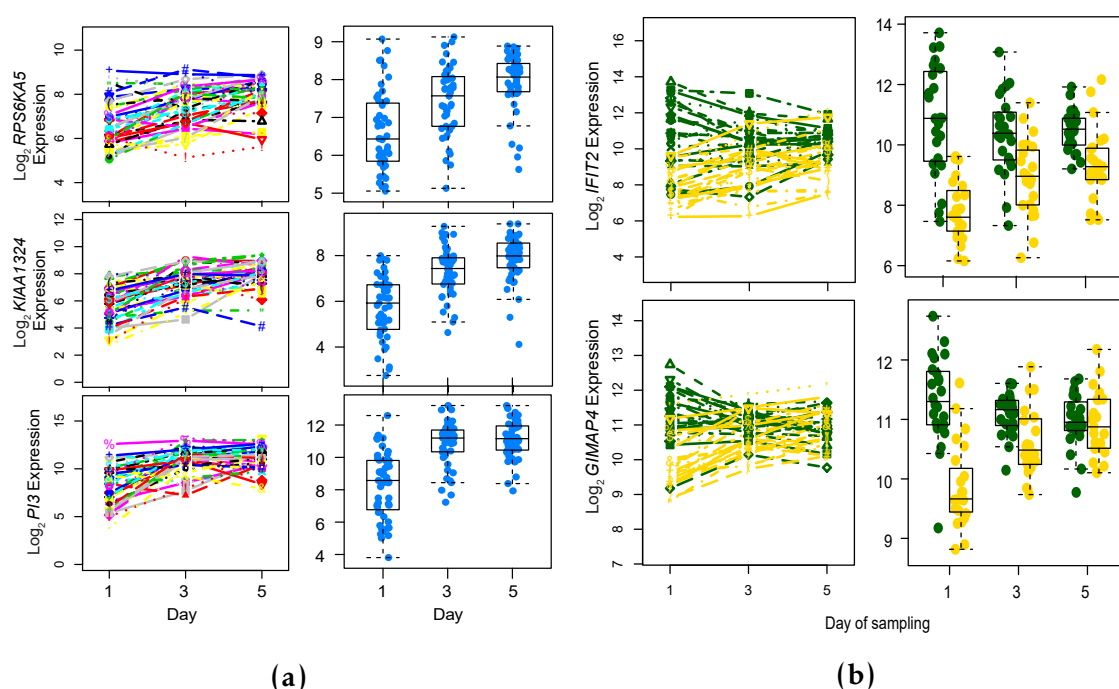


Figure 3.21: Sepsis time course profiles: example genes.

Gene expression is plotted over time in 47 sepsis patients for (a) selected genes showing large changes across the first five days in ICU and (b) with different time course profiles in CAP *green* and FP *yellow*. Each point represents a sample, with serial samples from the same patient connected by colour-coded lines.

samples from all three time points for 47 patients (24 CAP and 23 FP; Fig. 3.2c). The timecourse package was again used to identify genes with changing expression within patients over time.

Initially, longitudinal changes were investigated across all samples, meaning that genes with changes in expression over time but with consistent time course profiles across CAP and FP were prioritised (Table A.14). Highly prioritised genes included *KIAA1324* and *RPS6KA4*, which are involved in the cellular response to stress and in regulation of inflammatory genes, and *PI3*, which encodes an antimicrobial peptide (Fig. 3.21a).

Subsequently, as in the Radhakrishnan 2010 dataset, genes showing variation in their time course according to diagnosis were identified (Fig. 3.21b). These included *IFIT2*, an interferon-induced anti-viral protein, and *GIMAP4*, which is an immune-associated GTPase thought to regulate lymphocyte apoptosis. The

top 1% of genes that were prioritised were tested for enriched biological processes using XGR, with all genes detected in the Burnham 2016 dataset used as the background. This highlighted the defence response to viruses ($\text{FDR}=2.8\times 10^{-5}$), and the type 1 interferon signalling pathway ($\text{FDR}=1.4\times 10^{-4}$).

3.3 Results: cardiac surgery

3.3.1 Description of cardiac surgery cohort

As mentioned above, the Radhakrishnan 2010 and Davenport 2011 datasets were processed in parallel with paired pre- and post-operative samples taken from patients undergoing elective cardiac surgery. These samples enable the direct comparison of samples from the same patient to define the transcriptomic response to sterile inflammation caused by surgery. Cardiac and septic samples were randomised, with each cardiac pair of samples run on the same chip to avoid batch effects.

The 2010 cardiac surgery cohort (Svoren 2010) consisted of samples from 11 patients, whereas the 2011 cohort repeated these samples together with samples from subsequently recruited patients to give a total of 126 samples from 63 patients. Following quality control, samples from 10 and 56 patients were available in the Svoren 2010 and Svoren 2011 cohorts respectively. The cardiac surgery cohort was overall 73.9% male, and had a mean age of 66.4. Information on post-operative complications was only available for the first 16 patients recruited, of whom nine had no complications, four had atrial fibrillation, two had bibasal collapse, and one had a wound infection.

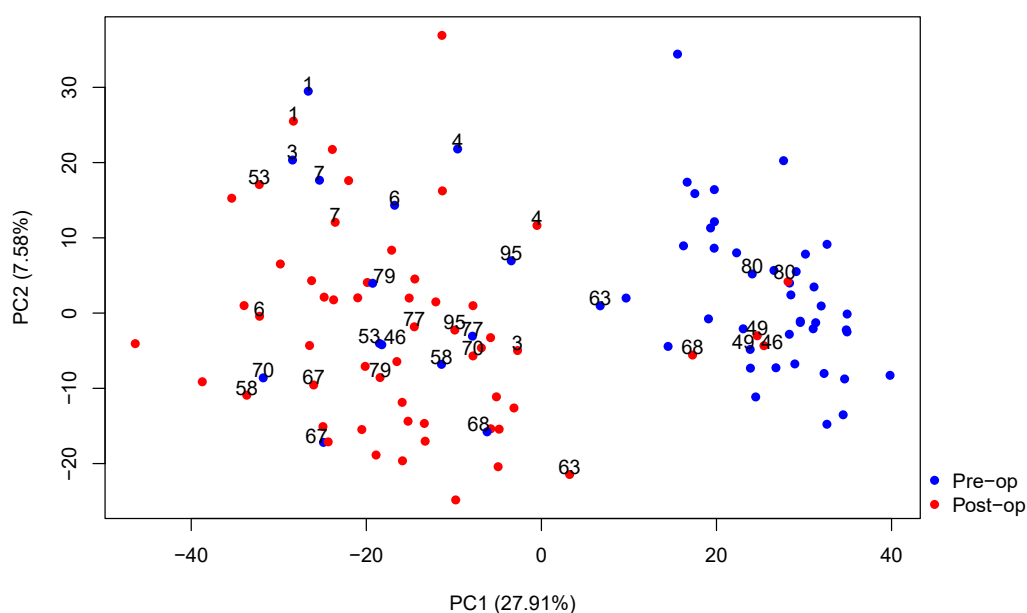


Figure 3.22: Principal component analysis in the Svoren 2011 cardiac surgery cohort. The first two principal components of the gene expression data are plotted and the proportion of variance explained by each principal component given. Each point represents a sample, coloured according to the operative status. Patients whose samples do not clearly fall into the clusters of pre- and post-operative samples are labelled by patient number.

3.3.2 Global variation in gene expression in surgical samples

The larger Svoren 2011 cohort is used to investigate changes in gene expression following cardiac surgery. Principal component analysis of these gene expression data demonstrates that the majority of samples cluster according to time point, *i.e.* pre-operative and post-operative (Fig. 3.22). However, for some patients the two samples either cluster together (*e.g.* Patients 1, 49, or the pre- and post-operative samples are in opposite groups to expected *e.g.* 46). It is possible that there were some sample mix-ups; however, the relatively large number of these instances suggests that some patients have a weaker response to surgery than others, and so do not exhibit such pronounced changes in gene expression. As only limited clinical information is available, it is not possible to investigate if this is related to outcomes.

3.3.3 Differential gene expression following surgery

Genome-wide gene expression was contrasted between the pre-operative and post-operative samples from each patient by linear regression in limma (Fig. 3.2d, 3.23, Table A.15). The 1472 genes that were significantly differentially regulated following surgery included inflammation-related genes such as *GFI1*, a transcriptional repressor that plays a role in haematopoiesis and is down-regulated in post-operative samples (Shi et al. 2013); *IFI30*, which is constitutively expressed in antigen-presenting cells, is induced by IFN γ , and plays a role in MHC class II-restricted antigen processing (Rausch and Hastings 2015); and *MARCO*, part of the innate antimicrobial immune system and upregulated following surgery (Fig. 3.24; Braun et al. (2011)).

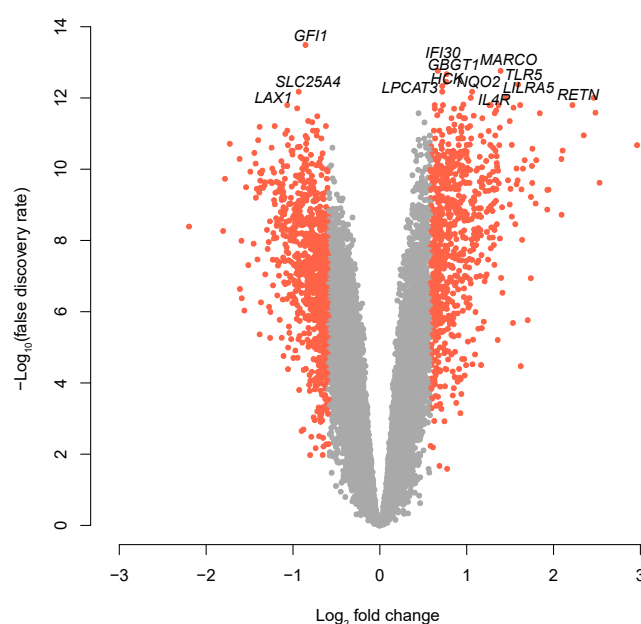


Figure 3.23: Comparison of gene expression in pre- and post-operative samples.

A volcano plot showing differences in gene expression within patients following elective cardiac surgery. Significance of differential expression is plotted against the log fold change, with a positive fold change indicating higher expression following surgery. Transcripts with FDR < 0.05 and fold change ≥ 1.5 are plotted in *red*, and the most significant genes are labelled.

Clustering all samples based on the 329 transcripts with a mean fold change of 2 or more between conditions reveals a similar pattern to the PCA; the majority

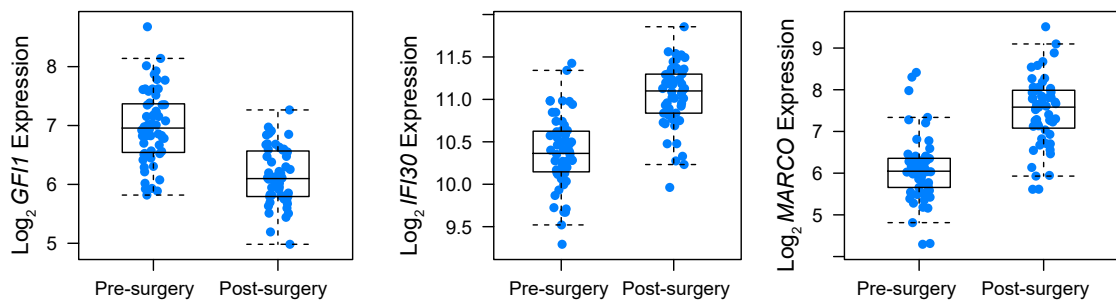


Figure 3.24: Boxplots of exemplar genes differentially expressed following cardiac surgery.

The log expression of three genes, *GFI1*, *IFI30*, and *MARCO*, is plotted for pre- and post-operative samples. These genes were all significantly differentially expressed and have roles in inflammation and immunity.

form groups of pre- and post-operative samples, but a small number of samples exhibit patterns of gene expression that are inconsistent with this (Fig. 3.25). Again, this includes samples from *e.g.* patient 49, which cluster together with the pre-operative samples, and patient 46, whose pre-operative sample clusters with post-operative samples and *vice versa*.

3.3.4 Gene ontology analysis

The genes significantly differentially expressed following surgery were tested for enriched GO Biological Processes using XGR, with all expressed genes used as the background. This identified significant enrichment of “apoptotic cell clearance” (FDR 0.0074) and “defence response to virus” (FDR 0.023). Enriched KEGG pathways included T cell differentiation (FDR=8.5x10⁻⁴), “cytokine-cytokine receptor interaction” (FDR 2.4x10⁻⁷), and “T cell receptor signalling pathway” (FDR 3.3x10⁻³). These demonstrate the effect of surgery on the immune system and the inflammatory response. Patients were not scheduled for surgery and so would not have been recruited if they had an infection, so the enrichment of the viral response could be a result of the more general inflammatory response.

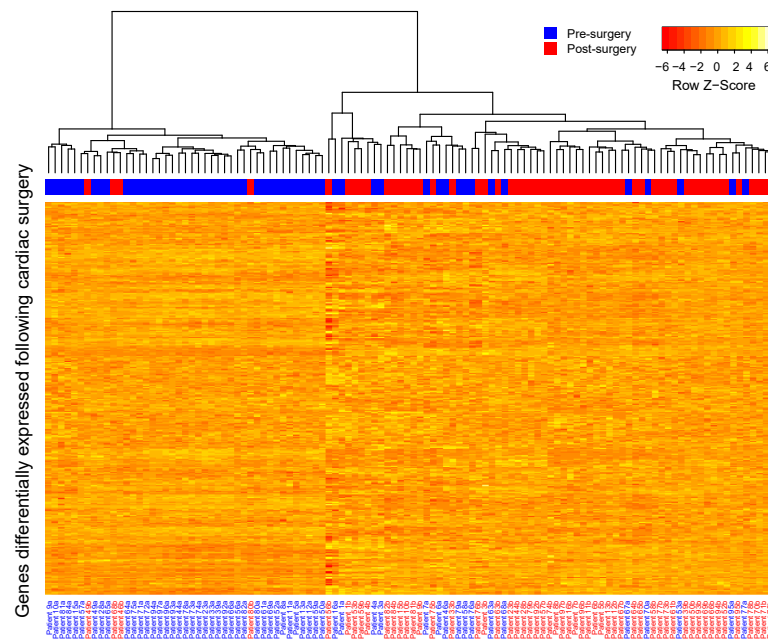


Figure 3.25: Heatmap of genes differentially expressed following cardiac surgery.

The 329 transcripts significantly differentially expressed following surgery with a fold change ≥ 2 are plotted (rows) for all samples (columns). Relative expression levels across each gene are indicated by cell colour (low relative expression *red*, high *white*). Samples with similar gene expression profiles are clustered together, and status (pre- or post-surgical) is indicated by the coloured bar at the top of the heatmap.

3.3.5 Global surgery response

Partial least squares (PLS) regression was used to assess the global difference in gene expression induced by surgery, with the sample state (*i.e.* pre- or post-surgery) as the response variable and all genes as explanatory variables. The components generated by this approach are similar to those calculated through PCA, but PLS takes the response variable into account by giving increased weight to correlated explanatory variables (Humburg et al. 2016).

This supervised approach demonstrates that there is a large change in overall gene expression following surgery. The first two components together account for 66.5% of the variation observed due to surgery, and so separate the majority of samples; however, the residual variance suggests that surgery does not have an entirely consistent effect on global gene expression (Fig. 3.26).

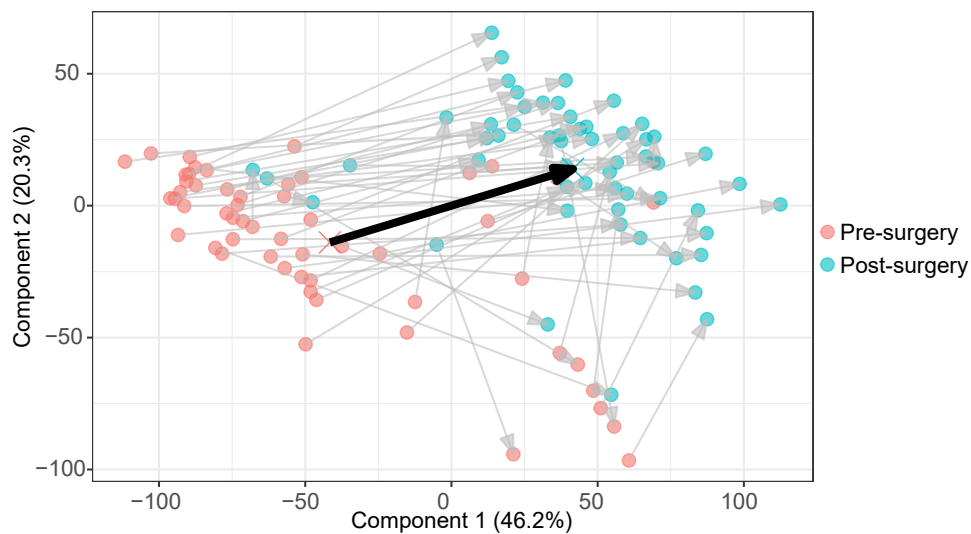


Figure 3.26: Global gene expression response to surgery.

Partial least squares regression reduces the transcriptomic response to surgery to two components, which account for 66.5% of the variance observed. These separate the majority of samples into distinct pre- and post-surgery clusters.

3.4 Results: comparison across SIRS sources

3.4.1 The sepsis response in patients with FP and CAP

The development of sepsis is not predictable, hence it is not possible to obtain pre-disease samples for comparison in a targeted manner. An alternative source of baseline samples must therefore be used to investigate transcriptomic changes induced by the sepsis response. Patients scheduled for elective cardiac surgery represent a population of patients with similar demographic and clinical features to patients who go on to develop sepsis, with the advantage that a non-septic, non-infectious pre-operative sample is available.

Sepsis patients in the Radhakrishnan 2010 cohort (FP $n=64$, CAP $n=73$) were compared with non-septic pre-surgical samples ($n=10$) assayed at the same time (Fig. 3.2e). Differential gene expression analysis demonstrated that sepsis was associated with differential regulation (≥ 1.5 FC, FDR < 0.05) of 4881 transcripts (18.0% of transcripts measured) (Fig. 3.27, Table A.16). When CAP and FP

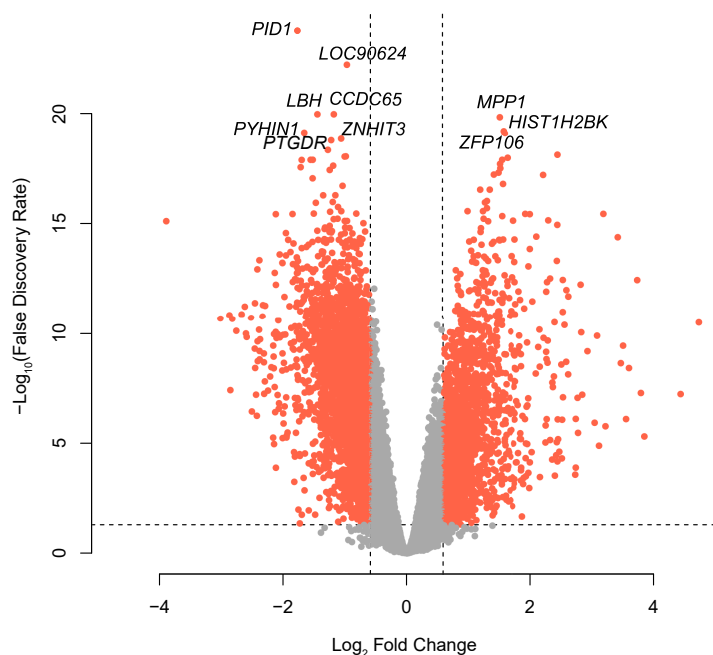
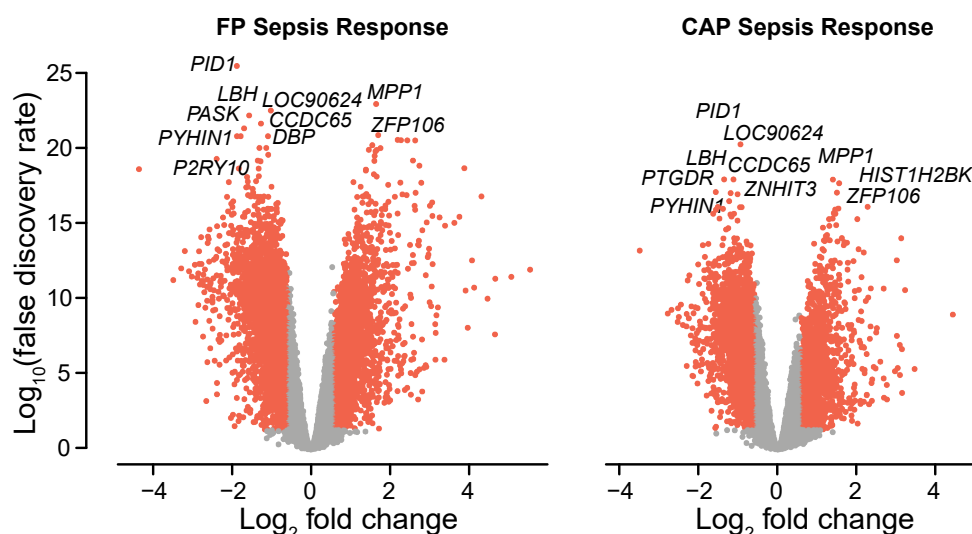


Figure 3.27: Comparison of septic and non-septic samples.

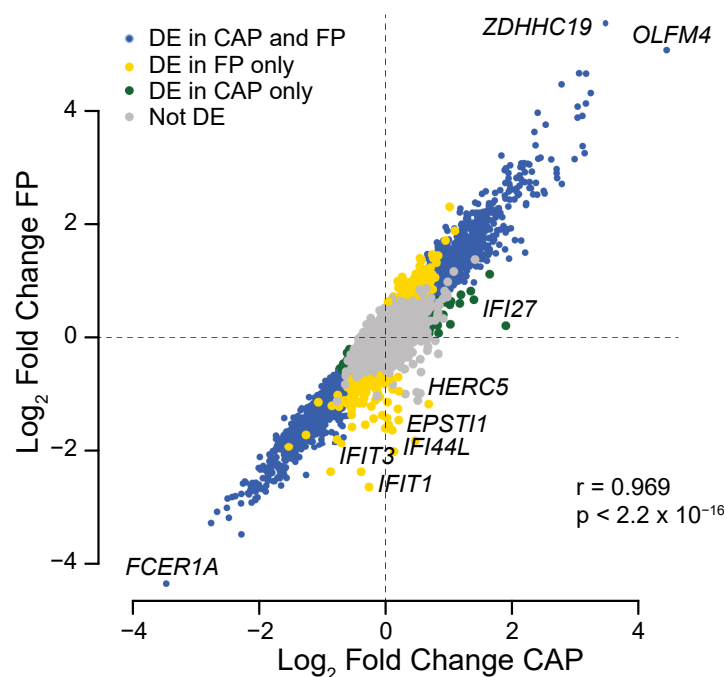
A volcano plot showing differences in gene expression between sepsis patients (CAP and FP) and non-septic patients scheduled for cardiac surgery. Significance of differential expression is plotted against the log fold change, with a positive fold change indicating higher expression in sepsis. Transcripts with $FDR < 0.05$ and fold change ≥ 1.5 are plotted in *red*, and the most significant genes are labelled.

samples were compared to the non-septic samples separately, the fold changes of each gene were highly correlated between the two analyses. This suggests that the majority of the sepsis response is independent of source of sepsis (Fig. 3.28).

Pathway analysis with IPA showed most significant enrichment for infectious disease (severe acute respiratory syndrome [$FDR=1.0 \times 10^{-30}$] and viral infection [$FDR=2.4 \times 10^{-25}$]) and cell death ($FDR=1.4 \times 10^{-9}$). EIF2 signalling was the top canonical pathway ($FDR=1.3 \times 10^{-16}$), which is consistent with the role EIF2 plays as a translational initiation factor in response to stress (viral and bacterial infection) (Table A.17). Similarly, enrichment analysis using XGR and all genes tested for differential expression as the background identified significant enrichment of GO biological processes relating to translational initiation, elongation, and termination ($FDR=2.1 \times 10^{-7}$), and the viral life cycle ($FDR=1.1 \times 10^{-7}$); enriched KEGG pathways identified included a large number of



(a) Comparison of septic and non-septic samples: CAP and FP.



(b) Correlation plot comparing the sepsis response in CAP and FP.

Figure 3.28: The sepsis response in CAP and FP.

(a) Volcano plots showing differences in gene expression between sepsis patients (divided into CAP and FP) and non-septic patients scheduled for cardiac surgery. Significance of differential expression is plotted against the log fold change, with a positive fold change indicating higher expression in sepsis. Transcripts with $FDR < 0.05$ and fold change ≥ 1.5 are plotted in red, and the most significant genes are labelled. (b) The log fold change for differential expression of all transcripts in the FP-non-septic contrast is plotted against the log fold changes for the same genes in the CAP-non-septic contrast. Blue colouring indicates significant differential expression in both contrasts; green in CAP only; yellow in FP only; grey indicates the gene is not significantly differentially expressed in sepsis in either subgroup.

immune-related terms, including T cell differentiation ($\text{FDR}=9.8\times 10^{-7}$), antigen processing and presentation ($\text{FDR}=9.2\times 10^{-5}$).

3.4.2 Comparison of the sepsis and cardiac surgery responses

Analysis of the sepsis and surgery responses as described above identify a number of common themes in the gene expression patterns. This is observed in particular at the pathway level, with both contrasts involving enrichment of T cell signalling and differentiation amongst other features. However, a more direct comparison of the two SIRS responses is possible with the data described in this chapter. The combined Davenport-Svoren 2011 gene expression cohort firstly allows the comparison of larger numbers of CAP sepsis samples to pre-surgical samples (265 vs 56), thus validating the sepsis response results. In addition, this sepsis response may be compared to the surgical response defined using the same cohort and therefore the same baseline samples (Fig. 3.2f).

Initially, the CAP sepsis response in the Davenport 2011 cohort was found to confirm the results from the Radhakrishnan 2010 cohort (Fig. 3.29), with a strong correlation in the fold changes of differentially expressed genes ($r=0.895$). Subsequently, the sepsis and surgical responses were compared directly. Of the 1472 genes involved in the cardiac response, 1176 were also differentially expressed in the CAP sepsis response. These shared genes were enriched for features relating to the inflammatory response, such as T cell differentiation, cytokine and T cell receptor signalling, and NK cell cytotoxicity. A further 2503 genes were significantly differentially expressed in the sepsis response. However, this is likely to be affected by the differences in power of the two analyses, as highlighted in the correlation plot (Fig. 3.30).

Comparison of the CAP and post-surgery samples identifies 1213 differentially expressed genes (Fig. 3.31, Table A.18), enriched for pathway annotations such as

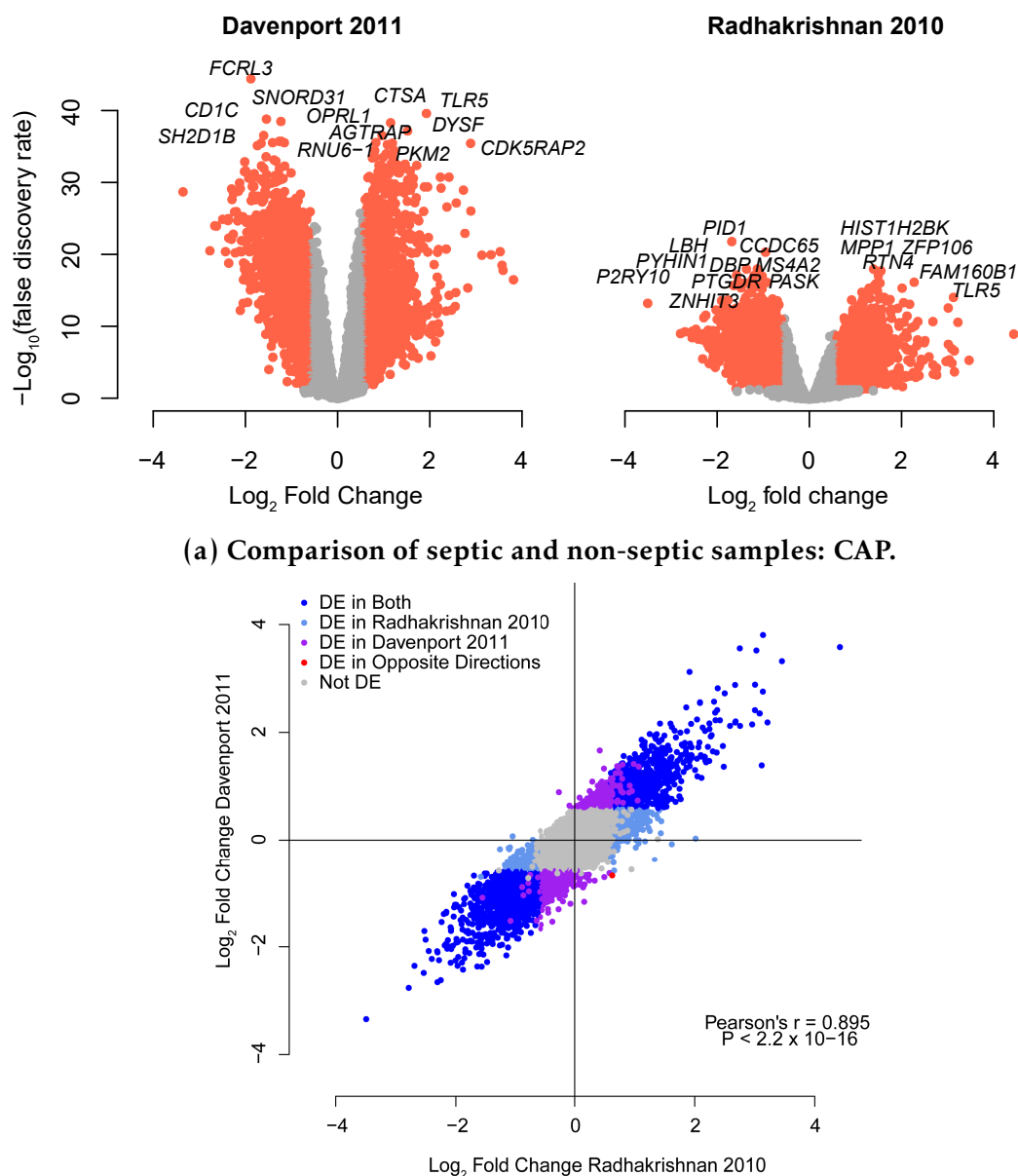
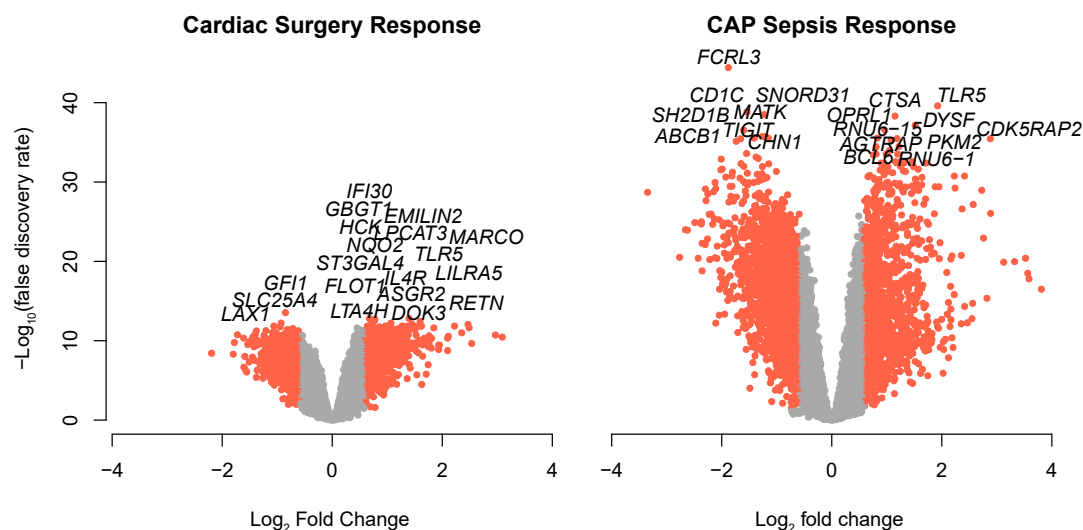
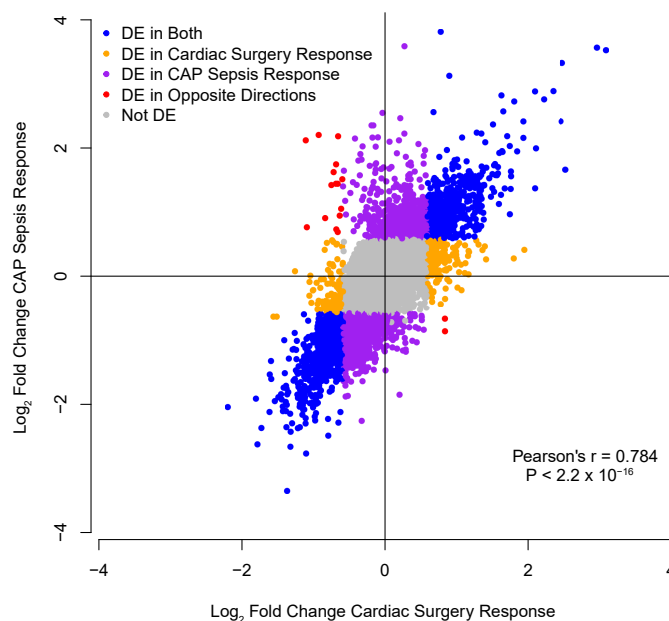


Figure 3.29: The CAP sepsis response in the Radhakrishnan 2010 and Davenport 2011 cohorts.

(a) Volcano plots showing differences in gene expression between CAP sepsis patients and non-septic patients scheduled for cardiac surgery in two cohorts. Significance of differential expression is plotted against the log fold change, with a positive fold change indicating higher expression in sepsis. Transcripts with FDR < 0.05 and fold change ≥ 1.5 are plotted in red, and the most significant genes are labelled. (b) The log fold change for differential expression of all transcripts in the CAP-non-septic contrast is compared across the two cohorts. Blue colouring indicates significant differential expression in both contrasts; light blue in the Radhakrishnan 2010 cohort only; purple in the Davenport 2011 cohort only; grey indicates the gene is not significantly differentially expressed in sepsis in either subgroup.



(a) Comparison of the sepsis and surgical responses.



(b) Correlation plot comparing the sepsis and surgical responses.

Figure 3.30: Comparison of the septic and surgery responses.

(a) Volcano plots showing differences in gene expression between CAP sepsis patients and non-septic patients, and pre- and post-surgical samples. Significance of differential expression is plotted against the log fold change, with a positive fold change indicating higher expression in sepsis or post-surgery. Transcripts with $FDR < 0.05$ and fold change ≥ 1.5 are plotted in *red*, and the most significant genes are labelled. (b) The log fold change for differential expression of all transcripts is compared across the sepsis and surgical responses. *Blue* colouring indicates significant differential expression in both contrasts; *orange* in the surgery response only; *purple* in the sepsis response only; *grey* indicates the gene is not significantly differentially expressed in sepsis in either subgroup; and *red* that they were differentially expressed in opposite directions.

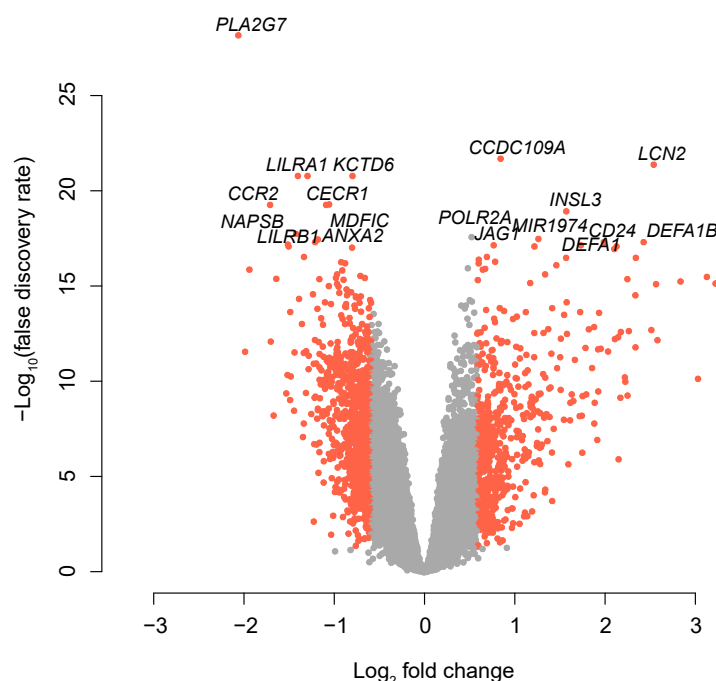


Figure 3.31: Comparison of sepsis and post-surgical samples.

Volcano plot showing differences in gene expression between CAP sepsis patients and post-surgical samples. Significance of differential expression is plotted against the log fold change, with a positive fold change indicating higher expression in sepsis. Transcripts with $FDR < 0.05$ and $FC \geq 1.5$ are plotted in *red*, and the most significant genes are labelled.

the antibacterial humoral response, apoptotic cell clearance, the innate immune response, and mitosis (Table A.19, background all genes detected in Davenport 2011 cohort). Directly contrasting FP and post-surgery samples was not possible, but the surgery response could be compared to the CAP and FP sepsis responses defined in the Radhakrishnan 2010 cohort. This demonstrates that the FP sepsis response is more strongly correlated with the cardiac surgery response than is the CAP sepsis response ($r=0.75$ vs 0.69 ; Fig. 3.32).

3.4.3 Comparison of the sepsis and traumatic injury responses

The “Inflammation and the Host Response to Injury” microarray data were accessed through GEO (Xiao et al. 2011). These include serial samples from 167 adult patients and 37 controls. In order to compare the trauma response to the sepsis response, this dataset was reduced to samples taken within the same time

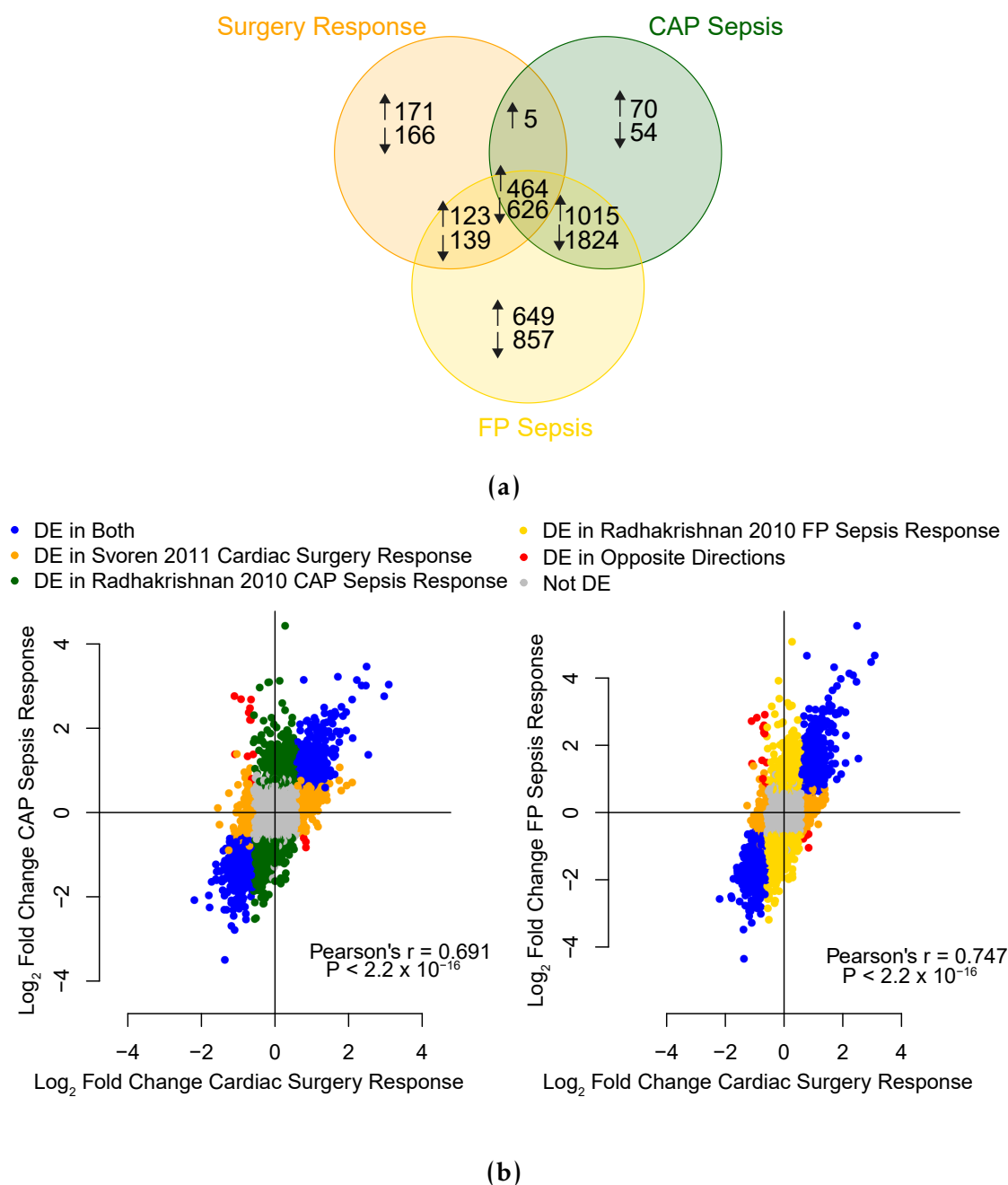


Figure 3.32: Comparison of the CAP, FP, and surgery responses.

(a) Venn diagram illustrating the number of genes significantly differentially expressed in the CAP and FP sepsis responses, and the cardiac surgery response. Genes are deemed to be shared if they are significantly differentially expressed with the same directionality. (b) The log fold change for differential expression of all transcripts is compared between the CAP, FP, and surgical responses. *Blue* colouring indicates shared significant differential expression; *orange* in the surgery response only; *green* in the CAP response only; *yellow* in the FP response only; *grey* that the gene is not significantly differentially expressed; and *red* differential expression in opposite directions.

frame as the sepsis cohort (five days following injury) and the 12,357 genes that were also measured in the sepsis data.

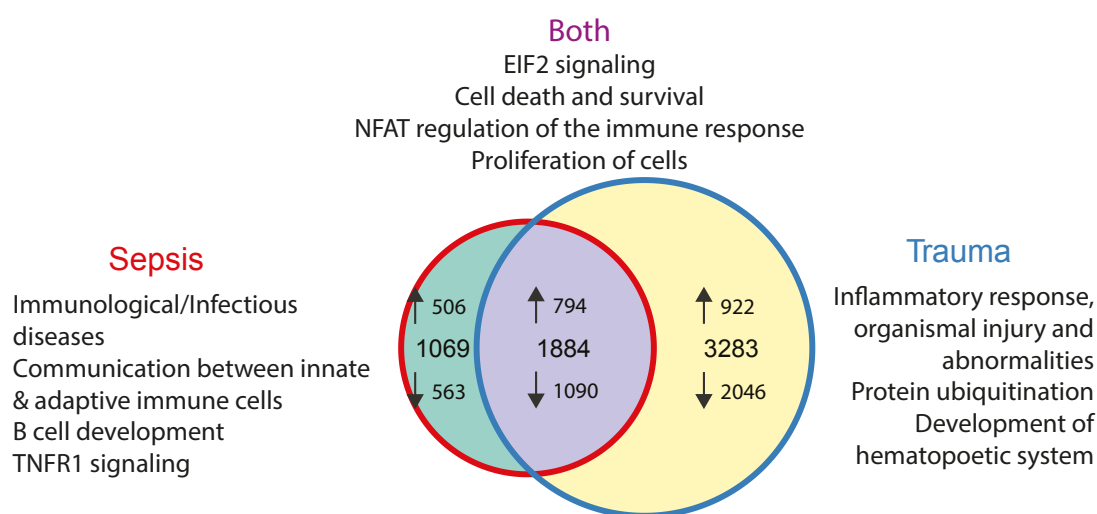


Figure 3.33: Comparison of the sepsis and traumatic injury transcriptomic responses. This Venn diagram illustrates the number of genes that were significantly differentially expressed in the sepsis and trauma response. Genes are deemed to be shared between the two if they were differentially expressed with the same directionality. Enriched pathways shared between and specific to the sepsis and trauma responses are given.

Global gene expression was contrasted between the healthy subjects and trauma patients, identifying 5,944 differentially expressed genes (FDR <0.05, FC >1.5). Due to more stringent criteria for identifying differential expression, this is fewer than reported in Xiao et al. 2011, where a “genomic storm” affecting > 80% of the genome was described. However, this reduced gene set included 1884 (64%) of the 3914 genes differentially expressed in the sepsis response that had also been measured in the trauma cohort, and fold changes were correlated (Fig. 3.33; $r=0.628$). Pathway analysis highlighted shared features of the sepsis and trauma responses, including eIF2 signalling, regulation of the immune response by nuclear factor of activated T cells (NFAT), and cell death and survival (Table A.20). As would be expected, enrichment of genes relating to infectious disease was restricted to the sepsis response, whereas trauma involved the inflammatory response and organismal injury. Although $\text{TNF}\alpha$ has been implicated in trauma, TNFR1 signalling was only enriched in the sepsis response.

3.5 Discussion

The Sepsis-3 definition contrasts the aberrant host response in sepsis with the appropriate immune response to an uncomplicated infection. It also highlights the broad nature of the syndrome, encompassing not only excess inflammation but also immunosuppression and dysregulation of pathways outside the immune system. The differential expression analyses described in this chapter investigated the sepsis transcriptome, with a view to exploring the molecular mechanisms underlying pathogenesis. I have particularly focussed on similarities and differences in gene expression between patient subgroups, and between sepsis and other inflammatory conditions.

3.5.1 Effect of cause of sepsis

The results described above illustrate the broad nature of the transcriptomic sepsis response, which here involved differential expression of 18% of the transcribed coding genome, and was to a large extent shared between sources of sepsis. There were however some differences in the response according to source of sepsis, with enriched networks relating to $\text{IFN}\alpha/\beta$ and the antimicrobial and inflammatory response. A number of viral response genes were also identified as being differentially expressed between CAP and FP, and source of sepsis was weakly associated with the first principal component of gene expression. Finally, the identification of a gene set that was predictive of infection source consistently across cohorts increases confidence in the differences in the gene expression response to different infection types.

Given the limited microbiology information available, biological and clinical interpretation of these transcriptomic differences is challenging. In FP, perforation of the peritoneal membrane results in a polymicrobial

infection, commonly including Enterobacteriaceae, *Pseudomonas aeruginosa*, and *Bacteroides fragilis*; in contrast, CAP is generally caused by a specific pathogen, such as *Streptococcus pneumoniae* or *Haemophilus influenzae* (Musher and Thorner 2014; Mayr et al. 2014). A subset of CAP patients also have a preceding or causative viral infection, and the results suggest that some modulation of the sepsis response occurs in these patients; however, it is likely that a greater number of patients had an unidentified viral infection, obscuring this association. Current work as part of the GAINs study includes refinement and expansion of the microbiological data, through clinician curation of patient information, PCR, and next-generation sequencing methods. This will greatly enhance power in comparisons of the host responses to differing infectious aetiologies.

Treatment choice, as well as the host response, may be specific to the source of infection. In CAP, narrow therapy targeting the specific pathogen is only possible following a positive blood culture, which occurs for about 40% of patients. The NICE guidelines recommend treatment with a combination of beta-lactam and a macrolide for high severity CAP; macrolides have immunomodulatory effects, and so could affect the transcriptomic response observed. However, patients will usually have begun antibiotics prior to ICU admission, for which time period no records are available and so this factor cannot be explored in this cohort. In FP, empiric broad spectrum antimicrobial treatments are recommended, along with surgical intervention for source control where appropriate. Corticosteroids are not routinely prescribed for either, but might be indicated by another condition, in which case they would again modulate the host response (National Institute for Health and Care Excellence 2014; National Institute for Health and Care Excellence 2017).

The FP sepsis response may be complicated by the causative insult, which is commonly blunt trauma or breakdown of surgical anastomosis. These events trigger the release of DAMPs, initiating a non-infectious inflammatory response

in combination with the immune response to infection. *HMGB2*, which was included in the diagnostic model, is an example of a DAMP downregulated in CAP relative to FP. Whilst *HMGB2* and related proteins are also active during infection, the higher expression of *HMGB2* in FP might be related to their pre-sepsis injury. The FP sepsis response is further complicated by the effects of surgery and anaesthesia. These are not restricted to those patients suffering from post-surgical complications, as all FP patients have undergone a laparotomy for diagnostic purposes. In comparisons to the cardiac surgery patients, the FP sepsis response did appear to show greater overlap with the response to surgery, which could indicate that a subset of the changes in gene expression relate to the effects of surgery. Ideally, patients without infection undergoing laparotomy would be used as a specific control for FP sepsis to further identify and explore sepsis-specific features of the transcriptomic response.

Previous studies have not consistently found a difference in the transcriptomic response as a result of the infection type (Tang et al. 2008; Pankla et al. 2009; Wynn et al. 2011; Parnell et al. 2012; Tsalik et al. 2016). However, our findings are supported by a recent study that found a similar discriminatory signature in ICU patients that can diagnose CAP, based on the *FAIM3:PLAC8* gene expression ratio (Scicluna et al. 2015). It has also been reported that CAP patients with mixed bacterial and viral infections have more severe illness and longer hospitalisation, indicating that these patients have a different molecular response to disease (Johansson et al. 2011). Finally, in a GWAS for survival in sepsis, significant associations were only identified when FP patients were excluded to leave just CAP sepsis (Rautanen et al. 2015). Together with the results described in this chapter, this suggests that causative infection has a subtle but potentially important impact on the sepsis response, which may be modulated by time from disease onset.

3.5.2 Longitudinal gene expression analysis

The course of disease progression varies between sepsis due to CAP and sepsis due to FP. Patients with FP usually have a rapid onset of illness, whereas patients with CAP are often unwell for many days prior to ICU admission. Comparison of the two groups on the basis of clinical covariates indicate that the CAP patients are more severely ill at the time of sampling. Whilst the FP GAINs patients had a lower overall mortality rate, it is possible that differences between the transcriptome of CAP and FP patients might reflect differing phases of disease being profiled in the first few days in ICU, which for FP might be an earlier, more rapidly progressing stage of the immune response than in CAP. The sepsis response might then become more homogeneous across subgroups with time.

The availability of small numbers of serial samples provided an opportunity to examine the longitudinal transcriptomic response in sepsis. Temporal changes were more pronounced in FP, with >8 times as many genes differentially expressed between admission and day five in ICU than in CAP patients. This may reflect the generally longer period between disease onset and ICU admission in CAP, leading to a more stable immune response state at the time of sampling. Additionally, this time period is more varying in the CAP patients, although it still ranges from 0-14 days in the FP subgroup. This means that day one samples do not correspond to the same time point in disease progression, and therefore patients are unlikely to have consistent patterns of changing gene expression between their day one and five samples. This confounding factor, which is expected to have a greater effect in CAP, could explain the relatively small numbers of genes that were differentially expressed over time within patients. This was partially addressed by using the time of the causative event in FP to identify differentially expressed genes over the course of disease progression. This approach identified potential biomarkers; for example,

measuring expression of *SNN* over time may be useful as a prognostic indicator for multi-organ failure.

Disease stage and temporal differences are demonstrably important factors in the resolution of heterogeneity in the sepsis response within and between patients, and the results described in this chapter build upon previous work exploring the longitudinal SIRS response. Time series analysing changes in gene expression in sepsis have of necessity largely focussed on animal models, endotoxin challenge, or the development of infectious complications following traumatic injury (Talwar et al. 2006; Prabhakar et al. 2005; Calvano et al. 2005; Cobb et al. 2009; Seok et al. 2013). Some more recent analyses have used time-matched cohorts to define a diagnostic gene set for sepsis (Sweeney et al. 2015). This is the first description of the FP-specific sepsis transcriptome, and as such is particularly useful for the investigation of the time course of disease. Further work should focus on expanding the time frame over which serial samples are taken, which might aid understanding of the long term immunosuppression observed in survivors of sepsis (Wang et al. 2012). Investigation of additional sources of sepsis would also be of great interest, as this would help to resolve the effects of the factors discussed above that might impact CAP and FP differently.

3.5.3 The relationship between sepsis and SIRS

Paired samples from patients taken before and after cardiac surgery provide the opportunity to define changes in gene expression within patients directly. This greatly adds to the confidence with which differentially expressed genes are identified, as the baseline expression levels for each patient are known. Additionally, the nature of this cohort means that all samples are comparable in terms of time from the inflammatory insult.

The analyses described, in particular the PLS regression, demonstrate the global

nature of the response to surgery and the involvement of many pathways associated with the innate immune and inflammatory response. The gene expression changes suggest some variation in the surgery response between patients; however, the cohort is too small to explore this further and to identify patients responding differently. Whilst major surgery can lead to multi-organ failure, all surgical patients described in this chapter survived, and little information is available on possible post-operative complications. Expansion of the cohort to include a greater variety of patients would permit exploration of this heterogeneity and the factors leading to adverse events.

The findings described in this chapter corroborate the previously described overlap between SIRS of differing sources (Xiao et al. 2011), and provide a more direct comparison of surgery and sepsis through parallel recruitment and processing to reduce the confounding effect of technical artefacts. Despite the differences in outcome between the surgery and sepsis cohorts, a great deal of commonality was observed across the contrasts between the acute and baseline samples. The transcriptomic response to sepsis was defined through comparison to pre-surgical samples, demonstrating that it involved a significant proportion of the transcribed genome, and included the vast majority of the genes differentially expressed following cardiac surgery. The sepsis response was also found to be comparable to the genomic storm following trauma, previously described by Xiao et al. 2011, with both showing the importance of transcriptomic regulation through the enrichment of EIF2 signalling. In this study, the trauma response had been found to be correlated with the endotoxin and burn response to a similar extent.

However, some aspects of the sepsis response were not shared with surgery and trauma. For example, as expected genes relating to the antibacterial humoral response were significantly upregulated in sepsis in comparison to post-surgical samples. Additionally, the antimicrobial gene *LCN2* had relatively higher

expression in sepsis, and has been previously proposed as a biomarker for sepsis (Vazquez et al. 2015). These results provide further evidence for its potential utility as an early diagnostic marker for infection. However, it is of note that the different sample processing and data generation methods used, in addition to the disparate sample sizes and the more controlled effects of timing in the trauma cohort, are expected to have an impact on this comparison.

Although the results presented in this chapter were shown to be robust to differential leukocyte count, greater resolution of the sepsis response would be possible through cell type-specific profiling. Additionally, this study was begun before RNA-seq became commonplace and microarrays were used for subsequent cohorts for consistency; however, future work would benefit from the greater depth of information obtained from sequencing. Specifically, genes not included on the array platform could be detected, along with alternative splicing isoforms, across a greater dynamic range.

3.6 Conclusions

The analysis presented in this chapter combines multiple transcriptomic datasets to compare the systemic inflammatory response across different sources. This demonstrates that a significant proportion of the sepsis transcriptomic response is shared between CAP and FP patients, although there is some specificity according to infection site. These findings may have implications for the management of sepsis, suggesting patient stratification and targeting treatment on the basis of cause of sepsis could be beneficial. In addition, future research and drug development may benefit from the use of more homogeneous cohorts. However, aspects of the transcriptomic response are shared across sepsis due to CAP and FP and sterile SIRS, indicating that findings in one SIRS subtype might be relevant to another and that similar treatment strategies could be considered.

VALIDATION AND CHARACTERISATION OF SEPSIS RESPONSE SIGNATURES

This chapter investigates inter-individual variation in the host response to sepsis, validating and exploring further previously described sepsis response signatures that identify subgroups of patients with clinically important features

4.1	Introduction	107
4.2	Results	117
4.3	Discussion	151
4.4	Conclusions	158

4.1 Introduction

Despite the completion of more than one hundred immunomodulatory drug trials, there are currently no specific therapies for sepsis, with treatment limited to antibiotics and organ support. One possible explanation for the difficulties in sepsis drug development is the confounding effect of disease heterogeneity (Marshall 2014; Cohen et al. 2015). This chapter aims to extend previous work using functional genomics approaches to investigate inter-individual variation in sepsis, and to identify opportunities for patient stratification and personalised medicine.

4.1.1 Heterogeneity in the sepsis response

As described in Chapter 3, the clinical diagnostic criteria for sepsis are broad and identify a heterogeneous patient population. The host immune response can be highly specific to an individual patient, modulated by their unique genetic and epigenetic background as well as a number of environmental factors. This was explored in relation to the source of sepsis in the previous chapter. The immune response is also very complex, comprising hundreds of factors with overlapping roles that may be activated or repressed to varying extents in different cell types throughout the course of the response to infection (van der Poll et al. 2017).

It is evident that dysregulation of this host response as seen in sepsis may vary considerably in terms of which players have key roles in pathogenesis and the overall impact of the resulting expression signature on outcomes. The development of effective targeted therapies therefore requires resolution of this heterogeneity in the individual host response to infection, and the identification of biomarkers that would enable patient stratification on the basis of the underlying response state (Pierrakos and Vincent 2010). Given the role of transcriptomic regulation in the immune response, investigating gene expression dynamics following infection could dramatically improve understanding of disease mechanisms within a given patient, allowing the development of targeted treatments for patient subgroups (Sweeney et al. 2016).

4.1.2 Current clinical risk scores

Several different measures of disease severity are currently used for sepsis patients in ICUs. Such severity scores can also be prognostic tools, such as the Sequential Organ Failure Assessment (SOFA) score (Vincent et al. 1996). This is a representation of organ dysfunction and rate of failure calculated using measures of organ function, for example creatinine for renal function and $\text{PAO}_2\text{:FIO}_2$ for

respiratory function. High or increasing SOFA scores indicate an increased risk of death; under the Sepsis-3 definitions, the quick SOFA (qSOFA) is also used as a diagnostic indicator (Singer et al. 2016). The Acute Physiology And Chronic Health Evaluation (APACHE) II score is a complementary assessment, calculated within the first 24 hours in ICU from standard clinical covariates including age, temperature, heart rate, and white cell count (Knaus et al. 1985). However, whilst the SOFA and APACHE II scores show predictive power for mortality in the septic population as a whole, they do not translate well to a risk of death on a patient-by-patient level.

4.1.3 Biomarkers and prediction models

A common research goal in sepsis is the identification of novel biomarkers, whose measurement in the clinic could aid in diagnosis or predict the likelihood of mortality or other specific outcomes (Sandquist and Wong 2014; Faix 2013). For example, serum procalcitonin (PCT) has been found to be elevated in bacterial sepsis (Nakamura et al. 2009). Measuring PCT by enzyme immunoassay has the advantage of speed over traditional diagnostic methods such as blood cultures, which typically take 12-36 hours (Kirn and Weinstein 2013).

Similarly, quantification of relevant gene expression might be a powerful approach in this context, as this could be performed in a matter of hours using a multiplex PCR system (e.g. bioMérieux FilmArray, (Salimnia et al. 2016)). Additionally, given the dynamic nature of gene expression, measuring a small number of key transcripts could characterise the immune response profile of the host at that point in time. This would allow treatment of the patient to be based upon a real time readout of their inflammatory state. An example of such a test is the ratio between *FAIM3* and *PLAC8* expression, which has been shown to diagnose CAP in critically ill patients, with important implications for informing treatment (Scicluna et al. 2015). In terms of outcome prediction, a number

of attempts have been made to characterise biomarkers and gene expression signatures in sepsis; for example, Tsalik et al. 2014 and Sweeney et al. 2016.

Omic technologies generate data for many thousands of genes and proteins, from which a greatly reduced set of informative variables must be extracted for translation to clinical applications. This large search space presents a challenge in biomarker and predictive model discovery, due to the large number of potential explanatory variables with relatively few observations. In this chapter the R package GeneRave is used to fit logistic regression models under these conditions, employing a sparsity prior to eliminate redundant predictors and fit a model with a small number of explanatory variables (Kiiveri 2008).

4.1.4 Unsupervised clustering analysis

An alternative approach to improve treatment in sepsis is patient stratification. As described in this thesis and as seen in clinical manifestation and the transcriptomic response to infection, sepsis is a highly heterogeneous syndrome. Therefore, a personalised medicine strategy might be beneficial, enabling the treatment of specific subgroups of patients with medicine that might be unhelpful or even harmful in other individuals (Sweeney and Wong 2016).

Such patient subgroups might be related to clear clinical phenotypes, for example, infection type (*i.e.* Gram positive vs Gram negative bacteria) is used to inform antibiotic choice when known. A number of additional criteria can be imagined, and indeed source of sepsis was explored in the previous chapter, revealing only relatively modest differences. However, this is a biased approach, limited by existing knowledge and hypotheses regarding pathophysiology. Conversely, unsupervised clustering provides an unbiased method for identifying latent structure in a dataset (Jacobs and Wong 2017).

The general aim is to identify subgroups of patients such that patients in the

same group are similar to each other but distinct from patients placed in another group. Many different methods exist to define clusters, none of which are viewed as optimal for all possible sets of conditions. Biological interpretation of the output is vital in determining whether or not subgroups are meaningful, given that the aim of such an approach is to inform therapy. Additionally, subgroups resulting from alternative methods may be used to explore the robustness of the clustering, and validation in an independent cohort is imperative (Ronan et al. 2016; Altman and Krzywinski 2017).

The subgroup analyses described below were carried out using the FactoMineR package (Lê et al. 2008). This uses two clustering methodologies: first, hierarchical clustering using Ward's method with Euclidean distances to initiate an appropriate number of clusters, which is verified by plotting the within group sum of squares for increasing numbers of clusters. Subsequently, the *k*-means algorithm is employed to consolidate subgroup membership.

4.1.5 Subgroups in paediatric sepsis

Sepsis subgroups have been investigated to the greatest extent in paediatric patients by Wong and colleagues. Unsupervised hierarchical clustering of transcriptomic data for septic shock patients was used to identify three subgroups, one of which had a higher mortality rate, illness severity, proportion of Gram-positive bacterial infections, and repression of adaptive immunity-related genes. The 6,000 genes differentially expressed between groups were reduced to a discriminatory set of 100 genes to validate these subclasses in new patients. Subsequently, 12 serum proteins were selected from this gene set for use in a biomarker risk model (Wong et al. 2009b; Wong et al. 2010b; Wong et al. 2012b). Their recent work has focussed on developing a personalised medicine strategy based on a transcriptomic signature of adaptive immunity and glucocorticoid receptor signalling that resolve two subgroups with differing

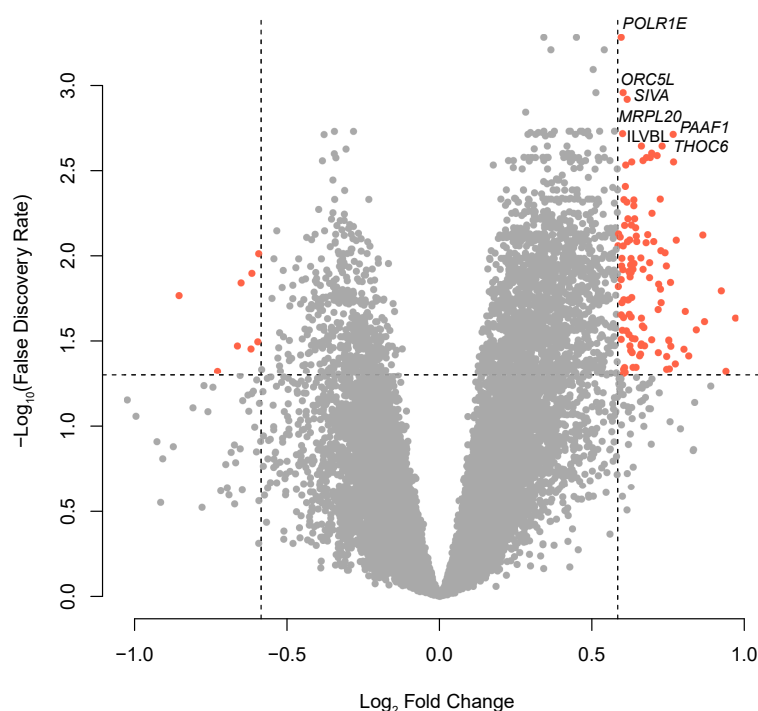
responses to corticosteroids (Wong et al. 2015b; Wong et al. 2016). Similar approaches have been applied to adult sepsis, but at the time of this study these were generally limited by small sample sizes and a lack of replication in prospectively recruited datasets.

4.1.6 Previous work in the Knight lab

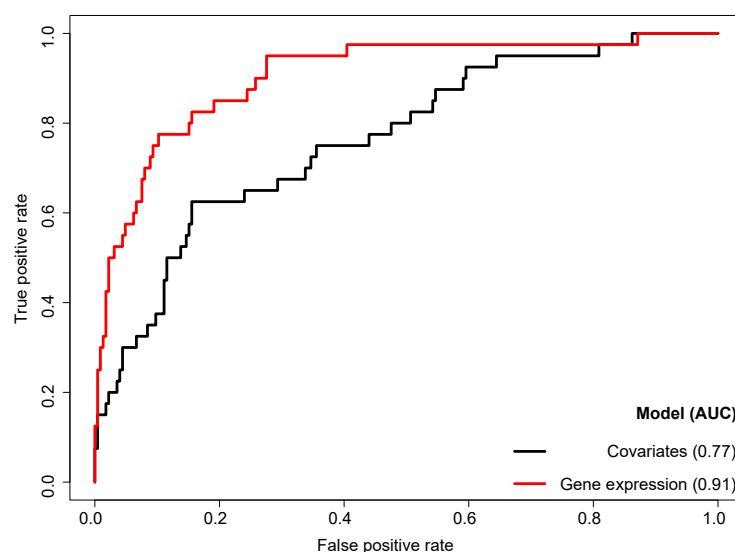
As described in chapter 3, global leukocyte gene expression had been previously profiled as part of the GAINs study in a cohort of 265 adult patients with sepsis due to CAP (Davenport 2011 dataset). These data had been used to investigate inter-individual variation in the transcriptome of sepsis patients, and to relate this to outcomes (Davenport 2014; Davenport et al. 2016).

Patients who died within the first 14 days of ICU stay were compared to patients still alive at 6 months on the basis of genome-wide gene expression. This identified 88 genes that were differentially expressed with an FDR <0.05 and a FC >1.5 (Fig. 4.1). With the hypothesis that some of these genes would represent a generic “mortality signature” useful for prognostication, the R package GeneRave (Kiiveri 2008) was used to identify a subset of ten genes able to predict survival (*B4GALT7*, *CHCHD10*, *CCDC71*, *SLC39A8*, *RIPK2*, *FHL2*, *MRPL11*, *PRL7*, *PPARG*, *THADA*). The gene expression model fitted was tested by leave-one-out cross validation, and had a low misclassification rate (MR) of 10.9% (AUC 0.91). For comparison, a model using clinical covariates (age, FIO₂, haematocrit, heart rate (low), systolic blood pressure (low)) was derived using the same approach, and was found to be less predictive of mortality than the gene expression data (Fig. 4.1, AUC 0.77).

A complementary approach was employed to investigate latent structure in the data. Unsupervised clustering (Lê et al. 2008) of the 10% most variably expressed transcripts identified two distinct groups within the cohort, which were denoted



(a) Comparison of 6-month survivors and 14-day non-survivors in the Davenport 2011 cohort.



(b) ROC curve for survival prediction model

Figure 4.1: Survival analysis in the Davenport 2011 cohort.

Generated using data from Davenport 2014. (a) A volcano plot showing differences in gene expression between patients with sepsis who died within 14 days of ICU admission and those who survived at 6 months. Significance of differential expression is plotted against the log fold change, with a positive fold change indicating higher expression in non-survivors. Probes with $FDR < 0.05$ and fold change ≥ 1.5 are plotted in red, and the most significant genes are labelled. (b) Receiver Operator Characteristic (ROC) curves show the sensitivity and specificity of gene expression and clinical covariate based models that aim to predict outcomes in sepsis.

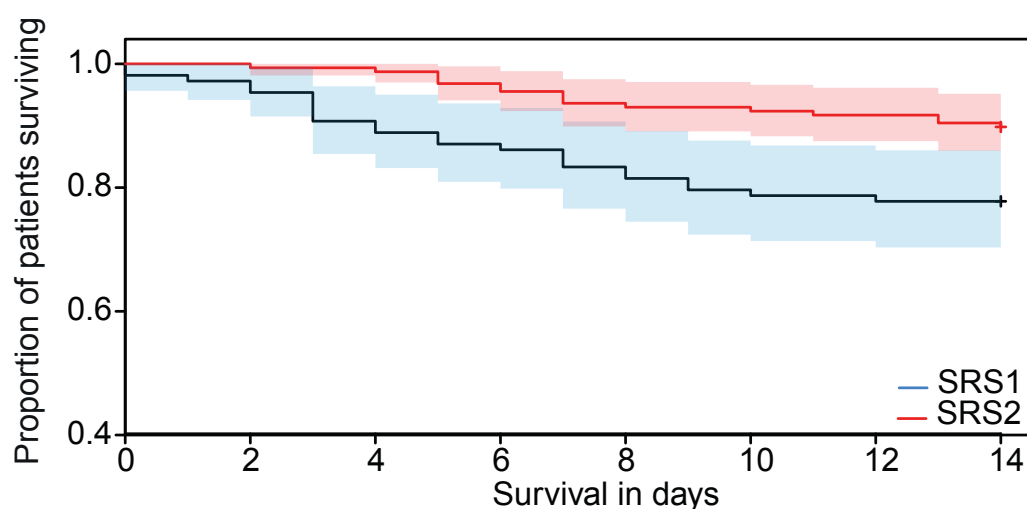
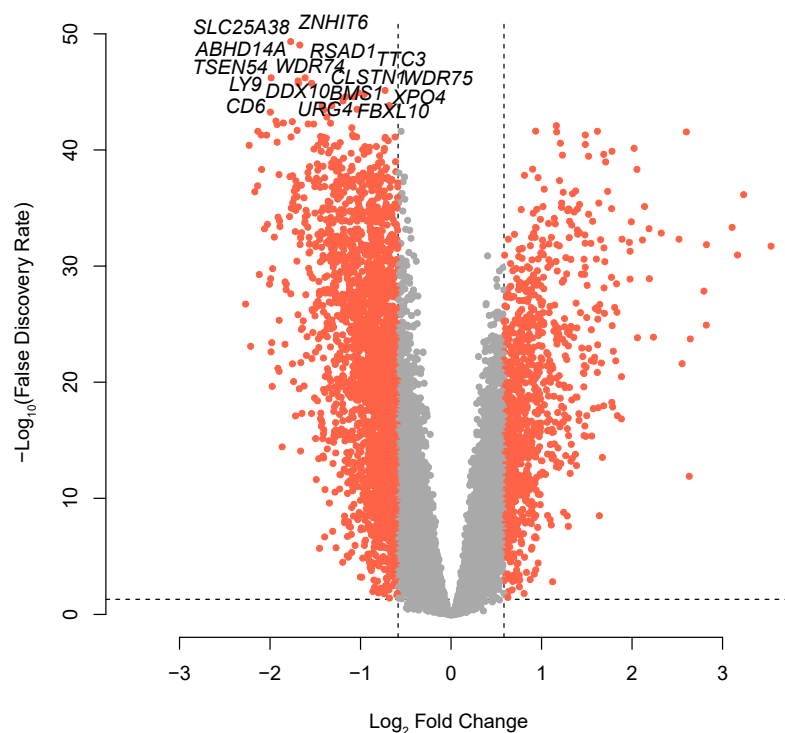


Figure 4.2: Kaplan-Meier plot for sepsis response signature groups.

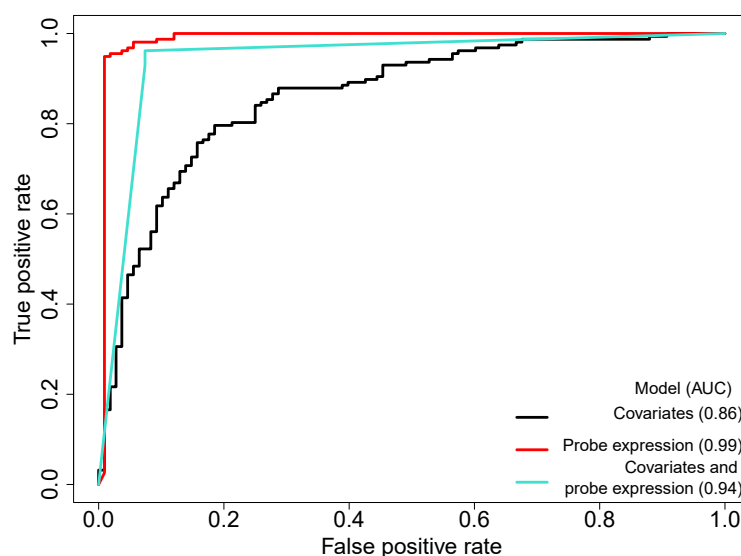
Subgroups defined using unsupervised analysis based on global gene expression variation differ significantly in terms of survival, particularly in the first 14 days following admission to ICU, with SRS2 having a lower mortality rate. The shaded areas represent the 95% confidence interval. Reproduced from Davenport 2014.

sepsis response signature (SRS) groups. These groups were informative for the immune response states and prognoses of individual patients. Specifically, the SRS1 group (n=108, 41% of patients) identified patients with a relatively immunosuppressed phenotype based on gene expression signatures and was associated with a significantly higher early mortality rate ($P=0.005$ for 14-day mortality; Fig. 4.2). Pathway enrichment analysis of the 3,080 genes differentially expressed between SRS groups ($FC > 1.5$, $FDR < 0.05$; Fig. 4.3) showed differences related to T cell activation, cell death, apoptosis, necrosis, cytotoxicity, and phagocyte movement.

SRS membership was associated with a number of indicators of more severe illness, including SOFA and APACHE II score, but not age, gender or microbiology. However, clinical covariates only showed limited efficacy at prediction of SRS group membership, either alone or in combination (age, lymphocyte count, neutrophil proportion, FiO₂, systolic blood pressure, bicarbonate, platelets, bilirubin, respiratory rate). A gene expression model comprising seven genes (*DYRK2*, *CCNB1IP1*, *TDRD9*, *ZAP70*, *ARL14EP*, *MDC1*,



(a) Comparison of gene expression between the SRS groups.



(b) ROC curves for sparse models predicting SRS.

Figure 4.3: SRS differential expression and predictive models in Davenport 2011

Generated using data from Davenport 2014. (a) Volcano plot showing differences in gene expression between CAP subgroups (SRS) resolved using hierarchical clustering. Significance of differential expression is plotted against the log fold change, with a positive fold change indicating higher expression in SRS1. Probes with $FDR < 0.05$ and fold change ≥ 1.5 are plotted in red, and the most significant genes are labelled. (b) ROC curves show the sensitivity and specificity of gene expression and clinical covariate models that aim to separate patients into SRS1 and 2.

EMR3) identified with GeneRave was fitted, which could be used more successfully for SRS classification (leave-one-out validation misclassification rate=3.8%, Fig. 4.3).

4.1.7 Endotoxin tolerance

The response to endotoxin is subject to homeostatic mechanisms to help ensure an appropriate response. Endotoxin tolerance (ET) is the phenomenon whereby cells previously exposed to lipopolysaccharide (LPS) display transient hypo-responsiveness to further stimulation (Beeson 1946). This is thought to be an adaptive mechanism to limit damage from excessive or inappropriate inflammation, and involves the precisely controlled regulation of several different gene sets (Seeley and Ghosh 2017). For example, pro-inflammatory cytokines including $\text{TNF}\alpha$ are subject to negative feedback and are repressed through epigenetic mechanisms such as chromatin remodelling, whereas antimicrobial genes and certain cytokines are induced.

It has been proposed that dysregulation of ET plays a role in sepsis, and may be an important mechanism causing sustained immunoparalysis even following recovery (Cavaillon and Adib-Conquy 2006). Whilst LPS is specifically a component of the cell wall of Gram negative bacteria, the cellular reprogramming and negative feedback involved in ET are also observed in Gram positive infections, which cause the majority of CAP sepsis cases. Additionally, some degree of cross-tolerance between Toll-like receptor ligands including LPS and TNF has been reported (Julian et al. 2015). ET has been shown to be positively associated with sepsis severity (Pena et al. 2014); moreover, inter-individual variation in the degree of tolerance developed may also be relevant to clinical heterogeneity (Fairfax et al. 2011). This chapter will therefore include analysis investigating the expression of genes previously implicated in ET in the context of sepsis, and in particular test for an association with the SRS groups.

4.1.8 Aims

The purpose of this chapter is to further investigate heterogeneity within sepsis due to CAP, and extend this analysis to additional sources of sepsis. I will build on the previous functional genomics work described above and investigate the findings in additional sepsis patients cohorts. Specific hypotheses tested in this chapter are:

1. that the previously described ten gene model for survival can predict 14 day mortality in a new CAP patient cohort
2. that the previously described sepsis response subgroups associated with immunosuppression and mortality can be resolved in the new CAP cohort
3. that an endotoxin tolerance-associated gene signature is enriched in SRS1
4. that there are subgroups comparable to the CAP SRS groups within a cohort of FP patients
5. that the SRS groups have similar gene expression profiles to published paediatric sepsis subgroups
6. that there are changes in sepsis response state over time
7. that the gene expression based model for SRS assignment can be refined to become more clinically applicable

4.2 Results

4.2.1 Application of survival models to new data

A set of ten genes had previously been demonstrated to predict 14-day survival in the Davenport 2011 cohort, and performed better than a model including five

clinical covariates (MR 10.9%, AUC 0.91; Davenport 2014). Initially, I aimed to validate this model in the Burnham 2014 CAP samples (n=106). As the phenotype of interest for validation was survival, I had selected CAP patients such that the new dataset had a 50% mortality rate, maximising power to detect differences in mortality. Following QC as described in the previous chapter, the validation data were normalised against the Davenport 2011 data using the VSN R package (Huber et al. 2002). This method compensates for batch effects and ensured that the gene expression values were scaled appropriately for application of the predictive models to the new data. However, the ten gene model had a high MR (34%) in my validation cohort, as did the covariate-based model (Fig. 4.4).

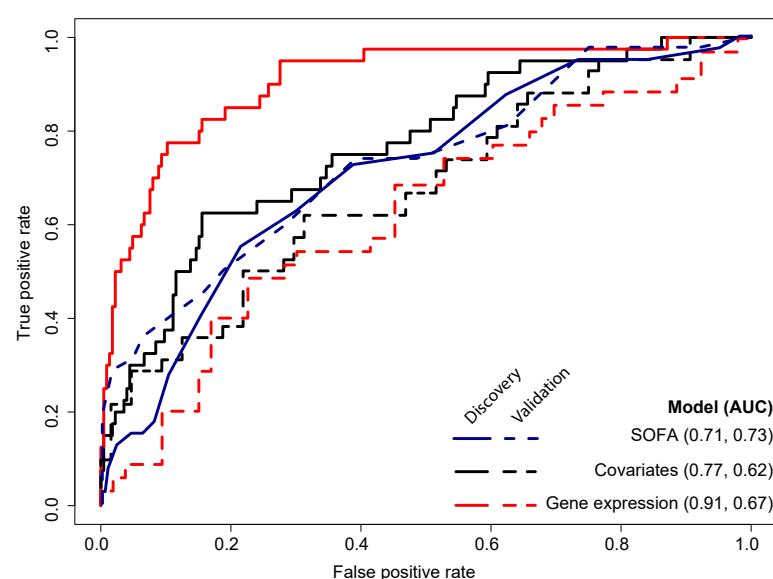


Figure 4.4: Validation of survival prediction model.

Receiver Operator Characteristic curves showing the sensitivity and specificity of SOFA score, gene expression, and clinical covariate-based models aiming to predict survival in sepsis. Although the latter models performed well in the Davenport 2011 cohort, the area under the curves (AUCs) in the Burnham 2014 cohort were poor.

4.2.2 Validation of the SRS groups in CAP

In the Davenport 2011 cohort, seven genes (*DYRK2*, *CCNB1IP1*, *TDRD9*, *ZAP70*, *ARL14EP*, *MDC1* and *ADGRE3*) had been identified that were predictive of SRS

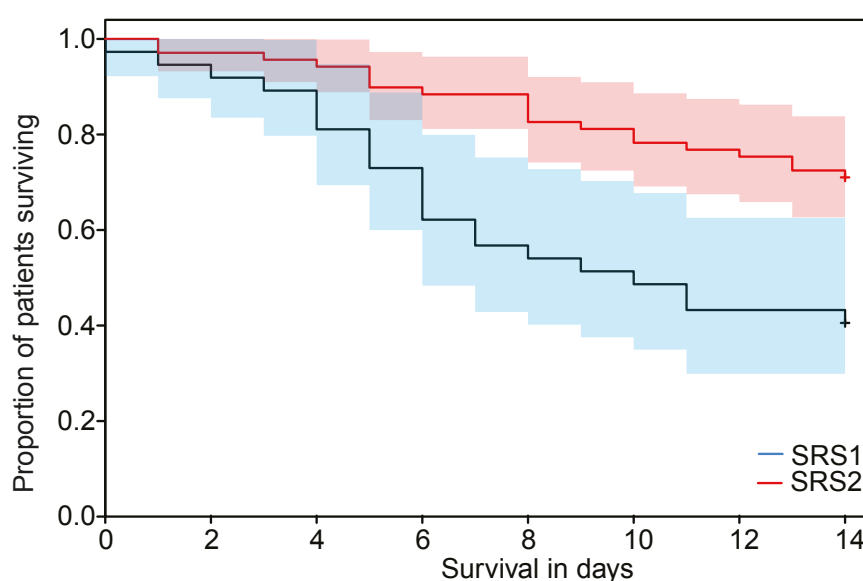


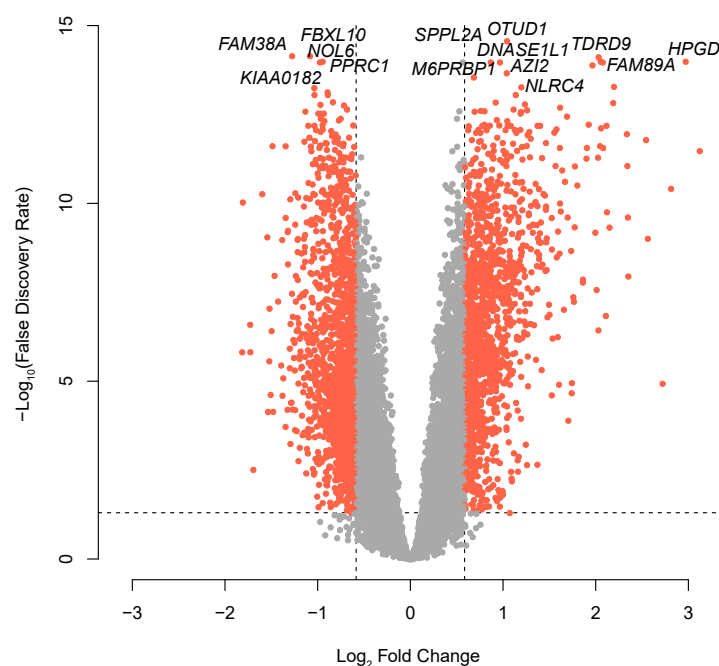
Figure 4.5: Kaplan-Meier plot for SRS (Burnham 2014 validation cohort).

SRS group membership was assigned in the Burnham 2014 CAP samples using a logistic regression model designed with GeneRave (Kiiveri 2008). Differential early mortality between SRS groups is shown over time, with SRS1 coloured *blue* and SRS2 *red*. The shaded areas represent the 95% confidence interval ($P=0.0007$).

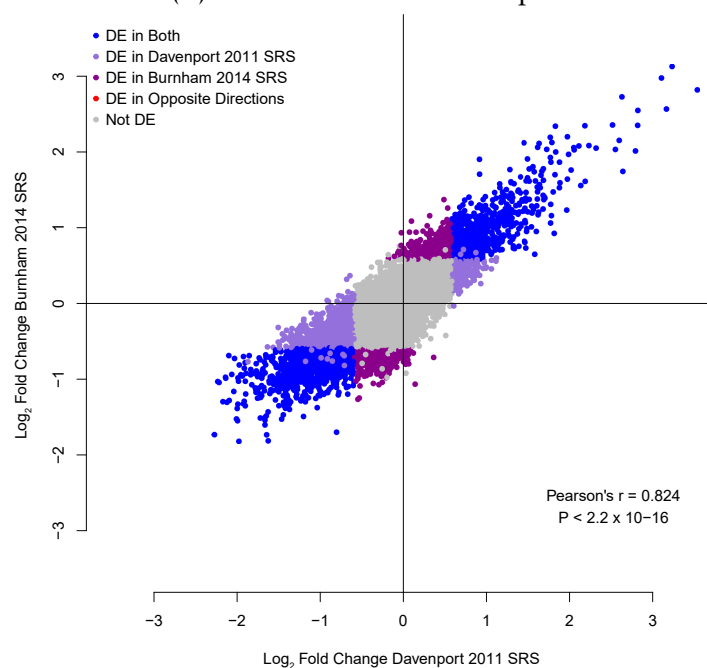
group membership (MR 3.8% in leave-one-out cross-validation). Given that the subgroups had been initially identified based on a larger gene expression signature, it was expected that a subset of marker genes would be able to achieve good internal prediction. However, it was unclear if a second cohort of prospectively recruited patients would have a similar substructure that could be resolved with the same genes.

SRS group membership was assigned for individuals in the Burnham 2014 CAP validation cohort using the seven gene expression model. This identified 37 patients (35%) as being members of the SRS1 subgroup and 69 patients (65%) as SRS2. I proceeded to test for associations with outcome and found that SRS1 group membership was associated with higher early mortality (log rank test $P=0.0007$ at 14 days; Fig. 4.5). This confirmed the result seen in the original cohort.

Differential gene expression analysis between SRS1 and 2 in the validation cohort again found a large number (2574) of significantly varying genes (Fig.



(a) SRS validation volcano plot



(b) Correlation between discovery and validation cohorts

Figure 4.6: Validation of the CAP SRS subgroups.

(a) Volcano plot showing differential gene expression between SRS1 and SRS2 in the CAP Burnham 2014 dataset, with group membership assigned using the seven gene model. A positive fold change indicates higher expression in SRS1, and probes with $FDR < 0.05$ and fold change ≥ 1.5 are plotted in red. (b) The log fold change for differential expression of all genes between SRS1 and SRS2 in the Burnham 2014 cohort is plotted against the fold changes for the same genes in the Davenport 2011 dataset. Blue colouring indicates significant differential expression in both datasets; lilac in the original cohort only; magenta in the new cohort only; grey indicates the gene is not differentially expressed.

4.6a, Table B.1). The fold changes of each transcript between the SRS groups were strongly correlated between the discovery and validation cohorts (Pearson's $r=0.82$ $P<2.2\times10^{-16}$; Fig. 4.6b). Similarly, the pathways enriched in the validation cohort showed high concordance with those for the Davenport 2011 cohort, including T cell exhaustion and hypoxia (Fig. 4.7, Table B.2).

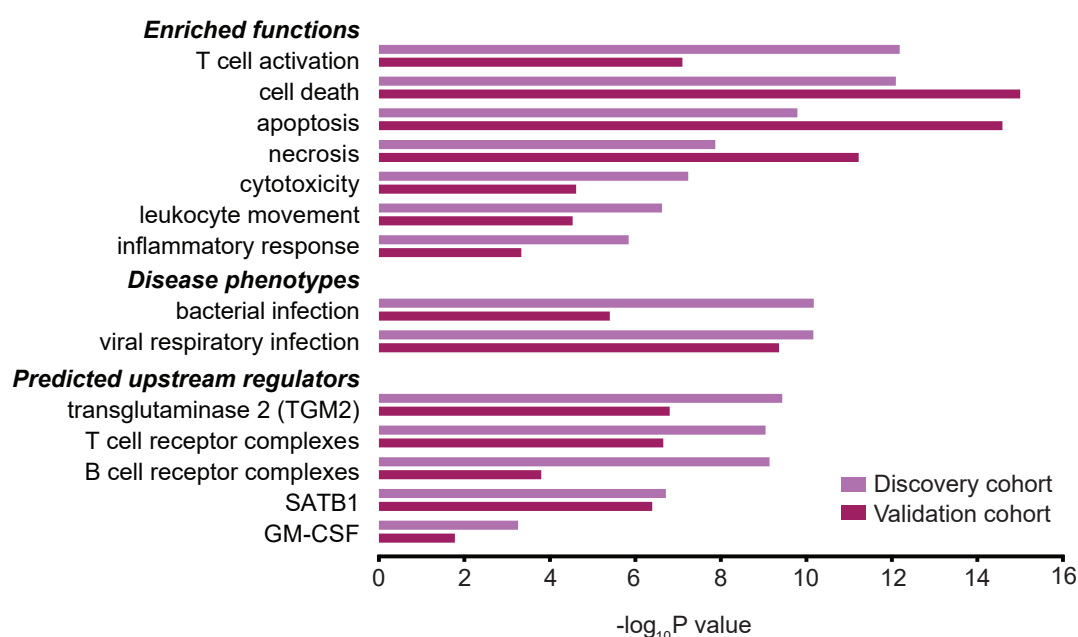


Figure 4.7: Validation of pathway enrichment in CAP SRS.

Canonical pathways and upstream regulators identified through enrichment analysis of genes differentially expressed between SRS1 and 2 are plotted by significance of enrichment in the Davenport 2011 and Burnham 2014 datasets.

These findings demonstrate that the Burnham 2014 cohort may also be divided into subgroups that show an association between relative immunosuppression and higher early mortality, and that the seven genes used in the model are representative of the overall transcriptomic signatures. However, given that the SRS groups were originally defined by gene expression data, it is perhaps not surprising that similar patterns are seen in the validation cohort. There could be an alternative substructure in the new cohort that is more significant than that resolved with the model derived from the discovery cohort. Unsupervised hierarchical clustering was therefore carried out on the validation cohort.

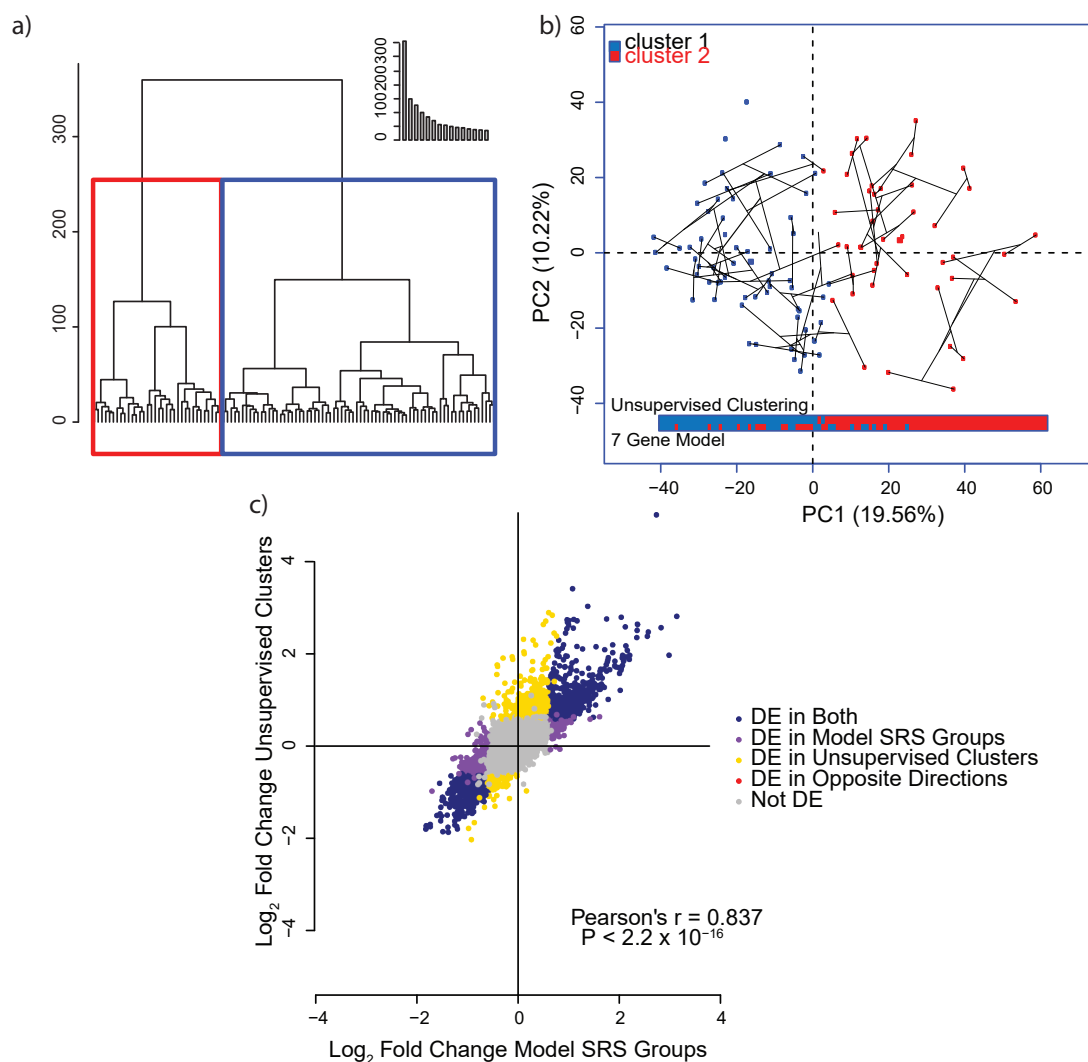


Figure 4.8: Comparison of SRS model prediction and unsupervised clustering

(a) Hierarchical clustering of the Burnham 2014 CAP samples (n=106) suggests that there are two distinct subgroups, (b) which are further resolved using *k*-means clustering in FactoMineR. The subgroups thus defined show 76% overlap with the SRS groups defined using the seven gene model. (c) The log fold change for differential expression of all genes between the unsupervised clustering groups is plotted against the fold changes for the sample genes in the SRS1 vs SRS2 contrast. *Blue* colouring indicates significant differential expression in both contrasts; *purple* in the SRS groups only; *yellow* in the unsupervised groups only; *grey* indicates the gene is not differentially expressed.

The data were divided into different numbers of subgroups and the within groups sum of squares calculated to identify the point where increasing the number of subgroups did not significantly reduce the intra-group dispersion. This suggested that it was appropriate to divide this dataset into two or three subgroups (Fig. 4.8a). When split into two groups, 76% of samples were assigned to the same groups as with the predictive gene set (Fig. 4.8b). Differential gene expression between these two groups was also highly correlated with that between the SRS groups assigned using the seven gene signatures (Pearson's r 0.84 $P < 2.2 \times 10^{-16}$; Fig. 4.8c).

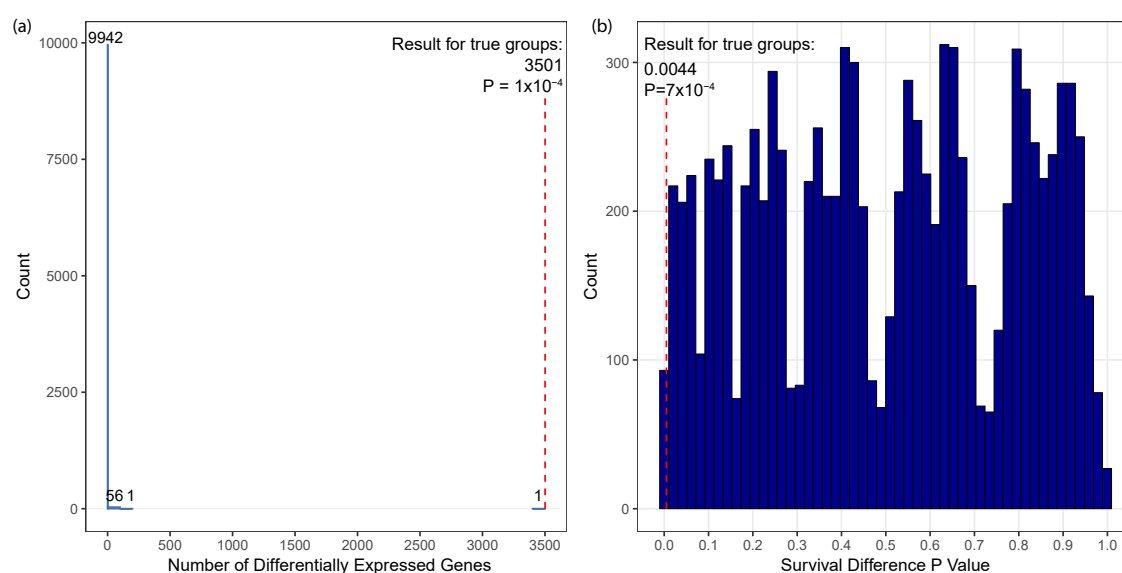


Figure 4.9: Permutation of SRS groups.

Generated using data from Davenport 2014. **(a)** From 9,999 permutations of SRS group membership, the majority did not have any differentially expressed genes, and the maximum number of differentially expressed transcripts was 13. This indicates that the 3501 transcripts differentially expressed between the SRS groups was significantly more than expected by chance ($p = 1 \times 10^{-4}$). **(b)** In survival comparisons, six permutations had a more significant p-value than that observed for the true SRS groups ($p = 0.0044$), giving an empirical p value for differential survival of 7×10^{-4} .

Given the intriguing biological associations observed, the sepsis response signatures described in the Davenport 2011 dataset and replicated in the Burnham 2014 cohort were examined further. This initially involved permutation analysis in the Davenport 2011 cohort to investigate if the SRS groups defined by unsupervised clustering were more clinically relevant than

random groupings of patients. SRS group membership was resampled without replacement, thus shuffling patient group assignment but keeping the size of the two groups the same as in the original analysis. Differential gene expression analysis was carried out between the resulting groups, and the survival comparison repeated. Over 9,999 iterations, there were no clusters with more than 13 significantly differentially expressed probes (true SRS groups $n=3501$). There were 44 permutations of the cluster assignment that resulted in significant differences in early mortality between the two groups, only 6 of which were more significant than for the true SRS groups. These 6 groupings did not show any significant differences in gene expression, suggesting that the SRS groupings defined using unsupervised cluster analysis represent a unique transcriptomic signature associating with early mortality (Fig. 4.9).

Additionally, in order to explore the impact of changes in the cell type composition on the gene expression differences observed, the differential expression analysis was repeated accounting for differences in relative mononucleocyte, polynucleocyte, and lymphocyte quantities. The genes identified as being differentially expressed with this approach included 2050 of the 2574 genes identified by the original analysis, and the fold changes for each gene tested were strongly correlated (Fig. B.1).

4.2.3 Endotoxin tolerance in the SRS groups

As described in Davenport 2014, pathway enrichment analysis of differentially expressed genes had demonstrated that the SRS groups displayed distinct inflammatory profiles. Notably, there was profound down-regulation of HLA class II genes in SRS1, the group with higher early mortality, along with genes involved in T cell activation and function. Network analysis had also highlighted differential gene signatures relating to T cell exhaustion, the hypoxic response, NF-kB, apoptosis, and histones.

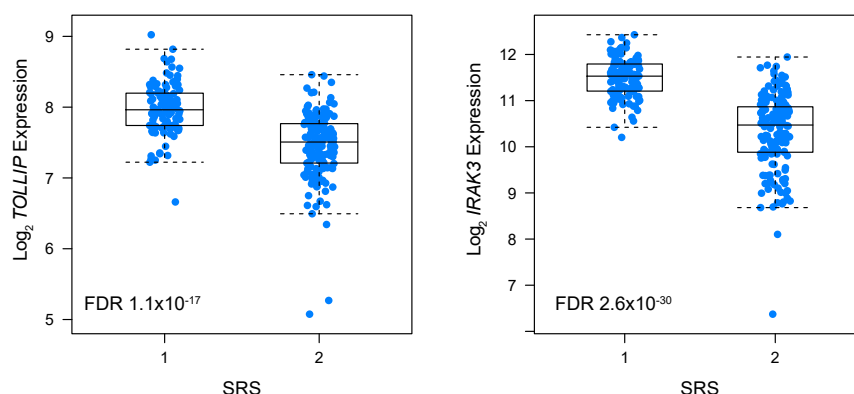


Figure 4.10: Boxplots of mediators of endotoxin tolerance in SRS (Davenport 2010).

The relative expression levels of *TOLLIP* and *IRAK3*, negative regulators of TLR signalling, are plotted in SRS1 (n=108) and 2 (n=157), and are shown to be significantly higher in SRS1.

These differences suggested a possible relationship between the sepsis response signatures and the extent to which endotoxin tolerance had developed. Some key mediators of tolerance were differentially expressed between the SRS groups. These included *IRAK3* and *TOLLIP*, negative regulators of TLR signalling that were relatively upregulated in SRS1 (Fig. 4.10). This is consistent with the hypothesis that SRS1 patients develop endotoxin tolerance to a greater degree, and thereby display a more immunosuppressed phenotype. This was further investigated through comparison of the sepsis cohort to published endotoxin tolerance datasets on the basis of gene expression.

Deriving an endotoxin tolerance gene signature

There have been a number of studies published where an *in vitro* model of endotoxin tolerance is employed to investigate genome-wide gene expression changes associated with tolerance (Table B.3). Generally, these involve the comparison of cells challenged with a single or two successive doses of LPS. Those using healthy human peripheral blood mononuclear cells (PBMCs) are of greatest interest for comparison to the sepsis gene expression patterns (Fresno et al. 2009; Pena et al. 2011; Yang et al. 2012; Allantaz-Frager et al. 2013; O'Carroll et al.

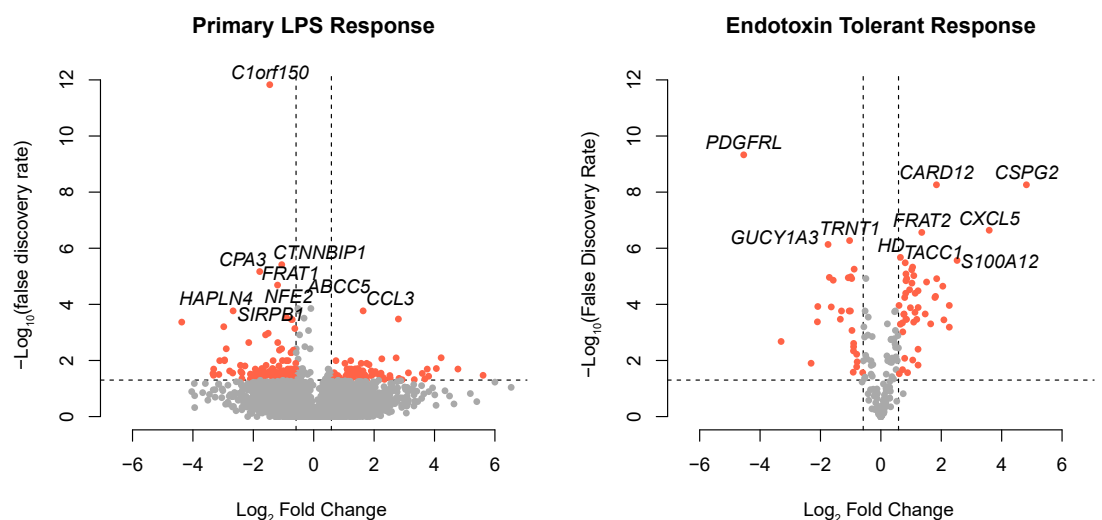
2014).

Defined simply, endotoxin tolerance is the difference in the response to single versus multiple exposures to LPS. However, descriptions of endotoxin tolerance usually focus on the reduced transcription or translation of pro-inflammatory genes upon a second stimulus. Several of the studies listed specifically investigated genes that were induced in the naïve cell response to LPS stimulation. For example, Foster et al. (2007) initially filtered out probes that were not upregulated in stimulated cells versus naïve cells, before categorising them as toleriseable or non-toleriseable depending on whether or not they are expressed at lower levels in tolerant cells than stimulated cells.

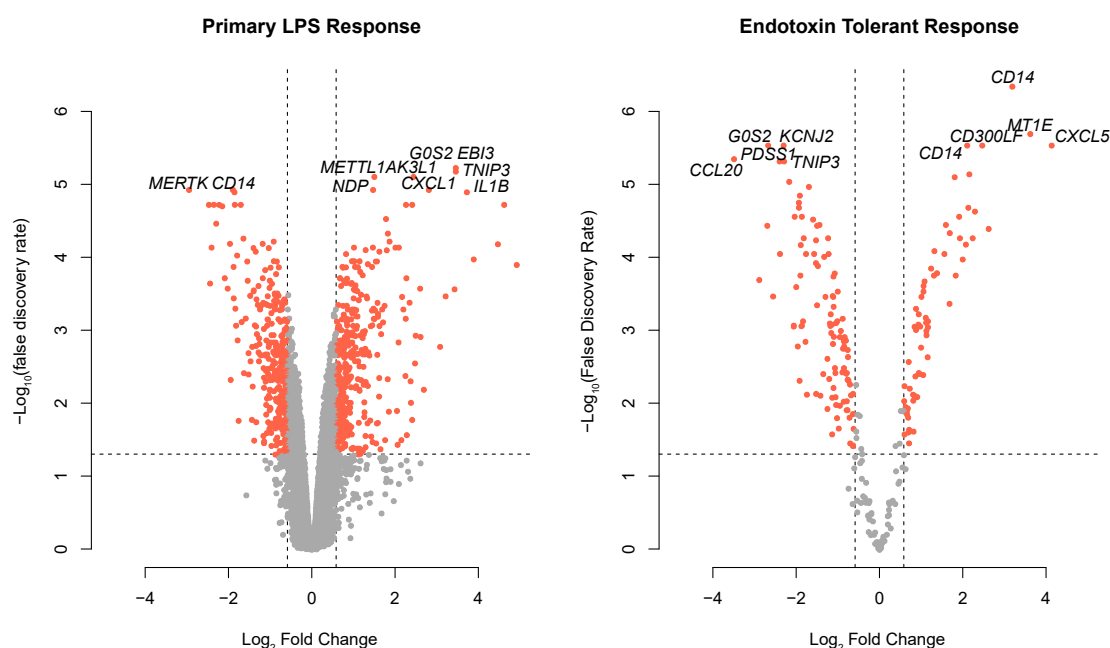
However, by limiting their analysis to LPS-induced genes, they exclude genes with potential roles in tolerance; for example, negative regulators of pro-inflammatory cytokines might be regulated in the opposite direction and play an important role in the development of tolerance. In order to enrich for genes relevant to the inflammatory response, analysis could be restricted to genes that are differentially expressed between naïve and treated cells. In contrast to the more specific filter mentioned above of genes induced with LPS treatment, this would also retain genes downregulated in response to endotoxin. An alternative approach would be to compare singly and doubly treated cells to naïve cells, and test for an interaction of treatment with number of doses.

Given the variety of different methods employed in the published studies of endotoxin tolerance, I downloaded processed data from the Gene Expression Omnibus (GEO) to analyse in parallel using the same approaches as described for the sepsis cohort. The Fresno et al. (2009) and Pena et al. (2011) datasets were selected for this analysis, due to their use of PBMCs from healthy human donors and the availability of processed data.

Focussing on the definition of tolerance as the alteration in the response to LPS stimulation due to prior exposure, the transcriptomic response to a single LPS



(a) del Fresno dataset



(b) Pena dataset

Figure 4.11: Endotoxin tolerance gene signatures.

Volcano plots for (a) the del Fresno dataset and (b) the Pena dataset, showing on the left the contrast between LPS treated and untreated cells, and on the right stimulated vs tolerant cells for the subset of genes differentially expressed in the first contrast. A positive fold change indicates higher expression in LPS (left) and tolerant (right) cells, and probes with $\text{FDR} < 0.05$ and fold change ≥ 1.5 are plotted in red.

treatment was first established by contrasting LPS-stimulated and naïve cells from the same subject. The analysis was thus limited to genes relevant to the inflammatory response. In this initial comparison, 195 probes from the Fresno et al. (2009) study and 668 from the Pena et al. (2011) data were differentially expressed (FDR <0.05, FC >1.5). The data were restricted to these genes and an endotoxin tolerance-associated gene signature derived by contrasting singly and doubly stimulated cells for all available timepoints. This defined lists of 73 (Del Fresno) and 336 (Pena) tolerance-associated genes (Fig. 4.11). The combined endotoxin tolerance gene expression signature resulting from this analysis consisted of 398 unique genes, 331 of which had also been measured in the sepsis cohort (Tables 4.1).

Dataset	# Genes in LPS Response Signature	# Genes in ET Signature	Overlap with SRS # DE Genes
(a) Fresno et al. 2009	189	73	
(b) Pena et al. 2011	605	336	
(c) Combined (a+b)	794	398	123

Table 4.1: Endotoxin tolerance gene signatures. Genes differentially expressed in the context of endotoxin tolerance were identified using (a-c) two datasets from the literature. The number of differentially expressed genes are given for each comparison ($p < 0.05$, ≥ 1.5 FC). Transcript levels were compared between naïve and LPS-stimulated cells, then significant genes from this were contrasted between cells treated once and twice with LPS. The number of genes also significant in the SRS1 vs 2 contrast is given.

There was limited overlap between the del Fresno and Pena gene lists (Table 4.1), which may be explained by a number of factors. In particular, the smaller Del Fresno gene set is likely due to the smaller sample size and thus reduced power to detect differentially expressed genes. Additionally, different platforms were used to quantify gene expression, and so not all genes have been measured in both datasets. Pathway analysis demonstrates some overlap at the level of enriched biological functions, for example “cytokine-cytokine receptor interaction”, “regulation of NF- κ B activity”, and TGF β are significantly enriched in both gene sets (FDR <0.05).

Enrichment of tolerance-associated genes in SRS1

Of the 331 genes in the tolerance associated expression signature that had also been measured in the Davenport 2011 cohort, 123 were found to be differentially expressed between the SRS groups (Table 4.1). In order to further investigate the tolerance gene set in the context of SRS, a gene set enrichment test that uses rotation to evaluate significance (ROAST) was used (Wu et al. 2010). This approach allows an additional layer of information to be included in the analysis, namely the weighting of the genes being tested by the fold changes observed in the tolerance data. This demonstrated that the tolerance-associated gene set was significantly enriched in the genes differentially expressed in SRS1 relative to SRS2 (10,000 rotations, $P < 1 \times 10^{-5}$). This finding was replicated in the Burnham 2014 CAP dataset (10,000 rotations, $P < 1 \times 10^{-5}$).

Pena et al. (2014) demonstrated that a tolerance-related gene set derived from their original data (Pena et al. 2011) may be used to distinguish septic from non-septic patients. This gene set furthermore showed greater enrichment in more severely ill patients. Whilst the expression signature they present is restricted to genes differentially expressed following two exposures but not one in contrast to the naïve state, these findings parallel the observation of immunosuppression associating with poorer outcome in SRS1. As with the more general tolerance signature described above, the Pena 2014 gene set was found to be significantly differentially expressed between the SRS groups (ROAST $P = < 1 \times 10^{-5}$).

4.2.4 Sepsis Response Signatures in FP

Having replicated the sepsis response signatures in CAP patients and observed consistent associations between immunological profiles and mortality rates, I then investigated if these subgroups were observed in sepsis due to a different infection type. To maximise sample size, the FP data from the Radhakrishnan

2010 and Burnham 2014 cohorts were combined to give genome-wide gene expression data for a total of 117 FP patients (147 samples). The combined cohort had a 28-day mortality rate of 13%, a mean age of 66 years, and 50.4% of patients were male. As the FP data were generated in two batches, systematic effects were corrected using ComBat (Fig. 4.12).

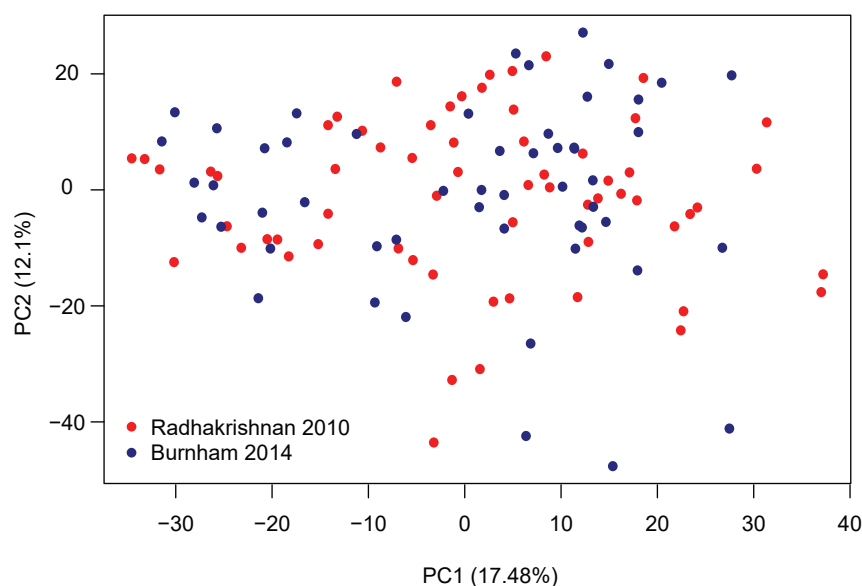


Figure 4.12: Principal component analysis of FP samples combined using ComBat.

Following correction of batch effects between the Radhakrishnan 2010 and Burnham 2014 FP cohorts, all available FP samples (n=147) are plotted by the first and second principal components of the top 10% most variably expressed probes. Samples are coloured according to batch, showing that the two datasets completely overlap with no evidence of a batch effect remaining.

Notably, this combined dataset includes a small number of serial samples from the same patient. Conversely, the original CAP cohort consisted of only the first available sample for each patient. Limiting the FP cohort to the first available sample would reduce the number of samples available; additionally, the inclusion of serial samples allows the exploration of dynamics in group membership. It is possible that the SRS groups reflect different time points in the sepsis response, despite not being related to day of sampling in the CAP cohort. As shown in the previous chapter, there are differences in disease progression between the diagnostic subgroups, and the knowledge of timing of disease

onset in FP provides a better indicator of disease stage than day of sampling to investigate this further. Following subgroup assignment, one sample from each patient (n=117) was selected to ensure sample independence in subsequent analyses. These were selected at random so that the samples included were evenly distributed over the time following disease onset.

Unsupervised clustering

An unsupervised approach was used initially, as this is agnostic to the existence of subgroups. The FactoMineR package was used as in the CAP patients to resolve distinct patterns of gene expression in the FP dataset, and identified two subgroups based on the top 10% most variable probes (n=2716). These groups were designated SRS1_FP (n=68, 46%) and SRS2_FP (n=79, 54%) (Fig. 4.13).

Cluster membership was not related to the microarray batch following use of ComBat to compensate for batch effects. One group, henceforth designated SRS1_FP to be consistent with the previous findings in CAP patients, was associated with higher early mortality (14 day mortality log rank test $P=0.0096$, 18.8% in SRS1_FP vs 4.3% in SRS2_FP; 4.13c).

Comparison of CAP and FP sepsis response signatures

In order to characterise the biological differences between SRS1_FP and SRS2_FP, global gene expression was contrasted between the two groups. This identified 1075 probes corresponding to 928 genes that were significantly differentially expressed (Table B.4). Of these genes, 854 genes remain significant following correction for different cell type proportions (Fig. B.2). The ROAST gene set enrichment test was applied to the SRS1_FP vs SRS2_FP contrast, demonstrating that the genes differentially expressed between the FP groups were enriched for the previously defined genes differentially expressed between the CAP SRS

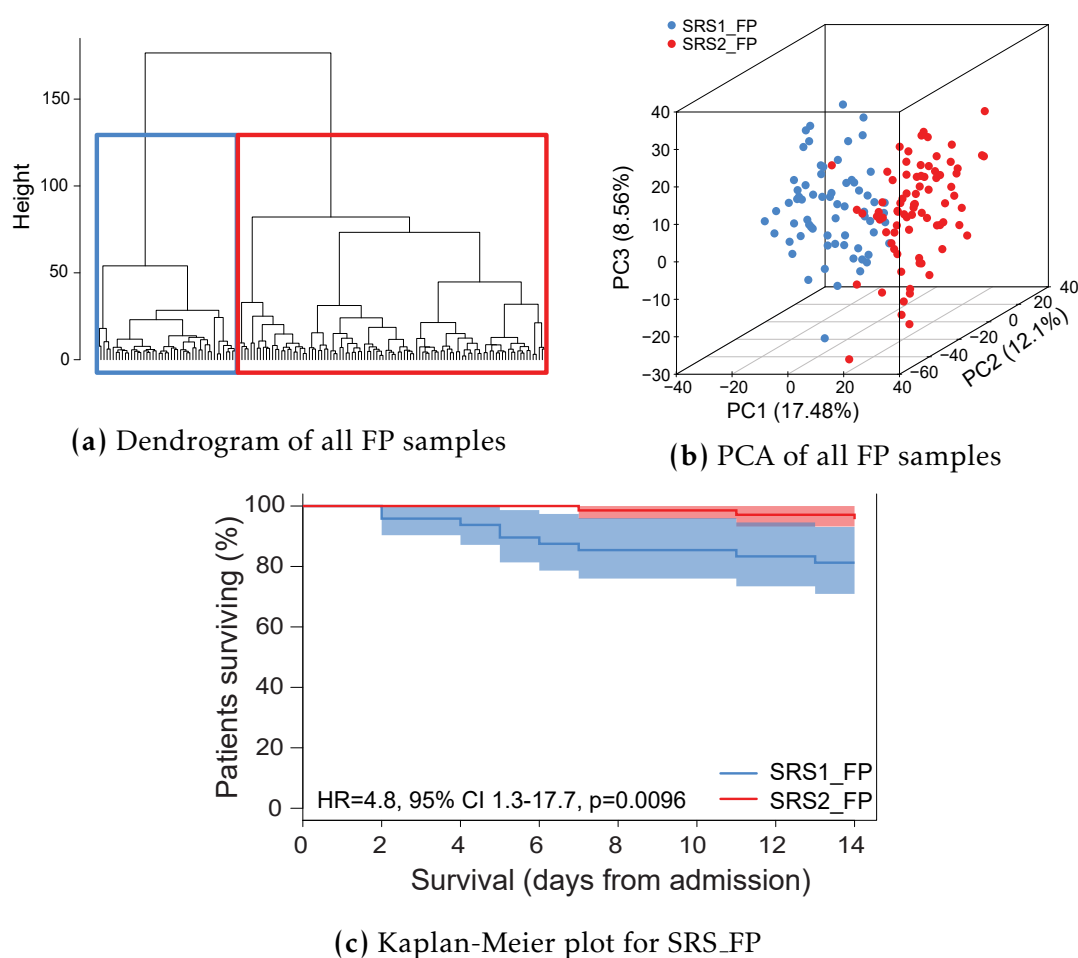
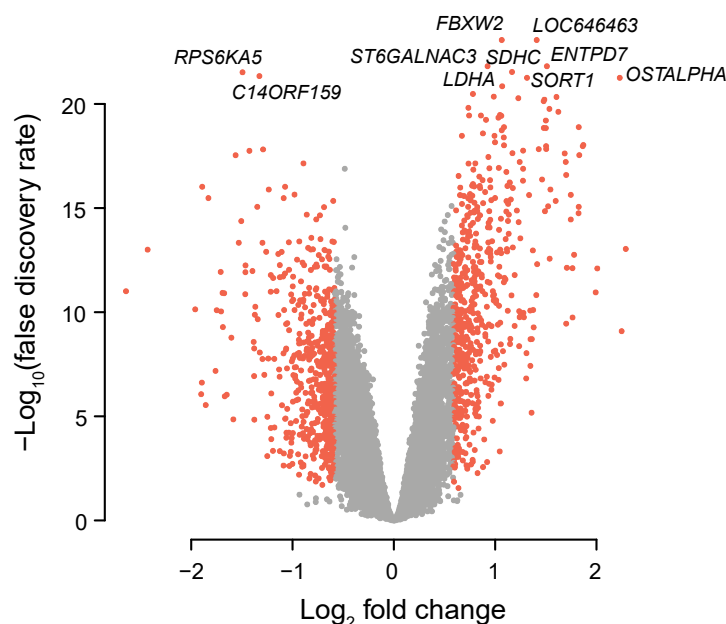


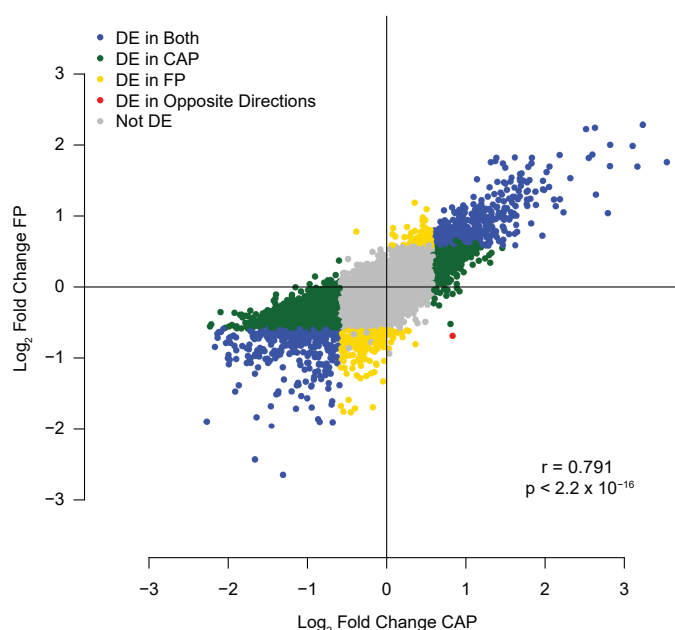
Figure 4.13: Unsupervised clustering using FactoMineR in FP patients.

The 147 FP samples are clustered based on the top 10% most variable probes. (a) Unsupervised hierarchical clustering reveals two subgroups in the data, which are then refined using *k*-means clustering. (b) The first three principal components are plotted with the proportion of variance explained by each component and samples coloured by SRS group membership. (c) The proportion of patients surviving in each SRS_FP group (reduced to one sample per patient) is plotted over the first 14 days following ICU admission. The shaded areas represent the 95% confidence interval.

groups. The fold changes for each gene were strongly correlated with those for the CAP SRS groups (Pearson's $r=0.791$), although there was only one third the number of significant probes (Fig. 4.14). The relatively small impact of cause of sepsis (Chapter 3) might be modulating SRS in CAP and FP to some extent; however, principal component analysis of the Burnham 2014 cohort demonstrates that SRS (assigned independently in CAP and FP) is much more strongly related to global variation in gene expression than source of sepsis (Fig. 4.15a)



(a) Volcano Plot of SRS1_FP vs SRS2_FP



(b) SRS CAP and FP correlation plot

Figure 4.14: Correlation of differential gene expression in SRS for CAP and FP.

(a) Volcano plot showing the genes differentially expressed between SRS1_FP and SRS2_FP. A positive fold change indicates upregulation in SRS1_FP, and significantly differentially expressed probes ($FC \geq 1.5$, $FDR < 0.05$) are coloured *red*. (b) The log fold change for differential expression between SRS1 and SRS2 in FP is plotted for each measured gene against log fold change for the same gene between SRS1 and SRS2 in CAP. *Blue* colouring indicates significant differential expression between SRS groups in both CAP and FP; *dark green* in CAP only; *yellow* in FP only; *red* indicates differential expression in both but in opposite directions; and *grey* indicates that the gene is not significantly differentially expressed.

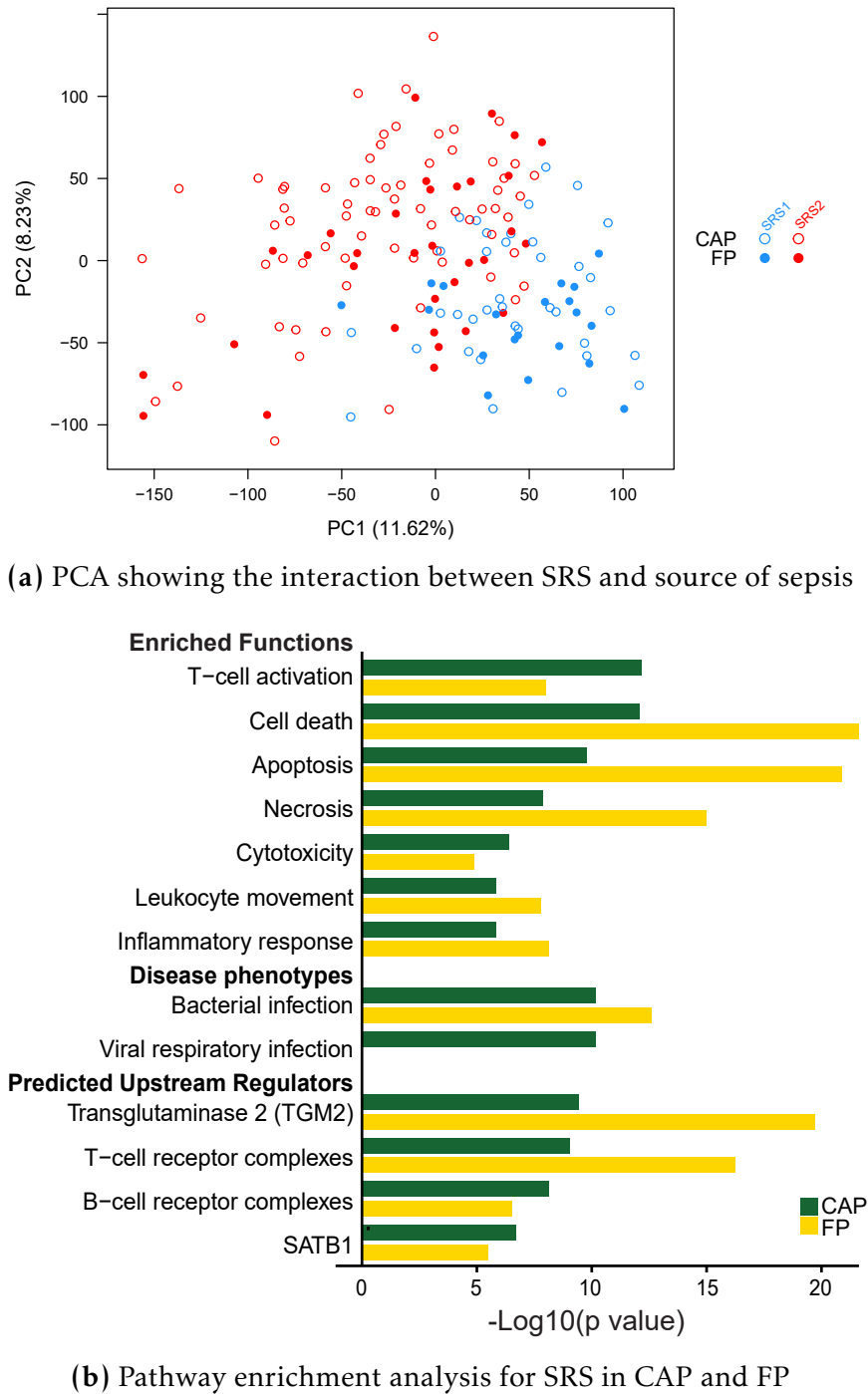


Figure 4.15: Interaction between SRS and source of sepsis
(a) The first two principal components of the Burnham 2014 gene expression data are plotted, with point type representing source of sepsis (CAP or FP) and colour denoting SRS group. For CAP samples, this was assigned using the seven gene model described above; for FP, by unsupervised clustering on the combined FP dataset. **(b)** Enriched functions, disease phenotypes, and predicted upstream regulators are identified for the genes differentially expressed between SRS1_FP and SRS2_FP. The significance of enrichment is compared to that for functions enriched in CAP SRS (Davenport et al. 2016).

The pathway enrichment output from IPA also showed overlap between SRS in CAP and FP; for example, cell death, apoptosis, necrosis, and T cell activation were significantly enriched functions in SRS1 for both causes of sepsis (Fig. 4.15b, Table B.5). Additionally, the endotoxin tolerance gene expression signature described above was enriched in SRS1.FP relative to SRS2.FP ($P < 1 \times 10^{-5}$).

Applying the CAP SRS model to FP patients

Given the similarities between the subgroups in CAP and FP, and to further explore substructure within the FP data, SRS group membership was assigned using the seven gene set established in the CAP dataset.

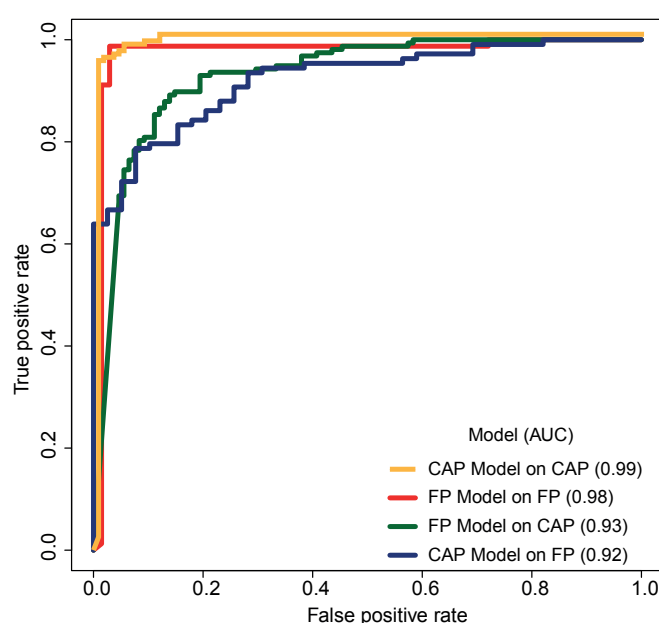


Figure 4.16: ROC curves for gene sets predicting SRS group membership.

The performance of SRS models derived using the FP and previously published CAP datasets are shown by receiver operator characteristic (ROC) curves and the area under the curve (AUC) given for each. Each model is tested on the dataset in which it was designed by leave-one-out cross-validation, and on new data of differing cause.

This approach has some limitations, stemming from the differences in methodology for the processing and normalising of the CAP and FP data. The SRS groups were validated in CAP by normalising the new Burnham 2014 data

against the Davenport 2011 data; the FP data were not normalised against the CAP cohort, and so the ranges of gene expression will be systematically, if subtly, different. Given the known expression differences between the CAP and FP, it is difficult to account for batch effects yet retain real biological variation between the CAP and FP cohorts. I therefore proceeded to apply the model fitted on the CAP data to the FP data with no further adjustments.

The SRS subgroups resulting from the application of the CAP model showed strong concordance with the results obtained by unsupervised analysis (MR 21.1%, AUC 0.923; Fig. 4.16). In general, a larger number of patients are assigned to SRS2, which could reflect systematically higher gene expression values in the FP cohort, or indicate subtle differences between the subgroup structures in CAP and FP (Fig. 4.17).

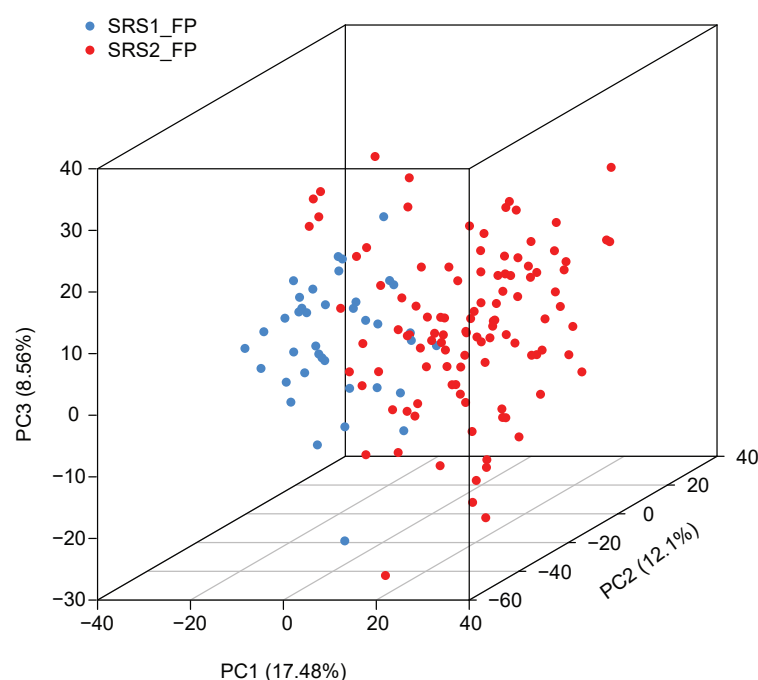


Figure 4.17: PCA of FP samples showing SRS group assignment.

The first three principal components of the top 10% most variable probes in the FP data are plotted with the proportion of variance explained by each component given. Samples are coloured according to SRS group membership assigned using the seven gene signature derived from the Davenport 2011 CAP cohort.

The two groups defined using the seven gene set showed a significant difference

in early mortality rates (log rank test $P=0.030$), and a differential gene expression signature strongly correlated with that from the CAP analysis (Pearson's $r=0.845$). The differences in gene expression between the FP SRS groups defined using the model are more similar to those previously reported in the CAP analysis, which is unsurprising given that the group assignment was carried out based on gene expression predictors selected from the CAP data. Gene set enrichment testing using ROAST demonstrates strong enrichment of the CAP differentially expressed genes, and SRS1 patients again demonstrated transcriptomic features of endotoxin tolerance, with pathway enrichment including OX40 signalling, B cell development, and T lymphocyte apoptosis (background all genes detected in FP cohort).

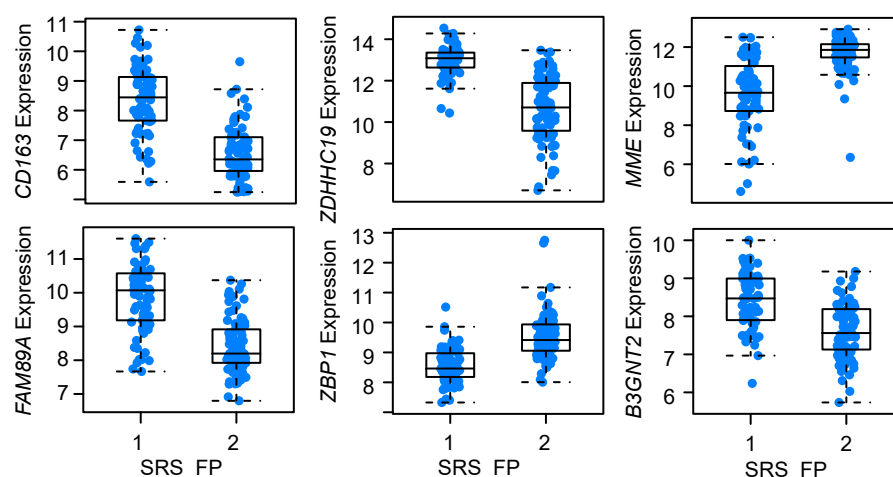
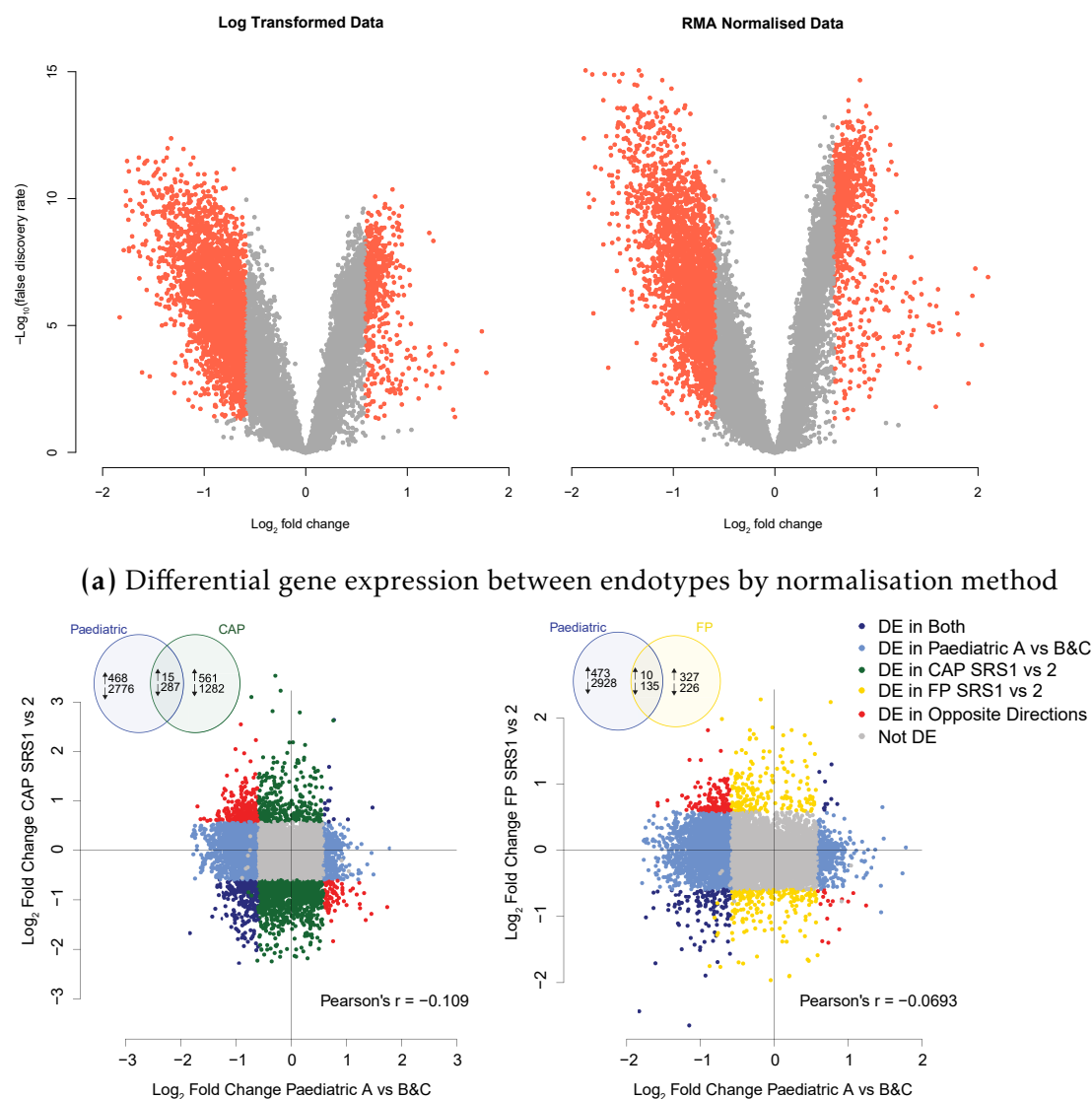


Figure 4.18: Boxplots of SRS_FP model genes. The expression of six genes, *CD163*, *ZDHHC19*, *MME*, *FAM89A*, *ZBP1*, and *B3GNT2*, is plotted for SRS1_FP and SRS2_FP patients in the combined FP dataset.

Defining a model for SRS in FP

Given that SRS groups might have some specificity according to source of sepsis, an SRS_FP specific model was derived using GeneRave (Kiiveri 2008). This comprised a six gene set (*CD163*, *ZDHHC19*, *MME*, *FAM89A*, *ZBP1*, *B3GNT2*) selected from the FP data predictive of SRS_FP group membership (Fig. 4.18). This predictive gene set had a 4.1% misclassification rate based on leave-one-



(b) Correlation between differential gene expression in paediatric endotypes and in SRS

Figure 4.19: Comparison of SRS with paediatric endotypes

(a) Volcano plots showing differential gene expression between paediatric endotype A and endotypes B and C, for log transformed data (left) and RMA normalised data (right). (b) Correlation between differential gene expression in paediatric endotypes (log data) and in (left) SRS in CAP and (right) in FP. Dark blue indicates significant differential expression in both contrasts; green and yellow in SRS CAP or FP only; light blue in paediatric endotypes only; red indicates significant differential expression in opposite directions; and grey no significant differential expression.

out cross-validation (AUC 0.975). When applied to the CAP cohort, it performs comparably to the CAP model on the FP data (MR 27.9%, AUC 0.931; Fig. 4.16). *CD163* encodes an acute phase regulated receptor expressed exclusively in monocytes and macrophages, which acts as an innate immune sensor for bacteria, and is an inducer of local inflammation (Fabriek et al. 2009). *ZBP1* binds foreign DNA and induces type-I interferon production (Takaoka et al. 2007).

4.2.5 Comparison to paediatric endotypes

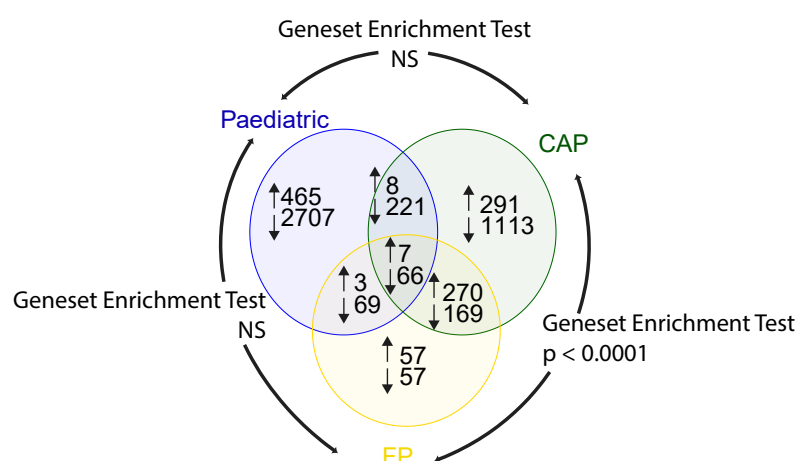


Figure 4.20: Gene set enrichment of SRS in paediatric endotypes

ROAST was used to assess gene set enrichment for the genes differentially expressed in CAP SRS and SRS_FP in the paediatric endotypes. This overlap was not significant, in contrast to the overlap between the CAP and FP SRS.

I next investigated whether there was overlap between the sepsis response signatures observed in GAIN patients and previously reported endotypes in paediatric sepsis. Wong et al. 2009b have previously identified three endotypes in children with sepsis, one of which (endotype A) showed poorer outcomes. In order to investigate similarities between paediatric endotype A and SRS1, I downloaded the paediatric expression data from the NCBI GEO database (accession: GSE26440). The original methods describe normalisation through robust multi-array average (RMA); in parallel, I log-transformed the data for better comparison to the GAIN data. Differential gene expression analysis

showed that these two approaches led to similar outputs for the contrast of endotype A (n=28) with endotypes B and C (n=70) (Fig. 4.19a).

The results from the log-transformed data were then compared to the SRS1 vs SRS2 gene sets for CAP and FP (Fig. 4.19b). Although there was some overlap in reported pathway enrichment (e.g., T- and B-cell receptor signalling), the gene expression patterns that distinguish SRS groups were not enriched in the paediatric endotype contrast (Fig. 4.20).

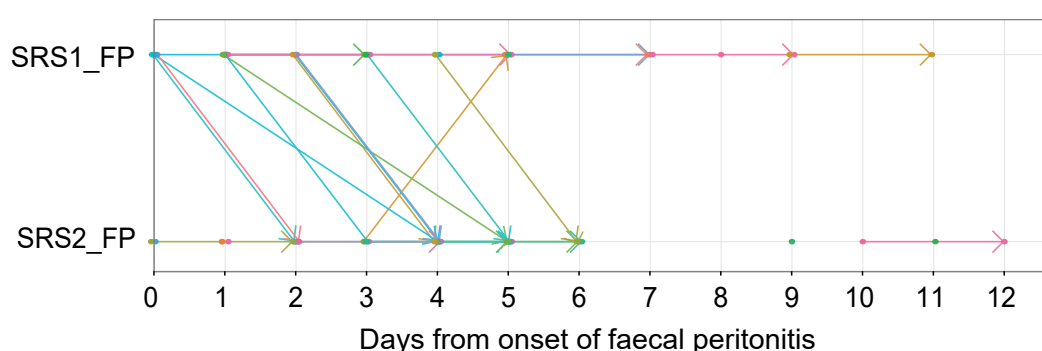


Figure 4.21: Timeline of SRS_FP group membership

SRS_FP group membership is illustrated for 24 patients with multiple gene expression time points. Each point represents a sample, with serial samples from the same patient connected by colour-coded lines. Samples are placed on a time line according to the number of days since disease onset and SRS group membership.

4.2.6 Sepsis response signature dynamics

Gene expression data for serial samples from FP patients provided the opportunity to investigate whether individual SRS group membership varied over time. Of the 24 FP patients with samples from multiple time points, 11 (46%) moved between SRS groups over days 1-5 of ICU admission (Fig. 4.21). Ten of these moved from SRS1 to SRS2, of whom only one died. Thirteen patients (54%) remained in the same SRS group; three of the seven patients remaining in SRS1 died, compared to no deaths amongst the six remaining in SRS2. This movement between SRS groups was reflected in large changes in gene expression within patients, illustrated by the six gene predictors for SRS_FP (Fig. 4.22).

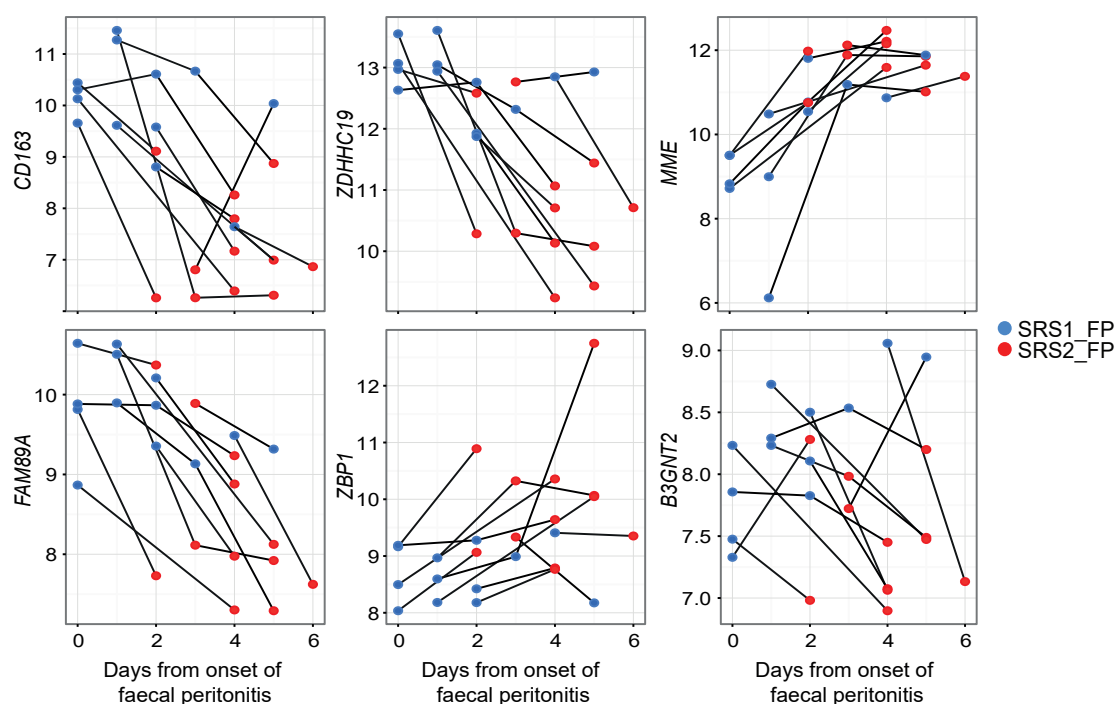


Figure 4.22: Expression of the SRS_FP model genes in patients moving between SRS groups.

Expression of the six genes selected as predictors of SRS_FP group membership (*CD163*, *ZDHHC19*, *MME*, *FAM89A*, *ZBP1*, and *B3GNT2*) in the 11 patients who move between SRS groups over time. Each point represents a sample, coloured by SRS group assignment and plotted according to time since disease onset. Samples from the same patient are linked, showing the changes in gene expression over time as patients move from one SRS group to another.

It is possible that the assignment of serial samples from the same patient to different SRS groups stems from the patient being on the border between the subgroups, leading to samples being assigned to different groups by chance. The proportion of patients moving between SRS groups, and the overall directionality of the movement make this unlikely to be the case for all patients, but it could be a contributing factor. Fig. 4.23 shows a plot of the first two principal components of the FP gene expression data, coloured according to SRS group assignment, and with serial samples from the same patients linked by arrows to show movement between SRS. This demonstrates that whilst some patients are borderline with fairly similar profiles over time, others are showing large changes from an SRS1 to an SRS2-like state.

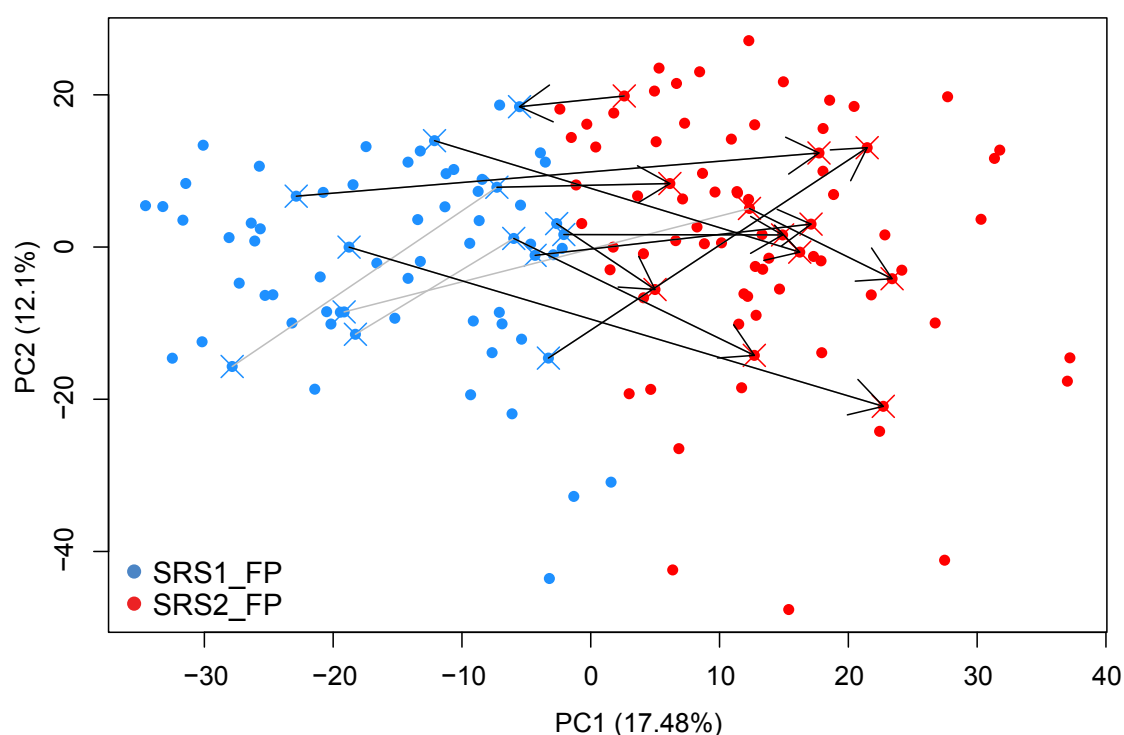
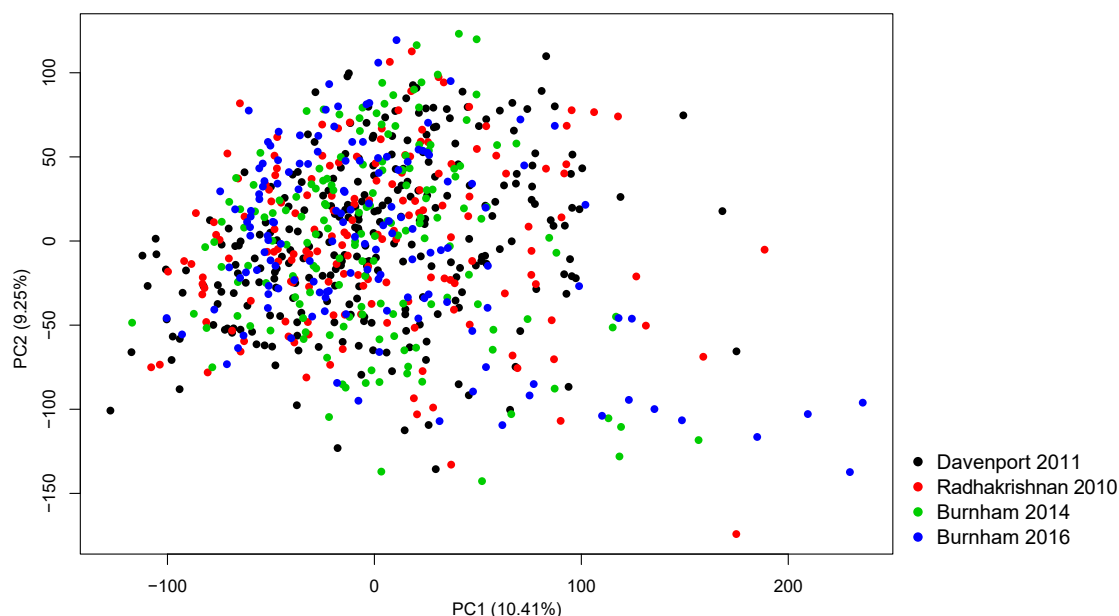


Figure 4.23: PCA investigating the evolution of SRS over time.

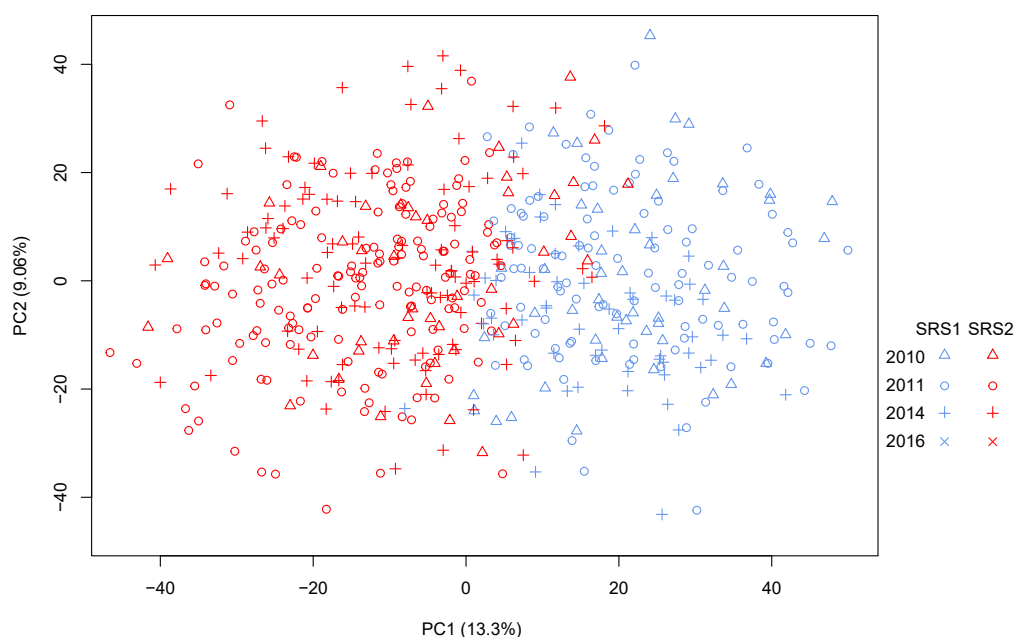
The first two principal components of the FP gene expression data are plotted, with the amount of variance explained by each given. Samples are coloured by SRS group assignment, and serial samples from patients who move between SRS groups are linked by arrows.

Validation of SRS dynamics

The Burnham 2016 cohort was designed to investigate gene expression dynamics in sepsis, comprising only patients with three serial samples available. The Radhakrishnan 2010 also contains some serial samples from CAP patients (n=39). These data were consequently used to validate and further explore the evolution of SRS membership over time. In order to assign SRS in these samples, firstly all available gene expression datasets were normalised together and batch effects removed using ComBat (Fig. 4.24a). The CAP and FP SRS prediction models were fitted using the original datasets from which they were derived, and then used to assign SRS to the new samples (Fig. 4.24b). This combined dataset comprises 732 samples from 552 patients, of which 279 samples (38%) were assigned to SRS1.



(a) Combination of all gene expression data using ComBat



(b) Sepsis response signatures in the combined dataset

Figure 4.24: Sepsis response signatures in the combined dataset.

(a) The first two principal components of all available GAInS gene expression data are plotted, with the amount of variance explained by each given. Samples are coloured according to the batch in which they were processed, and after correction using ComBat this is not associated with global variation in gene expression. (b) The first two principal components of the top 10% most variably expressed probes are plotted, with the amount of variance explained by each given. Samples are coloured by SRS group assignment, and point type represents batch, showing that the proportion of samples assigned to each SRS group is consistent across datasets.

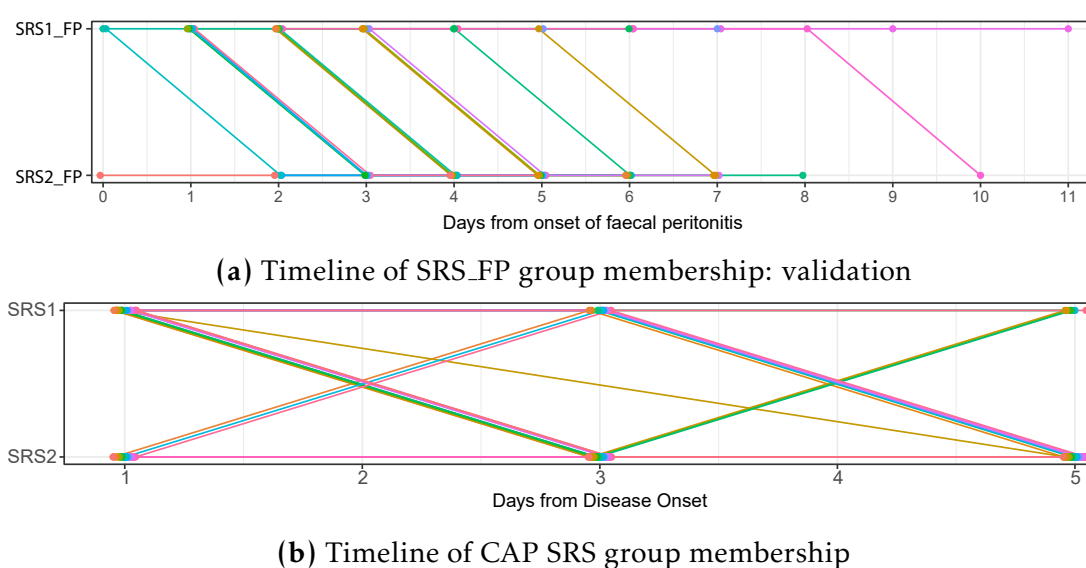


Figure 4.25: Evolution of SRS group membership

(a) SRS_FP group membership is illustrated for 24 patients with three serial samples profiled. Each point represents a sample, with serial samples from the same patient connected by colour-coded lines. Samples are placed on a time line according to the number of days since disease onset and SRS group membership. (b) SRS group membership is similarly illustrated for 64 CAP patients with serial samples profiled.

The evolution of SRS over time was investigated in the additional samples using the number of days since estimated disease onset. Two patients were recorded as having potentially inaccurate dates of disease onset, one being 31 days prior to ICU admission, and the other being seven days after recruitment to the study. These samples are therefore excluded. Nevertheless, these data replicated the findings from the Radhakrishnan 2010 data, with 15 of the 22 remaining patients moving between SRS1 and SRS2 and only 2 of these patients dying. Of the four patients who remain in SRS1 and the three in SRS2, three and one died respectively (Fig. 4.25a).

In order to see if similar dynamics were seen in CAP, all patients with more than one sample were compared over the first five days in the ICU. This demonstrated that a smaller proportion of patients moved between the SRS groups in CAP (29/64, 45.3%). Additionally, the directionality observed in the FP cohort from SRS1 to SRS2 was less pronounced in the CAP patients, with five patients moving into SRS1 (Fig. 4.25).

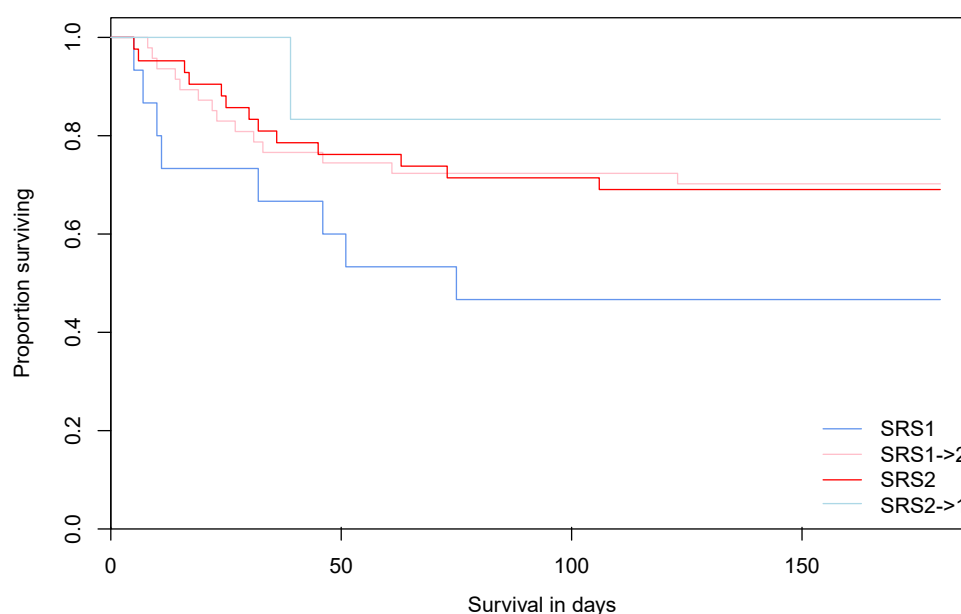


Figure 4.26: Kaplan-Meier plot for all patients with serial samples.

Patients with serial samples are classed by the SRS group to which their final time point is assigned, and whether or not they had previous time points in the other SRS group. This results in four groups: SRS1, SRS2→1, SRS2, and SRS1→2. The proportion of patients surviving in each of these groups is plotted over time, based on the time to death from the final sample taken for each patient. This shows that patients who move from SRS1 to SRS2 (*pale red*) have a more similar mortality rate to SRS patients (*red*) than to SRS1 (*blue*). Whilst patients moving from SRS2 to SRS1 (*pale blue*) do not show the same effect, there are very few patients in this group (n=6).

To look at differences in survival rate across the whole cohort relating to SRS group membership over time, all patients for whom serial samples were available were considered. These patients were categorised by their last known SRS status, and whether or not they had moved into that group. This results in four groups: patients who remain in the same group over time (SRS1, n=15; SRS2, n=42), and those who move between groups (SRS1→2, n=47; SRS2→1, n=6). These groups were then compared by applying the log rank test to these groups based on days to death for the final sample available for each patient. It is important to note that the requirement for serial samples exclude patients who died within the first few days in ICU; additionally, moving into SRS1 might preclude the possibility of later samples being taken given the higher mortality rate, thus biasing these analyses. This might explain the small size of the SRS2→1 group in particular. Acknowledging these limitations, this approach shows that the

patients who move from SRS1 into SRS2 have a more similar mortality rate to SRS2 patients (Fig. 4.26).

Finally, it is of particular interest to understand what might differ between patients who move from a SRS1-like state to a SRS2 gene expression profile, and those who remain in SRS1 over the sampling period. SRS group assignment was performed on the basis of gene expression profiles, so it was expected that overall gene expression would be very similar in SRS1 and SRS1→2. However, differential gene expression analysis was performed to test for the possibility that changes in the expression of a small number of genes might precede the more global changes associated with the movement from SRS1 to 2. Again, only patients with multiple samples were considered, and gene expression was compared using their latest time point that was assigned to SRS1 only.

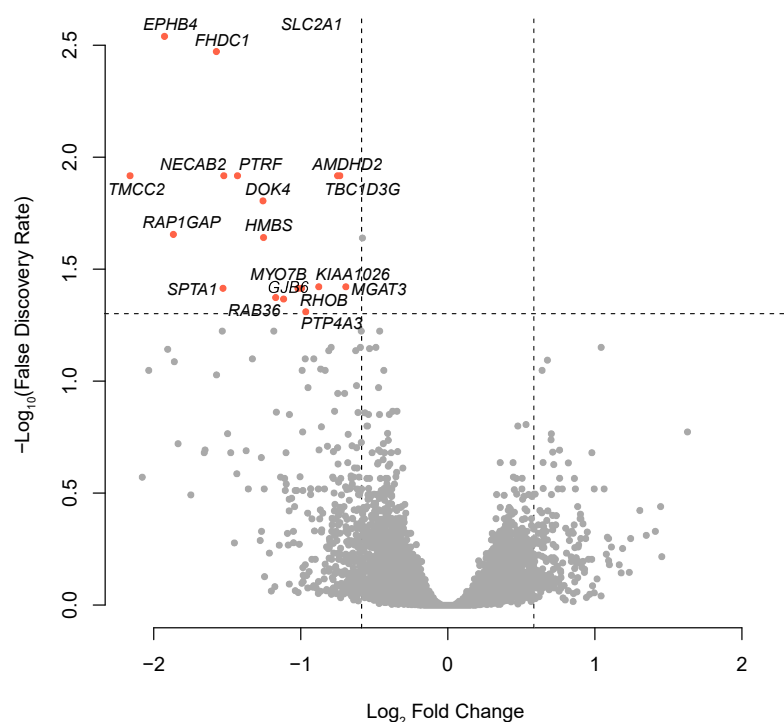


Figure 4.27: Differences in gene expression relating to SRS dynamics.

A volcano plot showing differences in gene expression between SRS1 patients who subsequently move to SRS2 and those who remain in SRS1. Significance of differential expression is plotted against the log fold change, with a positive fold change indicating higher expression in patients who move between groups. Probes with FDR < 0.05 and fold change ≥ 1.5 are plotted in red, and the most significant genes are labelled.

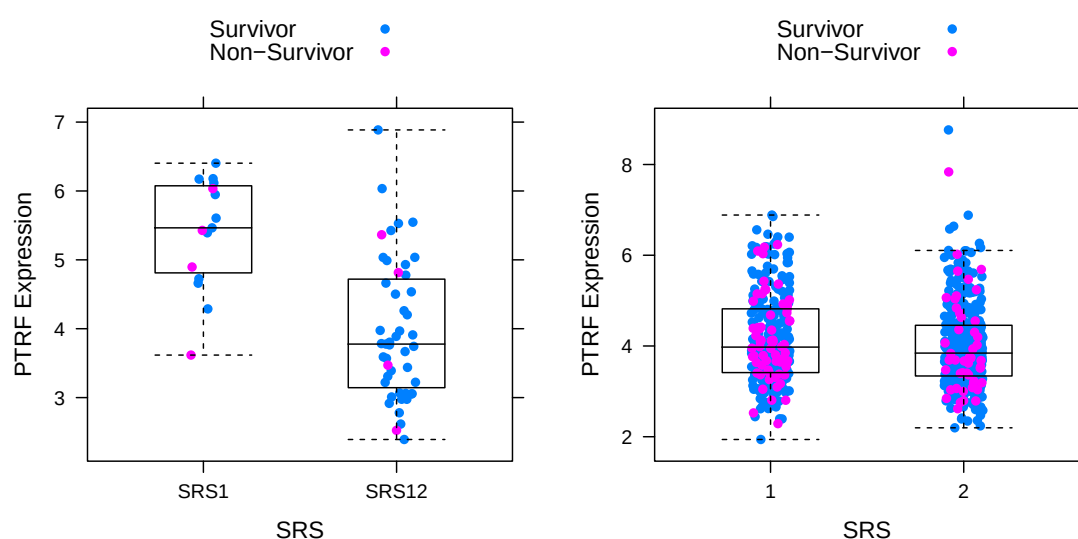


Figure 4.28: Expression of PTRF in the context of SRS dynamics.

PTRF is one of 19 genes that is significantly differentially expressed between SRS1 samples from patients who move into SRS2, and SRS1 samples from patients who remain in SRS1 (*left*). It is not however significantly differentially expressed between SRS1 and SRS2 (*right*).

Nineteen genes were significantly differentially expressed between the 15 patients remaining in SRS1 over time, and the 47 patients who moved from SRS1 to SRS2 (Fig. 4.27). These genes were enriched for the Gene Ontology Biological Process (GOBP) term “metabolic process” (FDR=0.00022) with the background of all detected genes that were tested for differential expression, and included *PTRF*, which encodes Polymerase I and transcript release factor. Knock down of this gene has been shown to alleviate sepsis in a mouse model of faecal peritonitis (caecal ligation and puncture) (Zheng et al. 2013). This gene is downregulated in patients before they move from SRS1 to SRS2 compared to patients who remain in SRS1, but is not differentially expressed between SRS1 and SRS2 or between survivors and non-survivors (Fig. 4.28). Another biologically interesting gene that was also downregulated in patients who move SRS groups was *DOK4*, an adaptor protein expressed in endothelial cells that has been implicated as a positive regulator of TNF α -mediated NF- κ B activation (Itoh et al. 2005).

4.2.7 Generalisation of the SRS model

Normalisation

In order to explore the potential translational utility of SRS in the clinic, a more generalised model is required. The first challenge is to transform the microarray gene expression data into quantitative values comparable to a point-of-care test in the clinical setting, *e.g.* quantitative PCR (qPCR) output for a small number of selected genes. This could be achieved by normalisation of the SRS model genes to a "housekeeper" gene, which should not show biological variation, but may fluctuate across samples according to systematic variation. Thus, it can correct for batch effects and the use of alternative gene expression assays. It is of course also vital that this gene is measured reliably and expressed constitutively.

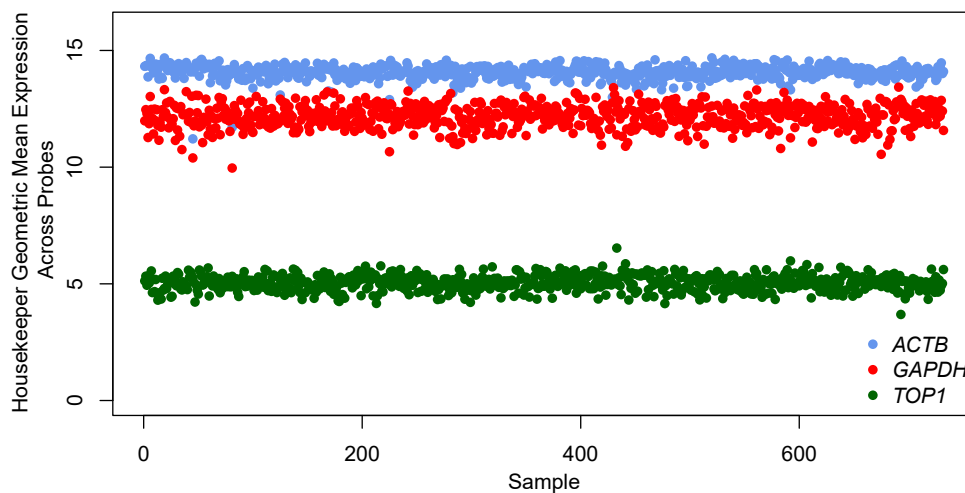


Figure 4.29: Expression of canonical housekeeper genes in the GAINs data.

ACTB, *GAPDH*, and *TOP1* are frequently used as housekeeper genes in qPCR experiments. The geometric mean expression of probes targeting these genes is plotted across the whole GAINs cohort, showing that only *ACTB* demonstrates high expression at a consistent level across all patients.

Housekeepers such as this are used in qPCR, with common choices being *GAPDH* and *ACTB*. Choosing one of these genes is appealing, as there would be increased confidence in a known housekeeper. Additionally, primer design and optimisation is unlikely to be an issue for a gene that is commonly measured

across a range of techniques. However, the well known housekeeper gene *GAPDH* had been found to be differentially expressed between survivors and non-survivors in the Radhakrishnan 2010 cohort (FDR=0.0064, Thorpe (2011)), and so is not appropriate in this context. *ACTB* and *TOP1* were found to be the most stably expressed commonly used housekeepers in the microarray data on the basis of the geometric mean of the probes targeting that gene (Fig. 4.29). The expression of *TOP1* was at the lower end of the detectable range, making *ACTB* a more suitable choice for normalisation.

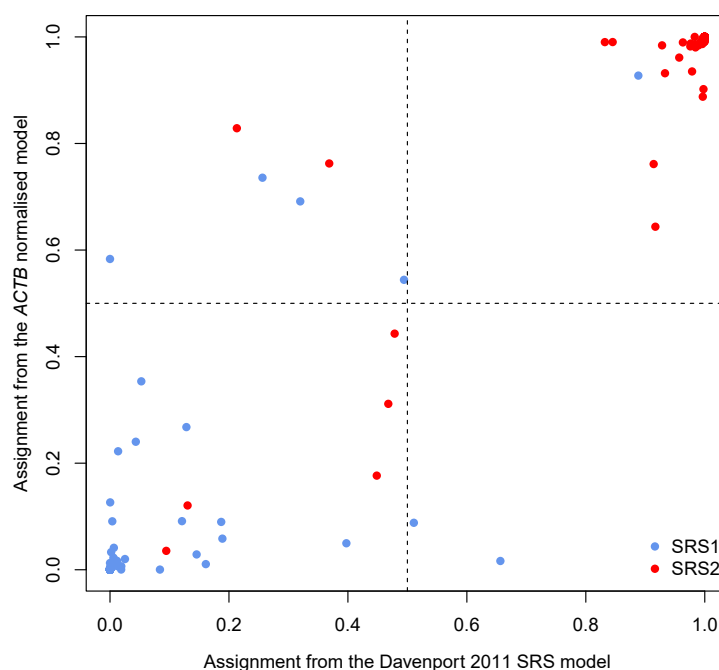


Figure 4.30: SRS model outputs with and without *ACTB* normalisation.

The SRS models produce a value between 0 and 1 for each sample, with samples below 0.5 being assigned to SRS1 and those above to SRS2. The output of the CAP SRS model are compared to the output of the *ACTB* normalisation model, and true SRS group membership indicated by point colour. The outputs of the two models are highly comparable, and both show little misclassification compared to the true group assignments for the Davenport 2011 cohort.

Normalising the CAP SRS model genes to *ACTB* in the Davenport 2011 cohort results in no loss of predictive power in leave-one-out validation. The group assignments from this method had only a 3.0% MR in comparison to the assignments from the original model, and 4.2% in comparison to the true group assignments (Fig. 4.30). Applying this model to the remaining CAP samples

(n=249) and the FP samples (n=218), the MR was 4.4% and 13.8% respectively.

Source of sepsis

Thus far, the CAP and FP datasets have been analysed independently. However, the SRS groupings within each cause of sepsis are highly comparable, and applying the CAP gene set to the FP cohort and *vice versa* still shows excellent prediction rates. Combining causes of sepsis would allow for greater analytical power, and reduce costs in clinical applications. Furthermore, there are a number of less common types of sepsis in addition to CAP and FP in which SRS groups might be seen. Therefore, a model for all-cause sepsis would be preferable.

In order to derive a more general gene set to assign SRS, I divided the combined GAINs dataset in half at random, such that training and test dataset of 257 CAP and 109 FP samples each were created. These two sets of patients had similar proportions of patients assigned to each SRS group (141 and 128 SRS1 samples respectively). The training data were reduced to probes mapping to known genes with at least moderate expression (>6.5) in 38% of samples (the proportion of SRS1 samples), and that were differentially expressed between SRS1 and SRS2 (n=1065 genes). GeneRave was used to select 9 genes (*RPS6KA5*, *DYRK2*, *CD3G*, *CCNB1IP1*, *ACER3*, *EI24*, *CSGALNACT2*, *B3GNT2*, and *LIPN*), which were then normalised to *ACTB* expression and used to fit a general model on the training data. Notably, *DYRK2* and *CCNB1IP1* were selected in the original CAP model, and *B3GNT2* in the SRS_FP model. This general model resolved SRS with an MR of 6.0% in the training data, and 9.0% in the test data.

Finally, for use in informing treatment decisions, a cut-off could be considered whereby SRS is only assigned for patients for whom the output of the SRS model places them confidently in one group or the other. For example, Fig. 4.31 shows the application of thresholds of 0.1 and 0.9 for SRS1 and SRS2 respectively, which would result in 144 of the 732 samples not being assigned to either group, thus

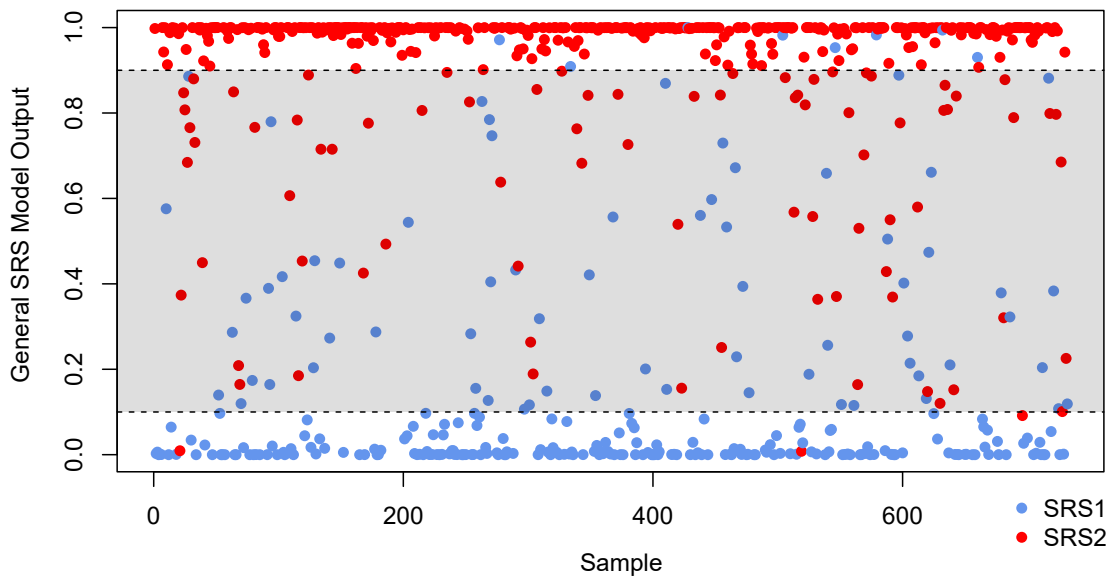


Figure 4.31: SRS model output for all-cause sepsis.

The SRS models produce a value between 0 and 1 for each sample, with samples below 0.5 being assigned to SRS1 and those above to SRS2. The output of the general SRS model is compared to the true SRS group membership, indicated by point colour. Little misclassification compared to the true group assignments was seen, but this could be reduced further by the application of thresholds (*grey shading*) to limit SRS assignment to patients with sufficiently divergent gene expression.

reducing the overall misclassification rate from 7.4% to 1.9%.

4.3 Discussion

A prevailing goal in sepsis research is to resolve heterogeneity in the host response, with a view to defining prognostic indicators or informing clinical management. In this chapter I have continued to use transcriptomic data to explore the dysfunctional host response in sepsis, and have extended previous exploratory work to a number of different datasets.

4.3.1 Survival prediction

The previously described gene expression signature for prediction of individual mortality performed poorly in the Burnham 2014 validation cohort. This is

possibly a result of overfitting, despite the use of a large training cohort, leave-one-out validation, and a relatively simple set of predictors. However, it may also reflect heterogeneity in the sepsis response and the variety of factors that contribute to mortality in individual patients. In the original analysis of the Davenport 2011 data, relatively few genes were differentially expressed between survivors and non-survivors. This might be due to a true lack of differences between patients who survive and those who do not; however, genes that are only differentially regulated in a subset of patients might not be identified. Moreover, the effect of time point described in the previous chapter could be playing a role, as time to death from the point of expression profiling varies across patients. Taking these factors into consideration, and given the complexity of the immune response, it seems likely that gene expression changes associated with mortality are not consistent across patients.

A similar approach has been adopted by previous investigators, most recently by Sweeney et al. 2016 who defined and compared several different predictive models using a large number of publicly available datasets. Performance varied across the models and datasets, but the most successful models performed reasonably well in test data. Such a model could be useful in the clinic to identify patients who require additional monitoring. Additionally, drug targets might be identified from genes differentially expressed between survivors and non-survivors. However, the clinical utility of survival prediction is relatively limited. In contrast, a gene signature that resolves the immune response state of a patient might present important opportunities for treatment in sepsis.

4.3.2 Transcriptomic sepsis response signatures

Patient subgroups previously identified in the Davenport 2011 cohort were validated in the prospectively recruited Burnham 2014 CAP cohort and in FP patients, replicating the association between an immunosuppressed phenotype

and increased mortality. Patient subgroups were robust to the choice of clustering methods (Davenport 2014), and their importance is further confirmed here by permutation analysis.

Whilst differing mortality rates highlight the biological relevance of the SRS groups, it is resolution of the host immune response state that might yield clinical benefit and was therefore explored in greater detail. Pathway analysis indicates features of relative immunosuppression, T cell exhaustion, and metabolic dysregulation occurring soon after the onset of sepsis, and contributing to early mortality. This supports recent research describing features of immunosuppression that manifest early in sepsis, contrasting with previous views of sepsis as solely a hyperinflammatory disease (Hotchkiss et al. 2013). Whilst late and post-recovery immunosuppression had previously been described, the focus of clinical trials on immunosuppressants was due to a lack of understanding of the prevalence of immunosuppression and the underlying immune response state of patients (Marshall 2014; Cohen et al. 2015).

The overlap of SRS1 with endotoxin tolerance gene signatures is consistent with previous work proposing that dysregulation of endotoxin tolerance contributes to immunosuppression and mortality in sepsis (Pena et al. 2014). Endotoxin challenge studies have demonstrated inter-individual variation in the development of tolerance related to genetic variation (Fairfax et al. 2011), which would be a plausible driver of heterogeneity in the sepsis response. These findings suggest that the degree of tolerance contributes to the differences observed between the SRS groups, although it is unlikely to be the only important factor involved.

Despite many differences in clinical covariates between the two groups, such data does not predict SRS successfully. However, more complex analyses considering changes in clinical measurements over time and more comprehensive pathogen profiling might reveal informative associations. As in the previous chapter,

the differential expression analyses were robust to differences in leukocyte cell count, yet cell type-specific expression data should be generated alongside immunological readouts to further understanding of the biology and clinical features of the SRS groups. Gene expression in sepsis is likely to be impacted by comorbidity, severity of the infection, stage of disease, and may also be modulated by genetic and epigenetic variation. Further work profiling these features in detail in the same patients, particularly epigenetic markers, could increase understanding of the interactions between the molecular and clinical aspects of the disease. Similarly, it is important to investigate if gene expression differences are propagated to the proteomic and functional level.

The subgroups defined in CAP and FP show considerable overlap, suggesting that these findings could apply to all causes of sepsis, and even in non-infectious SIRS. However, the small but important differences between CAP and FP described in the previous chapter necessitate careful consideration of the infectious source, and detailed exploration of more heterogeneous cohorts. The Molecular Diagnosis and Risk Stratification of Sepsis (MARS) consortium has recently described four molecular endotypes in all-cause sepsis, which were validated in the Davenport 2011 and Burnham 2014 cohorts (Scicluna et al. 2017). Specifically, four endotypes were identified using unsupervised consensus clustering and machine learning, one of which had a higher mortality rate and repression of immune-related genes. Applying the same methods to the GAINs data revealed a similar substructure, which showed considerable overlap with the SRS groupings (χ^2 $p=0.0005$).

No correlation was found between the SRS gene signatures and differentially expressed genes in the paediatric endotypes described by Wong and colleagues. Similarly, the Mars3 endotype was not reliably detected in paediatric sepsis, and no significant associations were found between mortality and endotype classification. This lack of overlap may reflect developmental differences

impacting pathophysiology and host immune function. This requires further validation in children of different ages, and investigation of subgroups in other cohorts of adult sepsis patients would increase confidence in the ubiquity of the SRS groups.

Importantly, serial samples allowed the exploration of SRS membership over time. A significant number of patients, particularly in FP, changed SRS group in the period for which gene expression data was available. The majority of these patients moved from SRS1 to SRS2, whilst persistence of the SRS1 gene expression profile was associated with a comparatively poor outcome. These findings suggest that serial determination of SRS group membership has potential as a biomarker to monitor the immune response state and the effects of immunomodulatory treatment. Intriguingly, a small number of genes including *PTRF* were found to be differentially expressed in SRS1 patients prior to movement to SRS2. *PTRF* has been implicated in sepsis development as a regulator of TLR4 signalling, and these findings highlight its potential as a biomarker or a therapeutic target (Zheng et al. 2013). Investigation of *PTRF* expression and protein levels in a larger number of serial samples is required to confirm this association, and further explore its implications.

4.3.3 Personalised medicine in sepsis

The data described in this thesis and in other cohorts demonstrate that features of both hyperinflammation and immunosuppression may be relevant in sepsis, and that the extent to which various aspects of the immune response are dysregulated varies across patients. This could explain why repeated trials of the same therapies return conflicting results, as the composition of the trial cohort will affect the efficacy of the treatment. Knowledge on the individual host response could therefore also be used to revisit completed clinical trials, refine patient selection for inclusion in new trials and other research such as GWAS, and enable

the development of novel targeted therapies in sepsis (Perner et al. 2017).

The SRS groups identified in both CAP and FP demonstrate that a small number of genes can be used to characterise the individual host response and resolve patients subgroups that differ in their immune response state. This presents an opportunity to develop these results into a tool for patient stratification and establish a precision medicine based approach to sepsis management, whereby appropriate therapies may be targeted to particular patient groups.

For example, SRS1 identifies patients with relative immunosuppression, who may be less able to eradicate infection and be more likely to succumb to hospital-acquired secondary infections. This leads to the hypothesis that they should be given therapy designed to boost their immune competence, such as immunomodulatory cytokines (IL7, IL15, GM-CSF, IFN γ) that have been suggested as potential treatments for sepsis (Hutchins et al. 2014). Conversely, patients in SRS2 show a relatively hyperinflammatory phenotype. There have been many clinical trials of immunosuppressive therapies in sepsis, under the hypothesis that pathogenesis is driven by an excessive inflammatory response (e.g. Opal et al. (2013) and Reinhart and Karzai (2001)). However, given the lower mortality rate in SRS2, these patients could be mounting a more appropriate response to infection. The link between the SRS profiles described in this chapter and the response to treatment has yet to be addressed in a comprehensive manner, so further work should focus on the interaction between sepsis response state and treatment efficacy.

The development of a test for use in the intensive care unit to determine the immune response state of the individual patient would therefore be extremely useful as a tool to help clinicians decide on an appropriate treatment and disease management strategy. Moreover, the data from serial sampling demonstrates that SRS are dynamic and individual patients show movement within and between groups over the course of disease progression. A set of representative predictor

genes could therefore also be used to follow the development of disease within a patient and be used to track the response to treatment. In this goal, however, it is important to consider that patients might be placed in one group although in actuality they are borderline. A point of care test would therefore ideally include some measure of confidence or strength of the signature, or as suggested above a threshold could be used, leaving some patients as unassigned.

Any point of care test should ideally be relevant to all-cause sepsis, and these findings suggest that the SRS subgroups are likely to be generally applicable. However, further exploration in additional datasets should explore the relevance in all-cause sepsis. In addition, the source-specific models described are derived from microarray data, and application to new data depends upon its normalisation against the original cohorts. This is adequate for the exploration and the validation of the SRS groups; for use in the hospital, a quantitative PCR approach would be more appropriate, and the test should only require measurements from the patient in question. Further validation and exploration of the sepsis response signatures could be achieved by application of the more general *ACTB*-normalised model to publicly available datasets from other platforms, such as Affymetrix microarrays or RNA-Seq. In parallel, the translation to qPCR could be tested by the measurement of the selected genes in GAIN patient samples already profiled using microarrays, allowing direct comparison of the outputs from the different methodologies. A similar approach is described in Pils et al. 2013 for the diagnosis of ovarian cancer, and preliminary qPCR results from LPS-stimulated monocytes suggest that the resulting values are comparable to the adjusted microarray gene expression data. Lastly, greater accuracy and reliability could potentially be achieved by the use of two housekeeper genes, or through the use of expression ratios for discriminatory genes.

Subgroups of therapeutic relevance have been identified in the context of sepsis

by a number of different groups. Each of these has used a different clustering strategy or data type, and the groupings described are not mutually exclusive in terms of clinical utility. As mentioned above, Wong *et al.* identified subgroups in paediatric sepsis with differing responses to corticosteroids. More recently, also in paediatric sepsis, Carcillo *et al.* 2017 used a combination of lab assays in a knowledge-driven approach to define inflammation-based subclasses that were associated with multi-organ failure. Knox *et al.* 2015 used self-organising maps to resolve four clusters of multi-organ failure in adult sepsis patients, and the NHLBI ARDS Network identified two subgroups within ARDS patients through latent class analysis of large numbers of clinical variables. The latter displayed a differential response to treatments including fluid management strategies (Calfee *et al.* 2014; Famous *et al.* 2017). The increasing number of such studies illustrates the movement towards using high-dimensional data to improve our understanding of disease, and the greater willingness to adopt personalised medicine in the ICU setting.

4.4 Conclusions

Sepsis is a complex disease with a persistently high mortality and a lack of specific therapies. In this chapter, I have explored sepsis heterogeneity using a transcriptomic approach in patients with FP and CAP, and demonstrated that this approach can improve phenotypic classification of sepsis patients by defining subgroups with differing immune response states. These SRS groups have potentially important therapeutic implications, specifically highlighting the opportunity for a precision medicine approach in sepsis.

THE PLASMA PROTEOME IN SEPSIS DUE TO FAECAL PERITONITIS

This chapter aims to investigate the sepsis response at the level of the proteome, comparing protein abundance across patient subgroups and integrating mass spectrometry data with transcriptomic data described in previous chapters

5.1	Introduction	159
5.2	Results	163
5.3	Discussion	187
5.4	Conclusions	193

5.1 Introduction

As described in the preceding chapters, transcriptomic profiling of sepsis patients is an extremely useful tool for investigating the underlying dysregulated host response, and can resolve clinically relevant subgroups of patients (Davenport et al. 2016; Burnham et al. 2017). However, whether or not signals detected at the RNA level are also present following translation is unclear.

This is a vital question when considering the implications of the transcriptomic findings. Gene expression data are generally interpreted with the assumption that transcript levels correspond well with protein levels, and so reflect the functional state of the cells being assayed. Additionally, the majority of drugs target a protein or proteins, and so knowledge of the proteome is important for

identifying putative therapeutic interventions. This being said, even if there is poor correspondence between transcriptomic and proteomic data, specific gene expression signatures might still be useful for diagnosis, prognostication, and patient stratification.

This chapter describes a pilot mass spectrometry experiment designed to profile the proteome of a cohort of sepsis patients, with a view to identifying proteins that are differentially abundant between patient groups, and to comparing this dataset with transcriptomic profiles from the same patients.

5.1.1 Mass spectrometry

Mass spectrometry is a high throughput technique used to analyse the mass of input molecules (Glish and Vachet 2003). The sample is ionised, frequently by electron bombardment, and the resulting ions are accelerated through an electric or magnetic field. The ions will be deflected according to their mass-to-charge ratio (m/z), and then detected using an instrument such as an electron multiplier. Mass spectrometry can be used for protein or metabolite identification and quantification on a single purified species or a complex mixture.

Depending on the experimental question, different types of mass spectrometer are used to generate different types of data. For example, a simple spectrometer will measure the mass of each intact peptide in the sample, generating a spectrum showing the relative abundance of detected ions against their m/z . A tandem mass spectrometry experiment (MS/MS), as described in this chapter, will additionally generate fragment ion spectra for each peptide by breaking the peptide bonds and measuring the masses of the resulting fragments. Each amino acid bond is typically cleaved once, meaning that a variety of child ions are produced. The resulting fragmentation mass spectra can be used as a means of identifying all the proteins present in a sample, by comparison to known masses

and fragmentation patterns (McLafferty 1981).

Typically, a human proteomics analysis will take as input a mixture of thousands of proteins extracted in bulk from a tissue sample, with each of these proteins being present in tens to hundred of thousands of copies (Kim et al. 2014). Given the complexity of this input, several steps are required for informative data to be generated. Firstly, direct MS analysis has an inherent size limit, meaning that proteins must be broken down into subpeptides prior to measurement. Additionally, before MS/MS analysis the sample is separated by chromatography to reduce complexity, allowing spectra from individual peptides to be produced. In the first round of MS, the masses of 20,000 intact peptides are measured per second, from which five distinct species are selected for fragmentation and detection with the second spectrometer.

5.1.2 Protein identification and quantification

Spectrum identification is made difficult by noise in the observed profiles, for example due to missed cleavages, unusual fragmentation events, and contaminants. Statistical methods must therefore be used to assess confidence in peptide identification.

The most common approach is to search a database of known spectra, for which multiple algorithms have been developed. Generally, for each observed spectrum S with a known precursor mass m , all entries with a mass in the range $[m - \delta, m + \delta]$ are extracted from the peptide database P , where δ represents the chosen mass tolerance. This gives a set of candidate peptides, each of which are compared to S using a score function (Zhang et al. 2012). Alternatively, in the case of *de novo* identification, an entire universe of possible amino acid sequences may be generated for matching. The highest scoring match is identified, and a score threshold applied to give high confidence peptide identifications.

Peptide abundance is calculated from the area under the peak in the mass spectrum for the intact peptide. Finally, these peptide measures are combined to give protein level data. Degenerate peptides occurring in multiple proteins are excluded at this stage (Fischer et al. 2012).

5.1.3 Aims

Plasma samples taken over the first five days following ICU admission are available for the majority of GAINs patients, from which protein can be extracted for mass spectrometry analysis. Unfortunately no samples were available to profile proteins present in the leukocyte cell population studied in the transcriptomics chapters. It is therefore expected that only secreted proteins will be detected in this experiment, and so correlation between the two types of data might be limited. However, analysis of the proteome will complement and extend the findings from the transcriptomic analysis, providing an additional layer of information for a subset of patients.

This pilot experiment aimed to test the following hypotheses:

1. that specific plasma proteins differ in abundance between patients with sepsis due to faecal peritonitis according to
 - (a) sepsis response signature subgroups
 - (b) survival
 - (c) time from disease onset
2. leukocyte gene expression is correlated with plasma protein abundance

5.2 Results

5.2.1 Cohort description

The instrument and pre-processing necessitated by this sample type limited the pilot experiment to 60 samples from 55 patients (Table 5.1). In order that samples might be placed on a time line relative to the estimated time of disease onset for longitudinal analysis, this cohort was composed of FP patients only. Samples were selected such that there were equal numbers from each of the SRS groups, with an even spread over the first six days following disease onset. Each patient also had gene expression from the same time point (Radhakrishnan 2010 and Burnham 2014 cohorts), and where possible genotyping data available to enable integration with the proteomic results.

	Mass spectrometry samples
Patients	55
Samples (D1, D3, D5)	60 (32, 19, 9)
Mean Age ($\pm$SD)	68.7 ($\pm$ 14.4)
% Male	56.4%
Mortality rate	23.6%

Table 5.1: Mass spectrometry cohort demographics

Demographic and clinical information are summarised for the whole proteomics cohort. Serial samples were taken on the first (D1), third (D3), and fifth (D5) day of ICU admission, and mortality status at 28 days following admission recorded.

5.2.2 Data generation

Samples were processed as described in chapter 2. Briefly, samples were delipidated and the top 12 most abundant proteins depleted. This was followed by protein precipitation, digestion, desalting, and drying. Mass spectra were then generated using nano-liquid chromatography electrospray ionisation tandem mass spectrometry.

5.2.3 Protein identification

All chromatograms were aligned based on retention score and mass to charge ratio (m/z) using data for a pool of all samples as a reference. The majority had an alignment score $>70\%$; five samples were manually realigned and one of these excluded from the analysis on the basis of poor alignment, leaving 59 samples.

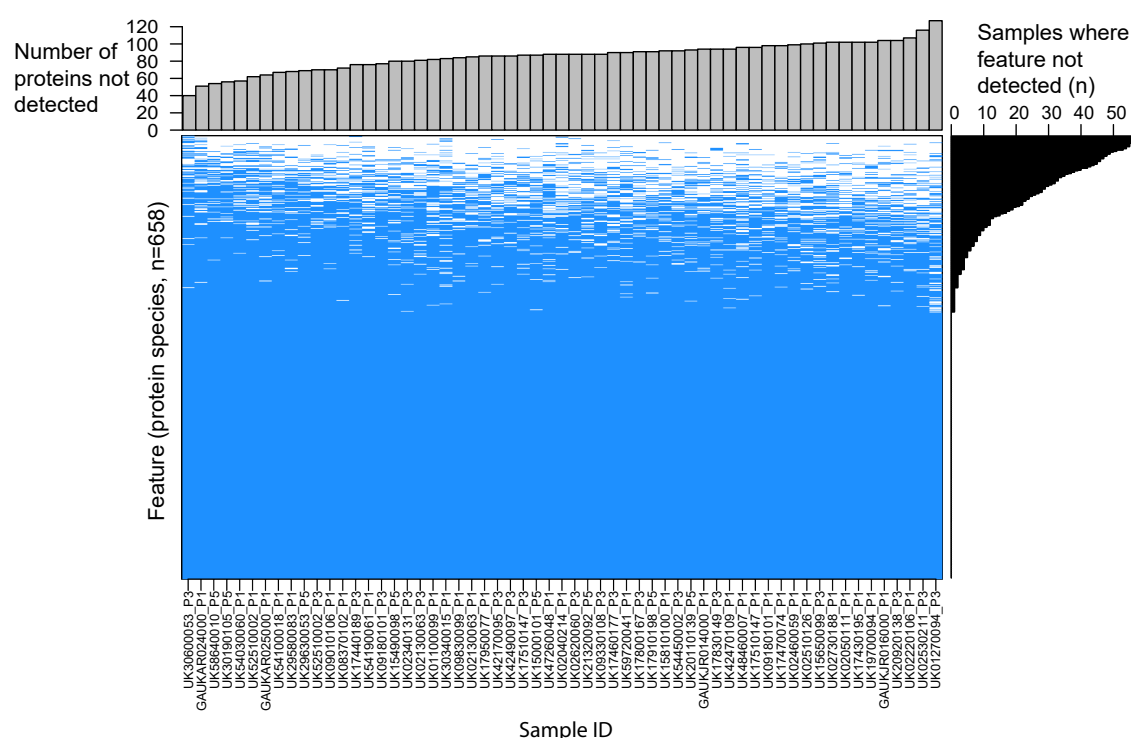


Figure 5.1: Heatmap of protein detection across samples.

A heatmap showing feature detection across the 59 plasma samples from 55 patients with sepsis due to FP. Each row in this heatmap represents a protein species, and each column a sample. Cells are coloured blue if the protein was detected with confidence in that sample, and white if not. The barplots at the top and right indicate the number of proteins not detected in a sample, and the number of samples in which a feature was not detected, respectively.

The spectra were matched to the Uniprot Reference database (The UniProt Consortium 2017) using the PEAKS search algorithm, with 560,911 individual queries performed (Zhang et al. 2012). The search parameters were as follows: the peptide and fragment mass tolerances were set as ± 10 ppm and ± 0.05 Da respectively, with a maximum of one missed cleavage and peptide charges of 2+, 3+, or 4+ permitted. False discovery rate was controlled by using a decoy search;

briefly, the database sequences were reversed and queried with the observed spectra to quantify the number of false positive matches. This result was then used to set a threshold for confidence in peptide identifications, which were filtered using an FDR cut-off of 1% and an ion score cut-off of 20.

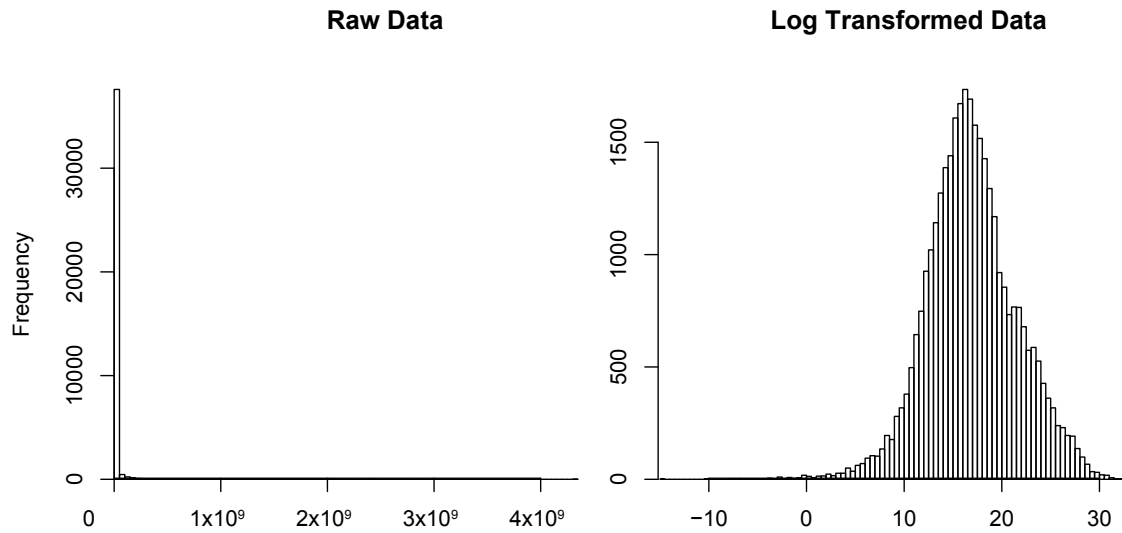
Non-degenerate peptides, *i.e* those contained within a single protein, were used for protein quantification. Across all samples, 658 proteins were identified by at least one non-degenerate peptide. Approximately two thirds of these features were detected in every sample, with the remaining 250 proteins missing in at least one sample (Fig. 5.1).

5.2.4 Normalisation and Quality Control

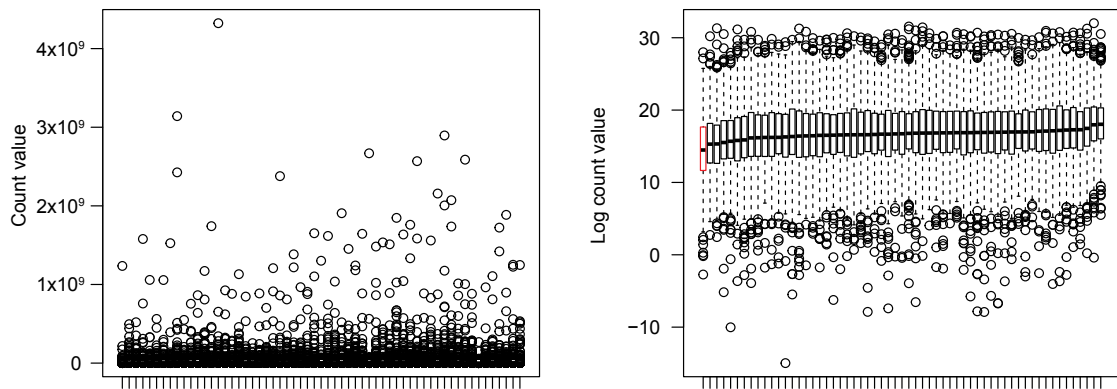
The raw data, which are essentially in the form of count data, were first $\log_2$ transformed for use in downstream analysis (Fig. 5.2). Examination of the log transformed data indicates that a small number of samples might have performed less well; for example, sample UK42170095_P3 was noted as having a lower median value than the rest of the cohort, although a range was seen (Fig. C.1).

During sample processing, known high-abundance proteins were depleted using the “Top12” kit (Thermo). Boxplots of these proteins across all samples show that this step was effective, although these proteins were still relatively abundant (Fig. 5.3). This could have resulted in lower abundance proteins not being detected due to the “Top12” proteins swamping the signal. To check for samples where this may have been an issue, the “Top12” variance was plotted against the median value for each sample (Fig. C.2); whilst two samples were slightly removed from the rest of the data, they were not considered outliers.

The data were then normalised by median centering and median absolute deviation (MAD) scaling. In order to account for differences in the Top12 protein variance across samples, the top 10% most variable features were not



(a) Histograms of raw and log transformed data



(b) Boxplots of raw and log transformed data

Figure 5.2: Effect of $\log_2$ transformation of raw count data.

The raw data are shown by a histogram (*top*) and by boxplots by sample (*bottom*). These data are similar to count data from sequencing experiments, and are not normally distributed, which impacts subsequent analysis. They are therefore $\log_2$ transformed, resulting in the distribution shown in the right hand histogram and boxplots. The boxplots are ordered by sample median value, with UK42170095_P3 coloured in red.

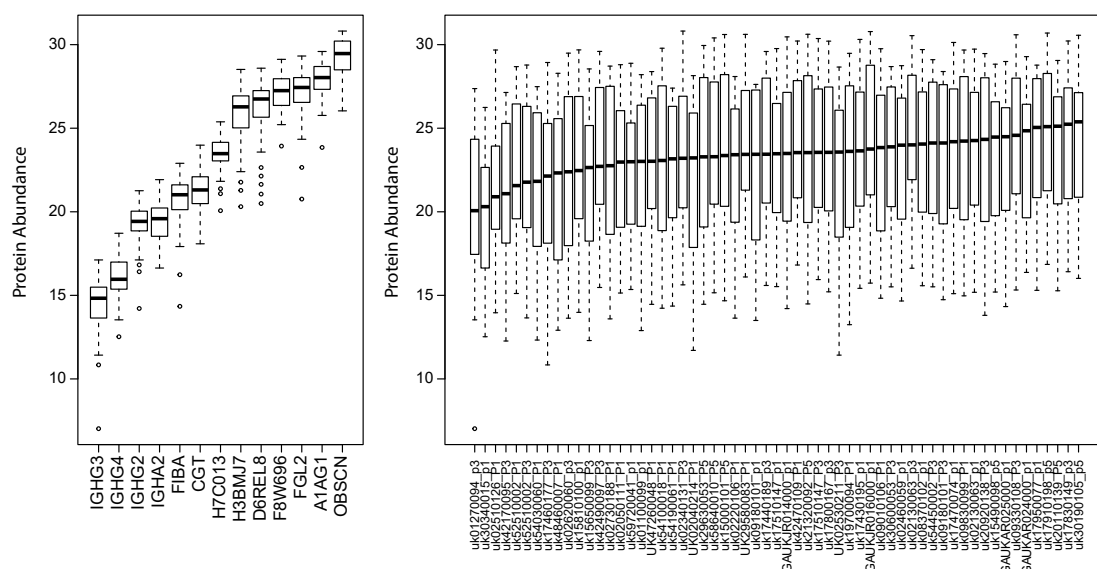


Figure 5.3: Sample quality control by variation in depleted proteins.

The effectiveness of the Top12 depletion was compared across samples by plotting the relative abundance of each protein as boxplots. (a) shows the range for each protein, and (b) shows the abundance of the Top12 proteins in each sample. This showed that there was some variation in the efficacy of the depletion; however, the Top12 proteins were within the range of other detected proteins.

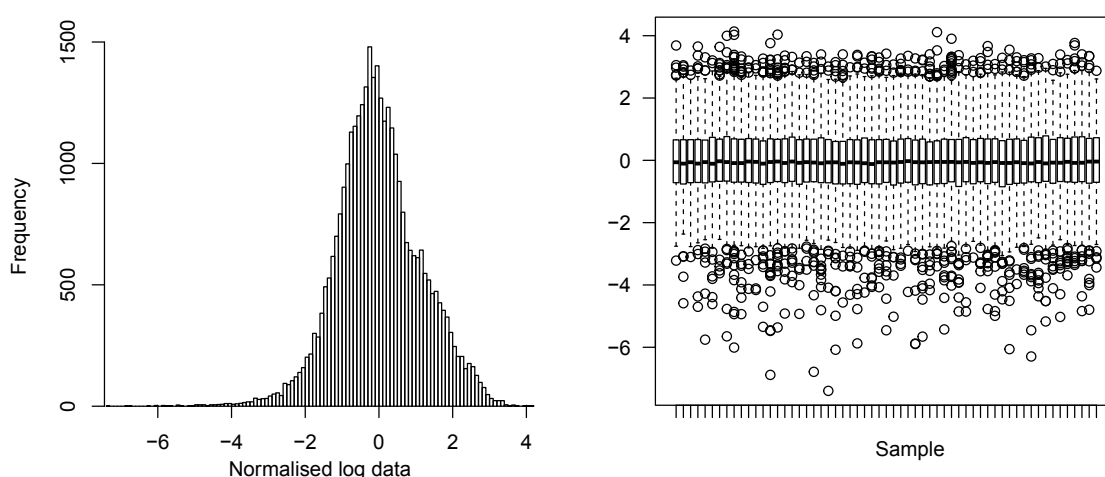


Figure 5.4: Histogram and boxplot of the normalised data.

The median centered and median absolute deviation scaled data are plotted as a histogram and boxplots to demonstrate the effect of normalisation.

used in the calculation of sample median and MAD (Philip Charles, personal communication). The resulting log normalised data were normally distributed, and centered around the sample median (Fig. 5.4). This approach assumes that the abundance of most proteins is constant across samples, so the intensity distributions are very similar across samples.

5.2.5 Missingness and Imputation

One important difference between mass spectrometry and microarray data is that the proteomic data may contain varying amounts of missing values, *i.e.* some proteins are detected only in a subset of samples. Whilst gene transcripts may also be differentially present across samples, this is reflected in the resulting microarray data through detection values, which may be used to filter out unreliable features. In the proteomics data described here, an alternative method must be used to permit further analysis of data containing missing values.

These data can be missing at random (MAR), or missingness may be biologically meaningful. If a protein is either not present in the sample, or present but in quantities below the detection limit of the instrument, then the missing value is non-random (MNAR). However, some missing values may represent proteins that have just not been detected by chance, despite being present at sufficient abundance. The methods chosen to deal with missingness are influenced by which of these two scenarios apply.

Figure 5.1 illustrates the pattern of missingness in the GAINs dataset described in this chapter. Of the 658 proteins identified, 395 were detected in all samples; 2 features were only detected in 1 sample, with a range of missingness between these extremes. On a sample level, between 40 and 128 proteins are not detected in each sample. Sample UK01270094_P3 has the most missing data, possibly relating to its outlying values in the quality control described above.

filtered out. Missing values for each remaining feature are then filled in by imputation. This imputation involves random draws from a distribution with mean and standard deviations estimated from the data available for that feature, by SRS group. Following this, the final dataset comprised 620 proteins across 59 samples (Fig. 5.6).

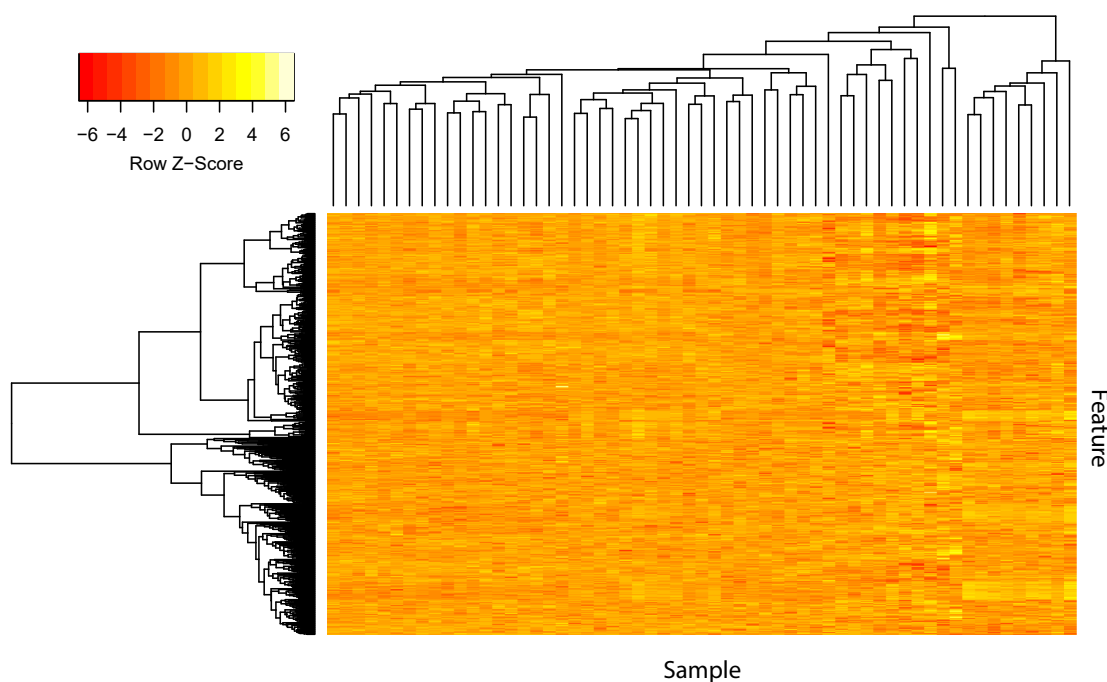


Figure 5.6: Heatmap of the final dataset.

This heatmap gives an overview of the final proteomics dataset. Each column represents a sample and each row a protein feature, with the colouring indicating its relative abundance across samples.

5.2.6 The sepsis plasma proteome: subcellular localisation

Of the 620 proteins detected reliably in plasma samples from sepsis patients, 544 (88%) were found in the Plasma Proteome Database (PPD), which is one of the largest information sources on proteins reported in human plasma and serum (Nanjappa et al. 2013). Additionally, the majority of the proteins detected were annotated as being found in the extracellular region (Fig. 5.7); enrichment analysis with the R package XGR found that this was the most enriched Cellular Component term with an FDR of 4.5×10^{-113} . Moreover, a number of proteins

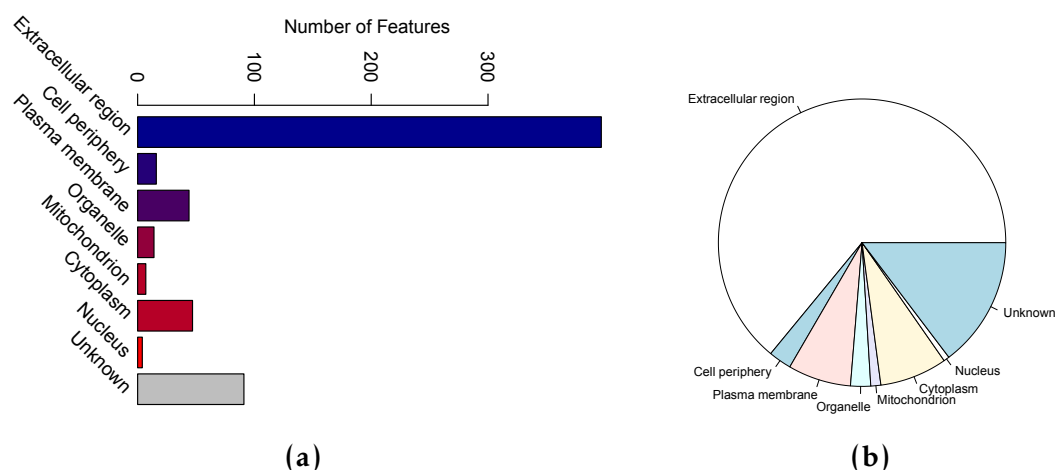


Figure 5.7: Number of detected proteins by subcellular location.

Proteins were placed in bins based on their Gene Ontology Cellular Compartment annotations. Each protein is only counted once, based on the most likely annotation for this sample type (ordered top to bottom on the barplot). For example, a protein annotated with “Extracellular region” and “Cytoplasm” is placed in the “Extracellular region” category. The total number of features in each category are shown as (a) a barplot; and (b) a pie chart.

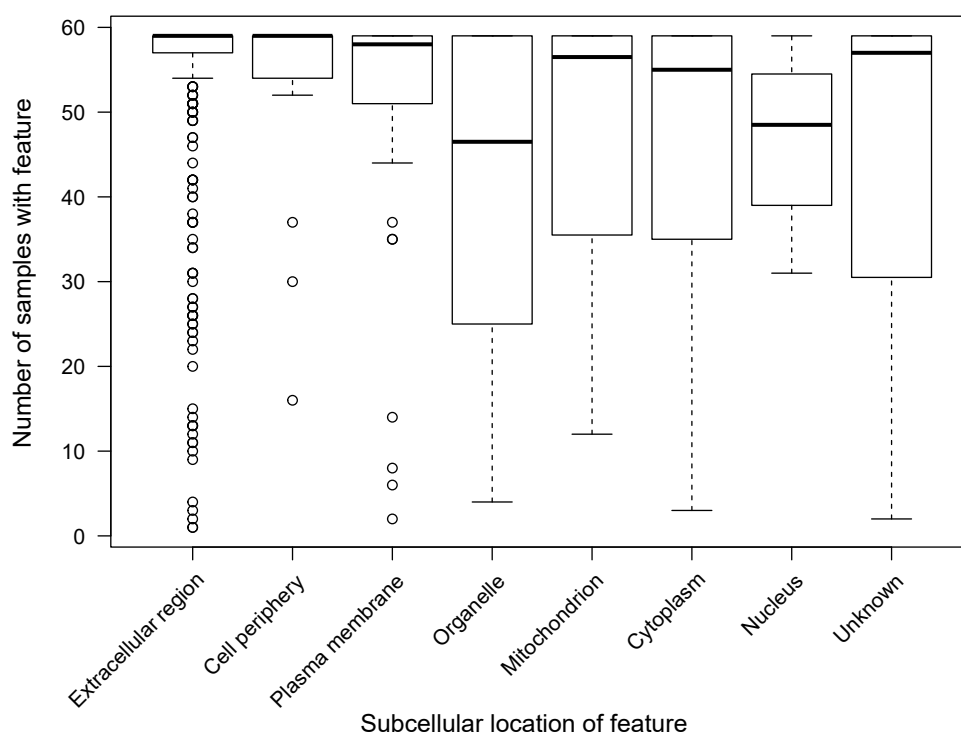


Figure 5.8: Detection frequency according to subcellular location.

Using the same binning method as described in Fig. 5.7, the number of samples in which a feature is detected is plotted against the subcellular location. Features located within the cell only show a lower detection rate than those localising to the cell membrane and extracellular region.

without this Gene Ontology (GO) annotation had been previously reported in plasma and were included in the PPD.

Interestingly, extracellular proteins were also more likely to be detected across all samples, with features described as being found in membrane bound organelles showing the lowest median detection rate (Fig. 5.8). Although abundance is not directly comparable between features, the value is semi-quantitative, and plotting this against subcellular localisation again shows the dominance of extracellular proteins in these samples (Fig. 5.9).

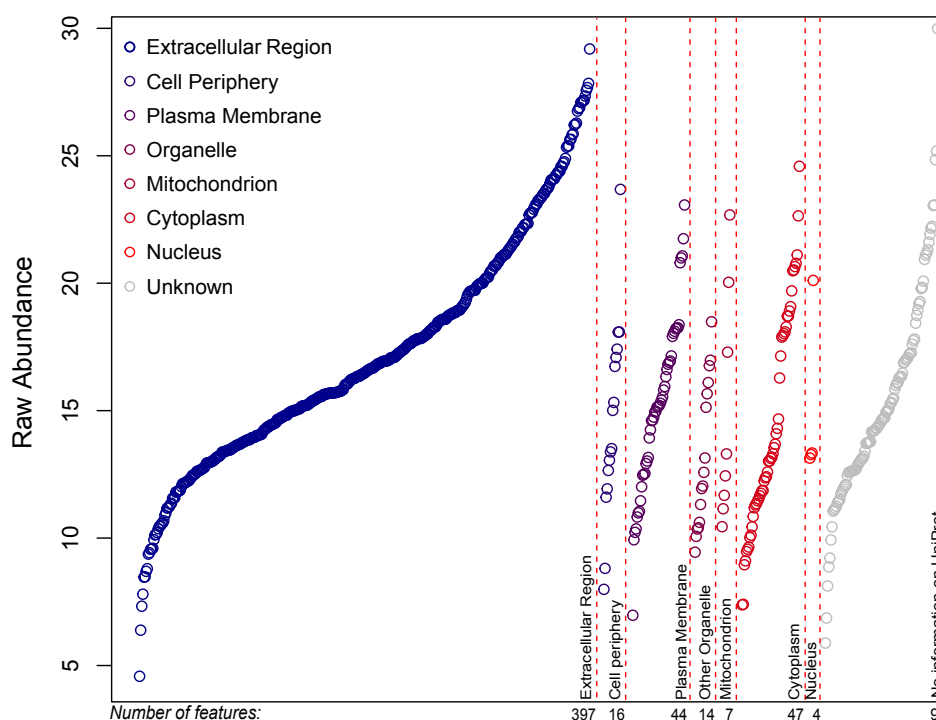


Figure 5.9: Raw abundance according to subcellular location.

Using the same binning method as described in Fig. 5.7, the log raw count for each feature (a semi-quantitative measure) is plotted according to its subcellular location. Features are ordered by their abundance within each location category.

5.2.7 The sepsis plasma proteome: enrichment analysis

Enrichment analysis was carried out using XGR in order to explore the biological functions represented by the proteins detected in sepsis plasma, compared to the background of all annotated proteins. Two different sources of annotations were

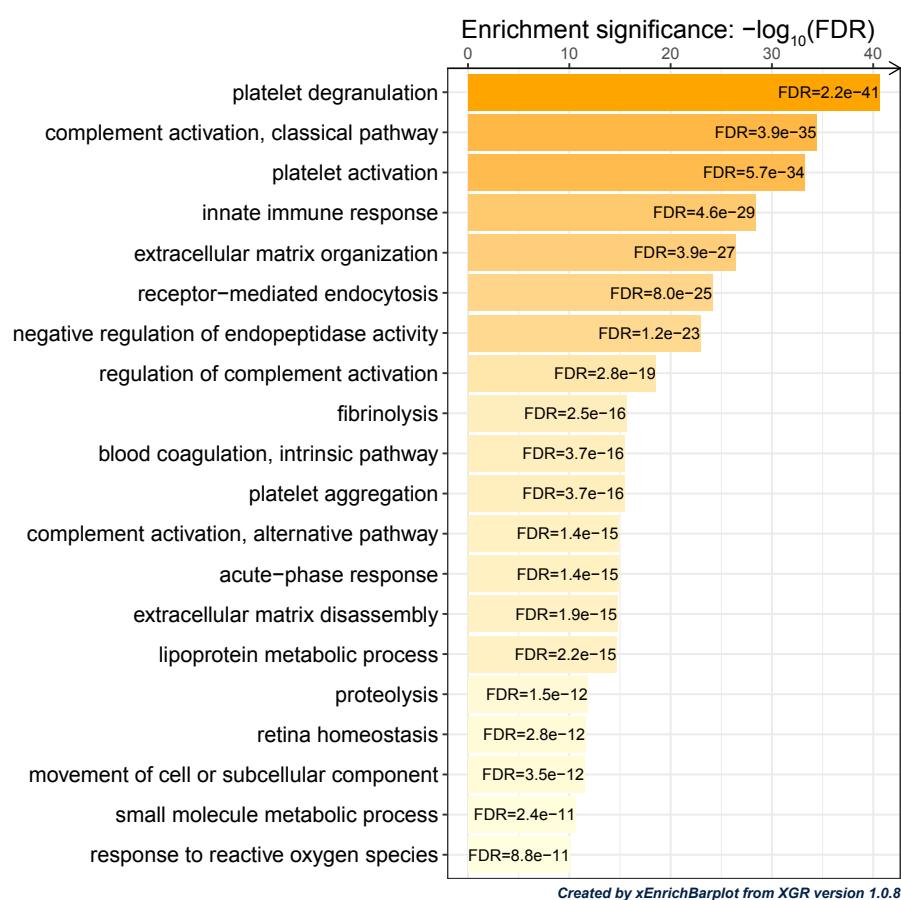


Figure 5.10: Pathway enrichment analysis using the Gene Ontology Biological Process ontology.

All proteins detected in the sepsis plasma mass spectrometry experiment were tested for enrichment of Gene Ontology Biological Process annotation terms. The hierarchical structure of the ontology was taken into account in the analysis, and the top 20 most significant results shown in the barplot.

used to get a general picture of the pathways in which the proteins identified were involved; the GO Biological Process ontology (GOBP), and the Molecular Signatures Database (Fig. 5.10, 5.11).

This revealed strong signals relating to the complement and coagulation cascades ($\text{FDR}=2.0 \times 10^{-48}$), and general functions expected for plasma samples e.g. haemostasis, extracellular matrix proteins. This is consistent with previous descriptions of the sepsis plasma proteome (Shen et al. 2006; Lissauer et al. 2007; Kiehntopf et al. 2011). Notably, platelet degranulation was a highly enriched GO Biological Process ($\text{FDR } 2.2 \times 10^{-41}$). On activation or as a result of cell lysis, platelets can release a number of proteins from storage granules such

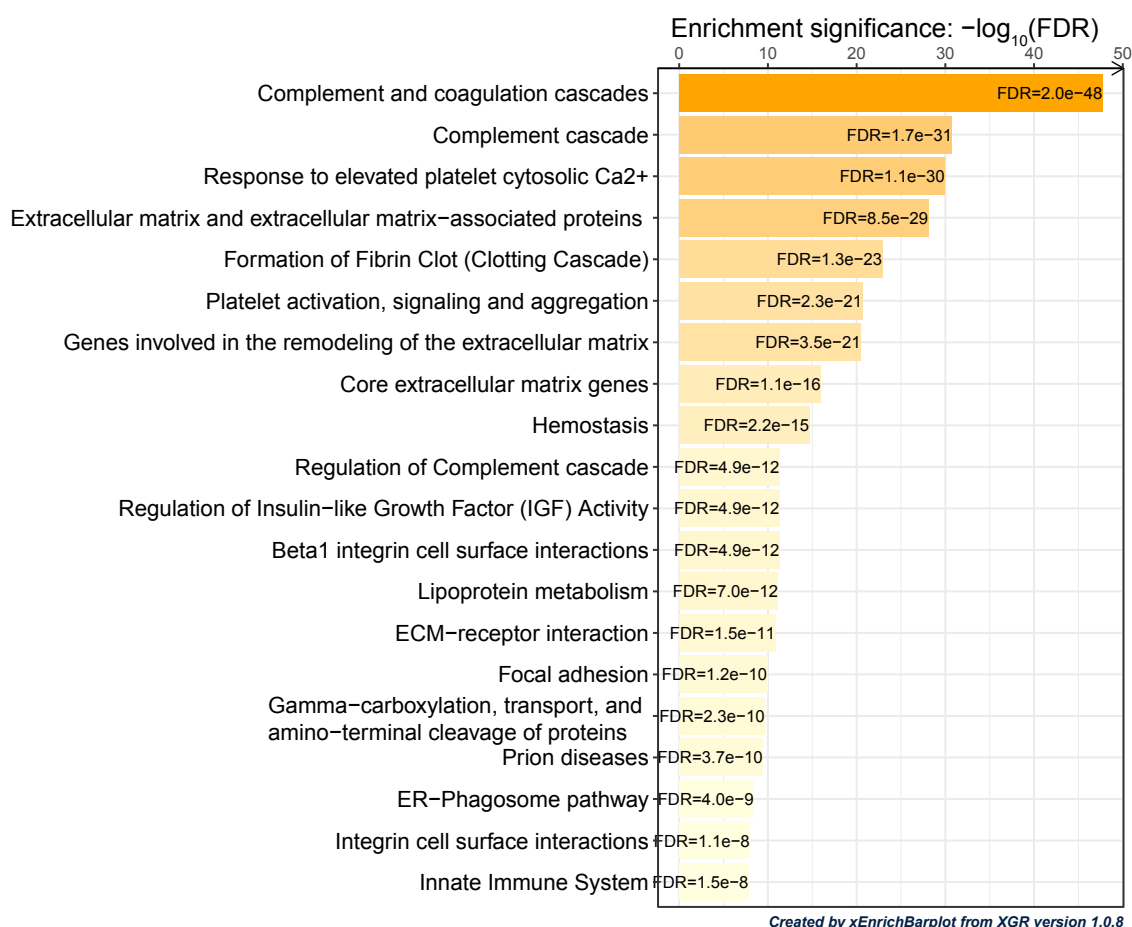


Figure 5.11: Pathway enrichment analysis using the Molecular Signatures Database. All proteins detected in the sepsis plasma mass spectrometry experiment were tested for enrichment of Molecular Signature Database annotation terms. Redundant terms were removed and the top 20 most significant results shown in the barplot.

as lysosomes. This could explain the detection of intracellular proteins in the plasma.

5.2.8 Comparison of samples

Principal component analysis was carried out based upon all features, in order to investigate variation between samples. Firstly, no separation of samples according to the sex of the patients was seen in a plot of the first two principal components of the proteomic data (Fig. 5.12). SRS membership was then plotted, and although SRS1 and 2 samples did show overlap in this projection, the centroids of the groups were separated.

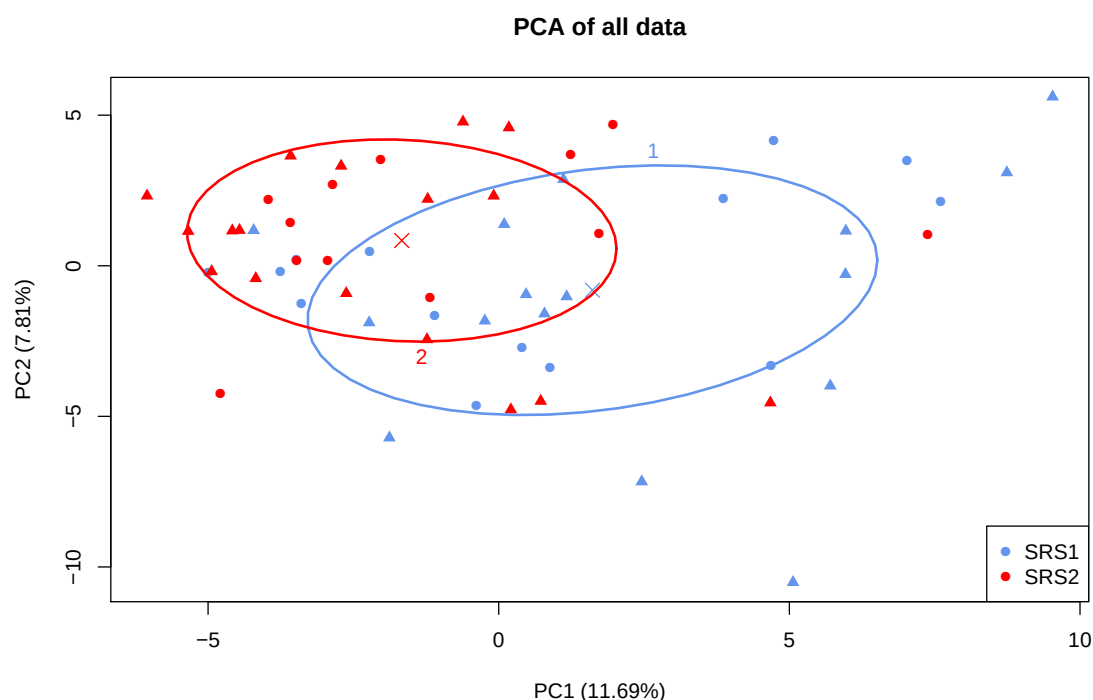
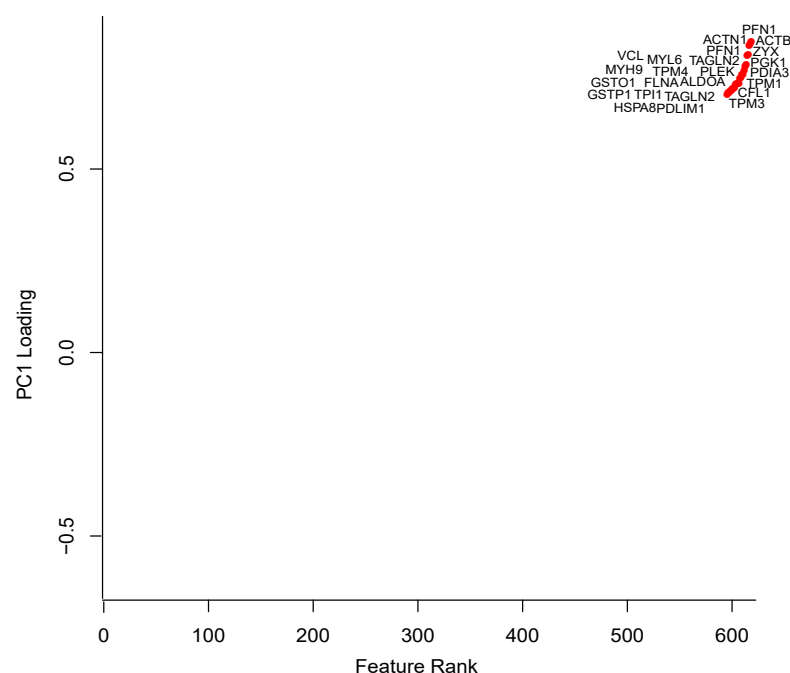


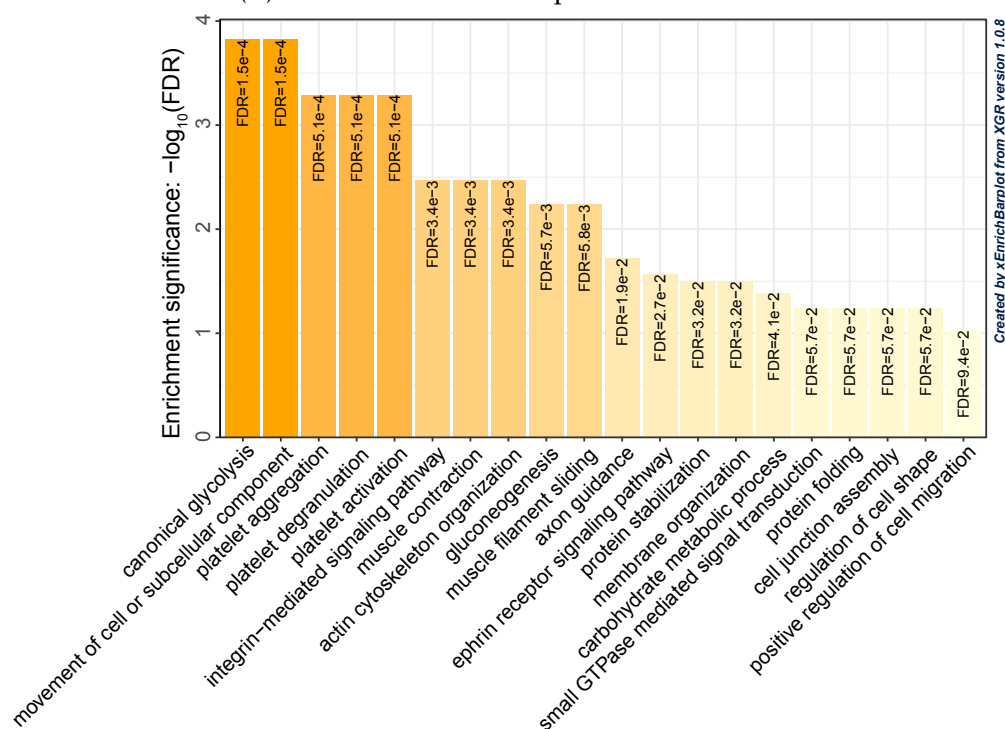
Figure 5.12: Plot of the first two principal components of the mass spectrometry data. Each point represents a sample, coloured according to SRS group assignment. Male patients are shown by triangles, and females by circles. Data ellipses are also drawn for each SRS group, and the centroids of these ellipses represented by crosses.

To explore this clustering further, the features contributing most to PC1 were identified. Here, the protein loadings are the correlations between the original variables (abundance of each protein) and the scaled principal components. Therefore, the proteins with the largest loadings are the most important to that component and thereby the variance described. The proteins with a correlation coefficient > 0.7 include several involved in glycolysis, e.g. PGK1, ALDOA, and TPI1, which echoes the association between SRS1 and the metabolic switch to glycolysis. All proteins with an absolute correlation loading > 0.5 ($n=94$) were then used for pathway enrichment analysis in XGR with the GO Biological Process ontology and the background of all detected proteins, and the most significantly enriched term was canonical glycolysis (Fig. 5.13).

Similarly, the correlation between all pairs of samples was calculated and visualised as a heatmap based upon the strength and significance of the



(a) Correlation between proteins and PC1



(b) Pathway enrichment analysis for proteins with a high loading in PC1

Figure 5.13: Correlation between features and the first principal component.

(a) The loadings for PC1 are plotted as an indication of the importance of each protein in overall variance across samples. Those with an absolute correlation coefficient >0.7 are labelled. **(b)** All proteins with an absolute correlation loading > 0.5 were used for pathway enrichment analysis with XGR and the Gene Ontology Biological Process ontology. The hierarchical structure of the ontology was taken into account in the analysis, and the top 20 most significantly enriched pathways shown.

correlation (Fig. 5.14). This was repeated using just the top 10% most variable proteins. Again, this shows that the majority of samples cluster according to their SRS group.

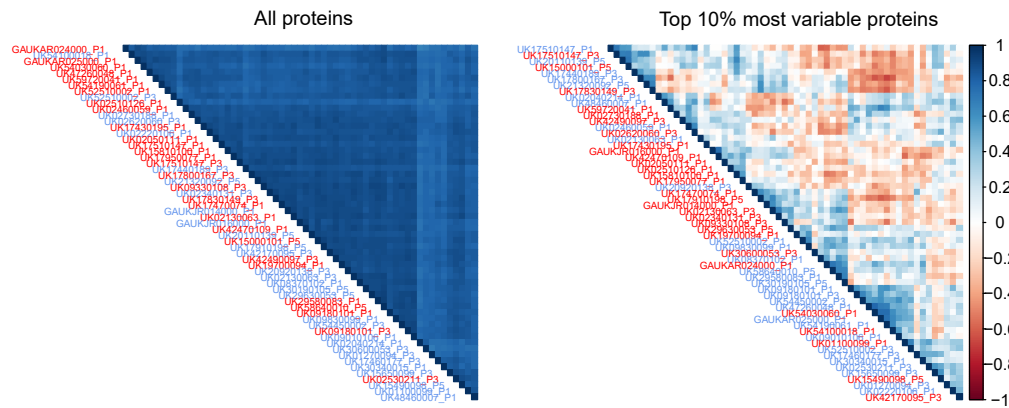


Figure 5.14: Visualisation of correlations between samples.

The correlation between all pairs of samples was calculated using all proteins (*left*) and the top 10% most variable proteins (*right*). For each pair of samples, the colour indicates the strength of correlation in terms of Pearson's correlation coefficient. Samples are ordered by hierarchical clustering by Ward's method. SRS1 patients are labelled in *blue*, and SRS2 in *red*.

5.2.9 Unsupervised clustering

Using the same methodology as in the transcriptomics data, unsupervised clustering was carried out on the mass spectrometry data set. Briefly, the R package FactoMineR was used to carry out agglomerative hierarchical clustering, which indicated two distinct groups were present in the data, followed by *k*-means clustering to consolidate group assignment (Fig. 5.15).

This clustering was compared to the SRS groups assigned by the transcriptomic data for the same patient and time point. This found that two thirds of samples (37/59) were grouped in the same way. Like SRS, these clusters showed a significant difference in early mortality ($p=0.00316$; Fig. 5.16)

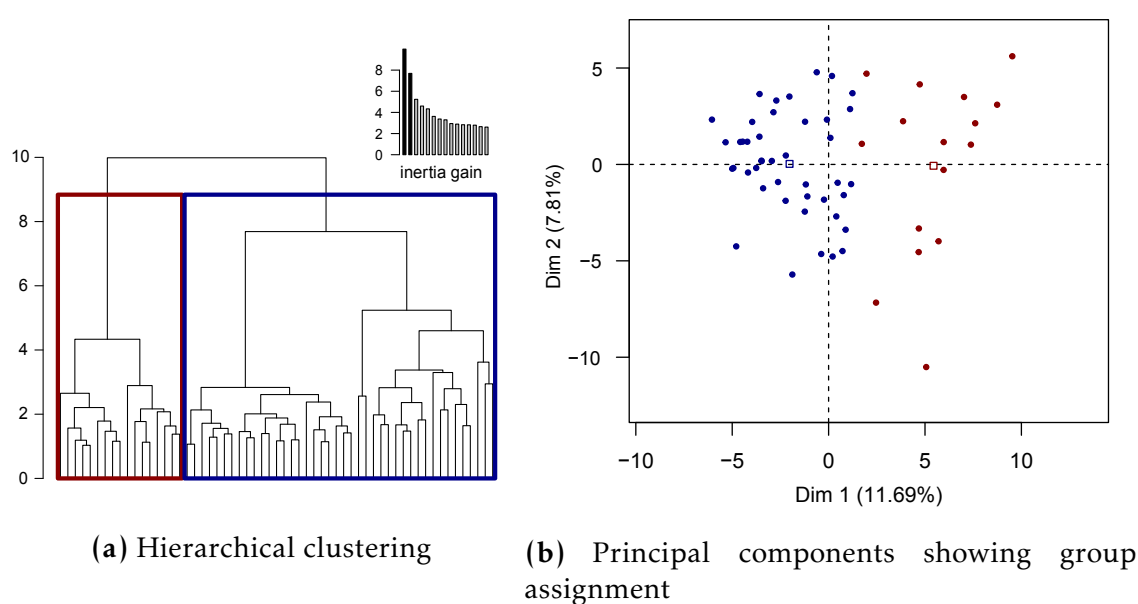


Figure 5.15: Unsupervised clustering of proteomics data.

Unsupervised clustering analysis was applied to the proteomics data using FactoMineR. **(a)** Two subgroups were identified in the data through hierarchical clustering and group membership was consolidated using *k*-means clustering. **(b)** The first two principal components of the data are plotted, with the amount of variance explained noted and the point colour indicating group assignment.

5.2.10 Differential protein abundance between SRS groups

In the principal component analysis described above, plotting PCs 1 and 2 showed some separation of SRS1 and 2, suggesting that a relatively large number of proteins differed in abundance between the two groups. Differential abundance analysis was carried out for SRS1 vs SRS2 by fitting linear models using limma, accounting for the sampling day and sex, and using only the first available sample for each patient. In addition, similar models were fitted to look at differential abundance by sex, and between survivors and non-survivors, again accounting for day of sampling and, in the latter contrast, sex.

Firstly, in order to check the efficacy of the imputation step, a histogram of the *p*-values from the SRS contrast analysis was plotted (Fig. C.3). This showed the appropriate distribution, suggesting that the imputation of missing values had not distorted the data. Using the same significance threshold as in the differential

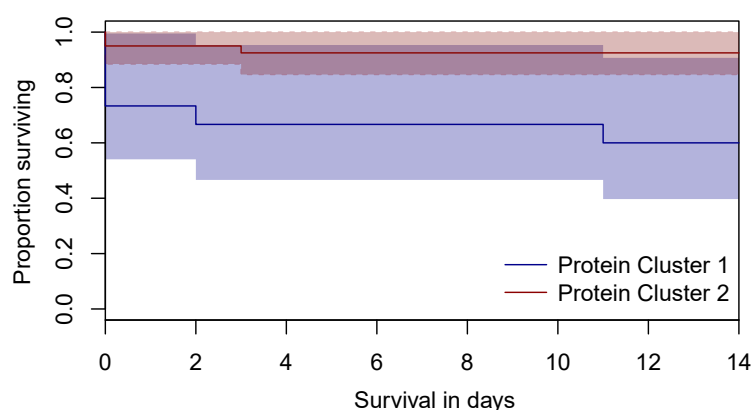


Figure 5.16: Survival differences between clusters based on proteomic data.

Survival rates for the first 14 days in ICU are plotted for the two subgroups defined using unsupervised clustering. Patients were divided into two groups of 15 and 40 patients respectively, and a significant difference was found in early mortality (6/15 vs 3/40 non-survivors).

gene expression analysis ($FDR < 0.05$), 38 proteins were found to be differentially abundant between SRS1 and SRS2 (Fig. 5.17), with six of these having a fold change > 1.5 . The majority (22/38) were upregulated in SRS1 compared to SRS2. In contrast, no differentially abundant proteins were identified between male and female patients, or between survivors and non-survivors.

However, previous studies from the TDI have shown that proteins with a Storey q -value < 0.05 can usually be validated (Roman Fischer, personal communication). This is a less stringent adjustment for multiple testing, as shown by volcano plots with either Benjamini-Hochberg FDR or Q -value representing significance level (Fig. 5.17). Using a Q -Value cut-off of 0.05, 81 proteins are significantly differentially abundant between SRS1 and 2, 10 of which have a fold change > 1.5 (Table C.1). Equal proportions are up- and downregulated in SRS1 (40 up, 41 down). This reflects the findings from the transcriptomic data, where there were roughly equal numbers of up- and downregulated (545 up, 530 down) genes in SRS1 compared to SRS2. Similarly, including differential cell counts in the model does not affect the number of differentially abundant proteins.

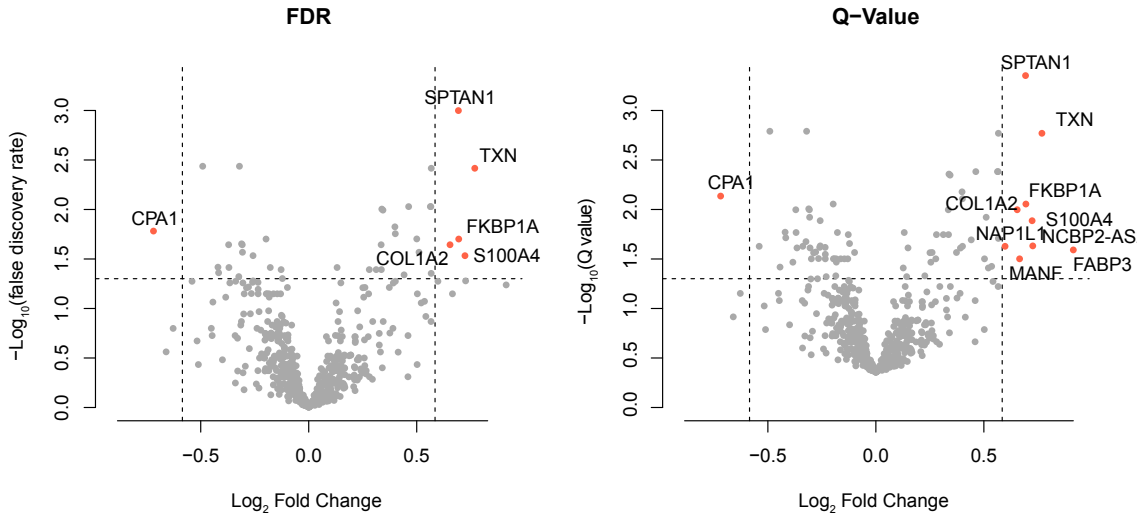


Figure 5.17: Volcano plots of protein abundance in SRS1 compared to SRS2. Each point represents a protein identified by mass spectrometry. Differential abundance in SRS1 versus SRS2 is shown by plotting FDR (*left*) or Q-Value (*right*) against the $\log_2$ fold change, with the dotted lines indicating the significance thresholds used. Proteins that are significantly differentially abundant using these thresholds are coloured red.

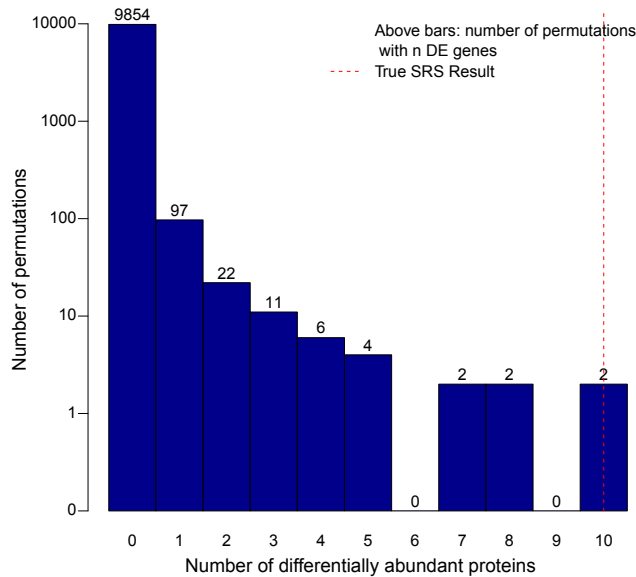


Figure 5.18: Permutation analysis for SRS differential abundance. SRS group membership was permuted 9,999 times and differential abundance analysis performed. The number of significantly differentially abundant proteins was recorded for each permutation and for the true group assignments, and was found to be greatest for the true SRS group assignments. These results are plotted as a histogram, with the true result indicated.

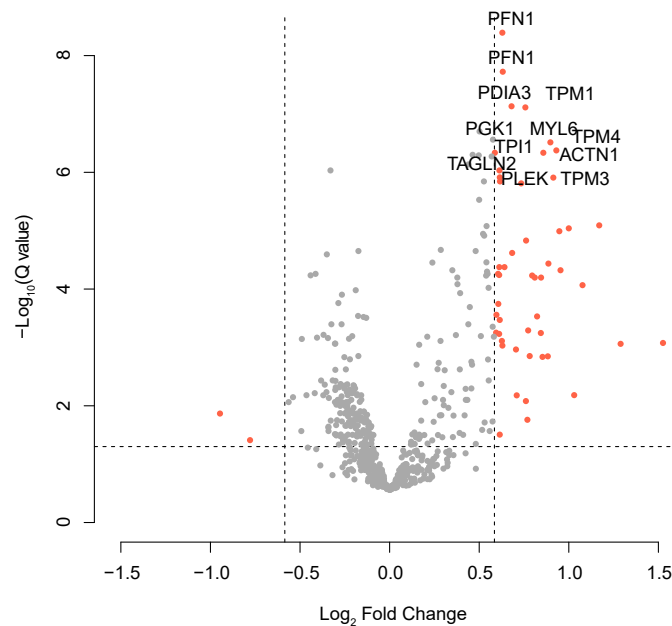
Given the small number of differentially abundant proteins with a large fold change, all significant features were used for pathway enrichment analysis with the background of all detected proteins. This identified enrichment of the GOBP term “positive regulation of I- κ B-kinase/NF- κ B signalling” (FDR=0.05) for SRS1 vs SRS2.

In order to increase confidence in the differences seen between SRS groups, permutation analysis was performed. SRS group membership was permuted 9,999 times, the limma analysis repeated, and the number of features significant at Q Value <0.05 recorded. The greatest number of differentially abundant features recorded (n=10) was for the true group assignment and one permutation (Fig. 5.18), suggesting that more proteomic differences are seen between the groups than would be expected by chance (p=0.0002).

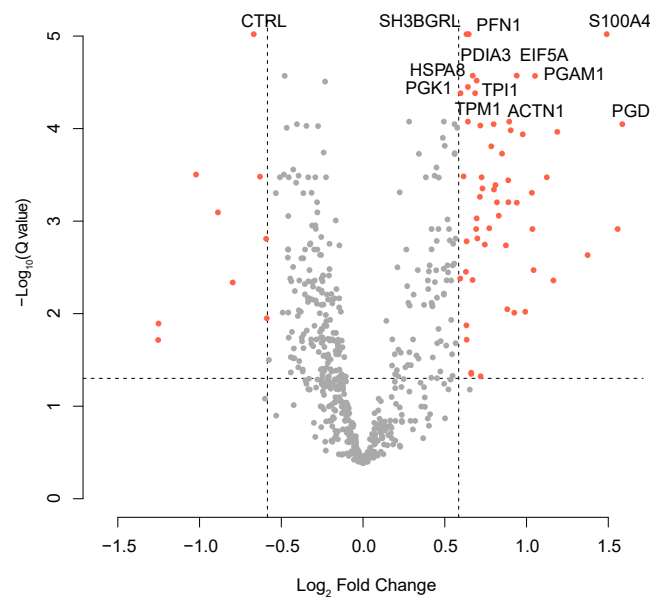
5.2.11 SRS and proteomic clusters

Differential abundance analysis was then performed between the protein clusters identified with the unsupervised clustering. This identified 51 species with a Q Value < 0.05 and a fold change > 1.5 (Fig. 5.19a, Table C.2). Enriched GOBP terms with the background of all identified proteins include canonical glycolysis (FDR 6.1×10^{-3}), as was seen in the principal component analysis and in the SRS transcriptomic analysis, and oxidation-reduction (FDR 6.1×10^{-3}) (Table C.3).

Finally, the cohort was restricted to those patients assigned to the same groups in both the transcriptomic and the proteomic clustering (n=37). This was intended as a means of identifying those patients with a more consistent sepsis response signature, observed at both the transcriptional and translational level. Of the 22 samples not assigned to the same group, 8 were not confidently assigned to an SRS group using the *ACT-B* normalised model described in Chapter 4. The differential abundance analysis was repeated with these consistent clusters, and



(a) Protein Cluster 1 vs Protein Cluster 2



(b) Consistent Cluster 1 vs Consistent Cluster 2

Figure 5.19: Volcano plots for clustering contrasts.

Each point represents a protein identified through mass spectrometry. Differential abundance between (a) the groups identified by unsupervised clustering of the protein data, and (b) consistent transcriptomic and proteomic clusters, are shown by plotting Q-Value against the $\log_2$ fold change. The dotted lines indicate the significance thresholds used, and proteins that are significantly differentially abundant using these thresholds are coloured red.

showed a greater number of differentially abundant proteins than for either the SRS contrast or the unsupervised clusters (65 features; Fig. 5.19b). Conversely, fewer enriched pathways were observed, possibly due to the large proportion of proteins that are included in the gene set of interest while correcting for the background of all tested proteins. Oxidation-reduction was again significantly enriched in this gene set, and though not significant after correction for multiple testing, glycolysis (FDR=0.064) and platelet aggregation (FDR=0.064) were both highly prioritised.

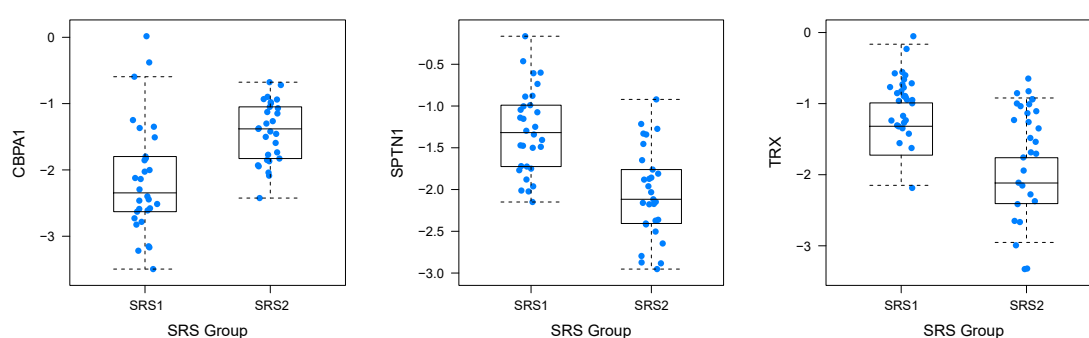


Figure 5.20: Proteins selected for predictive model.

The abundance of three proteins predictive of SRS group membership (CBPA1, SPTN1, and TRX) are plotted by SRS group.

5.2.12 Predictive biomarkers

As for the transcriptomic analysis, GeneRave was used to identify proteins that could be used to assign samples to SRS1 and 2. All proteins were used as potential predictors, resulting in a model based on three proteins. These were carboxypeptidase A1 (CBPA1, encoded by *CPA1*), thioredoxin (TRX, encoded by *TXN*), and spectrin alpha chain, non-erythrocytic 1 (SPTN1, encoded by *SPTAN1*). Notably, it has been hypothesised that TRX plays a protective role in acute inflammatory diseases, as it is an antioxidant and an inhibitor of neutrophil extravasation into sites of inflammation. Serum TRX levels have been found to be significantly higher in patients with sepsis compared with healthy individuals,

and in non-survivors than survivors of sepsis. However, in a mouse model for abdominal sepsis, treatment with recombinant TRX enhanced survival (Hofer et al. 2009; Chen et al. 2016; Trevelin et al. 2016).

TRX was similarly found to be present at higher levels in SRS1 patients, who have a higher mortality rate than SRS2 patients. SPTN1 was also more abundant in SRS1 patients, whereas CBPA1 was found at lower levels (Fig. 5.20). Using leave-one-out internal validation, this three protein model had a misclassification rate of 3.9% and an AUC of 0.921 (Fig. 5.21).

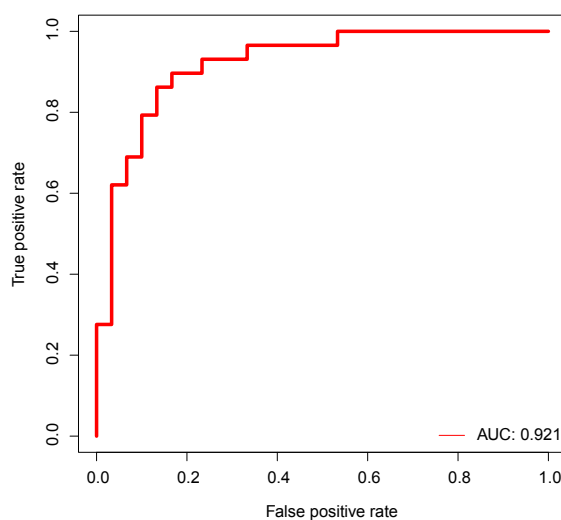


Figure 5.21: ROC curve for the three protein SRS model.

The accuracy of the protein model was calculated using leave-one-out validation. The sensitivity is plotted against the specificity, and the area under the curve given.

5.2.13 Time course analysis

FP patients were chosen for this pilot experiment due to knowledge of the time since onset of disease. In the transcriptomic datasets, time course analysis was carried out in limma based on number of days from disease onset. A comparable analysis was performed here, and no proteins were identified with significant changes in abundance over time. However, if SRS was included in the model, 27 proteins were found with an FDR <0.05 for temporal changes (Table C.4). This could provide further evidence for the importance of interactions between sepsis

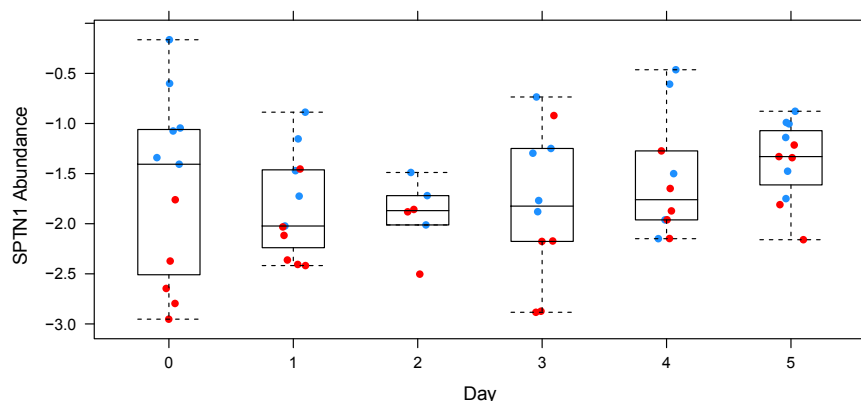


Figure 5.22: Abundance of SPTN1 over sepsis progression.

The abundance of SPTN1 in each sample is plotted according to the number of days from disease onset at the time of sampling. Each sample is coloured according to SRS group membership, showing the differential abundance between SRS1 and 2 and how this relates to changes in abundance over time.

response state and disease progression. For example, SPTN1, which was selected as a predictor for SRS, is also found to change over time; whilst differences between SRS1 and SRS2 are apparent over the whole time course, they are much greater immediately after disease onset (Fig. 5.22). However, the small sample sizes must be considered when interpreting these results.

5.2.14 Comparison with transcriptomics

Of the 620 features analysed in the proteomics dataset, 395 were matched to a gene in the transcriptomic data from the same patients. Comparing the gene expression with paired proteomics data by Spearman rank correlation does not show overall positive correlation between gene transcripts and proteins. The strongest relationship between a gene and the protein it encodes is for *SERPINB1*, a leukocyte elastase inhibitor (Spearman's $\rho=0.54$; Fig. 5.23).

It is plausible that the levels of proteins that are constitutively secreted will correlate more strongly with the gene expression measurements, as they would be less subject to modulation within the cell. Non-secreted proteins are detected

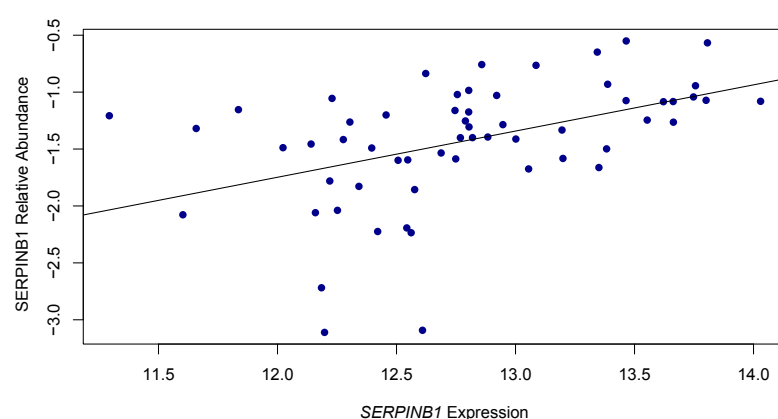


Figure 5.23: Correlation between SERPINB1 expression and protein abundance

The abundance of serum SERPINB1 is plotted against *SERPINB1* expression in leukocytes taken from the same patient on the same day. The two are fairly well correlated (Spearman's $\rho=0.54$)

in the serum as they are released through processes such as apoptosis; the abundance of such proteins will therefore vary in a complex manner that is unlikely to be predictable from gene expression in leukocytes. The Human Protein Atlas predicts 2918 proteins (15% of the human proteome) to be secreted based on the presence of a signal peptide and the exclusion of predicted membrane proteins (Uhlén et al. 2015). Of these, 194 were measured in the FP patients at both the gene expression and protein abundance levels; however, this subset were not more strongly correlated. Finally, excluding those proteins with missingness in the raw data does not improve the correlation between the transcriptomic and proteomic data.

Six of the ten features that were differentially abundant between SRS1 and SRS2 have gene expression data available. In the differential expression analysis described in Chapter 4, *FKBP1A* was also identified as being differentially expressed between SRS1 and 2 (Fig. 5.24). This protein is an immunophilin, which acts as an immune regulator and is also involved in protein folding and trafficking. Its upregulation in the more immunosuppressed SRS1 is consistent with its role as an inhibitor of TNF- β receptor signalling (Chen et al. 1997).

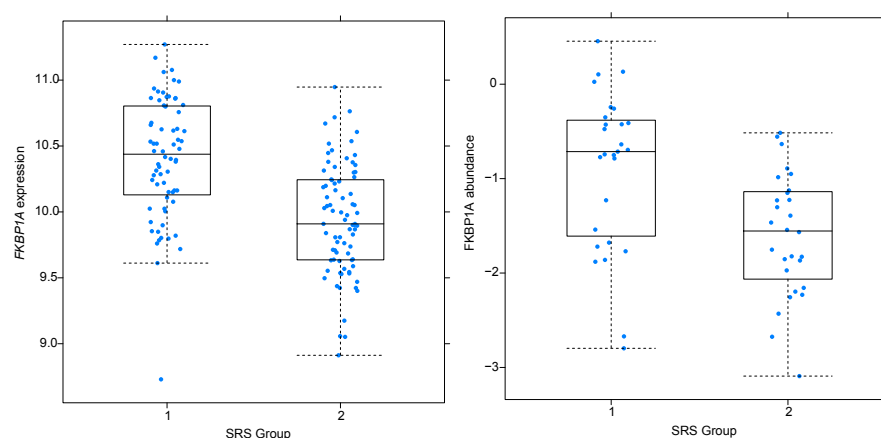


Figure 5.24: *FKBP1A* gene expression and protein abundance by SRS.

FKBP1A is found to be differentially expressed between SRS1 and SRS2 in FP patients, and is also differentially abundant in plasma, with consistent downregulation in SRS2.

5.3 Discussion

The transcriptome of sepsis patients has now been studied fairly extensively, by the GAINs Investigators and others (Davenport et al. 2016; Burnham et al. 2017; Tang et al. 2010; Wong et al. 2012a). Analysis of protein dynamics might be equally beneficial to our understanding of sepsis, and the correspondence of transcriptomic findings to the protein level had not been explored. In this chapter, I have profiled the plasma proteome of patients with sepsis due to FP, and described preliminary analyses investigating intra-individual and temporal variation in protein abundances that could be expanded upon in the future.

5.3.1 The sepsis proteome

A number of studies describe the utility of serum biomarkers in sepsis; for example, measurement of procalcitonin has been shown to distinguish between infectious and sterile SIRS (Lee 2013), and IL-8 has been described as a predictor of outcome in paediatric septic shock (Wong et al. 2008). Such biomarkers are of interest both as a source of information about the host immune response state,

and as potential therapeutic targets. A dataset profiling the adult sepsis plasma proteome is therefore an important resource for further hypothesis generation; in particular, given the heterogeneity in the sepsis response already described, the restriction of the cohort to a single source of sepsis adds to its value.

However, widespread variation is still observed across patients; for example, approximately one third of the proteins identified were not detected in every sample. The effect of this heterogeneity and complexity in the relatively small cohort profiled possibly contributes to the lack of differences detected between male and female patients, despite previously reported observations of large differences between sexes (Geyer et al. 2016). Expansion of the proteomics cohort to similar numbers as the transcriptomic data would greatly increase the analytical power to that described in the preceding two chapters. Additionally, inclusion of CAP samples and further stratification of patients by various clinical factors (treatments, comorbidities, disease severity) would allow other associations to be investigated. Lastly, plasma samples from pre-surgical patients could be profiled in parallel with septic samples for direct comparison of the patient proteomes, as described in the transcriptomic data. This would be particularly informative in exploring the relevance of the pathway enrichment results for all detected proteins; for example, the coagulation cascade might be active to some extent in healthy patients, but is also likely to be upregulated in the context of sepsis (Shen et al. 2006; Lissauer et al. 2007).

Moreover, the effect of tissue specificity on protein abundance and detection might be explored. Here, plasma was used as an easily accessible tissue type that is relevant to disease and future clinical testing; however, the profiling of multiple cell types from the same patients would give a much more detailed picture of the sepsis response. Whilst including differential cell counts in the differential abundance analysis did not have an effect on the results, some variability might be resolved by such an approach.

Lastly, in a larger cohort the data might be investigated at the sub-proteomic level, and the unidentified peptides explored. The databases used to identify peptides by mass return information about the quantities of different isoforms, which are not interrogated here as this is beyond the scope of this pilot experiment. However, more detailed investigation of differential isoform usage could be carried out, in a similar way to exon-level analyses in RNA-seq datasets. Finally, pathogenic proteins have been detected in human plasma through mass spectrometry, and mapping of non-human peptides to bacterial databases could refine current microbiological knowledge on these patients (Grassl et al. 2016).

5.3.2 SRS groups at the protein level

The contrast of SRS1 and 2 in terms of protein abundance demonstrated that, as in the gene expression data, much greater differences were observed than between survivors and non-survivors. Whilst these still only involved a relatively small number of proteins (1.6% of those tested, compared to 3.8% in the FP transcriptomic contrast), this might be related to the limited sample size used. In parallel, analysis of global variation and unsupervised clustering found that the SRS groups defined using gene expression data translated to significant variation at the protein level. Although the unsupervised clustering resolved slightly different groupings, with a number of SRS1 samples being grouped with the SRS2 samples in cluster 2, similarities in the proteomic profiles were observed. Again, further analysis in additional patients would be helpful to determine the true overlap in the transcriptomic and proteomic subgroups. Overall, these findings reinforce the evidence for the presence of subgroups within sepsis patients, the relevance of the SRS groups in particular, and the utility of -omics data to understand the host response to disease.

The differences observed in the plasma proteome can be interpreted in the light of the biological differences between the SRS groups. For example, cell death

was enriched in the transcriptomic contrast of SRS1 and SRS2, which could have caused an increase in the release of non-secreted proteins from cells in SRS1. Similarly, platelet degranulation is an enriched function in the combined cluster contrast, which again could explain the presence of increased quantities of intracellular species. However, specific assays would be necessary to validate such mechanisms, which would be relatively intractable *in vivo*.

The identification of three protein biomarkers that can assign SRS in FP patients with a high level of accuracy is promising for future applications of this approach. It is important to note that this protein panel has only been tested by the leave-one-out method due to sample size limitations, and is likely to perform less well in new samples. Additionally, the input for the model is in the form of relative measures of protein abundance, and it is therefore not directly translatable for use in prospectively recruited patients. For this application, absolute measurement of protein abundance would be required, or normalisation to a “housekeeper” protein as described in the transcriptomic model.

Combination of the transcriptomic and proteomic data might prove to be the most successful method for patient stratification, although the technical aspects of developing such an approach for use in the clinic would be more challenging. Clustering on combined data was not performed with this pilot dataset due to the small data size, but it would be of interest in future, larger cohorts. As mentioned above, samples that were less confidently assigned to SRS based on gene expression correspond with a number of patients who cluster differently in the protein data. Such discrepancies might be resolved, and confidence in group assignment increased, by use of a combinational approach that could take advantage of the complementary but differing information contained in the two data types. For example, in Pils et al. 2013, combining gene expression measurements with a serum protein biomarker panel significantly increased the ability to diagnose epithelial ovarian cancer.

5.3.3 The dynamic proteome

Investigation of protein abundance over time in this pilot dataset suggests that dynamic changes are prevalent, as observed in the transcriptomic data, and again these are dependent upon the immune state of the patient. However, the numbers of samples examined are too small for definitive conclusions to be made, particularly about individual proteins of interest. Changes over time could be further explored using a complementary approach, namely the profiling of serial samples from the same patient.

5.3.4 Integration of transcriptomics and proteomics

A basic assumption usually held when conducting gene expression-based experiments is that the results relate directly to protein quantity and are thus informative about cellular activity. However, in the FP patients described here, there is little to no correlation between leukocyte gene expression and plasma protein levels. This is in fact not unusual in such analyses; several studies have estimated that around 40% of variance in protein abundance can be explained by changes in gene expression (Vogel and Marcotte 2012). This varies across organisms and tissue types, but it highlights the effect of additional factors on protein concentration, such as translation efficiency and protein degradation.

In the data described here, the correlations were generally weaker than this, which may be explained by the different sample types profiled. Whilst leukocytes may be the source of much of the protein found in plasma, a number of intermediary biological processes are likely to have broken down the relationship between mRNA concentration in white blood cells and protein concentration in plasma. For future experiments, protein might be extracted from specific cell types in parallel with RNA processing in order to more directly compare gene expression and protein abundance. Additional noise might have been introduced

from the differing methods employed to process samples and generate the data. For example, protein extraction from plasma may be affected by haemolysis, or the lysis of red blood cells. An assay designed to assess the degree of haemolysis that has occurred might be used in the future to aid interpretation of results. However, as haemolysis can be caused by both improper plasma sampling and storage, and by some bacteria, it is difficult to understand its biological relevance in the context of infection (Sowemimo-Coker 2002).

In the context of SIRS, there have been several studies where both gene expression and protein concentration have been measured, with similar outcomes to those described in this chapter. For example, in Prabhakar et al. 2005 the endotoxin response is studied on the basis of cytokine gene expression and protein concentration. In general, they found that cytokine protein abundance correlated well with gene expression; however, they note that the temporal profiles of expression of some genes did not correlate well with the protein data. Secondly, Feezor et al. 2004 use leukocyte gene expression and plasma protein levels to predict SIRS in surgery patients. They do not attempt to formally compare the expression and proteomic data due to small sample sizes, but state that there was minimal correlation between changes in gene expression and the corresponding plasma protein concentrations. The authors suggest that this is because circulating cytokines are not generally derived from leukocytes. This parallels the findings described in this chapter.

The generally low correspondence observed between gene expression and protein abundance does not negate conclusions drawn from transcriptomic data. Much more basic research is necessary to dissect the complex relationship between transcriptomic regulation and biological activity, and to fully understand the consequences of differential gene expression. However, there are already examples of gene expression findings being translated to the protein level, despite the results described above.

In this pilot experiment, very few proteins were identified as being differentially abundant between groups, so it is not meaningful to specifically investigate correlation for this subset of species. However, increased mRNA-protein correlations has been described for differentially expressed genes in an ovarian cancer model (Koussounadis et al. 2015), increasing confidence in the relevance of conclusions drawn from expression analyses despite the overall poor correspondence with proteomic data. This is reinforced in the context of paediatric sepsis, where in the PERSEVERE programme Wong et al. 2012a selected 12 biomarkers for survival based on pre-existing gene expression data and measured serum protein levels. No direct comparison of gene expression and protein quantities was performed, but a subset of these biomarkers were used successfully in a decision tree to predict survival. This provides proof of principle for the translation of gene expression findings to protein assays. Hence, one possible means of extending the work described in this thesis would be to use targeted methods (e.g. ELISA) on large numbers of samples to assay specific proteins where the gene has been highlighted in the transcriptomics analysis.

5.4 Conclusions

Our understanding of the sepsis response has been greatly advanced in recent years through genome-wide gene expression profiling. These findings prompt the fundamental question of how gene expression results apply to the protein level. In this chapter, I have described the generation of proteomics data through mass spectrometry for 60 samples from patients with sepsis due to FP. This pilot experiment illustrates both the promise of an integrated -omics approach, and the limitations that should be considered in future work extending these analyses.

REGULATORY GENETIC VARIATION IN THE SYSTEMIC INFLAMMATORY RESPONSE

This chapter investigates the effect of genetic variation on the host response in sepsis and surgery, focussing particularly on the genetics of gene expression and how this relates to the heterogeneity described in previous chapters

6.1	Introduction	194
6.2	Results	203
6.3	Discussion	240
6.4	Conclusions	246

6.1 Introduction

There is substantial evidence that susceptibility to and outcome from infectious diseases are highly heritable traits (Sørensen et al. 1988). The GenOSept study therefore performed a genome-wide association study (GWAS) for survival in sepsis, and identified a significantly associated variant in the *FER* gene (Rautanen et al. 2015). However, the heterogeneity described in previous chapters make this an extremely challenging area of study, and as yet the functional explanation behind this association is not clear. This chapter describes investigation of regulatory genetic variation in sepsis, which could help elucidate the mechanisms by which the individual response to infection is controlled.

6.1.1 The genetics of gene expression

It has been estimated that more than 90% of single nucleotide polymorphisms (SNPs) reported in GWAS as showing association with complex phenotypes are in non-coding regions (Maurano et al. 2012). Given the lack of an impact on protein sequence and structure, many of these SNPs are presumed to have an effect through the regulation of gene expression. However, it is far from straightforward to demonstrate such an interaction. Whilst regulatory variants are frequently found close to their target gene, this is not always the case, and furthermore, the presence of multiple variants in strong linkage disequilibrium (LD) means that the SNP most strongly associated with a trait may not in fact be the causative variant (Cookson et al. 2009).

This may be addressed by mapping statistical associations between genetic loci and the expression of a gene, locally or for all variants and genes genome-wide. Expression quantitative trait loci (eQTL) are variants that are associated with the amount of a particular gene transcript (eGene), and so are hypothesised to have an impact on transcriptomic regulation. These eQTL may act locally (in *cis*), such as those found in transcription factor binding sites adjacent to the target gene, or distally (in *trans*), such as those modulating expression of a transcription factor or a microRNA (Fairfax and Knight 2014). Given that the phenotype of gene expression is much more closely linked mechanically to the genetic variant than the complex traits investigated in GWAS, eQTL studies may be sufficiently powered to detect common *cis* associations with hundreds rather than thousands of samples. eQTL mapping is therefore a useful tool to generate hypotheses about the mode of action of variants associated with a particular trait, and increase understanding of transcriptional regulatory mechanisms (Musunuru et al. 2010; Jostins et al. 2012; Ripke et al. 2014).

As mentioned above, the response to infection has been demonstrated to have

a strong genetic component. An adoption study found that adoptees had an increased risk of death from infections if one of their biological parents had died from an infection; this effect was not observed with adoptive parents (Sørensen et al. 1988). Further work estimated both an increased risk of acquiring an infection (1.18-fold) and risk of death from infection (9.36-fold) amongst full siblings of adoptees hospitalised with an infection, compared to adoptive siblings (Petersen et al. 2010). Genetic differences are therefore likely to be responsible for a large part of the heterogeneity and mortality observed in sepsis patients, possibly through the combinatorial effect of a large number of regulatory variants.

6.1.2 Context specificity of eQTL

The complexity of the human regulatory landscape require eQTL studies to take context specificity into account. Gene expression has been found to be highly dependent upon the cell type observed, meaning that different sets of eQTL may be found to be active in different tissues (Fairfax et al. 2012). Large studies such as the Genotype-Tissue Expression (GTEx) study describe eQTL in a variety of tissue types, providing publicly available resources that may be queried (Westra et al. 2013; Ardlie et al. 2015; Joehanes et al. 2017). Additionally, external stimuli can affect gene expression, so the action of regulatory variants and the eQTL detected may be context specific. For example, eQTL have been found that are only active following LPS stimulation (Fairfax et al. 2014; Lee et al. 2014), suggesting that infectious disease can have a profound effect on the regulatory landscape. This has been investigated through the mapping of eQTL using gene expression data from a disease context. For example, eQTL specific to the response to *Mycobacterium tuberculosis* found SNPs that played a role in anti-mycobacterial immunity (Barreiro et al. 2012). Similarly, associations between genetic variants and the gene expression response to rhinovirus infection highlighted loci that

could drive inter-individual variation (Çaliskan et al. 2015).

Sepsis is a disease that has a profound effect on physiology, metabolism, and immune function; as such, the regulatory landscape in sepsis is expected to differ from that of healthy individuals. Furthermore, the interplay between genetic and environmental factors can modulate gene expression differently across individuals, creating a complex and dynamic regulatory network.

6.1.3 Data processing for eQTL

Detection of eQTL with confidence requires high quality variant and gene expression data. These may be generated using a variety of platforms, e.g. sequencing or microarrays, but are usually reduced to SNP genotypes and gene level expression values for eQTL mapping. Whilst other variant types (insertions, deletions, structural variants) can also affect gene expression, these are not considered in this chapter. Similarly, RNA-seq data would allow the investigation of exon-level effects, but this was not possible with the cohort described here.

Genotyping data

Here, genotyping data were generated using DNA microarrays. QC steps are applied to identify and exclude poor quality SNPs and samples, with the aim of reducing spurious associations. The pipeline described in Anderson et al. 2010 was designed for GWAS data; however, the same considerations apply in the context of eQTL. The QC procedures described are carried out using PLINK 1.9 (Chang et al. 2015).

Firstly, per-individual checks identify discordance between the recorded sex and that ascertained from the genotyping data. As males only have one copy of the X chromosome, SNPs can only be heterozygous if they are found in the pseudoautosomal region of the Y chromosome; females however can be

heterozygous for any X chromosome SNP. This allows sex to be inferred from heterozygosity rates across all X chromosome markers. This can indicate issues such as contamination, errors in the clinical data, or in the sample plating. Similarly, deviation from the mean heterozygosity for autosomes can indicate contamination or inbreeding, and DNA quality can affect the genotype call rate. These measures can therefore be used to exclude poor quality samples.

Sample independence is a key assumption of genetic analyses, and is violated by hidden relatedness or duplicates. These can introduce biases in genotype frequencies and hence skew the results. Such cases may be identified by calculating the identity by state (IBS) for each pair of individuals based on the proportion of shared independent SNPs, which will be greater for relatives than for unrelated pairs. Second-degree relatives and above may thus be identified, and one individual from each pair excluded.

Finally, population stratification may be examined through multi-dimensional scaling and comparison to large population panels such as the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012). As allele frequencies may differ between populations, this can again be a confounding factor that must be addressed, either by excluding individuals of divergent ancestry or correction for population structure within the analysis.

SNP quality control steps include the removal of variants with a high missingness rate, with a higher threshold often used for low frequency SNPs, and the exclusion of rare variants (minor allele frequency (MAF) <1-5%) as these are difficult to call and often present as false-positives in association analyses. Deviation from Hardy-Weinberg equilibrium can indicate genotyping error, and so is also used as a filter for genotyping quality.

Gene expression

Further probe filtering is also advisable for eQTL mapping with microarray gene expression data to avoid spurious results. This can involve the exclusion of probes mapping to multiple sites in the genome, and those that overlap a known SNP of appreciable frequency in the population (e.g. $MAF \geq 5\%$).

6.1.4 Imputation

Genotyping data from DNA microarrays are limited to the SNPs included on the particular platform used. However, the number of variants covered and the overlap between different datasets may be increased through imputation. This technique makes use of the genetic linkage and statistical associations between neighbouring variants that result from a lack of recombination, meaning that a small number of SNPs can act as markers for additional variants. Genetic variation in the study sample is matched to a more densely genotyped reference panel, and the shared variation is used to estimate the likely genotypic states at untyped sites in the study sample.

A number of large reference panels are now publicly available, including the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012) and the Haplotype Reference Consortium (HRC) (McCarthy et al. 2016). The latter combines sequencing data from multiple cohorts of predominantly European ancestry (including the 1000 Genomes Project data), currently comprising 64,940 haplotypes at 39,235,157 SNPs. The large sample size provides greater accuracy for imputation, but the data are restricted to SNPs only. The Sanger Imputation Service provided by the Wellcome Trust Sanger Institute allows imputation into the HRC using the Positional Burrows-Wheeler Transform (PBWT) method.

6.1.5 eQTL mapping methods

Much like differential gene expression analyses, tests for eQTL involve the comparison of gene expression levels across groups of samples, defined by the genotype at a particular locus. This is frequently carried out by fitting a linear model using the number of copies of the minor allele as the independent variable. This allows the inclusion of covariates in the model to adjust for environmental or technical effects, thus increasing the power to detect eQTL (Westra et al. 2013). Whilst known covariates may be included in this way, this approach is based on prior knowledge and assumptions and is therefore biased. Instead, principal components of the gene expression and genotyping data may be included to account for confounding factors and reduce false positive rates.

MatruxeQTL is an R package that maps eQTL using a linear additive model, using matrix operations for computational efficiency. Alternatively, mapping eQTL using a linear mixed model permits the inclusion of multiple samples from the same patient to increase power, by introducing a random effect for patient. Interactions between genotype and environmental factors may also be tested for with this design (Davenport et al. 2017).

Cis and *trans*-eQTL are usually mapped separately, as the restriction on location for *cis*-eQTL reduces the number of tests performed and therefore the required sample size. Variable distance windows have been used in previous studies to define local eQTL, with the majority ranging from 250kb-1Mb.

6.1.6 Protein QTL

Differences in gene expression levels are thought to exert an effect on phenotype by causing differences in the amount of protein translated. Additionally, post-transcriptional regulation is prevalent and partially under genetic control. eSNPs

or other variants might therefore show an association with protein quantity.

Given the greater distance between regulatory variation and protein quantity, and the additional modulating factors likely to be involved, protein quantitative trait loci (pQTL) are expected to have smaller effect sizes. Larger sample sizes would therefore be required to detect pQTL genome-wide with correction for multiple testing. Proteome-wide studies are much less commonly performed than for eQTL, given the challenges in quantifying proteins (Horvatovich et al. 2014), but have identified associations in healthy individuals and in the context of disease (Melzer et al. 2008; Suhre et al. 2017; Kim et al. 2013; Di Narzo et al. 2017). With recent advances in technology, increasing numbers of individuals and features have been assayed (Sun et al. 2017; Chick et al. 2016; Wu et al. 2013).

6.1.7 Heritability of complex traits

The contribution of genetic factors to variation in a complex trait has traditionally been estimated using twin and adoptee studies; for example, the work by Sørensen et al. 1988 described above. Quantifying heritability is a central question in biology, as it is highly informative for risk prediction and for estimating the potential of genetic research in that trait. However, such studies take a long time and are not always practical.

Genome-wide Complex Trait Analysis (GCTA) is a software package that estimates the proportion of phenotypic variance that can be explained by all measured genetic variation (e.g. all genotyped SNPs) by fitting all SNPs simultaneously as a linear model (Yang et al. 2010). This approach does not identify individual associated SNPs, but can provide evidence for genetic control of the trait. It was originally designed to address the question of missing heritability from GWAS studies; this refers to the small proportion of variation in

a trait that can be explained by known GWAS associations despite much higher estimates of heritability having been reported. By estimating the combined effect of genotyped SNPs, GCTA aimed to better resolve the impact of SNPs with very small effects or those that are not well represented on genotyping arrays. It can now be used more generally to improve understanding of the genetic architecture of complex traits.

6.1.8 Aims

The main aim of this chapter is to define eQTL in the context of sepsis, and to explore these associations through the use of relevant data available from the literature. By identifying regulatory genetic variants that modulate gene expression in sepsis, I aim to explore the causes and consequences of inter-individual variation in the response to infection. My specific aims are to test the following hypotheses:

1. specific genetic variants are associated with the expression of nearby genes in the context of sepsis
2. eSNPs are enriched in context-specific regulatory regions
3. eQTL can be modulated by interactions between regulatory variants and
 - (a) source of sepsis
 - (b) SRS group
4. regulatory variants are associated with gene expression in surgery patients
5. genetic variants are associated with the abundance of specific plasma proteins in sepsis due to FP
6. QTL vary according to disease context (sepsis, surgery, healthy)
7. SRS group membership is associated with global genetic variation

6.2 Results

6.2.1 Genomic landscape of sepsis eQTL

Previous work in the Knight lab

eQTL have been previously mapped in sepsis, using 240 patients from the Davenport 2011 CAP cohort (Davenport 2014; Davenport et al. 2016). Cis-eQTL were defined for 3,795 genes (FDR <0.05; 26.7% of those tested) using an additive linear model (MatrixEQTL, Shabalin 2012). These genes with an eQTL (eGenes) were significantly enriched for pathways such as PI3K signalling, antigen presentation, and mitochondrial dysfunction, and included some key immune and metabolic response genes such as *TLR4*, *TNF*, *IL15*, *IFNAR2*, and *NFKB1*. Sepsis-specific eQTL were defined through comparison to a meta-analysis of whole blood eQTL in 5,311 healthy individuals (Westra et al. 2013). This identified 1,377 genes (36.3% of all sepsis eGenes) that had a significant eSNP (FDR <0.05) in the smaller sepsis cohort only.

No overlap was seen between these eQTL and the association reported for the *FER* gene with survival in sepsis by the GAINs Investigators in the GenOSept study (Rautanen et al. 2015). However, *FER* was below the threshold for detection in the gene expression dataset, so regulation by the GWAS SNP was undetectable under these conditions.

Sepsis eQTL in context

Initially, I explored the 3,795 cis-eQTL mapped in CAP in terms of genomic context and biological associations. Sepsis-specific eQTL SNPs were found to be located significantly further away from the transcriptional start site (TSS) of their target gene than those that were shared with the healthy cohort (Mann Whitney

test $P = 9.4 \times 10^{-38}$; Fig. 6.1). This is consistent with the hypothesis that context-specific regulatory variants are enriched at more distal regulatory elements.

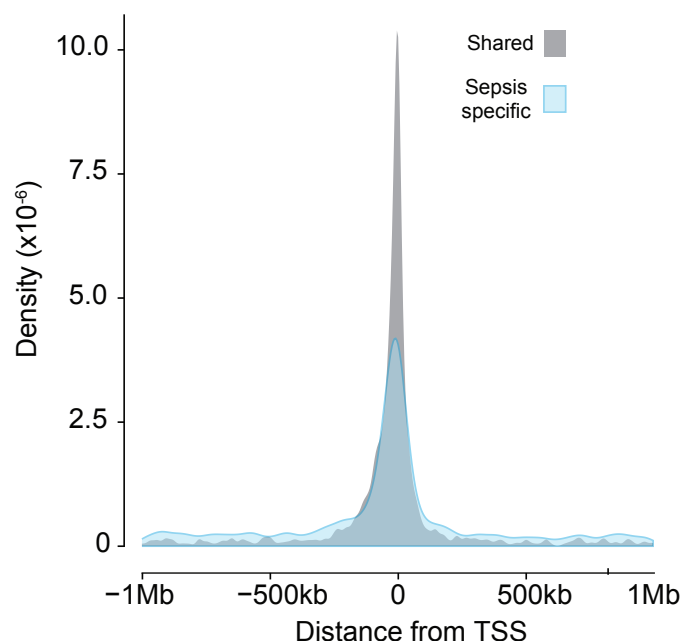


Figure 6.1: Distance between sepsis eSNPs and transcriptional start sites

The distance between each eSNP and the transcriptional start site of its associated eGene was calculated, and eQTL shared (*grey*) and specific (*blue*) to sepsis are found to differ in the absolute distance to the transcriptional start site.

In Chapter 4, an endotoxin tolerance gene signature comprising 331 genes was defined. Tolerance-related genes are likely to be under active regulation in the context of sepsis. I used the ET gene set to test the hypothesis that sepsis eGenes would be enriched for endotoxin tolerance-relevant genes, and found significant enrichment (chi-square $P < 0.0001$).

The epigenomic context of the eSNPs was then investigated using a relevant dataset, namely LPS-treated monocytes (Saeed et al. 2014). As noted above, transcriptomic regulation is highly dependent on the cell type, therefore these data are not a perfect description of the epigenetic background for the sepsis eQTL, which were mapped using leukocytes. Nevertheless, sepsis cis-eQTL eSNPs were significantly enriched in histone marks for active enhancer elements in LPS-activated cells (H3K4 methylation and H3K27 acetylation), in comparison

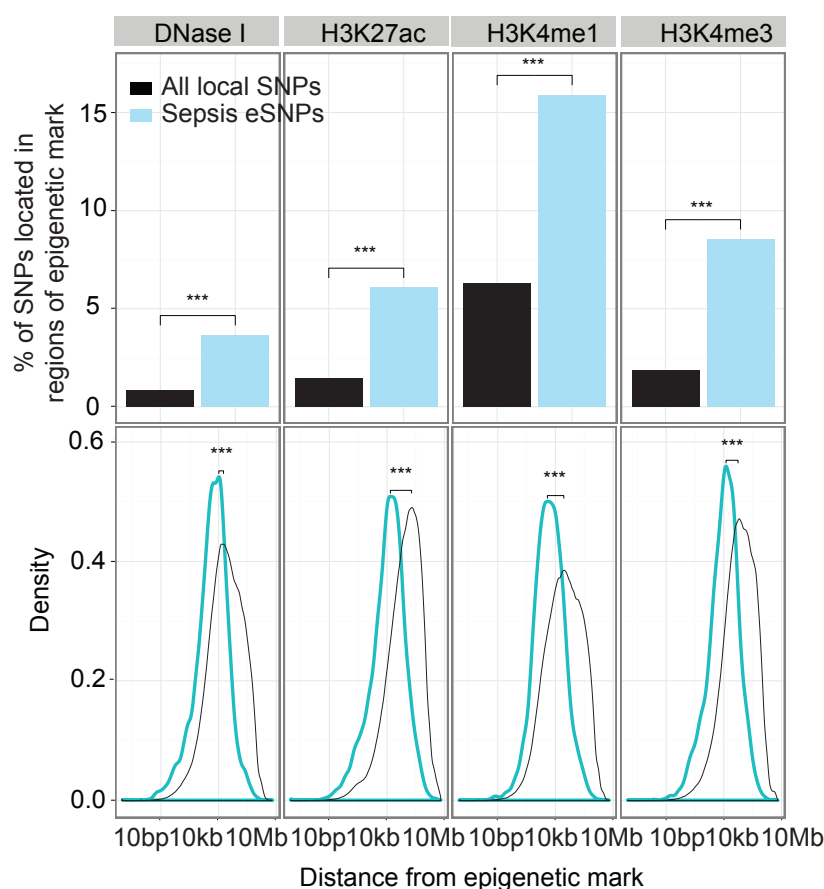


Figure 6.2: Enrichment of eSNPs in active regulatory regions.

Expression-associated SNPs with the lowest p values for each gene ($FDR < 0.05$) were annotated with the distance to a DNase I hypersensitive site (DNase I), and peak intervals defined for histone modifications (H3K27 acetylation, H3K4 monomethylation, and H3K4 trimethylation), in monocytes tolerised by exposure to LPS for 24 hours. The proportion of expression-associated SNPs that overlapped with each mark was compared to that for all local SNPs (Fisher's exact test), as was distance to the nearest mark (Mann-Whitney test). $*P < 2.2 \times 10^{-16}$.

to all SNPs in the cis-window (1Mb) of a probe (2.91% vs 0.86%, $P = 2.02 \times 10^{-26}$). Similarly, sepsis eSNPs were more likely to fall in regions of open chromatin (DNase-I hypersensitive sites) (3.65% vs 0.79%, $P = 1.98 \times 10^{-46}$; Fig. 6.2). These results together demonstrate that sepsis cis-eQTL are more commonly observed in regulatory genomic elements specific to endotoxin-activated cells, providing evidence for their regulatory function.

Sepsis eQTL and SRS

Given the large transcriptomic differences observed between SRS1 and SRS2 (described in Chapter 4), and the context specificity of transcriptomic regulation, it is possible that the two subgroups differ in their genetic architecture and epigenetic landscape. Additionally, some of the differences observed in the individual immune response are likely to stem from regulatory genetic variation. The eQTL mapped in the Davenport 2011 cohort were therefore investigated in combination with the sepsis response signatures defined in the same dataset to identify SRS1 or SRS2-specific associations.

The genes differentially expressed between SRS1 and SRS2 are enriched for eGenes, with 81% of eQTL genes being differentially expressed (ROAST $P < 0.0001$). This includes a large number of sepsis-relevant genes, such as *IL18RAP*, *CCR1*, *CCR3*, and *SIRT1*. It is possible that eGenes are better able to tolerate variation in expression levels, and so are also more likely to be differentially expressed. However, when eQTL were mapped in SRS1 and SRS2 patients separately, there were a number of examples of eGenes where a strong association was observed in one group only. For example, *MTOR* was down-regulated and showed evidence of an eQTL in SRS1 patients, the heat shock transcription factor *HSF2* also had an eQTL that was restricted to SRS1 patients, and for *LAX1* an eQTL was only observed in SRS2 patients (Fig. 6.3).

6.2.2 Extension of eQTL mapping to the larger GAINs cohort

The subsequent availability of additional transcriptomic data allowed me to perform further eQTL mapping in a larger cohort of patients. This increased power to detect associations, and permitted the investigation of modulation by clinical covariates. In particular, the effect of SRS on transcriptomic regulation could be explored further. In addition, the more recent transcriptomic data

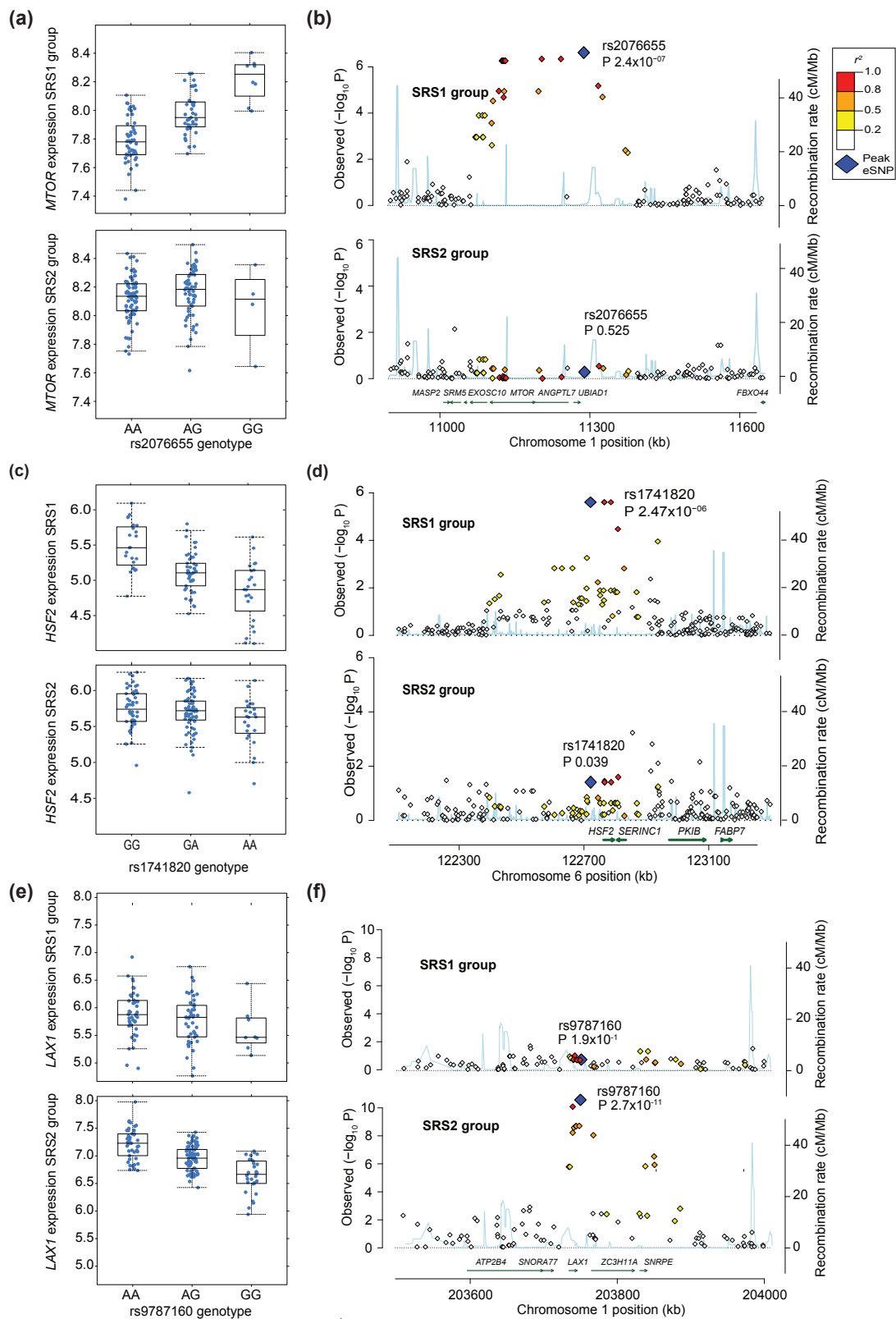


Figure 6.3: Examples of eQTL specific to one SRS group in CAP.

(a, b, c, d) SRS1 and (e, f) SRS2-specific eQTL are shown for *MTOR*, *HSF2*, and *LAX1*. (a, c, e) Boxplots by allele show the expression of each gene in SRS1 and SRS2, and (b, d, f) regional association plots show the significance of association of the eQTL in the context of linkage disequilibrium and nearby genes.

included both CAP and FP patients. The GWAS for survival in sepsis only found a significant association when the cohort was restricted to CAP patients (Rautanen et al. 2015), which together with the transcriptomic analysis described in Chapter 3 suggests that risk factors and the regulation of gene expression vary between patients with sepsis due to CAP and FP. eQTL were therefore also tested for interactions with source of sepsis.

Processing of data for sepsis eQTL mapping

In previous chapters, the various gene expression datasets have largely been analysed independently, each having been designed to answer a particular question. Here, to maximise power to detect eQTL and interaction effects, all the expression data (Radhakrishnan 2010, Davenport 2011, Burnham 2014, Burnham 2016) were combined using ComBat to correct for batch effects (described in Chapter 4). The final gene expression dataset comprised 732 samples from 552 patients, which was then reduced to the 592 samples from 468 patients with genotyping data available. Further probe filtering was carried out, such that probes containing a reported SNP (dbSNP) with a MAF >5%, or those not uniquely mapping to the human genome, were removed to give a final data set of 15,326 probes.

Data for 275 CAP patients had been generated in 2011 on Illumina OmniExpress arrays (644,000 SNPs post-QC). Subsequently, genotyping data were acquired for a further 655 patients with the Illumina CoreExome array (284,000 SNPs post-QC). These genotyping data were processed according to the methods described in Anderson et al. 2010. Samples were excluded on the basis of discordant sex information, proportion of missing genotypes >0.02, heterozygosity rate, and identity by descent; SNPs were excluded for a missing data proportion >0.05, minor allele frequency (MAF) <0.01, and Hardy-Weinburg equilibrium (HWE) $P < 1 \times 10^{-5}$. Following QC, data were available for 635 patients.

Genotyping data were therefore available for 910 GAIN patients in total. However, these data were generated in two batches, using different genotyping arrays, between which relatively small numbers of SNPs are shared (Fig. 6.4). In order to maximise coverage across all samples, imputation was performed against the Haplotype Reference Consortium (HRC) pilot release (McCarthy et al. 2016).

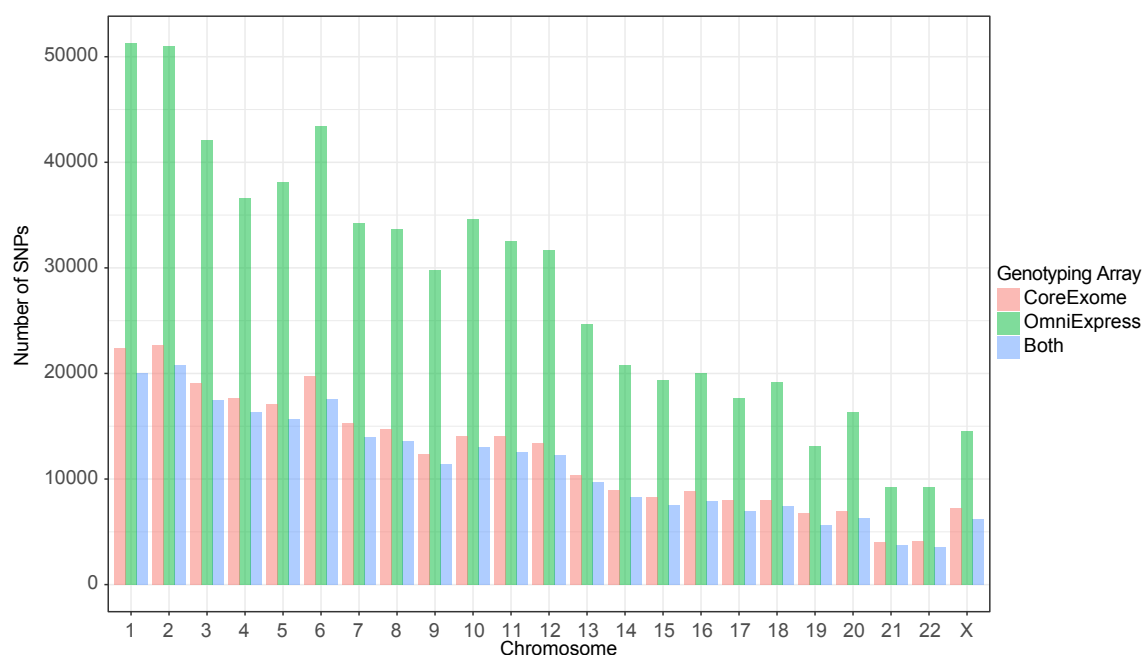


Figure 6.4: Sharing of SNPs between genotyping arrays

The number of SNPs included per chromosome on the Illumina OmniExpress array is compared to that for the Illumina CoreExome array. The number of SNPs shared between the arrays is also plotted.

The two genotyping datasets were prepared for imputation using the “HRC Pre-imputation Checks” tool (Will Rayner, WTCHG) and uploaded to the Sanger Imputation Service independently. Pre-phasing was carried out with EAGLE2 (Loh et al. 2016), and imputation into the HRC was performed using PBWT (Durbin 2014). The imputed data were then filtered on info score (>0.9), the datasets were combined, and reduced to those individuals with gene expression data available ($n=468$). The merged data were filtered on missingness $>5\%$, HWE p value <0.001 , and MAF $>5\%$ (Fig. 6.5). Whilst a large number of variants were discarded in this process, the final imputed data set comprised 18 times as many SNPs as the combined genotyping data ($n_{imp}=4,500,942$ vs $n_{geno}=251,827$; Fig.

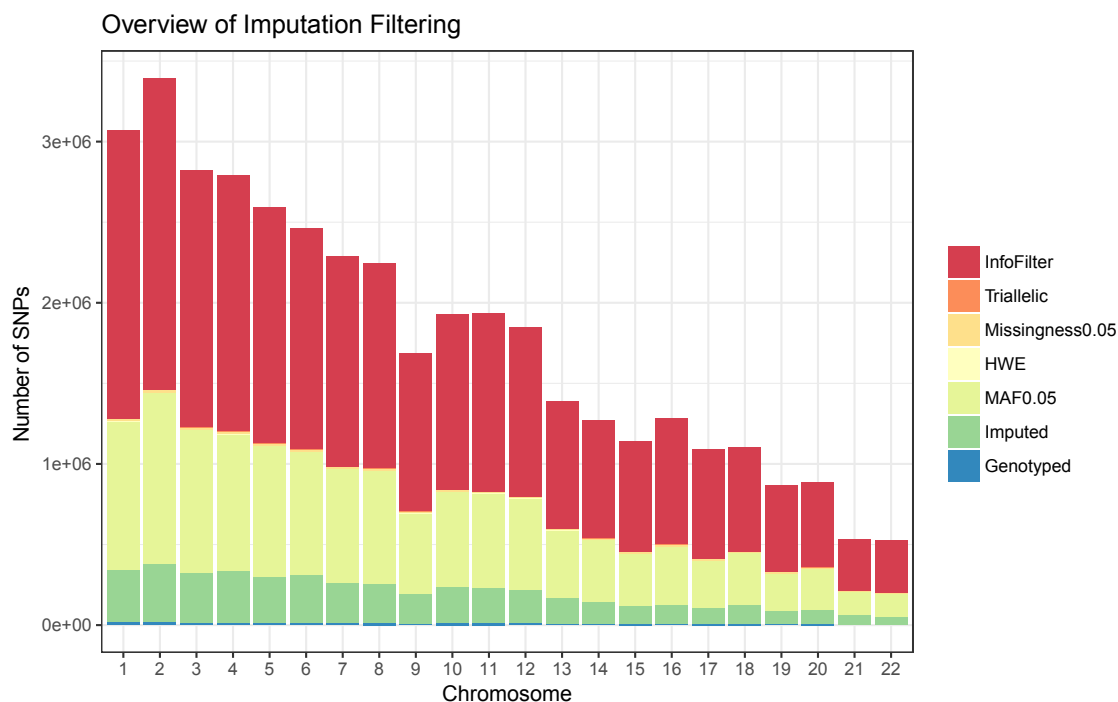


Figure 6.5: Filtering of imputed genotyping data.

The number of SNPs in the combined data set are shown by chromosome, with those removed based on info score, missingness, HWE, and MAF indicated by the colour of the bar.

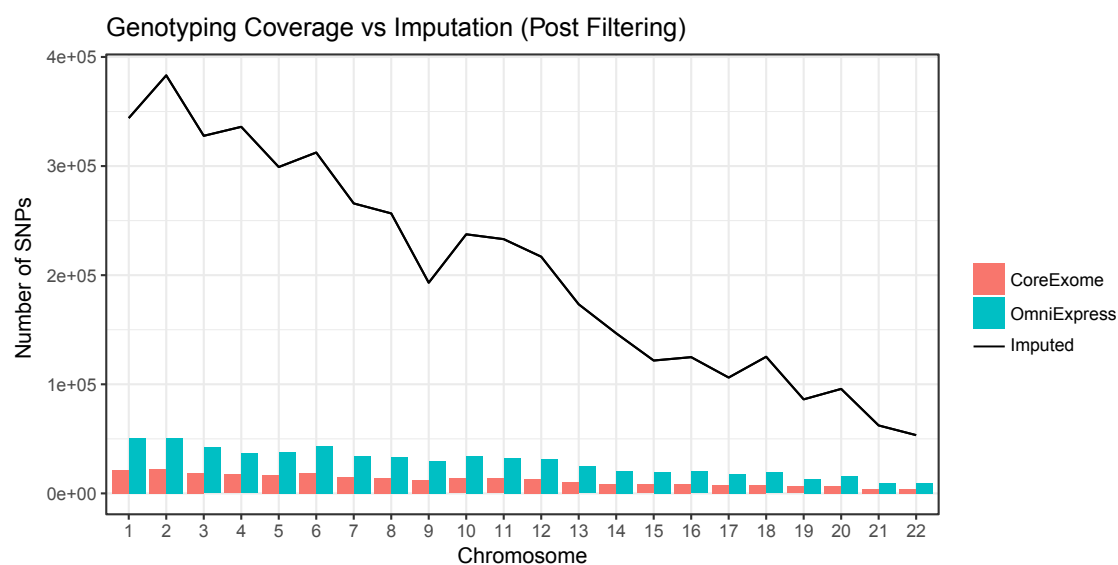


Figure 6.6: Comparison of genotyping and imputation coverage

The number of SNPs per chromosome for each genotyping array is compared to the number of SNPs available in the imputed data post-filtering.

6.6), and the conservative thresholds give high confidence in the resulting data.

The first eight principal components of the combined genotyping data were plotted to identify population stratification (Fig. 6.7). On the basis of these plots, the first six principal components were corrected for in the subsequent eQTL mapping.

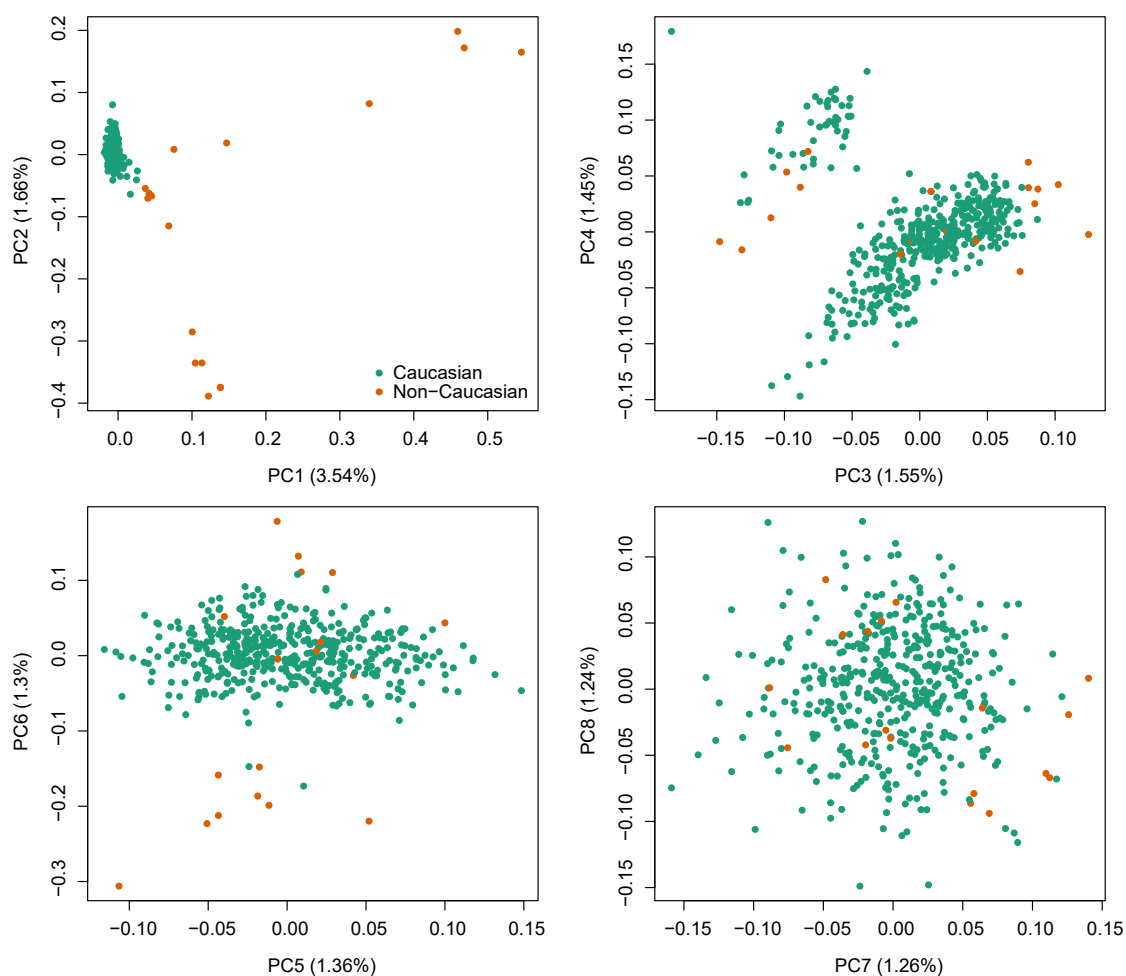


Figure 6.7: Principal component analysis: GAInS genotyping data.

The first eight principal components of the genotyping data for eQTL mapping are plotted, with the amount of variance explained by each noted. Samples are coloured according to Caucasian ethnicity based on correspondence with 1000 Genomes data.

Approach to mapping cis-eQTL in sepsis

As the majority of eSNPs are found within 250kb of the transcriptional start site, only SNPs within this reduced window were tested for local associations

(cis-eQTL). A linear mixed model was used to test each SNP-gene pair for a significant additive effect of genotype. This model included a random effect for patient in order that up to three serial samples (day one, three, and/or five of ICU admission) per patient might be included to increase power. Increasing numbers of gene expression principal components were included in order to identify the set that maximised power to detect eQTL. This question was also explored using a linear model, using only the first available sample from each patient (Fig. 6.8). This demonstrates the increase in power gained from the inclusion of serial samples. From these results, it was decided to include the first 30 gene expression PCs in the final analysis to adjust for systematic effects.

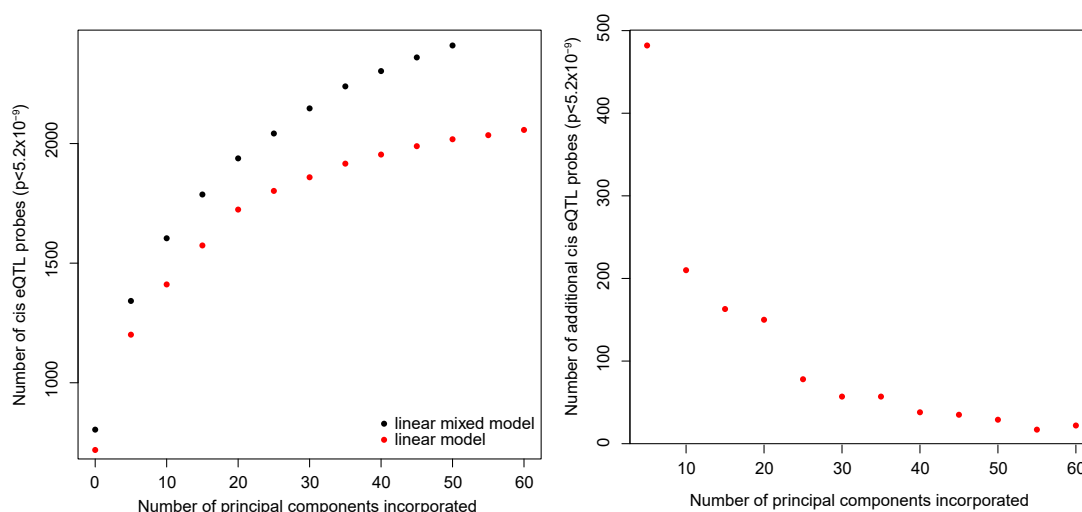


Figure 6.8: The effect of including sepsis gene expression PCs.

eQTL were mapped in the combined sepsis cohort using a linear model (*red*) on the first sample for each patient, and a linear mixed model (*black*) using all samples. Different numbers of gene expression principal components were included to correct for systematic confounding effects. More eQTL were identified when more principal components were incorporated (*left*), but the rate of increase slowed after 30 PCs (*right*).

In total, 9,615,518 tests were performed across 12,974 transcripts. To ensure high confidence in the eQTL identified, a strict Bonferroni multiple comparison correction was applied ($p = 0.05/9615518 = 5.2 \times 10^{-9}$).

Exploration of sepsis cis-eQTL

This identified 190,355 significant SNP-transcript associations, involving 2,147 unique transcripts, and with effect sizes ranging from -1.7 to 2.6 (Fig. 6.9, 6.10). Comparison with the differential expression analysis between septic and non-septic samples (Chapter 3) found that genes with an eQTL in sepsis were enriched for genes differentially expressed in the sepsis response (chi square $p < 2.2 \times 10^{-16}$). The ten most significant eQTL are shown in Table 6.1.

Gene Symbol	SNP	eQTL Beta	P value
<i>CHURC1</i>	rs6745	1.38	6.27×10^{-172}
<i>ZSWIM7</i>	rs2286795	-0.70	6.70×10^{-146}
<i>USMG5</i>	rs7911488	1.82	1.48×10^{-145}
<i>CLEC12A</i>	rs7309256	1.82	6.55×10^{-136}
<i>GLMP</i>	rs949464	0.77	7.32×10^{-136}
<i>ACCS</i>	rs2074038	1.99	8.18×10^{-135}
<i>PAQR6</i>	rs2274226	1.08	4.0×10^{-132}
<i>PPP2R3C</i>	rs57075379	-0.52	3.14×10^{-131}
<i>LIAT1</i>	rs11150881	1.61	4.61×10^{-131}
<i>PHACS</i>	rs2074038	1.78	1.99×10^{-128}

Table 6.1: Peak eQTL in sepsis patients

The top ten most significant eQTL are shown, with the peak SNP-gene pairs, the effect size, and the p value for each. The top 40 eQTL are shown in the appendix (Table D.1).

Comparison to the original CAP eQTL mapping finds good replication in this larger dataset. Despite the increased sample size in this analysis, fewer eQTL are detected than reported in the original cohort. This is due to the more stringent p-value threshold applied here; using an FDR cut-off of 0.05 as before finds eQTL for 8116 transcripts. The majority of the genes that had an eQTL previously reported and that were tested here are replicated at the more stringent p-value threshold (1694/2956), and the Z scores in the two analyses are significantly correlated (Pearson's $r=0.74$). Applying the same FDR <0.05 threshold increases the replication of previously reported eQTL to 99.1%. Given the requirement for high confidence in these initial results for the downstream analyses planned,

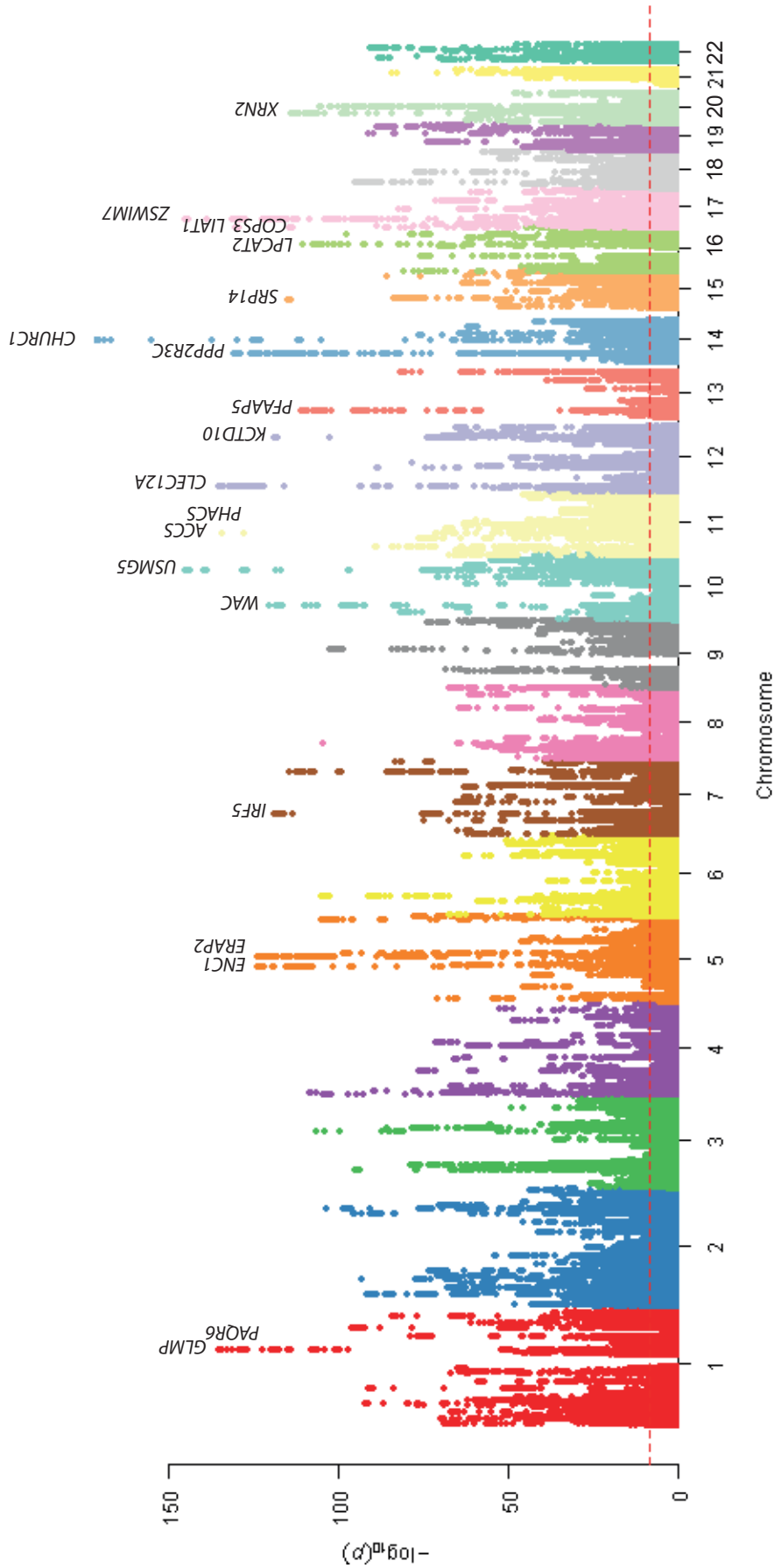


Figure 6.9: Manhattan plot of cis-eQTL in sepsis.

Cis-eQTL were mapped in sepsis for 2,629,291 SNPs and 12,974 transcripts using a linear mixed model. The significance of association for each SNP tested against nearby genes is plotted as $-\log_{10}(p \text{ value})$ against the chromosomal location of the SNP. The Bonferroni-corrected p -value cut-off is indicated by the red line at 5.2×10^{-9} , and the top 20 eGenes are labelled.

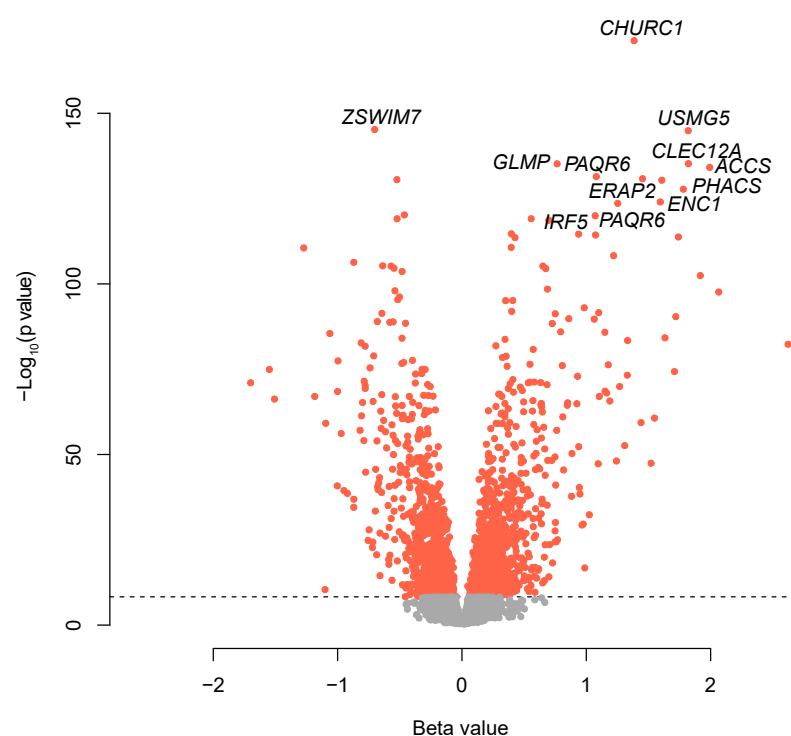


Figure 6.10: Volcano plot of cis-eQTL mapped in sepsis patients.

A volcano plot showing eQTL identified in sepsis patients across 12,974 transcripts. The significance of the peak SNP-gene association for each gene is plotted against the beta value, with a positive beta indicating higher expression with increasing number of copies of the minor allele. eQTL with a p value $< 5.2 \times 10^{-9}$ are plotted in red, and the most significant are labelled.

the stricter correction for multiple testing was used and subsequent findings are based on this set of eQTL.

In terms of individual eQTL, the most significant SNP-gene association in the CAP data was for *IRF5*, an interferon regulatory transcription factor. This example highlights the greater resolution achieved in the new data through the use of imputation. Again, a highly significant eQTL was observed here, with the peak association being with rs4731533 ($p = 2.8 \times 10^{-115}$), an imputed SNP that was tested in the new data only (Fig. 6.11). The peak SNP reported in CAP previously was rs10239340, a genotyped SNP that here has a p value for association of 2.6×10^{-109} . These two SNPs are in strong LD with each other, suggesting that their association with *IRF5* expression reflects the same underlying biological signal. The availability of imputed SNPs in larger numbers of sample permitted

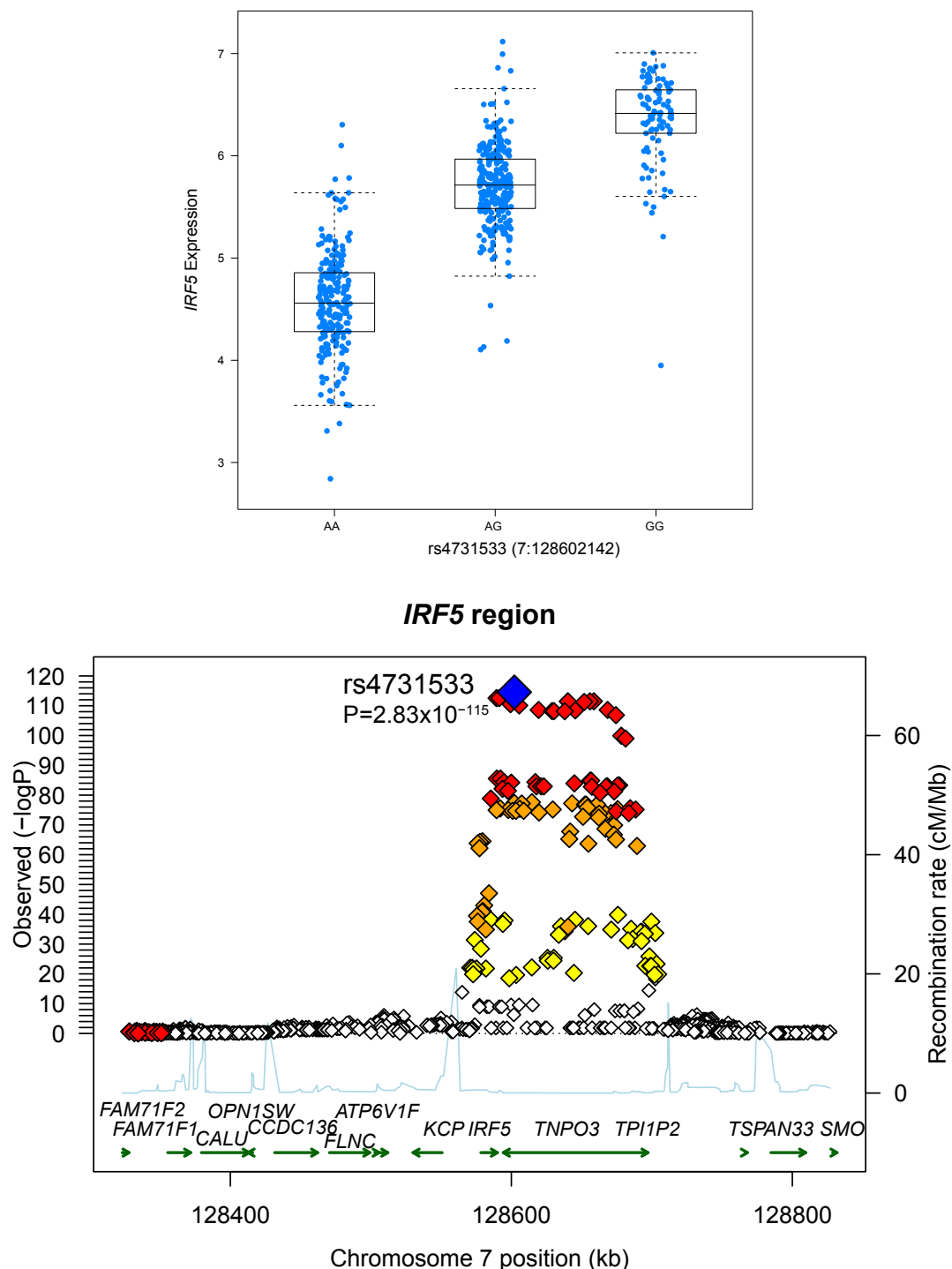


Figure 6.11: Replication of an eQTL for *IRF5*

(a) A boxplot showing the effect of the peak eSNP on expression of *IRF5*. Log gene expression adjusted for the covariates included in the eQTL model is plotted for each sample against the number of copies of the minor allele. (b) A regional association plot illustrating the *IRF5* eQTL. Each point represents a SNP, coloured according to LD with the peak eSNP, and with log p value plotted against genomic location. The genetic recombination map is overlaid, and nearby genes are labelled.

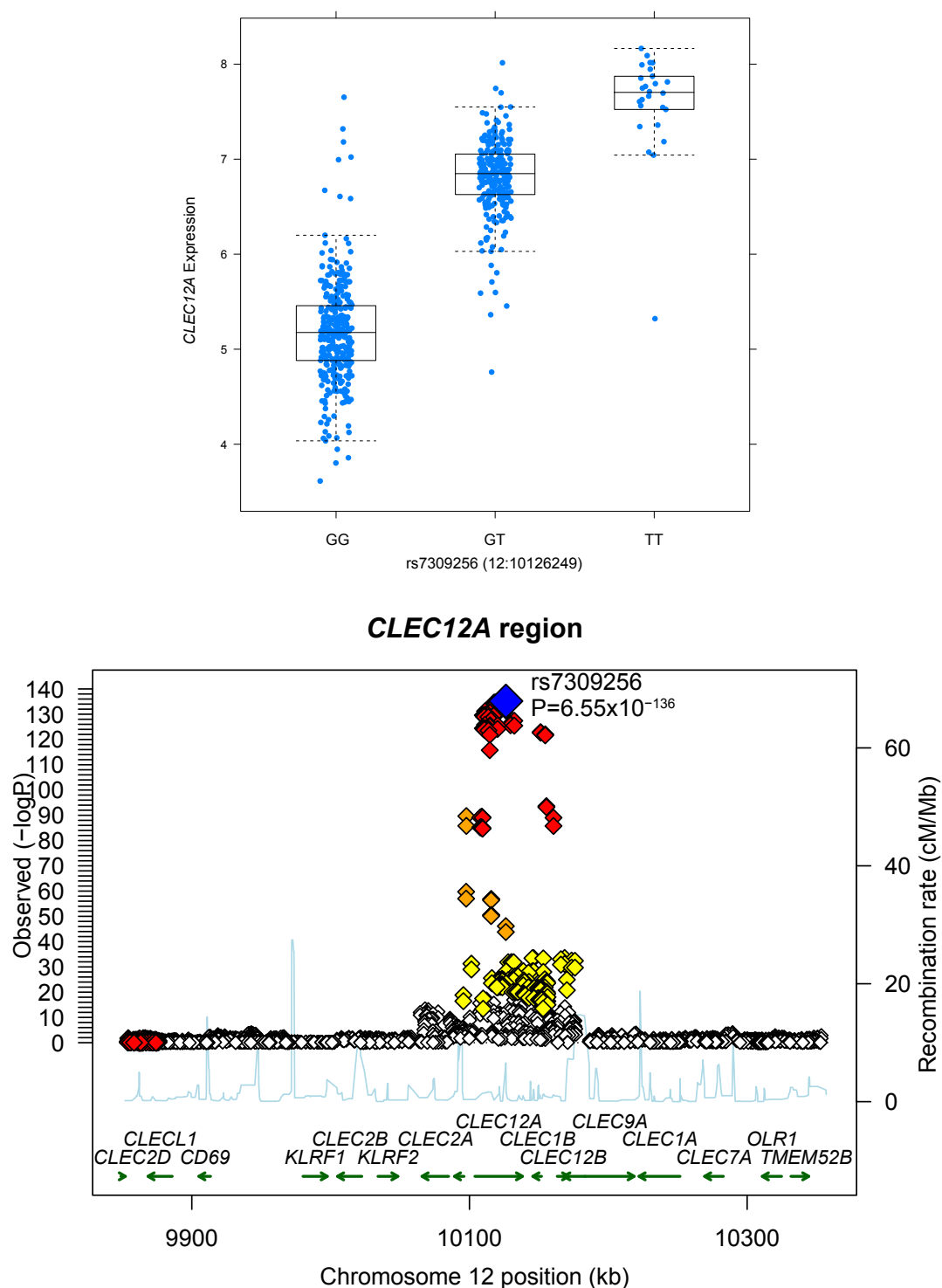


Figure 6.12: An eQTL for *CLEC12A* in leukocytes from sepsis patients.

(a) A boxplot showing the effect of the peak eSNP on expression of *CLEC12A*. Log gene expression adjusted for the covariates included in the eQTL model is plotted for each sample against the number of copies of the minor allele. (b) A regional association plot illustrating the *CLEC12A* eQTL. Each point represents a SNP, coloured according to LD with the peak eSNP, and with log p value plotted against genomic location. The genetic recombination map is overlaid, and nearby genes are labelled.

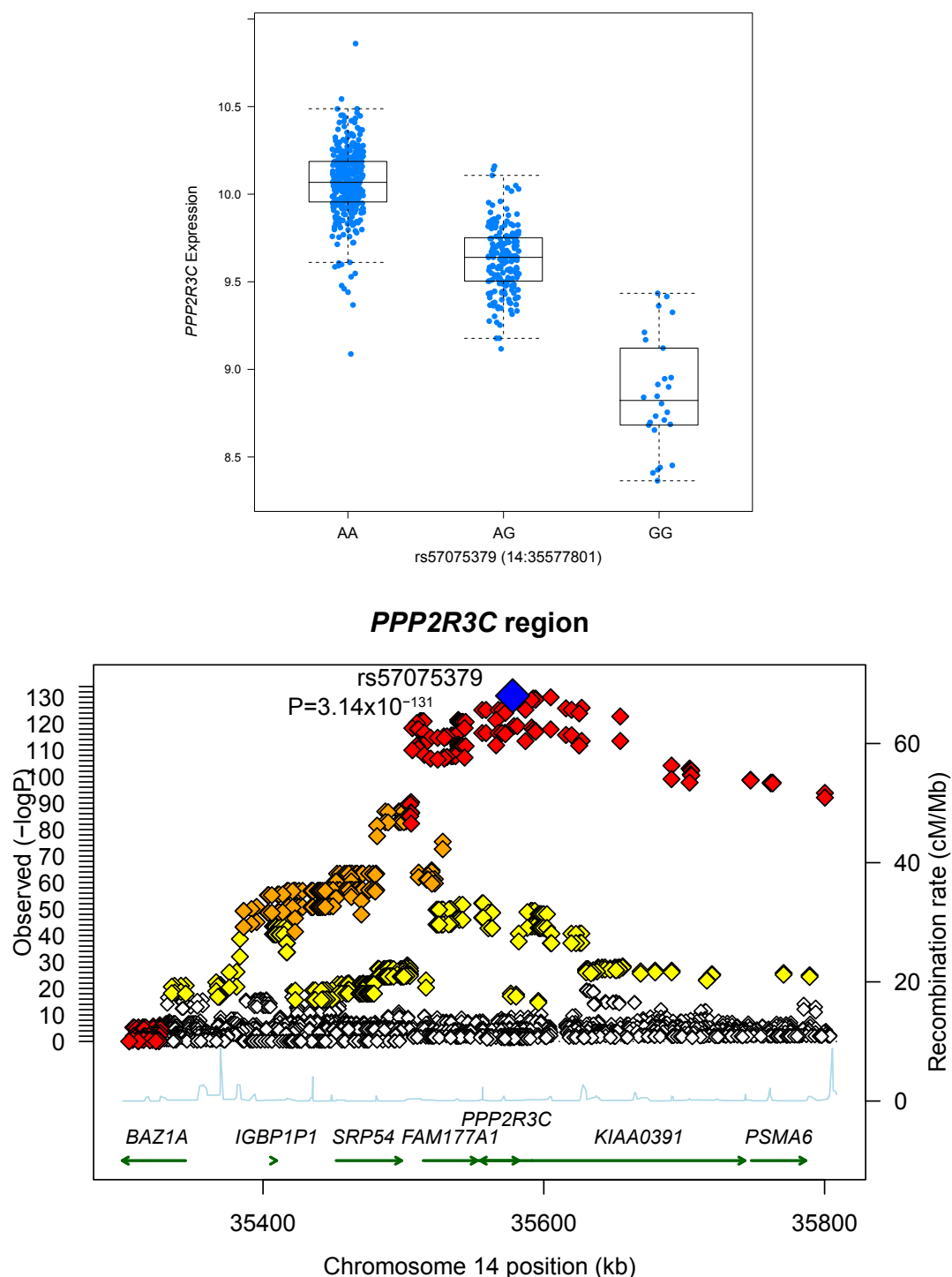


Figure 6.13: An eQTL for *PPP2R3C* in leukocytes from sepsis patients

(a) A boxplot showing the effect of the peak eSNP on expression of *PPP2R3C*. Log gene expression adjusted for the covariates included in the eQTL model is plotted for each sample against the number of copies of the minor allele. (b) A regional association plot illustrating the *PPP2R3C* eQTL. Each point represents a SNP, coloured according to LD with the peak eSNP, and with log p value plotted against genomic location. The genetic recombination map is overlaid, and nearby genes are labelled.

the refinement of this known association, indicating that rs4731533 is more likely to be the causal variant.

Additional examples of eQTL for biologically relevant genes include *CLEC12A*, which is a negative regulator of granulocyte and monocyte function (Fig. 6.12), and *PPP2R3C*, which is suggested to play a role in cell death of B cells (Fig. 6.13). The eGenes were tested for enrichment of GOBP annotations with XGR, including all tested genes as the background to correct for the effect of specific tissue type, but this did not identify any terms with an FDR <0.05. This is unsurprising, as the majority of genes are known to be under genetic control so no relationship with specific biological functions is expected. Conversely, those eQTL specific to the context of sepsis might be more functionally related.

The eSNPs identified may also be tested for enrichment using GWAS traits mapped to the Experimental Factor Ontology. This found enrichment of SNPs associated with autoimmune disease and specifically inflammatory bowel disease (IBD). This likely reflects the context in which eQTL were mapped; regulatory variants that are associated with autoimmune disease are more likely to show activity as eSNPs in disease-relevant cell types such as leukocytes (Fig. 6.14).

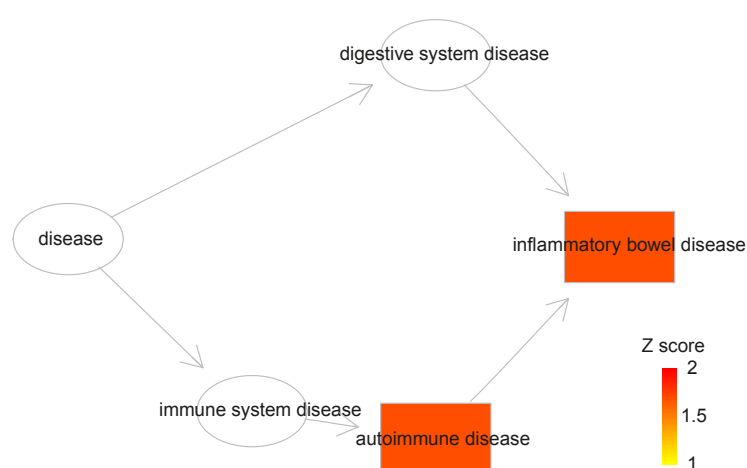


Figure 6.14: Enrichment of Experimental Factor Ontology terms in eSNPs

All sepsis eSNPs were annotated with Experimental Factor Ontology disease terms, and enrichment of those terms calculated. The structure of the ontology is represented as a tree, with the enriched terms highlighted.

Looking at the overlap of variants identified in GWAS with eSNPs mapped in the context of systemic inflammation may enhance understanding of disease mechanisms, particularly for autoimmune and inflammatory disorders. eQTL can help to identify the target gene, as well as providing information about directionality of effects. From the 18,864 genome-wide significant SNP-trait associations ($p \leq 5 \times 10^{-8}$) reported in the NCBI GWAS catalogue (downloaded 11/08/2017; MacArthur et al. (2017)), 1,429 SNPs show evidence of a sepsis eQTL. The association reported with survival in sepsis (Rautanen et al. 2015) involves the *FER* gene, which did not pass gene expression QC and so the mechanism of action could not be explored further. However, overlap with an eQTL can be used to identify GWAS results that did not reach genome-wide significance, potentially due to insufficient power rather than a lack of relevance to disease.

SNP	GWAS p-value	GWAS β	eGene	eQTL β	eQTL p-value
rs10922699	2.71×10^{-4}	1.51	LRRC8C	0.24	2.99×10^{-9}
rs11004435	2.11×10^{-4}	0.70	NCOA4	-0.12	1.13×10^{-24}
rs1061325	3.97×10^{-4}	0.73	CLTCL1	0.13	4.79×10^{-11}
rs1999653	2.73×10^{-4}	1.40	LY86	0.15	4.14×10^{-9}
rs10434903	2.50×10^{-4}	1.40	LY86	0.16	3.23×10^{-9}
rs10434904	2.51×10^{-4}	1.40	LY86	0.16	2.51×10^{-9}
rs2026486	2.53×10^{-4}	1.40	LY86	0.15	3.95×10^{-9}
rs1334709	2.70×10^{-4}	1.40	LY86	0.16	3.23×10^{-9}
rs10900991	1.05×10^{-4}	1.44	LY86	0.16	2.47×10^{-9}
rs10458188	2.52×10^{-4}	1.40	LY86	0.16	2.74×10^{-9}
rs481924	3.00×10^{-4}	0.72	LY86	0.15	4.85×10^{-9}
rs7765802	1.41×10^{-4}	1.43	LY86	0.19	3.21×10^{-14}
rs3804465	1.53×10^{-4}	1.43	LY86	0.19	3.21×10^{-14}

Table 6.2: Overlap of eQTL with sepsis GWAS associations.

The 13 SNPs associated with survival in sepsis ($p\text{-value} \leq 5 \times 10^{-4}$) that also had a significant sepsis eQTL. The GWAS and eQTL summary statistics are given (β =effect size).

SNPs that had been found to be associated with survival in sepsis with a $p\text{-value} \leq 5 \times 10^{-4}$ were interrogated in the eQTL data, and 13 also demonstrated a significant eQTL effect (Table 6.2). The eGenes implicated were *LRRC8C*,

NCOA4, *CLTCL1*, and *LY86*. *CLTCL1* encodes clathrin, which plays a major role in vesicular transport and therefore could affect the cell-mediated immune response (Jans et al. 2016). The eSNP associated with *CLTCL1*, rs1061325, is reported as a missense variant by Ensembl's Variant Effect Predictor. *LY86*, which was associated with ten variants in strong LD, of which rs3804465 and rs7765802 are both annotated as being in regulatory regions. This gene encodes a potential mediator of the innate immune response to LPS (Ogata et al. 2000). This illustrates how eQTL mapped in a disease context may be used initially to prioritise GWAS results, even those which do not reach genome-wide significance, on the basis of biological function.

6.2.3 Interaction of eQTL with context

Interactions with source of sepsis

Environmental factors have been demonstrated to have an effect on transcriptional regulation, and eQTL have previously been reported to be specific to context (Barreiro et al. 2012; Fairfax et al. 2014; Lee et al. 2014). Clinical variables already shown to be associated with changes in gene expression, such as the cause of sepsis and the sepsis response state, might therefore interact with genotype to have a magnifying or dampening effect on a subset of eQTL involved in key pathways. Identification of such modulators could increase understanding of the regulatory genome.

The 2,147 eQTL identified in sepsis were tested for an interaction with environmental factors, namely source of sepsis and SRS group. Firstly, eQTL where the peak eSNP did not have at least two minor allele homozygotes in each clinical subgroup were excluded (n=368 for source of sepsis). Interactions with source of sepsis were then identified for 85 eQTL ($P < 0.01$; Table 6.3, Table D.2). These interactions included eQTL that were more apparent in each of CAP and

Gene Symbol	SNP	eQTL Beta	Interaction Beta	P value
<i>S100P</i>	rs3822262	-0.62	0.32	9.17×10^{-8}
<i>SCCPDH</i>	rs7526237	-0.11	-0.45	8.10×10^{-7}
<i>OAS1</i>	rs10850097	-0.87	0.40	2.18×10^{-6}
<i>UEVLD</i>	rs10766486	-0.10	-0.22	1.38×10^{-5}
<i>LAS2</i>	rs4940331	-0.11	-0.31	1.52×10^{-5}
<i>BARD1</i>	rs6729681	0.52	0.37	4.08×10^{-5}
<i>ALOX5AP</i>	rs12429692	0.16	-0.11	6.01×10^{-5}
<i>IL18RAP</i>	rs917997	-0.45	0.23	8.78×10^{-5}
<i>ZNF641</i>	rs2450988	-0.64	0.19	9.14×10^{-5}
<i>FAM89A</i>	rs6700792	-0.30	-0.41	1.04×10^{-4}

Table 6.3: Top ten interactions with source of sepsis.

The top ten most significant interactions with source of sepsis are shown, with the peak SNP-gene pairs, the effect size for genotype and interaction, and the p value for each.

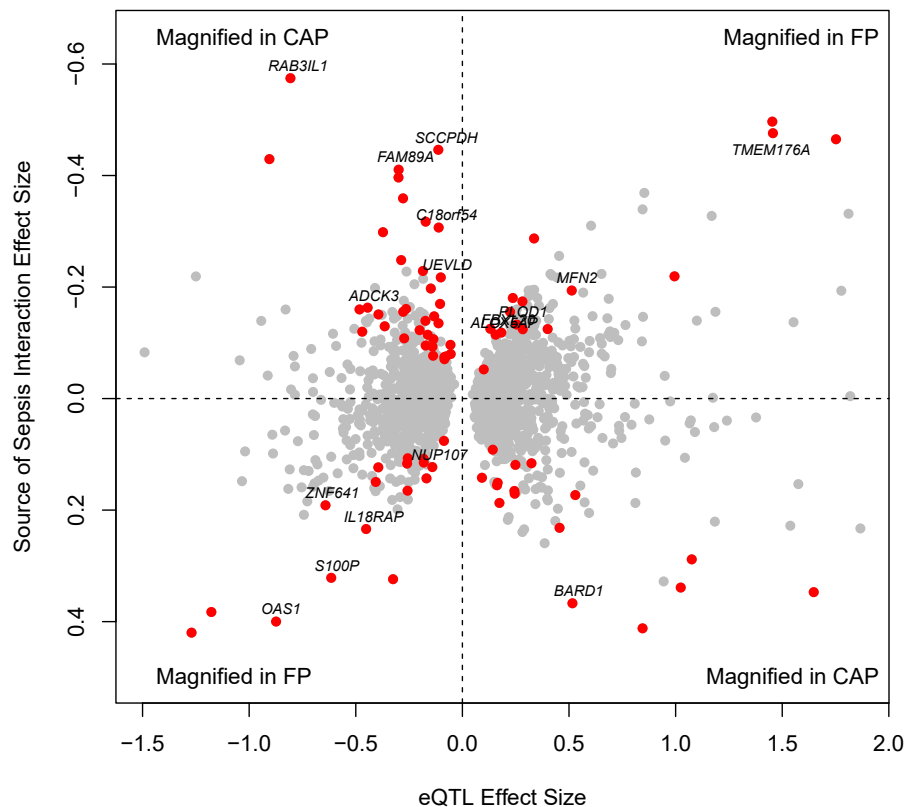


Figure 6.15: Interactions of sepsis eQTL with source of sepsis.

All significant eQTL were tested for an interaction with source of sepsis, *i.e.* CAP or FP. The interaction effect size is plotted against the baseline eQTL effect size, with significant interactions ($p < 0.01$) highlighted in *red* and those with the smallest p values labelled.

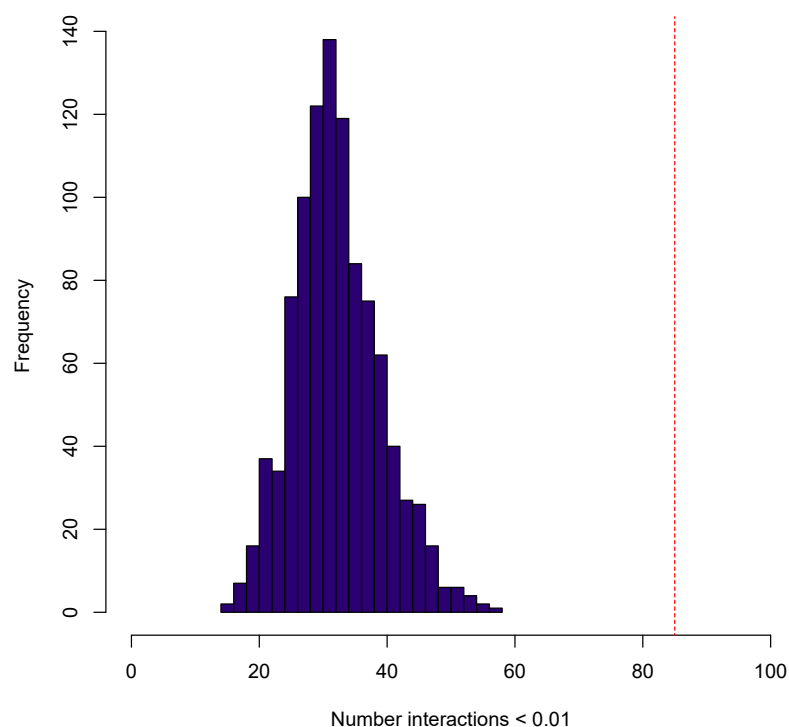


Figure 6.16: Interaction results for permutation of source of sepsis.

The number of significant diagnosis-eQTL interactions is counted for each of 1000 permutations, and the results plotted as a histogram. This shows that the result for the true groups ($n=85$) is significantly greater than would be expected by chance ($p \leq 0.001$).

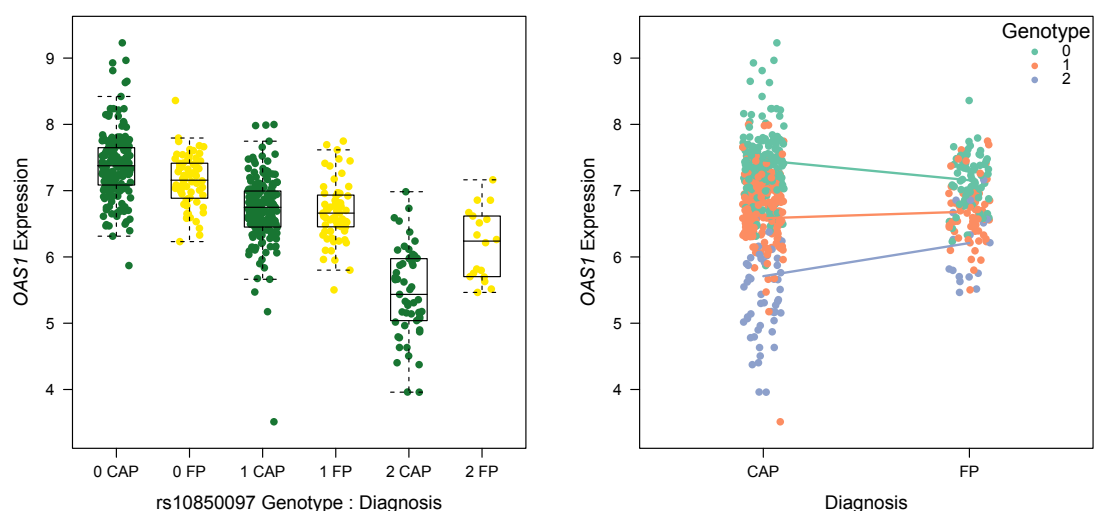


Figure 6.17: Interaction of *OAS1* eQTL with source of sepsis.

(a) The adjusted log expression of *OAS1* is plotted by genotype and source of sepsis, showing the larger effect of rs10850097 genotype in CAP. (b) *OAS1* expression is plotted in CAP and FP, showing the greater variance in CAP. Each sample is coloured according to its genotype, illustrating the larger effect in CAP.

FP (Fig. 6.15). Permutation of the diagnostic labels suggested that this was more interactions than would be expected by chance ($P=1/1000, \leq 0.001$; Fig. 6.16).

One example of an eQTL that was magnified in CAP compared to FP was *OAS1*, an interferon-induced gene related to the viral response ($P_{int}=2.18 \times 10^{-6}$; Fig. 6.17). Notably, eGenes showing an interaction with source of sepsis were enriched for the GOBP term “interferon-gamma-mediated signalling pathway” (FDR=0.013), as in the differential gene expression analysis (background all genes in the eQTL expression dataset). It could be hypothesised that this enrichment reflects the presence of viral infections in a subset of the CAP patients, as viral infection leads to increased interferon gamma signalling. The genes in this pathway might thus be under active regulation in viral CAP patients only, hence the interaction with source of sepsis described.

Interactions with SRS group

Gene Symbol	SNP	eQTL Beta	Interaction Beta	P value
<i>CYP26B1</i>	rs62147573	0.66	1.14	1.10×10^{-22}
<i>TAF1C</i>	rs1804500	0.38	-0.27	5.23×10^{-18}
<i>RAB3IL1</i>	rs2521561	-1.56	1.00	7.33×10^{-17}
<i>DNAJC27</i>	rs916485	0.54	-0.32	6.74×10^{-15}
<i>DDX1</i>	rs6755613	-0.24	0.15	1.23×10^{-14}
<i>AMACR</i>	rs34680	1.01	-0.59	6.75×10^{-14}
<i>MOSPD3</i>	rs7812123	-0.64	0.35	7.08×10^{-14}
<i>BLM</i>	rs389480	1.54	-0.65	1.06×10^{-13}
<i>BARDI</i>	rs6729681	0.86	-0.45	6.32×10^{-13}
<i>FAM89A</i>	rs6698364	-0.75	0.59	1.28×10^{-12}

Table 6.4: Top ten interactions with SRS group.

The top ten most significant interactions with SRS group are shown, with the peak SNP-gene pairs, the effect size for genotype and interaction, and the p value for each.

Interactions with SRS group were also explored. Similarly to the source of sepsis analysis, 232 eQTL were excluded from interaction testing due to a lack of minor allele homozygotes, and 259 eQTL interactions were found with SRS

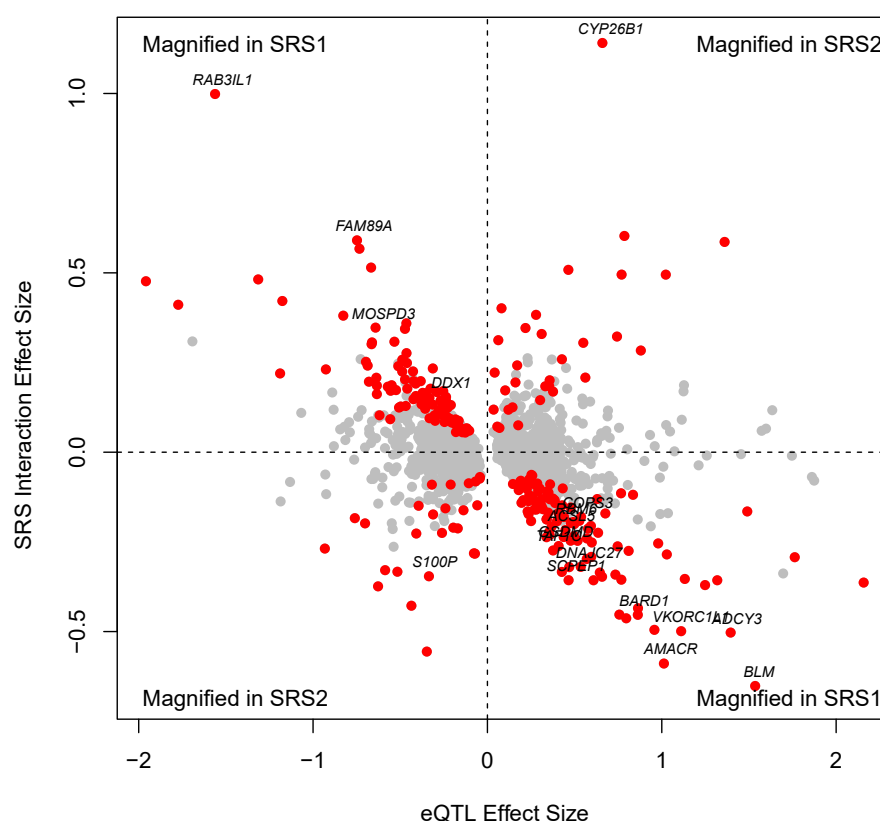


Figure 6.18: Interactions of sepsis eQTL with SRS group.

All significant eQTL were tested for an interaction with SRS group. The interaction effect size is plotted against the baseline eQTL effect size, with significant interactions highlighted in *red* and those with the smallest p values labelled.

group (Table 6.4, Table D.3, Fig. 6.18). Permutation analysis demonstrated that this was more than expected by chance (Fig. 6.19). These eGenes were enriched for the GOBP annotations “triglyceride biosynthetic process” (FDR=0.034), “protein K11-linked ubiquitination” (FDR=0.034), “protein catabolic process” (FDR=0.034), and “regulation of protein kinase activity” (FDR=0.039), with the background all genes in the eQTL expression dataset. These processes relate to the metabolic derangement that has been described in sepsis; for example, the specificity of the *MTOR* eQTL to SRS1 is replicated (8.4×10^{-5}). The link with SRS group is reinforced by the differential expression of a large number of genes relating to metabolism between SRS1 and SRS2.

An example of an eQTL showing an interaction with SRS group is *GSDMD*, which is magnified in SRS1 compared to SRS2 (Fig. 6.20a). *GSDMD* responds

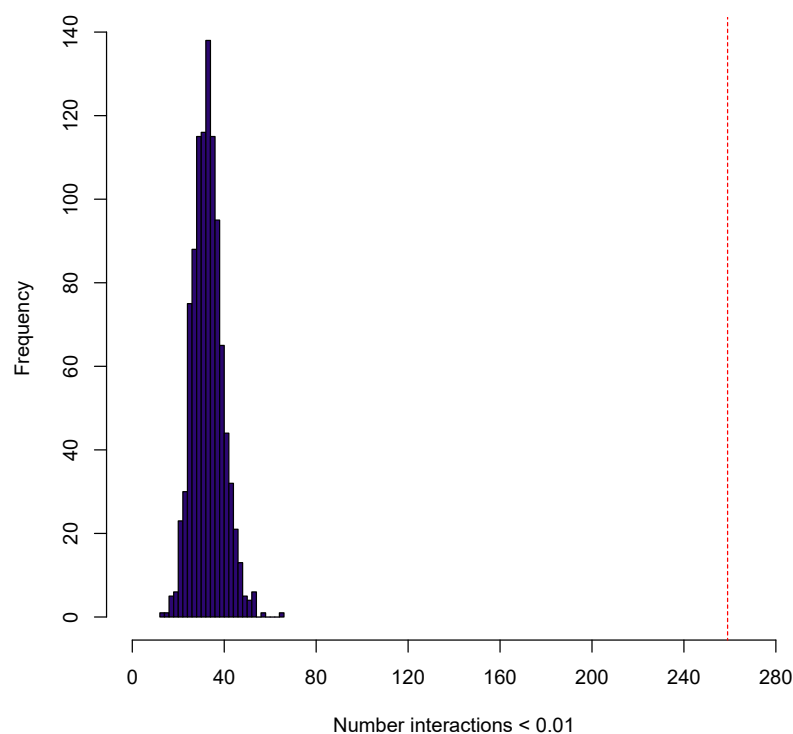


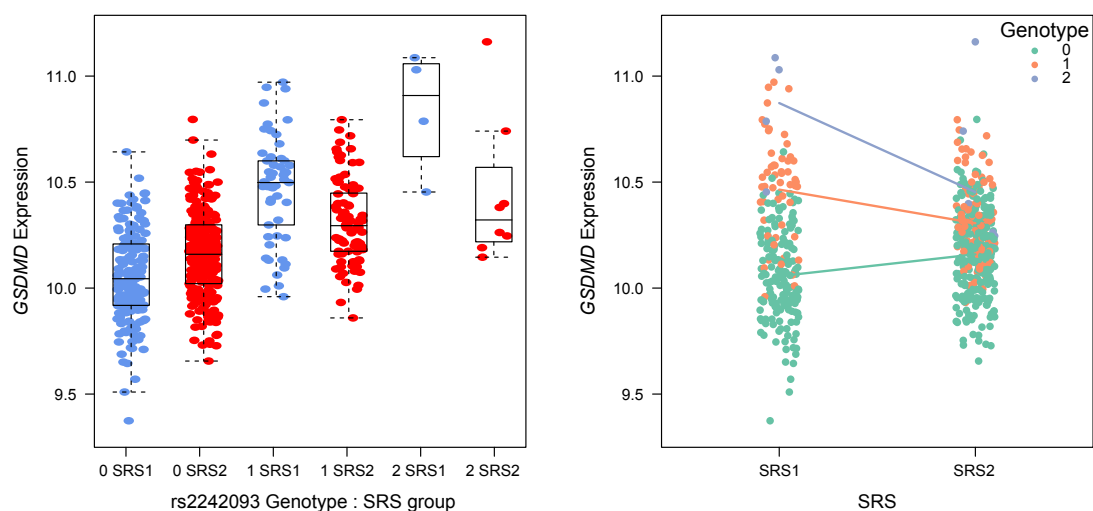
Figure 6.19: Interaction results for permutation of SRS group.

The number of significant SRS-eQTL interactions is counted for each of 1000 permutations, and the results plotted as a histogram. This shows that the result for the true groups ($n=259$) is significantly greater than would be expected by chance ($p \leq 0.001$).

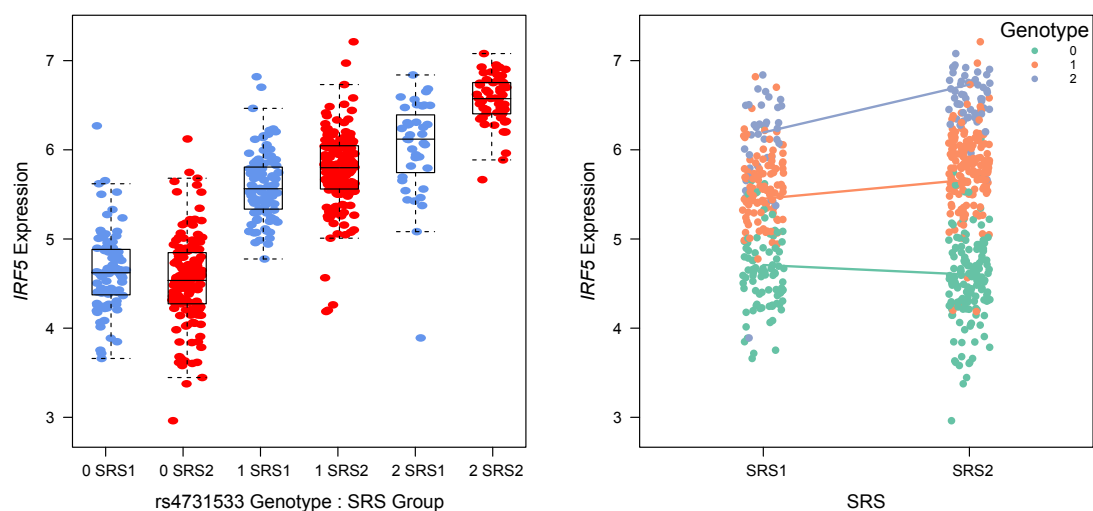
to microbial infection and danger signals, and promotes pyroptosis. Conversely, *IRF5* shows a stronger eQTL in SRS2 (Fig. 6.20b). It is likely that these effects relate to the cellular context in which the eQTL operate; for example, in the more inflammatory SRS2 *IRF5* is more actively upregulated, whereas in SRS1 the prominence of cell death-related pathways may affect *GSDMD* regulation. However, more functional validation would be required to determine the details of the mechanisms involved.

6.2.4 eQTL in cardiac surgery

Genotyping and gene expression data are available for a small number of cardiac surgery patients, enabling limited investigation of regulatory interactions in this context. Genotyping had been carried out by WTCHG Core Genomics for 63



(a) *GSDMD*: magnified in SRS1



(b) *IRF5*: magnified in SRS2

Figure 6.20: Examples of SRS-eQTL interactions.

The log adjusted expression of (a) *GSDMD* and (b) *IRF5* is plotted by genotype and SRS group, showing the larger effect of genotype in SRS1 and SRS2 respectively (*left*). (a) *GSDMD* and (b) *IRF5* expression is plotted in SRS1 and SRS2, with each sample coloured according to its genotype (*right*).

patients using Illumina OmniExpress BeadChips (730,525 SNPs). Genotyping QC was carried out as for the GAIN cohort, and three samples were excluded on the basis of identity by state (IBS) and heterozygosity. In the IBS check, two pairs of samples had been found to be nearly identical genetically, indicating sample mix-ups. Whilst no sample level sex information was available, sex could be inferred for both the genotyping and gene expression data. This identified one sample from each genotyping pair that did not correspond with its supposed corresponding gene expression data in terms of identified sex, and was therefore excluded. Together with the gene expression QC, this resulted in a total of 53 patients with both gene expression and genotyping data available. This cohort was restricted to Caucasian patients, so no genotyping principal components were included in the eQTL mapping.

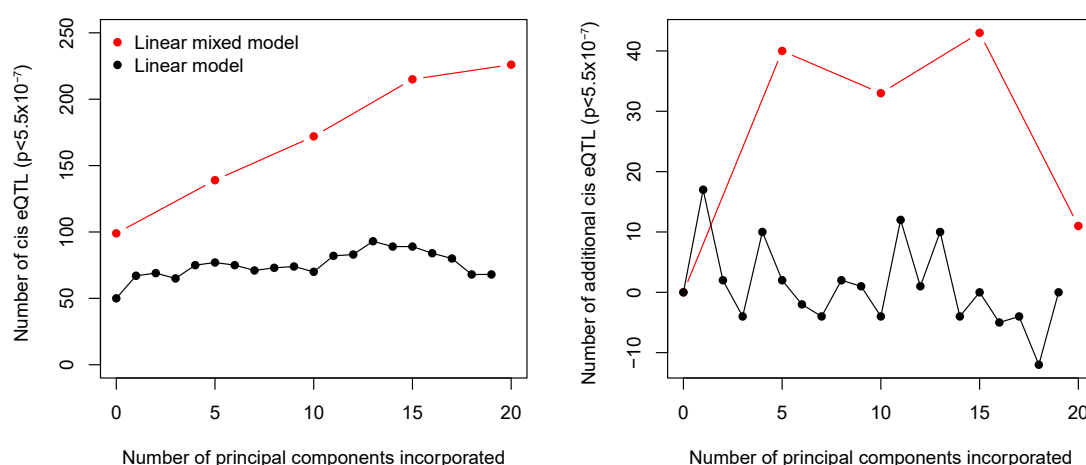


Figure 6.21: The effect of including gene expression PCs.

eQTL were mapped in the cardiac surgery cohort using a linear model (*black*) on the first sample for each patient, and a linear mixed model (*red*) using all samples. Different numbers of gene expression principal components were included to correct for systematic confounding effects. More eQTL were identified when more principal components were incorporated (*left*), but the rate of increase slowed after 15 PCs (*right*).

Given the low sample size, the gene expression data were restricted to the transcripts differentially expressed between the pre- and post-operative samples, thus reducing the number of tests performed in identifying QTL. These transcripts were tested for associations with SNPs within 250kb of

Gene Symbol	SNP	eQTL Beta	P value
<i>LPCAT2</i>	rs11647464	1.14	3.6×10^{-18}
<i>ERAP2</i>	rs7719705	-1.07	5.4×10^{-17}
<i>TACSTD2</i>	rs12118488	2.47	8.4×10^{-16}
<i>LILRA3</i>	rs103294	-1.14	9.2×10^{-16}
<i>MRVI1</i>	rs7115310	0.83	8.3×10^{-15}
<i>LRRC6</i>	rs11988034	-0.78	1.6×10^{-13}
<i>MYOM2</i>	rs10110395	1.40	1.7×10^{-13}
<i>S100P</i>	rs3822262	-1.08	6.1×10^{-13}
<i>OCIAD2</i>	rs2289065	-0.46	1.3×10^{-12}
<i>GSDMB</i>	rs907091	0.26	2.2×10^{-12}

Table 6.5: Peak eQTL in cardiac surgery patients

The top ten most significant eQTL are shown, with the peak SNP-gene pairs, the effect size, and the p value for each.

the transcriptional start site using a linear mixed model. Varying number of gene expression principal components were included in a linear model and a linear mixed model to determine the optimal number to correct for systematic confounding effects (Fig. 6.21). The first 15 gene expression principal components were included in the final model, as well as a random effect for patient. In total, 90,876 tests were performed across 832 transcripts, of which 43 showed evidence of an eQTL ($P < 5.5 \times 10^{-7}$; Fig. 6.22, 6.23, Table D.4).

Given that the eQTL mapping was restricted to genes that were differentially expressed in the surgery response, pathway enrichment analysis was not performed. However, individual genes with an eQTL included a number of biologically interesting genes, such as *ERAP2*, which is as already mentioned involved in MHC class I antigen presentation; and *LPCAT2*, which regulates platelet activating factor in inflammatory cells (Table 6.5, Fig. 6.24).

Cardiac surgery interaction eQTL

The most significant SNP-gene pair for each eQTL was then tested for an interaction with operative status, excluding 11 eQTL where there were fewer



Figure 6.22: Manhattan plot of eQTL in cardiac surgery patients. eQTL were mapped for 70,776 SNPs and 832 transcripts using a linear mixed model. The significance of association for each SNP tested against nearby genes is plotted as $-\log_{10}(p \text{ value})$ against the chromosomal location of the SNP. The Bonferroni-corrected p-value cut-off is indicated by the black line at 5.0×10^{-7} , and the top eGenes are labelled.

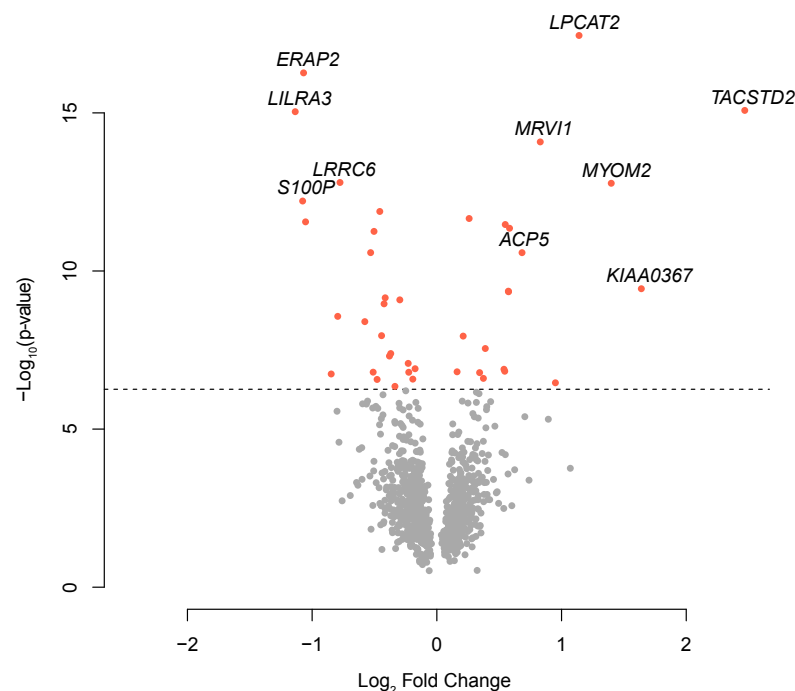


Figure 6.23: Volcano plot of eQTL mapped in cardiac surgery patients

The significance of the peak SNP-gene association for each gene is plotted against the beta value, with a positive beta indicating higher expression with increasing number of copies of the minor allele. eQTL with a p value $< 5.0 \times 10^{-7}$ are plotted in *red*, and the most significant are labelled.

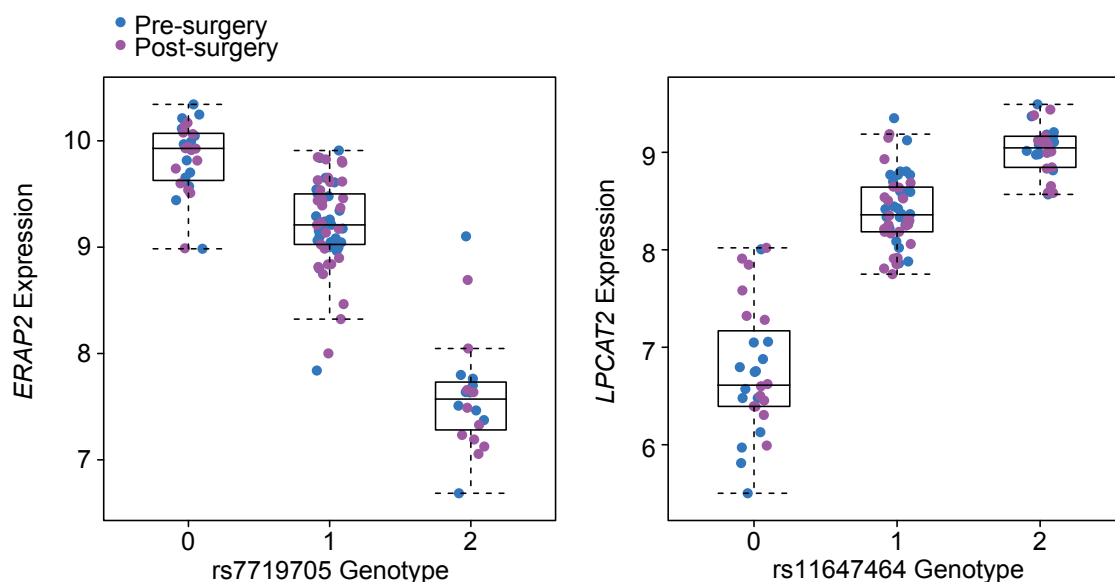


Figure 6.24: Boxplots of eQTL in cardiac surgery patients.

eQTL for *ERAP2* and *LPCAT2* are illustrated in these boxplots. Log gene expression adjusted for expression principal components is plotted by the number of copies of the minor allele for each patient, with the pre-operative samples coloured in *blue* and the post-operative samples in *pink*.

than two minor allele homozygotes. This identified 5 significant interactions ($p < 0.01$), of which four showed a dampening effect on the baseline eQTL with surgery (Fig. 6.25, Table 6.6). Over 1000 permutations of sample type, this many interactions or more were observed 13 times, meaning that this is more interactions than would be expected by chance ($p = 0.013$; Fig. D.1).

Gene Symbol	SNP	eQTL Beta	Interaction Beta	P value
<i>LILRA3</i>	rs103294	-0.83	-0.65	2.4×10^{-3}
<i>LRRC6</i>	rs11988034	-1.0	0.45	3.5×10^{-3}
<i>S100P</i>	rs3822262	-1.38	0.62	2.0×10^{-9}
<i>IL18RAP</i>	rs2110735	-0.66	0.31	2.2×10^{-3}
<i>PADI2</i>	rs2235910	-0.53	0.26	1.4×10^{-3}

Table 6.6: Interaction eQTL in cardiac surgery patients

The five significant interaction eQTL are shown, with the peak SNP-gene pairs, the effect size, and the p value for each.

The dampening of the eQTL post-surgery could indicate that genes involved in the surgery response are subject to strong differential regulation through inflammatory pathways, and this effect is much greater than that of regulatory variation. Examples of genes with such interactions are *IL18RAP*, a subunit of the receptor for the pro-inflammatory cytokine IL18, and *LILRA3*, which acts as a receptor for MHC class I antigens. (Fig. 6.26).

6.2.5 Protein QTL in FP

The previous chapter described a mass spectrometry experiment that generated proteomics data for 59 samples from FP patients. Genotyping data were also available for 49 of these patients, permitting a small scale protein QTL (pQTL) analysis. This sample size is severely limiting for pQTL mapping, but strong effects in specific genes might be detected, particularly if the number of tests performed is small.

As before, SNPs within 250kb of the gene transcriptional start site were tested for

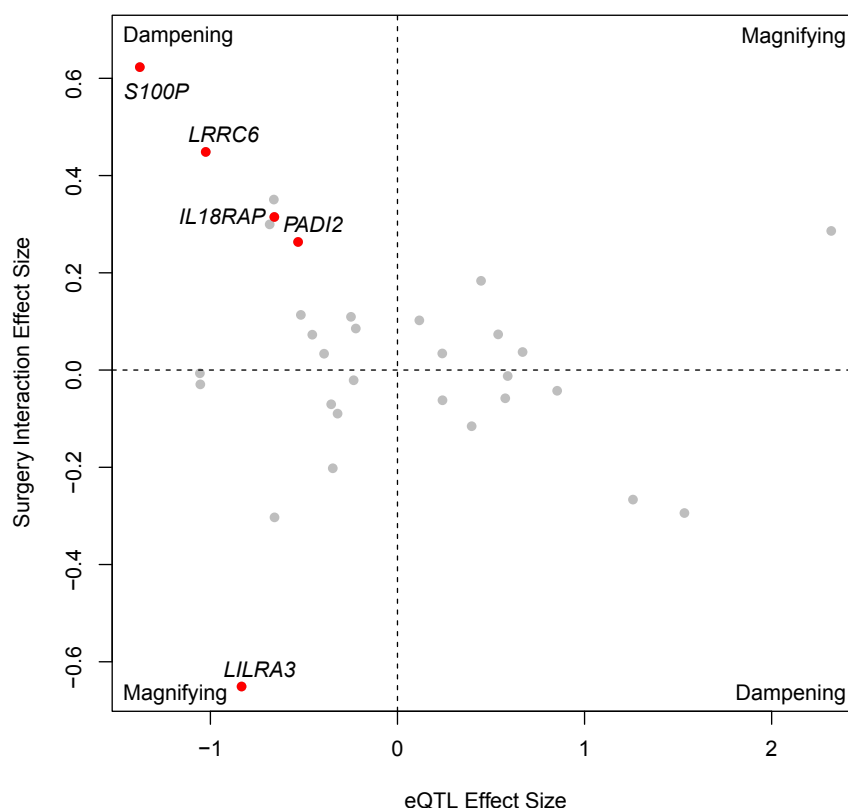


Figure 6.25: Interactions of cardiac surgery eQTL with operative status

All significant baseline eQTL were tested for an interaction with status. The interaction effect size is plotted against the baseline eQTL effect size, with significant interactions highlighted in red and those with the smallest p values labelled.

a pQTL. Unlike in the transcriptomic data, only genotyped SNPs were used and only one sample was included for each patient, so pQTL were identified using a linear model. Based on the output for varying numbers of proteomics PCs, 9 PCs were included in the final model (Fig. D.2). This identified 10 pQTL for 5 species ($p < 2.3 \times 10^{-6}$; Table 6.7, Fig. 6.27).

rs9898 is a missense variant in the HRG gene, perhaps reducing the confidence with which the protein is detected in patients with this variant. No other clear mechanisms were identified to explain the link between these proteins and the associated SNPs, although the peak SNP for the C4A pQTL was found to overlap a CTCF binding site (Variant Effect Predictor), and that for F12 is found in the

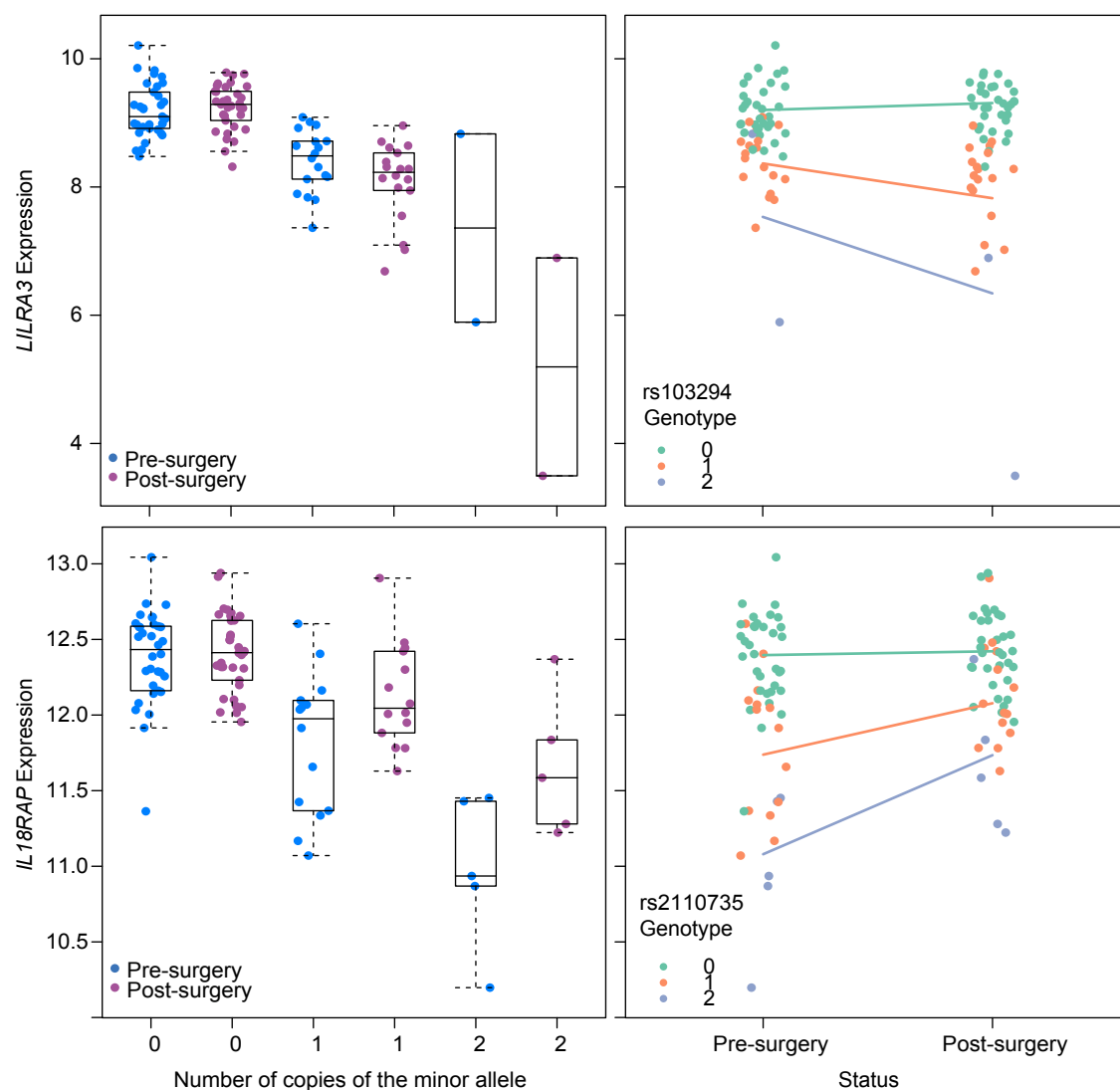


Figure 6.26: Examples of eQTL interacting with operative status

Expression of *LILRA3* (top) and *IL18RAP* (bottom) is plotted by genotype and status, showing the larger effect of rs103294 genotype in post-operative samples and of rs2110735 in pre-operative samples (left). *IL18RAP* and *LILRA3* expression is plotted in pre- and post-operative samples, with each sample coloured according to its genotype, illustrating the interaction effects (right).

Protein	SNP	Effect size	p-value
CFHR3	rs1410996	-0.46	1.56×10^{-9}
F12	rs1801020	-0.21	2.29×10^{-8}
C4A	rs3132935	0.80	1.22×10^{-6}
HRG	rs9898	0.14	1.36×10^{-6}
SAA2-SAA4	rs2045272	-0.18	2.14×10^{-6}

Table 6.7: Association of SNPs with protein abundance.

The effect size and p-values for the significant SNP-protein pairs are shown.

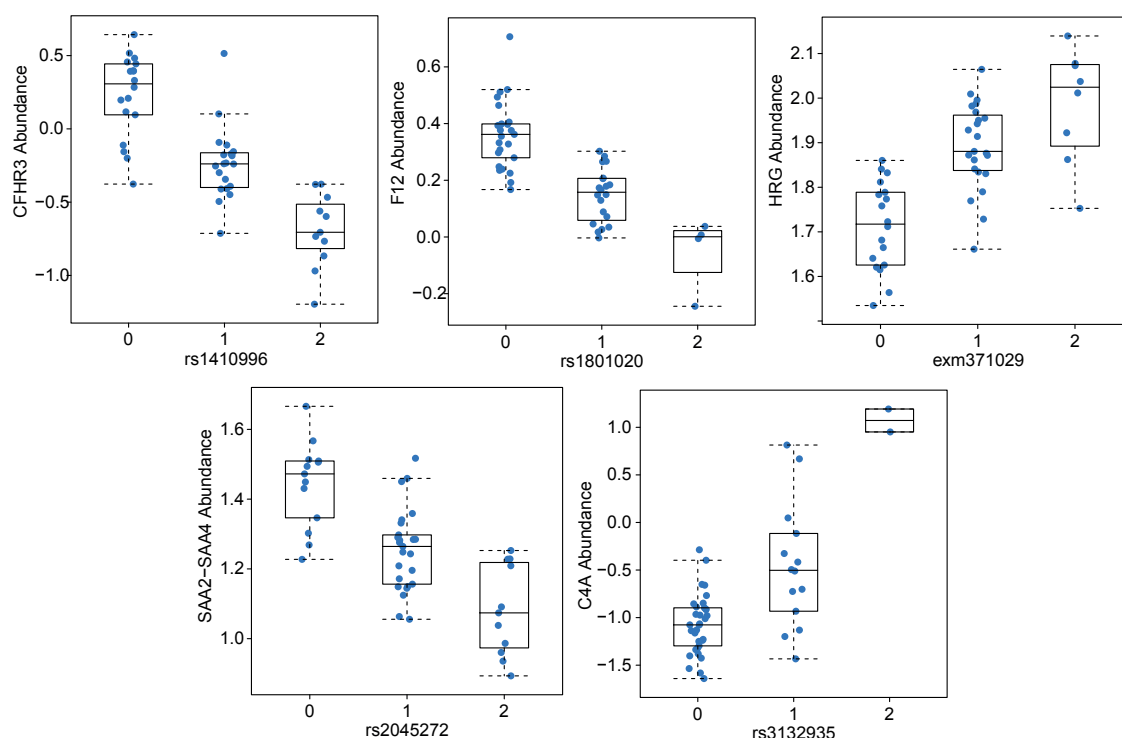


Figure 6.27: Associations between SNPs and protein abundance.

The five significant pQTL are shown in these boxplots. Protein abundance adjusted for principal components is plotted by the number of copies of the minor allele for each patient.

5' UTR and was predicted to have an effect on binding by RegulomeDB. Both proteins are synthesised as precursors and are subsequently modified; pQTL for these proteins could therefore represent differences in their processing rates. As only CoreExome genotyping data were used, it is likely that the pQTL found involve SNPs in LD with regulatory variants.

Limited correlation had been observed between gene expression and protein levels for sepsis patients. There are a number of confounding factors in this comparison, notably the different cell types used (leukocytes vs plasma). Additionally, regulatory variants that act at the protein level could result in dissociation of mRNA and protein concentrations. For these reasons, the analysis was not restricted to SNPs displaying evidence of an eQTL in sepsis. Of the five pQTL detected, only F12 had been tested for an eQTL in sepsis, and no significant associations were found. However, in a larger cohort of healthy volunteers

(Westra et al. 2013) rs1801020 did show a significant eQTL ($FDR < 0.05$) with F12. The minor A allele was associated with lower F12 expression, which is consistent with the pQTL here.

The pQTL identified were related to a number of significant GWAS associations. rs1410996 is reported to be associated with both age related macular degeneration and end-stage coagulation (Williams et al. 2013; Neale et al. 2010). Given that mutations in the *CFHR3* gene have been shown to reduce the risk of macular degeneration, the pQTL observed here with *CFHR3* provides further evidence for these links. Another SNP (rs6677604) showing an pQTL with *CFHR3* is reported for IgA nephropathy, where *CFHR3* has a known role (Xie et al. 2016). Additionally, rs2545801, which shows a significant association with F12, is associated with activated partial thromboplastin time (Tang et al. 2012); an association with F12 abundance has also been previously reported (Suhre et al. 2017). rs9898 is also associated with activated partial thromboplastin time (Tang et al. 2012).

6.2.6 Overlap of eQTL across conditions

In order to explore the effect of context on gene expression regulation, eQTL were compared across data sets. Firstly, the sepsis and cardiac surgery results were compared on the basis of all transcript-SNP pairs tested in both datasets. As the surgery eQTL mapping was restricted to those genes that were differentially expressed in surgery, this was 64,364 pairs. The allele assessed in each was compared to correct for differences in strand in the two datasets, and where this could not be determined (e.g. AT or CG SNPs) the pair was excluded from the comparison.

Z scores were positively correlated between the surgery and sepsis results (Pearson's $r=0.49$; Fig. 6.28), particularly for eQTL significant in surgery

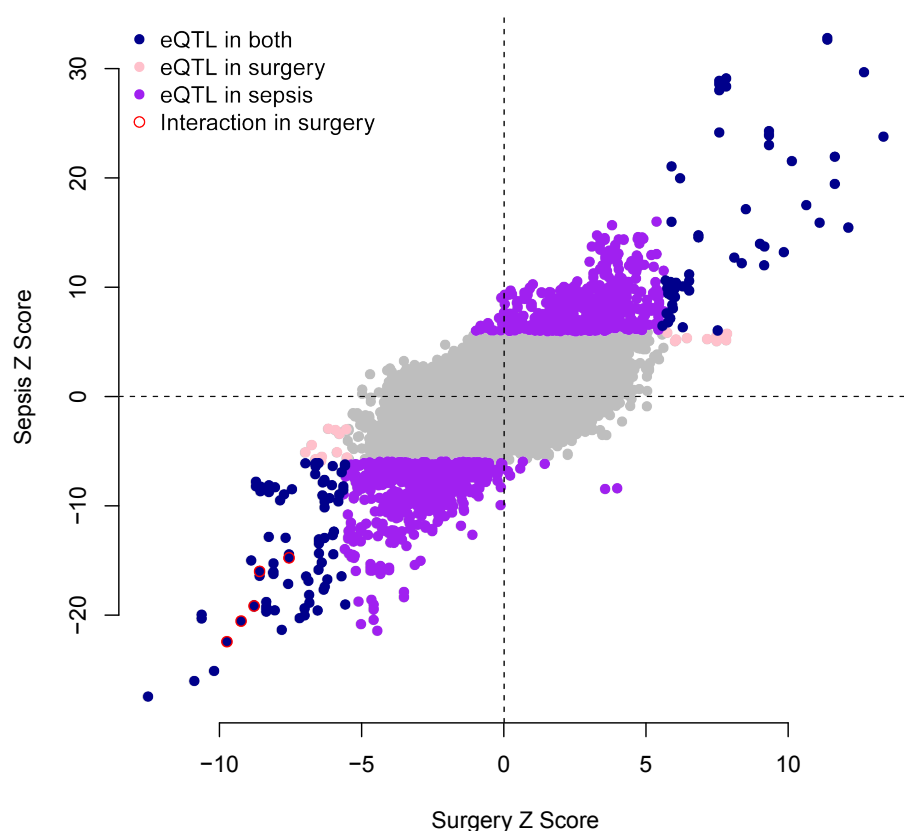


Figure 6.28: Correlation of eQTL Z scores in sepsis and surgery.

For each eQTL, a z score was calculated from the beta value and standard error. These were then compared for transcript-SNP pairs tested in both datasets.

(Pearson's $r=0.92$). All of the 163 eQTL that were significant in both datasets showed the same directionality. Despite the small sample size involved in the case of surgery, 11 eQTL identified in at least one data set had significantly different beta values ($FDR < 0.05$). Three of these were also significantly differentially expressed between sepsis and surgery. Whilst the genes involved did not have clear biological relevance, this illustrates the possible impact of context on transcriptomic regulation.

The sepsis eQTL were then compared to two external data sets where eQTL had been mapped using gene expression data generated using the same microarray platform (Westra et al. 2013; Fairfax et al. 2014). The Westra et al. data comprise eQTL mapped in whole blood from 5,311 healthy individuals, for transcript-SNP pairs with an $FDR < 0.5$. Again, transcript-SNP pairs tested in both sepsis

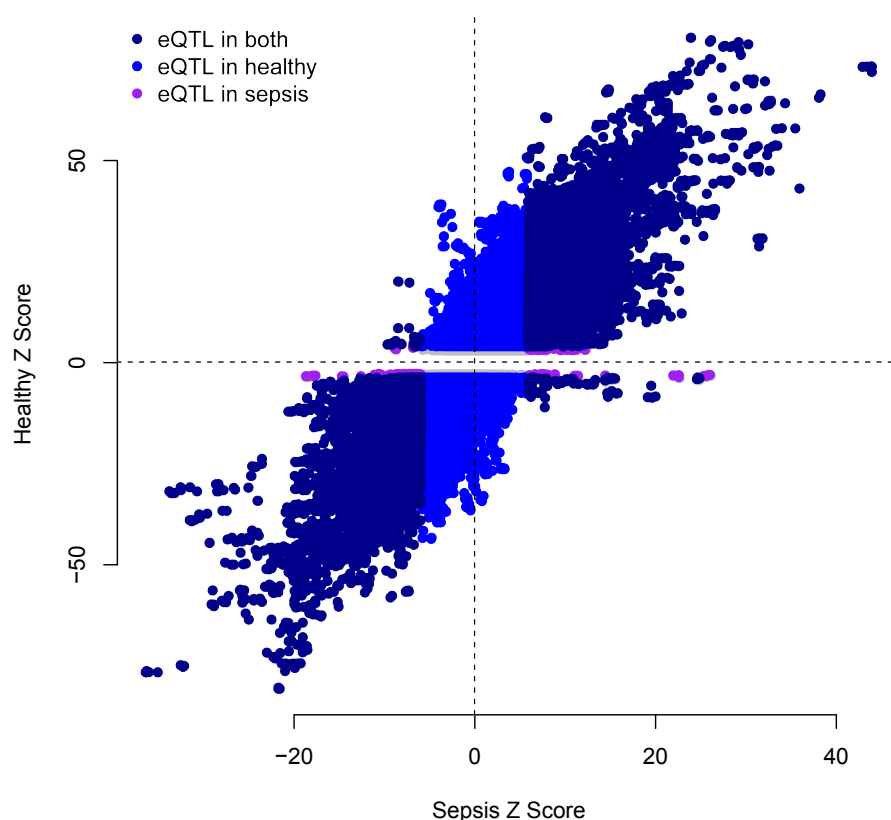


Figure 6.29: Correlation of eQTL Z scores in sepsis and healthy volunteers.

For each eQTL, a z score was calculated from the beta value and standard error. These were then compared for transcript-SNP pairs tested in both datasets..

and healthy volunteers were compared and found to be very strongly correlated (Pearson's $r=0.84$; 0.91 for eQTL pairs significant in both; Fig. 6.29).

Fairfax et al. mapped eQTL in monocytes from healthy volunteers following different stimulations, including LPS. Overall, the results are not strongly correlated with those for sepsis, indicating the prevailing effect of cell type on regulatory interactions. However, 36% of the pairs that were significant following 2 hours of LPS stimulation were also found to be significant in sepsis, compared to 34% of those significant in naïve monocytes (chi square p value for enrichment = 2.6×10^{-5}).

6.2.7 The heritability of SRS

In this thesis, I have described a number of different data types (genotyping, transcriptomic, proteomic), and have integrated these to some extent through analyses such as the eQTL mapping described in this chapter. The SRS subgroups that I described in Chapter 4 are likely to be a result of complex interactions between a variety of genetic and environmental factors, and be part of a wider network of molecular feedback loops. Whilst the sepsis response state of individual patients may be dynamic over time, it is plausible that the initial SRS developed upon infection is at least partially determined by genetic background. Some evidence for this hypothesis comes from the large number of genes that are differentially expressed between SRS and also demonstrate an eQTL.

SRS has been assigned only in patients with genome-wide gene expression data available ($n=552$), and of these, 468 patients also have associated genotyping data. This is a very small sample size in the context of GWAS. Given the complexity of the immune response, any genetic associations are likely to have small effect sizes and therefore are not expected to be detected with this sample size. Some idea of the genetic control of initial SRS state might however be estimated using GCTA. Restricted maximum likelihood analysis was performed to estimate the variance in SRS group assignment that is explained by the genotyping data available. This resulted in an estimate of no genetic effect; however, the GCTA-GREML power calculation suggested that a SNP-heritability greater than zero was highly unlikely with this sample size. Assuming a large true SNP-heritability of 0.5 and using a type 1 error rate of 0.05, the probability of detecting a SNP-heritability larger than 0 was 10.7%.

6.3 Discussion

The immune response to infection is highly complex, requiring the coordinated regulation of large numbers of genes, and is also highly variable across individuals. This variation is known to have a strong genetic component, likely to involve the differential regulation of key genes by specific polymorphisms. In this chapter, I have investigated the effect of genetic variation on the sepsis and surgery responses, focussing particularly on the genetics of gene expression and how this relates to the heterogeneity described in previous chapters.

6.3.1 eQTL methodology

The optimal approach for mapping eQTL is a topic under much debate. Several aspects of the approach used in this chapter were chosen due to the nature of the dataset under examination; however, further refinements should also be explored.

Firstly, the use of a linear mixed model allowed the inclusion of multiple samples from the same patient, which greatly increased the power to detect eQTL. Whilst the combined sepsis cohort comprises a reasonably large number of samples, the effect of increasing sample size further is illustrated by the greater number of eQTL detected in the Westra et al. 2013 data. Clinical studies are generally more limited in the number of patients that can be recruited, due to strict inclusion criteria. Given the availability of up to three samples per patient in the GAINs study, increasing power through the inclusion of serial samples is an appealing alternative.

This approach also enabled the identification of interaction effects, which was particularly important in the cardiac surgery dataset. Those regulatory effects that differed between the pre- and post-operative state were of greatest interest

here; however, mapping eQTL separately in each subgroup would not test directly for an effect of sample status. By using a linear mixed model, power was maximised in the detection of baseline eQTL and in exploring the effect of surgery by using paired samples from each patient. This analysis was even so severely limited by sample size, but illustrated the potential of such an approach in a larger cohort. To reduce the burden of multiple testing in the surgery cohort, the gene expression data were limited to those genes differentially expressed between pre- and post-operative samples, with the hypothesis that such genes would be under active regulation following surgery. However, the interaction effects observed were frequently associated with increased gene expression variance, as opposed to large changes in the mean expression levels. This could indicate that further effects might be observed for genes that are not differentially expressed, but are for example under tighter regulation following an inflammatory insult.

In general, the use of a linear model assumes that genotype has an additive effect on gene expression, which may not apply in all cases. Dominant effects may still be detected with this approach, but this would strongly depend on the effect size and the number of minor allele homozygotes. The latter factor can in general have a large impact on the confidence with which eQTL are identified, and for this reason, an additional filter based on the frequency of minor allele homozygotes was applied prior to interaction testing. Together with the stringent genotyping and imputation filtering, and the p-value threshold applied, this should yield high confidence results on which to perform functional validation. Other methods to adjust for multiple testing have been described; in particular, the approach used by the GTEx Consortium accounts for differences in the number of SNPs tested with each gene (Ardlie et al. 2015; Davis et al. 2016).

Another assumption implicit in the methods used is that the vast majority of local

regulatory variants are restricted to a 250kb window around the transcriptional start site. Whilst this reflects the results of previous studies, this could exclude the possibility of detecting *cis*-regulatory variants located at a greater distance from the TSS. For example, for very large genes SNPs in the 3' UTR might not be tested for an eQTL, despite the known importance of some such variants. Given the relatively restricted power to detect eQTL, a compromise was made between the total number of tests performed and the range of SNPs included.

The use of gene expression principal components as proxies for confounding factors has been described before and is generally accepted (Westra et al. 2013). Nevertheless, there is no clear consensus on the appropriate method to determine the optimal number of PCs to include in eQTL mapping. Inclusion of large numbers of PCs introduces very many variables into the model that are not expected to have an effect, so it is important to restrict this to a set of PCs that describe a reasonably large amount of the variation in the dataset. A secondary concern is that by including PCs in the model, biologically relevant effects may be regressed out. This is particularly pertinent to the analysis of the SRS groups, which are strongly associated with the gene expression PCs. However, only using PCs that show association with known batch effects does not permit the correction of latent factors. As the principal components describe genome-wide variation in gene expression, it is likely that only minor effects would be seen on individual genes.

Finally, the relative lack of overlap with the Fairfax et al. 2014 data reveals the importance of cell type in the regulatory landscape. The contributions of different cells to the associations already described could be deconvoluted to some extent using differential cell count information (Westra et al. 2015). However, further work should consider the profiling of multiple specific cell types from the same patients. The use of RNA-seq would permit more detailed investigation of the regulatory mechanisms involved, for example through

quantification of splicing events.

6.3.2 eQTL in sepsis and surgery

eQTL were identified in sepsis for a large number of immune related genes, which have been reported to show higher heritability for expression. These eQTL were also enriched for genes that were differentially expressed in the sepsis response. They included *CLEC12A*, *PPP2R3C*, and the chemokine receptors *CCR1* and *CCR3*. Additionally, an eQTL was observed for *IL18RAP*, which enhances IL18 binding; IL18 inhibitors have been proposed as a treatment for sepsis. Lastly, an eQTL for *LDHA* (lactate dehydrogenase A) observed in sepsis was also found specifically in monocytes treated with LPS (Fairfax et al. 2014). Serum lactate is used as a diagnostic biomarker in sepsis, and is associated with mortality and organ failure.

These findings suggest how the individual response to infection could be modulated according to the particular combinations of eSNPs active in each patient. Although the immune response state of patients was shown to be dynamic over time, genetic variation between individuals might affect their initial immune response. The enrichment of eQTL in the genes differentially expressed between SRS groups supports the hypothesis that sepsis response state is partially genetically determined. However, this enrichment could be because eGenes must be reasonably able to tolerate variation in expression levels, and so are also more likely to be differentially expressed. The large overlap of sepsis eQTL with GWAS associations is consistent with previous findings in a larger cohort, where 48% of GWAS hits were eQTL in whole blood (Joehanes et al. 2017), and the comparison with the sepsis GWAS demonstrates the utility of eQTL mapping in deciphering the effects of regulatory variants. However, due to LD and the abundance of eQTL identified, it is likely that some of these overlaps are coincidental. Further work should therefore integrate the GWAS and eQTL

results directly whilst accounting for LD (Nica et al. 2010), but this is beyond the scope of this thesis.

Those eQTL that are not present in healthy volunteers or that demonstrate an interaction with inflammatory status are of particular biological interest. The correlation between the eQTL identified in sepsis and in healthy volunteers showed highly significant overlap. However, the comparison was limited to those SNP-gene pairs reported by Westra et al. 2013, namely those with an FDR < 0.5. It was therefore not possible to test directly for differences between all the sepsis and healthy volunteer eQTL.

Ideally, sepsis-specific or inflammation-modulated eQTL would be identified through paired pre-sepsis samples, which are not currently obtainable as patients are recruited to GAinS upon diagnosis with sepsis. The cardiac surgery cohort provides a model for such an analysis, identifying a small number of eQTL that are affected by inflammation. Expansion of this dataset would greatly increase the power to detect such interactions, and given the large overlap between the sepsis and surgery responses, would also be informative for the response to infection. A similar approach has identified genetic factors that influence the response to vaccination (Franco et al. 2013). However, differences between the sepsis and surgery transcriptomes necessitate consideration of the specific effect of infection. As the surgery gene expression data were processed together with the Davenport 2011 GAinS cohort, eQTL could be mapped jointly in sepsis and surgery and interactions investigated, although sample size would again be a limiting factor.

Modulation by an environmental factor can identify key regulatory mediators and relevant pathways in a disease setting. For example, the enrichment of viral response genes in eQTL interacting with source of sepsis indicates the regulatory impact of the pathogen. This enrichment was also observed in the differential expression analysis. Additionally, the GWAS for survival in sepsis found a

significant association only when the cohort was restricted to CAP patients. This suggests a difference in the genetic factors driving mortality in CAP and FP, which could be investigated further in the light of the interaction eQTL.

Interactions with SRS are more complex, but may involve central players in the regulatory networks that drive the differences in the immune response observed. For example, *IRF5* shows evidence of a strong eQTL effect in sepsis, but this is enhanced in SRS2. Even small effects on a transcription factor such as this could result in widespread downstream effects. As more data become available, *trans*-eQTL might be explored in a similar way, and the loci involved could be mapped to known regulatory elements to connect eQTL in regulatory networks. Lastly, other perturbations could be investigated. For example, the action of drugs under development might be better understood through exploration of the genetic background of the patients together with their immune response states.

6.3.3 pQTL in sepsis

Whilst the scope of the pQTL analysis was restricted by the small sample size of this pilot data set, the biological relevance of the proteins involved indicates the potential of such an approach for future work exploring variation in the sepsis plasma proteome. For example, the SAA2-SAA4 protein is the result of naturally occurring transcriptional read through from the *SAA2* gene to the *SAA4* gene. These encode serum amyloid proteins, which associate with high density lipoproteins and are upregulated in inflammation. F12 is a coagulation factor, and C4A is a complement factor. CFHR3 is also involved in complement regulation. Finally, HRG is a histidine-rich glycoprotein that can bind a variety of ligands, such as heme and heparin. It has also been implicated in pathogen clearance, and coagulation.

It has been previously demonstrated that common SNPs are associated with the

abundance of human plasma proteins relevant to inflammation and immunity (Melzer et al. 2008). More recently, pQTL have been mapped on a larger scale, reporting 539 significant associations, of which 57 overlap with risk loci (Suhre et al. 2017). In this cohort, four of the ten significant pQTL identified overlapped with a GWAS hit. The numbers involved are too small to make any meaningful comment on the relative overlap of eQTL and pQTL with GWAS hits, or to compare the effect size of eQTL and pQTL. However, it is plausible that protein abundance is more directly relevant to complex traits than transcript quantity, and so a SNP impacting the former is more likely to overlap a GWAS association than the latter. Similarly to the weak correlation between transcriptomic and proteomic data, comparisons of eQTL and pQTL in model organisms do not always show substantial overlap (Ghazalpour et al. 2011; Foss et al. 2007; Foss et al. 2011). More recent studies in cell lines have identified many SNPs involved in both eQTL and pQTL, although a large number of pQTL were also not linked to mRNA levels (Hause et al. 2014; Battle et al. 2015; Wu et al. 2013). Regardless, the identification of pQTL is a powerful complementary method for improving our understanding of disease risk factors.

6.4 Conclusions

In this chapter, QTL were mapped in the context of sepsis and surgery to explore the interplay of regulatory variation with the gene expression responses and clinical heterogeneity described in previous chapters. Comparisons with GWAS results illustrated how eQTL can be used to generate hypotheses about the mode of action of risk variants, and external datasets describing epigenetic markers provide further evidence for regulatory interactions. Overall, this demonstrates the benefits of exploring the genetics of gene expression, and the utility of integrating multiple -omics datasets.

GENERAL DISCUSSION

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

7.1	The SIRS response	247
7.2	Implications of disease heterogeneity	249
7.3	Integration of multi-omic data	251
7.4	Limitations and future work	252
7.5	Conclusion	258

This thesis has applied an integrated -omics approach to address inter-individual heterogeneity in the immune and inflammatory responses, specifically in the context of sepsis. This has involved the generation of genotyping, gene expression, and proteomics data for a large cohort of sepsis patients with extensive clinical phenotyping. This has allowed me to explore the complexities and extent of variation across individuals in the sepsis response, and to validate a number of results in prospectively recruited patients. The broad findings are discussed below in the context of the current state of sepsis research, and potential areas upon which to focus in the future are suggested.

7.1 The SIRS response

The dysregulated host response to infection, surgery, and trauma is frequently considered a single unified syndrome (Bosmann and Ward 2013). The

stimulation of the same PRRs by a diverse range of PAMPs and DAMPs suggests that the resulting signalling and the subsequent response should be highly comparable, and that there is a common pathway leading to mortality and morbidity. Given the lack of specific treatments available, grouping critically ill patients together could help therapeutic management. Using simple and non-specific criteria promotes early diagnosis, and the supportive measures available are beneficial for all SIRS patients (Prescott et al. 2016).

However, in the goal to develop more specific therapies for sepsis and SIRS, a detailed understanding of the molecular mechanisms of disease may be vital. The comparisons between surgery, trauma, and sepsis patients described in Chapter 3 highlight the complex nature of SIRS, showing both a great deal of overlap in the transcriptomic response and features specific to each subgroup. The cross-talk between immune and inflammatory pathways makes it difficult to predict how a given patient will progress, but the known importance of genetic variation in susceptibility to specific infections suggest that subtleties in the immune response can be extremely important. Moreover, a large number of sepsis patients have developed an infection following surgery or trauma, increasing the difficulty of disentangling the shared and specific aspects of the SIRS response of different causes. This may permit immunomodulatory treatments to be developed that are appropriate for SIRS with and without infection; however, the target of such agents would evidently need to be carefully considered.

In recent years, a view of the SIRS response as simultaneous hyperinflammation and immunosuppression has replaced previous hypotheses describing excessive inflammation being compensated for by an anti-inflammatory response (Hotchkiss and Karl 2003). The importance of the anti-inflammatory aspects of sepsis has been recognised gradually (Deans et al. 2005; Hotchkiss et al. 2013), with the majority of drug development focussing on suppressing pro-inflammatory targets. An association between immunosuppression and

mortality in sepsis has been suggested by previous publications, and is strongly supported by the findings described in Chapter 4. However, whilst immunostimulatory treatments are currently in development, they are not necessarily expected to be appropriate for all patients.

7.2 Implications of disease heterogeneity

There are currently no targeted therapies available for sepsis. The work described in this thesis and in a number of papers (e.g. Wong et al. (2009b) and Scicluna et al. (2017)) demonstrate the extensive heterogeneity in the sepsis response relating to both the infectious process and the host-specific response. It is therefore highly unlikely that a drug targeting a particular aspect of the immune system will benefit all patients, strongly advocating for a precision medicine approach to disease management (Marshall 2014; Cohen et al. 2015).

A recent review from an international panel of experts on the current state of translational research suggests that the future research agenda takes the development of cancer therapeutics as a model (Perner et al. 2017). In cancer, a small number of immunomodulatory drugs have been approved by focussing on specific immunological signatures in subgroups of patients. Such an approach requires an understanding of treatment-relevant disease mechanisms that may differ between patients, the ability to assess these processes in the clinic, and the knowledge and availability of appropriate treatments.

Through the methods described in this thesis and in other work (Wong 2013), more homogeneous patient groups should be defined in sepsis based on biology as well as clinical features. From an understanding of disease mechanisms, treatment response subgroups could be identified (Famous et al. 2017; Wong et al. 2015a) that can be used to stratify patients and also identify patients where particular treatments should be avoided. As discussed in Chapter 4, the SRS

groups show promise in this context. Biomarker guided clinical trials could then be performed, similar to the RAMSES study where only patients with high IL-6 concentration were randomised to anti-TNF (Reinhart et al. 2001). Previous trials of promising drugs that were deemed to be failures due to the lack of a benefit in heterogeneous patient cohorts should be revisited, as they could in fact benefit a subset of individuals with a particular response state. For example, efficacy has been demonstrated for a number of potential treatments in animal models but not in humans. This could reflect the much more homogeneous nature of controlled models of sepsis, and argue for further investigation of such treatments in a patient-specific manner. Finally, it has been suggested that a machine learning approach could be established to select appropriate treatments based on clinical records (Tsoukalas et al. 2015). If such a system were implemented in the clinic, it could also integrate a number of genetic or molecular signatures that have been shown to be informative.

Disease heterogeneity should also be taken into account upstream of clinical translation. The GWAS for survival in sepsis performed by Rautanen et al. (2015) was restricted to CAP patients to increase power to detect risk variants. This suggests that the genetic variants that contribute to mortality may differ according to source of sepsis. The eQTL-environment interactions described in Chapter 6 provide further evidence that an understanding of such variants requires more homogeneous cohorts with precise patient definitions. For example, patients could be stratified by SRS and a GWAS performed in each subgroup in order to identify SNPs modulating sepsis risk in the context of relative hyperinflammation or immunosuppression.

7.3 Integration of multi-omic data

This thesis has integrated several different data types, namely genotyping, gene expression, and protein abundance. This has in particular involved the mapping of eQTL in a disease context, but other integrative analyses included the overlap of eQTL with GWAS results and the correlation of protein and mRNA levels. Additionally, information on the mode of action of specific variants was obtained through querying of databases describing the regulatory landscape using the Variant Effect Predictor (McLaren et al. 2016).

It is difficult to identify the target of regulatory variants, particularly in gene-rich regions. For example, rs2549794 had been reported in Crohn's disease as a highly significant risk variant. Nearby genes include *ERAP1* and *ERAP2*, both of which are strong candidate genes due to their joint role in trimming peptides for MHC class I antigen presentation. *ERAP1* had previously been implicated in ankylosing spondylitis (AS), albeit with a different pattern of variants. Given the clinical overlap between Crohn's and AS it seemed plausible that *ERAP1* was also the causal gene in Crohn's. However, evidence of a mechanistic link between rs2549794 and *ERAP2* came from observation of an eQTL involving the GWAS SNP (Franke et al. 2010).

Overlap was observed between sepsis eQTL and associations of uncertain significance in the sepsis survival GWAS (Rautanen et al. 2015). Associations between risk alleles and the expression of biologically relevant genes in a relevant tissue type and inflammatory context helps to prioritise variants. However, no fine mapping or conditional analyses were performed to demonstrate that the GWAS and eQTL associations were the same signal. Tools such as Regulatory Trait Concordance or Coloc could be used to increase confidence in the role of these variants in sepsis (Nica et al. 2010; Giambartolomei et al. 2014).

In general, full understanding of gene expression regulation and the impact

of regulatory variation requires profiling of all features involved, including epigenetic modifications and any intermediate species that may exist. The analyses described demonstrate the potential of a multi-omics approach, with integration of multiple types of data increasing confidence in the results and providing some evidence for functional effects. This was limited here by the availability of primary samples from patients, and power is expected to restrict findings to relatively large effect sizes. However, the value of this study comes from its focus on a particular patient population where the regulatory landscape is expected to differ significantly from that of healthy individuals. Any results are therefore known to be disease relevant, and findings specific to sepsis are of particular interest.

7.4 Limitations and future work

There are of necessity various limitations to the work described in this thesis, many of which have been discussed specifically at the end of each chapter. The major points that should be considered are summarised here.

Firstly, the inference of cell activity from gene expression results depends upon the assumption that transcript and protein abundance are strongly correlated. However, the relationship between the two is known to be non-linear, and may be disrupted by a number of post-transcriptional effects. Transcript stability and protein degradation may both be subject to regulation, and the function of a protein can be modulated by post-translational modifications. Finally, protein activity is often dependent on the cellular context in which it acts. Initial experiments and analysis to investigate this are described in Chapters 5 and 6, but as discussed therein additional data are required to further explore the combined sepsis transcriptome and proteome.

In particular, the importance of cell type has been mentioned throughout but

was not addressed directly due to limitations in the samples available. The use of the LeukoLOCK system for gene expression and plasma samples for proteomics maximised the number of patients who could be profiled, both from cost considerations and the practicalities of use in a large number of hospitals. However, in order to follow up and refine the findings from this cohort, future recruitment should be modified such that cell separation can be performed. This would only be possible for locally recruited patients due to the time sensitive nature of sample processing, but data for a small number of patients would increase resolution of selected context specific effects.

Additionally, the transcriptomics data were generated by microarray for compatibility and cost effectiveness. However, the Illumina microarray used has since been discontinued, and the use of RNA-seq for future samples would provide much more information on the nature of the transcripts present. For example, in the mass spectrometry data, different protein isoforms may be distinguished and quantified, reflecting processes such as differential splicing and post-translational modifications. Extending this to the gene expression data would reveal the effects of alternative splicing at the level of transcription, and increase our ability to describe the regulation of the expression of particular genes. This could be particularly informative for integration of eQTL and pQTL data; as suggested in Chapter 6, the associations described are difficult to explain in terms of mode of action due to the variety of potential mechanisms.

7.4.1 Functional characterisation

High throughput -omics methods generate large quantities of data in a non-targeted way, enabling unbiased exploration of disease and hypothesis generation. Subsequent work should focus on proving causal mechanistic relationships for specific associations of interest, and determining physiological consequences. In this study in particular, further work is required at an

immunological level to functionally characterise the SRS subgroups, and to follow up genes and proteins of biological interest. An illustration of how this goal may be addressed is given below in the context of metabolic effects.

Stimulation of leukocytes with LPS can be used to induce an inflammatory state, and has been described as a relevant cell model for sepsis (Fresno et al. 2009). An eQTL identified for the *LDHA* gene was selected for validation and functional characterisation due to the known disease relevance of lactate and the metabolic shift to glycolysis observed in sepsis. Healthy volunteers were recruited by genotype, monocytes were extracted and stimulated with LPS. *LDHA* expression was measured by qPCR, and the association between genotype and transcript quantity was replicated (Mark Smyth, data not shown). Although lactate was weakly correlated with *LDHA* expression on the day of measurement (Pearson's $r=0.44$), no association was seen between lactate and the eSNP.

Metabolism was an important feature of the differences between the SRS groups and has been implicated in trained immunity (Cheng et al. 2014). Further investigation might involve a cell model of endotoxin tolerance, parallel measurement of epigenetic markers, gene expression, and protein abundance for specific cytokines, and assays relating to metabolic activity such as lactate quantification by colorimetric assay, FACS to compare glucose content, and cellular oxygen consumption analysis. Experimental perturbation of gene expression and repeating these measurements could demonstrate the effects of particular eQTL. Metabolomics may also be carried out using stored samples from GAINs patients to explore the metabolic output of the cell, and mass cytometry might be applied to study particular signalling networks and the immune state.

More generally, animal models may be used to investigate the effects of specific genes. For example, caecal ligation and puncture, or lung contusion followed by bacterial inoculation may be used in mice to model sepsis due to FP and

CAP respectively. The GWAS association between *FER* and survival in CAP was investigated in mice, and transient overexpression of *FER* was found to improve survival rates through the recruitment of inflammatory monocytes and upregulation of antimicrobial genes (Dolgachev et al. 2016). The less strongly associated GWAS SNPs that were also eSNPs in sepsis might be investigated in a similar way.

7.4.2 Microbiology and metagenomics

Efforts to understand the impact of pathogen on disease heterogeneity are restricted by the limited understanding of microbiological aetiology. In the current GAINs cohort, 52% of the individuals with CAP have had no causative pathogen identified. Additionally, the gene expression results suggest a large proportion of these patients have an undetected viral infection, potentially alongside a bacterial infection, with a notable impact on the host response.

Work is ongoing to address the limitations of pathogen identification in sepsis and in GAINs in particular, through a combined high throughput sequencing and digital PCR approach on patient plasma. Better definition of infectious aetiology in this cohort would allow the transcriptomic data in particular to be used to explore the effect of pathogen on the sepsis response. In the clinical setting, more rapid diagnosis could enable more effective administration of appropriate antimicrobial therapies, which with the advent of widespread antibiotic resistance is increasingly important.

7.4.3 Epigenetics

Transient changes in chromatin structure and DNA methylation have been shown to play a major role in the induction of the inflammatory response (Bayarsaihan 2011). Widespread epigenetic remodelling is therefore likely

to accompany widespread differential gene regulation observed at the onset of sepsis. Additionally, sepsis survivors have been found to display immune reprogramming and increased mortality up to five years post-recovery, suggesting that sepsis induces long-term epigenetic changes (Quartin et al. 1997).

Profiling of genome-wide epigenetic marks would therefore be of great interest in understanding the regulation of gene expression in response to an infection. This could include assaying DNA methylation by microarray or bisulphite sequencing, chromatin immunoprecipitation for various histone modifications, and ATAC-seq to identify regions of open chromatin. Generation of such epigenetic maps in parallel with gene expression data for the same patient and time points would be particularly powerful. Use of serial samples would allow the investigation of the time frame of epigenetic modification in sepsis and after recovery, and ideally a pre-sepsis sample would also be profiled for comparison. Whilst the latter presents difficulties in terms of patient recruitment, it might be possible for a small number of patients through recruitment of high-risk patients and collaboration with other studies. For instance, BioAid is a multi-centre project aiming to recruit 5000-10000 adults with infectious disease, who could be followed throughout their illness. Importantly, RNA and plasma samples would be taken upon presentation, permitting comparisons with subsequent samples for those patients who develop sepsis (www.uclhospitals.brc.nihr.ac.uk/research/programmes/infection-immunity-and-inflammation-programme/funding-and-impact-iii).

In light of studies demonstrating epigenetic control of endotoxin tolerance (Chen et al. 2009; El Gazzar et al. 2008), such data would be particularly interesting in the context of the SRS groups. Polymorphisms that lead to differences in this regulatory landscape may contribute to inter-individual variation in response to infection (Pacis et al. 2014). Therefore epigenetic information could be integrated with GWAS, gene expression, and eQTL data to explore the impact of regulatory

variants in sepsis. Characterisation of epigenetic marks and non-coding RNAs may also identify factors affecting long-term survival following sepsis.

7.4.4 Clinical translation

The most promising finding in terms of translation into clinical use are the sepsis response signatures. In order for the gene expression classifiers to be used to stratify patients, further understanding of the biology is needed, and in particular the relationship between the immune state and the response to treatment.

An ongoing collaboration with Prof. Anthony Gordon (Imperial College, London) is investigating the SRS groups in a clinical trial cohort. **V**Asopressin versus **N**oradrenaline as **I**nitial therapy in **S**epsis **H**ock (VANISH) was a randomised, double-blind, placebo-controlled multi-centre trial, enrolling 400 patients with septic shock. The effect of vasopressin on renal failure was compared to that of noradrenaline, with a second level of randomisation to corticosteroids. It was concluded that there was no benefit on the primary outcome measure of number of days without renal failure (Gordon et al. 2016). However, it is plausible that substructure within the patient cohort had an impact on treatment efficacy and thereby confounded the results of the trial.

In order to explore the possibility that SRS group (or other subgroups within the cohort) could affect drug response, gene expression data has been generated by Prof. Gordon for 178 patients. This will allow the assignment of patients to SRS1 and SRS2, and the investigation of interactions between SRS group and treatment outcomes. Preliminary results suggest that such interactions may exist, but are complex and will require large clinical trial cohorts to fully understand.

Moreover, for use in the clinic an assay would need to be developed that is accurate, rapid and simple to use, and inexpensive. In general, advances

in technology present novel opportunities for the introduction of precision medicine for the treatment of sepsis, which could be based on a number of different stratifiers. For example, the Nanopore MinION device may be used to sequence microbial DNA and RNA, with automated sample preparation also available. NanoString allows multiplexed measurement of specific RNAs for up to 800 genes simultaneously, and the Sequenom MassARRAY platform could be customised to genotype a set of relevant markers.

7.5 Conclusion

The overall aim of this thesis was to explore the sepsis response using a functional genomics approach. In doing so, I have greatly expanded a resource of -omics data and an associated sample bank for future work. The integrated approach I have used here has enhanced understanding of the sepsis response, in particular in terms of disease heterogeneity and demonstrating the impact of source of sepsis or SIRS, differences in disease progression, and host genetics on the dysregulated immune response. These findings highlight opportunities for the development of targeted therapies and personalised medicine in sepsis.



THE TRANSCRIPTOMIC RESPONSE TO SEPSIS AND RELATED CONDITIONS

Clinical covariate	P Value	FDR	CAP	FP	
Sex	0.523	0.777	0.51	0.44	NS
Activated protein C	0.239	0.495	0.10	0.03	NS
Corticosteroid	0.216	0.490	0.16	0.27	NS
Renal failure	0.929	0.963	0.21	0.19	NS
Renal replacement therapy	1.000	1.000	0.08	0.08	NS
Ventilation	0.051	0.164	0.70	0.53	NS
Inotropes	0.735	0.867	0.48	0.45	NS
Comorbidities: Cardiac	0.797	0.867	0.40	0.48	NS
Comorbidities: Respiratory	0.160	0.403	0.34	0.25	NS
Comorbidities: Renal	0.366	0.496	0.05	0.11	NS
Comorbidities: Neurological	0.796	0.867	0.10	0.13	NS
Comorbidities: GI	0.002	0.020	0.04	0.27	**
Comorbidities: Immune	0.103	0.297	0.11	0.25	NS
Comorbidities: Infection	0.706	0.867	0.05	0.05	NS
Comorbidities: Endocrine	0.178	0.434	0.18	0.11	NS

Table A.1: Comparison of discrete clinical covariates between CAP and FP.

Discrete variables were compared by Chi-square test. The proportion for each is given for CAP and FP.

Clinical covariate	P Value	FDR	CAP	FP	
Mortality 14 days	0.768	0.867	0.11	0.09	NS
Mortality 28 days	0.414	0.673	0.18	0.13	NS
Mortality 6 months	0.384	0.666	0.23	0.17	NS

Table A.2: Comparison of survival between CAP and FP

Survival was compared using the log rank test. The mortality rate at each timepoint is given for CAP and FP.

Clinical covariate	P Value	FDR	CAP	FP	
Age	0.116	0.320	60.6	65.1	NS
Temperature (low)	0.018	0.088	36.4	36.1	*
Temperature (high)	0.050	0.164	37.9	37.6	NS
Systolic blood pressure (low)	0.600	0.828	114.7	117.9	NS
Systolic blood pressure (high)	0.591	0.828	128.6	131.6	NS
Mean arterial pressure (low)	0.226	0.490	65.0	67.7	NS
Mean arterial pressure (high)	0.728	0.867	99.8	98.8	NS
Heart rate (low)	0.422	0.673	82.1	79.9	NS
Heart rate (high)	0.652	0.840	111.9	113.5	NS
White cell count (high)	0.631	0.832	15.0	14.4	NS
White cell count (low)	0.429	0.673	12.9	11.9	NS
Platelets	0.390	0.666	226.3	247.1	NS
Haematocrit	4.41×10^{-6}	8.39×10^{-5}	35.4	30.0	***
Polynucleocytes (raw)	0.493	0.752	11.7	12.5	NS
Mononucleocytes (raw)	0.760	0.867	0.8	0.8	NS
Lymphocytes (raw)	0.016	0.083	1.1	0.8	*
Polynucleocytes (proportion)	0.029	0.111	84.10%	87.98%	*
Mononucleocytes (proportion)	0.308	0.595	6.73%	5.50%	NS
Lymphocytes (proportion)	0.008	0.059	9.17%	6.52%	**
Sodium (low)	0.961	0.978	135.8	135.7	NS
Sodium (high)	0.618	0.832	139.5	139.0	NS
Potassium low	0.096	0.295	3.7	3.9	NS
Potassium high	0.752	0.867	4.5	4.5	NS
pH	0.378	0.666	7.3	7.3	NS
Bicarbonate	0.008	0.059	25.0	22.1	**
Bilirubin	0.908	0.867	16.1	15.3	NS
Renal support (days)	0.131	0.345	1.1	0.4	NS
Urea	0.538	0.780	10.9	11.8	NS
Urine Volume	0.014	0.081	2091	1548	*
Creatinine high	0.352	0.658	110.7	124.6	NS
Creatinine low	0.228	0.490	99.9	116.0	NS
Respiratory support (days)	0.011	0.069	9.6	4.7	*
Respiratory rate	1.32×10^{-8}	7.65×10^{-7}	29.2	20.0	***
Oxygenation Index	9.45×10^{-8}	2.74×10^{-6}	19.3	30.9	***
PCO2	5.17×10^{-4}	7.50×10^{-3}	5.9	5.1	***
Inotropic support (days)	0.902	0.951	3.1	3.0	NS
SOFA	0.031	0.113	6.8	5.4	*
APACHE	5.10×10^{-3}	4.90×10^{-2}	18.3	15.3	**

Table A.3: Comparison of continuous clinical covariates between CAP and FP

Continuous variables were compared by t-test. The mean value for each is given for CAP and FP.

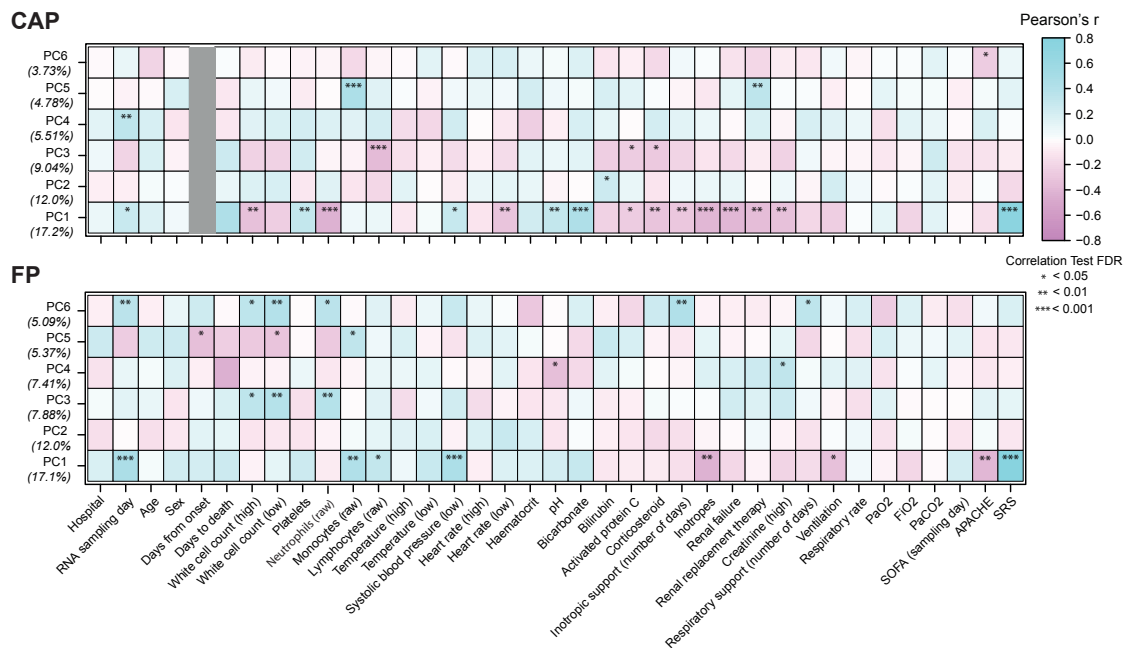


Figure A.1: Correlation between global gene expression and clinical covariates.

Heatmaps showing correlation between clinical covariates and the first six principal components of gene expression for CAP (top) and FP (bottom) samples from the Radhakrishnan 2010 cohort. Positive correlations are indicated in *blue* and negative correlations in *pink*.

Table A.4: Summary of correlation between principal components of gene expression data and clinical covariates.

Significant associations between clinical covariates and the first six principal components are shown.

Clinical Covariate	PC	Pearson's r	FDR
RNA sampling day	PC1	0.35	2.46×10^{-6}
Diagnosis	PC1	-0.42	6.12×10^{-9}
White cell count (high)	PC1	-0.18	3.42×10^{-2}
Platelets	PC1	0.23	5.58×10^{-3}
Polynucleocytes (raw)	PC1	-0.29	2.18×10^{-4}
Mononucleocytes (raw)	PC1	0.26	1.67×10^{-3}
Lymphocytes (raw)	PC1	0.24	4.31×10^{-3}
Temperature (low)	PC1	0.18	4.26×10^{-2}
Systolic blood pressure (low)	PC1	0.32	3.74×10^{-5}
Haematocrit	PC1	0.32	7.05×10^{-5}
pH	PC1	0.25	2.51×10^{-3}
Bicarbonate	PC1	0.43	2.21×10^{-8}
Corticosteroid	PC1	-0.24	3.16×10^{-3}
Inotropic support (number of days)	PC1	-0.20	1.98×10^{-2}
Inotropes	PC1	-0.35	2.94×10^{-6}
Renal failure	PC1	-0.26	1.45×10^{-3}

Continued on following page

Table A.4 - *continued from previous page*

Clinical Covariate	PC	Pearson's r	FDR
Renal replacement therapy	PC1	-0.22	8.62×10^{-3}
Creatinine (high)	PC1	-0.29	3.08×10^{-4}
Ventilation	PC1	-0.19	3.21×10^{-2}
Respiratory rate	PC1	0.18	4.26×10^{-2}
PaCO ₂	PC1	0.20	3.42×10^{-2}
APACHE	PC1	-0.19	2.71×10^{-2}
SRS	PC1	0.77	1.49×10^{-41}
Lymphocytes (raw)	PC2	0.19	3.21×10^{-2}
RNA sampling day	PC3	-0.19	2.73×10^{-2}
White cell count (high)	PC3	-0.19	2.64×10^{-2}
White cell count (low)	PC3	-0.21	1.72×10^{-2}
Lymphocytes (raw)	PC3	-0.27	8.59×10^{-4}
Heart rate (low)	PC3	-0.18	3.42×10^{-2}
Bilirubin	PC3	-0.22	1.69×10^{-2}
Activated protein C	PC3	-0.20	1.72×10^{-2}
Renal failure	PC3	-0.23	7.82×10^{-3}
Renal replacement therapy	PC3	-0.20	2.40×10^{-2}
Creatinine (high)	PC3	-0.28	6.18×10^{-4}
APACHE	PC3	-0.19	3.21×10^{-2}
RNA sampling day	PC4	0.23	6.43×10^{-3}
Platelets	PC4	0.21	1.61×10^{-2}
Sex	PC5	0.21	1.31×10^{-2}
Days from onset	PC5	-0.32	1.36×10^{-2}
Mononucleocytes (raw)	PC5	0.25	2.52×10^{-3}
RNA sampling day	PC6	0.31	4.64×10^{-5}
Days from onset	PC6	0.35	5.61×10^{-3}
White cell count (high)	PC6	0.39	1.06×10^{-7}
White cell count (low)	PC6	0.47	4.29×10^{-11}
Polynucleocytes (raw)	PC6	0.42	6.12×10^{-9}
Systolic blood pressure (low)	PC6	0.29	2.75×10^{-4}
Bicarbonate	PC6	0.23	1.12×10^{-2}
Inotropic support (number of days)	PC6	0.17	4.84×10^{-2}
SRS	PC6	0.19	2.55×10^{-2}

Table A.5: Summary of genes differentially expressed between CAP and FP.
40 most significantly differentially expressed probes.

Gene Symbol	Log ₂ fold change	Mean CAP	Mean FP	FDR
<i>EPSTI1</i>	2.25	10.17	7.91	2.97x10 ⁻⁸
<i>XAF1</i>	1.70	9.71	8.00	2.12x10 ⁻⁷
<i>IFIH1</i>	1.20	9.03	7.83	2.12x10 ⁻⁷
<i>HERC5</i>	1.82	10.33	8.51	2.12x10 ⁻⁷
<i>LYSMD2</i>	1.06	10.82	9.76	2.12x10 ⁻⁷
<i>OAS2</i>	1.58	10.04	8.46	2.12x10 ⁻⁷
<i>TDRD9</i>	-1.30	8.80	10.10	2.12x10 ⁻⁷
<i>LY6E</i>	1.56	10.56	8.99	4.12x10 ⁻⁷
<i>MSRA</i>	-0.83	9.02	9.85	4.12x10 ⁻⁷
<i>IFIT1</i>	2.35	10.02	7.67	4.12x10 ⁻⁷
<i>IFIT3</i>	1.96	9.91	7.95	4.12x10 ⁻⁷
<i>FLJ31222</i>	-0.62	6.22	6.84	4.12x10 ⁻⁷
<i>C20ORF3</i>	-0.65	11.05	11.71	4.50x10 ⁻⁷
<i>FBXW2</i>	-0.76	7.86	8.62	4.50x10 ⁻⁷
<i>ATP9A</i>	-1.11	9.80	10.91	5.45x10 ⁻⁷
<i>TDRD9</i>	-1.25	8.88	10.13	6.71x10 ⁻⁷
<i>BTN3A3</i>	0.84	7.40	6.56	6.71x10 ⁻⁷
<i>IFIT2</i>	1.56	11.90	10.34	6.71x10 ⁻⁷
<i>NQO2</i>	-0.84	9.94	10.78	7.40x10 ⁻⁷
<i>PARP14</i>	0.99	9.58	8.59	8.28x10 ⁻⁷
<i>FAM101B</i>	-0.75	9.08	9.83	8.80x10 ⁻⁷
<i>BTN3A1</i>	1.01	8.74	7.73	8.85x10 ⁻⁷
<i>GADD45A</i>	-1.10	10.08	11.18	8.85x10 ⁻⁷
<i>PSME2</i>	0.68	10.71	10.03	8.88x10 ⁻⁷
<i>REC8</i>	0.59	8.94	8.35	1.34x10 ⁻⁶
<i>CKAP4</i>	-0.72	11.72	12.44	1.59x10 ⁻⁶
<i>LILRA6</i>	-0.74	11.10	11.84	1.93x10 ⁻⁶
<i>PFKFB2</i>	-1.41	8.45	9.86	1.93x10 ⁻⁶
<i>IFI44L</i>	2.13	8.65	6.52	1.93x10 ⁻⁶
<i>ZDHHC19</i>	-2.09	9.84	11.93	1.94x10 ⁻⁶
<i>BIRC3</i>	0.75	9.12	8.37	2.32x10 ⁻⁶
<i>CA4</i>	-0.96	10.98	11.94	2.32x10 ⁻⁶
<i>EPHB1</i>	0.86	6.84	5.98	2.32x10 ⁻⁶
<i>MMP9</i>	-0.91	13.37	14.27	2.40x10 ⁻⁶
<i>WIP1</i>	-0.65	8.67	9.31	2.40x10 ⁻⁶
<i>KIF3C</i>	-0.62	6.08	6.70	2.40x10 ⁻⁶
<i>ISG15</i>	1.63	9.76	8.13	2.58x10 ⁻⁶
<i>PFKFB2</i>	-1.03	7.48	8.51	2.59x10 ⁻⁶
<i>IFI44</i>	1.58	9.94	8.36	3.01x10 ⁻⁶
<i>EFCBP1</i>	-1.32	5.82	7.15	3.19x10 ⁻⁶

Pathway	FDR
Interferon Signalling	6.46×10^{-10}
Antigen Presentation Pathway	1.62×10^{-8}
OX40 Signalling Pathway	2.69×10^{-6}
Type I Diabetes Mellitus Signalling	6.76×10^{-6}
Dendritic Cell Maturation	8.71×10^{-6}
Allograft Rejection Signalling	1.05×10^{-5}
Activation of IRF by Cytosolic PRRs	1.26×10^{-5}
Cdc42 Signalling	3.39×10^{-5}
Airway Pathology in Chronic Obstructive Pulmonary Disease	1.17×10^{-4}
Graft-versus-Host Disease Signalling	1.45×10^{-4}
Role of PRRs in Recognition of Bacteria and Viruses	1.66×10^{-4}
Atherosclerosis Signalling	1.78×10^{-4}
LXR/RXR Activation	1.86×10^{-4}
B Cell Development	2.04×10^{-4}
iCOS-iCOSL Signalling in T Helper Cells	2.95×10^{-4}
Nur77 Signalling in T Lymphocytes	3.16×10^{-4}
Communication between Innate and Adaptive Immune Cells	4.07×10^{-4}
Hematopoiesis from Pluripotent Stem Cells	5.89×10^{-4}
Calcium-induced T Lymphocyte Apoptosis	6.17×10^{-4}
Autoimmune Thyroid Disease Signalling	9.55×10^{-4}
Arginine Degradation I (Arginase Pathway)	1.00×10^{-3}
T Helper Cell Differentiation	1.26×10^{-3}
Primary Immunodeficiency Signalling	1.32×10^{-3}
Arginine Degradation VI (Arginase 2 Pathway)	2.45×10^{-3}
Hepatic Fibrosis / Hepatic Stellate Cell Activation	2.63×10^{-3}
CD28 Signalling in T Helper Cells	2.88×10^{-3}
Glucocorticoid Receptor Signalling	2.95×10^{-3}
Altered T Cell and B Cell Signalling in Rheumatoid Arthritis	3.16×10^{-3}
IL-6 Signalling	4.17×10^{-3}
Citrulline Biosynthesis	4.47×10^{-3}
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells	4.57×10^{-3}
Role of Cytokines in Mediating Communication between Immune Cells	4.68×10^{-3}
Role of NFAT in Regulation of the Immune Response	5.01×10^{-3}
FXR/RXR Activation	5.89×10^{-3}
Acute Phase Response Signalling	6.31×10^{-3}
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	7.08×10^{-3}
IL-12 Signalling and Production in Macrophages	7.41×10^{-3}
T Cell Receptor Signalling	7.76×10^{-3}

Table A.6: Summary of pathways enriched amongst genes differentially expressed between CAP and FP.

Gene Symbol	Log ₂ fold Change	Mean Viral	Mean Non-Viral	FDR
<i>IFI27</i>	3.47	9.22	5.77	7.14x10 ⁻⁵
<i>XAF1</i>	1.59	9.18	7.60	0.014
<i>SPATS2L</i>	1.20	6.84	5.66	0.014
<i>LOC389386</i>	0.86	5.88	5.02	0.014
<i>JUP</i>	1.65	6.43	4.77	0.018
<i>LGALS3BP</i>	1.11	4.87	3.77	0.020
<i>BTN2A2</i>	0.74	4.94	4.21	0.020
<i>SERPING1</i>	1.92	7.50	5.60	0.026
<i>GALM</i>	0.93	8.37	7.46	0.026
<i>MOV10</i>	0.88	7.60	6.73	0.026
<i>EPSTI1</i>	1.89	11.10	9.23	0.029
<i>TGIF2</i>	0.86	6.81	5.96	0.031
<i>LY6E</i>	1.49	11.07	9.60	0.033
<i>Hs.125087</i>	1.23	6.82	5.60	0.033
<i>OAS1</i>	1.20	9.25	8.06	0.033
<i>Hs.72010</i>	0.82	5.74	4.93	0.033
<i>CD2AP</i>	0.76	4.93	4.18	0.033
<i>IFI44L</i>	2.39	9.77	7.40	0.033
<i>IFI44</i>	1.49	10.54	9.06	0.033
<i>BATF2</i>	1.48	5.91	4.45	0.033
<i>OAS2</i>	1.39	5.72	4.34	0.033
<i>TIMM10</i>	1.35	8.89	7.56	0.033
<i>PARP12</i>	0.91	9.81	8.90	0.033
<i>HERC6</i>	0.83	7.70	6.88	0.033
<i>SCO2</i>	0.79	10.44	9.66	0.033
<i>SP140</i>	0.74	8.30	7.57	0.033
<i>ZCCHC2</i>	0.68	3.76	3.08	0.033
<i>PSME2</i>	0.58	11.07	10.49	0.033
<i>OAS3</i>	1.64	8.76	7.14	0.033
<i>LAP3</i>	0.81	9.21	8.41	0.035
<i>LOC652694</i>	1.28	8.48	7.20	0.040
<i>HES4</i>	0.98	7.68	6.70	0.041
<i>RSAD2</i>	1.74	8.24	6.52	0.043
<i>OTOF</i>	1.37	6.06	4.71	0.043
<i>IFIH1</i>	0.96	9.11	8.16	0.043
<i>LAMP3</i>	0.81	5.24	4.44	0.043
<i>RTP4</i>	1.37	7.30	5.94	0.044
<i>ISG15</i>	1.49	10.45	8.97	0.049
<i>C19orf12</i>	0.63	7.92	7.29	0.049
<i>ECHDC3</i>	-1.15	7.40	8.55	0.049

Table A.7: Summary of genes differentially expressed between viral and non-viral patients.

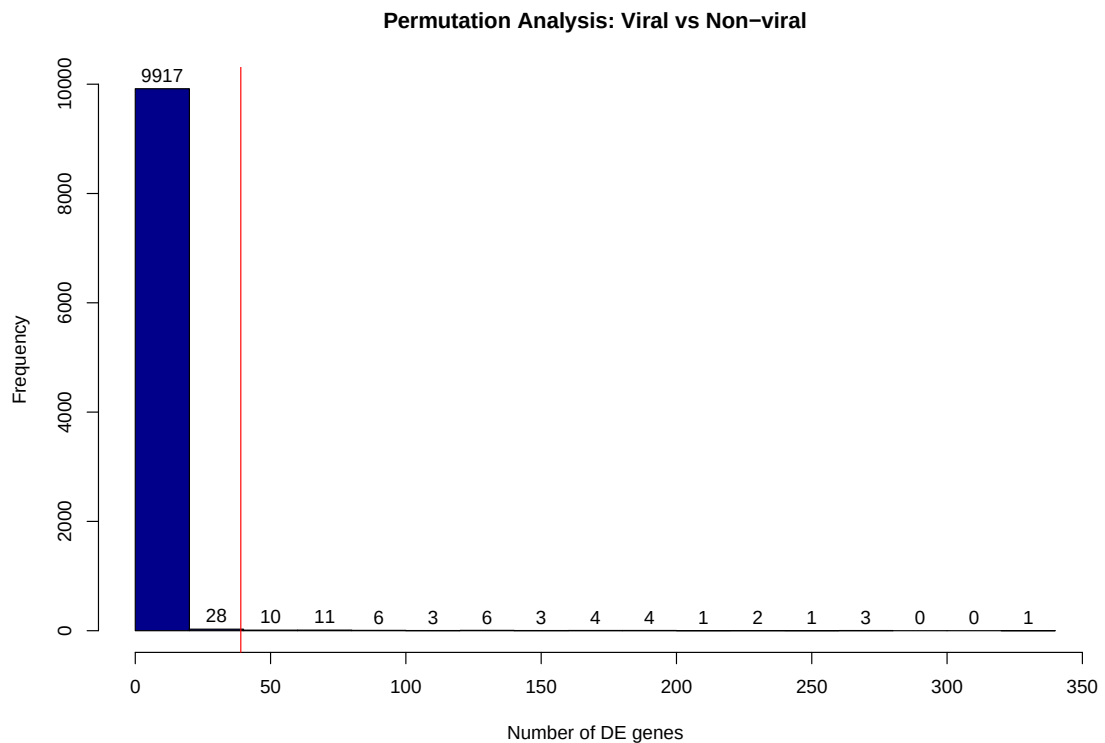


Figure A.2: Histogram of viral vs non-viral permutation results.

Viral status labels were permuted 9,999 times and differential gene expression analysis carried out. The number of significantly differentially expressed genes across all permutations is plotted, demonstrating that the true observation (39 genes) is significantly more than expected by chance ($p=0.0056$).

Table A.8: Summary of genes differentially expressed between CAP and FP (validation).

40 most significantly differentially expressed probes.

Gene Symbol	Log ₂ fold change	Mean CAP	Mean FP	FDR
<i>PATL2</i>	1.12	8.15	7.04	2.93×10^{-11}
<i>LOC197135</i>	0.96	7.62	6.66	2.35×10^{-10}
<i>BTN3A1</i>	1.38	9.12	7.74	2.35×10^{-10}
<i>CA4</i>	-1.12	11.04	12.17	7.92×10^{-10}
<i>IFIT3</i>	2.28	10.14	7.86	1.22×10^{-9}
<i>HERC5</i>	1.78	10.14	8.36	6.54×10^{-9}
<i>TAP2</i>	1.00	9.04	8.04	6.54×10^{-9}
<i>IFIT1</i>	2.44	9.52	7.08	1.18×10^{-8}
<i>HS.573264</i>	0.48	7.38	6.91	1.27×10^{-8}
<i>GFOD2</i>	-0.57	7.92	8.49	1.33×10^{-8}
<i>IFIT2</i>	1.78	11.88	10.10	1.72×10^{-8}
<i>ODF3B</i>	0.93	7.93	7.01	2.27×10^{-8}
<i>EPSTI1</i>	2.15	9.74	7.59	2.53×10^{-8}

Continued on following page

Table A.8 - *continued from previous page*

Gene Symbol	Log ₂ fold change	Mean CAP	Mean FP	FDR
<i>IFIH1</i>	1.21	8.71	7.50	2.95x10 ⁻⁸
<i>PARP14</i>	1.11	9.73	8.62	4.25x10 ⁻⁸
<i>POLB</i>	0.84	9.41	8.57	6.18x10 ⁻⁸
<i>XAF1</i>	1.66	9.66	8.00	6.95x10 ⁻⁸
<i>PDE4B</i>	0.99	9.30	8.31	8.53x10 ⁻⁸
<i>TCTN1</i>	0.57	6.79	6.23	8.53x10 ⁻⁸
<i>PSMB9</i>	0.68	8.11	7.43	1.84x10 ⁻⁷
<i>PECR</i>	-1.02	8.53	9.55	1.84x10 ⁻⁷
<i>TDRD9</i>	-1.49	8.94	10.43	2.86x10 ⁻⁷
<i>ZNF282</i>	-0.54	7.37	7.91	5.20x10 ⁻⁷
<i>GBP1</i>	1.53	8.84	7.32	5.65x10 ⁻⁷
<i>MMP9</i>	-0.99	13.19	14.18	5.65x10 ⁻⁷
<i>LYSMD2</i>	1.04	10.69	9.65	5.65x10 ⁻⁷
<i>TDRD9</i>	-1.51	8.74	10.25	6.65x10 ⁻⁷
<i>SCO2</i>	1.02	10.33	9.32	8.58x10 ⁻⁷
<i>GBP1</i>	1.49	9.65	8.16	9.76x10 ⁻⁷
<i>UBE2L6</i>	0.87	11.91	11.04	1.02x10 ⁻⁶
<i>PCMT1</i>	-0.48	11.56	12.04	1.15x10 ⁻⁶
<i>BTN3A3</i>	0.78	7.56	6.77	1.23x10 ⁻⁶
<i>GBP5</i>	1.39	9.81	8.41	1.24x10 ⁻⁶
<i>STAT1</i>	1.02	9.87	8.84	1.35x10 ⁻⁶
<i>ABCG1</i>	1.25	9.00	7.75	1.42x10 ⁻⁶
<i>ABCG1</i>	0.93	7.21	6.28	1.57x10 ⁻⁶
<i>IER5</i>	0.65	8.73	8.08	1.57x10 ⁻⁶
<i>PSME1</i>	0.48	12.52	12.04	1.57x10 ⁻⁶
<i>HS.445414</i>	0.68	8.51	7.83	1.82x10 ⁻⁶
<i>LOC282997</i>	0.68	7.41	6.74	1.82x10 ⁻⁶

Pathway	FDR
Interferon Signalling	6.31x10 ⁻¹¹
Role of PRRs in Recognition of Bacteria and Viruses	8.51x10 ⁻⁶
Antigen Presentation Pathway	1.12x10 ⁻⁵
Activation of IRF by Cytosolic PRRs	2.34x10 ⁻⁵
TREM1 Signalling	6.46x10 ⁻⁵
Airway Pathology in Chronic Obstructive Pulmonary Disease	1.55x10 ⁻⁴
Graft-versus-Host Disease Signalling	2.29x10 ⁻⁴
Atherosclerosis Signalling	3.47x10 ⁻⁴
Granulocyte Adhesion and Diapedesis	6.61x10 ⁻⁴
Arginine Degradation I (Arginase Pathway)	1.20x10 ⁻³
Autoimmune Thyroid Disease Signalling	1.38x10 ⁻³
Granzyme B Signalling	1.45x10 ⁻³
Bladder Cancer Signalling	1.51x10 ⁻³
Granzyme A Signalling	1.74x10 ⁻³
LXR/RXR Activation	1.86x10 ⁻³
Allograft Rejection Signalling	2.29x10 ⁻³
Neuroprotective Role of THOP1 in Alzheimer's Disease	2.57x10 ⁻³
Arginine Degradation VI (Arginase 2 Pathway)	2.95x10 ⁻³
Type I Diabetes Mellitus Signalling	3.39x10 ⁻³
Communication between Innate and Adaptive Immune Cells	4.27x10 ⁻³
Hepatic Fibrosis / Hepatic Stellate Cell Activation	4.79x10 ⁻³
Citrulline Biosynthesis	5.50x10 ⁻³
Agranulocyte Adhesion and Diapedesis	1.35x10 ⁻²
p53 Signalling	1.35x10 ⁻²
Histamine Biosynthesis	1.45x10 ⁻²
Complement System	1.48x10 ⁻²
Primary Immunodeficiency Signalling	1.58x10 ⁻²
Superpathway of Citrulline Metabolism	1.66x10 ⁻²
Role of IL-17F in Allergic Inflammatory Airway Diseases	2.09x10 ⁻²
FXR/RXR Activation	3.47x10 ⁻²
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	3.55x10 ⁻²
UVA-Induced MAPK Signalling	3.72x10 ⁻²
1D-myo-inositol Hexakisphosphate Biosynthesis V (from Ins(1,3,4)P3)	4.27x10 ⁻²
Differential Regulation of Cytokine Production in Intestinal Epithelial Cells by IL-17A and IL-17F	4.27x10 ⁻²
Tumoricidal Function of Hepatic Natural Killer Cells	4.57x10 ⁻²
Role of JAK1, JAK2 and TYK2 in Interferon Signalling	4.57x10 ⁻²

Table A.9: Summary of pathways enriched amongst genes differentially expressed between CAP and FP (validation).

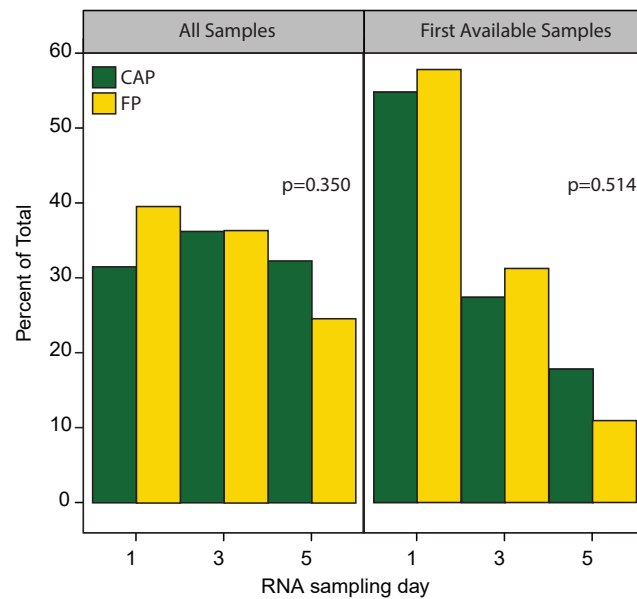


Figure A.3: Histogram of RNA sampling day for CAP and FP samples.

The percentage of samples in the Radhakrishnan 2010 cohort for each time point are given for CAP (green) and FP (yellow). The left hand plot shows the percentage for all available serial samples, and the right hand plot shows the proportions when only the first available sample for each patient is selected. The CAP and FP subgroups do not differ significantly in terms of the proportion of samples from each day of sampling (chi-square test p-values shown).

Table A.10: Summary of genes differentially expressed between day 1 and day 5 samples.

20 most significantly differentially expressed probes in CAP (top) and FP (bottom).

Gene	Log ₂ FC (CAP)	FDR CAP	Log ₂ FC (FP)	FDR FP
<i>MYB</i>	0.92	1.31x10 ⁻³	1.62	1.93x10 ⁻⁴
<i>OLR1</i>	1.44	2.62x10 ⁻³	2.82	9.88x10 ⁻⁶
<i>SLC2A5</i>	1.01	2.62x10 ⁻³	1.76	2.11x10 ⁻⁴
<i>C13ORF15</i>	0.89	2.62x10 ⁻³	1.61	2.16x10 ⁻⁴
<i>LOC653600</i>	1.72	2.62x10 ⁻³	2.32	4.34x10 ⁻³
<i>MS4A3</i>	1.36	2.62x10 ⁻³	1.80	6.16x10 ⁻³
<i>DEFA4</i>	1.56	2.62x10 ⁻³	2.18	7.21x10 ⁻³
<i>MFSD7</i>	-0.38	2.70x10 ⁻³	-0.09	6.77x10 ⁻¹
<i>CHI3L1</i>	1.14	3.13x10 ⁻³	2.07	4.60x10 ⁻⁶
<i>NCKAP1L</i>	-0.36	3.13x10 ⁻³	-0.60	2.71x10 ⁻⁴
<i>CEACAM6</i>	1.29	3.33x10 ⁻³	1.93	4.63x10 ⁻³
<i>CTSG</i>	1.43	5.98x10 ⁻³	1.91	1.07x10 ⁻²
<i>DEFA1B</i>	1.21	7.37x10 ⁻³	1.30	2.07x10 ⁻²
<i>CAMP</i>	1.17	7.38x10 ⁻³	2.47	2.00x10 ⁻⁵
<i>HS.224794</i>	-0.41	7.87x10 ⁻³	-0.79	3.52x10 ⁻⁴
<i>NIT2</i>	0.40	1.08x10 ⁻²	0.53	1.57x10 ⁻²

Continued on following page

Table A.10 - continued from previous page

Gene	Log ₂ FC (CAP)	FDR CAP	Log ₂ FC (FP)	FDR FP
<i>DEFA1B</i>	1.07	1.08x10 ⁻²	1.15	2.69x10 ⁻²
<i>SNRPN</i>	0.47	1.08x10 ⁻²	0.19	6.36x10 ⁻¹
<i>CD300LF</i>	-0.47	1.33x10 ⁻²	-0.96	1.97x10 ⁻⁵
<i>CYP1B1</i>	-0.93	1.33x10 ⁻²	-1.84	9.39x10 ⁻⁴
<i>RPS6KA5</i>	0.61	1.75x10 ⁻²	1.58	8.04x10 ⁻⁷
<i>PLEKHG1</i>	-0.38	5.97x10 ⁻²	-0.99	8.04x10 ⁻⁷
<i>SULF2</i>	0.63	8.74x10 ⁻²	2.09	8.04x10 ⁻⁷
<i>PRKDC</i>	0.46	9.43x10 ⁻²	1.28	8.04x10 ⁻⁷
<i>PLAC8</i>	-0.70	1.01x10 ⁻¹	-1.79	8.04x10 ⁻⁷
<i>MME</i>	0.63	1.15x10 ⁻¹	2.74	8.04x10 ⁻⁷
<i>C14ORF159</i>	0.32	1.19x10 ⁻¹	1.37	8.04x10 ⁻⁷
<i>EMR3</i>	0.34	3.06x10 ⁻¹	1.91	8.04x10 ⁻⁷
<i>PLAC8</i>	-0.72	1.20x10 ⁻¹	-1.81	1.11x10 ⁻⁶
<i>SMPDL3A</i>	-0.42	1.69x10 ⁻¹	-1.36	1.36x10 ⁻⁶
<i>DPEP3</i>	0.36	3.20x10 ⁻¹	1.36	1.87x10 ⁻⁶
<i>GPR84</i>	-0.85	5.23x10 ⁻²	-1.62	2.21x10 ⁻⁶
<i>MERTK</i>	-0.71	5.93x10 ⁻²	-2.08	2.21x10 ⁻⁶
<i>ENTPD7</i>	-0.45	1.93x10 ⁻¹	-1.57	2.21x10 ⁻⁶
<i>TNFAIP8L3</i>	-0.51	1.21x10 ⁻¹	-1.57	2.71x10 ⁻⁶
<i>EMR1</i>	-0.78	4.31x10 ⁻²	-1.59	2.90x10 ⁻⁶
<i>AP3B2</i>	-0.58	6.38x10 ⁻²	-1.31	3.43x10 ⁻⁶
<i>NDST2</i>	-0.33	8.69x10 ⁻²	-0.81	3.43x10 ⁻⁶
<i>RBP7</i>	0.21	4.50x10 ⁻¹	1.31	3.45x10 ⁻⁶
<i>FAM20A</i>	-0.58	4.66x10 ⁻²	-1.40	3.51x10 ⁻⁶

Table A.11: Summary of genes differentially expressed over time in CAP and FP.
 20 most highly ranked probes in CAP (*top*) and FP (*bottom*).

Gene	T2 Statistic (CAP)	Rank CAP	T2 Statistic (FP)	Rank FP
<i>LOC649848</i>	1.91	1	1.24	14292
<i>TNPO1</i>	0.47	2	4.57	2872
<i>PPP2CB</i>	3.88	3	9.01	19985
<i>MED6</i>	0.67	4	10.23	9618
<i>C16ORF58</i>	7.55	5	17.84	21171
<i>DRD3</i>	0.85	6	1.89	24497
<i>TDO2</i>	0.67	7	0.43	12142
<i>CPT2</i>	1.33	8	5.99	14091
<i>PCYOX1L</i>	4.17	9	0.65	2508
<i>LOC391825</i>	5.51	10	3.37	5850
<i>ENO1</i>	5.61	11	2.35	18126
<i>TMEM187</i>	6.96	12	2.01	3946

Continued on following page

Table A.11 - *continued from previous page*

Gene	T2 Statistic (CAP)	Rank CAP	T2 Statistic (FP)	Rank FP
<i>PLA2G2D</i>	0.24	13	17.79	3172
<i>PTGS1</i>	4.83	14	4.52	17189
<i>TRIM52</i>	0.54	15	0.55	16196
<i>YTHDC1</i>	2.08	16	13.39	18782
<i>HS.147030</i>	0.61	17	0.22	23486
<i>LOC730052</i>	0.85	18	2.56	7795
<i>C17ORF95</i>	3.99	19	0.80	17594
<i>ACAD8</i>	3.75	20	5.33	20339
<i>HS.568451</i>	0.52	40	0.02	1
<i>WTAP</i>	0.08	3346	0.40	2
<i>SPSB2</i>	0.22	18768	1.49	3
<i>LOC255326</i>	2.43	7023	7.37	4
<i>IFIT3</i>	7.03	18445	3.61	5
<i>GALC</i>	8.29	20891	16.35	6
<i>SIGLEC5</i>	2.38	25460	1.66	7
<i>EID2B</i>	4.51	18864	4.17	8
<i>IGF2</i>	2.08	6407	0.26	9
<i>LOC100129720</i>	2.10	15849	1.87	10
<i>LGALS8</i>	7.80	2662	7.56	11
<i>HS.527071</i>	0.54	20443	0.07	12
<i>YWHAG</i>	0.19	6769	9.14	13
<i>PLXNA3</i>	0.84	12952	3.84	14
<i>HS.143909</i>	0.29	11515	0.13	15
<i>PNO1</i>	1.01	24215	0.42	16
<i>ARRDC3</i>	0.79	7274	20.11	17
<i>HS.558271</i>	0.89	2422	1.14	18
<i>CNNM2</i>	1.84	8617	0.38	19
<i>IGHMBP2</i>	0.10	18851	1.62	20

Table A.12: Summary of genes differentially expressed from time of onset of FP.
40 most significantly differentially expressed probes.

Gene	F statistic	FDR
<i>GLT25D1</i>	13.11	3.45×10^{-3}
<i>C9orf102</i>	13.10	3.45×10^{-3}
<i>MACF1</i>	12.92	3.45×10^{-3}
<i>F8</i>	12.43	3.57×10^{-3}
<i>GALT</i>	12.40	3.57×10^{-3}
<i>AAK1</i>	11.88	5.18×10^{-3}
<i>KIAA0586</i>	11.60	5.69×10^{-3}
<i>SNN</i>	11.52	5.69×10^{-3}
<i>NCKAP1L</i>	11.33	6.18×10^{-3}
<i>ACER3</i>	11.07	7.12×10^{-3}
<i>TRPV2</i>	11.00	7.12×10^{-3}
<i>PARD6A</i>	10.87	7.12×10^{-3}
<i>LOC729566</i>	10.86	7.12×10^{-3}
<i>EPHB4</i>	10.66	8.20×10^{-3}
<i>PRDM8</i>	10.57	8.42×10^{-3}
<i>GK5</i>	10.20	1.09×10^{-2}
<i>FLJ00312</i>	10.19	1.09×10^{-2}
<i>FRY</i>	10.17	1.09×10^{-2}
<i>KIAA1602</i>	10.12	1.09×10^{-2}
<i>SLBP</i>	9.99	1.20×10^{-2}
<i>PHCA</i>	9.94	1.21×10^{-2}
<i>ZDHHC23</i>	9.77	1.39×10^{-2}
<i>AVIL</i>	9.68	1.46×10^{-2}
<i>CHI3L1</i>	9.57	1.56×10^{-2}
<i>SNORD6</i>	9.52	1.56×10^{-2}
<i>KIAA1737</i>	9.48	1.56×10^{-2}
<i>ARHGEF6</i>	9.48	1.56×10^{-2}
<i>ALDH3B1</i>	9.44	1.56×10^{-2}
<i>MME</i>	9.39	1.56×10^{-2}
<i>CMTM4</i>	9.37	1.56×10^{-2}
<i>PARL</i>	9.34	1.56×10^{-2}
<i>SGK1</i>	9.33	1.56×10^{-2}
<i>SULF2</i>	9.26	1.64×10^{-2}
<i>RHOB</i>	9.21	1.68×10^{-2}
<i>TBX19</i>	9.13	1.77×10^{-2}
<i>ISG20</i>	9.10	1.80×10^{-2}
<i>TST</i>	8.97	2.02×10^{-2}
<i>PKD2</i>	8.87	2.15×10^{-2}
<i>RUNX1</i>	8.85	2.15×10^{-2}
<i>SULF2</i>	8.84	2.15×10^{-2}

Pathway	FDR
Phagosome formation	2.88x10 ⁻³
Natural Killer Cell Signalling	3.98x10 ⁻³
Glioma Invasiveness Signalling	4.79x10 ⁻³
GM-CSF Signalling	6.03x10 ⁻³
IL-3 Signalling	8.71x10 ⁻³
Cardiac Hypertrophy Signalling	1.00x10 ⁻²
Glioblastoma Multiforme Signalling	1.10x10 ⁻²
Gq Signalling	1.10x10 ⁻²
Huntington's Disease Signalling	1.12x10 ⁻²
All-trans-decaprenyl Diphosphate Biosynthesis	1.17x10 ⁻²
Signalling by Rho Family GTPases	1.29x10 ⁻²
G Beta Gamma Signalling	1.55x10 ⁻²
Thiosulfate Disproportionation III (Rhodanese)	1.78x10 ⁻²
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	2.24x10 ⁻²
Glucocorticoid Receptor Signalling	2.29x10 ⁻²
ILK Signalling	2.29x10 ⁻²
Clathrin-mediated Endocytosis Signalling	2.45x10 ⁻²
Thrombin Signalling	2.57x10 ⁻²
Dermatan Sulfate Biosynthesis (Late Stages)	2.57x10 ⁻²
CD28 Signalling in T Helper Cells	2.69x10 ⁻²
Sphingosine-1-phosphate Signalling	2.69x10 ⁻²
Chondroitin Sulfate Biosynthesis (Late Stages)	2.82x10 ⁻²
Glycerol Degradation I	2.95x10 ⁻²
Galactose Degradation I (Leloir Pathway)	2.95x10 ⁻²
Heparan Sulfate Biosynthesis (Late Stages)	3.31x10 ⁻²
HMGB1 Signalling	3.31x10 ⁻²
Semaphorin Signalling in Neurons	3.80x10 ⁻²
Chondroitin Sulfate Biosynthesis	3.80x10 ⁻²
Heparan Sulfate Biosynthesis	4.27x10 ⁻²
Dermatan Sulfate Biosynthesis	4.27x10 ⁻²
Phospholipase C Signalling	4.27x10 ⁻²
Actin Nucleation by ARP-WASP Complex	4.37x10 ⁻²
ErbB2-ErbB3 Signalling	4.37x10 ⁻²
Salvage Pathways of Pyrimidine Deoxyribonucleotides	4.68x10 ⁻²
ErbB4 Signalling	4.68x10 ⁻²
Colorectal Cancer Metastasis Signalling	4.90x10 ⁻²

Table A.13: Summary of pathways enriched amongst genes differentially expressed from day of onset in FP.

Table A.14: Summary of genes differentially expressed over time in sepsis.
40 most highly ranked probes.

Gene	T2 Statistic	Rank
<i>CYP4F12</i>	147.7357	1
<i>RPS6KA5</i>	121.4114	2
<i>SULF2</i>	120.4925	3
<i>KIAA1324</i>	119.3089	4
<i>SULF2</i>	119.032	5
<i>NOV</i>	108.2554	6
<i>EPHB4</i>	106.4899	7
<i>PLAC8</i>	98.2722	8
<i>EMR3</i>	97.50025	9
<i>DHRS9</i>	97.14908	10
<i>PI3</i>	96.83894	11
<i>FGF13</i>	93.62463	12
<i>PHCA</i>	93.23158	13
<i>SULF2</i>	89.40005	14
<i>RUNX1</i>	88.72849	15
<i>ATG2A</i>	88.57827	16
<i>PLAC8</i>	88.12258	17
<i>EMR3</i>	87.17772	18
<i>NECAB2</i>	86.9786	19
<i>MARCKSL1</i>	86.7961	20
<i>PTP4A3</i>	86.29122	21
<i>CECR6</i>	84.42117	22
<i>PTP4A3</i>	83.97934	23
<i>PTP4A3</i>	83.82104	24
<i>SULT1A1</i>	83.69487	25
<i>GPR84</i>	83.44815	26
<i>CHST7</i>	81.4613	27
<i>CD300LB</i>	80.51865	28
<i>KCNH3</i>	79.84581	29
<i>SGK</i>	76.91026	30
<i>AZU1</i>	76.7039	31
<i>DHRS9</i>	75.54171	32
<i>ABCC5</i>	75.49961	33
<i>LOC646463</i>	75.30873	34
<i>TGM3</i>	74.65573	35
<i>MYB</i>	73.54402	36
<i>DPEP3</i>	72.55056	37
<i>MME</i>	71.55466	38
<i>LOC653381</i>	71.55448	39
<i>SGK1</i>	71.38512	40

Table A.15: Summary of genes differentially expressed between pre- and post-operative samples.

40 most significantly differentially expressed probes.

Gene	Log ₂ fold change	Mean expression	FDR
<i>GFI1</i>	-0.85	6.6	3.39x10 ⁻¹⁴
<i>IFI30</i>	0.67	10.7	1.80x10 ⁻¹³
<i>MARCO</i>	1.4	6.8	1.80x10 ⁻¹³
<i>EMILIN2</i>	0.77	9.9	2.23x10 ⁻¹³
<i>GBGT1</i>	0.76	8.2	3.64x10 ⁻¹³
<i>TLR5</i>	1.6	9.8	4.37x10 ⁻¹³
<i>HCK</i>	0.72	11.7	4.81x10 ⁻¹³
<i>LPCAT3</i>	0.72	8.6	6.92x10 ⁻¹³
<i>NQO2</i>	1.1	9.5	6.92x10 ⁻¹³
<i>SLC25A4</i>	-0.93	6.7	6.92x10 ⁻¹³
<i>LILRA5</i>	1.5	9.5	1.02x10 ⁻¹²
<i>ST3GAL4</i>	1.0	7.4	1.02x10 ⁻¹²
<i>C19ORF59</i>	2.5	11.7	1.02x10 ⁻¹²
<i>FLOT1</i>	0.96	10.4	1.58x10 ⁻¹²
<i>ASGR2</i>	1.3	7.7	1.64x10 ⁻¹²
<i>RETN</i>	2.2	8.0	1.64x10 ⁻¹²
<i>LAX1</i>	-1.1	7.2	1.64x10 ⁻¹²
<i>IL4R</i>	1.3	10.4	1.64x10 ⁻¹²
<i>DOK3</i>	1.3	6.7	1.64x10 ⁻¹²
<i>LTA4H</i>	0.73	12.4	1.64x10 ⁻¹²
<i>SYN2</i>	1.6	6.2	1.64x10 ⁻¹²
<i>LOC648984</i>	1.4	9.2	1.64x10 ⁻¹²
<i>GRN</i>	0.63	10.5	1.64x10 ⁻¹²
<i>LILRA5</i>	1.5	8.0	1.94x10 ⁻¹²
<i>SLC38A1</i>	-0.95	9.1	2.02x10 ⁻¹²
<i>ZNF438</i>	0.94	8.5	2.02x10 ⁻¹²
<i>MTF1</i>	0.86	9.5	2.07x10 ⁻¹²
<i>FKBP15</i>	0.64	8.37	2.07x10 ⁻¹²
<i>HK3</i>	1.3	10.8	2.33x10 ⁻¹²
<i>GLT25D1</i>	0.71	10.0	2.33x10 ⁻¹²
<i>ZDHHC19</i>	2.5	7.5	2.68x10 ⁻¹²
<i>ANXA3</i>	1.8	9.5	2.77x10 ⁻¹²
<i>LILRA6</i>	1.3	10.6	2.77x10 ⁻¹²
<i>GNA15</i>	0.66	8.8	3.15x10 ⁻¹²
<i>FCER1G</i>	0.92	12.8	3.25x10 ⁻¹²
<i>DHRS13</i>	1.1	7.7	3.25x10 ⁻¹²
<i>CDR2</i>	-0.72	7.6	3.36x10 ⁻¹²
<i>LOC442132</i>	0.91	5.3	3.36x10 ⁻¹²
<i>CTSD</i>	0.83	8.9	3.82x10 ⁻¹²
<i>RHOU</i>	0.98	9.3	3.82x10 ⁻¹²

Table A.16: Summary of genes differentially expressed between sepsis and pre-operative surgery samples.

40 most significantly differentially expressed probes.

Gene	Log ₂ FC	Mean Sepsis	Mean Control	FDR
<i>PID1</i>	-1.77	6.21	7.97	1.68x10 ⁻²⁴
<i>LOC90624</i>	-0.98	6.64	7.61	6.15x10 ⁻²³
<i>CCDC65</i>	-1.19	5.99	7.18	1.09x10 ⁻²⁰
<i>LBH</i>	-1.45	6.36	7.81	1.09x10 ⁻²⁰
<i>MPP1</i>	1.51	11.33	9.82	1.49x10 ⁻²⁰
<i>HIST1H2BK</i>	1.57	13.08	11.50	6.48x10 ⁻²⁰
<i>PYHIN1</i>	-1.66	6.51	8.17	7.56x10 ⁻²⁰
<i>ZFP106</i>	1.59	10.84	9.25	8.12x10 ⁻²⁰
<i>ZNHIT3</i>	-1.07	7.47	8.54	1.35x10 ⁻¹⁹
<i>PTGDR</i>	-1.23	5.99	7.22	1.59x10 ⁻¹⁹
<i>MS4A2</i>	-1.28	5.93	7.22	4.40x10 ⁻¹⁹
<i>TLR5</i>	2.44	11.34	8.90	7.31x10 ⁻¹⁹
<i>DBP</i>	-0.99	6.32	7.31	8.69x10 ⁻¹⁹
<i>HS.31532</i>	-1.01	6.64	7.65	8.88x10 ⁻¹⁹
<i>ZFP106</i>	1.64	10.68	9.05	1.01x10 ⁻¹⁸
<i>FAM160B1</i>	1.55	11.14	9.59	1.25x10 ⁻¹⁸
<i>PASK</i>	-1.53	6.35	7.88	1.25x10 ⁻¹⁸
<i>LOC647436</i>	-1.56	6.91	8.48	1.25x10 ⁻¹⁸
<i>P2RY10</i>	-1.71	5.97	7.68	1.27x10 ⁻¹⁸
<i>RTN4</i>	1.51	12.28	10.77	1.91x10 ⁻¹⁸
<i>PECI</i>	-1.20	6.56	7.75	2.31x10 ⁻¹⁸
<i>PTGDR</i>	-1.72	6.43	8.15	2.70x10 ⁻¹⁸
<i>CAPZA2</i>	1.52	12.11	10.59	3.02x10 ⁻¹⁸
<i>SEPT1</i>	-1.25	6.45	7.70	3.61x10 ⁻¹⁸
<i>KIAA1600</i>	1.49	10.74	9.25	4.78x10 ⁻¹⁸
<i>RTN3</i>	1.41	11.82	10.40	5.85x10 ⁻¹⁸
<i>HS.407903</i>	2.21	12.06	9.85	6.06x10 ⁻¹⁸
<i>LOC91561</i>	-1.53	7.48	9.01	8.63x10 ⁻¹⁸
<i>FOYN2</i>	1.56	11.10	9.54	1.55x10 ⁻¹⁷
<i>DCAF16</i>	-1.05	5.82	6.87	1.89x10 ⁻¹⁷
<i>EIF4G2</i>	1.19	11.73	10.54	2.82x10 ⁻¹⁷
<i>RNU6-1</i>	1.35	11.66	10.32	2.82x10 ⁻¹⁷
<i>ZNF791</i>	-1.16	7.80	8.96	5.06x10 ⁻¹⁷
<i>LOC644029</i>	-1.36	7.08	8.44	5.09x10 ⁻¹⁷
<i>RNU6-15</i>	1.30	11.70	10.40	9.20x10 ⁻¹⁷
<i>ABHD15</i>	-1.21	6.53	7.73	1.02x10 ⁻¹⁶
<i>HS.413494</i>	1.27	12.79	11.52	1.07x10 ⁻¹⁶
<i>LOC100129237</i>	-1.48	7.22	8.69	1.11x10 ⁻¹⁶
<i>ATPGD1</i>	-1.12	6.02	7.14	1.72x10 ⁻¹⁶
<i>LOC650832</i>	1.30	11.00	9.70	1.94x10 ⁻¹⁶

Pathway	FDR
EIF2 Signalling	1.26x10 ⁻¹⁶
Type I Diabetes Mellitus Signalling	2.29x10 ⁻⁷
iCOS-iCOSL Signalling in T Helper Cells	3.39x10 ⁻⁷
PKC Signalling in T Lymphocytes	5.75x10 ⁻⁷
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells	8.71x10 ⁻⁷
OX40 Signalling Pathway	1.26x10 ⁻⁶
Role of NFAT in Regulation of the Immune Response	1.32x10 ⁻⁶
mTOR Signalling	1.48x10 ⁻⁶
Calcium-induced T Lymphocyte Apoptosis	1.55x10 ⁻⁶
T Helper Cell Differentiation	2.82x10 ⁻⁶
Allograft Rejection Signalling	3.47x10 ⁻⁶
CD28 Signalling in T Helper Cells	4.27x10 ⁻⁶
Regulation of eIF4 and p70S6K Signalling	4.68x10 ⁻⁶
Nur77 Signalling in T Lymphocytes	5.25x10 ⁻⁶
Graft-versus-Host Disease Signalling	4.07x10 ⁻⁵
B Cell Development	5.37x10 ⁻⁵
Granzyme B Signalling	7.08x10 ⁻⁵
Autoimmune Thyroid Disease Signalling	9.33x10 ⁻⁵
Tumoricidal Function of Hepatic Natural Killer Cells	1.02x10 ⁻⁴
T Cell Receptor Signalling	1.05x10 ⁻⁴
Cdc42 Signalling	1.86x10 ⁻⁴
Altered T Cell and B Cell Signalling in Rheumatoid Arthritis	2.69x10 ⁻⁴
Natural Killer Cell Signalling	3.24x10 ⁻⁴
B Cell Receptor Signalling	4.47x10 ⁻⁴
Death Receptor Signalling	5.62x10 ⁻⁴
Apoptosis Signalling	5.62x10 ⁻⁴
Primary Immunodeficiency Signalling	5.75x10 ⁻⁴
CTLA4 Signalling in Cytotoxic T Lymphocytes	6.76x10 ⁻⁴
Dendritic Cell Maturation	7.76x10 ⁻⁴
Antigen Presentation Pathway	8.71x10 ⁻⁴
IL-4 Signalling	9.55x10 ⁻⁴
Hypoxia Signalling in the Cardiovascular System	1.00x10 ⁻³
Communication between Innate and Adaptive Immune Cells	1.38x10 ⁻³
Spermine and Spermidine Degradation I	2.19x10 ⁻³
Phospholipase C Signalling	2.88x10 ⁻³
ILK Signalling	4.07x10 ⁻³
TWEAK Signalling	4.27x10 ⁻³
Autophagy	4.27x10 ⁻³
Crosstalk between Dendritic Cells and Natural Killer Cells	5.13x10 ⁻³
Glioma Invasiveness Signalling	6.31x10 ⁻³

Table A.17: Summary of pathways enriched amongst genes differentially expressed in sepsis vs controls.

Table A.18: Summary of genes differentially expressed between septic and post-operative samples.

40 most significantly differentially expressed probes.

Gene	Log ₂ fold change	Mean expression	FDR
<i>PLA2G7</i>	-2.05332	5.9604893	7.72x10 ⁻²⁹
<i>CCDC109A</i>	0.846502	7.6525482	2.20x10 ⁻²²
<i>LCN2</i>	2.539199	11.443794	4.52x10 ⁻²²
<i>KCTD6</i>	-0.79105	6.2718018	1.78x10 ⁻²¹
<i>MARCH1.</i>	-1.3952	6.9391785	1.78x10 ⁻²¹
<i>LILRA1</i>	-1.2886	4.1084791	1.78x10 ⁻²¹
<i>C17ORF87</i>	-1.04936	6.8980588	5.48x10 ⁻²⁰
<i>CECR1</i>	-1.08295	3.684976	5.80x10 ⁻²⁰
<i>CCR2</i>	-1.70233	6.3828405	5.80x10 ⁻²⁰
<i>INSL3</i>	1.572988	5.9509559	1.26x10 ⁻¹⁹
<i>NAPSB</i>	-1.4158	6.7570154	1.92x10 ⁻¹⁸
<i>MIR1974</i>	1.26483	12.498972	3.50x10 ⁻¹⁸
<i>MDFIC</i>	-1.17422	4.5755387	3.75x10 ⁻¹⁸
<i>ANXA2</i>	-1.20517	7.9355196	5.18x10 ⁻¹⁸
<i>CD24</i>	1.988625	10.50505	5.18x10 ⁻¹⁸
<i>DEFA1B</i>	2.428043	12.499908	5.18x10 ⁻¹⁸
<i>LILRB1</i>	-1.50588	4.9709885	6.97x10 ⁻¹⁸
<i>JAG1</i>	0.769817	5.492401	7.51x10 ⁻¹⁸
<i>DEFA1</i>	1.734214	13.40448	7.51x10 ⁻¹⁸
<i>DEFA3</i>	2.128242	12.925472	8.67x10 ⁻¹⁸
<i>IPO11</i>	1.220463	8.9406181	8.67x10 ⁻¹⁸
<i>ID3</i>	-1.49787	4.8311506	8.67x10 ⁻¹⁸
<i>HIST2H2BF</i>	-0.79569	5.8870554	1.00x10 ⁻¹⁷
<i>DEFA1B</i>	2.103971	13.090453	1.18x10 ⁻¹⁷
<i>SHISA5</i>	0.694576	11.830639	3.07x10 ⁻¹⁷
<i>C9ORF127</i>	-1.32776	4.1586408	3.07x10 ⁻¹⁷
<i>HBQ1</i>	1.571041	9.1763353	3.43x10 ⁻¹⁷
<i>DEFA1B</i>	2.341589	12.644121	3.43x10 ⁻¹⁷
<i>UNC45A</i>	0.60514	7.9089881	4.29x10 ⁻¹⁷
<i>SLCO3A1</i>	0.786012	9.4348591	5.48x10 ⁻¹⁷
<i>KMO</i>	-0.91605	5.3868238	5.77x10 ⁻¹⁷
<i>CD1C</i>	-0.8725	3.7413868	6.62x10 ⁻¹⁷
<i>HDGF</i>	0.603476	8.1067932	6.66x10 ⁻¹⁷
<i>LOC388588</i>	1.46321	8.5139749	8.34x10 ⁻¹⁷
<i>CTSA</i>	0.676356	10.184618	1.29x10 ⁻¹⁶
<i>CREB5</i>	0.648555	11.678305	1.42x10 ⁻¹⁶
<i>CACNA2D3</i>	-1.93144	4.4661918	1.42x10 ⁻¹⁶
<i>VNN2</i>	-0.90414	3.5456434	1.56x10 ⁻¹⁶
<i>SMARCD3</i>	1.338765	7.3219693	2.52x10 ⁻¹⁶
<i>C17ORF87</i>	-0.96227	7.0250718	2.52x10 ⁻¹⁶

GO Biological Process	FDR
antibacterial humoral response	0.00051
regulation of transcription in G1/S transition of mitotic cell cycle	0.00051
apoptotic cell clearance	0.0088
innate immune response in mucosa	0.01
mitotic metaphase plate congression	0.027
mitotic cell cycle	0.027

Table A.19: Summary of pathways enriched amongst genes differentially expressed in sepsis vs surgery.

Pathway	FDR
EIF2 Signalling	7.41×10^{-10}
Integrin Signalling	7.76×10^{-10}
Tec Kinase Signalling	4.47×10^{-9}
Production of nitric oxide and ROS in macrophages	6.46×10^{-9}
Natural Killer Cell Signalling	1.07×10^{-8}
B Cell Receptor Signalling	1.10×10^{-8}
Molecular Mechanisms of Cancer	1.15×10^{-8}
iCOS-iCOSL Signalling in T Helper Cells	1.78×10^{-8}
Renin-Angiotensin Signalling	3.63×10^{-8}
CD28 Signalling in T Helper Cells	4.07×10^{-8}
Prolactin Signalling	4.17×10^{-8}
ERK/MAPK Signalling	5.89×10^{-8}
Glucocorticoid Receptor Signalling	7.41×10^{-8}
JAK/Stat Signalling	7.94×10^{-8}
Protein Ubiquitination Pathway	8.71×10^{-8}
Breast Cancer Regulation by Stathmin1	1.07×10^{-7}
IL-3 Signalling	1.48×10^{-7}
Thrombopoietin Signalling	2.29×10^{-7}
Fc Receptor-mediated Phagocytosis in Macrophages and Monocytes	2.40×10^{-7}
Huntington's Disease Signalling	3.72×10^{-7}
PKC Signalling in T Lymphocytes	4.57×10^{-7}
FLT3 Signalling in Hematopoietic Progenitor Cells	5.01×10^{-7}
T Cell Receptor Signalling	5.13×10^{-7}
Regulation of eIF4 and p70S6K Signalling	5.25×10^{-7}
Erythropoietin Signalling	9.12×10^{-7}
GM-CSF Signalling	9.55×10^{-7}
fMLP Signalling in Neutrophils	1.12×10^{-6}
Insulin Receptor Signalling	1.15×10^{-6}
CTLA4 Signalling in Cytotoxic T Lymphocytes	1.38×10^{-6}
IL-15 Signalling	1.66×10^{-6}

Table A.20: Summary of pathways enriched amongst genes differentially expressed in trauma vs controls.

VALIDATION AND CHARACTERISATION OF SEPSIS RESPONSE SIGNATURES

Table B.1: Summary of genes differentially expressed between SRS1 and SRS2 in CAP (validation).

40 most significantly differentially expressed probes.

Gene	Log ₂ FC	Mean SRS1	Mean SRS2	FDR
<i>OTUD1</i>	1.04	9.04	8.00	2.78x10 ⁻¹⁵
<i>FAM38A</i>	-1.27	7.89	9.16	7.38x10 ⁻¹⁵
<i>FBXL10</i>	-1.08	7.47	8.55	7.38x10 ⁻¹⁵
<i>C9orf103</i>	2.03	9.69	7.67	7.91x10 ⁻¹⁵
<i>HPGD</i>	2.97	9.26	6.29	1.05x10 ⁻¹⁴
<i>NOL6</i>	-0.95	7.04	7.99	1.05x10 ⁻¹⁴
<i>TDRD9</i>	2.05	9.79	7.74	1.05x10 ⁻¹⁴
<i>DNASE1L1</i>	0.96	9.28	8.32	1.10x10 ⁻¹⁴
<i>FAM89A</i>	2.07	7.42	5.35	1.10x10 ⁻¹⁴
<i>SPPL2A</i>	0.87	10.70	9.84	1.10x10 ⁻¹⁴
<i>LOC100130092</i>	0.87	9.53	8.66	1.11x10 ⁻¹⁴
<i>PPRC1</i>	-0.97	5.78	6.76	1.11x10 ⁻¹⁴
<i>TDRD9</i>	1.96	9.97	8.01	1.34x10 ⁻¹⁴
<i>AZI2</i>	1.04	8.91	7.88	2.23x10 ⁻¹⁴
<i>M6PRBP1</i>	0.68	12.62	11.93	2.92x10 ⁻¹⁴
<i>FAM89A</i>	2.19	9.02	6.83	5.34x10 ⁻¹⁴
<i>NLRC4</i>	1.19	9.81	8.61	5.52x10 ⁻¹⁴
<i>KIAA0182</i>	-1.04	6.50	7.53	5.79x10 ⁻¹⁴
<i>LOC284023</i>	-0.89	5.03	5.92	7.64x10 ⁻¹⁴
<i>DPP7</i>	-1.03	7.31	8.34	9.08x10 ⁻¹⁴
<i>FBL</i>	-0.89	8.02	8.91	9.08x10 ⁻¹⁴
<i>LILRA6</i>	1.13	7.07	5.94	9.08x10 ⁻¹⁴
<i>TYSND1</i>	-0.83	6.66	7.49	1.45x10 ⁻¹³
<i>SDHC</i>	0.92	9.42	8.51	1.48x10 ⁻¹³
<i>FAM89A</i>	2.19	9.08	6.90	1.53x10 ⁻¹³
<i>DDAH2</i>	1.23	9.69	8.46	1.66x10 ⁻¹³
<i>KTELC1</i>	-0.97	8.17	9.14	1.70x10 ⁻¹³
<i>EIF3B</i>	-0.72	9.88	10.59	1.78x10 ⁻¹³
<i>TBC1D8</i>	1.61	9.15	7.54	2.05x10 ⁻¹³
<i>RPL18A</i>	-0.68	12.79	13.48	2.32x10 ⁻¹³

Continued on following page

Table B.1 - *continued from previous page*

Gene	Log ₂ FC	Mean SRS1	Mean SRS2	FDR
<i>SERPINB1</i>	1.11	13.09	11.98	2.32x10 ⁻¹³
<i>VOPP1</i>	1.26	7.12	5.87	2.44x10 ⁻¹³
<i>UBE2F</i>	0.76	11.08	10.32	2.54x10 ⁻¹³
<i>C1orf162</i>	0.80	13.45	12.66	2.60x10 ⁻¹³
<i>ZNF783</i>	-1.13	4.60	5.74	2.65x10 ⁻¹³
<i>MAPRE1</i>	0.69	11.39	10.70	2.67x10 ⁻¹³
<i>OTUD1</i>	1.19	7.40	6.21	2.71x10 ⁻¹³
<i>ANKS1A</i>	1.19	9.82	8.64	3.02x10 ⁻¹³
<i>EIF3B</i>	-0.93	7.45	8.38	3.02x10 ⁻¹³
<i>PAQR8</i>	-0.99	6.10	7.09	3.02x10 ⁻¹³

Pathway	FDR
Role of NFAT in Regulation of the Immune Response	3.44x10 ⁻⁸
OX40 Signaling Pathway	3.76x10 ⁻⁸
B Cell Development	1.09x10 ⁻⁷
CD28 Signaling in T Helper Cells	2.46x10 ⁻⁷
Calcium-induced T Lymphocyte Apoptosis	2.66x10 ⁻⁷
PKC Signaling in T Lymphocytes	8.28x10 ⁻⁷
Primary Immunodeficiency Signaling	2.89x10 ⁻⁶
iCOS-iCOSL Signaling in T Helper Cells	3.43x10 ⁻⁶
T Cell Receptor Signaling	4.77x10 ⁻⁶
Cdc42 Signaling	5.19x10 ⁻⁶
Allograft Rejection Signaling	6.30x10 ⁻⁶
Nur77 Signaling in T Lymphocytes	1.40x10 ⁻⁵
Communication between Innate and Adaptive Immune Cells	2.50x10 ⁻⁵
Altered T Cell and B Cell Signaling in Rheumatoid Arthritis	3.88x10 ⁻⁵
CCR5 Signaling in Macrophages	6.70x10 ⁻⁵
Granzyme B Signaling	1.08x10 ⁻⁴
CTLA4 Signaling in Cytotoxic T Lymphocytes	1.29x10 ⁻⁴
Graft-versus-Host Disease Signaling	1.47x10 ⁻⁴
Type I Diabetes Mellitus Signaling	1.57x10 ⁻⁴
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells	2.25x10 ⁻⁴
Antigen Presentation Pathway	2.63x10 ⁻⁴
Phospholipase C Signaling	4.06x10 ⁻⁴
Autoimmune Thyroid Disease Signaling	6.29x10 ⁻⁴

Table B.2: Summary of pathways enriched amongst genes differentially expressed between SRS1 and SRS2 in CAP (validation).

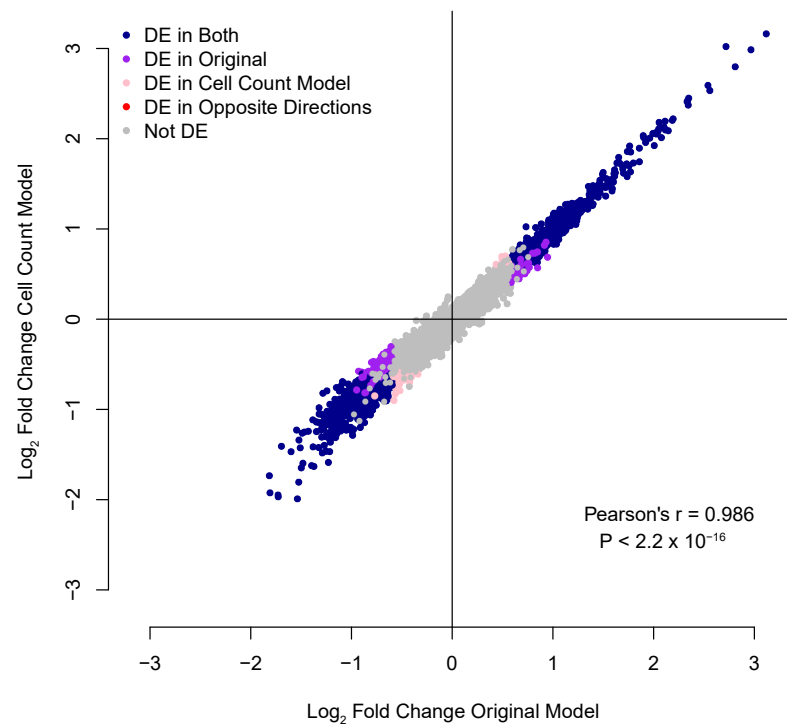


Figure B.1: Effect of differential cell count on differential expression analysis.

The log fold change for differential expression of all genes between SRS1 and SRS2 accounting for differences in relative mononucleocyte, polynucleocyte, and lymphocyte counts is plotted against the fold changes for the same genes in the original model. Blue colouring indicates significant differential expression in both models; purple in the original model only; pink in the cell count model only; grey indicates the gene is not significantly differentially expressed between SRS1 and SRS2 in either model.

Table B.3: In vitro studies of the transcriptomic endotoxin tolerance response.

Study	Samples	Platform	ET Analysis Approach	Accession
Foster et al. (2007)	Mouse macrophages (n=2): untreated (N), stimulated once (N+L) or twice (T+L) with LPS	Affymetrix Mouse Genome 430 2.0 arrays	Only LPS-induced genes considered. Tolerizeable genes defined as $(N+L)/(T+L) > 3$. Non-tolerizeable genes defined as $(N+L)/(T+L) \geq 1$	GSE7348
Mages et al. (2007)	Mouse bone marrow derived macrophages (n=3): prestimulated (1_X) or not (0_X) for 18h, 2h rest, cells either restimulated (X_1) or not (X_0)	Affymetrix GeneChip MOE430 2.0	All pairwise comparisons between treatments, presented as comparisons with untreated cells	GSE8621
Fresno et al. (2009)	Human monocytes (n=2): unstimulated, stimulated, and tolerant cells with three different exposure times	Illumina Sentrix HumanRef-8_V2 BeadChip	LPS-induced genes classified as tolerizeable and nontolerizeable in the ET phase, genes moving between these groups over time described	GSE15219
Pena et al. (2011)	Human PBMCs (n=4): untreated or challenged with LPS (10 ng/ml) in a single dose (LPS) or two doses at a 24h interval (LPS/LPS) and incubated for 4h	Illumina/GenomeET: BC Microarray Platform	differentially expressed between LPS/LPS and untreated cells.	GSE22248

Yang et al. (2012)	Human PBMCs (n=3) left untreated (N) or LPS stimulated for 16h (T), washed and 2h recovery; then given media (N, T) or LPS 6h (N+L, T+L)	NimbleGen array and GenePix Pro 6.0	Gene expression compared between untreated, stimulated, restimulated. T vs N most changes, but 356 non-tolerizeable genes also identified and focussed on as potential regulators of TLR-induced inflammation.	NA
Allantaz-Frager et al. (2013)	Healthy PBMCs (n=6): unstimulated (medium), 1 dose LPS (LPS unprimed), 2 doses of LPS (LPS primed) or 2 doses of LPS and interferon gamma (LPS primed + IFNg)	GeneChip Human Genome U133 Plus 2.0 Array (Affymetrix)	43 transcripts DE between LPS unprimed and LPS primed, 70 between medium and LPS unprimed - split into tolerizeable and non-tolerizeable. Selected transcripts associated with development of ET identified, expression restored after immunostimulation with IFN- γ . Validation by qRT-PCR in healthy donors and septic patients	GSE46914
O'Carroll et al. (2014)	Mouse bone marrow derived macrophages (n=3). Untreated (N), acute response to LPS (LPS activation group), LPS tolerance (T) and recovered (RM).	Agilent mouse 8x60K arrays	<i>k</i> -means clustering showed unique transcriptional signature for each treatment group, 10 gene expression profiles across groups. 30 up and downregulated genes compared with recovered macrophages. Focus on recovery from endotoxin tolerance, no gene list for tolerance given.	GEO accession number: GSE47783 raw and processed data

Table B.4: Summary of genes differentially expressed between SRS1 and SRS2 in FP.
40 most significantly differentially expressed probes.

Gene	Log ₂ FC	Mean SRS1_FP	Mean SRS2_FP	FDR
<i>FBXW2</i>	1.06	9.19	8.13	9.36x10 ⁻²⁴
<i>LOC646463</i>	1.41	9.62	8.22	9.36x10 ⁻²⁴
<i>ENTPD7</i>	1.51	8.11	6.60	1.69x10 ⁻²²
<i>SDHC</i>	0.92	9.81	8.89	1.69x10 ⁻²²
<i>ST6GALNAC3</i>	1.16	8.96	7.80	3.18x10 ⁻²²
<i>RPS6KA5</i>	-1.49	7.54	9.03	3.29x10 ⁻²²
<i>C14ORF159</i>	-1.33	8.40	9.73	5.00x10 ⁻²²
<i>SORT1</i>	1.31	11.14	9.83	6.05x10 ⁻²²
<i>OSTALPHA</i>	2.22	9.58	7.35	6.06x10 ⁻²²
<i>LDHA</i>	1.07	13.74	12.68	1.54x10 ⁻²¹
<i>ST6GALNAC3</i>	0.78	7.27	6.49	3.59x10 ⁻²¹
<i>LDHA</i>	0.98	13.45	12.46	4.83x10 ⁻²¹
<i>C5ORF32</i>	1.60	12.66	11.06	5.01x10 ⁻²¹
<i>PHCA</i>	1.22	11.30	10.08	5.73x10 ⁻²¹
<i>IDI1</i>	1.48	11.06	9.58	6.72x10 ⁻²¹
<i>CLEC4D</i>	1.47	11.15	9.67	7.95x10 ⁻²¹
<i>MAPRE1</i>	0.73	11.94	11.20	1.66x10 ⁻²⁰
<i>GADD45A</i>	1.53	10.99	9.46	1.87x10 ⁻²⁰
<i>PFKFB2</i>	1.62	9.38	7.76	2.58x10 ⁻²⁰
<i>UBE2H</i>	1.04	10.50	9.46	3.65x10 ⁻²⁰
<i>C1ORF162</i>	0.86	13.96	13.11	3.89x10 ⁻²⁰
<i>GALNT2</i>	0.74	7.13	6.39	3.89x10 ⁻²⁰
<i>SCPEP1</i>	1.06	11.93	10.87	4.00x10 ⁻²⁰
<i>SERPINB1</i>	1.03	13.31	12.28	4.92x10 ⁻²⁰
<i>SLC25A28</i>	0.91	11.22	10.31	6.24x10 ⁻²⁰
<i>UBE2H</i>	1.50	8.68	7.19	6.73x10 ⁻²⁰
<i>PAG1</i>	1.08	9.55	8.46	1.40x10 ⁻¹⁹
<i>MERTK</i>	1.82	8.74	6.92	1.40x10 ⁻¹⁹
<i>GADD45A</i>	1.48	11.97	10.50	1.53x10 ⁻¹⁹
<i>TNFAIP8L3</i>	1.50	7.75	6.26	1.55x10 ⁻¹⁹
<i>TIPARP</i>	1.06	10.46	9.40	1.98x10 ⁻¹⁹
<i>SNX3</i>	0.88	13.87	13.00	2.79x10 ⁻¹⁹
<i>FAM105A</i>	0.99	9.48	8.49	3.64x10 ⁻¹⁹
<i>CCDC53</i>	0.67	10.06	9.39	3.64x10 ⁻¹⁹
<i>MAPK6</i>	1.08	8.96	7.88	4.37x10 ⁻¹⁹
<i>ANKS1A</i>	0.99	10.43	9.43	7.44x10 ⁻¹⁹
<i>PFKFB2</i>	1.87	10.76	8.90	1.00x10 ⁻¹⁸
<i>GYG1</i>	1.08	13.91	12.82	1.04x10 ⁻¹⁸
<i>C9ORF103</i>	1.50	10.65	9.15	1.14x10 ⁻¹⁸
<i>GPR84</i>	1.86	10.55	8.69	1.14x10 ⁻¹⁸

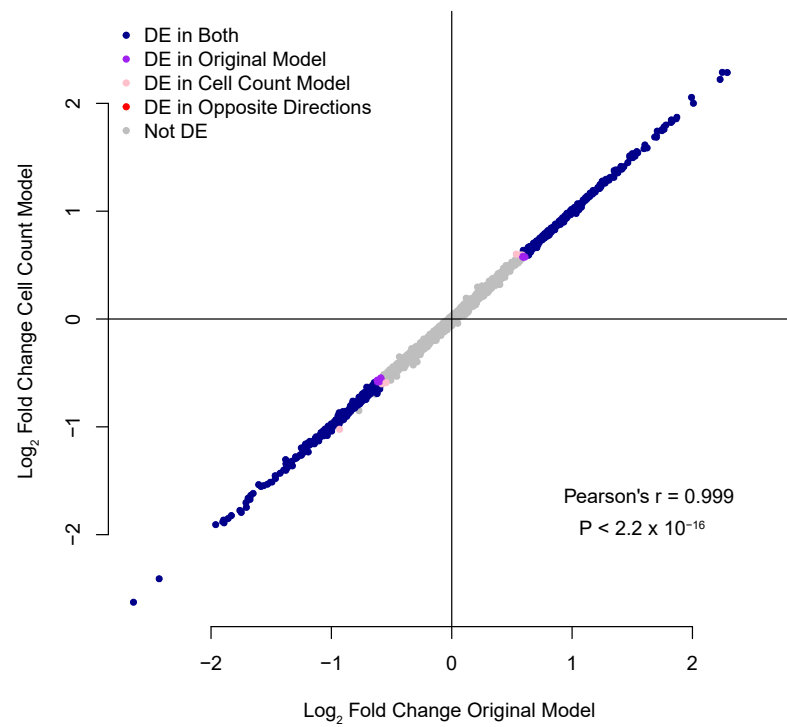


Figure B.2: Effect of differential cell count on differential expression analysis.

The log fold change for differential expression of all genes between SRS1_FP and SRS2_FP accounting for differences in relative mononucleocyte, polynucleocyte, and lymphocyte counts is plotted against the fold changes for the same genes in the original model. Blue colouring indicates significant differential expression in both models; purple in the original model only; pink in the cell count model only; grey indicates the gene is not significantly differentially expressed between SRS1 and SRS2 in either model.

Pathway	FDR
Altered T Cell and B Cell Signaling in Rheumatoid Arthritis	1.02x10 ⁻⁸
Communication between Innate and Adaptive Immune Cells	2.95x10 ⁻⁷
Type I Diabetes Mellitus Signaling	3.80x10 ⁻⁶
Calcium-induced T Lymphocyte Apoptosis	8.91x10 ⁻⁶
Role of NFAT in Regulation of the Immune Response	1.07x10 ⁻⁵
Allograft Rejection Signaling	2.40x10 ⁻⁵
TREM1 Signaling	2.75x10 ⁻⁵
Dendritic Cell Maturation	3.98x10 ⁻⁵
Autoimmune Thyroid Disease Signaling	5.89x10 ⁻⁵
Interferon Signaling	9.12x10 ⁻⁵
Nur77 Signaling in T Lymphocytes	9.55x10 ⁻⁵
Primary Immunodeficiency Signaling	1.12x10 ⁻⁴
OX40 Signaling Pathway	1.35x10 ⁻⁴
Graft-versus-Host Disease Signaling	1.66x10 ⁻⁴
Hematopoiesis from Pluripotent Stem Cells	1.82x10 ⁻⁴
B Cell Development	2.75x10 ⁻⁴
Role of PRRs in Recognition of Bacteria and Viruses	4.07x10 ⁻⁴
LXR/RXR Activation	4.90x10 ⁻⁴
CD28 Signaling in T Helper Cells	5.25x10 ⁻⁴
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells	5.62x10 ⁻⁴
Cdc42 Signaling	5.75x10 ⁻⁴
Antigen Presentation Pathway	5.89x10 ⁻⁴
iCOS-iCOSL Signaling in T Helper Cells	7.24x10 ⁻⁴
Eicosanoid Signaling	8.71x10 ⁻⁴
Phospholipase C Signaling	9.12x10 ⁻⁴
p38 MAPK Signaling	9.55x10 ⁻⁴
Adenine and Adenosine Salvage III	1.29x10 ⁻³
Purine Ribonucleosides Degradation to Ribose-1-phosphate	1.29x10 ⁻³
Role of macrophages, fibroblasts and endothelial cells in Rheumatoid Arthritis	1.78x10 ⁻³
IL-10 Signaling	1.91x10 ⁻³
p53 Signaling	2.45x10 ⁻³
Airway Pathology in Chronic Obstructive Pulmonary Disease	3.39x10 ⁻³
PPAR/RXR Activation	3.63x10 ⁻³
CCR5 Signaling in Macrophages	3.89x10 ⁻³
T Helper Cell Differentiation	3.89x10 ⁻³
ERK5 Signaling	3.89x10 ⁻³
D-myo-inositol (1,4,5,6)-Tetrakisphosphate Biosynthesis	4.79x10 ⁻³
D-myo-inositol (3,4,5,6)-tetrakisphosphate Biosynthesis	4.79x10 ⁻³
LPS/IL-1 Mediated Inhibition of RXR Function	6.03x10 ⁻³
CTLA4 Signaling in Cytotoxic T Lymphocytes	6.76x10 ⁻³

Table B.5: Summary of pathways enriched amongst genes differentially expressed between SRS1 and SRS2 in FP.

THE PLASMA PROTEOME IN SEPSIS DUE TO FAECAL PERITONITIS

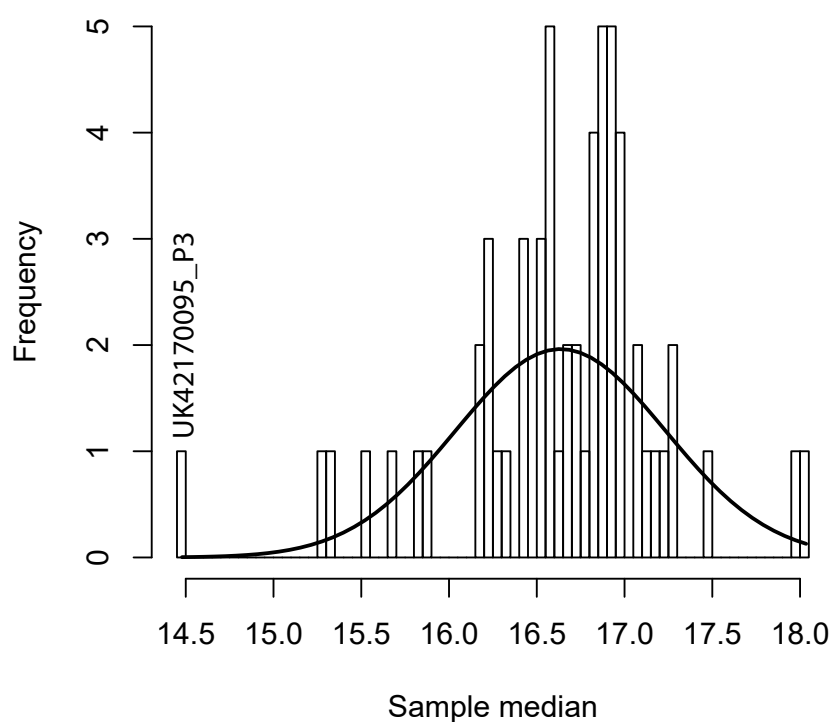


Figure C.1: Sample quality control by calculation of the sample median.
Samples with outlying median values were identified prior to secondary analysis.

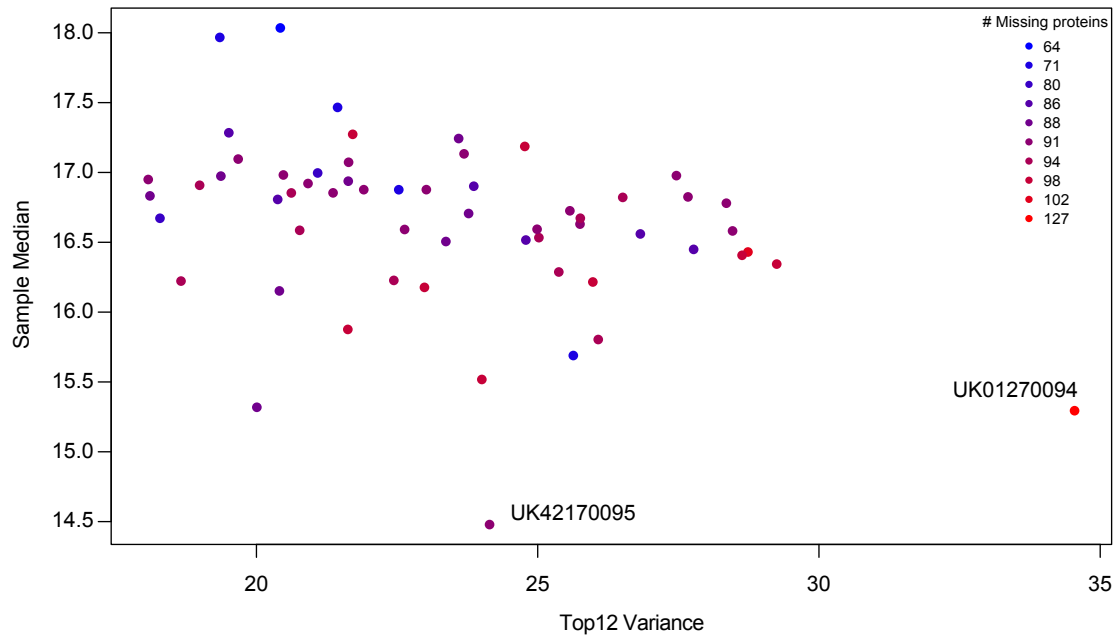


Figure C.2: Sample quality control by comparison of the variance of depleted proteins and the sample median.

Samples with outlying variances and means were identified prior to secondary analysis.

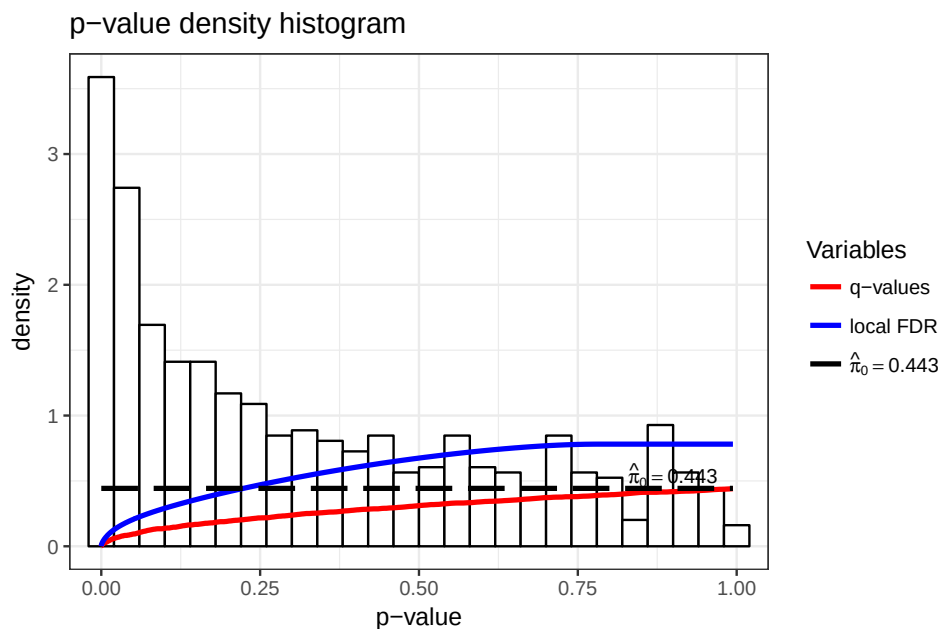


Figure C.3: Histogram of p values for differential abundance analysis.

The distribution of p-values from the SRS contrast is shown, confirming that the imputation step did not have a distorting effect on the data.

Symbol	Log FC	Q Value	Protein name
SPTAN1	0.69	4.44x10 ⁻⁴	spectrin alpha, non-erythrocytic 1
TXN	0.77	0.002	thioredoxin
CPA1	-0.72	0.007	carboxypeptidase A1
FKBP1A	0.69	0.009	FK506 binding protein 1A
COL1A2	0.65	0.010	collagen type I alpha 2 chain
S100A4	0.72	0.013	S100 calcium binding protein A4
NCBP2-AS2	0.73	0.023	NCBP2 antisense RNA 2 (head to head)
NAP1L1	0.60	0.024	nucleosome assembly protein 1 like 1
FABP3	0.91	0.026	fatty acid binding protein 3
MANF	0.67	0.032	mesencephalic astrocyte derived neurotrophic factor

Table C.1: Proteins differentially abundant between SRS1 and SRS2

The ten proteins that had a Q value <0.05 and a log fold change >1.5 are shown.

Table C.2: Summary of proteins differentially abundant between proteomic clusters.
40 most significantly differentially abundant proteins.

SwissProt ID	Log ₂ fold change	Mean abundance	Q Value
PROF1_HUMAN	0.63	0.52	4.05E-09
K7EJ44_HUMAN	0.63	-0.54	1.89E-08
PDIA3_HUMAN	0.68	-0.33	7.39E-08
TPM1_HUMAN	0.76	-0.63	7.72E-08
G3V1V0_HUMAN	0.90	0.05	3.06E-07
TPM4_HUMAN	0.93	0.10	4.21E-07
PGK1_HUMAN	0.59	-0.86	4.62E-07
ACTN1_HUMAN	0.86	-0.45	4.62E-07
TPIS_HUMAN	0.61	-0.99	9.28E-07
TAGL2_HUMAN	0.61	-0.40	1.22E-06
TPM3_HUMAN	0.91	-0.51	1.23E-06
PLEK_HUMAN	0.62	0.22	1.43E-06
ZYX_HUMAN	0.73	0.14	1.55E-06
S10A4_HUMAN	1.17	-1.77	8.11E-06
HSPB1_HUMAN	1.00	-1.11	9.11E-06
F5H5D3_HUMAN	0.95	-1.15	1.02E-05
PGAM1_HUMAN	0.76	-1.05	1.48E-05
A0A087X243_HUMAN	0.68	-0.64	2.41E-05
H3BM21_HUMAN	0.89	-0.96	3.66E-05
Q5TA02_HUMAN	0.64	-0.94	4.21E-05
B4E3S0_HUMAN	0.61	-0.94	4.21E-05
E7EPV7_HUMAN	0.95	-1.70	4.76E-05
PEBP1_HUMAN	0.61	-0.39	5.49E-05
CAP1_HUMAN	0.61	-0.73	5.75E-05

Continued on following page

Table C.2 - *continued from previous page*

SwissProt ID	Log ₂ fold change	Mean abundance	Q Value
PPIB_HUMAN	0.80	-1.59	5.86E-05
A0A1B0GUS4_HUMAN	0.81	-0.76	6.36E-05
Q5T123_HUMAN	0.85	-0.94	6.36E-05
PLF4_HUMAN	1.08	0.03	8.58E-05
SRGN_HUMAN	0.61	-1.13	0.00018
H0YGX7_HUMAN	0.60	-0.92	0.00028
THIO_HUMAN	0.82	-1.36	0.00029
PDLI5_HUMAN	0.61	-1.21	0.00034
GPV_HUMAN	0.77	-1.18	0.00051
ILEU_HUMAN	0.59	-1.44	0.00056
H0YH88_HUMAN	0.84	-1.07	0.00057
I3L397_HUMAN	0.61	-1.39	0.00059
H3BTN5_HUMAN	0.63	-0.83	0.00077
6PGD_HUMAN	1.53	-2.21	0.00084
NEUG_HUMAN	1.29	-1.87	0.00087
F13A_HUMAN	0.63	-1.14	0.00093

GO Biological Process	FDR
Movement of cell or subcellular component	5.50x10 ⁻⁵
Actin cytoskeleton organization	8.70x10 ⁻⁵
Muscle filament sliding	1.00x10 ⁻³
Canonical glycolysis	1.00x10 ⁻³
Oxidation-reduction process	1.20x10 ⁻³
Protein stabilization	6.30x10 ⁻³
Positive regulation of transcription from RNA pol II promoter	9.00x10 ⁻³
Platelet activation	9.50x10 ⁻³
Platelet degranulation	1.60x10 ⁻²

Table C.3: Summary of pathways enriched amongst genes differentially expressed between the proteomic clusters.

Table C.4: Summary of proteins differentially abundant over time.

All significantly differentially abundant proteins.

SwissProt ID	Gene	FDR
A0A0D9SF54_HUMAN	SPTAN1	0.012
MASP1_HUMAN	MASP1	0.012
H0YGX7_HUMAN	ARHGDIB	0.022
K1C9_HUMAN	KRT9	0.022
A0A087WTA8_HUMAN	COL1A2	0.022
I3L145_HUMAN	SHBG	0.022
CAZA1_HUMAN	CAPZA1	0.022
IGHA1_HUMAN	IGHA1	0.022
S10A4_HUMAN	S100A4	0.022
ILEU_HUMAN	SERPINB1	0.022
FKB1A_HUMAN	FKBP1A	0.022
SEPP1_HUMAN	SEPP1	0.022
THIO_HUMAN	TXN	0.022
PGAM1_HUMAN	PGAM1	0.022
I3L162_HUMAN	CTRL	0.022
LEG1_HUMAN	LGALS1	0.023
I3L397_HUMAN	EIF5A	0.023
G3V264_HUMAN	SERPINA5	0.028
G3V4W5_HUMAN	SHMT2	0.034
FETUB_HUMAN	FETUB	0.034
NUCB1_HUMAN	NUCB1	0.043
CBPB2_HUMAN	CPB2	0.043
A0A087WXX2_HUMAN	ALDOB	0.047
CBPA1_HUMAN	CPA1	0.047
TSK_HUMAN	TSKU	0.047
FAM3C_HUMAN	FAM3C	0.047

REGULATORY GENETIC VARIATION IN THE SYSTEMIC INFLAMMATORY RESPONSE

Table D.1: Peak eQTL in sepsis patients

The top 40 most significant eQTL are shown, with the peak SNP-gene pairs, the effect size, and the p value for each.

Gene	SNP	eQTL Beta	P value
<i>CHURC1</i>	rs6745	1.38	6.27×10^{-172}
<i>ZSWIM7</i>	rs2286795	-0.70	6.70×10^{-146}
<i>USMG5</i>	rs7911488	1.82	1.48×10^{-145}
<i>CLEC12A</i>	rs7309256	1.82	6.55×10^{-136}
<i>GLMP</i>	rs949464	0.77	7.32×10^{-136}
<i>ACCS</i>	rs2074038	1.99	8.18×10^{-135}
<i>PAQR6</i>	rs2274226	1.08	4.0×10^{-132}
<i>PPP2R3C</i>	rs57075379	-0.52	3.14×10^{-131}
<i>LIAT1</i>	rs11150881	1.61	4.61×10^{-131}
<i>PHACS</i>	rs2074038	1.78	1.99×10^{-128}
<i>ENC1</i>	rs6867790	1.59	1.07×10^{-124}
<i>ERAP2</i>	rs7703341	1.25	2.80×10^{-124}
<i>WAC</i>	rs332186	-0.46	6.83×10^{-121}
<i>PAQR6</i>	rs12128955	1.07	1.20×10^{-120}
<i>PPP2R3C</i>	rs57075379	-0.52	8.87×10^{-120}
<i>KCTD10</i>	rs6663	0.56	8.91×10^{-120}
<i>C7orf25</i>	rs647534	0.70	2.52×10^{-119}
<i>SRP14</i>	rs16924528	0.40	2.29×10^{-115}
<i>IRF5</i>	rs4731533	0.94	2.83×10^{-115}
<i>C17orf97</i>	rs11150881	1.07	5.09×10^{-115}
<i>XRN2</i>	rs6047417	1.74	1.88×10^{-114}
<i>COPS3</i>	rs4985761	0.43	3.01×10^{-114}
<i>PFAAP5</i>	rs9337	0.40	2.15×10^{-111}
<i>LPCAT2</i>	rs12934747	-1.27	3.14×10^{-111}
<i>CRIPAK</i>	rs113963956	1.22	5.80×10^{-109}
<i>GOLGB1</i>	rs9859702	-0.87	5.15×10^{-107}
<i>CPNE1</i>	rs143646210	-0.64	5.56×10^{-106}
<i>CPEB4</i>	rs7736263	0.65	6.40×10^{-106}
<i>PEX6</i>	rs2395943	-0.57	6.53×10^{-106}
<i>HMBOX1</i>	rs10088428	-0.55	2.69×10^{-105}

Continued on following page

Table D.1 - *continued from previous page*

Gene	SNP	eQTL Beta	P value
DGKQ	rs1377586	0.67	3.53x10 ⁻¹⁰⁵
PPIL3	rs2136600	-0.48	2.53x10 ⁻¹⁰⁴
KIAA0367	rs560094	1.92	3.96x10 ⁻¹⁰³
PAM	rs2431321	0.69	3.64x10 ⁻⁹⁹
CPNE1	rs61016611	-0.54	1.13x10 ⁻⁹⁸
NUDT7	rs80159064	2.06	2.50x10 ⁻⁹⁸
RNPEP	rs10800804	-0.50	7.50x10 ⁻⁹⁷
ASNSD1	rs10931430	-0.52	4.22x10 ⁻⁹⁶
WDR48	rs1053530	0.41	7.91x10 ⁻⁹⁶

Table D.2: Top 40 interactions with source of sepsis.

The top 40 most significant interactions with source of sepsis are shown, with the peak SNP-gene pairs, the effect size for genotype and interaction, and the p value for each.

Gene	SNP	eQTL Beta	Interaction Beta	P value
S100P	rs3822262	-0.62	0.32	9.17 x 10 ⁻⁸
SCCPDH	rs7526237	-0.11	-0.45	8.10 x 10 ⁻⁷
OAS1	rs10850097	-0.87	0.40	2.18 x 10 ⁻⁶
UEVLD	rs10766486	-0.10	-0.22	1.38 x 10 ⁻⁵
LAS2	rs4940331	-0.11	-0.31	1.52 x 10 ⁻⁵
BARD1	rs6729681	0.52	0.37	4.08 x 10 ⁻⁵
ALOX5AP	rs12429692	0.16	-0.11	6.01 x 10 ⁻⁵
IL18RAP	rs917997	-0.45	0.23	8.78 x 10 ⁻⁵
ZNF641	rs2450988	-0.64	0.19	9.14 x 10 ⁻⁵
FAM89A	rs6700792	-0.30	-0.41	0.00011
RAB3IL1	rs2521561	-0.81	-0.57	0.00011
TMEM176A	rs7806458	1.46	-0.48	0.00011
MFN2	rs28718032	0.51	-0.19	0.00016
FAM89A	rs6698364	-0.30	-0.40	0.00017
NUP107	rs2431655	-0.14	0.12	0.00018
FAM89A	rs11122198	-0.17	-0.32	0.00020
FAM89A	rs6698364	-0.28	-0.36	0.00027
PLOD1	rs56243319	0.25	-0.13	0.00029
ADCK3	rs3738725	-0.44	-0.16	0.00038
FBXL20	rs801419	0.18	-0.12	0.00040
FCN1	rs10858300	0.24	0.17	0.00041
LILRB4	rs370156	0.34	-0.29	0.00044
CYP26B1	rs62147573	1.45	-0.50	0.00049
ADCY3	rs11892869	1.02	0.34	0.00067
ELP3	rs34460067	0.25	0.12	0.00073
TNFRSF10A	rs13254617	0.13	-0.13	0.00077
LRRN1	rs1125735	-0.29	-0.25	0.00078

Continued on following page

Table D.2 - *continued from previous page*

Gene	SNP	eQTL Beta	Interaction Beta	P value
<i>IRF5</i>	rs4731533	0.99	-0.22	0.00083
<i>RBM6</i>	rs6771546	0.32	0.12	0.00099
<i>TOR1A</i>	rs7873555	-0.05	-0.08	0.0010
<i>CCDC77</i>	rs2041299	-0.20	-0.12	0.0011
<i>NUDT14</i>	rs3784239	-0.10	-0.17	0.0012
<i>MFGE8</i>	rs12900329	-0.37	-0.30	0.0012
<i>ATOX1</i>	rs2915824	-0.06	-0.10	0.0013
<i>ZNF333</i>	rs60385634	-0.26	0.17	0.0013
<i>C19orf60</i>	rs4808823	-0.16	-0.11	0.0014
<i>GPR162</i>	rs2857230	-0.33	0.32	0.0014
<i>ARFGAP3</i>	rs9611902	-0.14	-0.11	0.0014
<i>NME4</i>	rs77371232	0.53	0.17	0.0015
<i>ARPC4</i>	rs11131186	0.09	0.14	0.0016

Table D.3: Top 40 interactions with SRS group.

The top 40 most significant interactions with SRS group are shown, with the peak SNP-gene pairs, the effect size for genotype and interaction, and the p value for each.

Gene	SNP	eQTL Beta	Interaction Beta	P value
<i>CYP26B1</i>	rs62147573	0.66	1.14	1.10×10^{-22}
<i>TAF1C</i>	rs1804500	0.38	-0.27	5.23×10^{-18}
<i>RAB3IL1</i>	rs2521561	-1.56	1.00	7.33×10^{-17}
<i>DNAJC27</i>	rs916485	0.54	-0.32	6.74×10^{-15}
<i>DDX1</i>	rs6755613	-0.24	0.15	1.23×10^{-14}
<i>AMACR</i>	rs34680	1.01	-0.59	6.75×10^{-14}
<i>MOSPD3</i>	rs7812123	-0.64	0.35	7.08×10^{-14}
<i>BLM</i>	rs389480	1.54	-0.65	1.06×10^{-13}
<i>BARDI</i>	rs6729681	0.86	-0.45	6.32×10^{-13}
<i>FAM89A</i>	rs6698364	-0.75	0.59	1.28×10^{-12}
<i>ACSL5</i>	rs11813362	0.45	-0.22	1.93×10^{-12}
<i>S100P</i>	rs3822262	-0.34	-0.35	2.23×10^{-12}
<i>COPS3</i>	rs4985761	0.54	-0.18	2.73×10^{-12}
<i>GSDMD</i>	rs2242093	0.41	-0.26	5.24×10^{-12}
<i>FAM89A</i>	rs6700792	-0.73	0.57	7.50×10^{-12}
<i>RBM6</i>	rs6771546	0.48	-0.20	1.78×10^{-11}
<i>FAM89A</i>	rs6698364	-0.67	0.51	1.92×10^{-11}
<i>ADCY3</i>	rs11892869	1.40	-0.50	5.35×10^{-11}
<i>SCPEP1</i>	rs9303389	0.47	-0.36	6.31×10^{-11}
<i>VKORC1L1</i>	rs402418	1.11	-0.50	1.42×10^{-10}
<i>FCN1</i>	rs10858300	0.44	-0.24	6.65×10^{-10}
<i>PITPNA</i>	rs113289613	0.80	-0.46	2.10×10^{-9}
<i>LGALS2</i>	rs2281097	0.79	0.60	2.46×10^{-9}

Continued on following page

Table D.3 - continued from previous page

Gene	SNP	eQTL Beta	Interaction Beta	P value
<i>HIP1</i>	rs1185883	-0.08	-0.28	2.56x10 ⁻⁹
<i>ACSL5</i>	rs17129748	0.57	-0.30	4.86x10 ⁻⁹
<i>KRT23</i>	rs2872963	0.96	-0.50	9.99x10 ⁻⁹
<i>D4S234E</i>	rs2369110	-0.35	-0.56	1.13x10 ⁻⁸
<i>IRF5</i>	rs4731533	0.74	0.32	3.39x10 ⁻⁸
<i>PTER</i>	rs1055337	-0.59	-0.33	3.52x10 ⁻⁸
<i>RFWD3</i>	rs8052367	0.50	-0.20	7.50x10 ⁻⁸
<i>ARFGAP3</i>	rs9611902	-0.25	0.15	1.25x10 ⁻⁷
<i>ARPC4</i>	rs11131186	0.25	-0.19	1.44x10 ⁻⁷
<i>ADORA2B</i>	rs11078319	0.43	-0.33	1.76x10 ⁻⁷
<i>IL18RAP</i>	rs917997	-0.26	-0.23	2.06x10 ⁻⁷
<i>RIOK1</i>	rs67228999	-0.29	0.17	2.16x10 ⁻⁷
<i>FAM89A</i>	rs11122198	-0.46	0.36	3.35x10 ⁻⁷
<i>ASNSD1</i>	rs10931430	-0.63	0.19	3.43x10 ⁻⁷
<i>CDC16</i>	rs7324744	0.34	-0.11	4.08x10 ⁻⁷
<i>NEDD1</i>	rs11108747	-0.46	0.25	4.79x10 ⁻⁷
<i>FAM118B</i>	rs35063253	0.77	-0.36	4.96x10 ⁻⁷
<i>GLG1</i>	rs12927351	0.24	-0.12	7.86x10 ⁻⁷

Table D.4: Peak eQTL in cardiac surgery patients

The 43 eQTL are shown, with the peak SNP-gene pairs, the effect size, and the p value for each.

Gene	SNP	eQTL Beta	P value
<i>LPCAT2</i>	rs11647464	1.14	3.6 x 10 ⁻¹⁸
<i>ERAP2</i>	rs7719705	-1.07	5.4 x 10 ⁻¹⁷
<i>TACSTD2</i>	rs12118488	2.47	8.4 x 10 ⁻¹⁶
<i>LILRA3</i>	rs103294	-1.14	9.2 x 10 ⁻¹⁶
<i>MRVI1</i>	rs7115310	0.83	8.3 x 10 ⁻¹⁵
<i>LRRC6</i>	rs11988034	-0.78	1.6 x 10 ⁻¹³
<i>MYOM2</i>	rs10110395	1.40	1.7 x 10 ⁻¹³
<i>S100P</i>	rs3822262	-1.08	6.1 x 10 ⁻¹³
<i>OCIAD2</i>	rs2289065	-0.46	1.3 x 10 ⁻¹²
<i>GSDMB</i>	rs907091	0.26	2.2 x 10 ⁻¹²
<i>LOC401233</i>	rs6918464	-1.05	2.80E-12
<i>ROGDI</i>	rs2290562	0.55	3.39E-12
<i>LIPN</i>	rs10509554	0.58	4.46E-12
<i>IL18RAP</i>	rs2110735	-0.50	5.59E-12
<i>CENPK</i>	rs11746096	-0.53	2.63E-11
<i>ACP5</i>	rs2071484	0.68	2.64E-11
<i>KIAA0367</i>	rs560094	1.64	3.64E-10
<i>CPEB4</i>	rs7736263	0.57	4.35E-10

Continued on following page

Table D.4 - *continued from previous page*

Gene	SNP	eQTL Beta	P value
<i>S100A12</i>	rs3006475	0.58	4.54E-10
<i>PADI2</i>	rs2235910	-0.41	7.05E-10
<i>LGALS8</i>	rs3820564	-0.30	8.18E-10
<i>HABP4</i>	rs10820507	-0.42	1.09E-09
<i>ANKRD34B</i>	rs1677703	-0.79	2.72E-09
<i>RASGRP4</i>	rs10993	-0.58	4.00E-09
<i>RAB20</i>	rs7327721	-0.44	1.11E-08
<i>KCNJ15</i>	rs9984176	0.21	1.15E-08
<i>IL11RA</i>	rs2070074	0.39	2.83E-08
<i>FAM117B</i>	rs7568438	-0.37	4.08E-08
<i>DUSP14</i>	rs9890344	-0.38	4.90E-08
<i>SLFN5</i>	rs9908966	-0.23	8.35E-08
<i>LPIN1</i>	rs3795974	-0.17	1.23E-07
<i>HES4</i>	rs3128117	0.54	1.28E-07
<i>MERTK</i>	rs10180018	0.54	1.48E-07
<i>GLT25D1</i>	rs7251306	0.16	1.54E-07
<i>SERPINA1</i>	rs1243160	-0.51	1.58E-07
<i>NFE2</i>	rs35979828	-0.23	1.59E-07
<i>CTSF</i>	rs10791889	0.34	1.64E-07
<i>D4S234E</i>	rs6843595	-0.85	1.81E-07
<i>TMEM204</i>	rs3784828	0.37	2.49E-07
<i>ZBTB25</i>	rs11627387	-0.19	2.63E-07
<i>PDE9A</i>	rs2284965	-0.48	2.69E-07
<i>MERTK</i>	rs13419523	0.95	3.43E-07
<i>FAM117B</i>	rs7568438	-0.34	4.46E-07

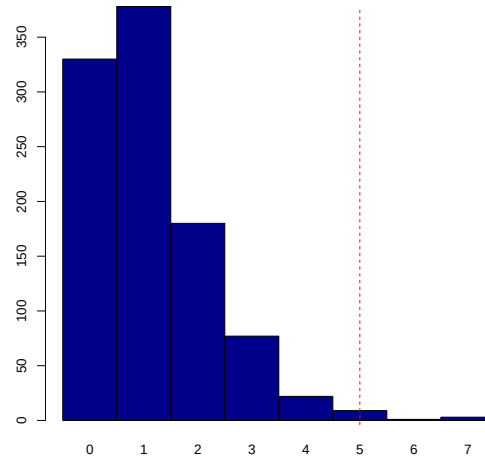


Figure D.1: Permutation of surgery interaction QTL.

Sample type was permuted 1000 times and the number of significant eQTL interactions recorded. For the true sample assignments, five interactions were reported (*red* line), which is more than would be expected by chance ($p=0.013$).

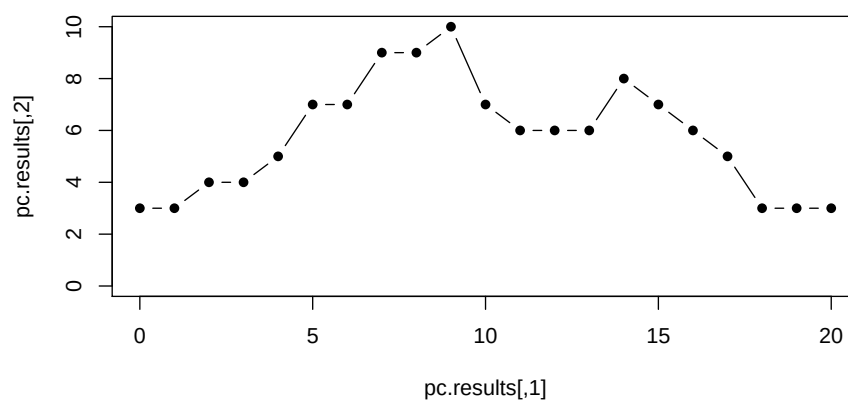


Figure D.2: Inclusion of PCs in pQTL mapping.

pQTL were mapped using a linear model, including increasing numbers of principal component to correct for systematic confounding effects. Most pQTL were detected when nine PCs were included in the model.

BIBLIOGRAPHY

- Adams, D., Altucci, L., Antonarakis, S. E., et al. (2012). "BLUEPRINT to decode the epigenetic signature written in blood." *Nature Biotechnology* 30: pp. 224–6.
- Albert, F. W. and Kruglyak, L. (2015). "The role of regulatory variation in complex traits and disease." *Nature Reviews Genetics* 16: pp. 197–212.
- Allantaz-Frager, F., Turrel-Davin, F., Venet, F., et al. (2013). "Identification of biomarkers of response to IFN γ during endotoxin tolerance: application to septic shock". *PLoS One* 8(7): e68218.
- Altman, N. and Krzywinski, M. (2017). "Points of Significance: Clustering". *Nature Methods* 14: pp. 545–546.
- Anderson, C. A., Pettersson, F. H., Clarke, G. M., et al. (2010). "Data quality control in genetic case-control association studies". *Nature Protocols* 5(9): pp. 1564–73.
- Ardlie, K. G., Deluca, D. S., Segrè, A. V., et al. (2015). "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans". *Science* 348: pp. 648–660.
- Baechler, E. C., Batliwalla, F. M., Karypis, G., et al. (2003). "Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus". *Proceedings of the National Academy of Sciences of the United States of America* 100 (5): pp. 2610–2615.
- Balemans, W., Ebeling, M., Patel, N., et al. (2001). "Increased bone density in sclerosteosis is due to the deficiency of a novel secreted protein (SOST)." *Human Molecular Genetics* 10: pp. 537–43.
- Banerjee, S. and Smith, I. (2010). "Management of small HER2-positive breast cancers". *The Lancet Oncology* 11(12): pp. 1193–9.
- Barreiro, L., Tailleux, L., and Pai, A. (2012). "Deciphering the genetic architecture of variation in the immune response to *Mycobacterium tuberculosis* infection". *Proceedings of the National Academy of Sciences of the United States of America* 109(4): pp. 1204–9.
- Battle, A., Khan, Z., Wang, S. H., et al. (2015). "Impact of regulatory variation from RNA to protein". *Science* 347(6222): pp. 664–7.
- Bayarsaihan, D. (2011). "Epigenetic mechanisms in inflammation". *Journal of Dental Research* 90: pp. 9–17.
- Beeson, P. (1946). "Development of tolerance to typhoid bacterial pyrogen and its abolition by reticulo-endothelial blockade." *Proceedings of the Society for Experimental Biology and Medicine* 61: pp. 248–50.
- Bensadoun, A. and Weinstein, D. (1976). "Assay of proteins in the presence of interfering materials". *Analytical biochemistry* 70(1): pp. 241–50.
- Bermejo-Martin, J. F., Ignacio, M., Rello, J., et al. (2010). "Host adaptive immunity deficiency in severe pandemic influenza". *Critical Care* 14: R167.
- Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., et al. (2010). "The NIH roadmap epigenomics mapping consortium". *Nature Biotechnology* 28(10): pp. 1045–1048.
- Bone, R., Balk, R., Cerra, F., et al. (1992). "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis." *Chest* 101(6): pp. 1644–55.

- Bonilla, F. and Oettgen, H. (2010). "Adaptive immunity". *Journal of Allergy and Clinical Immunology* 125(2 Suppl 2): S33–40.
- Bosmann, M. and Ward, P. A. (2013). "The inflammatory response in sepsis." *Trends in Immunology* 34: pp. 129–36.
- Braun, B., Slowik, A. and Leib, S. (2011). "The formyl peptide receptor like-1 and scavenger receptor MARCO are involved in glial cell activation in bacterial meningitis". *Journal of Neuroinflammation* 8: p. 11.
- Brunkow, M., Gardner, J., Van Ness, J., et al. (2001). "Bone dysplasia sclerosteosis results from loss of the SOST gene product, a novel cystine knot-containing protein." *American Journal of Human Genetics* 68: pp. 577–89.
- Burnham, K. L., Davenport, E. E., Radhakrishnan, J., et al. (2017). "Shared and Distinct Aspects of the Sepsis Transcriptomic Response to Fecal Peritonitis and Pneumonia." *American Journal of Respiratory and Critical Care Medicine* 196, No. 3: pp. 328–339.
- Çaliskan, M., Baker, S., Gilad, Y., et al. (2015). "Host genetic variation influences gene expression response to rhinovirus infection". *PLoS genetics* 11(4): e1005111.
- Calfee, C., Delucchi, K. and Parsons, P. (2014). "Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials". *The Lancet Respiratory Medicine* 2(8): pp. 611–20.
- Calvano, S. E., Xiao, W., Richards, D. R., et al. (2005). "A network-based analysis of systemic inflammation in humans." *Nature* 437: pp. 1032–7.
- Campbell, J. D., Spira, A., and Lenburg, M. E. (2011). "Applying gene expression microarrays to pulmonary disease". *Respirology* 16: pp. 407–418.
- Cao, Z. and Robinson, R. A. (2014). "The role of proteomics in understanding biological mechanisms of sepsis." *Proteomics. Clinical Applications* 8: pp. 35–52.
- Carcillo, J. A., Halstead, S. E., Hall, M. W., et al. (2017). "Three Hypothetical Inflammation Pathobiology Phenotypes and Pediatric Sepsis-Induced Multiple Organ Failure Outcome". *Pediatric Critical Care Medicine* 18: pp. 513–523.
- Carson, W. F., Cavassani, K. A., Dou, Y., et al. (2011). "Epigenetic regulation of immune cell functions during post-septic immunosuppression." *Epigenetics* 6: pp. 273–83.
- Cavaillon, J.-M. M. and Adib-Conquy, M. (2006). "Bench-to-bedside review: endotoxin tolerance as a model of leukocyte reprogramming in sepsis". *Critical Care* 10: p. 233.
- Cazalis, M. A., Lepape, A., Venet, F., et al. (2014). "Early and dynamic changes in gene expression in septic shock patients: a genome-wide approach." *Intensive Care Medicine Experimental* 2: p. 20.
- Chang, C., Chow, C., and Tellier, L. (2015). "Second-generation PLINK: rising to the challenge of larger and richer datasets". *Gigascience* 4: p. 7.
- Chapman, S. J. and Hill, A. V. (2012). "Human genetic susceptibility to infectious disease". *Nature Reviews Genetics* 13: pp. 175–188.
- Chen, G., Li, X., Huang, M., et al. (2016). "Thioredoxin-1 Increases Survival in Sepsis by Inflammatory Response Through Suppressing Endoplasmic Reticulum Stress." *Shock* 46: pp. 67–74.
- Chen, X., El Gazzar, M., Yoza, B. K., et al. (2009). "The NF-kappaB factor RelB and histone H3 lysine methyltransferase G9a directly interact to generate

- epigenetic silencing in endotoxin tolerance". *The Journal of Biological Chemistry* 284: pp. 27857–27865.
- Chen, Y., Liu, F., and Massague, J (1997). "Mechanism of TGF receptor inhibition by FKBP12". *The EMBO journal* 16(13): 3866–3876.
- Cheng, S. C., Quintin, J., Cramer, R. A., et al. (2014). "mTOR- and HIF-1-mediated aerobic glycolysis as metabolic basis for trained immunity." *Science* 345: p. 1250684.
- Chick, J. M., Munger, S. C., Simecek, P., et al. (2016). "Defining the consequences of genetic variation on a proteome-wide scale." *Nature* 534: pp. 500–5.
- Choi, J., Yu, U, Yoo, O., et al. (2005). "Differential coexpression analysis using microarray data and its application to human cancer". *Bioinformatics* 21(24): pp. 4348–55.
- Cobb, J., Moore, E. E., Hayden, D. L., et al. (2009). "Validation of the riboleukogram to detect ventilator-associated pneumonia after severe injury." *Annals of Surgery* 250: pp. 531–9.
- Cohen, J, Cristofaro, P, Carlet, J, et al. (2004). "New method of classifying infections in critically ill patients". *Critical Care Medicine* 32(7): pp. 1510–26.
- Cohen, J., Vincent, J., Adhikari, N. K., et al. (2015). "Sepsis: a roadmap for future research". *The Lancet Infectious Diseases* 15: pp. 581–614.
- Collins, F. S., Morgan, M., and Patrinos, A. (2003). "The Human Genome Project: lessons from large-scale biology." *Science* 300: pp. 286–90.
- Comstock, G. (1978). "Tuberculosis in twins: a re-analysis of the Proffit survey." *The American Review of Respiratory Disease* 117: pp. 621–4.
- Cookson, W., Liang, L., Abecasis, G., et al. (2009). "Mapping complex disease traits with global gene expression." *Nature Reviews Genetics* 10: pp. 184–94.
- Cosman, F., Crittenden, D. B., Adachi, J. D., et al. (2016). "Romosozumab Treatment in Postmenopausal Women with Osteoporosis." *New England Journal of Medicine* 375: pp. 1532–1543.
- Cvijanovich, N., Shanley, T. P., Lin, R., et al. (2008). "Validating the genomic signature of pediatric septic shock." *Physiological Genomics* 34: pp. 127–34.
- Dabney, A, Storey, J., and Warnes, G. (2010). *qvalue: Q-value estimation for false discovery rate control. R package version 2.8.0*, URL: <http://github.com/jdstorey/qvalue..>
- Daniels, R. (2011). "Surviving the first hours in sepsis: getting the basics right (an intensivist's perspective)". *The Journal of Antimicrobial Chemotherapy* 66 Suppl 2: p. 23.
- Davenport, E. E. (2014). "Functional genomics of variation in response to infection: insights into severe sepsis and common variable immune deficiency disorders". Thesis. University of Oxford.
- Davenport, E. E., Burnham, K. L., Radhakrishnan, J., et al. (2016). "Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study." *The Lancet Respiratory Medicine* 4: pp. 259–71.
- Davenport, E. E., Amariuta, T., Gutierrez-Arcelus, M., et al. (2017). "Identification of drug eQTL interactions from repeat transcriptional and environmental measurements in a lupus clinical trial". *bioRxiv*.
- Davis, J. R., Fresard, L., Knowles, D. A., et al. (2016). "An Efficient Multiple-Testing Adjustment for eQTL Studies that Accounts for Linkage

- Disequilibrium between Variants.” *American Journal of Human Genetics* 98: pp. 216–24.
- Davis, S. and Meltzer, P. S. (2007). “GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor”. *Bioinformatics* 23: pp. 1846–7.
- Deans, K. J., Haley, M., Natanson, C., et al. (2005). “Novel therapies for sepsis: a review”. *The Journal of Trauma* 58: pp. 867–874.
- DeCoux, A., Tian, Y., Y, D. K., et al. (2015). “Plasma Glycoproteomics Reveals Sepsis Outcomes Linked to Distinct Proteins in Common Pathways.” *Critical Care Medicine* 43: pp. 2049–2058.
- Deng, J. C., Cheng, G., Newstead, M. W., et al. (2006). “Sepsis-induced suppression of lung innate immunity is mediated by IRAK-M.” *The Journal of Clinical Investigation* 116: pp. 2532–42.
- Di Narzo, A., Telesco, S., and Brodmerkel, C (2017). “High-Throughput Characterization of Blood Serum Proteomics of IBD Patients with Respect to Aging and Genetic Factors”. *PLoS Genetics* 13(1): e1006565.
- Dolgachev, V., Goldberg, R, Suresh, M., et al. (2016). “Electroporation-mediated delivery of the FER gene in the resolution of trauma-related fatal pneumonia.” *Gene Therapy* 23: pp. 785–796.
- Dolinay, T, Kim, Y., and Howrylak, J (2012). “Inflammasome-regulated cytokines are critical mediators of acute lung injury”. *American Journal of Respiratory and Critical Care Medicine* 185(11): pp. 1225–34.
- Durbin, R (2014). “Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT)”. *Bioinformatics* 30(9): pp. 1266–72.
- Ebert, B. L. and Golub, T. R. (2004). “Genomic approaches to hematologic malignancies”. *Blood* 104: pp. 923–932.
- El Gazzar, M., Yoza, B. K., Chen, X., et al. (2008). “G9a and HP1 couple histone and DNA methylation to TNF α transcription silencing during endotoxin tolerance”. *The Journal of Biological Chemistry* 283: pp. 32198–32208.
- ENCODE Project Consortium (2012). “An integrated encyclopedia of DNA elements in the human genome”. *Nature* 489: p. 57.
- Fabrick, B., Bruggen, R van, and Deng, D. (2009). “The macrophage scavenger receptor CD163 functions as an innate immune sensor for bacteria”. *Blood* 113(4): pp. 887–92.
- Fairfax, B. P. and Knight, J. C. (2014). “Genetics of gene expression in immunity to infection”. *Current Opinion in Immunology* 30: pp. 63–71.
- Fairfax, B. P., Davenport, E. E., Makino, S., et al. (2011). “A common haplotype of the TNF receptor 2 gene modulates endotoxin tolerance”. *Journal of Immunology* 186: pp. 3058–3065.
- Fairfax, B. P., Makino, S., Radhakrishnan, J., et al. (2012). “Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles.” *Nature Genetics* 44: pp. 502–10.
- Fairfax, B. P., Humburg, P., Makino, S., et al. (2014). “Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression.” *Science* 343: p. 1246949.
- Faix, J. (2013). “Biomarkers of sepsis”. *Critical Reviews in Clinical Laboratory Sciences* 50(1): pp. 23–36.

- Famous, K., Delucchi, K., and Ware, L. (2017). "Acute respiratory distress syndrome subphenotypes respond differently to randomized fluid management strategy". *American Journal of Respiratory and Critical Care Medicine* 195(3): pp. 331–338.
- Fang, H., Knezevic, B., Burnham, K. L., et al. (2016). "XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits". *Genome Medicine* 8: p. 129.
- Feezor, R. J., Baker, H. V., Xiao, W., et al. (2004). "Genomic and proteomic determinants of outcome in patients undergoing thoracoabdominal aortic aneurysm repair". *Journal of Immunology* 172: pp. 7103–9.
- Fellay, J., Shianna, K. V., Ge, D., et al. (2007). "A whole-genome association study of major determinants for host control of HIV-1." *Science* 317: pp. 944–7.
- Fischer, R., Trudgian, D., Wright, C., et al. (2012). "Discovery of candidate serum proteomic and metabolomic biomarkers in ankylosing spondylitis". *Molecular and Cellular Proteomics* 1(2): p. M111.013904.
- Foss, E. J., Radulovic, D., Shaffer, S. A., et al. (2007). "Genetic basis of proteome variation in yeast". *Nature Genetics* 39: pp. 1369–1375.
- Foss, E. J., Radulovic, D., Shaffer, S. A., et al. (2011). "Genetic variation shapes protein networks mainly through non-transcriptional mechanisms." *PLoS Biology* 9: e1001144.
- Foster, S. L., Hargreaves, D. C., and Medzhitov, R. (2007). "Gene-specific control of inflammation by TLR-induced chromatin modifications". *Nature* 447: pp. 972–978.
- Franco, L., Bucasas, K., Wells, J., et al. (2013). "Integrative genomic analysis of the human immune response to influenza vaccination". *ELife* 2: e00299.
- Franke, A., PB, M. D., Barrett, J. C., et al. (2010). "Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci". *Nature Genetics* 42: pp. 1118–1125.
- Fresno, C., Garca-Rio, F., Gmez-Pia, V., et al. (2009). "Potent phagocytic activity with impaired antigen presentation identifying lipopolysaccharide-tolerant human monocytes: demonstration in isolated monocytes from cystic fibrosis patients". *The Journal of Immunology* 182: pp. 6494–6507.
- Geyer, P., Kulak, N., Pichler, G., et al. (2016). "Plasma proteome profiling to assess human health and disease". *Cell systems* 2(3): pp. 185–95.
- Ghazalpour, A., Bennett, B., Petyuk, V. A., et al. (2011). "Comparative analysis of proteome and transcriptome variation in mouse." *PLoS Genetics* 7: e1001393.
- Giambartolomei, C., Vukcevic, D., Schadt, E. E., et al. (2014). "Bayesian test for colocalisation between pairs of genetic association studies using summary statistics." *PLoS Genetics* 10: e1004383.
- Glish, G. L. and Vachet, R. W. (2003). "The basics of mass spectrometry in the twenty-first century". *Nature Reviews Drug Discovery* 2: pp. 140–150.
- Gogos, C., Kotsaki, A., Pelekanou, A., et al. (2010). "Early alterations of the innate and adaptive immune statuses in sepsis according to the type of underlying infection". *Critical Care* 14(3): R96.
- Goh, C and Knight, J. (2017). "Enhanced understanding of the hostpathogen interaction in sepsis: new opportunities for omic approaches". *The Lancet Respiratory Medicine* 5(3): pp. 212–223.

- Gordon, A. C., Mason, A. J., Thirunavukkarasu, N., et al. (2016). "Effect of Early Vasopressin vs Norepinephrine on Kidney Failure in Patients With Septic Shock: The VANISH Randomized Clinical Trial." *Journal of the American Medical Association* 316: pp. 509–18.
- Grassl, N, Kulak, N., and Pichler, G (2016). "Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome". *Genome Medicine* 8: p. 44.
- Gupta, S., Sakhuja, A., Kumar, G., et al. (2016). "Culture-Negative Severe Sepsis: Nationwide Trends and Outcomes." *Chest* 150: pp. 1251–1259.
- Hause, R. J., Stark, A. L., Antao, N. N., et al. (2014). "Identification and Validation of Genetic Variants that Influence Transcription Factor and Cell Signaling Protein Levels." *American Journal of Human Genetics* 95: pp. 194–208.
- Hofer, S, Rosenhagen, C, and Nakamura, H (2009). "Thioredoxin in human and experimental sepsis". *Critical Care Medicine* 37(7): pp. 2155–9.
- Horvatovich, P., Franke, L., and Bischoff, R. (2014). "Proteomic studies related to genetic determinants of variability in protein concentrations." *Journal of Proteome Research* 13: pp. 5–14.
- Horwitz, P., Tsai, E., Putt, M., et al. (2004). "Detection of cardiac allograft rejection and response to immunosuppressive therapy with peripheral blood gene expression". *Circulation* 110(25): pp. 3815–21.
- Hotchkiss, R. S. and Karl, I. E. (2003). "The pathophysiology and treatment of sepsis." *New England Journal of Medicine* 348: pp. 138–50.
- Hotchkiss, R. S., Monneret, G., and Payen, D. (2013). "Sepsis-induced immunosuppression: from cellular dysfunctions to immunotherapy." *Nature Reviews Immunology* 13: pp. 862–74.
- Howrylak, J. A., Dolinay, T., Lucht, L., et al. (2009). "Discovery of the gene signature for acute lung injury in patients with sepsis." *Physiological Genomics* 37: pp. 133–9.
- Hu, J., Yan, J., Rao, G., et al. (2016). "The Duality of Fgl2 - Secreted Immune Checkpoint Regulator Versus Membrane-Associated Procoagulant: Therapeutic Potential and Implications." *International Reviews of Immunology* 35: pp. 325–339.
- Huber, W., Heydebreck, A. von, Sltmann, H., et al. (2002). "Variance stabilization applied to microarray data calibration and to the quantification of differential expression". *Bioinformatics* 18: S96–S104.
- Humburg, P., Maugeri, N., Lee, W., et al. (2016). "Characterisation of the global transcriptional response to heat shock and the impact of individual genetic variation." *Genome Medicine* 8: p. 87.
- Hutchins, N. A., Unsinger, J., Hotchkiss, R. S., et al. (2014). "The new normal: immunomodulatory agents against sepsis immune suppression". *Trends in Molecular Medicine* 20: pp. 224–233.
- Ioannidis, I, B, M., Willette, M, et al. (2012). "Plasticity and virus specificity of the airway epithelial cell immune response during respiratory virus infection". *Journal of Virology* 86: pp. 5422–36.
- Itoh, S, Lemay, S, Osawa, M, et al. (2005). "Mitochondrial Dok-4 recruits Src kinase and regulates NF-kB activation in endothelial cells". *Journal of Biological Chemistry* 280(28): pp. 26383–96.

- Jacobs, L. and Wong, H. R. (2017). "Sepsis Subclasses: Be Careful of What You Wish for." *Pediatric Critical Care Medicine* 18: pp. 591–592.
- Jallow, M., Teo, Y., Small, K. S., et al. (2009). "Genome-wide and fine-resolution association analysis of malaria in West Africa". *Nature Genetics* 41: pp. 657–665.
- Jans, J., Hicham, e., Groot, R. de, et al. (2016). "Actin- and clathrin-dependent mechanisms regulate interferon gamma release after stimulation of human immune cells with respiratory syncytial virus." *Virology J* 13: p. 52.
- Joehanes, R, Zhang, X, and Huan, T (2017). "Integrated genome-wide analysis of expression quantitative trait loci aids interpretation of genomic association studies". *Genome Biology* 18(1): p. 16.
- Johansson, N, Kalin, M, and Hedlund, J (2011). "Clinical impact of combined viral and bacterial infection in patients with community-acquired pneumonia". *Scandinavian journal of Infectious Diseases* 43(8): pp. 609–15.
- Johnson, S. B., Lissauer, M., Bochicchio, G. V., et al. (2007a). "Gene expression profiles differentiate between sterile SIRS and early sepsis." *Annals of Surgery* 245: pp. 611–21.
- Johnson, W., Li, C., and Rabinovic, A. (2007b). "Adjusting batch effects in microarray expression data using empirical Bayes methods." *Biostatistics* 8: pp. 118–27.
- Jostins, L, Ripke, S, Weersma, R., et al. (2012). "Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease". *Nature* 491(7422): pp. 119–24.
- Julian, M. W., Strange, H. R., Ballinger, M. N., et al. (2015). "Tolerance and cross-tolerance following Toll-like receptor (TLR)-4 and-9 activation are mediated by IRAK-M and modulated by IL-7 in murine splenocytes". *PLoS One* 10: e0132921.
- Kalenka, A., Feldmann, R. E., Otero, K., et al. (2006). "Changes in the serum proteome of patients with sepsis and septic shock." *Anesthesia and Analgesia* 103: pp. 1522–6.
- Kangelaris, K. N., Prakash, A., Liu, K. D., et al. (2015). "Increased expression of neutrophil-related genes in patients with early sepsis-induced ARDS." *American Journal of Physiology. Lung Cellular and Molecular Physiology* 308: pp. L1102–13.
- Kaplan, J. M. and Wong, H. R. (2011). "Biomarker discovery and development in pediatric critical care medicine". *Pediatric Critical Care Medicine* 12: p. 165.
- Khor, C., Chau, T., Pang, J., et al. (2011). "Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1". *Nature Genetics* 43: pp. 1139–1141.
- Kiehntopf, M., Schmerler, D., Brunkhorst, F. M., et al. (2011). "Mass spectrometry-based protein patterns in the diagnosis of sepsis/systemic inflammatory response syndrome." *Shock* 36: pp. 560–9.
- Kiiveri, H. T. (2008). "A general approach to simultaneous model fitting and variable elimination in response models for biological data with many more variables than observations". *BMC Bioinformatics* 9: p. 195.

- Kim, J., Lee, J. N., Ye, J., et al. (2013). "Sufficient production of geranylgeraniol is required to maintain endotoxin tolerance in macrophages." *Journal of Lipid Research* 54: pp. 3430–7.
- Kim, M. S., Pinto, S. M., Getnet, D., et al. (2014). "A draft map of the human proteome." *Nature* 509: pp. 575–81.
- Kirn, T. and Weinstein, M. (2013). "Update on blood cultures: how to obtain, process, report, and interpret." *Clinical Microbiology and Infection* 19: pp. 513–20.
- Kitamura, T., Kabuyama, Y., Kamataki, A., et al. (2008). "Enhancement of lymphocyte migration and cytokine production by ephrinB1 system in rheumatoid arthritis." *American Journal of Physiology, Cell Physiology* 294: pp. C189–96.
- Knaus, W., Draper, E., Wagner, D., et al. (1985). "APACHE II: a severity of disease classification system." *Critical Care Medicine* 13: pp. 818–29.
- Knight, J. C. (2013). "Genomic modulators of the immune response". *Trends in Genetics* 29: pp. 74–83.
- Knox, D., Lanspa, M., Kuttler, K., et al. (2015). "Phenotypic clusters within sepsis-associated multiple organ dysfunction syndrome". *Intensive Care Medicine* 41(5): pp. 814–22.
- Koussounadis, A., Langdon, S., and Um, I. (2015). "Relationship between differentially expressed mRNA and mRNA-protein correlations in a xenograft model system". *Scientific Reports* 5: p. 10775.
- Kwan, A., Hubank, M., Rashid, A., et al. (2013). "Transcriptional instability during evolving sepsis may limit biomarker based risk stratification." *PLoS One* 8: e60501.
- Langley, R. J., Tsalik, E. L., Velkinburgh, J. C. van, et al. (2013). "An integrated clinico-metabolomic model improves prediction of death in sepsis." *Science Translational Medicine* 5: 195ra95.
- Lê, S., Josse, J., and Husson, F. (2008). "FactoMineR: An R package for multivariate analysis". *Journal of Statistical Software* 25: pp. 1–18.
- Lee, H (2013). "Procalcitonin as a biomarker of infectious diseases". *The Korean Journal of Internal Medicine* 28(3): pp. 285–91.
- Lee, M. N., Ye, C., Villani, A., et al. (2014). "Common Genetic Variants Modulate Pathogen-Sensing Responses in Human Dendritic Cells". *Science* 343(6175): p. 1246980.
- Lengauer, T. S., Oliver, S., Niko, B., et al. (2005). "ROCR: visualizing classifier performance in R". *Bioinformatics* 21(20): pp. 3940–1.
- Liangos, O., Domhan, S., Schwager, C., et al. (2010). "Whole blood transcriptomics in cardiac surgery identifies a gene regulatory network connecting ischemia reperfusion with systemic inflammation." *PLoS ONE* 5: e13658.
- Lissauer, M. E., Johnson, S. B., Siuzdak, G., et al. (2007). "Coagulation and complement protein differences between septic and uninfected systemic inflammatory response syndrome patients." *Journal of Trauma* 62: pp. 1082–9.
- Liu, X., Ren, H., and Peng, D. (2014). "Sepsis biomarkers: an omics perspective." *Frontiers in Medicine* 8: pp. 58–67.

- Lizio, M., Harshbarger, J., Shimoji, H., et al. (2015). "Gateways to the FANTOM5 promoter level mammalian expression atlas". *Genome Biology* 16:22.
- Loh, P., Danecek, P., Palamara, P., et al. (2016). "Reference-based phasing using the Haplotype Reference Consortium panel." *Nature Genetics* 48(11): pp. 1443–8.
- MacArthur, J., Bowler, E., Cerezo, M., et al. (2017). "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)." *Nucleic Acids Research* 45: pp. D896–D901.
- Mages, J., Dietrich, H., and Lang, R. (2007). "A genome-wide analysis of LPS tolerance in macrophages." *Immunobiology* 212: pp. 723–37.
- Man, M, Close, S., Shaw, A., et al. (2013). "Beyond single-marker analyses: mining whole genome scans for insights into treatment responses in severe sepsis." *Pharmacogenomics J.* 13: pp. 218–26.
- Marshall, J. C. (2014). "Why have clinical trials in sepsis failed?" *Trends in Molecular Medicine* 20: pp. 195–203.
- Martí-Carvajal, A. J., Solà, I., Lathyris, D., et al. (2012). "Human recombinant activated protein C for severe sepsis." *Cochrane Database Systematic Review*: p. CD004388.
- Maslove, D. M., Tang, B. M., and S, M. A. (2012). "Identification of sepsis subtypes in critically ill adults using gene expression profiling." *Critical Care* 16: R183.
- Mathias, B., Delmas, A. L., Tezcan, O., et al. (2017). "Human Myeloid-derived Suppressor Cells are Associated With Chronic Immune Suppression After Severe Sepsis/Septic Shock." *Annals of Surgery* 265: pp. 827–834.
- Maurano, M. T., Humbert, R., Rynes, E., et al. (2012). "Systematic localization of common disease-associated variation in regulatory DNA". *Science* 337: pp. 1190–1195.
- Mayr, F., Yende, S, and Angus, D. (2014). "Epidemiology of severe sepsis". *Virulence* 5(1): 411.
- McCarthy, S, Das, S, Kretzschmar, W, et al. (2016). "A reference panel of 64,976 haplotypes for genotype imputation". *Nature Genetics* 48(10): pp. 1279–83.
- McLafferty, F. (1981). "Tandem mass spectrometry." *Science* 214: pp. 280–7.
- McLaren, W., Gil, L., Hunt, S. E., et al. (2016). "The Ensembl Variant Effect Predictor." *Genome Biology* 17: p. 122.
- Melzer, D, Perry, J., Hernandez, D, et al. (2008). "A genome-wide association study identifies protein quantitative trait loci (pQTLs)". *PLoS Genetics* 4(5): e1000072.
- Misch, E., Berrington, W., and Vary, J. (2010). "Leprosy and the Human Genome". *Microbiology and molecular biology* 74(4): pp. 589–620.
- Monneret, G., Lepape, A., Voirin, N., et al. (2006). "Persisting low monocyte human leukocyte antigen-DR expression predicts mortality in septic shock." *Intensive Care Medicine* 32: pp. 1175–83.
- Musher, D. and Thorner, A. (2014). "Community-acquired pneumonia". *New England Journal of Medicine* 371: pp. 1619–1628.
- Musunuru, K., Strong, A., Maria, F., et al. (2010). "From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus." *Nature* 466: pp. 714–9.

- Nakamura, A, Wada, H, Ikejiri, M, et al. (2009). "Efficacy of procalcitonin in the early diagnosis of bacterial infections in a critical care unit". *Shock* 31(6): pp. 586–91.
- Nanjappa, V, Thomas, J., and Marimuthu, A (2013). "Plasma Proteome Database as a resource for proteomics research: 2014 update". *Nucleic Acids Research* 42: pp. D959–65.
- National Institute for Health and Care Excellence (2014). *Pneumonia in adults: diagnosis and management*. URL: <https://www.nice.org.uk/guidance/cg191>.
- National Institute for Health and Care Excellence (2017). *Surgical site infections: preprevent and treatment*. URL: <https://www.nice.org.uk/guidance/cg74/chapter/1-guidance>.
- Neale, B. M., Fagerness, J., Reynolds, R., et al. (2010). "Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (LIPC)". *Proceedings of the National Academy of Sciences of the United States of America* 107(17): pp. 7395–400.
- Ng, P. C., Ang, I. L., Chiu, R. W., et al. (2010). "Host-response biomarkers for diagnosis of late-onset septicemia and necrotizing enterocolitis in preterm infants." *The Journal of Clinical Investigation* 120: pp. 2989–3000.
- Nica, A. C., Montgomery, S. B., Dimas, A. S., et al. (2010). "Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations." *PLoS Genetics* 6: e1000895.
- O'Carroll, C., Fagan, A., Shanahan, F., et al. (2014). "Identification of a unique hybrid macrophage-polarization state following recovery from lipopolysaccharide tolerance". *Journal of Immunology* 192: pp. 427–436.
- Ogata, H, Su, I, Miyake, K, et al. (2000). "The toll-like receptor protein RP105 regulates lipopolysaccharide signaling in B cells." *Journal of Experimental Medicine* 192: pp. 23–9.
- Oon, S, Wilson, N., and Wicks, I (2016). "Targeted therapeutics in SLE: emerging strategies to modulate the interferon pathway". *Clinical and Translational Immunology* 5 (5): e79.
- Opal, S. M., Laterre, P. F., Francois, B., et al. (2013). "Effect of eritoran, an antagonist of MD2-TLR4, on mortality in patients with severe sepsis: the ACCESS randomized trial." *The Journal of the American Medical Association* 309: pp. 1154–62.
- Pachot, A., Lepape, A., Vey, S., et al. (2006). "Systemic transcriptional analysis in survivor and non-survivor septic shock patients: a preliminary study." *Immunology Letters* 106: pp. 63–71.
- Pacis, A., Ndlec, Y., and Barreiro, L. B. (2014). "When genetics meets epigenetics: deciphering the mechanisms controlling inter-individual variation in immune responses to infection". *Current Opinion in Immunology* 29C: pp. 119–126.
- Pankla, R., Buddhisa, S., Berry, M., et al. (2009). "Genomic transcriptional profiling identifies a candidate blood biomarker signature for the diagnosis of septicemic melioidosis." *Genome Biology* 10: R127.
- Parnell, G. P., S, M. A., Booth, D. R., et al. (2012). "A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia." *Critical Care* 16: R157.

- Parnell, G. P., Tang, B. M., Nalos, M., et al. (2013). "Identifying key regulatory genes in the whole blood of septic patients to monitor underlying immune dysfunctions." *Shock* 40: pp. 166–74.
- Paul, V, Tridente, A, Kaur, P, et al. (2015). "Critically ill patients with faecal peritonitis: a 5-year review in a tertiary centre". *Critical Care* 19: P374.
- Pena, O. M., Pistolic, J., Raj, D., et al. (2011). "Endotoxin tolerance represents a distinctive state of alternative polarization (M2) in human mononuclear cells". *Journal of Immunology* 186: pp. 7243–7254.
- Pena, O. M., Hancock, D. G., Lyle, N. H., et al. (2014). "An Endotoxin Tolerance Signature Predicts Sepsis and Organ Dysfunction at Initial Clinical Presentation". *EBioMedicine* 1(1): 647.
- Pennock, N., White, J., and Cross, E. (2013). "T cell responses: naive to memory and everything in between". *Advances in Physiology Education* 37(4): pp. 273–83.
- Perner, A., Gordon, A. C., Angus, D. C., et al. (2017). "The intensive care medicine research agenda on septic shock." *Intensive Care Medicine* 43: pp. 1294–1305.
- Petersen, L, Andersen, P, and Sørensen, T. (2010). "Genetic influences on incidence and case-fatality of infectious disease". *PLoS One* 5(5): e10603.
- Pierrakos, C and Vincent, J. (2010). "Sepsis biomarkers: a review". *Critical Care* 14(1): R15.
- Pils, D, Tong, D, Hager, G, et al. (2013). "A combined blood based gene expression and plasma protein abundance signature for diagnosis of epithelial ovarian cancer-a study of the OVCAD Consortium". *BMC Cancer* 13: p. 178.
- Prabhakar, U., Conway, T. M., Murdock, P., et al. (2005). "Correlation of protein and gene expression profiles of inflammatory proteins after endotoxin challenge in human subjects." *DNA Cell Biology* 24: pp. 410–31.
- Prescott, H. C., Calfee, C. S., Thompson, B., et al. (2016). "Toward Smarter Lumping and Smarter Splitting: Rethinking Strategies for Sepsis and Acute Respiratory Distress Syndrome Clinical Trial Design." *American Journal of Respiratory and Critical Care Medicine* 194: pp. 147–55.
- Prucha, M., Ruryk, A., Boriss, H., et al. (2004). "Expression profiling: toward an application in sepsis diagnostics." *Shock* 22: pp. 29–33.
- Purcell, S. and Chang, C. (2015). *PLINK [1.9]*. URL: www.cog-genomics.org/plink/1.9/.
- Qiu W, L. M. T. and Whitmore, G. A. (2006). "sizepower: Sample Size and Power Calculation in Micorarray Studies". *Journal of Biopharmaceutical Statistics* 22(1): pp. 30–42.
- Quartin, A. A., Schein, R. M., Kett, D. H., et al. (1997). "Magnitude and duration of the effect of sepsis on survival. Department of Veterans Affairs Systemic Sepsis Cooperative Studies Group". *The Journal of the American Medical Association* 277: pp. 1058–1063.
- Radhakrishnan, J. (2012). "Functional Genomics of Severe Sepsis and Septic Shock". Thesis. University of Oxford.
- Raju, M. S., V, J., Kamaraju, R. S., et al. (2016). "Continuous evaluation of changes in the serum proteome from early to late stages of sepsis caused by *Klebsiella pneumoniae*." *Molecular Medicine Reports* 13: pp. 4835–44.

- Rakyan, V. K., Down, T. A., Balding, D. J., et al. (2011). "Epigenome-wide association studies for common human diseases". *Nature Reviews. Genetics* 12: pp. 529–541.
- Rausch, M. and Hastings, K. (2015). "Diverse cellular and organismal functions of the lysosomal thiol reductase GILT". *Molecular Immunology* 68(2 Pt A): pp. 124–8.
- Rautanen, A., Mills, T. C., Gordon, A. C., et al. (2015). "Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study". *Lancet Respiratory Medicine* 3: pp. 53–60.
- Reinhart, K and Karzai, W (2001). "Anti-tumor necrosis factor therapy in sepsis: update on clinical trials and lessons learned." *Critical Care Medicine* 29: S121–5.
- Reinhart, K, Menges, T, Gardlund, B, et al. (2001). "Randomized, placebo-controlled trial of the anti-tumor necrosis factor antibody fragment afelimomab in hyperinflammatory response during severe sepsis: The RAMSES Study." *Critical Care Medicine* 29: pp. 765–9.
- Ripke, S., Neale, B. M., Corvin, A., et al. (2014). "Biological insights from 108 schizophrenia-associated genetic loci". *Nature* 511: p. 421.
- Rockman, M. V. and Kruglyak, L. (2006). "Genetics of global gene expression." *Nature Reviews Genetics* 7: pp. 862–72.
- Ronan, T., Qi, Z., and Naegle, K. M. (2016). "Avoiding common pitfalls when clustering biological data." *Science Signalling* 9: re6.
- Saeed, S., Quintin, J., Kerstens, H. H., et al. (2014). "Epigenetic programming of monocyte-to-macrophage differentiation and trained innate immunity". *Science* 345: p. 1251086.
- Salimnia, H., Fairfax, M. R., Lephart, P. R., et al. (2016). "Evaluation of the FilmArray Blood Culture Identification Panel: Results of a Multicenter Controlled Trial." *Journal of Clinical Microbiology* 54: pp. 687–98.
- Sandquist, M. and Wong, H. R. (2014). "Biomarkers of sepsis and their potential value in diagnosis, prognosis and treatment." *Expert Review of Clinical Immunology* 10: pp. 1349–56.
- Scherag, A., Schönebeck, F., Kesselmeier, M., et al. (2016). "Genetic Factors of the Disease Course after Sepsis: A Genome-Wide Study for 28Day Mortality." *EBioMedicine* 12: pp. 239–246.
- Schork, N. J., Murray, S. S., Frazer, K. A., et al. (2009). "Common vs. rare allele hypotheses for complex diseases." *Current Opinion in Genetics & Development* 19: pp. 212–9.
- Scicluna, B., Vught, L. van, Zwinderman, A. H., et al. (2017). "Classification of sepsis patients as blood genomic endotypes: a prospective cohort study." *The Lancet Respiratory Medicine* S2213-2600(17): pp. 30294–1.
- Scicluna, B. P., Klein Klouwenberg, P. M., Vught, L. A. van, et al. (2015). "A molecular biomarker to diagnose community-acquired pneumonia on intensive care unit admission." *American Journal of Respiratory Critical Care Medicine* 192: pp. 826–35.
- Seeley, J. J. and Ghosh, S. (2017). "Molecular mechanisms of innate memory and tolerance to LPS." *Journal of Leukocyte Biology* 101: pp. 107–119.

- Seok, J, Warren, H., and Cuenca, A. (2013). "Genomic responses in mouse models poorly mimic human inflammatory diseases". *Proceedings of the National Academy of Sciences of the United States of America* 110(9): pp. 3507–12.
- Severino, P., Silva, E., L, B. G., et al. (2014). "Patterns of gene expression in peripheral blood mononuclear cells and outcomes from patients with sepsis secondary to community acquired pneumonia." *PLoS ONE* 9: e91886.
- Shabalin, A. (2012). "Matrix eQTL: ultra fast eQTL analysis via large matrix operations". *Bioinformatics* 28(10): pp. 1353–8.
- Shen, Z., Want, E. J., Chen, W., et al. (2006). "Sepsis plasma protein profiling with immunodepletion, three-dimensional liquid chromatography tandem mass spectrometry, and spectrum counting." *Journal of Proteome Research* 5: pp. 3154–60.
- Shi, L., Kalupahana, N., and Turnis, M. (2013). "Inhibitory role of the transcription repressor Gfi1 in the generation of thymus-derived regulatory T cells". *Proceedings of the National Academy of Sciences of the United States of America* 110 (34): E3198E3205.
- Singer, M., Deutschman, C. S., Seymour, C. W., et al. (2016). "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)." *Journal of the American Medical Association* 315: pp. 801–10.
- Smirnov, D. A., Zweitzig, D. R., Foulk, B. W., et al. (2005). "Global gene expression profiling of circulating tumor cells". *Cancer research* 65 (12): pp. 4993–7.
- Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Statistical Applications in Genetics and Molecular Biology* 3: Article3.
- Soares, A. J., Santos, M., Trugilho, M., et al. (2009). "Differential proteomics of the plasma of individuals with sepsis caused by *Acinetobacter baumannii*." *Journal of Proteomics* 73: pp. 267–78.
- Solan, P. D., Dunsmore, K. E., Denenberg, A. G., et al. (2012). "A novel role for matrix metalloproteinase-8 in sepsis." *Critical Care Medicine* 40: pp. 379–87.
- Sørensen, T. I., Nielsen, G. G., Andersen, P. K., et al. (1988). "Genetic and environmental influences on premature death in adult adoptees". *The New England Journal of Medicine* 318: pp. 727–732.
- Sowemimo-Coker, S. (2002). "Red blood cell hemolysis during processing". *Transfusion Medicine Reviews* 16(1): pp. 46–60.
- Srinivasan, L., Page, G., Kirpalani, H., et al. (2017). "Genome-wide association study of sepsis in extremely premature infants." *Archives of Disease in Childhood. Fetal and Neonatal Edition* 102: F439–F445.
- Steinmetz, L. M. and Davis, R. W. (2004). "Maximizing the potential of functional genomics". *Nature Reviews. Genetics* 5: pp. 190–201.
- Styrkarsdottir, U., Halldorsson, B. V., Gretarsdottir, S., et al. (2009). "New sequence variants associated with bone mineral density." *Nature Genetics* 41: pp. 15–7.
- Suhre, K, Arnold, M, Bhagwat, A., et al. (2017). "Connecting genetic risk to disease end points through the human blood plasma proteome". *Nature Communications* 8: p. 14357.

- Sun, B., Maranville, J., Peters, J., et al. (2017). "Consequences Of Natural Perturbations In The Human Plasma Proteome". *bioRxiv*.
- Sutherland, A., Thomas, M., Brandon, R. A., et al. (2011). "Development and validation of a novel molecular biomarker diagnostic test for the early detection of sepsis." *Critical Care* 15: R149.
- Svoren, E. M. (2015). "Genetic regulation of the host response to cardiac surgery and cardiopulmonary bypass". PhD thesis. Queen Mary University.
- Sweeney, T. and Wong, H. (2016). "Risk Stratification and Prognosis in Sepsis: What Have We Learned from Microarrays?" *Clinics in Chest Medicine* 37(2): pp. 209–18.
- Sweeney, T., Perumal, T., Henao, R, et al. (2016). "Mortality prediction in sepsis via gene expression analysis: a community approach". *bioRxiv*.
- Sweeney, T. E. and Khatri, P. (2017). "Benchmarking Sepsis Gene Expression Diagnostics Using Public Data." *Critical Care Medicine* 45: pp. 1–10.
- Sweeney, T. E., Shidham, A., Wong, H. R., et al. (2015). "A comprehensive time-course-based multicohort analysis of sepsis and sterile inflammation reveals a robust diagnostic gene set." *Science Translational Medicine* 7: 287ra71.
- Takaoka, A, Wang, Z., Choi, M., et al. (2007). "DAI (DLM-1/ZBP1) is a cytosolic DNA sensor and an activator of innate immune response". *Nature* 448(7152): pp. 501–5.
- Talwar, S., Munson, P. J., Barb, J., et al. (2006). "Gene expression profiles of peripheral blood leukocytes after endotoxin challenge in humans." *Physiological Genomics* 25: pp. 203–15.
- Tang, B. M., S, M. A., Dawes, I. W., et al. (2008). "Gene-expression profiling of gram-positive and gram-negative sepsis in critically ill patients." *Critical Care Medicine* 36: pp. 1125–8.
- Tang, B. M., S, M. A., Dawes, I. W., et al. (2009). "Gene-expression profiling of peripheral blood mononuclear cells in sepsis." *Critical Care Medicine* 37: pp. 882–8.
- Tang, B. M., Huang, S. J., and McLean, A. S. (2010). "Genome-wide transcription profiling of human sepsis: a systematic review". *Critical Care* 14: R237.
- Tang, W, Schvienbacher, C, and Lopez, L. (2012). "Genetic associations for activated partial thromboplastin time and prothrombin time, their gene expression profiles, and risk of coronary artery disease". *The American Journal of Human Genetics* 91(1): pp. 152–62.
- Taylor, J. C., Martin, H. C., Lise, S., et al. (2015). "Factors influencing success of clinical genome sequencing across a broad spectrum of disorders." *Nature Genetics* 47: pp. 717–726.
- The 1000 Genomes Project Consortium (2012). "An integrated map of genetic variation from 1,092 human genomes". *Nature* 491(7422): pp. 56–65.
- The UniProt Consortium (2017). "UniProt: the universal protein knowledgebase". *Nucleic Acids Research* 45: D158D169.
- Thorpe, A. (2011). "Genetic variation and differential gene expression in patients with severe sepsis". MA thesis. University of Oxford.
- Thye, T., Vannberg, F. O., Wong, S. H., et al. (2010). "Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2". *Nature Genetics* 42: pp. 739–741.

- Trevelin, S. C., Dos Santos, C. X. X., Ferreira, R. G., et al. (2016). "Apocynin and Nox2 regulate NF- κ B by modifying thioredoxin-1 redox-state." *Scientific Reports* 6: p. 34581.
- Tsalik, E., Henao, R, Nichols, M, et al. (2016). "Host gene expression classifiers diagnose acute respiratory illness etiology". *Science Translational Medicine* 8(322): 322ra11.
- Tsalik, E. L., Langley, R. J., Dinwiddie, D. L., et al. (2014). "An integrated transcriptome and expressed variant analysis of sepsis survival and death". *Genome Medicine* 6: p. 111.
- Tsoukalas, A., Albertson, T., and Tagkopoulos, I. (2015). "From data to optimal decision making: a data-driven, probabilistic machine learning approach to decision support for patients with sepsis." *JMIR Medical Informatics* 3: e11.
- Turvey, S. and Broide, D. (2010). "Innate immunity". *Journal of Allergy and Clinical Immunology* 125(2): S24–32.
- Uhlén, M, Fagerberg, L, Hallström, B., et al. (2015). "Tissue-based map of the human proteome". *Science* 347(6220): p. 1260419.
- van de Vosse, E and Dissel, J. van (2009). "Genetic deficiencies of innate immune signalling in human infectious disease". *The Lancet Infectious Diseases* 9(11): pp. 688–98.
- van der Poll, T., Veerdonk, F. L. van de, Scicluna, B. P., et al. (2017). "The immunopathology of sepsis and potential therapeutic targets." *Nature Reviews Immunology* 17: pp. 407–420.
- Vazquez, D. E., Niño, D. F., Maio, A., et al. (2015). "Sustained expression of lipocalin-2 during polymicrobial sepsis". *Innate Immunity* 21: pp. 477–489.
- Vincent, J. L., Rello, J., Marshall, J., et al. (2009). "International study of the prevalence and outcomes of infection in intensive care units." *Journal of the American Medical Association* 302: pp. 2323–9.
- Vincent, J., Moreno, R, Takala, J, et al. (1996). "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine." *Intensive Care Medicine* 22: pp. 707–10.
- Visscher, P., Wray, N., Zhang, Q, et al. (2017). "10 years of GWAS discovery: biology, function, and translation". *The American Journal of Human Genetics* 101(1): pp. 5–22.
- Vogel, C and Marcotte, E. (2012). "Insights into the regulation of protein abundance from proteomic and transcriptomic analyses". *Nature Reviews. Genetics* 13(4): pp. 227–32.
- Walker, J. (1994). "The bicinchoninic acid (BCA) assay for protein quantitation". *Methods in Molecular Biology* 32: pp. 5–8.
- Wang, H., Zhang, P., Chen, W., et al. (2012). "Serum microRNA signatures identified by Solexa sequencing predict sepsis patients' mortality: a prospective observational study." *PLoS One* 7: e38885.
- Wehrens, R and Mevik, B. (2007). "The pls package: principal component and partial least squares regression in R". *Journal of Statistical Software* 18(2).
- Weledji, E. and Ngowe, M. (2013). "The challenge of intra-abdominal sepsis". *International Journal of Surgery* 11(4): pp. 290–5.

- Wessel, D and Flügge, U. (1984). "A method for the quantitative recovery of protein in dilute solution in the presence of detergents and lipids". *Analytical Biochemistry* 138: pp. 141–143.
- Westra, H., Peters, M., Esko, T, et al. (2013). "Systematic identification of trans eQTLs as putative drivers of known disease associations". *Nature Genetics* 45(10): pp. 1238–1243.
- Westra, H., Arends, D, Esko, T, et al. (2015). "Cell specific eQTL analysis without sorting cells". *PLoS Genetics* 11 (5): e1005223.
- Wiersinga, W. J. (2011). "Current insights in sepsis: from pathogenesis to new treatment targets". *Current Opinion in Critical Care* 17: pp. 480–486.
- Wiersinga, W. J., Leopold, S. J., Cranendonk, D. R., et al. (2014). "Host innate immune responses to sepsis." *Virulence* 5: pp. 36–44.
- Williams, F. M., Carter, A. M., Hysi, P. G., et al. (2013). "Ischemic stroke is associated with the ABO locus: the EuroCLOT study". *Annals of Neurology* 73(1): pp. 16–31.
- Wong, H. R. (2007). "Pediatric septic shock treatment: new clues from genomic profiling." *Pharmacogenomics* 8: pp. 1287–90.
- Wong, H. R. (2013). "Genome-wide expression profiling in pediatric septic shock." *Pediatric Research* 73: pp. 564–9.
- Wong, H. R., Cvijanovich, N., Allen, G. L., et al. (2009a). "Genomic expression profiling across the pediatric systemic inflammatory response syndrome, sepsis, and septic shock spectrum." *Critical Care Medicine* 37: pp. 1558–66.
- Wong, H. R., Cvijanovich, N., Lin, R., et al. (2009b). "Identification of pediatric septic shock subclasses based on genome-wide expression profiling." *BMC Medicine* 7: p. 34.
- Wong, H. R., Freishtat, R. J., Monaco, M., et al. (2010a). "Leukocyte subset-derived genomewide expression profiles in pediatric septic shock." *Pediatric Crit Care Medicine* 11: pp. 349–55.
- Wong, H. R., Wheeler, D. S., Tegtmeier, K., et al. (2010b). "Toward a clinically feasible gene expression-based subclassification strategy for septic shock: proof of concept." *Critical Care Medicine* 38: pp. 1955–61.
- Wong, H. R., Cvijanovich, N. Z., Allen, G. L., et al. (2011). "Validation of a gene expression-based subclassification strategy for pediatric septic shock." *Critical Care Medicine* 39: pp. 2511–7.
- Wong, H. R., Cvijanovich, N. Z., Hall, M., et al. (2012a). "Interleukin-27 is a novel candidate diagnostic biomarker for bacterial infection in critically ill children." *Critical Care* 16: R213.
- Wong, H. R., Cvijanovich, N. Z., Anas, N., et al. (2015a). "A Multibiomarker-Based Model for Estimating the Risk of Septic Acute Kidney Injury." *Critical Care Medicine* 43: pp. 1646–53.
- Wong, H. R., Cvijanovich, N. Z., Anas, N., et al. (2015b). "Developing a clinically feasible personalized medicine approach to pediatric septic shock." *American Journal of Respiratory and Critical Care Medicine* 191: pp. 309–15.
- Wong, H., Cvijanovich, N, and Wheeler, D. (2008). "Interleukin-8 as a stratification tool for interventional trials involving pediatric septic shock". *American Journal of Respiratory and Critical Care Medicine* 178(3): pp. 276–82.

- Wong, H., Salisbury, S., and Xiao, Q (2012b). "The pediatric sepsis biomarker risk model". *Critical Care* 16(5): R174.
- Wong, H., Atkinson, S., and Cvijanovich, N. (2016). "Combining prognostic and predictive enrichment strategies to identify children with septic shock responsive to corticosteroids". *Critical Care Medicine* 44(10): e1000–3.
- Wu, D., Lim, E., Vaillant, F., et al. (2010). "ROAST: rotation gene set tests for complex microarray experiments". *Bioinformatics* 26: pp. 2176–2182.
- Wu, L., Candille, S. I., Choi, Y., et al. (2013). "Variation and genetic control of protein abundance in humans." *Nature* 499: pp. 79–82.
- Wynn, J. L., Cvijanovich, N. Z., Allen, G. L., et al. (2011). "The influence of developmental age on the early transcriptomic response of children with septic shock." *Molecular Medicine* 17: pp. 1146–56.
- Xiao, W., Mindrinos, M. N., Seok, J., et al. (2011). "A genomic storm in critically injured humans". *The Journal of Experimental Medicine* 208: pp. 2581–2590.
- Xie, J, Kiryluk, K, Li, Y, et al. (2016). "Fine mapping implicates a deletion of *cfhr1* and *cfhr3* in protection from iga nephropathy in han chinese". *Journal of the American Society of Nephrology* 27(10): pp. 3187–3194.
- Yang, H., Wang, H., Levine, Y. A., et al. (2016). "Identification of CD163 as an antiinflammatory receptor for HMGB1-haptoglobin complexes." *JCI Insight* 1: e85375.
- Yang, J., Benyamin, B., P, M. B., et al. (2010). "Common SNPs explain a large proportion of the heritability for human height." *Nature Genetics* 42: pp. 565–9.
- Yang, Y., Sun, H., Mo, X., et al. (2012). "Prediction of novel genes associated with negative regulators of toll-like receptors-induced inflammation based on endotoxin tolerance". *Inflammation* 35: pp. 1889–1899.
- Zhang, J, Xin, L, Shan, B, et al. (2012). "PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification". *Molecular and Cellular Proteomics* 11(4): p. M111.010587.
- Zheng, Y., Lee, S., Liang, X., et al. (2013). "Suppression of PTRF alleviates the polymicrobial sepsis induced by cecal ligation and puncture in mice". *The Journal of Infectious Diseases* 208: pp. 1803–1812.
- Zhou, W., Wang, J., Li, Z., et al. (2016). "MicroRNA-2055b inhibits HMGB1 expression in LPS-induced sepsis." *International Journal of Molecular Medicine* 38: pp. 312–8.
- Zhou, Z., Xue, Q., Wan, Y., et al. (2011). "Lysosome-associated membrane glycoprotein 3 is involved in influenza A virus replication in human lung epithelial (A549) cells". *Virology Journal* 8: p. 384.
- Zhu, X., Pribis, J. P., Rodriguez, P. C., et al. (2014). "The central role of arginine catabolism in T-cell dysfunction and increased susceptibility to infection after physical injury." *Annals of Surgery* 259: pp. 171–8.