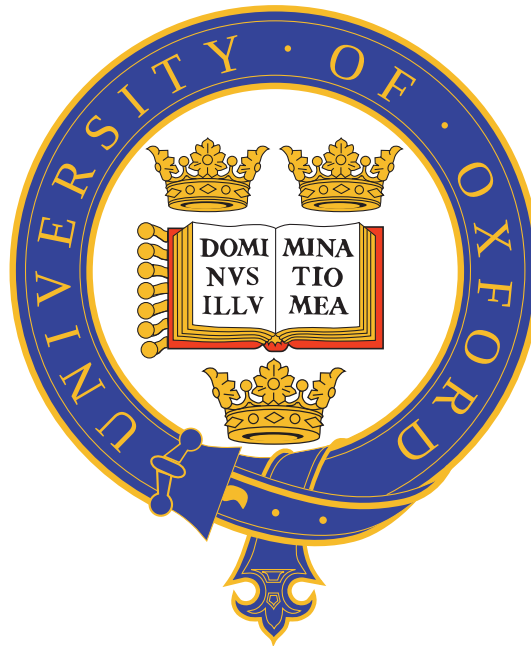




**Functional genomics of variation in
response to infection:
insights into severe sepsis and common
variable immune deficiency disorders**

Emma E. Davenport
Hertford College
University of Oxford

*A thesis submitted in partial
fulfilment of the requirements for the degree of
Doctor of Philosophy
Trinity Term, 2014*

Abstract

Functional genomics of variation in response to infection: insights into severe sepsis and common variable immune deficiency disorders

Emma E. Davenport, Hertford College, Trinity Term 2014

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

Functional genomics uses high throughput genome-wide technologies to investigate the functional consequences of genetic variants on gene expression and protein products. In the context of disease, using integrative functional genomic approaches to understand the genetic variation which underlies disease susceptibility and aetiology may contribute to better diagnosis, treatment and prevention. This thesis investigated genetic determinants of variation in response to infection by applying a functional genomics approach to investigate three clinical cohorts: patients with severe sepsis, an influenza challenge study and patients with common variable immune deficiency disorders.

The transcriptomic response to severe sepsis is reported here for the largest known adult severe sepsis community acquired pneumonia cohort. Two clusters within the cohort, based on gene expression signatures, which have different survival rates and identify individuals with distinct immune responses to sepsis, highlight the value of functional genomics for identifying heterogeneity within patient cohorts. This was further complemented by the resolution of gene expression signatures in healthy individuals following influenza challenge which identified individuals with moderate to severe disease. Shared gene expression signatures between the cohorts highlighted components of the immune response to viral infection important across multiple clinical settings and may assist with the identification of viral infections in the sepsis patients. Gene expression was mapped as a quantitative trait (eQTL). Comparison to data sets for healthy individuals and from innate immune stimulated cells identified regulatory variants specific to the context of sepsis.

Whole genome sequencing for a cohort of patients with common variable immune deficiency disorders was performed. This identified novel variants and pathways which may play a role in the underlying immunopathogenesis of disease. Integration with RNA-seq for a small number of patients allowed prioritisation of non-coding variants based on evidence of altered gene expression. Comparison to the sepsis cohort analysis identified key immune related genes involved in rare and common responses to bacterial infection.

This thesis has demonstrated the value of integrating multiple functional genomic techniques to further our understanding of the mechanisms underlying variation in the response to infection.

Acknowledgements

Firstly, I am grateful for the continuous support and guidance from my supervisor Julian Knight, all the members of the Knight lab and our collaborators.

In particular I would like to thank Peter Humburg for his infinite patience and assistance. I would also like to thank Katharine Plant, Tara Mills, Katie Burnham and Jennifer Shelton for all their help and advice over endless cups of tea (and all the proof-reading).

The unwavering support from my family, Mum, Dad, Rebecca and James, and my friends has kept me going. Finally, Douglas, your love, encouragement and tireless belief in me has made this experience possible.

Declarations

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

The Genomic Advances in Sepsis (GAinS) study was founded in 2005 and has involved patient recruitment from 35 intensive care unit sites around the UK. From October 2010, I have been responsible for liaising with research nurses, maintaining supplies, collecting and processing samples. For the patients presented in this thesis, samples for 91 had been previously processed but I carried out the remainder. Gene expression microarray data generation was carried out at the Wellcome Trust Sanger Institute. Genotyping data were generated at the WTCHG core facility. Helen Lockstone provided advice for microarray QC and data analysis. Tara Mills and Peter Humburg provided components for the genotyping QC and imputation pipeline. Peter Humburg also adapted Matrix eQTL for the eQTL analysis. The results of the GWAS for sepsis outcome were provided by Anna Rautanen.

The influenza challenge study was set up by Sarah Gilbert and colleagues. During this study, blood was collected for gene expression analysis. My role included extracting the RNA, microarray QC, and data analysis. Natasha Sahgal and Helen Lockstone provided advice for microarray QC and data analysis. Richard Antrobus assisted with aspects of the analysis.

The CVID cohort was recruited by Helen Chapel and colleagues. DNA extraction and B cell separation were conducted by Clinical Immunology. WTCHG core Bioinformatics (Jean-Baptiste Cazier, Helen Lockstone and Benjamin Wright), Hilary Martin, Pauline van Schouwenburg and Smita Patel provided assistance with analysis.

Submitted Abstracts

Identifying genes involved in common variable immune sequencing deficiency disorders: an integrated whole genome and RNA sequencing approach 2014

Oral presentation at ESID

van Schouwenburg PA, **Davenport** EE, Kienzler A-K, Marwah I, Wright B, Lucas M, Martin H, WGS500 Consortium, Chapel H, Knight JC and Patel SY

Integrating RNA-seq with whole genome sequencing to investigate the causes of common variable immune deficiency disorders 2014

Poster presentation at Genome Science conference

Davenport EE, van Schouwenburg P, Kienzler A-K, WGS500 Consortium, Chapel H, Patel S and Knight JC

Gene expression signatures in a clinical setting: survival in sepsis 2014

Invited speaker at 2nd Perioperative Care of the Cardiothoracic Surgical Patient Meeting

Davenport EE, Radhakrishnan J, Humburg P, Hutton P, Garrard C, Hinds CJ, Knight JC and The GAINs Investigators

Gene expression signatures in a clinical setting: survival in sepsis 2014

Talk at Quantitative Genomics 2014

Davenport EE, Radhakrishnan J, Humburg P, Hutton P, Garrard C, Hinds CJ, Knight JC and The GAINs Investigators

Gene expression signatures and regulatory variants in patients with severe sepsis due to CAP: associations with outcome 2013

Poster presentation at The Intensive Care Society State of the Art Meeting

Davenport EE, Radhakrishnan J, Humburg P, Mills TC, Gordon AC, Hutton P, Garrard C, Hinds CJ, Knight JC and The GAINs Investigators

Resolving regulatory genetic variants in severe sepsis due to CAP by mapping context specific expression quantitative trait loci 2013

Poster presentation at The American Society of Human Genetics conference

Davenport EE, Radhakrishnan J, Humburg P, Mills TC, Hutton P, Garrard C, Hinds CJ, Knight JC and The GAINs Investigators

Genome-wide imputation in multiple cohorts to study susceptibility to sepsis in Europe 2013

Poster presentation at The American Society of Human Genetics conference

Mills TC, Rautanen A, **Davenport** EE, Hutton P, Knight JC, Meitinger T,

Garrard C, Stueber F, Hinds CJ, Hill AVS, and the GenOSept Investigators

Mapping expression quantitative trait loci in a clinical setting: insights from sepsis 2013

Poster presentation at The Genomics of Common Diseases

Davenport EE, Radhakrishnan J, Humburg P, Mills TC, Hutton P, Garrard C, Hinds CJ, Knight JC and The GAINs Investigators

Functional genomics of severe sepsis in patients with CAP and FP 2012

Talk at Intensive Care Society State of the Art Meeting

Radhakrishnan J, Svoren E, **Davenport EE**, Ellis P, Langford C, Garrard C, Hinds CJ, Knight JC and The GAINs Investigators

A functional genomics approach to the identification of biomarkers of survival in severe sepsis due to CAP 2011

Talk at the Annual Congress of the European Society of Intensive Care Medicine

Radhakrishnan J, Svoren E, Ellis P, Langford C, Hutton P, **Davenport EE**, Thorpe A, Garrard C, Hinds CJ, Knight JC and The GAINs Investigators

Associated Publications

Transcriptomic profiling facilitates classification of response to influenza challenge

Journal Molecular Medicine (2015) 93, 105-114

Davenport EE*, Antrobus RD*, Lillie PJ, Gilbert S and Knight JC

Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study

The Lancet Respiratory Medicine (2015) 3, 53-60

Rautanen A, Mills TC, Gordon AC, Hutton P, Steffens M, Nuamah R, Chiche J-D, Parks T, Chapman SJ, **Davenport EE**, Elliott KS, Bion J, Lichtner P, Meitinger T, Wienker TF, Caulfield MJ, Mein C, Bloos F, Bobek I, Cotogni P, Sramek V, Sarapuu S, Kobilay M, Ranieri M, Rello J, Sirgo G, Weiss YG, Russwurm S, Schneider M, Reinhart K, Holloway PA, Knight J, Garrard CS, Russell JA, Walley KR, Stüber F, Hill AV and Hinds CJ

Application of whole genome and RNA sequencing to investigate the genomic landscape of common variable immunodeficiency

Journal of Allergy and Clinical Immunology – *In submission*

van Schouwenburg PA*, **Davenport EE***, Kienzler A-K, Marwah I, Wright B, Lucas M, Malinauskas T, Martin HC, WGS500 Consortium, Lockstone HE, Cazier J-B, Chapel HM, Knight JC and Patel SY

*These authors contributed equally

Other Publications

A common haplotype of the TNF receptor 2 gene modulates endotoxin tolerance

Journal of Immunology (2011) 186, 3058-3065

Fairfax BP, **Davenport EE**, Makino S, Hill AVS, Vannberg FO and Knight JC

Contents

Abstract	i
Acknowledgements	ii
Declarations	iii
Submitted Abstracts	iv
Associated Publications	vi
Contents	vii
List of Figures	xiii
List of Tables	xv
Abbreviations	xvi
1 Introduction	1
1.1 Aims and objectives	1
1.2 Functional genomics and gene regulation	2
1.2.1 Gene expression	2
1.2.2 Mechanisms of gene regulation	2
1.2.3 Methods for assaying gene expression	4
1.2.4 Insights from functional genomic approaches	6
1.3 Genetic variation and disease	7
1.3.1 Types of genetic variation and functional consequences . .	7
1.3.2 Single gene vs multifactorial diseases	8
1.3.3 Common disease common variant	9
1.3.4 Challenges in heritable traits of multifactorial disease . . .	12
1.3.5 Approaches for studying regulatory variants	13
1.3.6 Approaches for studying rare variants	18
1.4 Infectious disease	19
1.4.1 Immune response to infectious disease	19
1.4.2 Genetic variation and infectious disease	20
1.5 Sepsis	22
1.5.1 Causes of sepsis	23
1.5.2 Historic hypothesis of sepsis	23
1.5.3 Role of immune system	23
1.5.4 Current treatment	25
1.5.5 Clinical trials	25
1.5.6 A new hypothesis	26
1.5.7 Genetic studies of sepsis	26

CONTENTS

1.5.8	Functional genomics of sepsis	27
1.6	Influenza	30
1.7	Common variable immune deficiency disorders	31
1.7.1	Clinical complications	32
1.7.2	Current treatment	34
1.7.3	Challenges of CVID	34
1.7.4	Genetics of CVID	35
1.7.5	Candidate gene studies	38
1.7.6	Functional genomics of PIDs	39
1.8	Specific aims and objectives	40
2	Materials and Methods	42
2.1	Sepsis study	42
2.1.1	GAInS study patient recruitment and exclusion criteria . .	42
2.1.2	Sample collection	43
2.1.3	RNA extraction	43
2.1.4	Microarray data processing and analysis	44
2.1.5	Genomic DNA extraction	44
2.1.6	Genotyping	45
2.1.7	Imputation	45
2.1.8	eQTL QC	45
2.1.9	Statistical analysis	46
2.1.10	eQTL mapping	46
2.2	Influenza study	46
2.2.1	Study subjects and clinical trial procedures	46
2.2.2	RNA extraction and globin mRNA depletion	48
2.2.3	Microarray data processing and analysis	48
2.2.4	Statistical analysis	49
2.3	CVID study	49
2.3.1	Patient recruitment	49
2.3.2	Whole genome sequencing	49
2.3.3	Mapping and variant calling	50
2.3.4	Variant filtering	50
2.3.5	Sanger sequencing	52
2.3.6	RNA-seq	52
2.3.7	Statistical analysis	53
3	Functional genomic response in severe sepsis	54
3.1	Introduction	54
3.1.1	Community acquired pneumonia	54
3.1.2	Severity scores	55
3.1.3	Day of sampling	56
3.1.4	Tissue of interest and RNA extraction method	56
3.1.5	Use of controls	59
3.1.6	Microbiology	59
3.1.7	Prediction models	60
3.1.8	Unsupervised approach to gene expression	60

CONTENTS

3.1.9	Aims	61
3.2	Results	62
3.2.1	Sepsis cohort	62
3.2.2	Patient characteristics	62
3.2.3	Associations between clinical covariates and survival	64
3.2.4	Non-linear relationship between clinical covariates and survival	67
3.2.5	RNA extraction and gene expression microarray QC	70
3.2.6	Differential gene expression for survival	72
3.2.7	Pathway analysis for survival	75
3.2.8	Prediction models for survival	77
3.2.9	Differential gene expression for microbiology	82
3.2.10	Pathway analysis for viral infection	84
3.2.11	Prediction models for viral infection	85
3.2.12	Unsupervised approach to signatures in gene expression .	87
3.2.13	Associations between clinical covariates and cluster	91
3.2.14	Differential gene expression between clusters	93
3.2.15	Pathway analysis for clusters	94
3.2.16	Prediction models for clusters	97
3.2.17	Overlap between survival analysis and cluster membership	99
3.3	Discussion	102
3.3.1	Gene expression signatures relating to outcome	102
3.3.2	Severity scores	102
3.3.3	Microbiology	103
3.3.4	Cluster analysis	104
3.3.5	Limitations	105
3.3.6	Conclusion	106
4	Regulatory genetic variants in severe sepsis	107
4.1	Introduction	107
4.1.1	Genotyping QC	107
4.1.2	Imputation	109
4.1.3	Gene expression QC	109
4.1.4	Matrix eQTL	109
4.1.5	<i>Cis</i> and <i>trans</i> associations	110
4.1.6	PCs vs covariates	110
4.1.7	Integrating eQTL with sepsis GWAS	110
4.1.8	Context specific <i>cis</i> associations	113
4.1.9	Aims	113
4.2	Results	114
4.2.1	Genotyping QC	114
4.2.2	Gene expression QC	114
4.2.3	eQTL analysis with principal component adjustment	115
4.2.4	eQTL analysis with covariate adjustment	116
4.2.5	<i>Cis</i> associations	117
4.2.6	Pathway analysis	120
4.2.7	Imputation	120

CONTENTS

4.2.8	Integration with GWAS	121
4.2.9	Context specific <i>cis</i> associations	125
4.2.10	<i>Trans</i> associations	128
4.2.11	Comparison to cluster membership	130
4.3	Discussion	132
4.3.1	Integration with GWAS	132
4.3.2	Context specific <i>cis</i> associations	133
4.3.3	Limitations	134
4.3.4	Conclusion	134
5	Functional genomic response to influenza challenge	135
5.1	Introduction	135
5.1.1	Influenza	136
5.1.2	Influenza vaccine and challenge study	136
5.1.3	Aims	137
5.2	Results	138
5.2.1	RNA extraction and gene expression microarray QC	138
5.2.2	Differential expression for laboratory confirmed influenza	140
5.2.3	Clinical observations and variation in gene expression	142
5.2.4	Differential gene expression	145
5.2.5	Pathway analysis	150
5.2.6	Gene expression signature	152
5.2.7	Temporal response to influenza vaccination	154
5.2.8	Comparison to gene expression signatures in severe sepsis	156
5.3	Discussion	158
5.3.1	Variation in gene expression	159
5.3.2	Pathway analysis	159
5.3.3	Gene expression signature	159
5.3.4	Temporal response to influenza vaccination	160
5.3.5	Overlap of gene signatures between influenza challenge and severe sepsis	161
5.3.6	Conclusion	162
6	Functional genomic response in CVID	163
6.1	Introduction	163
6.1.1	Patient cohort	163
6.1.2	Previously described CVID genes	164
6.1.3	GWAS genes	164
6.1.4	Filtering variants	165
6.1.5	WGS500	166
6.1.6	Integrating WGS and RNA-seq	166
6.1.7	RNA-seq QC and normalisation	167
6.1.8	Aims	167
6.2	Results	168
6.2.1	Patient cohort clinical data	168
6.2.2	Whole genome sequencing	170
6.2.3	Patient exclusion	170

CONTENTS

6.2.4	Filtering criteria	170
6.2.5	Previously described CVID and PID genes	172
6.2.6	Variants upstream of candidate genes	178
6.2.7	GWAS genes	182
6.2.8	Further patient removal	184
6.2.9	Genome-wide variants	184
6.2.10	WGS pathways analysis	184
6.2.11	IgM analysis	185
6.2.12	Infections only vs complicated subphenotype	187
6.2.13	Grandmother-grandson pair	190
6.2.14	RNA-seq QC	191
6.2.15	Differential gene expression analysis	193
6.2.16	Differential gene expression previously described genes . .	195
6.2.17	RNA-seq pathway analysis	197
6.2.18	Integration of WGS and RNA-seq	199
6.2.19	Overlap with sepsis	204
6.3	Discussion	207
6.3.1	Filtering WGS variants	207
6.3.2	Pathway analysis	207
6.3.3	Integrating WGS and RNA-seq	208
6.3.4	Advantages of WGS	208
6.3.5	Limitations	209
6.3.6	Conclusion	209
7	General Discussion	210
7.1	Heterogeneity within clinical cohorts	211
7.2	Variant prioritisation	212
7.3	Context specificity	214
7.4	Integrative analysis	215
7.5	Clinical translation of findings	216
7.6	Future work	217
7.7	Concluding remarks	221
	Appendices	222
A	Functional genomic response in severe sepsis	224
B	Regulatory genetic variants in severe sepsis	240
C	Functional genomic response to influenza challenge	250
D	Functional genomic response in CVID	259
	Bibliography	291

List of Figures

1.1	Effect of regulatory variants on expression levels of genes	15
1.2	Cells involved in immune response	20
1.3	Nature and number of CVID complications	33
1.4	The effect of known CVID related mutations on the B cell response to antigen	36
2.1	Location of GAINs study recruitment sites	43
3.1	Leukocyte fraction of whole blood	58
3.2	Sepsis cohort characteristics	64
3.3	Density plots for non-linear relationship with survival	69
3.4	Principal component analysis	72
3.5	Volcano plot of differentially expressed probes for survival	74
3.6	<i>CD8A</i> differential expression	75
3.7	Pathway enrichment for survival	77
3.8	ROC curves for survival prediction models	81
3.9	Summary of ROC curves for each survival length	82
3.10	Volcano plot of differentially expressed probes for microbiology	84
3.11	Pathway enrichment for viral infection	85
3.12	ROC curves for viral prediction models	87
3.13	Clusters within the cohort	89
3.14	Kaplan-Meier plots for clusters	90
3.15	Volcano plot of differentially expressed probes for cluster	94
3.16	Pathway enrichment for clusters	95
3.17	Heatmap of IPA disease and functions activity between clusters	96
3.18	Box plots for cluster membership predictor genes	98
3.19	ROC curves cluster membership prediction	99
3.20	Overlap between survival analysis and cluster membership	101
4.1	Unique probes with <i>cis</i> associations	116
4.2	PC vs covariates	117
4.3	IRF5 eQTL	119
4.4	Pathway enrichment for genes with <i>cis</i> associations	120
4.5	Regional association plot for <i>ITIH4 cis</i> eQTL	122
4.6	Genes with <i>cis</i> eQTL unique to sepsis	126
4.7	Pathway enrichment for genes with <i>cis</i> associations unique to sepsis	127
4.8	Overlap of genes with <i>cis</i> associations between sepsis and stimulated monocytes	128
4.9	Enrichment of eQTL for genes differentially expressed between clusters	132
5.1	Principal component analysis	139
5.2	Study subjects clinical observations	140

LIST OF FIGURES

5.3	Volcano plot of differentially expressed probes for LCI	141
5.4	<i>SEPT4</i> expression over time	142
5.5	Principal component and hierarchical clustering plot	144
5.6	Volcano plot of differentially expressed probes for severe LCI . . .	146
5.7	Heatmaps for differentially expressed probes	149
5.8	Comparison of methods for determining differential expression of probes	150
5.9	Pathway enrichment for moderate/severe LCI	151
5.10	Time course for moderate/severe LCI expression signature	153
5.11	Differential expression for vaccination status pre-challenge	155
5.12	Comparison of gene expression signatures in influenza and sepsis	156
5.13	Upstream regulation by IFNG	157
6.1	Variant type and gene region	171
6.2	Network of variants one hop from CVID candidate genes	180
6.3	Network of variants for individual patients	182
6.4	Pathway enrichment for variants	185
6.5	CVID RNA-seq QC	192
6.6	Volcano plot of differentially expressed genes	194
6.7	Box plot for differentially expressed genes	195
6.8	Box plots for differentially expressed candidate genes	197
6.9	Pathway enrichment for differentially expressed genes	198
6.10	Overlap of enriched pathways for WGS and RNA-seq	199
6.11	WGS and RNA-seq integrated variants	201
6.12	<i>SH3BP2</i> compound heterozygous variant	203
6.13	Overlap of differentially expressed gene between CVID and sepsis	205
A.1	Kaplan-Meier plots for three and four clusters	236
B.1	<i>Cis</i> eQTL following adjustment for clinical covariates	241
B.2	Inclusion of PCs and clinical covariates	242
B.3	Correlations between PCs and clinical covariates	243

List of Tables

1.1	The clinical manifestations of SIRS	22
1.2	CVID diagnosis criteria	32
2.1	Filtering criteria for WGS variants	51
3.1	Severity scores	55
3.2	Sepsis cohort demographics and clinical covariates	63
3.3	Associations between continuous clinical covariates and survival .	65
3.4	Associations between count clinical covariates and survival	67
3.5	Differentially expressed probes for survival	73
3.6	IPA mapped genes	76
3.7	Prediction models for survival	79
3.8	Differentially expressed probes for microbiology	83
3.9	Prediction model for viral infection	86
3.10	Sepsis cohort demographics and severity scores by cluster	91
3.11	Associations between continuous clinical covariates and cluster . .	92
3.12	Associations between count clinical covariates and cluster	92
3.13	Prediction models for clusters	97
4.1	Sepsis outcome GWAS	112
4.2	Most significant SNP-probe associations	118
4.3	Overlap between sepsis eQTL sepsis outcome GWAS	122
4.4	Overlap between sepsis eQTL and GWAS catalog	124
4.5	Top master regulators of <i>trans</i> associations	130
5.1	Differentially expressed probes for LCI	141
5.2	Differentially expressed probes for severe LCI	146
5.3	Description of differentially expressed genes	147
5.4	Moderate/severe LCI prediction of Huang study subjects	154
6.1	Previously described genes in CVID	164
6.2	CVID patient cohort clinical information	169
6.3	Variants passing WGS filtering criteria	171
6.4	Number of patients each variant is observed in	172
6.5	Variants in previously described CVID genes	174
6.6	Variants in previously described PID genes	175
6.7	Number of genes upstream of candidate genes	179
6.8	Variants in GWAS genotype association regions	183
6.9	Filters for genome-wide variants	184
6.10	Enriched pathways for IgM+ variants	187
6.11	Enriched pathways for Polyclonal Lymphoproliferation subphenotype	189
6.12	Shared variants between grandmother and grandson.	191

LIST OF TABLES

6.13	Differential expression of PID candidate genes	196
6.14	Summary of WGS and RNA-seq integrated variants	200
6.15	Overlap between sepsis eQTL and CVID GWAS	206
A.1	Clinical covariates measured for severe sepsis cohort	225
A.2	Differential expression for 14 day non-survivors vs 6 month survivors	226
A.3	Differential expression for 28 day non-survivors vs 6 month survivors	228
A.4	Differential expression for 6 month non-survivors vs 6 month survivors	230
A.5	IPA 14 day non-survivors vs 6 month survivors	232
A.6	IPA 28 day non-survivors vs 6 month survivors	233
A.7	IPA 6 month non-survivors vs 6 month survivors	233
A.8	Differential expression for viral vs no viral infection	234
A.9	Differential expression for cluster 2 vs cluster 1	237
A.10	IPA cluster 2 vs 1	238
B.1	Genotyping QC sample removal.	240
B.2	<i>Cis</i> eQTL pathway enrichment	243
B.3	<i>Cis</i> eQTL unique to sepsis pathway enrichment	245
B.4	Master regulators of <i>trans</i> associations	246
C.1	Differential expression for LCI	251
C.2	Differential expression for moderate/severe LCI	253
C.3	IPA moderate/severe LCI 48 hours post-challenge	255
C.4	Differential expression for vaccination status pre-challenge	258
C.5	Influenza and sepsis overlap	258
D.1	Previously described genes for PIDs	260
D.2	CVID GWAS genes	262
D.3	WGS coverage per sample	263
D.4	WGS variant pathway enrichment	263
D.5	Differential expression RNA-seq CVID	267
D.6	RNA-seq pathway enrichment	268

Abbreviations

AID	Activation-induced cytidine deaminase
AIS	Acute ischemic stroke
AUC	Area under the curve
CAP	Community acquired pneumonia
CDCV	Common disease common variants
CNV	Copy number variant
CVID	Common variable immune deficiency
DAMP	Damage-associated molecular pattern
ENCODE	Encyclopedia of DNA Elements
eQTL	Expression quantitative trait loci
FC	Fold change
FDR	False discovery rate
FISH	Fluorescence in situ hybridization
GAinS	Genomic Advances in Sepsis
GenoSept	Genetics of Sepsis and Septic Shock
GWAS	Genome wide association study
HA	Haemagglutinin
haploChIP	Haplotype-specific chromatin immunoprecipitation
HI	Haemagglutination inhibition
HWE	Hardy-Weinberg equilibrium
IBD	Inflammatory bowel disease
ICU	Intensive care unit
IgAD	IgA deficiency
InDel	Insertion or deletion
IPA	Ingenuity pathway analysis
IVA	Ingenuity variant analysis
LCI	Laboratory confirmed influenza
LCL	Lymphoblastoid cell line
LD	Linkage disequilibrium
LIP	Lymphoid interstitial pneumonitis
LPS	Lipopolysaccharide
MAF	Minor allele frequency
MHC	Major histocompatibility complex
MR	Misclassification rate
NA	Neuraminidase
NGS	Next generation sequencing
OR	Odds ratio
PAMP	Pathogen-associated molecular pattern
PBMC	Peripheral blood mononuclear cells

LIST OF TABLES

PC	Principal component
PCA	Principal component analysis
pfu	Plaque forming units
PID	Primary immunodeficiency
PLAID	Phospholipase C γ 2-associated antibody deficiency and immune dysregulation
PRR	Pattern recognition receptor
QC	Quality control
RAR	Retinoic acid receptor
RIN	RNA integrity number
RMA	Robust multi-array
ROC	Receiver operator curve
ROS	Reactive oxygen species
SD	Standard deviation
siRNA	Small interfering RNA
SIRS	Systemic Inflammatory Response Syndrome
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SOCS	Suppressor of cytokine signalling
TF	Transcription factor
TLR	Toll Like Receptor
TMM	Trimmed Mean of M
UNG	Uracil-DNA glycosylase
WES	Whole exome sequencing
WGS	Whole genome sequencing
XL-SCID	X-linked severe combined immunodeficiency

Chapter 1

Introduction

1.1 Aims and objectives

The overall aim of this thesis is to understand the genetic determinants of variation in response to infection by adopting a functional genomics approach. A unique opportunity to address this question in terms of common and rare genetic variants arose through collaborative studies which adopted (1) a genome-wide association study (GWAS) approach to study the role of common genetic variants in severe sepsis due to community acquired pneumonia (CAP), as part of the Genetics of Sepsis and Septic Shock study (GenoSept, www.genosept.eu); and (2) a whole genome sequencing (WGS) analysis of rare variants in common variable immune deficiency disorders (CVID), which is the commonest cause of primary immunodeficiency in adults. Investigating genetic associations involving these diseases was proposed to be informative in identifying the functional genes and pathways that modulate immune response to infection.

This thesis will investigate the functional genomic response to infection for peripheral blood leukocytes using samples from these two patient cohorts together with an influenza challenge study. The analysis will be integrated with data from complementary genomic approaches including GWAS in order to define the functional genetic determinants of the individual response to infection.

1.2 Functional genomics and gene regulation

Functional genomics is an area of research which focuses on assigning biological meaning to genomic data. This is achieved through the use of high throughput genome-wide technologies to investigate the functional consequences of genetic variants for gene expression and protein products (Yang 2012). During the current era of next generation genomics, integrative analysis has become a desirable part of experimental design (Hawkins et al. 2010). By integrating genome-wide datasets, for instance those measuring the transcriptome, transcription factor (TF) binding, chromatin accessibility and histone modification, the functional genome can be mapped (Knight 2012). This will allow valuable insights into the genetic variation underlying disease.

1.2.1 Gene expression

Gene expression is the process by which information encoded in a gene is used to synthesise proteins. This can affect a phenotype by altering the amount or timing of protein production rather than by changing the type of protein produced. It is a complex trait that can, in principle, be influenced by genetic, environmental and epigenetic factors (Nicolae et al. 2010).

1.2.2 Mechanisms of gene regulation

Gene expression regulatory mechanisms include transcription initiation, chromatin condensation, alternative splicing, DNA methylation, mRNA stability and protein degradation. Generating a catalog of regulatory elements, particularly long range elements such as enhancers and insulators, has been a major challenge for the understanding of transcriptional control (Hawkins et al. 2010). However, the recent publications of the Encyclopedia of DNA Elements (ENCODE) Project (Encode Project Consortium 2012) the FANTOM5 enhancer

Introduction

atlas (Andersson et al. 2014) and the FANTOM5 promoter atlas (FANTOM Consortium et al. 2014) have provided new insights into the mechanisms of regulation.

Regulation of transcription

The simplest model of expression regulation involves TFs binding to TF binding sites in promoters or enhancers in a sequence specific manner. Both the identity of the TF as well as the binding affinity can affect initiation of transcription. Therefore variants that alter the TF binding site or the TF amino acid sequence can affect mRNA transcript levels (Stranger and Dermitzakis 2005). However, the DNA of eukaryotes is tightly bound to histones, forming chromatin, which adds complexity to the regulation of transcription. Genes which are actively transcribed are found in decondensed chromatin regions which can be identified through DNase I hypersensitivity assays. Histone acetylation is then required in order to allow transcription factors and RNA polymerase to bind to the DNA (Cooper 2000).

DNA modification

Gene expression can also be regulated by environmental interactions or external stimuli or substances, for example temperature, microorganisms, drugs or chemicals (Jaenisch and Bird 2003; Idaghdour and Awadalla 2012). Epigenetics encompasses the sequence-independent reversible heritable changes that can modulate gene expression. It includes DNA methylation and chromatin remodelling which provides a link between the environment and gene expression. For example histone tails can be chemically modified spontaneously or in response to environmental factors. Acetylation of histones alters the chromatin structure and usually results in the transcription of nearby genes being switched on (Choi and Kim 2007; Rivera and Ren 2013).

1.2.3 Methods for assaying gene expression

Gene expression can be measured at the genome-wide scale using either microarray or RNA-seq technology. The ability to measure gene expression across the genome, rather than at the candidate gene level, allows the identification of genes and pathways which have not been previously associated with the condition under study.

Microarray and RNA-seq

Whole-genome gene expression microarrays, a relatively cheap, high throughput technology, allow the expression levels of most annotated genes to be measured within a biological sample. Illumina's BeadArray technology (www.illumina.com) involves beadchips which contain thousands of silica beads. A probe with a sequence specific to a region of the genome is attached to each bead. cRNA is fluorescently labelled and then incubated with the beads. Complementary sequences bind and any unbound cRNA is washed off. Fluorescence emission is detected and is proportional to the amount of binding and level of gene expression.

Microarray quality control

There are a number of limitations to microarray analysis. Firstly quality control (QC) steps are required to remove the low-quality data for individual probes or samples. Normalisation methods are then required to adjust for any experimental variability such as sample preparation differences, hybridisation or batch effects from different chips (Stranger and Dermitzakis 2005). Additionally, microarrays rely on the principle of hybridization and therefore knowledge of gene sequence is required in order to design complementary probes. It is also difficult to compare expression levels across different experiments with

Introduction

complicated normalisation methods required to overcome batch effects (Wang et al. 2009).

RNA-seq

Microarray technology is now frequently replaced by next generation sequencing (NGS) technology such as RNA-seq. The sequencing reads are aligned to a reference genome or assembled *de novo* (Wang et al. 2009) and the level of gene expression can be determined by the proportion of reads which align to a given gene. This avoids the hybridization bias caused by SNPs within a probe and allows the discovery of novel sequences as it is not dependent on gene sequence knowledge. The technical biases inherent to microarrays are not present in RNA-seq experiments which reduces the sophistication of the normalisation methods required to compare transcriptomes across different tissues or conditions. However, systematic variation has still been reported such as differences between sample library sizes (Dillies et al. 2013). Cost remains high, particularly when a deeper coverage is required for example if exploring allele specific expression. It is more challenging to calculate the coverage of a transcriptome compared to a genome because the level to which genes are expressed is not generally known. Large amounts of data are produced which generates new bioinformatic challenges for data storage and read assembly, particularly short reads for repetitive regions . Additionally, the protocol includes the generation of cDNA libraries which requires several stages. These libraries are amplified which can lead to many identical short reads. It can be difficult to determine whether these reads are a true reflection of RNA abundance or if they are a PCR artefact (Cookson et al. 2009; Wang et al. 2009).

1.2.4 Insights from functional genomic approaches

Differential gene expression

One of the main goals of studies which generate genome-wide gene expression data is the identification of genes which are differentially expressed between two different conditions. In the context of disease, this can be informative for identifying disease specific signatures or relationships with certain outcomes. This has already had some success in clinical research for example Golub et al. (1999) were able to use gene expression to distinguish acute myeloid leukemia and acute lymphoblastic leukemia, van 't Veer et al. (2002) identified a gene expression signature which was predictive of poor outcome for breast cancer patients and in the context of pulmonary disease, gene expression signatures were used to define molecular subtypes within a disease as well as identify biomarkers to aid disease diagnosis and tailor therapy (Campbell et al. 2011).

Pathway analysis

Interpretation of the lists of genes generated by differential expression analysis can be enhanced by pathway analysis. This involves methods which group clusters of differentially expressed genes into pathways of biological functions (Emmert-Streib and Glazko 2011). This can be particularly helpful when dealing with heterogeneous conditions where differences are more likely to be seen at a pathway level rather than a gene level. Qiagen's Ingenuity Pathway Analysis (IPA[®], www.qiagen.com/ingenuity, QIAGEN Redwood City) is a web-based application designed for biological interpretation of genomic data. It is based on the Ingenuity Knowledge Base which is a repository of biological interactions and functional annotations curated by dedicated scientists. In addition to analysing a list of genes which are differentially expressed between two different conditions, it can also take into account the fold change difference allowing the

identification of up- or down-regulated pathways. For example, Dey-Rao et al. (2014) applied pathway analysis to their investigation of lupus erythematosus, an autoimmune condition affecting multiple organs including skin. They identified differentially expressed genes between cutaneous lupus compared to systemic disease. Pathways analysis revealed the enrichment of apoptosis and type 1 IFN signalling which are known to be important in lupus pathogenesis.

Although helpful for formulating new hypotheses, pathway analysis does have limitations. The analysis relies on information from publicly available datasets and this can result in bias towards genes/pathways with greater annotation. In the case of sequence variants pathway enrichment can also be affected by gene size with larger genes more likely to contain a variant of interest by chance (Wang et al. 2010).

1.3 Genetic variation and disease

Understanding the genetic variation underlying disease susceptibility and aetiology may contribute to better diagnosis, treatment and prevention of disease. In general, GWAS have limited value in terms of risk prediction but are important for the discovery of new genes and pathways involved in disease pathogenesis and potential drug targets (Kooperberg et al. 2010).

1.3.1 Types of genetic variation and functional consequences

Differences in DNA sequence together with structural genomic variants within human genomes constitute the genetic diversity seen within a population. Genetic polymorphisms are allelic variants that exist at a frequency of generally greater than 1% in a population. The most frequent type is the single nucleotide polymorphism (SNP) which can be a substitution, a deletion or an insertion of a single nucleotide (Ku et al. 2010).

The functional consequence of a particular variant is dependent on its type and position with respect to the affected gene (Haraksingh and Snyder 2013). Variants that occur in the coding region (exons) of a gene and change the amino acid sequence of the protein are known as non-synonymous while those that have no effect on the amino acid sequence are known as synonymous. Variants in coding regions are more easily assessed for functional consequence for example missense, nonsense or splicing. Variants that fall in non-coding regions such as promoters, enhancers and introns can affect the level of gene expression. The consequences of these variants are more complicated to predict as the relationship between a variant in the promoter or intergenic sequence and gene expression or phenotype is not as well understood (Glazier et al. 2002; Gaffney 2013). Alterations in gene expression, both over-expression and loss of expression, can result in disease. For instance the translocation of the *c-myc* gene into the immunoglobulin heavy chain locus results in over-expression of the c-Myc TF and is associated with Burkitt's Lymphoma (Hecht and Aster 2000).

1.3.2 Single gene vs multifactorial diseases

Diseases such as sickle-cell anaemia are caused by a variant in a single gene and tend to show Mendelian pattern of inheritance. In contrast, multiple genetic variants and environmental factors are required for multifactorial diseases such as type 1 and type 2 diabetes (Jin et al. 2012).

Technologies such as whole exome sequencing (WES) and WGS have recently been very successful in identifying the disease causing variant for single gene diseases. Different study designs can be adopted depending on the mode of inheritance observed for the disease (Bamshad et al. 2011). For example, selecting a small number of both affected and unaffected individuals from a family pedigree for sequencing allows the filtering of variants to include only those that segregate with the disease phenotype. This strategy was used in a study of

punctate palmoplantar keratoderma, a rare autosomal skin disorder, to identify a missense substitution in the gene *COL14A1* (Guo et al. 2012). Parent-child trios have assisted with identifying *de novo* mutations in the affected individual as illustrated by a study of severe early-onset epilepsy in which *de novo* non-synonymous mutations were identified in four of the six trios sequenced (Martin et al. 2014).

Common multifactorial diseases are generally more complicated and therefore different methods and study designs are required. It is normally suspected that multiple variants and genes will contribute to the disease phenotype making it harder to identify the underlying genetic aetiology (Glazier et al. 2002; Manolio 2010). In these situations, combining gene expression analysis with DNA sequencing can provide the additional evidence required to prioritise variants which may be disease causing (Wilkerson et al. 2014).

1.3.3 Common disease common variant

The Common Disease, Common Variant (CDCV) hypothesis predicts that the major contributors to common diseases are variants with a common frequency in the population but relatively low penetrance (Schork et al. 2009). Therefore few, if any of the variants are absolutely required to develop the disease (Manolio 2010).

Genome-wide association studies

GWAS rely on the CDCV hypothesis and are based on the principle of linkage disequilibrium (LD) (Visscher et al. 2012). LD is the non-random association of alleles at two or more loci. The particular combination of alleles observed is referred to as a haplotype and many common SNPs are transmitted across generations in haplotype blocks as a result of LD in regions with low rates of historical recombination. The boundaries of the blocks and the haplotypes

Introduction

observed within them are shared across populations which allows a small number of tagging SNPs to capture the variation in the region (Gabriel et al. 2002; Slatkin 2008). The International HapMap project, launched in 2002, has provided a publicly available database of common SNPs and haplotype structures across multiple populations (International HapMap Consortium 2003). The database now consists of 692 individuals from 11 populations and has facilitated the use of GWAS for the identification of genomic loci that influence human disease (International HapMap Consortium et al. 2010).

GWAS have so far focused on genotyping a representative panel of common SNPs with a minor allele frequency (MAF) of greater than 1-5%. The identification of the complement factor H gene as a major risk factor for age-related macular degeneration was one of the first successes of GWAS (Klein et al. 2005).

Inflammatory bowel disease (IBD) has shown the greatest success for the identification of associated loci through GWAS for a multifactorial disease to date. The GWAS meta-analysis of Crohn's disease and ulcerative colitis, two common forms of IBD, conducted by Jostins et al. (2012) reported 163 regions reaching genome-wide significance. However, only 29 associated SNPs were in strong LD with a missense coding variant and the total disease variance explained was only 13.6% for Crohn's disease and 7.5% for ulcerative colitis. This is not unique to this particular study. Most of the associated variants from GWAS (>90%) do not fall in coding regions (Maurano et al. 2012) and the studies have only explained a small fraction of the genetic component of many common human diseases.

Missing heritability

The literature cites a number of possible reasons for the missing heritability from GWAS. For example, rare and structural variants are poorly captured by existing

arrays. Epigenetics, gene-gene (epistatic) and gene-environment interactions may also play a role which current studies are not powered to detect. Limitations in study design such as imprecise phenotyping or control groups which do not provide appropriate comparisons could reduce the estimates of effect sizes. Similarly case-control designs prevent exclusion of latent diagnosis. Finally, genes that increase susceptibility to a particular disease may not be shared by as many people as expected (Schork et al. 2009; Manolio et al. 2009; Glazier et al. 2002; Maher 2008; McCarthy et al. 2008).

Finding the missing heritability

There are a number of approaches that could be undertaken to address this missing heritability.

Although the current SNP arrays cover a good proportion of the genome (>80%) there are still some unmapped regions and certain variants which are not well tagged for instance copy number variants (CNVs) and other structural variants (Cookson et al. 2009). Improving both laboratory and statistical methods for the detection of these variants would allow their contribution to disease susceptibility to be assessed. If these methods could be applied to studies that have already been conducted, it would allow data mining for further associations.

Common variants which are currently detectable may have low penetrance that cannot be statistically linked with a disease through GWAS. Increasing the number of individuals in studies may allow the discovery of further associations (Maher 2008) or carrying out meta-analysis from current studies could further improve the yield of associations.

Epigenetic changes can alter the accessibility of DNA for transcription and therefore the level of gene expression (Richards 2006). This can result in two individuals with the same genotype expressing different phenotypes. Collecting material from appropriate tissues suitable for epigenetic analysis, for example

cryopreserved nuclear preparations for chromatin studies, could help identify variants that act through this method (Manolio et al. 2009; Cirulli and Goldstein 2010).

Alternatively, genetic control may be due to rare variants with larger effects (the Common Disease, Rare Variant hypothesis) (Schork et al. 2009). Some of the common variants associated with disease may actually result from multiple rare variants that associate by chance more frequently with one allele at a common SNP than the other. Or, moderately penetrant variants which are too rare to currently detect may also play a role (Maher 2008). In order to test these theories, the rare variants must be identified. The development of high-throughput sequencing methods should assist with answering these questions for example, resequencing GWAS loci identified two rare coding variants (in *NOD2* and *CARD9*) associated with IBD which exceeded genome-wide significance (Rivas et al. 2011).

1.3.4 Challenges in heritable traits of multifactorial disease

In order for an association study to identify a disease susceptibility locus, an increased frequency of specific disease alleles in the affected individuals must be detected. This requires the locus to have a relatively simple allelic spectrum. For certain diseases, such as autosomal dominant and x-linked conditions, selection can reduce the frequency of the disease causing alleles (Reich and Lander 2001) making it harder to detect an association. High penetrance variants may not be involved in many common multifactorial diseases because it is likely that evolution would prevent their frequency from reaching common proportions in the population (Maher 2008). Due to the large number of tests required in a GWAS and therefore the increased chance of false positives, a stringent threshold for significance is used so large sample sizes are required (Hunter and Kraft 2007). Before functional studies are conducted, replication of the association in

an independent cohort is usually required to increase confidence in the statistical evidence (Todd 2006).

Once variants have been identified, the best way of translating the findings into clinical use is still under debate. The predictive power of the variants will depend on which populations they were discovered in as well as the effect sizes and the disease incidence (Manolio 2010). If it is found that effect sizes of common variants are very small for multifactorial diseases and therefore large areas of the genome would be required to explain the genetic component, it would prove challenging to use in a medical setting (Goldstein 2009).

1.3.5 Approaches for studying regulatory variants

Despite significant associations between genetic loci and disease phenotypes through GWAS, many studies are unable to translate the findings into an understanding of how the loci contribute to the disease, particularly when the signals map to non-coding regions (Cookson et al. 2009). More than 90% of associated variants from GWAS fall in non-coding regions of the genome but they are significantly enriched in DNaseI hypersensitivity sites suggesting a regulatory role (Maurano et al. 2012). Gene expression can therefore provide a link between genetic variation and disease phenotype (Liang et al. 2013).

Mapping quantitative traits

Gene expression can be mapped as a quantitative trait (Schadt et al. 2003). Expression quantitative trait loci (eQTL) analysis identifies associations between genotyping data and gene expression across the whole genome (Michaelson et al. 2009). It can provide an insight into the molecular mechanisms which link a gene to a disease phenotype but may also help identify networks of genes involved in the disease pathogenesis (Cookson et al. 2009). The identification of SNPs, associated with childhood asthma, strongly associating in *cis* with the transcript

levels of the gene *ORMDL3* is a successful example of this approach (Moffatt et al. 2007).

***Cis* and *trans* eQTL**

SNPs with an eQTL association can be classified as either *cis* or *trans* acting depending on the distance between the SNP and gene of interest. Although somewhat arbitrary, *cis* acting variants generally reside 250kb-1Mb from the gene. *Trans* SNPs are located elsewhere in the genome, including other chromosomes (Figure 1.1) (Williams et al. 2010). *Cis* acting eQTL tend to be enriched around the transcription start site of the gene and occur more frequently in exonic regions (Veyrieras et al. 2008; Nica et al. 2011). *Trans* effects are often weaker but may involve master regulators: factors that have multiple effects on gene expression (Morley et al. 2004). *Trans* SNPs may exert their effects through transcript stability as well as transcript production for example through protein-RNA interactions or small interfering RNAs (siRNAs) (Grosshans and Filipowicz 2008).

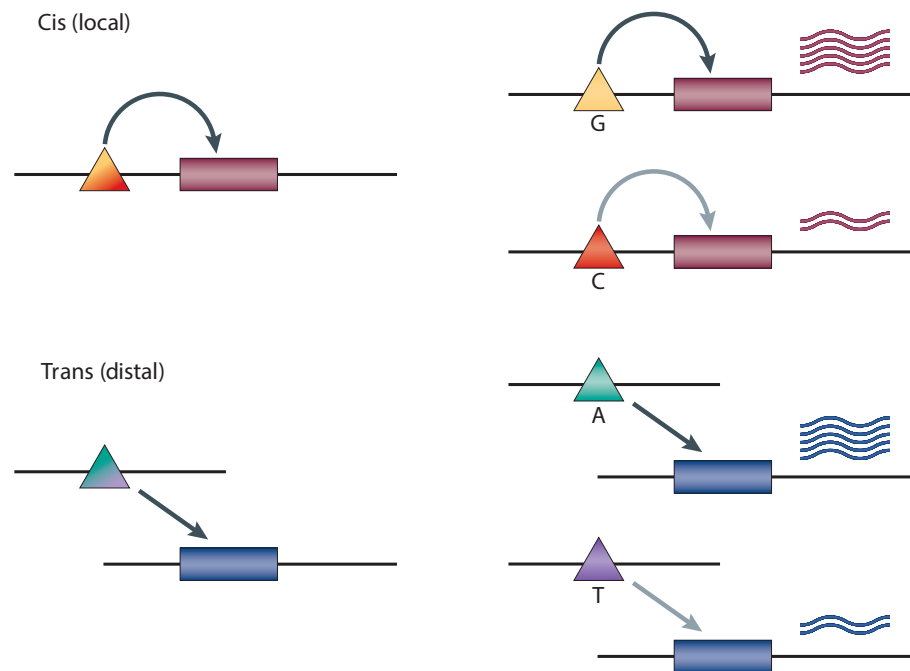


Figure 1.1: Effect of regulatory variants on expression levels of genes. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics (Cheung and Spielman 2009), copyright 2009. The effect of local (*cis*) or distal (*trans*) regulatory variants on levels of gene expression.

Cell type and context

Many of the initial eQTL studies were conducted using lymphoblastoid cell lines (LCLs), mainly because they provide an easily accessible single cell type. LCLs are immortalised by Epstein-Barr virus infection of B cells but genetic artefacts can be introduced from multiple passages of storage and re-growth (Cookson et al. 2009).

It is known that many genes have expression profiles that are dependent on tissue, developmental stage or physiological condition. Studies are now showing that many eQTL associations are tissue or cell specific (Fairfax et al. 2012; GTEx Consortium 2013). eQTLs which are cell specific tend to be located at a greater distance from the transcription start site of the gene, correlating with enhancer elements (Dimas et al. 2009; Nica et al. 2011).

Introduction

eQTL studies have now determined the influence of genetic variation on gene expression between different populations, between different individuals within one population, between different tissues and in response to different environmental factors (Dimas et al. 2009; Nica et al. 2011; Fairfax et al. 2012; Fairfax et al. 2014; Stranger et al. 2012; GTEx Consortium 2013). This highlights the importance of context for interpretation of associations. It is not necessarily trivial to predict which tissue or cell type will be most informative for different diseases so one approach is to measure gene expression in a number of different cell types (Nica et al. 2011). However, the appropriate tissue for a particular disease is not always readily available. This, along with the reagent costs, has limited many eQTL studies to several hundred individuals but highly significant associations have still been detected (Schadt et al. 2003; Cheung et al. 2003) suggesting the intermediate phenotype is less heterogeneous than the whole organism.

The activity of a regulatory variant can also change within a cell type depending on the cellular context as demonstrated by a recent study by Fairfax et al. (2014). This study demonstrated the context specificity of regulatory variants in monocytes following innate immune stimulation. The effect of context has not been as widely studied as that of cell or tissue type, particularly for disease cohorts (Westra and Franke 2014). Studies that have used disease cohorts tend to be smaller due to difficulties in accessing the appropriate patients and the costs involved in recruiting the cohort but successes are being reported. A cohort of 94 malaria-infected children and 61 age-matched controls recently revealed five peak SNP-gene associations in response to malaria infection. The SNPs had an effect in only one group or the effect was in the opposite direction. Many genes were identified which had interaction effects beneath genome-wide significance but were thought to play a role in modulating the course of infection due to the different magnitudes of effects observed between the two groups.

Several had been previously associated with malaria demonstrating the value of context specific eQTL for the facilitation of susceptibility gene mapping and understanding causal mechanisms (Idaghdour et al. 2012).

Function of regulatory regions

Carrying out eQTL analysis at a genome-wide level allows the identification of potentially functionally important regulatory variants without needing knowledge of the location or mechanisms of action of *cis* and *trans* regulatory regions (Stranger and Dermitzakis 2005).

Once a locus has been identified as an eQTL further work is still required to understand how the SNP or SNPs may regulate expression. Current knowledge of regulatory regions is still somewhat limited but databases are being curated with information from a variety of sources to help build a better picture of regulatory regions for example ENCODE (Encode Project Consortium 2012), FANTOM5 (Andersson et al. 2014; FANTOM Consortium et al. 2014) and RegulomeDB (Boyle et al. 2012).

Even with extensive fine mapping and regulatory annotation, it can still be challenging to identify the causal SNP for an eQTL association. This is due to LD with many non-causal SNPs associating with gene expression (Gaffney et al. 2012).

Gaffney et al. (2012) used HapMap LCLs to investigate the biological mechanisms underlying eQTL. They found that the most significant SNP for approximately 20% of eQTL occur within a DNaseI hypersensitivity site, a sign of open chromatin, even though this covers just 1% of the genome. In addition one of the most enriched regions were binding sites for immune related TFs. Other studies have also been able to link genetic variation to chromatin state (McVicker et al. 2013) which is helping to build a better picture of how non-coding regulatory sequences contribute to complex traits.

As reviewed by Knight (2014) a further approach for resolving functional variants, which complements eQTL mapping, is allele-specific gene expression. For coding SNPs, the allelic origin of a transcript can be determined from the abundance of each allele in the transcribed RNA (using RNA-seq). For non-transcribed variants, haplotype-specific chromatin immunoprecipitation (haploChIP) using antibodies for phosphorylated RNA polymerase II identifies transcriptional activity.

1.3.6 Approaches for studying rare variants

The advent of NGS has allowed the identification of rare variants, which may contribute to observed phenotypes (Manolio et al. 2009; Manolio 2010). NGS makes use of massively parallel sequencing (MPS) technology to sequence large quantities of DNA more rapidly and cheaply than the commonly used chain termination method (Sanger sequencing). The DNA sample is randomly fragmented, sequenced and then aligned to a reference sequence. This results in each target being sequenced multiple times as a series of overlapping short reads (Gundry and Vijg 2012).

WGS is already being used in the clinic for small numbers of patients with undiagnosed genetic disorders (Bick and Dimmock 2011). For the analysis of multifactorial diseases, these techniques generate a new set of challenges. In order to sequence the number of individuals necessary for statistically powered GWAS, the cost needs to decrease and new computational approaches to deal with the increased data sets must further evolve (Flicek and Birney 2009). Additionally, GWAS have tended to focus on statistical evidence for association rather than biological plausibility. It has been estimated that on average 74 germline single nucleotide variants (SNVs) occur *de novo* in an individual's genome (Veltman and Brunner 2012). Determining which are responsible for a phenotype could therefore be particularly problematic (Manolio et al. 2009).

1000 Genomes Project

The 1000 Genomes Project (1000 Genomes Project Consortium et al. 2012) aims to provide a comprehensive resource of human genetic variants with a frequency of at least 1%. This will provide a better reference panel for genetic analysis and will assist with ascertaining which variants are rare ($<1\%$) (Manolio et al. 2009).

1.4 Infectious disease

Infectious diseases are caused by microorganisms such as bacteria or viruses. In 2012, infectious disease was the leading cause of death in low-income countries and lower respiratory infections were the sixth most common cause in high-income countries (World Health Organization 2014b). Infectious diseases therefore remain a public health burden around the world affecting both health and economics (Fonkwo 2008). Understanding the genetic variation which underlies disease susceptibility and aetiology may contribute to better diagnosis, treatment and prevention.

1.4.1 Immune response to infectious disease

Both the innate and the adaptive immune system are triggered in response to infection. An inflammatory response is mounted by innate immune cells which release cytokines, chemokines and activate the adaptive immune system. Two strategies are used by the adaptive system: a humoral response involving antibodies produced by B cells to target extracellular microorganisms and a cell-mediated response involving T cells to target intracellular microorganisms (Lilic 2009). The cells involved in these two responses are summarised in Figure 1.2.

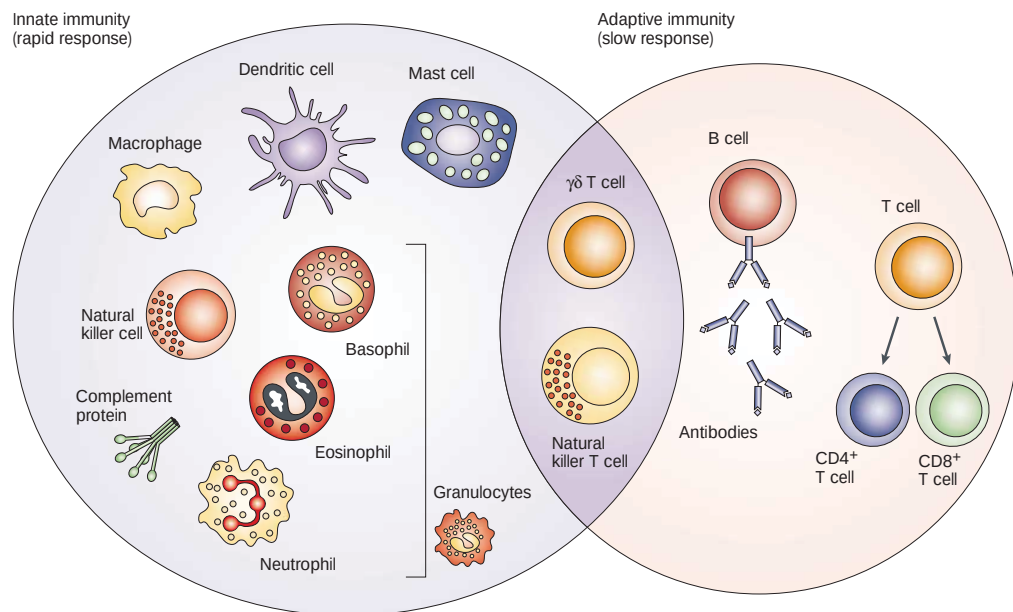


Figure 1.2: Cells involved in immune response. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer (Dranoff 2004), copyright 2004. The immune response to infection is composed of two branches: the innate and adaptive. The cells, soluble factors and cellular components of the innate immune response provide the first line of defence. The B and T cells of the adaptive response provide a slower but more specific response.

1.4.2 Genetic variation and infectious disease

Infectious diseases require exposure to a microbial agent and therefore, they are often regarded as examples of purely environmental diseases. However, a number of observations suggest a strong genetic component. Firstly, there is great variability in the response of individuals to a given infectious agent: some are never infected, only some of those affected will develop a disease and those that do develop symptoms display a large spectrum of clinical manifestations (Alcais et al. 2009). Adoption studies have demonstrated the heritability of infectious disease predisposition by studying premature death due to infection (Sørensen et al. 1988). Finally, there is evidence that the concordance rate of infectious disease is higher in monozygotic than dizygotic twins (Obel et al. 2010).

Introduction

Infectious disease susceptibility, disease outcome and individual response to therapy can all be influenced by genetic variation. Homozygous carriers of a non-synonymous SNP (rs334) in the β -globin chain of haemoglobin experience sickle cell disease. However, heterozygous individuals have a tenfold reduced risk of severe malaria, illustrating genetic influence on disease susceptibility (Allison 1954). A SNP upstream of *IL28B* is associated with spontaneous clearance of the hepatitis C virus, showing disease outcome can have a genetic component (Thomas et al. 2009). Severe hypersensitivity to an antiretroviral drug used to treat HIV (Abacavir) has been associated with HLA-B*5701 demonstrating the genetic influence on response to therapy (Mallal et al. 2002).

GWAS have been conducted for a number of infectious diseases (for example severe malaria, tuberculosis and leprosy) and many have shown evidence of associations (Chapman and Hill 2012). For example, the gram-negative bacteria *Neisseria meningitidis* causes meningococcal disease which can lead to sepsis. Davila et al. (2010) reported variants in the complement factor H region (CFH) are associated with susceptibility to meningococcal disease. The bacteria can bind CFH in order to evade complement-mediated killing providing a plausible biological mechanism of action for this association.

The Major Histocompatibility Complex (MHC) is a gene-dense, highly polymorphic region located on chromosome 6 encoding many of the proteins essential to immunity including those involved in antigen processing and presentation. It has been associated with more diseases than any other region of the genome. However, the degree of polymorphism, strong LD and high gene density make it difficult to determine the causal variants. This region is challenging to sequence using high-throughput technologies due to the issues of mapping short reads to such a polymorphic region and the choice of reference haplotype (Trowsdale and Knight 2013).

1.5 Sepsis

Sepsis is defined as the systemic inflammatory response to the presence of an infection and is classified as severe when associated with organ dysfunction, hypoperfusion abnormality or sepsis-induced hypotension. Systemic inflammatory response syndrome (SIRS) is used to describe the inflammatory response and it includes at least two of the clinical manifestations detailed in Table 1.1 (Bone et al. 1992). However this definition is limited in that the microbiological basis of the response is not considered and there is no graduation of severity. In the UK, severe sepsis accounts for 27% of ICU admissions (Padkin et al. 2003) and from a number of studies in both the UK and America, mortality has been estimated to be between 28% and 50% (Angus et al. 2001; Padkin et al. 2003; Sands et al. 1997; Zeni et al. 1997).

Table 1.1: The clinical manifestations of SIRS (Bone et al. 1992). SIRS is diagnosed in patients with at least two of these clinical manifestations.

SIRS clinical manifestations
Body temperature >38°C
Heart rate >90 beats per minute
Respiratory rate >20 breaths per minute or hyperventilation as indicated by a PACO ₂ of <32 mm Hg
An alteration in the white blood cell count (count >12,000/cu mm or <400/cu mm or the presence of >10% immature neutrophils)

1.5.1 Causes of sepsis

Sepsis can be caused by a multitude of different organisms and patients can have different anatomic sites of infection (Carlet et al. 2008). However, since the 1980s, Gram-positive bacteria have been the most common causative organisms of sepsis in the developed world (Martin et al. 2003), particularly in the community (Brun-Buisson et al. 1995), and the most common site for infection is the respiratory tract (Danai and Martin 2005). CAP is therefore a common cause of sepsis. The bacteria *Streptococcus pneumoniae* is the most common causative pathogen of CAP but viruses are also a common cause of pneumonia in adult patients. However, determining the causative organism remains a challenge in many patients (File 2004).

1.5.2 Historic hypothesis of sepsis

Although sepsis can be recognised clinically, it is challenging to define epidemiologically, biochemically or molecularly. Patients often present with high fevers, shock and respiratory failure so historically, it was hypothesised that excessive inflammation was the basis for an adverse outcome in sepsis (Hotchkiss et al. 2013).

Many studies have been conducted on single proteins or pathways based on the hypothesis that a series of reproducible and consistent physiological events occur as a result of a single molecule (Tang et al. 2009).

1.5.3 Role of immune system

It is thought that a number of factors have the potential to trigger SIRS. For example lipopolysaccharide (LPS) released from bacteria, products released after ischemia-reperfusion or after blunt trauma in which no infectious agent is present (Bosmann and Ward 2013).

Innate response and pathogen recognition

Following infection with a microbial agent, the innate immune response uses pattern recognition receptors (PRRs) to detect the presence of pathogen-associated molecular patterns (PAMPs) and damage-associated molecular patterns (DAMPs). PAMPs are conserved molecular structures found on the microbe while DAMPs are extracellular matrix components or intracellular constituents released from tissue damage or immune cell activation. This non-specific response is used to control the infection while the more specific adaptive immune response develops (Bianchi 2007). For example, LPS from Gram-negative bacteria is detected by TLR4, a member of the Toll-like receptor (TLR) family. This leads to the production of inflammatory mediators and is thought to be a key pathway in sepsis (Bosmann and Ward 2013).

Complement activation

The appearance of complement activation products in plasma (for example C3a and C5a) provides evidence for complement activation in patients with sepsis. From experimental models of sepsis in rodents, evidence suggests that complement activation contributes to adverse outcomes. Additionally in the early stages, neutrophils develop an exaggerated ability to respond to inflammatory mediators but as time progresses, they lose innate immune functions. This suggests that signalling pathways involving G protein-coupled receptors have become defective (Ward 2010).

Reactive oxygen species

Imbalances in the redox system during sepsis leads to the formation of an oxidant state (Andrades et al. 2011). Reactive oxygen species (ROS) are generated by activated neutrophils and macrophages during the inflammatory response. This can result in the oxidation of proteins, lipids and DNA leading to a reduction

in biochemical function. A link between SIRS and ROS has been suggested following the observation of an association between clinical improvement and the administration of antioxidant or antioxidant promoting factors (Bosmann and Ward 2013).

1.5.4 Current treatment

Sepsis is currently treated with antibiotics and supportive care (ventilation, blood pressure support, renal replacement therapy) but there are no specific therapies. The only licensed product, activated protein C, was recently withdrawn by the manufacturer due to risks of bleeding and little further evidence for survival benefits following treatment (Marti-Carvajal et al. 2012).

1.5.5 Clinical trials

Despite advances in the basic science of innate immunity, no potential targets have been translated into new treatments (Cohen et al. 2012). The concept of sepsis being a cytokine storm led to clinical trials of toll receptor blockers, endotoxin antagonists, TNF and interleukin 1 antagonists (Hotchkiss et al. 2013). For example a TLR4 antagonist did not progress beyond a phase 3 clinical trial because the primary efficacy endpoint was not achieved Opal et al. (2013) and toxic effects halted a phase 2-3 study of talactoferrin (human recombinant lactoferrin) (Cohen et al. 2012). These trials may have been confounded by the lack of reliable biomarker for selecting suitable patients to enroll, variable case definitions between studies resulting in different subgroups of disease being studied and many drugs suffer from a low signal to noise ratio. This supports the importance of identifying and studying clinically and microbiologically defined groups. If gene expression signatures allowed the identification of homogeneous subsets of patients, specific drugs could be used to target the relevant pathways (Cohen et al. 2012).

1.5.6 A new hypothesis

Despite the concerted effort to identify targets for therapeutic manipulation and many clinical trials, minimal improvement in mortality over the last 25 years has been made (Deans et al. 2005). The heterogeneous nature of the disease may be contributing to this failure (Wiersinga 2011) and highlights the need for further investigation into subgroups of sepsis patients which may benefit more from certain treatments. Another possible reason for the failure of clinical trials is an inadequate understanding of the pathophysiology underlying sepsis. The possibility of the involvement of the anti-inflammatory response in addition to the pro-inflammatory response lead to the development of therapeutic agents with a two phase model in mind. This consisted of an initial phase of SIRS followed by anti-inflammatory response syndrome. These agents have been unable to demonstrate any survival benefit and therefore it is now hypothesised that sepsis cannot be thought of as a simple pro- or anti-inflammatory phenomenon (Tang et al. 2010).

1.5.7 Genetic studies of sepsis

Chung and Waterer (2011) have reviewed polymorphisms identified from candidate genes for sepsis. These include genes involved in antigen recognition (MBL and TLRs), the pro-inflammatory response ($TNF\alpha$ and LTA) and the anti-inflammatory response ($IFN\gamma$ and $IL10$). $TNF\alpha$ is one of the most studied genes with respect to genetic susceptibility to disease. The SNP $TNF\alpha$ -308 has been included in many studies but on its own was not been found to be associated with increased susceptibility or outcome for CAP. However, including a second SNP ($TNF\alpha$ -238) to form a two-SNP haplotype proved a better predictor of mortality from sepsis.

The clinical significance of many of the polymorphisms identified in these genes remains unclear with many associations not replicating in additional

studies. Factors which may contribute to these inconsistent findings include heterogeneity in patient populations such as different ethnic groups, difference in the pathogenic bacteria, difference in treatment protocol, lack of statistical power from small samples sizes and different clinical end points.

1.5.8 Functional genomics of sepsis

It is now thought that the complex changes in cellular pathways that occur in sepsis cannot be represented by a single molecule. Microarray technology allows the study of genome-wide changes in gene expression in an unbiased way and the use of this technology in the context of sepsis has been reviewed by Tang et al. (2010). In total, 12 cohorts were identified which encompassed 784 individuals. Gene expression profiling was used to investigate disease profiles, severity and outcome. However, many of these studies were limited to small cohorts, represented phenotypically heterogeneous patients, and focused on paediatric rather than adult patients (Payen et al. 2008; Tang et al. 2009; Wong et al. 2011). Despite these difficulties, some similarities were identified across the cohorts. It was found that sepsis resulted in the activation of PRRs and an increase in the activity of signal transduction cascades (including NF- κ B, MAPK, JAK and STAT). In contrast, expression of traditional candidate markers of sepsis (such as TNF- α and IL1) varied considerably Tang et al. (2010).

Gene expression signatures in paediatric sepsis

Wong and colleagues have published the majority of work on paediatric sepsis. Their initial study of genome-wide gene expression identified a role for altered zinc homeostasis in poor outcome from children with septic shock (Wong et al. 2007). This was then extended to include patients with SIRS and sepsis as well as septic shock but the zinc findings were found to be specific to septic shock (Wong et al. 2009a).

Three subclasses of septic shock patients were identified through gene expression signatures. The first of these subclasses had higher illness severity, higher mortality rate and higher incidence of Gram-positive bacterial infection. These individuals also showed repression of genes corresponding to key signalling pathways of the adaptive immune system (Wong et al. 2009b). A gene expression signature of 100 genes was developed to distinguish the three subclasses. This signature was converted into a graphical expression mosaic, allowing clinicians to classify the patients based on the patterns of expression. This provided evidence for a clinically feasible stratification strategy with a quantifiable number of genes. However, the work has not yet been validated in an independent cohort of patients (Wong et al. 2010b)

Many of these initial studies were conducted on whole blood, representing a mixed population of blood cells. One of the main findings was repression of genes corresponding to the adaptive immune system. To determine whether these findings were confounded by the differential white blood cell counts, the analysis was repeated using neutrophil, monocytes and lymphocytes as representatives of leukocyte subpopulations. The differentially expressed genes identified were found to be comparable to those derived from whole blood, supporting the use of this tissue for further studies of sepsis (Wong et al. 2010a).

Wong and colleagues have provided comprehensive studies of paediatric sepsis patients, predominantly sepsis shock but many of these findings have not replicated when adults have been studied. This is illustrated by *IL8* gene which was upregulated in paediatric septic shock non-survivors compared to survivors or healthy controls (Wong et al. 2007) but had less predictive power in adults (Calfee et al. 2010).

Prediction models in sepsis

Tang et al. (2007) used a set of 50 genes to identify adult sepsis compared to critically ill patients without infection. Following isolation of RNA from neutrophils, they identified genes which were differentially expressed between the sepsis patients and controls and used this to develop a model for a training set of 44 patients. The model was applied to a validation cohort of 50 samples which had sensitivity of 0.95 (proportion of actual sepsis patients recorded as such) and predictive value of 0.91 (proportion of predicted sepsis patients who are true positives). Hierarchical clustering was performed on prediction model genes which revealed three clusters. The first cluster represented the inflammatory response which was downregulated in the sepsis patients. The second cluster represented those genes involved in immune regulation which again were inhibited in the sepsis patients. The final cluster was composed of genes involved in mitochondrial functions such as oxidative phosphorylation and ATP synthesis. These genes showed less inhibition in patients with sepsis. This study demonstrates the potential clinical utility in defining gene sets which can identify sepsis from SIRS and therefore aid treatment decisions. The limitations to this study are the single recruiting centre and the relatively small sample size. This study supports the use of genome-wide data for the investigation of sepsis compared to the traditional "one molecule, one diagnosis/intervention" method.

Tang repeated this study on peripheral blood mononuclear cells (PBMCs), identifying a 138 gene signature to distinguish sepsis from SIRS. Adjusting for variations in relative proportion of lymphocytes and monocytes did not change this list, demonstrating that gene expression from PBMCs can be just as representative for this type of study. The findings from this study complemented the work that had been carried out on neutrophils for instance the suppression of inflammatory response in the sepsis patients compared to SIRS. This may explain why anti-inflammatory therapies have failed (Tang et al. 2009).

Gene expression signatures in adult sepsis due to CAP

A study has recently been published which identifies gene expression signatures in adult patients with sepsis due to CAP (Severino et al. 2014). The study used just 10 patients with sepsis (five survivors and five non-survivors) with three healthy controls. Gene expression signatures between sepsis patients and healthy controls were identified but only modest difference in expression were observed between sepsis survivors and non-survivors (fold changes <1.7).

1.6 Influenza

A significant proportion of patients with severe sepsis due to CAP will have developed a post-influenza bacterial pneumonia (Metersky et al. 2012). The World Health Organisation estimates that seasonal influenza affects 5-15% of the European Region's 900 million population (World Health Organization 2013). In addition to the mortality and morbidity associated with seasonal influenza, there are also substantial social and economic costs which are both direct and indirect. For example, the costs associated with sickness absence and lost productivity can be 5-10 times greater than the visits to primary care physicians and hospitalisations (Ryan et al. 2006).

Influenza viruses are highly infectious RNA viruses that can cause acute respiratory illness. They affect all age groups and can recur in any individual, making them the most frequent cause of acute respiratory illness that requires medical intervention. Influenza A viruses can be divided into subtypes based on their surface glycoproteins (Cox and Subbarao 1999). There are 16 different subtypes of haemagglutinin (HA) and nine different subtypes of neuraminidase (NA). All the subtypes circulate in birds but historically, human influenza A viruses have been restricted to H1, H2 and H3 HA, and N1 and N2 NA subtypes (Epstein and Price 2010). Antigenic variation, which occurs primarily in these

glycoproteins, renders an individual susceptible to new strains, despite previous infection by other influenza viruses. Influenza vaccines usually induce high antibody titers to the highly polymorphic HA antigen and therefore must be re-formulated and re-administered annually. The production of new vaccines which provide protective immunity to highly conserved internal antigens would provide longer-lasting protection against multiple influenza subsets. One strategy for inducing immunity across influenza subtypes is through the induction of cytotoxic T cells with specificity for epitopes within conserved influenza proteins. Berthoud et al. (2011) have recently described a T cell inducing vaccine called MVA-NP+M1 and have demonstrated that it is safe and immunogenic in young adults and adults over the age of 50 (Antrobus et al. 2012). It may increase antibody responses to inactivated influenza vaccine when both vaccines are administered together (Antrobus et al. 2014).

1.7 Common variable immune deficiency disorders

Primary immunodeficiencies (PIDs) are a group of rare disorders that result from a failure in the immune system and CVID are the most frequently encountered PID in the clinic (Park et al. 2008). CVID are a group of diseases in which insufficient quantity and quality of immunoglobulin usually leads to susceptibility to recurrent bacterial infections, mainly of the respiratory and gastrointestinal tracts (Chapel and Cunningham-Rundles 2009). An immunodeficiency is recognized in most CVID patients in the second, third or fourth decade of life and the first diagnostic criteria were published in 1999 (Table 1.2) (Conley et al. 1999). These criteria have been used since then with slight variations for example, the minimum age of presentation is often increased to 4 years in order to avoid infants with immune defects which cause B cell differentiation and class-switch disorders (Chapel et al. 2008).

Table 1.2: CVID diagnosis criteria. Individuals fulfilling each of the criteria are diagnosed with CVID (Conley et al. 1999). SD = standard deviations.

Criteria for CVID diagnosis
Male or female patient >2 years of age
Serum IgG and IgA at least 2 SD below the mean for age
IgM is present or absent
Poor response to vaccines
Absent isohemagglutinins
Defined causes of hypogammaglobulinemia have been excluded
Normal or low B cell numbers

1.7.1 Clinical complications

The diagnosis of CVID is made by excluding known disorders of B cell failure for example B cell differentiation defects resulting in absent B cells, activation-induced cytidine deaminase (AID) and uracil-DNA glycosylase (UNG) deficiencies affecting B cell function and T cell switching pathways (Chapel et al. 2008). This results in a heterogeneous group of individuals with widely different clinical features and additional complications.

Within cohorts of CVID patients, beyond bacterial infections, the main clinical complications observed include autoimmunity, lymphocytic infiltration, enteropathy and malignancy (Figure 1.3). Patients with at least one clinical complication have a much poorer survival rate compared to those with an Infections Only phenotype (Chapel et al. 2008).

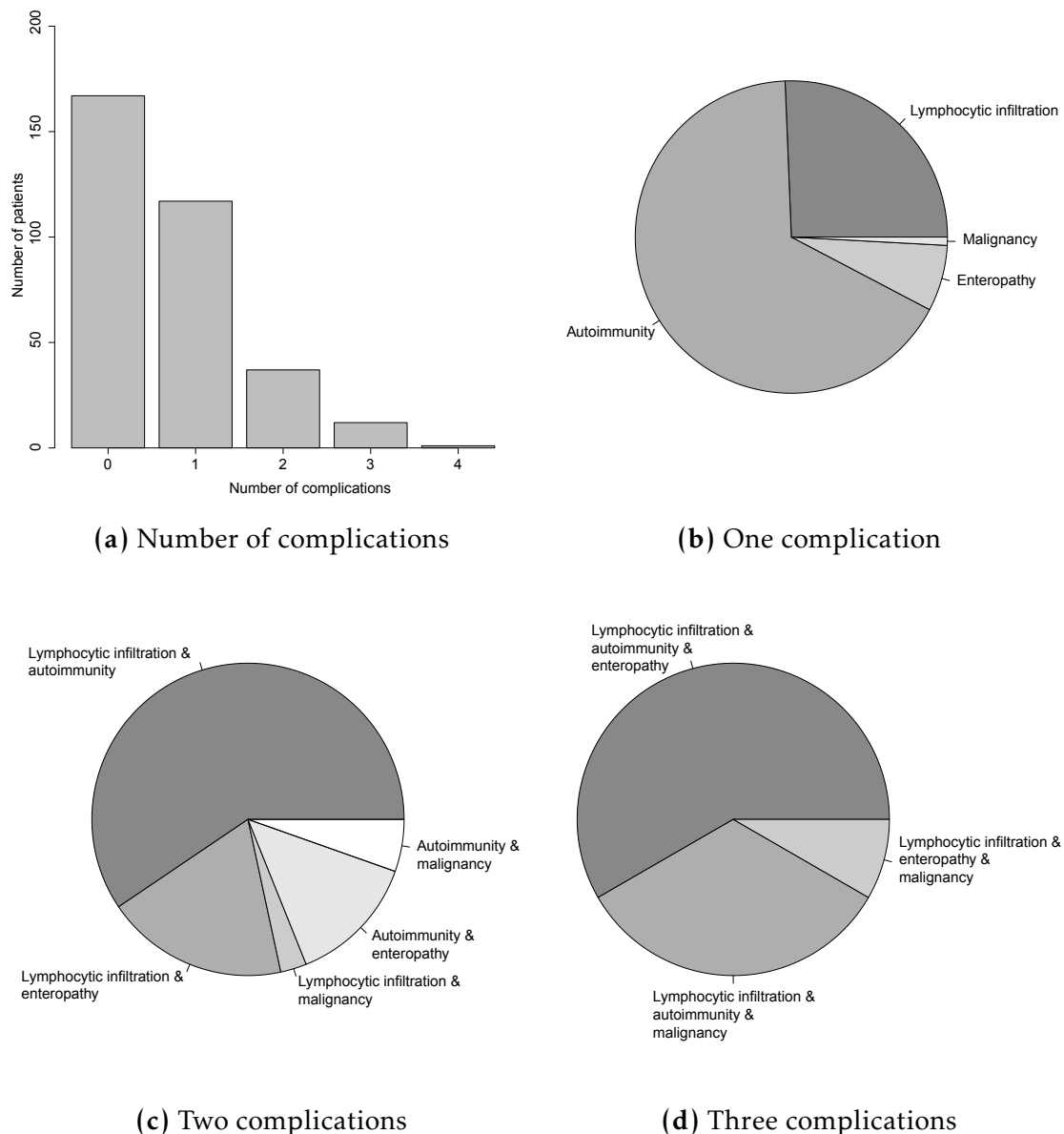


Figure 1.3: Nature and number of CVID complications. a) Number of complications observed in each individual. Individuals with no complications are referred to as infections only. b) Distribution of complications in patients with one complication. c) Distribution of complications in patients with two complications. d) Distribution of complications in patients with three complications. (Chapel et al. 2008)

The most common infection risk for patients with CVID is from extracellular bacteria, particularly those that affect the respiratory tract such as *Streptococcus pneumoniae* and *Haemophilus influenzae*. Although not common, sepsis, as a results of these bacterial infections, has been observed in a number of patients. (Cunningham-Rundles 2010).

Autoimmunity in the cohort can result from failure of checkpoints for autoreactivity during B cell development (for example clonal deletion, receptor editing or anergy) and the most common consequences are autoimmune thrombocytopenic purpura (reduced circulating platelets) and autoimmune haemolytic anaemia (caused by breakdown of red blood cells) (Park et al. 2008).

Patients with polyclonal lymphoproliferation, excessive production of lymphocytes, include those with lymphoid interstitial pneumonitis (LIP) (infiltration in the lungs), granuloma (collections of macrophages) and splenomegaly (enlargement of the spleen). These individuals are at increased risk of developing lymphoma (Chapel et al. 2012; Salzer et al. 2012).

These distinct subphenotypes within cohorts of CVID patients may be the result of different genetic variations and therefore focusing future analysis on specific subphenotypes may assist with the identification of the underlying genetic aetiology.

1.7.2 Current treatment

Intravenous immunoglobulin replacement is the standard treatment to reduce the occurrence of bacterial infections however, it does not reduce the additional complications of the disease (Cunningham-Rundles 2010).

1.7.3 Challenges of CVID

Due to recurrent clinical complications and life-long need for immunoglobulin therapy, CVID is the most common clinically significant immunodeficiency (Cunningham-Rundles and Knight 2007) but without a clear diagnostic test, diagnosis is often delayed for 6-8 years after the first characteristic symptoms (Cunningham-Rundles 2010). Repeated infections during this period of diagnostic delay can lead to irreparable organ damage. Improved diagnostics

would allow for improved intervention and reduced mortality (Orange et al. 2011).

1.7.4 Genetics of CVID

CVID is usually sporadic with familial inheritance occurring in only approximately 10-20% of cases (Cunningham-Rundles and Bodian 1999; Hammarström et al. 2000). In some of these families, linkage studies have identified possible loci for dominant CVID genes on chromosomes 4q and 16q (Chapel et al. 2008). In addition to these identified regions, mutations in a number of genes including *ICOS*, *CD19*, *TNFRSF13C* (BAFF-R) and *TNFRSF13B* (TACI) have been identified however they account for just 10% of CVID cases (Poodt et al. 2009). The proteins encoded by these genes contribute to different steps in the adaptive immune response to an antigen (Figure 1.4). These mutations illustrate that it is not yet clear whether CVID is caused by a failure of B cell maturation, immunoglobulin class switching or a failure of B cell activation due to defects in T cells.

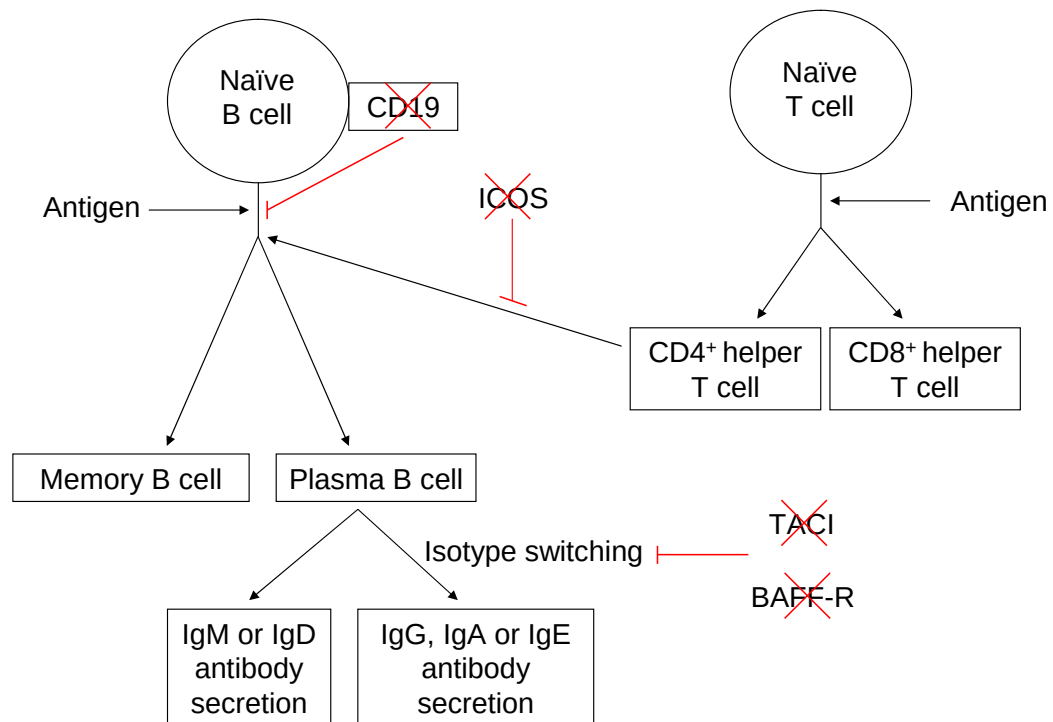


Figure 1.4: The effect of known CVID related mutations on the B cell response to antigen. After stimulation with antigen, a naïve B cell requires multiple other signals to effectively differentiate into an antibody secreting plasma B cell or memory B cell. A number of mutations in different locations of this pathway have been associated with CVID. Mutations in *ICOS*, encoding a T cell specific molecule, result in individuals with normal T cells but reduced memory B cells and plasma B cells (Grimbacher et al. 2003). Mature B cells from individuals with *CD19* mutations, a B cell surface molecule, show poor response to antigen (van Zelm et al. 2006). Mutations in *TNFRSF13C* (encoding BAFF-R) and *TNFRSF13B* (encoding TACI) impair isotype switching (Castigli et al. 2005; Warnatz et al. 2009).

ICOS

Grimbacher et al. (2003) first identified an inherited homozygous deletion in the *ICOS* gene while searching for CVID candidate proteins. They found four CVID patients from two families which lacked ICOS on the surface of activated T cells. These patients did not show any signs of T cell deficiencies but had reduced switched and memory B cells providing evidence for a critical involvement for ICOS in this process. This autosomal-recessive trait was later identified in a further five CVID patients from two more families and due to the identical

homozygous haplotypes shared between all individuals, was likely inherited from a common founder (Salzer et al. 2004).

Although the common feature of CVID is failure of B cells the variant identified in *ICOS* highlights the role that other components of the immune response may play in the pathogenesis of CVID. T cell abnormalities have been described in CVID cohorts with reduced naïve CD4 T cells being the most frequent (Bateman et al. 2012). CD4 T cells (T helper cells) are essential for B cell class switching and (Giovannetti et al. 2007) have demonstrated a correlation between reduced levels of these cells and disease severity. Continuing from this is the hypothesis that dendritic cells contribute to the disease phenotype due to their role in antigen presentation to T cells (Scott-Taylor et al. 2004).

CD19

Following flow-cytometric immunophenotyping of lymphocytes van Zelm et al. (2006) identified four CVID patients from two families with normal numbers of CD20+ B cells but reduced expression of CD19. CD19 is a positive regulator of B cell receptor signalling. Sequencing the gene revealed two homozygous variants that result in premature stop codons which caused a poor response to antigen from their mature B cells.

TNFRSF13B and TNFRSF13C

Naïve B cells express the immunoglobulins IgM and IgD. Following activation, the B cell can undergo isotype switching in order to express IgG, IgA or IgE. This involves the deletional recombination between regions upstream of the heavy chain of IgM and the genes encoding the heavy chains of IgG, IgA or IgE. This requires two signals: one from cytokines and the other from members of the TNFR family expressed on B cells for example TACI (encoded by *TNFRSF13B*) or BAFF-R (encoded by *TNFRSF13C*) (Castigli et al. 2005).

A number of heterozygous variants have been identified in *TNFRSF13B* (Castigli et al. 2005) however the frequency of these variants are not always significantly different between patients and controls (Pan-Hammarstrom et al. 2007). The variants in TACI may therefore result in predisposition to CVID rather than being causative.

Warnatz et al. (2009) identified a number of patients with reduced BAFF-R expression from a search for genetic defects affecting B-cell homeostasis. Two related individuals were revealed to have a homozygous deletion within the *TNFRSF13C* gene which resulted in reduced mature peripheral B cells.

1.7.5 Candidate gene studies

Following the identification of the variants in the genes described above, (Salzer et al. 2008) selected a further eight candidate genes for CVID (encoding APRIL, BCMA, IL10, IL10 α , IL10 β , IL21, IL21R and CCL18) based on linkage analysis for the respective gene loci. The protein coding regions of these genes were screened in families with recessive patterns of inheritance but no potentially disease causing mutations were identified.

GWAS genes

To address the underlying immunopathogenesis of CVID, Orange et al. (2011) conducted a GWAS in patients with CVID. In total, 363 patients were included along with 3,031 healthy controls. The samples were split into a discovery and validation cohort. The only association with genome-wide significance was to the MHC (rs3117426 with $p = 8.62 \times 10^{-10}$). The second most significant association, in locus 8p21.2 which harbors *ADAM28*, *ADAM7*, *ADAMDEC1* and *STC1*, did not reach genome-wide significance ($p = 6.24 \times 10^{-6}$) but was replicated in the validation cohort ($p = 0.0314$). The ADAM genes belong to a disintegrin metalloproteinase family and are involved in diverse biological processes

including immune responses. Multiple novel common and rare susceptibility loci for CVID were identified which is consistent with the heterogeneous nature of the disease. Despite this, diagnosis of CVID has not improved. The SNP association results were not powered for definitive conclusions.

1.7.6 Functional genomics of PIDs

The functional significance of mutations identified in PIDs have historically been determined through the use of a number of different genetic techniques. Fluorescence in situ hybridization (FISH) was used to identify chromosomal abnormalities for instance deletions of chromosome 22q11.2 in Di George syndrome, a disease characterised by variable features such as congenital heart disease and recurrent infections (Gong et al. 1996). Linkage mapping was used to determine the location of a disease associated allele relative to a polymorphic site of known location for example x-linked severe combined immunodeficiency (XL-SCID) was mapped to the proximal long arm of the X chromosome (Saint Basile et al. 1987) and a hemizygous mutation was later identified in the *IL2RG* gene (Puck et al. 1993).

An autosomal recessive STIM1 (a calcium sensor in the endoplasmic reticulum) deficiency in a patient with autoimmunity and recurrent infections was identified in 2010 providing a proof of principle that NGS technology (in this case WES) could be used to identify rare inborn errors of immunity in a single patient (Byun et al. 2010). This has been followed by identification of several other genetic defects in PIDs for example a gain-of-function STAT1 mutation in autosomal dominant chronic mucocutaneous candidiasis (Veerdonk et al. 2011; Liu et al. 2011).

1.8 Specific aims and objectives

1. **To investigate the functional genomic response in severe sepsis** (Chapter 3)

Sepsis is a common complex trait in which environmental and heritable factors are thought to be important. I aim to define the transcriptomic response to severe sepsis in patients with CAP for a large prospective patient cohort recruited as part of the Genomic Advances in Sepsis study (GAinS) and relate this to clinical covariates and survival in order to establish whether a functional genomic approach can be used to

- (a) resolve disease heterogeneity
- (b) provide insights into biological processes and pathways dysregulated in sepsis
- (c) improve prognostication in terms of predicting outcome

2. **To define regulatory genetic variants in severe sepsis** (Chapter 4)

Regulatory genetic variants modulating gene expression are responsible for most reported GWAS and are typically context specific. I aim to

- (a) map gene expression as a quantitative trait in peripheral blood leukocytes from patients with severe sepsis (Aim 1)
- (b) integrate this with the results of a GWAS for outcome in sepsis performed in the same patient cohort (GAinS/GenoSept)
- (c) establish the sepsis specificity of observed eQTL by comparison with data generated in other contexts including in resting state and following innate immune activation

3. To investigate the functional genomic response to influenza challenge and relate this to gene expression signatures in severe sepsis (Chapter 5)

Viral infection, notably with influenza, is commonly associated with bacterial pneumonia in patients with severe sepsis. I aim to define the transcriptomic response to influenza for healthy volunteers undergoing influenza challenge as part of a vaccine study and use this information to

- (a) define differentially expressed gene sets and pathways specific to influenza infection
- (b) relate any observed influenza signatures to data from the sepsis cohort (Aim 1) where patients have laboratory confirmed influenza infection

4. To investigate the functional genomic response in CVID and the relationship to the underlying genetic aetiology (Chapter 6)

PIDs provide the opportunity to resolve rare variants modulating the immune response, informing understanding of immunity. CVID encompasses the most common PID occurring in adults. I aim to

- (a) investigate genetic variants associated with CVID using a whole-genome sequencing approach
- (b) define the transcriptomic response in CVID compared to controls
- (c) use knowledge of differentially expressed genes in CVID to inform identification of functional genetic variants
- (d) relate disease associated genes/variants in CVID to sepsis (Aim 2)

Chapter 2

Materials and Methods

2.1 Sepsis study

2.1.1 GAINs study patient recruitment and exclusion criteria

Patients with severe sepsis were recruited as part of the GAINs study (<http://www.ukccg-gains.org>). After obtaining informed consent, adult patients (>18 years of age) admitted to ICU with severe sepsis due to CAP were enrolled by research nurses. CAP was defined as febrile illness and radiological features of pneumonia acquired in the community or within two days of hospital admission. Exclusion criteria included treatment with novel therapies, pregnancy, immunocompromised patients and neutropenia (not as a result of sepsis). Recruitment began in 2005 and 34 sites have been involved in collecting serial samples (Figure 2.1).



Figure 2.1: Location of GAINs study recruitment sites. 34 sites located in 22 cities/towns were involved in recruiting severe sepsis patients for the GAINs study.

2.1.2 Sample collection

Serial samples for RNA were obtained on days 1, 3 and/or 5 after admission by collecting 5 ml blood into Vacuette EDTA tubes (Becton Dickinson). The total blood leukocyte population was isolated using the LeukoLOCK filter system (Ambion). Blood was passed across a leukocyte enrichment filter using a vacutainer system. The filtered leukocytes were stabilized with RNALater. The filters were stored and transported from the sites to the WTCHG (Oxford) at -80°C. In addition buffy coat and whole blood were collected for DNA extraction.

2.1.3 RNA extraction

Purified RNA, depleted of globin mRNA, was extracted from the LeukoLOCK filters using the Total RNA Isolation Protocol (Ambion). The contents of the filter were lysed and eluted with a guanidine thiocyanate-based solution. Cellular

proteins and DNA were degraded using Proteinase K and DNase I respectively. The RNA was then purified using magnetic bead technology. Spectrophotometry (Nanodrop 2000; Thermo Scientific) was used to quantitate the RNA yield and the quality of a small subset was determined by on-chip electrophoresis (Biorad Bioanalyzer; Agilent).

2.1.4 Microarray data processing and analysis

High quality RNA was selected for 303 sepsis patients at the first available time point. Genome-wide gene expression analysis was carried out on 1000 ng RNA using the Illumina Human-HT-12 v4 Expression BeadChip gene expression platform comprising 47,231 probes (performed at microarray core facility, Wellcome Trust Sanger Institute, Cambridge). The report for analysis was generated by Illumina's Genomestudio software. Data backgrounds were subtracted and probes that did not have a detection value ≥ 0.95 in at least 5% of samples were removed. The raw data were transformed and normalised using the Variance Stabilization and Normalization method (Huber et al. 2002). QC checks including principal component analysis (PCA) to identify batch and array effects were carried out using R (R Core Team 2013).

2.1.5 Genomic DNA extraction

DNA from 91 patients in the cohort had previously been extracted from the buffy layer using the Qiagen DNA extraction protocol. DNA from the remaining patients was extracted from the buffy layer, or when not available, whole blood, using the automated Maxwell 16 Blood purification kit (Promega) or the QIAamp Blood Midi kit protocol (Qiagen). The Maxwell machine uses paramagnetic particles to purify the DNA after cell lysis while the QIAamp kit uses a spin column. The DNA yield was determined by fluorescence using the Quant-iT PicoGreen kit (Invitrogen).

2.1.6 Genotyping

295 samples were submitted for genome-wide genotyping. Genotyping was performed for 730,525 SNPs using the Illumina HumanOmniExpress BeadChip. PLINK (Purcell et al. 2007) was used for the genotyping QC. Sample QC included discordant sex information, proportion of missing genotypes (>0.02), heterozygosity rate, identity by descent (π hat >0.1875) and multi-dimensional scaling with 1000 Genomes populations. SNP QC included SNP missing data proportion (>0.02), MAF (<0.01) and HWE ($<1 \times 10^{-10}$).

2.1.7 Imputation

The genotypes were pre-phased with SHAPEIT2 (Delaneau et al. 2012). Missing genotypes and untyped SNPs were imputed using Impute2 (Howie et al. 2009; Howie et al. 2011) in 5Mb chunks against the 1000 Genomes Project Phase I v3 reference panel (1000 Genomes Project Consortium et al. 2012). SNPTEST was used to obtain summary statistics for each chunk (Marchini et al. 2007). SNPs with an info score <0.9 , MAF <0.05 or HWE <0.001 were removed from further analysis. This left 5,380,404 SNPs. The genotype probabilities were converted into dosage estimates and formatted for use with Matrix eQTL. Imputation was performed genome-wide but the eQTL fine mapping was restricted to regions which showed evidence of a genotyped *cis* eQTL.

2.1.8 eQTL QC

Genome-wide gene expression data had been generated for 303 patients in the sepsis cohort. QC and normalisation were carried out on all probes and for all individuals as described in Section 2.1.4. From the cohort, 240 patients had good quality genotyping and gene expression data available. Further QC steps were employed at the probe level before conducting the eQTL analysis. Probes with

sequences that mapped to more than one genomic location using BLAST or had a SNP present at a MAF of at least 1% in the European populations of the 1000 Genomes Project were excluded to reduce potential confounding effects.

2.1.9 Statistical analysis

Clinical covariates were compared with t-tests for continuous data and Chi-squared tests for count data. Differential gene expression analysis was carried out using the R package limma (Smyth 2004). Biological pathways and functions enrichment were generated through the use of Qiagen's IPA (www.qiagen.com/ingenuity, QIAGEN Redwood City). The GeneRave R package (CSIRO Bioinformatics R package version 3.0.8) was used to identify clinical covariates or genes for use in prediction models. Misclassification rates (MR) for models were determined using leave-one-out cross-validation (R Core Team 2013).

2.1.10 eQTL mapping

The Matrix eQTL R package (Shabalin 2012) was used to determine SNP-probe associations using an additive linear model. eQTL were deemed *cis* when the distance between the SNP chromosomal position and probe start site was less than 1Mb. Distances greater than this were defined as *trans*. PCA was performed on the gene expression matrix. The first 30 PCs (excluding those associated with genetic variants) were included.

2.2 Influenza study

2.2.1 Study subjects and clinical trial procedures

Twenty-two healthy volunteers, aged 18-45 years, were enrolled and 11 were vaccinated with 1.5×10^8 plaque forming units (pfu) of MVA-NP+M1. Thirty days

Materials and Methods

after vaccination, all 22 volunteers underwent intranasal challenge with H3N2 influenza (A/Wisconsin/67/2005) at a dose of 1 ml of $10^{5.25}$ TCID₅₀/ml in a quarantine facility. Volunteers were all challenged within a 2 hour period. All subjects had haemagglutination inhibition (HI) titres of <1:10 to the challenge virus on admission to the quarantine unit. Further details of the subjects and methods are described in Lillie et al (Lillie et al. 2012).

Following viral challenge, volunteers had a daily physical examination by a trial physician and self-reported their symptoms using a modified Jackson scoring system (Jackson et al. 1958; Zaas et al. 2009) twice daily, from 12 hours post challenge, for 6 days. This system categorises upper respiratory and systemic symptoms on a scale of 0-3 corresponding to: no symptoms, just noticeable, bothersome but can still do activities and bothersome and cannot do daily activities. Scores were summed over the duration of the challenge period to give a total symptom score. Those volunteers with scores of ≥ 4 and positive viral culture from nasal wash samples were categorised as having laboratory confirmed influenza (LCI). Severity of infection was graded as mild if the summed symptom score was 4-28, with scores of ≥ 29 classified as moderate/severe. Volunteers were kept in quarantine until day 6 post challenge and all had a negative rapid antigen test on a nasal wash sample prior to discharge. Seroconversion was determined by HI titre using serum samples obtained at 26 days post-challenge.

Whole blood was collected in PAXgene tubes prior to influenza challenge (day 30) and then at three further time points post-challenge (12, 24 and 48 hours). All nasal washing and blood collections for viral shedding and gene expression analysis were taken in the same order in which the volunteers were challenged.

2.2.2 RNA extraction and globin mRNA depletion

Total RNA was extracted from whole blood using the PAXgene Blood RNA Kit (Qiagen) according to the manufacturer's instructions. The abundance of globin mRNA present in whole blood reduces the sensitivity of expression array data, therefore globin mRNA depletion was carried out using the GLOBINclear™ Human kit (Ambion) following the manufacturer's instructions. Briefly, total RNA from whole blood is mixed with a biotinylated capture oligo mix specific for human globin mRNA. Streptavidin magnetic beads capture the globin mRNA. Placing the tube against a magnet then allows the globin mRNA depleted RNA to be transferred to a fresh tube. Globin depleted RNA is further purified using a rapid, magnetic bead-based clean up procedure. Globin mRNA depleted samples were quantified using a NanoDrop spectrophotometer and the quality of a subset was verified using a Bioanalyzer following the manufacturer's instructions (Bioanalyzer RNA 6000 Nano kit, Agilent).

2.2.3 Microarray data processing and analysis

Genome-wide gene expression analysis was carried out on 500 ng globin depleted RNA using the Illumina Human-HT-12 v4 Expression BeadChip gene expression platform comprising 47,231 probes (performed at Core Genomics facility, WTCHG). The report for analysis was generated by Illumina's GenomeStudio version 1.6.0. Data backgrounds were subtracted and probes that did not have a detection value ≥ 0.95 in at least one sample were removed. The raw data were transformed and normalised using the Variance Stabilization and Normalization method (Huber et al. 2002). QC checks including PCA to check for batch and array effects were carried out using R (R Core Team 2013).

2.2.4 Statistical analysis

Differential gene expression analysis was carried out on 36,073 probes using three different R packages: limma (Smyth 2004), maSigPro (Conesa and Nueda R package version 1.28.0) and timecourse (Tai 2007). Biological pathway enrichment for probes differentially expressed between the four moderate/severe LCI and the remaining samples, 48 hours post-challenge was determined through Qiagen's IPA (www.qiagen.com/ingenuity, QIAGEN Redwood City). The pamr R package (Hastie et al. 2013) was used to identify a small subset of genes that could accurately discriminate moderate/severe LCI from the remaining samples. The Huang et al. (2011) dataset were used for validation. The data were accessed through NCBI GEO database (Edgar et al. 2002) (accession GSE30550) and had been normalised using the robust multi-array (RMA) method.

2.3 CVID study

2.3.1 Patient recruitment

Patients were recruited through the Clinical Immunology service (Oxford) by Helen Chapel and colleagues. The cohort contains a grandmother and grandson pair and the remaining patients are unrelated. All patients were diagnosed with CVID between two and 52 years of age following the diagnostic criteria described in Table 1.2. Genomic DNA, peripheral blood cells, clinical and laboratory data have been collected and stored for each individual.

2.3.2 Whole genome sequencing

Genomic DNA from 35 CVID patients was quantitated by fluorescence using the Qubit Fluorometer (Invitrogen) and the quality assessed by running <500 ng on a 1% TAE agarose gel for 1 hour at 70 volts. WGS was performed on 3.5-7.5 ng

DNA (Core Genomics, WTCHG). Briefly, the genomic DNA was fragmented, end-paired, A-tailed and adapter-ligated before size selection and amplification for a multiplexed library preparation. The libraries were then paired end sequenced and the data aligned to the reference.

2.3.3 Mapping and variant calling

The samples were sequenced to an average coverage greater than 25x. They were mapped with Stampy (Lunter and Goodson 2011) and the SNPs and insertions or deletions (InDels) called with Platypus (Rimmer et al. 2014) (Core Bioinformatics, WTCHG). The resulting VCF files were analysed through the use of Qiagen's Ingenuity[®] Variant Analysis[™] software (www.qiagen.com/ingenuity, QIAGEN Redwood City).

2.3.4 Variant filtering

The following filters were applied to prioritise high quality calls for rare, non-synonymous and likely pathogenic variants (Table 2.1). Variants were further filtered depending on the objective of each individual analysis.

Table 2.1: Filtering criteria for WGS variants.

Filter	Thresholds	Rationale
Confidence	Call quality ≥ 20 in any sample Variant passed upstream pipeline filtering Read depth ≥ 10 in any sample Variant is outside top 0.2% most exonically variable 100 base windows in healthy public genomes	Exclude poor quality variant calls
No. reverse reads	≥ 1 reads containing variant	Reduce strand bias
No. forward reads	≥ 1 reads containing variant	Reduce strand bias
Common variants	Allele frequency $< 5\%$ in 1000 Genomes Project	Exclude common variants
WGS500 MAF	Allele frequency $< 5\%$ in remaining WGS500 samples	Exclude technical artefacts
Predicted deleterious	Pathogenic/likely pathogenic or Associated with gain of function or Associated with loss of function	Exclude variants which are unlikely to have effect on gene function

2.3.5 Sanger sequencing

Primers were designed with Primer3 to amplify the region containing the identified BTK variant (forward: GCATATCAGTGAGGGAATGG, reverse: ATCGGCAGAAAACGCTAGAA). 25ng DNA was amplified using a 25µl PCR reaction containing 2.5µl 10x Buffer, 1.5mM MgCl₂, 0.2mM dNTP, 250nM primer (each direction) and 0.5U Platinum Taq. Cycling conditions were an initial denaturation at 94°C for 2 minutes followed by 30 cycles of 94°C for 30 seconds, 56°C for 30 seconds and 72°C for 40 seconds.

4µl of the PCR product was cleaned using 0.1µl ExoI and 0.1µl CIP with 0.4µl NEB buffer 3 in a 8µl reaction. The reaction was incubated at 37°C for 30 minutes before inactivation at 95°C for 5 minutes.

3.5µl of template DNA from ExoCIP clean up was used in the sequencing PCR reaction. The reaction was set up using 0.25µl BigDye with 2µl BigDye Sequencing Buffer and 125nM forward or reverse primer in a total volume of 10µl. Cycling conditions were an initial denaturation at 96°C for 1 minute followed by 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes.

2.3.6 RNA-seq

RNA-seq data were generated for three CVID patients and three healthy controls as a pilot analysis. PBMCs were isolated using Lymphoprep (STEMCELL Technologies) and B-cells were enriched using CD19 microbeads (Miltenyi Biotec) according to the manufacturers instructions. RNA was extracted using the RNeasy Mini Kit (Qiagen) following the manufacturers instructions. RNA-seq was performed on 600-1000ng RNA (Core Genomics, WTCHG). Briefly, the mRNA fraction was selected from the total RNA before conversion to cDNA. dUTP was incorporated from second strand cDNA synthesis. The cDNA was end-repaired, A-tailed and adapter-ligated. Prior to amplification, the

samples underwent uridine digestion. The prepared libraries were size selected, multiplexed and QC'ed before end sequencing over one lane of a flow cell. Data were aligned to the reference and reads mapped to genes using HTSeq (Anders et al. 2014) (Core Bioinformatics, WTCHG).

QC to determine the proportion of reads mapping to genes and variation within the dataset was carried out using R (R Core Team 2013). The genes with mapped reads were filtered so only those with at least 10 reads in at least three samples were retained prior to normalisation. The data were normalised using the Trimmed Mean of M values (TMM) method (edgeR, Robinson et al. 2010).

2.3.7 Statistical analysis

Differential gene expression analysis was carried out on 16,546 genes using the edgeR R package (Robinson et al. 2010). Biological pathway enrichment was generated through the use of Qiagen's IPA (www.qiagen.com/ingenuity, QIAGEN Redwood City).

Chapter 3

Functional genomic response in severe sepsis

3.1 Introduction

Sepsis, the systemic inflammatory response to the presence of an infection, (Bone, 1992) remains a major area of unmet clinical need. Despite optimal antibiotic therapy and intensive care support, mortality rates have remained at an estimated 28-50% (Angus et al. 2001; Padkin et al. 2003; Sands et al. 1997; Zeni et al. 1997). New insights into disease pathophysiology and the development of novel interventions, as well as opportunities for early effective intervention and patient stratification are urgently required. The aim of this chapter is to investigate whether gene expression signatures could help yield these new insights in a cohort of patients with severe sepsis.

3.1.1 Community acquired pneumonia

While sepsis can be recognised clinically, there is no reliable diagnostic marker. The disease can be caused by a multitude of different organisms and patients can have different anatomic sites of infection (Carlet et al. 2008). This is likely to cause heterogeneity in study cohorts and may explain some of the lack of progress in improving mortality. Studying clinically and microbially defined groups may help to reduce this heterogeneity. Since the 1980s, Gram-positive

bacteria have been the most common causative organisms of sepsis (Martin et al. 2003), particularly in the community (Brun-Buisson et al. 1995), and the most common site for infection is the respiratory tract (Danai and Martin 2005). To focus on a common cause of severe sepsis, this chapter will use a cohort of patients with CAP admitted to intensive care units (ICUs) and recruited as part of the Genomic Advances in Sepsis (GAINs, www.ukccg-gains.org) study (Materials and Methods).

3.1.2 Severity scores

A number of different severity scores can be calculated for patients admitted to ICU. For the severe sepsis cohort, two severity scores were calculated: one which assesses the presence and severity of organ dysfunction and one which assesses disease severity and can be used to predict outcome. The SOFA score, which represents organ function and rate of failure, is calculated for each day of ICU stay and has a range 0-24. The APACHE II score, which represents acute physiology and chronic health evaluation, is calculated within the first 24 hours of ICU admission and has a range 0-71 (Vincent and Moreno 2010). The clinical covariates used to calculate the scores are shown in Table 3.1.

Table 3.1: Severity scores. Clinical covariates used to calculate two severity scores currently used in ICUs.

(a) SOFA score	
Clinical covariate (system)	
PAO ₂ /FIO ₂ (respiratory)	
Glasgow coma scale (neurological)	
Mean arterial pressure or administration of vasopressors (cardiovascular)	
Bilirubin (hepatic)	
Platelets (coagulation)	
Creatinine (renal)	

(b) APACHE II score
Clinical covariate
Age
Temperature
Mean arterial pressure
Arterial pH
Heart rate
Respiratory rate
Sodium
Potassium
Creatinine
Haematocrit
White cell count
Glasgow coma scale

3.1.3 Day of sampling

The precise onset of sepsis is difficult to pinpoint because individuals will be admitted to ICU at different stages of disease progression. In order to study early gene expression signatures of disease, samples were collected within 5 days of ICU admission. Day of sampling could complicate interpretation of analysis if gene expression changed over time. Previous studies in the lab showed samples from the same patients, taken at different time points, were highly correlated ($r^2=0.97$) and the day of sampling was not a major component of variance in the data set (Radhakrishnan 2012). As a day 1 sample for one patient may not be equivalent to day 1 for another, the first available time point for each patient was selected to build a large enough cohort for a well powered genetic study.

3.1.4 Tissue of interest and RNA extraction method

Whole blood and leukocyte fractions have been used as the tissue of interest in most sepsis studies because blood is a highly relevant tissue for the study of dysregulated immune response, and it is often impractical to take biopsies from the source of infection in critically ill patients (Maslove and Wong 2014). In

addition, tissue damage often occurs at different locations to the site of infection (Bosmann and Ward 2013) making it difficult to determine which is the most suitable tissue to study. It is still unclear whether or not findings from blood correlate with expression patterns from immune cells resident in other tissues, such as alveolar macrophages (Fitting et al. 2004; Ge et al. 1997), but this is beyond the scope of the study presented in this chapter.

Gene expression can be dependent on both the tissue sampled and the constituent cell types (Liu et al. 2008; Cobb et al. 2005). Ideally, the cell type of interest would be isolated in order to reduce confounding factors. However when recruiting patients for a multi-center study, logistics must also be evaluated when deciding which samples to collect for RNA extraction and what method to employ. Extraction of RNA from blood is a very convenient method but the abundance of globin mRNA in whole blood can reduce the sensitivity of expression microarray data. Feezor et al. (2004) also demonstrated that gene expression profiles from whole blood showed increased noise compared to isolated leukocytes. In contrast, cell separation requires a greater starting volume of blood and more complicated processing. This may affect the cell physiology and gene expression profiles particularly when carried out for multiple individuals at multiple recruitment centers (Feezor et al. 2004; Debey et al. 2004). Studies have not concluded which cell type is most informative for the study of sepsis and therefore a number of different cell fractions would be required. This would require so much additional work that cohort numbers would be low and not powered for the type of genetic analysis described in the following two chapters.

For this study, LeukoLOCK filters provided a compromise between the use of whole blood and specific cell types. These filters employ a leukocyte-depletion technology to isolate the total blood leukocyte population (Figure 3.1) without the need for further globin reduction steps. Minimal equipment is

required allowing multiple sites to recruit suitable numbers of patients in a timely manner. In addition, processing can occur rapidly at the bedside which preserves the expression profile. This is important as delays between collection and stabilisation of samples can result in changes to the gene expression profile (Debey et al. 2004; Tanner et al. 2002). The RNA is stable at -80°C for months without any effect on the expression profile (Cashion et al. 2011) easing transport logistics.

Wong et al. (2010a) have evaluated the correlation between gene expression signatures observed in whole blood and leukocyte subsets in paediatric sepsis compared to healthy controls. The pathway analysis for the differentially expressed genes correlated well suggesting whole blood can still be informative particularly for a global overview of the disease.

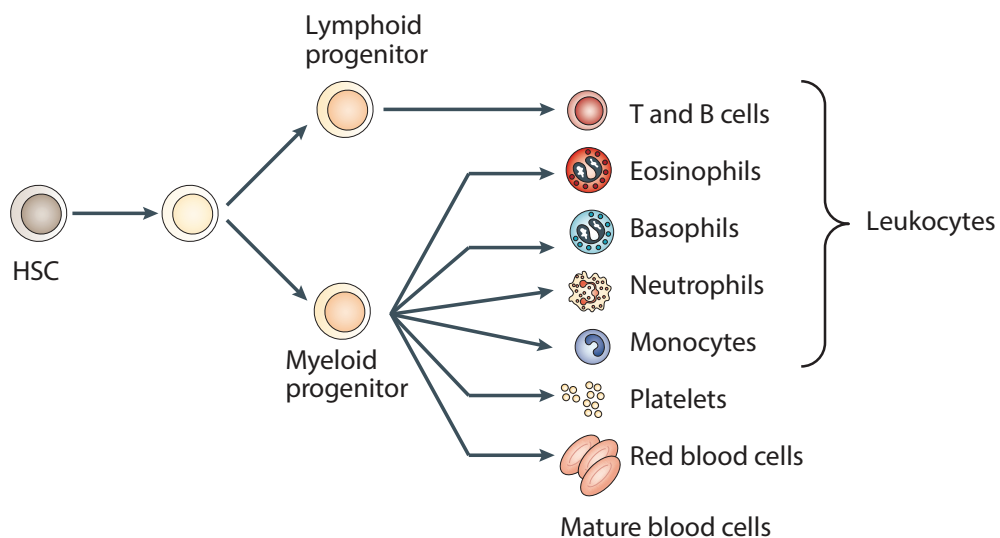


Figure 3.1: Leukocyte fraction of whole blood. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Cancer (Corey et al. 2007), copyright 2007. Haematopoietic stem cells (HSCs) give rise to lymphoid and myeloid progenitor cells. These proliferate and differentiate into the mature blood cells. The leukocyte fraction of whole blood includes the T and B cells from the lymphoid progenitor. The eosinophils, basophils, neutrophils and monocytes differentiate from the myeloid progenitor.

3.1.5 Use of controls

Previously described sepsis studies have used different cohorts for control comparisons, for example healthy individuals or SIRS patients without infections (Tang et al. 2007; Severino et al. 2014). The choice of control cohort is very dependent on the precise question being answered. They can be very helpful for defining gene expression signatures that differentiate between sepsis patients and other diseases or healthy individuals but it is often difficult to generate a control group that provides an appropriate comparison. For example, differences in age, sex or underlying co-morbidities could all lead to confounding results. The study presented in this chapter has therefore focused on identifying gene expression signatures within the sepsis cohort, for example between survivors and non-survivors, or other subtypes, to reduce this issue.

3.1.6 Microbiology

The question of which invading organisms are responsible for infection during sepsis remains a challenge for clinicians. This is particularly important with the current increase in antibiotic resistance (World Health Organization 2014a) as accurate organism identification would allow appropriate selection of antimicrobial agents. While Gram-positive bacteria have been the most common causative organisms of sepsis, a significant proportion of the patients with severe sepsis due to CAP will have developed a post influenza bacterial pneumonia (Metersky et al. 2012). Microbial cultures are commonly used for identification but the yield can be low for instance because of prior antibiotic therapy and the lag time can be long (up to 48 hours) (Llewelyn et al. 2001; Tang et al. 2007).

Comparing the host response between Gram-positive and Gram-negative is interesting for a number of reasons. Firstly, at the clinical level, the initial antimicrobial therapy is dependent on the identification of the bacteria (Tang et al. 2008). Secondly, Gram-positive sepsis is thought to have a different

pathogenesis to Gram-negative bacteria and prevalence is increasing (Martin et al. 2003). Finally, evidence suggests that different types of sepsis differ at the molecular level for example the importance of different TLR (Takeuchi and Akira 2001).

3.1.7 Prediction models

Prediction models can have prognostic value in the clinical setting. Many methods, for example generalized linear models or multiclass logistic regression, are not able to deal with a situation where the number of observations is much less than the number of potential explanatory variables. This is the case in genome-wide gene expression studies where the number of genes analysed far outnumbers the number of samples. The number of variables can be reduced, for instance by focusing on genes which are differentially expressed for a certain comparison but this can still result in a large number of variables with limited ways of reducing without sacrificing potentially informative variables. GeneRave is an R package (CSIRO Bioinformatics R package version 3.0.8) that employs a method focused on sparsity and therefore allows the elimination of redundant explanatory variables. In order to identify genes which discriminate between different phenotypes, a multiclass logistic regression model is fitted. This results in a model that only includes a few genes and is therefore more easily interpreted.

3.1.8 Unsupervised approach to gene expression

Differential gene expression analysis is usually performed to identify genes which have changed in abundance between experimental conditions. While informative, this type of analysis requires a knowledge of the contrasts of interest and can therefore result in a biased approach. To understand unanticipated structure in a dataset clustering methods can be employed. These methods use gene expression data to define clusters in the cohort such that there is

good separation between clusters but individuals within each cluster are similar. Different methods can be employed to determine the appropriate number of clusters and which individuals should be assigned to which cluster. The FactoMineR R package (Husson et al. 2013) uses hierarchical clustering to initiate the number of clusters followed by K-means to determine cluster membership. Given a particular number of clusters, the K-means function finds a partition of the samples that minimizes the total within-group sum of squares over all variables. Deciding on the "optimal" number of groups is often difficult and there is no method that can be recommended in all circumstances. One approach is to plot the within-group sum of squares for each partition given by applying the K-mean procedure and look for the point in the resulting curve at which increasing the number of clusters does not reduce the within-group sum of means (Everitt and Hothorn 2011).

3.1.9 Aims

The aim of this chapter is to investigate the functional genomic response in severe sepsis. Sepsis is a common complex trait in which environmental and heritable factors are thought to be important. I aim to define the transcriptomic response to severe sepsis in patients with CAP for a large prospective patient cohort (GAinS/GenoSept) and relate this to clinical covariates and survival in order to establish whether a functional genomic approach can be used to

1. resolve disease heterogeneity
2. provide insights into biological processes and pathways dysregulated in sepsis
3. improve prognostication in terms of predicting outcome

3.2 Results

3.2.1 Sepsis cohort

Data are presented for a cohort of patients of European ancestry with severe sepsis due to CAP recruited from 34 ICUs as part of the GAINs study. Recruitment has been ongoing from 2005 and as of September 2014, 659 patients with CAP have been recruited. Serial samples were obtained on days 1, 3 and/or 5 after admission to ICU. These included total blood leukocyte population (isolated using the LeukoLock filter system) for RNA, buffy coat for DNA, plasma and urine. In addition, extensive clinical phenotyping information was recorded including survival up to six months after ICU admission, microbiological data and treatments administered.

Genome-wide gene expression data were generated for 303 patients. Following QC, 265 patients remained and results for this cohort will be presented in this chapter.

3.2.2 Patient characteristics

The sepsis cohort demographics and clinical covariates are summarised in Table 3.2 and Figure 3.2. The cohort has a range of ages from 18 to 89 but the majority are elderly with a slightly greater proportion of males. The greatest proportion of deaths occur early after ICU admission. A single bacterial infection was identified in 27% of patients and a viral infection in 9%.

Table 3.2: Sepsis cohort demographics and clinical covariates.

(a) Demographics

Patient characteristic	Value
Mean age ($\pm$ SD)	62 years ($\pm$ 16)
Male	55%
Median APACHE II score (range)	18 (6-42)
Median SOFA score (range)	6 (0-21)

(b) Clinical covariates

Clinical covariate	No. patients (%)
14 day mortality	40 (15)
28 day mortality	56 (21)
6 month mortality	78 (29)
Corticosteroid treatment	71 (27)
Activated protein C treatment	20 (8)

(c) Microbiology

Microbiology	No. patients (%)
Gram+ bacterial infection	47 (18)
<i>Streptococcus pneumoniae</i> infection	33 (12)
Gram- bacterial infection	25 (9)
Viral infection	25 (9)
H1N1 viral infection	16 (6)

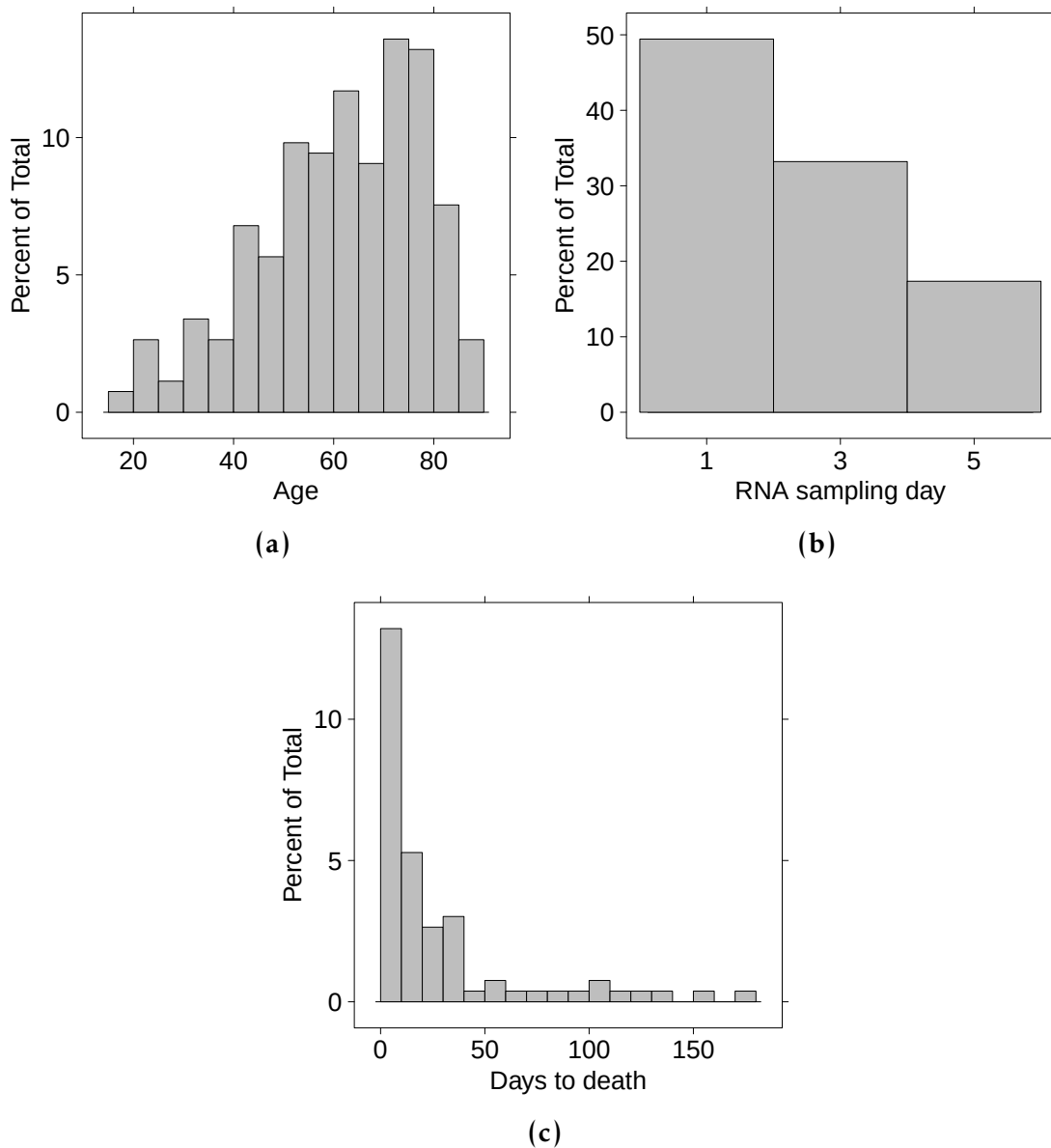


Figure 3.2: Sepsis cohort characteristics. a) Age of patients at ICU admission. b) Day of sampling used for analysis. c) Survival lengths for non-survivors (up to six months).

3.2.3 Associations between clinical covariates and survival

Survival from severe sepsis has been classically defined at 28 days from ICU admission (Vincent 2004). However, to generate a more comprehensive understanding of associations between clinical covariates and survival, 14 day (early) and 6 month (late) survival were also investigated. Associations between continuous clinical covariates and survival (Table 3.3) and between count clinical

covariates and survival (Table 3.4) were determined. Similar clinical covariates were associated with the different survival lengths within the cohort, for example increased age and severity score were associated with non-survival as expected. Examples of covariates associated with specific survival lengths include higher heart rate associating with non-survival within 14 days, higher creatinine levels with non-survival within 28 days and a greater number of renal support days with non-survival within six months. A full list of clinical covariates measured during the study are listed in Table A.1.

Table 3.3: Associations between continuous clinical covariates and survival. Associations between continuous clinical covariates and different lengths of survival time were determined by *t test*. The tables show covariates with a p value <0.05.

(a) 14 day survival

Covariate	p value	Survivors (mean value)	Non-survivors (mean value)
Age	2.70E-05	60.4	69.4
Lymphocytes (raw)	1.58E-02	1.04	0.79
Polynucleocytes (proportion)	2.78E-02	0.84	0.87
APACHE II	2.38E-03	17.8	21.8
SOFA (sampling day)	1.39E-04	6.0	8.9
SOFA (maximum recorded)	4.11E-04	7.5	9.9
Bicarbonate	3.45E-02	26.1	23.5
FIO ₂	1.49E-03	0.49	0.63
Haematocrit	2.76E-03	35.5	31.7
Heart rate (low)	3.90E-02	78.6	84.6
Heart rate (high)	1.14E-02	110.7	122.8
Mean arterial pressure (low)	5.66E-03	68.3	62.6
Systolic blood pressure (low)	1.72E-03	100.7	89.9

Functional genomic response in severe sepsis

(b) 28 day survival

Covariate	p value	Survivors (mean value)	Non-survivors (mean value)
Age	1.48E-06	59.7	69.5
Mononucleocytes (raw)	1.37E-02	0.87	0.68
Lymphocytes (raw)	1.23E-03	1.06	0.77
Mononucleocytes (proportion)	3.53E-02	0.07	0.06
APACHE II	1.49E-03	17.7	21.1
SOFA (sampling day)	1.59E-06	5.8	8.8
SOFA (maximum recorded)	2.20E-05	7.3	9.7
Bicarbonate	1.20E-02	26.2	23.8
FIO ₂	1.35E-02	0.49	0.58
Haematocrit	1.92E-03	35.6	32.4
Mean arterial pressure (low)	4.24E-03	68.4	63.7
Mean arterial pressure (high)	4.40E-02	101.5	96.3
Systolic blood pressure (low)	3.47E-03	100.8	92.6
Creatinine (low)	4.70E-02	110.9	144.6

(c) 6 month survival

Covariate	p value	Survivors (mean value)	Non-survivors (mean value)
Age	5.02E-10	58.3	70.2
Mononucleocytes (raw)	3.41E-02	0.88	0.71
Lymphocytes (raw)	4.10E-06	1.11	0.74
Polynucleocytes (proportion)	1.83E-02	0.84	0.87
Lymphocytes (proportion)	2.25E-02	0.09	0.07
APACHE II	8.76E-07	17.1	21.6
SOFA (sampling day)	2.31E-06	5.7	8.2
SOFA (maximum recorded)	5.60E-06	7.2	9.4
Bicarbonate	2.14E-02	26.3	24.3
FIO ₂	7.66E-03	0.49	0.57
Haematocrit	5.23E-05	36.1	32.1
Mean arterial pressure (low)	1.83E-04	69.1	63.4
Mean arterial pressure (high)	1.75E-02	102.0	96.6
Systolic blood pressure (low)	2.23E-02	100.8	94.9
Renal support (number of days)	2.96E-02	2.6	4.4

Table 3.4: Associations between count clinical covariates and survival. Associations between count clinical covariates and different lengths of survival time were determined by Chi-squared test. The tables show covariates with a p value <0.05.

(a) 14 day survival

Covariate	p value
Renal failure	4.45E-02
Inotropes	4.26E-04

(b) 28 day survival

Covariate	p value
Renal failure	1.36E-02
Inotropes	5.69E-05

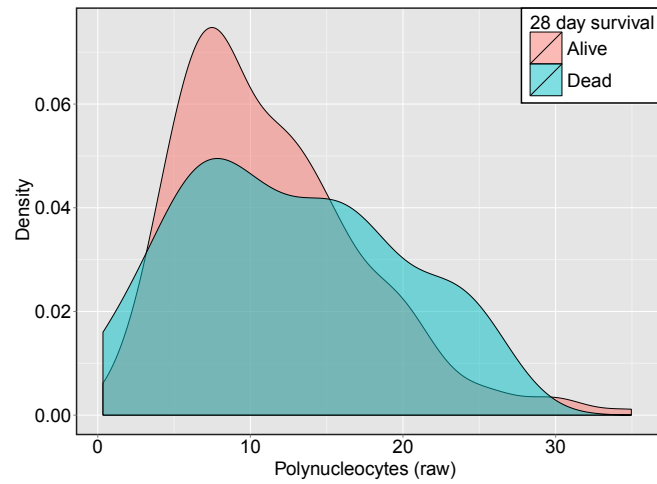
(c) 6 month survival

Covariate	p value
Inotropes	3.20E-03

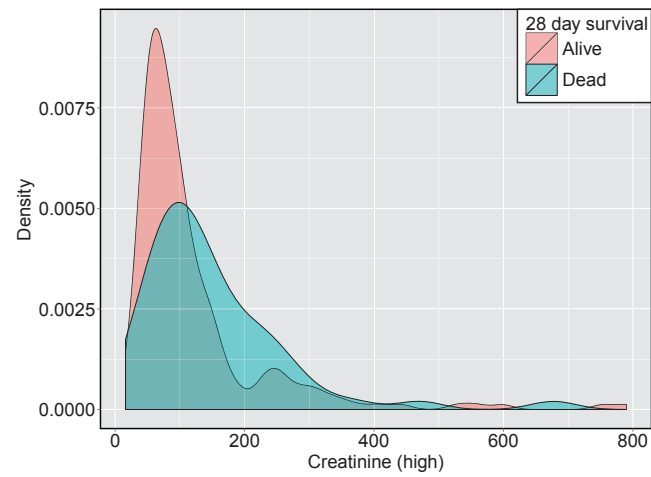
3.2.4 Non-linear relationship between clinical covariates and survival

It was hypothesised that for some clinical covariates, divergence from an optimum value in either direction could have an adverse effect on survival. Clinical covariates that did not show evidence of linear association with survival were further investigated for a non-linear relationship with survival. These covariates were adjusted to give the distance between the original value and the 28 day survivor median value. Associations between the adjusted values and 28 day survival were determined by Mann-Whitney-Wilcoxon test. The polynucleocyte cell count, highest recorded creatinine and white cell count all

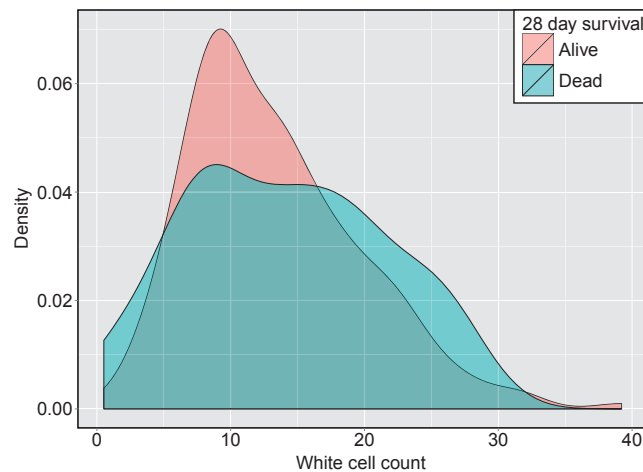
showed evidence of a non-linear relationship (Figure 3.3). The finding for white cell count is particularly interesting as it may explain why other studies such as Silvestre et al. (2009), where a linear relationship was assumed, have not found associations with survival.



(a)



(b)

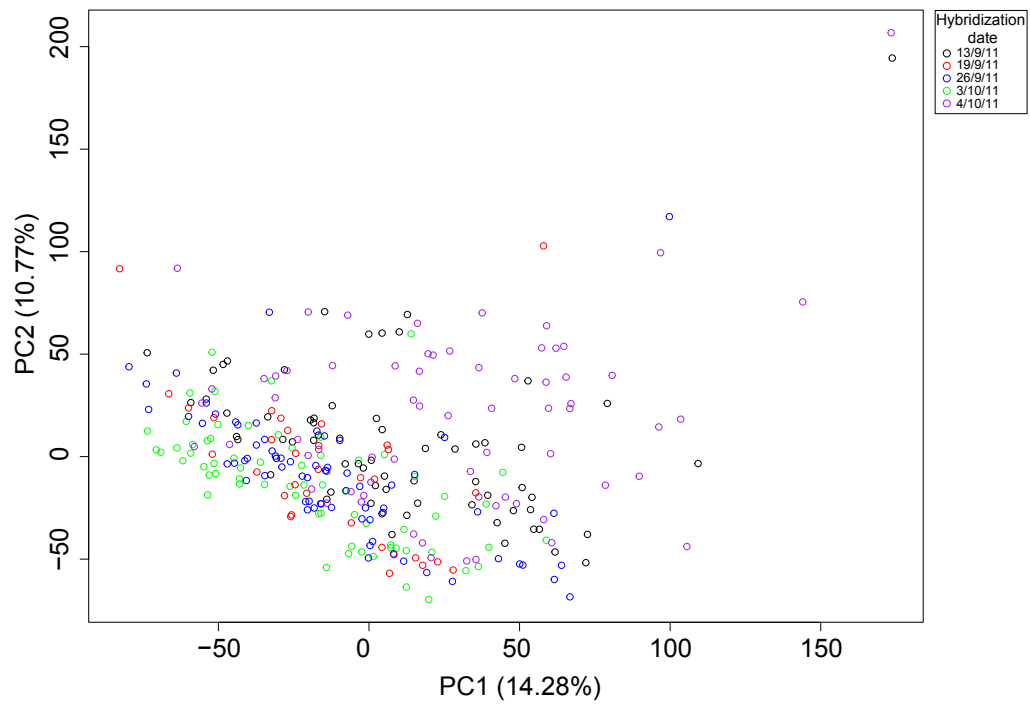


(c)

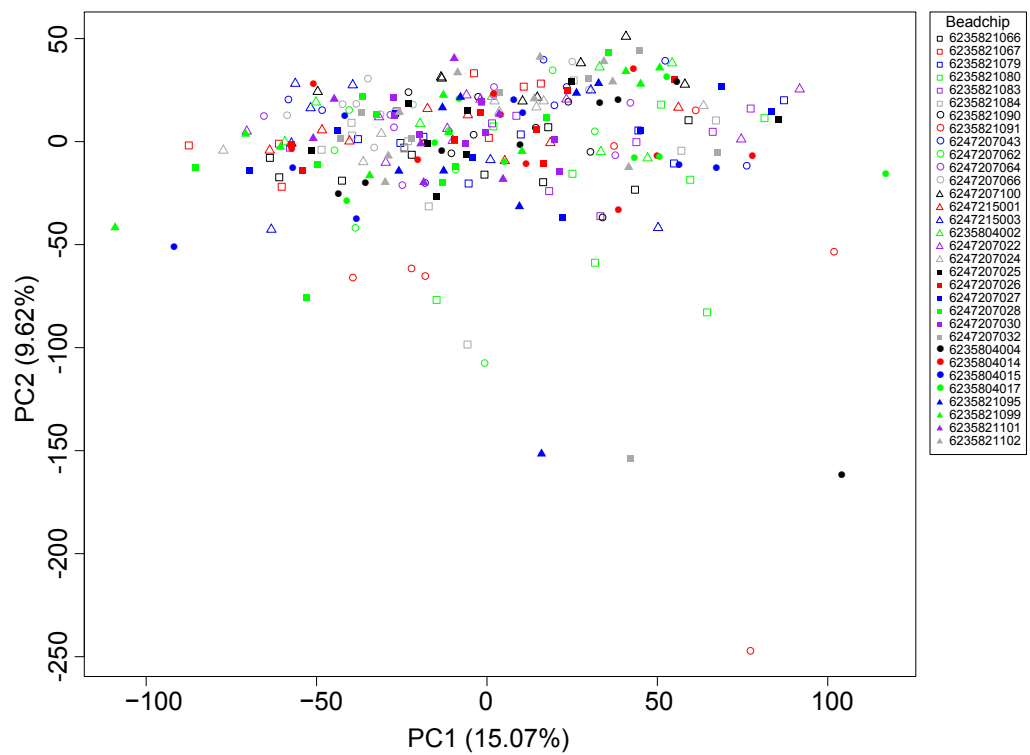
Figure 3.3: Density plots for non-linear relationship with survival. Associations between adjusted covariates and 28 day survival were determined by Mann-Whitney-Wilcoxon test. Three covariates showed evidence for a non-linear relationship as illustrated by the density plots. a) Raw polynucleocyte count on day of sampling ($p = 1.73\text{E-}02$). b) Highest creatinine level recorded on day of sampling ($p = 1.83\text{E-}02$). c) White cell count on day of sampling ($p = 1.62\text{E-}02$).

3.2.5 RNA extraction and gene expression microarray QC

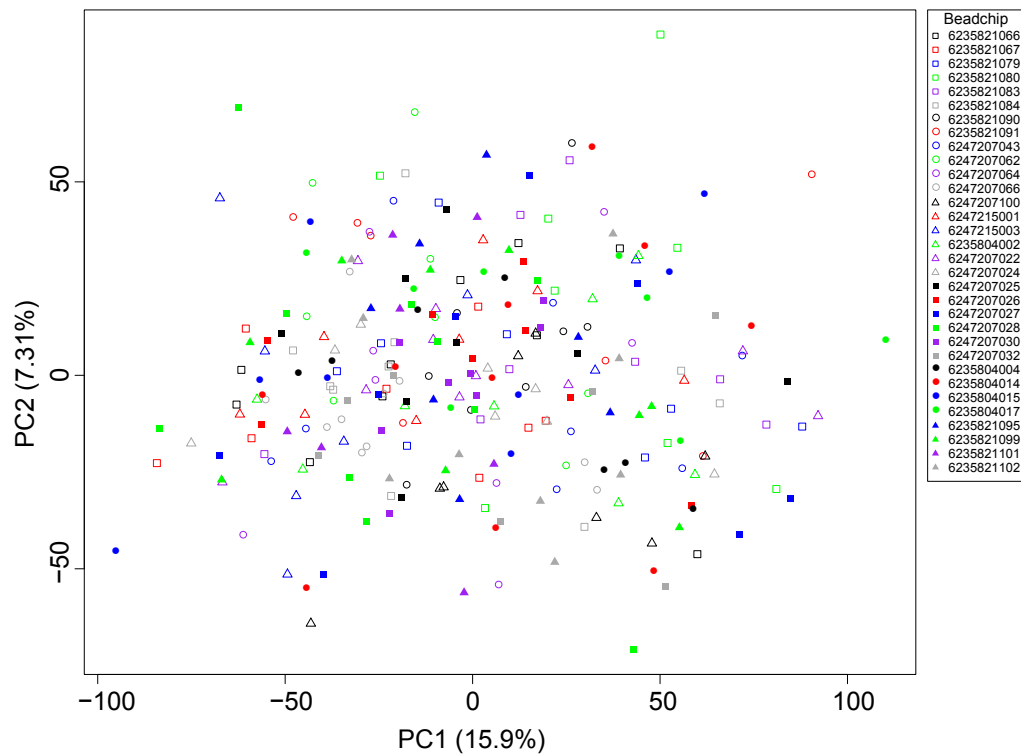
For each patient, RNA was extracted from the first available time point after ICU admission. Genome-wide gene expression data were generated for 303 patients with high quality RNA (average RNA integrity number (RIN) of 8) using the Illumina Human-HT-12 v4 Expression BeadChip. Data backgrounds were subtracted and the raw data were transformed and normalised using the Variance Stabilization and Normalization method (Huber et al. 2002). Following QC, 33 patients were removed due to batch effects and five patients were identified as outliers (Figure 3.4). 21,046 probes were removed as they did not have a detection p value ≥ 0.95 in at least 5% of samples.



(a) 303 samples



(b) 270 samples



(c) 265 samples

Figure 3.4: Principal component analysis. The first two principal components (PCs) are plotted with the proportion of variance explained by each component. a) When all 303 samples are included in the analysis, a batch effect between four chips and the remaining chips is identified due to a hybridization problem. b) The removal of four chips (33 samples) removes the batch effect. Five samples are identified as outliers. c) The removal of the outlying samples reduces the variation in the dataset.

3.2.6 Differential gene expression for survival

Differential gene expression was analysed for 26,185 probes using a linear regression model with an empirical Bayes approach to define the likelihood of observed differences in gene expression given overall variance in the dataset. Table 3.5 provides a summary of implementing this approach using the limma package (Smyth, 2004) for different lengths of survival time (14 days, 28 days and 6 months after ICU admission). The highest number of probes differentially expressed between survivors and non-survivors was seen for survival at 14 days post ICU admission (5.6% of probes). Differential expression decreased

as survival length increased (1.1% probes for 28 day survival and 0.6% for 6 month survival). For the 14 day and 28 day comparisons, the number of differentially expressed probes was further increased when the analysis was restricted to non-survivors vs patients that survived beyond 6 months (6.4% and 1.8% of probes respectively). For each comparison, approximately 75% of the differentially expressed probes were upregulated in survivors compared to non-survivors (Figure 3.5). The 50 most significant differentially expressed probes for each comparison compared to 6 month survivors are shown in Tables A.2, A.3 and A.4.

Figure 3.6 illustrates an example of a probe that is differentially expressed between 14 day non-survivors and 6 month survivors with the greatest fold change. The probe corresponds to the gene *CD8A*, which encodes the alpha chain of the CD8 cell surface glycoprotein found on cytotoxic T lymphocytes (Pruitt et al. 2014).

Table 3.5: Differentially expressed probes for survival. The number of probes with false discovery rate (FDR) <0.05 and fold change >1.2 for each comparison.

Comparison (no. patients)	Total no. patients	No. probes
14 day non-survivors (40) vs survivors (225)	265	1465
14 day non-survivors (40) vs 6 month survivors (187)	227	1685
28 day non-survivors (56) vs survivors (209)	265	285
28 day non-survivors (56) vs 6 month survivors (187)	243	466
6 month non-survivors (78) vs survivors (187)	265	161

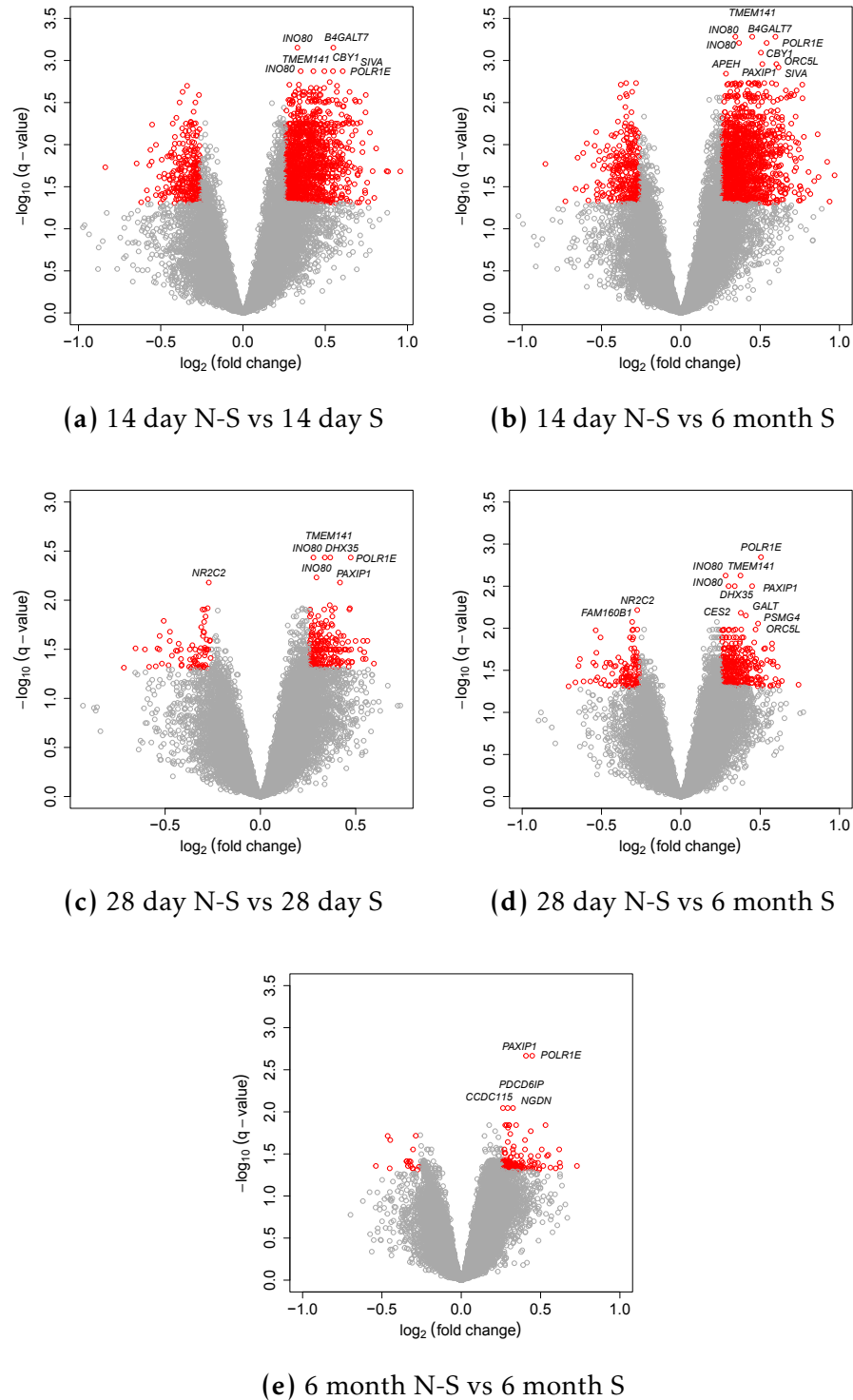


Figure 3.5: Volcano plot of differentially expressed probes for survival. The significance of differential expression between survivors (S) and non-survivors (N-S) is plotted against the fold change. Each point represents a single probe. Red points highlight probes with $FDR < 0.05$ and fold change > 1.2 . A positive fold change indicates upregulation in survivors. The most differentially expressed probes are labelled with their corresponding gene names. a) 14 day non-survivors vs survivors. b) 14 day non-survivors vs 6 month survivors. c) 28 day non-survivors vs survivors. d) 28 day non-survivors vs 6 month survivors. e) 6 month non-survivors vs survivors.

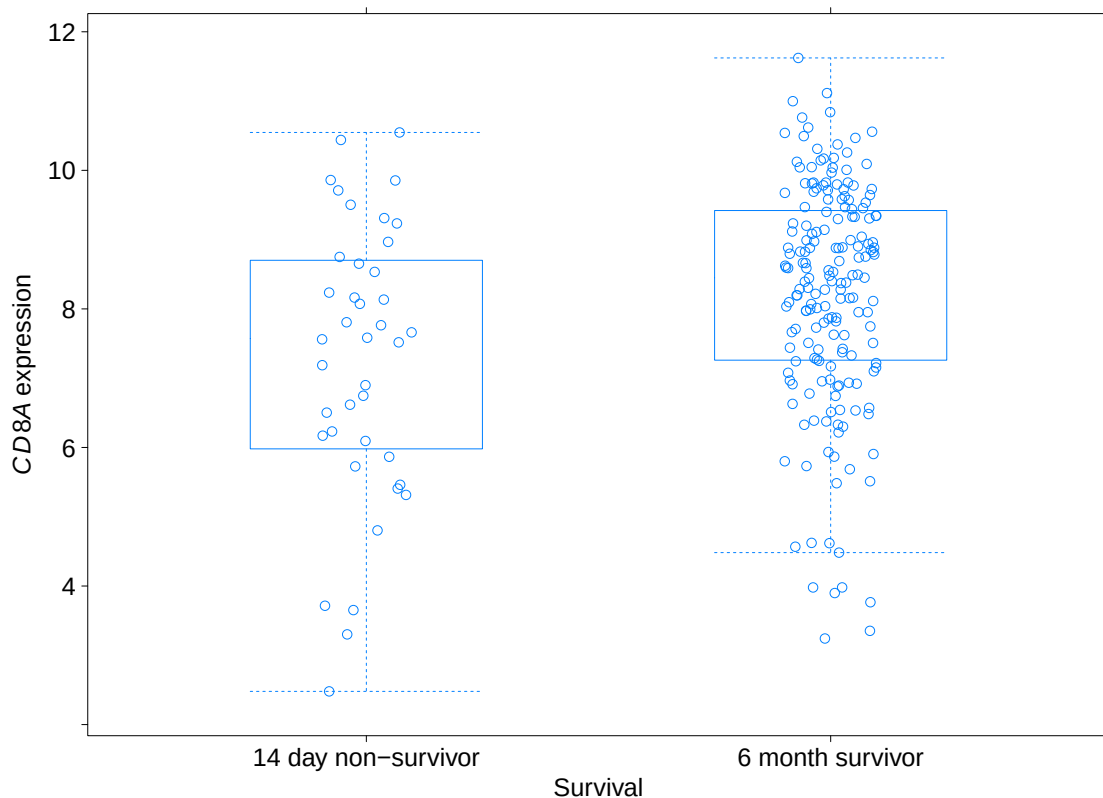


Figure 3.6: *CD8A* differential expression. *CD8A* is an example of a gene which was differentially expressed between 14 day non-survivors and 6 month survivors. The probe corresponding to this gene had the greatest fold change ($\log_2(\text{fold change}) = 0.97$) for this comparison.

3.2.7 Pathway analysis for survival

IPA was used to assess the functional significance of the differentially expressed probes from each comparison. The number of differentially expressed probes that map to genes in IPA and were therefore used for the analysis are shown in Table 3.6. The ten most significant pathways for each comparison were used to create the heatmap in Figure 3.7. The full list of significant pathways are shown in Tables A.5, A.6 and A.7. The 14 day survival comparison generated the greatest number of enriched pathways, which is likely due to the increased number of differentially expressed genes at this time point. The most striking result from this analysis is the EIF2 signalling pathway. This is the most significant pathway for 14 day non-survivors vs 6 month survivors (p value = $3.98\text{E-}21$) with 51 genes

Functional genomic response in severe sepsis

upregulated in the dataset out of the 201 in the pathway. This pathway is still significant but much less so at 28 days (p value = 2.17E-5) and six months (p value = 2.14E-2).

The disease and functions analysis in IPA shows that there is upregulation of genes involved in infectious disease, in particular viral infection (p value = 4.11E-4), in the 6 month survivors compared to the 14 day non-survivors. This is not observed in the comparison to 28 day non-survivors or the 6 month non-survivors. The proportion of patients with confirmed viral infections is the same in 14 day non-survivors and 6 month survivors.

Table 3.6: IPA mapped genes. The number of differentially expressed probes with FDR <0.05 and fold change >1.2 that map to genes in IPA.

Comparison	No. probes	No. genes
14 day non-survivors vs 6 month survivors	1,394	1,252
28 day non-survivors vs 6 month survivors	395	371
6 month non-survivors vs 6 month survivors	137	128

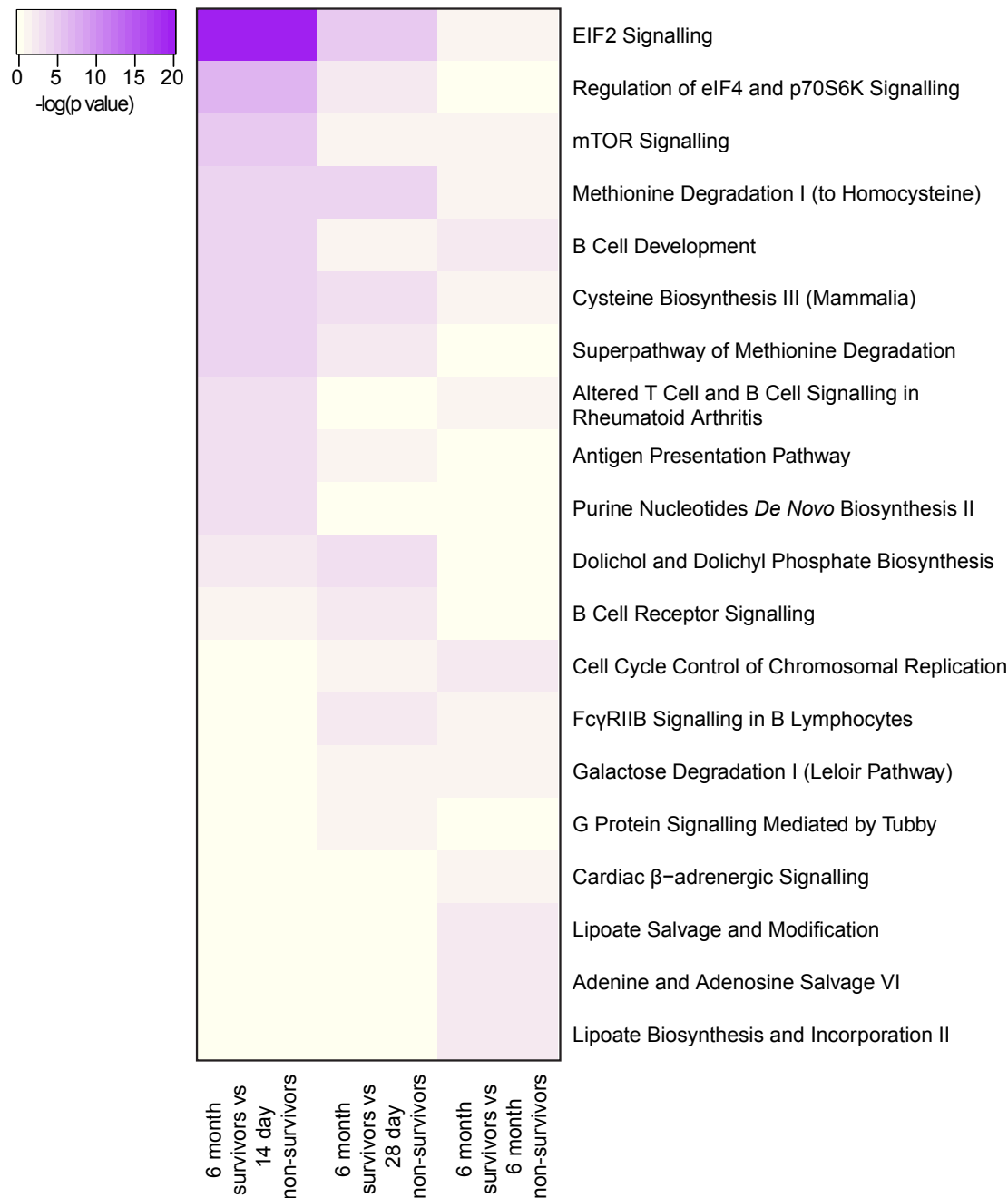


Figure 3.7: Pathway enrichment for survival. The ten most enriched pathways for each comparison are shown. The colour represents the significance of the enrichment ($-\log(p \text{ value})$).

3.2.8 Prediction models for survival

Prediction models can have prognostic value in the clinical setting. Subsets of clinical covariates and genes were selected for prediction models using the

GeneRave R package (CSIRO Bioinformatics R package version 3.0.8). This package employs a sparse regression method that allows sensible estimation of parameters when the number of predictors exceeds the number of observations.

Models were generated for survival (14 day, 28 day and 6 month). To reduce the number of possible predictors, only probes that were differentially expressed between non-survivors and 6 month survivors (FDR <0.05 and fold change >1.2), mapped to genes in IPA and were moderate to highly expressed ($\log_2(\text{probe expression}) > 6.5$) in a proportion of samples which equated to the smallest group of a comparison were used. Table 3.7 shows the results and lists the covariates and genes identified for each model. The MR for each model was determined using leave-one-out cross-validation (R Core Team 2013). Figures 3.8 and 3.9 illustrate the receiver-operator characteristic (ROC) curve for each model (ROCR R package, Sing et al. 2005).

The currently used ICU severity scores (SOFA and APACHE II) have similar predictive powers for all three survival lengths (14 day, 28 day and 6 month) (Figures 3.8a and 3.8b). The covariate only model is most successful for 6 month survival prediction while the gene expression model is most successful for 14 day survival (Figures 3.8c and 3.8d).

For each survival time point, the covariate model provides improved predictive power compared to the SOFA score. For 14 day survival, the gene expression model is most successful and is only marginally improved when the clinical covariates are added (Figure 3.9a). Similarly, the gene expression model for 28 day survival performs best but the area under the curve (AUC) is improved from 0.86 to 0.92 when the clinical covariates are also included (Figure 3.9b). Finally, for 6 month survival, the covariate model performs better than the gene expression model and but combining the two increases the AUC from 0.74 to 0.83 (Figure 3.9c). These results suggest that gene expression is most informative for patients who do not survive long after the onset of sepsis.

In comparison, the clinical covariates become more informative as the length between the sepsis insult and the survival time increases. This further highlights that gene expression is potentially informative for prediction of events within a short period of time from sampling while deaths that occur later may be more confounded by additional complications and therefore the clinical covariates are more helpful.

Table 3.7: Prediction models for survival. The GeneRave R package (CSIRO Bioinformatics R package version 3.0.8) identified the most informative clinical covariates or gene expression signatures for survival prediction. The MR was determined for clinical covariates only, gene expression only and the combination of covariates and expression using leave-one-out cross-validation.

(a) 14 day survival

SOFA MR (%)	APACHE II MR (%)	Covariates selected	Covariate MR (%)	Genes selected	Gene MR (%)	Covariate and gene MR (%)
15.5	14.7	Age FIO2 Haematocrit Heart rate (low) Systolic blood pressure (low)	14.0	<i>B4GALT7</i> <i>CHCHD10</i> <i>CCDC71</i> <i>SLC39A8</i> <i>RIPK2</i> <i>FHL2</i> <i>MRPL11</i> <i>RPL7</i> <i>PPARG</i> <i>THADA</i>	10.9	9.8

Functional genomic response in severe sepsis

(b) 28 day survival

SOFA MR (%)	APACHE II MR (%)	Covariates selected	Covariate MR (%)	Genes selected	Gene MR (%)	Covariate and gene MR (%)
21.9	20.4	Age Mononucleocytes (raw) Haematocrit Heart rate (low) Inotropes	21.1	<i>INO80</i> <i>TMEM141</i> <i>NR2C2</i> <i>DNAJA1</i> <i>MAGOH</i> <i>RDH14</i> <i>C19orf60</i> <i>CLEC1B</i> <i>PWP1</i> <i>NIPSNAP1</i> <i>S100A12</i> <i>MIB2</i>	16.2	13.2

(c) 6 month non-survivors

SOFA MR (%)	APACHE II MR (%)	Covariates selected	Covariate MR (%)	Genes selected	Gene MR (%)	Covariate and gene MR (%)
29.8	30.2	Age Lymphocytes (raw) Microbiology summary FIO2 Haematocrit Sodium (high) Inotropes	23.8	<i>PDCD6IP</i> <i>CD79B</i> <i>ZNHIT3</i> <i>GALT</i> <i>TFPI</i>	25.3	20.4

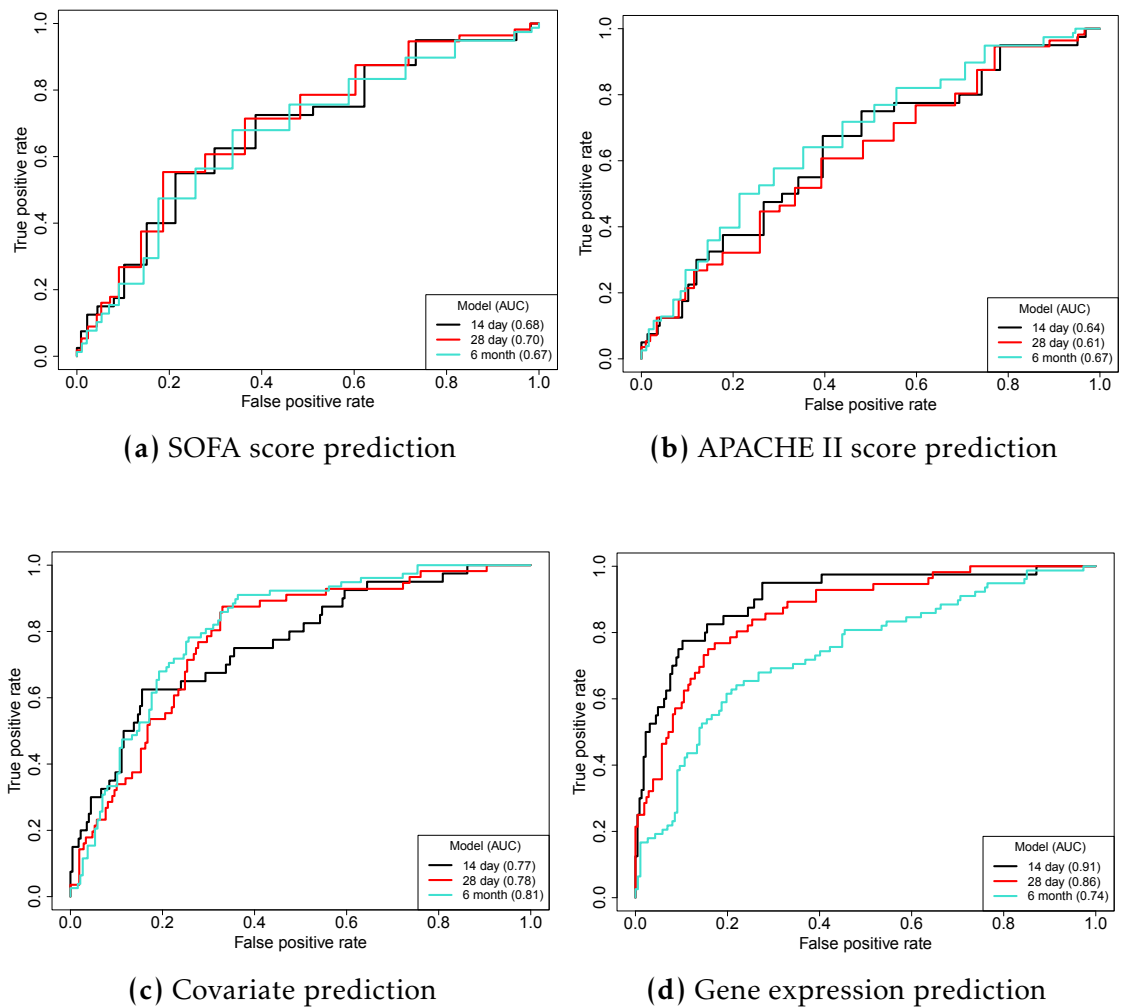


Figure 3.8: ROC curves for survival prediction models. A subset of clinical covariates and genes were identified to predict different lengths of survival. The performance for each model is illustrated by the ROC curve and AUC. a) SOFA score prediction for survival rates. b) APACHE II score prediction for survival rates. c) Covariate model prediction for survival rates. d) Gene expression model prediction for survival rates.

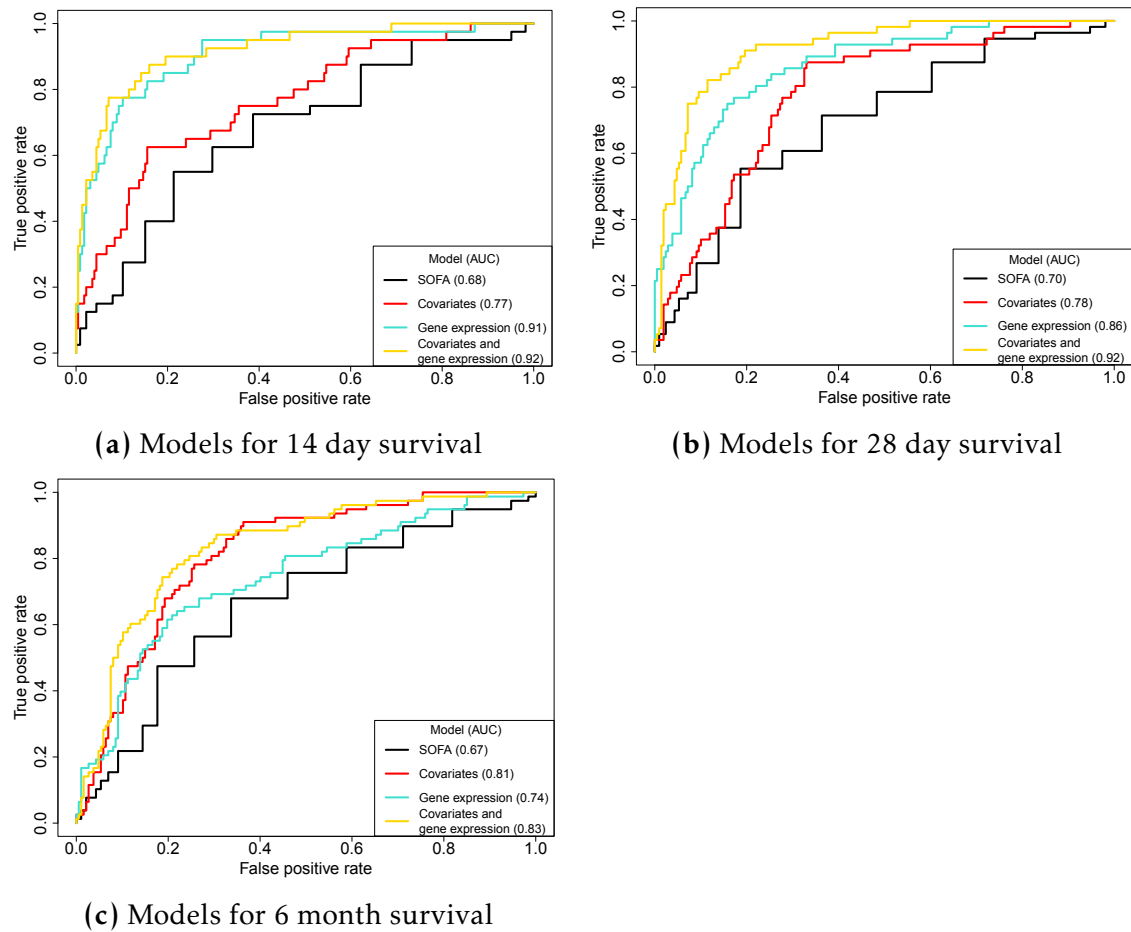


Figure 3.9: Summary of ROC curves for each survival length. A subset of clinical covariates and genes were identified to predict different lengths of survival. The performance for each model is illustrated by the ROC curve and AUC. a) Summary of models for 14 day survival. b) Summary of models for 28 day survival. c) Summary of models for 6 month survival.

3.2.9 Differential gene expression for microbiology

To determine whether or not the presence of different microorganisms have an effect on the host response during severe sepsis, differential gene expression was determined between subsets of patients. Viral infections were identified in 25 patients and of these, 16 had an H1N1 influenza A virus (this is of particular interest for the influenza challenge study described in Chapter 5). 66 probes were differentially expressed between the viral infection and non-viral infection patients (FDR <0.05 and fold change >1.2). There were no significantly

differentially expressed probes between the viral infections with confirmed H1N1 influenza and those without.

A bacterial infection was detected in 77 patients in the cohort. Comparing a confirmed viral infection to a confirmed bacterial infection (removing two patients with co-infection) identified two differentially expressed probes. Both of these probes were differentially expressed in the initial viral comparison so presumably this smaller number of probes is due to the reduced size of the cohort used for the comparison and therefore reduced power.

A single bacterial species was detected in 72 of the patients in the cohort. To determine whether the host response varies depending on whether a Gram-positive or Gram-negative infection is responsible, differential expression was determined between the 47 Gram-positive patients and 25 Gram-negative patients. No probes reached the threshold of differential expression for this comparison.

The results of these comparisons are summarised in Table 3.8 and the volcano plot for comparisons with significantly differentially expressed probes are shown in Figure 3.10. The 50 most significant differentially expressed probes for the initial viral comparison are summarised in Table A.8

Table 3.8: Differentially expressed probes for microbiology. The number of probes with FDR <0.05 and fold change >1.2 for each comparison. One patient was removed from the viral infection comparison due to lack of microbiology data. Two patients were removed from the viral vs bacterial infection comparison due to co-infection.

Comparison (no. patients)	Total no. patients	No. probes
Viral infection (25) vs no viral infection (239)	264	66
H1N1 infection (16) vs other viral infection (9)	25	0
Viral infection (23) vs bacterial infection (75)	98	2
Gram+ bacterial infection (47) vs Gram- (25)	72	0

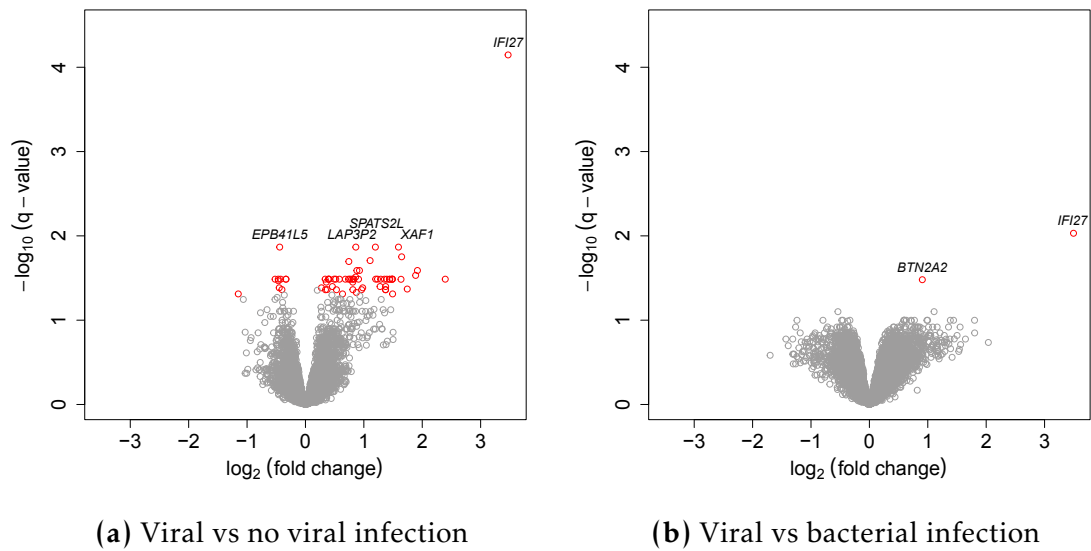


Figure 3.10: Volcano plot of differentially expressed probes for microbiology. The significance of differential expression between clinical covariates is plotted against the fold change. Each point represents a single probe. Red points highlight probes with FDR < 0.05 and fold change > 1.2 . The most differentially expressed probes are labelled with their corresponding gene names. a) Viral vs no viral infection. b) Viral vs bacterial infection.

3.2.10 Pathway analysis for viral infection

The 66 probes identified as being differentially expressed between viral infection compared to no viral infection map to 54 genes in IPA. These were used to determine pathway enrichment (Figure 3.11). The pathways involving the role of PRRs in recognition of viruses and the activation of interferon-regulatory factors (IRF) by cytosolic PRRs are of particular interest as the IRF family of TFs have been shown to play a role in antiviral response (Honda and Taniguchi 2006).

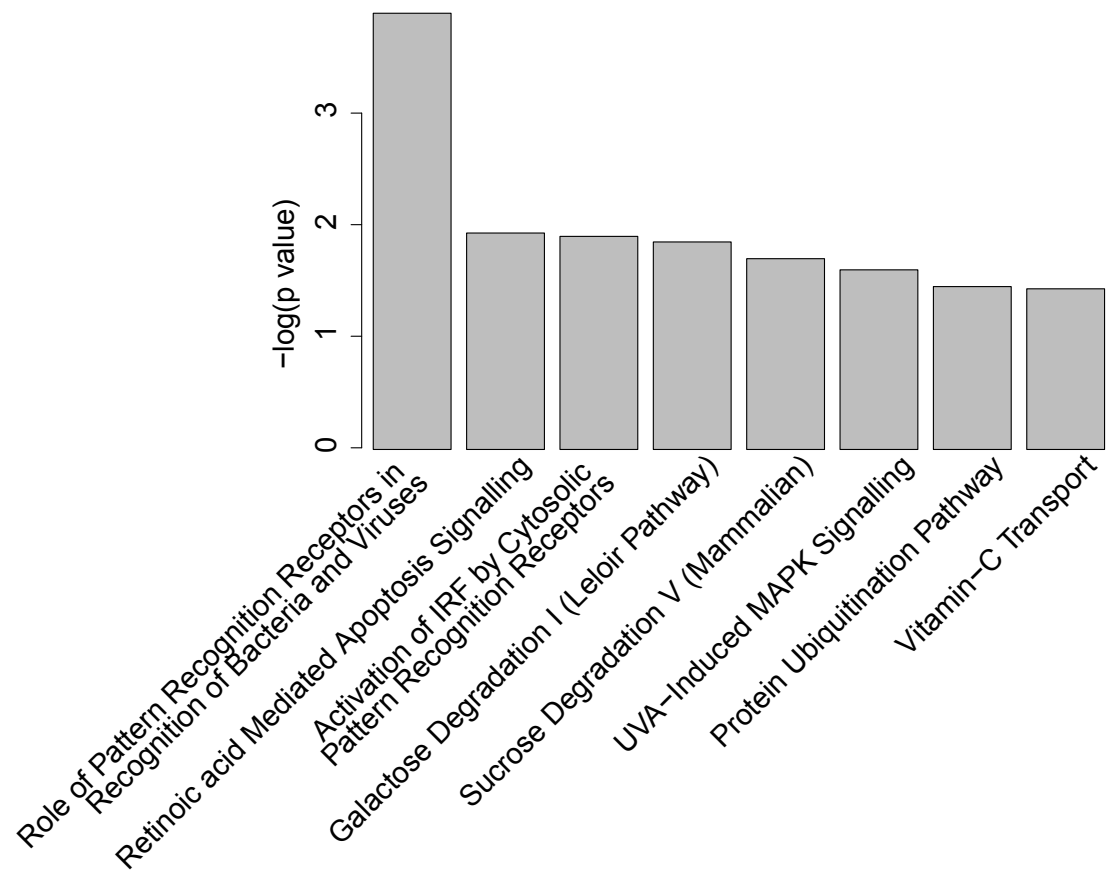


Figure 3.11: Pathway enrichment for viral infection. The differentially expressed probes between patients with viral infection and those without were used to determine pathway enrichment with IPA. The bars show the $-\log(p \text{ value})$ for the enrichment of each pathway. The ten most enriched pathways are shown.

3.2.11 Prediction models for viral infection

The identification of the invading organism in sepsis remains a challenge. Therefore subsets of clinical covariates and genes were selected for prediction models using the GeneRave R package (CSIRO Bioinformatics R package version 3.0.8). To reduce the number of possible predictors, only probes that were differentially expressed ($\text{FDR} < 0.05$ and fold change > 1.2), mapped to genes in IPA and were moderate to highly expressed (expression > 6.5) in a proportion of samples which equated to the smallest group of the comparison were used. Table 3.9 shows the covariates and genes selected for each model. The MR for each model was determined using leave-one-out cross-validation (R Core Team

2013). Figure 3.12 illustrates the ROC curve for each model (ROCR R package, Sing et al. 2005).

While the genes selected for the prediction model are markers for the general difference in response observed between patients with and without viral infection, two are of particular interest. *IFI27* encodes an interferon alpha-inducible protein which has previously been described as a marker of influenza virus infection in PBMCs from a cohort of paediatric patients (Ioannidis et al. 2012). *LAMP3* encodes a glycoprotein expressed when immature dendritic cells differentiate into interdigitating dendritic cells (Saint-Vis et al. 1998). This gene has been shown to be induced in human lung epithelial cells upon influenza A virus infection (Zhou et al. 2011). The biological relevance of these marker genes adds weight to the potential predictive power of this model in further patient cohorts.

Table 3.9: Prediction model for viral infection. The GeneRave R package (CSIRO Bioinformatics R package version 3.0.8) identified the most informative clinical covariates or gene expression signatures for viral infection. The MR was determined for clinical covariates only, gene expression only and the combination of covariates and expression using leave-one-out cross-validation.

Covariates selected	MR (%)	Genes selected	MR (%)	Combined MR (%)
Polynucleocytes (raw)	9.8	<i>IFI27</i>	9.1	10.2
FIO2		<i>TGIF2</i>		
Haematocrit		<i>LY6E</i>		
Temperature (low)		<i>CCNY</i>		
Respiratory rate		<i>DYNLL2</i>		
Inotropes		<i>LAMP3</i>		

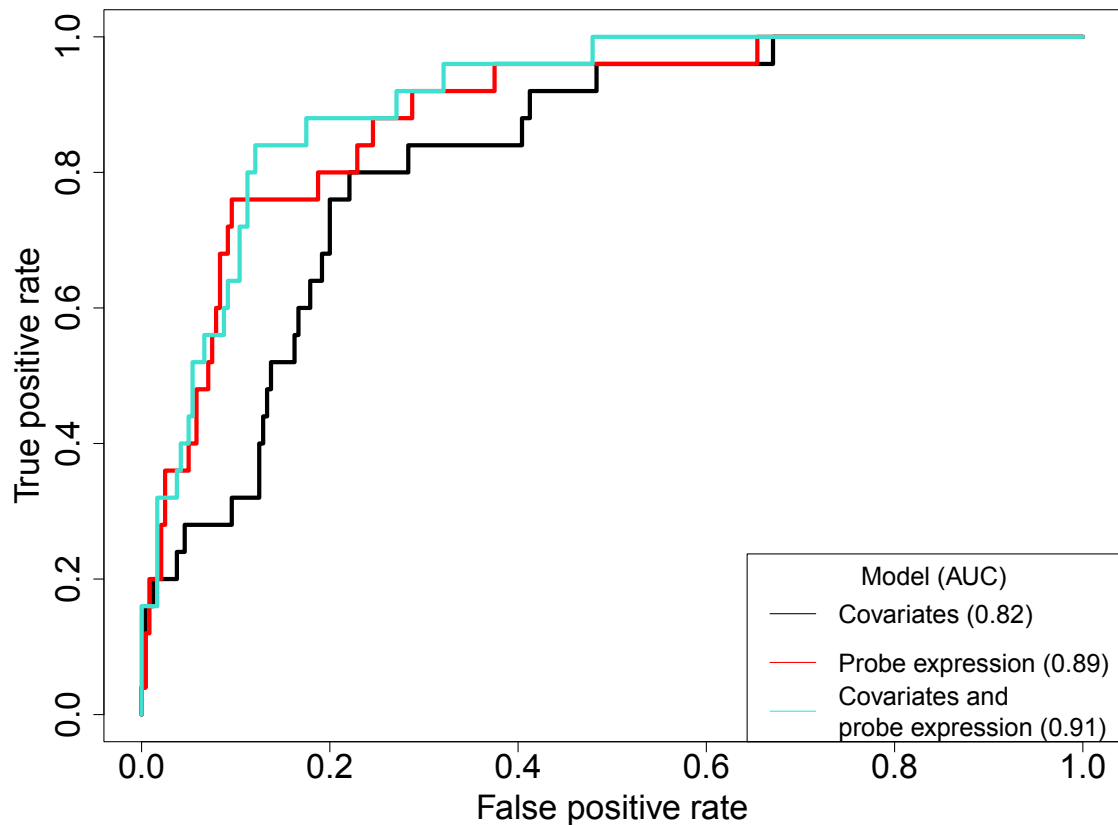


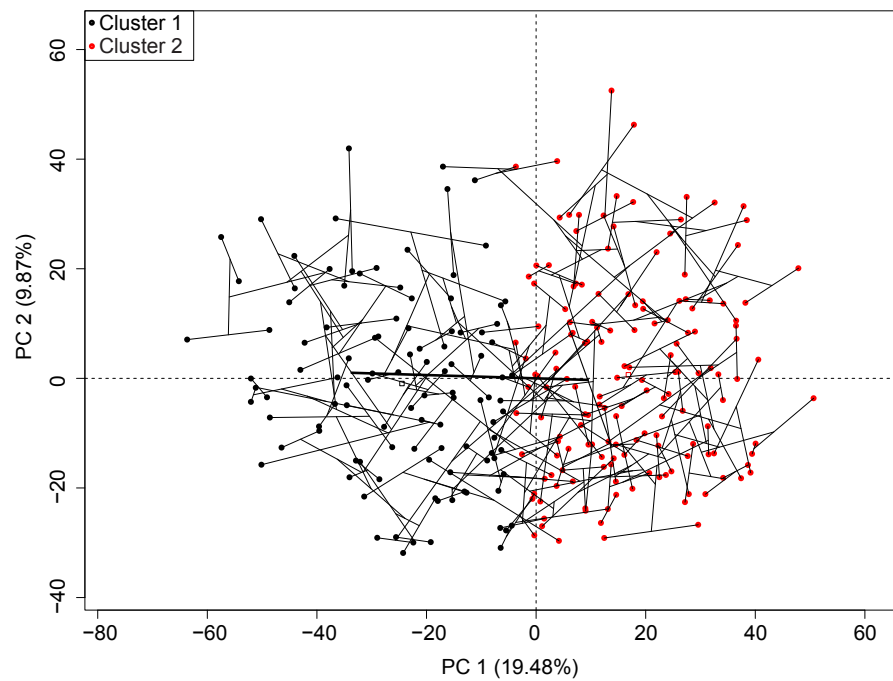
Figure 3.12: ROC curves for viral prediction models. A subset of clinical covariates and genes were identified to predict viral infection. The performance of each model is illustrated by the ROC curve and AUC.

3.2.12 Unsupervised approach to signatures in gene expression

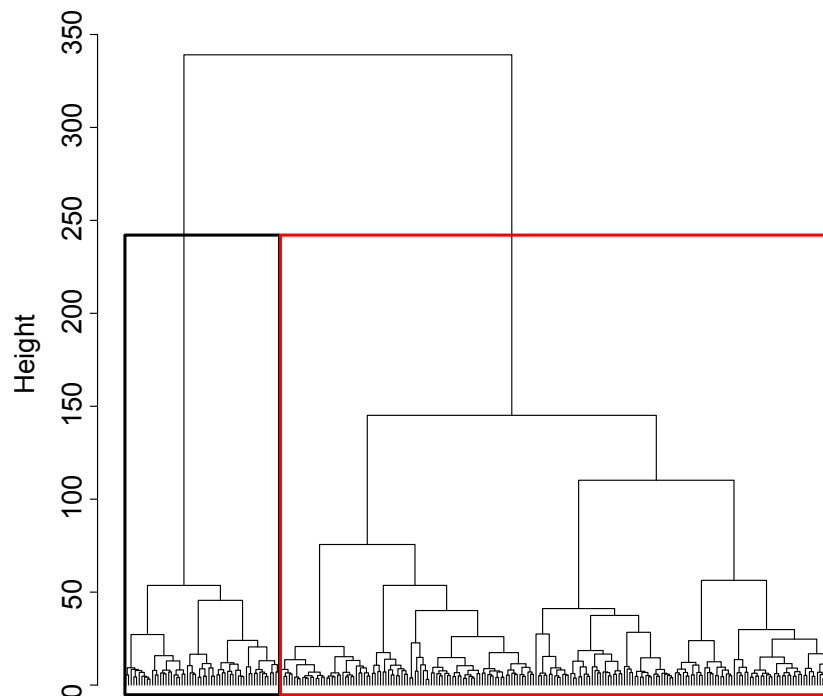
To understand unanticipated structure in the dataset, the FactoMineR R package (Husson et al. 2013) was used to define clusters in the cohort based on the most variably expressed probes. This package uses hierarchical clustering to initiate the number of clusters followed by K-means to determine cluster membership. Two clusters were observed in the data (Figure 3.13) and Kaplan-Meier plots reveal a difference in survival between the patients based on cluster membership (Figure 3.14).

To determine whether this was an optimal number of clusters to define in the dataset, the within-group sum of squares was calculated after applying the K-mean procedure for increasing numbers of clusters. The point at which

increasing the number of clusters in the dataset no longer reduced the within-group sum of squares occurred between 2 and 4 clusters. The minimum number of clusters were selected for further analysis but the observed difference in survival remained whether 3 or 4 clusters were defined in the dataset (Figure A.1).

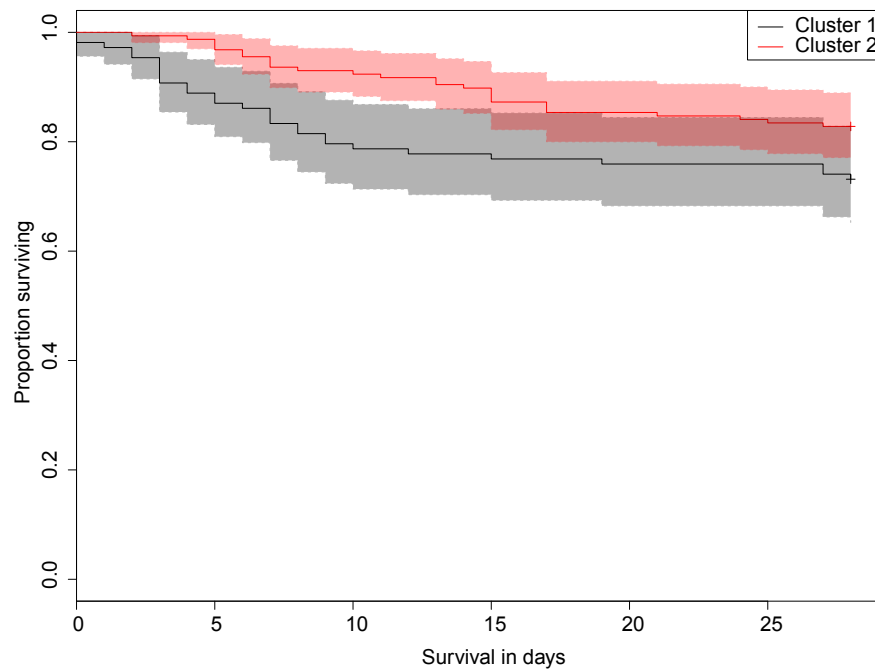


(a) Factor map

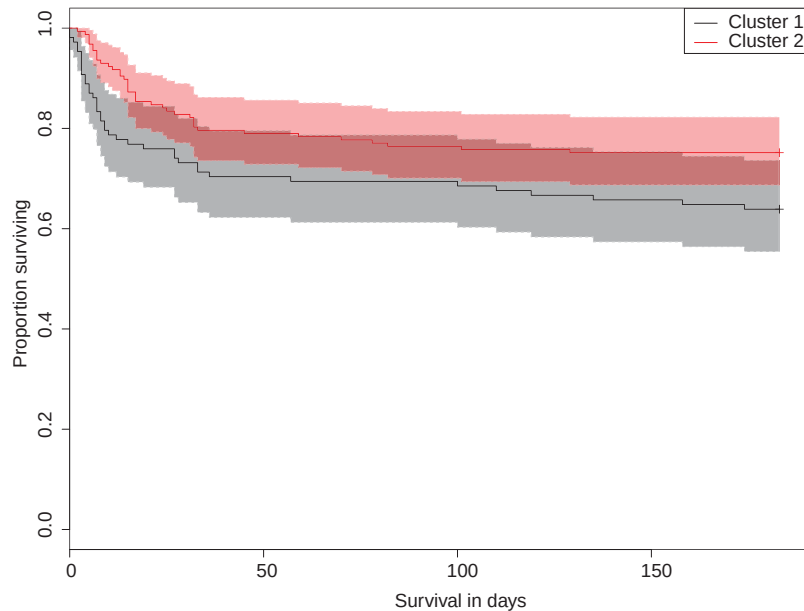


(b) Hierarchical clustering

Figure 3.13: Clusters within the cohort. Coefficient of variation was calculated to identify the 10% of probes which were most variable in the dataset (2,619). Two clusters were defined using FactoMineR (Husson et al. 2013). a) The first two PCs (dimensions) for each individual are plotted to produce the map which allows the two clusters to be defined. Each individual is coloured according to their cluster. b) The hierarchical tree illustrates the arrangement of clusters. The height of the branches indicates the strength of the separation between individuals and clusters.



(a) 28 day survival curves



(b) 6 month survival curves

Figure 3.14: Kaplan-Meier plots for clusters a) Kaplan-Meier plots (coloured by cluster) are significantly different between cluster up to 28 days (p value = 0.037). The shaded area represents the 95% confidence interval. b) Kaplan-Meier plots (coloured by cluster) are significantly different between cluster up to six months (p value = 0.032). The shaded area represents the 95% confidence interval.

3.2.13 Associations between clinical covariates and cluster

The demographics and severity scores for each cluster are shown in Table 3.10. The mean age and % male in the two clusters is quite similar suggesting these two covariates are not responsible for the difference in survival observed. The median severity scores are also similar but both scores have a higher maximum score observed in cluster 1 with the poorer outcome.

Tables 3.11 and 3.12 summarise the clinical covariates associated with cluster membership. These include white cell count, systolic blood pressure and heart rate. Of particular interest are the covariates that also show similar associations with survival. For example, increased raw lymphocyte count is associated with both survival at 14 day and cluster 2 membership (better outcome). This pattern is observed for a number of other covariates too for instance SOFA score, bicarbonate, FI0₂, heart rate and mean arterial pressure. Combining the results of the cluster analysis with the survival analysis may help prioritise the covariates which are truly important for survival.

Table 3.10: Sepsis cohort demographics and severity scores by cluster.

Patient characteristic	Cohort	Cluster 1	Cluster 2
Number of patients	265	108	157
Mean age ($\pm$ SD)	62 years ($\pm$ 16)	60 years ($\pm$ 15)	63 years ($\pm$ 17)
Male	55%	56%	54%
Median APACHE II score (range)	18 (6-42)	18 (6-42)	17 (6-34)
Median SOFA score (range)	6 (0-21)	6 (1-21)	5 (0-14)

Functional genomic response in severe sepsis

Table 3.11: Associations between continuous clinical covariates and cluster. Associations between continuous clinical covariates and cluster membership were determined by *t test*. The table shows covariates with a p value <0.05.

Covariate	p value	Cluster 1 (mean value)	Cluster 2 (mean value)
Polynucleocytes (raw)	2.71E-03	13.07	10.63
Mononucleocytes (raw)	1.40E-03	0.68	0.94
Lymphocytes (raw)	2.15E-09	0.71	1.20
Polynucleocytes (proportion)	9.65E-12	0.89	0.82
Mononucleocytes (proportion)	1.54E-05	0.05	0.08
Lymphocytes (proportion)	7.19E-10	0.06	0.11
White cell count	4.09E-02	14.53	12.78
SOFA (sampling day)	2.13E-07	7.9	5.4
SOFA (maximum recorded)	3.15E-05	9.0	7.1
Bicarbonate	6.93E-09	23.0	27.6
FIO2	5.93E-04	0.57	0.48
Heart rate (low)	5.36E-04	84.1	76.4
Heart rate (high)	3.25E-02	116.0	110.1
Mean arterial pressure (low)	1.80E-05	63.5	70.1
Mean arterial pressure (high)	2.76E-07	94.0	104.7
Systolic blood pressure (low)	1.94E-05	93.3	103.0
Systolic blood pressure (high)	2.81E-07	141.9	157.5
pH	1.83E-02	7.3	7.3
Platelets	1.44E-06	182.8	246.8
Respiratory rate	1.46E-02	27.7	30.9

Table 3.12: Associations between count clinical covariates and cluster. Associations between count clinical covariates and cluster were determined by Chi-squared test. The table shows covariates with a p value <0.05. N-S non-survivor. Ad administered.

Covariate	p value	Cluster 1 proportion (variable)	Cluster 2 proportion (variable)
RNA (sampling day)	5.38E-05	0.61 (day 1)	0.41 (day 1)
14 day survival	1.19E-02	0.22 (N-S)	0.10 (N-S)
Activated protein C	1.14E-02	0.13 (ad)	0.04 (ad)
Renal failure	2.14E-02	0.73	0.85
Inotropes	1.38E-03	0.46 (no dose)	0.69 (no dose)
		0.01 (low dose)	0 (low dose)
		0.16 (medium dose)	0.06 (medium dose)
		0.37 (high dose)	0.25 (high dose)

3.2.14 Differential gene expression between clusters

Within the cohort, 108 individuals were assigned to cluster 1 and the remaining 157 to cluster 2. Limma analysis revealed 9,138 differentially expressed probes with $FDR < 0.05$ and fold change > 1.2 . IPA recommends that pathway enrichment is calculated for fewer than 1,000 genes so the fold change threshold was increased to > 2 . This resulted in 946 differentially expressed probes ($FDR < 0.05$ and fold change > 2 , 845 genes) between cluster 2 and cluster 1. The differential expression threshold was increased for the cluster membership analysis to a fold change > 2 to generate a list appropriate for pathway enrichment. The majority of these probes are upregulated in cluster 2 (Figure 3.15). The 50 most significant differentially expressed probes are shown in Table A.9.

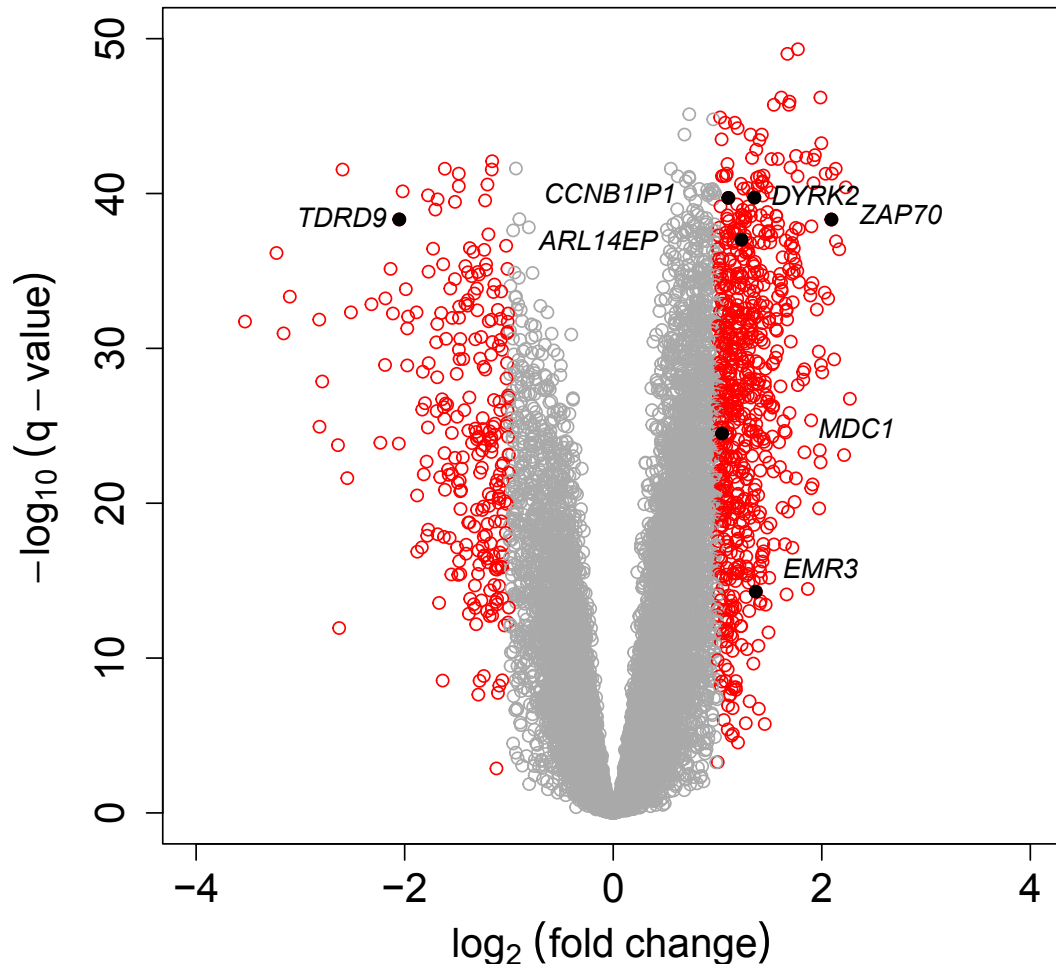


Figure 3.15: Volcano plot of differentially expressed probes for cluster. The significance of differential expression between the two clusters is plotted against the fold change. Each point represents a single probe. Red points highlight probes with FDR < 0.05 and a fold change > 2. A positive fold change indicates upregulation in cluster 2. The probes selected for the prediction model in Section 3.2.16 are highlighted in black and labelled with their corresponding gene names.

3.2.15 Pathway analysis for clusters

IPA was used to assess the functional significance of the probes differentially expressed between the two clusters. Of the 845 genes with differentially expressed probes, 778 mapped to IPA. The enriched pathways are shown in Figure 3.16 and Table A.10. A heatmap illustrates the increase or decrease of

IPA disease and functions between clusters (Figure 3.17). These results highlight distinct immune responses to sepsis occurring between the patients in the two clusters. The most striking result is the upregulation of genes involved in the activation and function of lymphocytes in the cluster 2 individuals (with the better survival rate) compared to cluster 1.

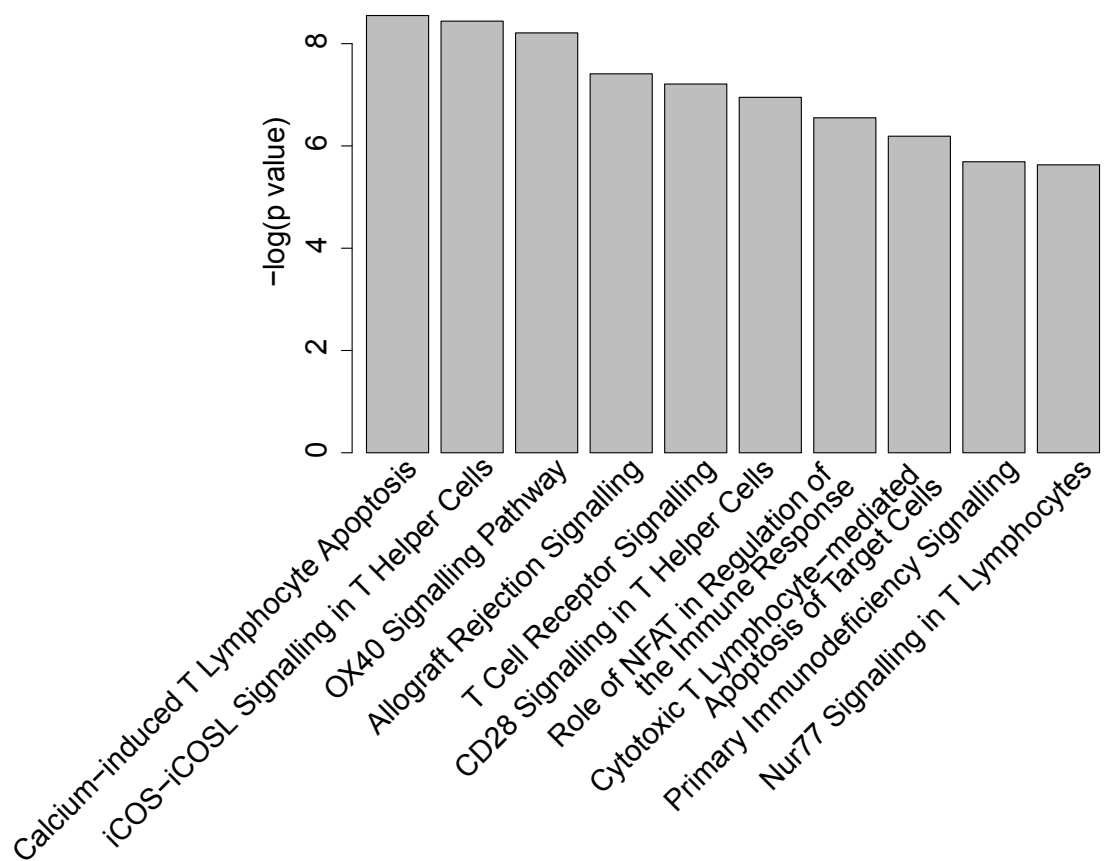


Figure 3.16: Pathway enrichment for clusters. The probes differentially expressed between cluster 2 and cluster 1 were used to determine pathway enrichment with IPA. The bars show the $-\log(p \text{ value})$ for the enrichment of each pathway. The ten most enriched pathways are shown.

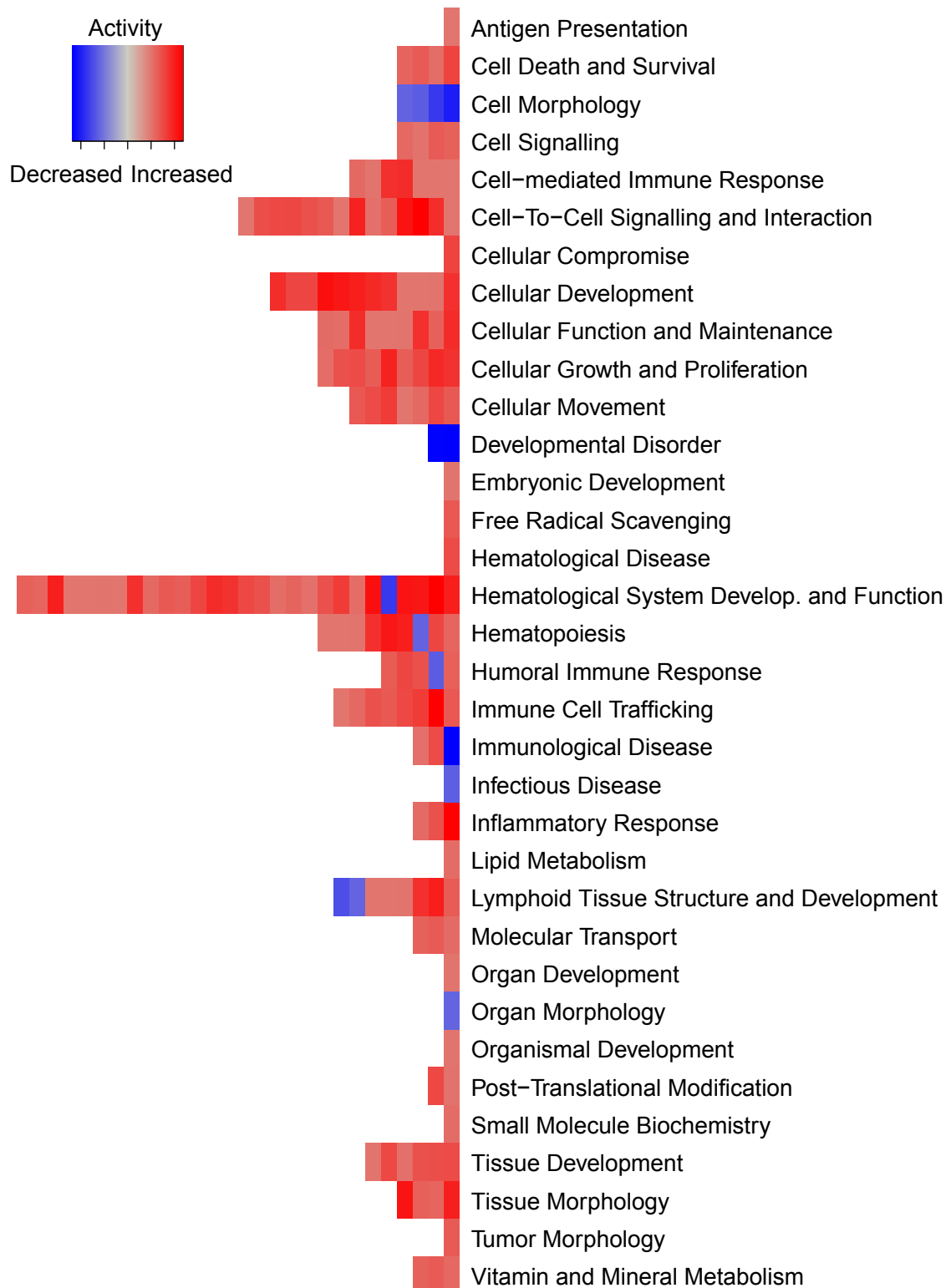


Figure 3.17: Heatmap of IPA disease and functions activity between clusters. IPA reports increase or decrease of disease and functions activity between clusters based on differentially expressed probes. Each square represents a disease and function which can be classified into broader categories.

3.2.16 Prediction models for clusters

Subsets of clinical covariates and genes were selected for cluster prediction models using the GeneRave R package (CSIRO Bioinformatics R package version 3.0.8). To reduce the number of possible predictors, only probes that were differentially expressed (FDR <0.05 and fold change >2), mapped to genes in IPA and were moderate to highly expressed (expression >6.5) in a proportion of samples which equated to the smallest group for the comparison were used. Table 3.13 shows the covariates and genes selected for each model. The MR for each model was determined using leave-one-out cross-validation (R Core Team 2013). Figure 3.18 shows the box plots for the cluster membership predictor genes and Figure 3.19 illustrates the ROC curve for each model (ROCR R package, Sing et al. 2005).

Table 3.13: Prediction models for clusters. The GeneRave R package (CSIRO Bioinformatics R package version 3.0.8) identified the most informative clinical covariates or gene expression signatures for cluster membership. The MR was determined for clinical covariates only, gene expression only and the combination of covariates and expression using leave-one-out cross-validation.

Covariates selected	MR (%)	Genes selected	MR (%)	Combined MR (%)
Age	19.6	<i>DYRK2</i>	3.8	6.4
Lymphocytes (raw)		<i>CCNB1IP1</i>		
Polynucleocytes (proportion)		<i>TDRD9</i>		
FIO2		<i>ZAP70</i>		
Systolic blood pressure (high)		<i>ARL14EP</i>		
Bicarbonate		<i>MDC1</i>		
Platelets		<i>EMR3</i>		
Bilirubin				
Respiratory rate				

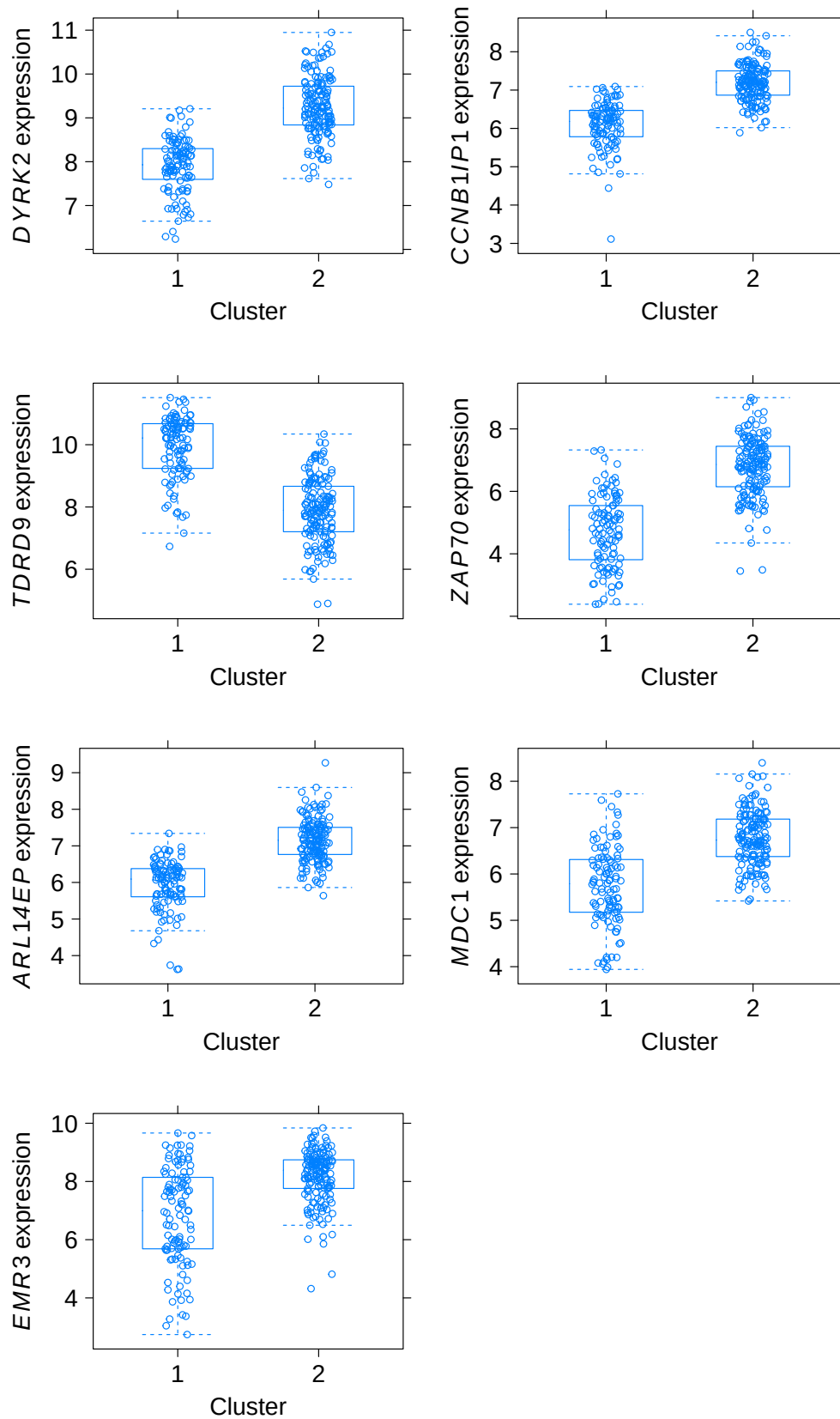


Figure 3.18: Box plots for cluster membership predictor genes. The expression signatures of seven genes were selected as the most informative predictors of cluster membership.

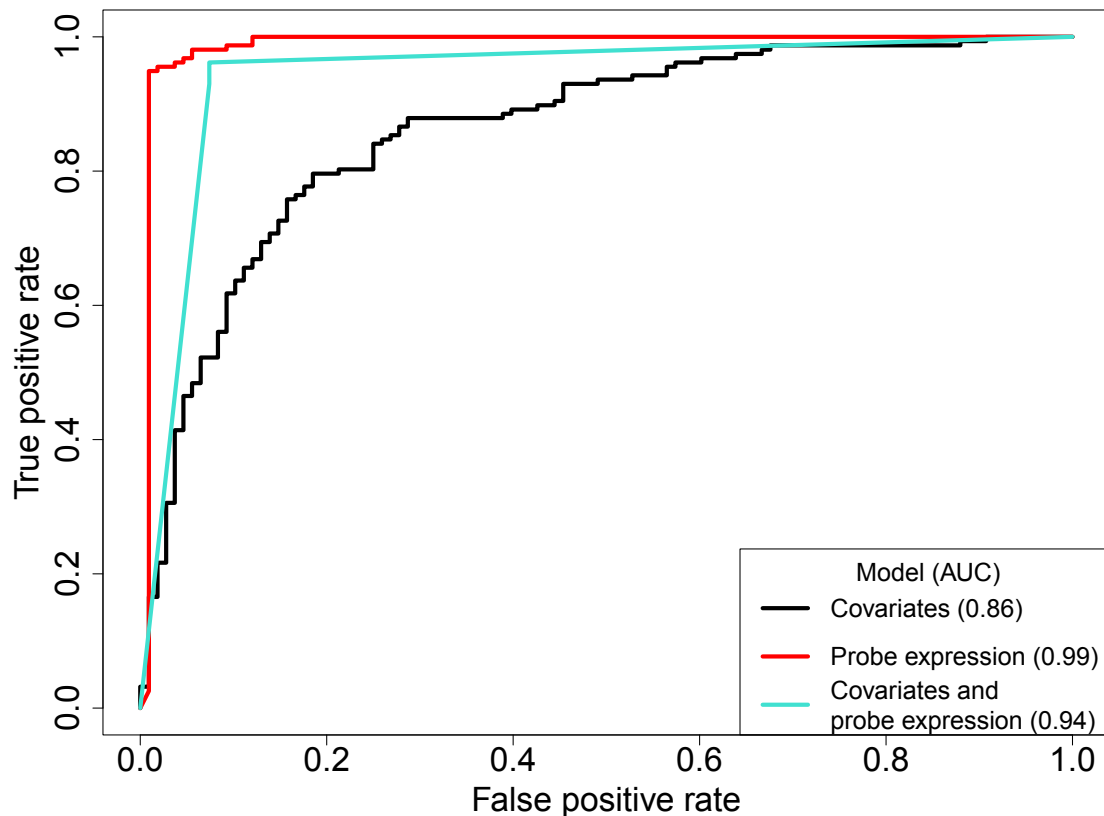


Figure 3.19: ROC curves cluster membership prediction. A subset of clinical covariates and genes were identified to predict cluster membership. The performance of each model is illustrated by the ROC curve and AUC.

3.2.17 Overlap between survival analysis and cluster membership

In order to compare the overlap of differentially expressed probes between the survival (14 day) and cluster membership analysis, the threshold for differential expression was set at FDR <0.05 and fold change >1.2. The probes differentially expressed between 14 day non-survivors and 6 month survivors are significantly enriched for probes differentially expressed between cluster 2 and cluster 1 (Chi-squared p value <0.001) (Figure 3.20a). Of the 1,495 probes shared between these two comparisons, only 1 probe had an opposite direction of fold change.

The differential expression threshold was increased for the cluster membership analysis to a fold change >2 to generate a list appropriate for pathway enrichment. The differentially expressed probes for each comparison were used to determine pathway enrichment with IPA. 43 pathways were

Functional genomic response in severe sepsis

enriched for the survival analysis and 47 for cluster membership. 13 pathways were shared between both analyses and are listed in Figure 3.20b.

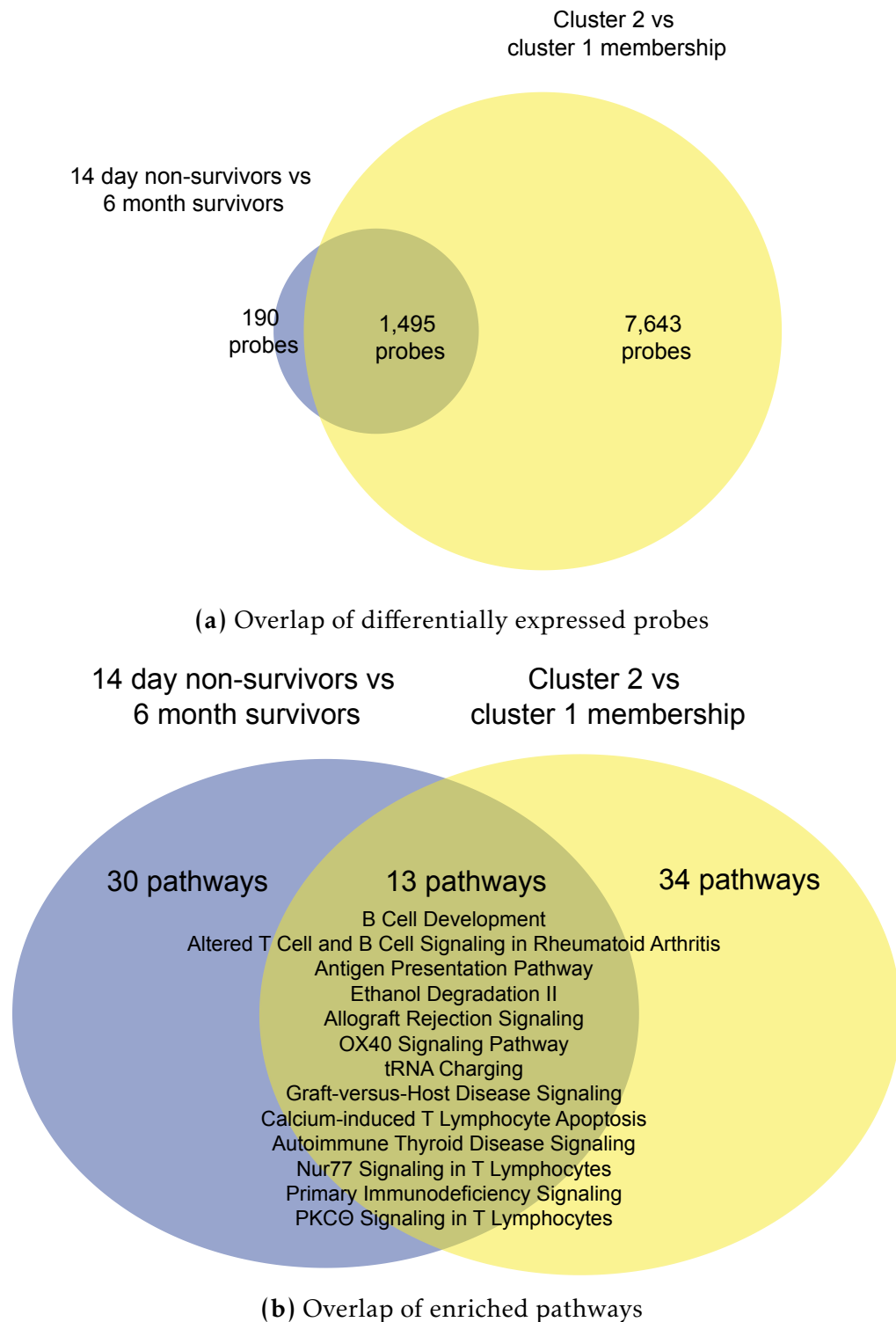


Figure 3.20: Overlap between survival analysis and cluster membership. a) Differential expression was defined as FDR <0.05 and fold change >1.2 for the both the survival and cluster membership analysis. 1,495 probes are differentially expressed in both analyses. b) The differential expression threshold was increased for the cluster membership analysis to a fold change >2 to generate a list appropriate for pathway enrichment. The differentially expressed probes for each comparison were used to determine pathway enrichment with IPA. 13 pathways were enriched for both analyses and are listed in the diagram.

3.3 Discussion

The work presented in this chapter defines the severe sepsis transcriptome and relates this to outcome and clinical covariates for the largest known adult sepsis CAP cohort.

3.3.1 Gene expression signatures relating to outcome

Extensive clinical phenotyping was recorded as part of the GAInS study. This allowed an analysis of associations with different survival lengths to highlight which covariates may be important for identifying the most severely ill patients. Many of the covariates identified were similar for different survival lengths (Tables 3.3 and 3.4) providing further evidence that alternative measures, for example gene expression, are required for distinguishing illness severity.

Differentially expressed genes were identified between 6 month survivors and non-survivors at each of the three survival lengths measured (14 days, 28 days and 6 months). The greatest number of differentially expressed probes were observed between the 14 day non-survivors and 6 month survivors (Figure 3.5) suggesting that differences in gene expression are most informative for outcome events which occur closer to the sampling time point. This was further supported by the prediction models which showed gene expression had greater prediction power for 14 day survival while clinical covariates were more successful for the 6 month time point (Figure 3.9). The fold changes observed for the outcome analysis were modest which suggests the somewhat arbitrary outcomes used for many sepsis studies (28 days) may not be the most informative measure.

3.3.2 Severity scores

The two ICU severity scores (APACHE II and SOFA) correlated with outcomes but these scores were developed for mixed groups of patients in ICU to identify

general illness severity rather than individual predictions. Their accuracy in particular subgroups of patients may therefore not be adequate (Vincent and Moreno 2010). This study identified 5-7 clinical covariates which were more informative for predicting severity of illness with non-survival as the outcome for each survival length (Figure 3.9).

3.3.3 Microbiology

No differentially expressed genes were identified between the patients with Gram-positive compared to those with Gram-negative bacterial infections in this cohort (Table 3.8). Previous studies have shown similar results for both neutrophils (Tang et al. 2008) and PBMCs (Tang et al. 2009). One problem with the clinical data collected for sepsis patients is the number of cases with negative bacterial cultures. 49% of this cohort did not have a positive bacterial culture, similar to Phua et al. (2013) which showed 40%. This calls into question the reliability of classification and therefore whether correct comparisons are being made in the cohort. However, one hypothesis is that there is in fact a general response to bacterial infections in sepsis and therefore the lack of difference between the different types of infection is not unexpected (Tang et al. 2009). This is supported by the indistinguishable clinical manifestations of Gram-negative and Gram-positive sepsis (Sriskandan and Cohen 1999).

This study has shown a significant difference in differentially expressed genes between individuals with a viral infection compared to those without (Figure 3.10). Ramilo et al. (2007) identified a panel of 35 classifier genes which could distinguish Influenza A viral infection from bacterial infection on a cohort of paediatric patients with acute infection. Two of the sepsis viral infection prediction genes (*IFI27* and *LY6E*) (Table 3.9) overlap with the classifiers from the paediatric study providing further insight into the genes which may be most informative for a prediction model that could be applied to other datasets.

3.3.4 Cluster analysis

While determining differential gene expression between survival times or other clinical covariates can be informative, it remains a biased approach based on current knowledge of the disease. Sepsis can be regarded as a relatively non-specific definition and the range of physiological and molecular phenotypes included suggests the existence of unrecognised subtypes (Maslove and Wong 2014). This study reduced heterogeneity by focusing on one cause of sepsis but further, unanticipated structure in the dataset was determined through cluster analysis (Section 3.2.12).

Wong et al. (2009b) identified three subclasses within a cohort of paediatric patients with septic shock. One of these subclasses had higher illness severity, higher mortality rates and repression of genes corresponding to key signalling pathways of the adaptive immune system, similar to individuals from cluster 1 presented in this chapter (Section 3.2.12). The three subclasses were distinguished by a signature of 100 genes but this has not been validated in an independent cohort.

Cluster analysis has also been conducted on adult sepsis cohorts but the results were not as striking as those presented in this thesis. Maslove and colleagues (2012) defined two clusters from neutrophil microarray data using two independent cohorts for discovery and validation (55 and 71 patients respectively). However, the clusters were identified using expression for a subset of 365 genes candidate genes which had previously been associated with sepsis. Pathway analysis of the genes differentially expressed between the clusters showed increased expression of genes involved in inflammation (chemokine and cytokine signalling pathways and Toll receptor signalling pathway) in one cluster compared to the other. The cohort was composed of both severe sepsis and septic shock patients and the proportion was not equal between the clusters which may have caused the differences observed.

3.3.5 Limitations

The gene expression data generated through a microarray study does not allow for analysis of post-translational modification of proteins or protein interactions that may play a role in the disease (Maslove and Wong 2014).

Sepsis is a dynamic disease and the underlying inflammatory changes are still not well understood. One major challenge in studying this disease in a clinical setting is determining the time point of disease on an individual basis. This study used the first available time point available within five days of ICU admission in an attempt to use clinically relevant time points but also generate a cohort large enough for genetic analysis. Due to ICU admission at different time points of disease progression it is unclear how best to improve this issue of sampling. The results should therefore be interpreted with this in mind however, particularly in the case of the cluster analysis, even if these clusters represent different time points in disease, identifying this on an individual basis is still helpful as different therapies could be applied based on this information.

The 28 day mortality rate in the cohort presented (21%) is lower than that in the literature (28-50%) (Angus et al. 2001; Padkin et al. 2003; Sands et al. 1997; Zeni et al. 1997). One possible explanation is enrichment of survivors due to consent being required before sampling which cannot always be achieved for the most severely patients who die very quickly upon ICU admission.

Cluster membership is determined by gene expression differences observed within the cohort under analysis. Therefore direct translation into the clinical setting could be more challenging. However, the difference in survival rates between clusters highlights that important information can be generated from unbiased approaches in terms on new insights into disease pathogenesis and heterogeneity.

3.3.6 Conclusion

This analysis illustrates the use of a functional genomic approach in the resolution of disease heterogeneity through the identification of clusters. These clusters may be the result of variation in the response to infection between different individuals or may be due to sampling different phases of response. This data rich cohort should provide a valuable resource for further transcriptomic research in the context of sepsis.

Chapter 4

Regulatory genetic variants in severe sepsis

4.1 Introduction

GWAS have identified significant associations between genetic loci and disease phenotypes but translating the findings into an understanding of how the loci contribute to disease, particularly when the signals map to inter-genic regions, has remained challenging (Cookson et al. 2009). Using gene expression as an intermediate phenotype between the genetic variation and the disease phenotype of interest can provide insight into the molecular mechanism linking the two together (Fairfax and Knight 2014). Gene expression can be mapped as a quantitative trait through eQTL analysis which identifies associations between genotyping data and gene expression across the whole genome (Cookson et al. 2009; Michaelson et al. 2009). The aim of this chapter is to integrate an eQTL analysis for the cohort of sepsis patients (described in Chapter 3) with the results of a GWAS for sepsis outcome, that includes the same patients, and determine the extent of context specificity by comparison with data generated in other contexts.

4.1.1 Genotyping QC

In order to map gene expression as a quantitative trait, the generation of genome-wide genotyping data is required. Following data generation, QC steps remove

substandard SNPs and samples to reduce the number of false-positive and false-negative associations. Anderson et al. (2010) summarise the QC steps that should be applied for case-control associations but many of these are equally important for eQTL analysis and therefore these steps can be used to form a pipeline for QC.

Implementing the QC steps for the samples before the SNPs maximises the number of markers remaining in the study and all the necessary steps can be conducted using PLINK software (Purcell et al. 2007).

For sample QC, particularly when using clinical samples, it is important to identify any discordance between the recorded sex and sex calculated from genotype. This can occur for a number of reasons including incorrect clinical data recording, contamination, sample mislabelling or DNA plating errors but can be identified from the homozygosity rate of the X-chromosome. To assess the quality of the sample DNA the missing genotype and heterozygosity rate can be used. Identity by descent (IBD) determines the relatedness of samples to ensure all samples are unrelated. This is important as it prevents the over-representation of certain genotypes. Finally, population stratification can be a source of confounding variables, particularly in an eQTL analysis where an association between a SNP and gene expression could arise because of allele frequency differences between populations. Multi-dimensional scaling identifies population outliers compared to the 1000 Genomes Project populations (1000 Genomes Project Consortium et al. 2012).

For the SNP QC, it is important to remove sub optimal SNPs but a balance must be struck between removing poor quality SNPs and losing potentially informative SNPs. A range of thresholds have been used for the SNP call rate from 95-99% with many studies choosing higher thresholds for low frequency SNPs. Deviation from Hardy-Weinberg equilibrium (HWE) can indicate genotyping errors. SNPs with low MAF are typically difficult to call

because of the low numbers of heterozygote and rare homozygote samples. Generally a MAF threshold of 1-2% is applied.

4.1.2 Imputation

Due to recombination hotspots, many common SNPs are transmitted across generations in blocks. This means a smaller number of SNPs can be used to tag (represent) the variation present (Gabriel et al. 2002; Slatkin 2008). Different cohorts are frequently genotyped using different platforms. This can result in few overlapping SNPs between datasets and therefore make comparisons more challenging. The missing genotypes can be imputed using known haplotypes from the HapMap Project (International HapMap Consortium et al. 2010) or 1000 Genomes Project (1000 Genomes Project Consortium et al. 2012). As well as allowing a more comprehensive comparison between datasets, imputation also allows the fine mapping of an association region to identify any SNPs with stronger associations than those genotyped (Manolio 2010).

4.1.3 Gene expression QC

For the gene expression data used in an eQTL analysis, similar QC and normalisation steps should be carried out to those described in Section 3.2.5 to reduce systematic biases introduced from experimental variability. In addition, probes which map to more than one position in the genome and those with sequences containing SNPs should be removed to prevent spurious results.

4.1.4 Matrix eQTL

eQTL analysis involves testing for association of large numbers of transcript-SNP pairs. This can result in a computationally heavy task which can limit the number of samples, SNPs and transcripts that are used in the analysis. The Matrix

eQTL R package (Shabalin 2012) provides a computationally efficient method of determining eQTL associations using an additive linear model.

4.1.5 *Cis* and *trans* associations

When a SNP-gene association is detected, the SNP is described as *cis* acting if it is located near the gene of interest. Variable distances have been used in the literature to define *cis* but for this study it will be defined as within 1Mb of the probe start site. Associations between the probe and SNPs outside of this window are therefore referred to as *trans*.

4.1.6 PCs vs covariates

Global gene expression can be confounded by variation as a result of environmental, technical, or physiological effects. Removing these effects can increase the power to detect eQTLs (Nica et al. 2011; Westra et al. 2013). There are two main methods to achieve this: taking account of PCs or known covariates. Using PCs provides a more unbiased approach as no previous knowledge is required but there is also an argument that if covariates are known to introduce variability then they should be included.

Liang et al (2013) found the top PCs for their gene expression data were highly correlated with a number of processes that can introduce variation into the dataset, for instance the date of RNA extraction and chip hybridization. Including these PCs in their eQTL analysis reduced false positives and increased power for eQTL detection.

4.1.7 Integrating eQTL with sepsis GWAS

Many SNP associations have been determined from GWAS but there are often challenges in identifying the specific modulated gene(s) and understanding how a particular locus may contribute to the disease of interest. An eQTL association

reveals a link between a SNP and expression of a specific gene. Integrating eQTL analysis with GWAS could provide unique insights into the molecular mechanisms which link genes to a disease. The sepsis patients recruited as part of the GAINs study are a subset of a larger, European study (GenOSept, www.genosept.eu). Professor Adrian Hill and colleagues (WTCHG, Oxford) have used this cohort to carry out GWAS for outcome from sepsis (Rautanen et al. 2015).

The sepsis outcome GWAS included 1,035 CAP patients recruited through the GenOSept study plus patients from two further independent Caucasian cohorts (VASST, $n = 217$ (Russell et al. 2008) and PROWESS, $n = 301$ (Bernard et al. 2001)) to allow a meta-analysis of 1,553 patients with sepsis due to pneumonia. The most significantly associated SNPs were then genotyped in a further 538 CAP patients which included the patients presented in Chapter 3, recruited as part of the GAINs study. The SNPs genotyped in the replication cohort are summarised in Table 4.1. The most significant association (rs 4957796, p value 5.6×10^{-8}) is in the region of the *FER* gene which is a non-receptor protein tyrosine kinase which influences leukocyte recruitment in response to LPS (Pruitt et al. 2014). However, the association did not reach the genome-wide significance threshold (5×10^{-8}). A further 865 associations were identified with a p value $< 1 \times 10^{-4}$. Identifying overlap between the eQTL study and the GWAS, particularly for SNPs with borderline significance may provide valuable evidence for directing future analysis.

Table 4.1: Sepsis outcome GWAS. The SNPs shown were selected from the discovery cohort for replication. *Although the p-value in the replication set is <0.05 the direction of the effect is in the opposite direction to that seen in the discovery set. Adapted from Rautanen et al. (2015). Chr = chromosome; OR = odds ratio.

Chr	Position	SNP ID	Discovery		Replication		Combined		Gene
			p value	OR	p value	OR	p value	OR	
2	134091649	rs13392963	1.60E-05	0.63	0.362	1.06	4.40E-04	0.73	NCKAP5
2	201260365	rs78318430	2.00E-05	2.96	0.025*	0.24	6.90E-04	2.26	SPATS2L
2	201308486	rs893357	3.40E-06	2.7	0.021*	0.48	0.0029	1.73	SPATS2L
4	91619360	rs72661871	7.80E-06	2.87	0.424	1.06	2.00E-04	2.03	CCSER1
4	91671916	rs72661895	6.40E-06	2.92	0.437	1.06	8.90E-05	2.21	CCSER1
5	108402140	rs4957796	9.70E-08	0.52	0.061	0.70	5.60E-08	0.56	FER
5	108406299	rs975056	3.30E-07	0.56	0.119	0.79	4.60E-07	0.61	FER
5	108417332	rs62375529	7.50E-07	0.56	0.062	0.72	3.70E-07	0.59	FER
5	134514728	rs553438	2.20E-05	0.63	0.252	1.13	1.10E-03	0.74	intergenic
5	134532593	rs639405	8.00E-06	0.62	0.399	0.95	1.60E-04	1.42	intergenic
6	103810003	rs79423885	8.10E-06	2.05	0.023	1.59	1.50E-06	1.89	intergenic
7	83670129	rs76881522	6.20E-06	1.77	0.48	0.99	9.00E-05	1.54	SEMA3A
7	123248738	rs35947027	7.60E-06	1.54	0.067	0.75	1.00E-03	1.33	ASB15
8	19554995	rs12114790	6.40E-05	1.43	0.085	1.23	3.60E-05	1.38	CSGALNACT1
13	38969457	rs17057959	1.50E-06	1.78	0.084	0.74	4.10E-04	1.45	intergenic
13	39102184	rs9566343	1.40E-06	1.71	0.222	0.84	7.10E-05	1.48	intergenic

4.1.8 Context specific *cis* associations

The expression profile of many genes are dependent on the tissue, developmental stage or physiological condition. Studies are showing that many eQTL associations are dependent on context for instance cell type or stimulated state.

Westra et al. (2013) performed an eQTL meta analysis for 5,311 healthy individuals with replication in a further 2,775. This was performed on peripheral blood samples and therefore provides a valuable resource of associations for comparison.

Cells involved in the immune response need to make phenotypic changes in response to alterations in environment for example encountering pathogens. It is therefore hypothesised that eQTL for genes involved in the immune system are likely to show context specificity (Fairfax and Knight 2014). This has been shown through a recent study led by members of our group in which context specific eQTL were identified following the exposure of monocytes from healthy individuals to innate immune stimuli (Fairfax et al. 2014).

Comparing the sepsis eQTL analysis with the two datasets described should yield new insights into the pathophysiology of sepsis.

4.1.9 Aims

The aim of this chapter is to define regulatory genetic variants in severe sepsis. Regulatory genetic variants modulating gene expression are responsible for most reported GWAS and are typically context specific. I aim to

1. map gene expression as a quantitative trait in peripheral blood leukocytes from patients with severe sepsis (Chapter 3)
2. integrate this with the results of a GWAS for outcome in sepsis performed in the same patient cohort (GAinS/GenoSept)

3. establish the sepsis specificity of observed eQTL by comparison with data generated in other contexts including in resting state and following innate immune activation

4.2 Results

Data are presented for the cohort of patients with severe sepsis described in Chapter 3. In order to map gene expression as a quantitative trait, genotyping was performed.

4.2.1 Genotyping QC

Genome-wide genotyping was performed for 730,525 SNPs using the Illumina HumanOmniExpress BeadChip. Data were generated for 295 patients with sepsis. Following QC, 20 samples were identified for removal from further analysis. Samples were removed following the identification of discordance between the recorded sex and the genotype, poor quality DNA from missing genotype and heterozygosity rate, relatedness to other samples or population outliers from multi-dimensional scaling with the 1000 Genomes Project populations (1000 Genomes Project Consortium et al. 2012). The QC reason for removing each sample is summarised in Table B.1. In addition, 86,135 SNPs were identified for removal, due to low call rate, deviation from HWE or low MAF, leaving 644,390 for further analysis.

4.2.2 Gene expression QC

Gene expression data were generated for 303 patients in the sepsis cohort with 265 passing QC. High quality genotyping data were available for 240 patients. Section 2.1.4 describes the QC and data normalisation steps carried out for the gene expression data which resulted in 26,185 probes used for the analysis

described in Chapter 3. Probes were then subsequently selected for eQTL analysis. Probes with sequences that mapped to more than one genomic location using BLAST or had a SNP present at a MAF of at least 1% in the European population of the 1000 Genomes Project were excluded to reduce potential confounding affects. This resulted in 17,347 probes used for the following analysis.

4.2.3 eQTL analysis with principal component adjustment

The eQTL analysis was performed for 17,237 probes and 644,390 SNPs with the R package Matrix eQTL (Shabalin 2012). This package uses an additive linear model and reports FDR for resulting associations between SNPs and probes. To reduce confounding variation in the gene expression data, *cis* eQTL mapping (distance between SNP and probe start site <1Mb) was performed after correction for increasing numbers of PCs. PCs correlating with genotype were not included. The first 30 PCs were included in the final analysis as this is the point which maximises the number of unique probes with *cis* eQTL associations (Figure 4.1).

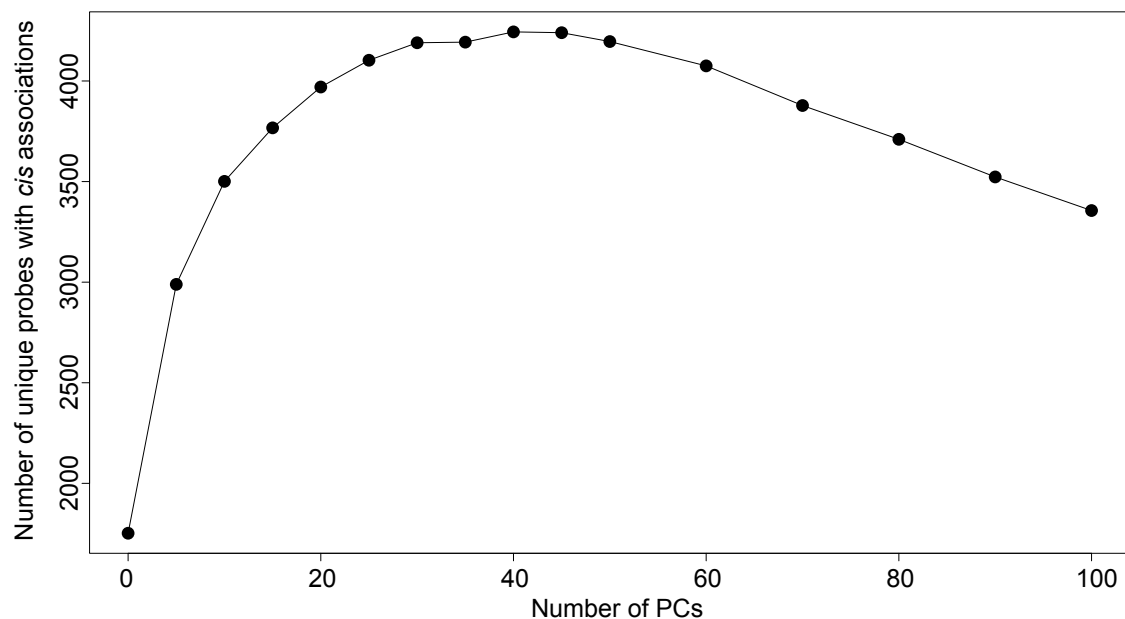


Figure 4.1: Unique probes with cis associations. Cis eQTL mapping was performed after correction for increasing numbers of PCs. PCs correlated with genotype were not used. The number of unique probes with a *cis* association is plotted.

4.2.4 eQTL analysis with covariate adjustment

Extensive clinical phenotyping was recorded for each sepsis patient and this can be used to account for variation within the gene expression data. The effect of taking account of clinical covariates, rather than PCs, was evaluated by performing eQTL mapping taking account of each covariate independently. The effect of accounting for each independent covariate on the number of *cis* eQTLs identified is summarised in Figure B.1. The proportion of polynucleocytes, proportion of lymphocytes and the SOFA score are the three covariates which result in the greatest increase in the number of *cis* eQTLs. Figure 4.2 summarises the comparison between the sequential addition of clinical covariates and the sequential addition of PCs. Matrix eQTL allows covariates to be used in combination with the PCs, although any PCs which correlate with the covariate will be filtered and removed. FigureB.2 demonstrates that taking account of both a clinical covariate as well as filtered PCs results in fewer eQTLs than

taking account of PCs alone. In contrast, if correlated PCs are not filtered and a clinical covariate is included, there is only a modest increase in the number of eQTLs. Further support for using the PC only approach was provided from determining correlations between the clinical covariates and the PCs. As Figure B.3 demonstrates, the majority of clinical covariates were correlated with the first few PCs. PCs only were therefore used in the following analyses.

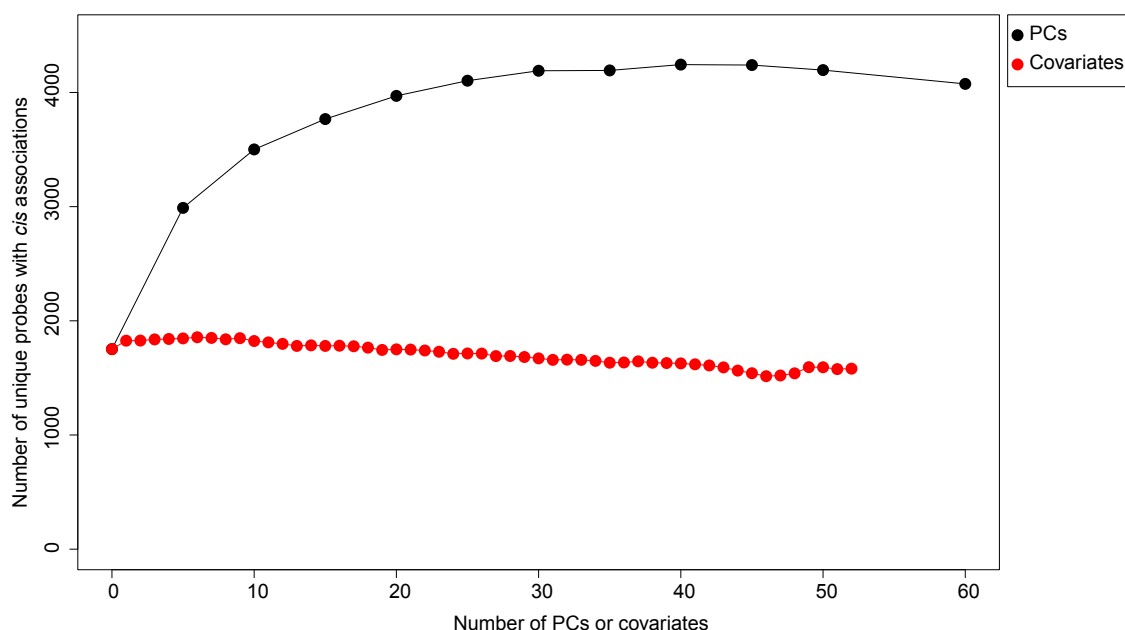


Figure 4.2: PC vs covariates. eQTL mapping was performed taking account of either PCs or clinical covariates. The number of unique probes with *cis* associations are shown for the sequential addition of PCs or clinical covariates.

4.2.5 *Cis* associations

Mapping *cis* eQTL with 30 PCs identified 40,755 SNP-probe associations (FDR <0.05) corresponding to 4,190 unique probes (3,795 genes). The 20 most significant SNP-probe associations are summarised in Table 4.2 and Figure 4.3 shows the association between rs10239340 and expression of *IRF5*. The expression of this gene has been shown to be induced by inflammatory stimuli in macrophages and leads to the expression of proinflammatory genes such as *IL12b*

and *IL23a* and repression of anti-inflammatory genes such as *IL10* (Krausgruber et al. 2011; Weiss et al. 2013).

Table 4.2: Most significant SNP-probe associations. Mapping *cis* eQTL with 30 PCs identified 40,755 SNP-probe associations (FDR <0.05). The most significant SNP-probe associations for the top 20 genes are shown.

SNP ID	Chr	Position	Probe	Gene	FDR
rs10239340	7	128668510	2230431	<i>IRF5</i>	4.22E-58
rs11191642	10	105055912	2320253	<i>USMG5</i>	1.86E-55
rs9849087	3	121465220	3130768	<i>GOLGB1</i>	1.86E-55
rs8070859	17	15887789	2000209	<i>ZSWIM7</i>	2.28E-54
rs7143764	14	65379283	6370612	<i>CHURC1</i>	2.28E-54
rs7313235	12	10132283	3170601	<i>CLEC12A</i>	1.16E-52
rs7736263	5	173339222	1500131	<i>CPEB4</i>	3.46E-51
rs1039320	5	73927752	7150563	<i>ENC1</i>	1.02E-50
rs7522860	1	156275281	6550367	<i>PAQR6</i>	2.48E-50
rs11150882	17	259648	4640161	<i>C17orf97</i>	2.93E-50
rs10063135	5	180432024	1470102	<i>BTNL3</i>	5.74E-48
rs3088238	20	60880653	1570014	<i>OSBPL2</i>	7.99E-46
rs7328733	13	33126737	990020	<i>PFAAP5</i>	2.83E-43
rs2025577	1	156314104	2640086	<i>C1orf85</i>	2.96E-43
rs1872241	6	41195903	1690670	<i>TREML4</i>	5.57E-43
rs639459	7	42951936	4010072	<i>C7orf25</i>	1.73E-42
rs4360063	5	96293153	2260731	<i>ERAP2</i>	1.86E-42
rs1393261	16	55618451	4120367	<i>LPCAT2</i>	8.59E-42
rs10051931	5	94001476	1070736	<i>ANKRD32</i>	8.59E-42
rs1736219	17	17127471	1030551	<i>COPS3</i>	8.59E-42

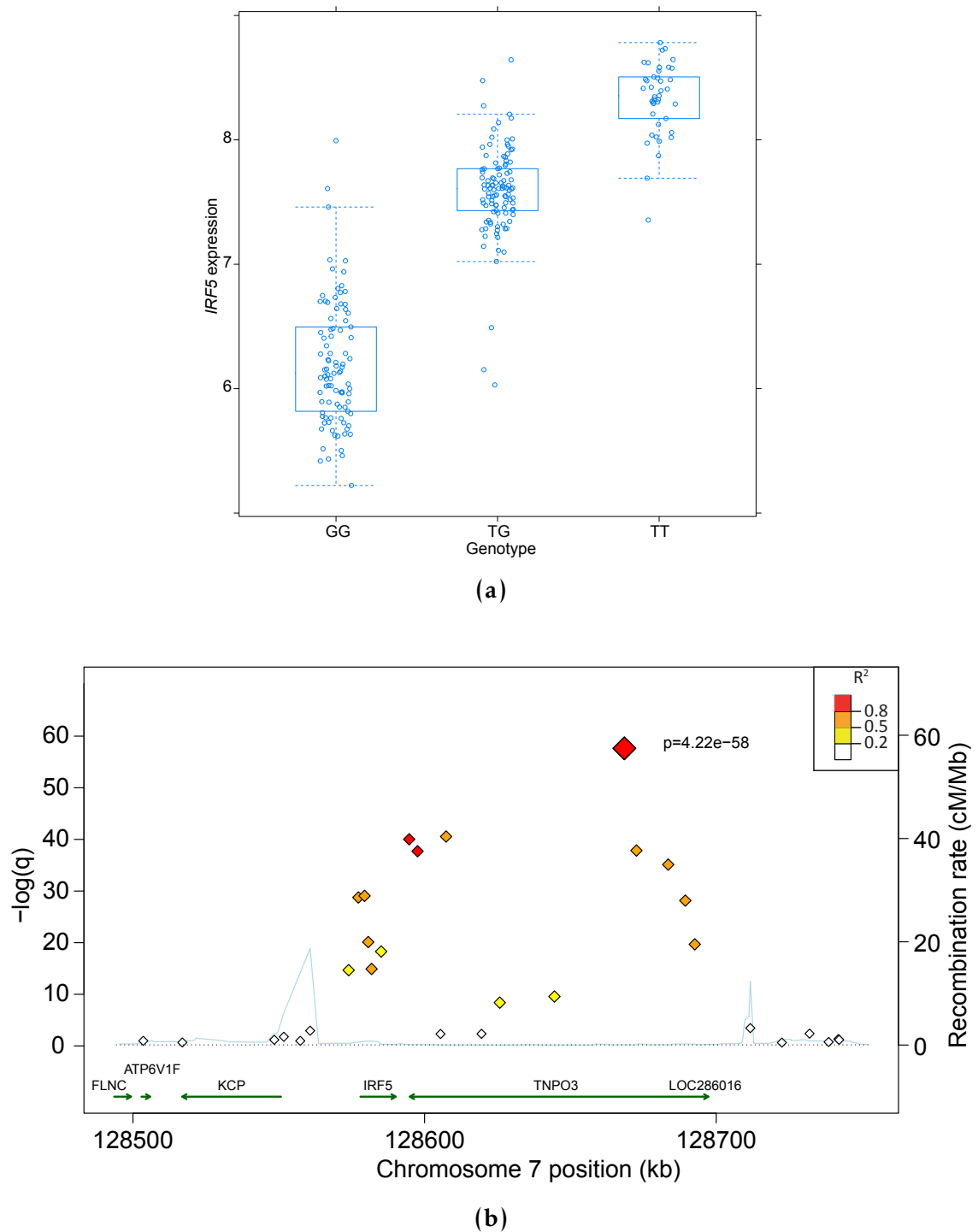


Figure 4.3: *IRF5* eQTL. rs10239340 is significantly associated with the expression of a probe mapping to *IRF5*. a) Expression of *IRF5* for each genotype of rs10239340. b) Regional association plot for the *IRF5* region with the FDR for genotyped SNPs plotted against their physical position on chromosome 7. rs10239340 is represented by the large red diamond. Estimated recombination rates from HapMap are shown. The SNP colour indicates LD with rs10239340.

4.2.6 Pathway analysis

Pathway enrichment for the 3,795 genes with evidence of a *cis* eQTL (FDR <0.05) was determined through IPA. 3,488 genes mapped in IPA and the enriched pathways are summarised in Figure 4.4 and Table B.2.

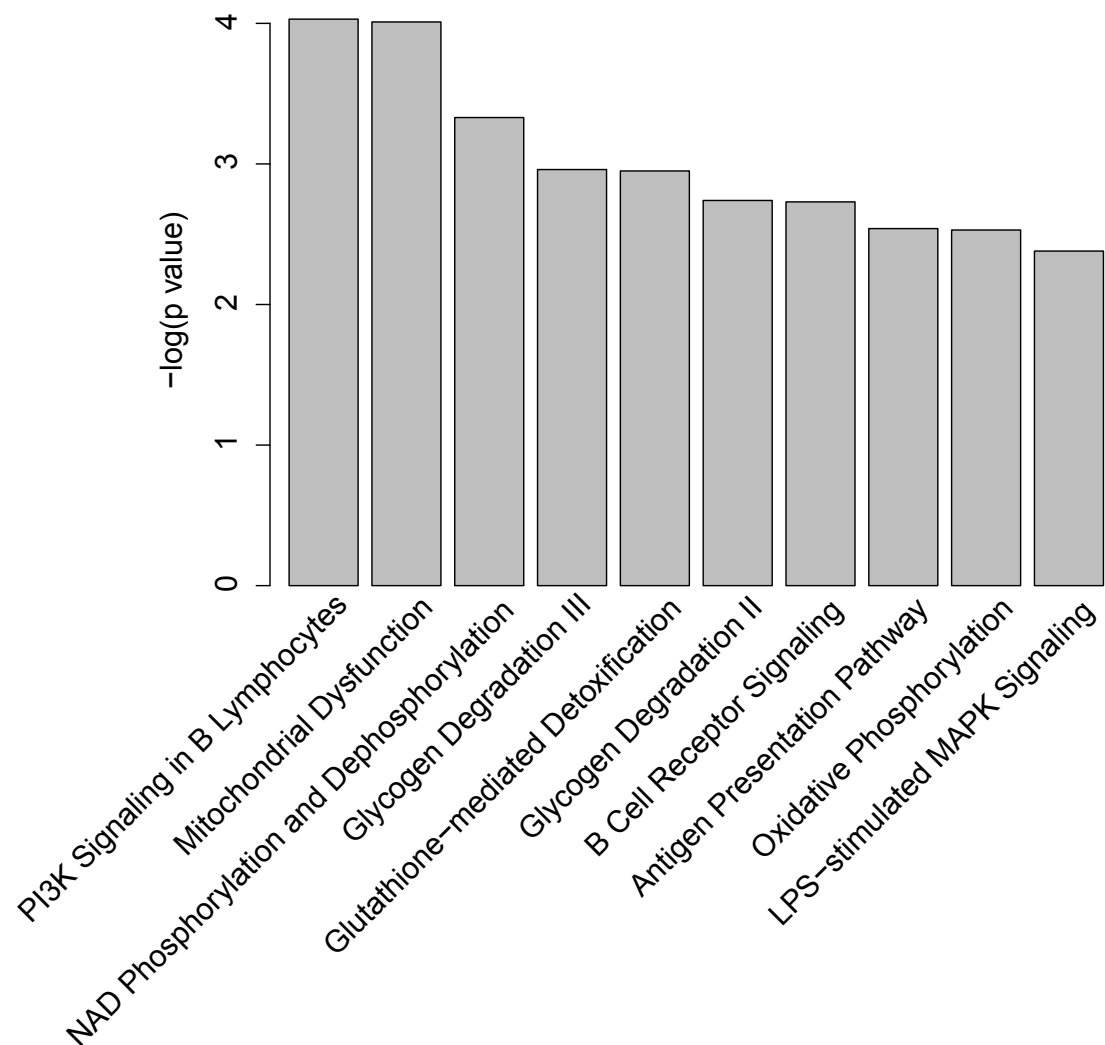


Figure 4.4: Pathway enrichment for genes with *cis* associations. Pathway enrichment for the 3,488 genes with evidence of a *cis* eQTL and which map to IPA. The ten most enriched pathways are shown.

4.2.7 Imputation

In order to compare the *cis* eQTL associations with GWAS and other eQTL datasets, missing and untyped SNPs were imputed with IMPUTE2 (Howie et al.

2009; Howie et al. 2011) against the 1000 Genomes Project Phase I v3 reference panel (1000 Genomes Project Consortium et al. 2012). Genotypes were imputed genome-wide but fine-mapping of eQTL regions were restricted to those with evidence of a *cis* eQTL from the genotyped data analysis. The rationale for this approach was the balance between increasing informative data to allow for comparisons with other datasets and reducing noise generated through the imputation process.

4.2.8 Integration with GWAS

Sepsis outcome

The sepsis outcome GWAS identified 16 associated SNPs for replication (Table 4.1). There was no evidence of a *cis* eQTL for these SNPs (FDR <0.05). However, five of the borderline significance SNPs (p value $<1 \times 10^{-4}$) showed evidence of an eQTL (based on the imputed *cis* eQTL results) (Table 4.3).

rs3199918 (chromosome 3 position 52581435) is associated with expression of *ITIH4*. *ITIH4* encodes a subunit of an inter- α -inhibitor protein (an anti-inflammatory protein) which has been identified as a marker of acute ischemic stroke (AIS) due to reduced expression in AIS patients (Kashyap et al. 2009). Administration of human inter- α -inhibitor proteins improves survival during sepsis in rat models (Yang et al. 2002).

Figure 4.5 summarises the regional association plot for the *ITIH4* *cis* eQTL. The SNP (rs3199918) from the GWAS association is represented by the large white diamond. This SNP is in low LD with the most significant SNP. However the pattern of LD in the region suggests further peaks may be involved and therefore additional investigation into this SNP should be carried out.

Regulatory genetic variants in severe sepsis

Table 4.3: Overlap between sepsis eQTL and GWAS for outcome. Five of the borderline significant SNPs (p value $<1 \times 10^{-4}$) from the GWAS for sepsis outcome also showed evidence of a *cis* eQTL in the sepsis cohort. OR = odds ratio.

Chr	Pos	SNP ID	Discovery		<i>cis</i> eQTL	
			p value	OR	FDR	gene
2	134078359	rs13414823	8.53E-05	0.64	0.031	<i>LYPD1</i>
2	134080719	rs58035542	6.47E-05	0.64	0.031	<i>LYPD1</i>
2	134080732	rs57942569	6.47E-05	0.64	0.031	<i>LYPD1</i>
3	52581435	rs3199918	8.66E-05	1.63	0.001	<i>ITIH4</i>
15	28356859	rs1129038	7.10E-06	0.49	0.003	<i>HERC2</i>

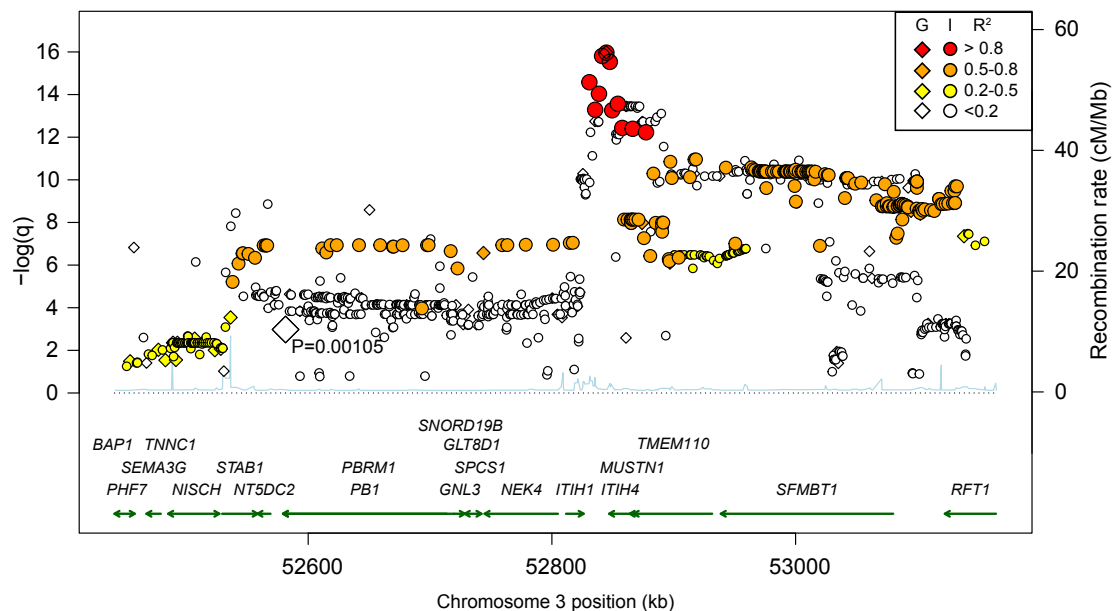


Figure 4.5: Regional association plot for *ITIH4* *cis* eQTL. The FDR for association with expression of *ITIH4* is plotted against the physical location on chromosome 3 for each SNP. rs3199918 is represented by the large white diamond. Estimated recombination rates from HapMap are shown. The SNP colour indicates LD with the most significantly associated SNP. Genotyped SNPs are represented by circles and imputed SNPs by diamonds.

GWAS catalog

The GWAS catalog provides a database of SNP-trait associations from published GWAS. As of July 2014, the database contained 17,080 entries (*A Catalog of Published Genome-Wide Association Studies*). Restricting these to trait associations

reaching genome-wide significance ($<5 \times 10^{-8}$) this corresponds to 5,010 unique SNPs.

Integrating the *cis* eQTL analysis with the GWAS catalog identified 743 of these SNPs that also showed evidence of a *cis* eQTL (FDR <0.05). These SNPs are associated with a range of traits but include many immune related diseases such as IBD, diabetes and systemic lupus erythematosus. The ten SNPs with the most significant *cis* eQTL and trait associations in the GWAS catalog are summarised in Table 4.4.

Table 4.4: Overlap between sepsis eQTL and GWAS catalog. The ten SNPs with the most significant *cis* eQTL and trait associations in the GWAS catalog are summarised. Chr = chromosome. Reported gene = gene reported by author. Mapped gene = nearest gene (both upstream and downstream if SNP is intergenic).

Pubmed ID	Disease/Trait	SNPs	Chr	Position	Reported Gene(s)	Mapped gene	<i>cis</i> eQTL		
							GWAS p value	FDR	Gene
21862451	Butyrylcholinesterase levels	rs4950806	1	201952574	RNPEP	RNPEP	9.00E-16	5.57E-36	RNPEP
20212171	C4b binding protein levels	rs3813948	1	207269858	C4BPB, C4BPA	C4BPB	4.00E-10	7.16E-32	C4BPA
21102463	Crohn's disease	rs2549794	5	96244549	ERAP2, LRAP	ERAP2	1.00E-10	2.21E-41	ERAP2
23128233	Inflammatory bowel disease	rs1363907	5	96252803	ERAP2, ERAP1, LNPEP	ERAP2	6.00E-13	4.3E-41	ERAP2
23128233	Crohn's disease	rs17695092	5	173337853	CPEB4	CPEB4	5.00E-09	2.83E-52	CPEB4
20935629	Waist-hip ratio	rs6861681	5	173362458	CPEB4	CPEB4	2.00E-09	2.79E-52	CPEB4
21768215	Urate levels	rs606458	11	64546391	MEN1, SF1, MAP4K2, PYGM, RASGRP2, CDC42BPG, NRXN2	SF1	6.00E-11	8.39E-34	SF1
23563607	Height	rs718444	13	33147052	APRIN	N4BP2L2-IT2, PDS5B	2.00E-10	1.05E-39	PFAAP5
20881960	Height	rs7332115	13	33147548	PDS5B, BRCA2	N4BP2L2-IT2, PDS5B	6.00E-10	1.05E-39	PFAAP5
23408906	Thyroid hormone levels	rs6499766	16	55604128	LPCAT2, CAPNS2	LPCAT2	5.00E-08	3.21E-44	LPCAT2

4.2.9 Context specific *cis* associations

Healthy volunteers

Westra et al. (2013) have performed an eQTL meta-analysis for 5,311 healthy individuals with replication in a further 2,775. The majority of the studies in the Westra discovery cohort generated genome-wide gene expression data using the Illumina-Human-HT12 v3 Expression BeadChip which is the previous version to the one used here for the sepsis study. In order to identify genes which show evidence of a *cis* eQTL unique to the context of sepsis, the comparison analysis was limited to genes with probes available on both Illumina platforms. 14,269 genes were used for the eQTL analysis in sepsis and 13,589 of these have a probe available on HT12 v3.

Figure 4.6 summarises the overlap of *cis* eQTL between the sepsis and healthy volunteer cohort. Genes which did not show evidence of a *cis* eQTL in this much larger healthy cohort were defined as unique to sepsis (1,508 genes). Pathway enrichment is summarised in Figure 4.7 and Table B.3.

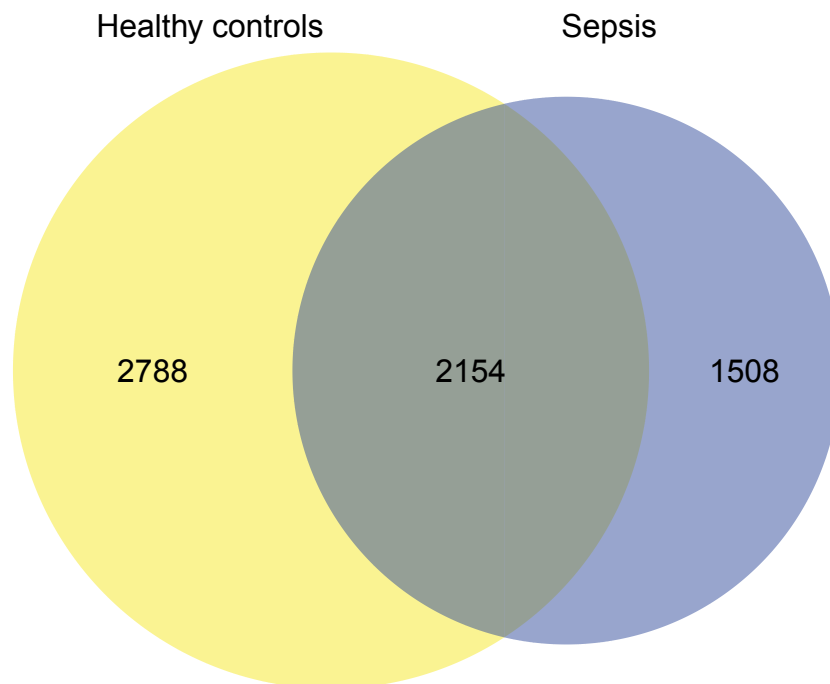


Figure 4.6: Genes with *cis* eQTLs unique to sepsis. The genes showing evidence of a *cis* eQTL in the sepsis cohort compared to the meta-analysis conducted by Westra et al. (2013) for healthy individuals. Genes which did not show evidence of a *cis* eQTL in this much larger healthy cohort were defined as unique to sepsis (1,508 genes).

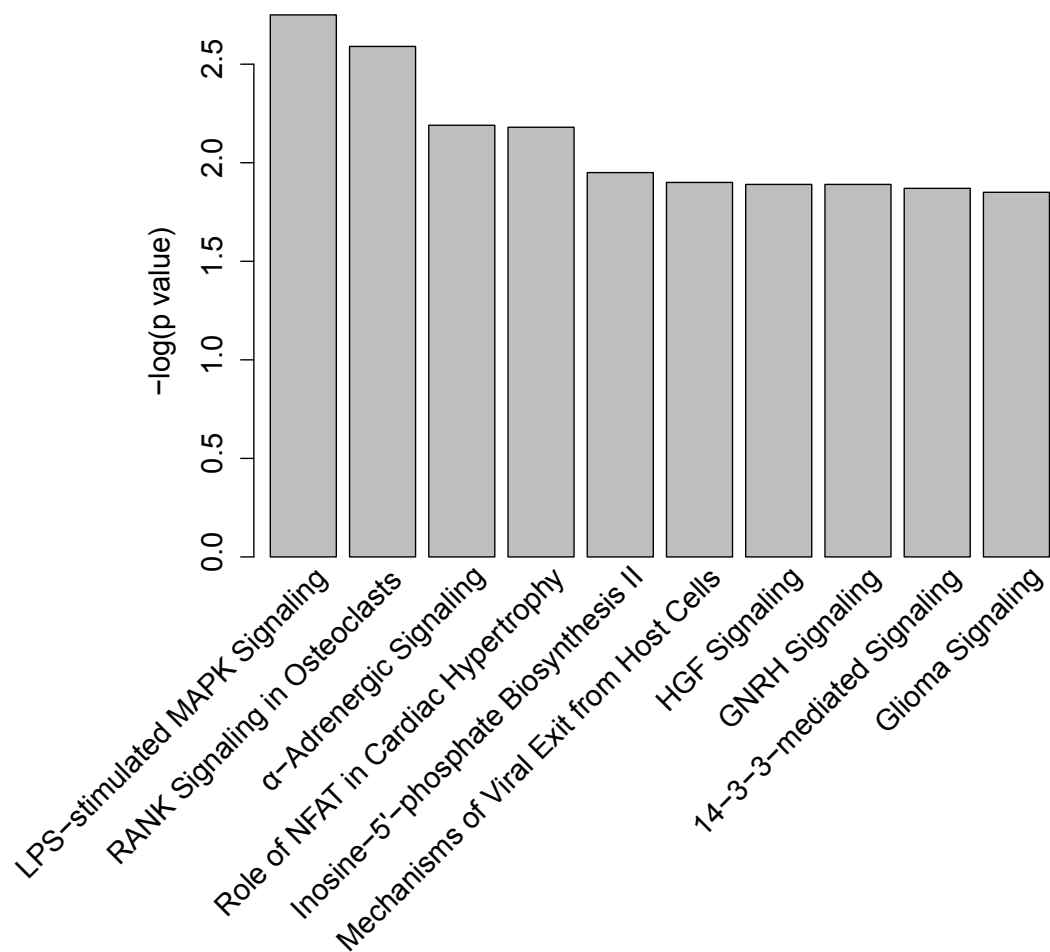


Figure 4.7: Pathway enrichment for genes with *cis* associations unique to sepsis. The ten most enriched pathways are shown.

The most enriched pathway for *cis* eQTL unique to sepsis was LPS-stimulated MAPK signalling with 12 of the 73 genes in the pathway showing a context specific eQTL. This suggested an interesting comparison could be made with a study recently conducted in our group highlighting context specific eQTL following innate immune stimulation (Fairfax et al. 2014).

Innate immune stimulation

Fairfax et al. (2014) have identified context specific eQTL when monocytes from healthy individuals were exposed to either LPS or $\text{IFN}\gamma$ which represent innate immune stimuli. 2,910 genes showed evidence of a *cis* association in at least

one of the stimulated states (LPS for 24 hours or IFN γ) but not the naïve state. The same Illumina platform was used to generate gene expression data for both studies but following QC and normalisation, 10,522 genes were measured in both studies. The comparison was restricted to these genes and the overlap between the studies is summarised in Figure 4.8.

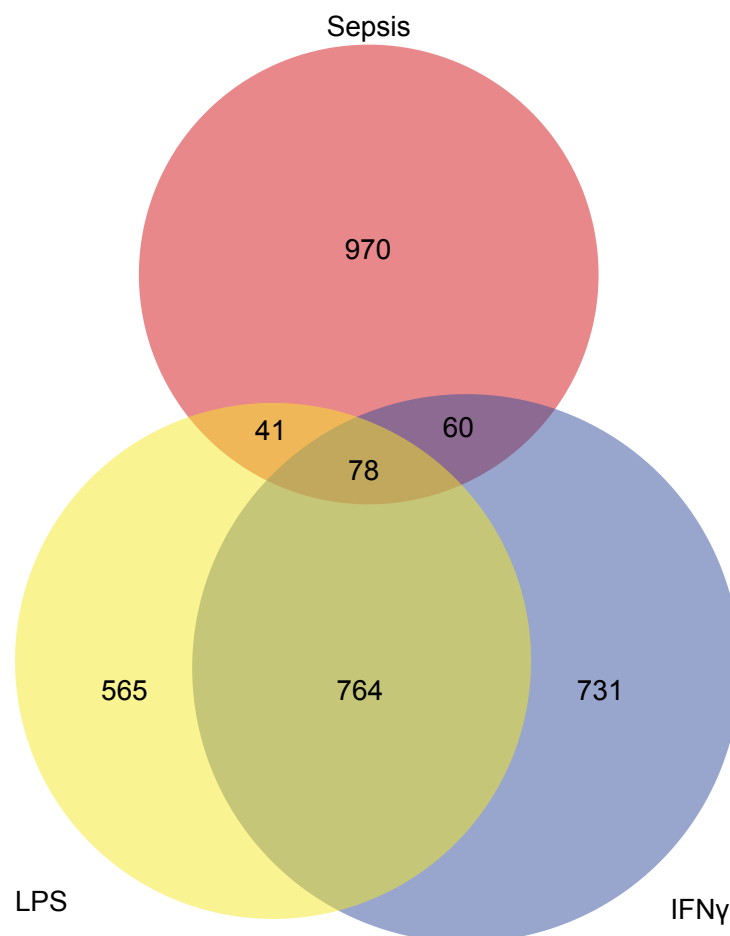


Figure 4.8: Overlap of genes with *cis* associations between sepsis and stimulated monocytes.

4.2.10 *Trans* associations

eQTL mapping was conducted to identify associations between probes and SNPs greater than 1Mb from the probe start site, classified as *trans* associations.

Greater power is needed to detect *trans* eQTLs (Westra et al. 2013; Fairfax and Knight 2014). The analysis was therefore restricted to the identification of master regulatory regions by using genes which have shown evidence of a *cis* eQTL and using genotyped SNPs only. A master regulatory region is suggested if a SNP associates with many genes in *trans* and with a gene in *cis*. This *cis* gene is then a likely upstream causative gene.

3,673 SNP-probe *trans* associations were identified (FDR <0.05). 112 of the *trans* SNPs showed evidence of a *cis* eQTL (FDR <0.05). The number of genes associating with these SNPs in *trans* plus the gene associating in *cis* are summarised in Tables 4.5 and B.4. SNP rs752845 shows the greatest number of *trans* associations (206 genes) and the *cis* association is with expression of *PIK3CD*. *PIK3CD*, encoding a phosphoinositide 3-kinase, phosphorylates inositol lipids and is involved in the immune response (Pruitt et al. 2014).

Table 4.5: Top master regulators of *trans* associations. 112 of the SNPs identified from the *trans* eQTL analysis also showed evidence of a *cis* eQTL. For each SNP, the number of genes associating in *trans* plus the genes associating in *cis* are summarised. The top master regulators (with at least six *trans* gene associations) are shown.

SNP	No. <i>trans</i> genes	<i>Cis</i> gene	<i>Cis</i> FDR
rs752845	206	<i>PIK3CD</i>	3.14E-05
rs4713488	127	<i>AGPAT1</i>	7.68E-09
rs951954	71	<i>BUB1</i>	4.31E-02
rs9517411	68	<i>UBAC2</i>	2.18E-03
rs2220679	54	<i>SPAG1</i>	4.26E-02
rs16904945	33	<i>PHF20L1</i>	3.17E-03
rs12148454	33	<i>VPS18</i>	1.63E-04
rs6748986	19	<i>EIF4E2</i>	2.89E-04
rs889518	14	<i>ATMIN</i>	3.72E-02
rs7915040	10	<i>EIF4EBP2</i>	3.75E-03
rs367146	10	<i>ITFG3</i>	5.79E-04
rs12397260	9	<i>ARSD</i>	8.39E-03
rs5983130	9	<i>ARSD</i>	1.19E-05
rs2534123	9	<i>ZFY</i>	1.98E-37
rs6725242	8	<i>PDIA6</i>	4.02E-02
rs7576288	8	<i>PDIA6</i>	4.02E-02
rs9786145	8	<i>PRKY</i>	1.07E-11
rs16930700	7	<i>CAMK2G</i>	1.21E-04
rs9786814	7	<i>EIF1AY</i>	3.43E-15
rs16986572	7	<i>FCAR</i>	2.83E-03
rs2874089	7	<i>UTY</i>	6.29E-10
rs7892889	7	<i>UTY</i>	7.95E-10
rs9786758	6	<i>JARID1D</i>	1.13E-13
rs2713254	6	<i>UTY</i>	2.87E-05

4.2.11 Comparison to cluster membership

The sepsis gene expression chapter identified two clusters within the cohort (based on gene expression signatures) which showed differences in survival rates (Section 3.2.12). The genes differentially expressed between the two clusters were compared to the results of the *cis* eQTL analysis. The comparison was restricted to the 14,269 genes which were used to determine the eQTL associations. 676 of these genes were differentially expressed between clusters and showed

enrichment for genes with evidence of a *cis* eQTL (Figure 4.9a) providing evidence of a genetic component underlying the observed heterogeneity within the cohort. However, the significance of this overlap may be overestimated due to the increased chance of a gene with great variation in expression being identified as both a differentially expressed gene and an eQTL compared to a gene with lower variation in expression. A similar comparison was conducted for the *cis* eQTLs identified as unique to the context of sepsis. The comparison was restricted to the 13,589 genes measured in both the sepsis and healthy control cohort. The 659 genes differentially expressed between clusters did not show enrichment for *cis* eQTL unique to sepsis (Figure 4.9b).

PLINK (Purcell et al. 2007) was used to investigate any associations between genotype and cluster 2 membership. The analysis was restricted to the 258 genes which were differentially expressed between clusters and showed evidence of a *cis* eQTL. The most associated SNP from each eQTL was used but none reached statistical significance after adjustment for multiple testing.

47 pathways were enriched for differentially expressed genes between clusters. Only five of these overlapped with pathways enriched for eQTL (73 pathways) and only two overlap with the pathways enriched for eQTL unique to sepsis (37 pathways). This overlap is not greater than that expected by chance.

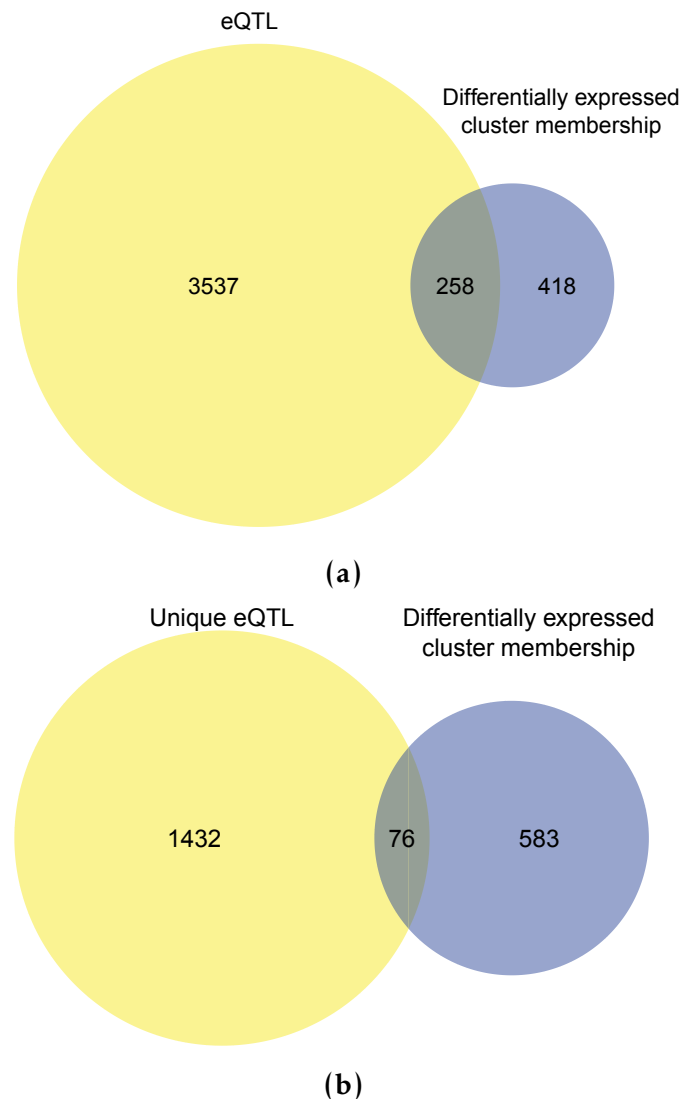


Figure 4.9: Enrichment of eQTL for genes differentially expressed between clusters. Two clusters were identified in the sepsis cohort based on gene expression signatures which show different survival rates. a) Genes differentially expressed between clusters are significantly enriched for *cis* eQTL (chi squared p value <0.0001). b) Genes differentially expressed between clusters are not significantly enriched for *cis* eQTL which are unique to the context of sepsis.

4.3 Discussion

4.3.1 Integration with GWAS

One of the challenges in interpreting the results of a GWAS is understanding how an associating variant may play a role in the disease of interest, particularly

when the variant lies in an intergenic region. Integrating eQTL analysis can provide insights into the molecular mechanisms underlying the association (Cookson et al. 2009). It can also assist with the identification of any potentially interesting variants which did not quite reach genome-wide significance. This was demonstrated by the integration of the sepsis eQTL with the results of the GWAS for sepsis outcome. The GWAS did not identify any regions which reached genome-wide significance but the eQTL highlighted a potentially interesting variant which showed evidence of an eQTL with *ITIH4*, a molecule which has previously been linked to a study of sepsis in rat models (Yang et al. 2002).

4.3.2 Context specific *cis* associations

Catalogs of eQTL analysis results are becoming publicly available which allows comparisons to be made between regulatory variants observed in different cells and tissues (Fairfax et al. 2012; Westra et al. 2013; Fairfax et al. 2014; GTEx Consortium 2013). An initial comparison to a cohort of healthy individuals (Westra et al. 2013) revealed a substantial number of genes with evidence of eQTL unique to the context of sepsis. This supports the hypothesis that eQTL may be context, as well as cell or tissue specific.

Pathways analysis of the genes with eQTL unique to sepsis highlighted the importance of signalling pathways in response to LPS. The data were compared with the results of an eQTL analysis following innate immune stimulation of monocytes (Fairfax et al. 2014). Only a small proportion of genes overlapped between the two analyses. It is likely that many of these are the result of comparing leukocyte fractions to a specific cell type but as further cell type specific data sets become available, the non-overlapping genes can be examined to determine whether other cell types are contributing to the regulatory variants identified in the context of this disease.

4.3.3 Limitations

This analysis was conducted on a mixed white cell population. While evidence is building for both tissue and cell specific eQTL this study demonstrates that experiments from blood can still give general insights into the regulatory variants important in the context of disease. These can then be used to generate specific hypotheses and help design follow up experiments to further characterise putative interactions in specific cell types (Idaghdour and Awadalla 2012).

The large cohort of healthy individuals provide an appropriate comparison for the identification of context specific eQTL but the perfect control would be gene expression data from a timepoint before sepsis (pre-disease state) for each individual. This approach is not currently feasible for sepsis but has shown success in the field of vaccine response. Franco et al. (2013) identified 20 genes which contributed to the differential immune response to vaccine.

In addition to eQTL specific to the context of sepsis, it would also be informative to understand any regulatory variants which are important for subgroups within the cohort such as survival or microbiology status. The cohort, however, is not powered to detect these.

4.3.4 Conclusion

eQTL were successfully mapped in a cohort of patients with severe sepsis. Integrating the analysis with results from a GWAS for sepsis outcome has provided evidence for replication of variants which did not quite reach genome-wide significance. Comparisons to other eQTL datasets have highlighted genes with context specific eQTL which may assist in understanding the genetic aetiology of sepsis.

Chapter 5

Functional genomic response to influenza challenge

5.1 Introduction

Influenza viruses are highly infectious RNA viruses that can cause acute respiratory illness. A significant proportion of patients with severe sepsis due to CAP (such as described in Chapter 3) will have developed a post-influenza bacterial pneumonia (Metersky et al. 2012). For clinicians, identifying the invading organism responsible for infection during sepsis remains a challenge and it is difficult to assess the timepoint of infection at which the patients are admitted to the ICU. Influenza challenge studies are a method frequently used to test influenza vaccine efficacy but they can also provide valuable data for understanding how infection, resulting from a defined challenge, develops over time.

The functional genomic host response to influenza has not been well studied in humans (Bortz and García-Sastre 2011). The work described in this chapter sought to characterise this further through an influenza vaccine and challenge study followed by integration with gene expression signatures identified in a cohort of severe sepsis patients with confirmed viral infection (Chapter 3).

5.1.1 Influenza

In adults and adolescents, influenza typically presents with clinical symptoms that include fever, headache, sore throat, myalgias and runny nose (Cox and Subbarao 1999). Many hosts, however, show no symptoms when exposed to comparable levels of virus and therefore exhibit a higher degree of tolerance (Carrat et al. 2008). During an influenza challenge study consenting subjects are vaccinated, quarantined and then inoculated with influenza via intranasal administration. This approach has the advantage that potentially susceptible individuals can be positively selected based on low pre-existing immunity to the challenge virus. The sample size is however restricted by the capacity of quarantine facilities and the high associated costs of performing a challenge experiment.

The accurate post-challenge classification of subjects by outcome is critical to determining vaccine efficacy. These outcomes are based on scoring systems with artificial thresholds for "symptomatic" and "asymptomatic" phenotypes and include subjective self-reported observations. Additionally, the biological evidence for infection (viral shedding and post-challenge serology) are often discordant. Increasing our understanding of susceptibility to the virus, the molecular basis of viral pathogenesis and the immune response will be crucial for the development of new vaccines and targeted diagnostics (Bortz and García-Sastre 2011).

5.1.2 Influenza vaccine and challenge study

Our collaborators have recently described a T cell inducing vaccine called MVA-NP+M1 (Berthoud et al. 2011). To test the efficacy of this vaccine they conducted an influenza vaccine and challenge study (FLU002) with 22 healthy individuals (Lillie et al. 2012). After enrolment, 11 volunteers were vaccinated with MVA-NP+M1. Thirty days after vaccination, all 22 volunteers underwent intranasal

challenge with H3N2 influenza. Self-reported symptoms were recorded twice daily from 12 hours post-challenge for six days. Additionally, whole blood was collected in PAXgene tubes prior to influenza challenge and then at three further time points over the following 48 hours (12, 24 and 48 hours post-challenge) (Materials and Methods).

In order to study the effects of influenza challenge on host biology, and to examine the impact of vaccination with MVA-NP+M1, genome-wide gene expression profiles were generated from samples taken during the FLU002 study. This chapter will describe the nature of the transcriptional response to influenza and describe how this information could be used to improve the biological classification of influenza challenge subjects. These results have implications for our understanding of the host-pathogen interactions for influenza, and have the potential to inform clinical diagnostic testing. This is of particular interest in the context of the sepsis cohort given the 2009 H1N1 influenza pandemic with a number of recruited patients (16 out of 265) shown to have H1N1 infection (Chapter 3).

5.1.3 Aims

The aim of this chapter is to investigate the functional genomic response to influenza challenge and relate this to gene expression signatures in severe sepsis (Chapter 3). Viral infection, notably with influenza, is commonly associated with bacterial pneumonia in patients with severe sepsis. I aim to define the transcriptomic response to influenza for healthy volunteers undergoing influenza challenge as part of a vaccine study and use this information to

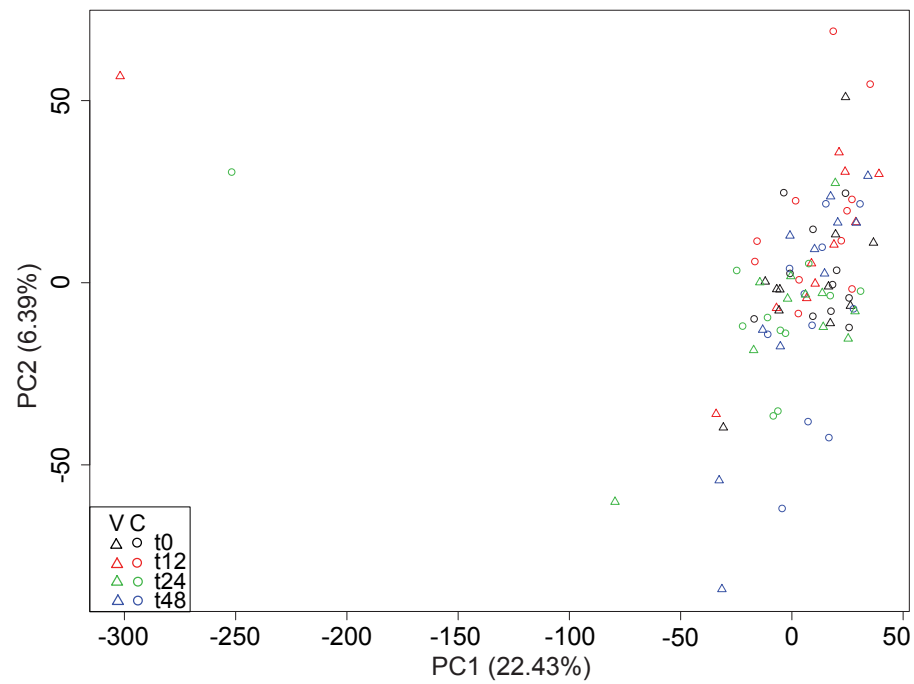
1. define differentially expressed gene sets and pathways specific to influenza infection
2. relate any observed influenza signatures to data from the sepsis cohort (Chapter 3) where patients have laboratory confirmed viral infection

5.2 Results

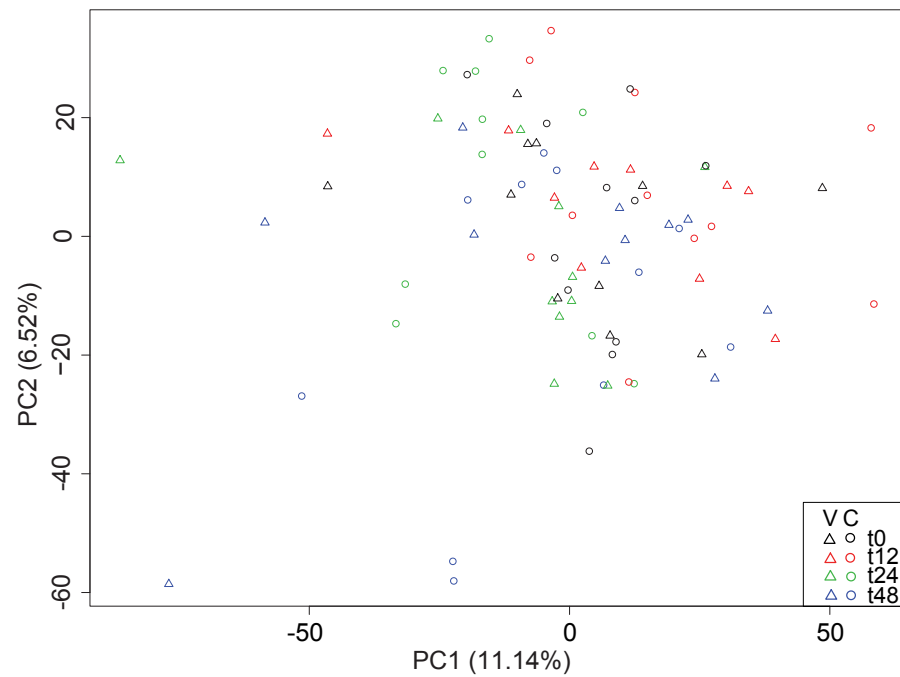
5.2.1 RNA extraction and gene expression microarray QC

For each volunteer, RNA was extracted from whole blood at four time points using the PAXgene Blood RNA Kit. This was followed by globin mRNA depletion. Genome-wide gene expression analysis was carried out using the Illumina Human-HT-12 v4 Expression BeadChip gene expression platform comprising 47,231 probes.

Data backgrounds were subtracted and the raw data were transformed and normalised using the Variance Stabilization and Normalization method (Huber et al. 2002). Following QC, including PCA, two samples were excluded as outliers (Figure 5.1). 11,158 probes were removed as they did not have a detection p value ≥ 0.95 in at least one sample leaving 36,073 probes for further analysis.



(a) 88 samples



(b) 86 samples

Figure 5.1: Principal component analysis. For each sample, the first two PCs are plotted with the proportion of variance explained by each component. V = vaccinee; C = control. a) When all 88 samples are included in the analysis, two are identified as clear outliers. b) The removal of the two outlying samples reduces the variation in the dataset.

5.2.2 Differential expression for laboratory confirmed influenza

Following influenza challenge, seven of the 22 subjects developed LCI as defined in the clinical trial protocol (Lillie et al. 2012) (Figure 5.2).

		Symptoms				
		None	Mild	Mod/Sev		
Shedding	-	I U	B F S		-	Seroconversion
	-	A E H J O Q	C V	L	+	
	+	K	M R		-	
	+		G	D N P T	+	

Figure 5.2: Study subjects clinical observations. Individuals are defined as having no symptoms, mild or moderate/severe based on the summed self-reported symptoms collected over 6 days. Positive shedding indicates viral shedding was detected on at least 1 day after challenge. Seroconversion was determined by HI titre using serum samples obtained at 26 days post-challenge. Laboratory confirmed influenza is defined as mild or moderate/severe symptoms and viral shedding (shown by grey shaded area). Vaccinated individuals are shown in red.

Differential gene expression was analysed using a linear regression model with an empirical Bayes approach to define the likelihood of observed differences in gene expression given overall variance in the dataset. Implementing this approach using the Limma package (Smyth 2004), a total of 68 probes (62 genes) were found to be differentially expressed (FDR <0.05 and fold change >1.2)

between the seven volunteers with LCI and the remaining subjects at 48 hours post-challenge (Table 5.1, Figure 5.3 and Table C.1). Using the same approach, no probes were differentially expressed at 12 or 24 hours post-challenge.

Table 5.1: Differentially expressed probes for LCI. The number of probes with FDR <0.05 and fold change >1.2 for each comparison. The pre-challenge expression value for each subject was included in the limma model for each comparison.

Comparison (no. subjects)	No. probes	No. genes
12 hours post-challenge LCI (7) vs not LCI (15)	0	0
24 hours post-challenge LCI (7) vs not LCI (15)	0	0
48 hours post-challenge LCI (7) vs not LCI (15)	68	62

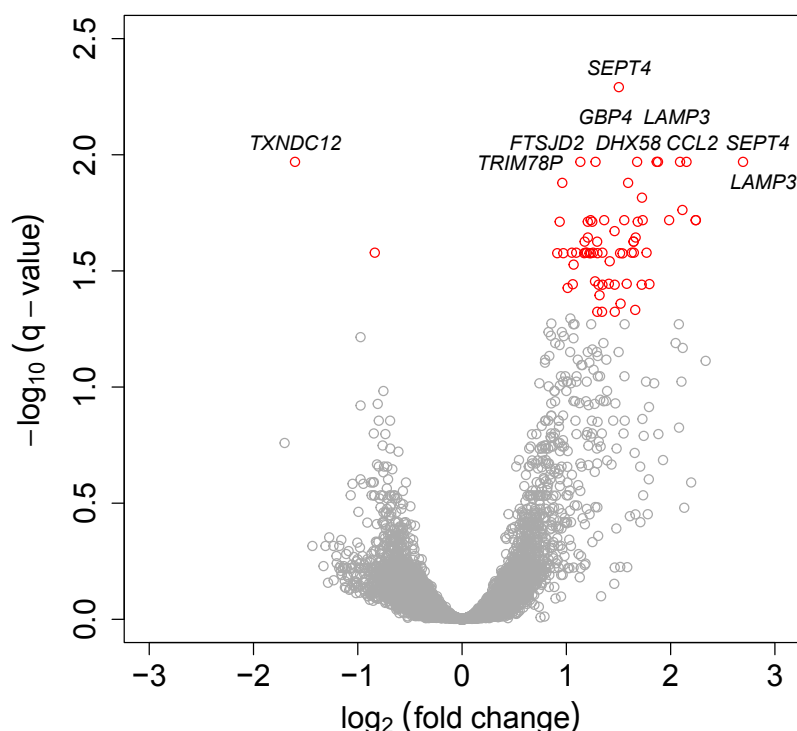


Figure 5.3: Volcano plot of differentially expressed probes for LCI. The significance of differential expression is plotted against fold change. Each point represents a single probe. Red points highlight probes with FDR <0.05 and fold change >1.2. A positive fold change indicates upregulation in the seven LCI samples 48 hours post-challenge compared to the other samples. The most significant differentially expressed probes are labelled with their corresponding gene names.

The most significant differentially expressed probe corresponds to the gene *SEPT4*. Plotting the expression values over time for this probe revealed that the difference in expression appears to be due to a subgroup of the LCI subjects rather than all seven (Figure 5.4).

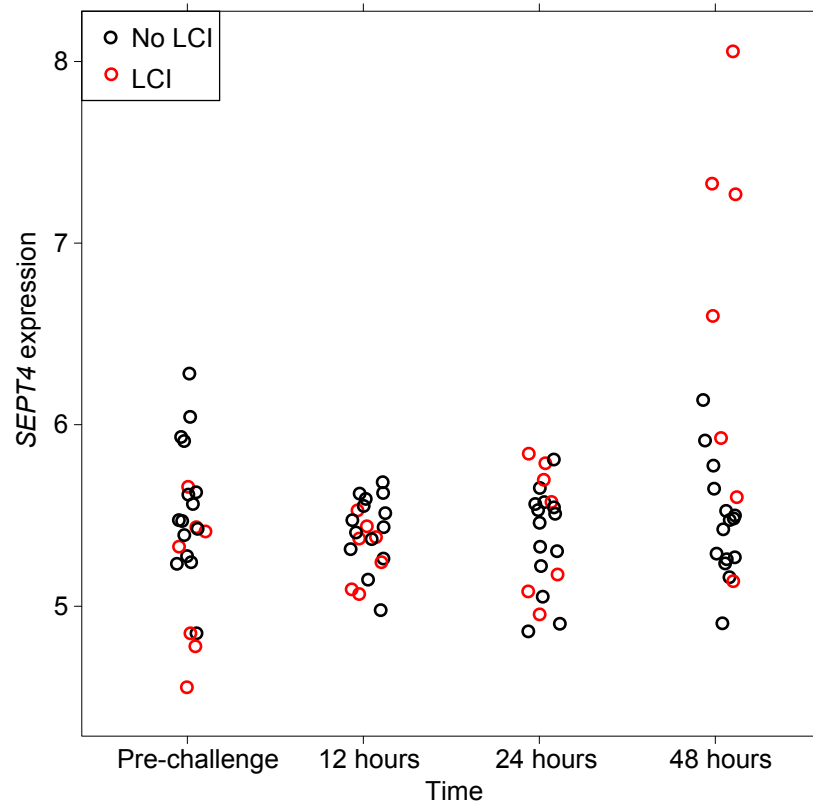


Figure 5.4: *SEPT4* expression over time. Expression of *SEPT4* over time with subjects distinguished by development of LCI.

5.2.3 Clinical observations and variation in gene expression

As illustrated by Figure 5.4, three of the LCI subjects do not appear to be contributing to the differential expression observed 48 hours post-challenge between LCI and non-LCI subjects. In addition, within the LCI group there was considerable variation in the summed modified Jackson scores (ranging from 4 to 38) describing the extent to which volunteers were symptomatic. It was therefore hypothesised that these differences in clinical phenotype might be reflected in patterns of gene expression.

Functional genomic response to influenza challenge

PCA and hierarchical clustering was performed to visualise variation in the dataset (Figure 5.5). This demonstrated that four subjects shared a similar profile of variance in gene expression 48 hours post-challenge. These four subjects were the only cases of LCI whose symptom scores were moderate/severe rather than mild and were positive for seroconversion (detection of specific antibody) as shown in Figure 5.2.

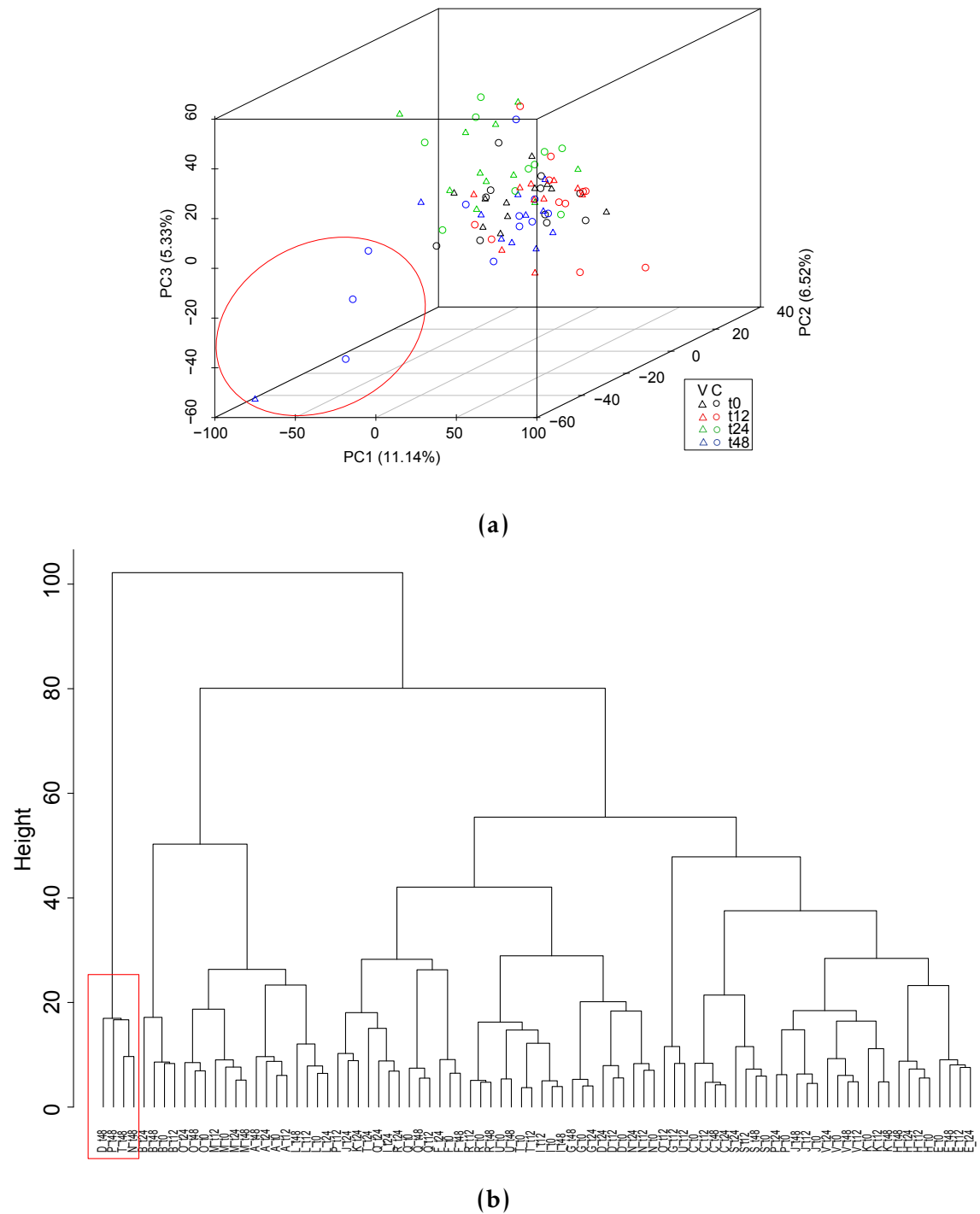


Figure 5.5: Principal component and hierarchical clustering plot. a) The first three PCs are plotted with the proportion of variance explained by each component. Samples from four subjects with moderate/severe LCI 48 hours post-challenge cluster together and show a different pattern of variance to the rest of the samples (highlighted by the red oval). b) Coefficient of variation was calculated to identify the 10% of probes that were most variable in the dataset (3,608 probes). The hierarchical tree, based on these probes, illustrates the relationship between samples. The height of the branches indicates the strength of the separation between individuals. The red rectangle highlights the four moderate/severe LCI samples 48 hours post-challenge.

5.2.4 Differential gene expression

Conducting differential gene expression analysis using the Limma package (Smyth 2004) a total of 1,103 probes were found to be differentially expressed between these four volunteers and the remaining subjects at 48 hours post-challenge (FDR <0.05 and fold change >1.2) (Table 5.2, Figure 5.6 and Table C.2). A description of the ten most significant differentially expressed genes is provided in Table 5.3. A number of the genes such as *SPATS2L* and *ANKRD22* do not have a known *in vivo* function but many, including *CCL2* and *USP18*, are linked to various components of the immune response.

As observed in the previous analysis, no probes were differentially expressed at 12 hours or 24 hours post-challenge. This specific effect on differential expression at 48 hours post-challenge is consistent with the observed temporal onset of symptoms (Lillie et al. 2012). The three mild LCI subjects showed no dynamic difference in gene expression compared to the subjects without LCI suggesting a similar gene expression profile across these groups.

In further analyses the 100 most significant differentially expressed probes (9%) were selected and the majority were found to be upregulated at 48 hours post-challenge (Figure 5.7). In contrast, subjects with LCI and mild symptoms did not show substantial changes in expression of the same probes. To further validate these findings, two alternative methods based on a two-step regression approach (maSigPro, Conesa and Nueda R package version 1.28.0) and on multivariate empirical Bayes statistics (timecourse, Tai 2007) were employed to generate an equivalent list of the 100 most significant differentially expressed probes. The overlap with the original probe list was 82% and 70% respectively (Figure 5.8).

Table 5.2: Differentially expressed probes for severe LCI. The number of probes with FDR <0.05 and fold change >1.2 for each comparison. The pre-challenge expression value for each subject was included in the limma model for each comparison.

Comparison (no. subjects)	No. probes	No. genes
12 hours post-challenge severe LCI (4) vs not (18)	0	0
24 hours post-challenge severe LCI (4) vs not (18)	0	0
48 hours post-challenge severe LCI (4) vs not (18)	1103	952

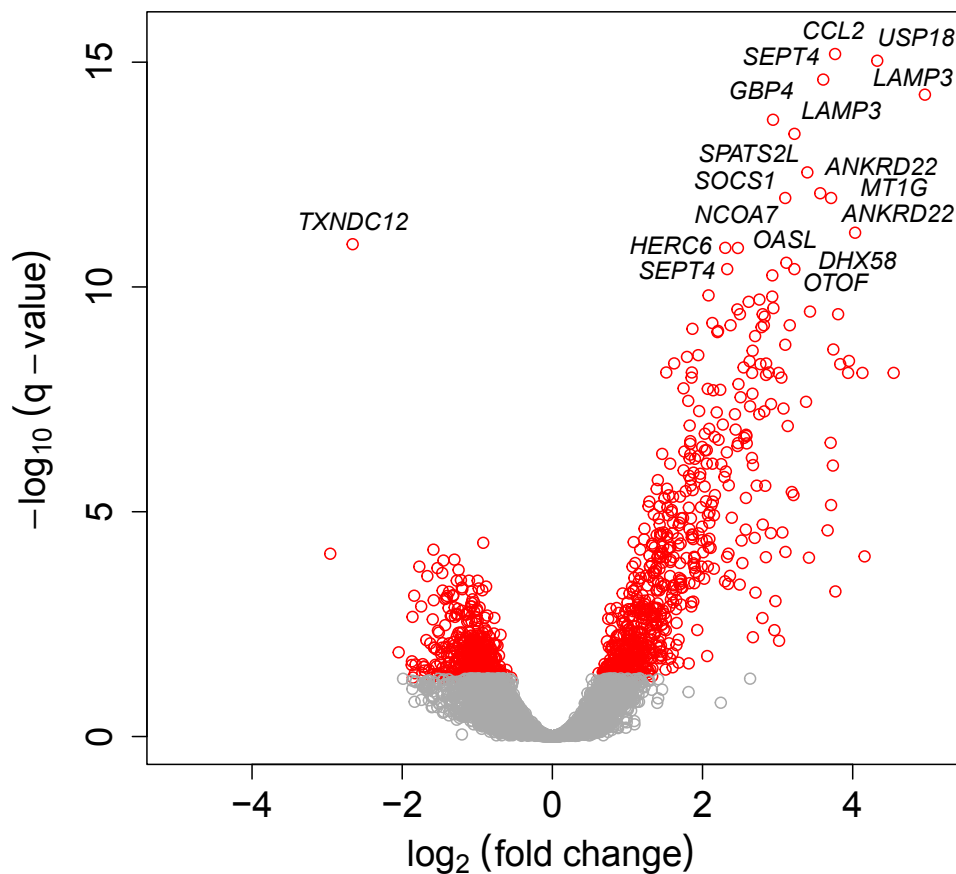


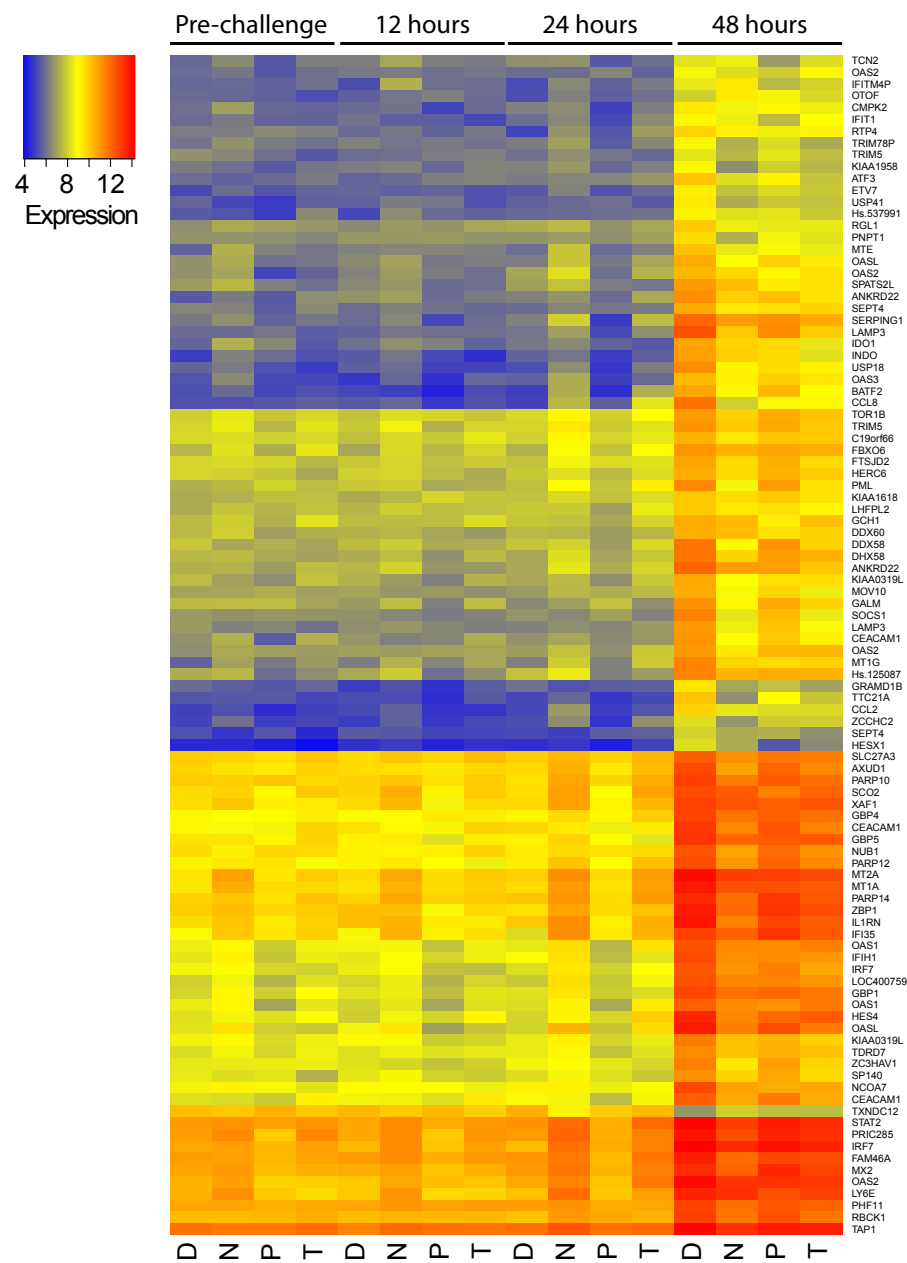
Figure 5.6: Volcano plot of differentially expressed probes for severe LCI. The significance of differential expression is plotted against fold change. Each point represents a single probe. Red points highlight probes with an adjusted p-value <0.05 and fold change >1.2. A positive fold change indicates upregulation in the four moderate/severe LCI samples 48 hours post-challenge compared to the other samples. The most differentially expressed probes are labelled with their corresponding gene names.

Functional genomic response to influenza challenge

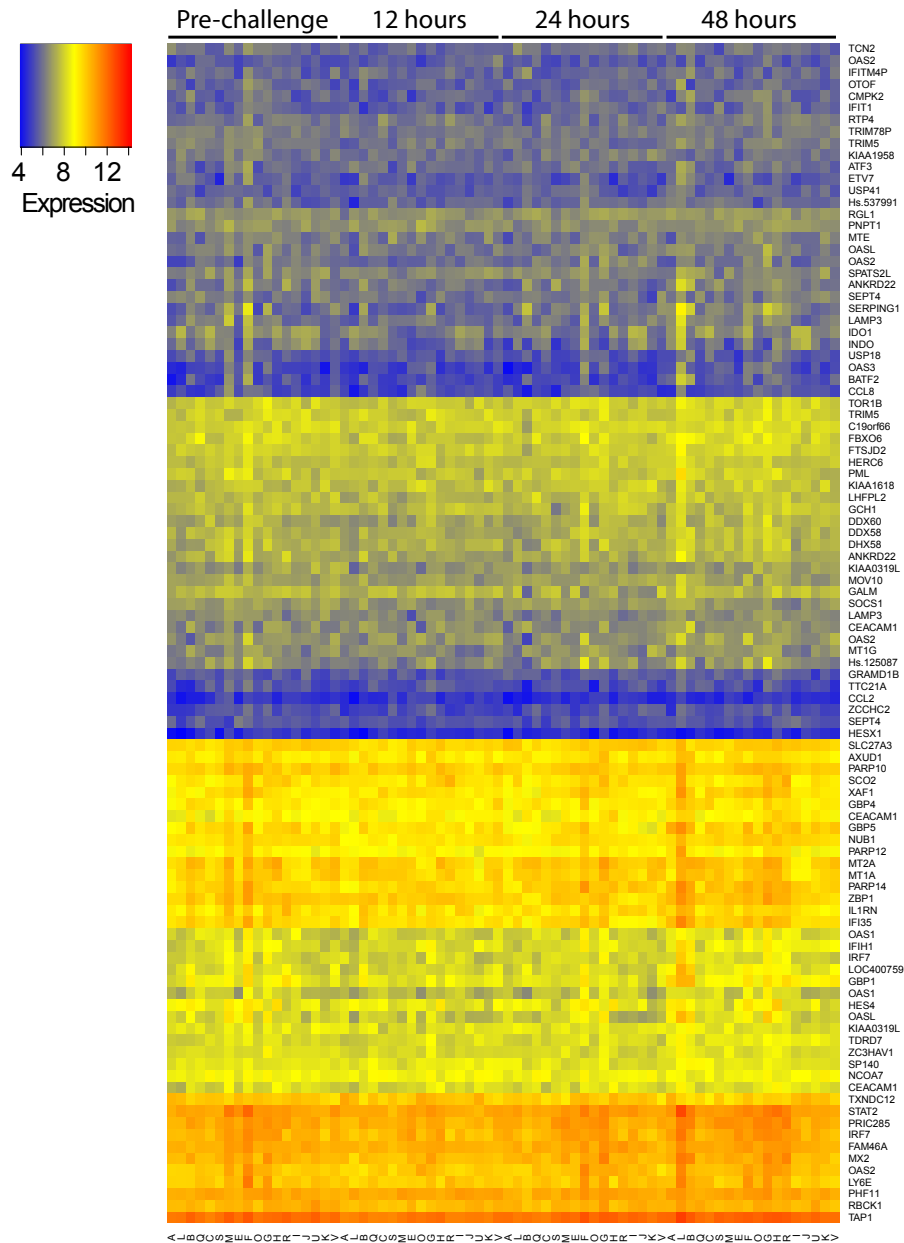
Table 5.3: Description of differentially expressed genes. A description of the 10 most significant differentially expressed genes between the four moderate/severe LCI samples 48 hours post-challenge compared to the remaining samples.

Gene	Description
<i>CCL2</i>	Encodes a CC chemokine family cytokine. It displays chemotactic activity for monocytes and basophils and binds to chemokine receptors CCR2 and CCR4 (Pruitt et al. 2014).
<i>USP18</i>	Encodes a ubiquitin-specific proteases (UBP) family member (cleaves ubiquitin from ubiquitinated protein substrates). Knockout mice are hypersensitive to interferon suggesting a role in downregulating interferon responses (Pruitt et al. 2014).
<i>SETP4</i>	Encodes a member of the septin family of nucleotide binding proteins. Involved in cell division regulation and disruption of function results in incomplete cytokinesis leading to large multinucleate cells (Pruitt et al. 2014).
<i>LAMP3</i>	Encodes a glycoprotein which is expressed when immature dendritic cells differentiate into interdigitating dendritic cells (Saint-Vis et al. 1998).
<i>GBP4</i>	Encodes a guanylate-binding protein induced by interferon and hydrolyses GTP to GDP and GMP (Vestal 2005).
<i>SPATS2L</i>	Encodes a protein of unknown <i>in vivo</i> function but has been identified as a novel bronchodilator response gene in an asthma GWAS (Himes et al. 2012).
<i>ANKRD22</i>	Encodes a protein of unknown <i>in vivo</i> function but differential expression has been observed in a study of inner and outer hair cells of the cochlear (Liu et al. 2014) and genome-wide RNAi screens suggests a role as a barrier to induced pluripotent stem cell reprogramming (Yang et al. 2014).
<i>SOCS1</i>	Encodes a member of the suppressor of cytokine signalling (SOCS) family. Expression can be induced by a number of cytokines for example IFN γ . It functions downstream of cytokine receptors and forms part of a negative feedback loop to attenuate cytokine signalling (Pruitt et al. 2014).
<i>MT1G</i>	Encodes a member of a family of proteins that bind both essential and toxic trace metals with high affinity. Zinc exerts great influence on certain immune functions through production release and action of cytokines (Pauwels et al. 1994).
<i>TXNDC12</i>	Encodes a member of the thioredoxin superfamily. It catalyzes disulphide bond formation and isomerization. Thought to play a role in defence against endoplasmic reticulum stress (Pruitt et al. 2014).

Functional genomic response to influenza challenge



(a) Four moderate/severe LCI individuals



(b) Remaining 18 individuals

Figure 5.7: Heatmaps for differentially expressed probes. The top 100 most significant differentially expressed probes between the four moderate/severe LCI and the remaining samples at 48 hours post-challenge were used to generate heatmaps of gene expression. a) Most of the probes are found to be upregulated at 48 hours post-challenge in the four moderate/severe LCI individuals. b) The same probes in the remaining individuals do not show any change in expression at 48 hours post-challenge.

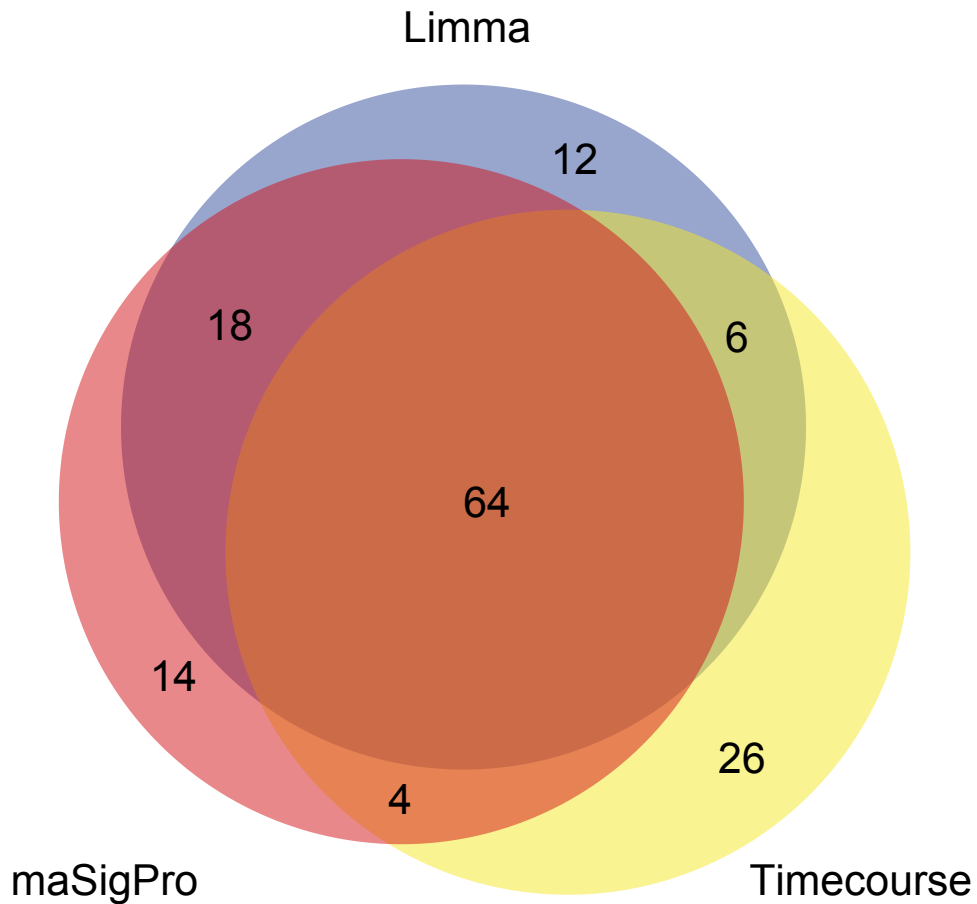


Figure 5.8: Comparison of methods for determining differential expression of probes. Three R packages (limma, maSigPro and timecourse) were used to determine the differentially expressed probes between the four moderate/severe LCI samples and the remaining samples 48 hours post-challenge. The top 100 most significant probes from each analysis were used to compare the overlap between methods.

5.2.5 Pathway analysis

To assess the functional significance of the differentially expressed probes, IPA was utilised (Figure 5.9, Table C.3). The analysis showed significant enrichment for genes involved in a number of pathways including interferon signalling, role of PRR in recognition of bacteria and viruses, and eIF2 signalling. Upregulated genes in these enriched pathways include those encoding the OAS proteins, which form part of the interferon signalling and PRR pathways.

The OAS proteins activate the latent form of RNaseL leading to RNA decay. Another example of an upregulated gene is *EIF2AK2*, part of the eIF2 signalling pathway. This gene produces the enzyme PKR, a serine-threonine kinase that phosphorylates the initiation factor eIF2. Following phosphorylation of the alpha-subunit, eIF2 binds to and sequesters EIF2 β (a guanine nucleotide exchange factor), which prevents the recycling of GDP and stops translation. The eIF2 signalling pathway therefore plays a role in host defence by reducing translation in the host cell, thereby reducing the rate at which new viral particles are generated (Sadler and Williams 2008).

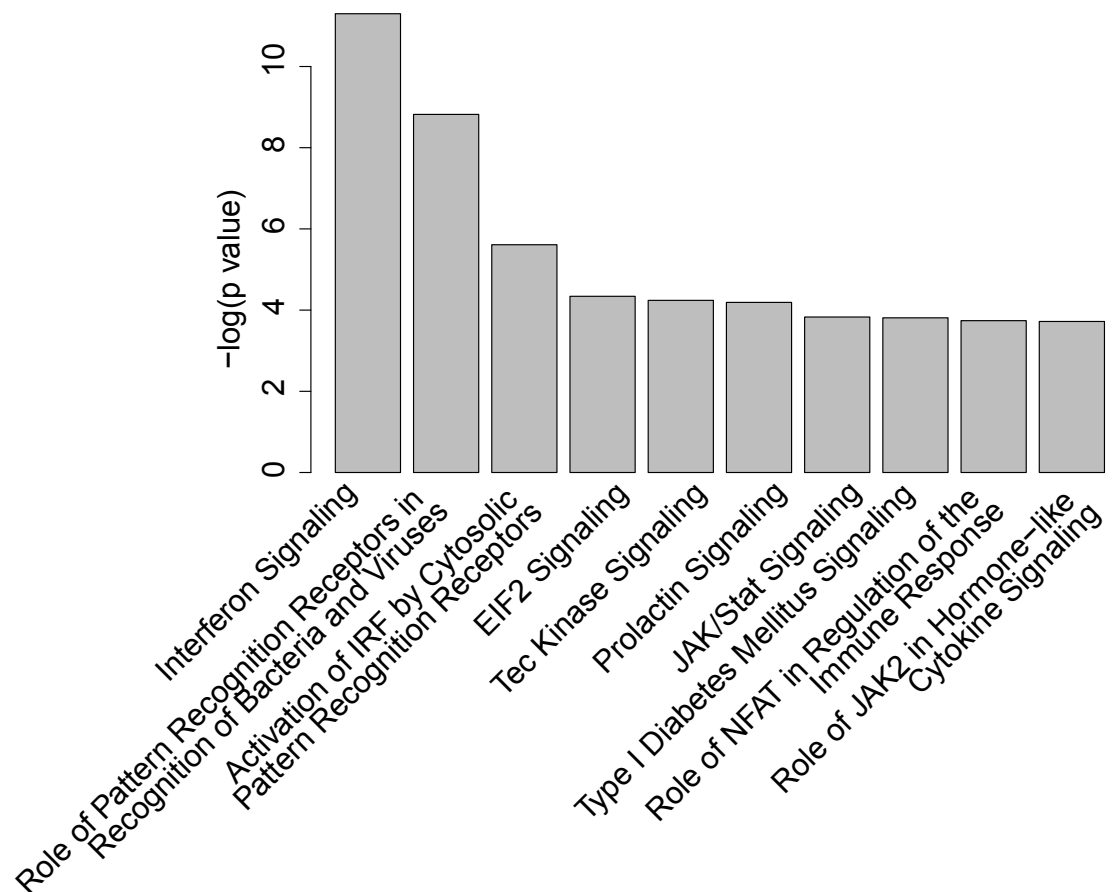


Figure 5.9: Pathway enrichment for moderate/severe LCI. The differentially expressed probes between the four moderate/severe LCI samples 48 hours post-challenge and the remaining subjects were used to determine pathway enrichment with IPA. The bars show the $-\log(p \text{ value})$ for the enrichment of each pathway. The ten most enriched pathways are shown.

5.2.6 Gene expression signature

Although the analysis of over 1,000 differentially expressed probes can provide insight into host-pathogen interactions, this magnitude of variables does not lend itself well to use in diagnostics. The tool ‘prediction analysis for microarrays’ (pamr) (Hastie et al. 2013) was therefore used to find the minimum number of genes that could be used to successfully classify the data. This tool identified six genes (*CCL2*, *SEPT4*, *LAMP3*, *RTP4*, *MT1G* and *OAS3*) which alone classified the samples into two groups with 100% accuracy. These two groups comprised those from subjects with moderate/severe LCI taken at 48 hours post-challenge and the remaining samples. The time course of gene expression values for the six genes is shown in Figure 5.10.

In order to assess the validity of this six gene signal, a data set from an independent but similarly conducted influenza challenge study, published by Huang et al. (2011) was used. This study used similar methods for RNA collection and extraction and the challenge was performed with the same strain of influenza (Wisconsin). Of the 17 subjects in this study, eight were described as having asymptomatic infection and nine as having symptomatic. Those with symptomatic infection had a median summed Jackson score of 39.

For each of the subjects in the Huang study, five samples, taken at different time points, were included: one sample prior to challenge and the remaining from consecutive post-challenge time points taken at 45, 53, 60 and 69 hours. The six gene set, identified from the FLU002 study, successfully classified samples from the eight asymptomatic volunteers with 100% accuracy. For 89% (8/9) of the symptomatic volunteers the gene set correctly classified at least one of the post-challenge samples (Table 5.4).

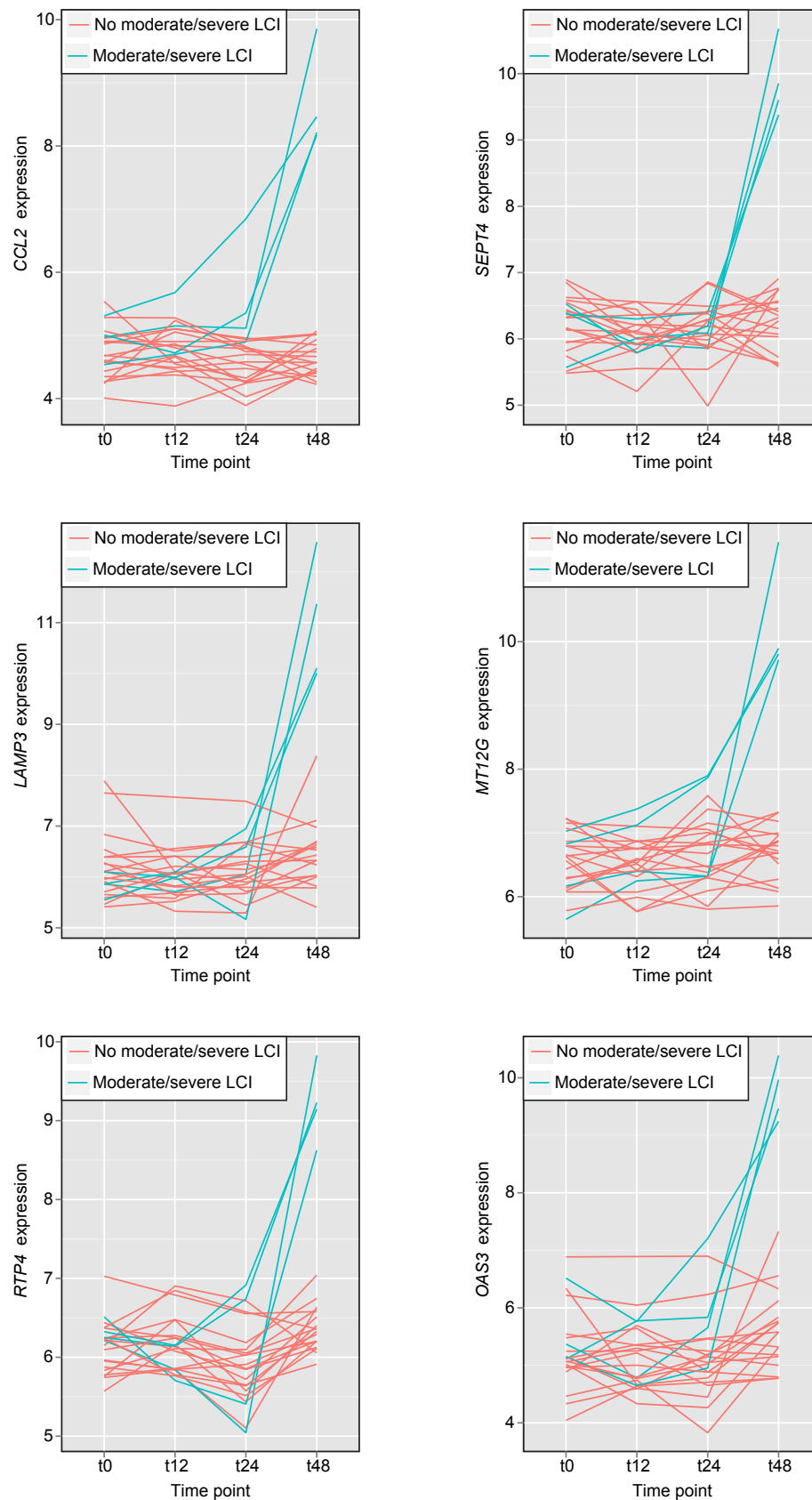


Figure 5.10: Time course for moderate/severe LCI expression signature. Six genes were selected to classify the four moderate/severe LCI subjects 48 hours post-challenge. The expression values for each gene are plotted over time with subjects distinguished by development of moderate/severe LCI.

Table 5.4: Moderate/severe prediction of Huang study subjects. Data from the 17 subjects described by Huang et al. (2011) were used to validate the predictive power of the six gene expression signature. One pre-challenge time point and four consecutive post-challenge time points (which cover the range over which volunteers became symptomatic) were selected. Sx = symptomatic subjects; Asx = asymptomatic subjects; S = sample would be classified as moderate/severe LCI by pamr; n = sample would be classified as pre-challenge or none/mild LCI by pamr.

Subject ID	Phenotype	Sample prediction by pamr					Group predicted
		0h	45h	53h	60h	69h	
1	Sx	n	S	S	n	n	Yes
2	Asx	n	n	n	n	n	Yes
3	Asx	n	n	n	n	n	Yes
4	Asx	n	n	n	n	n	Yes
5	Sx	n	S	S	S	S	Yes
6	Sx	n	S	S	S	S	Yes
7	Sx	n	S	S	S	n	Yes
8	Sx	n	n	n	S	S	Yes
9	Asx	n	n	n	n	n	Yes
10	Sx	n	n	n	n	S	Yes
11	Asx	n	n	n	n	n	Yes
12	Sx	n	S	S	S	S	Yes
13	Sx	n	n	n	S	S	Yes
14	Asx	n	n	n	n	n	Yes
15	Sx	n	n	n	n	n	No
16	Asx	n	n	n	n	n	Yes
17	Asx	n	n	n	n	n	Yes

5.2.7 Temporal response to influenza vaccination

Analysis was conducted to determine whether vaccination modulated the host responses to influenza challenge. Limma analysis revealed 23 probes (23 genes) are differentially expressed between vaccinees and controls pre-challenge (FDR <0.05 and fold change >1.2) (Figure 5.11 and Table C.4). Box plots for the most differentially expressed probes show high variation within the vaccination status groups suggesting vaccination status may not be the primary determinant of expression. In contrast, the temporal response to influenza challenge (with pre-challenge expression values for each subject included in the limma model) showed no difference between vaccinees and controls.

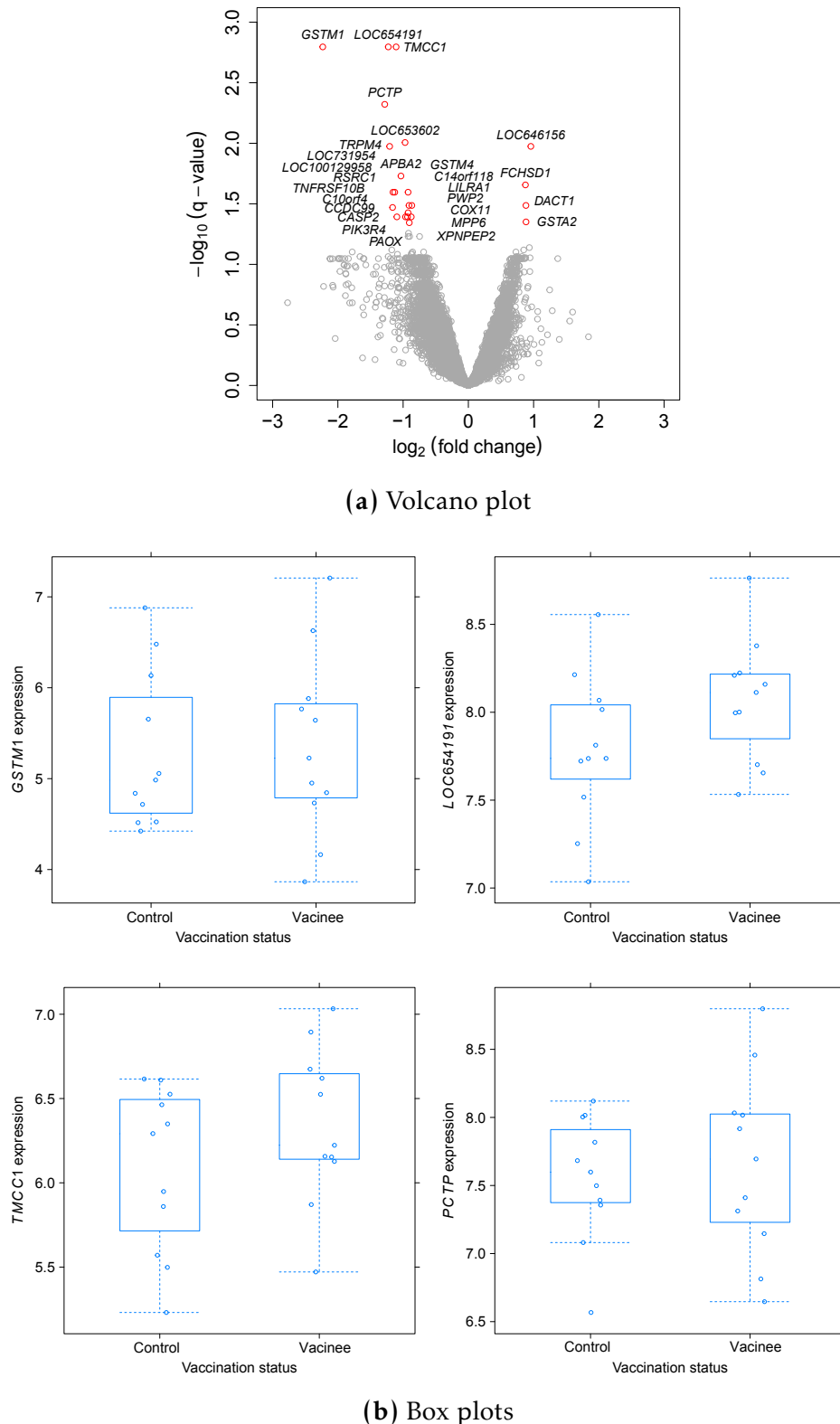


Figure 5.11: Differential expression for vaccination status pre-challenge. a) The significance of differential expression is plotted against fold change. Each point represents a single probe. Red points highlight probes with FDR < 0.05 and fold change > 1.2. A positive fold change indicates upregulation in vaccinated subjects pre-challenge. The most differentially expressed probes are labelled with their corresponding gene names. b) Box plots for the most differentially expressed probes highlight the high variation within the vaccination status groups.

5.2.8 Comparison to gene expression signatures in severe sepsis

Of the 265 severe sepsis patients described in Chapter 3, 25 had a laboratory confirmed viral infection and 16 of these had confirmed H1N1 influenza. Comparing the 25 severe sepsis individuals with viral infection to those without in the cohort revealed 66 differentially expressed probes. These probes are significantly enriched (chi squared p value <0.0001) for the FLU002 study probes differentially expressed between the four moderate/severe LCI and the remaining samples 48 hours post-challenge (Figure 5.12). Of the 41 differentially expressed probes (35 genes, Table C.5) shared between these two studies, one probe showed an opposite fold change but the remaining 40 are differentially expressed in the same direction with upregulation in those identified as having a viral infection.

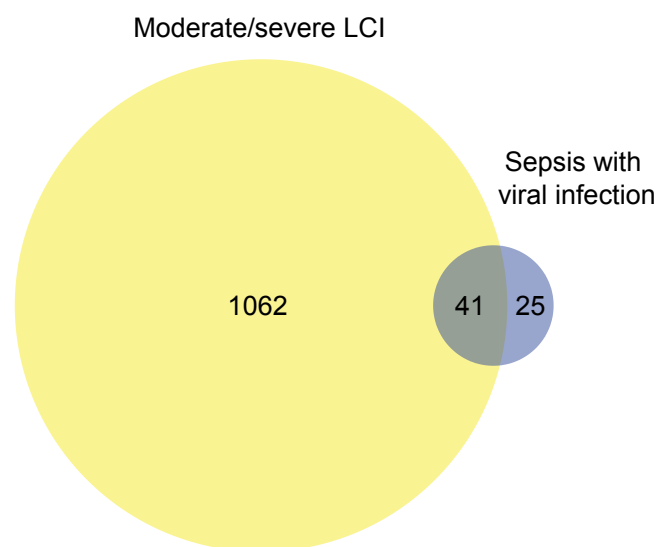


Figure 5.12: Comparison of gene expression signatures in influenza and sepsis. 66 probes were differentially expressed between severe sepsis patients with viral infection compared to those without. 1,103 probes were differentially expressed between the four moderate/severe LCI samples compared to the remaining FLU002 study samples 48 hours post-challenge. The differentially expressed genes for the sepsis cohort showed significant enrichment for probes also differentially expressed for the FLU002 study (chi squared p <0.0001).

Functional genomic response to influenza challenge

Based on the Ingenuity Knowledge Base, seven upstream regulators explain the upregulation in expression of the 35 genes observed in both studies. These include cytokines (IFNA2, IFNL1, IFNG and TNF), kinases (EIF2AK2), enzymes (TGM2) and a protein which interacts with RNA polymerase II. Figure 5.13 illustrates the upstream regulation by *IFNG*.

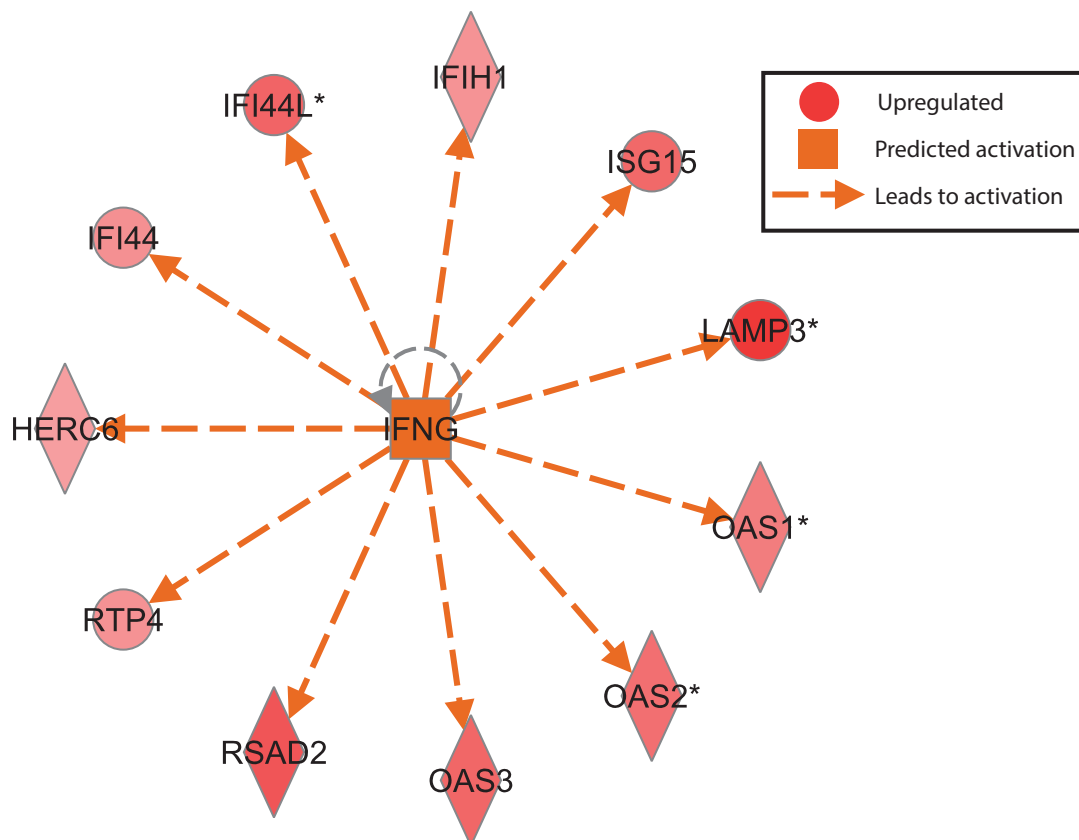


Figure 5.13: Upstream regulation by IFNG.

For both the severe sepsis cohort and influenza challenge, pathway enrichment was determined for the differentially expressed genes. Eight pathways were enriched for the genes differentially expressed between severe sepsis patients with and without a viral infection (Figure 3.11). Five of these pathways were also enriched for differentially expressed genes from the four moderate/severe LCI samples 48 hours post-challenge (role of PRRs, retinoic acid

mediated apoptosis signalling, activations of IRF, UVA-induced MAPK signalling and protein ubiquitination).

Of the six genes used for the influenza challenge gene signature, *LAMP3*, *SEPT4* and *OAS3* (alternative probe) were also differentially expressed in the sepsis patients with viral infections compared to those without. As described in Section 3.2.11 a prediction model was generated to distinguish the viral infection status of the severe sepsis patients. Unlike moderate/severe LCI, the R package *pamr* (Hastie et al. 2013) was unsuccessful in determining a small gene expression signature for viral infection status in the sepsis patients. Instead the model was generated using GeneRave (CSIRO Bioinformatics R package version 3.0.8). GeneRave employs a method focused on sparsity and therefore allows the elimination of redundant explanatory variables. This resulted in a prediction model of six genes (*IFI27*, *TGIF2*, *LY6E*, *CCNY*, *DYNLL2* and *LAMP3*). *LAMP3* is the only gene used in both the influenza challenge and sepsis viral infection status gene signatures.

The influenza challenge six gene signature was used to predict the viral infection status of the severe sepsis patients using *pamr*. The severe sepsis cohort contained 25 patients with a viral infection detected. The influenza challenge gene signature predicted nine patients had a viral infection but only two of these were true positives.

5.3 Discussion

The results of this chapter highlight the value of combining genome-wide gene expression profiling with clinical observations in order to improve the biological classification of influenza challenge studies and assist with clinical diagnostic testing. In addition, combining gene signatures from this study with a cohort of severe sepsis patients with confirmed viral infection provides evidence of a similar signature present in a more heterogeneous clinical phenotype.

5.3.1 Variation in gene expression

Using the PCs generated from the genome-wide gene expression profiles, four LCI samples taken 48 hours post-challenge were identified as clustering separately from the other samples. Upon further investigations, a pattern of predominately upregulated gene expression can be seen which is not observed in the remaining samples, including those previously defined as LCI according to the clinical trial protocol. The signature correlates with the clinical observations and therefore setting an arbitrary symptom score threshold to classify LCI may not be the most informative end result of an influenza challenge study. Instead, this signature may assist with redefining the symptom score threshold or could be used in combination with the symptom score to better classify moderate/severe LCI.

5.3.2 Pathway analysis

From the list of probes which are differentially expressed in the four moderate/severe LCI cases 48 hours post-challenge, IPA identified a number of pathways which are enriched within the dataset. Many of these pathways are involved in the antiviral response illustrating that these four individuals show a different immune response to those with mild LCI. Similar pathways, in particular those regulated by interferon, have previously been identified in a number of other studies however few have focused on the human *in vivo* response (Kawada et al. 2006; Wang et al. 2012).

5.3.3 Gene expression signature

The 48 hour time point gene expression signature was reduced to just six genes (*CCL2*, *SEPT4*, *LAMP3*, *RTP4*, *MT1G* and *OAS3*) using the R package *pamr*. The expression patterns of these six genes classified the samples into

two distinct groups (moderate/severe LCI or not) with 100% accuracy potentially providing a convenient, cost effective and quantitative method for determining post-challenge outcome during vaccine and challenge trials. The reproducibility of this gene set has been demonstrated using samples from an independent vaccine study with a high level of accuracy (100% for asymptomatic and 89% for symptomatic volunteers based on correct classification of at least one of the post-challenge samples).

The Huang et al. (2011) dataset illustrates the variability that can be seen between individuals in response to influenza challenge. For the eight samples that were successfully predicted as being symptomatic using the panel of six genes, there is no one single time point at which all eight samples are predicted as symptomatic. This suggests that the window for symptomatic classification can vary from individual to individual over a number of hours. In the FLU002 dataset, pamr is used to define gene expression profiles that predict moderate/severe LCI samples at 48 hours post-challenge. Therefore, in order to improve the accuracy of this prediction over a smaller time course, further work may be required to adjust the panel of genes used and to define the appropriate window of time to measure their expression values.

5.3.4 Temporal response to influenza vaccination

Only a small number of probes were differentially expressed between vaccinated and unvaccinated individuals pre-challenge and a difference in the temporal response to influenza was not detected. This is perhaps not surprising as initial samples were taken 30 days after vaccination and previous studies (Zak et al. 2012) with viral vectored vaccines have shown differential gene expression to be undetectable at 7 days post-vaccination. Further studies will be needed to confirm this preliminary evidence of vaccine efficacy, and these will be facilitated by the use of gene expression data to allow objective classification of outcomes.

5.3.5 Overlap of gene signatures between influenza challenge and severe sepsis

Of the severe sepsis cohort, described in Chapter 3, 9% have a laboratory confirmed viral infection. Overlap was observed at the differential gene expression and pathway level between these individuals and the moderate/severe LCI samples 48 hours post-challenge from the influenza challenge study. The 35 genes which were differentially expressed in both cohorts may provide a more robust signature for viral infection expressed in multiple clinical phenotypes.

Fewer genes were differentially expressed in the context of viral infection in the severe sepsis cohort than in the influenza challenge study. The severe sepsis phenotype is a much more heterogeneous phenotype and the time point of infection is unknown, in contrast to the clearly defined and carefully controlled influenza challenge study. In addition while a positive viral infection test was recorded for example the presence of H1N1 influenza, a negative result was not. Therefore non-viral infection patients may not be true negatives which could further increase the heterogeneity of the comparison.

The influenza challenge six gene signature was not successful in predicting viral infection in the severe sepsis cohort. As demonstrated by the prediction of the subjects from the Huang et al. (2011) study, time point may be important for the successful identification of infection and as this is an unknown variable in the severe sepsis cohort this could explain some of the failure. Additionally the flu challenge signature was developed in a much more homogeneous setting with healthy individuals subjected to a single challenge. The signal may therefore be masked in the sepsis cohort when additional infections are present and the general health of the individuals is poorer.

5.3.6 Conclusion

This analysis gives further insight into host-pathogen interactions that could be beneficial for early detection of influenza particularly during pandemics. However, the major value is in the clinical trial setting. Current biological phenotyping methods often show inconsistency but this work illustrates the use of a more quantitative method. The ability to better classify symptomatic individuals post influenza challenge will help accelerate future influenza vaccine efficacy studies. In addition the overlap in gene expression signatures and enriched pathways between the influenza challenge and the severe sepsis patients with a viral infection will help identify components of the immune response to viral infections which are important across multiple clinical settings and may aid in the identification of viral infections in the context of sepsis.

Chapter 6

Functional genomic response in CVID

6.1 Introduction

CVID are a group of diseases in which insufficient quantity and quality of immunoglobulin usually leads to susceptibility to recurrent bacterial infections (Chapel and Cunningham-Rundles 2009).

Heterogeneity in the disease phenotype in terms of clinical features and complications has made defining the genetic aetiology challenging, with candidate gene studies and one reported genome-wide association study having limited success (Salzer et al. 2008; Orange et al. 2011).

6.1.1 Patient cohort

Over the last 20 years, Professor Helen Chapel and colleagues at the Clinical Immunology Department in Oxford have created a unique research database containing detailed phenotyping based on clinical and laboratory data (Chapel et al. 2008). This, along with a repository of patient genomic DNA and peripheral blood cells, has allowed the selection of 35 CVID patients with well defined clinical phenotypes to be used in WGS as part of the Oxford Illumina WGS500 project. In a small subset, this has been complemented by functional genomic profiling to analyse gene expression in B cells. By grouping similar clinical phenotypes, it is hoped that mutations in the same gene or pathway will be identified in multiple CVID patients.

6.1.2 Previously described CVID genes

A number of genes have previously been described in CVID and these are summarised in Table 6.1. Many genes have also been previously described in PIDs (Table D.1).

Table 6.1: Previously described genes in CVID. Level 1 = Variant only observed in immune deficiency patients and supported by functional data. Level 2 = Variant also observed in healthy cohorts (at a lower frequency).

Gene	Evidence	Reference
<i>CD19</i>	Level 1	Kanegane et al. 2007, van Zelm et al. 2006
<i>CD20</i>	Level 1	Kuijpers et al. 2010
<i>CD21</i>	Level 1	Thiel et al. 2012
<i>CD81</i>	Level 1	van Zelm et al. 2006
<i>ICOS</i>	Level 1	Grimbacher et al. 2003, Salzer et al. 2004
<i>TNFRSF13C</i> (BAFF-R)	Level 1 and 2	Warnatz et al. 2009, Losi et al. 2005
<i>TNFRSF13B</i> (TACI)	Level 2	Salzer et al. 2005, Castigli et al. 2005, Pan-Hammarstrom et al. 2007
<i>HLA</i>	Level 2	Volanakis et al. 1992, Olerup et al. 1992
<i>CD27</i>	CVID like symptoms (more extensive phenotype)	Montfrans et al. 2012
<i>PLCG2</i>	CVID like symptoms (more extensive phenotype)	Ombrello et al. 2012
<i>LRBA</i>	CVID like symptoms (more extensive phenotype)	Lopez-Herrera et al. 2012
<i>PIK3CD</i>	CVID like symptoms (more extensive phenotype)	Lucas et al. 2014, Angulo et al. 2013

6.1.3 GWAS genes

To address the underlying immunopathogenesis of CVID, Orange et al. (2011) conducted a GWAS for patients with CVID. A number of the Oxford cohort were included in this study. In total, 363 patients were included along with 3,031 healthy controls. The samples were split into a discovery and validation cohort.

The only association of genome-wide significance was with the MHC (rs3117426 with $p = 8.62 \times 10^{-10}$). The second most significant association, in chromosome band 8p21.2 which harbours *ADAM28*, *ADAM7*, *ADAMDEC1* and *STC1*, did not reach genome-wide significance ($p = 6.24 \times 10^{-6}$) but was replicated in the validation cohort (p value range = 2.25×10^{-8} to 0.0314). In addition, 84 CNV and 98 duplications were identified in at least one CVID patient but not observed in any of the controls.

The identification of multiple novel common and rare susceptibility loci for CVID supports the hypothesis of diverse mechanisms leading to complex phenotypes but despite these insights, diagnosis of CVID has not improved. The most significant gene associations from the GWAS are summarised in Table D.2.

6.1.4 Filtering variants

Following the evolution of NGS, WGS is becoming more available in both the research and clinical setting. WGS enables sequencing of the entire genome allowing the screening for variants in known candidate genes as well as the identification of novel variants in both coding and non-coding regions. WGS results in the identification of a large number of variants in each individual. When working with sporadic cases in a cohort of individuals with a complex, heterogeneous disease, filters can be used to reduce this list to a more meaningful length. Filters can include a threshold for the confidence in variant call, the frequency in the 1000 Genomes Project and whether the variant is predicted to be deleterious to the protein function (Chou et al. 2012; Reumers et al. 2012). This can still leave a large number of variants and one major challenge is the biological interpretation of these variants. QIAGEN's IVA software (www.qiagen.com/ingenuity, QIAGEN Redwood City) allows the use of a filter that gives biological context for each variant. This is based on the Ingenuity Knowledge Base which is a repository of biological interactions and functional

annotations curated by scientists. The generated VCF files summarising the variants detected in each individual can be analysed directly through IVA and all the appropriate filters, from the call confidence through to biological context, can be applied. This can rapidly increase the speed in which large datasets, such as WGS for sporadic multifactorial diseases, can be analysed. However, by using the Knowledge Base, an element of bias will be introduced due to certain genes and pathways being better annotated than others. A major challenge with this type of data is determining how likely it would be to observe similar results in another cohort of individuals by chance and this is not easy to determine using IVA.

6.1.5 WGS500

WGS500 is a collaborative project between the University of Oxford and Illumina which aims to sequence the genomes of 500 individuals with a range of diseases including rare inherited diseases, immunological disorders and cancer. Analysis of the sequencing data is a collaborative effort between a designated team within core bioinformatics at the Wellcome Trust Centre for Human Genetics (WTCHG) and individuals from research groups submitting the clinical samples. At the time of analysis, a union file of WGS data for 274 individuals (239 non-CVID individuals) had been generated.

6.1.6 Integrating WGS and RNA-seq

Despite sequencing the entire genome, many WGS studies to date have typically focused on variants within the coding exome (Cooper and Shendure 2011). This is generally due to the challenges of functional prediction for non-coding variants. One method for interrogating non-coding variants is by combining WGS with global transcriptomic profiling from RNA-seq. The RNA-seq data can provide evidence of altered gene expression, for instance compared to healthy

controls, and therefore assist with prioritisation of potential candidate non-coding variants.

6.1.7 RNA-seq QC and normalisation

For RNA-seq experiments, QC is carried out to determine the proportion of reads mapping to genes and identify variation within the dataset. The technical biases inherent to microarray studies such as batch effect are not present in RNA-seq experiments but other sources of variation can occur such as differing library sizes between samples. Following data generation, the reads are mapped to a reference sequence and summarised at the gene level. A threshold is then set to remove the genes which are not classified as expressed (Oshlack et al. 2010).

A number of different methods can be used for RNA-seq data normalisation. These are generally based on either the library size concept or distribution adjustment of read counts. The TMM method which can be implemented through the edgeR R package (Robinson et al. 2010) has been shown to perform well for downstream differential gene expression analysis (Dillies et al. 2013). As well as accounting for library size, this method also allows for different compositions of RNA between samples in the experiment (Robinson and Oshlack 2010).

6.1.8 Aims

The aim of this chapter is to investigate the functional genomic response in CVID and the relationship to the underlying genetic aetiology. PIDs provide the opportunity to resolve rare variants modulating the immune response, informing understanding of immunity. CVID encompasses the most common PID occurring in adults. I aim to

1. investigate genetic variants associated with CVID using a WGS approach

2. define the transcriptomic response in CVID compared to controls
3. use knowledge of differentially expressed genes in CVID to inform identification of functional genetic variants
4. relate disease associated genes/variants in CVID to sepsis (cohort described in Chapters 3 and 4)

6.2 Results

6.2.1 Patient cohort clinical data

In order to investigate genetic variants associated with CVID, 33 individuals with sporadic CVID and one grandmother grandson pair were selected from the Oxford cohort for WGS. Extensive clinical information was recorded for each patient but for the purpose of this study, the cohort was summarised according to the clinical phenotype of complications and IgM status. The majority of the patients have an Infections Only subphenotype (26). The remaining nine have Polyclonal Lymphoproliferation which represents a more complicated case of CVID and includes patients with LIP, granuloma and splenomegaly. Twenty six of the cohort are IgM negative with the remaining 9 IgM positive (IgM status does not segregate with complications status). These classifications are summarised in Table 6.2.

Table 6.2: CVID patient cohort clinical information. Patients were classified as IgM- if the total serum IgM level at diagnosis was less than the quantity observed in healthy adult populations (<0.4 g/l).

Patient	Sex	Age at diagnosis	Clinical subphenotype	IgM status
D209	Male	43	Infections Only	Absent
D276	Male	2	Infections Only	Absent
D667	Male	27	Infections Only	Absent
D345	Female	49	Infections Only	Absent
D334	Male	15	Infections Only	Absent
C018	Male	10	Infections Only	Absent
C024	Female	48	Infections Only	Absent
C089	Female	18	Infections Only	Absent
D374	Male	35	Infections Only	Absent
C038	Male	17	Infections Only	Absent
D575	Male	9	Infections Only	Absent
C036	Female	20	Infections Only	Absent
C065	Male	24	Infections Only	Absent
D38	Male	52	Infections Only	Absent
D705	Female	50	Infections Only	Absent
C072	Female	35	Infections Only	Absent
D641	Female	47	Infections Only	Absent
C019	Male	37	Infections Only	Absent
C085	Female	38	Infections Only	Absent
C023	Male	33	Infections Only	Absent
D745	Male	5	Infections Only	Present
D218	Female	16	Infections Only	Present
C063	Female	49	Infections Only	Present
D765	Female	32	Infections Only	Present
C078	Male	29	Infections Only	Present
D232	Female	44	Infections Only	Present
C082	Female	38	Polyclonal Lymphoproliferation	Absent
D325	Female	39	Polyclonal Lymphoproliferation	Absent
C028	Female	49	Polyclonal Lymphoproliferation	Absent
C006	Male	37	Polyclonal Lymphoproliferation	Absent
D839	Male	40	Polyclonal Lymphoproliferation	Absent
C044	Female	34	Polyclonal Lymphoproliferation	Absent
C032	Male	35	Polyclonal Lymphoproliferation	Present
D269	Female	56	Polyclonal Lymphoproliferation	Present
C003	Female	54	Polyclonal Lymphoproliferation	Present

6.2.2 Whole genome sequencing

All 35 samples were sequenced, as part of the WGS500 project, to an average coverage greater than 25x (Table D.3). The reads were mapped with Stampy (Lunter and Goodson 2011) and the SNPs and InDels called with Platypus (Rimmer et al. 2014) (Core Bioinformatics, WTCHG).

6.2.3 Patient exclusion

During the course of the data generation phase, one patient returned to clinic following an absence of more than a year. Antibody levels were detected as normal and therefore it is suspected the diagnosed CVID was drug induced. This illustrates the dynamic process of phenotyping in the clinic and highlights one of the challenges involved in this type of research. This individual (D218) has been excluded from further analysis.

6.2.4 Filtering criteria

Following generation of WGS data, 14,819,871 variants were identified across the 34 patient samples. A number of filters were employed for all analyses described in this chapter in order to prioritise high quality calls for rare, non-synonymous and likely pathogenic variants (4,768 variants, Table 6.3). In addition, technical artefacts could be excluded using the MAF for the remaining samples in the WGS500 project. For the predicted deleterious variants, the type of variant and location in the gene region are summarised in Figure 6.1. The variants are predominantly SNVs which is not unexpected as the ability to identify structural variants from WGS data remains challenging (Hormozdiari et al. 2009). The predicted deleterious filter uses bioinformatics programs such as SIFT (Kumar et al. 2009) and Polyphen-2 (Adzhubei et al. 2010) to prioritise variants which

Functional genomic response in CVID

have a predicted impact on the protein, hence the enrichment for variants in the coding regions of genes.

The number of times a variant is observed in a CVID patient is summarised in Table 6.4. The majority of variants are observed in just one CVID case which highlights the complex nature of CVID. The variant observed in seven patients (20.6%) is a heterozygous variant in the gene *GRM2* encoding a GTPase which plays a role in the termination of mitochondrial protein translation (Pruitt et al. 2014).

Table 6.3: Variants passing WGS filtering criteria. Filters were applied to prioritise rare, predicted deleterious variants.

Filter	No. variants	No. genes
Confidence	11,304,176	20,594
No. reverse reads ≥ 1	11,118,902	20,583
No. forward reads ≥ 1	10,940,142	20,571
Common variants	4,808,176	20,132
WGS500 MAF	2,721,014	19,678
Predicted deleterious	4,768	3,737

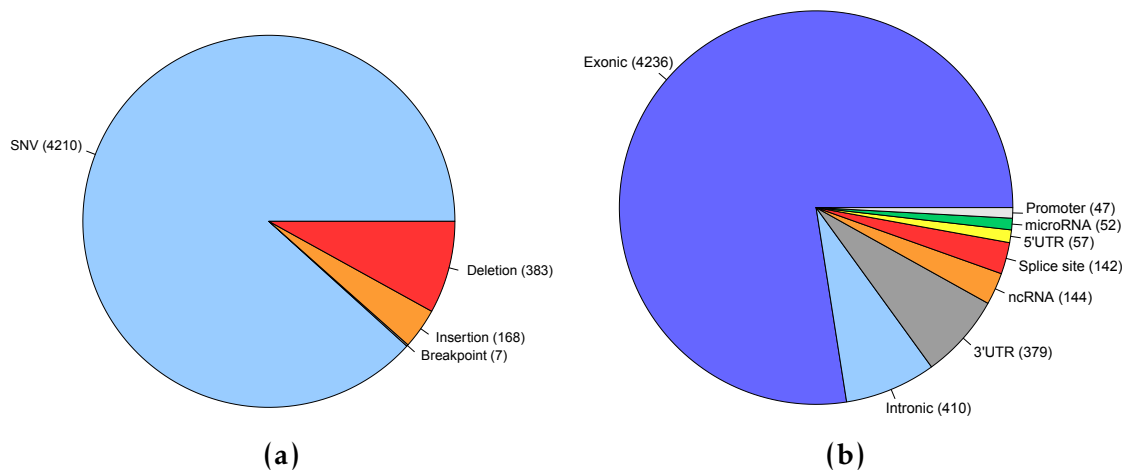


Figure 6.1: Variant type and gene region. Following filtering criteria, 4,768 predicted deleterious variants were identified. a) The proportion of each type of variant. b) The proportion of variants located in different gene regions.

Table 6.4: Number of patients each variant is observed in. Following filtering criteria, 4,768 predicted deleterious variants were identified. The number of CVID patients that each variant is observed in is listed.

No. patients variant is observed in	Frequency
1	3958
2	553
3	174
4	59
5	19
6	4
7	1

6.2.5 Previously described CVID and PID genes

The 4,768 predicted deleterious variants were first screened for genes that had previously been described in CVID or PID (Tables 6.1 and D.1). Two heterozygous deleterious variants were identified in genes from the CVID candidate list (Table 6.5). *TNFRSF13B* encodes TACI, a member of the TNF receptor family. The observed variant (C104R) is one of the most common sequence variants observed in patients with CVID. Homozygous or compound heterozygous variants in this gene have only been observed in individuals with antibody deficiencies but the heterozygous state is seen in healthy individuals, usually at a lower frequency. It is therefore hypothesised that heterozygous variants in this gene, as in this case, contribute to disease risk but do not cause the disease (Salzer et al. 2009). Likewise, homozygous variants in the gene *LRBA* are associated with individuals with CVID like symptoms but a more extensive phenotype. Individuals with observed heterozygous variants were asymptomatic and therefore it is less likely that the heterozygous state plays a role in CVID. These results confirm that variants in known candidate genes are unable to explain the underlying genetic aetiology in the CVID cohort and support the rationale for using WGS.

For the PID candidate gene list, 47 deleterious variants were observed (Table 6.6). The greatest number of deleterious variants were observed in the *NOD2* gene (present in six patients). This gene is associated with Blau syndrome in the autoinflammatory disorders category of PIDs (Al-Herz et al. 2011). Variants in *NOD2* have also been associated with Crohn's disease, a granulomatous condition of the digestive tract (Hugot et al. 2001). Granulomatous disease is one of the clinical complications observed in CVID patients which led to the hypothesis that Crohn's associated variants in *NOD2* may also be associated with the CVID subphenotype. In the Oxford CVID cohort, patients with granuloma are categorised as the Polyclonal Lymphoproliferation subphenotype but the six patients with a variant in *NOD2* are all classed as the Infections Only subphenotype. A study by (Packwood et al. 2010) (which included CVID patients from the Oxford cohort) did not find an association between *NOD2* variants and granulomatous disease in CVID.

Two individuals have a variant in *BTK* which is known to cause X-linked agammaglobulinemia. C063 is a female patient with a heterozygous variant. D276 is a male patient with a hemizygous insertion of a T nucleotide at position 100609671 on the X chromosome (pN526fs*11) resulting in a frame shift which introduces a stop codon. This patient had been screened for known *BTK* variants through NHS sequencing but this particular variant was not observed in the remaining WGS500 samples or 1000 Genomes which may explain why it was not identified previously. The presence of this variant was confirmed by Sanger sequencing.

Table 6.5: Variants in previously described CVID genes. The 4,768 predicted deleterious variants were screened for genes that had previously been described in CVID. Two heterozygous variants were observed. The number of times the variant is observed in both the CVID cohort and the remaining WGS500 cohort (239 individuals) is listed. Prop. = proportion.

Gene	Chr	Pos	Ref allele	Sample allele	Protein variant	CVID patient	Prop. CVID	No. other WGS500	Prop. other WGS500
<i>TNFRSF13B</i>	17	16852187	A	G	p.C104R	D38	0.03	5	0.02
<i>LRBA</i>	4	151392815	A	G	p.S2210P, p.S2221P	D374	0.03	0	0.00

Table 6.6: Variants in previously described PID genes. The 4,768 predicted deleterious variants were screened for genes that had previously been described in PIDs. Variants are heterozygous unless otherwise stated. The number of times the variant is observed in both the CVID cohort and the remaining WGS500 cohort (239 individuals) is listed. Prop. = proportion. Hemi = hemizygous. Hom = homozygous.

Gene	Chr	Pos	Ref allele	Sample allele	Protein Variant	No. CVID	Prop. CVID	No. other WGS500	Prop. other WGS500
<i>ATM</i>	11	1.08E+08	C	G	p.P1054R	1	0.029	11	0.046
<i>BTK</i> (hemi)	X	1.01E+08		T	p.N350fs*11;	1	0.029	0	0
					p.N526fs*11;				
<i>BTK</i>	X	1.01E+08	G	A	p.N560fs*11				
					p.R492C;	1	0.029	0	0
<i>C1QC</i>	1	22974051	C	G	p.R526C				
					p.H171Q	1	0.029	0	0
<i>C8B</i>	1	57411587	G	A	p.R286C;	1	0.029	0	0
					p.R338C;				
<i>C9</i>	5	39311336	A	T	p.R276C				
<i>C9</i>	5	39342243	T	A	p.F338L	1	0.029	2	0.008
					p.M45L	1	0.029	2	0.008
<i>CD46</i>	1	2.08E+08	G	A	p.D185N	1	0.029	0	0
<i>CD79A</i>	19	42383644	C	A	p.T102N;	2	0.059	1	0.004
					p.T140N				
<i>CSF2RA</i>	X	1428421	G	A	p.R358Q	1	0.029	4	0.017
<i>CSF2RA</i>	X	1428423	G	A	p.G359R	1	0.029	2	0.008
<i>CTSC</i>	11	88059362	A	G		1	0.029	1	0.004
<i>DCLRE1C</i>	10	14965056	A	T	p.L214M;	1	0.029	1	0.004
					p.L329M;				
					p.L209M				

Continued on next page

Table 6.6 – Continued from previous page

Gene	Chr	Pos	Ref allele	Sample allele	Protein Variant	No. CVID	Prop. CVID	No. other WGS500	Prop. other WGS500
<i>DKC1</i> (hom)	X	1.54E+08	AAG		p.K505del; p.K500del	1	0.029	2	0.008
<i>IFNGR2</i>	21	34787294	C	G	p.T58R	2	0.059	7	0.029
<i>IGLL1/IGLL5</i>	22	23237860	A	G	p.T136A; p.T211A	1	0.029	10	0.042
<i>IL12B</i>	5	1.59E+08	C	A	p.V298F	2	0.059	5	0.021
<i>IL21R</i>	16	27414502	G	A	p.R4H	2	0.059	4	0.017
<i>IL21R-AS1</i> ; <i>IL21R</i>	16	27460561	T	A	p.V525D; p.V547D	1	0.029	0	0
<i>ITGB2</i>	21	46323418	G	A	p.R121W	1	0.029	0	0
<i>KMT2D</i>	12	49426990	C	A	p.R383L	1	0.029	0	0
<i>KRAS</i>	12	25368405	A	T	p.C180*	1	0.029	0	0
<i>LPIN2</i>	18	2929103	G	A	p.L504F	1	0.029	2	0.008
<i>LRBA</i>	4	1.51E+08	A	G	p.S2221P; p.S2210P	1	0.029	0	0
<i>MASP2</i>	1	11105542	C	T	p.C156Y	2	0.059	2	0.008
<i>MCM4</i>	8	48887414	G	C	p.V753L	1	0.029	0	0
<i>MEFV</i>	16	3306917	C	T		2	0.059	8	0.033
<i>MRE11A</i>	11	94178974	A	G		1	0.029	0	0
<i>NCF2</i>	1	1.84E+08	T	A	p.N338L; p.N419I; p.N374I	2	0.059	1	0.004
<i>NCF4</i>	22	37260160	A	C	p.T36P	1	0.029	1	0.004
<i>NHEJ1</i>	2	2.2E+08	C	T	p.R57Q	1	0.029	1	0.004

Continued on next page

Table 6.6 – Continued from previous page

Gene	Chr	Pos	Ref allele	Sample allele	Protein Variant	No. CVID	Prop. CVID	No. other WGS500	Prop. other WGS500
NLRP2	19	55494283	C	G	p.T383R; p.T406R; p.T384R	3	0.088	2	0.008
NOD2	16	50746086	C	T	p.A755V	1	0.029	0	0
NOD2	16	50746238	G	C	p.G806R	1	0.029	0	0
NOD2	16	50756540	G	C	p.G908R	1	0.029	8	0.033
NOD2	16	50763778		C	p.L1007fs*2	3	0.088	8	0.033
PIK3R1	5	67597170		TATA		1	0.029	1	0.004
PMS2	7	6045634	T	C	p.I18V	2	0.059	9	0.038
PRKDC	8	48691175	G	C	p.L3869V; p.L3900V	1	0.029	0	0
PRKDC	8	48775035	A	G	p.Y1940H	1	0.029	0	0
RAG2	11	36614561	G	T	p.F386L	2	0.059	8	0.033
TAP2	6	32803458	A	T	p.L234Q	1	0.029	0	0
TRAF3	14	1.03E+08	C	G	p.Y342*; p.Y425*; p.Y400*	1	0.029	0	0
TYK2	19	10461390	G	A		1	0.029	1	0.004
UNC13D	17	73836610	C	T	p.G244E	1	0.029	0	0
VPS13B	8	1.01E+08	T	C	p.Y2663H; p.Y2638H	1	0.029	0	0
VPS13B	8	1.01E+08	A	G	p.N2993S; p.N2968S	1	0.029	5	0.021

6.2.6 Variants upstream of candidate genes

IVA allows the investigation of variants "1 hop" upstream from a biological context term or gene list of interest. Using the 11 previously described CVID genes (excluding HLA, Table 6.1), deleterious variants were identified in genes linked to CVID candidate genes in 33 of the 34 patients. In total 90 variants in 84 genes were identified.

Table 6.7 summarises the number of genes with a predicted deleterious variant upstream of each candidate gene. *PLCG2* has the greatest number of genes upstream. The gene encodes PLC γ 2, a signalling molecule expressed in B cells, natural killer cells and mast cells. Gain of function deletions have been associated with families with phospholipase C γ 2-associated antibody deficiency and immune dysregulation (PLAID), a disease with substantial phenotypic overlap with CVID. The deletions result in constitutive phospholipase activity but reduced downstream signalling (Ombrello et al. 2012).

Figure 6.2 displays a network that is created by linking each individual through the shared genes which contain predicted deleterious variants upstream of the CVID candidate genes. Most individuals have variants in more than one gene highlighting the likely polygenic underlying genetic aetiology of CVID. Forming this network separates the patients into three groups: two groups are internally connected through variants in shared genes but none of the patients in the third group share any genes with other members of the cohort. The symbol for each patient is coloured based on the subphenotype (Table 6.2) and one group is predominately composed of individuals with an Infections Only subphenotype but variable IgM status. This is a particularly striking result as it suggests similarities in genetic aetiology within the less complicated subphenotype of CVID. More extensive phenotyping information is available for each individual so this could be investigated further for the two other clusters.

At the patient level, three individuals had a high number of variants in genes upstream of the candidate genes. Figure 6.3 summarises the connections between these genes, the candidate genes and the CVID subphenotype. Generating these figures can help identify which genes and pathways may be important for an individual patient rather than the cohort as a whole.

Table 6.7: Number of genes upstream of candidate genes. The IVA biological context filter was implemented to identify genes with predicted deleterious variants upstream of the 11 CVID candidate genes. The number of variants and genes identified upstream of each candidate gene is listed.

Gene	No. variants	No. genes
<i>PLCG2</i>	31	29
<i>CD81</i>	25	27
<i>PIK3CD</i>	13	11
<i>CD19</i>	11	10
<i>CD20</i>	6	4
<i>LRBA</i>	6	6
<i>CD21</i>	5	4
<i>CD27</i>	5	4
<i>TNFRSF13C</i>	4	4
<i>TNFRSF13B</i>	4	4
<i>ICOS</i>	2	2

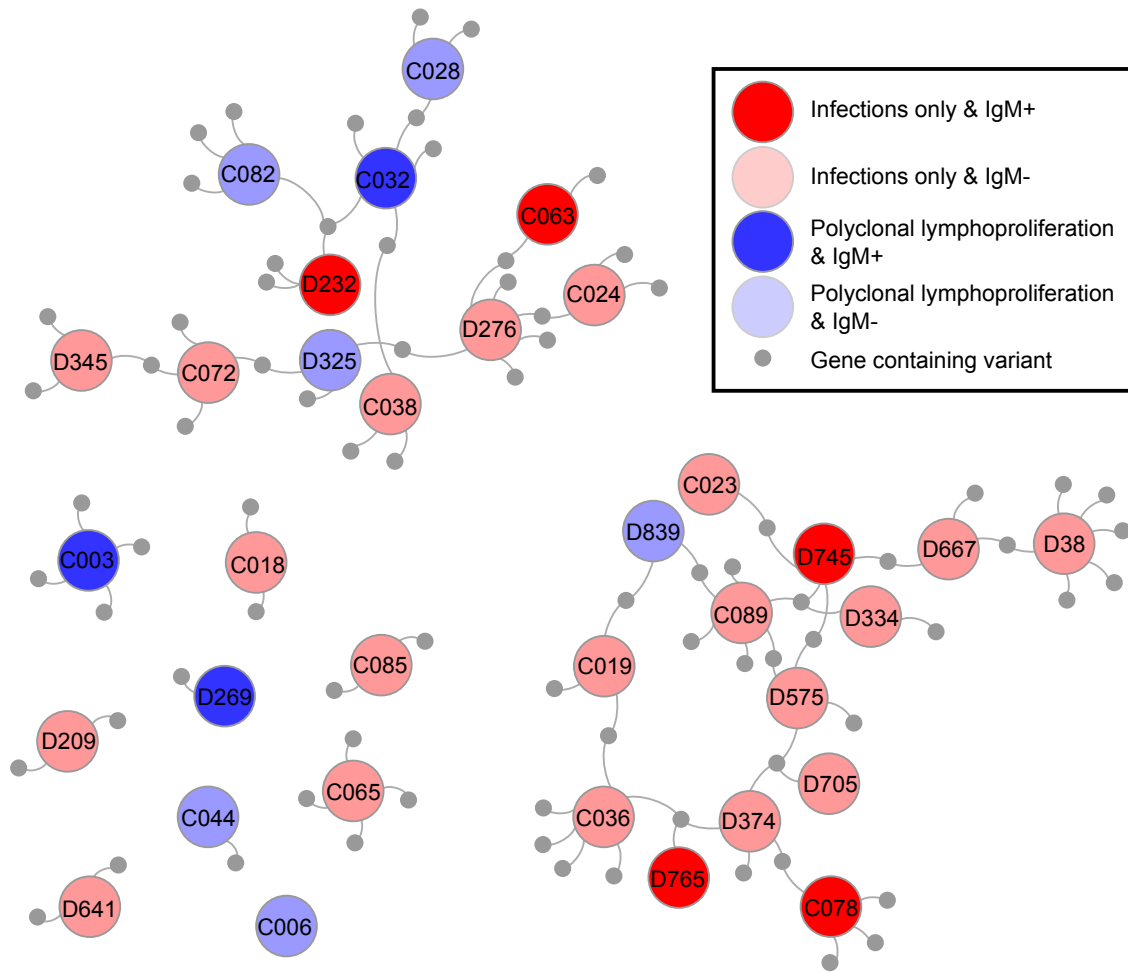
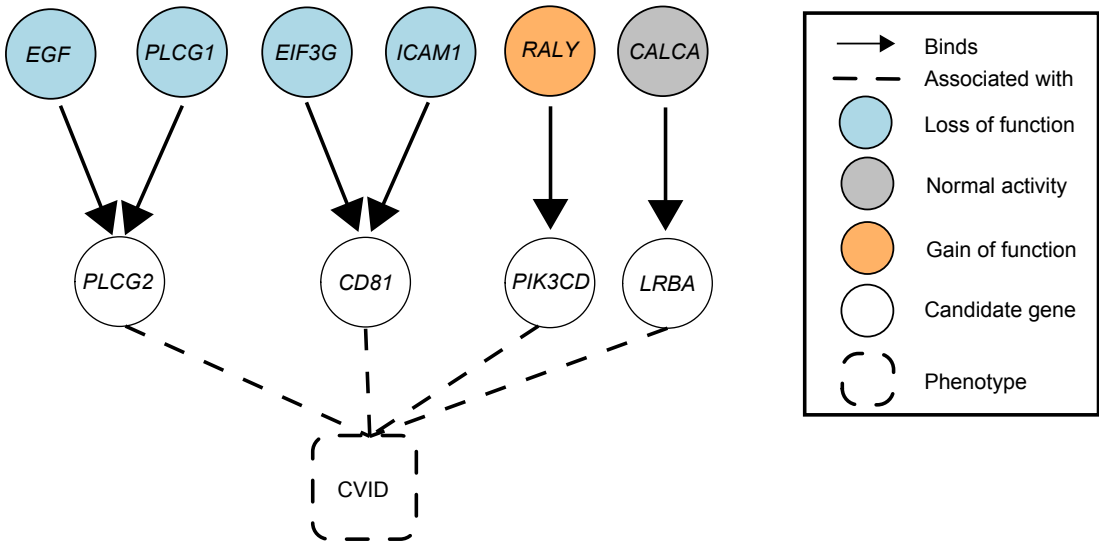
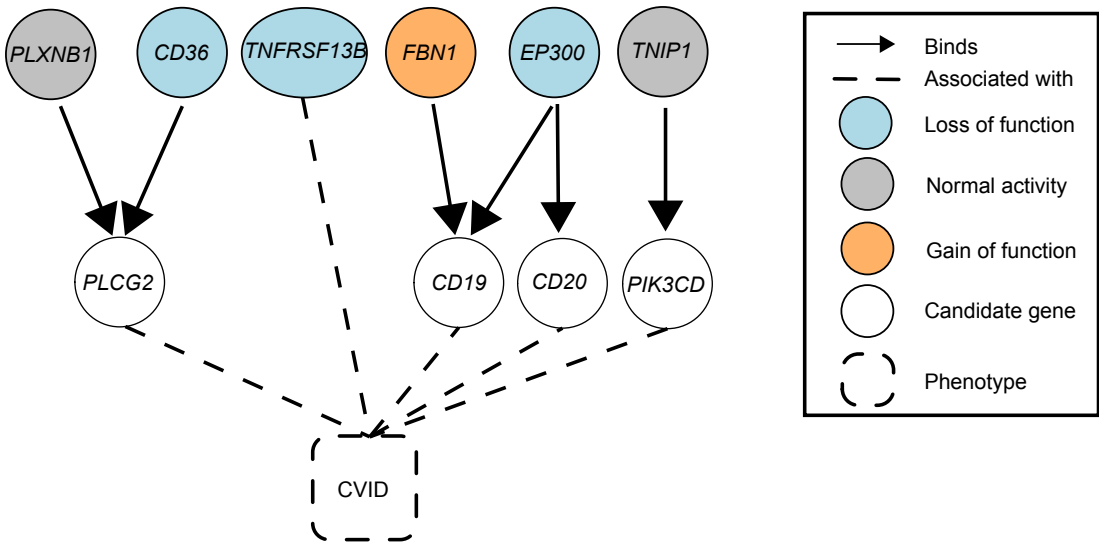


Figure 6.2: Network of variants one hop from CVID candidate genes. The IVA biological context filter was implemented to identify genes with predicted deleterious variants upstream of the 11 CVID candidate genes. A network was generated with Gephi (Bastian et al. 2009) to display the connections between each individual based on variants in shared genes. Each individual patient is represented by a large circle coloured by subphenotype. Genes with predicted deleterious variants are represented by small grey circles.



(a) Patient C036



(b) Patient C038

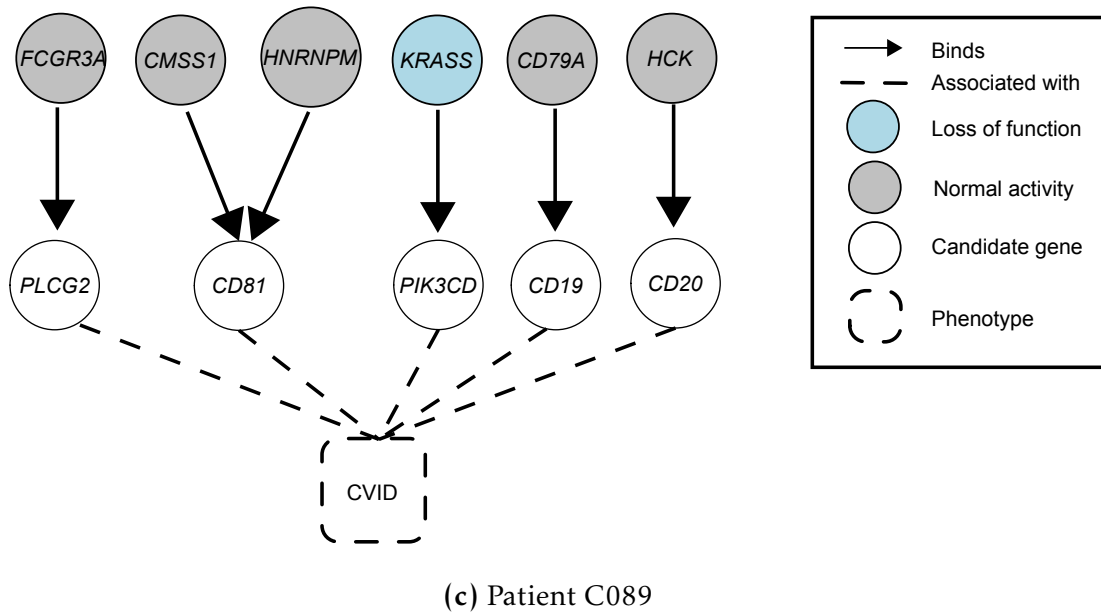


Figure 6.3: Network of variants for individual patients. The IVA biological filter was implemented to identify genes with predicted deleterious variants upstream of the 11 CVID candidate genes. The network displays the connections between these genes, the candidate genes and the CVID phenotype for three CVID patients. Genes containing the predicted deleterious variants are coloured by the predicted effect of the variant on the function of the gene. a) Patient C036. b) Patient C038. c) Patient C089.

6.2.7 GWAS genes

To address the underlying immunopathogenesis of CVID, Orange et al. (2011) conducted a GWAS for patients with CVID including 31 patients from the cohort presented in this chapter. The most significant genotype associations for the CVID phenotype and relevant subphenotypes are summarised in Table D.2. Genes with predicted deleterious variants which are found in the regions identified from the GWAS are summarised in Table 6.8. The variants in the *ADAMDEC1* gene are of particular interest as it encodes a member of the disintegrin metalloproteinase family and is thought to play a role in dendritic cell function and their interactions with T cells (Mueller et al. 1997).

Table 6.8: Variants in GWAS genotype association regions. The 4,768 predicted deleterious variants were filtered to prioritise those overlapping regions previously associated with CVID through GWAS. Variants are heterozygous unless otherwise stated. The number of times the variant is observed in both the CVID cohort and the remaining WGS500 cohort (239 individuals) is listed. Prop. = proportion. Hom = homozygous.

Gene	Chr	Pos	Ref allele	Sample allele	Protein Variant	No. CVID	Prop. CVID	No. other WGS500	Prop. other WGS500
ADAMDEC1	8	24250832	C	G	p.L10V; p.L89V	1	0.029	7	0.029
ADAMDEC1	8	24254818	T	C	p.I159T; p.I80T	1	0.029	0	0
ADAMDEC1	8	24261547	C	T	p.T451M; p.T372M	1	0.029	1	0.004
CDH23	10	73468866	G	T	p.D1040Y	1	0.029	0	0
CHL1 (Hom)	3	419552	C	G	p.D585E; p.D601E	1	0.029	0	0
HLA-E	6	30459130	A	C	p.Q276P	1	0.029	1	0.004
KCTD18	2	2.01E+08	GGCGCGGT		p.T387fs*30	1	0.029	2	0.008
PTCHD3	10	27687284	C	T	p.R748Q	1	0.029	0	0
					p.R267Q;				
SPATS2L	2	2.01E+08	G	A	p.R366Q; p.R336Q; p.R276Q	1	0.029	1	0.004
UBXN10	1	20517763	C	T	p.R237*	2	0.059	1	0.004
ZFHX3	16	72822564	TGCTGC		p.Q2289_Q2290del; p.Q3203_Q3204del	2	0.059	7	0.029
ZFHX3	16	72984595	C	A	p.A83S; p.A997S	1	0.029	11	0.046
ZFHX3	16	72993351	G	A	p.R232C	1	0.029	0	0

6.2.8 Further patient removal

The CVID cohort is composed of 32 sporadic CVID patients plus a grandmother grandson pair (D232 and C018). This pair will be analysed separately. Additionally, the individual with the variant in the *BTK* gene (D276) may have X-linked Agammaglobulinemia and will therefore also be removed from further analysis. The following genome-wide summary of variants will include the 31 sporadic CVID patients from the cohort.

6.2.9 Genome-wide variants

Filtering criteria were applied to the 14,457,884 variants identified across the cohort of 31 sporadic CVID patients to prioritise rare, predicted deleterious variants. Applying these filters results in 4,422 variants in 3,524 genes (Table 6.9).

Table 6.9: Filters for genome-wide variants.

Filter	No. variants	No. genes
All variants	14,457,884	20,627
Confidence	11,038,008	20,577
No. reverse reads ≥ 1	10,859,617	20,564
No. forward reads ≥ 1	10,687,782	20,554
Common variants	4,582,533	20,080
WGS500 MAF	2,537,263	19,584
Predicted deleterious	4,422	3,524

6.2.10 WGS pathways analysis

After applying the filtering criteria (Table 6.9) there are still an unmanageable number of variants. One approach for understanding how these variants may contribute to disease at the genome level is through the use of pathway analysis. Figure 6.4 highlights the 20 most enriched pathways and the top 100 are summarised in Table D.4. Of particular interest is the number of pathways

relating to T helper cells (including OX40, CD28 and iCOS-iCOSL signalling). T helper cells play multiple roles in the immune response including the activation of cytotoxic T cells and B cells (Luckheeram et al. 2012).

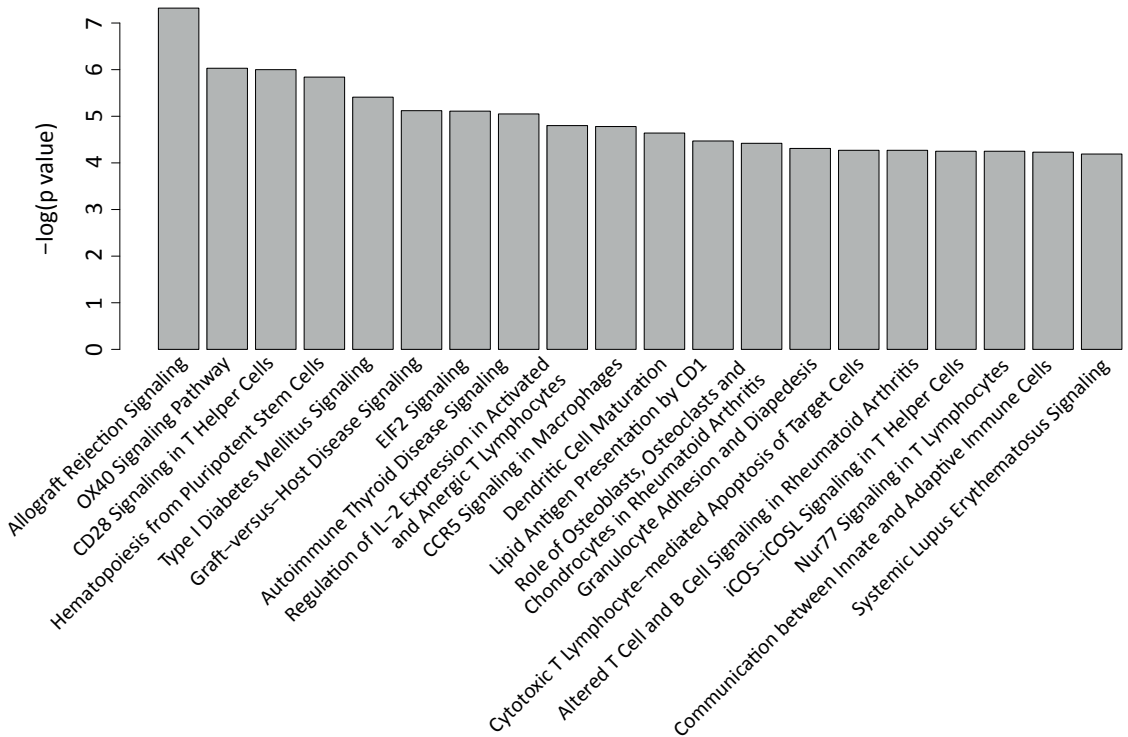


Figure 6.4: Pathway enrichment for variants. The rare, predicted deleterious variants observed across the 31 patients were used to determine pathway enrichment with IVA. The bars show the $-\log(p \text{ value})$ for the enrichment of each pathway. The 20 most enriched pathways are shown.

6.2.11 IgM analysis

CVID patients have reduced IgG and IgA but IgM can be present or absent. The cohort can be divided phenotypically depending on IgM status. For the 31 sporadic CVID patients, seven are IgM positive with the remaining 24 IgM negative. Naïve B cells express IgM and IgD and therefore if an individual has detectable IgM but not IgG, a class switch defect could be involved. The IgM positive samples were defined as cases and the IgM negative samples as controls. Filters to identify rare, predicted deleterious variants were applied as described

(Table 6.9) but included an additional filter for genetic analysis. Variants were excluded if a predicted deleterious variant was identified in any of the control samples (IgM negative) at the gene level. This resulted in 603 variants in 629 genes. Pathway enrichment for these genes is summarised in Table 6.10. There are no obvious pathways that suggest a defect in class switch recombination within this small group of patients. However, all seven patients have at least one predicted deleterious variant in the retinoic acid receptors (RAR) activation pathway. A mouse study has shown that one of these receptors (RAR γ) is involved in the regulation of inflammatory cytokine production in T lymphocytes (Dzhagalov et al. 2007).

Table 6.10: Enriched pathways for IgM+ variants.

Canonical pathway	p-value	No. genes	No. variants	No. CVID (/7)
Endothelin-1 Signalling	4.19E-03	13	13	6 (86%)
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	7.20E-03	3	3	3 (43%)
Systemic Lupus Erythematosus Signalling	1.15E-02	2	2	2 (29%)
VEGF Family Ligand-Receptor Interactions	1.52E-02	7	6	6 (86%)
Type II Diabetes Mellitus Signalling	1.57E-02	1	1	1 (14%)
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	1.82E-02	2	2	2 (29%)
MIF Regulation of Innate Immunity	2.10E-02	5	4	4 (57%)
Granulocyte Adhesion and Diapedesis	2.12E-02	1	1	1 (14%)
Fc Epsilon RI Signalling	2.71E-02	8	7	6 (86%)
Cdc42 Signalling	3.05E-02	2	2	2 (29%)
Glioblastoma Multiforme Signalling	3.51E-02	1	1	1 (14%)
RAR Activation	3.55E-02	11	11	7 (100%)
Corticotropin Releasing Hormone Signalling	3.60E-02	10	10	7 (100%)
MIF-mediated Glucocorticoid Regulation	4.01E-02	4	3	3 (43%)
CD28 Signalling in T Helper Cells	4.20E-02	1	1	1 (14%)
Phospholipases	4.39E-02	5	5	3 (43%)
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	4.84E-02	2	2	2 (29%)
Relaxin Signalling	4.88E-02	9	9	6 (86%)

6.2.12 Infections only vs complicated subphenotype

Polyclonal Lymphoproliferation represents a complicated case of CVID compared to the Infections Only subphenotype. For the 31 sporadic CVID patients, nine have Polyclonal Lymphoproliferation with the remaining 22 Infections Only. Polyclonal Lymphoproliferative patients were defined as cases

and the Infections Only patients as controls. Filters to identify rare, predicted deleterious variants were applied as described (Table 6.9) but included an additional filter for genetic analysis. Variants were excluded if a predicted deleterious variant was identified in any of the control samples (Infections Only) at the gene level. This resulted in 783 variants in 779 genes. Pathway enrichment for these genes is summarised in (Table 6.11). Seven of the patients have at least one predicted deleterious variant in a pathway involved in DNA repair (double strand break repair by non-homologous end joining or role of BRCA1 in DNA damage response). Lymphoma is one of the clinical complications observed in CVID patients and it may be more likely to arise in patients with Polyclonal Lymphoproliferation (Cunningham-Rundles 2010). Defects in DNA repair genes have been shown to result in the hyper IgM syndrome so it is hypothesised that more subtle DNA repair variants may play a role in CVID (Offer et al. 2010). A functional assay is currently being conducted by collaborators in Clinical Immunology for this cohort of patients to determine whether these observed variants result in reduced response to DNA damage.

Functional genomic response in CVID

Table 6.11: Enriched pathways for Polyclonal Lymphoproliferation subphenotype.

Canonical pathway	p-value	No. genes	No. variants	No. CVID (/9)
cAMP-mediated signalling	3.18E-03	17	18	8 (89%)
Cellular Effects of Sildenafil (Viagra)	6.75E-03	13	17	7 (78%)
DNA Double-Strand Break Repair by Non-Homologous End Joining	8.32E-03	4	4	3 (33%)
UVA-Induced MAPK Signalling	9.02E-03	9	9	6 (67%)
NF- κ B Signalling	1.19E-02	2	2	2 (22%)
Retinoic acid Mediated Apoptosis Signalling	1.48E-02	7	7	5 (56%)
Role of BRCA1 in DNA Damage Response	1.48E-02	7	7	7 (78%)
CD28 Signalling in T Helper Cells	1.52E-02	1	1	1 (11%)
Dendritic Cell Maturation	1.67E-02	3	3	2 (22%)
Signalling by Rho Family GTPases	1.86E-02	4	4	3 (33%)
Type I Diabetes Mellitus Signalling	2.08E-02	1	1	1 (11%)
Axonal Guidance Signalling	2.18E-02	10	10	8 (89%)
Rac Signalling	2.26E-02	1	1	1 (11%)
G-Protein Coupled Receptor Signalling	2.31E-02	17	18	8 (89%)
Tec Kinase Signalling	2.45E-02	3	3	3 (33%)
IL-8 Signalling	2.51E-02	3	3	3 (33%)
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	2.64E-02	6	7	5 (56%)
Communication between Innate and Adaptive Immune Cells	2.92E-02	1	1	1 (11%)
fMLP Signalling in Neutrophils	3.01E-02	1	1	1 (11%)
Gai Signalling	3.28E-02	10	10	5 (56%)
3-phosphoinositide Degradation	3.43E-02	2	2	2 (22%)
RhoGDI Signalling	3.89E-02	3	3	3 (33%)
Regulation of IL-2 Expression in Activated and Anergic T Lymphocytes	3.98E-02	1	1	1 (11%)
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	4.17E-02	4	5	5 (56%)
D-myo-inositol (1,4,5,6)	4.73E-02	2	2	2 (22%)
Tetrakisphosphate Biosynthesis	4.73E-02	2	2	2 (22%)
D-myo-inositol (3,4,5,6)	4.73E-02	2	2	2 (22%)
Tetrakisphosphate Biosynthesis	4.73E-02	2	2	2 (22%)
CXCR4 Signalling	4.85E-02	2	3	3 (33%)

6.2.13 Grandmother-grandson pair

In the cohort, patient D232 is the maternal grandmother of patient C018 (male). The mother of C018 was not diagnosed with CVID but does have IgA deficiency (IgAD). Patients with this disease lack IgA and studies have shown a higher prevalence than expected of CVID in families with IgAD suggesting a shared pathogenesis (Vorechovsky et al. 1999). D232 and C018 both have an Infections Only subphenotype but differ in their IgM status (C018 is IgM negative). The variability in subphenotype results in a complicated family structure but an analysis was carried out to determine whether any variants are shared between the grandmother and grandson. After applying the standard filters, as described (Table 6.9) there were 329 predicted deleterious variants. The results of investigating different possible inheritance patterns are summarised in Table 6.12. A single variant in the *MT-CO2* (mitochondrial) gene was observed in both patients. *MT-CO2* binds proteins encoded by two of the PID candidate genes: *ELANE* (Salipante et al. 2009) which is associated with cyclic neutropenia (Al-Herz et al. 2011) and *LIG4* (Woods et al. 2012) which is associated with DNA ligase IV deficiency (Al-Herz et al. 2011). Parent-of-origin penetrance differences have been observed in families with IgAD with greater numbers of affected offspring from women with IgAD than from men with IgAD. The predominance of maternal transmission and the different clinical manifestations provides further evidence for the variant in the mitochondrial DNA playing a role. Father-child pairs are observed but this could be explained by the complex nature of the diseases with further variants and genes being involved (Vorechovsky et al. 1999).

Table 6.12: Shared variants between grandmother and grandson. Within the cohort, patient D232 is the maternal grandmother of patient C018. Applying the standard filters, 329 predicted deleterious variants are observed across the two samples. The results of investigating different possible inheritance patterns are summarised

Filter	Variants	Genes
Heterozygous variant in both samples	36	36
Homozygous variant in both samples	1	1
X-linked variant	0	0

6.2.14 RNA-seq QC

An initial pilot study was carried out to generate RNA-seq data for CD19+ B cells from three CVID patients and three healthy controls. Data were aligned to the reference and reads mapped to genes using HTSeq (Anders et al. 2014) (Core Bioinformatics, WTCHG).

QC was carried out to determine the proportion of reads mapping to genes and variation within the dataset (Figure 6.5). Data were transformed using R (DEseq, Anders and Huber 2010) for QC plots. Transcripts were filtered so only those with >10 reads in at least 3 samples (smallest group size for comparison) were retained for further analysis. This resulted in a reduction from 32,426 to 16,546 transcripts. Normalisation was carried out using EdgeR (Robinson et al. 2010) for the subsequent differential gene expression analysis.

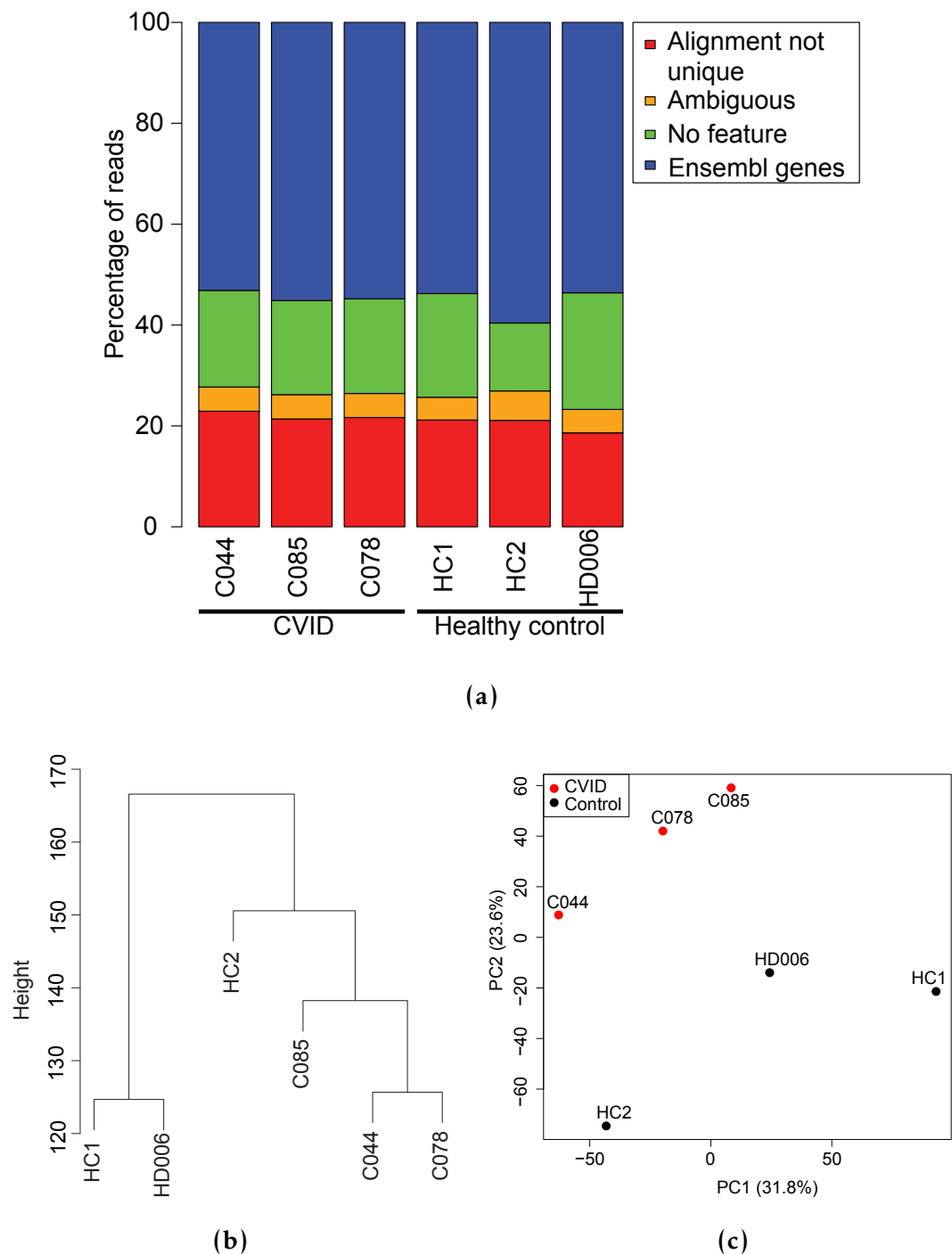


Figure 6.5: CVID RNA-seq QC. RNA-seq data were generated for three CVID patients and three healthy controls. a) A summary of the alignment of reads to the genome. b) A hierarchical tree illustrating the relationship between clusters of samples based on reads which map to genes. The height of the branches indicates the strength of separation between individuals. c) The first two PCs are plotted with the proportion of variance explained by each component (based on reads which map to genes).

6.2.15 Differential gene expression analysis

Differential gene expression between the three CVID patients and three healthy controls was determined for 16,546 transcripts using the edgeR R package (Robinson et al. 2010). 262 transcripts were differentially expressed (FDR <0.05 and fold change >2) and are summarised in Table D.5. The volcano plot illustrates the significance and fold change for differential expression compared to healthy controls (Figure 6.6). Two genes which are both downregulated in CVID compared to controls are of particular interest. *IGHA2* encodes immunoglobulin heavy constant alpha 2 and *IGJ* encodes immunoglobulin J polypeptide, a linker protein for immunoglobulin alpha and mu polypeptides. Box plots for these two genes are shown in Figure 6.7. This figure also includes box plots for the genes *SDR16C5* and *HBB*. Both these genes pass the thresholds for differential expression but the box plots highlight issues associated with studies with small samples sizes as *SDR16C5* may be confounded by low sequencing coverage and *HBB* by an outlying individual. The addition of further CVID patients and healthy controls to this analysis should reduce these false positive results and increase the power for detecting further differentially expressed genes.

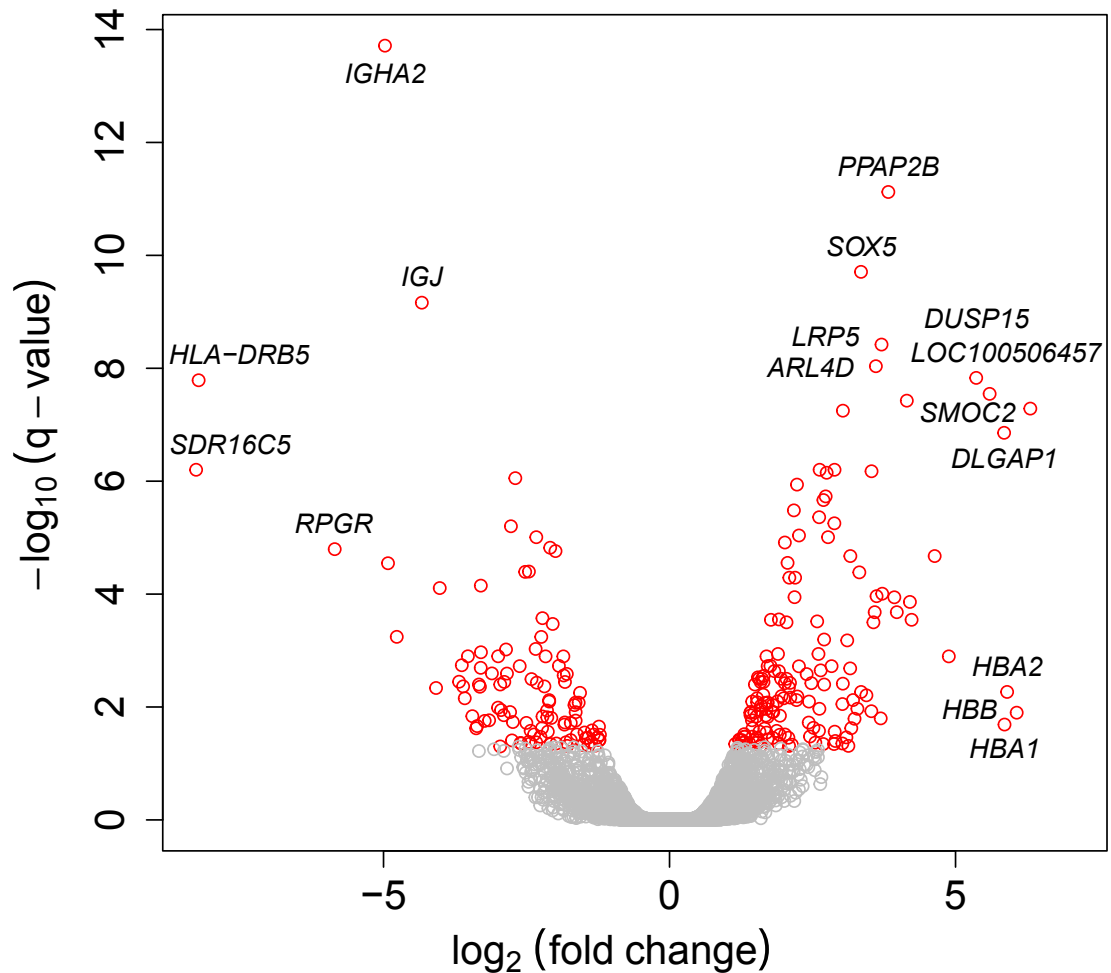


Figure 6.6: Volcano plot of differentially expressed genes. The significance of differential expression between CVID patients and healthy controls is plotted against fold change. Each point represents a single gene. Red points highlight genes with $\text{FDR} < 0.05$ and fold change > 2 . A positive fold change indicates upregulation in the CVID patients. The most differentially expressed genes are labelled.

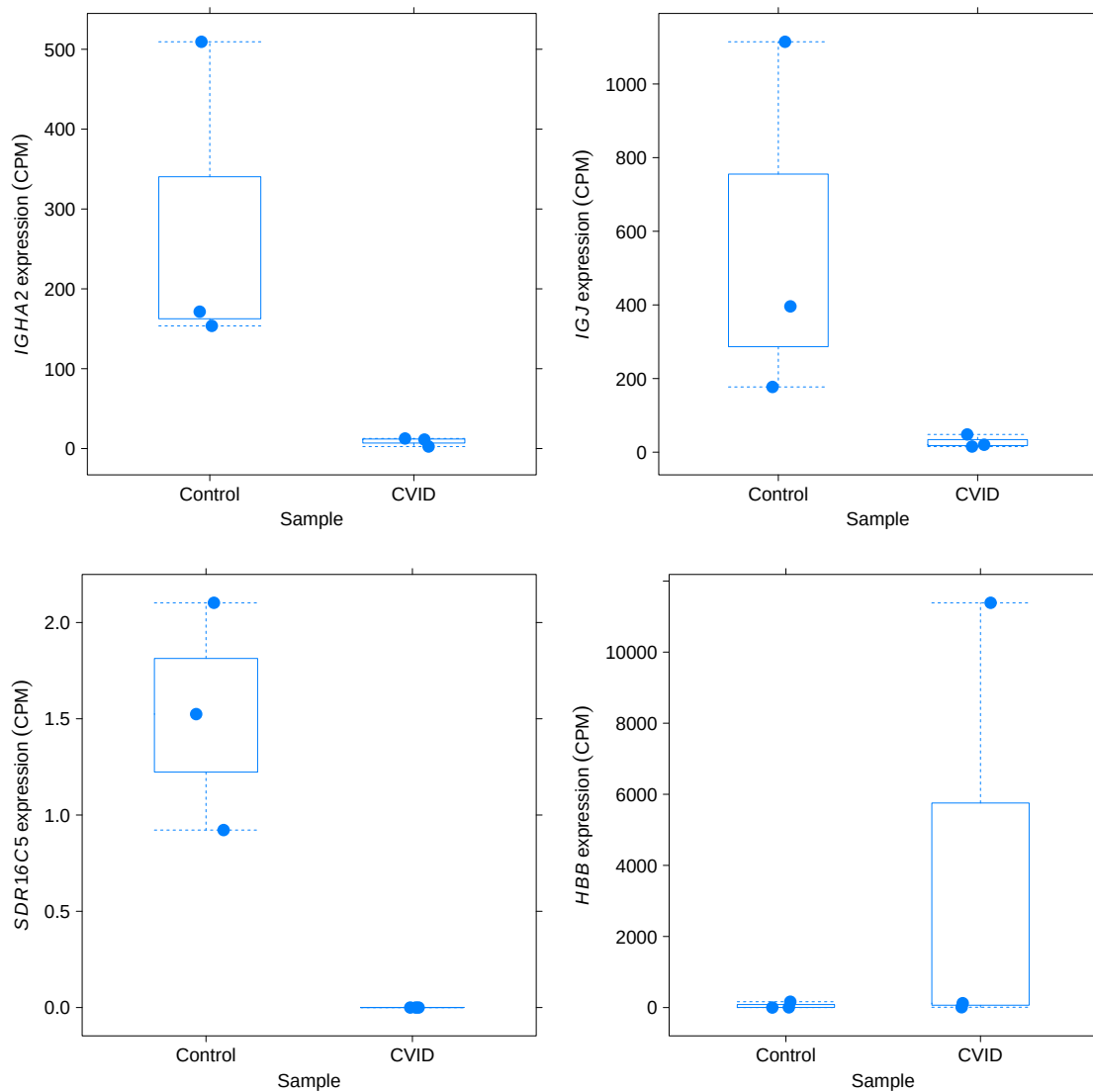


Figure 6.7: Box plot for differentially expressed genes. The expression value for each individual is plotted for four differentially expressed genes. *IGHA2* and *IGJ* are examples of genes which are downregulated in the CVID patients compared to the healthy controls. *SDR16C5* provides an example of a differentially expressed gene which may be confounded by low sequencing coverage. *HBB* is an example of a gene which may be confounded by an outlying individual.

6.2.16 Differential gene expression previously described genes

The differentially expressed genes were screened for those that had previously been described in CVID or PID (Tables 6.1 and D.1). Two HLA genes (HLA-DRB5 and HLA-DQB1) were differentially expressed but this result is likely confounded by mapping errors due to the highly polymorphic nature of this

region. Four PID candidate genes were differentially expressed and were all found to have increased expression in the CVID patients compared to controls (Table 6.13 and Figure 6.8). *IL10RA* encodes a receptor for interleukin 10 which mediates the immunosuppressive signal of interleukin 10, inhibiting synthesis of proinflammatory cytokines (Pruitt et al. 2014). *FAS*, a member of the TNF-receptor superfamily, encodes a receptor involved in regulation of programmed cell death (Pruitt et al. 2014). *SERPING1* encodes a gene involved in the regulation of the complement cascade (Pruitt et al. 2014). *STX11* encodes a member of the syntaxin family which are involved in the targeting and fusion of intracellular transport vesicles (Pruitt et al. 2014).

Apart from the HLA genes mentioned above, no further genes from significantly associated regions of the GWAS were found to be differentially expressed in this initial study.

Table 6.13: Differential expression of PID candidate genes. Four PID candidate genes were differentially expressed between CVID patients and healthy controls (FDR <0.05 and fold change (FC) >2).

Gene	FDR	log(FC)
<i>FAS</i>	3.11E-02	1.93
<i>IL10RA</i>	5.00E-02	1.59
<i>SERPING1</i>	1.59E-02	3.69
<i>STX11</i>	4.22E-02	1.54

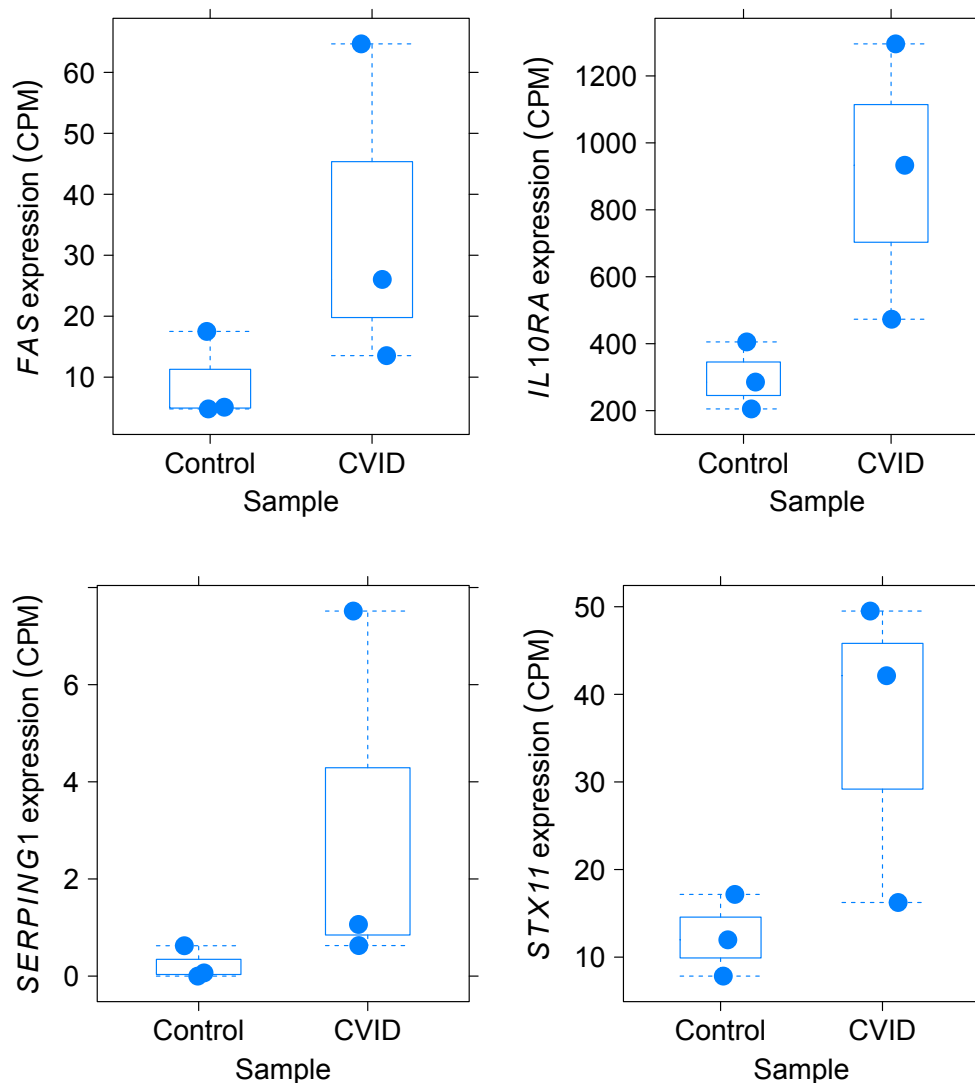


Figure 6.8: Box plots for differentially expressed candidate genes. The expression value for each individual is plotted for four PID candidate genes. All four genes have increased expression in the CVID samples compared to the controls.

6.2.17 RNA-seq pathway analysis

Pathway enrichment for the differentially expressed genes was determined through IPA. Figure 6.9 shows the 20 most enriched pathways for differential gene expression. The pathways which were also enriched for WGS predicted deleterious variants are highlighted in blue. The full list of enriched pathways are shown in Table D.6.

Functional genomic response in CVID

The Ingenuity Knowledge base contains 644 pathways but for the WGS variants, only the top 100 most significantly enriched pathways are reported in IVA. There is greater overlap between these pathways and the 52 pathways enriched for differential gene expression than expected by chance (Fisher's exact test p value <0.0001). The overlap of pathways is summarised in Figure 6.10. Pathways involving helper T cells are again highly represent but it is also of note that dendritic cell maturation is enriched for both predicted deleterious variants and differentially expressed genes.

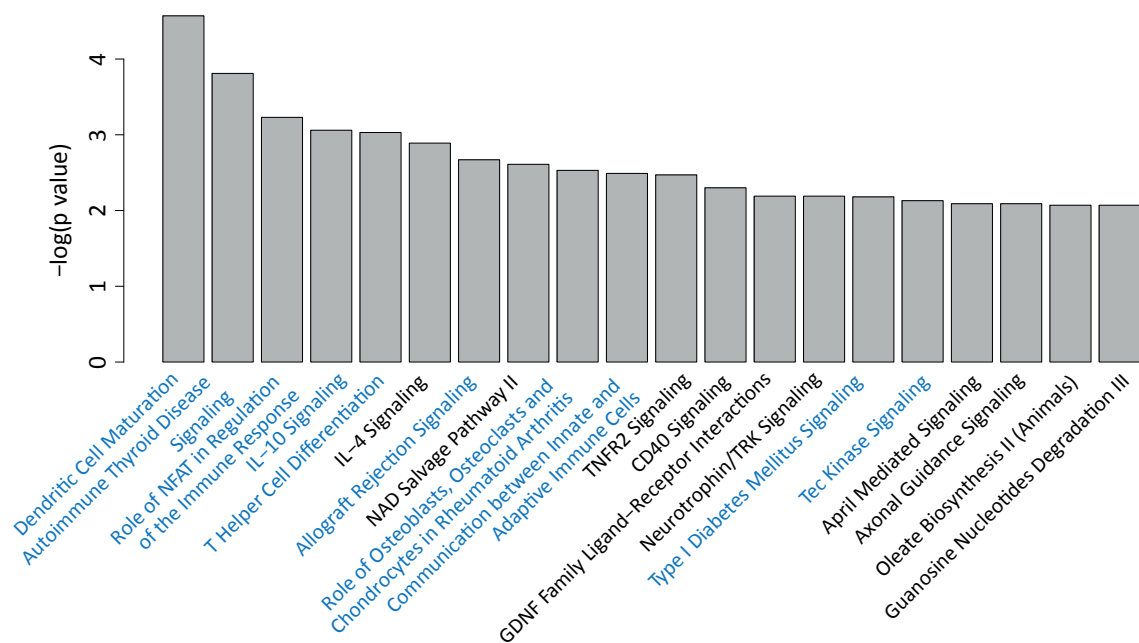


Figure 6.9: Pathway enrichment for differentially expressed genes. The differentially expressed genes between CVID patients and healthy controls were used to determine pathway enrichment with IPA. The bars show the $-\log(p \text{ value})$ for the enrichment of each pathway. The 20 most enriched pathways are shown. Pathways highlighted in blue were also enriched for WGS variants.

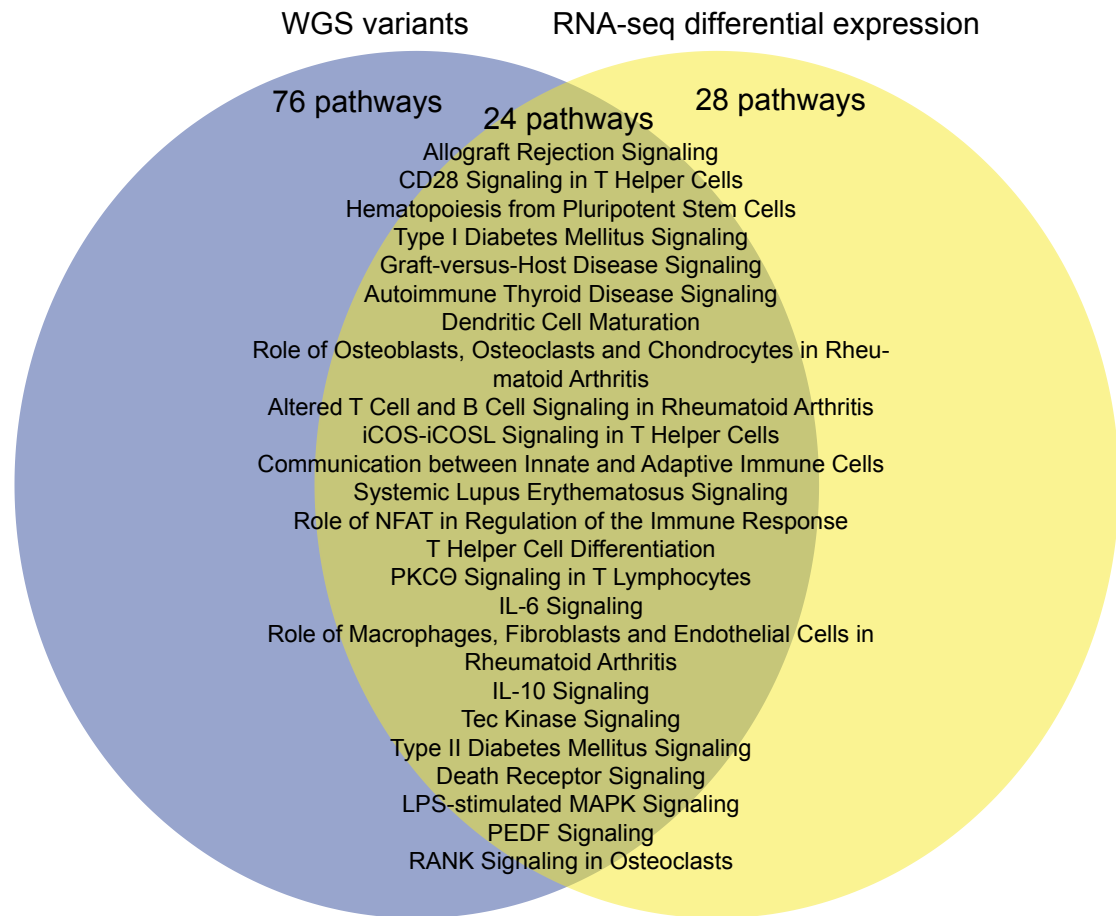


Figure 6.10: Overlap of enriched pathways for WGS and RNA-seq. The 100 most significantly enriched pathways for WGS predicted deleterious variants were reported by IPA. 52 pathways were enriched for differentially expressed genes from the RNA-seq analysis. The overlap of enriched pathways between these two datasets is greater than expected by chance (Fisher's exact test p value <0.0001).

6.2.18 Integration of WGS and RNA-seq

To resolve potentially disease causing variants, the RNA-seq gene expression data were integrated with the WGS to help prioritise non-coding variants. For each of the three patients with RNA-seq data, filters were applied to the WGS data as described in Table 6.9 with an alteration to the predicted deleterious filter to include variants which interrupt promoter and enhancer regions in addition to those with predicted deleterious effects to protein function. IVA utilises the JASPAR database (Mathelier et al. 2014) and ENCODE chromatin-

Functional genomic response in CVID

immunoprecipitation experiments (Encode Project Consortium 2012) to identify variants which disrupt predicted or known promoters or TF binding sites.

Fold change differences in gene expression for each patient compared to the healthy controls were determined. Integrating this with the WGS variant data allowed the identification of homozygous or compound heterozygous variants which resulted in reduced gene expression or heterozygous variants which resulted in increased gene expression (Table 6.14 and Figure 6.11)

Table 6.14: Summary of WGS and RNA-seq integrated variants. RNA-seq data were integrated with WGS data for three CVID patients. WGS variants were filtered to prioritise rare, predicted deleterious variants including those which interrupt promoter and enhancer regions. Gene expression for genes containing WGS variants were filtered by fold change (compared to median healthy control values). A genetic filter was applied to identify homozygous, compound heterozygous, or heterozygous variants. A biological context filter was applied to identify genes connected to CVID/PID candidate genes. Homozygous and compound heterozygous loss of function variants were classified with a FC <0.5 and heterozygous variants gain of function variants were classified with a fold change >2. Compound het = compound heterozygous. FC = fold change

(a) Homozygous loss of function variants.

Filter	Patient C085		Patient C078		Patient C044	
	No. variants	No. genes	No. variants	No. genes	No. variants	No. genes
Predicted deleterious	784	718	819	748	760	689
Expression FC <0.5	50	51	47	42	24	22
Homozygous	0	0	5	5	1	1
Biological context	0	0	1	1	0	0

(b) Compound heterozygous loss of function variants.

Filter	Patient C085		Patient C078		Patient C044	
	No. variants	No. genes	No. variants	No. genes	No. variants	No. genes
Predicted deleterious	784	718	819	748	760	689
Expression FC <0.5	50	51	47	42	24	22
Compound het	16	15	8	8	3	4
Biological context	0	0	3	4	0	0

(c) Heterozygous gain of function variants.

Filter	Patient C085		Patient C078		Patient C044	
	No. variants	No. genes	No. variants	No. genes	No. variants	No. genes
Predicted deleterious	784	718	819	748	760	689
Expression FC <2	30	36	39	41	46	46
Heterozygous	29	36	38	40	44	44
Biological context	0	0	10	12	2	2

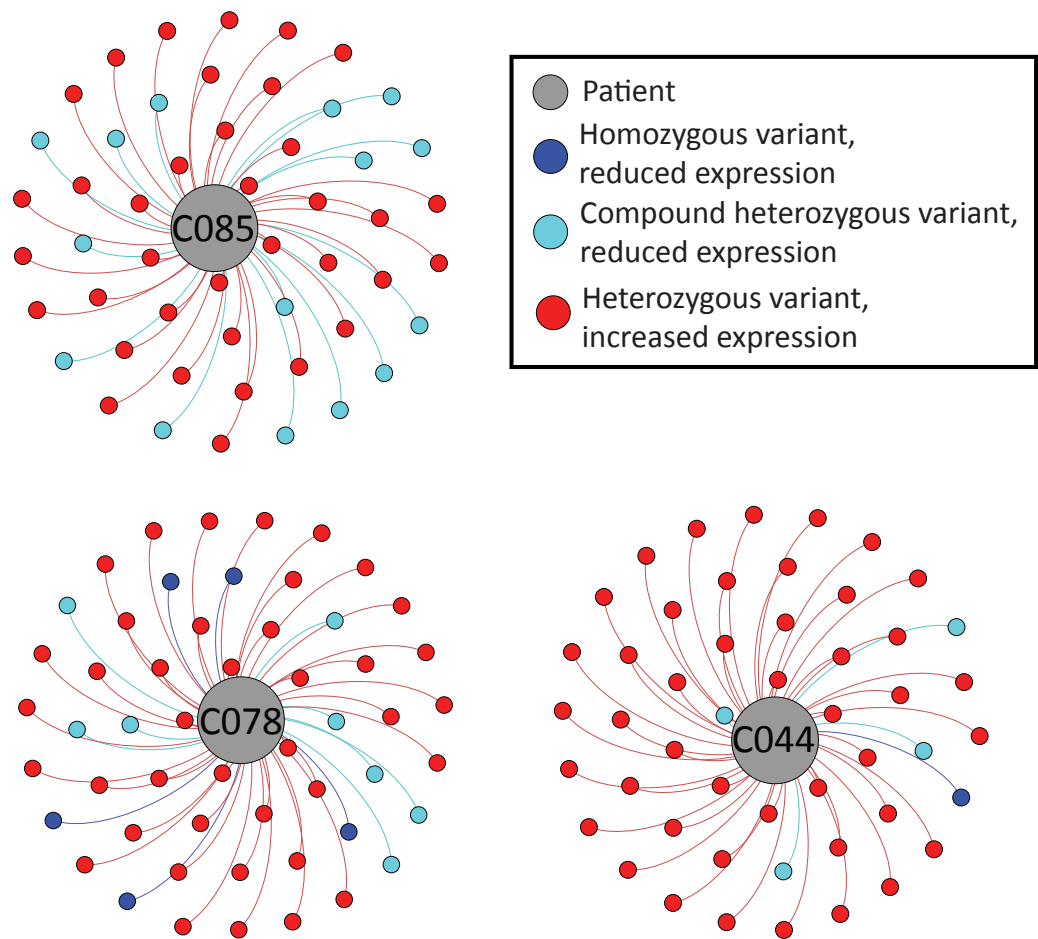


Figure 6.11: WGS and RNA-seq integrated variants. The WGS variants and RNA-seq gene expression values were integrated to identify variants in genes which also show evidence of altered expression in the patient compared to healthy controls (fold change >2). The type of variant and altered expression is illustrated for each of the three patients.

Patient C085

Patient C085 is a female patient with an Infections Only and IgM absent subphenotype. Combining the RNA-seq with the WGS data provided evidence for a number of compound heterozygous variants with reduced gene expression and heterozygous variants with increased expression. Following the application of the biological context filter with IVA, none of these genes are connected to the previously described CVID or PID candidate genes suggesting that the underlying genetics for this particular individual may involve genes which have not been previously described in the context of CVID. This is further supported by Figure 6.2 which shows this individual has two predicted deleterious variants in genes upstream of the CVID candidate genes but is not connected by these genes to any other individuals in the cohort.

Patient C078

Patient C078 is a male patient with an Infections Only and IgM present subphenotype. This patient has a number of variants in genes, with corroborating expression fold changes, which are connected to the CVID or PID candidate genes. For example, a homozygous variant is predicted to result in promoter loss for the *NOD1* gene and the RNA-seq data confirm a reduction in expression. This gene is connected upstream to three PID candidate genes (*CASP8*, *IKBKG* and *XIAP*). A compound heterozygous variant was identified in the *SH3BP2* gene which also showed reduced expression. In T, NK and basophilic cells, this gene acts as a cytoplasmic adaptor protein to positively regulate transcriptional activity (Pruitt et al. 2014). The connections between this gene and the CVID/PID phenotypes are summarised in Figure 6.12. From the network figure of predicted deleterious variants in genes upstream of the CVID candidate genes (Figure 6.2) this patient clusters with a group of predominantly Infections Only with many shared genes. The integration of the RNA-seq data with the WGS

variant data appears to support the hypothesis that genes connected to known CVID genes may play a significant role in the underlying genetic aetiology of a subset of the patients.

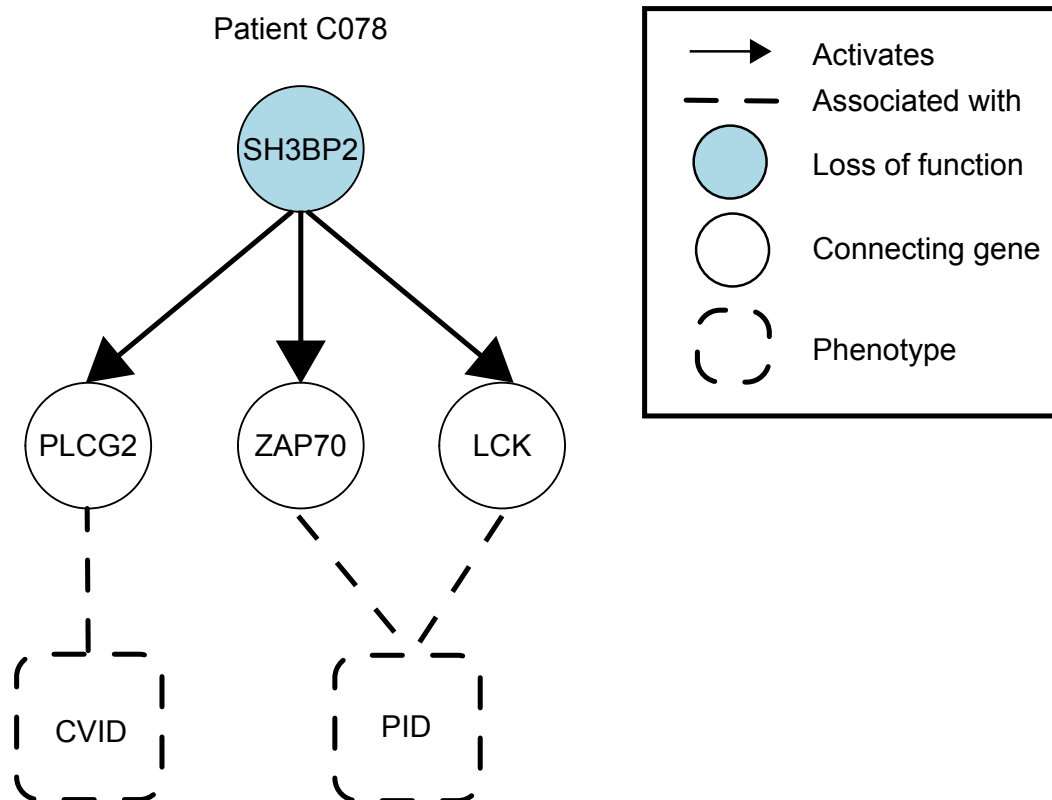


Figure 6.12: *SH3BP2* compound heterozygous variant. Patient C078 possesses a heterozygous variant in the *SH3BP2* gene. Expression of this gene is 5x lower in C078 than the median expression for the healthy controls. This gene is upstream of a CVID candidate gene and two PID candidate genes. The binding of *SH3BP2* to *PLCγ* is involved in the regulation of NK cell-mediated cytotoxicity (Jevremovic et al. 2001)

Patient C044

Patient C044 is a female patient with Polyclonal Lymphoproliferation and absent IgM. This patient has heterozygous variants in two PID candidate genes (*IL21R* and *ITGB2*). These variants are both in the promoter regions of the respective genes and are predicted to cause promoter loss. In contrast, the RNA-seq data indicates increased gene expression compared to healthy controls. From the WGS variant analysis, this patient only had one predicted deleterious variant in a gene

upstream of a CVID candidate gene. Like patient C085, this individual is not connected by this gene to any others in the cohort (Figure 6.2). Further analysis of the variants with corroborating gene expression data may assist in determining new disease causing genes.

6.2.19 Overlap with sepsis

Although less common, bacterial infections which cause sepsis are observed in some CVID patients and often the same organisms are responsible for the infections. To understand any key genes involved in both the rare (CVID) and common (sepsis) response to bacterial infection, the results of the gene expression and eQTL analysis for the severe sepsis cohort (described in Chapters 3 and 4) were compared to those for the CVID cohort.

A key finding from investigating the functional genomic response in severe sepsis was the identification of two clusters in the cohort based on variation in gene expression (Section 3.2.12). These two clusters had significantly different survival rates, providing a more informative comparison for differences in gene expression than the traditional 28 day survival threshold. The genes identified as differentially expressed between severe sepsis patients in cluster 1 (poor outcome) and cluster 2 (good outcome) were compared to those differentially expressed between CVID patients and controls. Assuming the improved outcomes of cluster 2 sepsis patients are the result of successfully fighting the bacterial infections, genes important for this response may be upregulated in these individuals and downregulated in the CVID patients. Figure 6.13 summarises the overlap of differentially expressed genes between the two cohorts.

A number of these genes are of particular interest for example *DOK3* gene suppresses B cell activation by inhibiting the B-cell receptor activation of the JNK signal transduction pathway (Robson et al. 2004). This gene is downregulated in

cluster 2 sepsis patients (good outcome) but upregulated in the CVID patients. In contrast, *CD96* shows the opposite pattern of expression in the two cohorts. This gene is thought to play a role in the adhesion of T cell and NK cells but a recent study has also shown that it acts as a negative regulator of cytokine responses of NK cells (key components of the innate immune response) (Chan et al. 2014).

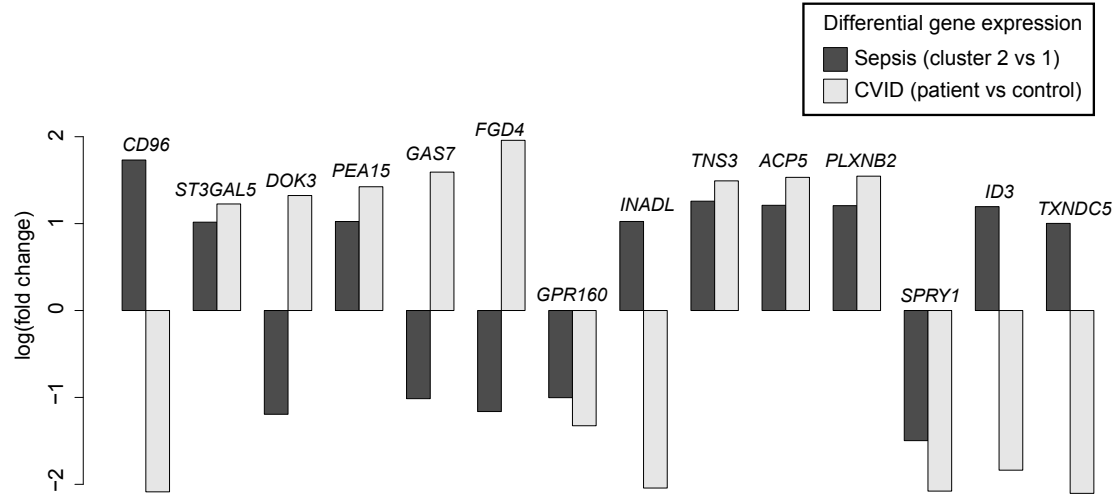


Figure 6.13: Overlap of differentially expressed gene between CVID and sepsis. Gene expression signatures in the CVID cohort were compared to those identified in the severe sepsis cohort. Two clusters in the sepsis cohort, based on gene expression signatures, show difference in survival rates and immune response to infection. Genes differentially expressed between cluster 2 sepsis patients (better outcome) and cluster 1 patients were compared to the differentially expressed genes between CVID patients and controls.

The regions containing the most significant genotype associations from the CVID GWAS were compared to the eQTLs determined in the sepsis cohort (Chapter 4). In total, six of the CVID GWAS loci showed evidence of eQTL in the sepsis cohort (Table 6.15). The region on chromosome 6 containing the MHC showed the greatest overlap however this was not unexpected due to the high degree of genetic variation and great gene density. This particular region will therefore require further analysis before conclusions can be made. Genes which also contain a predicted deleterious variant in the CVID cohort are highlighted in bold. Outside of the MHC region, there are just two genes for which this applies: *CDH23* and *IQGAP2*. *CDH23* encodes a calcium dependent cell-cell

Functional genomic response in CVID

adhesion glycoprotein which is associated with hearing loss (Miyagawa et al. 2012) whereas *IQGAP2* interacts with cell adhesion and signalling molecules to regulate cell morphology and motility (Pruitt et al. 2014) so may be more likely to be involved with the CVID phenotype.

Table 6.15: Overlap between sepsis eQTL and CVID GWAS. Regions of association with CVID (Orange et al. 2011) were compared to eQTL identified in the sepsis cohort. The genes highlighted as bold also contain a predicted deleterious variant in the CVID cohort.

CVID GWAS region	Gene with sepsis <i>cis</i> eQTL		
chr1: 20496782-20618204	<i>ECE1</i>		
chr5: 76092225-76093207	<i>IQGAP2</i>		
chr6: 25462824-25466889	<i>CMAH</i>		
chr6: 27128806-32413317	<i>ABCF1</i>	<i>HCG9</i>	<i>PPT2</i>
	<i>AGER</i>	<i>HIST1H2BD</i>	<i>PSORS1C3</i>
	<i>AGPAT1</i>	<i>HLA-DOB</i>	<i>RDBP</i>
	<i>AIF1</i>	<i>HLA-DPB1</i>	<i>RFP</i>
	<i>ATF6B</i>	<i>HLA-E</i>	<i>RGL2</i>
	<i>BAT2</i>	<i>HLA-F</i>	<i>RPP21</i>
	<i>BAT3</i>	<i>HMGN4</i>	<i>SKIV2L</i>
	<i>BTN2A1</i>	<i>HSPA1A</i>	<i>TAP2</i>
	<i>BTN3A1</i>	<i>HSPA1B</i>	<i>TAPBP</i>
	<i>C6orf48</i>	<i>IER3</i>	<i>TNF</i>
	<i>CLIC1</i>	<i>LOC100129668</i>	<i>TRIM10</i>
	<i>CSNK2B</i>	<i>LOC642073</i>	<i>TRIM26</i>
	<i>DDAH2</i>	<i>LOC646366</i>	<i>TRIM27</i>
	<i>DOM3Z</i>	<i>LST1</i>	<i>TRIM39</i>
	<i>DPCR1</i>	<i>LY6G5C</i>	<i>VAR2</i>
	<i>FLOT1</i>	<i>MICB</i>	<i>ZNF193</i>
	<i>GABBR1</i>	<i>MRPS18B</i>	<i>ZNRD1</i>
	<i>GPSM3</i>	<i>MSH5</i>	
	<i>HCG27</i>	<i>PPP1R10</i>	
chr8: 23690631-24625718	<i>NKX3-1</i>		
	<i>DOCK5</i>		
chr10: 73413824-73414577	<i>CDH23</i>		

6.3 Discussion

The work in this chapter illustrates some of the challenges associated with WGS in clinical samples but also the power of combining multiple sequencing techniques to provide a better understanding of multifactorial disease.

6.3.1 Filtering WGS variants

The study highlights the volume of information that is generated during WGS. When dealing with small family pedigrees with obvious inheritance patterns it is often possible to use unaffected individuals as an internal control to reduce the number of potentially causative variants as they are not expected to be observed in these individuals (Guo et al. 2012; Martin et al. 2014). In contrast, multifactorial diseases such as CVID, with unclear inheritance, do not have obvious internal controls. The analysis pipeline for the identification of predicted deleterious variants was improved by the use of IVA software with the quick and easy application of filter thresholds. The most informative filter for the prioritisation of variants was the WGS500 project MAF. Applying a threshold of 5% for this filter frequently reduced the variant list by approximately 50%, highlighting the advantage of sequencing these sporadic cases alongside many other individuals in order to remove variants caused by artefacts of the sequencing pipeline.

6.3.2 Pathway analysis

Pathway analysis revealed 24 pathways were enriched for both WGS predicted deleterious variants and RNA-seq differentially expressed genes. This overlap was greater than expected by chance. Despite measuring gene expression in B cells, pathways involving helper T cells and dendritic cell maturation were enriched.

Impaired dendritic cell maturation has been observed in a subset of CVID patients. Dendritic cells can provide key signals required for naïve CD4⁺ T cells to differentiate into T helper cells which in turn are important for B cell class switch recombination (Scott-Taylor et al. 2004). Th17 helper T cells (which differentiate from naïve CD4 T cells) also share cues with B cell differentiation so although they are effector cells for IgM/IgA B cells, upstream defects may account for the pathway enrichment observation (Yazdani et al. 2014). Overall the results from the pathway analysis suggest variants interrupting interactions between B cells, T cells and dendritic cells may account for some of the underlying immunopathogenesis of CVID.

6.3.3 Integrating WGS and RNA-seq

The genes and pathways with predicted deleterious variants or differentially expressed genes can give an overall view of variants across the genome that may be contributing to the CVID phenotype. However, this only focuses on the variants within the coding exome. The integration of the RNA-seq data with the WGS allowed non-coding variations to be identified which showed evidence of altered gene expression in the particular CVID patient compared to healthy controls.

The consequences of regulatory variants are harder to predict than variants falling in coding regions because the relationship between the variant and gene expression is not as well understood (Glazier et al. 2002; Gaffney 2013). Although individual examples were identified, function follow up will be required to understand specific mechanisms of action.

6.3.4 Advantages of WGS

This study highlighted the value of WGS over previous sequencing methods. The patient with the *BTK* mutation had been screened for known variants but at the

time the resources were not available to detect novel variants. While WGS may produce potentially unnecessary amounts of additional data, the clinical data for this particular individual did not fit classic X-linked Agammaglobulinemia. Investigations are currently ongoing including a western blot to detect protein level. The advantage of WGS is the ability to look for further variants that may be contributing to the disease if the *BTK* mutation is not conclusive.

6.3.5 Limitations

The analyses presented have focused on SNPs and small InDels. Structural variants may contribute to the underlying genetic aetiology of disease (Manolio et al. 2009) but they are challenging to identify in WGS data (Hormozdiari et al. 2009).

A number of potentially disease causing variants have been identified for particular patients through WGS and RNA-seq but further functional assays will be required to confirm the biological effect of the variants.

The results of pathway enrichment can be biased by genes which have been well studied and are therefore more thoroughly annotated (Wang et al. 2011). For IVA the many research fields relating to cancer have created a bias in annotation (IVA, personal communication).

6.3.6 Conclusion

Novel variants and pathways have been identified which may play a role in the underlying immunopathogenesis of the disease. The integration with RNA-seq for a small number of patients allowed the prioritisation of non-coding variants based on evidence of altered gene expression.

Chapter 7

General Discussion

This thesis has demonstrated the value of integrating multiple functional genomic techniques to further our understanding of the mechanisms underlying variation in response to infection. A functional genomic approach was applied to investigate three clinical cohorts: patients with severe sepsis, an influenza challenge study and patients with COVID. This provided further insights into the variation in response to infection for each of the cohorts.

The key finding from the transcriptomic response to severe sepsis was the identification of two clusters, based on gene expression signatures, which have different survival rates and identify individuals with distinct immune responses to sepsis. Mapping gene expression as an eQTL and comparing to other datasets for healthy individuals and innate immune stimulated cells identified regulatory variants specific to the context of sepsis. This was further complemented by gene expression signatures distinguishing severity of illness following influenza challenge in healthy individuals. Finally gene and pathways which may play a role in the underlying immunopathogenesis of COVID were identified through WGS and RNA-seq. Comparison to the sepsis cohort analysis identified key immune related genes involved in rare and common responses to bacterial infection.

The most interesting result to follow up is the biology underlying the two clusters identified in the severe sepsis cohort. In particular, determining whether the observed differences in immune response are the result of variation in response

to infection at the individual level or due to sampling different phases of response may contribute to the improvement of diagnosis and treatment of the disease.

This chapter discusses the themes arising from the thesis that have broader impact in the context of the wider research field.

7.1 Heterogeneity within clinical cohorts

The misclassification of phenotype has been shown to reduce the power to detect genetic associations with disease, particularly in case-control studies. However, phenotype or genetic heterogeneity within multifactorial traits can have a similar effect (Manchia et al. 2013). Heterogeneity is a theme central to all three cohorts presented in this thesis. For each cohort, well defined clinical phenotypes were selected but despite these efforts, heterogeneity was still observed.

Within the cohort of severe sepsis patients, the greatest heterogeneity was observed following the unsupervised approach to signatures in gene expression. Two clusters within the cohort, based on gene expression signatures, had different survival rates and identified individuals with distinct immune responses to sepsis which could not be explained by clinical data alone (Table 3.13). The genes differentially expressed between these two cohorts were enriched for those with evidence of a *cis* eQTL (Figure 4.9) providing further evidence for a genetic component to the underlying heterogeneity. This finding is of particular importance to the field of sepsis research as it has been hypothesised that heterogeneity within cohorts has contributed to the failure of new treatments (Wiersinga 2011). This study has shown that heterogeneity remains even when patients are group based on a similar clinical phenotype.

A potentially confounding factor for the severe sepsis cohort is the timepoint of infection. By definition, the CAP patients develop infection in the community and therefore each individual patient will be admitted to ICU at a different timepoint of disease. In this situation, insights from challenge

studies can provide information to help understand the signatures observed. However, heterogeneity was similarly observed within the 22 healthy individuals undergoing influenza challenge as part of the FLU002 study (Chapter 5). Seven individuals were originally defined as LCI based on self-reported symptoms and clinical observations. Great variation in symptoms score was observed within the seven individuals and this was found to be reflected in the transcriptomic response to challenge. Gene expression signatures allowed the classification of four of these individuals with moderate/severe LCI which could be a more appropriate classification for challenge studies (Section 5.2.6). This demonstrated that even under much more controlled conditions, variation in response are still observed.

WGS was conducted for a cohort of CVID patients (Chapter 6). The cohort was defined based on two broad clinical phenotypes (Infections Only and Polyclonal Lymphoproliferation) plus IgM status (Chapel et al. 2008). The analysis conducted to compare predicted deleterious variants observed within the phenotypes was not as revealing as anticipated (Sections 6.2.11 and 6.2.12). Instead, the network analysis based on shared connections of variants provided greater insight into heterogeneity within the cohort (Figure 6.2). This analysis revealed three groups in the cohort and this more unbiased approach may be useful for furthering our understanding of the heterogeneity which cannot necessarily be identified from the clinical categories alone.

The analysis from all three cohorts suggests significant heterogeneity remains even when well defined phenotypes are used for patient recruitment. Evidence from each cohort suggests a role for genetics in the heterogeneity observed.

7.2 Variant prioritisation

Analysis at the genome-wide level for multifactorial diseases results in large numbers of genes or variants which must be further prioritised in order to

identify those that may play a role in the underlying genetic aetiology. For both the sepsis eQTL (Chapter 4) and the CVID WGS (Chapter 6) study, additional datasets (available both publicly and through collaborations) assisted with the prioritisation process.

For both cohorts, it was important to identify variants observed at higher frequencies in the patients compared to healthy individuals. Publicly available datasets cataloguing both common variation and regulatory variants allowed these comparisons to be made. Genes with evidence of a *cis* eQTL unique to the context of sepsis were identified by comparison to the large eQTL study conducted by Westra et al. (2013) for 5,311 healthy individuals. The 1000 Genomes Project (1000 Genomes Project Consortium et al. 2012), which provides a comprehensive resource of human genetic variants with a frequency of at least 1% was used to determine which of the variants observed in the CVID WGS data were rare. Variants were excluded if they had MAF >5% in the 1000 Genomes Project.

Many of the successful WGS analyses published to date have made use of unaffected family members to help prioritise potentially disease causing variants in the affected individuals (Guo et al. 2012; Martin et al. 2014). Family members were not available for the CVID cohort and due to the sporadic nature of the disease, high levels of shared variants or genes were not expected between the individuals. An alternative approach was therefore required to reduce the large number of variants identified across so many individuals. In addition to filters frequently used in other studies such as the 1000 Genomes Project frequency and predicted deleteriousness of the variant (Chou et al. 2012; Reumers et al. 2012) the project benefited from data available through the collaborative nature of the data generation. The genomes were sequenced as part of the WGS500 project which includes individuals with a range of diseases. The MAF for each variant observed across the WGS500 samples was calculated and allowed the

removal of variants detected at high frequencies. This removed variants which were likely a result of pipeline artefacts. Applying a threshold of 5% for this filter frequently reduced the variant list by approximately 50%, substantially reducing noise within the remaining variants.

7.3 Context specificity

Many genes have expression profiles that are dependent on tissue, developmental stage or physiological condition. Many studies have now shown that the number of shared eQTLs between different cell types and tissues varies widely. (Fairfax et al. 2012; GTEx Consortium 2013; Fairfax et al. 2014). This has highlighted the importance of cell type or tissue when conducting genetic experiments. However, determining the most representative cell type is not necessarily trivial as demonstrated by the RNA-seq data generated for the CVID cohort. CVID is due to primary antibody production failure and is generally thought to be the result of failure of B cell maturation or activation (Chapel et al. 2008). RNA-seq data for CD19+ B cell were generated for a small number of patients. A particularly striking result from the analysis was the enrichment of pathways involving T cell and dendritic cells, suggesting a small panel of cell types would have been more informative for this analysis (Figure 6.10). This also demonstrates the benefit of conducting pilot studies as they can highlight ways in which the study could be altered in order to generate maximum output from the data.

The effect of context on regulatory variants has not been so greatly studied, particularly in disease cohorts. Studies that do exist tend to be small due to the accessibility and cost of appropriate cohorts (Westra and Franke 2014). For infectious disease, a recent eQTL study conducted in 94 malaria-infected children and 61 age-matched controls only identified five peak SNP-gene associations in response to malaria infection but this could be due to the small

sample size. 1,508 (41%) of the genes with a *cis* eQTL in the sepsis cohort were not observed in a cohort of healthy individuals and therefore defined as specific to sepsis (Figure 4.6), highlighting the value of larger cohorts for the identification of context specific eQTL. The identification of such a high percentage of unique eQTL not only supports the idea of context being as important as cell type or tissue type but also highlights the potential advantage of using the same cohort of patients for integrative analysis with GWAS loci. If context is important for disease then comparing GWAS loci to publicly available datasets of healthy individuals may not provide the most informative comparison as the eQTL may not be present in that state. This approach of using the same individuals for both GWAS and eQTL analysis discovered SNPs associated with childhood asthma were also strongly associated with the transcript levels of the gene *ORMDL3* (Moffatt et al. 2007).

7.4 Integrative analysis

GWAS have identified a number shared genetic loci between different immune-related diseases including asthma, rheumatoid arthritis and IBD (Zhernakova et al. 2009). It was proposed that integrating analysis from the cohorts presented in the thesis would be informative in identifying the functional genes and pathways that modulate immune response to infection. This was the case for the sepsis cohort and the influenza challenge in which viral infection gene expression signatures were observed in both studies (Sections 3.2.11 and 5.2.6). The challenge study model was not successful in predicting viral infection in the sepsis cohort but the significant enrichment of differentially expressed between viral infection status in sepsis and moderate/severe LCI status in the challenge study (Figure 5.12) should help to develop a more robust signature suitable for different cohorts of individuals.

Variants within the same gene have also been shown to be associated with both rare and common disease. For example a rare variant in *IL17RA* gene results in Omenn syndrome, a combined immunodeficiency (Giliani et al. 2006). Common variants in the same gene are associated with common autoimmune diseases such as multiple sclerosis and type 1 diabetes. These common variants alter the splicing efficiency and result in a less severe phenotype than Omenn syndrome (Goris and Liston 2012). The most common infection risk for patients with CVID is from extracellular bacteria, particularly those that affect the respiratory tract such as *Streptococcus pneumoniae* (Cunningham-Rundles 2010). This Gram-positive bacteria is also the leading cause of sepsis due to CAP (File 2004). It was therefore hypothesised that there may be key genes involved in both the rare (CVID) and common (sepsis) response to bacterial infection.

The *DOK3* gene provides an example of the potential informativeness of integrative analysis between rare and common disease. This gene suppresses B cell activation by inhibiting the B-cell receptor activation of the JNK signal transduction pathway (Robson et al. 2004). Sepsis patients that belong to cluster 2 show a better outcome and distinct immune response to sepsis characterised by upregulation of genes involved in the activation and function of lymphocytes. *DOK3* is downregulated in sepsis patients that belong to cluster 2 (Section 3.2.12) but upregulated in the CVID patients (Figure 6.13).

7.5 Clinical translation of findings

Identifying genetic loci associated with disease will help generate new hypotheses for pathogenic mechanisms and potential therapeutic targets but translating these findings into the clinical setting remains one of the biggest challenges in this field of research.

In order for gene expression to be of any clinical use, assays must be developed which can offer rapid, consistent, interpretable and inexpensive

results. In particular, the vast quantities of data generated from genome-wide gene expression analysis must be condensed to a more manageable gene signature which can be measured rapidly within hours or minutes and does not require batching with other samples (Maslove and Wong 2014). A number of studies have reported success in the development of assays which meet some of these criteria suggesting information generated from this work may one day be of clinical use.

Wong et al. (2010b) have developed a 100 gene signature which can distinguish three subclasses of paediatric septic shock patients. This was of particular interest because one subgroup showed higher severity of illness and higher mortality rates than the other two and may therefore benefit from timely treatment. This signature was converted into a graphical expression mosaic allowing clinicians to classify the patients based on the patterns of expression.

An even more promising result comes from a study of trauma patients. Many of these patients develop organ injury which results in a complicated recovery that is difficult to predict based on physiological and anatomical prognosticators. Cuenca et al. (2013) developed a simple genomic score to predict clinical outcome. This was based on 63 genes which differed in expression between the uncomplicated and complicated clinical outcomes over 28 days. The RNA abundance was quantitated using nanoString technology (www.nanostring.com/lifesciences). This is an exciting new technology which allows the digital quantification of multiplexed RNA for up to 800 genes simultaneously, ideal for diagnostic use.

7.6 Future work

Sepsis prediction model validation

Recruitment of CAP sepsis patients for the GAINs study is ongoing and the cohort now consists of over 650 patients. The most striking result from

the analysis of the transcriptomic response to severe sepsis (Chapter 3) was the identification of two clusters within the cohort, based on gene expression signatures (Section 3.2.12). These clusters showed different survival rates and identified individuals with distinct immune responses to sepsis. In order to replicate this finding, genome-wide gene expression data are currently being generated for a further 106 CAP patients. This validation cohort has been enriched for non-survivors (49%) in order to increase power. Cluster membership will be assigned using the seven gene prediction model described in Section 3.2.16 and any difference in survival will be determined.

Sepsis microbiology

Microbiology data for the sepsis cohort were used to define groups of patients depending on whether Gram-positive, Gram-negative or viral infections were identified. No differentially expressed genes were identified between the patients with Gram-positive compared to Gram-negative within the cohort (Table 3.8). However, this may be confounded by the challenge of identifying the invading organism responsible for infection. Additional clinical data are available for these patients including the specific microorganism and method of detection but 49% of the cohort did not have a positive bacterial culture and similar numbers have been reported for other studies such as Phua et al. (2013) which showed 40%. Microbial cultures are commonly used for identification but the yield can be low for instance because of prior antibiotic therapy (Llewelyn et al. 2001; Tang et al. 2007). As the cost of WGS continues to fall, the application of this method to the detection of microorganisms is an appealing prospect. This will generate its own set of challenges, particularly with respect to the small size of the bacterial or viral genome compared to the host, but could be particularly informative for the identification of novel or unsuspected pathogens (Barzon et al. 2013).

In the shorter term, the gene expression signature that distinguishes viral infection status within the cohort (Section 3.2.11) would be interesting for further analysis. Within the validation cohort of 106 CAP patients, only three have confirmed viral infection. However, recruitment for this study is ongoing and therefore once this number increases, the model could be validated. Alternatively, small cohorts of adult CAP sepsis patients are beginning to become publicly available, for example the study published this year by Severino et al. (2014), and these may provide a suitable independent cohort for validation.

Sepsis cell specific expression signatures

The LeukoLOCK system for sample collection in the sepsis gene expression signature work provided an efficient method that did not require complicated pre-processing and therefore minimised potential expression artefacts that can occur during lengthy cell isolation procedures. Additionally, the majority of hospital sites do not have the facilities required for cell separation. However, now that a substantial cohort has been recruited and this initial analysis is complete, specific hypotheses can be generated and investigated in a small subset of individuals. The recruitment centres have been reduced to include only those with access to facilities for cell separation. This will allow gene expression signatures to be explored in specific cell fractions.

Sepsis context specific eQTL

Determining eQTLs which are specific to the context of sepsis is of particular interest. Following the initial comparison with the Netherlands healthy cohort, agreement has been secured for access to this data. Gene expression QC and genotype imputation will be conducted for each cohort independently but the eQTL mapping will be restricted to probes and SNPs available for both cohorts. This will allow a more comprehensive comparison between eQTLs present in

the context of sepsis and those observed in healthy volunteers. In addition, the analysis will be extended to include the identification of secondary peaks of association or associations with opposing directions of effect.

The leukocyte fraction of whole blood is composed of a number of different cells types. If a particular cell type only forms a small proportion of the total count, cell-specific eQTLs may not be detected. Following the analysis of eQTLs in the leukocyte fraction, this dataset could be extended by identifying cell type specific eQTL in a smaller cohort of patients. Methods are also available for the bioinformatic deconvolution of subtype specific signals. Westra et al. (2014) used a gene expression signature to predict the percentage of neutrophils in whole blood. eQTL analysis was performed with a term for interaction between the proxy cell count and SNP to identify eQTL which showed a stronger effect size in neutrophils compared to lymphoid cells. Using this approach may help identify the cell types most important in the context of sepsis.

CVID integration of WGS with RNA-seq

The variants determined from combining the CVID WGS with RNA-seq data clearly demonstrate the advantages of gene expression data when investigating non-coding variants. The RNA-seq data represents a small subset of the cohort and increasing this number should yield new insights into the underlying pathophysiology of the disease. In particular, increasing the numbers so that comparisons can be made between subsets of individuals, for instance Infections Only with IgM absent or present may help identify the underlying genetic causes of the clinical differences observed between these patients. Many of the patients presented in this cohort regularly return to clinic, which will allow the isolation of RNA from fresh cells. The results of both the WGS and RNA-seq analysis also highlight the potential role T cells and dendritic cells play in the pathogenesis of disease and therefore a panel of cell types may be more informative.

CVID clinical subphenotype

The CVID cohort were split into two broad categories (Chapel et al. 2008) however extensive further clinical data are available for all individuals. Combining this data particularly with the subgroups identified from the network analysis may provide further understanding of the subgroups observed.

7.7 Concluding remarks

The overall aim of this thesis was to understand the genetic determinants of variation in response to infection by adopting a functional genomics approach. Three clinical cohorts were investigated: patients with severe sepsis, an influenza challenge study and patients with CVID. The results from each of these individual cohorts will impact the respective fields of research. For example, the presented transcriptomic response to severe sepsis represents the largest known adult severe sepsis CAP cohort. In addition, integrating results from the different cohorts has identified shared variants and genes which may contribute to the observed variation in response to infection for example the differential expression of genes in response to common bacterial infection (sepsis) and rare (CVID). This should further out understanding of the mechanisms underlying variation in response to infection.

Appendices

A	Functional genomic response in severe sepsis	224
B	Regulatory genetic variants in severe sepsis	240
C	Functional genomic response to influenza challenge	250
D	Functional genomic response in CVID	259
	Bibliography	291

List of Figures

A.1	Kaplan-Meier plots for three and four clusters	236
B.1	<i>Cis</i> eQTL following adjustment for clinical covariates	241
B.2	Inclusion of PCs and clinical covariates	242
B.3	Correlations between PCs and clinical covariates	243

List of Tables

A.1	Clinical covariates measured for severe sepsis cohort	225
A.2	Differential expression for 14 day non-survivors vs 6 month survivors	226
A.3	Differential expression for 28 day non-survivors vs 6 month survivors	228
A.4	Differential expression for 6 month non-survivors vs 6 month survivors	230
A.5	IPA 14 day non-survivors vs 6 month survivors	232
A.6	IPA 28 day non-survivors vs 6 month survivors	233
A.7	IPA 6 month non-survivors vs 6 month survivors	233
A.8	Differential expression for viral vs no viral infection	234
A.9	Differential expression for cluster 2 vs cluster 1	237
A.10	IPA cluster 2 vs 1	238

B.1	Genotyping QC sample removal.	240
B.2	<i>Cis</i> eQTL pathway enrichment	243
B.3	<i>Cis</i> eQTL unique to sepsis pathway enrichment	245
B.4	Master regulators of <i>trans</i> associations	246
C.1	Differential expression for LCI	251
C.2	Differential expression for moderate/severe LCI	253
C.3	IPA moderate/severe LCI 48 hours post-challenge	255
C.4	Differential expression for vaccination status pre-challenge	258
C.5	Influenza and sepsis overlap	258
D.1	Previously described genes for PIDs	260
D.2	CVID GWAS genes	262
D.3	WGS coverage per sample	263
D.4	WGS variant pathway enrichment	263
D.5	Differential expression RNA-seq CVID	267
D.6	RNA-seq pathway enrichment	268

Appendix A

Functional genomic response in severe sepsis

Table A.1: Clinical covariates measured for severe sepsis cohort.

Continuous clinical covariates	Count clinical covariates
Age	Sex
Polynucleocytes (raw)	Activated protein C
Mononucleocytes (raw)	Corticosteroid
Lymphocytes (raw)	Microbiology summary
Polynucleocytes (proportion)	Viral infection
Mononucleocytes (proportion)	H1N1 infection
Lymphocytes (proportion)	Ventilation
White cell count	Renal failure
APACHE II	Renal replacement therapy
SOFA (sampling day)	Inotropes
SOFA (maximum recorded)	14 day survival
Bicarbonate	28 day survival
Bilirubin	6 month survival
FIO ₂	RNA sampling day
PCO ₂	
PAO ₂	
Haematocrit	
Temperature (low)	
Temperature (high)	
Heart rate (low)	
Heart rate (high)	
Mean arterial pressure (low)	
Mean arterial pressure (high)	
Systolic blood pressure (low)	
Systolic blood pressure (high)	
Sodium (low)	
Sodium (high)	
Potassium (low)	
Potassium (high)	
pH	
Platelets	
Respiratory rate	
Inotropic support (number of days)	
Renal support (number of days)	
Creatinine (high)	
Creatinine (low)	
Days in ICU	
Days to death	

Table A.2: Differential expression for 14 day non-survivors vs 6 month survivors. The 50 most significant differentially expressed probes.

Probe	Gene	log(FC)	FDR
3890048	<i>INO80</i>	0.34	5.23E-04
5130113	<i>TMEM141</i>	0.45	5.23E-04
520332	<i>B4GALT7</i>	0.54	6.18E-04
3170538	<i>INO80</i>	0.37	6.18E-04
510100	<i>CBY1</i>	0.50	8.07E-04
3800110	<i>ORC5L</i>	0.60	1.10E-03
3830372	<i>PAXIP1</i>	0.51	1.10E-03
5550673	<i>SIVA</i>	0.62	1.21E-03
450475	<i>APEH</i>	0.28	1.44E-03
940010	<i>CHCHD10</i>	0.53	1.86E-03
5050082	<i>PUF60</i>	0.34	1.86E-03
6660195	<i>RPUSD3</i>	0.43	1.86E-03
670671	<i>TMEM134</i>	0.43	1.86E-03
5690711	<i>TMEM134</i>	0.46	1.86E-03
1580215	<i>FBXL6</i>	0.49	1.86E-03
1010253	<i>NT5C2</i>	-0.28	1.86E-03
20192	<i>DHX35</i>	0.38	1.86E-03
1980632	<i>DNAJA1</i>	-0.34	1.86E-03
1340193	<i>DHX30</i>	0.47	1.86E-03
6400138	<i>PRMT1</i>	0.57	1.86E-03
6900278	<i>FYCO1</i>	0.34	1.89E-03
1990520	<i>GPS1</i>	0.33	1.89E-03
1240521	<i>PSMG4</i>	0.47	1.89E-03
5910338	<i>TMEM203</i>	0.30	1.89E-03
10196	<i>C21orf33</i>	0.60	1.92E-03
2350576	<i>DHDDS</i>	0.45	1.94E-03
5310050	<i>CCDC71</i>	0.37	1.94E-03
3290170	<i>PAAF1</i>	0.77	1.94E-03
4830328	<i>CYB561D2</i>	0.29	1.94E-03
50553	<i>SIGMAR1</i>	0.49	1.94E-03
1850500	<i>C3orf75</i>	0.37	1.94E-03
3450253	<i>ZNHIT3</i>	0.32	1.94E-03
110347	<i>PDK3</i>	-0.38	1.94E-03
4610048	<i>AAMP</i>	0.37	1.94E-03
6060201	<i>DDX18</i>	0.49	1.98E-03
2230403	<i>MTIF2</i>	0.35	1.98E-03
1940228	<i>FAM58A</i>	0.42	2.27E-03
520521	<i>MRPL20</i>	0.66	2.27E-03
1010347	<i>POLR2I</i>	0.36	2.27E-03
4890382	<i>ILVBL</i>	0.73	2.27E-03
7150382	<i>NR2C2</i>	-0.31	2.36E-03

Continued on next page

Table A.2 – *Continued from previous page*

Probe	Gene	log(FC)	FDR
5130414	<i>C9orf23</i>	0.31	2.36E-03
5870576	<i>C21orf33</i>	0.54	2.36E-03
3450364	<i>THOC6</i>	0.70	2.50E-03
1470561	<i>FAM173A</i>	0.42	2.50E-03
6900025	<i>POLR2D</i>	0.49	2.50E-03
1340739	<i>FAM160B1</i>	-0.35	2.51E-03
2190221	<i>EIF3L</i>	0.41	2.51E-03
3710039	<i>MRPL54</i>	0.43	2.51E-03
830341	<i>TBRG4</i>	0.71	2.58E-03

Table A.3: Differential expression for 28 day non-survivors vs 6 month survivors. The 50 most significant differentially expressed probes.

Probe	Gene	log(FC)	FDR
6480064	<i>POLR1E</i>	0.51	1.43E-03
3890048	<i>INO80</i>	0.28	2.37E-03
5130113	<i>TMEM141</i>	0.38	2.37E-03
3830372	<i>PAXIP1</i>	0.45	3.16E-03
3170538	<i>INO80</i>	0.30	3.16E-03
20192	<i>DHX35</i>	0.34	3.16E-03
7150382	<i>NR2C2</i>	-0.27	6.08E-03
5690750	<i>GALT</i>	0.38	6.57E-03
1240521	<i>PSMG4</i>	0.41	7.06E-03
1340739	<i>FAM160B1</i>	-0.31	8.42E-03
3800110	<i>ORC5L</i>	0.49	8.79E-03
5220672	<i>KIAA1600</i>	-0.30	1.04E-02
510100	<i>CBY1</i>	0.38	1.04E-02
6660195	<i>RPUSD3</i>	0.34	1.04E-02
3450253	<i>ZNHIT3</i>	0.27	1.04E-02
5550673	<i>SIVA</i>	0.47	1.04E-02
6650561	<i>ZNHIT3</i>	0.30	1.04E-02
1980632	<i>DNAJA1</i>	-0.27	1.04E-02
5310050	<i>CCDC71</i>	0.30	1.06E-02
110347	<i>PKD3</i>	-0.31	1.06E-02
1340039	<i>TFPI</i>	-0.54	1.06E-02
670671	<i>TMEM134</i>	0.33	1.06E-02
5910095	<i>CCDC115</i>	0.27	1.06E-02
5130414	<i>C9orf23</i>	0.26	1.06E-02
6060201	<i>DDX18</i>	0.39	1.29E-02
6350189	<i>MGC4677</i>	-0.32	1.29E-02
5050082	<i>PUF60</i>	0.26	1.29E-02
6900278	<i>FYCO1</i>	0.26	1.29E-02
1010347	<i>POLR2I</i>	0.29	1.29E-02
5690711	<i>TMEM134</i>	0.35	1.29E-02
6650452	<i>DET1</i>	0.31	1.29E-02
5420400	<i>FLJ39827</i>	0.38	1.29E-02
2000228	<i>FEM1C</i>	-0.30	1.29E-02
2450132	<i>THAP7</i>	0.33	1.29E-02
4850731	<i>TFPI</i>	-0.51	1.29E-02
5860605	<i>PDCD6IP</i>	0.28	1.29E-02
6280458	<i>BCL6</i>	-0.31	1.36E-02
520332	<i>B4GALT7</i>	0.37	1.45E-02
1450156	<i>MAGOH</i>	0.26	1.45E-02
10196	<i>C21orf33</i>	0.46	1.48E-02
2070730	<i>NGDN</i>	0.32	1.52E-02

Continued on next page

Table A.3 – *Continued from previous page*

Probe	Gene	log(FC)	FDR
50553	<i>SIGMAR1</i>	0.37	1.52E-02
830341	<i>TBRG4</i>	0.57	1.52E-02
5820170	<i>SIVA1</i>	0.35	1.52E-02
1660170	<i>HNRNPA0</i>	0.28	1.54E-02
1340193	<i>DHX30</i>	0.35	1.69E-02
6290681	<i>TOMM20</i>	0.30	1.69E-02
430369	<i>TEX264</i>	0.34	1.73E-02
1940228	<i>FAM58A</i>	0.32	1.73E-02
2230403	<i>MTIF2</i>	0.27	1.73E-02

Table A.4: Differential expression for 6 month non-survivors vs 6 month survivors.
The 50 most significant differentially expressed probes.

Probe	Gene	log(FC)	FDR
6480064	<i>POLR1E</i>	0.45	2.16E-03
3830372	<i>PAXIP1</i>	0.41	2.16E-03
5910095	<i>CCDC115</i>	0.26	9.01E-03
2070730	<i>NGDN</i>	0.33	9.01E-03
5860605	<i>PDCD6IP</i>	0.29	9.01E-03
6450594	<i>CD79B</i>	0.53	1.44E-02
6650561	<i>ZNHIT3</i>	0.28	1.44E-02
4570576	<i>Hs.558212</i>	0.30	1.44E-02
2140687	<i>ZNF45</i>	0.30	1.44E-02
4150608	<i>LOC644619</i>	0.29	1.44E-02
1240521	<i>PSMG4</i>	0.35	1.44E-02
5130113	<i>TMEM141</i>	0.29	1.54E-02
4860468	<i>ARL1</i>	0.44	1.70E-02
5690750	<i>GALT</i>	0.31	1.83E-02
6350189	<i>MGC4677</i>	-0.29	1.92E-02
1340039	<i>TFPI</i>	-0.46	1.93E-02
3800110	<i>ORC5L</i>	0.40	2.16E-02
4850731	<i>TFPI</i>	-0.45	2.16E-02
6760133	<i>PRMT3</i>	0.29	2.28E-02
50553	<i>SIGMAR1</i>	0.33	2.58E-02
160639	<i>EBAG9</i>	0.33	2.80E-02
7150296	<i>LOC100130562</i>	0.62	2.80E-02
7160408	<i>Hs.269011</i>	-0.30	2.80E-02
830341	<i>TBRG4</i>	0.51	2.80E-02
10196	<i>C21orf33</i>	0.40	2.80E-02
6660195	<i>RPUSD3</i>	0.27	2.80E-02
670671	<i>TMEM134</i>	0.27	3.08E-02
7040220	<i>RDH14</i>	0.28	3.08E-02
6060201	<i>DDX18</i>	0.33	3.08E-02
2230538	<i>LRRN3</i>	0.55	3.24E-02
1440343	<i>TMEM209</i>	0.38	3.31E-02
1030239	<i>SCRN1</i>	0.49	3.35E-02
2140382	<i>PASK</i>	0.44	3.35E-02
3940053	<i>ZC3H14</i>	0.35	3.35E-02
1820703	<i>PIGP</i>	0.54	3.35E-02
7380398	<i>DAK</i>	0.31	3.45E-02
3610333	<i>LOC441013</i>	0.31	3.47E-02
5690711	<i>TMEM134</i>	0.28	3.71E-02
4540463	<i>C1orf156</i>	0.26	3.71E-02
2360156	<i>RABL2B</i>	0.28	3.87E-02
2810626	<i>LOC648294</i>	0.30	3.87E-02

Continued on next page

Table A.4 – *Continued from previous page*

Probe	Gene	log(FC)	FDR
5390121	<i>GDPD3</i>	-0.33	3.87E-02
160091	<i>LOC399804</i>	0.43	3.87E-02
6400524	<i>LOC390705</i>	-0.34	3.87E-02
6270475	<i>DUS3L</i>	0.45	3.87E-02
1710639	<i>RBM4B</i>	0.27	3.87E-02
5420400	<i>FLJ39827</i>	0.31	3.87E-02
7050553	<i>Hs.79881</i>	-0.34	3.87E-02
2510463	<i>METTL13</i>	0.29	3.87E-02
6520100	<i>Hs.478682</i>	0.26	3.87E-02

Table A.5: IPA 14 day non-survivors vs 6 month survivors. Pathway enrichment for differentially expressed probes between 14 day non-survivors and 6 month survivors.

Ingenuity Canonical Pathways	p value
EIF2 Signalling	3.98E-21
Regulation of eIF4 and p70S6K Signalling	9.12E-08
mTOR Signalling	1.51E-05
Methionine Degradation I (to Homocysteine)	3.09E-05
B Cell Development	5.89E-05
Cysteine Biosynthesis III (mammalia)	7.59E-05
Superpathway of Methionine Degradation	1.02E-04
Altered T Cell and B Cell Signalling in Rheumatoid Arthritis	3.16E-04
Antigen Presentation Pathway	3.55E-04
Purine Nucleotides De Novo Biosynthesis II	3.63E-04
Fatty Acid β -oxidation I	3.72E-04
Ethanol Degradation II	4.68E-04
Allograft Rejection Signalling	6.76E-04
OX40 Signalling Pathway	8.13E-04
Acetate Conversion to Acetyl-CoA	2.40E-03
tRNA Charging	2.51E-03
Dolichol and Dolichyl Phosphate Biosynthesis	4.17E-03
Cardiolipin Biosynthesis II	4.17E-03
Purine Nucleotides Degradation II (Aerobic)	4.68E-03
Graft-versus-Host Disease Signalling	1.12E-02
5-aminoimidazole Ribonucleotide Biosynthesis I	1.20E-02
D-glucuronate Degradation I	1.20E-02
Inosine-5'-phosphate Biosynthesis II	1.20E-02
Adenosine Nucleotides Degradation II	1.32E-02
Oxidative Ethanol Degradation III	1.32E-02
Oxidative Phosphorylation	1.48E-02
Mitochondrial Dysfunction	1.55E-02
Leucine Degradation I	1.66E-02
Calcium-induced T Lymphocyte Apoptosis	1.95E-02
Autoimmune Thyroid Disease Signalling	2.00E-02
Mitochondrial L-carnitine Shuttle Pathway	2.09E-02
Ethanol Degradation IV	2.09E-02
Dolichyl-diphosphooligosaccharide Biosynthesis	2.29E-02
Spermine and Spermidine Degradation I	2.29E-02
Valine Degradation I	2.51E-02
Nur77 Signalling in T Lymphocytes	2.69E-02
Primary Immunodeficiency Signalling	2.95E-02
Assembly of RNA Polymerase II Complex	3.98E-02
NAD Salvage Pathway II	4.27E-02
PKC θ Signalling in T Lymphocytes	4.47E-02
Fatty Acid Activation	4.68E-02

Continued on next page

Functional genomic response in severe sepsis

Table A.5 – Continued from previous page

Ingenuity Canonical Pathways	p value
Urate Biosynthesis/Inosine 5'-phosphate Degradation	4.68E-02
TCA Cycle II (Eukaryotic)	4.90E-02

Table A.6: IPA 28 day non-survivors vs 6 month survivors. Pathway enrichment for differentially expressed probes between 28 day non-survivors and 6 month survivors.

Ingenuity Canonical Pathways	p value
EIF2 Signalling	2.19E-05
Methionine Degradation I (to Homocysteine)	1.86E-04
Cysteine Biosynthesis III (mammalia)	3.02E-04
Dolichol and Dolichyl Phosphate Biosynthesis	3.55E-04
Superpathway of Methionine Degradation	2.57E-03
Regulation of eIF4 and p70S6K Signalling	5.50E-03
FcγRIIB Signalling in B Lymphocytes	7.24E-03
B Cell Receptor Signalling	9.12E-03
G Protein Signalling Mediated by Tubby	2.00E-02
mTOR Signalling	2.19E-02
Colanic Acid Building Blocks Biosynthesis	2.75E-02
Antigen Presentation Pathway	2.95E-02
Extrinsic Prothrombin Activation Pathway	3.55E-02
S-methyl-5-thio- α -D-ribose 1-phosphate Degradation	3.72E-02
GDP-L-fucose Biosynthesis II (from L-fucose)	3.72E-02
GDP-L-fucose Biosynthesis I (from GDP-D-mannose)	3.72E-02
RAR Activation	4.47E-02

Table A.7: IPA 6 month non-survivors vs 6 month survivors. Pathway enrichment for differentially expressed probes between 6 month non-survivors and 6 month survivors.

Ingenuity Canonical Pathways	p value
Lipoate Salvage and Modification	6.17E-03
Adenine and Adenosine Salvage VI	6.17E-03
B Cell Development	8.71E-03
Cell Cycle Control of Chromosomal Replication	1.20E-02
Lipoate Biosynthesis and Incorporation II	1.23E-02
EIF2 Signalling	2.14E-02
mTOR Signalling	2.57E-02
FcγRIIB Signalling in B Lymphocytes	2.63E-02
Galactose Degradation I (Leloir Pathway)	3.02E-02
Cardiac β -adrenergic Signalling	4.79E-02

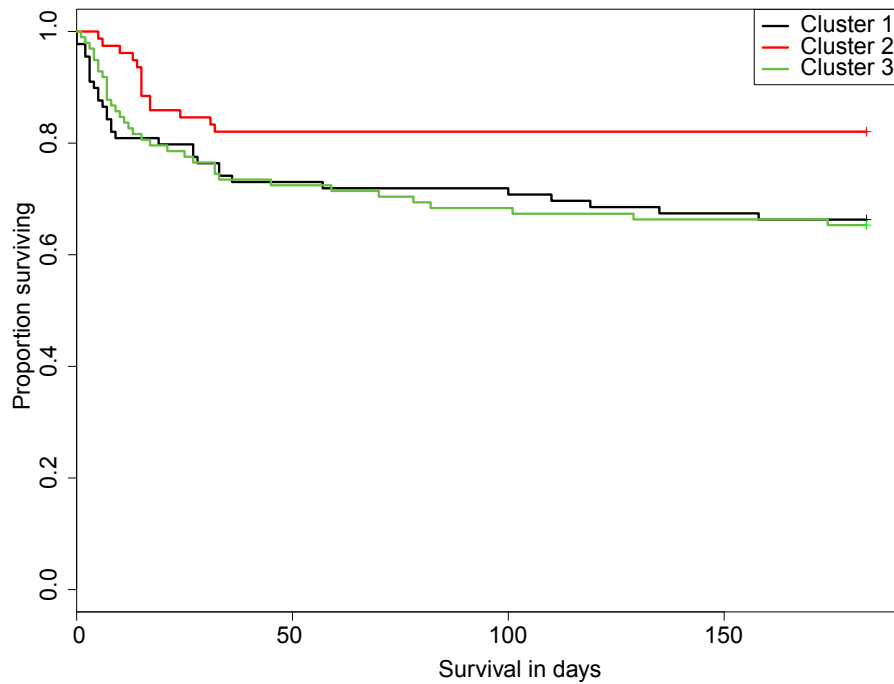
Table A.8: Differential expression for viral vs no viral infection. The 50 most significant differentially expressed probes.

Probe	Gene	log(FC)	FDR
3990170	<i>IFI27</i>	3.47	7.14E-05
7610440	<i>XAF1</i>	1.59	1.36E-02
7380221	<i>SPATS2L</i>	1.20	1.36E-02
4570164	<i>LOC389386</i>	0.86	1.36E-02
6480491	<i>EPB41L5</i>	-0.44	1.36E-02
4880370	<i>JUP</i>	1.65	1.77E-02
4290730	<i>LGALS3BP</i>	1.11	1.96E-02
70458	<i>BTN2A2</i>	0.74	2.02E-02
2030309	<i>SERPING1</i>	1.92	2.58E-02
2480450	<i>GALM</i>	0.93	2.58E-02
6580484	<i>MOV10</i>	0.88	2.58E-02
5700725	<i>EPSTI1</i>	1.89	2.94E-02
1340020	<i>TGIF2</i>	0.86	3.08E-02
4890270	<i>LY6E</i>	1.49	3.26E-02
3990010	<i>Hs.125087</i>	1.23	3.26E-02
7040035	<i>OAS1</i>	1.20	3.26E-02
6760286	<i>Hs.72010</i>	0.82	3.26E-02
3140646	<i>CD2AP</i>	0.76	3.26E-02
2690619	<i>SNORA45</i>	0.49	3.26E-02
70601	<i>CUL1</i>	0.34	3.26E-02
2030328	<i>UBXD7</i>	-0.47	3.26E-02
3870338	<i>IFI44L</i>	2.39	3.26E-02
2570300	<i>IFI44</i>	1.49	3.26E-02
4730059	<i>BATF2</i>	1.48	3.26E-02
450189	<i>XAF1</i>	1.44	3.26E-02
240722	<i>OAS2</i>	1.39	3.26E-02
3310376	<i>TIMM10</i>	1.35	3.26E-02
7320561	<i>OAS2</i>	1.29	3.26E-02
2340072	<i>PARP12</i>	0.91	3.26E-02
6860482	<i>HERC6</i>	0.83	3.26E-02
7510537	<i>SCO2</i>	0.79	3.26E-02
3140041	<i>SP140</i>	0.74	3.26E-02
2190397	<i>ZCCHC2</i>	0.68	3.26E-02
830403	<i>PSME2</i>	0.58	3.26E-02
70541	<i>REC8</i>	0.52	3.26E-02
3930377	<i>TRIM38</i>	0.41	3.26E-02
4560754	<i>GRIPAP1</i>	0.40	3.26E-02
6400025	<i>C16orf61</i>	0.39	3.26E-02
5310605	<i>CCNY</i>	-0.33	3.26E-02
3140382	<i>SCAP</i>	-0.34	3.26E-02
4860379	<i>C5orf21</i>	-0.43	3.26E-02

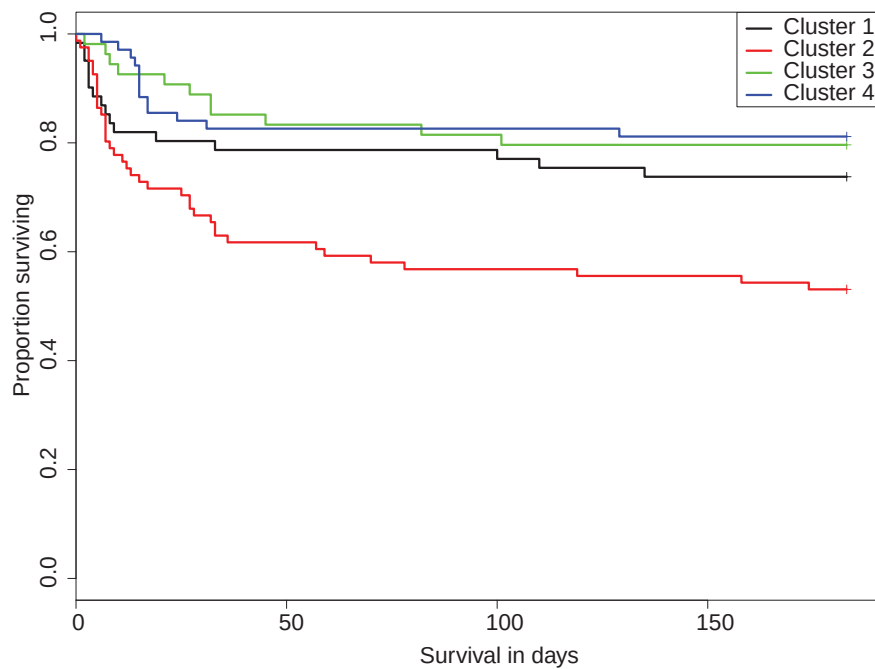
Continued on next page

Table A.8 – *Continued from previous page*

Probe	Gene	log(FC)	FDR
4230543	<i>LOC644532</i>	-0.52	3.26E-02
990768	<i>OAS3</i>	1.64	3.27E-02
1090390	<i>OAS1</i>	1.44	3.27E-02
4120037	<i>CBL</i>	-0.46	3.41E-02
3290292	<i>LAP3</i>	0.81	3.54E-02
4390360	<i>PARP6</i>	0.36	3.54E-02
2230204	<i>OAS2</i>	1.37	3.99E-02
580136	<i>LOC652694</i>	1.28	3.99E-02
3060474	<i>SLC23A2</i>	0.46	3.99E-02



(a)



(b)

Figure A.1: Kaplan-Meier plots for three and four clusters. Coefficient of variation was calculated to identify the 10% of probes which were most variable in the dataset (2,619). Cluster membership was defined using FactoMineR (Husson et al. 2013). a) Kaplan-Meier plots (coloured by cluster) are significantly different when three clusters are defined in the dataset (p value = 0.034). b) Kaplan-Meier plots (coloured by cluster) are significantly different when four clusters are defined in the dataset (p value = 0.0002).

Table A.9: Differential expression for cluster 2 vs cluster 1. The 50 most significant differentially expressed probes.

Probe	Gene	log(FC)	FDR
2680639	<i>SLC25A38</i>	1.77	4.83E-50
150681	<i>ZNHIT6</i>	1.67	9.57E-50
3170703	<i>LY9</i>	1.99	6.32E-47
540382	<i>RSAD1</i>	1.61	6.32E-47
2190537	<i>WDR74</i>	1.69	1.16E-46
4890092	<i>TSEN54</i>	1.69	1.87E-46
1430360	<i>ABHD14A</i>	1.54	1.87E-46
290343	<i>TTC3</i>	1.03	1.22E-45
1190717	<i>BMS1</i>	1.17	2.68E-45
4220364	<i>C19orf2</i>	1.07	2.68E-45
1470184	<i>CLSTN1</i>	1.19	6.06E-45
1190246	<i>FBXL10</i>	1.32	1.56E-44
3930553	<i>DDX10</i>	1.42	1.56E-44
2940343	<i>XPO4</i>	1.04	3.18E-44
3610647	<i>URG4</i>	1.40	3.53E-44
6590228	<i>CD6</i>	2.00	5.61E-44
4590390	<i>FAM38A</i>	1.37	1.47E-43
3850767	<i>SBK1</i>	1.93	3.31E-43
3930541	<i>PLEKHA1</i>	1.75	3.71E-43
1110180	<i>RORA</i>	1.85	4.80E-43
3780762	<i>FAM44B</i>	1.33	5.03E-43
3290044	<i>TTC27</i>	1.58	5.87E-43
5820333	<i>RPUSD2</i>	1.52	5.96E-43
4220053	<i>SKAP1</i>	1.92	6.54E-43
5390064	<i>ANKS1A</i>	-1.16	8.23E-43
3130358	<i>Hs.25892</i>	1.10	1.24E-42
10594	<i>PECI</i>	1.70	2.09E-42
4280239	<i>PFKFB2</i>	-1.61	2.51E-42
5050347	<i>CD5</i>	2.13	2.51E-42
3190112	<i>SERPINB1</i>	-1.16	2.83E-42
7040181	<i>PFKFB2</i>	-2.60	2.83E-42
4040187	<i>C6orf190</i>	2.10	5.19E-42
670528	<i>GYG1</i>	-1.48	5.19E-42
1010653	<i>POLR1C</i>	1.09	5.19E-42
6400603	<i>PVRIG</i>	2.04	5.32E-42
6180133	<i>ITPR3</i>	1.44	6.91E-42
4390382	<i>KIAA0182</i>	1.06	7.18E-42
6550243	<i>ESYT1</i>	1.08	7.18E-42
1450075	<i>ILF3</i>	1.04	7.57E-42
2760239	<i>RASGRP1</i>	1.77	7.73E-42
5050500	<i>ZBTB9</i>	1.40	8.09E-42

Continued on next page

Table A.9 – Continued from previous page

Probe	Gene	log(FC)	FDR
5570309	THOC1	1.38	9.56E-42
7160100	PPRC1	1.43	1.62E-41
620692	SLC25A42	1.42	1.97E-41
1430711	KLF12	1.92	2.16E-41
620543	SLC4A7	1.23	2.35E-41
6200736	HRB	-1.21	2.63E-41
7380327	FBXO31	1.39	3.35E-41
620619	TBC1D8	-1.48	3.35E-41
7000270	PYHIN1	2.23	4.00E-41

Table A.10: IPA cluster 2 vs 1. Pathway enrichment for differentially expressed probes between cluster 2 and cluster 1.

Ingenuity Canonical Pathways	p value
Calcium-induced T Lymphocyte Apoptosis	2.82E-09
iCOS-iCOSL Signalling in T Helper Cells	3.63E-09
OX40 Signalling Pathway	6.17E-09
Allograft Rejection Signalling	3.89E-08
T Cell Receptor Signalling	6.17E-08
CD28 Signalling in T Helper Cells	1.12E-07
Role of NFAT in Regulation of the Immune Response	2.82E-07
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells	6.46E-07
Primary Immunodeficiency Signalling	2.04E-06
Nur77 Signalling in T Lymphocytes	2.34E-06
B Cell Development	2.63E-06
Type I Diabetes Mellitus Signalling	4.68E-06
Autoimmune Thyroid Disease Signalling	7.94E-06
PKC Signalling in T Lymphocytes	1.12E-05
CTLA4 Signalling in Cytotoxic T Lymphocytes	3.55E-05
Crosstalk between Dendritic Cells and Natural Killer Cells	8.51E-05
Antigen Presentation Pathway	1.00E-04
Graft-versus-Host Disease Signalling	1.82E-04
Hematopoiesis from Pluripotent Stem Cells	2.00E-04
Cdc42 Signalling	2.04E-04
Altered T Cell and B Cell Signalling in Rheumatoid Arthritis	3.24E-04
Communication between Innate and Adaptive Immune Cells	1.02E-03
CCR5 Signalling in Macrophages	1.10E-03
Natural Killer Cell Signalling	2.09E-03
Dendritic Cell Maturation	3.55E-03
Airway Pathology in Chronic Obstructive Pulmonary Disease	3.55E-03
Granzyme B Signalling	3.80E-03

Continued on next page

Table A.10 – *Continued from previous page*

Ingenuity Canonical Pathways	p value
T Helper Cell Differentiation	4.27E-03
Phospholipase C Signalling	5.62E-03
TREM1 Signalling	9.77E-03
Regulation of IL-2 Expression in Activated and Anergic T Lymphocytes	1.45E-02
Tumoricidal Function of Hepatic Natural Killer Cells	1.70E-02
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	2.00E-02
tRNA Charging	2.09E-02
p38 MAPK Signalling	2.19E-02
Adenine and Adenosine Salvage III	2.40E-02
Purine Ribonucleosides Degradation to Ribose-1-phosphate	2.40E-02
Factors Promoting Cardiogenesis in Vertebrates	3.02E-02
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	3.31E-02
Ethanol Degradation II	3.55E-02
Role of Tissue Factor in Cancer	3.63E-02
Induction of Apoptosis by HIV1	3.72E-02
Citrulline Biosynthesis	4.17E-02
Thiamin Salvage III	4.17E-02
Histamine Biosynthesis	4.17E-02
Noradrenaline and Adrenaline Degradation	4.37E-02
Sperm Motility	4.68E-02

Appendix B

Regulatory genetic variants in severe sepsis

Table B.1: Genotyping QC sample removal. Following QC, 20 samples were identified for removal from further analysis. The reasons for removal are listed for each samples. MDS = multi-dimensional scaling.

Patient	QC removal reason
UK02450065	Sex check and MDS
GAUKJR009000	Sex check
GAUKRV027000	Sex check
UK48290012	Sex check
UK58650004	Sex check
GAUKRV020000	Relatedness
GAUKRV021000	Relatedness
GAUKJR007000	Missingness
UK01440089	MDS
UK02120166	MDS
UK02740182	MDS
UK02880195	MDS
UK08070088	MDS
UK09200089	MDS
UK17410207	MDS
UK29150050	MDS
UK30430058	MDS
UK45110077	MDS
UK52830004	MDS
UK54760010	MDS

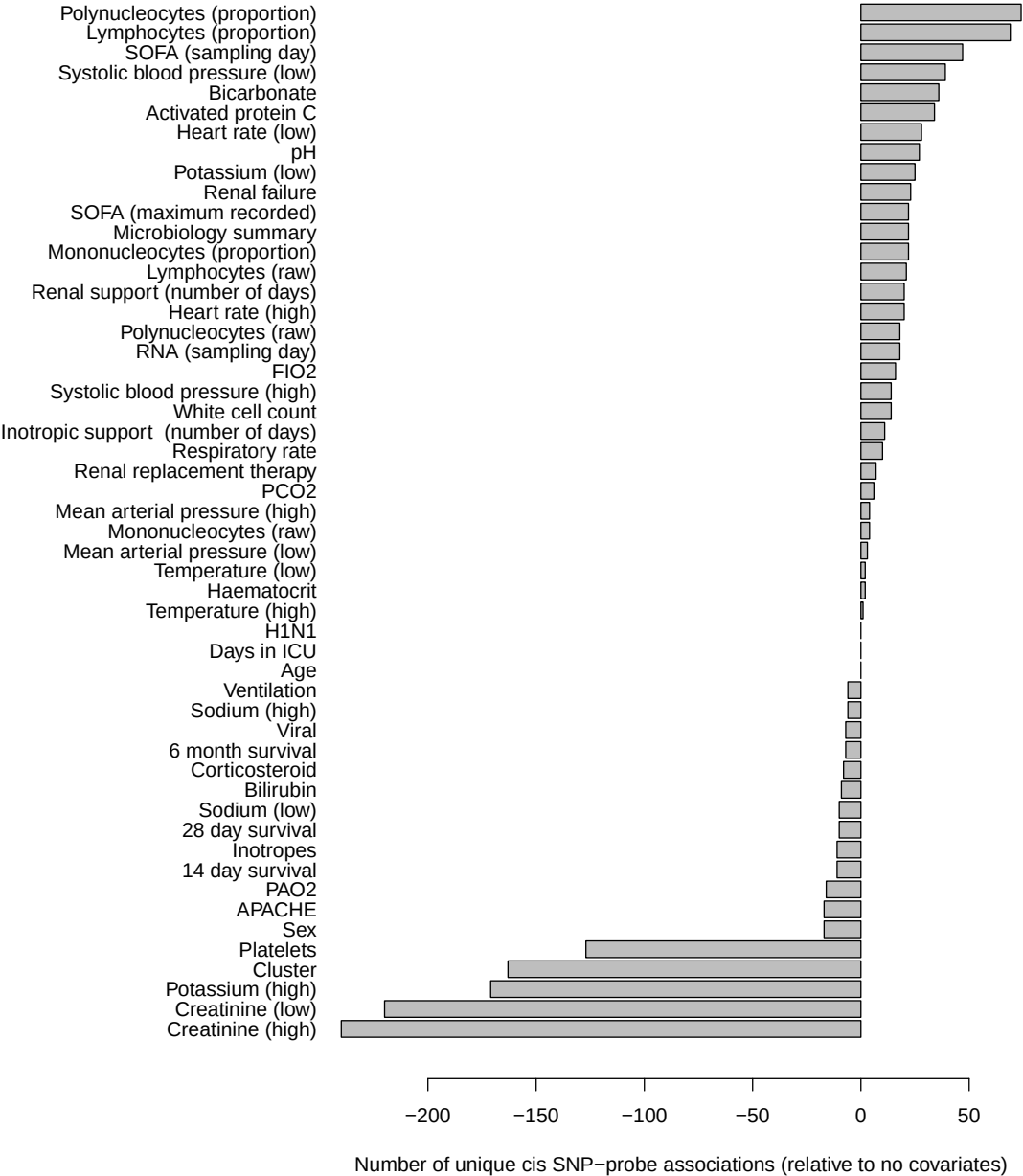


Figure B.1: *Cis* eQTL following adjustment for clinical covariates. The number of unique cis SNP-probe associations from eQTL mapping taking account of each clinical covariate independently.

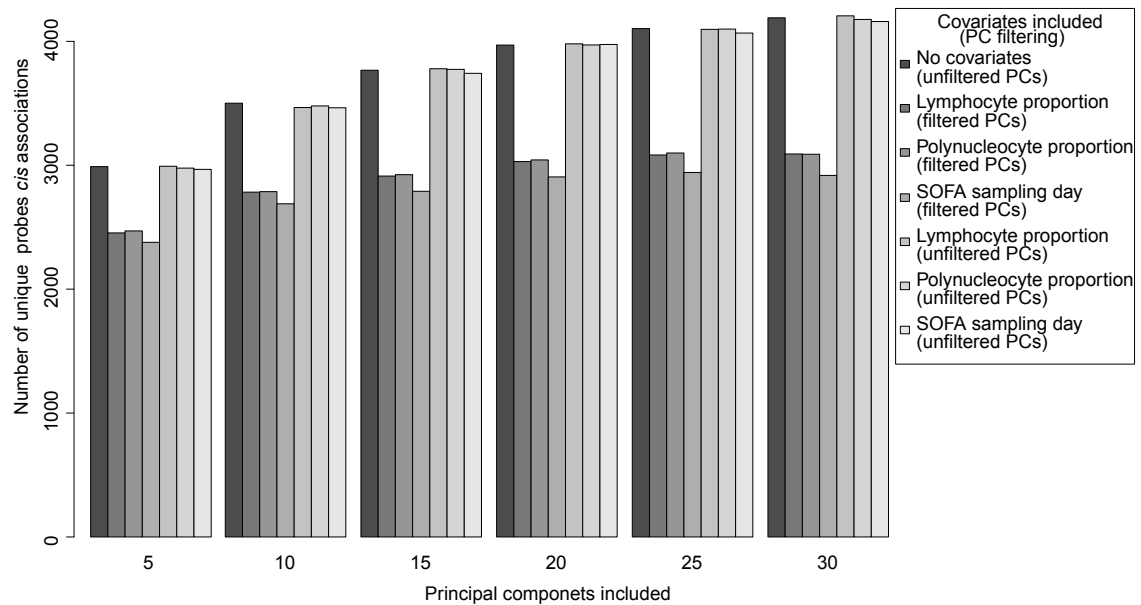


Figure B.2: Inclusion of PCs and clinical covariates. The effect of including PCs and clinical covariates on the number of unique probes with a *cis* association. Taking account of both a clinical covariate and PCs filtered to remove those associated with a clinical covariate reduces the number of eQTL compared to taking account of PCs alone. Including both a clinical covariate and unfiltered PCs results in a modest increase in the number of eQTL detected.

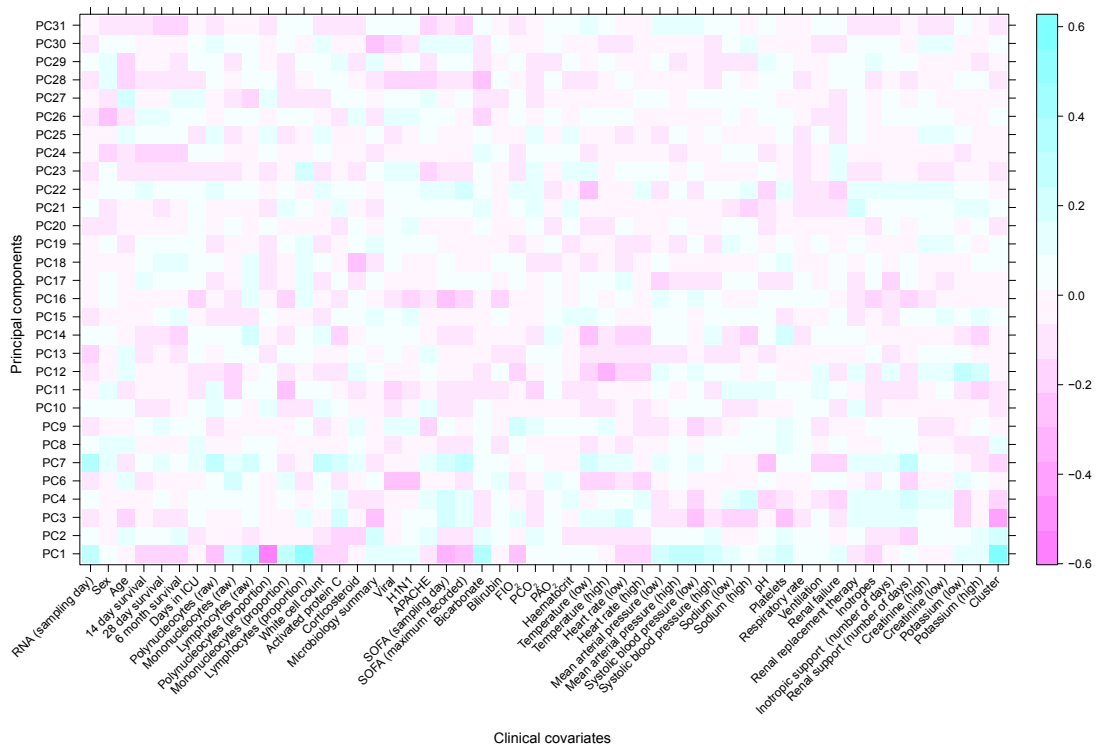


Figure B.3: Correlations between PCs and clinical covariates.

Table B.2: *Cis* eQTL pathway enrichment.

Ingenuity Canonical Pathways	p value
PI3K Signalling in B Lymphocytes	9.33E-05
Mitochondrial Dysfunction	9.77E-05
NAD Phosphorylation and Dephosphorylation	4.68E-04
Glycogen Degradation III	1.10E-03
Glutathione-mediated Detoxification	1.12E-03
Glycogen Degradation II	1.82E-03
B Cell Receptor Signalling	1.86E-03
Antigen Presentation Pathway	2.88E-03
Oxidative Phosphorylation	2.95E-03
LPS-stimulated MAPK Signalling	4.17E-03
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	4.68E-03
2-oxobutanoate Degradation I	4.68E-03
Galactose Degradation I (Leloir Pathway)	4.68E-03
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	6.76E-03
Role of PKR in Interferon Induction and Antiviral Response	8.51E-03
Valine Degradation I	9.12E-03
TREM1 Signalling	9.33E-03

Continued on next page

Table B.2 – Continued from previous page

Ingenuity Canonical Pathways	p value
Type II Diabetes Mellitus Signalling	1.15E-02
Superoxide Radicals Degradation	1.20E-02
Assembly of RNA Polymerase I Complex	1.32E-02
Amyloid Processing	1.32E-02
Hypoxia Signalling in the Cardiovascular System	1.45E-02
14-3-3-mediated Signalling	1.48E-02
PKC θ Signalling in T Lymphocytes	1.66E-02
Nucleotide Excision Repair Pathway	1.74E-02
RANK Signalling in Osteoclasts	1.74E-02
3-phosphoinositide Degradation	1.78E-02
Huntington's Disease Signalling	1.86E-02
p70S6K Signalling	1.86E-02
Tumoricidal Function of Hepatic Natural Killer Cells	2.04E-02
Branched-chain α -keto acid Dehydrogenase Complex	2.09E-02
Methylmalonyl Pathway	2.09E-02
Reelin Signalling in Neurons	2.29E-02
Apoptosis Signalling	2.29E-02
Glutathione Redox Reactions I	2.34E-02
Phosphatidylglycerol Biosynthesis II (Non-plastidic)	2.34E-02
Ethanol Degradation IV	2.34E-02
NAD Salvage Pathway II	2.57E-02
p53 Signalling	2.63E-02
Ceramide Signalling	2.63E-02
Prostate Cancer Signalling	2.63E-02
DNA Double-Strand Break Repair by Non-Homologous End Joining	2.88E-02
GNRH Signalling	3.02E-02
HGF Signalling	3.09E-02
Role of NFAT in Cardiac Hypertrophy	3.24E-02
Spliceosomal Cycle	3.31E-02
Guanine and Guanosine Salvage I	3.31E-02
Lipoate Biosynthesis and Incorporation II	3.31E-02
Sulfate Activation for Sulfonation	3.31E-02
Breast Cancer Regulation by Stathmin1	3.39E-02
Death Receptor Signalling	3.39E-02
NF- κ B Activation by Viruses	3.39E-02
CD27 Signalling in Lymphocytes	3.47E-02
Dopamine-DARPP32 Feedback in cAMP Signalling	3.55E-02
NRF2-mediated Oxidative Stress Response	3.80E-02
Salvage Pathways of Pyrimidine Ribonucleotides	3.80E-02
Antioxidant Action of Vitamin C	3.80E-02
CDP-diacylglycerol Biosynthesis I	4.07E-02
Oxidative Ethanol Degradation III	4.07E-02
Salvage Pathways of Pyrimidine Deoxyribonucleotides	4.17E-02

Continued on next page

Table B.2 – *Continued from previous page*

Ingenuity Canonical Pathways	p value
IL-8 Signalling	4.17E-02
D-myo-inositol-5-phosphate Metabolism	4.17E-02
fMLP Signalling in Neutrophils	4.27E-02
Superpathway of Methionine Degradation	4.27E-02
iCOS-iCOSL Signalling in T Helper Cells	4.47E-02
Tetrapyrrole Biosynthesis II	4.47E-02
UVB-Induced MAPK Signalling	4.79E-02
3-phosphoinositide Biosynthesis	4.79E-02
Renin-Angiotensin Signalling	4.79E-02
Sphingosine-1-phosphate Signalling	4.79E-02
Glioma Signalling	4.79E-02
Estrogen Receptor Signalling	4.79E-02
Gap Junction Signalling	4.90E-02

Table B.3: *Cis* eQTL unique to sepsis pathway enrichment.

Ingenuity Canonical Pathways	p value
LPS-stimulated MAPK Signalling	1.78E-03
RANK Signalling in Osteoclasts	2.57E-03
α -Adrenergic Signalling	6.46E-03
Role of NFAT in Cardiac Hypertrophy	6.61E-03
Inosine-5'-phosphate Biosynthesis II	1.12E-02
Mechanisms of Viral Exit from Host Cells	1.26E-02
HGF Signalling	1.29E-02
GNRH Signalling	1.29E-02
14-3-3-mediated Signalling	1.35E-02
Glioma Signalling	1.41E-02
Melanocyte Development and Pigmentation Signalling	1.51E-02
Glycogen Degradation II	1.55E-02
Eumelanin Biosynthesis	2.14E-02
Arginine Degradation I (Arginase Pathway)	2.14E-02
Prostate Cancer Signalling	2.69E-02
ErbB4 Signalling	2.69E-02
Glycogen Degradation III	2.75E-02
G α 12/13 Signalling	2.88E-02
NGF Signalling	3.31E-02
P2Y Purigenic Receptor Signalling Pathway	3.31E-02
Tetrapyrrole Biosynthesis II	3.47E-02
IL-17 Signalling	3.47E-02
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	3.47E-02

Continued on next page

Table B.3 – Continued from previous page

Ingenuity Canonical Pathways	p value
PKC θ Signalling in T Lymphocytes	3.72E-02
STAT3 Pathway	3.80E-02
CD27 Signalling in Lymphocytes	3.80E-02
4-1BB Signalling in T Lymphocytes	4.17E-02
Cell Cycle: G1/S Checkpoint Regulation	4.17E-02
NAD biosynthesis II (from tryptophan)	4.37E-02
PI3K Signalling in B Lymphocytes	4.37E-02
UVC-Induced MAPK Signalling	4.47E-02
IL-17A Signalling in Airway Cells	4.47E-02
Calcium-induced T Lymphocyte Apoptosis	4.57E-02
MIF-mediated Glucocorticoid Regulation	4.68E-02
SAPK/JNK Signalling	4.79E-02
Non-Small Cell Lung Cancer Signalling	4.90E-02
Urea Cycle	4.90E-02

Table B.4: Master regulators of *trans* associations. 112 of the SNPs identified from the *trans* eQTL analysis also showed evidence of a *cis* eQTL. For each SNP, the number of genes associating *trans* plus the gene associating in *cis* are summarised.

SNP	No. <i>trans</i> genes	<i>Cis</i> gene	<i>Cis</i> FDR
rs752845	206	<i>PIK3CD</i>	3.14E-05
rs4713488	127	<i>AGPAT1</i>	7.68E-09
rs951954	71	<i>BUB1</i>	4.31E-02
rs9517411	68	<i>UBAC2</i>	2.18E-03
rs2220679	54	<i>SPAG1</i>	4.26E-02
rs16904945	33	<i>PHF20L1</i>	3.17E-03
rs12148454	33	<i>VPS18</i>	1.63E-04
rs6748986	19	<i>EIF4E2</i>	2.89E-04
rs889518	14	<i>ATMIN</i>	3.72E-02
rs7915040	10	<i>EIF4EBP2</i>	3.75E-03
rs367146	10	<i>ITFG3</i>	5.79E-04
rs12397260	9	<i>ARSD</i>	8.39E-03
rs5983130	9	<i>ARSD</i>	1.19E-05
rs2534123	9	<i>ZFY</i>	1.98E-37
rs6725242	8	<i>PDIA6</i>	4.02E-02
rs7576288	8	<i>PDIA6</i>	4.02E-02
rs9786145	8	<i>PRKY</i>	1.07E-11
rs16930700	7	<i>CAMK2G</i>	1.21E-04
rs9786814	7	<i>EIF1AY</i>	3.43E-15
rs16986572	7	<i>FCAR</i>	2.83E-03
rs2874089	7	<i>UTY</i>	6.29E-10

Continued on next page

Table B.4 – Continued from previous page

SNP	No. <i>trans</i> genes	<i>Cis</i> gene	<i>Cis</i> FDR
rs7892889	7	UTY	7.95E-10
rs9786758	6	JARID1D	1.13E-13
rs2713254	6	UTY	2.87E-05
rs754799	5	ATP5D	4.52E-04
rs17316227	5	PRKY	1.64E-04
rs871461	4	GTF3C5	2.00E-02
rs9786139	4	PRKY	2.06E-06
rs2436278	4	ZFY	1.11E-03
rs7611956	3	TSC22D2	1.07E-02
rs9840738	3	TSC22D2	1.07E-02
rs2440113	3	UPF2	1.91E-03
rs295308	3	XRN1	1.27E-02
rs9357436	2	CDC5L	2.27E-02
rs17200768	2	CLIC1	7.48E-04
rs4629186	2	DYSF	3.15E-03
rs34459399	2	EIF1AY	1.98E-06
rs12179975	2	IER3	1.13E-02
rs9500864	2	IER3	1.13E-02
rs1384	2	LYZ	3.54E-23
rs2168029	2	LYZ	5.87E-26
rs7224469	2	NDEL1	5.88E-03
rs9893451	2	NDEL1	5.88E-03
rs4806866	2	S1PR4	1.83E-03
rs3779925	2	SLA	3.17E-02
rs10914540	2	TMEM39B	1.05E-02
rs7544661	2	TMEM39B	3.11E-03
rs968274	2	VNN2	7.93E-03
rs2794961	1	ACBD6	1.74E-04
rs438534	1	BTN2A1	7.52E-04
rs11125073	1	CALM2	3.71E-02
rs10219333	1	CAT	3.09E-02
rs2832192	1	CCT8	1.37E-06
rs2525557	1	COPS6	5.95E-04
rs10014526	1	CRIPAK	2.50E-15
rs10024529	1	CRIPAK	1.08E-21
rs11933855	1	CRIPAK	2.09E-15
rs11937228	1	CRIPAK	2.50E-15
rs1250114	1	CRIPAK	1.24E-16
rs1250116	1	CRIPAK	1.11E-18
rs1265923	1	CRIPAK	1.11E-19
rs1680032	1	CRIPAK	6.23E-21
rs17164229	1	CRIPAK	2.23E-16
rs1732115	1	CRIPAK	6.23E-21

Continued on next page

Table B.4 – Continued from previous page

SNP	No. <i>trans</i> genes	<i>Cis</i> gene	<i>Cis</i> FDR
rs28429103	1	CRIPAK	1.04E-20
rs3817604	1	CRIPAK	1.04E-20
rs730830	1	CRIPAK	9.40E-20
rs884421	1	CRIPAK	1.11E-18
rs10305740	1	CTSK	1.77E-02
rs10305751	1	CTSK	1.77E-02
rs4971044	1	CTSK	1.77E-02
rs11900588	1	DYSF	1.61E-02
rs7127097	1	EIF4G2	8.60E-04
rs11538131	1	FAM134C	4.42E-02
rs10012074	1	KIAA1530	1.36E-08
rs10784774	1	LYZ	3.83E-23
rs10878948	1	LYZ	4.03E-05
rs10878951	1	LYZ	4.03E-05
rs10878964	1	LYZ	6.21E-19
rs10878965	1	LYZ	4.03E-05
rs11177571	1	LYZ	4.03E-05
rs11177610	1	LYZ	6.25E-05
rs2231700	1	LYZ	2.07E-13
rs2305641	1	LYZ	3.51E-18
rs317688	1	LYZ	1.04E-06
rs317689	1	LYZ	6.84E-07
rs4761154	1	LYZ	6.21E-19
rs529268	1	LYZ	2.22E-14
rs589448	1	LYZ	2.75E-17
rs645026	1	LYZ	1.51E-05
rs6581886	1	LYZ	8.46E-13
rs6581889	1	LYZ	1.66E-19
rs7135092	1	LYZ	8.14E-18
rs7297153	1	LYZ	1.12E-04
rs7300266	1	LYZ	8.21E-18
rs7308481	1	LYZ	8.21E-18
rs7980936	1	LYZ	2.08E-13
rs4962150	1	MED22	1.79E-07
rs10419790	1	MRPL54	4.99E-05
rs9512980	1	POMP	1.74E-02
rs11710277	1	RBM6	1.28E-02
rs2023953	1	RBM6	9.79E-03
rs11170	1	RNF160	1.76E-06
rs1537093	1	RNF160	1.70E-05
rs2251087	1	RNF160	4.26E-05
rs4531866	1	SPOPL	3.32E-03
rs10079560	1	STK10	7.28E-03

Continued on next page

Table B.4 – *Continued from previous page*

SNP	No. <i>trans</i> genes	<i>Cis</i> gene	<i>Cis</i> FDR
rs17024393	1	<i>STXBP3</i>	9.55E-03
rs11235911	1	<i>UCP2</i>	5.27E-03
rs11235951	1	<i>UCP2</i>	5.27E-03
rs10456357	1	<i>ZNF193</i>	1.35E-03
rs388017	1	<i>ZNF193</i>	3.25E-02

Appendix C

Functional genomic response to influenza challenge

Table C.1: Differential expression for LCI. The 50 most significant differentially expressed probes.

Probe	Gene	log(FC)	FDR
5900543	<i>SEPT4</i>	1.50	5.12E-03
7160274	<i>TXNDC12</i>	-1.60	1.07E-02
1980524	<i>GBP4</i>	1.68	1.07E-02
4670458	<i>SEPT4</i>	2.15	1.07E-02
3930681	<i>DHX58</i>	1.87	1.07E-02
520059	<i>LAMP3</i>	1.88	1.07E-02
2940059	<i>FTSJD2</i>	1.28	1.07E-02
7320546	<i>LAMP3</i>	2.69	1.07E-02
1030333	<i>CCL2</i>	2.09	1.07E-02
150240	<i>TRIM78P</i>	1.13	1.07E-02
4880010	<i>ZBP1</i>	1.59	1.32E-02
1510196	<i>PHF11</i>	0.96	1.32E-02
4760703	<i>DDX58</i>	1.73	1.53E-02
1170300	<i>MT1G</i>	2.11	1.73E-02
1740360	<i>USP18</i>	2.24	1.91E-02
3990010	<i>Hs.125087</i>	2.24	1.91E-02
3140041	<i>SP140</i>	1.23	1.91E-02
4850228	<i>LOC647135</i>	1.36	1.91E-02
7510537	<i>SCO2</i>	1.56	1.91E-02
240722	<i>OAS2</i>	1.99	1.91E-02
1990300	<i>SOCS1</i>	1.73	1.91E-02
4200240	<i>SP100</i>	1.21	1.94E-02
360132	<i>LHFPL2</i>	1.25	1.94E-02
1440615	<i>OTOF</i>	1.68	1.94E-02
5310427	<i>PATL1</i>	0.94	1.94E-02
2190397	<i>ZCCHC2</i>	1.46	2.13E-02
3310450	<i>LOC648366</i>	1.21	2.27E-02
6400176	<i>IRF7</i>	1.66	2.27E-02
2070059	<i>PARP10</i>	1.30	2.37E-02
2360348	<i>CMPK2</i>	1.64	2.37E-02
7320561	<i>OAS2</i>	1.65	2.37E-02
5220204	<i>C19orf66</i>	1.18	2.37E-02
6290743	<i>APOBEC3F</i>	1.09	2.64E-02
6180056	<i>TOR1B</i>	1.26	2.64E-02
5700020	<i>BTBD15</i>	-0.84	2.64E-02
2340332	<i>SRGAP2L</i>	1.18	2.64E-02
5490470	<i>MX2</i>	1.20	2.64E-02
1470382	<i>IRF7</i>	1.63	2.64E-02
2510523	<i>RBCK1</i>	1.05	2.64E-02
7380221	<i>SPATS2L</i>	1.77	2.64E-02
6200402	<i>MT1A</i>	1.65	2.64E-02

Continued on next page

Table C.1 – *Continued from previous page*

Probe	Gene	log(FC)	FDR
6580484	<i>MOV10</i>	1.35	2.64E-02
6860482	<i>HERC6</i>	1.23	2.65E-02
5960343	<i>PRIC285</i>	1.30	2.65E-02
630397	<i>KIAA0319L</i>	1.52	2.65E-02
130048	<i>KIAA1958</i>	1.22	2.65E-02
6100521	<i>PHF11</i>	0.91	2.65E-02
3180039	<i>RGL1</i>	1.17	2.65E-02
4060632	<i>C17orf87</i>	0.97	2.65E-02
2690435	<i>IFIT1</i>	1.54	2.66E-02

Table C.2: Differential expression for moderate/severe LCI. The 50 most significant differentially expressed probes.

Probe	Gene	log(FC)	FDR
1030333	<i>CCL2</i>	3.77	6.59E-16
1740360	<i>USP18</i>	4.33	9.29E-16
4670458	<i>SEPT4</i>	3.61	2.44E-15
7320546	<i>LAMP3</i>	4.96	5.28E-15
1980524	<i>GBP4</i>	2.94	1.91E-14
520059	<i>LAMP3</i>	3.22	3.93E-14
7380221	<i>SPATS2L</i>	3.40	2.83E-13
4150270	<i>ANKRD22</i>	3.57	8.24E-13
1990300	<i>SOCS1</i>	3.10	1.05E-12
1170300	<i>MT1G</i>	3.71	1.05E-12
130181	<i>ANKRD22</i>	4.03	6.22E-12
7160274	<i>TXNDC12</i>	-2.66	1.12E-11
6860482	<i>HERC6</i>	2.30	1.35E-11
630091	<i>NCOA7</i>	2.47	1.36E-11
6280543	<i>OASL</i>	3.12	2.90E-11
5900543	<i>SEPT4</i>	2.33	4.01E-11
3930681	<i>DHX58</i>	3.22	4.01E-11
1440615	<i>OTOF</i>	2.93	5.53E-11
2940059	<i>FTSJD2</i>	2.08	1.54E-10
4760703	<i>DDX58</i>	2.93	1.64E-10
4880010	<i>ZBP1</i>	2.76	1.91E-10
1450427	<i>RTP4</i>	2.61	2.14E-10
6200402	<i>MT1A</i>	2.94	2.96E-10
1570156	<i>OAS2</i>	2.46	3.16E-10
240722	<i>OAS2</i>	3.43	3.53E-10
5260070	<i>HES4</i>	3.81	4.04E-10
4780128	<i>ATF3</i>	2.80	4.04E-10
6110020	<i>USP41</i>	2.50	4.04E-10
6980133	<i>Hs.537991</i>	2.83	4.58E-10
3140041	<i>SP140</i>	2.13	6.34E-10
6580484	<i>MOV10</i>	2.37	7.11E-10
7320561	<i>OAS2</i>	2.82	7.11E-10
450615	<i>MT2A</i>	3.16	7.11E-10
6400176	<i>IRF7</i>	2.79	7.82E-10
2510523	<i>RBCK1</i>	1.87	8.53E-10
2070059	<i>PARP10</i>	2.21	9.51E-10
6180056	<i>TOR1B</i>	2.20	1.01E-09
7550242	<i>HESX1</i>	2.70	1.24E-09
2230204	<i>OAS2</i>	3.10	1.93E-09
3990010	<i>Hs.125087</i>	3.74	2.46E-09
3800398	<i>FBXO6</i>	2.67	2.62E-09

Continued on next page

Table C.2 – *Continued from previous page*

Probe	Gene	log(FC)	FDR
3180039	<i>RGL1</i>	1.94	3.29E-09
1440300	<i>SLC27A3</i>	1.79	3.60E-09
6620121	<i>CCL8</i>	3.95	4.43E-09
5700753	<i>CEACAM1</i>	2.63	4.49E-09
7330392	<i>TAP1</i>	1.62	5.02E-09
3060735	<i>IL1RN</i>	2.85	5.05E-09
2480450	<i>GALM</i>	2.77	5.21E-09
10360	<i>OAS3</i>	3.83	5.21E-09
630397	<i>KIAA0319L</i>	2.55	6.17E-09

Functional genomic response to influenza challenge

Table C.3: IPA moderate/severe LCI 48 hours post-challenge. The differentially expressed probes between the four moderate/severe LCI samples 48 hours post challenge and the remaining samples were used to determine pathway enrichment with IPA. Pathways with an enrichment $-\log(p \text{ value}) > 1.3$ ($p \text{ value} < 0.05$) are shown. The ratio is calculated from the number of differentially expressed genes in the dataset divided by the number of genes in the pathway.

Ingenuity Canonical Pathways	p value
Interferon Signalling	5.01E-12
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	1.51E-09
Activation of IRF by Cytosolic Pattern Recognition Receptors	2.45E-06
EIF2 Signalling	4.57E-05
Tec Kinase Signalling	5.75E-05
Prolactin Signalling	6.46E-05
JAK/Stat Signalling	1.48E-04
Type I Diabetes Mellitus Signalling	1.55E-04
Role of NFAT in Regulation of the Immune Response	1.82E-04
Role of JAK2 in Hormone-like Cytokine Signalling	1.91E-04
Type II Diabetes Mellitus Signalling	3.16E-04
Antigen Presentation Pathway	3.31E-04
Natural Killer Cell Signalling	6.76E-04
T Cell Receptor Signalling	8.13E-04
Growth Hormone Signalling	8.71E-04
HGF Signalling	1.05E-03
TREM1 Signalling	1.29E-03
CTLA4 Signalling in Cytotoxic T Lymphocytes	1.38E-03
Glucocorticoid Receptor Signalling	1.51E-03
mTOR Signalling	1.78E-03
Calcium-induced T Lymphocyte Apoptosis	1.82E-03
Role of JAK1 and JAK3 in γ c Cytokine Signalling	1.82E-03
iCOS-iCOSL Signalling in T Helper Cells	1.82E-03
Primary Immunodeficiency Signalling	2.69E-03
Thrombopoietin Signalling	3.72E-03
CCR5 Signalling in Macrophages	3.89E-03
Role of Hypercytokinemia/hyperchemokine in the Pathogenesis of Influenza	5.01E-03
GADD45 Signalling	5.37E-03
Fc γ Receptor-mediated Phagocytosis in Macrophages and Monocytes	6.92E-03
IL-9 Signalling	7.41E-03
Oncostatin M Signalling	7.41E-03
Salvage Pathways of Pyrimidine Ribonucleotides	7.94E-03
Systemic Lupus Erythematosus Signalling	7.94E-03
Regulation of eIF4 and p70S6K Signalling	8.13E-03
Acute Phase Response Signalling	9.33E-03

Continued on next page

Table C.3 – *Continued from previous page*

Ingenuity Canonical Pathways	p value
PKC θ Signalling in T Lymphocytes	1.00E-02
IL-12 Signalling and Production in Macrophages	1.00E-02
Erythropoietin Signalling	1.10E-02
Retinoic acid Mediated Apoptosis Signalling	1.10E-02
Cholecystokinin/Gastrin-mediated Signalling	1.15E-02
Role of JAK family kinases in IL-6-type Cytokine Signalling	1.20E-02
Phospholipase C Signalling	1.23E-02
N-acetylglucosamine Degradation II	1.23E-02
Protein Ubiquitination Pathway	1.26E-02
DNA Double-Strand Break Repair by Homologous Recombination	1.32E-02
Dendritic Cell Maturation	1.38E-02
Mechanisms of Viral Exit from Host Cells	1.48E-02
IL-15 Production	1.58E-02
Communication between Innate and Adaptive Immune Cells	1.82E-02
Renin-Angiotensin Signalling	1.82E-02
Granzyme B Signalling	1.91E-02
HER-2 Signalling in Breast Cancer	2.09E-02
Role of RIG1-like Receptors in Antiviral Innate Immunity	2.14E-02
Role of BRCA1 in DNA Damage Response	2.19E-02
RAN Signalling	2.24E-02
VDR/RXR Activation	2.34E-02
NF-KB Signalling	2.51E-02
IGF-1 Signalling	2.57E-02
Endoplasmic Reticulum Stress Pathway	2.63E-02
CD28 Signalling in T Helper Cells	2.63E-02
p53 Signalling	2.69E-02
GM-CSF Signalling	2.69E-02
Complement System	3.09E-02
Nitric Oxide Signalling in the Cardiovascular System	3.16E-02
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	3.16E-02
B Cell Receptor Signalling	3.16E-02
p70S6K Signalling	3.16E-02
Huntington's Disease Signalling	3.47E-02
Role of NFAT in Cardiac Hypertrophy	3.55E-02
Adenine and Adenosine Salvage VI	3.63E-02
Pathogenesis of Multiple Sclerosis	4.07E-02
Salvage Pathways of Pyrimidine Deoxyribonucleotides	4.07E-02
IL-2 Signalling	4.37E-02
G Beta Gamma Signalling	4.37E-02
Polyamine Regulation in Colon Cancer	4.47E-02
IL-3 Signalling	4.47E-02
PI3K Signalling in B Lymphocytes	4.47E-02

Continued on next page

Functional genomic response to influenza challenge

Table C.3 – *Continued from previous page*

Ingenuity Canonical Pathways	p value
Crosstalk between Dendritic Cells and Natural Killer Cells	4.68E-02
UVA-Induced MAPK Signalling	4.90E-02

Table C.4: Differential expression for vaccination status pre-challenge.

Probe	Gene	log(FC)	FDR
3610139	<i>GSTM1</i>	-2.23178	1.60E-03
3990241	<i>LOC654191</i>	-1.22742	1.60E-03
6760471	<i>TMCC1</i>	-1.10761	1.60E-03
1300075	<i>PCTP</i>	-1.2816	4.77E-03
4230768	<i>LOC653602</i>	-0.97007	9.86E-03
870437	<i>TRPM4</i>	-1.20658	1.06E-02
6400435	<i>LOC646156</i>	0.957732	1.06E-02
3710053	<i>APBA2</i>	-1.03331	1.86E-02
3460241	<i>FCHSD1</i>	0.874938	2.21E-02
1050300	<i>LOC100129958</i>	-0.92395	2.54E-02
6900592	<i>GSTM4</i>	-1.12822	2.54E-02
6580019	<i>LOC731954</i>	-1.15539	2.54E-02
3780093	<i>LILRA1</i>	-0.91	3.26E-02
1690367	<i>C14orf118</i>	-0.86662	3.26E-02
3520241	<i>DACT1</i>	0.882552	3.26E-02
6760730	<i>RSRC1</i>	-1.16138	3.39E-02
780491	<i>C10orf4</i>	-0.92438	3.75E-02
520528	<i>CCDC99</i>	-1.09568	4.05E-02
1710014	<i>TNFRSF10B</i>	-0.93259	4.05E-02
2970133	<i>PWP2</i>	-0.87504	4.05E-02
1440019	<i>CASP2</i>	-0.96348	4.05E-02
2120743	<i>GSTA2</i>	0.884013	4.46E-02
2630411	<i>COX11</i>	-0.90537	4.53E-02

Table C.5: Influenza and sepsis overlap.

Genes				
<i>BATF2</i>	<i>Hs.72010</i>	<i>LGALS3BP</i>	<i>OAS3</i>	<i>SERPING1</i>
<i>CUL1</i>	<i>IFI44</i>	<i>LOC389386</i>	<i>OTOF</i>	<i>SP140</i>
<i>EPSTI1</i>	<i>IFI44L</i>	<i>LOC652699</i>	<i>PARP12</i>	<i>SPATS2L</i>
<i>GALM</i>	<i>IFIH1</i>	<i>LY6E</i>	<i>PSME2</i>	<i>TIMM10</i>
<i>HERC6</i>	<i>ISG15</i>	<i>MOV10</i>	<i>RSAD2</i>	<i>TRIM38</i>
<i>HES4</i>	<i>LAMP3</i>	<i>OAS1</i>	<i>RTP4</i>	<i>XAF1</i>
<i>Hs.125087</i>	<i>LAP3</i>	<i>OAS2</i>	<i>SCO2</i>	<i>ZCCHC2</i>

Appendix D

Functional genomic response in CVID

Table D.1: Previously described genes for primary immunodeficiency diseases. (Bousfiha et al. 2013; Al-Herz et al. 2011; Milner and Holland 2013; Stepensky et al. 2013; Keller et al. 2013).

Genes			
ACTB	CFI	IL2RG	PTPRC
ADA	CIITA	IL36RN	RAB27a
ADAM17	COLEC11	IL7R	RAC2
AICDA	CORO1A	IRAK4	RAG1
AIRE	CSF2RA	IRF8	RAG2
AK2	CTSC	ISG15	RFX 5
AP3B	CXCR4	ITCH	RFXANK
APOL1	CYBA	ITGB2	RFXAP
ATM	CYBB	ITK	RHOH
BLM	DCLRE1C	JAK3	RMRP
BLNK	DKC1	KRAS	RNF168
BTK	DNMT3B	LCK	ROBLD3
CEBPE	DOCK8	LIG4	SBDS
C15	ELANE	LPIN2	SERPING1
C16orf57	EVER1	LRBA	SH2D1A
C1QA	EVER2	LYST	SLC35C1
C1QB	FADD	MAGT1	SMARCA1
C1QC	FAS	MASP1	SP110
C1R	FASLG	MASP2	SPINK5
C1S	FCN3	MBL2	STAT1
C2	FERMT3	MCM4	STAT3
C3	FOXP1	MEFV	STAT5B
C4	FOXP3	MIRL	STIM-1
C5	FPR1	MLL2	STK4
C6	G6PC3	MRE11	STX11
C7	G6PD	MSH5	STXBP2
C8A	G6PT1	MTHFD1	TAP1
C8B	GATA2	MVK	TAP2
C9	GFI1	MYD88	TAPBP
CARD11	HAX1	NBS1	TAZ
CARD9	HOIL 1	NCF1	TBK1
CASP10	ICOS	NCF2	TERC
CASP8	IFNGR1	NCF4	TERT
CD19	IFNGR2	NHEJ1	TINF2
CD20	IGHM	NHP2	TLR3
CD21	IGKC	NKX2-5	TNFRSF1A
CD247	IGLL1	NLRP3	TRAC
CD27	IKBA	NOD2	TRAF3
CD35	IKBkappa	NOP10	TRIF
CD3D	IKBKG	NRAS	TYK2
CD3E	IKZF1	ORAI1	UNC119

Continued on next page

Table D.1 – *Continued from previous page*

Genes			
CD3G	IL10	PFC	UNC13D
CD40	IL10RA	PIGA	UNC93B1
CD40L	IL10RB	PIK3R1	UNG
CD46	IL12B	PLC2G	VPS13B
CD79A	IL12RB1	PLDN	WAS
CD79B	IL17F	PMS2	WIPF1
CD81	IL17RA	PNP	XIAP
CD8A	IL1RN	PRF1	ZAP70
CFD	IL21R	PRKDC	ZBTB24
CFH	IL2RA	PSTPIP1	

Functional genomic response in CVID

Table D.2: CVID GWAS genes. Summary of genes from most significant associated regions based on genotype association for the CVID phenotype and relevant subphenotypes.

CVID phenotype	LIP phenotype	IgM- phenotype
<i>MHC</i>	<i>ZNF645</i>	<i>LDLRAP1</i>
<i>ADAM28</i>	<i>AOX1</i>	<i>BC034799</i>
<i>ADAM7</i>	<i>DNAPTP6</i>	<i>F2RL1</i>
<i>ADAMDEC1</i>	<i>KCTD18</i>	
<i>STC1</i>	<i>SGOL2</i>	
<i>AK095968</i>	<i>TUSC3</i>	
<i>CDH23</i>	<i>AK094504</i>	
<i>SDK1</i>	<i>TMEM28</i>	
<i>FLJ32784</i>	<i>HLA-DRA</i>	
<i>UBXN10</i>		
<i>PTCHD3</i>		
<i>FGGY</i>		
<i>PPFIA2</i>		
<i>MMP27</i>		
<i>ZFHX3</i>		
<i>KLHL38</i>		
<i>GRM7</i>		
<i>BC020876</i>		
<i>LRRC16A</i>		
<i>MTRF1L</i>		
<i>CHL1</i>		
<i>DEPDC6</i>		
<i>FLJ16124</i>		
<i>BC043580</i>		

Table D.3: WGS coverage per sample.

Sample	Coverage (x)
C003	36.39377
C006	34.2921
C018	31.22145
C019	28.47816
C023	32.38855
C024	29.87695
C028	29.06513
C032	29.51614
C036	29.51712
C038	31.0639
C044	29.93026
C063	29.58286
C065	27.48816
C072	32.28586
C078	31.94046
C082	30.18512
C085	28.79322
C089	28.91554
D209	29.91589
D218	28.09291
D232	28.08959
D276	30.59691
D269	28.82057
D325	40.29593
D334	29.97318
D345	29.8109
D374	31.37916
D38	31.26916
D575	27.52372
D641	29.1277
D667	32.39151
D705	28.4802
D745	32.25072
D765	28.75778
D839	36.70012

Table D.4: WGS variant pathway enrichment.

Ingenuity Canonical Pathways	p value
Allograft Rejection Signalling	4.74E-08
OX40 Signalling Pathway	9.29E-07

Continued on next page

Table D.4 – *Continued from previous page*

Ingenuity Canonical Pathways	p value
CD28 Signalling in T Helper Cells	1.01E-06
Hematopoiesis from Pluripotent Stem Cells	1.45E-06
Type I Diabetes Mellitus Signalling	3.90E-06
Graft-versus-Host Disease Signalling	7.66E-06
EIF2 Signalling	7.76E-06
Autoimmune Thyroid Disease Signalling	8.86E-06
Regulation of IL-2 Expression in Activated and Anergic T Lymphocytes	1.58E-05
CCR5 Signalling in Macrophages	1.64E-05
Dendritic Cell Maturation	2.31E-05
Lipid Antigen Presentation by CD1	3.43E-05
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	3.83E-05
Granulocyte Adhesion and Diapedesis	4.92E-05
Cytotoxic T Lymphocyte-mediated Apoptosis of Target Cells	5.36E-05
Altered T Cell and B Cell Signalling in Rheumatoid Arthritis	5.36E-05
iCOS-iCOSL Signalling in T Helper Cells	5.61E-05
Nur77 Signalling in T Lymphocytes	5.64E-05
Communication between Innate and Adaptive Immune Cells	5.88E-05
Systemic Lupus Erythematosus Signalling	6.39E-05
Role of NFAT in Regulation of the Immune Response	8.62E-05
SAPK/JNK Signalling	1.34E-04
T Helper Cell Differentiation	2.47E-04
Regulation of Actin-based Motility by Rho	4.97E-04
Glucocorticoid Receptor Signalling	4.98E-04
NF- κ B Signalling	5.26E-04
Calcium-induced T Lymphocyte Apoptosis	5.66E-04
fMLP Signalling in Neutrophils	5.78E-04
mTOR Signalling	8.02E-04
PKC θ Signalling in T Lymphocytes	8.02E-04
IL-6 Signalling	8.83E-04
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	1.06E-03
IL-10 Signalling	1.31E-03
Tec Kinase Signalling	1.89E-03
Molecular Mechanisms of Cancer	2.11E-03
Type II Diabetes Mellitus Signalling	2.28E-03
Human Embryonic Stem Cell Pluripotency	2.32E-03
Role of IL-17A in Arthritis	2.39E-03
CXCR4 Signalling	2.63E-03
Semaphorin Signalling in Neurons	2.91E-03
Crosstalk between Dendritic Cells and Natural Killer Cells	3.03E-03
HMGB1 Signalling	3.35E-03
IL-8 Signalling	3.75E-03

Continued on next page

Table D.4 – *Continued from previous page*

Ingenuity Canonical Pathways	p value
Death Receptor Signalling	3.78E-03
Hepatic Cholestasis	3.82E-03
Phospholipases	3.83E-03
Dolichyl-diphosphooligosaccharide Biosynthesis	4.25E-03
FXR/RXR Activation	4.49E-03
Mitochondrial Dysfunction	4.72E-03
LPS-stimulated MAPK Signalling	4.84E-03
Cdc42 Signalling	4.95E-03
Role of JAK2 in Hormone-like Cytokine Signalling	5.64E-03
Pyridoxal 5'-phosphate Salvage Pathway	5.98E-03
CTLA4 Signalling in Cytotoxic T Lymphocytes	6.26E-03
Superpathway of Cholesterol Biosynthesis	6.34E-03
Superpathway of Methionine Degradation	6.70E-03
Androgen Signalling	6.79E-03
Gαq Signalling	6.83E-03
Small Cell Lung Cancer Signalling	7.48E-03
Epithelial Adherens Junction Signalling	7.66E-03
G Beta Gamma Signalling	7.74E-03
Purine Nucleotides De Novo Biosynthesis II	8.21E-03
NAD biosynthesis II (from tryptophan)	8.60E-03
Cellular Effects of Sildenafil (Viagra)	9.16E-03
Cholecystokinin/Gastrin-mediated Signalling	1.07E-02
Pancreatic Adenocarcinoma Signalling	1.15E-02
PPAR Signalling	1.16E-02
Salvage Pathways of Pyrimidine Ribonucleotides	1.16E-02
TGF-β Signalling	1.27E-02
IL-17A Signalling in Airway Cells	1.34E-02
PEDF Signalling	1.34E-02
VEGF Signalling	1.37E-02
Regulation of the Epithelial-Mesenchymal Transition Pathway	1.40E-02
Huntington's Disease Signalling	1.44E-02
Role of PKR in Interferon Induction and Antiviral Response	1.45E-02
BMP signalling pathway	1.51E-02
Phosphatidylglycerol Biosynthesis II (Non-plastidic)	1.72E-02
Pyrimidine Deoxyribonucleotides De Novo Biosynthesis I	1.75E-02
IGF-1 Signalling	1.77E-02
Role of PI3K/AKT Signalling in the Pathogenesis of Influenza	1.79E-02
Estrogen-Dependent Breast Cancer Signalling	1.79E-02
RANK Signalling in Osteoclasts	1.80E-02
Bile Acid Biosynthesis, Neutral Pathway	1.80E-02
Antiproliferative Role of Somatostatin Receptor 2	1.91E-02
Gα12/13 Signalling	2.04E-02
p53 Signalling	2.09E-02
Prostate Cancer Signalling	2.12E-02

Continued on next page

Table D.4 – *Continued from previous page*

Ingenuity Canonical Pathways	p value
Melatonin Degradation I	2.13E-02
T Cell Receptor Signalling	2.20E-02
Superpathway of Geranylgeranyldiphosphate Biosynthesis I (via Mevalonate)	2.24E-02
IL-9 Signalling	2.24E-02
Role of JAK family kinases in IL-6-type Cytokine Signalling	2.26E-02
Melanoma Signalling	2.29E-02
Erythropoietin Signalling	2.40E-02
Breast Cancer Regulation by Stathmin1	2.48E-02
Glioma Signalling	2.55E-02
Chronic Myeloid Leukemia Signalling	2.57E-02
IL-17A Signalling in Gastric Cells	2.60E-02
Glutamate Receptor Signalling	2.62E-02
IL-15 Signalling	2.63E-02

Functional genomic response in CVID

Table D.5: Differential expression RNA-seq CVID. The 50 most significant differentially expressed genes.

Gene	log(FC)	FDR
<i>IGHA2</i>	-4.98	1.93E-14
<i>PPAP2B</i>	3.83	7.54E-12
<i>SOX5</i>	3.35	1.96E-10
<i>IGJ</i>	-4.33	6.89E-10
<i>LRP5</i>	3.71	3.80E-09
<i>ARL4D</i>	3.61	9.19E-09
<i>DUSP15</i>	5.36	1.48E-08
<i>HLA-DRB5</i>	-8.23	1.63E-08
<i>AC096559.1</i>	5.60	2.84E-08
<i>SDC3</i>	4.15	3.75E-08
<i>SMOC2</i>	6.31	5.18E-08
<i>RP11-22B23.1</i>	3.03	5.64E-08
<i>DLGAP1</i>	5.85	1.39E-07
<i>THEMIS2</i>	2.62	6.28E-07
<i>SDR16C5</i>	-8.28	6.28E-07
<i>LRRC16B</i>	2.88	6.28E-07
<i>FAM167A</i>	3.53	6.67E-07
<i>ENC1</i>	2.75	7.09E-07
<i>FAM177B</i>	-2.70	8.84E-07
<i>SEMA7A</i>	2.23	1.15E-06
<i>SLC25A29</i>	2.73	1.86E-06
<i>TBX21</i>	2.69	2.16E-06
<i>TBKBP1</i>	2.17	3.29E-06
<i>KLK1</i>	2.62	4.35E-06
<i>ITGAX</i>	2.88	5.57E-06
<i>IGHG4</i>	-2.77	6.29E-06
<i>AEBP1</i>	2.26	9.16E-06
<i>CNTNAP2</i>	-2.33	9.82E-06
<i>TAS1R3</i>	2.77	9.82E-06
<i>ADAM15</i>	2.02	1.22E-05
<i>CD96</i>	-2.09	1.51E-05
<i>RP3-462D8.2</i>	-5.85	1.60E-05
<i>TCTN1</i>	-1.99	1.73E-05
<i>DENND2A</i>	4.64	2.12E-05
<i>GIPR</i>	3.16	2.12E-05
<i>TUBB6</i>	2.06	2.80E-05
<i>RPS4Y2</i>	-4.92	2.84E-05
<i>PARM1</i>	-2.46	3.99E-05
<i>P2RY14</i>	-2.52	4.04E-05

Continued on next page

Table D.5 – Continued from previous page

Gene	log(FC)	FDR
<i>NDRG4</i>	3.32	4.12E-05
<i>RP11-731F5.2</i>	2.20	5.12E-05
<i>ZFPM1</i>	2.09	5.12E-05
<i>COCH</i>	-3.30	7.08E-05
<i>IFNG-AS1</i>	-4.02	7.82E-05
<i>ANK2</i>	3.72	9.85E-05
<i>WASF3</i>	3.62	1.09E-04
<i>MGAT3</i>	2.19	1.13E-04
<i>CHDH</i>	3.93	1.14E-04
<i>CNTNAP3</i>	4.20	1.38E-04
<i>TMEM255B</i>	3.59	2.09E-04

Table D.6: RNA-seq pathway enrichment.

Ingenuity Canonical Pathways	p value
Dendritic Cell Maturation	2.69E-05
Autoimmune Thyroid Disease Signalling	1.55E-04
Role of NFAT in Regulation of the Immune Response	5.89E-04
IL-10 Signalling	8.71E-04
T Helper Cell Differentiation	9.33E-04
IL-4 Signalling	1.29E-03
Allograft Rejection Signalling	2.14E-03
NAD Salvage Pathway II	2.45E-03
Role of Osteoblasts, Osteoclasts and Chondrocytes in Rheumatoid Arthritis	2.95E-03
Communication between Innate and Adaptive Immune Cells	3.24E-03
TNFR2 Signalling	3.39E-03
CD40 Signalling	5.01E-03
GDNF Family Ligand-Receptor Interactions	6.46E-03
Neurotrophin/TRK Signalling	6.46E-03
Type I Diabetes Mellitus Signalling	6.61E-03
Tec Kinase Signalling	7.41E-03
April Mediated Signalling	8.13E-03
Axonal Guidance Signalling	8.13E-03
Oleate Biosynthesis II (Animals)	8.51E-03
Guanosine Nucleotides Degradation III	8.51E-03
CD28 Signalling in T Helper Cells	8.91E-03
PKC θ Signalling in T Lymphocytes	8.91E-03
B Cell Activating Factor Signalling	9.33E-03
Urate Biosynthesis/Inosine 5'-phosphate Degradation	9.77E-03

Continued on next page

Table D.6 – *Continued from previous page*

Ingenuity Canonical Pathways	p value
Reelin Signalling in Neurons	1.10E-02
Systemic Lupus Erythematosus Signalling	1.12E-02
Graft-versus-Host Disease Signalling	1.38E-02
Chondroitin Sulfate Biosynthesis (Late Stages)	1.38E-02
γ -linolenate Biosynthesis II (Animals)	1.45E-02
Adenosine Nucleotides Degradation II	1.45E-02
Altered T Cell and B Cell Signalling in Rheumatoid Arthritis	1.45E-02
RANK Signalling in Osteoclasts	1.51E-02
Hematopoiesis from Pluripotent Stem Cells	1.82E-02
Role of Macrophages, Fibroblasts and Endothelial Cells in Rheumatoid Arthritis	1.91E-02
Primary Immunodeficiency Signalling	1.91E-02
Purine Nucleotides Degradation II (Aerobic)	2.00E-02
Choline Degradation I	2.19E-02
Chondroitin Sulfate Biosynthesis	2.24E-02
Dermatan Sulfate Biosynthesis	2.45E-02
Death Receptor Signalling	2.57E-02
Induction of Apoptosis by HIV1	2.63E-02
Germ Cell-Sertoli Cell Junction Signalling	2.82E-02
iCOS-iCOSL Signalling in T Helper Cells	2.95E-02
B Cell Receptor Signalling	3.02E-02
Role of Tissue Factor in Cancer	3.31E-02
Type II Diabetes Mellitus Signalling	3.80E-02
IL-6 Signalling	3.80E-02
Renal Cell Carcinoma Signalling	4.07E-02
PEDF Signalling	4.27E-02
LPS-stimulated MAPK Signalling	4.57E-02
Production of Nitric Oxide and Reactive Oxygen Species in Macrophages	4.68E-02
B Cell Development	4.79E-02

Bibliography

- 1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., and McVean, G. A. (2012). An integrated map of genetic variation from 1,092 human genomes. **Nature** 491: 56–65.
- Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., Kondrashov, A. S., and Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. **Nat Methods** 7: 248–249.
- Al-Herz, W., Bousfiha, A., Casanova, J. L., Chapel, H., Conley, M. E., Cunningham-Rundles, C., Etzioni, A., Fischer, A., Franco, J. L., Geha, R. S., et al. (2011). Primary immunodeficiency diseases: an update on the classification from the international union of immunological societies expert committee for primary immunodeficiency. **Front Immunol** 2: 54.
- Alcais, A., Abel, L., and Casanova, J. L. (2009). Human genetics of infectious diseases: between proof of principle and paradigm. **J Clin Invest** 119: 2506–2514.
- Allison, A. C. (1954). Protection afforded by sickle-cell trait against subtertian malarial infection. **Br Med J** 1: 290–294.
- Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. **Genome Biology** 11: R106.
- Anders, S., Pyl, P. T., and Huber, W. (2014). HTSeq–A Python framework to work with high-throughput sequencing data. **bioRxiv preprint**. DOI: 10 . 1101 / 002824.
- Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. **Nat Protoc** 5: 1564–1573.
- Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. **Nature** 507: 455–461.
- Andrades, M. E., Morina, A., Spasic, S., and Spasojevic, I. (2011). Bench-to-bedside review: sepsis - from the redox point of view. **Crit Care** 15: 230.
- Angulo, I., Vadas, O., Garcon, F., Banham-Hall, E., Plagnol, V., Leahy, T. R., Baxendale, H., Coulter, T., Curtis, J., Wu, C., et al. (2013). Phosphoinositide 3-kinase delta gene mutation predisposes to respiratory infection and airway damage. **Science** 342: 866–871.
- Angus, D. C., Linde-Zwirble, W. T., Lidicker, J., Clermont, G., Carcillo, J., and Pinsky, M. R. (2001). Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. **Crit Care Med** 29: 1303–1310.

BIBLIOGRAPHY

- Antrobus, R. D., Lillie, P. J., Berthoud, T. K., Spencer, A. J., McLaren, J. E., Ladell, K., Lambe, T., Milicic, A., Price, D. A., Hill, A. V., et al. (2012). A T cell-inducing influenza vaccine for the elderly: safety and immunogenicity of MVA-NP+M1 in adults aged over 50 years. **PLoS One** 7: e48322.
- Antrobus, R. D., Berthoud, T. K., Mullarkey, C. E., Hoschler, K., Coughlan, L., Zambon, M., Hill, A. V., and Gilbert, S. C. (2014). Coadministration of Seasonal Influenza Vaccine and MVA-NP+M1 Simultaneously Achieves Potent Humoral and Cell-Mediated Responses. **Mol Ther** 22: 233–238.
- Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. **Nat Rev Genet** 12: 745–755.
- Barzon, L., Lavezzo, E., Costanzi, G., Franchin, E., Toppo, S., and Palu, G. (2013). Next-generation sequencing technologies in diagnostic virology. **J Clin Virol** 58: 346–350.
- Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. *International AAAI Conference on Weblogs and Social Media*.
- Bateman, E. A., Ayers, L., Sadler, R., Lucas, M., Roberts, C., Woods, A., Packwood, K., Burden, J., Harrison, D., Kaenzig, N., et al. (2012). T cell phenotypes in patients with common variable immunodeficiency disorders: associations with clinical phenotypes in comparison with other groups with recurrent infections. **Clin Exp Immunol** 170: 202–211.
- Bernard, G. R., Vincent, J. L., Laterre, P. F., LaRosa, S. P., Dhainaut, J. F., Lopez-Rodriguez, A., Steingrub, J. S., Garber, G. E., Helterbrand, J. D., Ely, E. W., et al. (2001). Efficacy and safety of recombinant human activated protein C for severe sepsis. **N Engl J Med** 344: 699–709.
- Berthoud, T. K., Hamill, M., Lillie, P. J., Hwenda, L., Collins, K. A., Ewer, K. J., Milicic, A., Poyntz, H. C., Lambe, T., Fletcher, H. A., et al. (2011). Potent CD8+ T-cell immunogenicity in humans of a novel heterosubtypic influenza A vaccine, MVA-NP+M1. **Clin Infect Dis** 52: 1–7.
- Bianchi, M. E. (2007). DAMPs, PAMPs and alarmins: all we need to know about danger. **J Leukoc Biol** 81: 1–5.
- Bick, D. and Dimmock, D. (2011). Whole exome and whole genome sequencing. **Curr Opin Pediatr** 23: 594–600.
- Bone, RC, Balk, RA, Cerra, FB, Dellinger, RP, Fein, AM, Knaus, WA, Schein, RM, and Sibbald, WJ (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. **Chest** 101: 1644–1655.
- Bortz, E. and García-Sastre, A. (2011). Predicting the pathogenesis of influenza from genomic response: a step toward early diagnosis. **Genome Med** 3: 67.
- Bosmann, M. and Ward, P. A. (2013). The inflammatory response in sepsis. **Trends Immunol** 34: 129–136.

BIBLIOGRAPHY

- Bousfiha, A. A., Jeddane, L., Ailal, F., Al Herz, W., Conley, M. E., Cunningham-Rundles, C., Etzioni, A., Fischer, A., Franco, J. L., Geha, R. S., et al. (2013). A phenotypic approach for IUIS PID classification and diagnosis: guidelines for clinicians at the bedside. **J Clin Immunol** 33: 1078–1087.
- Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., et al. (2012). Annotation of functional variation in personal genomes using RegulomeDB. **Genome Research** 22: 1790–1797.
- Brun-Buisson, C., Doyon, F., Carlet, J., Dellamonica, P., Gouin, F., Lepoutre, A., Mercier, J. C., Offenstadt, G., and Rgnier, B. (1995). Incidence, risk factors, and outcome of severe sepsis and septic shock in adults. A multicenter prospective study in intensive care units. French ICU Group for Severe Sepsis. **JAMA** 274: 968–974.
- Byun, M., Abhyankar, A., Lelarge, V., Plancoulaine, S., Palanduz, A., Telhan, L., Boisson, B., Picard, C., Dewell, S., Zhao, C., et al. (2010). Whole-exome sequencing-based discovery of STIM1 deficiency in a child with fatal classic Kaposi sarcoma. **J Exp Med** 207 (11): 2307–2312.
- Calfee, C. S., Thompson, B. T., Parsons, P. E., Ware, L. B., Matthay, M. A., and Wong, H. R. (2010). Plasma interleukin-8 is not an effective risk stratification tool for adults with vasopressor-dependent septic shock. **Crit Care Med** 38: 1436–1441.
- Campbell, J. D., Spira, A., and Lenburg, M. E. (2011). Applying gene expression microarrays to pulmonary disease. **Respirology** 16: 407–418.
- Carlet, J., Cohen, J., Calandra, T., Opal, S. M., and Masur, H. (2008). Sepsis: time to reconsider the concept. **Crit Care Med** 36: 964–966.
- Carrat, F., Vergu, E., Ferguson, N. M., Lemaitre, M., Cauchemez, S., Leach, S., and Valleron, A. J. (2008). Time lines of infection and disease in human influenza: a review of volunteer challenge studies. **Am J Epidemiol** 167: 775–785.
- Cashion, A. K., Umberger, R. A., Goodwin, S. B., and Sutter, T. R. (2011). Collection and storage of human blood and adipose for genomic analysis of clinical samples. **Res Nurs Health** 34: 408–18.
- Castigli, E., Wilson, S. A., Garibyan, L., Rachid, R., Bonilla, F., Schneider, L., and Geha, R. S. (2005). TACI is mutant in common variable immunodeficiency and IgA deficiency. **Nat Genet** 37: 829–834.
- Chan, C. J., Martinet, L., Gilfillan, S., Souza-Fonseca-Guimaraes, F., Chow, M. T., Town, L., Ritchie, D. S., Colonna, M., Andrews, D. M., and Smyth, M. J. (2014). The receptors CD96 and CD226 oppose each other in the regulation of natural killer cell functions. **Nat Immunol** 15: 431–438.
- Chapel, H. and Cunningham-Rundles, C. (2009). Update in understanding common variable immunodeficiency disorders (CVIDs) and the management of patients with these conditions. **Br J Haematol** 145: 709–727.

BIBLIOGRAPHY

- Chapel, H., Lucas, M., Lee, M., Bjorkander, J., Webster, D., Grimbacher, B., Fieschi, C., Thon, V., Abedi, M. R., and Hammarstrom, L. (2008). Common variable immunodeficiency disorders: division into distinct clinical phenotypes. **Blood** 112: 277–286.
- Chapel, H., Lucas, M., Patel, S., Lee, M., Cunningham-Rundles, C., Resnick, E., Gerard, L., and Oksenhendler, E. (2012). Confirmation and improvement of criteria for clinical phenotyping in common variable immunodeficiency disorders in replicate cohorts. **J Allergy Clin Immunol** 130: 1197–1198.
- Chapman, S. J. and Hill, A. V. (2012). Human genetic susceptibility to infectious disease. **Nat Rev Genet** 13: 175–188.
- Cheung, V. G. and Spielman, R. S. (2009). Genetics of human gene expression: mapping DNA variants that influence gene expression. **Nat Rev Genet** 10: 595–604.
- Cheung, V. G., Conlin, L. K., Weber, T. M., Arcaro, M., Jen, K. Y., Morley, M., and Spielman, R. S. (2003). Natural variation in human gene expression assessed in lymphoblastoid cells. **Nat Genet** 33: 422–425.
- Choi, J. K. and Kim, S. C. (2007). Environmental effects on gene expression phenotype have regional biases in the human genome. **Genetics** 175: 1607–13.
- Chou, J., Ohsumi, T. K., and Geha, R. S. (2012). Use of whole exome and genome sequencing in the identification of genetic causes of primary immunodeficiencies. **Curr Opin Allergy Clin Immunol** 12: 623–8.
- Chung, L. P. and Waterer, G. W. (2011). Genetic predisposition to respiratory infection and sepsis. **Crit Rev Clin Lab Sci** 48: 250–68.
- Cirulli, E. T. and Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. **Nat Rev Genet** 11: 415–425.
- Cobb, J. P., Mindrinos, M. N., Miller-Graziano, C., Calvano, S. E., Baker, H. V., Xiao, W., Laudanski, K., Brownstein, B. H., Elson, C. M., Hayden, D. L., et al. (2005). Application of genome-wide expression analysis to human health and disease. **Proc Natl Acad Sci U S A** 102: 4801–6.
- Cohen, J., Opal, S., and Calandra, T. (2012). Sepsis studies need new direction. **Lancet Infect Dis** 12: 503–505.
- Conesa, A. and Nueda, M. J. (R package version 1.28.0). maSigPro: Significant Gene Expression Profile Differences in Time Course Microarray Data. URL: <http://bioinfo.cipf.es/>.
- Conley, M. E., Notarangelo, L. D., and Etzioni, A. (1999). Diagnostic criteria for primary immunodeficiencies. Representing PAGID (Pan-American Group for Immunodeficiency) and ESID (European Society for Immunodeficiencies). **Clin Immunol** 93: 190–197.
- Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. **Nat Rev Genet** 10: 184–194.

BIBLIOGRAPHY

- Cooper, G. M. (2000). *The Cell: A Molecular Approach*. Sunderland: Sinauer Associates.
- Cooper, G. M. and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. **Nat Rev Genet** 12: 628–640.
- Corey, S. J., Minden, M. D., Barber, D. L., Kantarjian, H., Wang, J. C., and Schimmer, A. D. (2007). Myelodysplastic syndromes: the complexity of stem-cell diseases. **Nat Rev Cancer** 7: 118–129.
- Cox, N. J. and Subbarao, K. (1999). Influenza. **Lancet** 354: 1277–1282.
- CSIRO Bioinformatics (R package version 3.0.8). GeneRave: Sparse response modelling for high dimensional data.
- Cuenca, A. G., Gentile, L. F., Lopez, M. C., Ungaro, R., Liu, H., Xiao, W., Seok, J., Mindrinos, M. N., Ang, D., Baslanti, T. O., et al. (2013). Development of a genomic metric that can be rapidly used to predict clinical outcome in severely injured trauma patients. **Crit Care Med** 41: 1175–1185.
- Cunningham-Rundles, C. (2010). How I treat common variable immune deficiency. **Blood** 116: 7–15.
- Cunningham-Rundles, C. and Bodian, C. (1999). Common variable immunodeficiency: clinical and immunological features of 248 patients. **Clin Immunol** 92: 34–48.
- Cunningham-Rundles, C. and Knight, A. K. (2007). Common variable immune deficiency: reviews, continued puzzles, and a new registry. **Immunol Res** 38: 78–86.
- Danai, P. and Martin, G. S. (2005). Epidemiology of sepsis: recent advances. **Curr Infect Dis Rep** 7: 329–334.
- Davila, S., Wright, V. J., Khor, C. C., Sim, K. S., Binder, A., Breunis, W. B., Inwald, D., Nadel, S., Betts, H., Carrol, E. D., et al. (2010). Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. **Nat Genet** 42: 772–6.
- Deans, K. J., Haley, M., Natanson, C., Eichacker, P. Q., and Minneci, P. C. (2005). Novel therapies for sepsis: a review. **J Trauma** 58: 867–874.
- Debey, S., Schoenbeck, U., Hellmich, M., Gathof, B. S., Pillai, R., Zander, T., and Schultze, J. L. (2004). Comparison of different isolation techniques prior gene expression profiling of blood derived cells: impact on physiological responses, on overall expression and the role of different cell types. **Pharmacogenomics J** 4: 193–207.
- Delaneau, O., Marchini, J., and Zagury, J. F. (2012). A linear complexity phasing method for thousands of genomes. **Nat Methods** 9: 179–181.
- Dey-Rao, R., Smith, J. R., Chow, S., and Sinha, A. A. (2014). Differential gene expression analysis in CCLE lesions provides new insights regarding the genetics basis of skin vs. systemic disease. **Genomics** 104: 144–55.

BIBLIOGRAPHY

- Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, G., Castel, D., Estelle, J., et al. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. **Brief Bioinform** 14: 671–683.
- Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., Ingle, C., Beazley, C., Gutierrez Arcelus, M., Sekowska, M., et al. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. **Science** 325: 1246–1250.
- Dranoff, G. (2004). Cytokines in cancer pathogenesis and cancer therapy. **Nat Rev Cancer** 4: 11–22.
- Dzhagalov, I., Chambon, P., and He, Y. W. (2007). Regulation of CD8+ T lymphocyte effector function and macrophage inflammatory cytokine production by retinoic acid receptor gamma. **J Immunol** 178: 2113–2121.
- Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. **Nucleic Acids Res** 30: 207–210.
- Emmert-Streib, F. and Glazko, G. V. (2011). Pathway analysis of expression data: deciphering functional building blocks of complex diseases. **PLoS Comput Biol** 7: e1002053.
- Encode Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. **Nature** 489: 57–74.
- Epstein, S. L. and Price, G. E. (2010). Cross-protective immunity to influenza A viruses. **Expert Rev Vaccines** 9: 1325–1341.
- Everitt, B. and Hothorn, T. (2011). An introduction to applied multivariate analysis with R. New York: Springer.
- Fairfax, B. P. and Knight, J. C. (2014). Genetics of gene expression in immunity to infection. **Curr Opin Immunol** 30C: 63–71.
- Fairfax, B. P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F. O., and Knight, J. C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. **Nat Genet** 44 (5): 502–510.
- Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., Jostins, L., Plant, K., Andrews, R., McGee, C., et al. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. **Science** 343: 1246949.
- FANTOM Consortium, Riken Pmi, Clst, Forrest, A. R., Kawaji, H., Rehli, M., Baillie, J. K., Hoon, M. J. de, Lassmann, T., Itoh, M., Summers, K. M., et al. (2014). A promoter-level mammalian expression atlas. **Nature** 507: 462–470.
- Feezor, R. J., Baker, H. V., Mindrinos, M., Hayden, D., Tannahill, C. L., Brownstein, B. H., Fay, A., MacMillan, S., Laramie, J., Xiao, W., et al. (2004). Whole blood and leukocyte RNA isolation for gene expression analyses. **Physiol Genomics** 19: 247–254.

BIBLIOGRAPHY

- File T. M., Jr. (2004). Streptococcus pneumoniae and community-acquired pneumonia: a cause for concern. **Am J Med** 117 Suppl 3A: 39S–50S.
- Fitting, C., Dhawan, S., and Cavaillon, J. M. (2004). Compartmentalization of tolerance to endotoxin. **J Infect Dis** 189: 1295–1303.
- Flicek, P. and Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. **Nat Methods** 6: S6–S12.
- Fonkwo, P. N. (2008). Pricing infectious disease. The economic and health implications of infectious diseases. **EMBO Rep** 9 Suppl 1: S13–17.
- Franco, L. M., Bucasas, K. L., Wells, J. M., Nino, D., Wang, X., Zapata, G. E., Arden, N., Renwick, A., Yu, P., Quarles, J. M., et al. (2013). Integrative genomic analysis of the human immune response to influenza vaccination. **Elife** 2: e00299.
- Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. **Science** 296: 2225–2229.
- Gaffney, D. J. (2013). Global properties and functional complexity of human gene regulatory variation. **PLoS Genet** 9: e1003501.
- Gaffney, D. J., Veyrieras, J. B., Degner, J. F., Pique-Regi, R., Pai, A. A., Crawford, G. E., Stephens, M., Gilad, Y., and Pritchard, J. K. (2012). Dissecting the regulatory architecture of gene expression QTLs. **Genome Biol** 13: R7.
- Ge, Y., Ezzell, R. M., Clark, B. D., Loiselle, P. M., Amato, S. F., and Warren, H. S. (1997). Relationship of tissue and cellular interleukin-1 and lipopolysaccharide after endotoxemia and bacteremia. **J Infect Dis** 176: 1313–1321.
- Giliani, S., Bonfim, C., Saint Basile, G. de, Lanzi, G., Brousse, N., Koliski, A., Malvezzi, M., Fischer, A., Notarangelo, L. D., and Le Deist, F. (2006). Omenn syndrome in an infant with IL7RA gene mutation. **J Pediatr** 148: 272–274.
- Giovannetti, A., Pierdominici, M., Mazzetta, F., Marziali, M., Renzi, C., Mileo, A. M., De Felice, M., Mora, B., Esposito, A., Carello, R., et al. (2007). Unravelling the complexity of T cell abnormalities in common variable immunodeficiency. **J Immunol** 178: 3932–3943.
- Glazier, A. M., Nadeau, J. H., and Aitman, T. J. (2002). Finding genes that underlie complex traits. **Science** 298: 2345–2349.
- Goldstein, D. B. (2009). Common genetic variation and human traits. **N Engl J Med** 360: 1696–1698.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. **Science** 286: 531–537.
- Gong, W., Emanuel, B. S., Collins, J., Kim, D. H., Wang, Z., Chen, F., Zhang, G., Roe, B., and Budarf, M. L. (1996). A transcription map of the DiGeorge and velo-cardio-facial syndrome minimal critical region on 22q11. **Hum Mol Genet** 5: 789–800.

BIBLIOGRAPHY

- Goris, A. and Liston, A. (2012). The immunogenetic architecture of autoimmune disease. **Cold Spring Harb Perspect Biol** 4.
- Grimbacher, B., Hutloff, A., Schlesier, M., Glocker, E., Warnatz, K., Drger, R., Eibel, H., Fischer, B., Schffer, A. A., Mages, H. W., et al. (2003). Homozygous loss of ICOS is associated with adult-onset common variable immunodeficiency. **Nat Immunol** 4: 261–268.
- Grosshans, H. and Filipowicz, W. (2008). Molecular biology: the expanding world of small RNAs. **Nature** 451: 414–416.
- GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. **Nat Genet** 45: 580–585.
- Gundry, M. and Vijg, J. (2012). Direct mutation analysis by high-throughput sequencing: From germline to low-abundant, somatic variants. **Mutat Res** 729: 1–15.
- Guo, B. R., Zhang, X., Chen, G., Zhang, J. G., Sun, L. D., Du, W. D., Zhang, Q., Cui, Y., Zhu, J., Tang, X. F., et al. (2012). Exome sequencing identifies a COL14A1 mutation in a large Chinese pedigree with punctate palmoplantar keratoderma. **J Med Genet** 49: 563–568.
- Hammarström, L., Vorechovsky, I., and Webster, D. (2000). Selective IgA deficiency (SIgAD) and common variable immunodeficiency (CVID). **Clin Exp Immunol** 120: 225–231.
- Haraksingh, R. R. and Snyder, M. P. (2013). Impacts of variation in the human genome on gene regulation. **J Mol Biol** 425: 3970–3977.
- Hastie, T., Tibshirani R. and Narasimhan, B., and Chu, G. (2013). pamr: Pam: prediction analysis for microarrays. R package version 1.54.1. URL: <http://CRAN.R-project.org/package=pamr>.
- Hawkins, R. D., Hon, G. C., and Ren, B. (2010). Next-generation genomics: an integrative approach. **Nat Rev Genet** 11: 476–486.
- Hecht, J. L. and Aster, J. C. (2000). Molecular biology of Burkitt's lymphoma. **J Clin Oncol** 18: 3707–3721.
- Himes, B. E., Jiang, X., Hu, R., Wu, A. C., Lasky-Su, J. A., Klanderman, B. J., Ziniti, J., Senter-Sylvia, J., Lima, J. J., Irvin, C. G., et al. (2012). Genome-wide association analysis in asthma subjects identifies SPATS2L as a novel bronchodilator response gene. **PLoS Genet** 8: e1002824.
- Hindorff, L. A., MacArthur, J., Morales, J., Junkins, H. A., Hall, P. N., Klemm, A. K., and Manolio, T. A. A Catalog of Published Genome-Wide Association Studies. URL: www.genome.gov/gwastudies (visited on 07/16/2014).
- Honda, K. and Taniguchi, T. (2006). IRFs: master regulators of signalling by Toll-like receptors and cytosolic pattern-recognition receptors. **Nat Rev Immunol** 6: 644–658.
- Hormozdiari, F., Alkan, C., Eichler, E. E., and Sahinalp, S. C. (2009). Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. **Genome Res** 19: 1270–1278.

BIBLIOGRAPHY

- Hotchkiss, R. S., Monneret, G., and Payen, D. (2013). Immunosuppression in sepsis: a novel understanding of the disorder and a new therapeutic approach. **Lancet Infect Dis** 13: 260–268.
- Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. **G3 (Bethesda)** 1: 457–470.
- Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. **PLoS Genet** 5: e1000529.
- Huang, Y., Zaas, A. K., Rao, A., Dobigeon, N., Woolf, P. J., Veldman, T., ien, N. C., McClain, M. T., Varkey, J. B., Nicholson, B., et al. (2011). Temporal dynamics of host molecular responses differentiate symptomatic and asymptomatic influenza a infection. **PLoS Genet** 7: e1002234.
- Huber, W., Heydebreck, A. von, Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. **Bioinformatics** 18 Suppl 1: S96–104.
- Hugot, J. P., Chamaillard, M., Zouali, H., Lesage, S., Cezard, J. P., Belaiche, J., Almer, S., Tysk, C., O’Morain, C. A., Gassull, M., et al. (2001). Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn’s disease. **Nature** 411: 599–603.
- Hunter, D. J. and Kraft, P. (2007). Drinking from the fire hose—statistical issues in genomewide association studies. **N Engl J Med** 357: 436–439.
- Husson, F., Josse, J., Le, S., and Mazet, J. (2013). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R.
- Idaghdour, Y. and Awadalla, P. (2012). Exploiting gene expression variation to capture gene-environment interactions for disease. **Front Genet** 3: 228.
- Idaghdour, Y., Quinlan, J., Goulet, J. P., Berghout, J., Gbeha, E., Bruat, V., Malliard, T. de, Grenier, J. C., Gomez, S., Gros, P., et al. (2012). Evidence for additive and interaction effects of host genotype and infection in malaria. **Proc Natl Acad Sci U S A** 109: 16786–16793.
- International HapMap Consortium (2003). The International HapMap Project. **Nature** 426: 789–796.
- International HapMap Consortium, Altshuler, D. M., Gibbs, R. A., Peltonen, L., Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., et al. (2010). Integrating common and rare genetic variation in diverse human populations. **Nature** 467: 52–58.
- Ioannidis, I., McNally, B., Willette, M., Peeples, M. E., Chaussabel, D., Durbin, J. E., Ramilo, O., Mejias, A., and Flano, E. (2012). Plasticity and virus specificity of the airway epithelial cell immune response during respiratory virus infection. **J Virol** 86: 5422–5436.
- Jackson, G. G., Dowling, H. F., Spiesman, I. G., and Boand, A. V. (1958). Transmission of the common cold to volunteers under controlled conditions. I. The common cold as a clinical entity. **AMA Arch Intern Med** 101: 267–278.

BIBLIOGRAPHY

- Jaenisch, R. and Bird, A. (2003). Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. **Nat Genet** 33 Suppl: 245–254.
- Jevremovic, D., Billadeau, D. D., Schoon, R. A., Dick, C. J., and Leibson, P. J. (2001). Regulation of NK cell-mediated cytotoxicity by the adaptor protein 3BP2. **J Immunol** 166: 7219–7228.
- Jin, W., Qin, P., Lou, H., Jin, L., and Xu, S. (2012). A systematic characterization of genes underlying both complex and Mendelian diseases. **Hum Mol Genet** 21: 1611–1624.
- Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Schumm, L. P., Sharma, Y., Anderson, C. A., et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. **Nature** 491: 119–124.
- Kanegane, H., Agematsu, K., Futatani, T., Sira, M. M., Suga, K., Sekiguchi, T., Zelm, M. C. van, and Miyawaki, T. (2007). Novel mutations in a Japanese patient with CD19 deficiency. **Genes Immun** 8: 663–670.
- Kashyap, R. S., Nayak, A. R., Deshpande, P. S., Kabra, D., Purohit, H. J., Taori, G. M., and Dagainawala, H. F. (2009). Inter-alpha-trypsin inhibitor heavy chain 4 is a novel marker of acute ischemic stroke. **Clin Chim Acta** 402: 160–163.
- Kawada, J., Kimura, H., Kamachi, Y., Nishikawa, K., Taniguchi, M., Nagaoka, K., Kurahashi, H., Kojima, S., and Morishima, T. (2006). Analysis of gene-expression profiles by oligonucleotide microarray in children with influenza. **J Gen Virol** 87: 1677–1683.
- Keller, M. D., Ganesh, J., Heltzer, M., Paessler, M., Bergqvist, A. G., Baluarte, H. J., Watkins, D., Rosenblatt, D. S., and Orange, J. S. (2013). Severe combined immunodeficiency resulting from mutations in MTHFD1. **Pediatrics** 131: e629–634.
- Klein, R. J., Zeiss, C., Chew, E. Y., Tsai, J. Y., Sackler, R. S., Haynes, C., Henning, A. K., SanGiovanni, J. P., Mane, S. M., Mayne, S. T., et al. (2005). Complement factor H polymorphism in age-related macular degeneration. **Science** 308: 385–389.
- Knight, J. C. (2012). Resolving the variable genome and epigenome in human disease. **J Intern Med** 271: 379–391.
- Knight, J. C. (2014). Approaches for establishing the function of regulatory genetic variants involved in disease. **Genome Med** 6: 92.
- Kooperberg, C., LeBlanc, M., and Obenchain, V. (2010). Risk prediction using genome-wide association studies. **Genet Epidemiol** 34: 643–652.
- Krausgruber, T., Blazek, K., Smallie, T., Alzabin, S., Lockstone, H., Sahgal, N., Hussell, T., Feldmann, M., and Udalova, I. A. (2011). IRF5 promotes inflammatory macrophage polarization and TH1-TH17 responses. **Nat Immunol** 12: 231–238.

BIBLIOGRAPHY

- Ku, C. S., Loy, E. Y., Salim, A., Pawitan, Y., and Chia, K. S. (2010). The discovery of human genetic variations and their use as disease markers: past, present and future. **J Hum Genet** 55: 403–415.
- Kuijpers, T. W., Bende, R. J., Baars, P. A., Grummels, A., Derks, I. A., Dolman, K. M., Beaumont, T., Tedder, T. F., Noesel, C. J. van, Eldering, E., et al. (2010). CD20 deficiency in humans results in impaired T cell-independent antibody responses. **J Clin Invest** 120: 214–222.
- Kumar, P., Henikoff, S., and Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. **Nat Protoc** 4: 1073–1081.
- Liang, L., Morar, N., Dixon, A. L., Lathrop, G. M., Abecasis, G. R., Moffatt, M. F., and Cookson, W. O. (2013). A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. **Genome Res** 23: 716–726.
- Lilic, D (2009). Immune response to infection. **Anaesthesia & Intensive Care Medicine** 10: 218–220.
- Lillie, P. J., Berthoud, T. K., Powell, T. J., Lambe, T., Mullarkey, C., Spencer, A. J., Hamill, M., Peng, Y., Blais, M. E., Duncan, C. J., et al. (2012). Preliminary assessment of the efficacy of a T-cell-based influenza vaccine, MVA-NP+M1, in humans. **Clin Infect Dis** 55: 19–25.
- Liu, H., Pecka, J. L., Zhang, Q., Soukup, G. A., Beisel, K. W., and He, D. Z. (2014). Characterization of transcriptomes of cochlear inner and outer hair cells. **J Neurosci** 34: 11085–11095.
- Liu, L., Okada, S., Kong, X. F., Kreins, A. Y., Cypowyj, S., Abhyankar, A., Toubiana, J., Itan, Y., Audry, M., Nitschke, P., et al. (2011). Gain-of-function human STAT1 mutations impair IL-17 immunity and underlie chronic mucocutaneous candidiasis. **J Exp Med** 208: 1635–1648.
- Liu, X., Yu, X., Zack, D. J., Zhu, H., and Qian, J. (2008). TiGER: a database for tissue-specific gene expression and regulation. **BMC Bioinformatics** 9: 271.
- Llewelyn, M., Cohen, J., and International Sepsis, Forum (2001). Diagnosis of infection in sepsis. **Intensive Care Med** 27 Suppl 1: S10–32.
- Lopez-Herrera, G., Tampella, G., Pan-Hammarstrom, Q., Herholz, P., Trujillo-Vargas, C. M., Phadwal, K., Simon, A. K., Moutschen, M., Etzioni, A., Mory, A., et al. (2012). Deleterious mutations in LRBA are associated with a syndrome of immune deficiency and autoimmunity. **Am J Hum Genet** 90: 986–1001.
- Losi, C. G., Silini, A., Fiorini, C., Soresina, A., Meini, A., Ferrari, S., Notarangelo, L. D., Lougaris, V., and Plebani, A. (2005). Mutational analysis of human BAFF receptor TNFRSF13C (BAFF-R) in patients with common variable immunodeficiency. **J Clin Immunol** 25: 496–502.

BIBLIOGRAPHY

- Lucas, C. L., Kuehn, H. S., Zhao, F., Niemela, J. E., Deenick, E. K., Palendira, U., Avery, D. T., Moens, L., Cannons, J. L., Biancalana, M., et al. (2014). Dominant-activating germline mutations in the gene encoding the PI(3)K catalytic subunit p110delta result in T cell senescence and human immunodeficiency. **Nat Immunol** 15: 88–97.
- Luckheeram, R. V., Zhou, R., Verma, A. D., and Xia, B. (2012). CD4(+)T cells: differentiation and functions. **Clin Dev Immunol** 2012: 925135.
- Lunter, G. and Goodson, M. (2011). Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. **Genome Res** 21: 936–939.
- Maher, B. (2008). Personal genomes: The case of the missing heritability. **Nature** 456: 18–21.
- Mallal, S., Nolan, D., Witt, C., Masel, G., Martin, A. M., Moore, C., Sayer, D., Castley, A., Mamotte, C., Maxwell, D., et al. (2002). Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. **Lancet** 359: 727–732.
- Manchia, M., Cullis, J., Turecki, G., Rouleau, G. A., Uher, R., and Alda, M. (2013). The impact of phenotypic and genetic heterogeneity on results of genome wide association studies of complex diseases. **PLoS One** 8: e76295.
- Manolio, T. A. (2010). Genomewide association studies and assessment of the risk of disease. **N Engl J Med** 363: 166–176.
- Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. **Nature** 461: 747–753.
- Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. **Nat Genet** 39: 906–913.
- Marti-Carvajal, A. J., Sola, I., Gluud, C., Lathyris, D., and Cardona, A. F. (2012). Human recombinant protein C for severe sepsis and septic shock in adult and paediatric patients. **Cochrane Database Syst Rev** 12: CD004388.
- Martin, G. S., Mannino, D. M., Eaton, S., and Moss, M. (2003). The epidemiology of sepsis in the United States from 1979 through 2000. **N Engl J Med** 348: 1546–1554.
- Martin, H. C., Kim, G. E., Pagnamenta, A. T., Murakami, Y., Carvill, G. L., Meyer, E., Copley, R. R., Rimmer, A., Barcia, G., Fleming, M. R., et al. (2014). Clinical whole-genome sequencing in severe early-onset epilepsy reveals new genes and improves molecular diagnosis. **Hum Mol Genet** 23: 3200–3211.
- Maslove, D. M. and Wong, H. R. (2014). Gene expression profiling in sepsis: timing, tissue, and translational considerations. **Trends Mol Med** 20: 204–213.

BIBLIOGRAPHY

- Mathelier, A., Zhao, X., Zhang, A. W., Parcy, F., Worsley-Hunt, R., Arenillas, D. J., Buchman, S., Chen, C. Y., Chou, A., Ienasescu, H., et al. (2014). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. **Nucleic Acids Res** 42: D142–147.
- Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic localization of common disease-associated variation in regulatory DNA. **Science** 337: 1190–1195.
- McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. **Nat Rev Genet** 9: 356–369.
- McVicker, G., Geijn, B. van de, Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., Lewellen, N., Myrthil, M., Gilad, Y., and Pritchard, J. K. (2013). Identification of genetic variants that affect histone modifications in human cells. **Science** 342: 747–749.
- Metersky, M. L., Masterton, R. G., Lode, H., File T. M., Jr., and Babinchak, T. (2012). Epidemiology, microbiology, and treatment considerations for bacterial pneumonia complicating influenza. **Int J Infect Dis** 16 (5): e321–331.
- Michaelson, J. J., Loguercio, S., and Beyer, A. (2009). Detection and interpretation of expression quantitative trait loci (eQTL). **Methods** 48: 265–276.
- Milner, J. D. and Holland, S. M. (2013). The cup runneth over: lessons from the ever-expanding pool of primary immunodeficiency diseases. **Nat Rev Immunol** 13: 635–648.
- Miyagawa, M., Nishio, S. Y., and Usami, S. (2012). Prevalence and clinical features of hearing loss patients with CDH23 mutations: a large cohort study. **PLoS One** 7: e40366.
- Moffatt, M. F., Kabisch, M., Liang, L., Dixon, A. L., Strachan, D., Heath, S., Depner, M., Berg, A. von, Bufe, A., Rietschel, E., et al. (2007). Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. **Nature** 448: 470–473.
- Montfrans, J. M. van, Hoepelman, A. I., Otto, S., Gijn, M. van, Corput, L. van de, Weger, R. A. de, Monaco-Shawver, L., Banerjee, P. P., Sanders, E. A., Zijde, C. M. Jol-van der, et al. (2012). CD27 deficiency is associated with combined immunodeficiency and persistent symptomatic EBV viremia. **J Allergy Clin Immunol** 129: 787–793e786.
- Morley, Michael, Molony, Cliona M., Weber, Teresa M., Devlin, James L., Ewens, Kathryn G., Spielman, Richard S., and Cheung, Vivian G. (2004). Genetic analysis of genome-wide variation in human gene expression. **Nature** 430: 743–747.

BIBLIOGRAPHY

- Mueller, C. G., Rissoan, M. C., Salinas, B., Ait-Yahia, S., Ravel, O., Bridon, J. M., Briere, F., Lebecque, S., and Liu, Y. J. (1997). Polymerase chain reaction selects a novel disintegrin proteinase from CD40-activated germinal center dendritic cells. **J Exp Med** 186: 655–663.
- Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., Travers, M., Potter, S., Grundberg, E., Small, K., et al. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. **PLoS Genet** 7: e1002003.
- Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., and Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. **PLoS Genet** 6: e1000888.
- Obel, N., Christensen, K., Petersen, I., Sorensen, T. I., and Skytthe, A. (2010). Genetic and environmental influences on risk of death due to infections assessed in Danish twins, 1943–2001. **Am J Epidemiol** 171: 1007–1013.
- Offer, S. M., Pan-Hammarstrom, Q., Hammarstrom, L., and Harris, R. S. (2010). Unique DNA repair gene variations and potential associations with the primary antibody deficiency syndromes IgAD and CVID. **PLoS One** 5: e12260.
- Olerup, O., Smith, C. I., Bjorkander, J., and Hammarstrom, L. (1992). Shared HLA class II-associated genetic susceptibility and resistance, related to the HLA-DQB1 gene, in IgA deficiency and common variable immunodeficiency. **Proc Natl Acad Sci U S A** 89: 10653–10657.
- Ombrello, M. J., Remmers, E. F., Sun, G., Freeman, A. F., Datta, S., Torabi-Parizi, P., Subramanian, N., Bunney, T. D., Baxendale, R. W., Martins, M. S., et al. (2012). Cold urticaria, immunodeficiency, and autoimmunity related to PLCG2 deletions. **N Engl J Med** 366: 330–338.
- Opal, S. M., Laterre, P. F., Francois, B., LaRosa, S. P., Angus, D. C., Mira, J. P., Wittebole, X., Dugernier, T., Perrotin, D., Tidswell, M., et al. (2013). Effect of eritoran, an antagonist of MD2-TLR4, on mortality in patients with severe sepsis: the ACCESS randomized trial. **JAMA** 309: 1154–1162.
- Orange, J. S., Glessner, J. T., Resnick, E., Sullivan, K. E., Lucas, M., Ferry, B., Kim, C. E., Hou, C., Wang, F., Chiavacci, R., et al. (2011). Genome-wide association identifies diverse causes of common variable immunodeficiency. **J Allergy Clin Immunol** 127: 1360–1367e1366.
- Oshlack, A., Robinson, M. D., and Young, M. D. (2010). From RNA-seq reads to differential expression results. **Genome Biol** 11: 220.
- Packwood, K., Drewe, E., Staples, E., Webster, D., Witte, T., Litzman, J., Egner, W., Sargur, R., Sewell, W., Lopez-Granados, E., et al. (2010). NOD2 polymorphisms in clinical phenotypes of common variable immunodeficiency disorders. **Clin Exp Immunol** 161: 536–41.
- Padkin, A., Goldfrad, C., Brady, A. R., Young, D., Black, N., and Rowan, K. (2003). Epidemiology of severe sepsis occurring in the first 24 hrs in intensive care units in England, Wales, and Northern Ireland. **Crit Care Med** 31: 2332–2338.

BIBLIOGRAPHY

- Pan-Hammarstrom, Q., Salzer, U., Du, L., Bjorkander, J., Cunningham-Rundles, C., Nelson, D. L., Bacchelli, C., Gaspar, H. B., Offer, S., Behrens, T. W., et al. (2007). Reexamining the role of TACI coding variants in common variable immunodeficiency and selective IgA deficiency. **Nat Genet** 39: 429–430.
- Park, M. A., Li, J. T., Hagan, J. B., Maddox, D. E., and Abraham, R. S. (2008). Common variable immunodeficiency: a new look at an old disease. **Lancet** 372: 489–502.
- Pauwels, M., Weyenbergh, J. van, Soumillon, A., Proost, P., and De Ley, M. (1994). Induction by zinc of specific metallothionein isoforms in human monocytes. **Eur J Biochem** 220: 105–110.
- Payen, D., Lukaszewicz, A. C., Belikova, I., Faivre, V., Gelin, C., Russwurm, S., Launay, J. M., and Sevenet, N. (2008). Gene profiling in human blood leucocytes during recovery from septic shock. **Intensive Care Med** 34: 1371–1376.
- Phua, J., Ngerng, W., See, K., Tay, C., Kiong, T., Lim, H., Chew, M., Yip, H., Tan, A., Khalizah, H., et al. (2013). Characteristics and outcomes of culture-negative versus culture-positive severe sepsis. **Crit Care** 17: R202.
- Poodt, A. E., Driessen, G. J., Klein, A. de, Dongen, J. J. van, Burg, M. van der, and Vries, E. de (2009). TACI mutations and disease susceptibility in patients with common variable immunodeficiency. **Clin Exp Immunol** 156: 35–39.
- Pruitt, K. D., Brown, G. R., Hiatt, S. M., Thibaud-Nissen, F., Astashyn, A., Ermolaeva, O., Farrell, C. M., Hart, J., Landrum, M. J., McGarvey, K. M., et al. (2014). RefSeq: an update on mammalian reference sequences. **Nucleic Acids Res** 42: D756–D763.
- Puck, J. M., Deschenes, S. M., Porter, J. C., Dutra, A. S., Brown, C. J., Willard, H. F., and Henthorn, P. S. (1993). The interleukin-2 receptor gamma chain maps to Xq13.1 and is mutated in X-linked severe combined immunodeficiency, SCIDX1. **Hum Mol Genet** 2: 1099–1104.
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., Bakker, P. I. de, Daly, M. J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. **Am J Hum Genet** 81: 559–575.
- R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL: <http://www.R-project.org/>.
- Radhakrishnan, J. (2012). Functional Genomics of Severe Sepsis and Septic Shock. PhD thesis. University of Oxford.
- Ramilo, O., Allman, W., Chung, W., Mejias, A., Ardura, M., Glaser, C., Wittkowski, K. M., Piqueras, B., Banchereau, J., Palucka, A. K., et al. (2007). Gene expression patterns in blood leukocytes discriminate patients with acute infections. **Blood** 109: 2066–2077.

BIBLIOGRAPHY

- Rautanen, A., Mills, T. C., Gordon, A. C., Hutton, P., Steffens, M., Nuamah, R., Chiche, J-D., Parks, T., Chapman, S. J., Davenport, E. E., et al. (2015). Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study. **Lancet Respiratory Medicine** 3: 53–60.
- Reich, D. E. and Lander, E. S. (2001). On the allelic spectrum of human disease. **Trends Genet** 17: 502–510.
- Reumers, J., De Rijk, P., Zhao, H., Liekens, A., Smeets, D., Cleary, J., Van Loo, P., Van Den Bossche, M., Catthoor, K., Sabbe, B., et al. (2012). Optimized filtering reduces the error rate in detecting genomic variants by short-read sequencing. **Nat Biotechnol** 30: 61–68.
- Richards, E. J. (2006). Inherited epigenetic variation—revisiting soft inheritance. **Nat Rev Genet** 7: 395–401.
- Rimmer, A., Hang, P., Mathieson, I., Iqbal, Z., Twigg, S. R. F., Consortium, WGS500, Wilkie, A. O. M., McVean, G., and Lunter, G. (2014). Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. **Nat Genet** 46: 912–918.
- Rivas, M. A., Beaudoin, M., Gardet, A., Stevens, C., Sharma, Y., Zhang, C. K., Boucher, G., Ripke, S., Ellinghaus, D., Burt, N., et al. (2011). Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. **Nat Genet** 43: 1066–1073.
- Rivera, C. M. and Ren, B. (2013). Mapping human epigenomes. **Cell** 155: 39–55.
- Robinson, M. D. and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. **Genome Biol** 11: R25.
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. **Bioinformatics** 26: 139–140.
- Robson, J. D., Davidson, D., and Veillette, A. (2004). Inhibition of the Jun N-terminal protein kinase pathway by SHIP-1, a lipid phosphatase that interacts with the adaptor molecule Dok-3. **Mol Cell Biol** 24: 2332–2343.
- Russell, J. A., Walley, K. R., Singer, J., Gordon, A. C., Hebert, P. C., Cooper, D. J., Holmes, C. L., Mehta, S., Granton, J. T., Storms, M. M., et al. (2008). Vasopressin versus norepinephrine infusion in patients with septic shock. **N Engl J Med** 358: 877–887.
- Ryan, J., Zoellner, Y., Gradl, B., Palache, B., and Medema, J. (2006). Establishing the health and economic impact of influenza vaccination within the European Union 25 countries. **Vaccine** 24: 6812–6822.
- Sadler, A. J. and Williams, B. R. (2008). Interferon-inducible antiviral effectors. **Nat Rev Immunol** 8: 559–568.
- Saint Basile, G. de, Arveiler, B., Oberle, I., Malcolm, S., Levinsky, R. J., Lau, Y. L., Hofker, M., Debre, M., Fischer, A., Griscelli, C., et al. (1987). Close linkage of the locus for X chromosome-linked severe combined immunodeficiency to polymorphic DNA markers in Xq11-q13. **Proc Natl Acad Sci U S A** 84: 7576–7579.

BIBLIOGRAPHY

- Saint-Vis, B. de, Vincent, J., Vandenabeele, S., Vanbervliet, B., Pin, J. J., Ait-Yahia, S., Patel, S., Mattei, M. G., Banchereau, J., Zurawski, S., et al. (1998). A novel lysosome-associated membrane glycoprotein, DC-LAMP, induced upon DC maturation, is transiently expressed in MHC class II compartment. **Immunity** 9: 325–336.
- Salipante, S. J., Rojas, M. E., Korkmaz, B., Duan, Z., Wechsler, J., Benson, K. F., Person, R. E., Grimes, H. L., and Horwitz, M. S. (2009). Contributions to neutropenia from PFAAP5 (N4BP2L2), a novel protein mediating transcriptional repressor cooperation between Gfi1 and neutrophil elastase. **Mol Cell Biol** 29: 4394–405.
- Salzer, U., Maul-Pavicic, A., Cunningham-Rundles, C., Urschel, S., Belohradsky, B. H., Litzman, J., Holm, A., Franco, J. L., Plebani, A., Hammarstrom, L., et al. (2004). ICOS deficiency in patients with common variable immunodeficiency. **Clin Immunol** 113: 234–240.
- Salzer, U., Chapel, H. M., Webster, A. D., Pan-Hammarström, Q., Schmitt-Graeff, A., Schlesier, M., Peter, H. H., Rockstroh, J. K., Schneider, P., Schiffer, A. A., et al. (2005). Mutations in TNFRSF13B encoding TACI are associated with common variable immunodeficiency in humans. **Nat Genet** 37: 820–828.
- Salzer, U., Neumann, C., Thiel, J., Woellner, C., Pan-Hammarström, Q., Lougaris, V., Hagena, T., Jung, J., Birmelin, J., Du, L., et al. (2008). Screening of functional and positional candidate genes in families with common variable immunodeficiency. **BMC Immunol** 9: 3.
- Salzer, U., Bacchelli, C., Buckridge, S., Pan-Hammarström, Q., Jennings, S., Lougaris, V., Bergbreiter, A., Hagena, T., Birmelin, J., Plebani, A., et al. (2009). Relevance of biallelic versus monoallelic TNFRSF13B mutations in distinguishing disease-causing from risk-increasing TNFRSF13B variants in antibody deficiency syndromes. **Blood** 113: 1967–76.
- Salzer, U., Warnatz, K., and Peter, H. H. (2012). Common variable immunodeficiency - an update. **Arthritis Res Ther** 14: 223.
- Sands, K. E., Bates, D. W., Lanken, P. N., Graman, P. S., Hibberd, P. L., Kahn, K. L., Parsonnet, J., Panzer, R., Orav, E. J., Snyderman, D. R., et al. (1997). Epidemiology of sepsis syndrome in 8 academic medical centers. **JAMA** 278: 234–240.
- Schadt, Eric E., Monks, Stephanie A., Drake, Thomas A., Lusi, Aldons J., Che, Nam, Colinayo, Veronica, Ruff, Thomas G., Milligan, Stephen B., Lamb, John R., Cavet, Guy, et al. (2003). Genetics of gene expression surveyed in maize, mouse and man. **Nature** 422: 297–302.
- Schork, N. J., Murray, S. S., Frazer, K. A., and Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. **Curr Opin Genet Dev** 19: 212–219.
- Scott-Taylor, T. H., Green, M. R., Eren, E., and Webster, A. D. (2004). Monocyte derived dendritic cell responses in common variable immunodeficiency. **Clin Exp Immunol** 138: 484–490.

BIBLIOGRAPHY

- Severino, P., Silva, E., Baggio-Zappia, G. L., Brunialti, M. K., Nucci, L. A., Rigato O., Jr., Silva, I. D. da, Machado, F. R., and Salomao, R. (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. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. **Bioinformatics** 28: 1353–1358.
- Silvestre, J., Povia, P., Coelho, L., Almeida, E., Moreira, P., Fernandes, A., Mealha, R., and Sabino, H. (2009). Is C-reactive protein a good prognostic marker in septic patients? **Intensive Care Med** 35: 909–913.
- Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCr: visualizing classifier performance in R. **Bioinformatics** 21: 3940–3941.
- Slatkin, M. (2008). Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. **Nat Rev Genet** 9: 477–485.
- Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. **Stat Appl Genet Mol Biol** 3: Article3.
- Sørensen, T. I., Nielsen, G. G., Andersen, P. K., and Teasdale, T. W. (1988). Genetic and environmental influences on premature death in adult adoptees. **N Engl J Med** 318: 727–732.
- Sriskandan, S. and Cohen, J. (1999). Gram-positive sepsis. Mechanisms and differences from gram-negative sepsis. **Infect Dis Clin North Am** 13: 397–412.
- Stepensky, P., Keller, B., Buchta, M., Kienzler, A. K., Elpeleg, O., Somech, R., Cohen, S., Shachar, I., Miosge, L. A., Schlesier, M., et al. (2013). Deficiency of caspase recruitment domain family, member 11 (CARD11), causes profound combined immunodeficiency in human subjects. **J Allergy Clin Immunol** 131: 477–485.
- Stranger, B. E. and Dermitzakis, E. T. (2005). The genetics of regulatory variation in the human genome. **Hum Genomics** 2: 126–131.
- Stranger, B. E., Montgomery, S. B., Dimas, A. S., Parts, L., Stegle, O., Ingle, C. E., Sekowska, M., Smith, G. D., Evans, D., Gutierrez-Arcelus, M., et al. (2012). Patterns of cis regulatory variation in diverse human populations. **PLoS Genet** 8: e1002639.
- Tai, Y. C. (2007). timecourse: Statistical Analysis for Developmental Microarray Time Course Data. R package version 1.28.0. URL: [http : / / www . bioconductor . org](http://www.bioconductor.org).
- Takeuchi, O. and Akira, S. (2001). Toll-like receptors; their physiological role and signal transduction system. **Int Immunopharmacol** 1: 625–635.
- Tang, B. M., McLean, A. S., Dawes, I. W., Huang, S. J., and Lin, R. C. (2007). The use of gene-expression profiling to identify candidate genes in human sepsis. **Am J Respir Crit Care Med** 176: 676–684.

BIBLIOGRAPHY

- Tang, B. M., McLean, A. S., Dawes, I. W., Huang, S. J., Cowley, M. J., and Lin, R. C. (2008). Gene-expression profiling of gram-positive and gram-negative sepsis in critically ill patients. **Crit Care Med** 36: 1125–1128.
- Tang, B. M., McLean, A. S., Dawes, I. W., Huang, S. J., and Lin, R. C. (2009). Gene-expression profiling of peripheral blood mononuclear cells in sepsis. **Crit Care Med** 37: 882–888.
- Tang, B. M., Huang, S. J., and McLean, A. S. (2010). Genome-wide transcription profiling of human sepsis: a systematic review. **Crit Care** 14: R237.
- Tanner, M. A., Berk, L. S., Felten, D. L., Blidy, A. D., Bit, S. L., and Ruff, D. W. (2002). Substantial changes in gene expression level due to the storage temperature and storage duration of human whole blood. **Clin Lab Haematol** 24: 337–341.
- Thiel, J., Kimmig, L., Salzer, U., Grudzien, M., Lebrecht, D., Hagen, T., Draeger, R., Voelxen, N., Bergbreiter, A., Jennings, S., et al. (2012). Genetic CD21 deficiency is associated with hypogammaglobulinemia. **J Allergy Clin Immunol** 129: 801–810.
- Thomas, D. L., Thio, C. L., Martin, M. P., Qi, Y., Ge, D., O’Huigin, C., Kidd, J., Kidd, K., Khakoo, S. I., Alexander, G., et al. (2009). Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. **Nature** 461: 798–801.
- Todd, J. A. (2006). Statistical false positive or true disease pathway? **Nat Genet** 38: 731–733.
- Trowsdale, J. and Knight, J. C. (2013). Major histocompatibility complex genomics and human disease. **Annu Rev Genomics Hum Genet** 14: 301–323.
- van ’t Veer, L. J., Dai, H., Vijver, M. J. van de, He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., Kooy, K. van der, Marton, M. J., Witteveen, A. T., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. **Nature** 415: 530–536.
- van Zelm, M. C., Reisli, I., Burg, M. van der, Castao, D., Noesel, C. J. van, Tol, M. J. van, Woellner, C., Grimbacher, B., Patio, P. J., Dongen, J. J. van, et al. (2006). An antibody-deficiency syndrome due to mutations in the CD19 gene. **N Engl J Med** 354: 1901–1912.
- Veerdonk, F. L. van de, Plantinga, T. S., Hoischen, A., Smeekens, S. P., Joosten, L. A., Gilissen, C., Arts, P., Rosentul, D. C., Carmichael, A. J., Graaf, C. A. Smits-van der, et al. (2011). STAT1 mutations in autosomal dominant chronic mucocutaneous candidiasis. **N Engl J Med** 365: 54–61.
- Veltman, J. A. and Brunner, H. G. (2012). De novo mutations in human genetic disease. **Nat Rev Genet** 13: 565–575.
- Vestal, D. J. (2005). The guanylate-binding proteins (GBPs): proinflammatory cytokine-induced members of the dynamin superfamily with unique GTPase activity. **J Interferon Cytokine Res** 25: 435–443.
- Veyrieras, J. B., Kudaravalli, S., Kim, S. Y., Dermitzakis, E. T., Gilad, Y., Stephens, M., and Pritchard, J. K. (2008). High-resolution mapping of expression-QTLs yields insight into human gene regulation. **PLoS Genet** 4: e1000214.

BIBLIOGRAPHY

- Vincent, J. L. (2004). Endpoints in sepsis trials: more than just 28-day mortality? **Crit Care Med** 32: S209–213.
- Vincent, J-L. and Moreno, R. (2010). Clinical review: Scoring systems in the critically ill. **Crit Care** 14: 207.
- Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five Years of GWAS Discovery. **Am J Hum Genet** 90: 7–24.
- Volanakis, J. E., Zhu, Z. B., Schaffer, F. M., Macon, K. J., Palermos, J., Barger, B. O., Go, R., Campbell, R. D., Schroeder H. W., Jr., and Cooper, M. D. (1992). Major histocompatibility complex class III genes and susceptibility to immunoglobulin A deficiency and common variable immunodeficiency. **J Clin Invest** 89: 1914–1922.
- Vorechovsky, I., Webster, A. D., Plebani, A., and Hammarstrom, L. (1999). Genetic linkage of IgA deficiency to the major histocompatibility complex: evidence for allele segregation distortion, parent-of-origin penetrance differences, and the role of anti-IgA antibodies in disease predisposition. **Am J Hum Genet** 64: 1096–109.
- Wang, J., Nikrad, M. P., Travanty, E. A., Zhou, B., Phang, T., Gao, B., Alford, T., Ito, Y., Nahreini, P., Hartshorn, K., et al. (2012). Innate immune response of human alveolar macrophages during influenza A infection. **PLoS One** 7: e29879.
- Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. **Nat Rev Genet** 11: 843–854.
- Wang, L., Jia, P., Wolfinger, R. D., Chen, X., and Zhao, Z. (2011). Gene set analysis of genome-wide association studies: methodological issues and perspectives. **Genomics** 98: 1–8.
- Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. **Nat Rev Genet** 10: 57–63.
- Ward, P. A. (2010). The harmful role of c5a on innate immunity in sepsis. **J Innate Immun** 2: 439–445.
- Warnatz, K., Salzer, U., Rizzi, M., Fischer, B., Gutenberger, S., Bohm, J., Kienzler, A. K., Pan-Hammarstrom, Q., Hammarstrom, L., Rakhmanov, M., et al. (2009). B-cell activating factor receptor deficiency is associated with an adult-onset antibody deficiency syndrome in humans. **Proc Natl Acad Sci U S A** 106: 13945–13950.
- Weiss, M., Blazek, K., Byrne, A. J., Perocheau, D. P., and Udalova, I. A. (2013). IRF5 is a specific marker of inflammatory macrophages in vivo. **Mediators Inflamm** 2013: 245804.
- Westra, H. J. and Franke, L. (2014). From genome to function by studying eQTLs. **Biochim Biophys Acta**.
- Westra, H. J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. **Nat Genet** 45: 1238–1243.

BIBLIOGRAPHY

- Westra, H-J, Arends, D., Esko, T., Peters, M. J., Schurmann, C., Schramm, K., Kettunen, J., Yaghootkar, H., Fairfax, B., Andiappan, A. K., et al. (2014). Cell specific eQTL analysis without sorting cells. **bioRxiv**. doi: 10.1101/002600.
- Wiersinga, W. J. (2011). Current insights in sepsis: from pathogenesis to new treatment targets. **Curr Opin Crit Care** 17: 480–486.
- Wilkerson, M. D., Cabanski, C. R., Sun, W., Hoadley, K. A., Walter, V., Mose, L. E., Troester, M. A., Hammerman, P. S., Parker, J. S., Perou, C. M., et al. (2014). Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. **Nucleic Acids Res** 42: e107.
- Williams, A., Spilianakis, C. G., and Flavell, R. A. (2010). Interchromosomal association and gene regulation in trans. **Trends Genet** 26: 188–197.
- Wong, H. R., Shanley, T. P., Sakthivel, B., Cvijanovich, N., Lin, R., Allen, G. L., Thomas, N. J., Doctor, A., Kalyanaraman, M., Tofil, N. M., et al. (2007). Genome-level expression profiles in pediatric septic shock indicate a role for altered zinc homeostasis in poor outcome. **Physiol Genomics** 30: 146–155.
- Wong, H. R., Cvijanovich, N., Allen, G. L., Lin, R., Anas, N., Meyer, K., Freishtat, R. J., Monaco, M., Odoms, K., Sakthivel, B., et al. (2009a). Genomic expression profiling across the pediatric systemic inflammatory response syndrome, sepsis, and septic shock spectrum. **Crit Care Med** 37: 1558–1566.
- Wong, H. R., Cvijanovich, N., Lin, R., Allen, G. L., Thomas, N. J., Willson, D. F., Freishtat, R. J., Anas, N., Meyer, K., Checchia, P. A., et al. (2009b). Identification of pediatric septic shock subclasses based on genome-wide expression profiling. **BMC Med** 7: 34.
- Wong, H. R., Freishtat, R. J., Monaco, M., Odoms, K., and Shanley, T. P. (2010a). Leukocyte subset-derived genomewide expression profiles in pediatric septic shock. **Pediatr Crit Care Med** 11: 349–355.
- Wong, H. R., Wheeler, D. S., Tegtmeier, K., Poynter, S. E., Kaplan, J. M., Chima, R. S., Stalets, E., Basu, R. K., and Doughty, L. A. (2010b). Toward a clinically feasible gene expression-based subclassification strategy for septic shock: proof of concept. **Crit Care Med** 38: 1955–1961.
- Wong, H. R., Cvijanovich, N. Z., Allen, G. L., Thomas, N. J., Freishtat, R. J., Anas, N., Meyer, K., Checchia, P. A., Lin, R., Shanley, T. P., et al. (2011). Validation of a gene expression-based subclassification strategy for pediatric septic shock. **Crit Care Med** 39: 2511–2517.
- Woods, N. T., Mesquita, R. D., Sweet, M., Carvalho, M. A., Li, X., Liu, Y., Nguyen, H., Thomas, C. E., Iversen E. S., Jr., Marsillac, S., et al. (2012). Charting the landscape of tandem BRCT domain-mediated protein interactions. **Sci Signal** 5: rs6.
- World Health Organization (2013). Influenza. URL: <http://www.euro.who.int/en/what-we-do/health-topics/communicable-diseases/influenza/facts-and-figures> (visited on 11/06/2013).
- World Health Organization (2014a). Antimicrobial resistance. URL: <http://www.who.int/mediacentre/factsheets/fs194/en/> (visited on 10/01/2014).

BIBLIOGRAPHY

- World Health Organization (2014b). The top 10 causes of death. URL: <http://www.who.int/mediacentre/factsheets/fs310/en/index2.html> (visited on 10/04/2014).
- Yang, C. S., Chang, K. Y., and Rana, T. M. (2014). Genome-wide Functional Analysis Reveals Factors Needed at the Transition Steps of Induced Reprogramming. **Cell Rep** 8: 327–337.
- Yang, S., Lim, Y. P., Zhou, M., Salvemini, P., Schwinn, H., Josic, D., Koo, D. J., Chaudry, I. H., and Wang, P. (2002). Administration of human inter-alpha-inhibitors maintains hemodynamic stability and improves survival during sepsis. **Crit Care Med** 30: 617–622.
- Yang, X. (2012). Use of functional genomics to identify candidate genes underlying human genetic association studies of vascular diseases. **Arterioscler Thromb Vasc Biol** 32: 216–22.
- Yazdani, R., Hakemi, M. G., Sherkat, R., Homayouni, V., and Farahani, R. (2014). Genetic defects and the role of helper T-cells in the pathogenesis of common variable immunodeficiency. **Adv Biomed Res** 3: 2.
- Zaas, A. K., Chen, M., Varkey, J., Veldman, T., Hero A. O., 3rd, Lucas, J., Huang, Y., Turner, R., Gilbert, A., Lambkin-Williams, R., et al. (2009). Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. **Cell Host Microbe** 6: 207–217.
- Zak, D. E., Andersen-Nissen, E., Peterson, E. R., Sato, A., Hamilton, M. K., Borgerding, J., Krishnamurty, A. T., Chang, J. T., Adams, D. J., Hensley, T. R., et al. (2012). Merck Ad5/HIV induces broad innate immune activation that predicts CD8(+) T-cell responses but is attenuated by preexisting Ad5 immunity. **Proc Natl Acad Sci U S A** 109: E3503–3512.
- Zeni, F., Freeman, B., and Natanson, C. (1997). Anti-inflammatory therapies to treat sepsis and septic shock: a reassessment. **Crit Care Med** 25: 1095–1100.
- Zhernakova, A., Diemen, C. C. van, and Wijmenga, C. (2009). Detecting shared pathogenesis from the shared genetics of immune-related diseases. **Nat Rev Genet** 10: 43–55.
- Zhou, Z., Xue, Q., Wan, Y., Yang, Y., Wang, J., and Hung, T. (2011). Lysosome-associated membrane glycoprotein 3 is involved in influenza A virus replication in human lung epithelial (A549) cells. **Virol J** 8: 384.