Functional Genomics of Severe Sepsis and Septic Shock

Jayachandran Radhakrishnan

Christ Church

University of Oxford

Hilary Term 2012

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

an unexamined life is not worth living

dedicated to the memory of my father

Abstract

Sepsis is the systemic inflammatory response to an infection. Severe sepsis with multi organ failure is one of the commonest causes of admission to intensive care units, and is associated with poor early and late outcomes. The pathophysiology of sepsis is complex, and poorly understood. This is reflected in the limited and contentious treatment options for sepsis.

Genetic factors have been shown to be associated with the risk of and subsequent outcomes from infection. However, clear associations with bacterial sepsis are rare, and even when associations are present their functional effects are often unknown.

Gene expression signatures in sepsis are investigated in this project using serial samples obtained from patients admitted to intensive care units with community-acquired pneumonia or faecal peritonitis.

The evolving gene expression signatures that define the response to sepsis were identified with large changes seen in genes coding for ribosomal proteins *RPS4Y1* and *RPS26P54*. The differences in the sepsis response between the two diagnostic classes were examined. The gene expression predictors of mortality in sepsis were determined and include genes from the class II MHC *HLA-DRB4*, *HLA-DRB5* and the T cell differentiation protein *MAL*. The effects of important covariates on gene expression were investigated and their impact on survival related expression determined. The findings were confirmed in a validation cohort. A novel clustering of samples representing distinct inflammatory patterns in a clinically homogeneous population of sepsis patients was identified and related to differences in clinical behaviour. The biological relevance of the differentially expressed genes was ascertained by identifying enriched gene sets.

The gene expression changes in sepsis were examined in the context of related clinically relevant immune phenomena: the sterile systemic inflammatory response in patients undergoing elective cardiac surgery and the phenomenon of endotoxin tolerance in PBMCs derived from healthy volunteers.

The results highlight the complexities of clinical sepsis and identify hypotheses for future investigations.

Acknowledgements

The work presented in this thesis is the result of investigations conducted at the Wellcome Trust Centre for Human Genetics in Oxford, from 2006 to 2010.

I wish to thank Prof. Charles Hinds for encouraging me to investigate the genetic basis of sepsis and providing constant help throughout this endeavour, and Dr Julian Knight for his support and supervision.

I wish to thank Dr. Benjamin Fairfax and Dr. Frederick Vannberg for allowing me access to the endotoxin tolerance data, Dr. Eduardo Svoren for recruiting patients undergoing cardiac surgery, and providing support with the lab work, and Ms. Helen Lockstone for guidance with the bioinformatics aspects of this work. I also acknowledge the help provided by members of the Knight group at the Wellcome Trust Centre for Human Genetics.

I wish to thank Ms. Paula Hutton, the research co-ordinator for the project, for the difficult task of coordinating the recruiting centres and providing logistical support. I also have a deep sense of gratitude to the principle investigators and research nurses for the recruitment.

The work presented in this thesis was supported by a Clinical Fellowship Grant from the Wellcome Trust, and an Young Investigator award from the Intensive Care Society. I am indebted to these organisations for acknowledging the value of the project and providing the essential financial support.

I hold deep gratitude for my wife in what were clearly very trying circumstances, and for the constant support she has provided me.

Declarations

The Genomic Advances in Sepsis (GAinS) study was founded by Prof. Charles Hinds of Queen Mary University of London and Dr. Chris Garrard of the John Radcliffe Hospital, Oxford, as a result of frustration due to the consistent failure of clinical studies to provide substantial improvements in the clinical care of patients with sepsis. The initial study focussed only on candidate gene associations, but subsequently expanded into a genome wide association study. With the involvement of Dr. Julian Knight, my supervisor, the project also began to focus on functional genomics. The study was well-established and recruitment was ongoing when I joined the project.

My involvement in the project included logistics (maintaining supplies, collection of samples, and helping with recruitment), sample processing (with Dr. Narelle Magueri and Dr. Eduardo Svoren), analysis of microarray data (with the guidance of Ms. Helen Lockstone), analysis of eQTL data (with the help of Ms. Emma Davenport) and the design of some of the subsequent functional work (with the help of Ms. Emma Davenport and Ms. Ashley Thorpe). Microarray hybridisation was performed by Mr. Peter Ellis at the Sanger Institute, Cambridge.

The endotoxin tolerance project investigating gene expression changes following stimulation with endotoxin in macrophages derived from healthy volunteers was conceived and led by Dr. Benjamin Fairfax, Dr. Frederick Vannberg, and Dr. Julian Knight. The recruitment of subjects, extraction of macrophages, and subsequent stimulation was exclusively done by them. My involvement in the project consisted of processing of samples, particularly the extraction of RNA, and analysis of the microarray data. Dr. Fairfax and Dr. Knight subsequently expanded the study to investigate differences in gene expression in the context of cell types. My involvement in this part of the study was restricted to data analysis.

Dr. Eduardo Svoren conducted the study of gene expression changes in cardiac surgery. along with Dr. Julian Knight and Prof. Hinds. I contributed to the design of the study as a model of sterile inflammatory response, which could then be interpreted in the context of my own investigations. Recruitment was exclusively done by Dr. Svoren. I helped with some of the sample processing. Data analysis was done by Dr. Svoren, which I have independently replicated, so as to extend the analysis to be more informative in the context of sepsis. Dr. Svoren was supported by a grant from the UK Intensive Care Society.

I am responsible for a large part of the wet lab work in the sepsis cohort presented in this work, data management, and the entire statistical work, the latter with guidance from Ms. Helen Lockstone.

Contents

List of Tables viii

List of Figures xii

I Background 1

1 Review of Literature 2
- 1.1 Introduction . . . 2
- 1.2 Definitions . . . 2
- 1.3 Clinical importance of sepsis . . . 3
- 1.4 The Genetic Basis of Sepsis . . . 6
- 1.5 The Immune Response and Clinical Sepsis . . . 8
- 1.6 Functional Genomics of Sepsis . . . 16
- 1.7 Methodological Issues . . . 26

2 Aims 28

II Methods 30

3 Materials and Methods 31
- 3.1 The Sepsis Cohort . . . 31
- 3.2 The Cardiac Cohort . . . 33
- 3.3 Volunteer Cohort . . . 34
- 3.4 Sample Collection, Storage, and Transport . . . 35
- 3.5 DNA Processing . . . 36
- 3.6 RNA Processing . . . 37
- 3.7 Microarray Methods . . . 38

4 Statistical Methods 41
- 4.1 Background . . . 41
- 4.2 Demographics and Covariates . . . 41
- 4.3 Analysis of Microarray Data . . . 42
- 4.4 Class Prediction . . . 48

III The Genomic Response in Sepsis **51**

5 The Sepsis Cohort **52**
5.1 Background . . . 52
5.2 Demographics . . . 52
5.3 Summary . . . 56

6 Normalization **61**
6.1 Background . . . 61
6.2 Data Import . . . 62
6.3 Preprocessing . . . 63
6.4 Finalization . . . 69
6.5 Summary . . . 74

7 Covariates **76**
7.1 Background . . . 76
7.2 Relationships to Survival . . . 77
7.3 Relationships to Gene Expression . . . 81
7.4 Summary . . . 82

8 Exploratory Analysis **84**
8.1 Background . . . 84
8.2 Assumption Checks . . . 85
8.3 Hierarchical clustering of principal components . . . 87
8.4 Characterization of Clusters . . . 92
8.5 Validating the Clusters . . . 101
8.6 Summary . . . 109

9 The Genomic Response in Sepsis **112**
9.1 Background . . . 112
9.2 Community-acquired pneumonia response . . . 112
9.3 Faecal Peritonitis Response . . . 121
9.4 Validation . . . 126
9.5 Summary . . . 130

10 The Genomic Response and Survival in Sepsis **132**
10.1 Background . . . 132
10.2 Mortality in CAP . . . 133
10.3 Mortality in FP . . . 145
10.4 Mechanisms of mortality . . . 149
10.5 Morbidity Outcomes . . . 152
10.6 Validation . . . 156
10.7 Summary . . . 157

11 Genomic Effects of Covariates **160**
11.1 Demographic covariates . . . 161
11.2 Organ function markers . . . 163
11.3 Therapy . . . 170
11.4 Cell count covariates . . . 176
11.5 Microbiological Covariates . . . 186

11.6 Sampling covariates . . . 191
11.7 Summary . . . 192

12 Genomic Signature based Predictions 196
12.1 Background . . . 196
12.2 Diagnostic predictions . . . 197
12.3 Prognostic predictions . . . 208
12.4 Summary . . . 216

IV Sepsis in Context 218

13 Genomics of Cardiac Surgery 219
13.1 Background . . . 219
13.2 Materials and Methods . . . 220
13.3 Genomics of cardiac surgery . . . 221
13.4 Relation to sepsis . . . 232
13.5 Limitations . . . 243
13.6 Summary . . . 243

14 Genomic Signatures of Endotoxin Tolerance 245
14.1 Background . . . 245
14.2 Materials and Methods . . . 245
14.3 Exploratory analysis . . . 246
14.4 Genomic response to LPS . . . 247
14.5 Genomic signatures of tolerance . . . 256
14.6 Relationship to sepsis . . . 261
14.7 Summary . . . 269

V Conclusions and Discussion 271

15 Conclusions and Discussion 272
15.1 Conclusions . . . 272
15.2 Future Work . . . 280

Appendices 286

Supplementary Tables 287

Supplementary Figures 307

Bibliography 317

List of Tables

1.1 ACCP/SCCM consensus definition of sepsis . . . 3
1.2 ACCP/SCCM/ATS/ESICM consensus definition of sepsis . . . 4

3.1 Comparison of Illumina BeadChip platforms . . . 40

5.1 Demographics of the CAP derivation cohort . . . 53
5.2 Vital signs in the CAP derivation cohort . . . 53
5.3 Cell counts in the CAP derivation cohort . . . 54
5.4 Biochemical markers in the CAP derivation cohort . . . 54
5.5 Markers of renal function in the CAP derivation cohort . . . 54
5.6 Respiratory function in the CAP derivation cohort . . . 55
5.7 Comorbidities in the CAP derivation cohort . . . 55
5.8 Outcome risk in the CAP derivation cohort . . . 55
5.9 Treatment factors in the CAP derivation cohort . . . 55
5.10 Microbiology of the CAP derivation cohort . . . 56
5.11 Demographics of the FP derivation cohort . . . 57
5.12 Vital signs of the FP derivation cohort . . . 57
5.13 Cell counts in the FP derivation cohort . . . 58
5.14 Biochemical markers in the FP derivation cohort . . . 58
5.15 Markers of renal function in the FP derivation cohort . . . 58
5.16 Respiratory function in the FP derivation cohort . . . 59
5.17 Comorbidities in the FP derivation cohort . . . 59
5.18 Outcome risk in the FP derivation cohort . . . 59
5.19 Treatment factors in the FP derivation cohort . . . 59

7.1 Predictor variables for survival in the logistic regression model . . . 78
7.2 Logistic regression for survival at two weeks . . . 79
7.3 Logistic regression for survival at four weeks . . . 80

8.1 Demographic differences between clusters . . . 89
8.2 Differences in vital signs between clusters . . . 89
8.3 Cluster differences in white cell counts . . . 90
8.4 Biochemical differences between cluster . . . 90
8.5 Cluster differences in renal function . . . 90
8.6 Cluster differences in respiratory function . . . 91
8.7 Outcome differences between clusters . . . 91
8.8 Differentially expressed probes between clusters in the regression model . . . 93
8.9 Genes down-regulated in cluster 3 . . . 96

8.10 Differentially expressed probes between clusters in the Bayesian model . . 97
8.11 Genes down-regulated in cluster 3 in the Bayesian hierarchical model . . . 99
8.12 DAVID functional annotation of cluster differences 100
8.13 DAVID functional annotation of Bayesian cluster differences 100
8.14 Demographic differences between clusters in FP 106
8.15 Cluster differences in vital signs in FP 106
8.16 Haematological differences between clusters in FP 107
8.17 Biochemical differences between clusters in FP 107
8.18 Cluster specific differences in renal function in FP 107
8.19 Cluster specific differences in outcome in FP 107

9.1 Number of genes differentially expressed in sepsis due to CAP in each cluster 114
9.2 Top ten differentially expressed genes in each cluster in CAP sepsis 114
9.3 DAVID functional annotation of the sepsis response 117
9.4 Enriched pathways for cluster 1 vs baseline comparison in CAP sepsis . . . 119
9.5 Enriched pathways for cluster 2 vs baseline comparison in CAP sepsis . . . 120
9.6 Enriched pathways for cluster 3 vs baseline comparison in CAP sepsis . . . 121
9.7 Differentially regulated genes in the sepsis response in FP 123
9.8 Top genes of the FP response 124
9.9 Enriched pathways in FP response 125

10.1 Number of genes differentially expressed in each cluster in CAP survival . . 134
10.2 Genes consistently associated with non-survival in all clusters 136
10.3 Pathways enriched in non-survivors in cluster 1 of CAP 138
10.4 Pathways enriched in non-survivors in cluster 1 of CAP 140
10.5 Enriched pathways for survival in cluster 3 143
10.6 Genes in CAP survival 143
10.7 Enriched pathways for survival in CAP 147
10.8 Pathways enriched in non-survivors in the FP cohort 150
10.9 Genes associated with survival in CAP in the Bayesian analysis 153
10.10 Genes significantly associated with increased duration of ventilatory support in CAP sepsis 155

11.1 Number of genes associated with sex in the Bayesian analysis 163
11.2 Genes associated with sex in the Bayesian analysis 164
11.3 Number of genes associated with renal failure in the Bayesian analysis . . . 165
11.4 Genes associated with renal failure in Bayesian analysis 166
11.5 Number of genes associated with PF ratio in the Bayesian analysis 168
11.6 Genes associated with PF ratio in the Bayesian analysis 169
11.7 Number of genes associated with steroid use in sepsis in the Bayesian analysis 171
11.8 Genes induced with steroid use in CAP 172
11.9 Genes repressed with steroid use in CAP 173
11.10 Number of genes associated with APC use in the Bayesian analysis 175
11.11 Genes significantly associated with activated protein C use in the Bayesian analysis 176
11.12 Number of genes associated with relative neutrophil counts in the Bayesian analysis 179
11.13 Genes associated with relative neutrophil counts 180

11.14 Number of genes associated with relative monocyte counts in the Bayesian analysis . . . 181
11.15 Genes associated with relative monocyte counts . . . 183
11.16 Genes with discordant effects between probes in both the relative neutrophil and monocyte count contrasts in the Bayesian analysis . . . 183
11.17 Genes with discordant effects between relative neutrophil and monocyte counts in the Bayesian analysis and induced in neutrophils . . . 184
11.18 Genes with discordant effects between relative neutrophil and monocyte counts in the Bayesian analysis and repressed in neutrophils . . . 185
11.19 Microbiological categories . . . 186
11.20 Gene expression differences between microbiological categories . . . 187

12.1 Classification of clusters using support vector machines . . . 199
12.2 Classification of clusters using random forests . . . 201
12.3 Diagnostic classification using support vector machines . . . 205
12.4 Diagnostic classification using random forests . . . 205
12.5 SVM classification of survival in CAP . . . 209
12.6 RF classification of survival in CAP . . . 211
12.7 SVM classification of survival in FP . . . 215
12.8 RF classification of survival in CAP . . . 216

13.1 Significant probes in cardiac surgery . . . 224
13.2 Clusters associated with the immediate response to cardiac surgery . . . 227
13.3 Clusters associated with the late response to cardiac surgery . . . 228
13.4 Clusters associated with resolution . . . 229
13.5 Genes differentially regulated in sepsis when compared with the immediate response to cardiac surgery . . . 233
13.6 Gene sets enriched in sepsis when compared to the immediate response to cardiac surgery . . . 235
13.7 Differentially regulated genes in sepsis when compared with the late response to cardiac surgery . . . 236
13.8 Gene sets enriched in sepsis when compared with the late response to cardiac surgery . . . 236
13.9 Pathway correlation with immediate response to cardiac surgery and CAP response . . . 240
13.10 Pathway correlation with immediate response to cardiac surgery and FP response . . . 240
13.11 Pathway correlation with the late response to cardiac surgery and CAP response . . . 242
13.12 Pathway correlation with the late response to cardiac surgery and FP response 242

14.1 Significant probes in LPS stimulation . . . 249
14.2 Genes of the low dose response . . . 251
14.3 Genes of the high dose response . . . 252
14.4 Genes unique to LD response . . . 253
14.5 Genes differentiating LD and HD response . . . 255
14.6 Genes associated with endotoxin tolerance . . . 258
14.7 Gene sets associated with endotoxin tolerance . . . 262

14.8 Correlation of pathways between the endotoxin tolerance phenotype and CAP survival . . . 268
14.9 Correlation of pathways between the endotoxin tolerance phenotype and FP survival . . . 269

1 Recruiting ITUs . . . 287
2 LOC genes referenced in the text, but do not have an official symbol or an ortholog . . . 288
3 Demographics of the CAP validation cohort . . . 288
4 Vital signs in the CAP validation cohort . . . 289
5 Cell counts in the CAP validation cohort . . . 289
6 Biochemical markers in the CAP validation cohort . . . 290
7 Markers of renal and respiratory function in the CAP validation cohort . . . 290
8 Outcome risk in the CAP validation cohort . . . 290
9 Gene up-regulated in cluster-wise comparisons in CAP sepsis . . . 291
10 Genes uniquely down-regulated in cluster-wise comparisons in CAP sepsis 292
11 Genes down-regulated in all contrasts in cluster-wise comparisons in CAP sepsis . . . 293
12 Genes discriminating atypical pneumonias from gram positive infections . 294
13 Genes discriminating atypical pneumonias from gram positive infections . 295
14 Genes discriminating Gram negative pneumonias from Gram positive infections . . . 296
15 Genes induced in Gram positive pneumonias compared to viral infections . 297
16 Genes repressed in Gram positive pneumonias compared to viral infections 298
17 Genes induced in atypical pneumonias compared to Gram negative infections 299
18 Genes repressed in atypical infections compared to Gram negative infections 300
19 Genes induced in Gram negative pneumonias when compared to viral pneumonias . . . 301
20 Genes repressed in Gram negative pneumonias when compared to viral pneumonias . . . 302
21 Genes discriminating atypical pneumonias from viral infections . . . 303
22 Pathway correlation with tolerance in CAP cluster 1 survival . . . 304
23 Pathway correlation with tolerance in CAP cluster 2 survival . . . 305
24 Pathway correlation with tolerance in CAP cluster 3 survival . . . 306

List of Figures

3.1 The LeukoLOCK filter system 36

6.1 Expression signals from control probes 64
6.2 Expression signals from house keeping genes 65
6.3 Density distribution of derivation data set 70
6.4 Distribution of the validation data set 71
6.5 MA plots 72
6.6 Technical Replicates 73
6.7 Principle component analysis of the derivation data 74

7.1 Effect of haematocrit on two week survival 79
7.2 Effect of PF ratio on two week survival 80
7.3 Multi-Factor Analysis — First two components 82
7.4 Multi-Factor analysis — second and third components 83

8.1 HCPC of CAP samples 88
8.2 Venn diagram of cluster 3 94
8.3 Heatmap of repressed genes in cluster 3 95
8.4 Venn diagram of cluster 3 98
8.5 Pathways associated with clusters based on regression statistics 102
8.6 Pathways associated with clusters based on Bayesian estimates 103
8.7 CAP clusters in validation data 104
8.8 FP clusters in derivation data 105
8.9 Clusters in Paediatric Sepsis 110

9.1 Correlation in gene expression between clusters and baseline 115
9.2 Venn diagram of induced genes in CAP sepsis with cluster-wise comparisons to baseline 116
9.3 Venn diagram of repressed genes in cluster-wise comparisons in CAP sepsis 116
9.4 Volcano plot of the cluster invariant CAP sepsis response 122
9.5 Volcano plot of the sepsis response in FP 123
9.6 Sepsis response correlations across diagnostic subsets for all probes 126
9.7 Sepsis response correlations across diagnostic subsets for significant probes 127
9.8 Overlap of the sepsis response across diagnostic subsets 128
9.9 Validation of the CAP sepsis response 129
9.10 Validation of the FP sepsis response 129

10.1 Volcano plot of survival in CAP sepsis for each cluster 135

10.2 Effect sizes of genes differentially expressed in non-survivors in CAP sepsis for each cluster . . . 137
10.3 Correlation between clusters of the genes differentially expressed in non-survivors . . . 139
10.4 Overlap of genes induced in non-survivors in CAP sepsis between clusters . 141
10.5 Overlap of genes repressed in non-survivors in CAP sepsis between clusters 142
10.6 Correlation between the most significant probes in clusters 1 and 3 . . . 144
10.7 Genomic changes associated with CAP survival . . . 145
10.8 Volcano plot of cluster invariant differentially expressed genes in non-survivors in CAP . . . 146
10.9 Volcano plot of change in mean expression in non-survivors in FP . . . 148
10.10 CAP FP mortality correlation for all probes . . . 149
10.11 CAP FP mortality correlation for shared differentially expressed probes . . 151
10.12 Probability density distribution of morbidity outcomes . . . 154
10.13 Representative genes associated with increased ventilatory requirements . 156
10.14 Correlation of genes differentially expressed in CAP non-survivors in the derivation and the validation cohorts . . . 157

11.1 Genes associated with sex and survival . . . 165
11.2 Genes associated with renal failure and survival in Bayesian analysis with discordant effects . . . 167
11.3 Genes associated with steroid use and survival in the Bayesian analysis . . . 174
11.4 Genes associated with APC use and survival . . . 177
11.5 Distribution of effect sizes of relative neutrophil counts in the Bayesian analysis . . . 179
11.6 Distribution of effect sizes of relative monocyte counts in the Bayesian analysis 182
11.7 Overlap of genes associated with both relative neutrophil and monocyte counts in the Bayesian analysis . . . 184
11.8 Correlation of the relative neutrophil and monocyte counts coefficients in the Bayesian regression model . . . 185
11.9 Correlation between the Gram positive and negative signatures against viral infections . . . 189
11.10 Oligoadenylate synthetase expression in different infections . . . 190

12.1 PCA of clusters . . . 198
12.2 Misclassification rate of cluster prediction model using SVM . . . 200
12.3 Misclassification rate of cluster prediction model with random forests . . . 201
12.4 Error rates in predicting clusters with random forests . . . 202
12.5 Classifier importance in cluster prediction model with random forests . . . 203
12.6 PCA of diagnosis . . . 204
12.7 Misclassification rate of diagnosis prediction model using support vector machines . . . 206
12.8 Misclassification rate of diagnosis prediction model with random forests . 206
12.9 Error rates of diagnosis prediction model with random forests . . . 207
12.10 Classifier importance in diagnosis prediction model with random forests . 208
12.11 PCA of CAP survival . . . 210
12.12 Misclassification rate of CAP survival prediction model using support vector machines . . . 211
12.13 Misclassification rate of survival prediction model in CAP with random forests 212

12.14 Classifier importance in CAP survival prediction model with random forests 213
12.15 PCA of FP survival 214
12.16 Misclassification rate of FP survival prediction model using support vector machines 215
12.17 Misclassification rate of survival prediction model in FP with random forests 217

13.1 HCPC of the most variant probes in the cardiac derivation cohort 222
13.2 Differential gene expression after cardiac surgery 223
13.3 Effect size and significance of gene expression following cardiac surgery. A number of probes show greater than 1.5 fold change in the mean expression. Compared to the sepsis data set, more probes show greater statistical significance for similar changes in mean expression. 225
13.4 Venn diagram of probes in the immediate and late responses 225
13.5 Genes with directional effects in cardiac surgery 226
13.6 HCPC of the probes associated with the immediate response to cardiac surgery 227
13.7 HCPC of probes associated with the late response to cardiac surgery 228
13.8 HCPC of probes associated with the resolution of inflammatory response in cardiac surgery 229
13.9 Pathways associated with cardiac surgery 231
13.10 Venn diagram of gene expression overlap in sepsis compared to each of the cardiac post-operative samples 237
13.11 Correlation between the differentially regulated genes in sepsis compared to each of the cardiac postoperative samples 238
13.12 Effect sizes in the first day of sepsis compared to the first postoperative day 239

14.1 Time course for TNF tolerance 247
14.2 TNF Desensitisation 248
14.3 Effect sizes of the gene expression response to different doses of LPS 250
14.4 Venn diagram of genes shared between cells stimulated with different doses of LPS 254
14.5 Bimodal expression of *TLR8* with LPS stimulation 256
14.6 Bimodal expression of *TNF* and *IL6* with LPS stimulation 257
14.7 Correlation between probes that are differentially expressed in the response to low dose and high dose LPS 259
14.8 Pathways associated with LPS response 260
14.9 Volcano plot of the endotoxin tolerance phenotype 261
14.10 Overlap of the survival and tolerance response 263
14.11 Correlation between the survival and tolerance response for shared genes . 263
14.12 Overlap between the survival and tolerance response for shared probes identified in the Bayesian analysis 265
14.13 Correlation between the survival and tolerance response for shared probes identified in the Bayesian analysis 265
14.14 Correlation between the survival and tolerance response for all probes identified in the Bayesian analysis 266

15.1 cis associations for *TLR4* expression 282

1 Distribution of representative repressed genes in cluster 3 307
2 Largest differences between clusters 1 and 2 308
3 HCPC of the most variant probes in the cardiac validation cohort 309

4 Leading edge genes related to the mitochondrial inner membrane proteins in cardiac surgery 310
5 Leading edge genes related to the mitochondrial outer membrane proteins in cardiac surgery 311
6 Leading edge gene related to insulin synthesis and secretion in cardiac surgery 311
7 Leading edge genes related to the complement cascade in cardiac surgery . 312
8 Leading edge genes related to the regulation of apoptosis in cardiac surgery 312
9 Leading edge genes in the steroid metabolic process for the endotoxin tolerance phenotype 313
10 Leading edge genes in the antigen presentation pathway for the endotoxin tolerance phenotype 314
11 Leading edge genes related to heat shock proteins in the antigen presentation pathway for the endotoxin tolerance phenotype 315
12 Correlation between the survival and tolerance response for all genes . . . 316

Abbreviations

ACCP	American College of Chest Physicians
ANOVA	Analysis of Variance
ARDS	Acute Respiratory Distress Syndrome
CAP	Community-acquired pneumonia
CLP	Caecal Ligation and Puncture
COREC	Central Office of Research and Ethics Committees
CPB	Cardiopulmonary bypass
eQTL	expression Quantitative Trait Localisation
FDR	False Discovery Rate
FDR	False discovery rate
FP	Faecal peritonitis
GAinS	Genomic Advances in Sepsis
GSEA	Gene Set Enrichment Analysis
ICU	Intensive Care Unit
LPS	Lipopolysaccharide
LPS	Lipopolysaccharide
MCMC	Markov Chain Monte Carlo
MMP	Matrix metalloprotease
MRC	Medical Research Council
PBMC	Peripheral blood mononuclear cells
PBS	Phosphate Buffered Saline
PCA	Principle Component Analysis
PRR	Pattern recognition receptor
RBC	Red Blood Cells
RSN	Robust spline normalization
SCCM	Society of Critical Care Medicine
SIRS	Systemic Inflammatory Response Syndrome
SNP	Single nucleotide polymorphism
SVM	Support Vector Machine
TNF	Tumour necrosis factor
TRALI	Transfusion associated lung injury
UKCCG	UK Critical Care Genomics Group
VST	Variance stabilizing transformation

Part I

Background

1

Review of Literature

1.1 Introduction

Sepsis is the systemic inflammatory response to infection. The pathophysiology of sepsis is complex and involves the recognition of conserved molecular patterns on invading microbes, with subsequent release of cytokines, resulting in shock, multiple organ failure, and death (Wray *et al.,* 2001). There is abundant literature on sepsis; the investigations have been multi faceted and have involved diverse model systems and experimental designs. To keep the literature review manageable, this chapter focusses mainly on the following aspects: the definitions and impact of sepsis, evidence for the genetic basis of sepsis, aspects of the immune response relevant to clinical sepsis, genomic investigations into the pathophysiology of sepsis, and aspects related to the study methodology. As far as possible, only literature pertaining to bacterial sepsis is considered.

1.2 Definitions

Sepsis was not clearly defined prior to 1992, resulting in heterogeneous definitions that precluded meaningful comparisons of clinical cohorts and studies. In 1992, the American College of Chest Physicians (ACCP) and the Society of Critical Care Medicine (SCCM) introduced a consensus definition for sepsis based on non-specific clinical signs in the presence of an infection (Table 1.1) (Bone *et al.,* 1992) The non-specific clinical signs are diagnostic of the Systemic Inflammatory Response Syndrome (SIRS) . This

Two out of the four following criteria, in the presence of clinical or microbiological evidence of infection
Temperature >38°Cor <36°C
Heart rate >90 beats/min
Respiratory rate >20 breaths/min
WCC >12,000 cells/μL or 4000 cells/μL

Table 1.1: ACCP/SCCM consensus definition of sepsis

definition allows for rapid identification of acutely ill patients in clinical settings, but has some significant limitations. In particular, there is no gradation of severity, and the microbiological basis and pathophysiological events underlying the sepsis response are not considered.

The ACCP/SCCM criteria for sepsis was revised in 2001 to include a greater range of organ dysfunction (Table 1.2) and enable some gradation of sepsis severity (Levy *et al.,* 2003). The definitions were finalised after evaluating features such as signs and symptoms of sepsis, cell markers, cytokines, microbiological data, and coagulation parameters with a primary focus on identifying organ dysfunction. However, the 1992 definition is still commonly used to identify sepsis in clinical studies.

Severe sepsis is defined as sepsis together with organ dysfunction. Septic shock is defined as sepsis with refractory hypotension despite adequate fluid resuscitation.

1.3 Clinical importance of sepsis

Incidence

Sepsis is common and global. The cumulative incidence of sepsis in the United States from 1979 to 2000 was 8.7% (G. S. Martin *et al.,* 2003). In an observational study of more than 300,000 patients over 4 years in China, the incidence of sepsis was 2.72% (Y. Yang *et al.,* 2010) In contrast, the annual incidence of Type 2 diabetes mellitus in the French population was only 0.69% (Bellili *et al.,* 2010).

Documented or suspected infection along with some of
General Variables
Temperature >38.3°C or <36.0°C
Heart rate >90 beats/min
Tachypnea
Altered mental status
Significant oedema or positive fluid balance (>20mL/kg over 24h)
Hyperglycemia in the absence of diabetes (plasma glucose >7.7mmol/L)
Inflammatory variables
Leukocytosis or leukopenia
Normal white cell count with >10% immature forms
Plasma CRP >2 standard deviations above normal value
Plasma pro-calcitonin >2 standard deviations above normal value
Haemodynamic variables
Arterial hypotension
SvO_2 >70%
Cardiac index >3.5 L/min/m^2
Organ dysfunction variables
Arterial hypoxaemia
Acute oliguria
Creatinine increase >0.5 mg/dL
Coagulation abnormalities
Ileus
Thrombocytopenia (platelet count <100,000/μL)
Hyperbilirubinemia (total bilirubin >70 mmol/L)
Tissue perfusion variables
Hyperlactemia (>1 mmol/L)
Decreased capillary refill or mottling

Table 1.2: ACCP/SCCM/ATS/ESICM consensus definition of sepsis

The incidence of sepsis is increasing, but with a changing pattern of infection. In a study of sepsis incidence from 1979 to 2000 based on discharge data from 750 million admissions in the United States, the pattern of sepsis was observed to change from predominant Gram negative infections in early 80s to predominant Gram positive infections in the late 80s (G. S. Martin *et al.,* 2003). Over this time period the incidence of fungal sepsis increased by 207%.

As with sepsis, the incidence of severe sepsis is also high and increasing. Severe sepsis was present in 27% of ICU admissions in the first 24 hours after admission (Harrison *et al.,* 2006). This is comparable to a 19% incidence of severe sepsis in Canadian ICUs (C. M. Martin *et al.,* 2009) In the United States, this incidence rose from 23.5% to 28.7% from 1996 to 2004 (Harrison *et al.,* 2006).

Mortality

Sepsis is associated with high mortality despite advances in supportive management. The mortality rate in patients admitted to Canadian ICUs with severe sepsis was about 38% (C. M. Martin *et al.,* 2009); a similar figure was also reported from the UK (Harrison *et al.,* 2004) Mortality was higher in patients with more than two failing organs, non-white populations (G. S. Martin *et al.,* 2003) and among men (C. M. Martin *et al.,* 2009).

Between 1979 and 2000, mortality rates due to sepsis fell from 27.8% to 17.9% (G. S. Martin *et al.,* 2003). A similar pattern of decreasing mortality was also observed in severe sepsis amongst 343,860 admissions to general critical care units in England, Wales, and Northern Ireland (Harrison *et al.,* 2006). The mortality decreased from 48.3% in 1996 to 44.7% in 2004, but the overall number of deaths increased.

The increased mortality risk following sepsis persists after discharge from hospital. In a cohort of 66 ICU patients with sepsis, the cumulative two year survival rate was only 33% (Korosec Jagodic *et al.,* 2006) The standardised mortality rate for patients who have been admitted to ICUs with sepsis has been reported to be significantly higher than in the general population for up to 15 years post discharge (Williams *et al.,* 2008) The quality of life is also significantly decreased in patients who survive sepsis (Combes *et al.,* 2003; Cuthbertson *et al.,* 2010).

1.4 The Genetic Basis of Sepsis

The high incidence of sepsis, the persistent high mortality despite supportive therapy, the lack of specific therapeutic measures, and the controversies over canonical treatment strategies suggest the need for further investigations into the pathophysiology of sepsis.

Evidence for the genetic basis of sepsis comes from three sources: epidemiological studies, family based studies, and genetic association studies. The studies are heterogeneous with respect to causative agents and study populations and relate sepsis risk and sepsis related outcomes to genetic diversity.

Epidemiological associations

The heritable risk of death due to severe infections has been quantified by the Danish Adoptees study that investigated the genetic and environmental influences on adult mortality in 960 families and adopted children (Sørensen *et al.,* 1988). Death of a biological parent due to infection before the age of 50 was associated with a 5.81 (2.47–13.7) fold increase in relative risk of death in the adoptee due to infection. The death of an adopting parent was not associated with any additional mortality risk in the adoptee, suggesting an inherited mechanism. Interestingly the genetic risk of mortality due to infection is greater than the genetic risk of mortality due to cardiac disease (relative risk 4.52, 1.32–15.4). A similar but weaker pattern was observed when the biological parent died between 50 and 70 years of age (relative risk 3.02, 0.72–12.8).

Family based studies

Twin-studies and family-based studies are not feasible in bacterial sepsis, but evidence for genetic susceptibility is available for specific infections. For example, in a study of *Plasmodium falciparum* malaria in Gambian twins, the clinical presentation but not disease susceptibility was found to be influenced by genetic variation (A. P. Jepson *et al.,* 1995). A large proportion of the phenotypic variance in this group of patients was contributed by non-MHC loci (A. Jepson *et al.,* 1997) Sibling pair studies have also been used to identify associations with chronic infections such as leprosy and HBV

(Hill, 2006) Similarly, linkage studies using mouse models have associated NRAMP1 with tuberculosis infections (Bellamy *et al.,* 1998).

Association studies

There is a profusion of studies associating sepsis risk and outcomes to sequence variants. Nearly all of these studies use one of two approaches. The first involves sequencing of genes identified as being dysfunctional by other methods in extreme phenotypes. The second involves candidate gene approaches using a case control study design. These associations have been extensively reviewed (Arcaroli *et al.,* 2005; Brouwer *et al.,* 2009; Kaslow *et al.,* 2008; Papathanassoglou *et al.,* 2006). Associations to sepsis susceptibility or outcomes have been reported for a number of components of the immune system including TLR4(Arbour *et al.,* 2000; Lorenz *et al.,* 2002), CD14(Koenig *et al.,* 2002), MBL(Eisen *et al.,* 2006; Gordon *et al.,* 2006), IRAK4(Picard *et al.,* 2003), TNF(Mira *et al.,* 1999), and the MHC (Nooh *et al.,* 2007).

There is at least one genome wide association study for sepsis susceptibility in bacterial sepsis. A genome wide association study was performed in 475 children with meningococcal disease using 4703 controls (Davila *et al.,* 2010) Two single nucleotide polymorphisms (SNP), one in complement factor H (*CFH*) and the other in CFH-related protein 3 (*CFHR3*), were associated with sepsis risk. One interesting feature in this study was the modest significance levels in the initial study cohort. Two replications were performed in 553 Western European patients and 415 Spanish patients respectively. Only a small subset of SNPs (79 and 11 respectively) were genotyped in the replication cohorts. The associations continued to be modest in each of the individual analysis, but achieved genome wide significance levels in the combined analysis. The appropriateness of this approach is not clear, but is partly offset by the strong biological plausibility of the reported associations; failure of complement mediated killing contributes to the virulence in *Neisseria meningitidis* infections (Kugelberg *et al.,* 2008) Sibling studies attribute a third of the risk in meningitis to genetic factors (Haralambous *et al.,* 2003).

1.5 The Immune Response and Clinical Sepsis

This section focusses on features of the immune response in relation to some of the clinical aspects of sepsis: the relationship to hypoxia and coagulation, the concept of cell priming and endotoxin tolerance, mitochondrial dysfunction, metabolic aspects of sepsis, and the use of steroids.

Hypoxia

The genomic response to sepsis can be influenced by a variety of clinical events each of which have independent effects on gene expression. For example, respiratory failure is common in sepsis(Cartin-Ceba *et al.,* 2011), while hypoxia is associated with specific gene expression effects, predominantly mediated by the hypoxia inducible transcription factor HIF-1 (Gregg L. 2001) This is a key molecule with a number of modulatory effects on gene expression, that has also been shown to be expressed in a non-specific manner in skin infections in humans and in murine models of *Staphylococcus aureus* infections (Werth *et al.,* 2010) HIF-1α accumulation has been demonstrated to be the result of downstream TLR4 signalling (Vadim V. 2008) Mouse lacking HIF-2α are resistant to the effects of endotoxin (Imtiyaz *et al.,* 2010) Similarly, HIF-1α knock-down abrogates production of IL6 in myeloid cells following exposure to LPS (Lall *et al.,* 2008). These effects may be specific to certain cell types; for example the knock-down of *HIF-1*α in T cells improves survival in a mouse model of sepsis (Thiel *et al.,* 2007).

The effects of HIF-1 are closely linked to NF-κB, a transcription factor with a key role in the inflammatory response. In mouse models IKK-β acting through NF-κB is essential for *HIF-1* transcription(Rius *et al.,* 2008), while HIF-1α activates NF-κB (Walmsley *et al.,* 2005) Increased production of TNF by murine macrophages in hypoxia is mediated by HIF-1α (F. Q. Liu *et al.,* 2008), and augmentation of the cytokine response (TNF-α, IL-1, IL-4, IL-6,IL-12) has also been demonstrated in non-hypoxic mice models (Peyssonnaux *et al.,* 2007) HIF-1 has additional important effects and links the multiple organ failure in sepsis to bio-energetic failure by augmenting the expression of the mitochondrial ATPase inhibitory protein IF1 in rat liver cells (L. Huang *et al.,* 2011) There are also

organ specific effects; HIF-1 reduces hypoxia induced vascular leak through repression of adenosine kinase in endothelial cells (Morote-Garcia *et al.,* 2008).

Coagulation

Sepsis is associated with defects in coagulation(Saracco *et al.,* 2011), resulting from a combination of endothelial dysfunction and dysregulated coagulation factors.

A mild pro-thrombotic state is an intrinsic component of the inflammatory response. Microthrombotic episodes in capillaries result in compartmentalization of bacteria and prevent systemic spread of infection. However, diffuse microvascular thrombosis impairs tissue perfusion and contributes to tissue injury. The use of activated protein C (APC) in sepsis was largely based on the hypothesis of improved micro-vascular flow. (Dhainaut *et al.,* 2003) Although APC has been withdrawn, investigations into the protein C system has identified novel anti-inflammatory effects of protein C in sepsis (Della Valle *et al.,* 2012).

The interactions between inflammation and coagulation have been recently reviewed (van der Poll *et al.,* 2012) Neutrophils are key players in the coagulation response, through the production of serine proteases: elastase (ELANE) and cathepsin G. The proteases lyse the tissue factor pathway inhibitor resulting in activation of factor X, IX, and XII (Massberg *et al.,* 2010).

There have been a number of attempts to favourably modify the coagulation system at different levels of the protein C pathway. Recently, a small study in mouse model of sepsis identified improved outcomes with the use of low molecular weight heparins (Iba *et al.,* 2012) In another study in mice, infusion of hemin, an inducer of heme-oxygenase 1, reversed the pro-thrombotic effects of sepsis, and increased the expression of activated protein C (Fei *et al.,* 2012).

The importance of other molecules, involved in both inflammation and coagulation, are increasingly recognised. For example, VEGF contributes to the pro-thrombotic state in sepsis and genetic repression of VEGF production reduces the thrombin-antithrombin interaction in mouse models of sepsis (X. Liu *et al.,* 2010) Similarly, the presence of the 4G polymorphism in *PA-1* (plasminogen activator inhibitor) in Caucasian

patients with severe sepsis due to pneumonia is associated with increased odds of multiple organ failure (OR 3.0, CI 1.3–6.7) and septic shock (OR 2.6, CI 1.1–6.0) (Madách *et al.,* 2010).

Endothelial dysfunction is an important element of the pathophysiology of sepsis. The matrix metalloproteases (MMP) are molecules derived from leucocytes and endothelial cells that are involved in the host pathogen response. Plasma levels of MMP1 increase during sepsis and contribute to increased endothelial permeability and death in mouse LPS models of sepsis (Tressel *et al.,* 2011).

Priming of immune cells

The immune response is dependent on cell state at the time of infection. Therefore, the immune response is modulated not only by events that occur during the infection, but also by preceding events. The "two-hit" hypothesis that a more severe inflammatory response follows an initial injury has been favoured by clinicians, and has been linked to diverse clinical events such as transfusion associated lung injury (TRALI) (Looney *et al.,* 2004) and nosocomial infections following acute respiratory distress syndrome (ARDS) (Meduri, 2001).

Splenocytes derived from mice with burn injury have been shown to be primed to further infectious insults (Paterson *et al.,* 2003) After priming by the injury, mice exposed to LPS show increased production of TNFα, IL-6, and IL-1β compared to animals subjected to sham injury. The priming was not associated with increased surface expression of TLR2 and TLR4 and was also observed in TLR4 deficient animals. Additional evidence for this phenomenon has been provided at the molecular level in leucocytes derived from patients, especially in the context of thermal injury (Murphy *et al.,* 2005) However, this paradigm has been challenged in a recent genomic study, and a pattern of persistent dysregulation of the immune response has been suggested (Xiao *et al.,* 2011).

Endotoxin tolerance

The concept of cell priming is related to the phenomenon of endotoxin tolerance, where a prior exposure to lipopolysaccharide (LPS) results in transient hypo-responsiveness to a second challenge to LPS. Endotoxin tolerance is consistently observed in in-vivo systems, cell model systems, animal models of sepsis, and after non-infective insults like trauma and cardiopulmonary bypass(Grundmann *et al.,* 2000) and involves a stereotyped reprogramming of the immune response (Shnyra *et al.,* 1998) The mechanisms underlying endotoxin tolerance are unclear but clearly involve multiple cell types and cytokines (Medvedev *et al.,* 2000; Shnyra *et al.,* 1998; Wysocka *et al.,* 2001).

Endotoxin tolerance was first described by Paul Beeson who noted that rabbits with a prior exposure to typhoid vaccine failed to develop a febrile response following a second exposure (Beeson, 1946) At the cellular level, endotoxin tolerance was defined initially by a reduction in plasma levels of TNF; subsequently reduced production was also identified in other cytokines such as IL-6, IL-8, and IFN-γ (Astiz *et al.,* 1995; Rayhane *et al.,* 1999) Tolerance is now understood to involve a complex rearrangement of cellular events, with altered patterns of gene expression (including epigenetic modifications) and cytokine production (T. F. Liu *et al.,* 2011; Melo *et al.,* 2010; Xiong *et al.,* 2011).

There is an abundance of literature on the molecular mechanisms of endotoxin tolerance, but a complete understanding remains elusive. The molecular mechanisms of endotoxin tolerance are complex and include changes in protein expression and chromatin remodelling at the promoters of *TNF* and *IL1*β (Chan *et al.,* 2005; X. Chen *et al.,* 2009; Gazzar *et al.,* 2007). In macrophage models of endotoxin tolerance, PBMCs derived from human volunteers showed suppression of class II activity. The overall decrease in activity was due to a combination of decreased expression of the invariant chain, *CD74*, *HLA-DM*, reduced synthesis of class II MHC proteins, and decreased cell surface expression of class II MHC molecules (Wolk *et al.,* 2003).

Exposure to endotoxin results in the production of pro-inflammatory cytokines such as TNF, while simultaneously reducing the surface expression of TNF receptors in monocytes and granulocytes (by approximately four fold) (van der Poll *et al.,* 1995).

The latter is possibly a mechanism to protect tissues from excessive circulating levels of TNF and might contribute to the phenomenon of endotoxin tolerance.

Paradoxically, endotoxin tolerance is associated with an increased surface expression of TLR4. (Calvano *et al.,* 2003) In a study of monocytes derived from 33 patients admitted to a surgical intensive care unit, both the transcript levels of *TLR4*, and the cell surface expression of TLR4 showed 2.4 and 1.7 fold change in expression respectively, when compared to controls. No changes were observed in the expression of *TLR2* and *CD14*. The increased expression of TLR4 was also seen in patients with SIRS irrespective of the presence of infection, but any inference is limited by small sample sizes (n = 3 for SIRS without infection). This increase in surface TLR4 appears to be non-specific and is also seen in other models of injury such as mouse models of burn injury and is mediated by the p38/MAPK signalling system (Maung *et al.,* 2005). The increased TLR4 reactivity post injury underpins the "two hit" hypothesis (Murphy *et al.,* 2005).

As with the molecular mechanisms, the relationship of endotoxin tolerance to clinical sepsis is difficult to define, despite some clear associations. In clinical terms, endotoxin tolerance may be protective. An interesting study in pigs relates organ damage in sepsis to endotoxin tolerance. Tolerant pigs show reduced cardiac and renal dysfunction but not pulmonary dysfunction compared to non-tolerant pigs (Castegren *et al.,* 2012) Survival benefits have also been demonstrated in murine models of tolerance in sepsis (Shi *et al.,* 2011) Tolerance has also been related to clinical features of sepsis such as endothelial dysfunction(Draisma *et al.,* 2010) and adrenocortical insufficiency (Polito *et al.,* 2011).

Tolerance has been demonstrated at a cellular level in patients with sepsis. Whole blood derived from sepsis patients displays tolerance to endotoxin when compared to controls. The suppression of pro-inflammatory cytokine release persisted for at least 10 days after sepsis onset (Ertel *et al.,* 1995) The reduction in concentrations of TNF and IL6 seems to result from a combination of reduced transcription and decreased half life. The clinical phenomenon has been shown to be mediated by different cell types including neutrophils(Pillay *et al.,* 2012) and monocytes (Kanaan *et al.,* 2011) Molecular events associated with tolerance in human sepsis include alterations in TLR4

signalling, including downstream molecules like IRAK4,(Q. Yang *et al.,* 2011a) nitric oxide synthesis,(Draisma *et al.,* 2010) and heat shock proteins (Su *et al.,* 2010).

In contrast to the abundance of literature investigating the molecular mechanism of tolerance, few studies have examined the genetic basis of tolerance. A single nucleotide polymorphism (SNP) in the promoter of *TNFR2* (rs522807) is associated with decreased secondary production of TNF. Increased production of TNFR2 (TNF receptor) was associated with decreased production of TNF in response to LPS. Interestingly, TNF produced as the primary response to LPS, increases the expression of TNFR2. The primary response, therefore, establishes the context for the secondary response (Fairfax *et al.,* 2011).

Metabolic effects

The immune system has strong interactions with tissue metabolism. Polymorphisms of *IL6* (174G/C) are associated with obesity(Berthier *et al.,* 2003) and insulin resistance,(Kubaszek *et al.,* 2003) while diabetes mellitus is associated with a pro-inflammatory state (Devaraj *et al.,* 2010) Modulation of metabolic mechanisms can be either due to the systemic effects of inflammation(Kovarik *et al.,* 2012) or secondary to hepatic dysfunction as a result of inflammation (Q. Yang *et al.,* 2011b).

The relationship between metabolism and immunity involves core processes of the immune response. RIP40 is a nuclear cofactor interacting with NR family members like PPARs, ERRs, ERs. RIP40 reduces cellular energy expenditure through control of mitochondrial activity. In mouse bone-marrow derived macrophages, RIP40 deficiency is associated with reduced expression of pro-inflammatory genes in both resting and stimulated cells (TLR2, TLR3 and TLR4 ligands). Changes in gene expression are also associated with decreased production of cytokines IL1b, IL6 and TNFa. The effect is mediated at the promoter region of genes responding to the NFkB transcription complex, through interaction mainly with the RD1 domain of the RelA subunit of NFkB as a co-activator with the histone acetyltransferase p300/CBP, but also with RelB and to a lesser extent with c-Rel. A synergistic effect was best seen at low concentrations of RelA (Zschiedrich *et al.,* 2008).

Immune processes are also related to lipid metabolism. Sphingosine and ceramide are intermediates of sphingolipid metabolism. Ceramide is a signal transducer that promotes apoptosis. Sphingosine, the precursor of S1P promotes cellular growth and differentiation. Dysfunctional uptake of cholesterol in inflammation is a causative model for steroid insufficiency (Cai *et al.,* 2008).

Mitochondrial dysfunction

Mitochondrial dysfunction is common in sepsis and contributes to tissue oxidative stress (Apostolova *et al.,* 2011) Mitochondrial abnormalities play a key role in the organ dysfunction of a number of organ systems including gut,(Andersson *et al.,* 2012) heart,(Chopra *et al.,* 2011) and kidneys (Pathak *et al.,* 2012).

Other roles for mitochondria in the innate immune response are now being recognised. There is particular interest in damage associated molecular patterns (DAMPs) resulting in a systemic inflammatory response in a manner analogous to the pathogen associated molecular patterns (PAMPs). In one study, the levels of mitochondrial genome derived DNA (mtDNA) were elevated more than 1000 fold for more than 24 hours in patients with blunt trauma when compared to volunteers. (Zhang *et al.,* 2010) The mtDNA induced an inflammatory response with neutrophil and p38/MAPK activation, as well as release of MMP8 and IL8 from activated neutrophils. This response is qualitatively similar to the immune response seen in infection.

Mitochondria also play a crucial role in the intrinsic pathway of apoptosis. A central event in the intrinsic pathway is mitochondrial outer membrane permeabilization, resulting in the release of pro-apoptotic proteins from the mitochondrial inner membrane space (Tait *et al.,* 2010) The apoptotic effects of mitochondrial dysfunction and inflammation secondary to DAMPs have been demonstrated in clinical sepsis (Garrabou *et al.,* 2012).

Consistent with clinical experience, mitochondrial dysfunction is not restricted to immune cells. Testicular cells in a mouse model of sepsis (intra-peritoneal LPS) showed an early induction (three hours) of stress proteins (HMGB1, HMGB2, and HSP-60) (Metukuri *et al.,* 2010). This was followed by a 74% decrease in Bcl-2(anti-

apoptotic)/Bax(pro-apoptotic) ratio and an increase in p53 levels. Mitochondrial dysfunction was observed from six hours, with the leakage of Cytochrome c. There was also an increase in caspase-3 activity. These findings are consistent with the important role of mitochondrial outer membrane permeability in apoptosis.

The predominant pattern of apoptosis in severe sepsis is unclear. Death pathways are activated by TLR4 signalling. However, the activation of the heat shock proteins results in an anti-apoptotic state. One of the mechanisms by which LPS causes apoptosis is by increasing the activity of sphingosine kinase 1 (SK1), which catalyses the formation of sphingosine-1-phosphate (S1P), which initiates some immune processes. In a murine macrophage model (RAW 264.7), treatment with LPS/SK1 reduced viability by 85% (Ding *et al.,* 2010) Over-expression of HSP-70 increased viability by 25%. HSP-70 increased SK1 transcript and protein levels. In contrast, mitochondrial dysfunction in severe sepsis results in a pro-apoptotic state (Metukuri *et al.,* 2010).

Steroid use

The presence of adrenal insufficiency is an independent risk factor for mortality in sepsis (Annane *et al.,* 2000), but the use of steroids as adjunctive therapy in sepsis is controversial and the role of steroids in sepsis management is constantly being reappraised. Current recommendations limit steroid therapy to septic patients with hypotension refractory to vasopressors.

The immunosuppressive effects of glucocorticoids are well documented but findings are often contradictory. For example, treatment with dexamethasone has been shown to paradoxically increase the production of macrophage migration inhibitory factor in mice CLP models of sepsis,(Fingerle-Rowson *et al.,* 2003) a finding refuted by others (Bruhn *et al.,* 2006). There was a tendency towards decreased production of pro-inflammatory cytokines, but this was not statistically significant.

The glucocorticoid receptor is a member of the nuclear receptor super-family. On forming a complex with glucocorticoids in the cytoplasm, the nuclear receptor complex translocates into the nucleus and disrupts the p65/interferon regulatory complexes, thereby interfering with TLR4 signalling. Other ligands for the nuclear receptor super-

family include PPARγ, vitamin D3 and oestrodial. The responsive gene sets of each of the nuclear receptors overlap, but with distinct molecular mechanisms of action (Ogawa *et al.,* 2005).

Steroid deficiency in sepsis might itself be a result of the inflammatory process. The scavenger receptor B-I (SR-BI) is necessary for HDL uptake into cells and steroid synthesis. SR-BI also binds to LPS similar to PRRs, and protects cells from LPS mediated damage. SR-BI null mice display uncontrolled inflammation with LPS and glucocorticoid insufficiency due to lack of glucocorticoid induction. SR-BI is not needed for basal synthesis but seems to be important for steroid response in shock. In SR-BI null mice stimulated with LPS greater and persistent production of cytokines was observed along with 100% mortality when compared to controls (Cai *et al.,* 2008). Mortality was improved by steroid replacement. ACTH stimulation had no effect suggesting adrenal failure resulting from the failure to replenish intracellular cholesterol.

1.6 Functional Genomics of Sepsis

Whole Genome Studies

There is very limited literature investigating the functional genomics of sepsis, especially in clinical contexts.

Paediatric sepsis

The majority of published work on paediatric sepsis is from a research group based in Cincinnati Children's Hospital Medical Centre. The reports are based on overlapping data sets from a study design that is analogous to the design of the studies reported in this thesis. The study population consisted of children (less than 10 years of age) admitted to paediatric ICUs with sepsis, severe sepsis, septic shock, or SIRS. Samples were collected at multiple time points. The sample obtained in the first 24 hours of ICU admission was labelled as day 1 sample. mRNA transcripts were obtained from whole blood using the PAXGene kit. Samples derived from healthy children were used as controls.

In the first study of 42 patients (non-survivors = 9) and 15 controls, a panel of 2482 genes were differentially regulated with at least two fold change in expression (Wong *et al.,* 2007) Gene sets related to IL-6, IL-10, TLR, B-cell receptor signalling, and complement cascade were expressed while gene sets related to T-cell receptor signalling and antigen presentation pathways were repressed. Induction of *MT*. metallothionein, an intra-cellular zinc binding protein, were associated with increased mortality.

A further study was performed on 18 control patients, 22 with SIRS, 32 with sepsis and 67 with septic shock (Wong *et al.,* 2009a). Apart from samples on day 1, additional samples were collected on day 3 in patients with persisting pathology (n = 20 sepsis, 39 septic shock) and 24 patients with resolved SIRS. Compared to SIRS, patients with sepsis and septic shock expressed gene sets involved in inflammation and immunity on both days. Gene sets related to the adaptive immune system were repressed in septic shock on both days. Other important gene clusters related to class II MHC activity and TGF-β signalling. Interestingly, the number of genes that were differentially regulated on day 1 was greater in SIRS than in sepsis. This observation contrasts with the clinical notion of the SIRS-sepsis spectrum where sepsis is further removed from normality than SIRS and could be indicative of previously unsuspected difference in underlying pathophysiologies.

Based on gene expression profiling, three subclasses representing different patterns of inflammation could be identified (Wong *et al.,* 2009b) A panel of 6934 genes were differentially regulated across the clusters. One of the subclasses was associated with Gram positive infections and higher mortality rates, while a second subclass was associated with steroid use. A panel of 100 genes was used to predict the cluster membership. The model performed with 91% accuracy with internal validation but no external validation was performed. No attempt was made to correlate the subclasses to other covariates. An attempt has been made to use these predictors clinically with good success rates (Wong *et al.,* 2010b).

To determine the validity of gene expression profiles determined from whole blood mRNA, transcripts from whole blood were compared to leucocyte subsets (neutrophils, monocytes, lymphocytes) in 5 controls and 13 patients with septic shock (Wong *et al.,*

2010a). A panel of 2355 genes were differentially regulated between sepsis and controls in whole blood samples. Using cell type specific signature probes, these genes were found to be represented in all three cell types. For the individual subsets, 58% of the expressed probes and 69% of the repressed probes were commonly expressed in all three leucocyte subsets.

Adult sepsis

There are three studies from the Department of Intensive Care Medicine, University of Sydney investigating gene expression profiles in adult patients with sepsis.

A panel of 94 genes were differentially expressed when compared to controls in the PBMCs of 96 patients admitted to ICU (sepsis n = 72) (Tang *et al.,* 2008). The authors did not restrict recruitment to specific diagnostic groups. The differentially expressed genes included those involved in the immune response and mitochondrial function. There was no difference in the gene expression profiles between Gram positive and Gram negative infections.

The same authors examined gene expression profiles in neutrophils from 94 patients admitted to ICU with sepsis (sepsis n = 71) (Tang *et al.,* 2007). Again there was no recruitment restrictions based on underlying cause of sepsis. The time of sampling was not specified. A panel of 50 genes were differentially regulated between sepsis and healthy controls. A predictive model for the diagnosis of sepsis was developed from 44 samples and validated against the remaining 50 samples with a predictive accuracy of 88%. The differentially regulated genes comprised three functional clusters involving mitochondrial function, immune regulation and inflammatory response.

In a later study gene expression profiles were determined in PBMCs from 70 patients admitted to ICU with either sepsis or SIRS using microarray hybridization (Tang *et al.,* 2009). The samples were split into training and validation sets of 35 each. A panel of 138 genes differentiated sepsis from SIRS, after correcting for lymphocyte and monocyte counts. A predictive model for sepsis was derived which had an predictive accuracy of 80%. The differentially expressed genes included functional clusters involved in apoptosis, lymphocyte activation, inflammatory response, and immune function. Only

three genes were shared between this panel and the panel of genes derived from gene expression in neutrophils. As with earlier studies Gram positive infections could not be differentiated from Gram negative infections.

The interesting feature of these three studies is the very low gene yield, especially since large changes would be expected between sepsis and control samples. The use of controls derived from patients admitted to ICU with non-septic conditions might have seriously biased the results. These patients are likely to have inflammatory events which would have reduced effect sizes.

Other infections

Whole genome gene expression profiling of whole blood cells has been used to identify the genomic signatures associated with dengue shock syndrome in 48 Cambodian children. A regression model, which corrected for age, gender, and technical covariates, was used to analyse the data. A panel of 2959 genes with a false discovery rate (FDR) of 0.1 were identified that differentiated dengue shock from other dengue variants. Pathway analysis identified canonical inflammatory pathways associated with dengue shock syndrome along with some metabolic pathways (Devignot *et al.,* 2010). Interesting features of this study include the large number of genes used to create the signature, and the small effect sizes in the clusters of the samples derived using the genes in this panel. A predictive model was not attempted, and is unlikely to have had adequate predictive power. In addition, the expression data was not normalized to relative cell counts. The authors do however acknowledge the predominance of neutrophil specific effects in some of the gene sets.

In contrast, a small panel of 393 transcripts derived from whole blood constitute the genomic signature for human tuberculosis, albeit for the relatively extreme phenotype of moderate or high disease load (Berry *et al.,* 2010). An even smaller panel of 86 transcripts discriminates tuberculosis from other infectious diseases. Part of the differential expression was attributed to the change in relative numbers of cells. The pattern of differential expression suggests a large role for interferon driven pathways in active tuberculosis. As with the genomic signature for the dengue shock syndrome, no

predictive model was developed. Interestingly, a subset of patients with latent disease displayed gene expression patterns similar to active tuberculosis, which would have reduced the predictive power of a classification model.

Covariates

The commonly used risk scores in sepsis use a number of covariates such as age, sex, and organ function to estimate a long-range frequentist mortality risk (Knaus *et al.,* 1985). If the variation in sepsis response observed among individuals and differences in sepsis related outcomes have a genetic basis, it is reasonable to expect these covariates to modulate gene expression in sepsis. Despite the strong biological plausibility, very few studies specifically address the effect of these covariates in sepsis, even though the modulatory effects have been described in non-sepsis contexts.

The effect of sex on whole genome gene expression in *Coxiella burnetti* infections was investigated in 40 C57BL/6 mice. (Textoris *et al.,* 2010) mRNA transcripts were extracted from liver cells after 24 hours of *Coxiella burnetti* infection. A panel of 2777 probes were differentially regulated with *C. burnetti* infection. Surprisingly, 86% of the probes were sex specific. The effect of sex hormones were subtle and influenced cell adhesion and signal transduction pathways. Interestingly, there was an association with the circadian rhythm pathway in female but not male mice.

Gene Reprogramming

The innate immune response to bacterial infections consist of stereotyped changes involving gene reprogramming (McCall *et al.,* 2007). Two elegant studies describe these sequential changes in expression using whole genome microarray based expression data.

In the first study, the gene expression signatures of peripheral blood mononuclear cells (PBMCs) exposed to different bacterial antigens was used to demonstrate a stereotyped gene expression program that was specific to the bacterial antigen (Boldrick *et al.,* 2002) PBMCs derived from human volunteers were exposed to heat killed strains of *Staphylococcus aureus*, *Escherchia coli*, and *Bordetella pertussis* and live strains of *B. pertussis*. Sequential changes in gene expression were identified by microarray hybridization

at eight time points over 24 hours. A panel of 206 genes related to leucocyte recruitment and activation, sharing in common the transcription factor NF-κB, constituted the shared response to bacterial antigens. A panel of 96 genes were commonly repressed but with a delayed effect. This panel included genes encoding antigen presenting molecules and chemokines. Interestingly, increasing the antigen dose resulted in a gene expression response gradient. Also, while exposure to killed strains resulted in the persistent induction of TNF, cells exposed to live strains displayed a rapid induction followed by persistent repression. In summary there are four key findings from this study: a common temporal pattern of gene reprogramming with different bacterial strains, a dose-response relationship to bacterial antigens, a consistent pattern of donor specific differences in expression, and a tolerance like phenomena with live strains but not killed strains.

This focussed reprogramming has also been demonstrated for genes sensitive to TLR4 signalling in mice bone marrow macrophages (Foster *et al.,* 2007). Based on microarray hybridisation of transcripts from cells exposed to LPS, differentially expressed genes were classified into two groups: tolerizable (genes that could not be induced in pre-stimulated macrophages) and non-tolerizable (genes that could be induced in pre-stimulated macrophages). While each of these groups included a number of functional gene categories, pro-inflammatory genes dominate the tolerizable subset while anti-microbial genes dominate the non-tolerizable subset. The differential regulation of these two groups was mediated by epigenetic modifications including histone modifications and nucleosome remodelling. These findings suggest that the tolerant state reduces potential damage from uncontrolled inflammation while maintaining responsiveness to pathogens.

Candidate Gene Studies

There are very few studies that focus primarily on functional genetics using a candidate gene approach. Most of the available literature is from associations of structural variations to clinical phenotypes with additional corroborations from changes in mRNA transcript or protein levels. While the phenotypic effects of any polymorphism are likely

to be based on changes in transcript and protein levels and function,(Pastinen *et al.,* 2006) this review is restricted to data where such changes have been reported. For obvious reasons, the candidates in most of these studies are genes directly related to the immune system.

Cytokines

The production of TNF is a hallmark of the systemic immune response. Excessive production of TNF in response to endotoxin is associated with increased mortality in primates(Hinshaw *et al.,* 1990), and the use of anti-TNF antibodies improves organ function and survival in primates, although not in humans (Hinshaw *et al.,* 1992; Reinhart *et al.,* 2001). This could be due to the failure of primate models of sepsis to adequately reflect human sepsis and the inability to initiate therapy at the appropriate time in patients due to the delays in diagnosing sepsis clinically.

Polymorphisms in *TNF*, and circulating levels of TNF have been associated with sepsis risk and outcome. There has been considerable interest in TNF based on its role as a key mediator in sepsis response, its ability to replicate sepsis symptoms, and its defining role in endotoxin tolerance. (Spooner *et al.,* 1992)

In a study of 110 patients homozygosity of *TNFB2* to the Nco1 polymorphism was associated with severe post-traumatic sepsis (OR 5.22, CI 1.6–17.9), and higher TNFα levels in patients who developed severe sepsis (Majetschak *et al.,* 1999). The risk was adjusted for age and injury severity but not for other covariates. The reported change in TNFα levels were modest, and despite statistical significance there was considerable overlap.

The association of TNF polymorphisms to clinical sepsis risk and outcomes has not been consistent. In a study of 31 SNPs in 14 candidate genes related to the immune response in 774 adult patients in medical ICUs, the presence of two marker haplotypes in *TNFA* (-308A, -238A), was associated with 85.7% mortality compared to 38.9% mortality in other haplotypes, along with shorter time to death (Henckaerts *et al.,* 2009) This result was based on just 7 patients with the two-marker haplotype. A similarly modest association with mortality was noted with the *MASP2* variant (D120G), the presence of

which was correlated with lower MASP2 levels on admission. Stronger associations were seen between a combination of a *NOD2* variant with a *TLR4* 299G variant and bacteremia (OR 4.34, CI 1.89–9.98) and mortality (OR 2.27, CI 1.09–4.74).

The expression of a number of cytokines have been linked to sepsis mortality. Over-expression of pro-inflammatory cytokines TNF, and IL6 have been associated with poor outcome (Brown *et al.,* 2006) In contrast, in mice caecal ligation and puncture model of sepsis, treatment with IL-33 reversed the pro-inflammatory changes and improved survival (Alves-Filho *et al.,* 2010).

Pattern recognition receptors

The observation that a number of different mechanisms can contribute to an increased risk of sepsis reinforces the notion that sepsis as a complex disease, in which different mechanisms can result in the same or similar phenotypes. For example, rare deficiencies in IRAK-4 have been associated with increased risk of Gram positive infections in children (Picard *et al.,* 2003; Puel *et al.,* 2005). Interestingly, TLR mediated induction of interferon in response to viral infections also requires IRAK-4 (K. Yang *et al.,* 2005).

The use of functional genomics has also clarified some of the inconsistent associations between structural variants and sepsis risk. For example, two non-synonymous polymorphisms of *TLR4*, Asp299Gly and Thr399Ile were inconsistently associated with hypo-responsiveness to endotoxin and septic shock. Despite the change in protein affinity in *TLR4*, the polymorphisms were not associated with changes in cytokine expression in PBMCs derived from healthy volunteers following exposure to endotoxin or respiratory syncytial virus (Douville *et al.,* 2010).

The pattern of regulation of the TLR signalling system varies considerably in patients with severe sepsis when compared to sepsis and in PBMCs when compared to polymorphonuclear leucocytes. (Salomao *et al.,* 2009) These results were obtained from just 15 patients (5 with severe sepsis, 5 with septic shock, 5 with sepsis) and 5 controls. Transcript levels were estimated by real time PCR in cells isolated by Ficoll gradient. For PBMCs the pattern of differential expression (defined as 2 fold change) was similar between severe sepsis versus controls and sepsis versus controls. The findings were in concordance with

other reports and showed an induction of TLR transcripts and repression of downstream transcripts such as cytokines and *NFKB1*. In PMNs, the number of expressed genes was considerable lower in severe sepsis when compared to controls than in sepsis, while the number of repressed genes was larger. The changes in severe sepsis were comparable to those seen in septic shock. The repressed genes in PMNs mainly involved the MAPK system in contrast to PBMCs. Despite the small numbers, the results show cell type specific differences in expression and provide further evidence of a more dysfunctional immune response in severe sepsis and septic shock when compared to sepsis.

The activity of TLR4 is now known to be modulated by endogenous ligands which are yet to be fully characterized. MRP8 and MRP14 are calcium binding S100 proteins, the heterodimers of which are endogenous ligands of TLR4. These proteins are released during phagocyte activation and amplify the inflammatory response. Mice deficient in MRP14 are protected from endotoxin shock (Vogl *et al.,* 2007) In addition, these proteins also have apoptotic effects, mediated by both caspase-dependent and caspase-independent mechanisms (Viemann *et al.,* 2007).

Mannose binding lectin is a pattern recognition receptor that binds to carbohydrate ligands in bacterial cell walls. Variants in the *MBL2* exon and promoter have been variably associated with susceptibility to severe sepsis, septic shock and mortality. In particular the A/O and O/O exon 1 polymorphism has been associated with low plasma MBL levels and increased susceptibility to severe sepsis, but not with mortality (Gordon *et al.,* 2006) These findings are in contrast with a study of MBL genotypes and plasma levels in 195 patients with bacteremia or pneumonia, where low MBL levels were associated with increased risk of severe sepsis and mortality, although no consistent association was identified with the *MBL2* genotypes (Eisen *et al.,* 2006) A subsequent meta-analysis of 4 studies identified a two fold increase in the odds of death in severe bacterial infections (Eisen *et al.,* 2008) While the increase in odds of mortality for ITU patients was only 1.58, this increased to 5.62 for patients with *Streptococcus pneumonia*. infections, providing evidence for the specificity of the immune response to the invading pathogen.

The NOD receptors are intra-cellular pattern recognition receptors and are organised as inflammasomes, a complex of proteins that regulate the post-translational activa-

tion of the pro-apoptotic protein caspase-1 (Bryant *et al.,* 2009) The relative expression of inflammasome, cytokine and the TLR signalasome transcripts was investigated by qPCR in monocytes derived from patients admitted to ITU with sepsis, those without sepsis and control patients (Fahy *et al.,* 2008) Compared to the critically ill without sepsis, monocytes from patients with sepsis showed decreased production of *CASP1*, *NALP1*, *PYPAF7*, *TNF*, *IL1β*, *IRAK1* and increased expression of *TLR2*. For suppressed transcripts, the levels were comparable to control subjects. The plasma levels of cytokines TNF, IL6 were decreased in non-survivors, along with decreased expression of *NALP1*. The results indicate a hypo-inflammatory state in septic shock, paradoxically associated with anti-apoptosis.

Heat shock proteins

Heat shock proteins are highly conserved chaperone proteins expressed during cellular stress. HSP70 inhibits the production of pro-inflammatory cytokines by macrophages. HSP70 also inhibits apoptosis by activating sphingosine kinase 1 (Ding *et al.,* 2010).

A rare haplotype of HSP70, HSPA1B-179C>T was associated with decreased levels of *HSPA1B* transcript along with a corresponding decrease in protein levels. The same haplotype was also associated with increased levels of *TNF* transcripts with a corresponding increase in protein levels. The expression of *HSP70* was independent of the expression of *TNF* (Kee *et al.,* 2008). Work done in our group has also identified a haplotypic block in *cis* linkage to *HSPA1B*, that explains 29% and 31% of the variance related to HSPA1B levels in lymphoblastoid cell lines at resting state or following heat shock (Maugeri *et al.,* 2010). The haplotypic block shows strong population specific differences, and is commonly seen in individuals with African ancestry.

Covariates

There is convincing evidence for significant modulation of the immune response by age and gender. Commonly used risk prediction scores associate increasing age and the male sex with increased mortality risk.

In mice models of sepsis, female mice have a better immunological response,(Zellweger *et al.,* 1997) are less susceptible to haemorrhagic shock,(Remmers *et al.,* 1998) with a stronger cytokine response(Angele *et al.,* 1999) compared to male mice but have similar mortalities with LPS administration (Laubach *et al.,* 1998) Clinical studies are slightly inconclusive with studies showing either a greater mortality in males(Schröder *et al.,* 2000) or no difference between the sexes (Wichmann *et al.,* 2000).

Increasing age has been associated with immune dysfunction and increase mortality risk. While increasing organ dysfunction with age might contribute to the added mortality risk, age is modelled as an independent risk factor in the risk prediction scores suggesting a specific age related component. In mice models, increasing age is associated with decreasing production of cytokines,(Vega *et al.,* 2004) and declining function of B, T, and antigen presenting cells (Miller, 1996) The decreasing production of cytokines with age has also been demonstrated in humans (Gon *et al.,* 1996).

1.7 Methodological Issues

The choice of samples and methods to investigate functional genomics in clinical settings is not well defined. In particular, the choice of cell types, the effect of RNA isolation methods, and the characteristics of the microarray platform can significantly alter the findings. The potential of functional approaches is, however, generally accepted and it has been suggested that the use of perturbation in gene networks as traits in association studies instead of the genes would be more informative in complex diseases (Y. Chen *et al.,* 2008).

The "Glue Grant" initiative by the National Institute of General Medical Sciences, has attempted to quantify the reproducibility and variance of commonly used methods. Replicate samples derived from healthy volunteers and hybridised to Affymetrix Gene Chips showed strong correlations among the expressed genes (Pearson correlation coefficient 0.985), thereby validating the microarray based methods in investigating genomic responses (Feezor *et al.,* 2004). Gene expression was also studied in samples derived from healthy volunteers which were then stimulated with Staphylococcus en-

terotoxin B and in samples derived from trauma victims. In particular, correlation in gene expression was investigated in each patient group for transcripts derived from whole blood and leucocyte enriched samples. The variance due to the RNA isolation method was greater than the variation due to the nature of the stimulus and trauma in the respective cohorts. In addition, the use of leucocyte enriched samples was associated with greater discriminatory power for the identification of differentially expressed probes.

Genomic signatures relating to clinically relevant phenotypes have been identified with varying degrees of success. Best results have been seen in cancers(Parmigiani *et al.,* 2004) and in auto-immune disease, while predictive models in complex diseases such as sepsis show modest success. In these studies the signatures have been derived from a variety of cell types.

A panel of 1913 genes derived from purified $CD8^+$ cells has been used to successfully predict disease exacerbations in systemic lupus erythematosus and autoimmune vasculitis. (McKinney *et al.,* 2010). The prognostic groups identified by the gene expression signature correlated with different time courses to disease relapse, suggesting an underlying mechanistic basis for the observed phenotypic differences. Intriguingly, the same panel of genes also identified similar clusters in healthy individuals based on a bimodal distribution of expression in the implicated genes.

There are a number of concerns when using mixed leucocyte populations, instead of specific cell types, to identify gene expression signatures. Mixed cell populations are known to misrepresent the genomic changes in specific cell types, but there are sound scientific and logistical justifications for their applicability, including peer use. Statistical methods to identify the contribution of individual components to the cumulative gene expression from a mixed source where the proportion of each component is known have been recently described,(Clarke *et al.,* 2010) but have not been widely used. Previously published studies using mixed cell populations have not addressed this issue in a convincing manner.

2

Aims

Sepsis is a complex pathophysiological entity, with a persistently high mortality and morbidity. Genetic factors are known to influence the risk of sepsis and subsequent clinical outcomes. Therefore, the description of sepsis as a pattern of genomic events in response to an infection will enhance our understanding of sepsis and identify potential diagnostic and prognostic biomarkers, and therapeutic targets.

In the context of this hypothesis, the overall aim of this work is to characterise the patterns of gene expression in severe sepsis and septic shock in peripheral leucocytes derived from patients admitted to Intensive Care Units with severe sepsis due to community-acquired pneumonia and faecal peritonitis.

The specific aims include:

- To identify the genomic response to severe sepsis by comparing samples derived from septic patients to samples derived from appropriate controls,
- To identify genomic signatures associated with survival in severe sepsis and to derive predictive models for survival based on differentially expressed transcripts,
- To identify the genomic effects of important covariates and their influence on survival,
- To investigate the equivalence of diagnostic labels in terms of genomic effects by examining the transcript expression concordance between community-acquired pneumonia and faecal peritonitis, and

- To place the genomic effects of sepsis in the context of other immune phenomena, in particular endotoxin tolerance and the sterile systemic inflammatory response syndrome.

Part II

Methods

3

Materials and Methods

3.1 The Sepsis Cohort

Samples for the study of functional genomics in sepsis were obtained from patients recruited as part of GAinS Project.

The GAinS/GenoSept Project

The GAinS project (`http://www.ukccg-gains.org`) was founded by Prof. Charles Hinds (CH) of Queen Mary's University, and Dr. Chris Garrard (CG) of the John Radcliffe Hospitals NHS Trust, under the auspices of the UK Critical Care Genomics Group to provide a high quality resource for investigating the genetic basis of severe sepsis and septic shock.

The GAinS project is a multi-centre study with patients recruited from intensive care units (ICUs) around the United Kingdom. The primary aim of the GAinS project is to identify genetic variants associated with disease risk and outcome in patients admitted to ICU with severe sepsis and septic shock through an unbiased genome wide association framework. Samples for functional genomics are available in a subset of these patients.

Inclusion Criteria

To avoid biases due to heterogeneity in diagnosis and to improve the strength of associations, recruitment was restricted to adult patients admitted to ICU with severe sepsis due to community-acquired pneumonia or faecal peritonitis.

Standard diagnostic criteria were used. All diagnoses were independently confirmed by a second physician (CH, CG, or Dr. A Gordon of UCH, London).

Sepsis was defined using the ACCP/SCCM criteria(Bone *et al.,* 1992)(Table 1.1). Severe sepsis was defined as sepsis with organ dysfunction while septic shock was defined as persistent hypotension despite adequate fluid resuscitation.

Community acquired pneumonia (CAP) was defined as a febrile illness associated with cough, sputum production, breathlessness, leukocytosis, and radiological features of pneumonia acquired in the community or within less than 2 days of hospital admission (Angus *et al.,* 2002).

Faecal peritonitis (FP) was defined as inflammation of the peritoneal membrane secondary to faecal contamination as diagnosed at laparotomy. Faecal peritonitis due to diverticular disease, breakdown of surgical anastomosis, and penetrating trauma were included, but upper gastrointestinal perforation was excluded.

Exclusion Criteria

Young patients (less than 18 years of age) were excluded from the study. Other exclusion criteria included lack of consent by next of kin or the patient when competent, prior recruitment to a study involving novel therapeutic agents, pregnancy, advanced directives to withhold therapy and patients with immunocompromise such as high dose hydrocortisone therapy, treatment with immunosuppressive agents, and severe neutropenia due to any cause excluding severe sepsis.

Recruitment and the collection of samples and data was done by local investigators. There were no treatments restrictions and the study was purely observational, although all ICUs that collected samples for the functional genomics studies adhered to the "Surviving Sepsis" campaign treatment guidelines.

Consent and Ethics

Written informed consent was obtained in accordance with the UNESCO guidelines for consent in patients unable to provide consent. Patients were recruited for study on the basis of inclusion criteria on admission to the intensive care unit. Study related

information was provided by qualified personnel both verbally and in written format. When possible, consent was obtained from the patient. Where this was not possible due to the severity of illness, consent was obtained from the appropriate professional legal representative, and confirmed when the patient had recovered adequately to make an informed decision. In the event of withdrawal of consent, all samples and data were destroyed.

Ethics approval was obtained from the Central Office of Research and Ethics Committees (COREC) and the local ethics and R&D committees of the participating centres. Samples and data were anonymised using a coded numbering system. Data management was Caldicott compliant and was maintained in separate data storage systems at two different locations. Since the study was observational, there was no direct impact on patient care, or on the legal and human rights of the patients. Since the findings of the study have no direct relevance to patient care, participants do not have access to the genetic data.

The study was conducted in accordance with all relevant ethical and regulatory requirements including the Medical Research Council (MRC) guidelines "Human Tissue and Biological Samples for use in Research" November 1999 and the MRC statement "Human Tissue and Biological Samples for Use in Medical Research" (2001) and issued by the Health Departments and Local Health Authorities.

3.2 The Cardiac Cohort

Samples obtained from patients undergoing cardiac surgery involving cardio-pulmonary bypass (CPB) were used as controls and to investigate the genetic basis of severe inflammation in the absence of an infective component. This population is commonly used because of ease of availability, comparable age distributions, co-morbidities and clearly defined time of injury. The control cohort was comparable to the sepsis cohort with respect to age and sex distribution. Robust matching of controls to cases was not possible to the limitations of sample size.

Inclusion and Exclusion Criteria

Patients scheduled to undergo cardiac surgery (coronary artery bypass grafts and simple valve procedures) were recruited for the study. Patients with complex valvular heart disease were excluded. Minimally invasive procedures, "off-pump" coronary artery grafts, and complex cardiac procedures, and lack of consent were the other exclusion criteria.

Consent and Ethics

Informed consent was obtained from all patients before enrolment in the study. Relevant information was provided by a competent clinician both verbally and in written format.

Ethics approval was obtained from the Riverside Research and Ethics Committee and the local ethics and R&D committees of the Bart's and the London NHS Trust. Samples and data were anonymised and were handled in compliance with Caldicott requirements. The study was observational with no impact on patient care or legal rights.

3.3 Volunteer Cohort

The phenomenon of endotoxin tolerance, and the response of particular cell types to endotoxin was investigated in PBMCs derived from healthy volunteers (Fairfax *et al.*, 2010).

96 healthy volunteers from the Oxfordshire area were recruited to the study. Subjects with a self reported history of infectious disease in the two weeks prior to sampling and pregnant subjects were excluded from the study.

Ethics approval for this study was provided by the Oxfordshire Research Ethics Committee. Written, informed consent was obtained by the principal investigator. All data was anonymised, and sample handling was in accordance with the appropriate regulations.

3.4 Sample Collection, Storage, and Transport

For the sepsis study, samples were collected from patients on days 1, 3, and 5 of ICU admission. Day 1 was defined as the first 24 hours of ICU admission. Additional samples on days 3 and 5 were collected, if the patient survived and remained in the treating unit.

Samples collected on day 1 included buffy coats for DNA extraction, plasma, urine, and samples for RNA extraction. All samples were repeated on days 3 and 5 except for buffy coat samples.

Blood samples were collected from previously established central venous or arterial cannulae with appropriate precautions. Using sterile conditions, blood was collected in two 4.5mL EDTA vacutainer tubes (BD Medical) for buffy coat and plasma samples. The collected blood was centrifuged at 20,000g for 10 minutes at 4°C. The supernatant was sub-aliquoted into four sterile cryovials with approximately 0.5 mls in each vial. The buffy coat was carefully aspirated into a separate cryovial. No effort was made to avoid red cell contamination so as to maximise white cell yield. Downstream protocols were modified to account for blood contamination. Samples were appropriately labelled and stored in a cryobox at -20°C.

Samples for RNA extraction were collected at the bed side using the Ambion LeukoLOCK system(Invitrogen) (Idaghdour *et al.,* 2008) This system consists of leucocyte enrichment filters which trap and lyse white cells, and stabilise the mRNA transcripts. By removing the red blood cells (RBC) , the system avoids globin mRNA bias. Blood for RNA extraction was collected in a 4.5 ml EDTA tube. The leukoLOCK system was assembled as shown in Figure 3.1. The assembly results in a gentle vacuum aspiration of blood across the filtering membrane, by a 10 ml red top vacutainer tube. The filter is then flushed with phosphate buffered saline solution (PBS) . to expel the blood and RNALater to stabilise the RNA in the membranes. RNAlater(Invitrogen) is a proprietary solution containing formaldehyde, that stabilises the cellular RNA transcripts. The filters were appropriately labelled and stored at -20°C avoiding direct exposure to ice.

Urine samples were collected by aspirating urine directly from the indwelling urinary catheters. Urine samples were not collected in anuric patients. 5mL of urine was

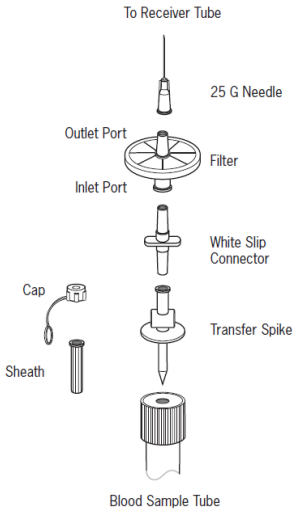


Figure 3.1: Assembly of the LeukoLOCK filter system. *(Adapted from the Invitrogen LeukoLOCK protocol manual)*

collected which was then sub-aliquoted into 2 cryovials. The samples were appropriately labelled and stored at -20°C.

Samples were transported to the Wellcome Trust Centre for Human Genetics in Oxford from the recruiting centres using specialised courier, when the number of samples was large enough to make collection financially sensible. Samples were transported in sealed dry ice containers. On arrival samples were catalogued and stored at -80°C.

3.5 DNA Processing

DNA Extraction

DNA was extracted from buffy coat samples using the manufacturer recommended protocol (DNAeasy blood and tissue kit, Qiagen). The technique involves cell lysis with sequential removal of unwanted cell components such as proteins and RNA, resulting in concentrated high quality DNA.

The sample was mixed with RBC lysis solution in a 1:3 ratio. After adequate time for RBC lysis, the solution was centrifuged and the supernatant discarded. The debris was re-suspended in cell lysis solution, followed by the removal of RNA by degradation with RNase A. Proteins in the solution were precipitated using a protein precipitation solu-

tion and removed by centrifugation. DNA in solution was precipitated by isopropanolol and washed with 70% ethanol. DNA was stored after re-suspension in a hydration solution and quantified using spectrophotometry (Nanodrop 2000, ThermoScientific).

3.6 RNA Processing

RNA Extraction

RNA was extracted from the LeukoLOCK filters using the manufacturer recommended protocol (Ambion LeukoLOCK RNA extraction kit, Invitrogen). The technique involves the elution of lysed cell components from the filter using a solvent, followed by sequential removal of cell components including protein and DNA.

The material retained in the leucocyte enrichment filters was flushed with a pH adjusted, guanidine thiocyanate based lysis solution. The solution lyses white cells and stabilises the mRNA transcripts. After proteinase K digestion of proteins, RNA was precipitated with isopropanalol and bound to magnetic beads. The beads were recovered, and washed to remove reagent traces. The bound nucleotides were then released in solution and DNA digested by Turbo DNase. The beads were recovered using a magnetic rack and washed. The RNA bound to the beads was eluted into cryovials.

Quality Control

The extracted RNA was quantified using spectrophotometry (Nanodrop 2000, ThermoScientific). The spectrophotometer uses transmitted light to estimate the concentration of RNA in a small volume of solution (typically 1μL). Modified Beer-Lambert law is used to convert the absorbence at a particular wavelength to concentration. The choice of wavelength is determined by the molecule to be measured.

The quality of extracted RNA was checked using on-chip gel electrophoresis (Bioanalyzer 2000, Agilent). The process involves a chip containing micro-channels filled with a sieving polymer. Samples in each well are subjected to voltage driven electrophoresis by means of independent micro-electrodes. The migrating molecules are identified by

laser-induced fluorescence and converted to traditional densitometric gel images by software.

3.7 Microarray Methods

Sample Selection

A hybridization experiment was performed using samples from 15 CAP survivors where samples were available at all 3 time points, 15 non-survivors, 15 FP survivors with complete sample sets and 2 FP non-survivors (the entire cohort of non-survivors available at the time), along with 45 samples from 15 patients in the cardiac cohort. The non-survivor cohort lacked the complete set of samples at all three time points either because of death or logistical issues. For the final analysis this data was used as the validation data set.

All samples that passed quality control analysis and had adequate concentrations (237 samples from 128 patients) were used for the derivation data set, along with 33 samples from 11 patients in the cardiac cohort.

Hybridisation

Hybridisation to the Illumina BeadArray microarray platform was performed at the Sanger Institute. mRNA samples for hybridization were provided in a 96 well format, at a concentration of 15 ngμL and 25 μL volume. An automated two step reverse transcription process and transcript amplification is required prior to hybridisation. Samples from the 96 well plates were allocated randomly to the microarray chips to avoid geographic biases in expression estimates.

The BeadArray technology consists of 3 μm silicon chips on which beads are randomly assembled into microscopic wells with Van der Waals and hydrostatic forces. Each bead consists of an address sequence, and multiple copies (several thousand) of the detection sequence. There is one to one correspondence between the address sequence and the probe sequence, thus allowing quantitation of the transcript levels in the sample. There is redundancy in probe detection with multiple beads sharing the same probe

specificity; the numbers vary with the version with 10–15 beads sharing probe specificity in the Human HT12-v4 array and 20–30 beads sharing probe specificity in the Human WG6-v3 array. This unique design of redundant probes in the Illumina platform allows the quantification of relative expression as a mean expression value across all probes and a standard deviation value which is a measure of hybridisation efficiency, non-specific hybridisation and detection biases. Quantitation is performed using the Illumina iScan system, which achieves greater than 99% call rate using high performance lasers and optics.

Samples from the derivation cohort were hybridised to the Illumina HT-12 v4 BeadChips, while samples from the validation cohort were hybridised to the Illumina WG-6 v3 human BeadChips. The characteristics of the two platforms are summarised in Table 3.1. In contrast to the WG-6 v3 BeadChips, the HT-12 v4 BeadChip only contains 99.98% of the listed probes. The random selection of the unrepresented probes in each array reduces the impact of this bias, as does the greater detection accuracy.

The difference in platforms between the two cohorts was the result of improvements in technology and reduction in costs over the course of this project. While the platform differences precludes direct comparison between the cohorts, it allows estimation of platform robust expression changes in contrasts of interest.

Probes	Description	Human WG-6 v3	Human HT-12 v4
RefSeq Content			
NM	Coding transcript, well established annotation	27455	28688
XM	Coding transcript, provisional annotation	7870	11121
NR	Non coding transcript, well established annotation	446	1752
XR	Non coding transcript, provisional annotation	196	2209
Source	RefSeq source release	Rel 36.2	Rel 38
Supplementary Content			
UniGene	Experimental mRNA sequences, that align to EST clusters	12837	3461
Total		**48804**	**47231**

Table 3.1: Comparison of the Illumina BeadChip HT-12 v4 and WG-6 v3 BeadChip platforms (*adapted from Illumina data sheets*)

4

Statistical Methods

4.1 Background

An overview of the statistical methods used in this work is presented in this chapter. Context for these methods with additional details are provided in subsequent chapters. All analysis was performed in the open source statistical package "R" (R Development Core Team, 2011).

4.2 Demographics and Covariates

Basic Comparisons

Demographic and covariates were compared with *t* tests for continuous data, Wilcoxon rank-sum test for ordinal data and Chi-square tests for count data. *p* values were not corrected for multiple comparisons at this stage, since a more robust solution was planned in the downstream analysis.

Covariate Correlations

Correlations between covariates can have considerable implications in regression analysis and can lead to spurious results. The handling of covariates in clinical studies of sepsis has been problematic, and this work uses a systematic approach to examine covariate effects. Three covariate specific analyses were performed: logistic regression, multi-

factor analysis, and Bayesian hierarchical models. The latter is part of a larger analysis and is described later in this chapter.

The effect of covariates on mortality (two week, ICU and hospital) was determined, initially using univariate models. All predictors with significant effects in the univariate models were then used in a multivariate logistic regression model. This approach cannot identify predictors with significant interactions, but is mandated by the large number of covariates with limited data.

The effect of covariates on gene expression was determined using canonical correlation analysis (Soneson *et al.,* 2010; Witten *et al.,* 2009) This differs from standard regression models and Bayesian models by correlating the covariance matrix of all the covariates to the covariance matrix of the entire gene expression data set. Canonical correlation analysis correlates each covariate to each probe while also correlating each covariate and each probe to one another. So if the covariates and probe data was represented as two matrices, the analysis identifies the largest linear combination of covariates that is associated with the largest linear combination of the expression data. The largest linear combination is the first principal component of a matrix, and in essence the analysis involves the projection of the result of principal components of each matrix into a global principal component analysis. The multi-factor analysis builds on canonical correlation analysis by normalising the data to the first eigenvalue, thereby removing the disproportionate effects of factors with large variance.(Bécue-Bertaut *et al.,* 2008) The relationships identified are entirely correlational, and no causal effects can be inferred. This analysis was performed using the package *FactoMineR*.(Lê *et al.,* 2008).

4.3 Analysis of Microarray Data

A summary of the analysis is presented in this section, with a more detailed exposition of the method and rationale in chapter 6 on page 61.

Pre-processing

Data from the hybridization studies were provided by the Sanger institute as plain text files without any pre-processing. The data was directly imported into *R* using the *lumi* package (Du *et al.,* 2008). The data was normalized using the robust spline algorithm after applying a variance stabilizing transformation.

Both the training and validation data set were normalized using the same techniques. An iterative procedure was used to detect and remove outliers. Data was re-normalized after the removal of outliers. Probes with low detection confidence were removed before further analysis.

Exploratory Analysis

Exploratory analysis was performed by hierarchical agglomerative clustering of the principle components of the most variant probes (Lê *et al.,* 2008). Probes were ranked by their coefficient of variance (ratio of mean expression value to the variance). Transcripts with a coefficient of variance greater than 0.9 were considered the most variant.

Hierarchical clustering is an analytical method to identify groups of observations which are similar to each other. Similarity is determined on the basis of a distance metric and linkage criteria. The distance measure differentiates one observation from another while the linkage metric differentiates one group of observations from another. The Euclidean distance which is the distance between two points in a n-dimensional space was used as the distance metric for this analysis. The distance between the centroid of each group of observations was used as the linkage metric. The agglomerative clustering procedure starts with a single observation and then adds other observations in relation to the first using the distance metric, while simultaneously maximizing the linkage metric.

Clustering can be performed on normalized expression values. However, in genomic data sets clustering on the principle components of the expression data results in better yields. Principle component analysis (PCA) is a transformation procedure which converts a set of observations into a set of orthogonal components, where each component is a linear combination of the predictor variables. This is particularly useful in genomic data sets, where the expression of each transcript contributes to different components of

variation. PCA resolves these components into a linear combination of each transcript expression. The first principal component identifies the greatest variation, while the second component relates to the second greatest variation in the data set and so on. In genomic data sets susceptible to interference with a number of "noise" parameters, the last few principle components should map to the noise, while the first few components should capture the most relevant variance in the data set. Analysis can then be restricted to these components.

Differentially Expressed Genes

Linear Models

The use of simple methods like *t* tests to identify differentially expressed genes in genomic data perform reasonably well in comparison to more complex methods like analysis of variance (ANOVA), despite being highly conservative (Jeanmougin *et al.,* 2010). However, the applicability of simple tests in complex study designs is not clear, and use of robust methods is more appropriate.

Linear regression models allow for the simultaneous exploration of the effect of multiple predictors on the dependent variable. This allows the quantification of the individual effect of each predictor by partitioning variance across the predictors. Linear models are limited by their assumptions of homoscedascity (uniform variance of the dependent variable across all values of the predictor variable) and require adequate sample sizes and balance in design for accurate results. In addition, predictor effects in mixed models can only be estimated approximately and are difficult to interpret. Linear regression was done with the *limma* package (Smyth, 2005). The package fits a generalised linear model to the expression of each probe for any specified contrast. An empirical Bayes approach is then used to shrink the estimated variance of a single probe towards a pooled estimate from all the probes, to produce a more stable inference. This procedure estimates the likelihood of the observed difference in expression between contrasts given the overall variance in the data set. Significance is reported using the FDR (FDR).(Smyth, 2004).

Bayesian Hierarchical Models

Bayesian models overcome the limitations of the linear models with the added advantage of a simpler interpretation at the cost of computational complexity. In the simplest format, Bayesian models are similar to simple regression models, where each predictor is associated with a regression coefficient. In simple models, each predictor is associated with a single coefficient, which translates to a set of parallel regression lines corresponding to each level of predictor. The lines have the same slope corresponding to the regression coefficient.

While useful, the simple regression models may not be representative of the study design in complex data sets. For example, in the sepsis data set, different values of microbiological predictors might be reasonably expected to have different effects, mapping to regression lines with different slopes. A hierarchical model allows for such complex inferences, by treating the regression coefficient as a random variable. The determinants of the regression coefficients, also known as the hyperparameters, are estimated from the data. This approach allows for more accurate results, especially in small data sets with a large number of predictors, and overcomes the limitation of the frequentist regression analysis.

Mixed effect multi-level hierarchical models were constructed with the covariates of interest. The effects were estimated by Markov Chain Monte Carlo (MCMC) methods with the Gibbs sampler using the *jags* software with its *R* interface (*rjags*, *R2jags*, and *coda* packages) (Plummer, 2003) 30,000 iterations were run in 3 chains and convergence was verified by the *Rhat* measure and confirmed with statistics from the *coda*. package. Non-informative prior probabilities were used for the training data set which approximates to standard mixed-effect linear models of the frequentist approach. Probes were considered to be significant if the Bayesian 95% credible interval did not include zero.

Pathway Methods

Techniques described in the previous sections generate lists of differentially expressed genes ranked by significance or change in expression. Interpretation of these lists is considerably enhanced by analytical methods which identify clusters of differentially

expressed genes related to biological function. These methods have been variably termed as pathway or network analysis.

DAVID

The DAVID functional classification tool (`http://david.abcc.ncifcrf.gov/`) is an online resource that reduces large gene sets into functional clusters using annotation terms derived from 14 annotation sources (D. W. Huang *et al.,* 2009a,b) The approach is semi-quantitative, where the number of genes of the gene set that are present in the list of differentially expressed genes is used to estimate the Fisher exact *p* value. An enrichment score (the geometric mean of the *p* values) is estimated for each annotation cluster. Larger enrichment scores indicate clusters with the lowest *p*. values, and the greater likelihood of biological significance. The algorithm is fuzzy and allows for a gene to be present in more than one cluster while automatically estimating the optimal number of clusters. In addition, a stringency threshold can be set which filters members of the clusters on the basis of a similarity estimate.

There are three major limitations to the DAVID functional classification algorithm that restrict its applicability. The first is the dependence on the quality of the annotation data from which the results are derived. As a consequence of this dependence significant pathways can only be identified when the pathways have already been defined. This limitation is shared by all network analysis tools, which are therefore blind to novel networks. The second is the failure to use gene specific metrics (either significance or fold change in expression) in the analysis. Since the significance measure of a gene is ignored, each gene in the differentially expressed gene list is assigned equal importance. The third limitation is closely related to the algorithm which mandates the use of the most differentially expressed genes in the analysis. Genes with modest effects are necessarily ignored.

Gene Set Enrichment Analysis

The Gene Set Enrichment Analysis (GSEA) is a quantitative approach to the identification of biological relevance in gene lists. The GSEA uses a system of rewards and penalties

to generate enrichment scores for the given gene sets based on the differences in relative expression or any other test statistic that appropriately ranks the genes (Subramanian *et al.,* 2005). In contrast to the DAVID tool, GSEA uses the entire data set, including genes with modest, minimal, or no effects. The GSEA algorithm overcomes two of the major limitations of the DAVID analysis, but the dependence on annotation source remains.

The algorithm works by ranking genes in the order of significance. The *t* statistic was used as the discriminating statistic in this analysis. Since the magnitude of the *t* statistic indicates the effect size and the sign of the statistic indicates the direction of effect, ranking by the *t* statistic results in a gene list where the genes with the biggest effects are present at the peripheries of the list, with each end corresponding to the opposite direction of change. A running sum is calculated by considering the genes from this list in sequence. For each gene in the list considered in order, if the gene is present in the annotated gene set of interest, the absolute value of the running sum is increased in proportion to the absolute value of the *t* statistic. If the gene is not present in the gene set the absolute value of the running sum is decreased in proportion to the total number of genes in the list. A gene with a high absolute *t* statistic (a greater effect size) will have a greater effect on the running sum compared to a gene with a lower absolute *t* statistic. The enrichment score for the pathway is defined as the maximum deviation of the running sum from zero. A large enrichment score can therefore be a result of a small number of genes with large effect sizes or a large number of genes with modest effect sizes. Since modest changes in genes with regulatory effects can result in large phenotypic changes, the identification of gene sets with large number of genes with modest effects is of importance.

Pathways are lists of genes related to each other either as a sequence of reactants in a biological pathway, involved in a physiological system or in genomic proximity. Pathways for the pathway analysis were obtained from the curated gene lists of the Molecular Signature database (MSigdb, `http://www.broadinstitute.org/gsea/msigdb/index.jsp`). This database annotates gene lists in five clusters; genomic location, motifs, ontology based gene sets, computational gene sets and curated gene sets. Analysis was restricted to the curated gene sets (approximately 2600), which are based

on functions defined from the literature. Only pathways with at least 15 genes were included for analysis. GSEA was used to estimate enrichment score for each pathway for each contrast of interest. Significance levels were determined by permutation analysis (n = 1000) involving phenotype label switching and calculating enrichment scores for the permuted data sets. Significance thresholds were determined by identifying the 95^{th} quantile of the enrichment score distribution. Pathways were considered significant if the permuted *p* values were below 0.05.

4.4 Class Prediction

The prediction of phenotypes based on differentially expressed genes serves two purposes. Predictive models have clinical applications, especially for diagnostic and prognostic variables. Secondly, predictive models validate the robustness of the regression coefficients. Predictive power does not necessarily follow from significant regression coefficients, even with large effect sizes, and the lack of predictive power reduces confidence in the regression estimates.

Class prediction requires a number of steps: feature selection, testing class models, and testing prediction accuracy. A number of algorithms are available for classification of existing gene data and prediction of classes for new data. The appropriate choice of algorithm varies with the data set and no definitive guidelines are available for optimal algorithm choice. Two algorithms with better performance in genomic data sets were used to derive the predictive models.

Support Vector Machines

Given a set of points distributed in a n-dimensional space and derived from two classes, the Support Vector Machine (SVM) algorithm identifies a plane that separates the two classes by maximising the margin between the closest points of the two classes. The margins are termed as the support vectors.

Support vectors were estimated from the derivation data set and validated in the validation data set using the *e1071* package (Dimitriadou *et al.,* 2011). The algorithm re-

quires the optimization of a hyperparameter. The optimal value for the hyperparameter was estimated by trial and error for each model. Genes used in the predictive model were based on a ranked list of differentially expressed genes derived from linear models. The number of genes required for an adequate model was determined by constructing models with 5, 10, 20, 50, 100, and multiples of 100 up to the maximum number of differentially expressed genes. Leave one out cross validation was used to estimate the model. For validation the predictors in the model were restricted to genes present in both data sets. The performance of the model in the validation data set was estimated by calculating the predictive accuracy. Sensitivity and specificity was also calculated for models with binary outcomes.

Random Forests

The random forest algorithm is an ensemble classifier that uses multiple models for classification, with a resultant performance that is greater than any of the constituent models (Caruana *et al.,* 2006). The random forest algorithm uses boot-strapping to construct forests consisting of decision trees. Each decision branch in a tree is based on a random selection of variables, the number of which must be specified in the algorithm. Each observation is assigned to the class to which it belongs in most of the trees in the forest.

Compared to SVM, random forests are accurate and efficient even with very large data sets with large number of predictors. The algorithm generally requires lower number of predictors compared to other prediction methods and has been successfully used in genomic data sets (Goldstein *et al.,* 2010). Validation is built into the algorithm and error is estimated during the model construction. In addition, random forests do not require prior identification of important predictors, and predictors with the greatest predictive power are identified during model construction.

The random forests were constructed in the derivation data set using the *randomForest* package (Liaw *et al.,* 2002). The optimal number of genes required for an adequate predictive model was determined by using different number of genes for the decision

nodes. For each model, genes with the most predictive power were identified. Model effectiveness was checked by calculating misclassification rates.

Part III

The Genomic Response in Sepsis

5

The Sepsis Cohort

5.1 Background

The sepsis cohort was derived from patients admitted to Intensive Care Units (Supplementary Table 1) with severe sepsis or septic shock due to community acquired pneumonia or faecal peritonitis. Relevant, anonymized clinical information was collected using an online data entry system and stored in a relational database. This chapter briefly describes the attributes of the sepsis cohort.

5.2 Demographics

The demographics and other covariates of the CAP sepsis cohort is summarised in Tables 5.1 to 5.10. The demographics for the CAP validation cohort is summarised in Supplementary Tables 3 to 8.

Comparisons were made between survivors and non-survivors using appropriate statistical tests described in section 4.2 on page 41. In the CAP cohort, non-survivors had a lower systolic blood pressure and higher heart rate. There was a tendency towards higher temperatures in non-survivors. Non-survivors had a higher relative neutrophil count, while survivors had a higher relative monocyte count. Non-survivors also showed a tendency towards lower haematocrit. Non-survivors did not differ from survivors in the commonly used biochemical markers including markers of renal function. Non-survivors had significantly lower PF (P_aO_2/F_iO_2) ratio. Survivors and non-survivors were comparable in terms of comorbidities, infecting organisms, additional treatments,

	Overall	Survivors	Non-Survivors	p
Patients (n)	73	60	13	
Samples (n)	129	105	24	
Day 1 samples (n)	40	30	10	
Day 3 samples (n)	47	38	9	
Day 5 samples (n)	42	37	5	
Age in years (Median/IQR)	61 (23)	60.5 (20.8)	64 (28)	0.76
Sex (M:F)	1:0.97	1:1	1:0.86	0.95

Table 5.1: Demographics of the CAP derivation cohort. p values are only shown where relevant.

	Overall	Survivors	Non-Survivors	p
Temperature (°C)				
High	37.83 (0.94)	37.75 (0.91)	38.20 (1.0)	0.06
Low	36.46 (0.81)	36.50 (0.78)	36.25 (0.93)	0.23
SBP (mm Hg)				
High	150 (27)	152 (27)	144 (23)	0.19
Low	96 (18)	98 (19)	87 (15)	0.005
MAP (mm Hg)				
High	100 (19)	101 (19)	99 (18)	0.67
Low	66 (12)	67 (13)	63 (9)	0.13
HR (bpm)				
High	115 (23)	113 (23)	124 (21)	0.03
Low	83 (17)	82 (16)	88 (17)	0.17

Table 5.2: Vital signs of the CAP derivation cohort. SBP = systolic blood pressure, MAP = mean arterial pressure, HR = heart rate

and the number of days of organ support required. Interestingly, the SAPS II score, which predicts risk of mortality, was not different between the survivors and non-survivors. The overall predicted mortality rate is 55.3% but the observed mortality rate is only 17.9%.

The demographics and other covariates of the FP sepsis cohort are summarised in Tables 5.11 to 5.19. Comparisons were made between survivors and non-survivors using appropriate statistical tests described in section 4.2 on page 41. The interpretation of the results must be guarded due to the small number of non-survivors in the FP cohort.

	Overall	Survivors	Non-Survivors	p
WCC (1000 cells/mm^3)				
High	15.4 (8.4)	14.9 (8.7)	17.4 (6.6)	0.12
Low	13.7 (7.6)	13.3 (7.8)	15.3 (6.7)	0.23
DCC (%)				
Neutrophils	83.1 (10.7)	82 (11.4)	88 (5.1)	0.0008
Monocytes	7.0 (6.3)	7.6 (6.6)	4.5 (3.7)	0.003
Lymphocytes	9.9 (6.2)	10.4 (6.6)	8.0 (4.1)	0.03
Platelets	223 (114)	228 (108)	202 (139)	0.41
Haematocrit	26 (17)	29 (16)	18.7 (17)	0.08

Table 5.3: Cell counts in the CAP derivation cohort. WCC = white cell count, DCC = differential cell count

	Overall	Survivors	Non-Survivors	p
Sodium (mmol/L)				
High	139.3 (4.6)	138.9 (4.6)	140.6 (4.4)	0.3
Low	136.1 (5.4)	136.0 (6.0)	136.2 (2.7)	0.89
Potassium (mmol/L)				
High	4.5 (0.6)	4.5 (0.6)	4.7 (0.9)	0.56
Low	3.8 (0.6)	3.8 (0.5)	3.8 (0.8)	0.96
Bilirubin (mmol/L)	26.7 (26.4)	15.3 (25.3)	23.3 (31.3)	0.32
pH	7.33 (0.10)	7.35 (0.13)	7.27 (0.13)	0.11
Bicarbonate (mEq/L)	26.5 (6.7)	27.0 (6.3)	24.5 (8.2)	0.2

Table 5.4: Biochemical markers in the CAP derivation cohort

	Overall	Survivors	Non-Survivors	p
Urea (mmol/L)	10.7 (8.5)	10.7 (8.6)	10.5 (8.2)	0.91
Urine Volume (ml)	2148 (1319)	2181 (1267)	2001 (1551)	0.61
Creatinine (mmol/L)				
High	106 (79)	102 (71)	124 (108)	0.35
Low	99 (73)	95 (60)	117 (112)	0.37

Table 5.5: Markers of renal function in the CAP derivation cohort

	Overall	Survivors	Non-Survivors	p
Respiratory Rate (/min)	29 (9)	29 (9)	27 (13)	0.55
PF ratio	20 (10)	21 (10)	15 (7)	0.003
P_aCO_2 (KPa)	6.1 (1.6)	6.1 (1.5)	6.1 (2.2)	0.97

Table 5.6: Respiratory function in the CAP derivation cohort

	Overall	Survivors	Non-Survivors	p
Cardiac	27	22	5	0.87
Respiratory	24	18	6	0.4
Renal	4	4	0	0.78
Neurological	7	6	1	0.78
Gastrointestinal	2	1	2	0.78
Endocrine	11	10	1	0.71

Table 5.7: Comorbidities in the CAP derivation cohort

	Overall	Survivors	Non-Survivors	p
SAPS II	54 (15)	53 (13)	56 (21)	0.82
Predicted Death Rate	55.3 (33.0)	53.0 (33.0)	59.8 (44.7)	0.82

Table 5.8: Outcome risk in the CAP derivation cohort

	Overall	Survivors	Non-Survivors	p
Therapy				
Steroids	13	11	2	0.86
APC	8	7	1	0.92
Number of days of support				
Respiratory	7 (12)	6 (14)	7 (5.6)	0.46
Renal	1.3 (3.9)	1.4 (4.2)	0.7 (2.0)	0.69
Cardiovascular	1 (6)	1 (5)	2.5 (8)	0.37

Table 5.9: Treatment factors in the CAP derivation cohort. APC = activated protein C

	Overall	Survivors	Non-Survivors	p
Pneumococcus	19	15	4	0.73
Haemophilus	1	1	0	
Pseudomonas	4	4	0	
Staphylococcus	5	5	0	
Legionella	4	3	1	
Mycoplasma	1	1	0	
Mixed	4		1	
Other	16	13	3	
Viral	10	10	0	

Table 5.10: Microbiology of the CAP derivation cohort

Compared to the CAP cohort, the FP cohort has significant differences in demographics between survivors and non-survivors. In particular, mortality is higher in the elderly and females, but the latter association is not significant. Non-survivors in the FP cohort have lower temperatures and mean arterial pressures. There is a tendency towards lower systolic blood pressure. Non-survivors also have higher sodium and lower pH levels. Oxygenation index is significantly low in non-survivors. There are also significant differences in the incidence of comorbidities between survivors and non-survivors in the FP cohort. A larger proportion of non-survivors were treated with steroids. Non-survivors also required inotropic support for more days than survivors. There is a clear difference in the predicted death rate and the SAPS II scores between the two groups. However, the overall mortality rate is 12.5% against a predicted mortality of 66%.

5.3 Summary

The demographics of the derivation cohort was summarised in this chapter. Attributes for the CAP and FP cohorts were estimated separately, since diagnostic equivalency is not assumed in this analysis. The differences in the covariates between survivors and non-survivors were calculated. The overall mortality in the CAP and the FP cohorts was 17.9% and 12.5% respectively.

	Overall	Survivors	Non-Survivors	p
Patients (n)	64	56	8	
Samples (n)	94	83	11	
Day 1 samples (n)	37	31	6	
Day 3 samples (n)	34	31	3	
Day 5 samples (n)	22	21	1	
Age in years (Median/IQR)	68 (27.3)	66.5 (30.3)	80.5 (8.8)	0.006
Sex (M:F)	1:1.3	1:1.5	1:3	0.45

Table 5.11: Demographics of the FP derivation cohort. p values are only shown where relevant.

	Overall	Survivors	Non-Survivors	p
Temperature (°C)				
High	37.51 (0.80)	37.51 (0.79)	37.52 (.92)	0.98
Low	36.18 (0.69)	36.25 (0.69)	35.66 (0.47)	0.002
SBP (mm Hg)				
High	154 (22)	154 (22)	155 (25)	0.88
Low	104 (20)	105 (20)	95 (16)	0.08
MAP (mm Hg)				
High	101 (19)	102 (19)	92 (18)	0.004
Low	69 (13)	70 (14)	60 (7)	0.0003
HR (bpm)				
High	113 (21)	112 (20)	122 (26)	0.24
Low	80 (14)	80 (14)	77 (14)	0.45

Table 5.12: Vital signs of the FP derivation cohort. SBP = systolic blood pressure, MAP = mean arterial pressure, HR = heart rate

	Overall	Survivors	Non-Survivors	p
WCC (1000 cells/mm^3)				
High	14.6 (7.5)	14.2 (6.9)	17.9 (10.5)	0.28
Low	12.6 (7.7)	13.2 (7.2)	15.3 (10.7)	0.38
DCC (%)				
Neutrophils	87.3 (8.9)	87.1 (9.1)	88.5 (7.4)	0.59
Monocytes	5.9 (7.8)	6.0 (8.1)	5.8 (5.8)	0.93
Lymphocytes	6.7 (3.8)	6.9 (3.9)	5.7 (2.8)	0.23
Platelets	234 (143)	241 (147)	184 (108)	0.18
Haematocrit	21 (14)	23 (14)	14 (13)	0.16

Table 5.13: Cell counts in the FP derivation cohort. WCC = white cell count, DCC = differential cell count

	Overall	Survivors	Non-Survivors	p
Sodium (mmol/L)				
High	139.0 (5.5)	138.1 (5.2)	143.7 (4.8)	0.04
Low	136.4 (5.8)	135.4 (5.4)	141.2 (5.4)	0.03
Potassium (mmol/L)				
High	4.5 (0.4)	4.5 (0.4)	4.4 (0.3)	0.82
Low	3.8 (0.5)	3.8 (0.5)	3.8 (0.2)	0.93
Bilirubin (mmol/L)	15.0 (13.5)	15.2 (13.8)	13.9 (11.8)	0.75
pH	7.33 (0.10)	7.34 (0.09)	7.25 (0.08)	0.03
Bicarbonate (mEq/L)	22.8 (5.3)	22.9 (5.5)	21.6 (4.3)	0.40

Table 5.14: Biochemical markers in the FP derivation cohort

	Overall	Survivors	Non-Survivors	p
Urea (mmol/L)	12.5 (8.5)	12.3 (8.8)	14.1 (5.9)	0.42
Urine Volume (ml)	1663 (1333)	1719 (1385)	1248 (777)	0.11
Creatinine (mmol/L)				
High	122 (89)	119 (92)	144 (62)	0.26
Low	116 (87)	112 (89)	150 (61)	0.12

Table 5.15: Markers of renal function in the FP derivation cohort

	Overall	Survivors	Non-Survivors	p
Respiratory Rate (/min)	22 (8)	22 (8)	21 (10)	0.81
PF Ratio	30 (13)	31 (13)	20 (7)	0.0004
P_aCO_2 (KPa)	5.1 (1.0)	5.1 (1.0)	5.0 (1.1)	0.95

Table 5.16: Respiratory function in the FP derivation cohort

	Overall	Survivors	Non-Survivors	p
Cardiac	28	22	6	1×10^{-7}
Respiratory	14	13	1	4×10^{-12}
Renal	8	6	2	2×10^{-16}
Neurological	7	7	0	2×10^{-16}
Gastrointestinal	13	11	2	1×10^{-13}
Endocrine	7	5	2	2×10^{-16}

Table 5.17: Comorbidities in the FP derivation cohort

	Overall	Survivors	Non-Survivors	p
SAPS II	59 (13)	58 (12)	65 (7)	0.01
Predicted Death Rate	66.1 (25.0)	64.0 (24.6)	76.9 (11.6)	0.01

Table 5.18: Outcome risk in the FP derivation cohort

	Overall	Survivors	Non-Survivors	p
Therapy				
Steroids	14	11	3	0.0009
APC	2	2	0	0.16
Number of days of support				
Respiratory	2 (4)	2 (4)	4 (1.5)	0.12
Renal	0.5 (1.4)	0 (0)	0 (0)	0.96
Cardiovascular	2 (2.8)	2 (3.5)	3 (0.5)	0.06

Table 5.19: Treatment factors in the FP derivation cohort. APC = activated protein C

There are some significant differences in covariates between survivors and non-survivors that are commonly shared in the CAP and FP cohorts. In particular, non-survivors tend to have lower perfusion pressures, temperature anomalies, and worse PF ratios. Other covariates are significantly different between survivors and non-survivors but only in one cohort. In particular, in CAP there are strong associations with relative white cell counts, while in FP there are strong associations with sodium and pH levels. Interestingly, there is considerable error in mortality prediction using the commonly used SAPS II score. The SAPS II score overestimates mortality by 3 fold in the CAP cohort and more than 5 fold in the FP cohort.

6

Microarray Data: Quality Control and Normalization

6.1 Background

Whole genome gene expression profiles were obtained by hybridising mRNA samples to the Illumina BeadChip arrays. The Illumina BeadArray platform has a unique design (section 3.7 on page 38), that allows for detection of relative expression levels with greater accuracy.

Adequate preprocessing of gene expression data is of critical importance to allow comparison of data from different arrays (Dunning *et al.,* 2008) Microarray data is susceptible to batch effects related to the experimental conditions during the hybridization process. Data from the hybridisation experiment are expression values relative to an undefined baseline. The aims of pre-processing are to reduce noise and to allow comparison of data between arrays. Preprocessing involves a number of steps aimed at reducing bias by homogenising variance and improving detection rates by preserving the fold change differences. Unfortunately both aims are antagonistic and an appropriate pre-processing strategy involves determining a suitable balance between these imperatives, individualised for a given data set.(Schmid *et al.,* 2010).

This chapter describes data import, the pre-processing strategy along with rationale for the methods used, and the generation of the final data set used in this work.

6.2 Data Import

The Illumina Data Format and Analysis Tools

The Illumina iScan system outputs relative expression data for each array, which can then be read by the Illumina BeadStudio software (`http://www.illumina.com/gsp/genomestudio_software.ilmn`). The GX (Gene Expression) module of the software is designed to analyse gene expression data and was used to verify data integrity and perform basic quality control checks. While the GX module is useful for the initial processing, the analysis pipeline is rudimentary and no further processing was performed on the BeadStudio platform. Raw data was exported into text files which were used in all downstream analyses. Results for each plate of 96 samples were made available as an independent data file. Data for the derivation and the validation data sets were processed separately.

The data file exported from BeadStudio consists of one line for each probe. Values for each sample are displayed in 8 columns, and include the minimum, maximum, and average signal for the probe, along with the number of beads containing the probe, the standard deviation of the signal among the beads and a detection p value. The latter is a confidence measure that the transcript is expressed above the background noise. The background noise is estimated by the BeadStudio software from the negative control probes incorporated in the array. The remaining two columns are not relevant to this study design.

All analysis was performed with the R statistical package (R Development Core Team, 2011) with additional packages from the Bioconductor repository (Gentleman *et al.,* 2004). The Bioconductor *lumi* package (Du *et al.,* 2008) was used to import data into the R system; the package imports gene expression data along with relevant phenotype and annotation information and converts the gene expression data from Illumina BeadArray format into BioConductor *ExpressionSet* format. This format stores the gene expression data along with phenotype and annotation data as a single dataset, and maintains a record of all modifications performed on the data.

All data from each experimental run was combined into a single data set. Thus each data set consists of samples from the sepsis cohort, including both CAP and FP samples, and the cardiac cohort. A separate data set was created for the derivation and validation data set.

Control Data

The Illumina BeadArray platform includes oligonucleotides that act as control probes. This data is output separately and includes six types of controls: biotin, CY3, house keeping genes, labelling genes, low stringency genes and negative probes. The biotin controls are signal generation controls reflecting successful secondary staining. CY3 controls consist of CY3 labelled probes that generate a signal gradient dependent on concentration but independent of RNA quality. House keeping genes consist of probes that should be expressed in every sample. Labelling genes consist of probes that measure the success of the reverse transcription reaction. Low stringency genes consist of probes containing two mismatched bases. With adequate stringency control, these probes should have a low yield. Negative controls consist of random oligonucleotide sequences with no corresponding sequence in the genome. The signal from these probes determine the background noise.

The distribution of these probe groups in the validation data set is shown in Figure 6.1. Each probe group contains a set of related probes. For example the house keeping group is composed of 7 probes; the distribution of these probes are shown in Figure 6.2. Examination of the control data shows strong array specific effects and the presence of outliers.

6.3 Preprocessing

The expression value for a probe is a measure of the optical intensity reflecting hybridisation at the probe. As such it is a relative measure with an indeterminate baseline unique to each array. Therefore, expression values for a probe from different arrays cannot be directly compared. Data from different arrays must be normalized before comparisons

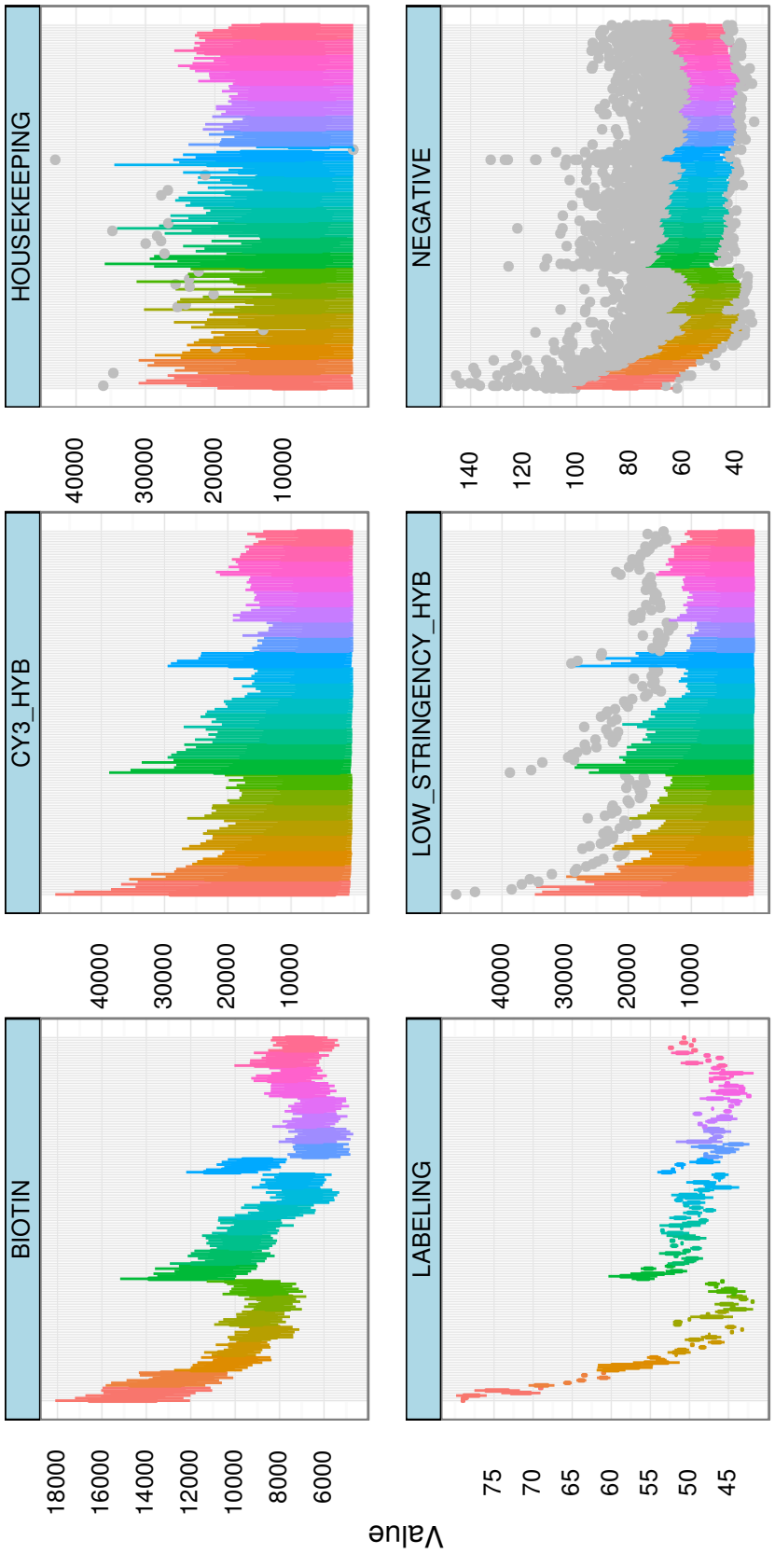


Figure 6.1: Distribution of control probes before normalization. Each panel shows the relative, non-normalized expression of a group of genes for all samples. The genes are detailed in the text. Samples are coloured according to the array to which they were hybridized. There is considerable variation in the expression values both within and across arrays. The differences in expression between arrays results in array specific biases. The clustering of colours, for example the green and blue peaks in the Biotin group, indicate array specific effects.

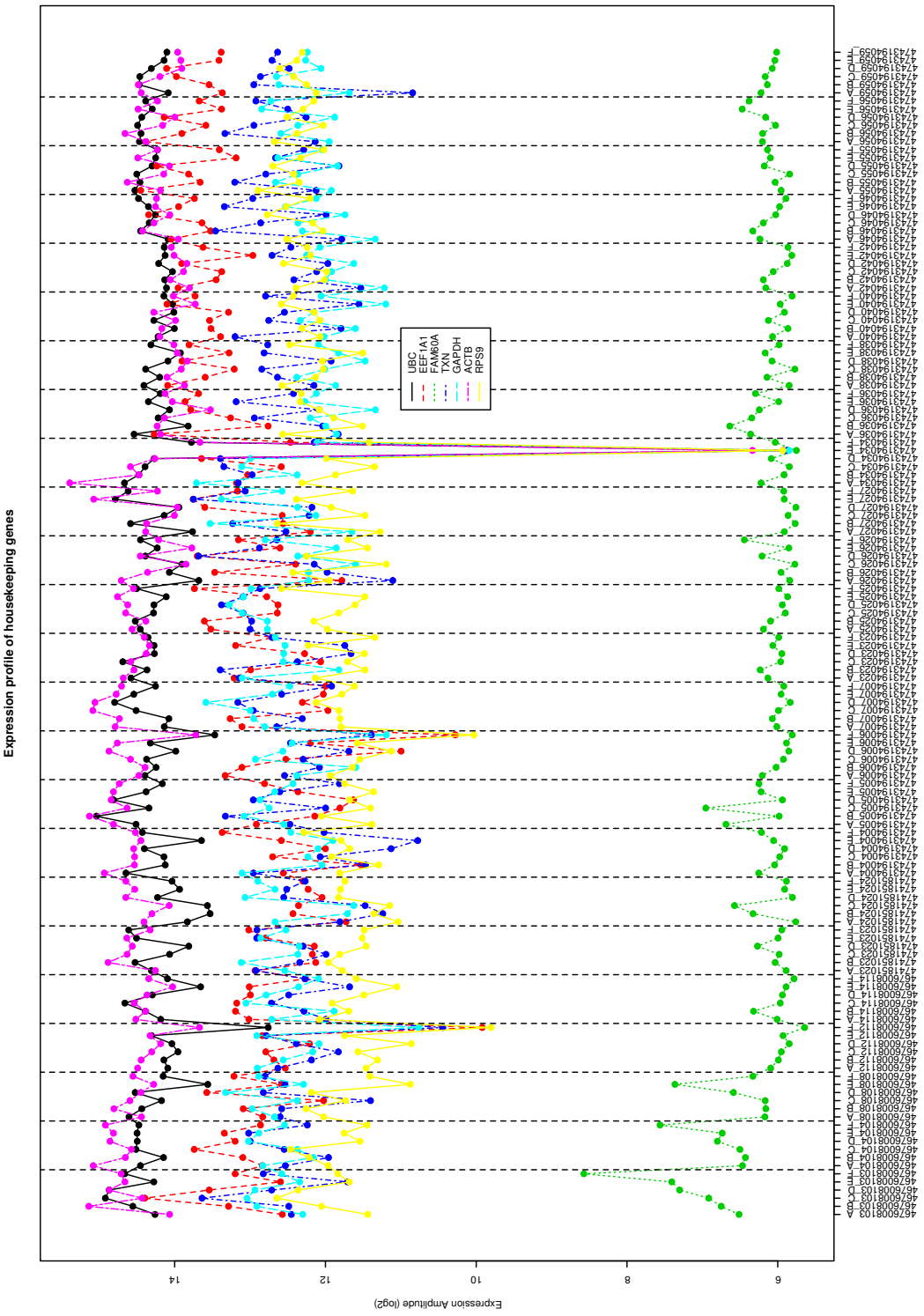


Figure 6.2: Expression of the house keeping genes before normalization. The relative, non-normalized expression of 7 house-keeping genes included in the array are displayed for all the samples. House keeping genes have a relatively constant expression across cells irrespective of experimental conditions. There are large fluctuations from the baseline for each of the house-keeping gene reflecting the lack of an absolute baseline in microarray studies. In addition, there is one large outlier with very low expression values for all the house keeping genes suggesting hybridisation failure. Data from this sample was removed during normalization.

across arrays. Normalization converts the raw expression values across multiple arrays to comparable values across the entire data set.

Three distinct preprocessing steps are required to convert the data into a usable format: background correction, transformation, and normalization. For each cohort (derivation and validation) the entire data set was processed together; the data set was not segregated into diagnostic subsets or control samples before preprocessing and normalization.

Background Correction

Background correction removes the background noise derived from non-specific hybridisation in the data set. Illumina arrays include negative control probes that do not correspond to known oligonucleotides which allow for accurate estimation of the background noise. Illumina recommends that raw data values are background corrected before further analysis. Background processing was done using the GX module of the BeadStudio software before the data was exported for further processing.

Transformation

Pre-processing involves a set of transformations of the data which allows for data from different arrays to be compared while preserving the differences in expression values of the probes. One of the major problems in microarray data, especially with small data sets, is the heterogeneity of variance between samples (Schmid *et al.,* 2010). Large differences in the variance of data sets increases noise and results in false negatives. Also, in most biological data sets, the variance of the data set depends on the mean value of the data set. Typically variance depends on the mean expression value, with increasing mean expression associated with increasing variance and uncertainty. The dependence of variance on the mean violates one of the assumptions of regression analysis. A variance stabilizing transformation is therefore necessary for small data sets. However, these transformations reduce the fold change difference between probes and decrease gene yield (false negatives, with the failure to identify probes with true differences in expression).

Variance stabilizing transformation (VST) is a method of data transformation which removes the dependence of variance on the mean and results in homogeneous variance across data sets (Schmid *et al.,* 2010) A number of transformation methods are available with different properties. The redundancy of beads in the Illumina array design with error estimates for hybridization of each probe allows for a specific VST which utilises the extra information to improve the transformation characteristics. However, this method has been shown to perform poorly with HT-12 arrays, probably because of lesser redundancy features of the array.(ibid.). One of the alternatives is a simple logarithmic transformation, which does not stabilise variance but converts the exponentially distributed raw data values to normal scale, while preserving probe expression rank.

The derivation data set is a large data set and therefore has less problems with variance heterogeneity. This data set was log2 transformed; the choice also determined by the normalization strategy described in the next section. The validation data set is a smaller data set and is better suited to variance stabilizing transformation but was also log2 transformed to allow direct comparisons. Since only a limited number of comparisons were to be performed in the validation data set on probes identified from the derivation data set, the heterogeneity of variance was not anticipated to cause major problems.

Normalization

As with transformation, a number of normalization strategies with different properties are available, with no clear guidelines on the appropriate choice of algorithm (ibid.).

The aim for normalization is to allow comparison of data across different arrays; normalization converts the raw expression values which are relative to an arbitrary baseline unique to each array, to an expression value relative to an arbitrary baseline shared across all arrays. Normalization approaches depend on the fact that the expression of most probes remain unchanged across data sets. If the expression of these probes across all arrays can be converted to a common baseline, all other values values can be expressed relative to this baseline.

In practice there are a number of choices for the baseline probes. One simple option is to base the reference point on known housekeeping genes, but this choice consistently under-performs compared to other options (Schmid *et al.,* 2010). Commonly the mean expression of all probes is chosen as the baseline value. Since most probes are expressed unchanged across arrays, this usually gives a consistent value across arrays and has better performance characteristics.

There are also additional choices in the methods for converting the expression data relative to the baseline. The choices can be differentiated on the basis of a number of parameters, two of which are particularly relevant — the preservation of fold changes between contrasts of interest and the preservation of expression rank — identical ranking of probes based on expression values before and after normalization.

Commonly used methods include the quantile normalization where each array is classified into quantiles which are then made comparable across arrays by matching each quantile of each array to the corresponding quantile in a reference array (Bolstad *et al.,* 2003). The reference array is usually derived from the cumulative distribution of all arrays. The quantile distribution preserves the rank order of genes but results in discontinuous intensity transformations. A popular alternative is spline normalization which results in continuous transformations but does not preserve the rank order of the genes.

The authors of the *lumi* package recommend robust spline normalization (RSN) for Illumina arrays. The RSN method integrates both approaches while circumventing these problems. A quantile normalization is first performed and putative differentially expressed genes are identified. A monotonic spline normalization is then performed with a locally weighted regression (loess) function, where the contributions of the differentially expressed genes are down-weighted, using weights determined from the quantile normalization. This approach results in preserving the rank order of genes while simultaneously maintaining the effect sizes of the differentially expressed genes.

An iterative procedure was used to normalize the data sets, using the RSN algorithm. In each iteration, outlier samples were identified after the normalization procedure. To detect outlier samples, the means of the samples excluding the extreme 10% samples was

calculated. The euclidean distance of the mean of each sample to this centre (mean of means) was determined. Outliers were defined as samples whose distances were at least 2 standard deviations away from the mean distance to the centre. Normalization was repeated after removing the outliers and the procedure continued till no outliers could be identified.

For the derivation data set, 2 iterations were necessary. Seven samples were removed as outliers, including one control sample. For the validation data set, one sample was removed after 2 iterations. This sample corresponds to the outlier visualised in Figure 6.2.

6.4 Finalization

Filtration

Probes whose expression levels could not be differentiated from background noise with statistical confidence (defined as detection *p* value less than 0.01) were excluded from further analysis. The filtration of the data set reduces analysis complexity and the issues with multiple comparisons.

After filtration about 29,117 probes were retained in the derivation data set and 21,102 probes were retained in the validation data set. In similar data sets reported in the previous literature, probes which do not correspond to known genes were also excluded from analysis, to improve the interpretation of results (Checchia *et al.,* 2008; Dixon *et al.,* 2007; Tofil *et al.,* 2007). This approach presupposes definite knowledge of mechanisms and was not done in this analysis. All probes were retained to identify novel transcripts that have not been previously reported in sepsis.

Quality Control

The normalization efficacy and the correlation between replicates were verified by examining the distribution of expression values. The effect of normalization on the derivation and the validation data sets is shown in Figures 6.3 and 6.4 respectively.

Both figures show the same information while retaining adequate visual clarity. The box and whisker plots (Figure 6.4) provides greater visual confirmation of normalization effects, but has reduced clarity for data sets with a large number of samples, where an overlapping density plot (Figure 6.3) is more appropriate. The post-normalization plots show alignment of the median expression values and relative homogeneity of variance across samples.

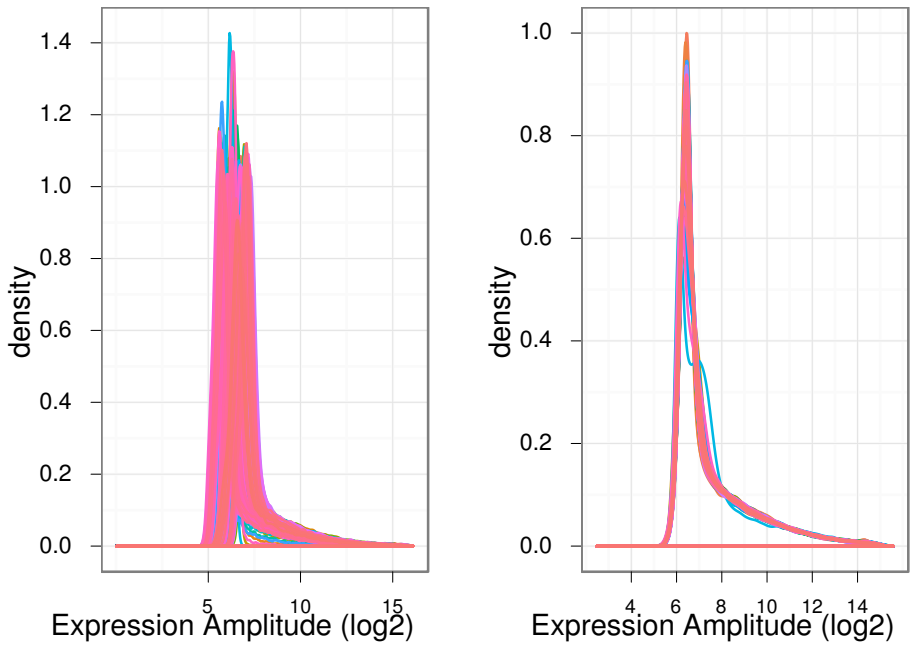


Figure 6.3: Distribution of expression value densities in the derivation data set before (left panel) and after (right panel) normalization. There is better alignment of values after normalization.

The quality of each array can be confirmed by examining the MA plots which relate the intensity ratio (M) of each probe to the average intensity (A) (Y. H. Yang *et al.*, 2002) A representative plot is shown in Figure 6.5. The median values of the MA plot is an index of the efficacy of normalization. A properly normalized sample has a median value approaching zero and indicates that the variance is not dependent on the mean expression value. The very low median values in the data set validates the normalization process.

Three technical replicates were included in both the derivation and the validation data sets. Correlation coefficients of 0.97 or greater were observed in each of the paired

Figure 6.4: Distribution of expression values in the validation data set before (top panel) and after (bottom panel) normalization. There is better alignment of values after normalization and more homogeneity in the variance.

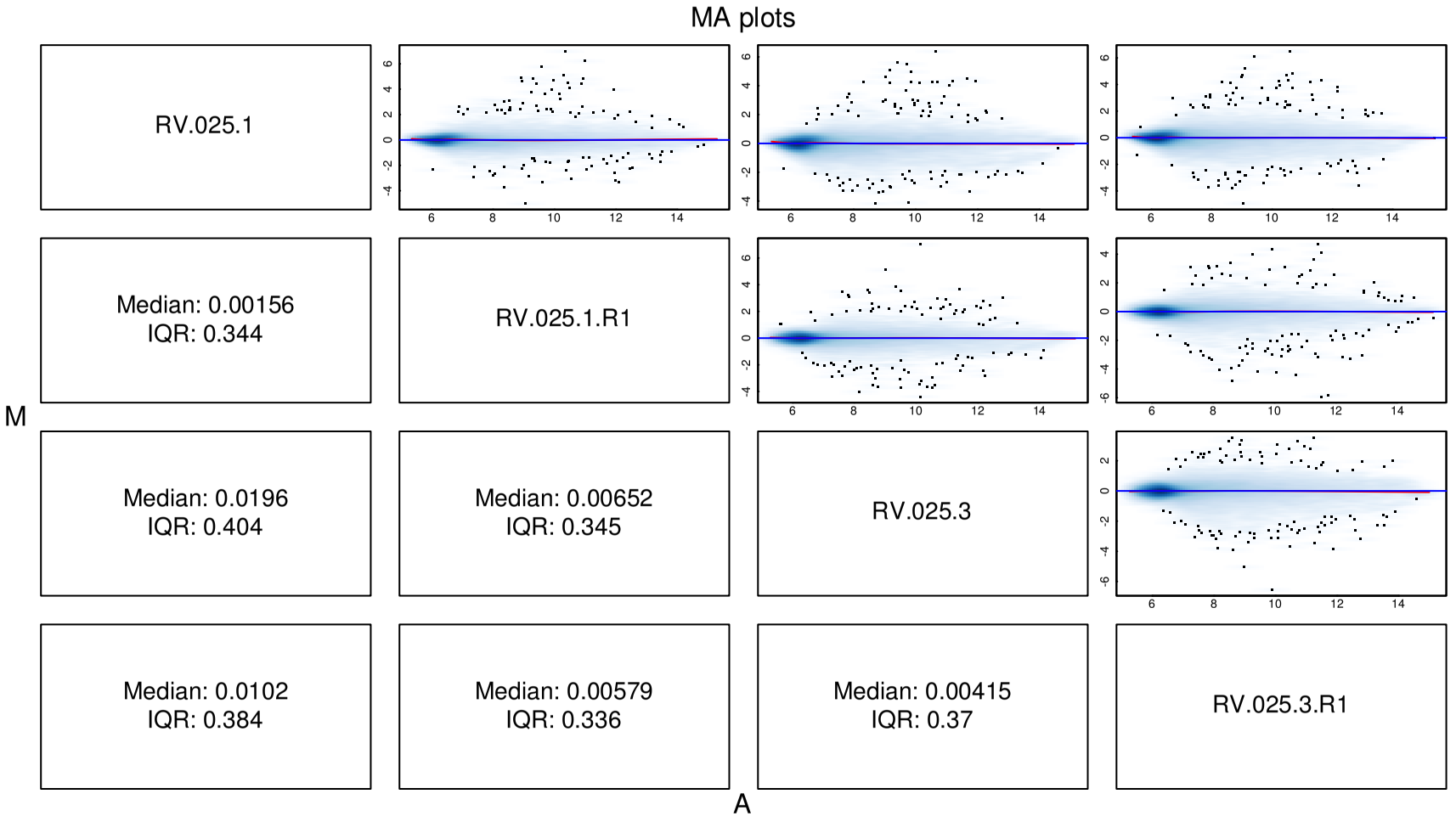


Figure 6.5: Plot of intensity ratio (M) to average intensity (A) for a random selection of samples in the derivation data set. The median values (red line) in each plot is an index of normalization efficiency. A value of 0 suggests adequate normalization.

samples. The correlation between replicates in the derivation data set are shown in Figure 6.6.

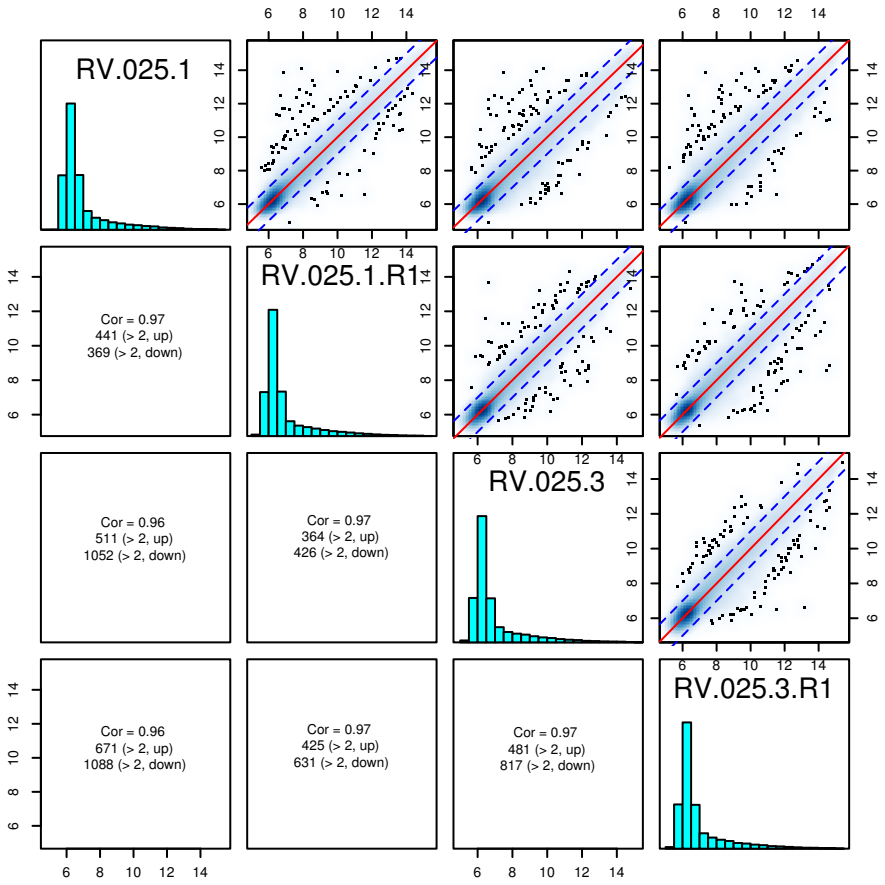


Figure 6.6: Correlation between technical replicates in the derivation data set. Each panel plots the expression value of each probe in each of the two samples compared. RV.025.1 and RV.025.1.R1 are technical replicates as are RV.025.3 and RV.025.3.R3. Both samples are from the same patient but at two different time points. As expected strong correlation is observed between the technical replicates. There is also strong correlation between samples from the same patient.

The absence of systematic technical biases in the expression data was confirmed by principal component analysis of the normalized data sets. The PCA plot of the derivation data set shown in Figure 6.7, reflects the absence of array specific effects seen in the pre-normalization data.

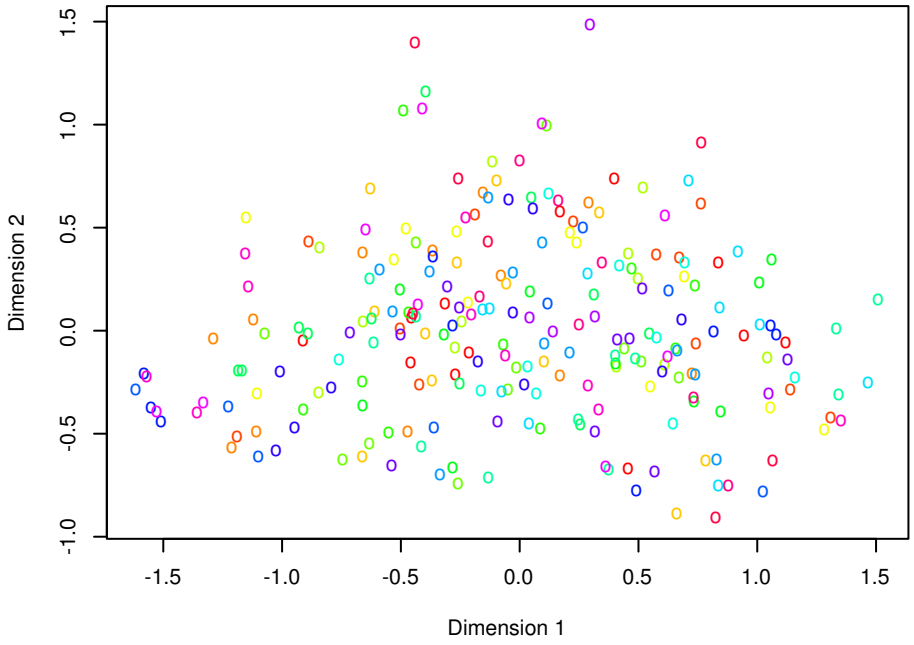


Figure 6.7: Principle components of the expression values of all the probes in the derivation data set after normalization and filtering. Dimension 1 is the first principal component and corresponds to the greatest variance in the data set. Dimension 2 is the second principal component and corresponds to the second largest variance in the data set. Each point represents one sample and is coloured by the chip to which it was hybridised. There is no discernible pattern of distribution of colours.

Final Data Sets

The final data sets consist of samples which had been adequately normalized and filtered. Outliers, samples with unusual distributions, and replicates were removed from the final data sets. 29,117 probes in 233 samples from the derivation data set and 21,102 probes in 101 samples from the validation data set were retained in final data sets.

6.5 Summary

Whole genome gene expression data derived by microarray hybridisation technology provides large amounts of useful information relating to genomic states related to disease or other stimuli. However, to obtain this data in a useful format requires careful pre-processing of the data.

The sepsis data set was background corrected to remove spurious information, log2 transformed to retain fold change differences and normalized using a robust spline algorithm to maintain rank order. Seven samples from the derivation data set and one sample from the validation data set had unusual pattern of gene expression and were removed as outliers. The quality of the normalization process was verified by ensuring the homogeneity of variance, and the lack of dependence of variance on the mean.

The normalized data set was filtered to remove non-informative probes. The normalized and filtered data does not show biases related to array hybridization. The final data set, consisting of 29,117 probes from 233 samples of the derivation cohort, and 21,102 probes from 101 samples of the validation cohort, along with associated phenotype and annotation data was aggregated and stored in a useful data format.

7

Covariates

7.1 Background

Sepsis is a complex clinical phenomenon modulated by a number of factors. Even though some covariates such as increasing age and male sex have been associated with poor outcomes in sepsis, the handling and interpretation of covariates in clinical studies of sepsis has been problematic, even with well designed randomised controlled trials. Outcomes in sepsis result from complex physiological events triggered by infection, and relating to individual organ function, their inter-relationships and effects on cellular events. Outcome is also influenced by therapy and clinical management, including the appropriate choice and end points for fluid therapy, appropriate choice and timing of antibiotics, appropriate choice of inotropes and vasopressors, appropriate use of renal replacement therapy, appropriate choice, timing and route of nutrition and appropriate choice of respiratory support. Much of the clinical management in sepsis is uncertain, reflecting the complexity of the underlying disease.

One of the difficult problem with sepsis covariates, especially in the context of gene expression studies, is the issue of cell counts. An altered white cell count is one of the diagnostic criteria for sepsis, but the pattern of changes in white cell subtypes is neither quantified nor considered in most clinical investigations in sepsis. Changes in white cell counts do not seem to influence sepsis outcomes except in severe neutropenia. However, gene expression in leucocytes is sensitive to changes in differential cell counts, with some evidence that the cumulative mixed white cell expression is poorly correlated with cell type specific changes.

The traditional approach in sepsis studies has been to largely ignore the covariate complexity. Of the many possible covariates only a few have been considered, usually for controlling group equivalency in randomised controlled trials using risk scores. Sepsis risk prediction scores such as APACHE II(Knaus *et al.,* 1985) and SAPS II(Le Gall *et al.,* 1993), use a compendium of easily measured clinical and biochemical factors to predict mortality rates for patient groups. The covariates include a mixture of biochemical markers and measures of baseline physiological function and chronic disease status. The main utility of these markers lie in risk prediction for groups of patients and have limited utility in prognostication for any individual patient, though it is commonly misinterpreted as such.

In terms of mechanistic explanations, only age and sex have been given any serious consideration, and hard data exists for sex related effects on sepsis outcomes.

This chapter examines the covariates associated with sepsis, particularly their relationship to survival, and an overview of their relationship to gene expression. The effect of specific covariates is examined in detail in the following chapters.

7.2 Relationships to Survival

Information about 64 covariates were collected in a usable form for this analysis. The relationship of the covariates to survival was examined by logistic regression analysis. Since the number of predictor variables is high compared to the number of observations, a two stage approach was used where each predictor was first independently associated with survival and in the second stage a regression model was constructed with all predictors that were statistically significant in the first stage. This approach is limited by the inability to identify predictors that have significant interaction effects. To improve the power of the analysis, data from both the CAP and FP cohorts were used in the regression model.

The outcome variables for the linear regression model were survival at two weeks, four weeks, at ICU discharge and at hospital discharge. The predictor variables are listed in Table 7.1.

Demographics	**Organ Support**	**Microbiology**
Age	Number of days of renal, respiratory, and inotropic support	Isolated organism
Sex		
Diagnosis		
Therapy	**Vitals**	**Haematology**
Activated Protein C	Temperature	White cell count
Steroids	Systolic blood pressure	Differential cell count
Inotropes	Mean arterial pressure	Platelets
	Heart rate	Haematocrit
Biochemistry	**Acid-Base**	**Renal function**
Sodium	pH	Renal failure
Potassium	Bicarbonate	Renal replacement therapy
Bilirubin		Urea
		Creatinine
		Urine volume
Respiratory function	**Comorbidity**	**Scores**
Ventilation	Cardiac	SAPS II
P_aO_2	Respiratory	
F_iO_2	Renal	
P_aCO_2	Neurological	
	Endocrine	
	Gastrointestinal	
	Immune	
	Infection	
	Endocrine	

Table 7.1: Covariates used as predictors in the univariate and multivariate regression models

For survival at two weeks following ICU admission, 15 variables were significantly associated with outcome on independent comparisons. These were age, lowest recorded temperature in the 24 hour period, lowest systolic blood pressure, lowest mean arterial pressure, number of days of inotrope support, highest recorded heart rate, haematocrit, highest sodium levels, pH, bicarbonate levels, creatinine levels both high and low, partial pressure of oxygen in the arterial blood, inspired oxygen concentrations, and neutrophil count.

With a full model only three of these variables were significantly related to outcome (Table 7.2). Each unit increase in haematocrit decreases the odds of death by 1.04 (Figure 7.1) and each unit increase in PF ratio decreases the odds of death by 1.13

	Estimate	Std Error	p
Intercept	1.75	0.96	0.07
Haematocrit	-0.05	0.02	0.02
Oxygenation Index	-0.13	0.05	0.006

Table 7.2: Logistic regression model for survival at two weeks. Oxygenation index is defined as the ratio of arterial partial pressure of oxygen to the inspired oxygen concentration.

(Figure 7.2). The failure to deliver adequate oxygen to tissues is associated with death, with changes in PF ratio resulting in greater effects than change in haematocrit. The reduction in the number of significant predictors from 15 in the univariate model to three in the multivariate model suggests significant correlation between the predictors.

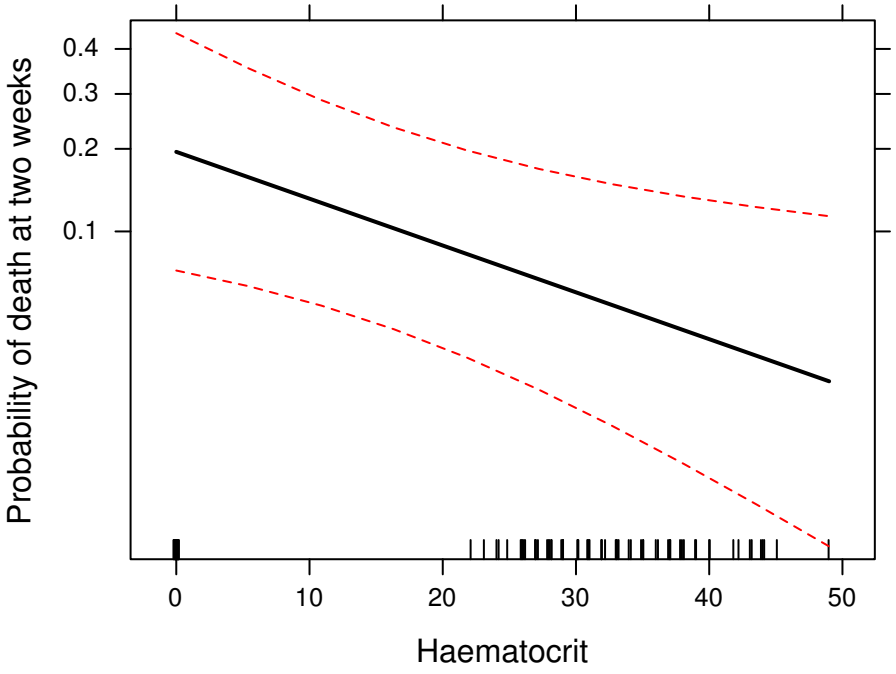


Figure 7.1: Effect of haematocrit on two week survival. Tick marks in the horizontal axis show the observed haematocrit values.

For survival at four weeks following ICU admission, 14 variables were significantly associated with survival. These include temperature — both high and low, lowest systolic blood pressure, lowest mean arterial pressure, number of days of inotropic support, highest heart rate, white cell counts, haematocrit, sodium levels, pH, PF ratio and

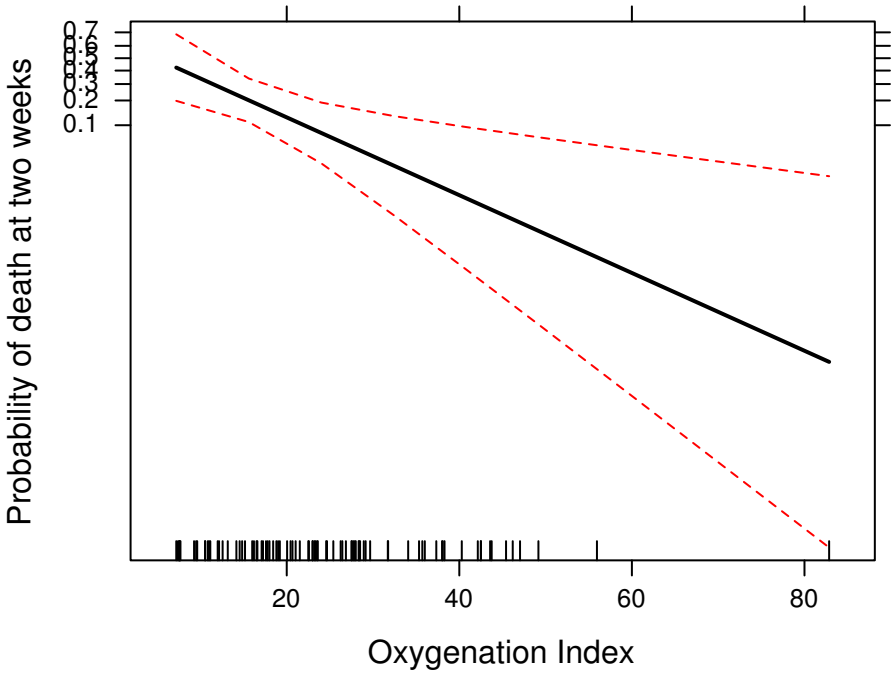


Figure 7.2: Effect of PF ratio on two week survival. Tick marks in the horizontal axis show the observed PF ratios.

	Estimate	Std Error	p
Intercept	1.67	0.82	0.04
Haematocrit	-0.04	0.01	0.02
Oxygenation Index	-0.10	0.03	0.005

Table 7.3: Logistic regression for survival at two weeks. Oxygenation index is defined as the ratio of arterial partial pressure of oxygen to the inspired oxygen concentration.

neutrophil count. After fitting the full model, haematocrit and PF ratio are significantly associated with outcome (Table 7.3). The effect sizes are similar to those seen for the two week survival model.

Similar results are seen with the logistic regression models for hospital survival and survival to ICU discharge.

While this analysis is significantly limited by power, there are a number of inferences that can be mode. Most of the covariates are not directly related to survival as reported previously. This is probably due to small sample size in this analysis. Some covariates are

consistently associated with survival on independent comparisons. This consistency is at least partly related to the inevitable correlation in outcomes (mortality at two weeks always implies mortality at 4 weeks). The pattern of significant covariates suggest that low perfusion pressures, compromised oxygenation, and markers of impaired tissue perfusion are associated with poor outcomes. There is also an association with neutrophil counts and high sodium levels, but the clinical relevance of these observations is unclear.

7.3 Relationships to Gene Expression

Covariates can be the cause or the result of changes in gene expression. The relationship of the covariates to gene expression is of considerable interest, especially the relationship of groups of covariates to one another and gene expression. These correlations can be investigated with multi-factor analysis, the principles of which are described in section 4.2 on page 41.

The result of the multi-factor analysis is shown in Figures 7.3 and 7.4. The analysis was performed with continuous covariates and expression data with a coefficient of variance greater than 0.9 (n = 292). Each covariate is represented as a vector against the background of gene expression. The direction of the vector shows the bias in gene expression data because of that covariate and the length of the vector represents the magnitude of the effect.

With multi-factor analysis, covariates seem to form two clusters. One cluster consists of high systolic blood pressure, high mean arterial pressure, high pH, and high bicarbonate levels which are markers of good tissue perfusion, and usually associated with good outcomes. In contrast, the other cluster consists of lower perfusion pressures, low pH and renal failure suggesting poor perfusion and organ dysfunction, usually predictors of poor outcome. This is confirmed by the correlation of these variables to the number of days of respiratory, renal, and inotropic support.

Platelets seem to have a large effect, but the effect is orthogonal to all expression data and is likely to be irrelevant. Changes in white cell counts are also associated with gene expression. In particular, high white cell count and neutrophilia are correlated with

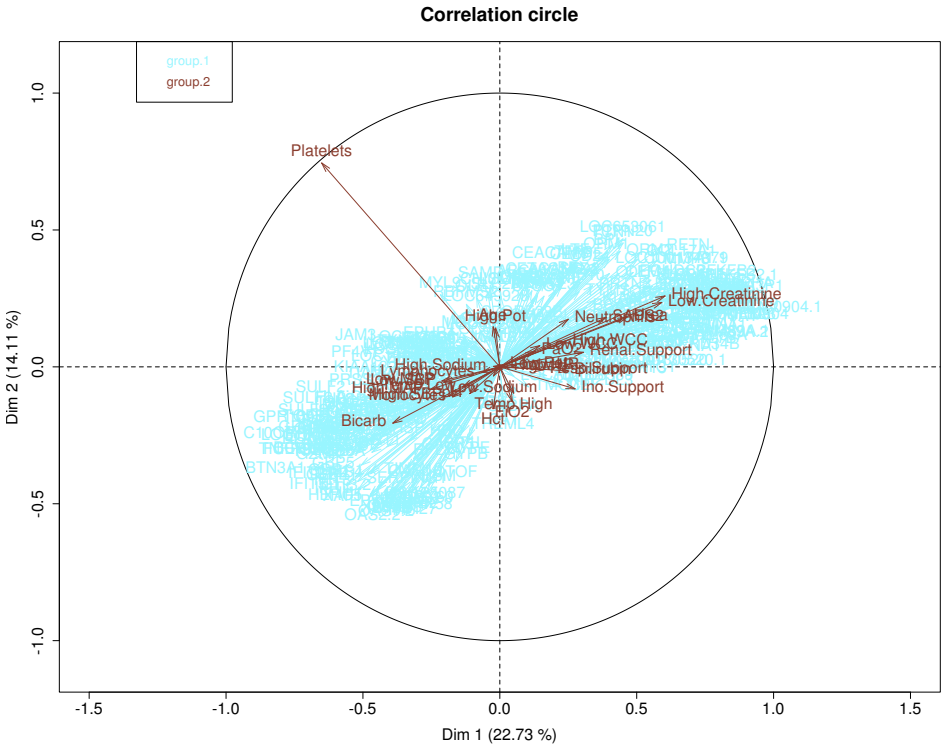


Figure 7.3: Multi factor analysis of gene expression and clinical covariates. The analysis involves simultaneous principle component like decompositions of the gene expression data and the clinical covariates. The analysis clusters genes based on similarities of expression and overlaps a similar cluster of clinical covariates, thus relating gene expression clusters to covariate clusters. The figure shows that the covariate clusters overlap with the gene clusters, suggesting a relationship between molecular mechanisms and observed clinical phenotypes, in particular parameters related to organ perfusion. Most of the variation occurs in the first two dimensions.

other variables associated with organ dysfunction and adverse outcomes. In contrast monocytosis and lymphocytosis are associated with effects in the opposite direction.

While causal inferences are not possible with the correlation analysis, the results suggest a strong influence of total and differential cell counts on gene expression data. At least part of this influence is related to organ function and outcome. Markers of organ function are also likely to be independently correlated with gene expression.

7.4 Summary

The complexity of clinical sepsis is due to the combination of the complexity of underlying pathophysiology and the complex interactions of a number of covariates. The

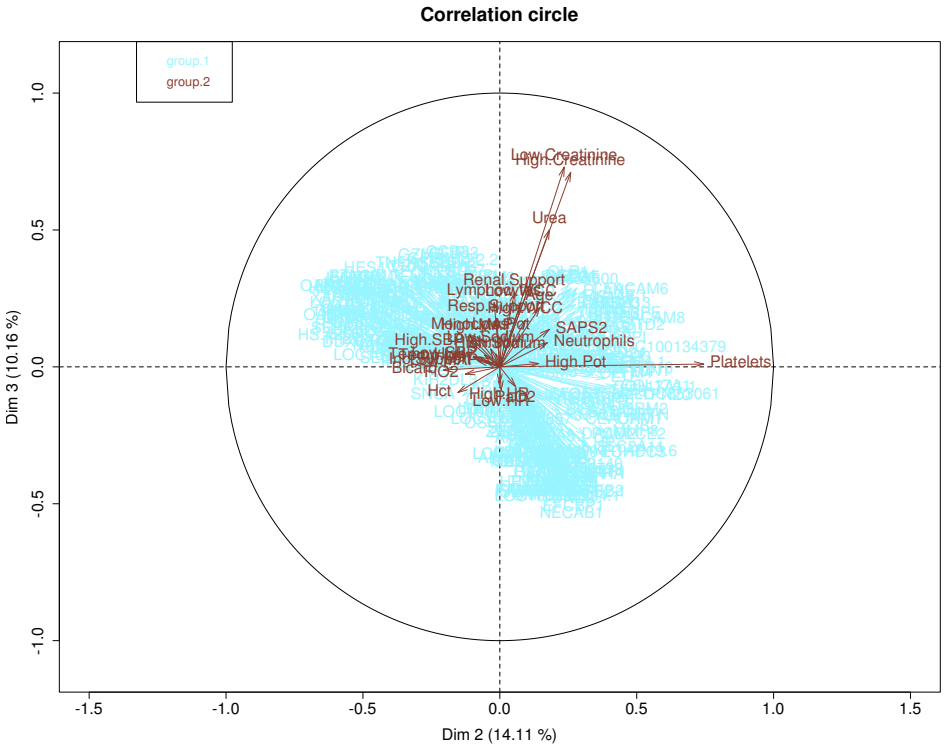


Figure 7.4: Mutli factor analysis of gene expression and clinical covariates showing the variation along the second and third dimensions of the multi factor decomposition. Most of the variation appears to be related to change in platelet counts.

traditional approach to covariate complexity is to restrict the use of covariates to specific predictive applications.

The sepsis data set consists of gene expression data along with extensive covariate data. Simple regression analysis to mortality at two weeks shows the consistent influence of haematocrit and PF ratio on outcome. Other covariates such as perfusion pressure and tissue perfusion indices are also associated with outcome in individual analysis but not in a full regression model.

The covariates associated with outcome are also related to changes in gene expression. Large changes in gene expression are associated with markers related to tissue perfusion, organ function and white cell counts. There is consistency between the associations to expression phenotype and clinical phenotype. Correction for these covariates is necessary to avoid bias, and a robust analytical approach was used to identify the gene expression changes associated with these covariates (chapter 11 on page 160)

8

Exploratory Analysis

8.1 Background

The core of the analytical process for gene expression data involves the identification of differentially expressed genes for contrasts of interest. Lists of differentially expressed genes are typically identified on the basis of regression models reflecting the study design. However, inferences from regression models fail when the assumptions underlying the analysis are not met. In particular, regression analysis assumes normal distribution of the errors and homogeneity of variance in the groups compared.

Inferences from the regression model are also invalid in the presence of systematic bias in the data. Systematic differences in gene expression can occur either as a result of technical factors related to the experiment, in which case they need to be controlled as nuisance parameters, or due to biological factors related to disease pathophysiology. The identification of the latter provides insight into disease mechanisms; these factors also act as confounders in some of the analytical methods. Exploratory analysis is also important in identifying and characterizing unanticipated structure in the data.

This chapter interrogates the sepsis data sets for some common assumptions and systematic biases. Assumptions were checked over the entire data set, while exploratory analysis was performed independently for the CAP and FP cohorts. The validation data set was always analysed separately.

8.2 Assumption Checks

Normality

Parametric statistical tests assume normal distributions. Log transformation of the raw expression data converts the distribution from exponential to normal. Despite the transformation, violation of normality assumption is a common problem with microarray expression data.

The expression of each probe was checked for normality using the Shapiro-Wilk test (Shapiro *et al.,* 1965) Normality assumptions were violated in 62.9% (n = 18,317) of the probes of the derivation data set and 53.1% (n = 11,220.) of the validation data set. The tests for normality are based on the probability distribution of the expression data for each probe. It is important to distinguish the normal probability distribution from the microarray normalization process from the previous chapter, which aligns the expression values from different arrays to a common baseline. The normalization process does not generally change the probability distribution, except during the log transformation.

Homoscedascity

Along with normality assumptions, linear models also assume that variances are homogeneous across groups of interest (homoscedascity). For example, when considering survival, the variance of the expression values of a given probe should be similar in both survivors and non-survivors. Differences in variance implies that the change in the predictor changes the error in the independent variable. Outliers are a cause of heteroscedascity and can result in undue influence of the outlier in the regression estimate.

The assumption of homogeneity of variance across the primary contrast of interest, survival, was checked for all probes using the Bartlett test (Bartlett, 1937) 12.3% (n = 3596) of the probes in the derivation data set and 46.7% (n = 9852) of the probes in the validation data set are not homogeneous across the survival groups. The Bartlett test is sensitive to deviations from normality. With the Levene test, which is robust to normality violations,

non-homogeneous variance is seen in 7.6% (n = 2216) and 28.1% (n = 5938.) of the probes in the derivation and validation data sets respectively.

These assumption failures, in theory, should seriously limit the use of standard regression models in the analysis of genomic data. However, there is evidence, based on real world and simulated data sets, that the standard tests perform well in genomic data despite assumption violations (Jeanmougin *et al.,* 2010). The approach used in this work is to use standard linear models for canonical analysis and mitigate the effect of these problems using Bayesian hierarchical models.

Systematic Biases

The data was checked for systematic biases from factors that should not influence biological function. The variation in gene expression with the recruiting centre and the array hybridisation was checked by fitting a linear regression model. None of the probes showed significant differences in expression (at least 1.2 fold change in mean expression between arrays or centres) after correction for multiple comparisons (FDR < 0.05) for both these contrasts. This excludes a systematic effect on gene expression by these extraneous factors.

There were significant effects due to differences in the hybridisation platform between the derivation (HT-12 v3) and validation (WG6-v3) data set. With principal component analysis, the biggest source of variation between the two data sets was the hybridisation platform, despite normalising both data sets together. The first principal component is contributed entirely by the differences in the platform and contributes to more than 75% of the variation in the data set. The swamping of the relevant biological variation by this noise factor made a combined analysis impossible. All further analyses were performed on data sets normalised individually and no further efforts were made to combine the data sets.

8.3 Hierarchical clustering of principal components

One of the aims of exploratory analysis is the identification of unanticipated structure in the data set. Unsupervised clustering algorithms were used to interrogate the data set. The analysis was performed independently on the CAP and FP cohorts to avoid clustering based on diagnosis.

Clustering in community-acquired pneumonia

Unsupervised Clustering

Principal components from the most variant probes in the CAP derivation data set were clustered using a hierarchical-agglomerative algorithm described in section 4.3 on page 42. Probes were selected on the basis of coefficient of variance; the most variant genes were defined as probes with a coefficient of variation greater than 0.9. Clustering was performed on the first two principal components. The algorithm allows for the optimal choice of cut off in the hierarchical tree based on statistical evidence with a p value less than 0.05.

The results show the presence of three well defined clusters (Figure 8.1). Each point corresponds to one sample. The height of the branches indicate the strength of the cluster. There is clear support for the separation of cluster 1 from the other two clusters and more modest evidence for the separation of cluster 2 from 3. The first two principal components account for 30.88% of the overall variance.

Each component in the principal component analysis is a linear combination of a number of predictor variables. In this analysis, the components are due to the linear combination of the expression of a group of probes. In other words, each cluster represents a distinct pattern of gene expression, based on a set of probes with the largest contributions to the overall variance.

Relation to covariates

Exploratory analysis of the probes with the greatest variance in the CAP cohort shows the presence of three distinct gene expression patterns. These patterns could be the result

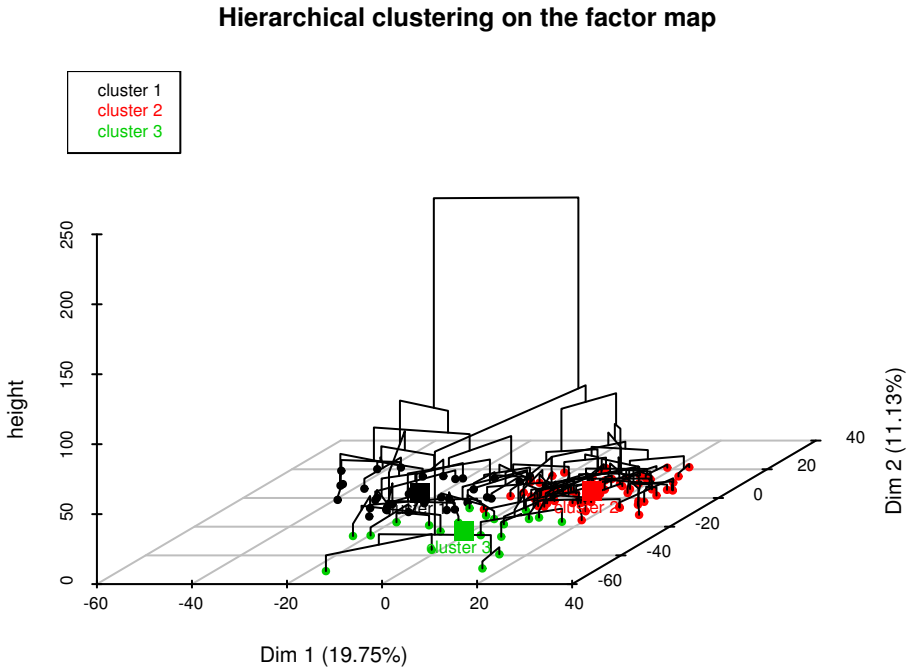


Figure 8.1: Hierarchical clustering of the principal components of the most variant probes from the CAP sub-cohort of the derivation data set. Larger height between the branches indicate stronger evidence for cluster difference. Dim 1 and Dim 2 are the first and second principal components respectively.

of a biological phenomenon that is unrelated to other biological factors, or more likely related by varying degrees to measured or unmeasured biological factors (for example, the underlying genetic variation). In this context, it is useful to examine the relationship of these clusters to the measured covariates.

It is interesting to note that the multi-factor analysis in section 7.3 on page 81 uses a similar approach to the hierarchical-agglomerative clustering algorithm, and overlays the covariates on the gene expression data. Results from the multi-factor analysis shows correlation between covariates related to organ dysfunction and white cell count and the principal components of the gene expression data. The relationships of the covariates to clusters identified in section 8.3 on the preceding page are summarised in Tables 8.1 to 8.7.

Cluster 1 is differentiated from other clusters by the following characteristics: lower systolic blood pressure, lower mean arterial pressure, lower neutrophil count, lower

	Cluster 1	Cluster 2	Cluster 3	p
Patients (n)	21	40	12	
Samples (n)	32	75	20	
Day 1 samples (n)	14	20	6	
Day 3 samples (n)	13	24	9	
Day 5 samples (n)	5	31	5	
Age in years (Median/IQR)	55 (23)	63.5 (25.8)	70 (10.8)	0.33
Sex (M:F)	1:0.6	1.4:1	1:1	0.35

Table 8.1: Demographic differences between clusters. p values are only shown where relevant.

	Cluster 1	Cluster 2	Cluster 3	p
Temperature (°C)				
High	38.0 (1.1)	37.7 (0.8)	38.1 (1.1)	0.06
Low	36.4 (1.1)	36.4 (0.7)	36.5 (0.9)	0.87
SBP (mm Hg)				
High	135 (18)	157 (25)	151 (37)	0.0005
Low	88 (19)	101 (17)	90 (14)	0.0008
MAP (mm Hg)				
High	88 (11)	106 (18)	101 (24)	0.00002
Low	61 (10)	69 (12)	62 (13)	0.01
HR (bpm)				
High	115 (21)	113 (25)	116 (17)	0.85
Low	89 (19)	79 (19)	85 (12)	0.008

Table 8.2: Differences in vital signs between clusters. SBP = systolic blood pressure, MAP = mean arterial pressure, HR = heart rate

	Cluster 1	Cluster 2	Cluster 3	p
WCC (1000 cells/mm^3)				
High	18.3 (10.4)	13.0 (5.6)	19.6 (11.2)	0.0006
Low	15.5 (9.0)	11.9 (4.9)	18.3 (11.5)	0.002
DCC (%)				
Neutrophils	79.8 (11.3)	88 (7.9)	89 (4.7)	0.00006
Monocytes	4.7 (4.8)	8.6 (6.8)	4.1 (2.4)	0.002
Lymphocytes	7.2 (5.1)	11.6 (6.4)	6.9 (4.3)	0.0004
Platelets	170 (12)	248 (107)	220 (124)	0.005
Haematocrit	27 (15)	25 (18)	29 (18)	0.83

Table 8.3: Differences in white cell counts between clusters. WCC = white cell count, DCC = differential cell count

	Cluster 1	Cluster 2	Cluster 3	p
Sodium (mmol/L)				
High	139.4 (3.8)	139.6 (4.5)	138.3 (6.4)	0.82
Low	135 (5.5)	137.2 (5.3)	135.1 (5.1)	0.47
Potassium (mmol/L)				
High	4.5 (0.6)	4.5 (0.6)	4.6 (0.8)	0.88
Low	3.7 (0.7)	3.8 (0.5)	3.8 (0.8)	0.89
Bilirubin (mmol/L)	22.9 (30.7)	10.5 (0.8)	37.5 (54.9)	0.001
pH	7.29 (0.10)	7.34 (0.13)	7.35 (0.13)	0.41
Bicarbonate (mEq/L)	22.3 (7.2)	28.5 (5.8)	25 (6.4)	0.0001

Table 8.4: Differences in biochemical covariates between clusters

	Cluster 1	Cluster 2	Cluster 3	p
Urea (mmol/L)	13.9 (8.8)	8.9 (8.4)	12.5 (6.6)	0.02
Urine Volume (L)	1579 (1350)	2256 (1236)	2629 (1379)	0.01
Creatinine (mmol/L)				
High	155 (114)	81 (52)	126 (66)	0.00002
Low	141 (99)	77 (53)	121 (68)	0.0002

Table 8.5: Differences in renal function markers between clusters

	Cluster 1	Cluster 2	Cluster 3	p
Respiratory Rate (/min)	26 (9)	29 (10)	30 (8)	0.25
PF ratio	18 (9)	21 (11)	18 (6)	0.33
P_aCO_2 (KPa)	5.9 (1.7)	6.2 (1.7)	6 (1.3)	0.65

Table 8.6: Differences in respiratory function between clusters

	Cluster 1	Cluster 2	Cluster 3	p
SAPS II	49 (14.5)	63 (12)	57 (6)	0.0002
Predicted Death Rate	43.8 (27.1)	73.6 (26.7)	61.9 (13.1)	0.0002
ICU mortality	8	10	2	
In hospital mortality	10	19	4	
Four week mortality	8	12	4	
Two week mortality	2	14	3	
Time to death (days)	9.5 (5.5)	20 (22)	22 (4)	0.08

Table 8.7: Differences in outcome between clusters

platelets, lower bicarbonate, lower urine output, lower urea, and higher creatinine levels. Cluster 2 is differentiated from the other clusters by normal temperature, lower heart rate, lower white cell count, higher monocyte and lymphocyte count, higher bicarbonate, low urea and creatinine. Cluster 3 occupies a middle position with some values aligned closely with other clusters, while others take a value between the two clusters.

The clusters are also associated with differences in mortality. Firstly cluster 1 has the lowest SAPS II score (49) with a predicted mortality rate of 44%, while cluster 2 has the highest SAPS II score (63) with a predicted mortality rate of 74%. Cluster 3 has a SAPS II score of 57 with a predicted mortality rate of 62%. This suggests that the cluster structure or differences in pattern of gene expression are associated with mortality in a way similar to the commonly used risk scoring system.

Secondly, the observed in-hospital mortality rates for the three clusters are 50%, 50%, and 30% against the predicted rate of 44%, 74%, and 62% for clusters 1, 2, and 3 respectively. From Figure 8.1 it is clear that on the basis of gene expression there is clear

discrimination between clusters 1 and 2, which have identical mortality rates. Cluster 3 is discriminated from cluster 2 with lesser confidence, but has clear outcome differences from cluster 2. On the basis of the covariates clusters 1 and 3 form opposite ends of the spectrum, suggesting that the apparent homogeneous clinical syndrome of sepsis consists of considerable heterogeneity in cell function.

Thirdly, the different gene expression signatures are correlated with differences in time to death, with a trend towards statistical significance ($p = 0.08$). The median time to death in cluster 1 is 10 days, and cluster 2 is 22 days. The median time to death in cluster 3 is also 22 days, but with low sample numbers the relevance is uncertain.

8.4 Characterization of Clusters

Linear Models

The gene expression signatures of the clusters associated with the first two principal components were identified by constructing linear models using the procedure described in section 4.3 on page 44.

Three way comparisons were made between the three identified clusters using regression models that did not correct for covariates. The results are summarised in Table 8.8 on the following page.

Stringency thresholds relating to the false discovery rates and the fold change differences are arbitrary and mostly determined by convention. Genomic studies routinely report a 1.5 fold change as the minimum significant threshold. Recent combined proteomic and transcriptomic investigation fission yeast cells has identified functional effects with only one copy of mRNA per cell(Marguerat *et al.,* 2012). However, given the signal-noise ratio profile of the microarray platform, a 2 fold change difference in expression is used as a stringent threshold and a 1.5 fold change in expression is used as a minimum threshold. While presenting results, a pragmatic approach is used; thresholds are selected as appropriate for the topic while simultaneously limiting the number of significant genes to manageable numbers. A 1.2 fold change in expression is used as a

threshold for hypothesis generation with an implicit acceptance of the increased type I error.

Despite a stringent threshold of at least 2 fold change in expression and FDR less than 0.01, there are significant differences between the clusters. Based on Figure 8.1 on page 88, large differences are expected between cluster 1 and the other clusters and small differences are expected between cluster 2 and 3; these expectations are confirmed in the Table 8.8.

Contrast	Counts	Up-regulated	Down-regulated	Total
Cluster 2 to 1	Probes	73	64	137
	Genes	65	56	121
Cluster 3 to 2	Probes	1	34	35
	Genes	1	30	31
Cluster 3 to 1	Probes	23	286	309
	Genes	23	250	273

Table 8.8: Number of probes/genes with at least 2 fold change difference in expression and a FDR less than 0.01. The results are based on linear models which do not correct for any covariates.

The differentiation of cluster 3 from other clusters is of interest since this cluster is associated with lower mortality. The overlap of the differentially expressed genes in the two comparisons that include cluster 3 is shown in Figure 8.2. Since the Figure shows genes that differentiate cluster 3 from the other two clusters, the overlapping region of 19 genes constitute a panel that can discriminate cluster 3 from all other clusters. These genes along with their biological roles are listed in Table 8.9. The immediate function of some of these genes in immunity is unclear, but the down-regulation of *CASP8* (apoptosis related cysteine peptidase), *LYZ* (lysozyme), and *TNFAIP3* (TNF induced protein) suggests an inflammatory state that is not as pronounced as in the other clusters, probably reflecting the resolution of inflammation. It is unlikely that the observed differences reflect a lesser degree of inflammation, perhaps as a response to a less virulent infectious load, given that the patients in cluster 3 are clinically indistinguishable from the other clusters. However, the inference of inflammatory status

from isolated gene expression patterns is problematic, and additional proteomic data will be required to confirm these hypotheses.

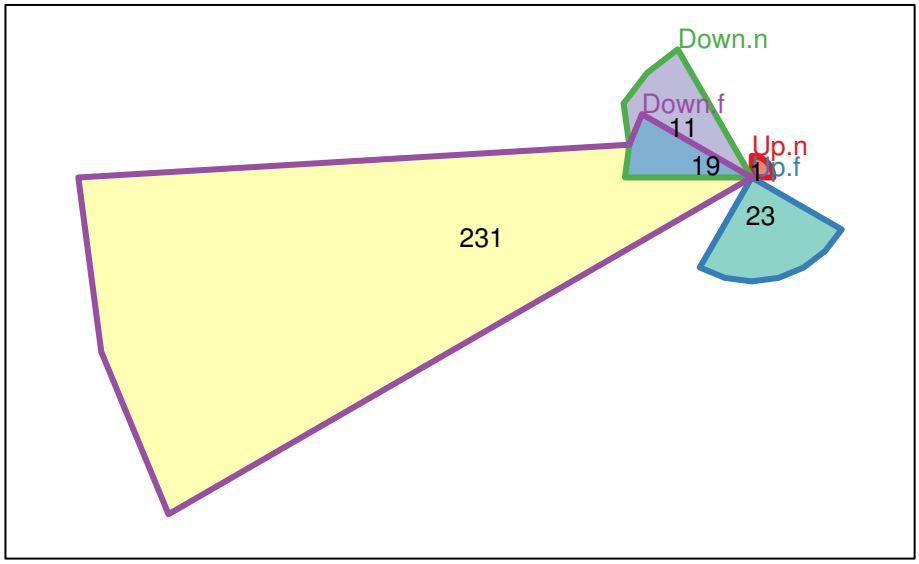


Figure 8.2: Venn diagram of genes that differentiate cluster 3 from other groups with at least a 2 fold change difference a FDR less than 0.01. The aim is to identify common patterns of expression in the two contrasts involving cluster 3 (n = Near contrast, comparison of cluster 3 to cluster 2, and f = Far contrast, comparison of cluster 3 to cluster 1). Each region, bounded by a coloured line and a matching label, is proportional to the number of genes that belong to that group. The legend is formatted as follows. The first part shows the direction of regulation and the second part refers to the contrast.

The distribution of some of these genes are plotted in Supplementary Figure 1. The relationships of these genes are also summarised in a heatmap (Figure 8.3). The cluster on top of the heatmap groups the samples on the probe expression value. The panel of genes accurately capture only part of the samples in cluster 3, suggesting poor predictive power.

Bayesian Analysis

The regression analysis presented in the previous section has two significant limitations: first as shown in section 8.2 on page 85 the assumptions underlying linear models do

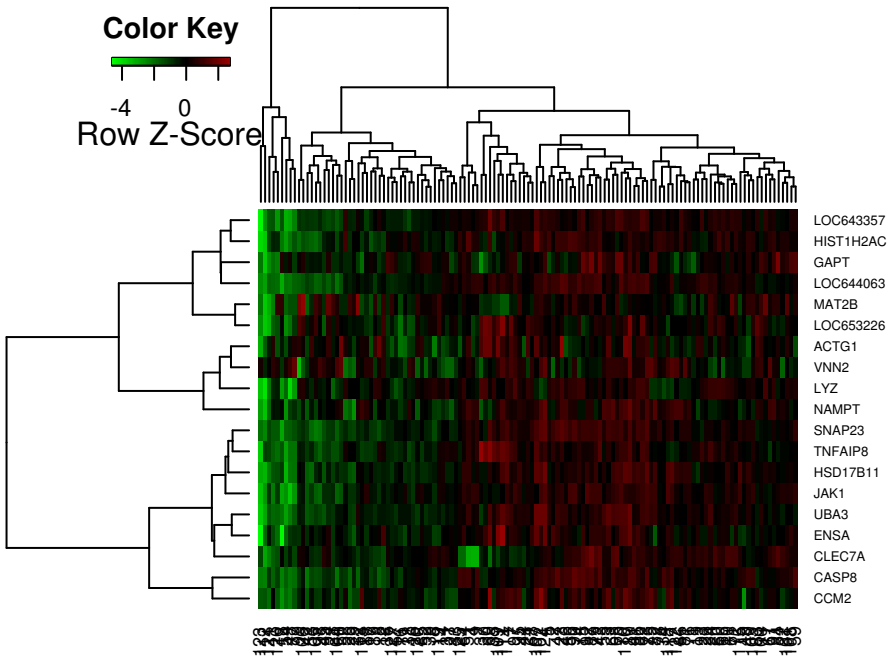


Figure 8.3: Distribution of some of the genes consistently repressed in cluster 3 compared to other clusters. Expression of each probe has been scaled and centred before plotting. Each column corresponds to a sample. Cluster 3 samples are not clearly differentiated in the dendrogram at the top of the heatmap.

not really hold, despite routine application of this method for analysing gene expression data; second the linear models do not correct for the presence of covariates.

Selecting covariates that need correction before data interpretation is problematic for two reasons: tagging a covariate for correction implies some degree of causality, while the covariate might both cause and be modified by gene expression changes. Also the interpretation of covariate corrected data in clinical contexts is difficult.

Given these problems with linear models, it is clear that their applicability in deriving mechanistic models of disease is limited. For this purpose, a more robust Bayesian analysis was performed as described in section 4.3 on page 45.

A hierarchical Bayesian model was constructed where the expression of the probe was modelled on the following covariates: differential cell counts, age, sex, microarray, day of sampling, survival status, centre, cluster membership, infecting organism, oxygen index, renal failure, activated protein C use, steroid use, total white cell count and

Gene	Chromosome	Function
ACTG1	17	gamma-actin
CASP8	2	Apoptosis related cysteine peptidase
CCM2	7	scaffold protein in MAPK signalling
CLEC7A	12	C type lectin domain family
ENSA	1	regulatory subunit of KATP channel
GAPT	5	GRB2 binding adaptor protein
HIST1H2AC	6	Histone H2A
HSD17B11	4	Steroid dehydrogenase
JAK1	1	Janus Kinase 1
LOC643357	7	not in current assembly
HNRNPKP4	3	HNRNPKP4 — nuclear ribonucleoprotein
SRP9P1	10	SRP9P1 — signal recognition particle
LYZ	12	Lysozyme C
MAT2B	5	Methionine adenosyl transferase
NAMPT	7	Nicotinamide phophoribosyltransferase
SNAP23	15	Synpatosomal associated protein
TNFAIP8	5	TNFa induced protein
UBA3	3	Ubiquitin activating enzyme
VNN2	6	Vanin family

Table 8.9: The common panel of genes that differentiate cluster 3 from other clusters. All genes are down-regulated in cluster 3.

haematocrit. Covariates were selected for inclusion in the model if at least one of three criteria was satisfied: a clear causal association such as age, sex, recruiting centre and therapy choices, an association identified in a previous analysis, or systemic effects with high likelihood of modulation of gene expression such as white cell counts and markers of organ dysfunction. Some of these associations were identified in the canonical correlation analysis in 7.3 on page 81. MCMC simulation was used to identify probes with significant changes in expression for any covariate of interest. A probe was considered significant if there was more than 1.5 fold change in mean expression levels and the 95% Bayesian credible interval did not include zero.

Cluster membership was modelled as a hierarchical variable; instead of one co-efficient associated with changes in cluster membership, each cluster was allowed to

have an independent mean and the difference in means was estimated for each probe. Since some of the covariates are correlated with some of the variation in the principal components of gene expression, the cluster structure is expected to reduce in magnitude when covariates are included in the model. However, the cluster structure persists despite robust correction for plausible covariates, perhaps suggesting an intrinsic genetic cause of different patterns of inflammation. The number of significantly differentially expressed probes are summarised in Table 8.10.

Contrast	Counts	Up-regulated	Down-regulated	Total
Cluster 2 to 1	Genes	79	92	171
Cluster 3 to 2	Genes	111	148	258
Cluster 3 to 1	Genes	232	317	547

Table 8.10: Number of probes/genes with at least 2 fold change difference in expression between clusters, with a 95% credible interval that does not include zero. The results are from a Bayesian hierarchical model that includes a number of covariates listed in the text.

The results show a paradoxical increase in the gene yield with a preserved pattern of predominant down-regulation. This is likely a result of more accurate partitioning of variance between the covariates and the improved accuracy with the Bayesian analysis for small sample sizes. The difference is particularly striking when cluster 3 is contrasted with cluster 2; the gene yield increased from 31 to 258 for the same fold change difference threshold between the simple regression model without the covariates and the Bayesian hierarchical model with the covariates.

The set of genes that discriminate cluster 3 from other clusters are shown in Figure 8.4. A more stringent threshold was used to keep the gene yields comparable with the regression analysis. A common panel of 19 genes were identified, of which two were induced and 17 repressed. The final panel of 14 genes after pruning for contentious transcripts (withdrawn from the hg16 repository) are listed in Table 8.11. The pattern of gene expression could be consistent with resolving inflammation, but would require validation with proteomic data.

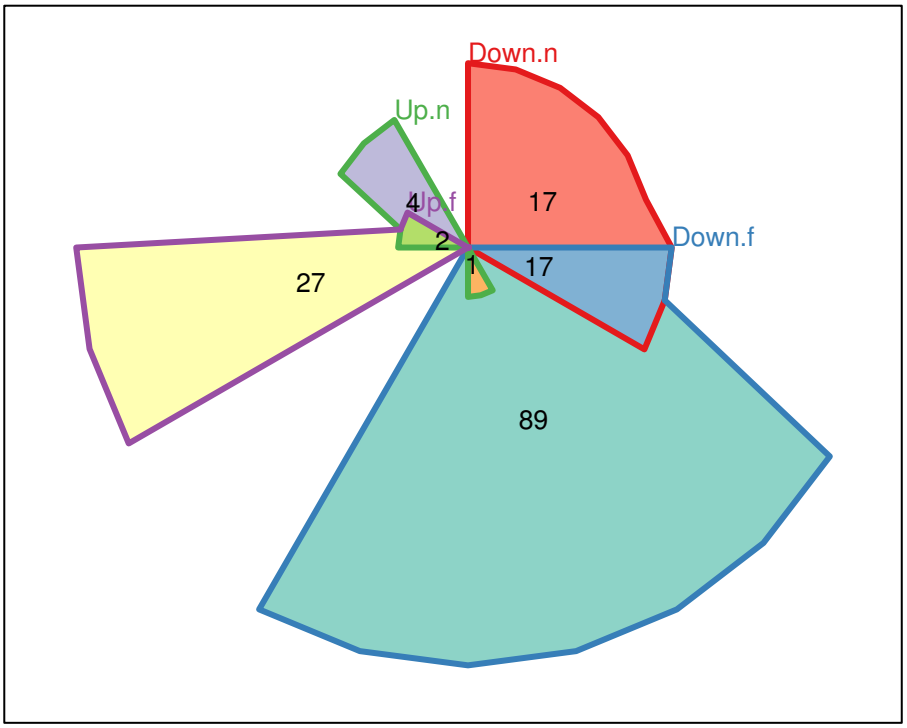


Figure 8.4: Venn diagram of genes that differentiate cluster 3 from other groups with at least a 2.5 fold change difference and with a 95% credible interval that does not include zero. The aim is to identify expression patterns that consistently differentiate cluster 3 from the other two clusters. Each region bounded by a coloured line is proportional to the number of genes that belong to that group. The legend is formatted as follows. The first part shows the direction of regulation and the second part refers to the contrast. n = Near contrast, comparison of cluster 3 to cluster 2, and f = Far contrast, comparison of cluster 3 to cluster 1.

Clusters 1 and 2 have significantly different gene expression signatures but similar mortality rates, suggesting that adverse outcomes could be the result of two different processes. The mechanistic differences between clusters 1 and 2 are therefore of interest.

With a 2 fold change threshold, 79 genes are induced in cluster 2 when compared to cluster 1, while 92 genes are repressed. The induced genes include members of the class II MHC: *HLA-DMA*, *HLA-DMB*, *HLA-DPA1*, *HLA-DRA*, *HLA-DRB4*, *HLA-DRB5*, *HLA-DRB6*, interferon induced proteins: *IFIT1*, *IFIT2*, *IFIT3*, interleukin molecules: *IL32*, *IL7R*, *IL8*, CD antigens: *CD2*, *CD247*, *CD27*, *CD3D*, *CD6*, *CD74*, *CD86*, *CD8A*, and lymphocyte antigen *LY86*. The four genes with the greatest fold changes in expression are shown in Supplementary Figure 2. The bimodal expression of core immune genes in the clusters, along with the inferences about cluster 3 has a significant implication; the clusters

Gene	Chromosome	Function
CDC42EP3	2	negative regulator of CDC42
CHMP5	9	degrades surface receptor proteins
CXCR4	2	receptor for stromal cell derived factor. Association with hypogammaglobulinemia and infections
GNG10	9	Guanine binding G protein
HIF1A	14	Hypoxia inducible factor
JAK1	1	Janus Kinase 1
LOC100133565	12	ubiquitin like gene
HNRNPKP4	3	HNRNPKP4 nuclear ribonucleoprotein
MAT2B	5	Methionine adenosyl transferase
NAMPT	7	Nicotinamide phophoribosyltransferase
PPM1A	14	negative regulator of stress response
TRIM22	11	mediates interferon effects
UBA3	3	Ubiquitin activating enzyme
VNN2	6	Vanin family

Table 8.11: The common panel of genes that differentiate cluster 3 from other clusters from the Bayesian hierarchical model. Two of the down-regulated genes and both the up-regulated genes are no longer present in the hg16 assembly.

do not merely represent distinct patterns of gene expression, but distinct patterns of inflammation.

Pathway Analysis

The functional significance of the differentially expressed genes identified by linear models and Bayesian hierarchical models can be interpreted using pathway analysis.

DAVID Functional Annotation

In the simplest analysis, a pathway or gene set is considered significantly enriched if a certain proportion of the genes in that gene set are significant. The measure of significance or the effect size is irrelevant; only that the gene be significant. The web based tool DAVID can identify the enriched pathways given a list of significant genes using a modified Fisher exact p value.

Gene lists derived from the regression analysis and the Bayesian analysis were used to identify the enriched gene sets. While the former reflects the cumulative physiological effects the latter identifies the effects of individual physiological components.

Genes with a minimum 2 fold change threshold and a FDR of 0.05 were used for the functional clustering (Table 8.12). The pathway yield is generally low but there is strong evidence for differences in class II MHC regulation between clusters 1 and 2. The results of the Bayesian analysis are shown in Table 8.13. Again the yield is very low, but there is strong evidence of class II MHC regulatory differences between clusters 1 and 2.

Contrast	Genes	Functional Annotation
Cluster 2 vs 1	159	Interleukin like receptors (n = 8, ES = 4.52) class II MHC (n = 6, ES = 3.8) C-type lectin receptor (n = 11, ES= 3.69)
Cluster 3 vs 2	68	
Cluster 3 vs 1	335	Proteasome (n = 6, ES = 2.54)

Table 8.12: DAVID functional annotation for genes associated with cluster differences with at least a 2 fold change difference and a FDR of 0.05. ES = enrichment score. There are no functional clusters of significance in the second contrast.

Contrast	Genes	Functional Annotation
Cluster 2 vs 1	171	class II MHC (n = 8, ES = 4.91) Interleukin like receptors (n = 21, ES = 3.57)
Cluster 3 vs 2	547	
Cluster 3 vs 1	248	

Table 8.13: DAVID functional annotation for genes associated with cluster differences in the Bayesian hierarchical model with at least a 2 fold change in mean expression and with the 95% credible interval not including zero. ES = enrichment score. Only functional clusters meeting the significance thresholds are listed.

Gene Set Enrichment Analysis

The DAVID functional annotation tool ignores the effect sizes for the genes; genes with the highest effect sizes are used and genes with moderate effect sizes are ignored. Lists of significant genes can be interrogated for pathway enrichment in a quantitative manner

using GSEA. This analysis gives greater weight to genes with larger effects but includes all genes, irrespective of significance levels.

With GSEA an enrichment score is calculated for each gene set. The significance level for each gene set is estimated using permutation analysis involving phenotype switching. The process involves recalculation of test statistic for each probe and estimation of an enrichment score for each pathway in each permutation. This is computationally prohibitive for the Bayesian analysis and only pathways with extreme enrichment scores (defined as greater than the 95^{th} quantile) are listed.

Gene sets with a permuted p value less than 0.05 were considered to be significantly enriched. The pathways enriched with the regression statistic is shown in Figure 8.5 and with Bayesian mean is shown in Figure 8.6. A number of pathways are identified with minimal overlap with the pathways identified from DAVID functional annotation. More pathways are enriched from the Bayesian analysis, though the statistical significance of the pathways is uncertain. There is strong biological plausibility for most pathways, even though only a few like the mitochondria and JNK signalling mechanisms have been previously considered in the context of inflammation and sepsis. Particularly intriguing is the enrichment of pathways related to viral infections when comparing clusters 1 and 2. Both these clusters have high mortality rates but different inflammatory patterns; part of the late sepsis mortality could perhaps be due to concomitant, undetected viral infection.

8.5 Validating the Clusters

The clustering of samples based on patterns of gene expression identified in the exploratory analysis was validated in other data sets — the CAP validation data set, the faecal peritonitis cohort and an external data set of paediatric sepsis.

CAP Validation Cohort

The CAP validation cohort consists of samples from 15 survivors and 15 non-survivors (defined by 28 day mortality) hybridised to the Illumina WG6 v3 BeadArray platform.

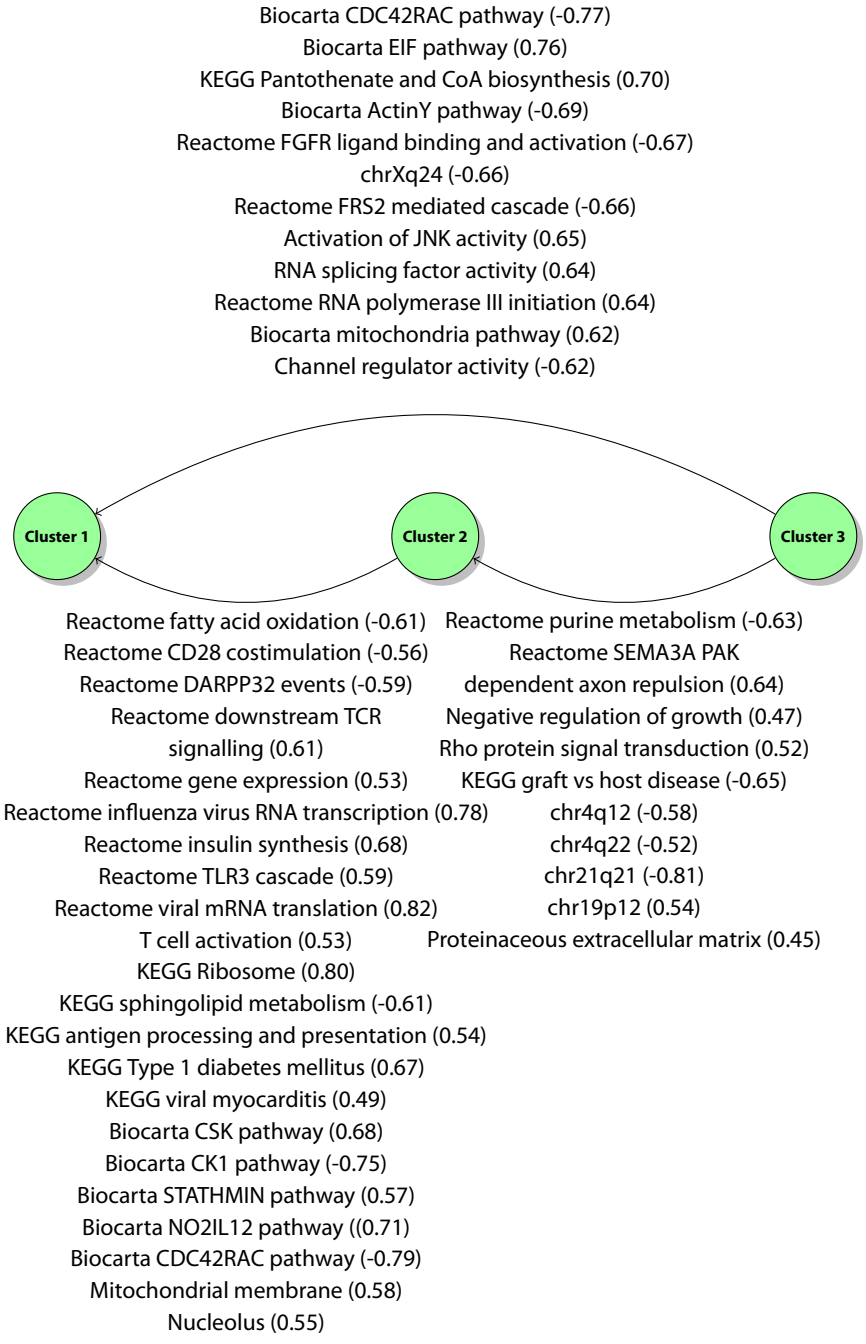
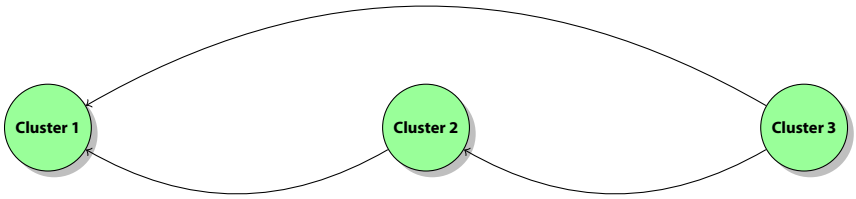


Figure 8.5: Pathways associated with clusters based on regression statistics. All pathways shown have a p value less than 0.05 with permutation involving phenotype switching and estimation of mean enrichment scores.

Down regulated
Reactome E2F transcriptional targets
Aerobic respiration
KEGG Pantothenate and CoA biosynthesis
Oxygen binding
Biocarta GATA3 pathway
chr3p12
chr3q24
chr21q21
chr8q23
chr4q32
Up regulated
Reactome P130CAS linkage to MAPK signalling
Reactome PD1 signalling
Reactome phosphorylation of CD3 and TCR zeta chains
Reactome translocation of ZAP70 to immunological synapse
Biocarta AMI pathway
chr11q
Myosin complex



Down regulated
Reactome classical antibody mediated complement activation
Reactome PKA activation
Reactome Na Cl dependent transporters
Myeloid leucocyte differentiation
KEGG Pantothenate and CoA biosynthesis
Oxygen binding
Monooxygenase activity
Biocarta CK1 pathway
chr4q24
SA PTEN pathway
Up regulated
Reactome generation of second messenger molecules
Reactome PD1 signalling
Reactome phosphorylation of CD3 and TCR zeta chains
Biocarta IL17 pathway
Biocarta STEM pathway
chr11q

Down regulated
Reactome generation of second messenger molecules
Reactome peptide chain elongation
Reactome translocation of ZAP70 to immunological synapse
KEGG ribosome
KEGG asthma
KEGG graft vs host disease
Biocarta IL17 pathway
Biocarta STEM pathway
chr21q21
chrXq24
Up regulated
Energy reserve metabolic process
Phospholipase C activity
Acetylcholine binding
Biocarta ERK5 pathway
Biocarta UCALPAIN pathway
chr11q
Myosin complex

Figure 8.6: Pathways associated with clusters based on Bayesian estimates. All pathways have an enrichment score in the 0.5% quantile in both extremes.

Supervised clustering on the principal components was performed on a subset of this data. This subset was defined on the basis of genes present in both the Illumina WG6 v3 and the Illumina HT-12 v4 platforms.

A panel of 2277 unique genes that differentiate any two of the clusters with a 1.5 fold change difference and a FDR of 0.05 were extracted from the derivation data set. Of these genes 1947 were present in the validation data set. Hierarchical-agglomerative clustering was performed on the expression data of the probes corresponding to these genes (Figure 8.7). Three clusters with a pattern of clustering similar to the derivation data set can be seen in the validation data set. There is greater overlap between the clusters compared to the derivation data. The first two principal components explain 51% of the overall variance in the expression of the 1947 probes. The decrease in clustering efficiency and the explanatory power are probably due to smaller sample sizes and the reduction in the number of genes used in the clustering.

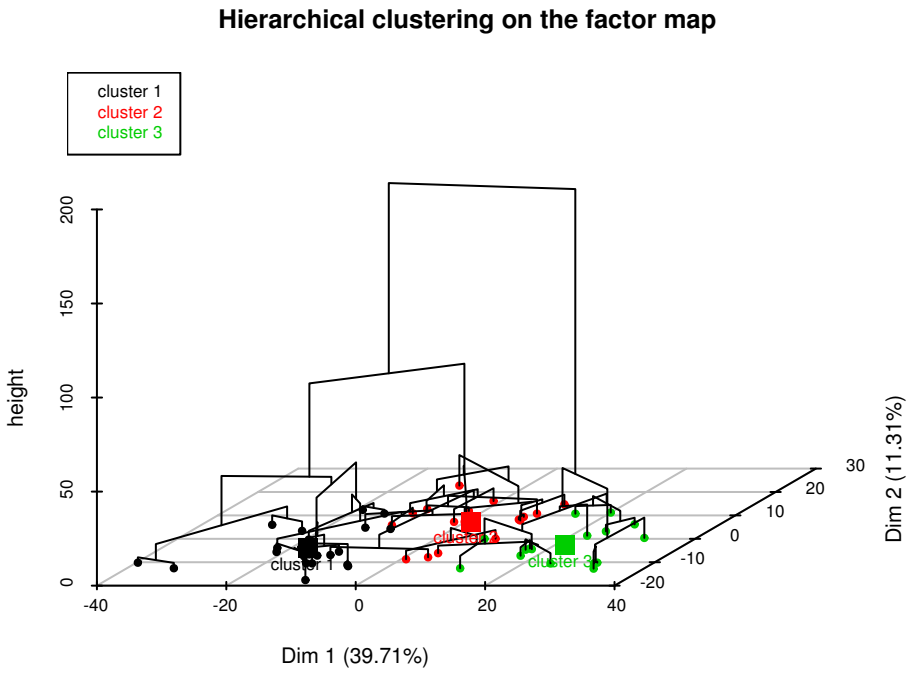


Figure 8.7: Hierarchical agglomerative clustering of the validation data set. Supervised clustering performed on expression data from 1947 genes identified from the derivation data set.

FP Cohort

The FP derivation cohort and CAP derivation cohort were hybridised to the sample platform in the sample experiment. The data was normalized together and is directly comparable.

A panel of 2597 probes that differentiate any two of the clusters with a 1.5 fold change difference and a FDR of 0.05 were extracted from the derivation data set. Supervised hierarchical agglomerative clustering was performed on the principal components of the expression of these probes in the FP derivation data set. Three clusters similar in distribution to the CAP derivation clusters are seen (Figure 8.8). 43% of the overall variation is explained by the first two principal components.

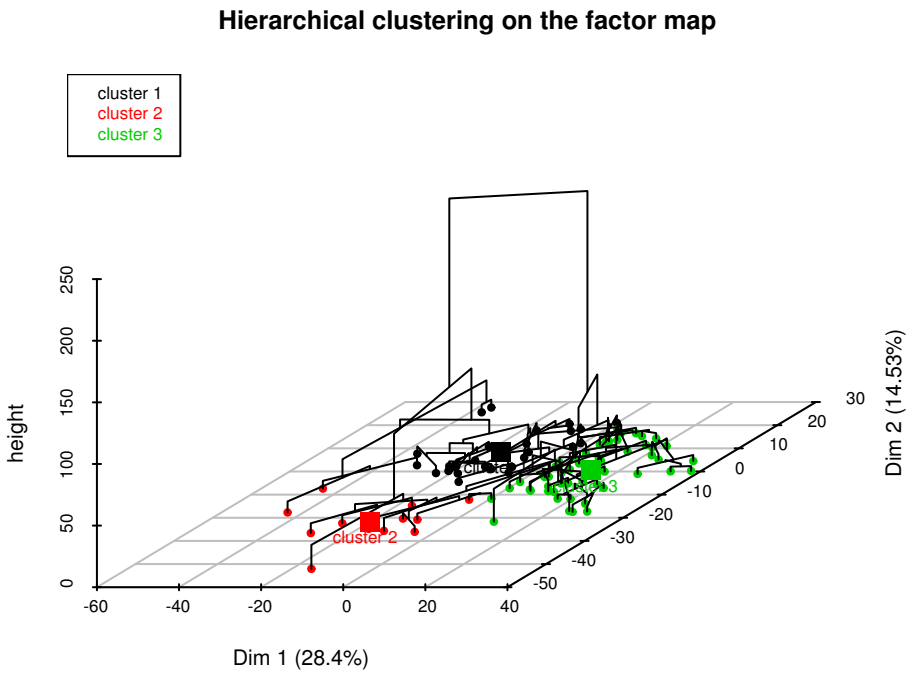


Figure 8.8: Hierarchical agglomerative clustering of the faecal peritonitis derivation data set. Supervised clustering performed on expression data from 2597 probes identified from the derivation data set.

With the CAP derivation cohort, the clusters were shown to have significant associations with covariates relating to organ function. The relationship of the covariates to the clusters identified in the faecal peritonitis group is seen in Tables 8.14 to 8.19. Comparisons were only made to covariates identified as significant in the CAP derivation

	Cluster 1	Cluster 2	Cluster 3
Patients (n)	14	9	41
Samples (n)	28	12	54
Day 1 samples (n)	4	15	9
Day 3 samples (n)	15	1	5
Day 5 samples (n)	9	18	9

Table 8.14: Demographic differences between clusters in FP

	Cluster 1	Cluster 2	Cluster 3	p
Temperature (°C)				
High	37.6 (1.0)	37.4 (0.9)	37.5 (0.7)	0.56
SBP (mm Hg)				
High	158 (18)	158 (26)	150 (26)	0.25
Low	107 (21)	110 (15)	100 (20)	0.12
MAP (mm Hg)				
High	103 (13)	102 (14)	100 (17)	0.81
Low	72 (13)	67 (11)	64 (14)	0.51
HR (bpm)				
Low	83 (18)	73 (17)	80 (11)	0.15

Table 8.15: Differences in vital signs between clusters in FP

cohort. The numbering of clusters is arbitrarily selected by the clustering algorithm; Cluster 3 in the FP cohort is analogous to cluster 2 in the CAP cohort. The following discussion takes this into account.

The differences in covariate associations between the CAP and FP cohorts are as follows. There are no differences between the clusters in FP for any of the vital signs.

Total white cell counts are comparable among the FP clusters in contrast to the CAP clusters. Differences in the relative neutrophil and monocyte count are seen in both diagnostic groups. The pattern of differences is maintained for relative neutrophil counts, but is different for the relative monocyte counts. Cluster 1 in the CAP cohort shows low relative neutrophil count compared to other clusters. Cluster 1 shows greater lymphocyte than monocyte counts. In the FP cohort, the neutrophil count is the lowest among the three clusters, but there is greater monocyte than lymphocyte count. Clusters

	Cluster 1	Cluster 2	Cluster 3	p
WCC (1000 cells/mm^3)				
High	12.5 (5.4)	17.6 (10.1)	15.0 (7.6)	0.13
Low	11.6 (4.9)	15.4 (10.6)	12.5 (8.0)	0.35
DCC (%)				
Neutrophils	81.8 (11.9)	90.1 (4.9)	89.7 (5.9)	0.0004
Monocytes	10.3 (12.5)	4.8 (3.7)	3.9 (2.9)	0.002
Lymphocytes	7.9 (3.3)	5.1 (1.7)	6.4 (4.2)	0.09
Platelets	265 (110)	270 (274)	211 (111)	0.18

Table 8.16: Differences in cell counts between clusters in FP

	Cluster 1	Cluster 2	Cluster 3	p
Bilirubin (mmol/L)	15.3 (12.4)	12.6 (14.2)	15.5 (14.1)	0.80
Bicarbonate (mEq/L)	23.4 (4.8)	23.5 (8.0)	22.4 (5.2)	0.71

Table 8.17: Differences in biochemical markers between clusters in FP

	Cluster 1	Cluster 2	Cluster 3	p
Urea (mmol/L)	10.2 (7.6)	15.4 (8.0)	13.2 (9.0)	0.20
Urine Volume (L)	2023 (1434)	1263 (1005)	1578 (1328)	0.20
Creatinine (mmol/L)				
High	93 (68)	153 (91)	130 (95)	0.09
Low	89 (67)	149 (88)	125 (95)	0.12

Table 8.18: Differences in renal function between clusters in FP

	Cluster 1	Cluster 2	Cluster 3	p
SAPS II	52.0 (12.5)	58.5 (12.8)	60.0 (14.0)	0.03
Predicted Death Rate	50.7 (27.5)	65.1 (27.2)	68.1 (27.0)	0.03
ICU mortality	0	0	5	
In hospital mortality	2	2	14	
Four week mortality	2	1	8	
Two week mortality	0	0	8	
Time to death (days)	15 (0)	20 (38)	7(0)	0.80

Table 8.19: Differences in outcome between clusters in FP

2 and 3 in the CAP cohort have equivalent neutrophil counts but inverse monocyte and lymphocyte counts. A similar patter in seen in the FP cohort. There is no difference between platelet counts in the FP clusters in contrast to the CAP clusters.

The strong differences in renal function between the CAP clusters are lost in the FP clusters, though there is a tendency towards significance for the highest observed creatinine levels. The differences between bilirubin and bicarbonate are also absent in the FP clusters.

There are differences in SAPS II scores between the clusters in each cohort suggesting some association between inflammatory patterns and mortality risk. The pattern of increase in risk is preserved across the clusters in both cohorts. The time to death is not different among the FP clusters even though the magnitude of difference between the clusters is high. The pattern is also different compared to the CAP clusters but is difficult to interpret because of the low mortality rate.

Comparing the covariate relationship to clusters in the CAP and FP cohorts two strong conclusions can be made. There is a relationship between clusters and relative cell counts, with each inflammatory pattern corresponding to a differing proportions of neutrophils, monocytes and lymphocytes. There is also a relationship to outcome risk, suggesting that part of the clinical risk can be attributed to the changes in gene expression.

There are a number of speculative reasons for the differences between the clusters in the FP and CAP cohorts. The simplest reason could be that FP and CAP result in fundamentally different gene expression patterns as a consequence of the different sources of infection and the nature of the microbiological insult. However, it could also be that the FP and CAP samples capture gene expression events at different time points in the natural history of the inflammatory response. Patients with faecal peritonitis tend to have a short course of illness before admission to ITU.They also tend to have early ITU admission and resuscitation. The infective component of the inflammation is modified by a super-added major surgical insult. These differences could account for the observed differences in gene expression. The differences in mortality rate between the two cohorts support these inferences. A third reason could be that the some of

the covariate associations in the CAP cohort are spurious and the cluster structure is independent of the covariates. This is supported by the overlap of differentially expressed genes between the clusters from the simple regression analysis in which there is no correction for covariates and the Bayesian analysis which corrects for many of the covariates. The improved associations with the Bayesian analysis suggests that the covariates are noise that reduces the true structure in the data.

Other Datasets

The presence of distinct inflammatory patterns was also validated in gene expression data derived from other sources. The publicly available data set (GSE13904) used for this exercise is from a cohort of paediatric sepsis patients (Wong *et al.*, 2009a) The paediatric samples were hybridized to the Affymetrix platform. Of the 2277 genes that were differentially expressed between any two clusters in the derivation data set, 1685 genes (corresponding to 4048 probes) were present in the paediatric data set. Hierarchical agglomerative clustering was performed on the principal components of the expression of these genes (Figure 8.9.). 86.8% of the variation in the expression in these probes can be attributed to the first two principal components. There is strong evidence for at least two clusters. Interestingly these clusters are differentiated by the bimodal expression of core immune function genes such as class II MHC and interferons. The persistence of this difference in paediatric sepsis lends credence to the speculation that the observed difference in inflammatory pattern could be the result of underlying genetic variation.

8.6 Summary

The aim of exploratory analysis is to validate assumptions used in downstream analysis and find unanticipated patterns in the data.

Gene expression data is known to lack robustness with respect to normality assumptions. Nearly two-thirds of all the probes in the derivation and validation data sets fail the normality assumption and the homogeneity of variation assumption. This has implications in the interpretation of regression models based on gene expression data.

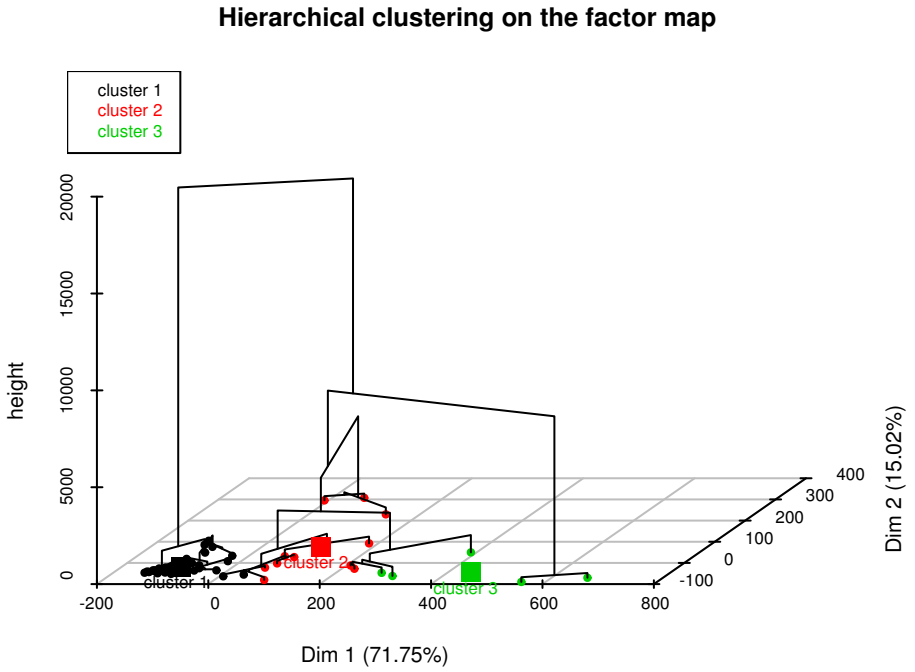


Figure 8.9: Hierarchical agglomerative clustering of the paediatric sepsis data set (GSE13904). Supervised clustering performed on expression data from 4048 probes corresponding to 2277 probes identified from the CAP derivation data set.

Gene expression data from the CAP cohort shows three clusters based on hierarchical agglomerative clustering of the principal components of the most variant probes. Nearly a third of the variation in the expression of these probes is explained by the first two principal components. This structure is associated with a number of covariates including total white cell counts, differential cell counts, and markers of organ function, especially renal impairment. There is also an association with mortality risk scores and outcomes, as well as a less tenuous association with the time to death. Supervised clustering of the CAP validation and FP derivation data sets, and an external data set of paediatric sepsis reveals similar structure in these data sets. In the FP data set, the only covariates associated with the clusters are the differential cell counts, but the pattern of association is consistent between the FP and CAP cohort. The association between the clusters and mortality risk scores is maintained.

The clusters were characterised by a combination of regression models and Bayesian hierarchical models. Panels of genes that were differentially regulated between the clusters were identified along with pathways and gene sets related to these genes. One of the clusters is associated with a low mortality and has predominantly repressed pattern of expression. The other two clusters have comparably high mortality but have significant differences in expression in core inflammatory pathways including the class II MHC. Intriguingly one of the clusters is characterised by up-regulation of genes related to viral infections. The robustness of the cluster subdivisions in paediatric sepsis supports the theory of distinct inflammatory patterns in sepsis caused by underlying structural genetic variation.

9

The Genomic Response in Sepsis

9.1 Background

The sepsis response consists of a pre-programmed sequence of events in response to an infection. The response is complex and involves orchestrated interactions between cells, the host genome and proteome, and the genome of the invading pathogen. This chapter describes the genomic response to infection by comparing gene expression in sepsis to baseline gene expression derived from controls. The responses in community-acquired pneumonia (CAP) and faecal peritonitis (FP) are analysed separately and shared and unique features are described. The biological pathways and gene sets involved in the response are identified.

9.2 Community-acquired pneumonia response

There are two covariates that could make interpretation of the sepsis response difficult. The first is the structure identified in 8.3 on page 87 representing distinct patterns of inflammation, and the second is the day of sampling and its relationship to evolution of the inflammatory state.

As shown in the previous chapter, the sub-structure in the data is associated with differences in gene expression including key components of the immune response like the class II MHC molecules. Ignoring these differences would result in spurious inferences. The solution is to make two comparisons: one compares each cluster independently to the baseline to identify cluster specific events and the second compares all

samples to baseline with the cluster membership specified as a confounding variable in the regression model. The latter identifies events that are robust to cluster membership.

The day of sampling is a much more contentious problem. Samples were collected on days 1, 3 and 5 of ICU admission. So "day 1" can be assumed to correspond to a clinical time point when there is a failure of physiological homoeostasis resulting in the inability to maintain organ function, especially adequate oxygenation and vasomotor stability. Whether this clinical phenotype correlates to a specific pattern of gene expression is unknown. Moreover, as described in section 11.6 on page 191, samples from a given patient were highly correlated, and the day of sampling was not a major component of the variance in the data set. Therefore, for this analysis the day of sampling was ignored, and repeated samples from individual patients were treated as biological replicates.

Cluster specific responses

Regression Models

Linear models were constructed with the cluster memberships as a predictor variable. Baseline samples were obtained prior to induction of anaesthesia from the cardiac surgery cohort(section 13.3 on page 221). The difference in gene expression between each cluster and the baseline samples was specified as an independent contrast.

Probes with at least 1.5 fold change in expression and an adjusted p value (FDR) of less than 0.05 were considered significant. To identify robust changes more stringent thresholds were used and a minimum fold change of 2.0 and a FDR of 0.01 were considered significant. The number of probes and genes that meet these thresholds are summarised in Table 9.1. A panel of less than 1000 genes constitute the sepsis response in each cluster. The predominant pattern is that of repression or gene silencing in all clusters. Cluster 2 is least differentiated from controls while cluster 1 is the most differentiated.

The top ten genes in each of the three contrasts are listed in Table 9.2. Clearly there is considerable overlap between the top ten genes. Over the entire probe set, clusters 2 and 3 have similar changes with a correlation coefficient of 0.899. Clusters 1 and 3 show

		Stringent threshold			Relaxed threshold		
Contrast	**Type**	**Up**	**Down**	**Total**	**Up**	**Down**	**Total**
Cluster 1 vs baseline	Probes	58	272	330	258	712	970
	Genes	56	254	310	243	648	891
Cluster 2 vs baseline	Probes	17	45	62	97	164	261
	Genes	17	42	59	92	155	247
Cluster 3 vs baseline	Probes	23	104	127	162	469	631
	Genes	23	100	123	158	435	592

Table 9.1: Number of differentially expressed probes and genes between each of the clusters and the control samples. Stringent threshold = 2.0 fold change in expression and FDR < 0.01, Relaxed threshold = 1.5 fold change and FDR < 0.05

the least similar changes with a correlation coefficient of 0.829, while clusters 1 and 2 have a correlation coefficient of 0.869 (Figure 10.3).

Contrast 1		**Contrast 2**		**Contrast 3**	
Gene	**logFC**	**Gene**	**logFC**	**Gene**	**logFC**
SIGLEC14	5.9506	SIGLEC14	5.5599	SIGLEC14	5.5682
XYLT1	-2.5155	LOC100133662	-4.4554	LOC100133662	-4.2942
PASK	-2.2962	LOC644936	-4.1668	HS.508889	-1.2323
LOC100133662	-4.1319	HS.508889	-1.1698	XYLT1	-2.2550
HS.508889	-1.2024	RPS4Y1	-5.0861	IL23A	-1.3168
RPS26P54	-2.6897	XYLT1	-2.1559	PASK	-2.1190
TMEM204	-1.9348	GPR128	-2.0081	GPR128	-1.9695
IL23A	-1.2686	RPS26P54	-2.4614	RPS26P54	-2.5413
TXK	-2.4397	HLA-DRB1	3.9534	HNRNPA1P2	-3.3366
RPS4Y1	-4.4460	PASK	-1.9329	RPS4Y1	-4.5746

Table 9.2: Ten most differentially expressed genes in cluster-wise comparisons to baseline in sepsis due to CAP. Genes are ranked by statistical significance. Contrast 1 = Cluster 1 vs baseline, Contrast 2 = Cluster 2 vs baseline, Contrast 3 = Cluster 3 vs baseline. LOC644946 = beta actin pseudogene, LOC100133662 = withdrawn PubMed record

The similarities and differences between each of the responses are shown in Figures 9.2 and 9.3. The similarities show the shared component of the response while the differences represent the unique component of the response. The genes that constitute the shared and unique components are listed in Supplementary Tables 9 to 11. The

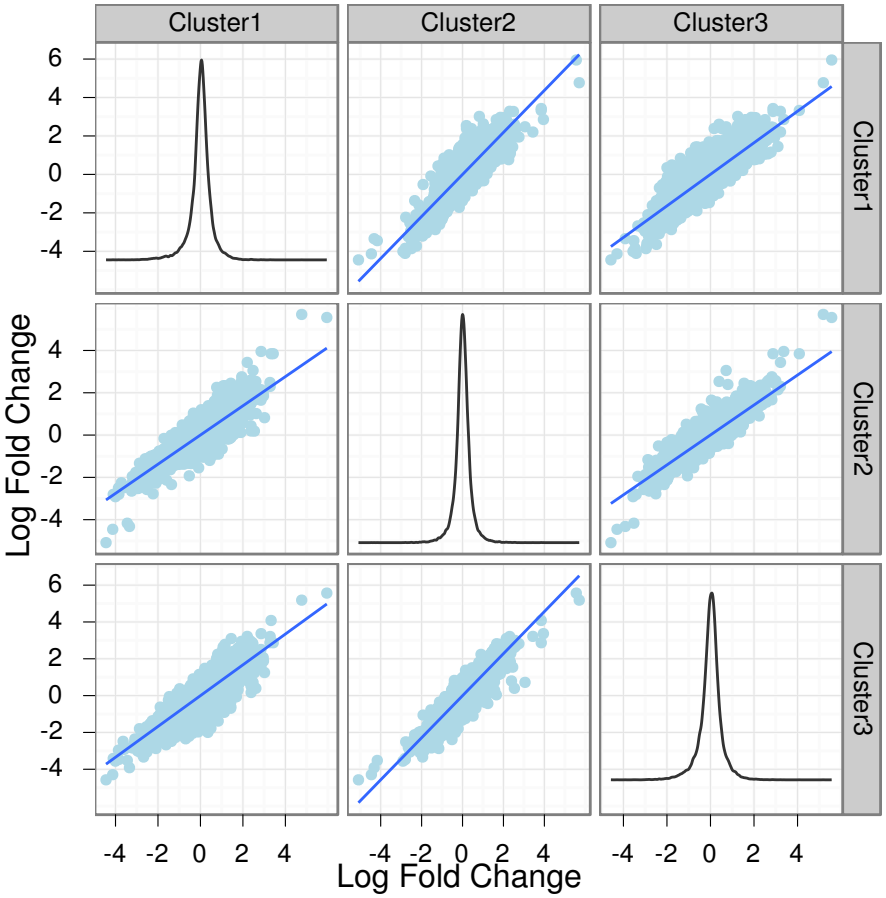


Figure 9.1: Correlation of mean change in expression values for cluster-wise comparisons with baseline samples. Panels in the right diagonal show the density distribution of the mean expression differences in the respective cluster vs baseline comparison. The other panels show a scatter plot of mean change in expression values in the clusters compared. The blue line in these panels is a regression line. Contrast 1 = Cluster 1 vs baseline, Contrast 2 = Cluster 2 vs baseline, Contrast 3 = Cluster 3 vs baseline

direction of gene expression change is same in all the responses, suggesting a consistent response profile in sepsis with the differences largely due to magnitude of expression rather than the pattern of expression.

Pathway Analysis

DAVID Functional Annotation The proportion of genes in a given gene set (defined on the basis of biological function, genomic location, or shared attributes) that are considered significant can be used as a measure of importance of the gene set. This is expressed as an enrichment score, the geometric mean of the *p* values for each gene calculated using a Fisher exact test. The pathways enriched in the DAVID functional

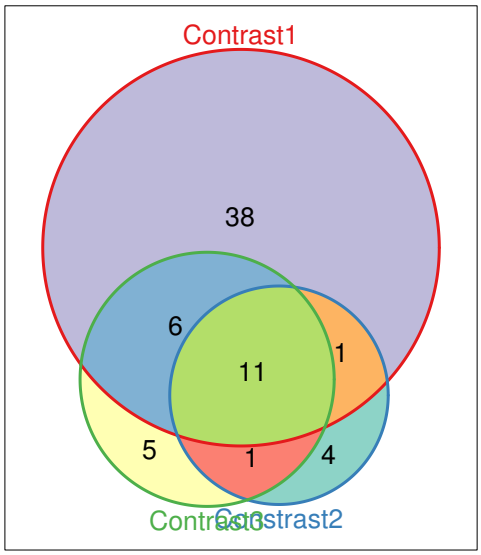


Figure 9.2: Overlap of genes induced with at least 2.0 fold change in mean expression and FDR < 0.01 in cluster-wise comparisons to baseline in CAP. Contrast 1 = Cluster 1 vs baseline, Contrast 2 = Cluster 2 vs baseline, Contrast 3 = Cluster 3 vs baseline. 38, 5, and 4 genes are uniquely induced in contrasts 1, 2 and 3 respectively. 11 genes (area shaded in green) are induced in all three contrasts.

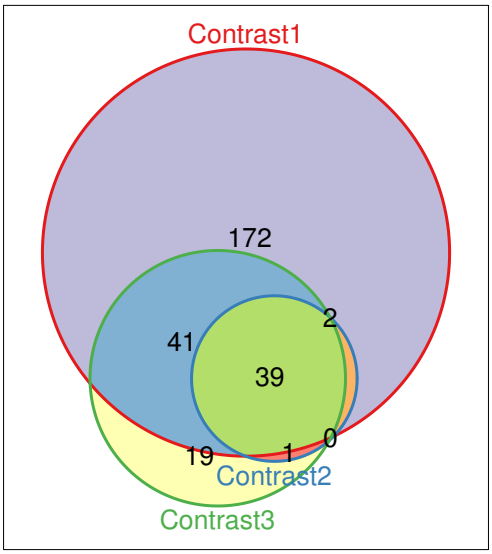


Figure 9.3: Overlap of genes down regulated with at least 2.0 fold change in mean expression and FDR < 0.01 in the CAP sepsis in cluster-wise comparisons. Contrast 1 = Cluster 1 vs baseline, Contrast 2 = Cluster 2 vs baseline, Contrast 3 = Cluster 3 vs baseline. 172 and 19 genes are uniquely repressed in contrasts 1 and 3 respectively. All genes repressed in contrast 2 are also repressed in at least one of the other contrasts. 39 genes (area shaded in green) are consistently repressed in all three contrasts.

clustering tool are listed in Table 9.3. Clearly the mechanisms of sepsis are very similar across the clusters. The predominant function of the differentially expressed genes in the sepsis response is related to the activity of ribosomal proteins. Ribosomal proteins are important components of gene regulatory networks,(Hu *et al.*, 2007) and it is interesting to note that at the genomic level, the sepsis response is defined in terms of regulatory networks instead of more canonical events such as cytokine expression.

Contrast	DAVID gene set
Cluster 1	Ribosomal protein (n = 41, ES = 18.53) RNA splicing (n = 5, ES = 4.07) rRNA processing (n = 14, ES = 3.68)
Cluster 2	Ribosomal proteins (n = 5, ES = 1.15)
Cluster 3	Ribosomal proteins (n = 20, ES = 7.64) RNA processing (n = 6, ES = 3.55) RNA splicing (n = 5, ES = 3.17) Transcriptional regulation (n = 5, ES = 2.88)

Table 9.3: DAVID functional annotation of the cluster-wise sepsis response in CAP. ES = enrichment score. Higher scores indicate better evidence for the annotation. n = number of differentially expressed genes from the data set that belong to the gene set.

Gene Set Enrichment Analysis Gene set enrichment can also be estimated more quantitatively by estimating enrichment scores in which each gene is weighted using a discriminatory test statistic. The modified gene set enrichment algorithm was used to calculate enrichment scores for each of the gene sets. Significance levels were estimated using a permutation process involving phenotype switching and estimating the distribution of enrichment scores for each gene set independently. A permuted p value of 0.05 was considered significant.

The gene sets enriched for each contrast are shown in Tables 9.4 to 9.6. A negative enrichment score indicates a repressed pathway while a positive enrichment score indicates an induced pathway. Larger values of the enrichment score indicate the presence of genes with large effect sizes while moderate scores indicate large number of genes with moderate effects.

The cluster 1 sepsis response is characterised by suppression of ribosomal processing. There is also a reduction in insulin synthesis and secretion resulting from altered gene expression in beta cells. This is consistent with clinically observed insulin resistance in sepsis (Robinson *et al.,* 2011) There is induction of TLR4 receptor cascade and cytokine production. Genes in the q22 karyosomal band in chromosome 5 are induced while clusters in chromosomes 2, 3, 12, and Y are repressed. There is increased expression of cytokine related genes with a simultaneous repression of IL17 and IL7 pathways. The VEGF pathway, normally induced in ischaemia, is repressed. The last finding is unexpected given the low PF ratio in cluster 1 (Table 8.6.)

The number of gene sets enriched in the cluster 2 response is smaller, probably as a result of the sensitivity of GSEA to differences in mean expression. Some gene sets, especially cytokine secretion and chromosomal clusters are shared with cluster 1, while others like BMP signalling are unique to cluster 2.

Gene sets in cluster 3 overlap those of cluster 1 and cluster 2. In particular, the viral RNA translation pathway, repressed in cluster 1 but not in cluster 2, is significantly enriched in cluster 3. This difference is consistent with the observations in section 8.4 on page 99. Specifically, in comparison to the other clusters, the viral replication gene set is induced in cluster 2, possibly as a result of a concomitant viral infection. Gene sets related to lipid transport and carbohydrate transport are induced in cluster 3; the significance of these changes is unclear.

Cluster invariant response

The set of 11 and 39 genes that are consistently up and down regulated with at least 2.0 fold change in mean expression and FDR less than 0.01 in all cluster responses shown in Figures 9.2 and 9.3 on page 116 represent the shared elements of the sepsis response. However, this list is generated based on cluster membership and has limited validity in terms of statistical association. The accurate identification of the shared response requires the construction of a linear model where the variance due to cluster membership is appropriately partitioned away to reveal the true underlying shared response.

Gene Set	Enrichment Score
Reactome ribosome processing	-0.80
Reactome influenza viral RNA transcription	-0.77
Reactome insulin synthesis and secretion	-0.70
Reactome peptide chain elongation	-0.83
Reactome beta cell gene expression	-0.78
Chromatin assembly	-0.63
RNA processing	-0.48
Covalent chromatin modification	-0.68
RNA splicing via trans-esterification reactions	-0.56
Histone modification	-0.71
KEGG primary immunodeficiency	-0.65
Intramolecular oxidoreductase activity	-0.70
Biocarta IL17 pathway	-0.75
Biocarta IL7 pathway	-0.63
Biocarta NKT pathway	-.0.57
Biocarta VEGF pathway	-0.57
chr12q14	-0.58
chr2p11	-0.49
chryq11	-0.61
chr3q12	-0.63
Reactome gene expression	0.55
Reactome Toll like receptor 4 cascade	0.62
Regulation of cytokine secretion	0.68
Regulation of protein secretion	0.59
Potassium ion transport	0.49
Synaptic vesicle	0.63
chr5q22	0.71

Table 9.4: Gene sets enriched in the cluster 1 vs baseline comparison in CAP sepsis

Gene Set	Enrichment Score
Reactome CD28 dependent PI3K signalling	-0.61
KEGG seleno-amino acid metabolism	-0.64
KEGG T cell receptor signalling	-0.65
Biocarta IL17 pathway	-0.72
Biocarta NKT pathway	-.0.60
Biocarta VEGF pathway	-0.61
chr2p11	-0.51
chryq11	-0.74
chr3q12	-0.58
chr14q13	-0.60
Chromatin remodelling complex	-0.64
Reactome signalling by BMP	0.60
Energy reserve metabolic process	0.60
Protein secretion	0.55
Exocytosis	0.56
Hormone metabolic precess	0.57
Cytokine secretion	0.63
Carbohydrate transmembrane transporter	0.69
Biocarta AMI pathway	0.71
Biocarta ALK pathway	0.53

Table 9.5: Gene sets enriched in the cluster 2 vs baseline comparison in CAP sepsis

Regression analysis was performed to identify the sepsis response by specifying cluster membership as an additional covariate. The coefficients due to the covariates were ignored and only the effects of the main contrast examined. 970 probes showed at least 1.5 fold change in gene expression and a FDR less than 0.05. The predominant pattern is that of repression (Figure 9.4); 712 probes are repressed and 258 probes are induced. This corresponds to 648 and 243 genes respectively.

The number of induced and repressed genes is similar to the response seen in cluster 1. There is considerable overlap between the genes that are differentially expressed in the cluster invariant response and cluster specific responses, with the former nearly subsumed by the latter. This suggests that the genomic response to infection is much

Gene Set	Enrichment Score
Reactome CD28 dependent PI3K signalling	-0.56
Reactome ribosome processing	-0.77
Reactome Influenza viral RNA transcription and replication	-0.73
Reactome insulin synthesis and secretion	-0.67
Reactome protein metabolism	-0.58
Reactome purine metabolism	-0.64
Reactome regulation of beta cell gene expression	-0.72
RNA splicing via trans-esterification reactions	-0.54
KEGG seleno-amino acid metabolism	-0.57
GTP binding	-0.57
Biocarta ETS pathway	-0.64
Biocarta NKT pathway	-0.59
chr2p11	-0.54
chryq11	-0.65
chr3q12	-0.69
Reactome striated muscle contraction	0.58
Lipid transport	0.56
Carbohydrate transport	0.64

Table 9.6: Gene sets enriched in the cluster 3 vs baseline comparison in CAP sepsis

larger than the differences that are seen between clusters. The large changes seen in the genomic sepsis response overwhelms the modest intra-cluster differences.

9.3 Faecal Peritonitis Response

There is also evidence of clustering in the FP gene signatures as shown in subsection 8.5 on page 105. However, there is considerable overlap between the clusters and the relationship of the clusters to outcome is less clear. Therefore, the sepsis response in faecal peritonitis was examined in the full data set with no attempt to differentiate between the clusters.

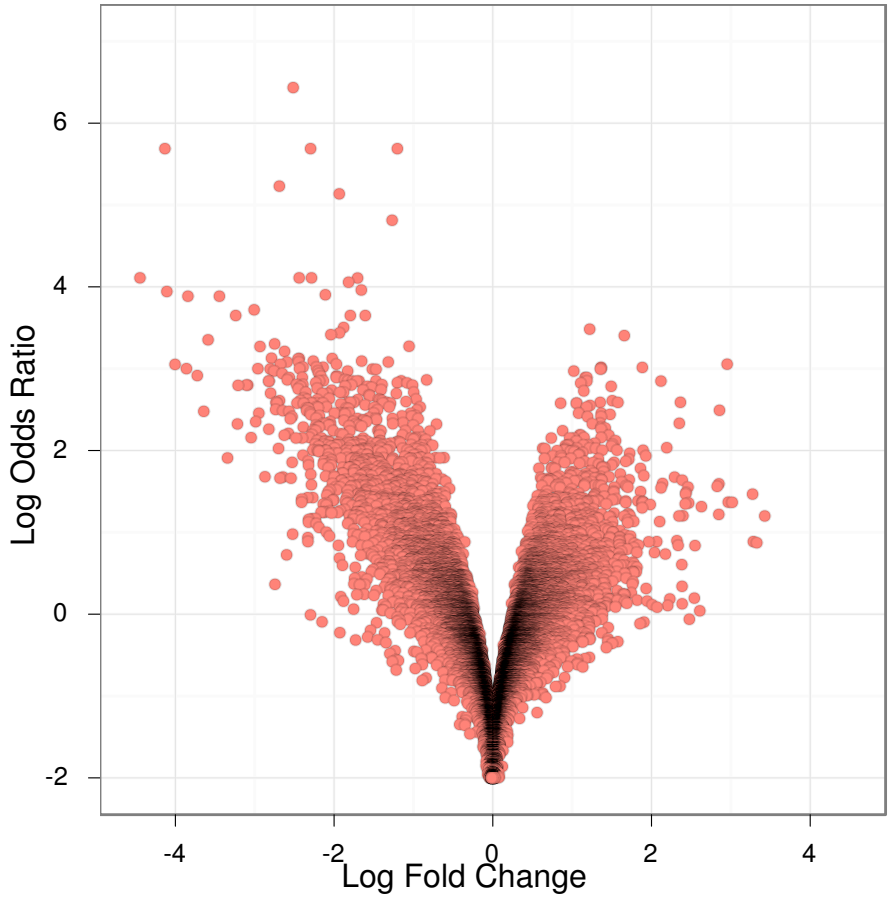


Figure 9.4: Probe distribution of the cluster invariant sepsis response in CAP. Each point is a probe. Significant probes have a larger absolute log fold change and greater odds ratios (outer and higher points in the volcano). The volcano is asymmetrical with a wider and taller eruption on the left, suggesting predominant repression.

Regression Models

Simple regression models were constructed for each probe with the comparison of FP samples to baseline samples as the contrast. The predominant pattern of response is repression with between 64% and 70% of the genes repressed dependent on the filtration threshold (Table 9.7). The asymmetric distribution of significance and fold change differences is comparable to the CAP cohort(Figure 9.5). The top 10 genes are listed in Table 9.8. The list includes genes involved in signalling systems, transcription factors and immune related genes. *ATG16L1* has a reported association with inflammatory bowel disease,(Glubb *et al.,* 2011) *Helicobacter pylori* infections,(Raju *et al.,* 2012) and is involved in autophagy (Lapaquette *et al.,* 2012).

Threshold	Type	Up	Down	Total
Stringent	Probes	36	85	121
	Genes	35	82	117
Relaxed	Probes	179	310	480
	Genes	162	291	453

Table 9.7: Number of differentially regulated genes/probes in the sepsis response in faecal peritonitis. Stringent threshold = 2.0 fold change in mean expression and FDR < 0.01, Relaxed threshold = 1.5 fold change in mean expression and FDR < 0.05.

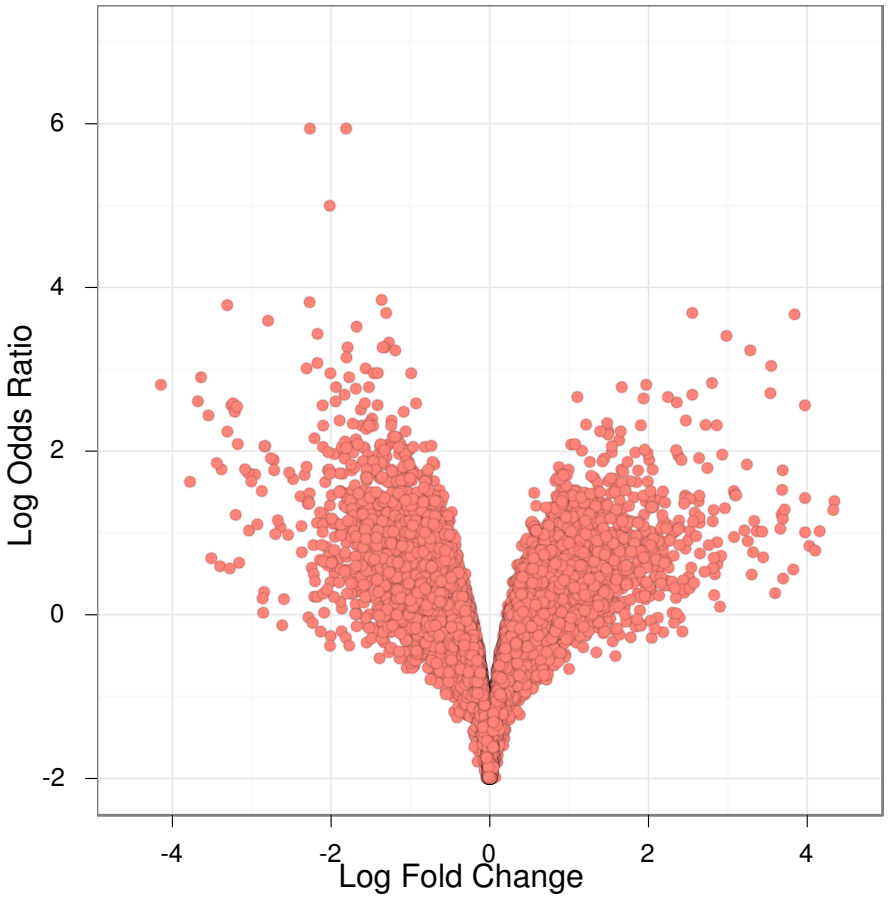


Figure 9.5: Volcano plot of the sepsis response in FP. Each point is a probe. Significant probes have a larger absolute log fold change and greater odds ratios (outer and higher points in the volcano). The volcano is asymmetrical with a wider eruption on the right but a taller eruption on the left. This suggests that more probes show induction but a larger proportion of the repressed probes are statistically significant. The former could be due to greater variance in probe expression, while the latter due to changes in probes that normally show tighter regulation.

Gene	log FC	Chromosome	Function
GPR128	-2.6	3	G protein coupled receptor
HLA-A29.1	-5.2	6	HLA-A antigen
NKX3-1	-2.3	8	NK3 homeobox
RAPGEF6	-2.0	5	Guanine nucleotide exchange factor
LOC1000133662	-4.5		withdrawn record
RPS4Y1	- 5.2	Y	Component of 40S ribosomal subunit
ATG16L1	-1.4	2	Autophagy related protein
FCRL3	-2.3	1	Immunoglobulin receptor superfamily
RPS26P54	-3.3	18	Ribosomal protein pseudogene
PPARG	2.5	3	PPAR nuclear receptor

Table 9.8: Top 10 genes of the faecal peritonitis response

Pathway Analysis

The functional basis of the faecal peritonitis response was investigated with GSEA using the same methodology as described in the CAP patients. Gene sets with a permuted *p* value of less than 0.05 were considered significantly enriched and are listed in Table 9.9. Faecal peritonitis is associated with repression of AKT and CTLA4 signalling and induction of the complement cascade.

Comparisons with CAP

To identify the shared component of the sepsis response across diagnostic labels, the sepsis response in faecal peritonitis was compared to the sepsis response in community-acquired pneumonia. Quantitative comparisons were made at the level of genes and gene sets.

Genes with at least 1.5 fold change difference in expression and a FDR of 0.05 in either cohort were used for comparisons. Fold change correlations over the entire probe set is shown in Figure 9.6 and over the shared significant probe set is shown in Figure 9.7. The overall correlation is only about 74%. However, the significant probe sets have a correlation coefficient of 96.8% suggesting that the largest changes in the sepsis response are similar.

Gene Set	Enrichment Score
Reactome CD28 dependent PI3K AKT signalling	-0.65
Reactome chemokine receptors	-0.54
Reactome CTLA4 inhibitory signalling	-0.57
Covalent chromatin modification	-0.65
Ras GTPase binding	-0.59
Carbohydrate transmembrane transporter activity	0.65
Biocarta IL17 pathway	-0.78
Biocarta VEGF pathway	-0.54
chryq11	-0.66
Reactome complement cascade	0.73
Reactome initial triggering of complement	0.79
Cytokine secretion	0.63
KEGG porphyrin metabolism	0.52
KEGG drug metabolism and cytochrome P450	0.51
KEGG complement and coagulation cascades	0.47
Biocarta complement pathway	0.72
chr4q28	0.53

Table 9.9: Gene sets enriched in the sepsis response in faecal peritonitis

The correlation coefficient is derived from probes that are significant in both cohorts. This overlap is shown in Figure 9.8. The shared component includes 23% of the genes in the CAP response and 45% of genes in the FP response. CAP is associated with a greater response than FP, in terms of the number of differentially expressed genes for the same threshold. However, the magnitude of response seen in the shared significant probe is nearly identical in both contrasts. The latter suggests a shared sepsis component that is quantitatively and qualitatively similar across diagnostic subsets. The former could represent the unique component of the CAP response or differences in response attributable to different sampling times. The observed differences could also result from the differences in the focus of infection, pathogen load and type, and organ effects such as impaired oxygenation in CAP.

The functional basis of the shared and unique components is clearly shown by the overlap of enriched gene sets with the gene set enrichment analysis. CD28 depen-

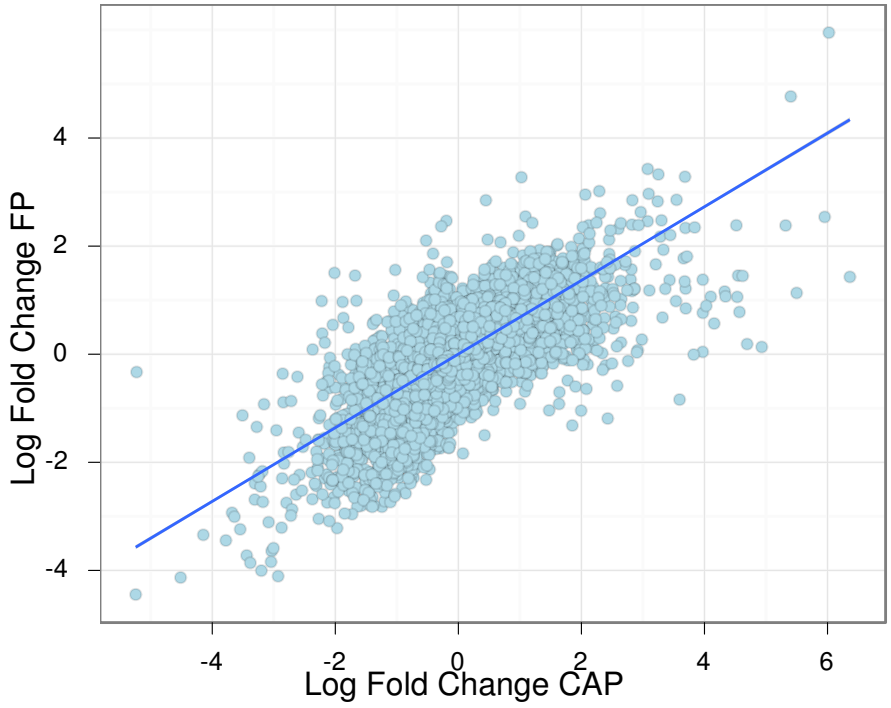


Figure 9.6: Correlation between the probes in FP and CAP sepsis response for all probes. Each point represents the mean expression change when compared to baseline for a probe. The regression line is shown in blue.

dent AKT signalling, Biocarta IL17 and VEGF pathways, chromatin modification, and karyosomal cluster q11 in chromosome Y are consistently repressed in both cohorts. Similarly, cytokine secretion is induced in both cohorts. induction of the complement cascade and repression of the CTLA4 pathway is unique to the FP cohort while repression of the viral translational systems are unique to the CAP cohort.

9.4 Validation

The genomic response to sepsis in the CAP and FP cohorts were validated in the appropriate validation cohorts described in section 5.2 on page 52. Data from both the cohorts cannot be directly compared because of differences in the hybridization platform (subsection 8.2 on page 86). In addition, the FDR in the regression model is a Bayesian estimate of the significance level based on the variation in the entire data set. Changes

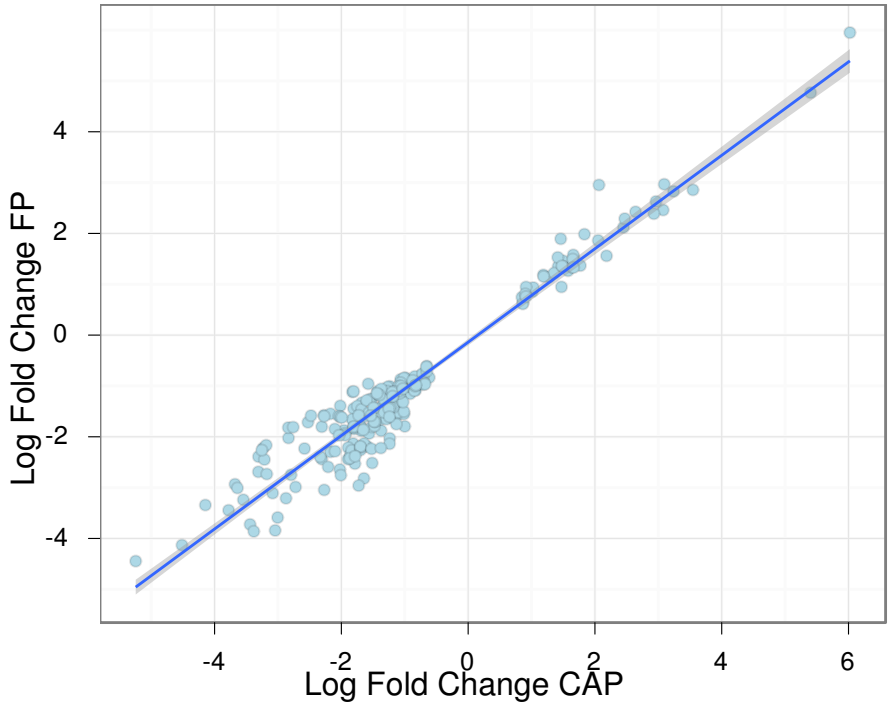


Figure 9.7: Correlation between the probes in FP and CAP for significant probes.Each point represents the mean expression change when compared to baseline for a probe. The regression line is shown in blue. Significant probes were defined as probes with at least 1.5 fold change in mean expression and FDR < 0.05 in both the contrasts.

in the number of probes, probe characteristics, and hybridisation efficiency alters this estimate.

To overcome these problems the following approach was used to validate the findings. Significantly expressed genes in the derivation data set were first identified. The validation data set was reduced to these genes, so both data sets consisted of comparable information. A regression analysis was performed in the validation cohort using a model identical to the model used in the derivation cohort. This eliminates methodological bias in the results. Significance was determined based on unadjusted *p* values, since only specific hypothesis were tested.

CAP Validation

Of the 891 unique genes in the derivation cohort only 621 genes are represented in the validation data set. The reduction in gene number is due to differences in probes

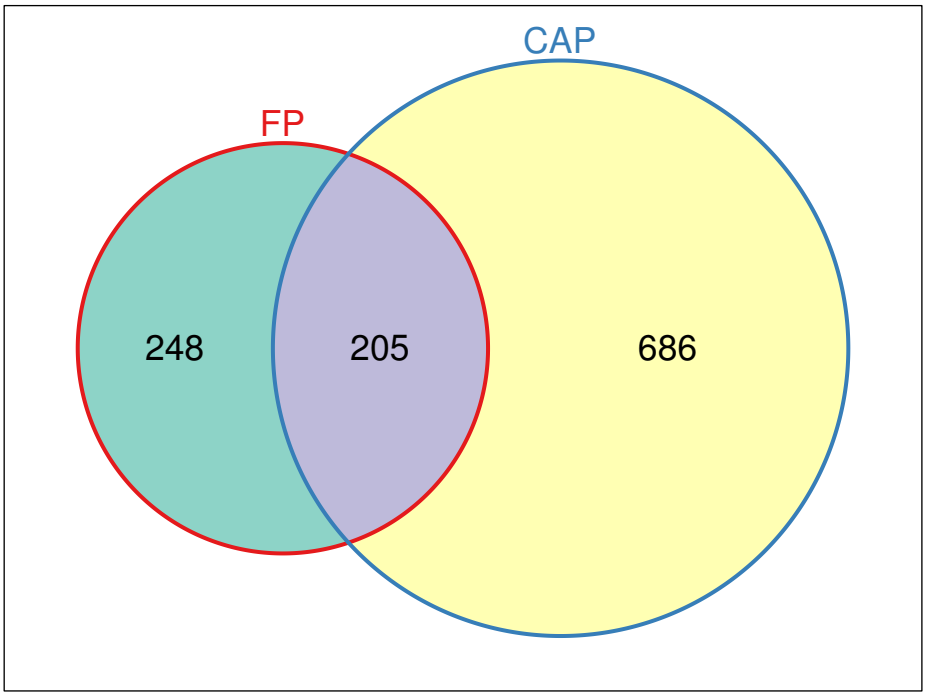


Figure 9.8: Venn diagram showing the overlap of genes that are differentially expressed in the sepsis response in FP and CAP.

represented on the platforms used (Table 3.1). The changes in gene expression of these genes in each cohort compared to the control group are only moderately correlated (correlation coefficient 65.7%) as seen in Figure 9.9. Of the 621 genes tested, only 247 (39%) achieve statistical significance.

FP Validation

Of the 453 unique genes in the faecal peritonitis derivation cohort, 359 are present in the validation cohort. The reduction in gene number is due to differences in probes represented on the platforms used (Table 3.1). As with the validation for community acquired pneumonia, there is only a moderate correlation (coefficient 62.8%) between the expression data sets (Figure 9.10). Onlu 44% of the genes (n = 158) show statistically significant differences in expression in the validation cohort.

The correlation between the sepsis responses between the derivation and validation data sets are modest for both the CAP and FP cohorts. There are a number of common

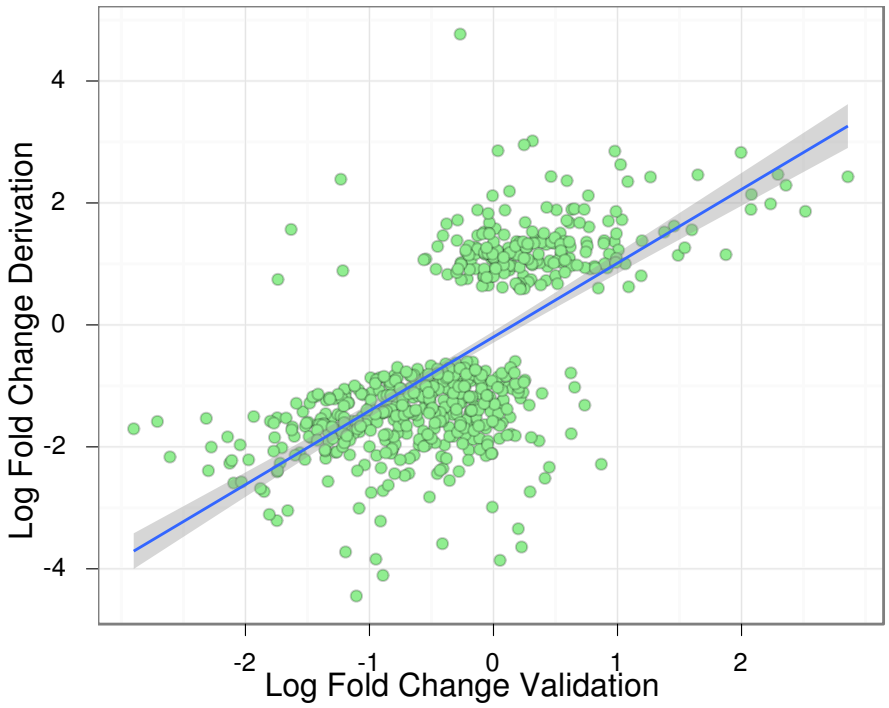


Figure 9.9: Correlation between the genes determining the CAP response in the derivation and validation cohort. The plot shows 621 genes that are commonly present in both platforms.

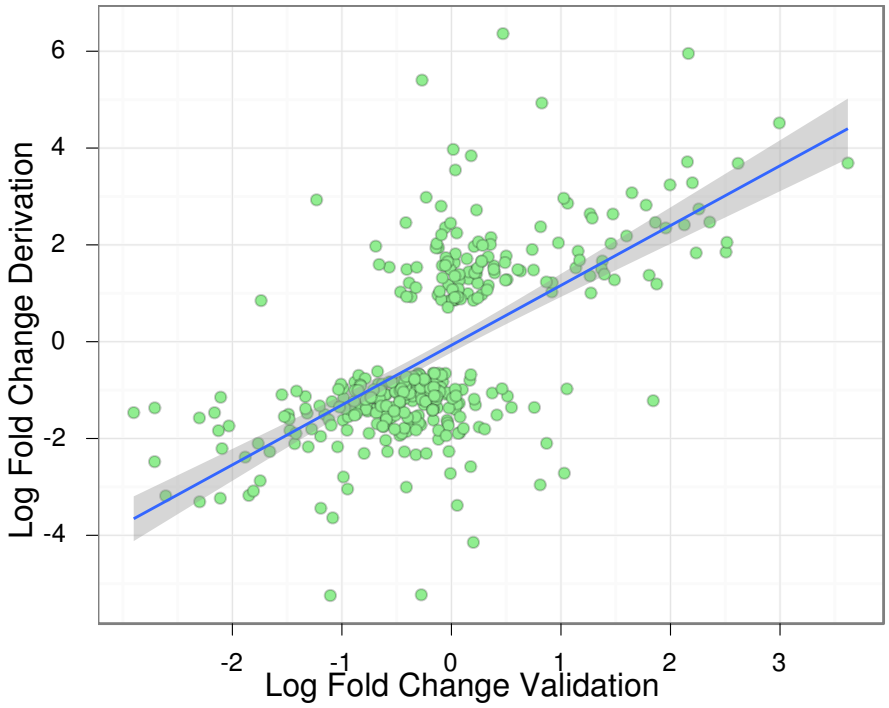


Figure 9.10: Correlation between the genes determining the FP response in the derivation and validation cohort. The plot shows 359 genes that are commonly present in both platforms.

observations. First, nearly a quarter of the genes are lost between the comparisons due to hybridisation platform differences. It is difficult to extrapolate the observed results to the missing genes. Secondly, for the observed results in both cohorts, there is a reasonable correlation between gene expression patterns (more than 60%), despite more modest statistical significance (about 40%). This could be the result of lower sample size in the validation cohort. This conclusion is partly supported by the observation in Figures 9.9 and 9.10, where there is an asymmetry in the distribution of the points around the regression line. In other words, the effect sizes are systematically under-estimated in the validation cohort when compared to the derivation cohort for both the diagnostic data sets. A systematic bias due to platform specific biases in hybridization efficiencies is also possible, and cannot be ruled out.

9.5 Summary

The genomic response to sepsis consists of a pre-programmed sequence of orchestrated changes. The predominant pattern of gene expression change observed in severe sepsis due to community-acquired pneumonia is gene silencing.

Clusters representing distinct inflammatory patterns were identified in the CAP cohort in the exploratory analysis. These clusters share elements of the response with additional components unique to each cluster. The shared component is strong with more than 80% correlation in the gene expression pattern across clusters when compared to baseline samples. The shared pattern was related to changes in ribosomal protein gene regulatory networks. Cluster specific differences in gene expression are modest but are related to important immune related systems such as interleukins, chromatin remodelling, TLR4 signalling, and processes not conventionally related to immune functions such as hormone metabolic process, and genes related to carbohydrate and lipid metabolism.

Changes related to sepsis are of much larger magnitude than the intra-cluster differences. The shared sepsis response in community acquired pneumonia involves induction of cytokine pathways and repression of IL17 and VEGF pathways, chromatin

modification and ribosomal processing gene sets. Pathways unique to some clusters but not others include the viral mRNA translational system, suggesting the presence of concomitant viral infection.

The gene expression patterns in FP, like CAP, consists of gene-silencing but with fewer significant genes. This could be related to disease intensity, time of sampling or reflect fundamental differences in pathophysiology. Some of the top genes like *ATG16L1* have been implicated in intestinal pathologies. Like CAP, the gene expression in FP consists of induction of cytokine related pathways and repression of the IL17 and VEGF systems and chromatin modifying genes. However, FP is also uniquely associated with complement activation and not with viral mRNA translational gene sets.

The genomic response to sepsis was validated in the validation cohorts. Typically about a quarter of the gene hits in the derivation cohort are missing in the validation cohort due to platform specific probe distribution issues. The remaining genes show modest correlation with a consistent pattern of low effect sizes in the validation cohort, related probably to small sample sizes. About 40% of the genes have statistical significance in the validation cohorts. As expected the loss of significance is the result of loss of power with small sample sizes, and genes with large differences in expression show a consistent effect across cohorts.

10

The Genomic Response and Survival in Sepsis

10.1 Background

The identification of the genomic signatures of survival in sepsis is one of the primary aims of this work. This chapter examines the genomic signatures of survival in severe sepsis due to community-acquired pneumonia and faecal peritonitis.

The definition of mortality in sepsis varies and relates to difficulties in identifying mortality attributable to sepsis. Clinical studies have variably used mortality at 28 days and mortality during ITU or hospital stay as end points (Dhainaut *et al.,* 2003; Goodin *et al.,* 2011; Lagu *et al.,* 2011). All the end points have clinical validity but very little scientific justification. For this data set, mortality at 28 days might be too distant an event to be predicted by the genomic signature observed on the day of admission to ITU. The use of early mortality (usually defined at 7 days after admission) might miss outcome events with loss of information in the analysis. Mortality at 14 days was chosen as a reasonable end point for this data set. The use of 14 day mortality is conservative and should result in fewer false positive results with the disadvantage of a reduced sample size.

While mortality is an adequate and valuable clinical end point, the application of mortality as an end point in genomic studies is problematic, especially when trying to define the mechanistic basis of poor outcomes. Mortality in sepsis is the result of diverse and sometimes unrelated events. These include varying patterns of acute organ dysfunction in the context of pre-existing organ dysfunction, general failure of physiological systems, and non-biological factors such as end-of-life decisions, as well as

sudden unexpected events like pulmonary embolism and cardiac arrhythmias that are remote from the immune response and its consequences. This phenotypic heterogeneity is difficult to handle in genomic studies.

The value of survival signatures is increased if the patterns of gene expression can be related to surrogate morbidity measures such as number of days in hospital and number of days of organ support. The relationship of organ function to mortality has strong biological plausibility and has been clearly demonstrated in several large studies (Knaus *et al.,* 1985; Le Gall *et al.,* 1993; Vincent *et al.,* 1996).

Mortality at 14 days after ICU admission was used as the primary outcome variable in this chapter while the number of days of respiratory, renal, and inotropic support during the ICU stay and organ function scores are used as the surrogate outcome variables. The analysis was performed in the derivation cohort and the results confirmed in the validation cohort.

10.2 Mortality in CAP

Exploratory analysis of CAP samples revealed a structure in the data based on gene expression that is associated with mortality risk (section 8.3 on page 87). In particular one subgroup had a in hospital mortality risk of 30% while the other groups had a mortality risk of 50% but with different time intervals to death (median duration of 10 and 22 days respectively). The difference between the clusters involves differences in expression of important immune related genes such as those belonging to the class II MHC. Therefore, the mortality signature in each group was determined independently.

Cluster specific mortality

Linear Models

Differences in gene expression between survivors and non-survivors in each cluster were examined independently. All comparisons are reported as changes in non-survivors when compared to survivors. A single analysis incorporating all models was performed,

instead of three separate analyses with the data subset by clusters. This allows for the FDR to be estimated more accurately from the overall variance in the data.

The results are summarised in Table 10.1. With stringent thresholds, changes in expression of a small panel of genes differentiate the non-survivors from survivors. The differences are greatest in cluster 1 and least in cluster 3, a pattern which matches the relationship of clusters to baseline samples (Figure 10.1).

		Stringent threshold			Relaxed threshold		
Contrast	**Type**	**Up**	**Down**	**Total**	**Up**	**Down**	**Total**
Cluster 1	Probes	195	112	307	506	310	816
	Genes	173	101	204	452	274	726
Cluster 2	Probes	94	77	171	243	195	438
	Genes	82	70	152	216	179	395
Cluster 3	Probes	32	17	49	77	45	122
	Genes	30	16	46	71	42	113

Table 10.1: Number of differentially expressed probes and genes between non-survivors and survivors for each cluster in CAP sepsis. Stringent threshold = 2.0 fold change in mean expression and FDR < 0.01, Relaxed threshold = 1.5 fold change in mean expression and FDR < 0.05. Each contrast compares non-survivors to survivors for a given cluster.

The effect size of these changes are of interest since large effect sizes imply greater confidence in the results and genes with large differences between the groups could function as potential biomarkers. The distributions of the fold change differences are shown in Figure 10.2. The median log2 fold change is approximately two for all clusters in either direction. This corresponds to at least four fold change in relative expression for 50% of the probes. There is a slight predominance of induced genes.

The relationship between the probes in each list allows the identification of shared and unique components, among the clusters, of the differential gene expression associated with mortality in CAP. Over the entire probe set, the least correlation is seen between clusters 1 and 2 with a correlation coefficient of 75.2% (Figure 10.3). The greatest correlation is seen between clusters 1 and 3 (coefficient of 83.5%). The correlation between clusters 2 and 3 is 82.8%.

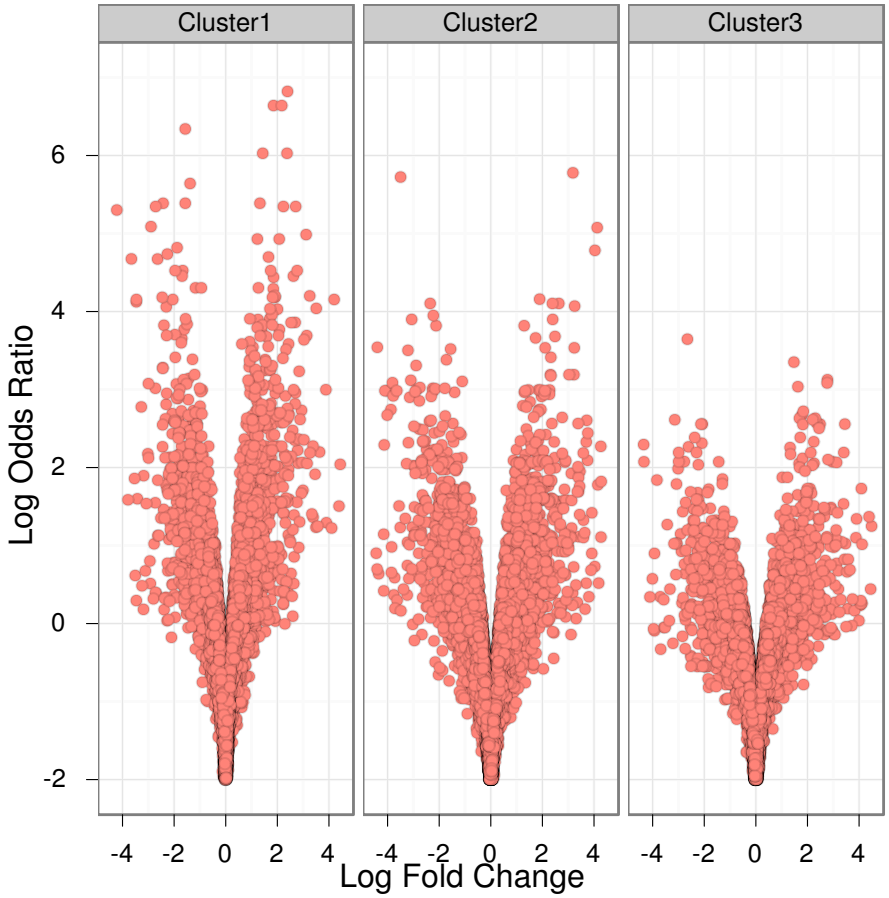


Figure 10.1: Volcano plot of survival in CAP sepsis for each cluster. While the effect sizes are similar across all three clusters, clusters 1 and 2 contain more statistically significant differences in probe expression. The lack of significance in cluster 3 is probably related to the low sample size in that cluster.

The overlap of genes identified with the stringent threshold (2.0 fold change in mean expression and FDR < 0.01) among the clusters are shown in figures 10.4 and 10.5. The gene set in cluster 3 is nearly a complete subset of the cluster 1 gene set, with highly correlated expression among its genes (Figure 10.6).

A panel of 37 genes (22 induced, 15 repressed) are consistently associated with non-survival among all three clusters. These genes along with their biological roles are listed in Table 10.2. Overall the picture is one of increased expression of Ig receptors and increased energy metabolism with decrease in class II MHC activity. There are intriguing associations with genes associated with mitochondrial activity, especially in relation to apoptosis and glucose metabolism, and cardiomyocyte related proteins in the context of depressed myocardial function in clinical sepsis.

Induced			Repressed		
Gene	**Chr**	**Role**	**Gene**	**Chr**	**Role**
ARHGEF19	1	Rho guanine exchange factor	C14ORF159	14	Nuclear encoded mitochondrial gene
C1QC	1	Complement component C1Q	C21ORF7	21	TBFb activated kinase
C7ORF68	7	HILPDA Hypoxia inducible gene	C7ORF28B	7	Vacuolar protein
CLEC11A	19	Haematopoietic growth factor	CTSS	1	Protease for class II presentation
ENO1	1	Alpha enolase	CX3CR1	3	Chemokine receptor
FABP4	8	Fatty acid binding protein	EAPP	14	modulation of E2F transcription
GAPDH	12	Carbohydrate metabolism	FCGR3B	1	Fc fragment of IgG
GPI	19	Energy metabolism	FGL2	7	Fibrinogen like protein
KCTD12	13	Potassium channel domain	HGD	3	amino acid catabolism
KIR2DL1	19	Killer cell Ig receptor	KLF4	9	endothelial zinc finger protein
KIR2DL4	19	Killer cell Ig receptor	HTATSF1P2	6	HIV1 Tat specific factor pseudogene
KIR2DL5A	19	Killer cell Ig receptor	LOC731682	6	HLA-DQA - class II MHC
KIR2DL5B	19	Killer cell Ig receptor	NEBL	10	cardiac muscle related protein
KIR3DL1	19	Killer cell Ig receptor	SCGB1C1	11	Secretoglobin
KIR3DL3	19	Killer cell Ig receptor			
LAMC1	1	Extracelluar glycoprotein			
MPO	17	Myeloperoxidase			
PLIN2	9	Marker of lipid accumulation			
PRSSL1	19	KLK10 Serine protease			
RHAG	6	Rhesus associated glycoprotein			
VAT1	17	Membrane protein of cholinergic vesicles			

Table 10.2: Genes with consistent differences in mean expression in all three clusters in CAP. The directional changes are reported for non-survivors in comparison to survivors.

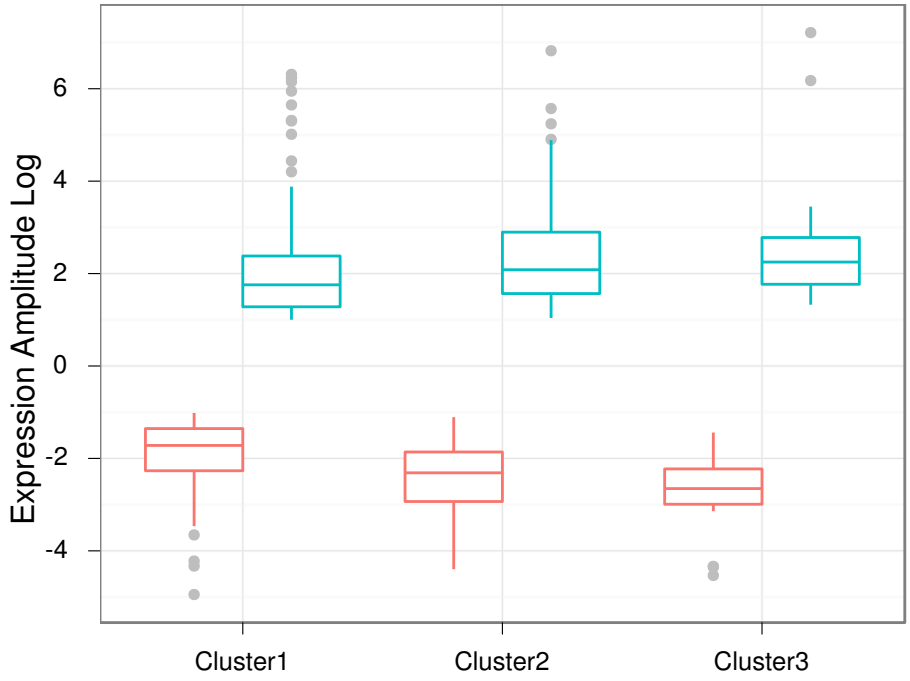


Figure 10.2: Distribution of effect sizes (change in mean expression) in non-survivors compared to survivors in CAP sepsis for each cluster (repressed genes in red and induced genes in blue). Nearly 50% of probes in each cluster show a 4 fold change in expression in non-survivors when compared to survivors.

Pathway Analysis

GSEA was used to quantitatively identify gene sets enriched for survival differences. A permutation procedure involving phenotype label switching was used to determine significance. The permutation procedure was not stratified for clusters. Gene sets with a permuted *p* value less than 0.05 were considered significantly enriched.

Pathways associated with survival in cluster 1 are shown in Table 10.3 along with the enrichment scores. A positive enrichment score indicates an induced pathway and vice versa. Larger enrichment scores indicate genes with large changes in mean expression. The major themes revealed by the pathways are increased cell mitotic activity, along with DNA elongation and repair mechanisms and chromatin activity. Non-survivors also have an induced hypoxic response and alterations in glucose metabolism. There is also a global suppression of the immune response including the adaptive immune response signalling.

Gene Set	Enrichment Score
Reactome activation of the pre-replicative complex	0.76
Reactome amino acid transport across plasma membrane	0.63
Reactome DNA strand elongation	0.77
Reactome telomere extension	0.75
Reactome Fanconi anaemia pathway	0.78
Reactome gluconeogenesis	0.62
Reactome glycolysis	0.82
Reactome mitosis	0.59
Reactome DNA polymerase repair	0.77
Cytokinesis	0.67
Negative regulation of DNA metabolism	0.69
Fatty acid oxidation	0.81
Mitotic sister chromatid segregation	0.79
Sister chromatid segregation	0.79
Response to hypoxia	0.64
KEGG pentose phosphate pathway	0.70
Intramolecular oxidoreductase activity	0.78
Biocarta MCM pathway	0.76
Reactome downstream TCR signalling	-0.59
Reactome generation of second messenger molecules	-0.71
Reactome BMP signalling	-0.63
Adaptive immune response	-0.62
Activation of the immune response	-0.62
Exocytosis	-0.65
KEGG Type 1 diabetes mellitus	-0.71
KEGG Asthma	-0.74
KEGG Allograft rejection	-0.79
Biocarta DC pathway	-0.63
SA MMP Cytokine connection	-0.82

Table 10.3: Gene sets enriched in the non-survivors of the first cluster in the CAP derivation cohort

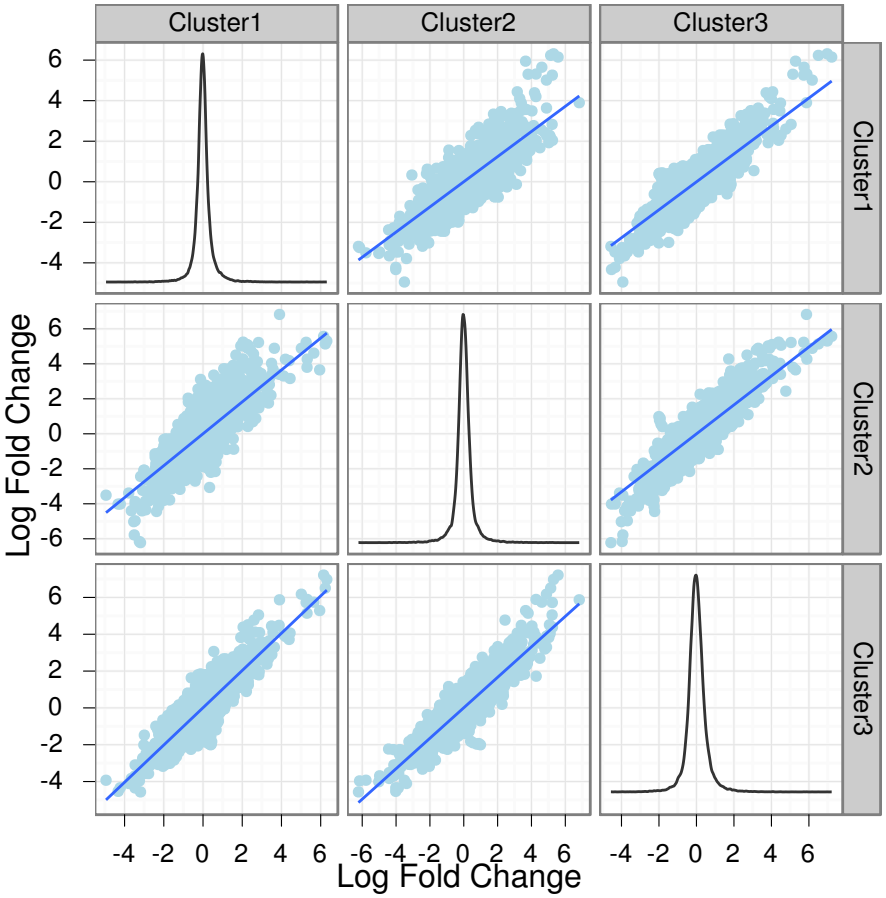


Figure 10.3: Correlation of the mean expression differences of all probes between non-survivors and survivors for each cluster in sepsis due to CAP. The right diagonal panel plots the density distribution of the mean expression differences for each cluster. Other panels show the scatter plot relating two clusters. The line in blue is the regression line for each pairwise comparison.

Pathways associated with mortality in cluster 2 are shown in Table 10.4. The induction of glucose metabolism related gene sets and increased mitotic and DNA related activity are shared with cluster 1, as are repression of the immune response, lymphocyte activation, and antigen processing and presentation gene sets.

Pathways associated with mortality in cluster 3 are shown in Table 10.5. Again there is induction of glucose metabolism and suppression of the immune response pathways. For cluster 3 the strongest association is seen with the Reactome classical antibody mediated complement activation gene set. This is interesting as the complement system was not identified as a significant mechanism of the sepsis response for CAP in the

Gene Set	Enrichment Score
Reactome amino acid transport	0.59
Reactome Gluconeogenesis	0.66
Reactome Glycolysis	0.78
Reactome PD1 signalling	0.80
Reactome pyruvate metabolism	0.68
Reactome regulation of insulin by ACh	0.56
Reactome regulation of insulin by FFA	0.61
Reactome phospholipase C cascade	0.73
Fatty acid oxidation	0. 66
Homoeostasis of number of cells	0.57
KEGG pentose phosphate pathway	0.62
KEGG glycolysis and gluconeogenesis	0.55
Intramolecular oxidoreductase activity	0.72
Reactome CD28 costimualation	-0.55
Reactome downstream TCR signalling	-0.61
Reactome generation of second messenger molecules	-0.75
Cell recognition	-0.79
Adaptive immune response	-0.55
Activate of immune response	-0.55
KEGG beta alanine metabolism	-0.59
KEGG antigen processing and presentation	-0.50
KEGG cytosolic DNA sensing pathway	-0.55
KEGG Type 1 diabetes mellitus	-0.71
KEGG ASthma	-0.69
KEGG Allograft rejection	-0.75
Chemokine receptor binding	-0.54
Biocarta CSK pathway	-0.70
Biocarta LAIR pathway	-0.66
Biocarta CTL pathway	-0.70
Biocarta MTA3 pathway	0.70
Biocarta cytokine pathway	-0.68
Biocarta DC pathway	-0.62
Biocarta NO2 IL12 pathway	-0.71

Table 10.4: Gene sets enriched in the non-survivors in the second cluster of the CAP cohort

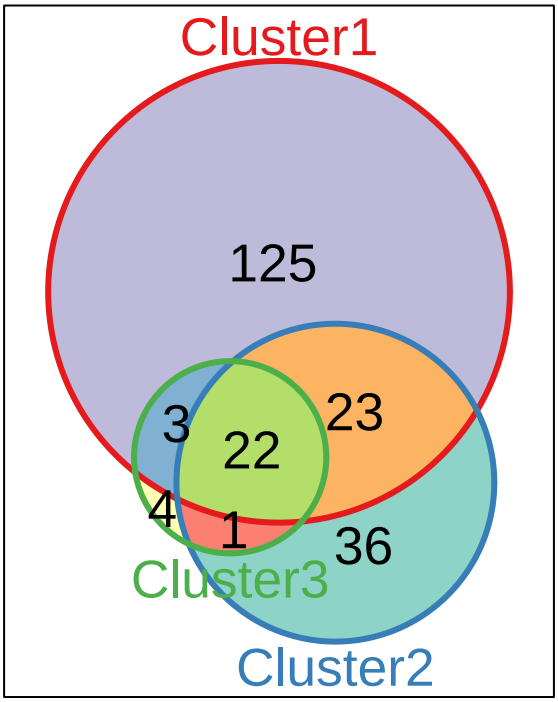


Figure 10.4: Overlap of genes induced in non-survivors compared to survivors in CAP sepsis between clusters. The genes show at least 2.0 fold change in mean expression and FDR < 0.01. 125, 36, and 4 genes are uniquely induced in non-survivors in clusters 1, 2, and 3 respectively. 22 genes (area shaded in green) are consistently induced in all clusters.

previous chapter. However, the complement system was enriched in FP sepsis. The relevance of this observation is uncertain.

Cluster invariant mortality

Linear Model

Survival signatures robust to cluster variance can be more accurately estimated by constructing linear models with survival as the primary outcome variable and the cluster membership as additional covariates. The larger sample size from the use of the full data set (all the CAP samples irrespective of cluster membership) should result in improved power and detection rates.

After controlling for repeated sampling and cluster membership, 883 probes corresponding to 772 unique genes were differentially regulated in non-survivors when compared to survivors. 524 probes were induced and 359 repressed. This corresponds

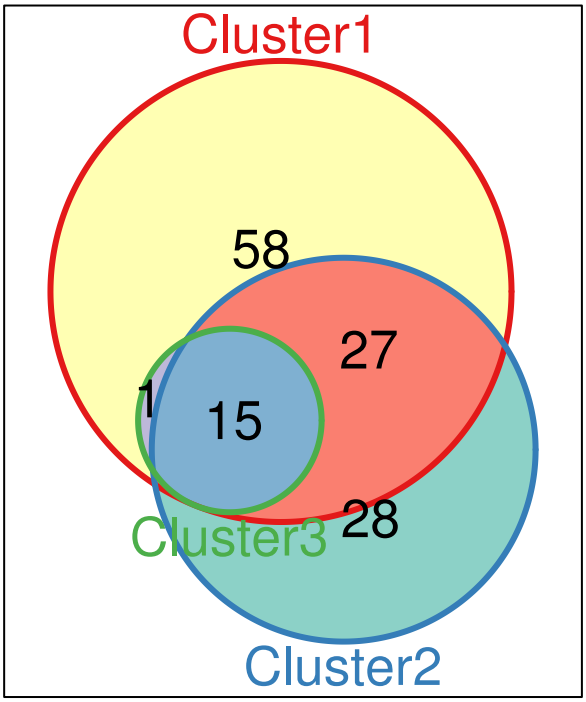


Figure 10.5: Overlap of genes repressed in non-survivors compared to survivors in CAP sepsis between clusters. The genes show at least 2.0 fold change in mean expression and FDR < 0.01. 58 and 28 genes are uniquely repressed in non-survivors in clusters 1 and 2 respectively. 22 genes (area shaded in blue) are consistently repressed in all clusters.

to 463 and 310 genes respectively. The genomic locations of these genes are shown in Figure 10.7. As with cluster specific differences, large effect sizes can be seen (Figure 10.8) with a pre-dominant gene silencing effect.

The 10 genes with the largest differences in expression are listed in Table 10.6. Each of these genes are associated with a 30 fold change in relative expression. All genes except *HLA-DRB5* are induced. There is over-representation of genes from chromosome 19. A number of these genes are neutrophil specific (*CTSG*, *MPO*, *DEFA4* among others) and are in particular related to neutrophil azurophilic granules. While some of the changes could be attributed to changes in relative cell counts biasing the measured transcript levels, a persistent bactericidal response appears to be associated with poor outcomes in CAP. A better approach to this problem of bias due to total and differential white cell counts is presented in section 10.4 on page 149.

Gene Set	Enrichment Score
Reactome classical antibody mediated complement activation	0.90
Reactome DNA strand elongation	0.67
Reactome gluconeogenesis	0.68
Reactome glycolysis	0.85
Reactome nucleotide metabolism	0.53
Reactome pyruvate metabolism	0.67
Reactome steroid metabolism	0.54
Carbohydrate catabolism	0.62
KEGG pentose phosphate pathway	0.69
KEGG glycolysis and gluconeogenesis	0.58
KEGG DNA replication	0.70
Chromatin binding	0.69
Adaptive immune response	-0.57
cAMP mediated signalling	-0.53
KEGG allograft rejection	-0.73
Double stranded RNA binding	-0.56
Biocarta calcineurin pathway	-0.60
Biocarta TH1 TH2 pathway	-0.55

Table 10.5: Gene sets enriched in the non-survivors of the third cluster in the CAP derivation cohort.

Gene	Chr	Function
CTSG	14	Cathepsin G
MPO	17	Myeloproxidase
ELANE	19	Neutrophil elastase
LOC653600	8	alpha 1 defensin
CEACAM6	19	Cell adhesion molecule
OLFM4	13	Olfactomedin 3
AZU1	19	Azurophil granule antibiotic protein
DEFA4	8	alpha defensin
PRTN3	19	Neutrophil membrane phagocytosis protein
HLA-DRB5	6	Class II MHC

Table 10.6: Genes associated with CAP survival with at least 30 fold change in relative expression. There is a strong preponderance of neutrophil specific effects, reflecting higher neutrophilic granule activity in non-survivors.

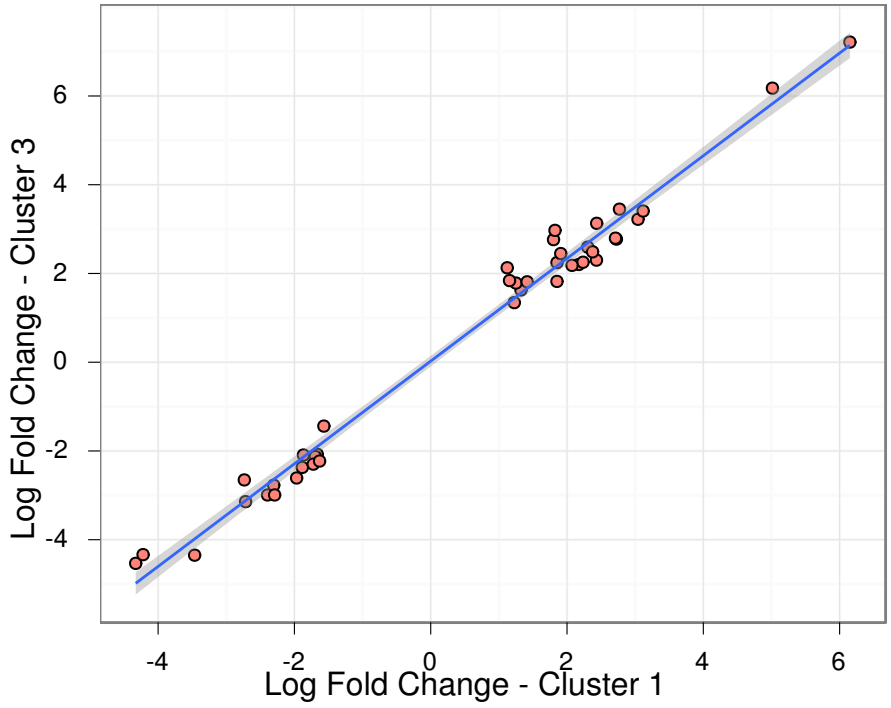


Figure 10.6: Correlation between the probes differentially expressed in non-survivors with at least 1.45 fold change in expression and FDR < 0.01 in both clusters 1 and 3. Pearson correlation coefficient = 99.2%

Pathway Analysis

Gene sets associated with cluster invariant survival response were identified using GSEA. Statistical significance was determined using permutation analysis involving phenotype switching. Gene sets with a permuted *p* value less than 0.05 were considered significant.

Gene sets significantly enriched in non-survivors in CAP that are robust to cluster differences are shown in Table 10.7. Death in CAP is associated with a repression of the cellular ribosomal assembly and function. There is induction of the TLR4 cascade and cytokine secretion, while there is repression of T cell activation. There is also repression of energy regulation processes of the outer mitochondrial membrane. Overall, death in CAP is associated with induction of components of the innate immune response and repression of the T cell components of the acquired immune system.

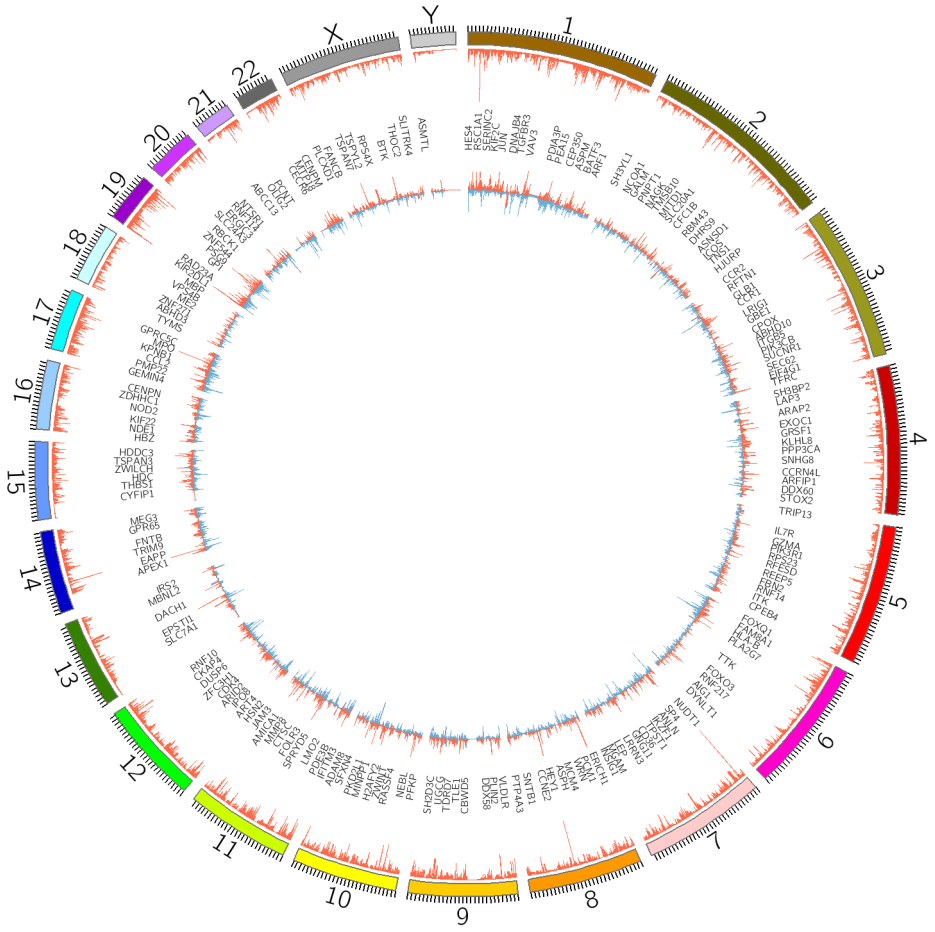


Figure 10.7: Circulogram showing genomic locations of genes signficantly associated with survival in CAP. From the outside the tracts are as follows: 1. stylised representation of the karyogram, 2. p values in -log10 scale for all genes in the genome, with peaks corresponding to greater statistical significance, 3. Gene names pruned to avoid illegibility due to text overlap, 4. fold change differences for all genes in the genome, up-regulated genes in red and down-regulated genes in blue. Clusters of significant genes are seen in chromosome 1, chromosome 19, and the short arm of chromosome 6.

10.3 Mortality in FP

Linear Models

The predictors of mortality in faecal peritonitis were identified using simple regression models. This analysis has necessarily reduced power due to the lower mortality rates in FP (10% vs 30% in hospital mortality rates). In contrast to CAP, only 52 probes corresponding to 51 unique genes were differentially regulated in non-survivors in FP when compared to survivors. The effect sizes are also mostly modest, even though the largest magnitudes are comparable (Figure 10.9). Notable genes that are differentially

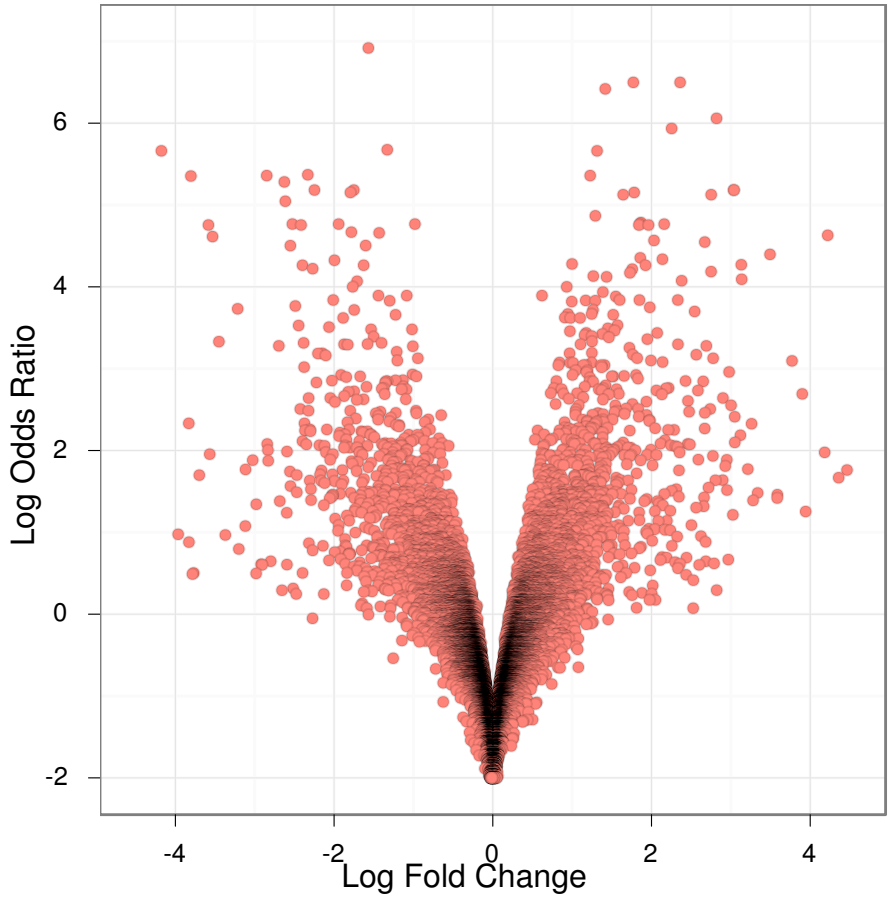


Figure 10.8: Volcano plot of cluster invariant differences in mean expression in non-survivors in the CAP cohort

expressed include members of the class II MHC and TLR7. Also, in contrast to CAP, the pre-dominant pattern is repression, although this may be an artefact of low gene yield. However, the pattern of gene expression in gene groups like the MHC is preserved across both cohorts.

The overlap of the predictors of mortality in CAP and FP are of interest. Overlapping genes provide evidence for a common mechanism of mortality in severe sepsis and could possibly indicate a single or a small set of failures in cellular processes. Over the entire probe set the the correlation between the probes in each cohort for the survival comparison is very poor (Figure 10.10). When the analysis is restricted to the probes that achieve statistical significance in both groups, there is better concordance with a correlation coefficient of 93% (Figure 10.11). However, only 7 probes are shared

Gene Set	Enrichment Score
Reactome TLR4 cascade	0.62
Reactome FRS2 mediated cascade	0.67
Regulation of cytokine secretion	0.68
Metal ion transport	0.40
Cytokine secretion	0.71
KEGG ether lipid metabolism	0.50
Biocarta VEGF pathway	0.58
Reactome ribosome formation	-0.80
Reactome gene expression	-0.55
Reactome influenza viral RNA transcription	-0.76
Reactome insulin synthesis and secretion	-0.70
Reactome peptide chain elongation	-0.82
Covalent chromatin modification	-0.67
T cell activation	-0.48
KEGG ribosome	-0.80
KEGG primary immunodeficiency	-0.64
Biocarta IL17 pathway	-0.75
Biocarta IL7 pathway	-0.63
Biocarta NKT pathway	-0.59
Mitochondrial outer membrane	-0.59
snRNP complex	-0.69

Table 10.7: Gene sets enriched in the CAP non-survivors

between both data sets. These correspond to the genes *GZMH* (cytotoxic T-lymphocyte proteinase), *HLA-DRA*, *KCTD12* (potassium channel tetramerisation domain containing protein), *LOC644936* (actin β pseudogene), *HLA-DRB1*, *HLA-DRB5*, and *TLR7*. While a definitive conclusion is difficult with such small numbers, the strong biological role of class II MHC in the immune response along with a high prior likelihood that the MHC is involved in immune dysfunction, suggests that repression of the MHC is a key component of mortality in sepsis that is robust to the source of infection.

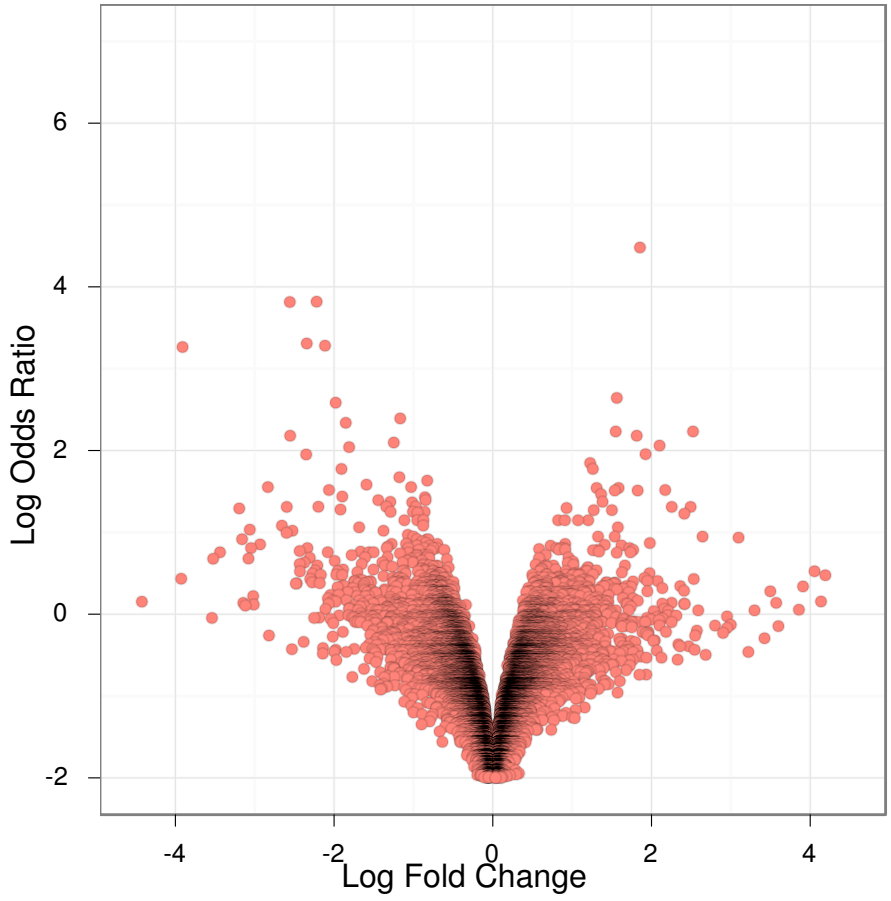


Figure 10.9: Volcano plot of change in mean expression values of all probes in non-survivors compared to survivors in the FP cohort. The volcano is shorter compared to the corresponding volcano for the CAP cohort.

Pathway Analysis

GSEA was used to identify the biological pathways associated with poor outcomes in faecal peritonitis. Permutation procedure involving phenotype switching was used to determine statistical significance. Significantly enriched pathways (defined as a permuted p value less than 0.05) are listed in Table 10.8. Non-survival is associated with the induction of steroid metabolism and decreased activation of complement. Glucose metabolism is dysregulated along with changes in chromatin structure. There is repression of chemokine and cytokine activity, antigen processing and presentation and T cell activity. Overall, in FP, death is associated with a suppressed immune response. While some

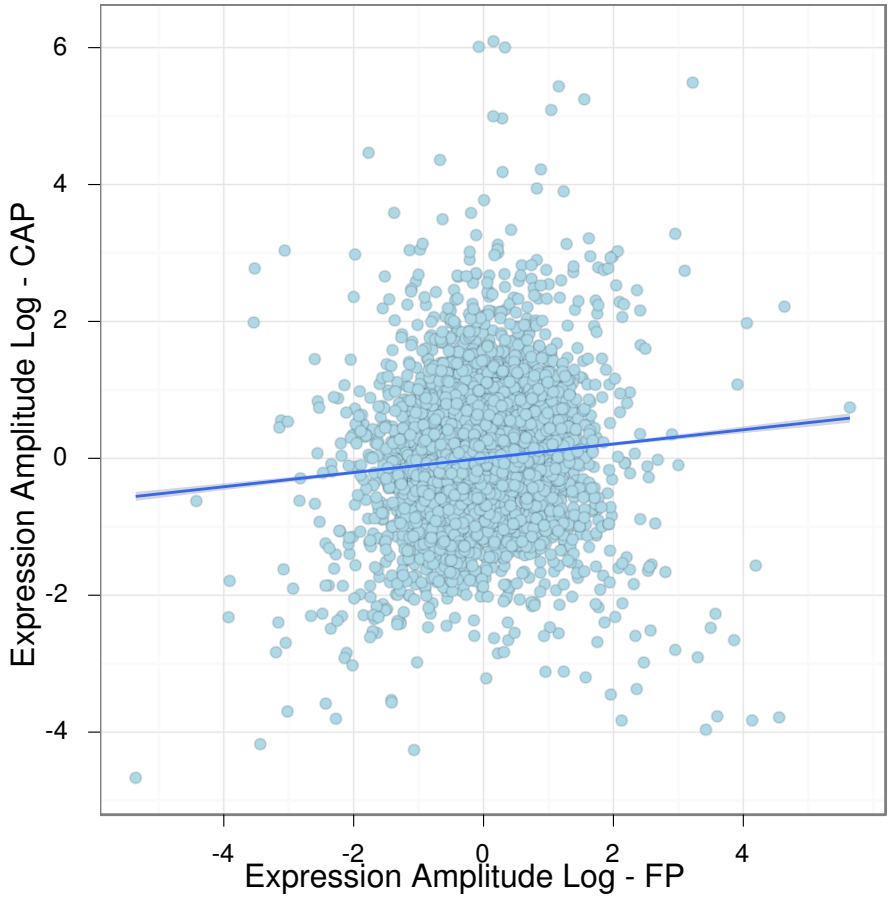


Figure 10.10: Correlation between mortality predictors in CAP and FP. The correlation coefficient is only 9.8%.

mechanisms like chromatin modification and cytokine activity are similar to CAP, the molecular mechanisms of death are different in CAP and FP.

10.4 Mechanisms of mortality

The linear models presented above do not correct for any covariates. The results identified in earlier sections suggest that differential cell counts might be a significant covariate and the mRNA transcript levels between survival cohorts may be biased by relative neutrophil counts. To identify pathophysiological mechanisms the differences due to the covariates need to be accounted for before making any inferences for survival specific effects.

Gene Set	Enrichment Score
Reactome cholesterol biosynthesis	0.72
Reactome DARPP32 events	0.62
Reactome glucose metabolism	0.57
Reactome glycolysis	0.70
Reactome steroid metabolism	0.55
Myeloid leucocyte differentiation	0.69
Chromatin modification	0.48
RAS GTPase binding	0.55
Biocarta NO1 pathway	0.60
Biocarta ECM pathway	0.57
Biocarta CK1 pathway	0.73
Biocarta MAL pathway	0.62
Biocarta WNT pathway	0.57
Reactome complement cascade	-0.80
Reactome downstream TCR signalling	-0.64
Reactome initial triggering of complement	-0.88
Reactome PD1 signalling	-0.82
Positive regulation of immune response	-0.53
KEGG antigen processing and presentation	-0.54
KEGG type 1 diabetes mellitus	-0.70
KEGG asthma	-0.86
KEGG allograft rejection	-0.75
KEGG graft vs host disease	-0.73
KEGG viral myocarditis	-0.62
Chemokine activity	-0.53
Biocarta COMP pathway	-0.81
Biocarta TH1TH2 pathway	-0.72

Table 10.8: Gene sets enriched in non-survivors in the FP cohort.

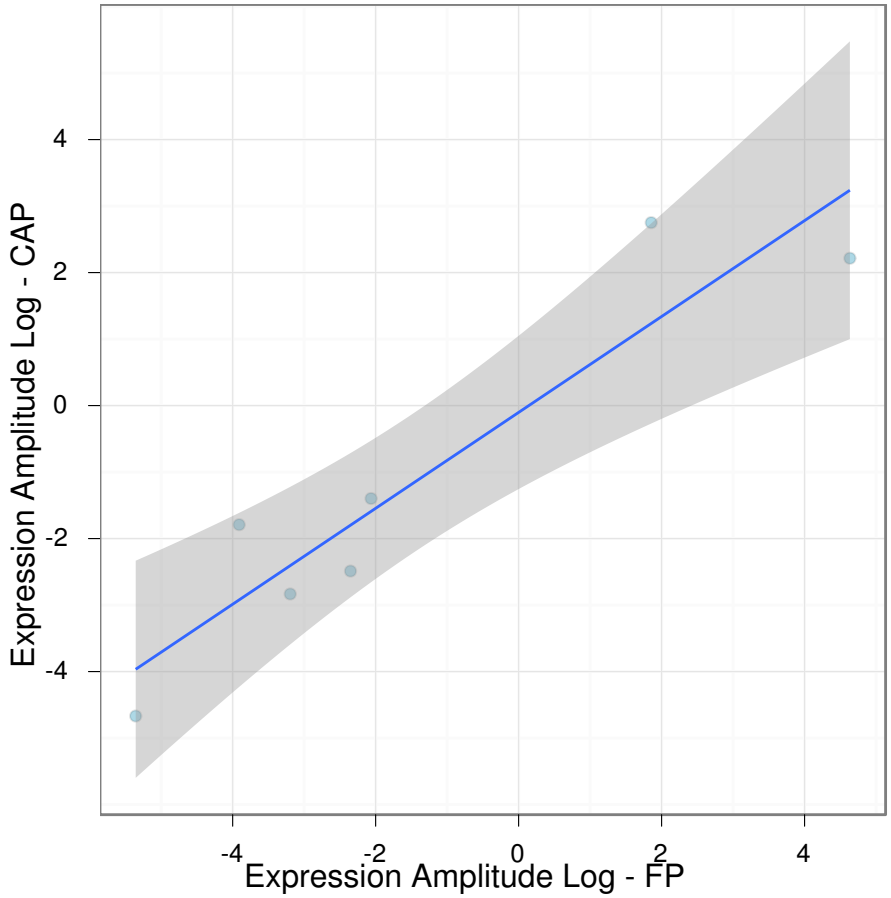


Figure 10.11: Correlation between mortality predictors in CAP and FP that achieve statistical significance in both cohorts (1.5 fold change in mean expression and FDR < 0.05). The correlation coefficient is 93% but only 7 probes are shared between the cohorts.

A Bayesian hierarchical model was constructed to model probe expression with a number of covariates including survival, organ function markers and demographic variables. Survival was modelled as a hierarchical variable, with the mean of each cluster determined independently by MCMC simulation. The analysis was performed on the entire CAP derivation data set. Probes with a minimum 1.5 fold change in mean expression with a Bayesian credible interval not including zero were considered significant.

After correcting for relevant covariates 6 genes are differentially expressed with at least a 2 fold change in mean expression. Of these one *PRKAR1A* is repressed, while the others *LOC100133662*, *LOC644936*, *RPS4Y1*, *EIF1AY*, *HLA-DRB5* are induced. Intriguingly

LOC100133662 is a discontinued PubMed record for an unknown protein and *LOC644936* codes for a pseudogene of actin.

With a less stringent threshold of at least 1.5 fold change in mean expression between non-survivors and survivors, 24 genes are differentially expressed (Table 10.9). Interesting findings include the induction of two Y chromosomal genes *RPS4Y1* and *EIF1AY* despite correction for sex related effects, suggesting independent effects of sex on mortality. There are also directional changes in class II MHC, where *HLA-DRB5* is induced while *HLA-DRB4* is repressed, suggesting that mortality is associated with nuanced and complex changes in MHC activity. For example, this observation could be the result of over-representation of certain *HLA-DRB4* and *HLA-DRB5* haplotypes in survivors and non-survivors observed in other clinical settings (Huo *et al.*, 2011) In conjunction with the identification of differential expression of class II MHC molecules as the basis for cluster differentiation in section 8.3 on page 87, these observations indicate that structural variants influence the response to sepsis and outcome, at least in CAP. The strong role of ribosomal proteins in mortality is reinforced as is the consistent induction of ATP-cassette binding proteins from different genomic regions. The interesting association with viral activity reported earlier is confirmed.

10.5 Morbidity Outcomes

The difficulties associated with the use of mortality as an outcome measure in sepsis such as phenotypic heterogeneity and multiple confounding factors have been described earlier. These problems can be circumvented by using alternative outcome measures that are less susceptible to these problems. In particular morbidity indices such as number of days of organ support, and duration of ICU stay are useful crude, but clinically relevant measures of illness severity.

To overcome the issue of parameter cut-off where the the morbidity measure is circumscribed in non-survivors the analysis was restricted to survivors. 113 samples from the CAP derivation cohort were analysed.

Gene	Chr	Product	Coefficient	95% CI
LOC100133662		withdrawn record	1.15	0.53 – 1.78
LOC644113		withdrawn record	0.62	0.34 – 0.90
LOC644936	5	beta actin pseudogene	1.33	0.54 – 2.14
MAL	2	T cell differentiation protein	0.65	0.07 – 1.20
RNU1-3	1	small nuclear RNA	0.91	0.02 – 1.78
RNU1-5	1	small nuclear RNA	0.88	0.02 – 1.72
RNU1A3	1	small nuclear RNA	0.67	0.04 – 1.30
RNU1G2	1	small nuclear RNA	0.95	0.07 – 1.86
RPS4Y1	Y	ribosomal protein S4	1.52	0.71 – 2.32
TGM3	20	formation of epidermis	0.60	0.14 – 1.08
ABCA1	9	ATP binding cassette transporter	0.61	0.19 – 1.02
ABCG1	21	ATP binding cassette transporter	0.64	0.13 – 1.15
CPA3	3	mast cell carboxypeptidase	0.76	0.40 – 1.11
EIF1AY	Y	eukaryotic translation initiation factor	1.26	0.47 – 2.08
FCER1A	1	Fc fragment of IgE	0.84	0.10 – 1.61
HDC	15	histidine decarboxylase	0.87	0.27 – 1.48
HLA-DRB5	6	class II MHC	1.38	0.24 – 2.52
LOC649143		class II MHC pseudogene	-0.68	-1.35 – -0.02
LOC728093	5	POM121 like protein	-0.74	-1.24 – -0.25
OAS1	12	innate immune response to viral infection	-0.75	-1.52 – -0.01
PRKAR1A	17	cAMP dependent protein kinase	-1.09	-1.69 – -0.47
CD163	12	macrophage scavenger receptor	-0.79	-1.44 – -0.11
CLEC4D	12	C type lectin superfamily	-0.67	-1.18 – -0.19
HLA-DRB4	6	class II MHC	-0.99	-1.80 – -0.20

Table 10.9: Genes significantly associated with survival in CAP in the Bayesian analysis. Genes were considered significant if the mean change in expression was at least 1.5 fold and the 95% credible interval does not include zero. Genes in the lower panel are repressed.

The distribution of the outcome variables is shown in Figure 10.12. There are three outlier values from one patient who required 96 days of respiratory support. This individual was removed from further analysis.

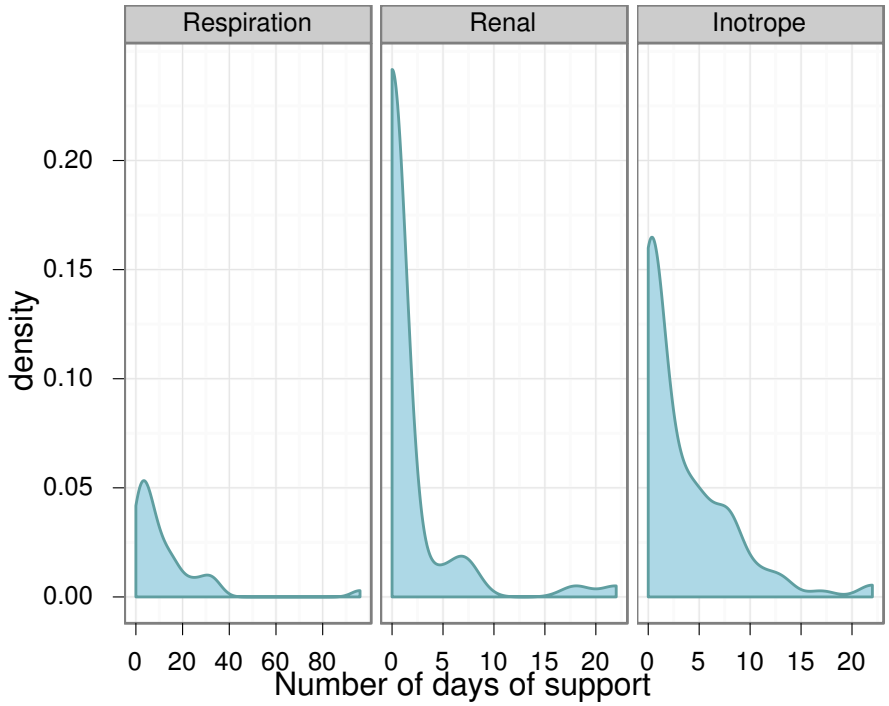


Figure 10.12: Probability density distribution of morbidity outcomes. Density plots are histograms normalized to contain an area under the curve of one.

Only 7 survivors required inotropic support, which is inadequate to fit a linear model. Similarly only one survivor received renal support. Therefore, renal support was also not considered further because of lack of data.

Respiratory Support

Linear regression analysis was performed with the number of days of respiratory support as a continuous variable. Probes with a regression coefficient greater than 1.5 corresponding to fold change in expression in log2 scale and a FDR less than 0.01 were considered significant. In contrast to the previous analysis where the coefficients indicate the difference in mean expression between the groups, the coefficients in this analysis have

a standard regression inference. Each coefficient is the expected change in the number of days of respiratory support for a unit change in expression value of the probe.

A total of 1287 probes corresponding to 1151 genes were associated with the requirements for respiratory support. The top ten hits are shown in Table 10.10 and a representative distribution shown in Figure 10.13. The overlap between the top associations and the genes identified from the morbidity analysis show a good concordance, suggesting common mechanisms for each outcome. Some interesting findings include the presence of three probes for *DEFA1B* in the top 10 hits indicating good concordance. *ARG1* codes for arginase 1, which has two isoforms. One of the isoforms has immune functions, but the induced gene is reported to be specific to liver and involved in the urea cycle (Morris, 2002) *IL1R2*. binds to interleukins 1α and 1β and is inhibitory. Overall, increased neutrophil activity, reduced apoptosis indicative of an active inflammatory state, and dysregulation of glucose metabolism are associated with increased ventilatory requirements.

Gene	Chr	Function	logFC
ARG1	6	Liver arginase	0.60
IL1R2	2	Interleukin receptor 1 Type 2	0.67
FOLR3	11	Folate receptor 3	0.64
DEFA1B	8	Defensin alpha 1b	0.73
C19ORF59	19	Mast cell membrane protein	0.67
S100P	4	Calcium binding protein	0.68
DEFA3	8	Defensin alpha 3	0.72
MMP9	20	Matrix metallopeptidase	0.69
OLFM4	13	Antiapoptotic factor	0.59
PFKFB3	10	phosphofructo-2-kinase	0.63

Table 10.10: Genes significantly associated with increased duration of respiratory support in severe sepsis due to CAP.

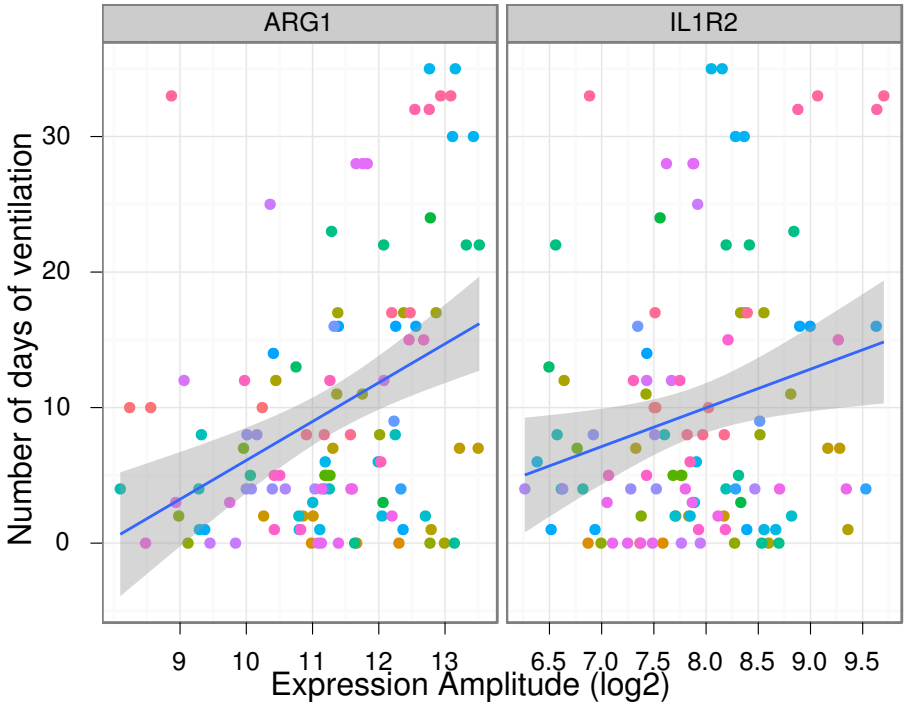


Figure 10.13: Representative genes associated with respiratory support. Colours represent samples from an individual. The is strong correlation in the expression values from the samples of any particular individual. Despite a reasonable model fit (indicated by the blue line and the shaded confidence intervals), predictive power is poor. The same expression level is seen in a patient with no ventilatory support and the patient with more than 30 days of ventilatory support.

10.6 Validation

The gene expression signatures of survival in CAP was validated in the validation cohort. Since only specific hypotheses were tested, analysis was restricted to genes identified as significant in the derivation cohort and also present in the validation cohort. There were only two deaths in the FP validation cohort. Therefore, validation analysis was restricted to the CAP cohort.

Validation of CAP Survival

Of the 772 genes that were significantly differentially expressed in the derivation cohort and robust to cluster memberships, 691 were represented in the validation cohort. After fitting a linear model, 76 genes (10.9%) had at least 1.5 fold change in mean expression and achieved statistical significance. With a relaxed threshold of 1.2 fold change in mean

expression, an additional 12 genes (12.7% in total) were identified with a *p* value less than 0.05. The correlation between the expression of these genes in both the data sets is only 43% but this is statistically significant ($p < 0.001$). Of the 88 genes, 29 are directionally discordant (Figure 10.14).

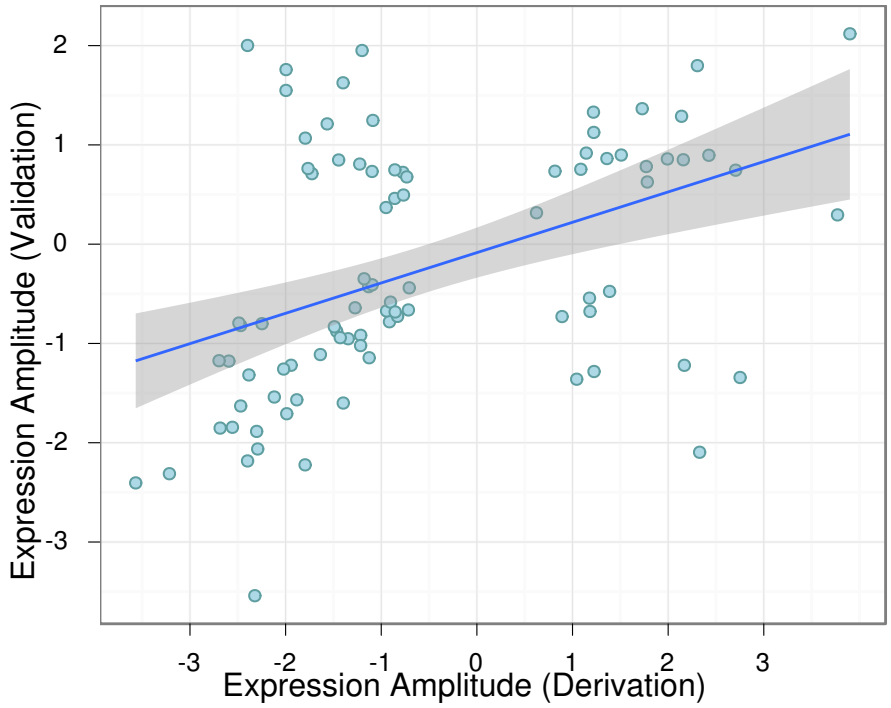


Figure 10.14: Validation of CAP Survival. Correlation between differentially expressed in non-survivors in both the derivation and validation data sets for the CAP cohort.

In CAP patients, the correlation between the gene expression patterns associated with outcome in the derivation and validation cohort is modest. In contrast to the similar validation comparison for the genomic response to sepsis (section 9.4 on page 126), there is no systematic under-estimation of effects in the validation cohort. These observations confirm that mortality in sepsis as a noisy end-point, and indicate that the likelihood of developing ad adequate predictive model based on gene expression data in low.

10.7 Summary

The analysis presented in this chapter shows that there are important differences between the gene expression signatures associated with outcomes in CAP and FP.

In addition, even within the clinically homogeneous label of community-acquired pneumonia, distinct clusters defined by gene expression profiling are associated with phenotype differences including predicted and observed mortality rates. Nevertheless, dysregulation of glucose metabolism, alterations in ribosomal activity and the involvement of class II MHC genes, albeit in different patterns, are consistent themes across all comparisons.

Overall mortality in FP is associated with an increase in steroid metabolism and immune suppression, although gene yield is low with only modest effect sizes. Since these changes are concomitant a causal relationship between steroid synthesis and suppression of immune function is biologically plausible. Non-survival in FP is also associated with a failure to induce the complement cascade. Since the complement cascade is a unique component of the sepsis response in faecal peritonitis, this change is of considerable significance.

In contrast to FP, effect sizes are large in CAP, but the genes with biggest effect sizes are related to induction of neutrophil azurophilic granule activity which could reflect either a persistent inflammatory response in CAP or transcript bias due to differences in relative neutrophil counts. While the former contrasts with the FP response, the Bayesian analysis lends some credence to the latter hypothesis. It is striking that while correlations between the overall expression signatures in CAP and FP is poor, there is a strong correlation between the most differentially expressed genes, particularly in relation to class II MHC activity.

The gene expression signature associated with mortality in CAP shares some components with the mortality associated signatures in FP, particularly chromatin modification and repression of the T cell response. However, CAP deaths are uniquely associated with suppression of ribosomal assembly, alterations in initiation of transcription and changes in translational machinery. There are also unique association to the mitochondrial outer membrane oxidative machinery but no associations to the complement cascade.

There are also cluster specific differences in gene expression signatures of mortality and the effect sizes are large in each cluster. Cluster 1 and 3 seem to have similar

expression signatures associated with mortality even though these clusters are widely separated on the first principal component. Cluster 3 shows changes in complement cascade similar to the FP cohort. All clusters show dysregulation of glucose metabolism in non-survivors and suppression of adaptive immune responses and T cell activity. Cluster 1 shows unique associations with DNA repair mechanisms, mitotic activity and telomere related activity. There is also induction of hypoxia activity in non-survivors which may be clinically relevant.

The pattern of differential gene expression associated with mortality is also related to the number of days of ventilatory support. In particular, an active inflammatory response is associated with an increased number of days of respiratory support.

11

Genomic Effects of Covariates

Sepsis is a syndrome with multi layered complexity involving multiple, interacting pathophysiological and clinical events. The relationship of some of the covariates to survival and overall gene expression was examined in chapter 7.1 on page 76.

For clinicians, covariates are usually seen as important only so far as they relate to tangible clinical outcomes, especially when such relationships can be expressed as simple associations. The main use of covariates in sepsis is in clinical risk prediction scores, where the outcome risk is estimated from a logistical regression derived formula using predictive covariates (Knaus *et al.,* 1985). The predominant application of these scores is to identify cohorts with identical risk where the effect of a therapeutic intervention can be studied. While clinically useful this approach has low explanatory power and is a poor reflection of underlying pathological processes. For example severity of illness scores assume that the biological mechanisms causing morbidity and death are homogeneous, an assumption not substantiated by gene expression data or clinical experience.

One of the problems in the handling of covariates is identifying the relevant covariates and the limited statistical power when the number of covariates overwhelms sample size. To identify the consequences of the covariates, especially their independent effects on gene expression data, a Bayesian regression analysis with hierarchical mixed models was used (section 4.3). This approach allows for the estimation of independent effects using MCMC simulation and can be used when the number of predictors is high relative to the number of observations, with the downside of approximate estimations and computational complexity.

The analysis presented in this chapter is restricted to the CAP derivation cohort because of the computational burden. The effects of each covariate on gene expression and on survival related genes are examined.

11.1 Demographic covariates

Age

Increasing age is known to adversely influence outcome in sepsis(Bucala *et al.,* 1984; Ershler *et al.,* 2000; Jensen *et al.,* 2011) and is also known to influence gene expression (De Magalhães *et al.,* 2009). Both the APACHE II and SAPS II scores include age as a risk factor for mortality with a non-linear relationship between increasing age and predicted risk.

The coefficients produced by the Bayesian model can be interpreted in a similar fashion as conventional regression models. The coefficients are the expected change in the response variable for a unit change in the predictor variable when all other factors are held constant. For age, the coefficients are the expected change in gene expression for a unit change in age or an change in age of 1 year. This difference is too low to be clinically meaningful and it is more useful to use larger differences such as seen between decades.

No genes show a significant change in expression with age for a mean change in expression of 1.5 fold. Only four genes have statistically significant changes in gene expression over a decade change in age with a relaxed threshold of 1.2 fold change in mean expression. One gene (*CFD*, complement factor D) is induced, while the other three (*VAV3* - guanine nucleotide exchange factor, *ZDHHC19* - zinc finger, *IFI27* - mitochondrial interferon inducible protein) are repressed. All genes (except *ZDHHC19*) have immune effects.Over 20 years these four genes show at least a 1.5 fold change in mean expression. None of these genes are associated with survival in the Bayesian analysis.

These minimal changes in gene expression associated with age are surprising given the known additional mortality risk associated with increasing age. This observation

may be due to the age related mortality being closely linked to changes in organ function, which are independently modelled in the Bayesian analysis. The low yield might also be due to small effect sizes which are not captured in the age range seen in the CAP derivation cohort.

Sex

As is the case with age, sex differences influence survival in sepsis and are known to affect gene expression (Bhasin *et al.,* 2008) In particular, survival rates are lower in males and post-menopausal females.(Adrie *et al.,* 2007; Eachempati *et al.,* 1999). Interestingly, while sex is included as a predictor of mortality in the APACHE II score, it is not used as a predictor variable in the SAPS II score.

The coefficient for sex in the Bayesian analysis is the change in gene expression in males compared to females, when all other covariates are unchanged. With a 1.5 fold change in mean expression, 15 genes are repressed and 16 genes are induced. With a relaxed threshold of 1.2 fold change in expression the predominant pattern of change is repression (Table 11.1).

The genes are listed in Table 11.2. As expected, sex related modulation of gene expression is not restricted to genes in the sex chromosomes. Also as expected a number of Y chromosome related genes are correctly identified in males, even though the relevance of some of the pseudogenes is unclear. Interestingly, there is down regulation of *HDC* (histidine decarboxylase) and *FCER1A* (Fc fragment of IgE) in males (both involved in the allergic response), while there is induction of prostaglandin metabolism (*HPGD*) in females. The genes *UTY* and *JARID1D* encode minor histocompatibility antigens, compatible with modulation of immune function by sex. There is also sex specific expression of transcription factors *EIF1AY* in males and *NFXL1*, *KLF9* in females, along with histone function modulation (*SAP30*, *ZBTB16*) in females. *PGM5* is induced in females and is involved in the synthesis of galactose, suggesting sex related modulation of glucose metabolism. At least some parts of mitochondrial function (*ECHDC3*) are modulated by sex. Interestingly, mitochondrial function is known to be influenced by

both age and sex with differing susceptibility to oxidative stress (Guevara *et al.,* 2011, 2009).

Fold change	Up	Down	Total
2.0	7	1	8
1.5	16	15	31
1.2	79	200	279

Table 11.1: Gene expression changes associated with sex in the Bayesian regression model for different fold change thresholds. Genes were considered significant only if the 95% credible interval did not include zero.

Of the 16 genes induced in males with at least 1.5 fold change in gene expression, 5 genes (*LOC100133662*, *RPS4Y1*, *EIF1AY*, *FCER1A*, *HDC*) are also induced in non-survivors when compared to survivors in the Bayesian analysis with a 1.5 fold change in expression after correcting for other covariates including cluster memberships (Figure 11.1). This constitutes about 30% of the significant genes in the sex and survival comparisons. The expression of these genes could be a mechanistic explanation of adverse outcomes in males with sepsis.

11.2 Organ function markers

Renal Failure

Renal dysfunction is a common feature of severe sepsis and is usually due to a combination of pre-renal (hypoperfusion) and renal factors. Renal failure is also an independent predictor of mortality in severe sepsis and biomarkers of renal function (creatinine levels) are used as predictive factors in both the APACHE II and SAPS II risk prediction scores. Changes in the physiological milieu due to renal dysfunction can cause changes in gene expression.

The impact of renal function on gene expression was evaluated by including the presence of renal failure as a covariate in the Bayesian regression model. With a 1.5 fold change in mean expression as a threshold, 76 genes achieve statistical significance. The

Gene	Chr	Function	Coefficient	95% CI
LOC100133662		discontinued record	2.93	2.52 – 3.33
METTL7B	12	methyltransferase like protein	0.60	0.21 – 0.96
MFF	2	Mitochondrial fission factor	0.68	0.27 – 1.09
PRKY	Y	X linked protein kinase pseudogene	0.74	0.58 – 0.90
RPS4Y1	Y	component of 40S ribosomal subunit	3.95	3.42 – 4.48
RPS4Y2	Y	component of 40S ribosomal subunit	1.23	1.03 – 1.43
TACSTD2	1	cell surface calcium signalling receptor	0.61	0.03 – 1.17
UTY	Y	minor MHC antigen	1.17	0.96 – 1.37
CYORF15A	Y	TXLNG2P taxilin gamma pseudogene	1.66	1.41 – 1.90
CYORF15B	Y	TXLNG2P taxilin gamma pseudogene	0.62	0.50 – 0.76
EIF1AY	Y	eukaryotic translation initiation factor	3.56	2.99 – 4.11
FCER1A	1	IgE receptor	0.71	0.20 – 1.22
HDC	15	Histidine decarboxylase	0.71	0.28 – 1.12
IFI27	14	Interferon alpha inducible protein	0.95	0.15 – 1.78
JARID1D	Y	KDM5D minor MHC antigen	1.58	1.36 – 1.80
MAOA	X	Monoamine oxidase A	-0.89	-1.53 – -0.28
NFXL1	4	Nuclear transcription factor	-0.71	-1.08 – -0.36
OLAH	10	oleoyl-ACP hydrolase	-0.75	-1.25 – -0.27
PGM5	9	phosphoglucomutase	-0.75	-1.21 – -0.26
SAP30	4	component of histone deacetylase	-0.66	-0.98 – -0.37
TPST1	7	tyrosyl protein sulphotransferase	-0.82	-1.20 – -0.42
XIST	X	X inactivation centre transcript	-1.34	-1.55 – -1.14
ZBTB16	11	interacts with histone deacetylase	-0.73	-1.07 – -0.40
ACPL2	3	acid phosphatase like	-0.68	-0.98 – -0.38
ADORA3	1	cell proliferation and death signalling	-0.79	-1.19 – -0.39
DAAM2	6	activator of morphogenesis	-0.64	-1.02 – -0.24
ECHDC3	10	mitochondrial enoyl CoA hydratase	-0.74	-1.17 – -0.30
GBP5	1	guanylate binding protein	-0.61	-1.09 – -0.14
HPGD	4	prostaglandin metabolism	-0.66	-1.17 – -0.13
KLF9	9	transcription factor	-0.60	-0.92 – -0.28

Table 11.2: Genes associated with sex specific expression in Bayesian analysis. Genes in the top panel are induced in males and in the bottom panel induced in females. All genes show at least 1.5 fold change in mean expression. The coefficient is in log2 scale.

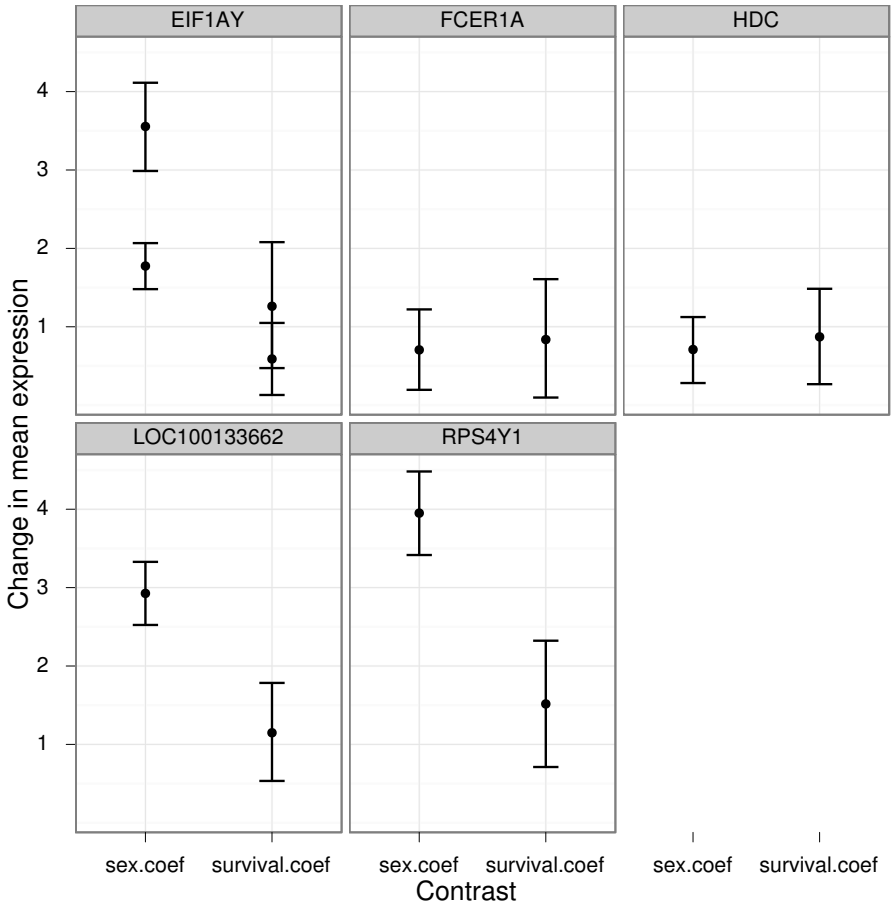


Figure 11.1: Genes associated with sex and survival in the Bayesian analysis. The plot shows the changes in mean expression in males compared to females and the 95% credible intervals. Multiple bars in the plot are due to multiple probes for the same gene. sex.coef = coefficient for sex in the Bayesian analysis, survival.coef = coefficient for survival in the Bayesian analysis.

predominant pattern is that of repression with renal failure, which is seen consistently across different thresholds of change in mean expression (Table 11.3).

Fold change	Up	Down	Total
2.0	1	9	10
1.5	24	52	76
1.2	286	388	674

Table 11.3: Number of genes associated with renal failure in Bayesian analysis. Genes were considered significant only if the 95% credible interval did not include zero.

The genes that show at least 2 fold change in mean expression with renal failure are shown in Table 11.4. The results show a surprising but consistent repression of genes

encoding immunoglobulin IgG chains. Interestingly a number of these records have been discontinued in the PubMed indexing system. There is also repression of *KIAA1324* (aliased as *EIG121*), an oestrogen induced gene which has been used as a biomarker for hyper-oestrogenic states (Deng *et al.,* 2005). This could indicate modulation of renal failure related changes by sex, but this conclusion is speculative given the unknown function of the gene.

With a normal threshold of 1.5 fold change in mean expression, additional features are revealed. Interesting findings include induction of *VWA5A* (von Willebrand factor domain A) and repression of *F13A1* (coagulation factor XIII), *HSPB1* (heat shock protein 1), *LYZ* (lysozyme), immunoglobulin associated molecules *CD79A* and *CD79B*, and polypeptide chains of cytochrome family 4F *CYP4F3* and *CYP4F12*. The changes link renal failure to a number of other biological systems such as immune response, coagulation and metabolism.

Gene	Chr	Function	Coefficient	95% CI
C4BPA	1	Complement component binding protein	1.32	0.33 – 2.32
LOC642113		IgG κ chain (discontinued record)	-1.31	-2.33 – -0.30
LOC647450		IgG κ chain (discontinued record)	-1.10	-2.13 – -0.06
LOC647506		IgG κ chain (discontinued record)	-1.15	-2.27 – -0.07
LOC649923		IgG chain C region (discontinued record)	-1.20	-2.19 – -0.23
LOC652493		IgG κ chain V-I region (discontinued record)	-1.09	-2.17 – -0.01
EPHB4	7	Ephrin receptor (tyrosine kinase family)	-1.08	-1.49 – -0.65
IGJ	4	Immunoglobulin J polypeptide	-1.33	-2.47 – -0.16
IGLL1	22	Immunoglobulin λ like polypeptide	-1.39	-2.28 – -0.45
KIAA1324	1	Oestrogen induced gene of unknown function	-1.17	-1.77 – -0.55

Table 11.4: Genes associated with renal failure in Bayesian analysis. Genes in the top panel are induced in renal failure and in the bottom panel repressed. Coefficients are in log2 scale.

Since renal failure is an independent risk factor for mortality in sepsis, the relationship between the genes associated with death and the genes associated with renal failure is interesting. With a 1.2 fold change in mean expression as threshold for both the survival and renal failure comparisons, only two genes have similar expression changes in both groups. *PLIN2* is induced in both groups and is a marker of lipid accumulation

in a number of disease processes. *XRN2* is repressed in both groups and is involved in homologous recombination and RNA metabolism. The significance of these changes is uncertain. Interestingly, a number of genes show discordant changes in both groups. With a 1.2 fold change in mean expression as threshold, 22 genes with induced expression in death (16% of the total hits) have repressed expressions in renal failure (8.0%), while 23 genes show the opposite effect (23% and 5.6% of the total hits in survival and renal comparisons respectively). The six genes that show discordant effects and have at least 1.5 fold change in gene expression in both comparisons are shown in Figure 11.2.

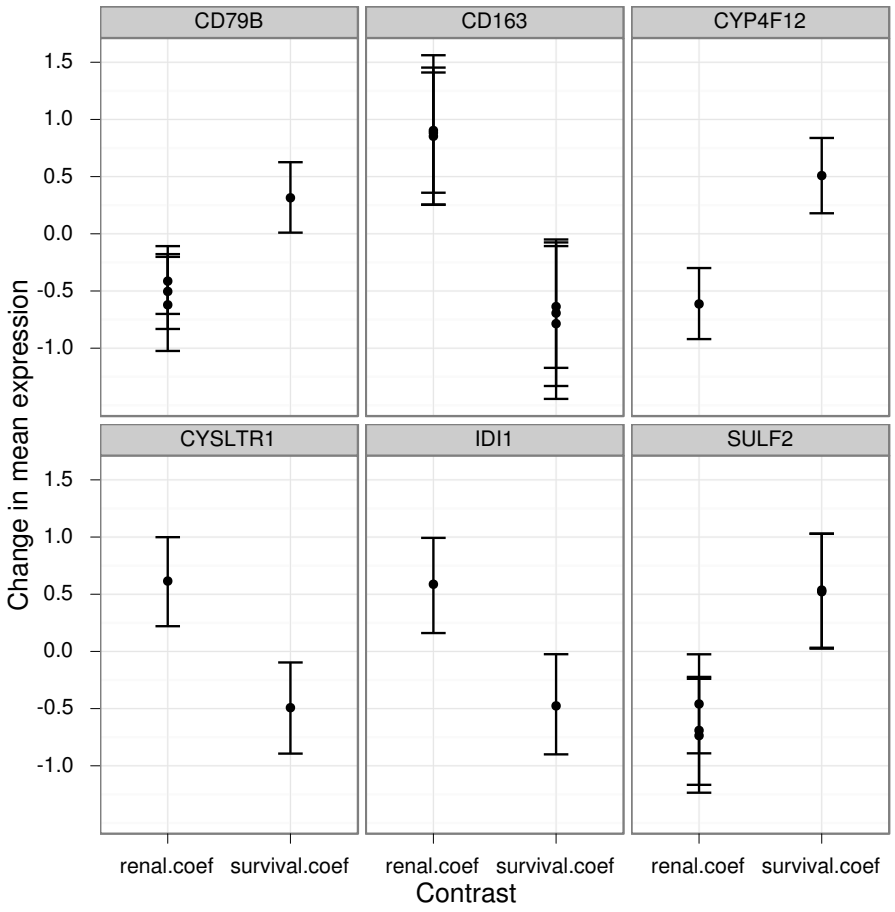


Figure 11.2: Genes significantly associated with renal failure and survival in the Bayesian analysis but with discordant effects. *CD163* haemoglobin scavenger receptor, *CYSLTR1* leukotriene D4 receptor, *IDI1* IPP isomerase, *SULF2* sulfatase-2, *CD79B* immunoglobulin associated molecule, *CYP4F12* cytochrome P450 family 4. renal.coef = regression coefficient for renal failure, survival.coef = regression coefficient for non-survival

Oxygenation Index

Respiratory failure is a common complication of severe sepsis. In the absence of pre-existing lung disease, respiratory failure in sepsis is characterized by an increased alveolar-arterial oxygen gradient, manifest clinically as decreased partial pressures of oxygen in arterial blood and increasing oxygen requirements to maintain an adequate arterial oxygen partial pressure. In ventilated patients, the latter is also influenced by the positive end expiratory pressure (PEEP) applied to the lungs. The PF ratio is defined as the ratio of the arterial partial pressure of oxygen to the fractional inspired concentration of oxygen and is a measure of lung injury. The index ignores the effects of PEEP which reduces the applicability of the index, but is still commonly used in clinical practice to guide ventilatory management.

As with age, clinically relevant ranges for PF ratio were used to determine significant differences in expression. A difference in the PF ratio of 20 (with arterial partial pressure measured in KPa) corresponds to the diagnostic criteria used to differentiate acute lung injury from acute respiratory distress syndrome. With a 1.5 fold mean expression difference threshold, 21 genes were significantly differentially expressed (Table 11.5). There is a consistent pattern of repression with increasing PF ratio irrespective of the chosen threshold.

Oxygenation index	**10**			**20**			**25**		
Fold Change	**Up**	**Down**	**Total**	**Up**	**Down**	**Total**	**Up**	**Down**	**Total**
2.0	0	0	0	0	1	1	2	4	6
1.5	0	0	0	5	16	21	11	39	50
1.2	5	30	35	141	224	365	290	350	640

Table 11.5: Number of genes significantly associated with PF ratio for different thresholds of mean expression. The values of PF ratio represent clinically used end points to represent acute respiratory distress syndrome, acute lung injury and normal lung function respectively.

The genes with a 1.5 fold difference in mean expression for a change in PF ratio of 20 are shown in Table 11.6. Increasing PF ratio is associated with higher expression of immune related receptor *FCER1A* and repression of *IL18R1*. Effective oxygenation also down-regulates *CA1* (carbonic anhydrase), perhaps reflecting a better acid-base

balance. Interestingly, *ARG1* (arginase), *VNN1* (vannin, involved in haematopoietic cell trafficking), and *IL18R1* are part of the gene set that differentiate the patient clusters (defined on the basis of gene expression profiles, section 8.3 on page 87) in the first principal component of variance. This suggests that part of the cluster differentiation is oxygenation related, which is reflected in the covariate association with the clusters.

Gene	Chr	Function	Coefficient	95% CI
CCR3	3	chemokine (C-C motif) receptor 3	0.04	0.02 – 0.07
KLRB1	12	killer cell lectin-like receptor subfamily B, member 1	0.03	0.01 – 0.05
LGALS2	22	lectin, galactoside-binding, soluble, 2	0.04	0.02 – 0.06
FCER1A	1	Fc fragment of IgE	0.03	0.01 – 0.06
CLC	19	Charcot-Leyden crystal protein	0.04	0.02 – 0.07
C19orf59	19	chromosome 19 open reading frame 59		
ARG1	6	arginase, liver	-0.04	-0.07 – -0.02
VNN1	6	vannin 1	-0.06	-0.08 – -0.03
ZDHHC19	3	zinc finger, DHHC-type containing 19	-0.04	-0.07 – -0.01
CST7	20	cystatin F (leukocystatin)	-0.04	-0.05 – -0.02
DSC2	18	desmocollin 2	-0.03	-0.05 – -0.01
CA1	8	carbonic anhydrase I	-0.03	-0.07 – -0.00
IL18R1	2	interleukin 18 receptor 1	-0.04	-0.06 – -0.02
CDK5RAP2	9	CDK5 regulatory subunit associated protein 2	-0.05	-0.07 – -0.03
GRB10	7	growth factor receptor-bound protein 10	-0.03	-0.05 – -0.02
ALAS2	X	aminolevulinate, delta-, synthase 2	-0.03	-0.06 – -0.01
HPGD	4	hydroxyprostaglandin dehydrogenase 15-(NAD)	-0.04	-0.07 – -0.02
CD177	19	CD177 molecule	-0.03	-0.06 – -0.01
OLAH	10	oleoyl-ACP hydrolase	-0.03	-0.06 – -0.01

Table 11.6: Genes significantly associated with PF ratio in the Bayesian analysis. Genes in the upper panel are induced and in the lower panel are repressed with increased oxygenation. Coefficients are expressed in the log2 scale and represent expected change in expression value for a unit change in PF ratio.

Poor tissue oxygenation is a risk factor for adverse outcome. The common elements in the genetic signatures of survival and oxygenation response could reflect a genome level mechanism of this association. With a 1.5 fold change in mean expression as threshold for each of the comparisons, there are no significant genes that are shared between increasing PF ratio and increased mortality risk. The lack of association be-

tween oxygenation and survival is surprising given the strong clinical association and biological plausibility. This could be due to the cluster differentiation capturing most of the oxygenation effects and the possibility that mortality in poor oxygenation states is explained by organ dysfunction (such as cardiac dysfunction) that is not manifest as changes in leucocyte gene expression.

11.3 Therapy

Steroids

There are three groups of endogenous steroid hormones derived from the three layers of the adrenal cortex: aldosterone, hydrocortisone, and sex steroids. In clinical use the term steroid refers to hydrocortisone and this convention is used in this section.

The role of steroids in sepsis is controversial (Mason *et al.,* 2009) Mortality in sepsis was thought to be due to an inappropriate and uncontrolled hyper-inflammatory state resulting in tissue damage.(Chong *et al.,* 2011) The immunosuppressive effects of steroids were believed to be therapeutic. Randomised control studies failed to show survival benefits with high dose steroid therapy but indicated that physiological doses of steroids could be useful in certain subgroups characterised by acquired adrenal insufficiency, manifesting as either high or persistent vasoactive drug requirements. It was thought that clinically relevant adrenal insufficiency could be diagnosed with a short synacthen test which would then identify subgroups who might benefit from steroid supplementation. However, the existence and diagnostic criteria for relative adrenal insufficiency, as well as the effects of "low-dose" hydrocortisone on outcomes have been challenged. The recent surviving sepsis guidelines does not recommend the routine use of steroids in sepsis. (Surviving Sepsis guidelines, `http://www.survivingsepsis.org/guidelines/`.))

Hydrocortisone and related compounds are liphophilic and cross cell membrane boundaries with ease. These compounds are ligands to cytosolic receptors; the combined ligand-receptor complex translocates into the nuclear membrane and acts as a transcription factor that modulates the expression of a number of genes across the genome.

With a 1.5 fold change in mean expression as threshold, 135 genes show significant differences in expression with steroid use. Of these, 97 are induced and 38 are repressed. With the use of more stringent thresholds (2.0 fold change in mean expression), the predominant pattern of induction with steroid use is maintained, but with a relaxed threshold of 1.2 fold change the predominant change is repression (Table 11.7). While steroids are traditionally associated with immunosuppression, the strongest changes at the genomic level seem to be induction.

Fold change	Up	Down	Total
2.0	31	7	38
1.5	97	38	135
1.2	571	345	916

Table 11.7: Number of genes associated with steroid use in CAP sepsis in the Bayesian hierarchical models with different thresholds of mean expression change.

The genes that show at least 2.0 fold change in mean expression with steroid use after controlling for other covariates are shown in tables 11.8 and 11.9. The use of steroids is associated with increased expression of neutrophil related genes (after correcting for altered neutrophil counts). There is also a surprising association with haematopoietic genes *MS4A3*, *LTF*, and *TCN1*.

The effect of steroids on gene expression and its relation to survival related gene expression is of interest because of the controversies surrounding the clinical association. The Bayesian hierarchical regression analysis offers a unique insight into these interactions. If steroid use was associated with survival benefits, discordant changes would be expected between the survival and steroid gene sets, whereby improved survival should be seen with steroid use. With a 1.5 fold change in mean expression as threshold in both comparisons, four genes show discordant changes (Figure 11.3). Of these three (*LST1* - lymphocyte antigen, *FCER1A*, *HDC*) are repressed with steroid use and are associated with death, while one (*CD163* - haemoglobin scavenger receptor) is induced with steroid use and is associated with survival. Interestingly, three genes (*OAS1* - oligoadenylate synthetase, *TNFSF13B* - TNF ligand superfamily, *CYSLTR1* -

Gene	Chr	Function	Coefficient	95% CI
MPO	17	myeloperoxidase	1.40	0.51 – 2.31
TACSTD2	1	tumour-associated calcium signal transducer 2	1.46	0.66 – 2.29
TCN1	11	transcobalamin I (vitamin B12 binding protein)	1.18	0.32 – 2.03
MS4A3	11	membrane-spanning (hematopoietic cell-specific)	1.50	0.60 – 2.36
DEFA4	8	defensin, alpha 4, corticostatin	1.74	0.62 – 2.84
CD163	12	CD163 molecule	1.14	0.50 – 1.76
LTF	3	lactotransferrin	1.09	0.04 – 2.13
FLT3	13	fms-related tyrosine kinase 3	1.05	0.68 – 1.41
SLC2A5	1	glucose/fructose transporter	1.14	0.51 – 1.79
AZU1	19	azurocidin 1	1.55	0.65 – 2.40
PRTN3	19	proteinase 3	1.28	0.48 – 2.09
ELANE	19	elastase, neutrophil expressed	1.52	0.46 – 2.60
ADORA3	1	adenosine A3 receptor	1.34	0.77 – 1.91
CHI3L1	1	chitinase 3-like 1 (cartilage glycoprotein-39)	1.14	0.45 – 1.85
OLFM4	13	olfactomedin 4	1.44	0.27 – 2.63
MMP8	11	matrix metallopeptidase 8	1.41	0.34 – 2.46
CEACAM6	19	cell adhesion molecule 6	1.61	0.61 – 2.63
RNASE3	14	ribonuclease, RNase A family, 3	1.19	0.36 – 2.01
CTSG	14	cathepsin G	1.72	0.68 – 2.74
ABCA13	7	ATP-binding cassette	1.20	0.51 – 1.91
BPI	20	bactericidal/permeability-increasing protein	1.25	0.33 – 2.20
BEX1	X	brain expressed, X-linked 1	1.31	0.55 – 2.09
CECAM8	19	cell adhesion molecule 8		
VSIG4	X	V-set and immunoglobulin domain containing 4	1.02	0.63 – 1.42
MAOA	X	monoamine oxidase A	2.03	1.18 – 2.89
OLAH	10	oleoyl-ACP hydrolase	1.41	0.74 – 2.11
ARL4A	7	ADP-ribosylation factor-like 4A	1.17	0.69 – 1.66
MFGE8	15	milk fat globule-EGF factor 8 protein	1.25	0.72 – 1.79

Table 11.8: Genes induced with steroid use in sepsis in the Bayesian regression model. Coefficients are in log2 scale.

Gene	Chr	Function	Coefficient	95% CI
CX3CR1	3	chemokine (C-X3-C motif) receptor 1	-1.22	-1.80 – -0.64
XAF1	17	XIAP associated factor 1	-1.04	-1.77 – -0.30
OASL	12	2′-5′-oligoadenylate synthetase-like	-1.02	-1.68 – -0.36
MX1	21	myxovirus resistance 1, IFN-inducible	-1.11	-1.97 – -0.25
FCER1A	1	Fc fragment of IgE, high affinity	-1.09	-1.77 – -0.40
HERC5	4	hect domain and RLD 5	-1.01	-1.89 – -0.10
OAS1	12	2′-5′-oligoadenylate synthetase 1, 40/46kDa	-1.01	-1.75 – -0.28

Table 11.9: Genes repressed with steroid use in sepsis in the Bayesian hierarchical model. Coefficients are in the log2 scale.

cysteinyl leukotriene receptor) are repressed in both groups (associated with survival, and repressed with steroid use). There are no associations in the opposite direction.

Activated Protein C

As with steroids, the use of activated protein C in sepsis has been contentious. Sepsis is a pro-thrombotic state in which capillary occlusion at the micro-circulatory level results in micro-infarcts and tissue damage. This was proposed to be due to a deficiency of functional protein C, an endogenous anti-coagulant, which could be corrected by the administration of activated protein C. The treatment was contentious because of the high therapeutic costs, the risks of bleeding, and limited evidence for survival benefits following treatment. Following the failure to identify any survival benefit, activated protein C has recently been withdrawn and is no longer indicated in sepsis (Dhainaut *et al.,* 2003).

Despite the lack of therapeutic benefits, the use of activated protein C is associated with considerable changes in gene expression. With a 1.5 fold change in mean expression as threshold, 88 genes have significant difference in mean expression with activated protein C use. Of these, 63 genes are induced, while 25 are repressed. With different thresholds, there is a consistent pattern of induction of gene expression with the use of activated protein C (Table 11.10).

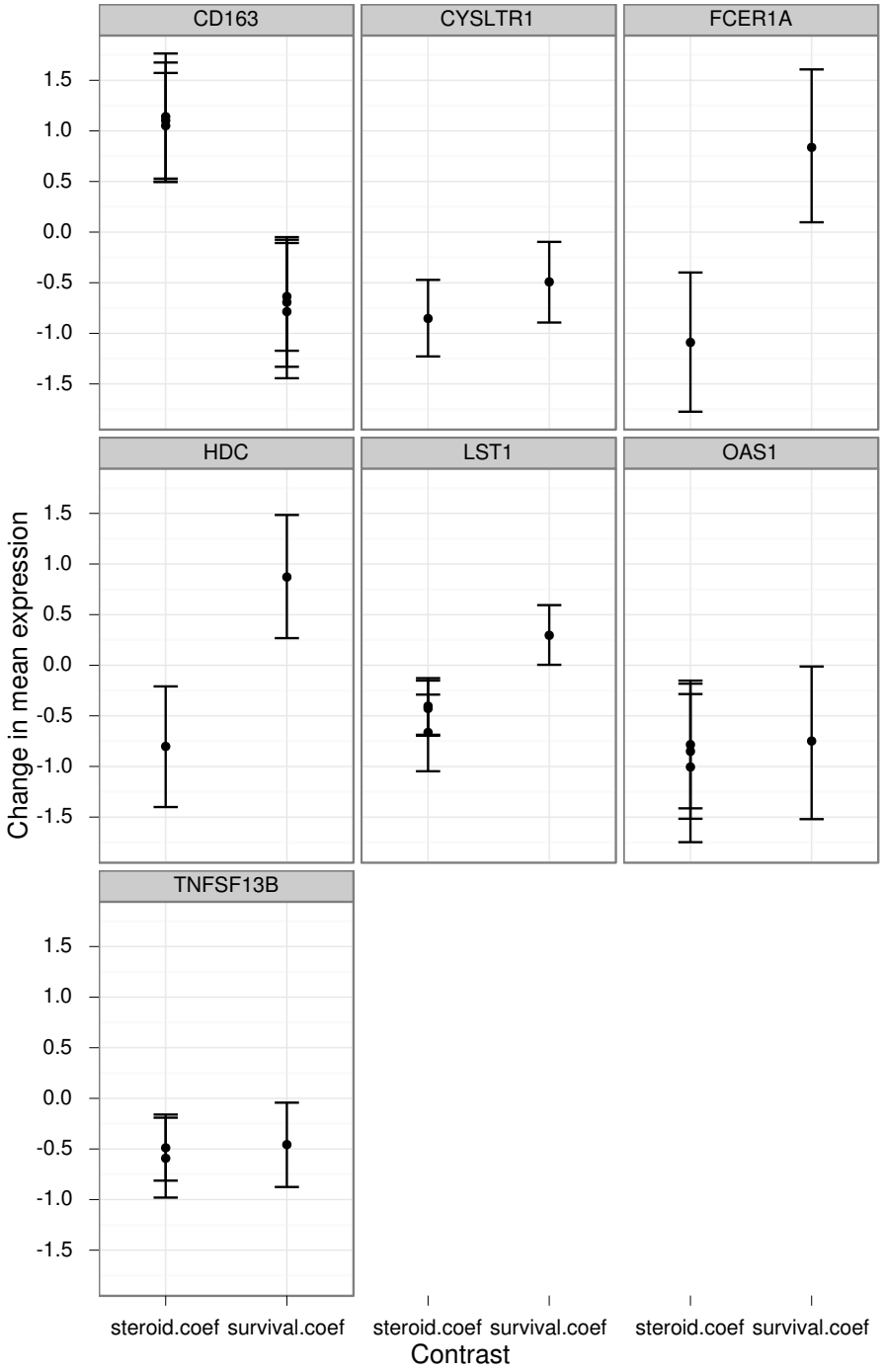


Figure 11.3: Genes associated with steroid use and survival in the Bayesian analysis. The plots show the change in mean expression with the 95% credible intervals. Multiple bars for a coefficient indicates multiple probes for the gene. steroid.coef = regression coefficient for steroid use, survival.coef = regression coefficient for non-survival

Fold change	Up	Down	Total
2.0	10	3	13
1.5	63	25	88
1.2	415	308	717

Table 11.10: Number of genes significantly associated with APC use in the Bayesian analysis for different thresholds of change in mean expression.

The genes that show at least 2.0 fold change in gene expression with activated protein C use are listed in Table 11.11. The use of activated protein C is associated with induction of genes encoding ATP binding cassettes. When a 1.5 fold change in gene expression is used, additional associations include induction of genes encoding the class II MHC molecule *HLA-F*, interleukin 1β *IL1B*, apoptosis related protein *CASP5*, and C type lectin domain containing proteins *CLEC7A*, and *CLEC12A*. In addition, with a 1.5 fold mean expression change threshold, there is repression of lysozyme *LYZ*, complement components *C1QC*, and *C1QB*, low density lipoprotein receptor *LDLR*, and class II MHC molecule *HLA-DRB4*. The use of activated protein C is associated with significant change in expression of a number of immunity related genes. In particular, there seems to repression of the complement system and mixed effects in the class II MHC. These changes do not seem to be directly related to changes in tissue perfusion caused by activated protein C.

As with steroids, the use of activated protein C should be associated with discordant changes in the survival related genes if treatment is associated with a survival benefit. A total of 11 genes show such discordant changes; Of these 3 genes (*LOC644936*, *RPS29*, *CD79B*) are repressed with APC use and are induced in death. The remaining 11 show changes in the opposite direction (Figure 11.4). There are also concordant changes, where the use of activated protein C causes gene expression changes similar to the changes seen in death. Three genes show concordant effects, of which two (*ABCA1*, *ABCG1*) are induced in death and with APC use. *HLA-DRB4* is repressed in survivors and also with APC use. Considering these changes, it is reasonable to infer a complex effect of APC on sepsis survival that is unrelated to changes to tissue perfusion. The effects on survival seen with APC use could be related to changes to the inflammatory pattern, which

Gene	Chr	Function	Coefficient	95% CI
PRKAR1A	17	cAMP-dependent kinase, regulatory	1.23	0.51 – 1.98
KCNMA1	10	K^+ large conductance Ca^{2+}-activated channel	1.14	0.38 – 1.89
CLEC12A	12	C-type lectin domain family 12, member A	1.08	0.15 – 1.99
MME	3	membrane metallo-endopeptidase	1.04	0.27 – 1.84
C4BPA	1	complement component 4 binding protein	1.70	0.49 – 2.86
ABCG1	21	ATP-binding cassette	1.02	0.42 – 1.65
ORM1	9	orosomucoid 1	1.07	0.06 – 2.12
ABCA1	9	ATP-binding cassette	1.10	0.60 – 1.61
TMEM119	12	transmembrane protein 19	1.07	0.42 – 1.72
ERAP2	5	endoplasmic reticulum aminopeptidase 2	-1.17	-1.99 – -0.32
CLC	19	Charcot-Leyden crystal protein	-1.49	-2.50 – -0.46
LOC644936	5	beta actin, pseudogene	-1.25	-2.27 – -0.27

Table 11.11: Genes significantly associated with activated protein C use in sepsis in the Bayesian analysis.

supports the notion that the benefit of APC would be best seen in appropriate subgroups, defined by gene expression rather than clinical criteria.

11.4 Cell count covariates

Total white cell counts

The total white cell count can either be reduced or increased in severe sepsis. However, this should not impact gene expression for two reasons. First, the amount of mRNA hybridized to the microarray is relatively constant, and second the normalization process removes any residual bias by readjusting the values to a reference that is relatively constant across the microarrays. The Bayesian hierarchical model includes the total white cell count as a predictor to identify any problems with this process. Even with a relaxed threshold of 1.2 fold change in mean expression, there are no probes with expression values related to total white cell counts. This reinforces the confidence in the hybridisation techniques and the analytical methods.

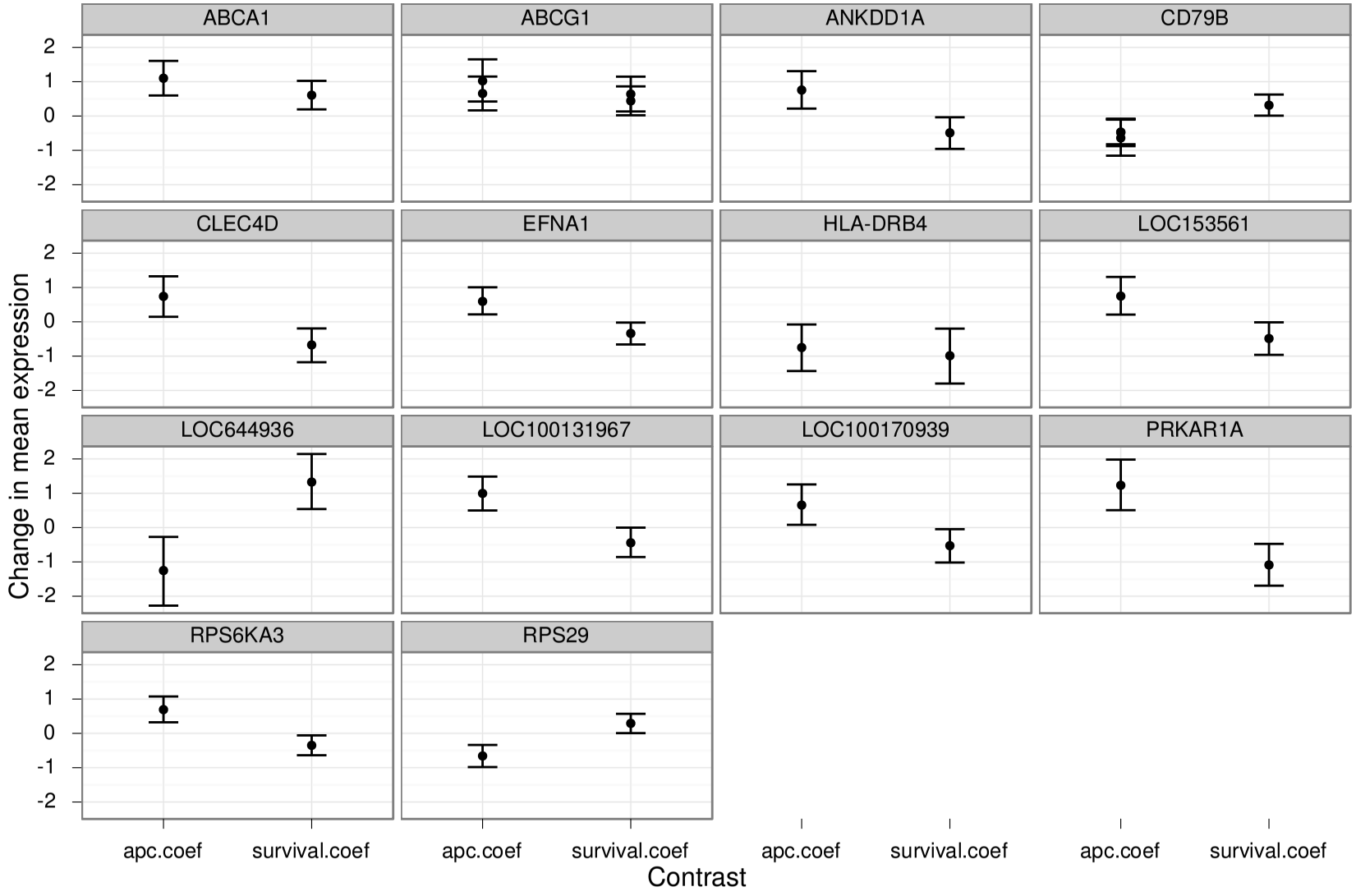


Figure 11.4: Genes associated with APC use and survival in the Bayesian analysis. Plots show change in mean expression for that coefficient along with the 95% credible interval. apc.coef = regression coefficient for APC use, survival.coef = regression coefficient for non-survival

Relative white cell counts

In contrast to the total white cell counts where the effect of changes in cell counts are removed by the technical and statistical approach, the systematic effects of the variations in relative cell counts are preserved. Cell type specific gene expression has been well described in a variety of physiological and pathological processes. Gene expression changes specific to the cell types are important for understanding fundamental biological and pathophysiological mechanisms, while the changes in overall expression may be more relevant to practical applications. The Bayesian hierarchical model includes the relative neutrophil and monocyte counts as covariates in the regression model. Since the relative lymphocyte count is dependent on these two counts, the inclusion of relative lymphocyte counts results in a singular matrix, and therefore, cannot be directly inferred.

Relative neutrophil counts

The sepsis response can manifest as an increase in neutrophil counts; therefore the relative neutrophil counts would be expected to have large effects on gene expression. This is clearly demonstrated in Table 11.12; of all the covariates, relative neutrophil counts have the greatest effects in terms of the number of associated probes. Even with a stringent threshold of at least 2.0 fold change in mean expression, 4965 probes show significant differences. Of these 3114 probes are repressed and 1862 are induced. Two notable features include the large effect sizes (Figure 11.5)) and the consistent pattern of predominant repression for any given effect size. Of all the covariates, relative neutrophil counts are associated with the largest effect sizes. Of the 6144 probes that show a significant change with a conventional 1.5 fold difference in mean expression, 1524 (24.8%) show greater than 5 fold change in expression, and 155 (2.5%) probes show greater than 10 fold change in expression.

The genes with the largest effect sizes are listed in Table 11.13. As expected a number of immune related genes are associated with relative neutrophil counts.

Fold change	Up	Down	Total
10.0	155	281	436
5.0	544	980	1524
4.0	773	1384	2157
2.0	1862	3114	4965
1.5	2565	3618	6144

Table 11.12: Number of genes associated with relative neutrophil counts in the Bayesian analysis for different thresholds of change in mean expression.

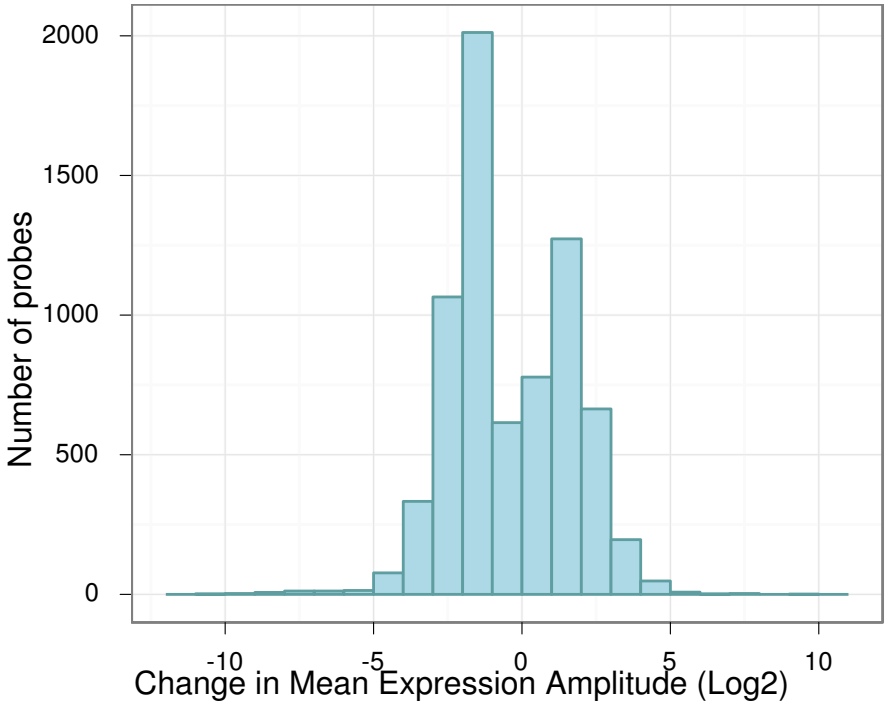


Figure 11.5: Effect sizes of relative neutrophil counts in the Bayesian regression model. The plot includes the probes with all effect sizes where the 95% credible interval does not equal zero.

Gene	Chr	Function	Coefficient	95% CI
SIGLEC10	19	sialic acid binding Ig-like lectin 10	5.50	2.21 – 8.75
RFX2	19	regulatory factor (HLA class II expression)	5.28	2.61 – 7.74
VAV3	1	VAV 3 guanine nucleotide exchange factor	9.34	1.82 – 15.69
PI3	20	peptidase inhibitor 3, skin-derived	6.07	0.13 – 12.27
CASP5	11	caspase 5, apoptosis-related cysteine peptidase	5.56	2.08 – 9.04
DSC2	18	desmocollin 2	7.20	4.41 – 9.86
ABCA1	9	ATP-binding cassette, sub-family A (ABC1)	7.04	4.71 – 9.41
CDK5RAP2	9	CDK5 regulatory subunit associated protein 2	7.85	3.97 – 11.96
LPCAT2	16	lysophosphatidylcholine acyltransferase 2	6.17	2.41 – 9.92
BCAT1	12	branched chain amino-acid transaminase 1	5.85	2.16 – 9.59
MPO	17	Myeloperoxidase	-7.91	-13.67 – -2.06
DEFA4	8	defensin, alpha 4, corticostatin	-9.75	-16.01 – -2.93
ELANE	19	elastase, neutrophil expressed	-10.87	-17.53 – -3.60
CEACAM6	19	CEA-related cell adhesion molecule 6	-7.87	-13.79 – -1.20
IGLL1	22	immunoglobulin lambda-like polypeptide 1	-8.37	-14.34 – -2.89
CTSG	14	cathepsin G	-8.80	-14.78 – -3.05
IGJ	4	immunoglobulin J polypeptide	-8.59	-15.52 – -2.06
CEACAM8	19	CEA-related cell adhesion molecule 8	-8.32	-15.20 – -1.39

Table 11.13: Genes associated with relative neutrophil counts. Genes in the top panel are induced and in the bottom panel are repressed. Each gene shows nearly 100 fold change for a 1% change in relative neutrophil count.

Relative monocyte counts

While neutrophils are important effector cells in sepsis, the monocyte-macrophage system is the primary modulator of the innate immune response. As with relative neutrophil counts, changes in relative monocyte counts would be expected to have a large impact on gene expression.

With a conventional threshold of 1.5 fold change in mean expression, 3619 probes show significant differences in expression. Of these 1457 are induced and 2157 are repressed. Across all thresholds, the predominant pattern of expression is repression (Table 11.14), but compared to relative neutrophil counts the proportion of induced probes is higher. Change in relative monocyte counts are associated with large effect sizes (Figure 11.6). Of the 3619 probes that have a 1.5 fold change in mean expression,

1951 (53.9%) show greater than 5 fold change in mean expression, and 915 (25.5%) show greater than 10 fold change in mean expression. In addition, for a higher proportion of genes with large effect sizes within each group, the number of genes showing larger effect size is higher in monocytes when compared to neutrophils. For example 924 probes show greater than 10 fold change in expression with change in relative monocyte counts, while the corresponding number is 155 for neutrophils. These findings are compatible with the regulatory role of monocytes, where small changes in monocytes result in amplified end effects.

Fold change	Up	Down	Total
10.0	387	537	924
5.0	777	1174	1951
4.0	915	1427	2341
2.0	1442	2126	3556
1.5	1475	2157	3619

Table 11.14: Number of genes associated with relative monocyte counts in the Bayesian analysis for different thresholds of changes in mean expression.

The genes with the largest effect sizes are listed in Table 11.15. As expected a number of immune related genes are associated. When combined with Table 11.13, some interesting observations can be made. Among the most induced genes in both groups, *VAV3* (guanine nucleotide exchange factor), *PI3* (peptidase inhibitor), and *ABCA1* (ATP cassette) are shared. Since monocytes are modulatory and neutrophils are effectors, it is likely that these genes are primarily expressed in neutrophils but modulated by monocytes. Similar inferences can be made about the repressed genes *DEFA4*, *ELANE*, *CECAM6*, and *CECAM8*.

The overlap of the neutrophil and monocyte relative count associated changes in gene expression are shown in Figure 11.7. 1121 genes are induced with at least 1.5 fold change in mean expression in both groups. This constitutes 44% of the genes induced in the neutrophil contrast and 77% of genes in the monocyte contrast. 1927 genes repressed with at least 1.5 fold change in mean expression in both contrasts. This constitutes 53.8% of the neutrophil gene yield and 90% of the monocyte gene yield.

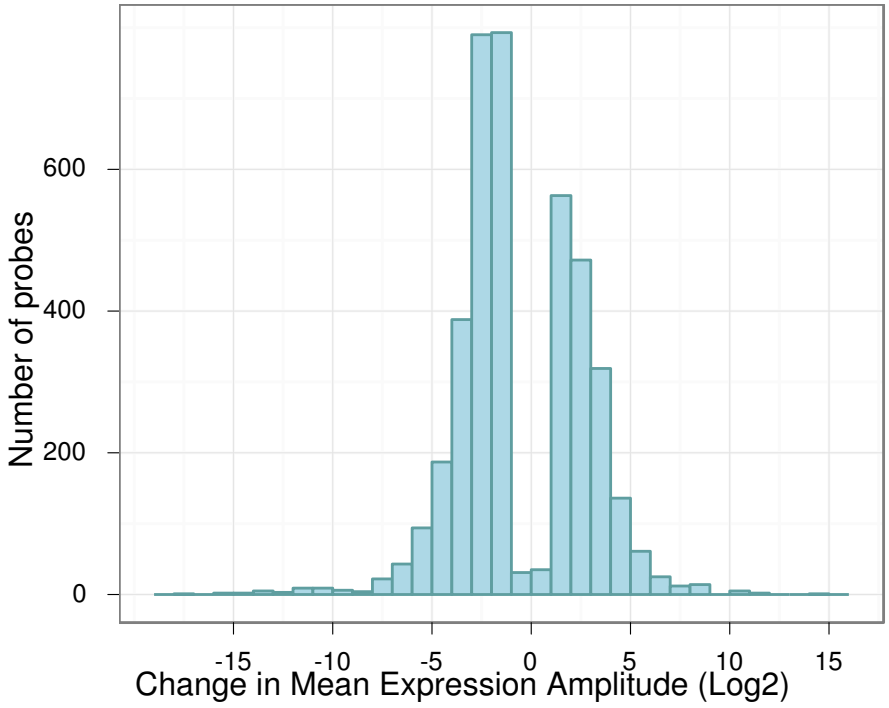


Figure 11.6: Effect sizes of relative monocyte counts with the Bayesian regression model for all probes. The plot includes the probes with all effect sizes where the 95% credible interval does not equal zero.

Interestingly there are some discordant effects, both within and across groups. 14 genes in the neutrophil group show discordant effects for different probes of the same gene. These could either be false-positive effects or the result of alternate splicing or differential transcript stability based on transcript sequence. In addition, 12 genes show discordant effects at the probe level in each group in both the monocyte and neutrophil comparisons. These genes are listed in Table 11.16. The consistency across the contrasts, shifts the weight of evidence to biological effects rather than statistical artefacts.

There are also a small number of genes which show discordant effects between the groups. 23 genes are induced with neutrophil and repressed with monocytes (Table 11.17). 20 genes show opposite effects (Table 11.18). 12 genes are shared in each group, and correspond to the 12 genes described earlier (Table 11.16).

The overall correlation (Figure 11.8) between the coefficients for probes that have at least 1.5 fold change in mean expression in both neutrophil and monocyte relative cell count associations is high (correlation coefficient of 96.7%).

Gene	Chr	Function	Coefficient	95% CI
MS4A7	11	membrane-spanning 4-domains, subfamily A	11.10	5.51 – 16.25
CD36	7	CD36 molecule (thrombospondin receptor)	10.76	5.81 – 15.39
CSF1R	5	colony stimulating factor 1 receptor	8.90	3.32 – 14.06
VAV3	1	VAV 3 guanine nucleotide exchange factor	14.93	3.20 – 25.24
PI3	20	peptidase inhibitor 3, skin-derived	10.70	0.99 – 20.90
KYNU	2	kynureninase	8.90	3.67 – 14.03
DSC2	18	desmocollin 2	10.96	6.12 – 15.62
ABCA1	9	ATP-binding cassette, sub-family A (ABC1)	11.14	7.22 – 15.03
TIAM2	6	T-cell lymphoma invasion and metastasis 2	10.12	3.77 – 15.60
BCAT1	12	branched chain amino-acid transaminase 1	8.77	3.01 – 14.80
TACSTD2	1	tumour-associated calcium signal transducer 2	-11.73	-19.57 – -3.63
DEFA4	8	defensin, alpha 4, corticostatin	-15.12	-25.86 – -3.87
DEFA1B	8	defensin, alpha 1B	-13.31	-22.32 – -4.42
ELANE	19	elastase, neutrophil expressed	-15.76	-26.95 – -4.14
MMP8	11	matrix metallopeptidase 8 (neutrophil collagenase)	-14.73	-25.59 – -3.81
CEACAM6	19	CEA related cell adhesion molecule 6	-11.88	-21.59 – -1.18
CTSG	14	cathepsin G	-14.52	-24.41 – -4.79
BPI	20	bactericidal/permeability-increasing protein	-12.76	-21.67 – -3.47
CEACAM8	19	CEA related cell adhesion molecule 8	-13.42	-24.43 – -2.30

Table 11.15: Genes associated with relative monocyte counts. Genes in the top panel are induced and in the bottom panel are repressed. Each gene shows nearly 100 fold change for a 1% change in relative monocyte count.

Gene	Chr	Function
TMEM189-UBE2V1	20	TMEM189-UBE2V1 readthrough
HNRNPA1L2	13	heterogeneous nuclear ribonucleoprotein A1-like 2
PNKD	2	paroxysmal nonkinesigenic dyskinesia
PILRB	7	paired immunoglobulin-like type 2 receptor beta
KIAA0101	15	KIAA0101
ANAPC1	2	anaphase promoting complex subunit 1
PDCD7	15	programmed cell death 7
HYDIN	16	hydrocephalus inducing homolog
GMPR2	14	guanosine monophosphate reductase 2

Table 11.16: Genes with discordant effects between probes in both the relative neutrophil and monocyte count contrasts in the Bayesian analysis

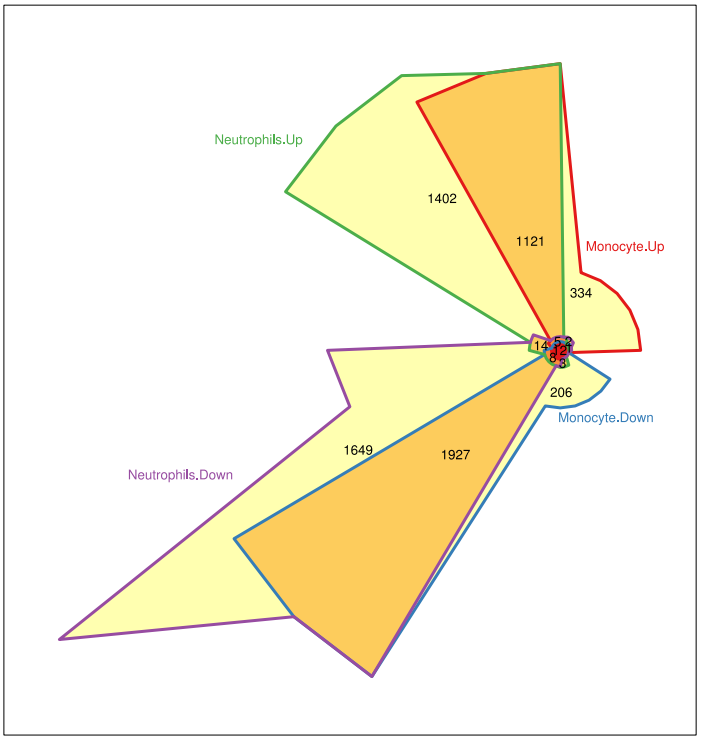


Figure 11.7: Overlap of genes associated with both relative neutrophil and monocyte counts in the Bayesian analysis. 1402 genes are uniquely induced with increase in relative neutrophil counts. 334 genes are uniquely induced with increase in relative monocyte counts. Similarly 1649 and 206 genes are uniquely repressed for neutrophils and monocytes respectively. 1121 genes are induced in common with increase in relative numbers of both the cell types, while 1927 are repressed.

Gene	Chr	Function
CSMD1	8	CUB and Sushi multiple domains 1
GALK2	15	galactokinase 2
GABBR1	6	gamma-aminobutyric acid (GABA) B receptor, 1
YIPF5	5	Yip1 domain family, member 5
CCR6	6	chemokine (C-C motif) receptor 6
TRIM4	7	tripartite motif containing 4
ADAM15	1	ADAM metallopeptidase domain 15
ZKSCAN5	7	zinc finger with KRAB and SCAN domains 5

Table 11.17: Genes with discordant effects between relative neutrophil and monocyte counts in the Bayesian analysis and induced in neutrophils

Gene	Chr	Function
ELF2	4	E74-like factor 2 (ets domain transcription factor)
GPS1	17	G protein pathway suppressor 1
MTMR10	15	myotubularin related protein 10
CDCA7	2	cell division cycle associated 7
VNN2	6	vannin 2
THADA	2	thyroid adenoma associated
XKR6	8	XK, Kell blood group complex subunit-related family, member 6
OTUD6B	8	OTU domain containing 6B

Table 11.18: Genes with discordant effects between relative neutrophil and monocyte counts in the Bayesian analysis and repressed in neutrophils

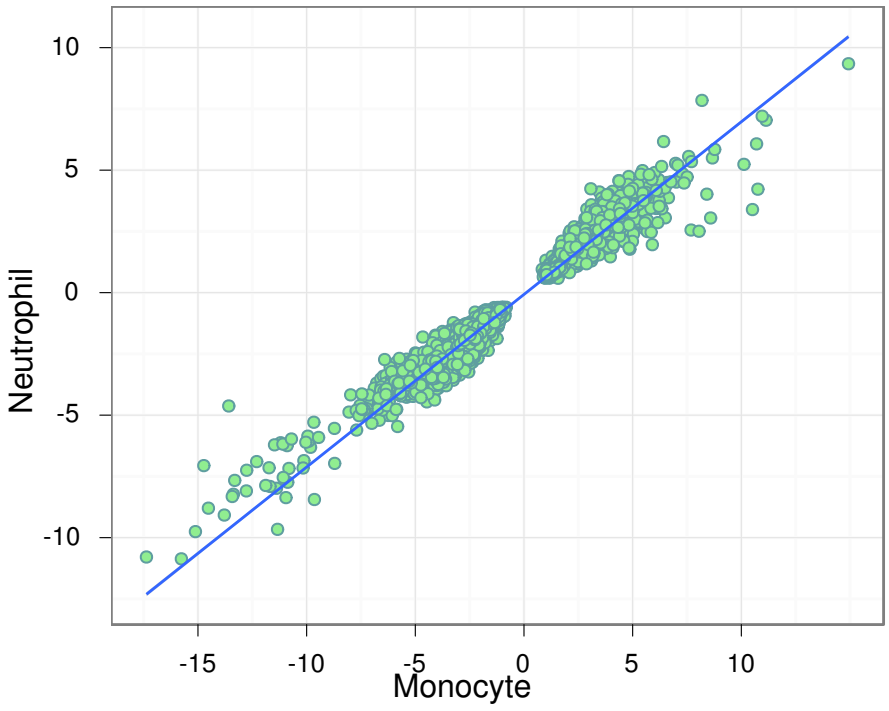


Figure 11.8: Correlation of the relative neutrophil and monocyte counts coefficients in the Bayesian regression model. Coefficients are in the log2 scale.

11.5 Microbiological Covariates

The molecular events surrounding the sepsis response can be characterized as the interaction of two genomes: the host and the invading pathogen. Characteristics of the invading organism can considerably influence the immune response. Therefore, an accurate description of the immune response should involve detailed characterization of the microbe including species and strain. A more manageable discussion would necessarily involve some degree of reductionism based on inclusive categories of microorganisms.

To identify the variation in the host genomic response in sepsis attributable to the nature of the invading pathogen, a Bayesian hierarchical regression analysis was performed, with the microbiological agent as a predictor. This was modelled as a hierarchical variable, where each category of the variable was allowed to have its own mean. The parameters for each category of the variable are determined by hyperparameters determined from the data. The difference in the mean expression of each probe for each category was the primary feature of interest.

Based on the reported microbiological data, 7 categories were defined (Table 11.19). The genomic response in each microbiological category is contrasted with others in the following section.

Group	Components	Number of Patients
Atypical	*Mycoplasma pneumoniae, Legionella spp*	5
Gram Negative	*Pseudomonas aeruginosa, Haemophilus influenzae*	5
Gram Positive	*Staphylococcus aureus, Streptococcus pneumoniae*	24
Mixed		4
Other		16
Unknown		17
Viral		10

Table 11.19: Microbiological categories

The differentially expressed genes for the comparisons of interest are shown in Table 11.20. The genes with the largest changes are listed in Supplementary Tables 12 to 21.

Category	Fold Change	Atypical			Gram negative			Gram positive			Mixed		
		Up	Down	Total	Up	Down	Total	Up	Down	Total	Up	Down	Total
Gram negative	2.0	38	47	85									
	1.5	125	111	235									
Gram positive	2.0	16	31	47	4	11	15						
	1.5	72	117	189	14	53	67						
Mixed	2.0	69	94	163	29	46	75	40	48	88			
	1.5	294	269	563	164	111	275	162	168	330			
Viral	2.0	43	21	64	55	43	98	32	14	46	51	46	97
	1.5	164	101	265	181	170	351	198	175	373	123	93	216

Table 11.20: Numbers of differentially expressed genes between the categories of infection in the Bayesian analysis. The category in the column is the baseline for each comparison.

There are some interesting associations in these comparisons. For example, *HLA-C* (class I MHC) is induced in atypical pneumonias while *HLA-DRB4* (class II MHC) is induced in Gram positive pneumonias. In addition, atypical pneumonias are associated with the induction of interferon related genes and oligoadenylate synthetases. Interferon related genes and oligoadenylate synthetases are also induced in viral pneumonias, along with the induction of the pro-apoptotic gene *CARD17*. In contrast, Gram negative pneumonias are associated with induction of the class II MHC molecule *HLA-DRB5* and small nuclear RNAs, along with *IL32* and *PROS1* (protein S, cofactor for protein C). Some neutrophil specific genes (*ELANE*, *MPO*) are repeatedly identfied in a number of comparisons, which is intriguing since the analysis also includes the relative cell counts as covariates.

The gene expression signatures differentiating Gram negative from viral infections and Gram positive from viral infections were compared to identify a shared component attributable to the bacterial component of these infections. Only a small proportion of the genes were shared between both comparisons with a threshold of at least 1.5 fold change in mean expression (n = 22, 6.2% of genes in the Gram negative comparison, 8.3% of genes in the Gram positive comparison). In this small subset of overlapping genes, the mean expression differences are highly correlated (Figure 11.9, Pearson correlation coefficient 95%). The small overlap suggests that each of the gene expression response in each of the three categories (Gram positive, Gram negative, and viral) are fairly unique.

Comparative changes

Pairwise differentiation of gene expression in microbiological categories has been presented earlier in this section. Some genes are present in more than one gene list. Examination of these genes can be instructive with regard to the association of gene expression to specific microbiological categories.

The oligoadenylate synthetases *OAS1*, *OAS2*, *OAS3* are present in a number of gene lists. In addition, the gene group shows evidence for common regulatory mechanism; all genes in this group seem to have similar expression differences between groups. The expression of these genes across all the microbiological category comparisons are shown

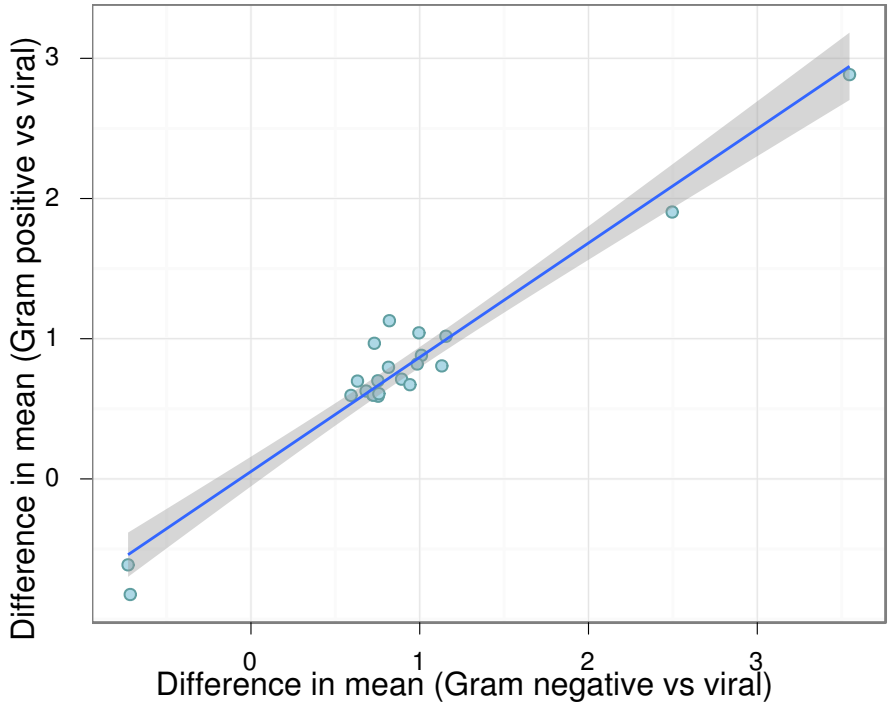


Figure 11.9: Correlation between the Gram positive and negative signatures against viral infections. The small fraction of genes that are shared in both contrasts show high correlation (Pearson correlation coefficient 95%)

in Figure 11.10. The expression of oligoadenylate synthetases are consistently induced in atypical pneumonias and viral pneumonias suggesting similarities in regulatory mechanism in these infections. There is also fairly strong correlation among multiple probes suggesting a robust and dependable effect. Interestingly, there are no significant differences in expression of *OAS3* when these two groups are compared to Gram negative infections, suggesting differences in regulation of this gene in this infective category.

Some interesting observations can be made when this approach is used in genes that are not specifically listed as a major differentiating gene. For example, *TLR4* transcript levels are induced in viral infections when compared to Gram positive and mixed infections, but this difference is not seen in atypical and Gram negative infections.

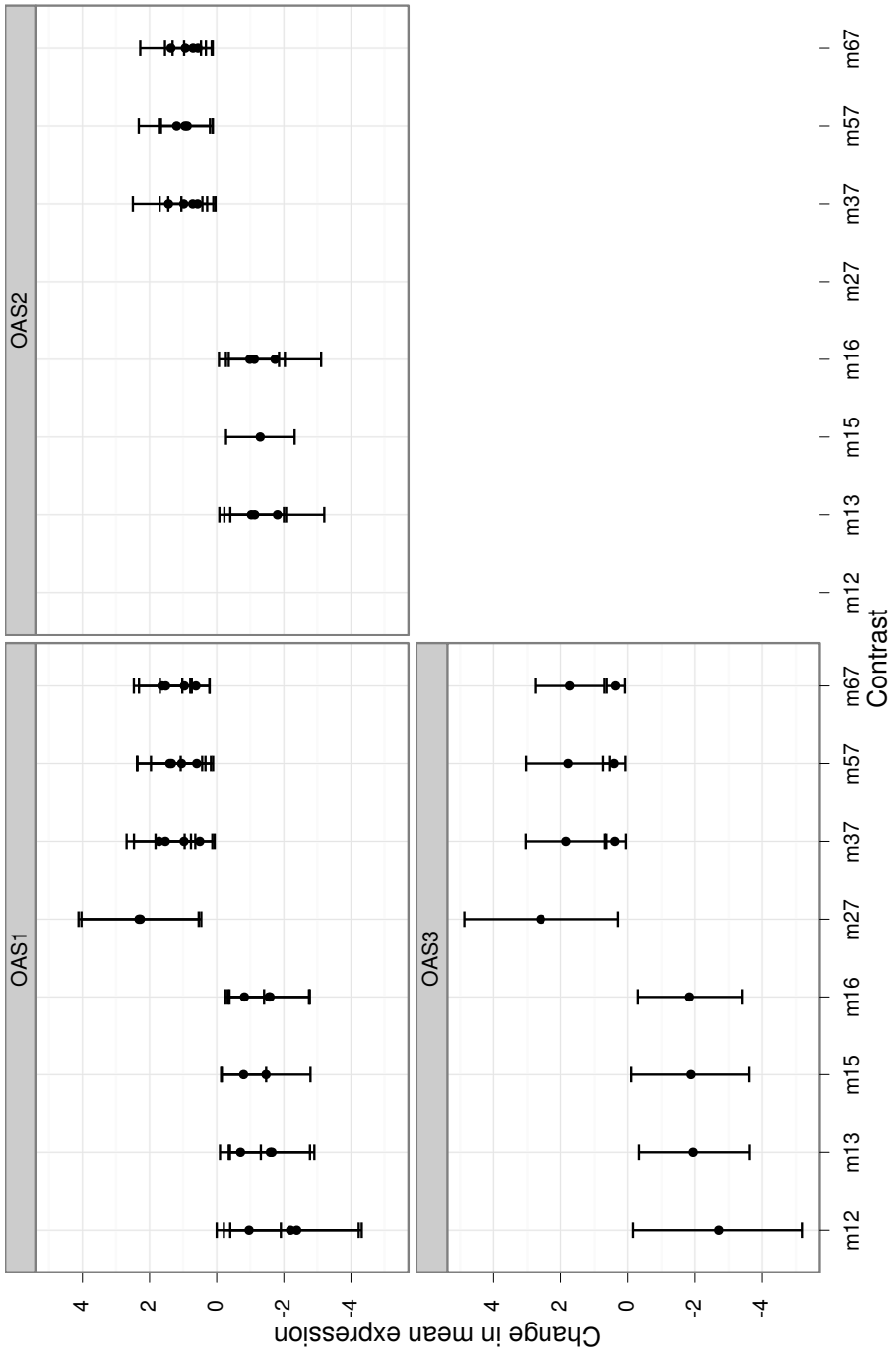


Figure 11.10: The effect of microbiology on oligoadenylate synthetase expression. Coefficients are labelled as mxy, where expression in group x is compared to expression group y. 1 = atypical pneumonia, 2 = Gram negative, 3 = Gram positive, 4 = mixed, 5 = other, 6 = unknown, 7 = viral. Multiple bars for any coefficient indicate multiple probes for that gene. Bars are absent when the coefficient did not reach statistical significance. Coefficient values are in log2 scale. All three genes are down-regulated in atypical pneumonias compared to other infections. When compared to viral pneumonias *OAS1* and *OAS3* are up-regulated in bacterial infections. For the same comparison *OAS2* is notably absent when Gram positive pneumonias are compared to viral pneumonias, suggesting differential regulatory mechanisms.

11.6 Sampling covariates

Day of sampling

Samples collected from each patient on different days of admission have so far been treated as biological replicates. Samples were collected on days 1, 3, and 5 of ITU admission. Day 1 represents a clear clinical milestone representing physiological decompensation requiring ICU admission. The hypothesis that this could also represent a unique pattern of gene expression was not substantiated in the exploratory analysis. The clusters identified in the exploratory analysis are not correlated with the day of sampling, suggesting that the clusters do not represent an evolving inflammatory response.

The correlation of gene expression to the clinically defined time of sampling was investigated in the Bayesian analysis by using the day of sampling as a predictor variable. The inclusion of the day of sampling as a predictor also has the advantage of removing systematic biases in gene expression due to the sampling time.

With a 1.5 fold change in mean expression as a threshold, only one gene showed statistically significant change in expression with the time of sampling. The repressed gene *IFI44L* encodes an interferon induced protein, and is likely to be a spurious association (statistical false positive) despite biological plausibility.

With a relaxed threshold of 1.2 fold change in mean expression as a threshold, 215 genes show significant difference in expression. Of these 138 are induced and 77 are repressed. The list includes a number of genes that have important immune roles like *STAT2*, *BMP6*, *IFIT3*, *IFIT5*, and *DEFA4*, but the significance of these changes is uncertain.

Hybridization and Chip effects

Systematic biases due to hybridization, especially batch effects relating to samples hybridized to the same chip are common in microarray experiments. The normalization process reduces but does not always remove these effects. The inclusion of the array identification as a covariates quantifies and removes this bias. In the Bayesian model, there are no differentially expressed genes related to the hybridizing array, even with a relaxed threshold of 1.2 fold change in mean expression. This result increases confidence

in the normalization process and the ability of the Bayesian hierarchical regression analysis to identify small changes in mean expression between groups.

Recruiting centre

Systematic biases in gene expression could also be possible due to unmeasured differences between the recruiting centres. Possible differences could include heterogeneity of patient cohorts and management strategy among recruiting centres. The inclusion of the recruiting centre as a predictor variable quantifies the bias and increases the accuracy of the associations to other covariates. With a liberal threshold of 1.2 fold change in mean expression, none of the genes show a significant difference in expression among recruiting centres. The homogeneity of patients across the centres is confirmed by this result.

11.7 Summary

The analysis of gene expression in sepsis is complicated by the pathophysiological complexity of this condition and the interactions with a number of covariates, some clinical and others technical, each of which has the potential to modulate or systematically alter the observed gene expression changes. While the observed cumulative gene expression effects in each patient is relevant to that patient and likely determines, or reflects the clinical course and outcome, the independent impact of each covariate for the patient cohort is important. The independent effect due to a covariate provides a mechanistic explanation of the biological influence of that covariate, and provides insights into the contribution of that covariate to gene expression changes associated with clinically relevant outcomes. It is possible that a better understanding of the effect of covariates amenable to clinical manipulation could lead to directed therapeutic efforts to improve clinical outcomes.

The simplest of the covariate effects are seen with the "noise" covariates: day of sampling, technical biases due to the experimental methods, and systematic biases due to the recruiting centre. The latter two have no demonstrable effects in the normalized

data reinforcing the robustness of the normalization methods and increasing confidence in other results. The day of sampling, while a clinically relevant milestone, is a poor marker of the underlying inflammatory state. The effects due to the sampling day are minimal with uncertain biological relevance. The decision to use repeated samples from a patient as biological replicates is justified by these results.

The effects of demographic covariates, age and sex, are more complex. While increasing age is a risk factor for poor outcome in severe sepsis, there are minimal associations with gene expression changes in this data set. This suggests that the impact of increasing age on survival could be due to age related decrease in organ function, which is not captured by changes in gene expression in peripheral blood leucocytes. In contrast, sex, also known to influence sepsis related outcomes, is associated with some convincing gene expression changes. The effects, while modest, occur in genes with important immunity related functions.

There is enough information in the data set to identify the effects of impaired in renal and respiratory function. Both are independent predictors of mortality in severe sepsis. Renal failure is associated with a predominant pattern of repression in gene expression. More than 70 genes show strong effects with an additional 600 genes showing modest effects. Changes are seen in the complement system and the expression of immunoglobulin IgG chains. There are additional effects on lysozyme expression, coagulation system and the lipid regulatory system.

The changes seen in respiratory dysfunction are opposite to that of renal function in terms of directionality of effects, with increased oxygenation or better respiratory function associated with a predominant repressed expression signature. There are some effects related to acid-base balance inferred from large changes in carbonic anhydrase expression. There is some overlap between the genes related to oxygenation and the genes predicting the first principal component (or the greatest variance) of the cluster differentiation, providing a molecular level link analogous to the association of the clinical covariates. There is also some overlap between the PF ratio and the requirement for respiratory support, which reinforces the internal consistency in the data set and the analytical methods.

Even larger changes are seen with the two therapeutic interventions of considerable clinical interest: steroid and activated protein C, the former commonly used, the latter withdrawn, but both equally contentious. Steroid use is conventionally associated with immunosuppression, but in the sepsis cohort examined here, the use of steroids results in a predominantly up-regulated pattern of expression. In the context of the predominant gene silencing in the sepsis response, this could represent the relative normalization of the immune response. There are unexpected associations with the haematopoietic system, along with more conventional associations with neutrophil function.

As is the case with steroids, activated protein C is also associated with large changes in gene expression. These changes do not directly relate to the expected benefits of anti-coagulation and improved tissue micro-perfusion and occur in core immune related genes such as the complement system and the class II MHC. There are additional associations with lysozyme, apoptosis, and C type lectin domain receptors.

Relative cell counts have strong effects both in the terms of number of genes affected and the effect sizes. In contrast, gene expression is not influenced by total white cell counts, reflecting the lack of bias due to hybridisation technique and an effective normalization procedure. Nearly a quarter of known genes show associations with relative neutrophil counts. For relative monocyte counts the association is seen in about one sixth of all known genes. More than 400 genes in the neutrophil group and more than 900 genes in the monocyte group show more than 10 fold change in gene expression for a unit change in relative counts. There is considerable overlap between the most significant effects, with the overlapping genes showing strong correlations in effect sizes in both groups. This is likely due to the modulatory effect of monocytes on neutrophil function.

Interesting effects are also seen with microbiological covariates. Pairwise comparison of each category of infections allows the examination of effects specific to the broadly defined microbiological category. For any pairwise comparison between 40 and 90 genes show at least a 2.0 fold change in mean expression. Both atypical and Gram positive pneumonias show induction of MHC genes but with varying patterns. Atypical

pneumonias up-regulate class I MHC molecule *HLA-C* while Gram positive pneumonias induced class II MHC molecules. Gram positive infections are differentiated from Gram negative pneumonias by changes in expression of metallo-endopeptidase and small nuclear RNA among other genes. Viral infections are characterised by induction of interferon inducible genes. Bacterial infections are characterized by induction of class II MHC molecules.

Some genes show consistent changes in expression across different comparisons. These changes could reflect components that are unique to that microbiological category. For example Gram negative infections are associated with consistent induction of *IL32* (interleukin 32). Similarly oligoadenylate synthetases are consistently induced in viral and atypical pneumonias but not in other infections.

12

Genomic Signature based Predictions

12.1 Background

The genomic signatures associated with different patient cohorts and subgroups were described in the previous chapters. The allocation of patients to various cohorts on the basis of gene expression signatures is described in this chapter.

The clinical application of genome based research in sepsis may result in diagnostic, prognostic, and therapeutic benefits. The first two of these requires the accurate assignment of a clinically relevant label according to the gene expression pattern see in an individual patient.

In addition, the development of predictive models can help distinguish robust associations from weaker associations. A robust association will result in a strong predictive model, while weak associations are associated with poor predictive power despite statistical significance. It has been recommended, at least in clinical studies, that all statistical significant associations should be validated with predictive models.

Clinically relevant discriminants in community-acquired pneumonia include cluster membership and outcome. The differentiation of sepsis from non-sepsis states is also relevant, but has limited clinical value, since sepsis is easily diagnosed on the basis of clinical features. The differentiation of diagnostic labels is more relevant; despite easy clinical diagnosis, the development of accurate predictive models would lend credence to the inference of unique genomic responses included in the common clinical label of sepsis.

The identification of predictive models in sepsis has been fraught with problems especially when predicting survival. The large number of predictors, each with modest effects, and the heterogeneous nature of the outcome variable, result in predictive models that have good sensitivity but poor specificity. In addition, there are a number of statistical alternatives for the generation of predictive models, and for the selection of features to include in the model. Each approach results in different models with different performance characteristics, which makes inference and clinical application difficult. Even when a reasonable model has been constructed the clinical translation of the model is problematic.

12.2 Diagnostic predictions

The identification of cluster membership and the differentiation of faecal peritonitis from community-acquired pneumonia from gene expression signatures are discussed in this section.

Cluster membership

The exploratory analysis of the gene expression data from patients with community-acquired pneumonia revealed the presence of three clusters which are associated with differences in predicted and actual death rates, time to death and relative cell counts.

Principle component analysis

Using simple linear models, 433 genes were identified that distinguish the three clusters. Genes with at least two fold change in gene expression and an adjusted p value less than 0.05 were considered significant. The results of the principle component analysis of the expression data set of these genes is shown in Figure 12.1. The figure shows that despite large effect sizes, statistical significance, and nearly all the variance explained by one axis, the discriminatory power of the genes is poor and is accounted for by a small number of genes. The predictive power of a model discriminating cluster membership based on gene expression pattern is therefore expected to be poor.

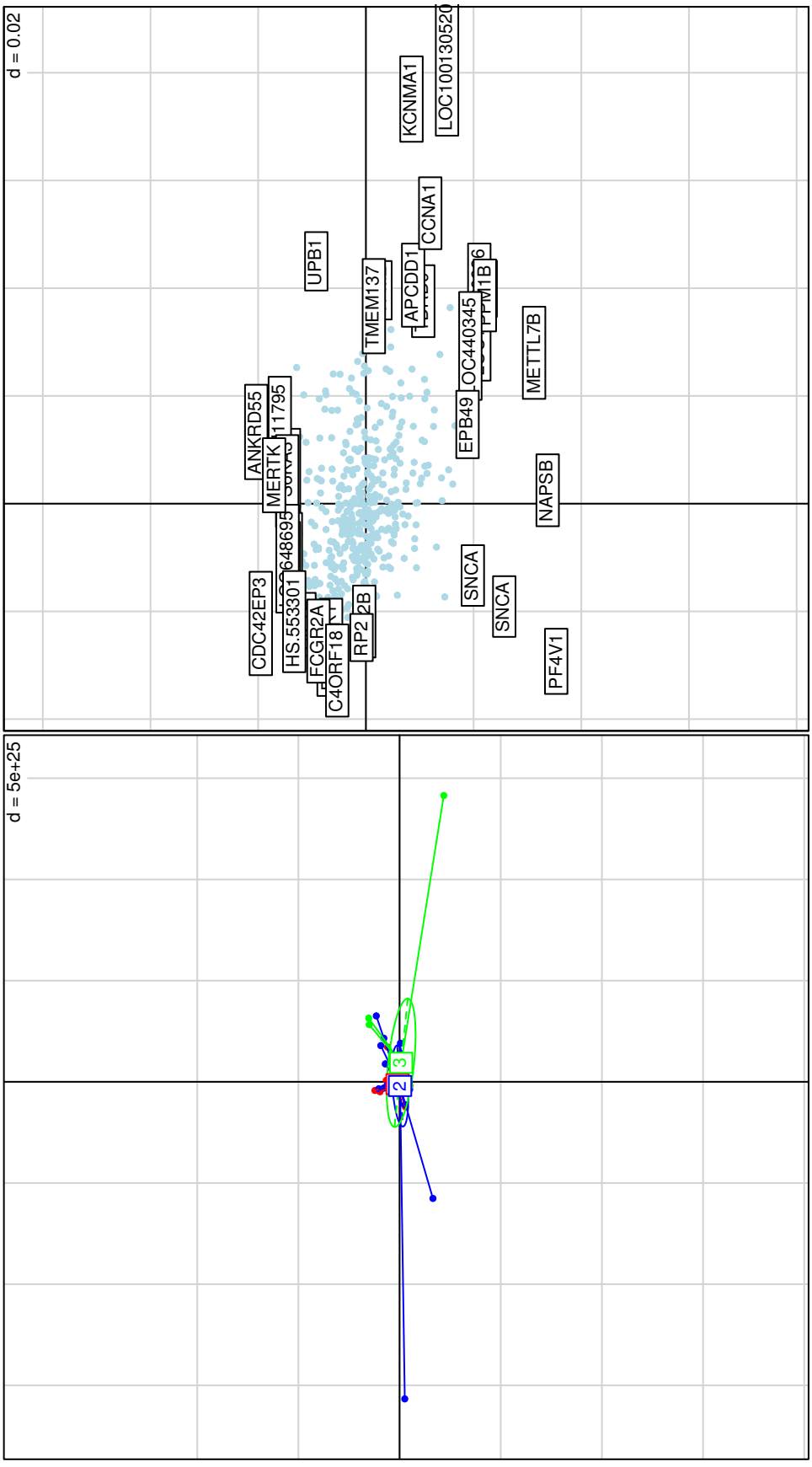


Figure 12.1: PCA based on genes discriminating the clusters. The centroid of each cluster is shown in the first panel along with the member data points. The centroids are close together suggesting small differences between clusters. The second panel shows the genes with the largest discriminatory effects.

Support Vector Machines

A predictive model to identify class membership was developed using support vector machines. Leave one out cross validation was used to develop the model, whereby in each iteration one gene was left out. A simple multi-class linear regression analysis was performed for each iteration and the remaining genes ranked by significance. The model was generated with the 100 most informative genes determined by the iterative procedure. A tuning step was performed to identify the optimal cost hyper-parameter for the support vector. A classification model was then generated and the predictive power assessed. The results are shown in Table 12.1. The overall performance is modest with a misclassification rate internal to the data set of 5.5%.

		Predicted		
		Cluster 1	**Cluster 2**	**Cluster 3**
Observed	**1**	31	0	1
	2	0	74	1
	3	1	4	15

Table 12.1: Classification of clusters using support vector machines

The power of the predictive models can vary with the number of predictive variables. The relationship between the misclassification rate and the number of genes in the cluster prediction model is shown in Figure 12.2. The misclassification rate appears to be stable with rates of 5 to 9% beyond a threshold of at least 5 predictive genes.

Random Forests

Random forests use a different approach where a forest with a specified number of trees are grown. Each tree is a decision tree where each branch is a decision based on a specified number of predictors or classifiers. The classifiers at each branch are selected randomly and the best classification is used at each division. The final nodes of the tree represent the classes that need to be predicted.

Random forests have three major advantages compared to other classification methods. They perform well with complex data sets and produce accurate decision trees with a lower number of classifiers. The process of generation of trees also allows the

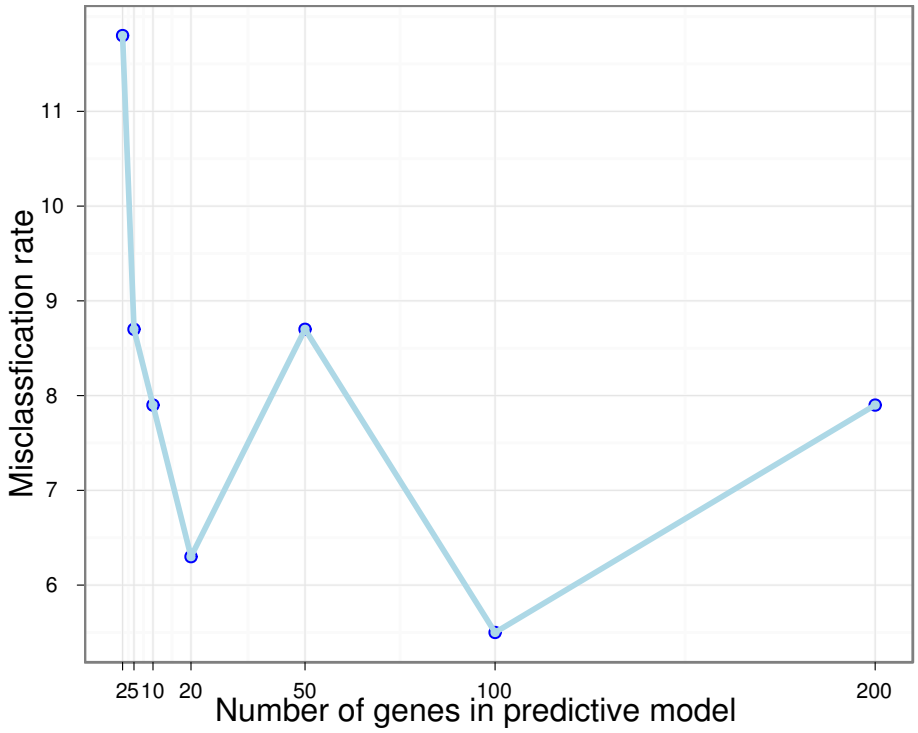


Figure 12.2: Dependency of misclassification rate on the number of predictive genes for the prediction of cluster memberships using support vector machines.

identification of classifiers with the greatest importance. In addition there is no need for internal cross validation with random forests as the bootstrapping process of the algorithm estimates the out of bag (OOB) error during each run. This comes at the cost of the random nature of the process, with some discrepancy between different runs of the algorithm.

The performance of a random forest with a decision tree based on 100 classifiers at each stage is shown in Table 12.2. The overall misclassification rate is 7.9%. The dependence of model performance on the number of classifiers is shown in Figure 12.3. With at least 20 classifiers at each decision node, there is considerable stability of misclassification rates between 8 and 10%.

The final classification using random forests was performed with a random forest consisting of 10,000 trees, with each decision split based on 20 classifier variables. The error estimates during the classification are shown in Figure 12.4. There is stability in the error estimates with the iterations. The overall misclassification rate is about 8.0%. Predicting cluster 3 is the most error prone with a consistent 30% misclassification, an

		Predicted		
		Cluster 1	**Cluster 2**	**Cluster 3**
Observed	**1**	28	3	1
	2	0	75	0
	3	2	4	14

Table 12.2: Classification of clusters using random forests

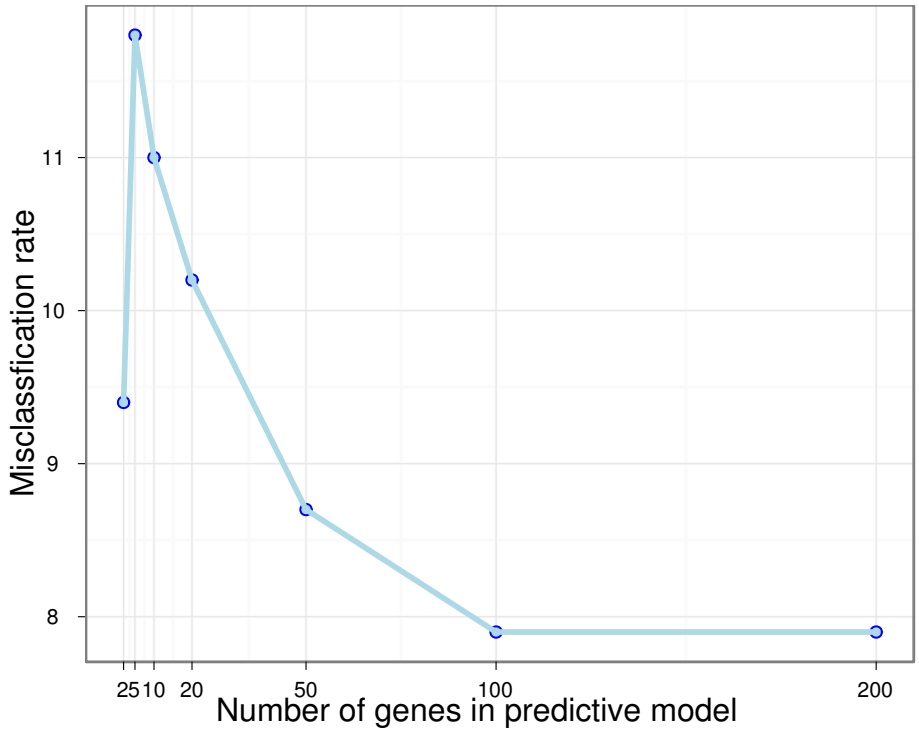


Figure 12.3: Dependency of misclassification rate on the number of predictive genes for the prediction of cluster memberships using random forests.

observation possibly related to the small numbers in this cluster. Predicting cluster 2 is the least error prone with accurate predictions throughout. Error rates with cluster 1 prediction varies between 9.4 and 12.5%.

Two related criteria can be used to identify the important classifiers in random forests. The first uses the mean decrease in prediction accuracy as the result of removing that classifier from the model. The second uses the Gini index, a measure of dispersion, where larger values indicate greater importance. The variables of importance identified using each of these criteria in the cluster prediction model are shown in Figure 12.5. There is good concordance between the two measures of importance. While there are

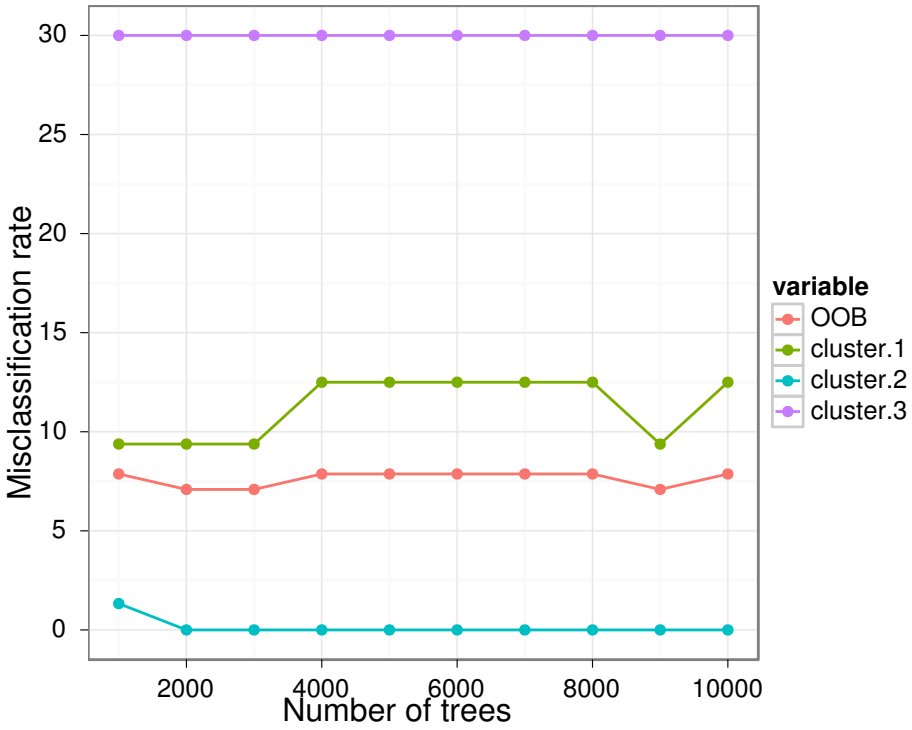


Figure 12.4: Error rates in predicting clusters in the CAP cohort with random forests. OOB = out of bag error rate, an internal validation score in the random forest algorithm.

classifiers of clear biological importance like *IRAK3* (IL1 receptor associated kinase), and classifiers with related functions like *GYG1* (glycogenin) and *EXOSC4* (exosome complex component), the importance measure is purely statistical and gives limited information about the biological relevance.

Diagnostic labels

The clinical differentiation of CAP from FP is straightforward. Therefore, there is no direct clinical value of using genomic signatures to predict the diagnosis. However, the presence of robust predictors for the diagnostic labels, increases the confidence that there are true biological differences between sepsis of different aetiologies.

Principle component analysis

Using regression models, 271 genes were identified that distinguish the two diagnostic cohorts with at least 1.5 fold change in gene expression and an adjusted p value less than

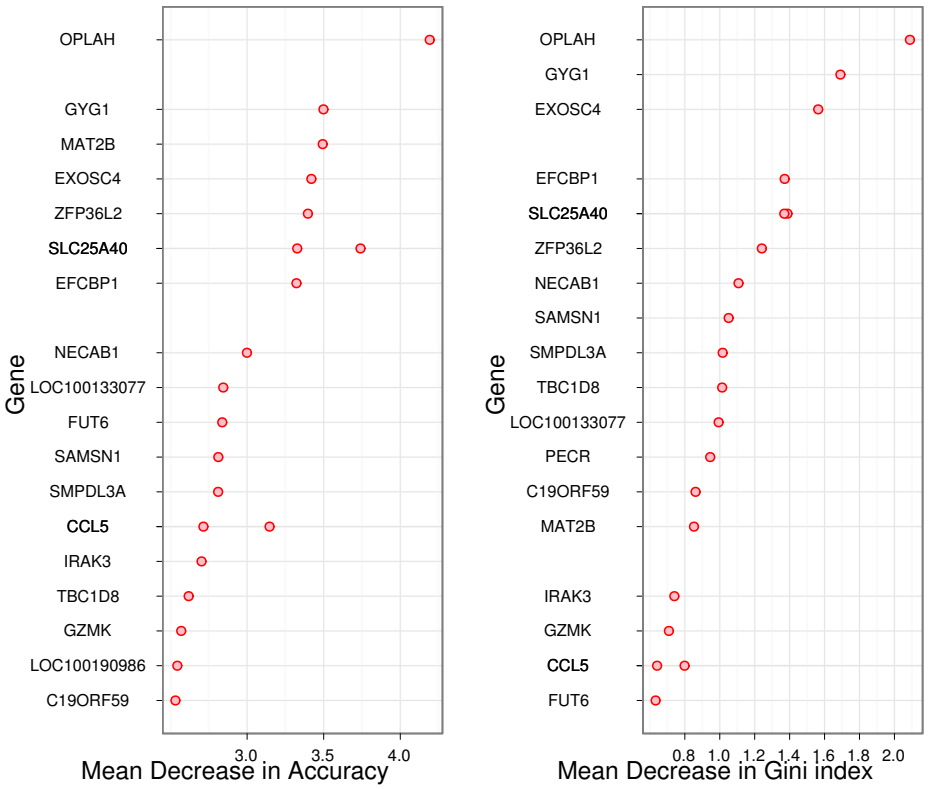


Figure 12.5: Classifier importance in cluster prediction model with random forests

0.05. The principle component analysis of the expression of these genes is shown in 12.6. Most of the variance is accounted for by the first principal component as expected. There is good separation of the centroids of each group, but there is considerable overlap of the individual samples of both the groups.The predictive model can be expected to have reasonable predictive power despite smaller effect sizes of most of the genes (only 33 of the 271 genes have more than two fold change in gene expression). This is in direct contrast to the cluster predictive models. However, similar to the cluster predictive models, the number of important classifiers is small.

Support Vector Machines

A predictive model to identify class membership was developed using support vector machines using the same process described in the previous section. In brief, leave one out cross validation was used to identify the most informative genes. A final classification model was generated after tuning the appropriate hyper-parameters. The results are

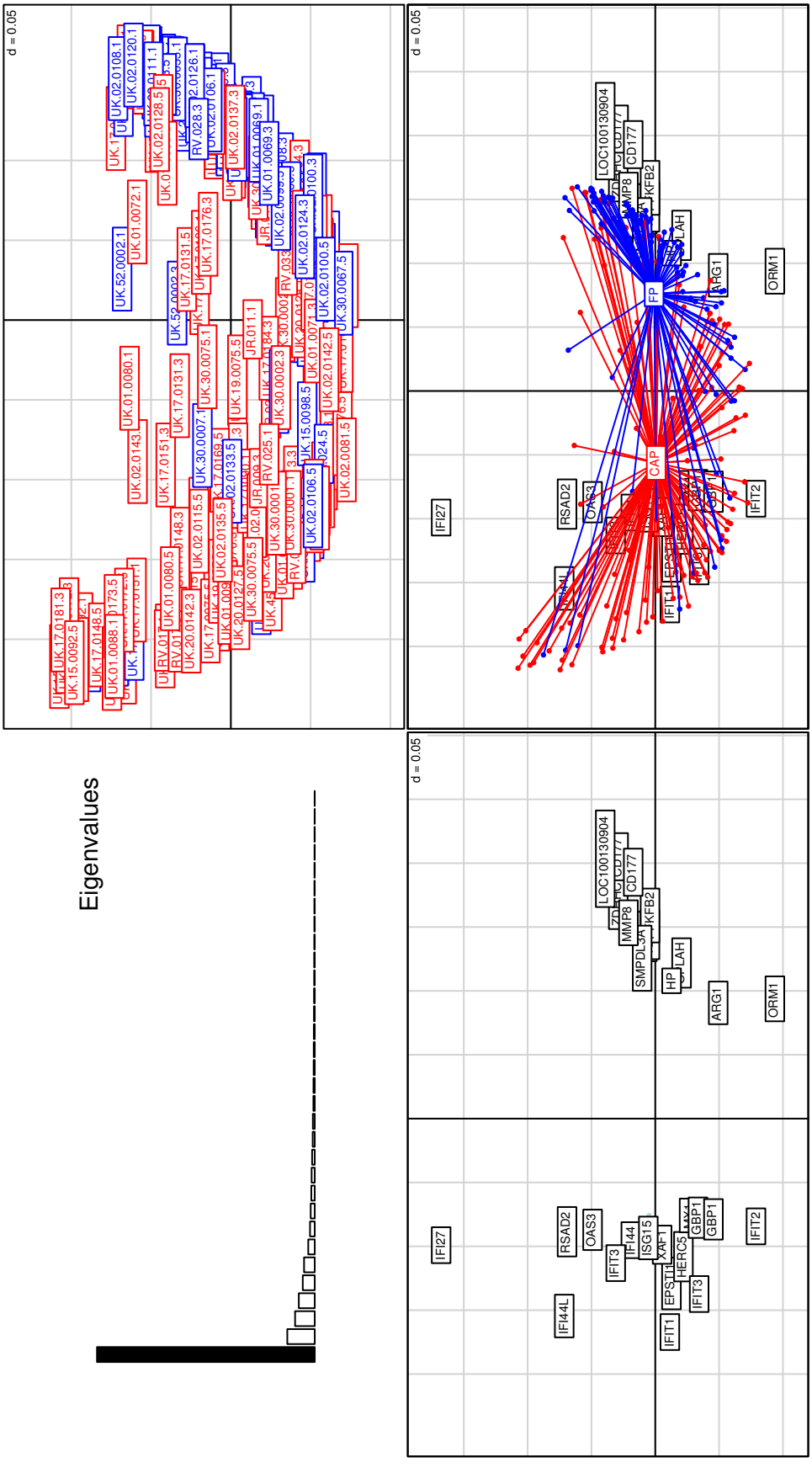


Figure 12.6: PCA based on genes discriminating the diagnoses. Panel 1 shows the eigenvalues, where the first value accounts for most of the variance in the data set. The second panel shows the distribution of the samples based on the expression of these genes. The third panel shows the distribution of the genes. The fourth panel is an overlap of panel two and three and shows the centroids of both clusters, the sample data points and the genes contributing to the cluster discrimination.

shown in Table 12.3 for a model generated with 100 classifiers with at least 1.5 fold change in gene expression. The overall performance is modest with a misclassification rate internal to the data set of 21.3%. The sensitivity is 75.5% and specificity is 81.1%.

		Predicted	
		CAP	**FP**
Observed	**CAP**	103	24
	FP	23	71

Table 12.3: Diagnostic classification using support vector machines

The relationship of the model parameters to the number of genes in the prediction model is shown in Figure 12.7. The sensitivities, specificities, and misclassification rates are stable over different numbers of classifiers. The slight reduction in power observed in the sensitivity and specificity is due to the overlap of samples observed in the principle component analysis. The overall sensitivity and specificity is good suggesting reasonable predictive power. Compared to the cluster identification model, the major gain is in sensitivity while specificity has remained high in both models. This pattern of lack of sensitivity with high specificity is a common feature of prediction of clinical phenotypes using gene expression data.

Random Forests

The performance of a random forest with a decision tree based on 100 classifiers at each stage is shown in Table 12.4. The overall misclassification rate is 22.6%. The dependence of model performance on the number of classifiers is shown in Figure 12.8. The random forest algorithm produces more stable estimates compared to support vector machine with lower classifier numbers.

		Predicted	
		CAP	**FP**
Observed	**CAP**	103	24
	FP	26	68

Table 12.4: Diagnostic classification using random forests

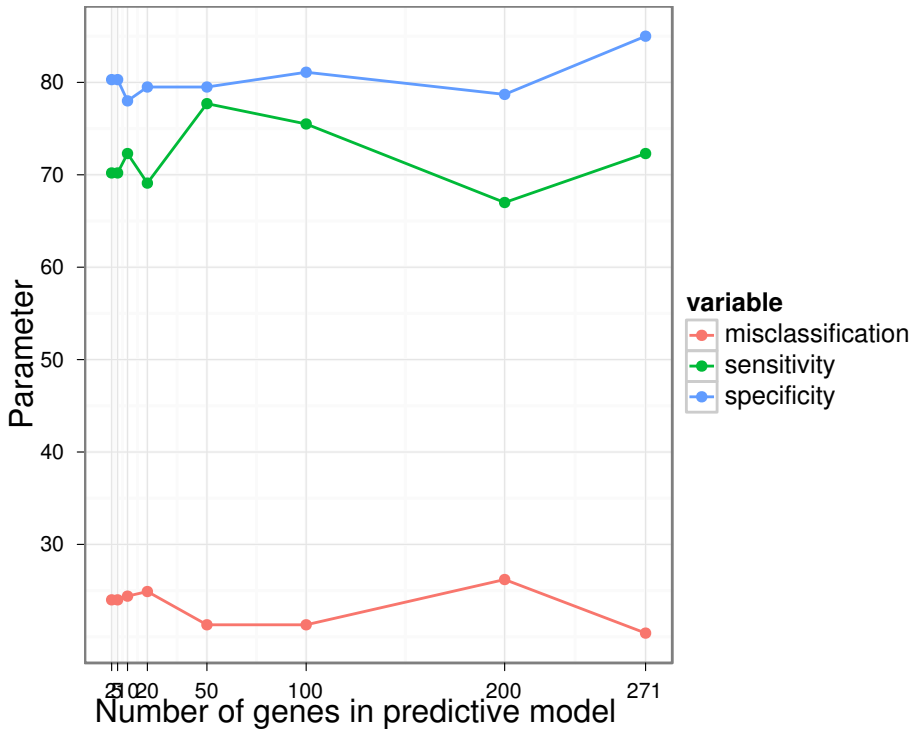


Figure 12.7: Dependency of misclassification rate on the number of predictive genes for the prediction of diagnosis using support vector machines.

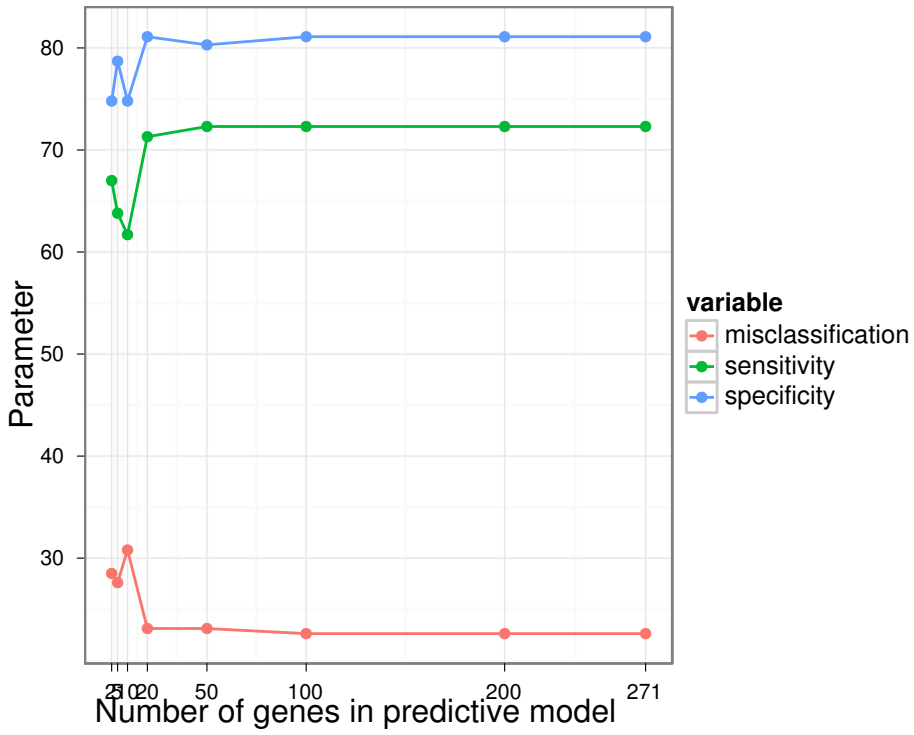


Figure 12.8: Dependency of misclassification rate on the number of predictive genes for the prediction of diagnosis using random forests.

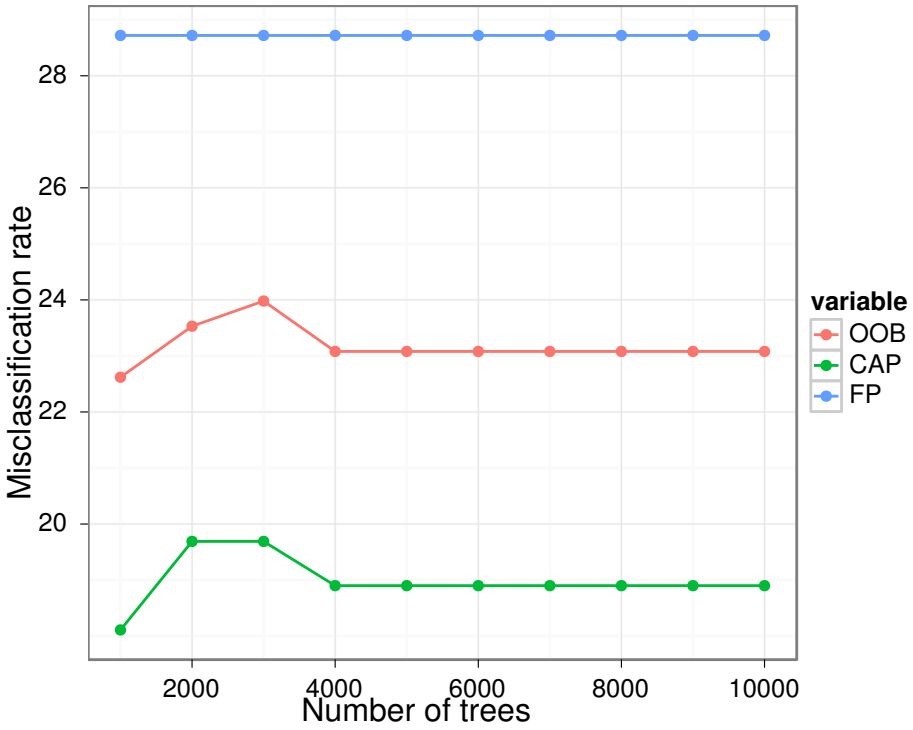


Figure 12.9: Error rates in predicting diagnosis with a random forest model using 10000 trees and 16 classifier variables.

The final classification using random forests was performed with a random forest consisting of 10,000 trees, with each decision split based on 16 classifier variables. The error estimates during the classification are shown in Figure 12.9. There is stability in the error estimates with the iterations. The overall misclassification rate is about 23%. Predicting FP is the most error prone with a consistent 28.72% misclassification. Predicting CAP is the least error prone with a misclassification rate of 18.9%.

The classifiers of importance identified using the mean decrease in accuracy and Gini index in the diagnosis prediction model are shown in Figure 12.10. There is good concordance between the two measures of importance. As with the cluster predictions, while most of the important classifiers are biologically relevant in sepsis, the interpretation of their importance is mainly statistical.

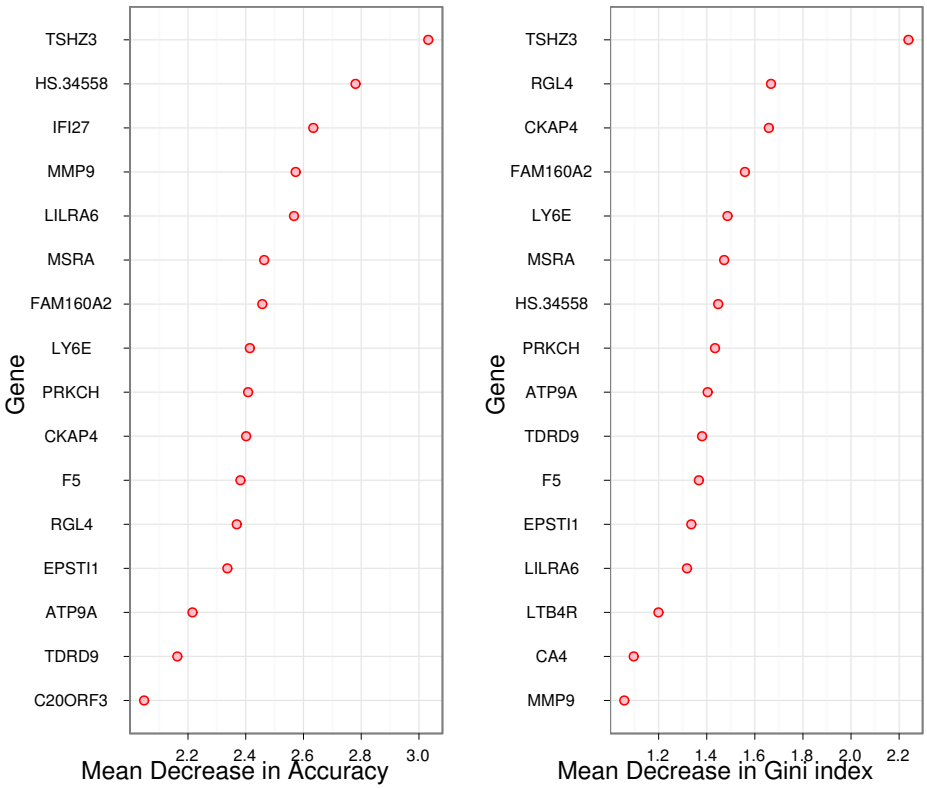


Figure 12.10: Classifier importance in diagnosis prediction model with random forests

12.3 Prognostic predictions

For clinical application, prediction of mortality risk at diagnostic level rather than cluster sub-group level is more useful.

CAP

The overall approach used in predicting mortality risk in community-acquired pneumonia is similar to the approach used in the diagnostic predictions.

Principle component analysis

Using simple linear models, 354 genes were associated with survival differences with at least 1.5 fold change in relative expression levels and an adjusted p value less than 0.05. The regression models use cluster membership as a covariate, so that the genes identified are robust to the CAP sub-groups.

The principle component analysis of the expression of these genes is shown in 12.11. Most of the variance is accounted for by the first principal component. However, there is considerable overlap of the samples in each group along with overlap of the centres of each group. Therefore, survival predictability is expected to be poor, despite strong biological plausibility of the most significant genes. The genes predicting survival show nearly equal variance in both the first and second principal components, suggesting additional unknown contributors to the variance.

Support Vector Machines

A predictive model for mortality risk in CAP was developed using support vector machines with leave one out cross validation. A tuning step was used to optimize the hyperparameters for the support vector machine. The features of a predictive model based on 100 classifiers with at least 1.5 fold change in gene expression and an adjusted *p* value less than 0.05 are shown in Table 12.5. The overall performance is very poor. The model has a 100% sensitivity but with just 7.1% specificity. The misclassification rate is only 10.2%, but nearly all the samples from patients who died are misclassified as alive. The model has extremely limited utility.

		Predicted	
		Dead	**Alive**
Observed	**Alive**	0	113
	Dead	1	13

Table 12.5: SVM classification of survival in CAP

The relationship of the model parameters to the number of genes in the prediction model is shown in Figure 12.12. The model performance is poor over the entire range of classifier numbers. The best performing model includes 50 classifiers, and has a sensitivity of 99.1%, specificity of 35.7%, and a misclassification rate of 7.9%. Paradoxically, increasing the number of classifiers decreases predictive accuracy, suggesting a worse noise-signal profile in the classifiers.

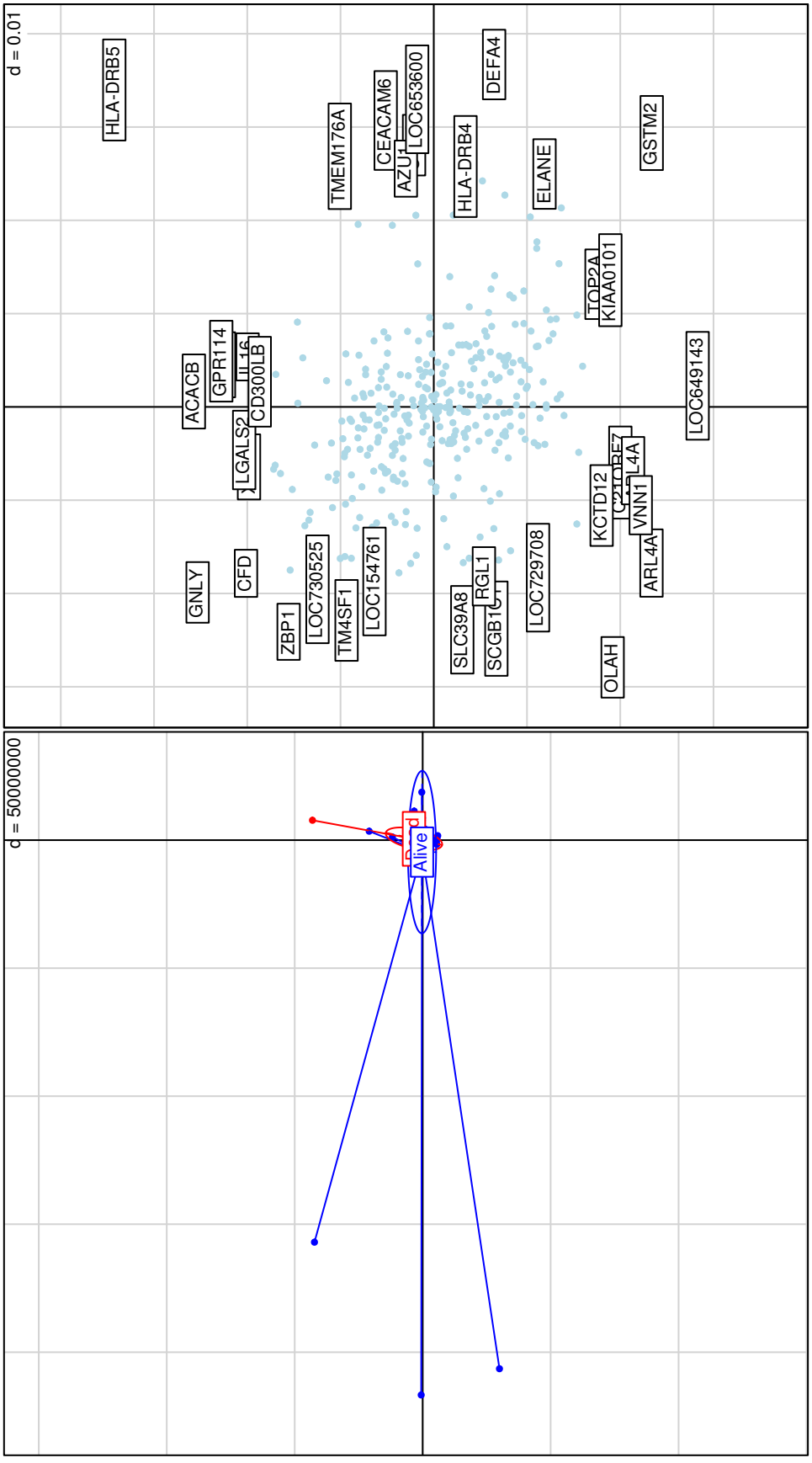


Figure 12.11: PCA based on genes discriminating survival in CAP. Panel 1 shows the centroids of each group (survivors and non-survivors). The centroids and data points overlap considerably. Panel 2 shows the distribution of genes along the first principal component. The dispersion of genes along the first component in small, suggesting low discriminatory power. The dispersion in the second component is as large as the first component, suggesting additional components to the variance.

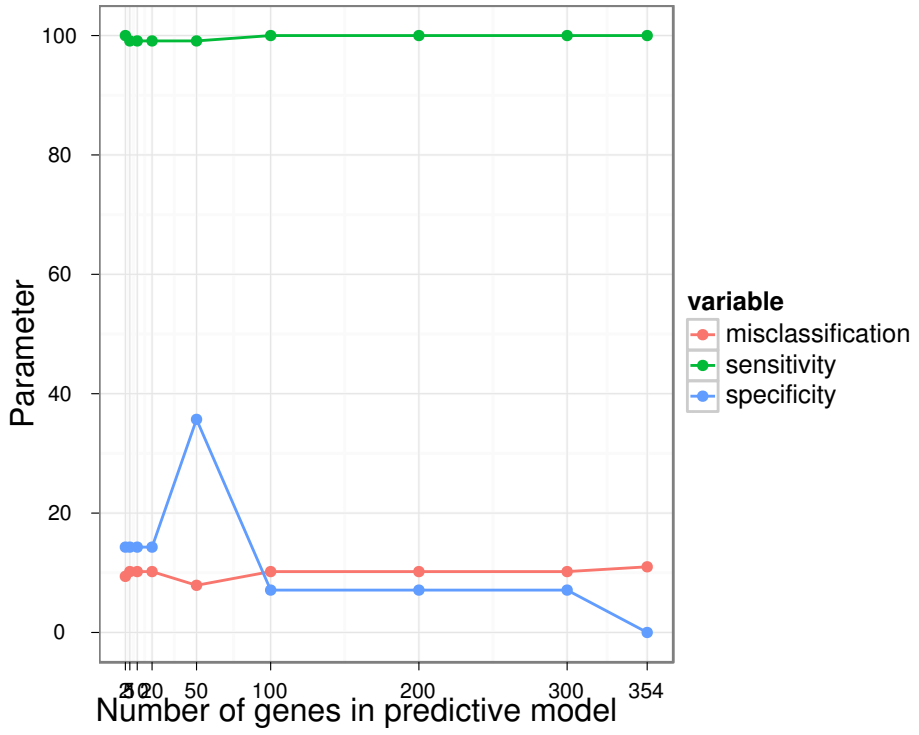


Figure 12.12: Dependency of misclassification rate on the number of predictive genes for the prediction of survival in CAP using support vector machines.

Random Forests

The performance of a random forest with a decision tree based on 100 classifiers at each stage is shown in Table 12.6. As with SVM, the model performance is poor with a misclassification rate of 10.2%. The sensitivity is 99.1% and specificity is 14.3%. The dependence of model performance on the number of classifiers is shown in Figure 12.13. The random forest algorithm produces more stable estimates compared to SVM with lower classifier numbers.

		Predicted	
		Dead	**Alive**
Observed	**Alive**	1	112
	Dead	2	12

Table 12.6: Classification of survival in CAP with random forests

The final classification using random forests was performed with a random forest consisting of 10,000 trees, with each decision split based on 18 classifier variables.

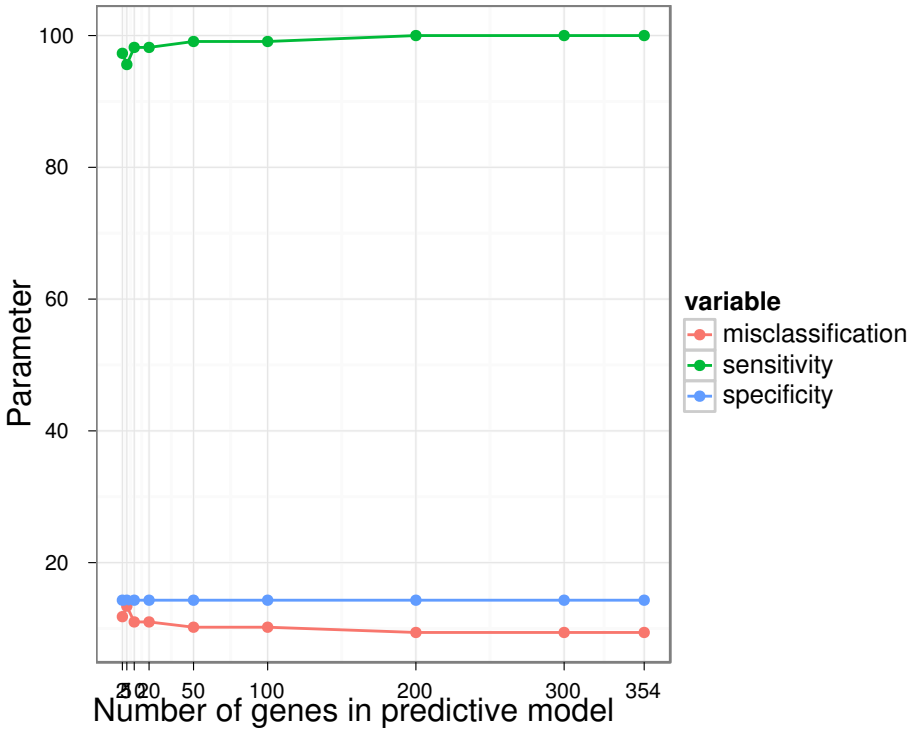


Figure 12.13: Dependency of misclassification rate on the number of predictive genes for the prediction of survival in CAP using random forests.

There is stability in the error estimates with the iterations, with no change in error estimate after the first 1000 trees. The overall misclassification rate is about 9.45%. All survivors are correctly identified, but 12 of the 14 non-survivors are identified incorrectly corresponding to an error rate of 85.71%. The classifiers of importance identified using the mean decrease in accuracy and Gini index in the diagnosis prediction model are shown in Figure 12.14. There is good concordance between the two measures of importance.

FP

The cluster structure in faecal peritonitis was not as convincing as the structure in community-acquired pneumonia. The gene expression signatures of mortality in FP identified in previous chapters, therefore do not include the cluster membership as predictor variables. For consistency, the same approach was used in identifying predictive models.

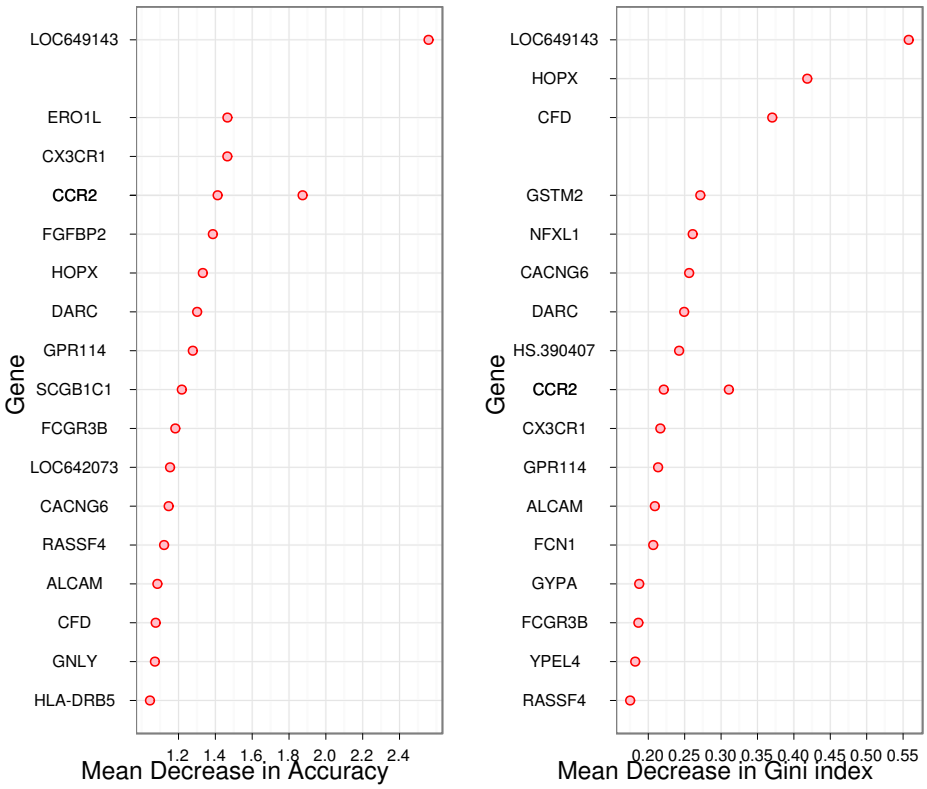


Figure 12.14: Classifier importance in CAP survival prediction model with random forests

Principle component analysis

Using simple linear models, 19 genes were associated with outcome with at least 1.5 fold change in relative expression levels and an adjusted p value less than 0.05. The principle component analysis of the expression of these genes is shown in Figure 12.15. In contrast with the previous comparisons, the majority of the variance is accounted for by the first two principal components. There is considerable overlap between survivors and non-survivors, suggesting poor predictive power in the model. Interestingly, there is evidence for three clusters within the samples, caused predominately by the expression of class I and II MHC genes. This is consistent with the findings in the exploratory analysis.

Support Vector Machines

A predictive model for mortality risk in faecal peritonitis was developed using support vector machines with leave one out cross validation. A tuning step was used to optimize the hyperparameters for the support vector machine. The features of a predictive model

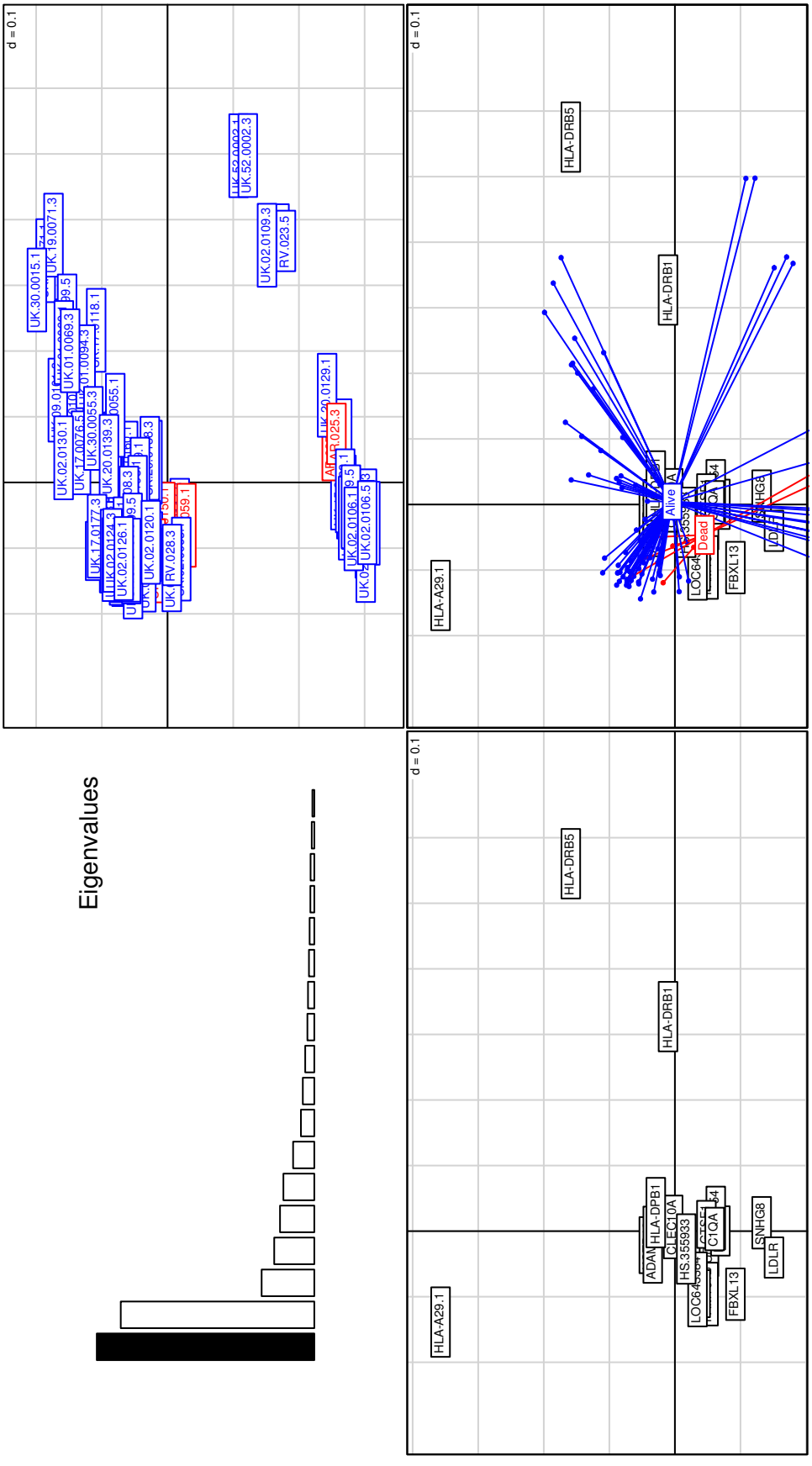


Figure 12.15: PCA based on genes discriminating survival in FP. Panel 1 shows the eigenvalues associated with each principal component. The first two components make significant contributions to the overall variance. Panel 2 shows the distribution of sample data points along the principal components. Panel 3 shows the distribution of the genes along with their relative importance. Panel 4 overlays panels 2 and 3 and shows the centroids of each group, along with the data points, and the genes with the greatest discriminatory power.

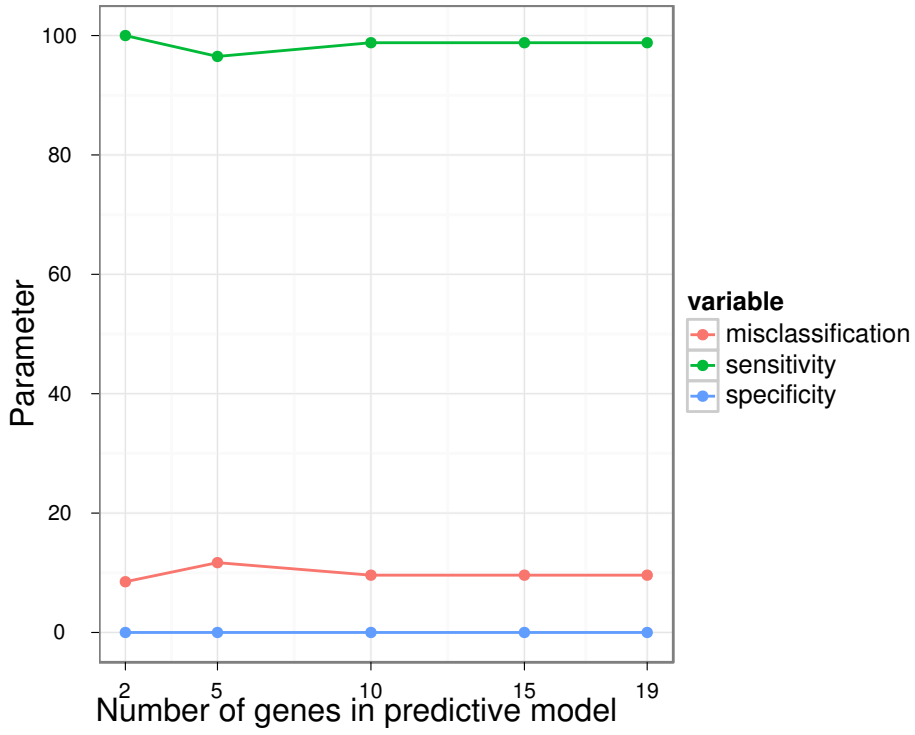


Figure 12.16: Dependency of misclassification rate on the number of predictive genes for the prediction of survival in FP using support vector machines.

based on 19 classifiers with at least 1.5 fold change in gene expression and an adjusted p value less than 0.05 are shown in Table 12.7. The overall performance is as poor as the CAP predictive model. The model has 98.8% sensitivity but a specificity of 0. The misclassification rate is 9.6%, but as with the CAP predictive model, there is no discrimination of non-survivors from survivors.

		Predicted	
		Dead	**Alive**
Observed	**Alive**	1	85
	Dead	0	8

Table 12.7: SVM classification of survival in FP

The relationship of the model parameters to the number of genes in the prediction model is shown in Figure 12.16. The model consistently has no specificity and has no predictive power for any choice of number of classifiers.

Random Forests

The performance of a random forest with a decision tree based on 19 classifiers at each stage is shown in Table 12.8. Compared to SVM, the model performance is equally poor. The model does not predict any mortality, and the conventional model parameters of sensitivity, specificity and misclassification rate are irrelevant. The dependence of model performance on the number of classifiers is shown in Figure 12.17. The random forest algorithm produces slightly better estimates compared to support vector machine with lower classifier numbers. However, given the low mortality rates, the interpretation of the results is difficult and probably irrelevant.

		Predicted	
		Dead	**Alive**
Observed	**Alive**	0	86
	Dead	0	8

Table 12.8: Dependency of misclassification rate on the number of predictive genes for the prediction of survival in FP using random forests.

12.4 Summary

The development of predictive models has diagnostic and prognostic applications in clinical practice. In addition, a strong predictive model differentiates robust statistical associations from weak associations.

The clusters identified in community-acquired pneumonia are associated with relevant clinical outcomes like time to death. Using support vector machines, a misclassification rate of 6% can be achieved with approximately 20 classifiers. Cluster 2 is most accurately identified, while cluster 3 membership is predicted with least accuracy. Using random forests, a misclassification rate of 8% is associated with 20 classifiers. Cluster 2 has the best predictive accuracy, followed by cluster 1 and cluster 3. The most important classifiers can be identified with the random forest approach and includes a number of biologically relevant genes.

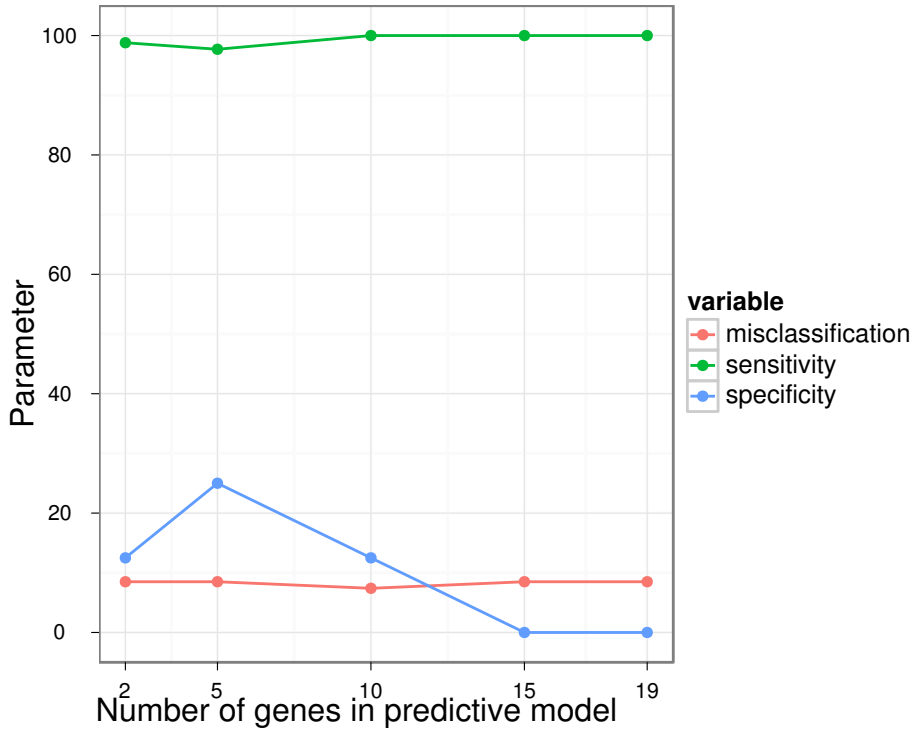


Figure 12.17: Dependency of misclassification rate on the number of predictive genes for the prediction of survival in FP using random forests.

The differentiation of CAP from FP can be achieved with 20 classifiers, with a misclassification rate of 21% and with sensitivity and specificity greater than 75%. Both the support vector machine and random forest approaches have similar accuracies. FP can be predicted with approximately 70% accuracy, while CAP can be predicted with 80% accuracy.

The prediction of survival in CAP and FP is far less successful. The predictors for CAP are robust to cluster memberships, while the predictors for FP do not include cluster membership as a predictive variable. Both support vector machine and random forest approaches perform poorly in both CAP and FP survival predictions. The misclassification rate is about 10%, but the rate is spuriously low and is related to the low overall mortality rate in the samples. Both approaches have extremely high specificities and extremely low sensitivities in each group, suggesting complete lack of discriminatory power.

Part IV

Sepsis in Context

13

Genomics of Cardiac Surgery

13.1 Background

A severe systemic inflammatory response (SIRS) is seen not only in severe sepsis and septic shock, but can also result from traumatic or surgical tissue damage without an infectious component. The relationship between the gene expression changes in these two kinds of responses can provide insights into the biology of inflammation and immunity.

Open heart surgery involves the cessation of cardiopulmonary activity when the function of the heart and lungs is taken over by roller pumps and an extra-corporeal circuit that includes an oxygenator. The exposure of immune cells to the foreign surface triggers a severe systemic inflammatory response. Cardiac surgery provides a unique model of severe, relatively sterile, systemic inflammatory response as a result of a transient, but well defined stimulus in subjects who are normal except for cardiac disease. Rapid clinical recovery with resolution of inflammatory parameters follows surgery, and serial changes in inflammation over time can be compared to the baseline.

This analysis is restricted to the identification of the genomic signatures of the sterile SIRS induced by cardiac surgery and cardiopulmonary bypass compared to the genomic signatures of sepsis.

13.2 Materials and Methods

Recruitment, collection of samples, and part of the RNA processing was performed by Dr. Eduardo Svoren. The sample collection, processing, microarray hybridization, and the analysis mirror those from the sepsis study.

Patients scheduled to undergo cardiac surgery were recruited for the study after informed consent and ethical approval (n = 16 for validation cohort and n = 10 for derivation cohort). Patients scheduled for minimally invasive procedures and complex, multiple procedures were excluded. Samples were collected at three time points: a baseline sample just before induction of anaesthesia, at the end of the procedure when the patient was transferred to the ITU and 24 hours after the procedure. These samples are referred to as baseline, post op and day 1 samples.

Samples collected included Buffy coats for DNA extraction, plasma, urine, and samples for mRNA extraction using the Ambion LeukoLOCK kit. Sample processing and handling was as described previously for the sepsis samples. Samples were hybridised to the Illumina HT-12 array (validation data set) and Illumina WGA-6 array (derivation data set).

To identify the gene expression response to cardiac surgery a three way comparison was performed between the three time points. A simple linear model was used to compare expression between groups of interest. The validation data was independently analysed using the same methods. All analyses were performed at the probe level but comparison between the validation and derivation cohorts was performed using gene level data because of lack of direct comparability between the two different platforms.

GSEA was used to identify biologically relevant pathways. Significance levels were determined by permutation as described earlier.

Genes predictive of the inflammatory state were derived using support vector machines. Panels of genes predicting the three time points were identified.

To identify the differences in gene expression patterns between SIRS and sepsis, the patterns of gene expression in the cardiac data were compared to those in the sepsis data set. For these comparisons uncorrected data sets were used to maximise comparability and clinical relevance. Only the derivation cohorts were used for comparisons.

- The inflammatory patterns identified from the non-supervised hierarchical clustering of sepsis samples were compared to each of the two post operative samples.
- Sepsis samples collected at ITU admission were compared to the day 1 samples from the cardiac cohort.
- The pattern of gene expression in the most significant pathways distinguishing the post operative samples was examined in the sepsis samples.

13.3 Genomics of cardiac surgery

45 samples from 15 patients and 30 samples from 10 patients were analysed in the derivation and the validation cohorts respectively. Three samples from one patient in the derivation cohort failed quality checks and were removed from the dataset. Quality control assessment and normalization were performed as described for the sepsis data set. After normalization and filtration based on detection threshold 18,245 and 23,044 probes were retained for analysis in the derivation and validation cohort respectively.

Exploratory Analysis

Hierarchical agglomerative clustering was performed on the principal components on the most variant probes (defined as probes with co-efficient of variance greater than 0.9). Clustering was performed independently on the validation and the derivation cohorts.

The results of clustering on the principal components of the derivation cohort show three well defined clusters (Figure 13.1) which correspond to the time of sampling. Clustering was performed on 1825 probes corresponding to 1601 genes. The maximal difference (first principal component) differentiates the baseline samples from the post operative samples while the second principal component differentiates the two post surgical samples. 58.95% of the total variance is explained by the first two principal components. The post op samples show greater variance than the baseline and day 1 samples. The day 1 samples show a trend towards a normalising immune response despite the similarities with the post op samples.

The results of clustering on the principal components of the validation cohort derived from 2305 probes is shown in Supplementary Figure 3. The distribution of clusters is similar to the derivation cohort but there is a lack of resolution between the post surgical samples, with a considerable overlap reflecting inadequate sample size. 55.95% of the overall variance is explained by the first two principal components.

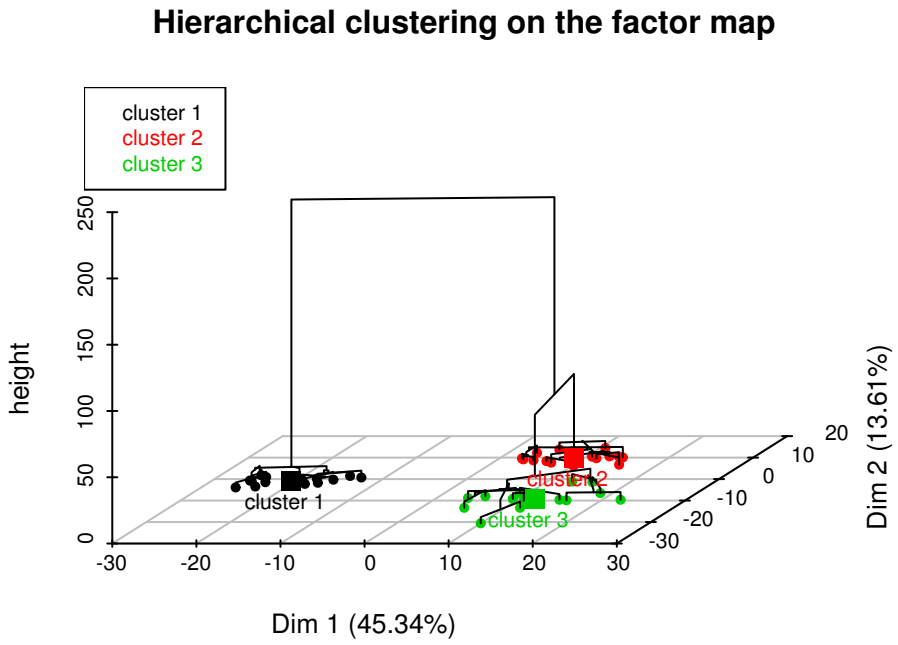


Figure 13.1: Hierarchical agglomerative clustering on the principal components for the derivation cohort in the cardiac dataset. There is clear separation of three clusters corresponding to the time of sampling. The first principal component corresponds to the operation status, while the second component discriminates the two post operative samples.

Gene expression signatures of sterile inflammation

The differences between the post surgical samples and the controls show a snapshot of the inflammatory state in the immediate aftermath of the operation and the first postoperative day (Figure 13.2). The differences between the post-op sample and control and between the day 1 and the post op sample show the evolution of the inflammatory state over time.

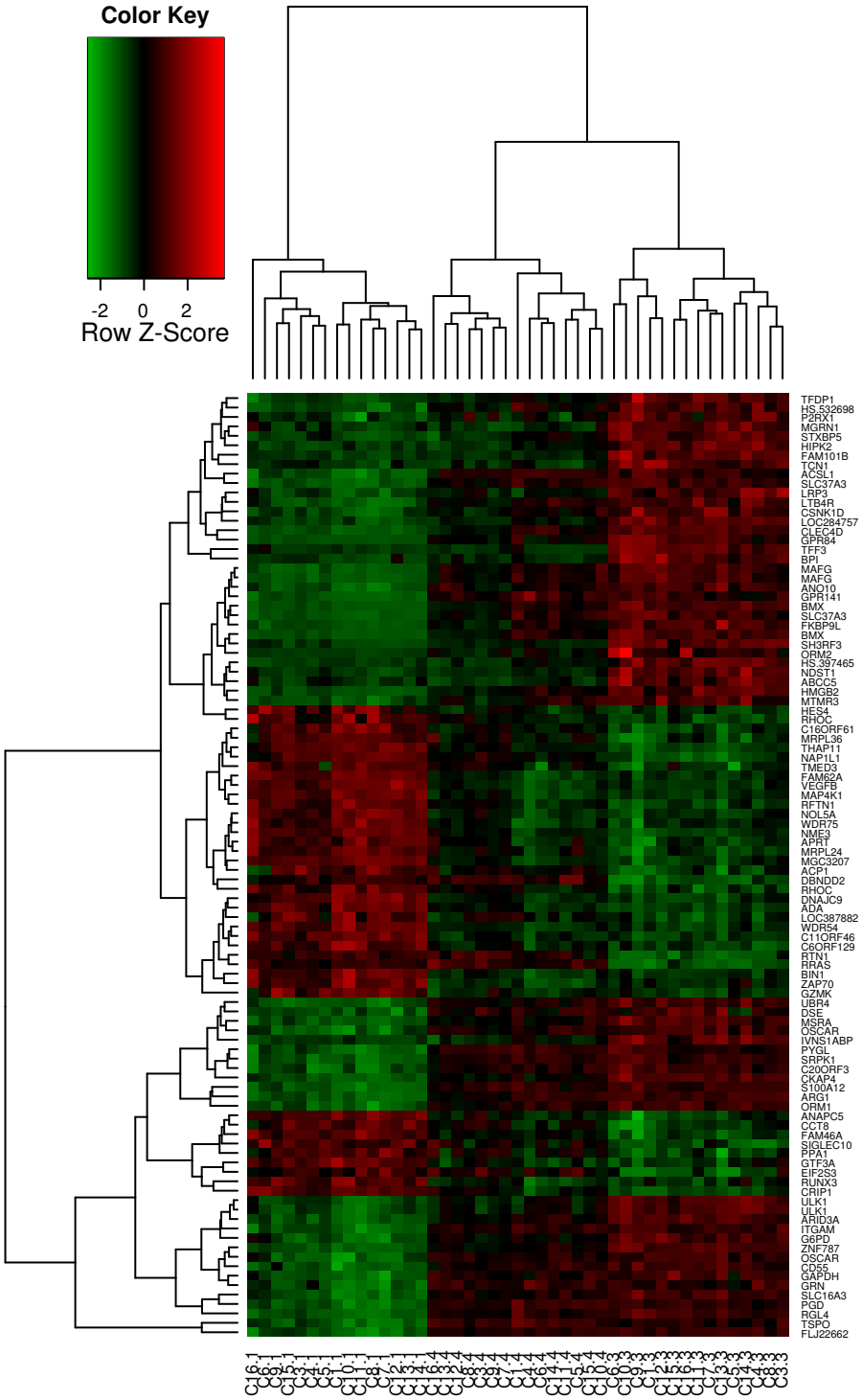


Figure 13.2: Differential gene expression after cardiac surgery. Each row corresponds to a probe and each column to a sample. Samples cluster by time of sampling. The genes contributing to the signature show clear separation into modules based on the similarity in expression patterns. Compared to the baseline (leftmost column cluster), the immediate post operative state is characterised by large changes in gene expression, both induction and repression (rightmost column). There is relative resolution of the response in the first postoperative day (central column).

Differences in gene expression were estimated with robust linear models resistant to the effect of outliers. Probes with at least 1.5 fold change in expression and a FDR of 0.05 were considered significant. The stress of cardiac surgery resulted in altered gene expression in 1260 probes in the immediate period (immediate changes) and in 1041 probes on day 1 (late changes). The evolution of responses from the immediate to late (resolution) was associated with changes in 307 probes (Table 13.1). The pattern of gene expression suggests reprogramming involving both induction and repression of gene expression.

Epoch	Counts	induced	repressed	Total
Immediate	Probes	603	657	1260
	Genes	525	594	1119
Late	Probes	558	483	1041
	Genes	470	436	906
Resolution	Probes	168	139	307
	Genes	147	124	271

Table 13.1: Number of probes/genes with at least 1.5 fold change difference in expression and a FDR less than 0.05 in the cardiac cohort. Immediate = Changes in the immediate postoperative period compared to the baseline, Late = Changes in the first postoperative day compared to the baseline, Resolution = Changes in the first postoperative day compared to the immediate postoperative period.

The distribution of effect sizes for all three comparisons in shown in Figure 13.3. Approximately 24% of the probes have at least 2 fold change in expression in the immediate and late response. During the evolution of the immediate to the late response 21% of the probes display a 2 fold change in expression.

The immediate and late responses following cardiac surgery each have a unique component and a shared component (Figure 13.4). 283 of the induced genes are common to both responses (53% and 60% of the total complement of induced genes in the immediate and late responses respectively), while 340 genes are repressed in both responses (57% and 79% of the total complement of repressed genes in the immediate and late responses respectively). Four genes (*CD24* - B cell membrane sialoglycoprotein, and the neutrophil defensins *DEFA1B*, *DEFA1*, *DEFA3*) that are significantly induced during the immediate phase are significantly repressed in the late phase (Figure 13.5).

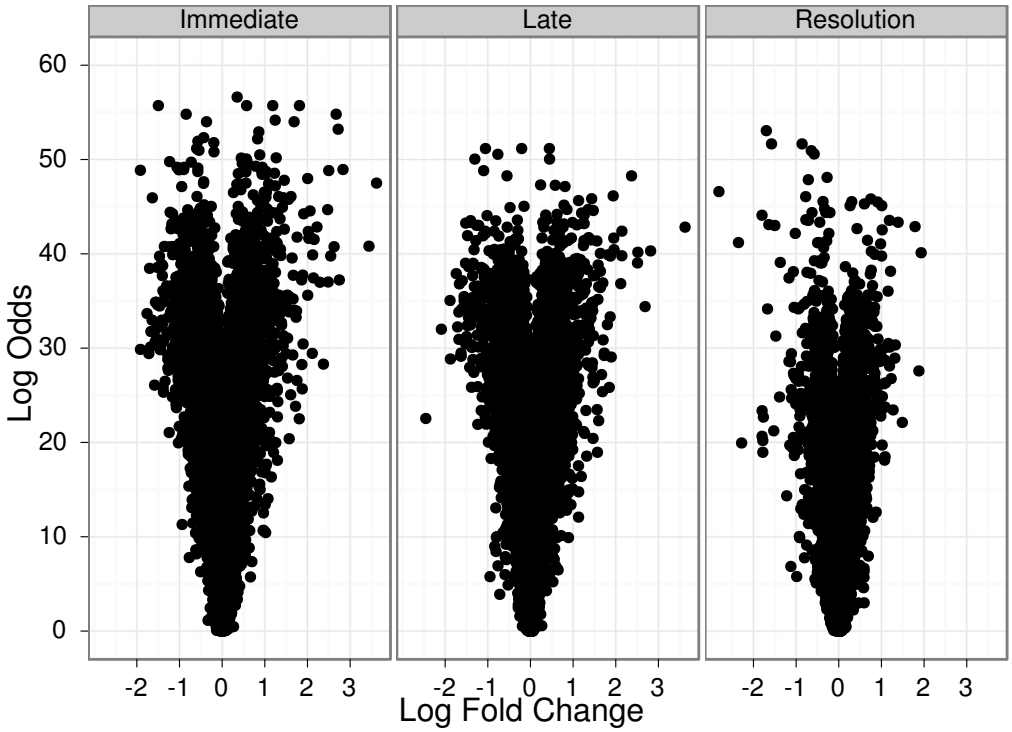


Figure 13.3: Effect size and significance of gene expression following cardiac surgery. A number of probes show greater than 1.5 fold change in the mean expression. Compared to the sepsis data set, more probes show greater statistical significance for similar changes in mean expression.

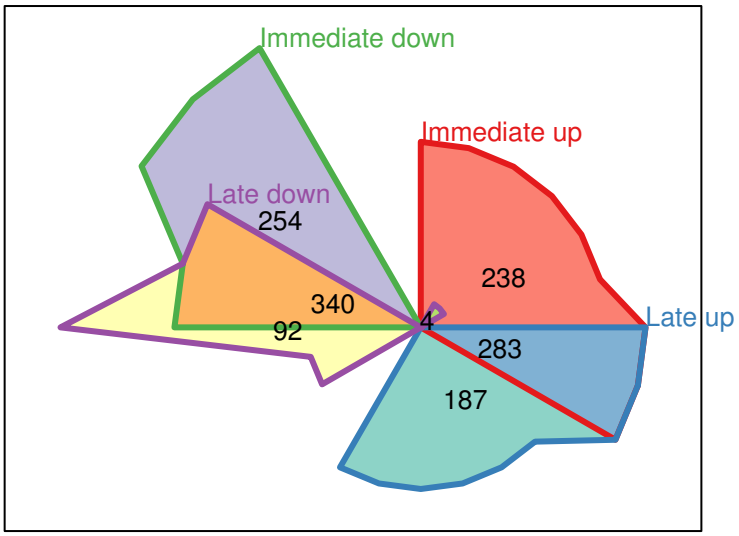


Figure 13.4: The shared and unique components of immediate and late responses to cardiac surgery. The area enclosed by each coloured shape is proportional to the number of genes showing significant difference in expression at that time point when compared to the baseline. 283 genes are induced, while 340 are repressed at both post-operative time points when compared to the baseline. A large proportion of the response is unique to each time point. Four genes that are induced in the immediate postoperative period, are repressed in the first postoperative day.

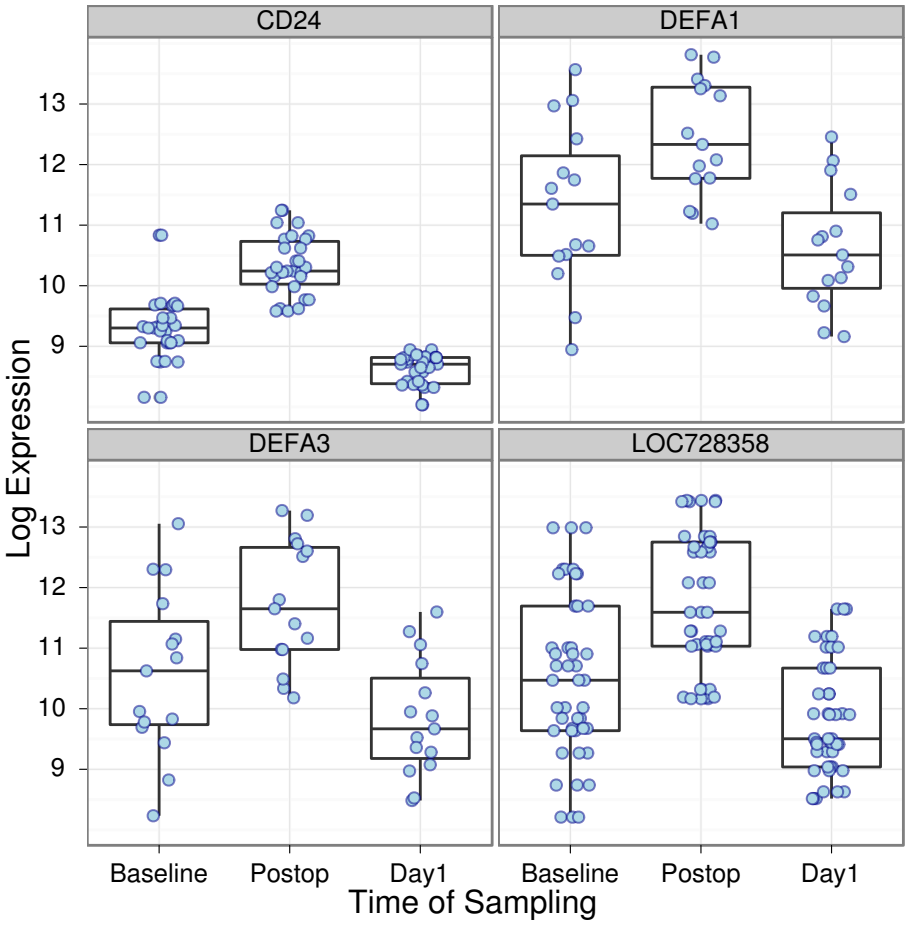


Figure 13.5: Genes with significant differences in expression compared to baseline and with opposite regulation at each time point. *DEFA1* (Chr 8) neutrophil defensin α1, *DEFA3* (Chr 8) neutrophil defensin α3, *CD24* (Chr 6) B cell membrane sialoglycoprotein, *DEFA1B* (Chr 8) neutrophil defensin 1β

Hierarchical clustering of the probes that define the immediate response shows the presence of four clusters that explain more than 95% of the variance (Figure 13.6). While the first principal component accounts for the cluster structure the second principal component maps to the direction of change in expression. The first principal component explains more than 87% of the variance suggesting that the cluster similarity is determined largely by features other than the direction of change in expression. The clusters are summarised in Table 13.2

Hierarchical clustering of the probes that define the late response shows three clusters that also explain more than 95% of the variance (Figure 13.7). As with the immediate response the second component corresponds to the change in direction of expression. The clusters are summarised in Table 13.3

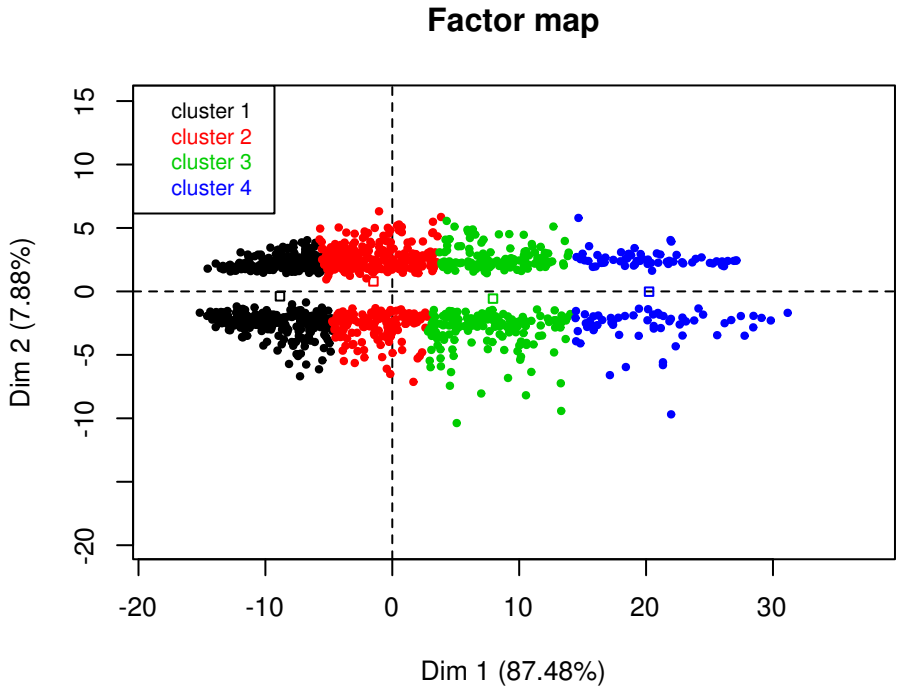


Figure 13.6: Hierarchical agglomerative clustering of the principal components of the probes that are associated with the immediate response to cardiac surgery. The second principal component corresponds to the direction of change in expression but only accounts for 7% of the variance.

Cluster	Total	Up	Down	Functional Annotation
1	451	250	201	Immunoglobulin receptor (n = 19, ES = 3.79) Lectin type receptors (n = 6, ES = 3.55) Transmembrane ion transporter (n = 11, ES = 3.16) G-protein leukotriene, eicosanoid receptors (n = 16, ES = 2.73)
2	417	142	275	Mitochondrial membrane, respiratory chain proteins (n = 3, ES = 3.44)
3	270	152	118	Ribosomal proteins (n = 17, ES = 7.14) Eukaryotic translation initiation factor (n = 4, ES = 4.18) Lipocalin, IL receptor antagonists (n = 6, ES = 2.09)
4	122	59	73	Ribosomal proteins (n = 24, ES = 18.6) Eukaryotic translation elongation factor (n = 4, ES = 4.43)

Table 13.2: Biological relevance of the genes clusters associated with the immediate response. Relevant pathways were identified using the DAVID functional annotation tool with medium stringency classification. Only pathways with an enrichment score greater than 2 are listed. n = number of genes, ES = enrichment score

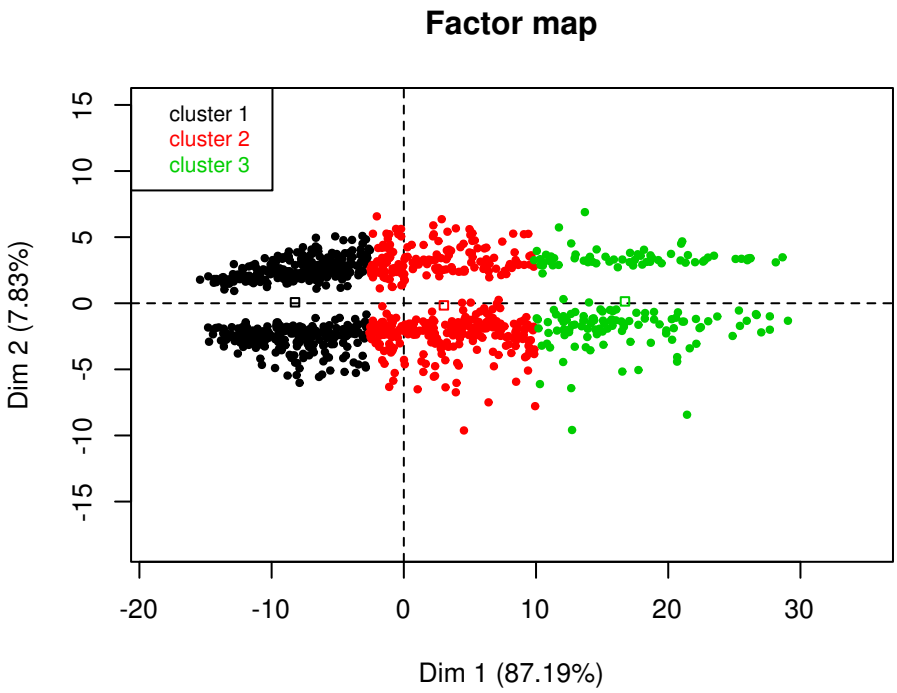


Figure 13.7: Hierarchical agglomerative clustering of the principal components of the probes that are associated with the late response. As with the immediate response, the second principal component corresponds to the direction of change in expression but only accounts for 7% of the variance.

Cluster	Total	Up	Down	Functional Annotation
1	493	223	260	Ig and IL receptors (n = 17, ES = 3.5) Ras/guanylate proteins (n = 5, ES = 2.45) Transmembrane proteins with GRAM domain (n = 8, ES = 2.27) Lectin receptors (n = 5, ES = 2.2) Serine/Threonine kinases (n = 10, ES = 2.03) G protein coupled receptors (n = 10, ES = 2.0)
2	373	219	154	Ribosomal protein (n = 8, ES = 3.02) Transmembrane Ig receptors (n = 18, ES = 2.91)
3	175	106	69	Ribosomal proteins (n = 18, ES = 11.69)

Table 13.3: Biological relevance of clusters of genes associated with the late response to cardiac surgery. Relevant pathways were identified using the DAVID functional annotation tool with medium stringency classification. Only pathways with an enrichment score greater than 2 are listed.

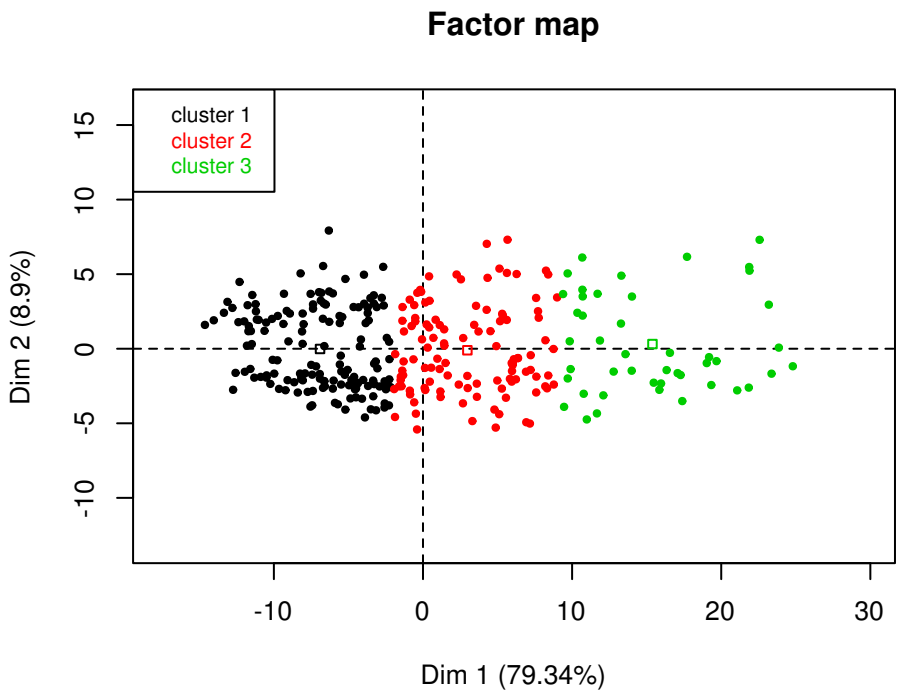


Figure 13.8: Hierarchical agglomerative clustering of the principal components of the probes that differ between the immediate and late response to cardiac surgery. The first principal component has a marginal decrease in explanatory power compared to earlier clusters, but the significance is uncertain.

Cluster	Total	Up	Down	Functional Annotation
1	149	77	72	Lectin receptor family (n = 5, ES = 2.38)
2	113	62	51	
3	45	29	16	

Table 13.4: Biological relevance of gene clusters associated with resolution of immediate response to the late response. Relevant pathways were identified using the DAVID functional annotation tool with medium stringency classification. Only pathways with an enrichment score greater than 2 are listed. Cluster 2 is associated with plausible pathways (Histone clusters) which are below the ES threshold. Cluster 3 is not associated with any known pathways.

Probes associated with the conversion of the immediate response to late response cluster into 3 groups, and are summarised in Table 13.4.

The biological basis of sterile inflammation

The significance of genes differentially expressed between contrasts of interest including those that did not achieve the expression and FDR thresholds was explored with pathway

analysis using GSEA. Analysis was restricted to approximately 2700 gene sets relating to genomic position, KEGG, Reactome and Gene Ontology, derived from the Molecular Signature database.

The serial changes in biological pathways over the time period are shown in Figure 13.9. The gene set enrichment analysis uses all genes weighted by their significance and shares some pathways with the DAVID functional annotation which uses only the most significant genes. In particular, the GSEA approach confirms the importance of ribosomal proteins and mitochondrial membrane proteins in the immune response. The GSEA analysis also captures the evolution of the immune response over time; while the mitochondrial inner membrane activity is important in the immediate response, the mitochondrial outer membrane activity becomes more important as the immediate response evolves into the late response. There are also some association with clinically observable events such as the relationship of inflammation to insulin synthesis and glycogen metabolism.

Leading edge analysis was used to identify the genes determining the importance of biological pathways. The analysis was restricted to some representative pathways of biological and clinical interest.

Mitochondrial function

Both the inner and outer membrane proteins play a role in the sterile inflammatory response.

The mitochondrial inner membrane gene set is significantly different between baseline and the immediate postop period. 31 genes constitute the leading edge of this pathway. Some of these genes are shown in Supplementary Figure 4. The genes show varying levels of constitutive expression, but have a low variance suggesting tight regulatory control. All genes are repressed in the immediate post-operative period, with some genes returning to normal and others persistently repressed on the first post-operative day.

The mitochondrial outer membrane gene set is significantly different between the immediate post op and first post operative day. The significance is due to the expression

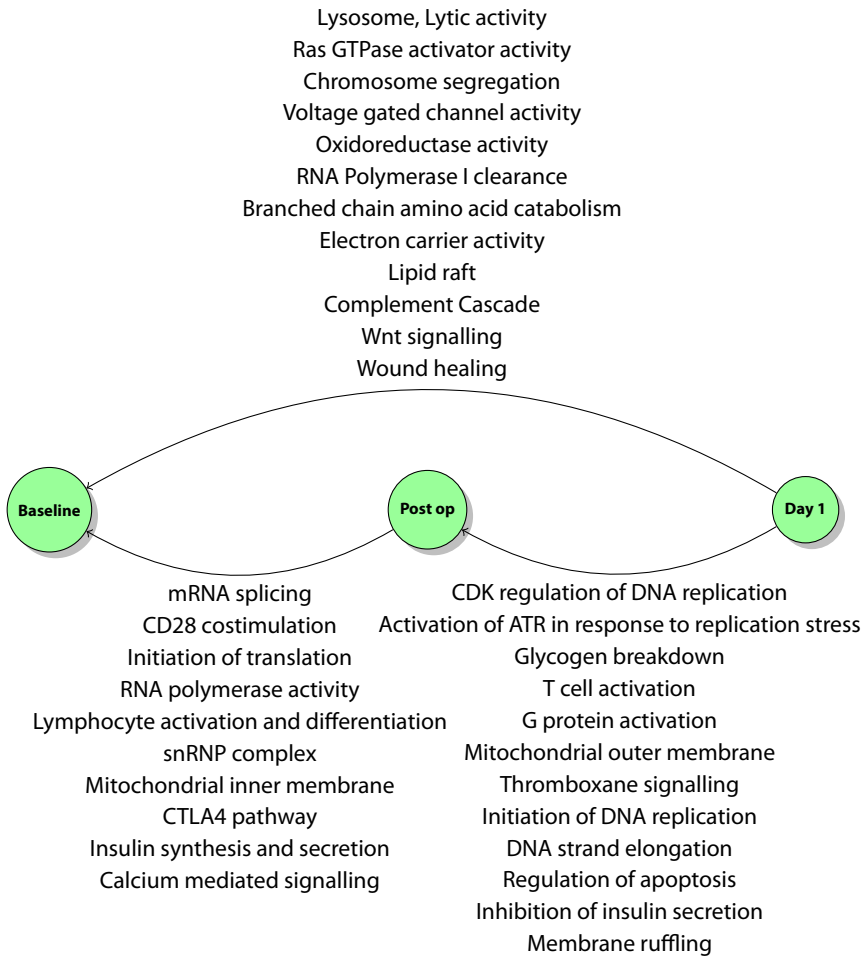


Figure 13.9: Pathways associated with cardiac surgery. All pathways shown have a FDR < 0.15, determined by permutation involving phenotype switching and estimation of mean enrichment scores.

of just 3 genes that constitute the leading edge (Supplementary Figure 5). Compared to the immediate postop period all the genes are repressed but when compared to the baseline each gene has a different pattern of expression.

Insulin and glycogen metabolism

Insulin resistance and disorders of glucose metabolism are common after surgical and septic insults. 62 genes constitute the leading edge panel of the Reactome insulin synthesis and secretion pathway (Supplementary Figure 6). 60 of these genes are ribosomal proteins. The remaining two, *VAMP2* and *FAU*, encode fusion proteins related to vesicle fusion with the cell membrane.

Complement cascade

The complement cascade is differentially regulated on day 1 compared to both the baseline and the immediate post operative period. The difference is attributable to just two genes, both of which show an initial transient suppression followed by a large induction (Supplementary Figure 7).

Regulation of apoptosis

Regulation of apoptosis is a key mechanism in the regulation of the immune response. 20 genes constitute the leading edge subset of the Reactome apoptosis pathway. A selection of these genes is shown in Supplementary Figure 8.

Predicting the sterile inflammatory response

The primary aim of prediction was to identify a gene expression signature that differentiates the postop from the day 1 samples, thereby delineating the inflammatory state. A predictive model was developed using random forests, that uses selection methods appropriate for larger data sets with more predictors than observations. Initially a full model was constructed by growing 10,000 trees, including both the baseline and the inflammatory state samples. The model was validated using the validation data set. Gene level analysis, restricted to genes present in both data sets was performed. Multi-class ROC analysis of this model had an AUC of 89.1% with a 16.7% misclassification rate when applied to the validation data set.

13.4 Relation to sepsis

Relationship to the septic inflammatory patterns

To allow direct comparison between the sepsis and cardiac data sets, the combined data set was normalized together using the approach described in earlier chapters. After filtration for a detection threshold p of 0.05, 29,409 probes were retained for analysis.

To identify the common patterns of genomic response in sterile and septic inflammation, the sepsis samples were compared to the cardiac samples. Independent comparisons were made to each time point of the cardiac sample.

Immediate Response

With a 1.5 fold change in mean expression threshold and a FDR of 0.05, 1263 probes were differentially expressed between the sepsis and the immediate post-operative samples, of which 672 probes were repressed. These correspond to 1129 genes, of which 534 were induced and 596 were repressed.

The genes with the largest effect sizes are listed in Table 13.5. Sepsis is associated with induction of *TLR5* and adhesion molecules but repression of *IL8*. Other interesting hits include the 2 fold induction of CARD domain containing caspase *NLRC4* and the 3 fold induction of protein S *PROS1* in sepsis.

Gene	Chr	Role
IFITM3	11	interferon induced transmembrane protein 3
ELANE	19	elastase, neutrophil expressed
ZDHHC19	3	zinc finger, DHHC-type containing 19
OLFM4	13	olfactomedin 4
TLR5	1	toll-like receptor 5
CEACAM6	19	carcinoembryonic antigen-related cell adhesion molecule 6
ITGA2B	17	integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)
CDK5RAP2	9	CDK5 regulatory subunit associated protein 2
CCL4L1	17	chemokine (C-C motif) ligand 4-like 1
IL8	4	interleukin 8

Table 13.5: Genes with the largest differences in expression in sepsis when compared to the immediate response to cardiac surgery, p 0.01 and 4 fold change in expression

Gene sets significantly enriched for the difference between sepsis and immediate post-op genomic responses were identified using GSEA. The significance of the enriched gene sets was determined by permutation analysis involving phenotype switching. The significantly enriched gene sets are shown in Table 13.6. Compared to the sterile inflammation in the immediate aftermath of cardiac surgery, sepsis is associated with

induction of genes involved in the complement cascade, but repression of TLR and IFN cascades, and chemokine receptor activity. There is also altered regulation of genes involved in lipid, carbohydrate, and anion transport.

Late Response

With a 1.5 fold change in mean expression threshold and a FDR of 0.05, 979 probes were differentially expressed between the sepsis and the immediate post-operative samples, of which 580 probes were repressed. These correspond to 886 genes, of which 367 were induced and 520 were repressed.

The genes with the largest effect sizes are listed in Table 13.7. In contrast to the comparison in the previous section, the genes with the biggest changes are all induced. A number of the genes are neutrophil specific; the induction seen could be the result of persistent neutrophilia in sepsis patients compared to the re-normalizing neutrophil counts at the first post-operative day after cardiac surgery.

The overlap between the genes associated with sepsis to immediate response and the sepsis to late response contrasts with a 1.5 fold change in relative expression and a FDR of 0.05 is shown in Figure 13.10. Overall, 600 genes have significant differences in expression in both contrasts. The expression of these genes are highly correlated across the contrasts, with a Pearson correlation coefficient of 94.8%.

Gene sets that are significantly enriched for the difference between sepsis and the late post-op response are shown in Table 13.8. Compared to the late genomic events following cardiac surgery, sepsis is associated with induction of genes associated with cell division and motility and repression of apoptotic pathways. This would suggest a persistent inflammatory response in contrast to the resolution of inflammation in SIRS. However, given the negative scores in the inflammation and graft versus host disease pathways, an altered pattern of inflammation might be a more accurate description.

The gene sets associated with the two contrasts comparing sepsis to SIRS show minimal overlap. This probably reflects the dynamics of the SIRS response, with a faster resolution kinetics contrasting with a persistent inflammatory response in sepsis.

Gene Set	Enrichment Score
chr4p16	-0.45
chr3q28	-0.63
chr6p23	-0.68
Reactome chemokine receptors	-0.58
Reactome TRAF6 mediated induction of IFNab cascade	-0.50
Reactome MAPK activation in TLR cascade	-0.48
Spliceosome assembly	-0.71
Inorganic anion transport	-0.80
Locomotory behaviour	-0.47
RNA splicing via trans-esterification	-0.57
G protein coupled receptor binding	-0.54
Chemokine activity	-0.56
Chemokine receptor binding	-0.57
chr1q22	0.45
Biocarta COMP	0.70
Reactome complement cascade	0.70
Reactome trafficking of AMPA receptors	0.61
Energy reserve metabolic process	0.65
Centrosome organization and biogenesis	0.68
Lipid transport	0.64
Carbohydrate transport	0.60
Intracellular receptor mediated signalling	0.57
Hormone metabolic process	0.56
Cytokine secretion	0.61
Adherens junction	0.57
Integrin complex	0.62
Exopeptidase activity	0.53
Carbohydrate transmembrane transporter activity	0.68
Enzyme inhibitor activity	0.38
Actin filament binding	0.60

Table 13.6: Gene sets enriched in sepsis when compared to the immediate response to cardiac surgery

Gene	Chr	Role
MPO	17	myeloperoxidase
MS4A3	11	membrane-spanning 4-domains (hematopoietic cell-specific)
DEFA4	8	defensin, alpha 4, corticostatin
DEFA1	8	defensin, alpha 1
DEFA1B	8	defensin, alpha 1B
DEFA3	8	defensin, alpha 3, neutrophil-specific
LCN2	9	lipocalin 2
LTF	3	lactotransferrin
ELANE	19	elastase, neutrophil expressed
OLFM4	13	olfactomedin 4
MMP8	11	matrix metallopeptidase 8 (neutrophil collagenase)
CEACAM6	19	carcinoembryonic antigen-related cell adhesion molecule 6
CTSG	14	cathepsin G
BPI	20	bactericidal/permeability-increasing protein
CEACAM8	19	carcinoembryonic antigen-related cell adhesion molecule 8

Table 13.7: Genes with the largest differences in mean expression in sepsis samples when compared with the late response to cardiac surgery, p 0.01 and 4 fold change in expression

Gene Set	Enrichment Score
chr15q14	0.52
Biocarta AMI	0.70
KEGG taste transduction	0.54
Cytokinesis	0.65
Cell division	0.65
Sensory perception of chemical stimulus	0.81
Microtubule cytoskeleton	0.43
Biocarta Inflammation	-0.69
KEGG graft vs host disease	-0.69
DNA damage response with induction of apoptosis	-0.61
Negative regulation of translation	-0.54

Table 13.8: Gene sets enriched in sepsis when compared with the late response to cardiac surgery

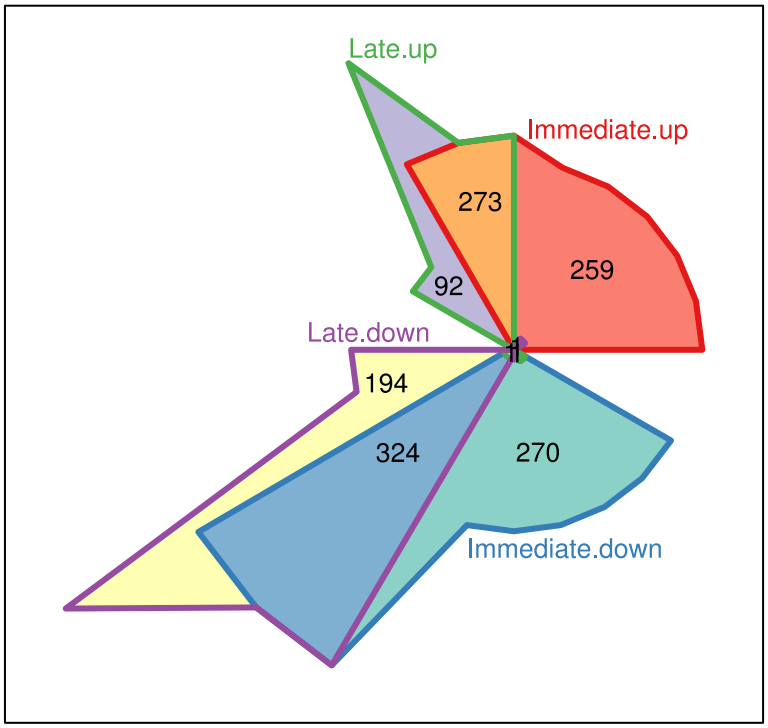


Figure 13.10: Venn diagram of gene expression overlap in sepsis compared to each of the cardiac post-operative samples. Immediate = Differentially regulated genes in sepsis when compared with the immediate response to cardiac surgery, Late = Differentially regulated genes in sepsis when compared with the late response to cardiac surgery, up = induced, down = repressed. 273 genes are consistently induced, while 324 are consistently repressed in sepsis when compared to the SIRS response of cardiac surgery, irrespective of the time of sampling.

Relationship to time of sampling

The significance of time of sampling in sepsis is unclear. The day of sampling in sepsis represents a physiological state of failing homoeostatic mechanisms but does not correspond to a homogeneous gene expression signature. In contrast, the time of sampling in cardiac surgery has a clearly defined relationship to the surgical insult and the onset of the inflammatory response. If an unjustified assumption that the day 1 of ITU admission corresponds to the first postoperative day of cardiac surgery is made, then there should be some correspondence between the underlying gene expression patterns.

To test this hypothesis, samples from the first postoperative day of cardiac surgery were compared with the samples from the first day of ITU admission. A simple regression analysis was performed. With the conventional 1.5 fold change in mean expression values as threshold and a FDR threshold of 0.05, 1392 probes corresponding to 1241 unique

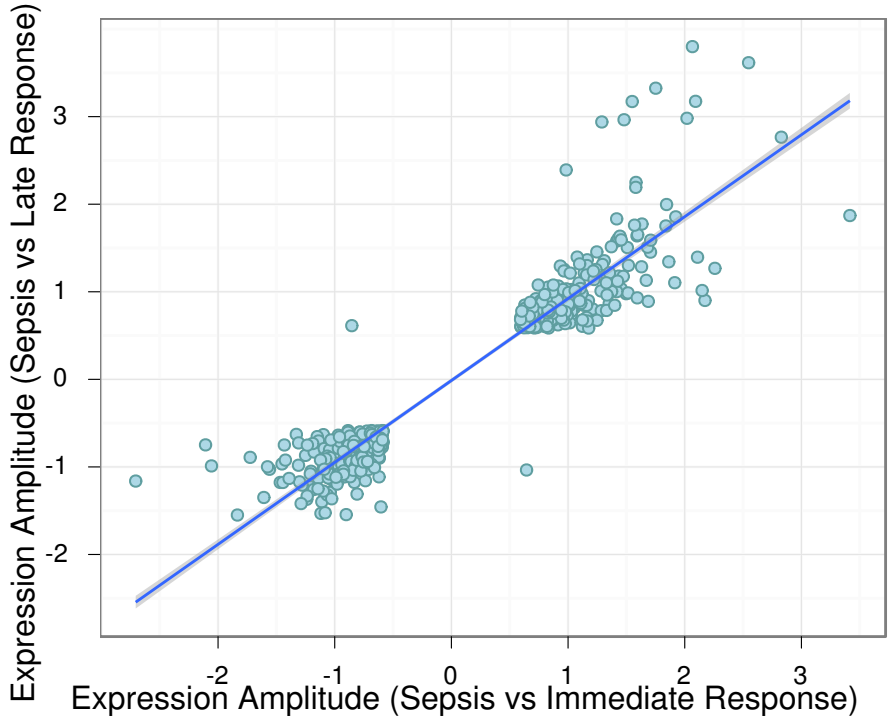


Figure 13.11: Correlations in the gene expression in differentially regulated genes when sepsis samples are compared to the immediate and late responses to cardiac surgery.

genes had significant differences in expression between the two groups. Of these 542 probes were induced and 850 probes were repressed. This corresponds to 494 and 747 genes respectively.

The distribution of effect sizes suggests a significant proportion of genes with large effect sizes (Figure 13.12). The top hits are similar to the sepsis to late response comparisons suggesting that time of sampling in the cardiac and sepsis cohorts are not equivalent. Given this inference, there is unlikely to be a biological relevance to this contrast.

Pathways

The impact of infection on the sterile systemic inflammatory syndrome was detailed in the preceding analysis. The results suggest significant differences between sepsis and SIRS resulting from cardiac surgery. Gene sets related to TLR signalling and complement cascade were differentially regulated between sepsis and SIRS. In this context, it is

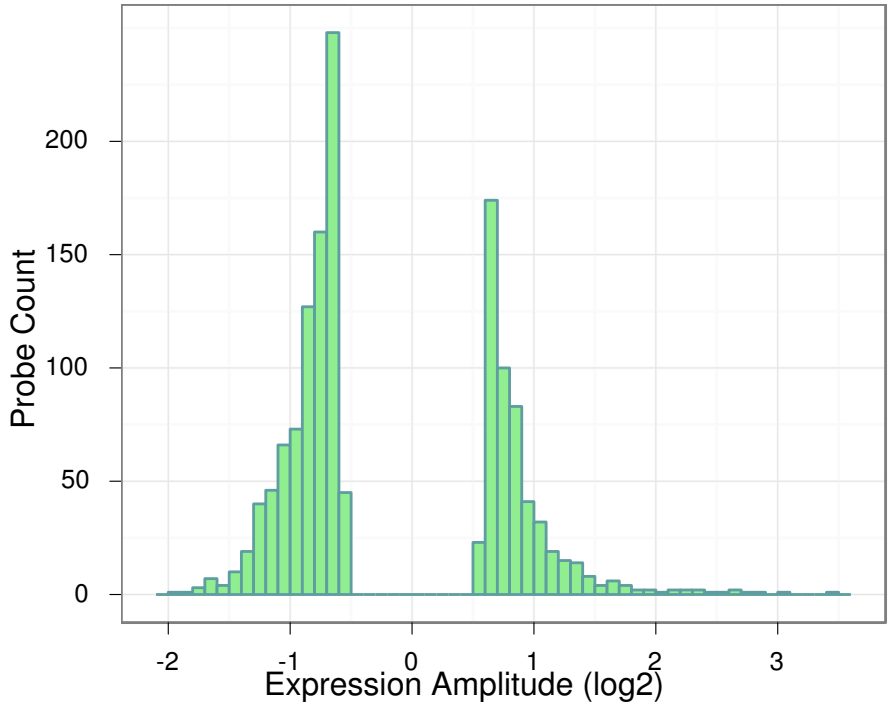


Figure 13.12: Gene expression changes in patients with sepsis on the day of admission when compared to the inflammatory response in the first postoperative day following cardiac surgery

useful to contrast the differences between the gene sets associated with cardiac surgery and sepsis. Many possible permutations of contrasts in sepsis and cardiac surgery are possible, but this analysis is restricted to the pathways enriched in comparisons of post-operative samples to the baseline, which are compared to the analogous contrasts in sepsis comparing sepsis samples to baseline cardiac surgery samples.

Pathways significantly associated with cardiac surgery were analysed. The leading edge genes contributing to the enrichment score of the gene set was extracted for the cardiac surgery and the sepsis response contrasts. The significance of the pathway in relation to sepsis response was estimated by counting the number of genes that are shared between both the leading edge gene lists, and correlating the mean expression difference for the shared genes.

The behaviour of gene sets associated with the immediate response in the cardiac cohort in the context of severe sepsis due to CAP and FP are shown in tables 13.9 and 13.10 respectively. Overall, there is minimal correlation between the sepsis response in CAP and FP with the immediate response following cardiac surgery.

Gene set	Cardiac	CAP	Shared	p	Estimate
Reactome mRNA splicing	35	39	14	0.91	0.032
Reactome costimulation by the CD28 family	22	23	14	0.46	-0.214
Reactome translation initiation complex formation	37	43	35	0.19	-0.229
Regulation of translational initiation	13	10	8	0.74	-0.139
RNA polymerase activity	10	7	5	0.10	0.809
Regulation of lymphocyte activation	15	7	7	0.08	-0.697
Lymphocyte differentiation	12	5	4	0.82	0.176
Lymphocyte activation	24	13	11	0.10	-0.518
Small nuclear ribonucleoprotein complex	13	9	8	0.29	0.428
Mitochondrial inner membrane	33	30	20	0.32	0.234
Biocarta CTLA4 pathway	10	9	7	0.34	0.427
Reactome insulin synthesis and secretion	64	66	53	0.30	-0.145
KEGG calcium signalling pathway	16	24	6	0.69	-0.213

Table 13.9: Pathway correlation with the genomic response in CAP and the immediate genomic response to cardiac surgery. The numbers indicate the number of leading edge genes contributing to the enrichment of the gene set. Shared = number of genes that are shared in the leading edge gene subset for each of the contrasts, Estimate = Spearman correlation coefficient

Gene set	Cardiac	Sepsis	Shared	p	Estimate
Reactome mRNA splicing	35	46	12	0.37	0.284
Reactome costimulation by the CD28 family	22	21	10	0.85	0.067
Reactome translation initiation complex formation	37	26	20	0.38	-0.208
RNA polymerase activity	10	7	4	0.64	-0.355
Regulation of lymphocyte activation	15	13	10	0.28	0.376
Lymphocyte differentiation	12	4	3	0.87	-0.200
Lymphocyte activation	24	14	11	0.77	-0.100
Small nuclear ribonucleoprotein complex	13	8	6	0.47	0.374
Mitochondrial inner membrane	33	17	11	0.86	0.060
Biocarta CTLA4 pathway	10	7	5	0.95	-0.043
Reactome insulin synthesis and secretion	64	56	42	0.73	0.054

Table 13.10: Pathway correlation with the genomic response in FP and the immediate genomic response to cardiac surgery. The numbers indicate the number of leading edge genes contributing to the enrichment of the gene set. Shared = number of genes that are shared in the leading edge gene subset for each of the contrasts, Estimate = Spearman correlation coefficient

In the CAP cohort, there are cluster specific correlations to the immediate response to cardiac surgery. Cluster 1 shows two strong correlations; the KEGG complement and coagulation cascade and the voltage gated channel activity gene sets both have a correlation coefficient of -99.7%. However, in both cases the result is only based on 3 shared genes. There are no significant associations with other clusters. However, cluster 3 shows a tendency towards significance with the electron carrier gene set. The correlation is moderate (Pearson correlation coefficient 57.2%) and is contributed by 5 genes.

The behaviour of gene sets associated with the late response in the cardiac cohort in the context of severe sepsis due to CAP and FP are shown in tables 13.11 and 13.12 respectively. Only a few gene sets show significant correlations, but the correlated gene sets are different for the CAP and FP responses. For CAP, the KEGG lysosome gene set is correlated with 27 shared genes (Pearson correlation coefficient 44.6%, p 0.02). A stronger correlation is seen with Ras GTPase activator gene set with 7 shared genes showing strong correlations (Pearson correlation coefficient 78.2%, p 0.04). In the sepsis response in FP, KEGG lysosome gene set is not significantly correlated, despite 22 shared genes. In contrast to the CAP response, the genes in the FP response show a negative correlation. Significant correlation is seen in the Wnt signalling gene set based on gene ontology (Pearson correlation coefficient 98.6%, p 0.01), but the result is only based on 4 genes. Interestingly, the KEGG Wnt signalling gene set show no correlation, despite sharing 14 genes, and the association has dubious validity.

In CAP, there are cluster specific differences in gene set associations with the late response to cardiac surgery. For cluster 1, KEGG lysosome gene set is associated in a pattern that mirrors the CAP response association. This association is not seen in the other clusters. Cluster 2 shows correlations with the electron carrier activity gene set (Pearson correlation coefficient 78.7%, p 0.06), with a tendency towards significance. This association is based on 6 shared genes. Cluster 3 has no significant associations.

The results suggest that the overall sepsis response is heavily biased by cluster 1, and re-iterates the different gene expression patterns in the clusters. The electron carrier activity gene set is correlated with cluster 3 in the immediate response, but to cluster 2

Gene set	Cardiac	Sepsis	Shared	p	Estimate
KEGG lysosome	47	39	27	0.02	0.446
Ras GTPase activator activity	7	20	7	0.04	0.782
Voltage gated channel activity	5	13	3	0.20	0.953
Oxidoreductase activity	45	28	16	0.39	0.230
Reactome RNA polymerase I promoter clearance	21	15	11	0.96	0.019
Reactome branched chain amino acid catabolism	9	8	4	0.36	0.643
Electron carrier activity	15	8	6	0.44	0.395
KEGG complement and coagulation cascades	13	12	5	0.31	0.572
Wnt signalling	13	16	8	0.67	-0.181
Reactome signalling by Wnt	19	13	3	0.43	-0.783
Wound healing	16	14	8	0.64	-0.197

Table 13.11: Pathway correlation with the genomic response in CAP and the late genomic response to cardiac surgery. The numbers indicate the number of leading edge genes contributing to the enrichment of the gene set. Shared = number of genes that are shared in the leading edge gene subset for each of the contrasts, Estimate = Spearman correlation coefficient

Gene set	Cardiac	Sepsis	Shared	p	Estimate
KEGG lysosome	47	37	22	0.30	-0.233
Regulation of catalytic activity	43	48	5	0.39	0.499
Ras GTPase activator activity	7	9	4	0.54	-0.465
Oxidoreductase activity	45	54	18	0.09	0.410
Reactome branched chain amino acid catabolism	9	4	3	0.82	-0.274
Electron carrier activity	15	21	8	0.40	0.348
KEGG complement and coagulation cascades	13	21	6	0.78	-0.145
Wnt signalling	13	16	4	0.01	0.986
KEGG Wnt signalling pathway	28	24	14	0.64	0.137
Reactome signalling by Wnt	19	20	3	0.97	0.041

Table 13.12: Pathway correlation with the genomic response in FP and the late genomic response to cardiac surgery. The numbers indicate the number of leading edge genes contributing to the enrichment of the gene set. Shared = number of genes that are shared in the leading edge gene subset for each of the contrasts, Estimate = Spearman correlation coefficient

in the late response. Cluster 1 shows consistent correlations to the immediate and late response for the KEGG lysosome gene set. This suggests the possibility that at least some parts of the cluster structure might have temporal relationships.

13.5 Limitations

While cardiac surgery provides a good model of sterile inflammatory response, the approach in this chapter has a number of limitations.

1. Biased patient population — While the age group and sex distribution of the cardiac cohort matches that of the sepsis cohort, these patients have coronary artery disease which is known to be associated with inflammatory changes and abnormal macrophage behaviour. The predisposition to cardiac disease results in a systematic bias.

2. No correction for cell counts — At least a part of the changes in gene expression are due to changes in the absolute white cell counts and the relative proportion of immune cells.

3. Relevance of time of sampling in sepsis patients — In the cardiac cohort, there is a consistent relationship between samples from a given time point reflecting the time course of evolution of the immune response. In sepsis samples, there is considerable variation between the onset of illness and time of sampling, the duration of which cannot be estimated. There is less coherence between samples and the assignment of time point labels to the samples is of unclear value.

13.6 Summary

Cardiac surgery is associated with a significant systemic inflammatory response. The clinical course of recovery following cardiac surgery has a consistent temporal pattern. Thus cardiac surgery is a good clinical model for studying the systemic inflammatory response against which the clinical sepsis response can be contrasted.

Hierarchical agglomerative clustering of the principle components of the most variant genes in the cardiac cohort shows three distinct clusters clearly differentiating the time of sampling. Regression models comparing samples from different time points, show large changes with more than 2 fold change in nearly a quarter of the genes. The greatest changes are associated with strong predictive power, with a multi-class predictive model showing 83% predictive accuracy. The samples capture a dynamic response following cardiac surgery, where a strong initial inflammatory response in the immediate post-operative period gives way to a different inflammatory pattern at day 1 with some evidence of resolution of response.

The inflammatory pattern is composed of changes in the mitochondrial inner membrane, CTLA4, RNA polymerase, and voltage gated channel activity among others. The resolution of the inflammatory response is associated with T cell activation, and changes in the gene sets related to the mitochondrial outer membrane, and the regulation of apoptosis.

The sepsis response shows clear and strong differences from the systemic inflammatory response. Clear differences are seen between *TLR5*, *IL8*, and *CECAM6* expression in the immediate response, while the late response shows predominant patterns of expression probably related to altered neutrophil counts. The difference between sepsis samples and the immediate response is associated with significant changes is gene sets related to chemokine activity, cell adhesion, substrate and ion transport, and TLR cascade. The late response is associated with fewer pathways related to cell adhesion, motility, and inflammation. 600 genes are shared between both the contrasts, with highly correlated expression values.

There are only minimal associations between the CAP and FP sepsis responses. More consistent but moderate correlations are seen with CAP clusters. In particular, cluster 1 shows an analogous pattern of expression to the immediate and late responses to cardiac surgery in genes involved in the KEGG lysosome pathway. Intriguingly, the electron carrier gene set shows associations with cluster 2 and 3 at different time points, suggesting the possibility of the clusters resulting from the evolution of the genomic inflammatory response.

14

Genomic Signatures of Endotoxin Tolerance

14.1 Background

Endotoxin tolerance is an immune phenomenon in which the pattern of immune response to a second exposure to bacterial lipopolysaccharide (LPS) differs from the first. The phenomenon was first described in humans but was subsequently identified in other species and cell line systems. Tolerance is usually defined on the basis of suppression of TNF protein release, but decreased production of other cytokines such as IL6 have also been demonstrated (Flohé *et al.*, 1999).

It has been suggested that endotoxin tolerance is a feature of clinical sepsis, but the role of tolerance in sepsis risk and outcome is not clear. Tolerance could result in an inadequate immune response in the face of persisting infection, whereas a lack of tolerance might be associated with an inappropriate hyper-inflammatory state resulting in tissue damage.

The genomic basis of tolerance and its relationship to the genomic events in sepsis is examined in this chapter.

14.2 Materials and Methods

After ethics approval and informed consent, 96 healthy volunteers were recruited for the study. Subjects with self-reported infection in the preceding two weeks and pregnant subjects were excluded from the study.

50 mls of peripheral venous blood was obtained from each volunteer using EDTA vacutainer system. The blood was layered over a Ficoll (Sigma-Aldrich) gradient and centrifuged for 30 minutes at 400g at 4°C. The leucocytes were aspirated from the Ficoll interface and washed twice with Hanks buffered saline solution and once with culture media. Cells were counted with a haemocytometer using tryptan blue staining and resuspended in RPMI 1640 medium (Sigma) supplemented with L-glutamine, penicillin, streptomycin, and foetal calf serum.

PBMC samples from each volunteer were separated into four aliquots. The first aliquot was suspended in fresh media and was used as the control sample. Cells in two of the remaining three aliquots were stimulated with LPS at two concentrations, 0.2 mcg/ml and 2 mcg/ml (low dose and high dose, denoted as LD and HD respectively) for 6 hours. Cells in the fourth aliquot were first stimulated with low dose LPS for 6 hours, washed, and restimulated with high dose LPS for another 6 hours (denoted as LDHD). After the appropriate time period, each aliquot was centrifuged.

The supernatant was retained for protein analysis, while mRNA from the cells was extracted using a fast spin-column (Qiagen RNA mini). mRNA samples were checked for quality using the Bioanalyzer 2000 platform and an aliquot of the samples was hybridised to the Illumina Human WGA6-v3 platform.

14.3 Exploratory analysis

TNF response to LPS

Subjects were classified as tolerant and non-tolerant by the suppression of TNF protein production in pre-exposed cells as compared to naive cells.

There is some discrepancy in the literature regarding the most suitable dose of LPS and the duration of exposure required to differentiate the tolerant phenotype. The time required for adequate discrimination of the phenotypes was established using cell lines derived from the HapMap CEU population. Cells were suspended in fresh culture medium containing low dose LPS. After 6 hours cells were resuspended in either fresh culture medium or culture medium containing high dose LPS. The TNF production

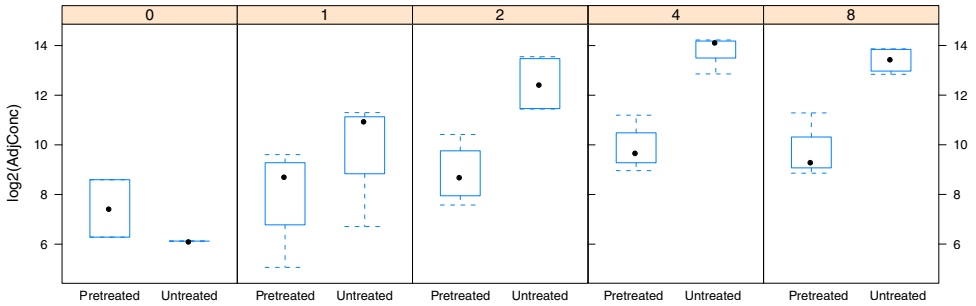


Figure 14.1: Change in TNF concentrations over time in cells pretreated with LPS and naive cells showing the time course for TNF tolerance. Numbers at the top of each panel refer to the number of hours following stimulation with LPS. The production of TNF plateaus at 4 hours following stimulation with LPS.

in each group was quantified at baseline, 1, 2, 4, and 8 hours post stimulation using ELISA. The results (Figure 14.1) show adequate discrimination between 4 and 6 hours post stimulation.

The effect of repeat LPS stimulation on TNF production in pre-treated cells derived from healthy volunteers is shown in Figure 14.2. Using a log scale, the fractional TNF concentration at 6 hours post-stimulation is approximately normally distributed with prominent tails due to the presence of outliers at both extremes. The left and right extremes represent the tolerant and non-tolerant phenotypes respectively. Twelve volunteers from either extreme were selected as representative of the phenotype. mRNA derived from these volunteers was hybridized to Illumina WGA6-v3 BeadArray platform, and forms the data set used in the rest of this chapter.

Hierarchical clusters

The microarray data was normalized and checked for quality using the methodology described in previous chapters. After filtration for a detection threshold p value of 0.05, 16,031 probes were retained for analysis.

14.4 Genomic response to LPS

The general approach to the comparison of phenotypes of interest has been described in earlier chapters. In brief, a regression model was constructed for the expression of each probe with the specified experimental design. The contrasts of interest are estimated by

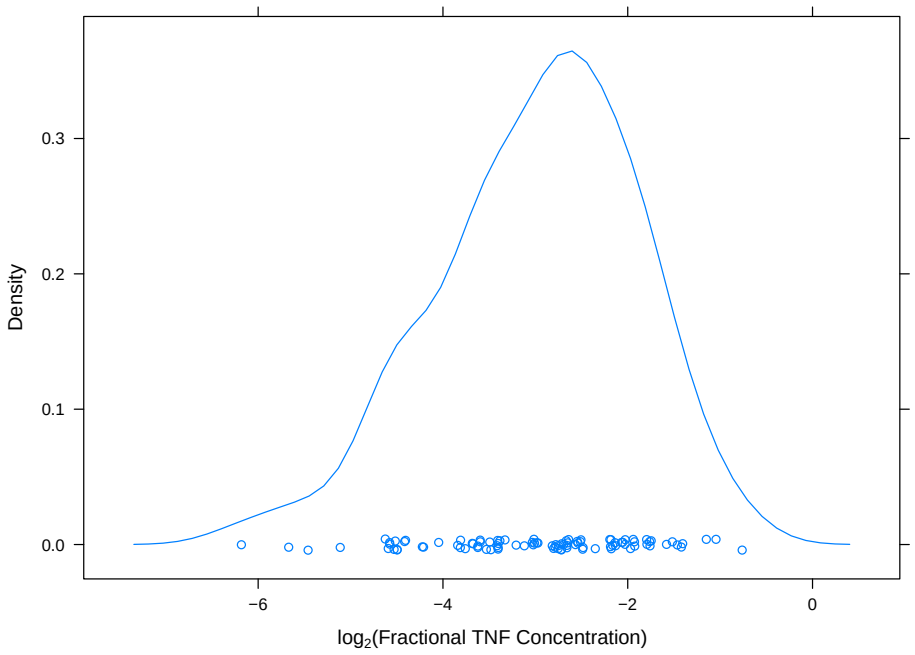


Figure 14.2: Fractional change in the TNF concentrations. Each circle represents an individual. Fractional change is defined as the ratio of the concentration observed in pre-treated cells to that observed in naive cells following high dose stimulation. Decreasing values in the abscissa indicate greater decreases in TNF concentration and greater tolerance. Each unit decrease in the abscissa corresponds to a 2 fold decrease in concentration.

comparing the appropriate coefficients. An adjusted p value (FDR) is calculated using a Bayesian approach which compares the variance of the expression of the probe of interest, to the overall variance of all the probes. Lists of significant genes are placed in a biological context using pathway analysis with GSEA.

Regression models

Using a robust regression model, with a predefined significance threshold of at least 1.5 fold change in mean expression and a FDR of 0.05, the expression of 330 probes define the low dose response, and 851 probes define the high dose response (Table 14.1). The dose response relationship, defined as the difference in responses to high dose and low dose stimulation, consists of 261 probes. The predominant genomic response of PBMCs to LPS consists of repression, a pattern which occurs irrespective of the stimulating dose. There is a clear dose response relationship where increasing dose, increases the

number of probes involved in the response (Figure 14.3). The change in expression due to increasing LPS dose is also predominantly repressed.

Contrast	Counts	induced	repressed	Total
Low dose	Probes	116	214	330
	Genes	105	197	302
High dose	Probes	342	509	851
	Genes	321	463	784
Dose response	Probes	108	153	261
	Genes	99	149	248

Table 14.1: Number of probes/genes with at least 1.5 fold change difference in expression and a FDR less than 0.05 in the endotoxin tolerance experiment. Low dose = gene expression in cells exposed to low dose LPS compared to baseline, high dose = gene expression in cells exposed to high dose LPS compared to baseline, dose response = gene expression in cells exposed to high dose LPS compared to expression in cells exposed to low dose LPS.

The top genes for the low dose and high dose response are listed in Tables 14.2 and 14.3 respectively. The list includes a number of immune and regulatory genes such as RNAses, cytokines, chemokines, and receptors. Some of these genes have been linked to features of clinical sepsis. Their relevance in sepsis is examined in section 14.6 of this chapter.

The overlap of the responses to LD and HD stimulation is shown in Figure 14.4. The genomic response to increasing doses of LPS is not a simple increase in the number of genes, and appears to be a graded increase related to the endotoxin load. While a large proportion of genes are shared between the two doses, 11 induced probes, and 16 repressed probes are unique to the LD stimulation. These genes are listed in Table 14.4, and include genes with immune functions such as *PPARG* (peroxisome proliferator-activated receptor gamma), *TLR8*, *ADAMDEC1* (disintegrin metalloproteinase). The consistent induction of oligodendrocyte transcription factors from chromosome 21 and consistent repression of two probes for *PRF1* (lymphocyte pore forming protein) suggests that these results are not spurious.

The repression of *TLR8* (which binds to viral ligands) on exposure to endotoxin is particularly interesting. The distribution of *TLR8* expression following different stim-

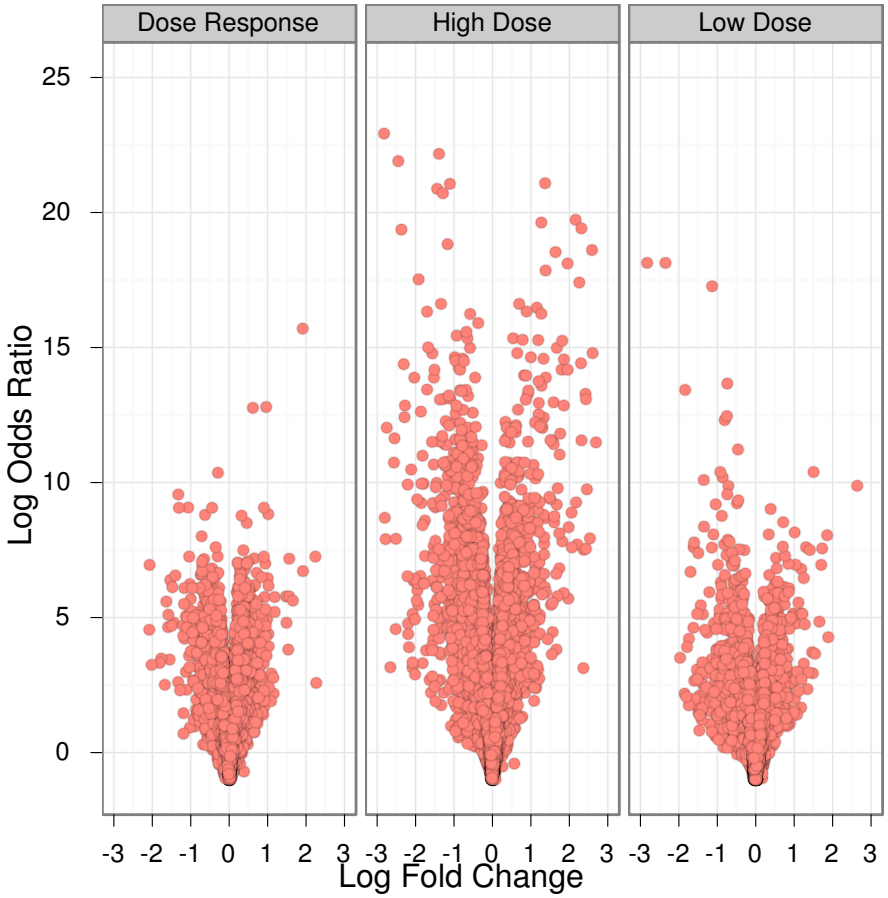


Figure 14.3: Volcano plot showing the effect sizes of the gene expression response to different doses of LPS. The right panel shows the gene expression changes in cells exposed to low dose LPS when compared to baseline gene expression. The middle panel shows genes expression in cells exposed to high dose LPS when compared to baseline gene expression. Higher dose of of LPS is associated with greater fold changes and statistical significance. The change in expression between the high dose and low dose response can be quantified as the dose response (left panel).

ulations is shown in Figure 14.5. The expression of *TLR8* clearly differs with the desensitisation response. However, the expression of *TLR8* is bimodal even in cells with no prior exposure to endotoxin, suggesting that the mechanisms underlying tolerance are active in cells that have not been exposed to LPS. Also the pattern of expression is reversed in the tolerance phenotype, and the cells with higher expression when exposed to endotoxin in the naive state, showing greater suppression when exposed to endotoxin in the pretreated state. Interestingly, the same pattern of bimodal distribution and the reversal of expression in the tolerance phenotype is also seen in the genes that define tolerance phenotype (Figure 14.6). The pattern of expression of these genes also suggests

Gene	Chr	Function
FPR3	19	formyl peptide receptor 3
FCN1	9	ficolin (collagen/fibrinogen domain containing) 1
TGFBI	5	transforming growth factor, beta-induced, 68kDa
RNASE6	14	ribonuclease, RNase A family, k6
RNASE1	14	ribonuclease, RNase A family, 1 (pancreatic)
FBP1	9	fructose-1,6-bisphosphatase 1
CDC42EP5	19	CDC42 effector protein (Rho GTPase binding) 5
TM4SF1	3	transmembrane 4 L six family member 1
TNIP3	4	TNFAIP3 interacting protein 3
TNFRSF21	6	tumor necrosis factor receptor superfamily
SERPINB7	18	serpin peptidase inhibitor, clade B (ovalbumin)
CFB	6	complement factor B
MUCL1	12	mucin-like 1
IL23A	12	interleukin 23, alpha subunit p19
CD274	9	CD274 molecule
SLAMF1	1	signalling lymphocytic activation molecule family member 1
KIAA1199	15	KIAA1199
IL2RA	10	interleukin 2 receptor, alpha

Table 14.2: Genes of the low dose response (3 fold change, p 0.01). Genes in the top panel are induced, and in the bottom panel are repressed.

that the low dose and high dose phenotypes consists of independent expression profiles and not merely similar expression patterns with increasing effect sizes. This could either be due to different patterns of expression determined by endotoxin dose, or an evolving response pattern where the low dose response is a later event in the natural history of gene expression response to endotoxin.

The correlation between the expression amplitudes with low dose and high dose stimulation are shown in Figure 14.7. The figure plots the expression amplitudes of 851 probes identified to be significant for the high dose response, and includes probes that were not statistically significant for the low dose response. There is a strong correlation between the expression amplitudes with both doses (Pearson correlation coefficient 91.27%). In general, increasing the dose increases the effect size of the response. However, in 25 probes the effect size is greater with low dose stimulation than high dose stimula-

Gene	Chr	Function
FPR3	19	formyl peptide receptor 3
CD163	12	CD163 molecule
HLA-DRB5	6	major histocompatibility complex, class II, DR beta 5
FCN1	9	ficolin (collagen/fibrinogen domain containing) 1
CAMK1	3	calcium/calmodulin-dependent protein kinase I
TGFBI	5	transforming growth factor, beta-induced, 68kDa
FUCA1	1	fucosidase, alpha-L- 1, tissue
RNASE6	14	ribonuclease, RNase A family, k6
RNASE1	14	ribonuclease, RNase A family, 1 (pancreatic)
STAB1	3	stabilin 1
CSF2	5	colony stimulating factor 2 (granulocyte-macrophage)
ANGPTL4	19	angiopoietin-like 4
IL6	7	interleukin 6 (interferon, beta 2)
F3	1	coagulation factor III (thromboplastin, tissue factor)
SERPINB7	18	serpin peptidase inhibitor, clade B (ovalbumin)
MUCL1	12	mucin-like 1
IL23A	12	interleukin 23, alpha subunit p19
CCL20	2	chemokine (C-C motif) ligand 20
IL1A	2	interleukin 1, alpha
CCL3L1	17	chemokine (C-C motif) ligand 3-like 1
PTGS2	1	prostaglandin-endoperoxide synthase 2 (cyclooxygenase)
SERPINA1	14	serpin peptidase inhibitor (alpha-1 antiproteinase, antitrypsin)
HES4	1	hairy and enhancer of split 4 (Drosophila)

Table 14.3: Genes of the high dose response (5 fold change, p 0.01). Genes in the top-panel are induced, and in the bottom-panel are repressed.

Gene	Chr	Function
PPARG	3	peroxisome proliferator-activated receptor gamma
SULF2	20	sulfatase 2
OLIG2	21	oligodendrocyte lineage transcription factor 2
OLIG1	21	oligodendrocyte transcription factor 1
TKTL1	X	transketolase-like 1
GPR35	2	G protein-coupled receptor 35
NT5DC2	3	5'-nucleotidase domain containing 2
COL7A1	3	collagen, type VII, alpha 1
VENTX	10	VENT homeobox
CTSD	11	cathepsin D
COL23A1	5	collagen, type XXIII, alpha 1
LILRB2	19	leucocyte immunoglobulin-like receptor, subfamily B
LMNB1	5	lamin B1
EMR1	19	egf-like module containing
STOM	9	stomatin
NAPSB	19	napsin B aspartic peptidase pseudogene
TSPAN13	7	tetraspanin 13
FBXO6	1	F-box protein 6
LDLRAD3	11	low density lipoprotein receptor class A domain containing 3
APCDD1	18	adenomatosis polyposis coli repressed 1
TLR8	X	toll-like receptor 8
PDCD1LG2	9	programmed cell death 1 ligand 2
ARNT2	15	aryl-hydrocarbon receptor nuclear translocator 2
PRF1	10	perforin 1 (pore forming protein)
ADAMDEC1	8	ADAM-like, decysin 1
HAPLN3	15	hyaluronan and proteoglycan link protein 3

Table 14.4: Genes unique to LD response. The immunological function of this unique component supports the notion that the dose response to LPS is not just a change in magnitude of gene expression.

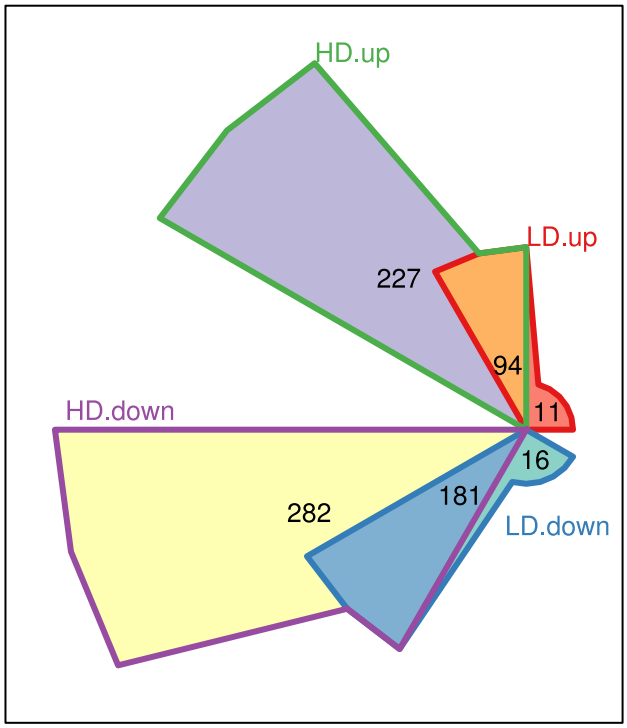


Figure 14.4: Venn diagram showing the unique and shared components of the gene expression response in cells stimulated to different doses of LPS. Exposure to high dose LPS is associated with a quantitatively greater changes in gene expression. LD = low dose, HD = high dose, up = induced genes, down = repressed genes. 94 genes are induced with both treatments, while 181 are repressed. 11 genes are uniquely induced with low dose treatment, while 16 are uniquely repressed. This unique component should not be present, if the response to high dose LPS is just an exaggeration of the response to low dose LPS.

tion. The statistical significance of these genes were not independently calculated, since they are better identified by the dose-response contrast of the regression model.

The regression analysis identifies 248 genes that are differentially expressed between the two endotoxin dose responses, with at least 1.5 fold change in mean expression and a FDR less than 0.05. Of these 149 genes are repressed, while 99 are induced when the expression following high dose stimulation is compared to low dose stimulation. The genes with the largest changes in expression are listed in Table 14.5. The results confirm the earlier inference that the high dose response is an unique phenotype and not just the low-dose response with different effect sizes.

Gene	Chr	Function
CD163	12	CD163 molecule
CD36	7	CD36 molecule (thrombospondin receptor)
HLA-DRB5	6	major histocompatibility complex, class II, DR beta 5
CXCL9	4	chemokine (C-X-C motif) ligand 9
CXCL10	4	chemokine (C-X-C motif) ligand 10
CCL8	17	chemokine (C-C motif) ligand 8
PKIB	6	protein kinase (cAMP-dependent, catalytic) inhibitor beta
CSF2	5	colony stimulating factor 2 (granulocyte-macrophage)
CLEC1B	12	C-type lectin domain family 1, member B
ANGPTL4	19	angiopoietin-like 4
IL6	7	interleukin 6 (interferon, beta 2)
F3	1	coagulation factor III (thromboplastin, tissue factor)
SEPT5	22	septin 5
TUBB1	20	tubulin, beta 1 class VI
NRGN	11	neurogranin (protein kinase C substrate, RC3)
IL23A	12	interleukin 23, alpha subunit p19
CCL20	2	chemokine (C-C motif) ligand 20
CA12	15	carbonic anhydrase XII
TFPI2	7	tissue factor pathway inhibitor 2
IL1A	2	interleukin 1, alpha
GP9	3	glycoprotein IX (platelet)
CXCL2	4	chemokine (C-X-C motif) ligand 2
HES4	1	hairy and enhancer of split 4 (Drosophila)

Table 14.5: Genes differentiating the low dose from high dose response. Genes in the top panel are induced and the bottom panel are repressed in the high dose response with at least 2 fold change in expression.

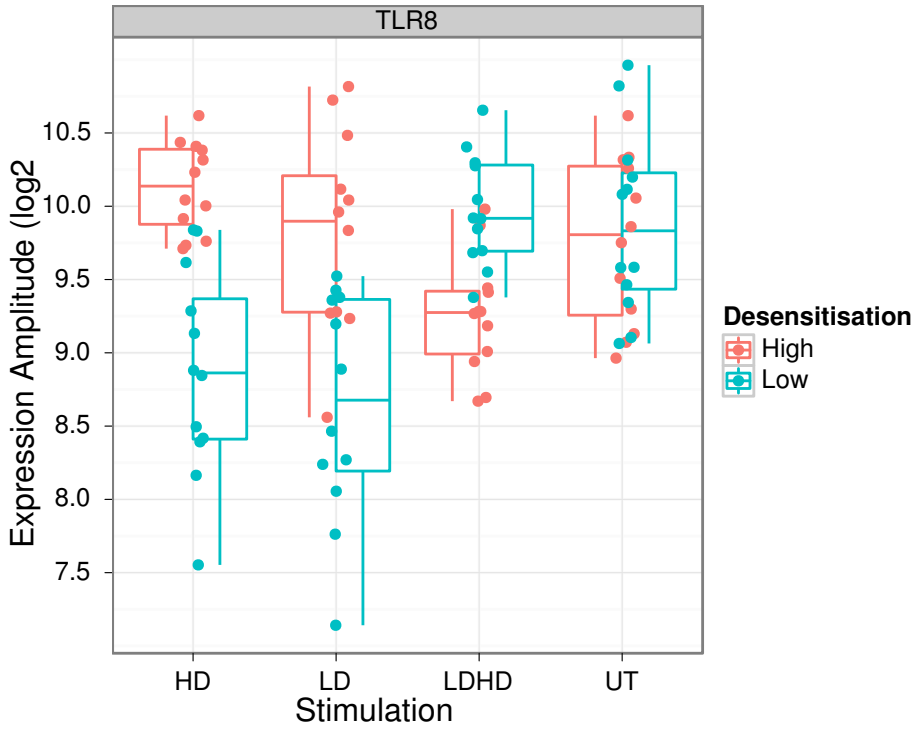


Figure 14.5: Bimodal expression of *TLR8* with LPS stimulation. UT = naive cells, LD = stimulation with low dose LPS, HD = stimulation with high dose LPS, LDHD = stimulation with high dose LPS in cells pretreated with low dose LPS, Low = low desensitisers (non-tolerant cells), High = high desensitisers (tolerant cells).

Pathway analysis

Pathway analysis was performed using the gene set enrichment algorithm. Gene sets relating to a biological feature were assigned an enrichment score calculated from an appropriate test statistic for each gene in that gene set. The significance levels for the enrichment score was determined by permutation analysis involving phenotype switching. The significantly enriched gene sets that are associated with the low and high dose responses and the dose-response relationship to endotoxin are shown in Figure 14.8.

14.5 Genomic signatures of tolerance

The endotoxin tolerance phenotype is defined as the difference between expression in the high desensitisers (tolerant) and the low desensitisers (non-tolerant).

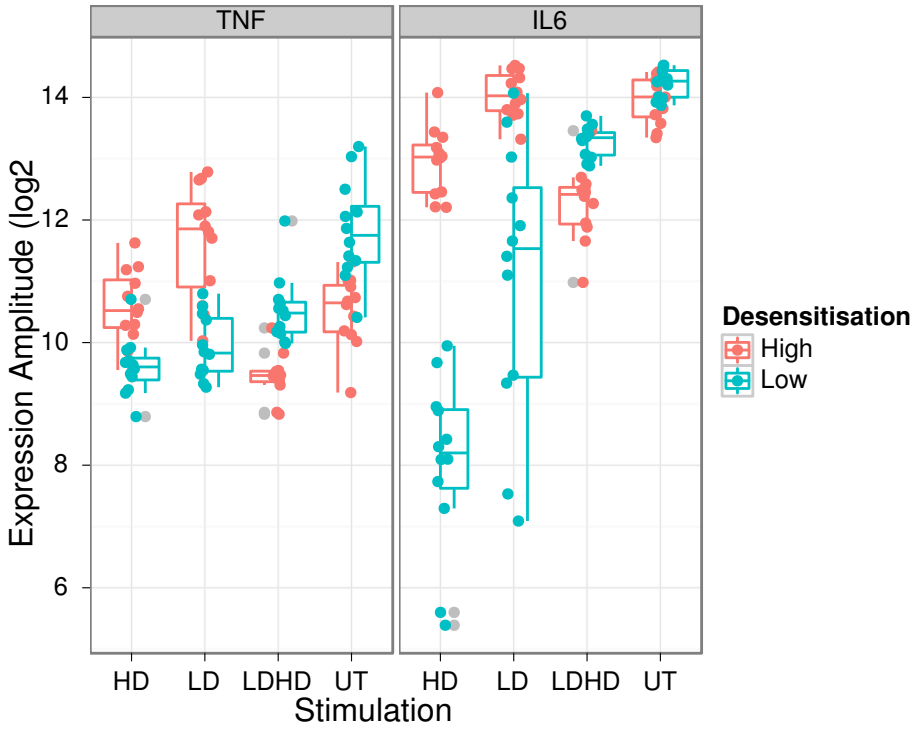


Figure 14.6: Bimodal expression of *TNF* and *IL6* expression with LPS stimulation. UT = naive cells, LD = stimulation with low dose LPS, HD = stimulation with high dose LPS, LDHD = stimulation with high dose LPS in cells pretreated with low dose LPS, Low = low desensitisers (non-tolerant cells), High = high desensitisers (tolerant cells).

Regression model

With a 1.5 fold change in mean expression and a FDR of 0.05 as threshold, the gene expression signature for the endotoxin tolerance phenotype consists of 470 probes, of which 292 are induced and 178 are repressed. This corresponds to 446 genes of which 276 are induced and 170 are repressed. The phenotype is associated with large effect sizes (Figure 14.9).

The genes with the biggest effect sizes are listed in Table 14.6. The list includes a combination of immune related genes, chemokines, and transcription factors. Other notable findings include the induction of *PROCR* (in view of the controversy surrounding activated protein C use in sepsis), the repression of gene cluster in chromosome Y, and class II MHC genes *HLA-DRA* and *HLA-DPA1* albeit with a lower mean expression difference threshold.

Gene	Chr	Function
RBP1	3	retinol binding protein 1, cellular
XIST	X	X (inactive)-specific transcript (non-protein coding)
CLDN23	8	claudin 23
NEFH	22	neurofilament, heavy polypeptide
C19orf59	19	chromosome 19 open reading frame 59
LRRC38	1	leucine rich repeat containing 38
SERPINB7	18	serpin peptidase inhibitor, clade B (ovalbumin)
SCG5	15	secretogranin V (7B2 protein)
MUCL1	12	mucin-like 1
CXCR7	2	chemokine (C-X-C motif) receptor 7
EMR3	19	EGF-like module containing, mucin-like, hormone receptor-like 3
PMP22	17	peripheral myelin protein 22
PAPLN	14	papilin, proteoglycan-like sulfated glycoprotein
PTGES	9	prostaglandin E synthase
CA12	15	carbonic anhydrase XII
PROCR	20	protein C receptor, endothelial
TFPI2	7	tissue factor pathway inhibitor 2
TGM2	20	transglutaminase 2
IL17RB	3	interleukin 17 receptor B
CACNA2D3	3	calcium channel, voltage-dependent, alpha 2/delta subunit 3
NT5E	6	5′-nucleotidase, ecto (CD73)
PPBP	4	pro-platelet basic protein
MT1F	16	metallothionein 1F
TPST1	7	tyrosylprotein sulfotransferase 1
RPS4Y1	Y	ribosomal protein S4, Y-linked 1
UTY	Y	ubiquitously transcribed tetratricopeptide repeat gene, Y-linked
EIF1AY	Y	eukaryotic translation initiation factor 1A, Y-linked
PRKY	Y	protein kinase, Y-linked, pseudogene
TTTY15	Y	testis-specific transcript, Y-linked 15 (non-protein coding)
FCN1	9	ficolin (collagen/fibrinogen domain containing) 1
PTX3	3	pentraxin 3, long
FEZ1	11	fasciculation and elongation protein zeta 1 (zygin I)
HAMP	19	hepcidin antimicrobial peptide
EGR2	10	early growth response 2

Table 14.6: Genes associated with endotoxin tolerance (4 fold change, p 0.01). Genes in the top panel are induced in the tolerant phenotype, while genes in the bottom panel are repressed.

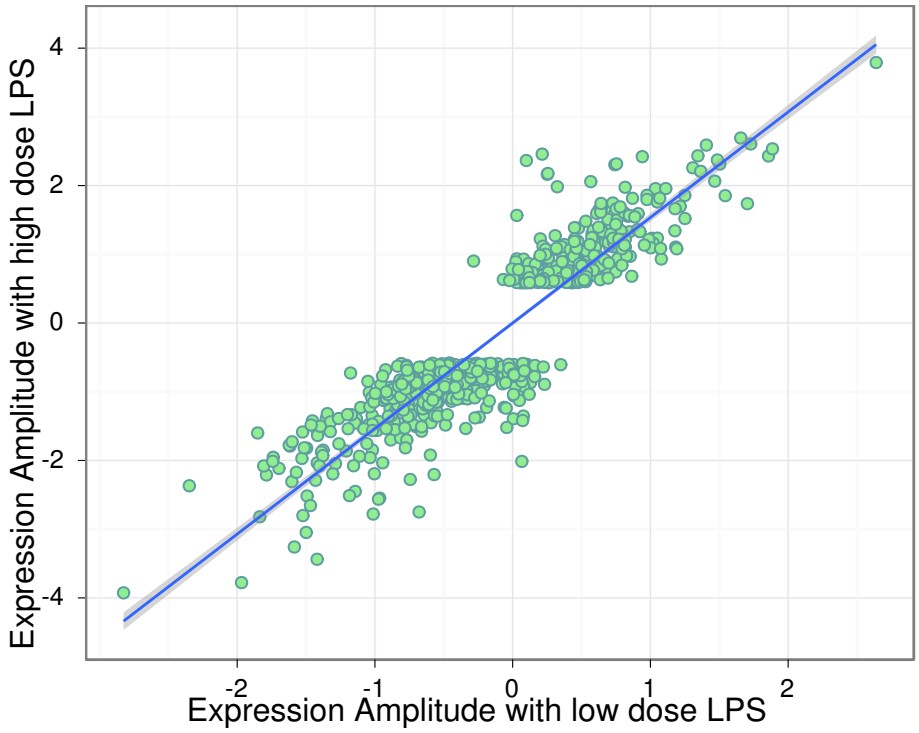


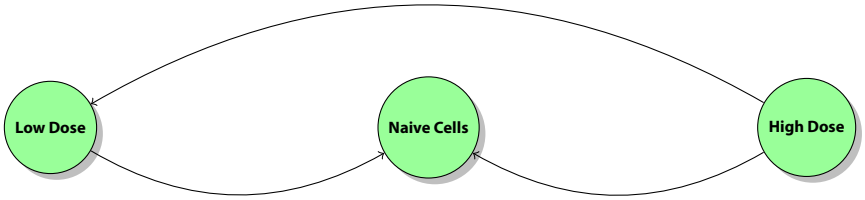
Figure 14.7: Correlation between probes that are differentially expressed in the response to low dose and high dose LPS.

Pathway analysis

Gene sets significantly enriched for the endotoxin tolerance phenotype were determined using the gene set enrichment algorithm. Significance was determined using permutation involving phenotype switching. The significant gene sets are listed in Table 14.7. A number of these gene sets are also associated with sepsis; these gene sets are examined in the next section.

Leading edge analysis identifies genes that contribute to the enrichment score of each gene set. The leading edge genes includes genes with large and modest effects. The leading edge genes of the steroid metabolic process are shown in Supplementary Figure 9. 33 genes from the antigen processing and presentation pathway constitute the leading edge of this gene set. The leading edge gene list includes a number of genes from the MHC. The expression of these genes with different stimulations is shown in Supplementary Figure 10. The list also contains three heat shock chaperone proteins

Biocarta PPARA
Biocarta VIP
Reactome adherens junctions
Reactome nuclear import of REV protein
Reactome PKA activation
Reactome REV mediated nuclear export of HIV1 RNA
Reactome snRNP assembly
Reactome transport of mature RNA
Reactome VPR mediated nuclear import of PICS
Reactome transport of RNP into host nucleus
Reactome regulation of glucokinase
Hormone secretion
DNA recombination
Ribonuclease activity
Growth factor activity



Biocarta Caspase pathway
Biocarta cell cycle pathway
Biocarta P53 hypoxia pathway
Biocarta HIV NEF pathway
KEGG pentose phosphate pathway
KEGG JAK STAT signalling pathway
Reactome glucose and
sugar SLC transporters
Reactome lipoprotein metabolism
Reactome effects of PIP2 hydrolysis
IL4 receptor in B lymphocytes
Negative regulation of signal transduction
Protein Homodimerization activity
Kinase inhibitor activity

Biocarta cell cycle pathway
KEGG cysteine and methionine metabolism
KEGG VEGF signalling
KEGG JAK STAT signalling
Reactome ADP signalling
Reactome glycogenolysis
Reactome lipoprotein metabolism
Reactome PD1 signalling
Reactome IGF activity
Reactome signal amplification
IL4 receptor in B lymphocytes
Protein kinase inhibitor activity
Mono-oxygenase activity
Oxidoreductase activity

Figure 14.8: Pathways associated with the response to LPS.

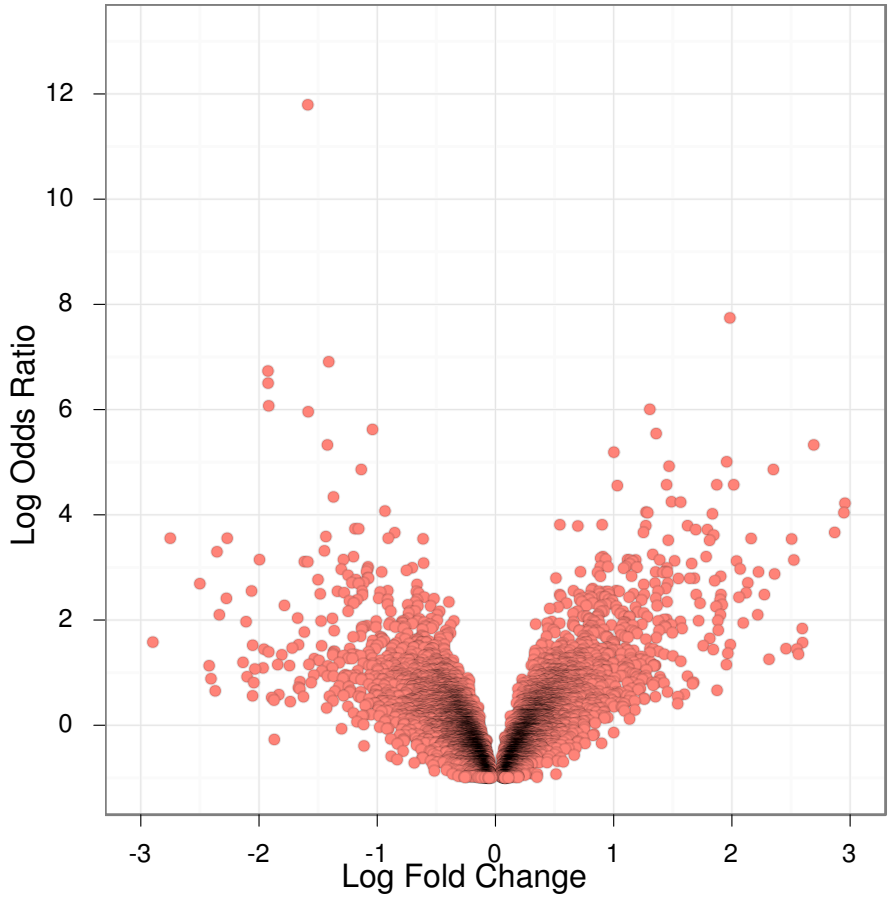


Figure 14.9: Magnitude of the expression changes in the panel of probes that define the endotoxin tolerance phenotype.

(Supplementary Figure 11). The figures show a combination of genes with high and moderate effects.

14.6 Relationship to sepsis

The gene lists and pathways of significance listed in earlier sections have some overlap with the gene lists and pathways associated with sepsis related features listed in earlier chapters. To identify the relationship of sepsis to endotoxin tolerance, two simple approaches are used. First, genes associated with endotoxin tolerance are tested for significant differences in mean expression in contrasts of interest in sepsis, especially

Gene set	ES
Biocarta UCALPAIN	0.70
KEGG arachidonic acid metabolism	0.54
Reactome dual incision reaction in TC NER	0.52
Reactome phase 1 functionalization of compounds	0.51
Steroid metabolic process	0.49
Cell cell adhesion	0.48
Apical junction complex	0.70
Anion transmembrane transporter activity	0.58
Sulphuric ester hydrolase activity	0.86
Serine type endopeptidase inhibitor activity	0.73
Biocarta CTL	-0.81
KEGG antigen processing and presentation	-0.59
KEGG Type I DM	-0.63
KEGG allograft rejection	-0.65
KEGG graft vs host disease	-0.74
KEGG viral myocarditis	-0.50
Reactome lymphoid-non lymphoid interactions	-0.58
Reactome MTOR signalling	-0.54
Cellular defense response	-0.54
Antigen binding	-0.64

Table 14.7: Gene sets associated with endotoxin tolerance

survival. Second, the behaviour of pathways associated with endotoxin tolerance are examined in the context of sepsis, again mainly with respect to survival.

Genes

With a regression model, using a 1.5 fold mean expression change and a FDR of 0.05 as threshold, 883 and 470 probes define the survival and tolerance signatures respectively (corresponding to 772 and 446 unique genes). Of these, only 41 genes are shared between both the lists (Figure 14.10). The overall correlation between the expression values is poor (Pearson correlation coefficient 11.2%), even though some genes seem to correlate closely (Figure 14.11).

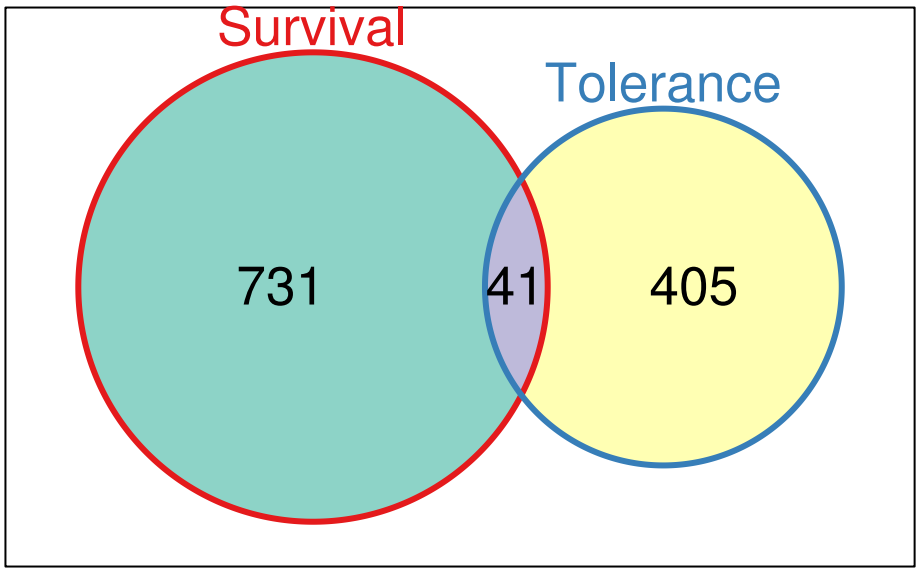


Figure 14.10: Overlap of the probes that are differentially regulated in sepsis survival and the endotoxin tolerance response.

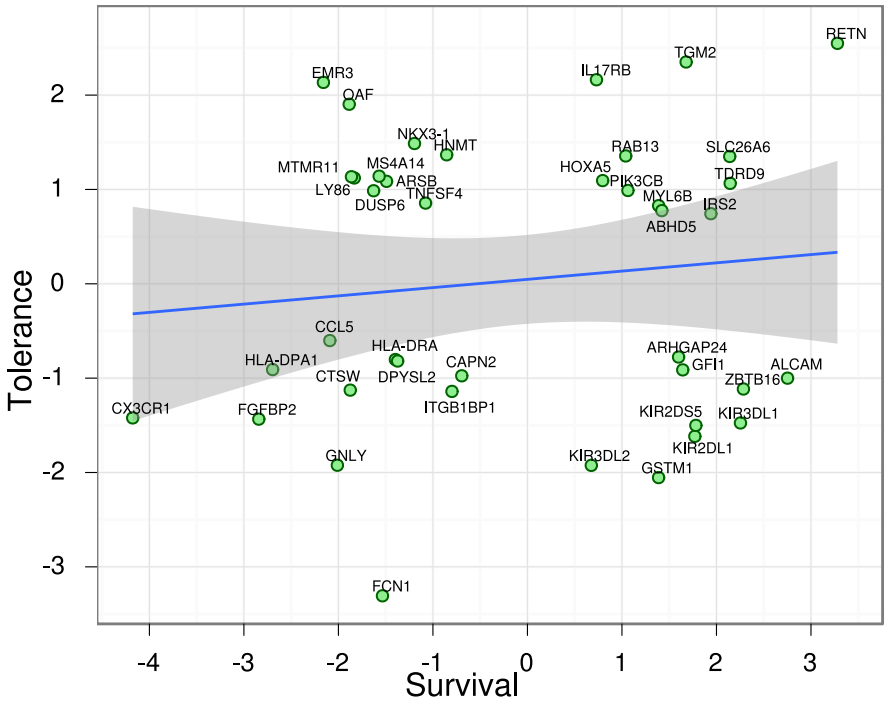


Figure 14.11: Correlation between the expression of the genes that are differentially expressed in both sepsis survival and the endotoxin tolerance response

Of the 1177 genes that are significantly different in either of the two comparisons, 940 genes are present in both data sets. The expression of these genes across the two data sets for the contrasts of interest are shown in Supplementary Figure 12. There is no correlation between the gene expression signatures of survival in sepsis and tolerance.

The lack of correlation between endotoxin tolerance and survival in sepsis could reflect a true lack of association, whereby endotoxin tolerance is not an important determinant of outcome in clinical sepsis. However, this could be a false negative inference for the following reasons. The endotoxin tolerance phenotype described here is cell type specific and was determined under controlled experimental conditions. In contrast to the sepsis dataset, there are no significant confounders in the endotoxin tolerance dataset. In addition, gene expression in the sepsis cohort is modulated by the in-vivo interactions between immunological systems. In contrast, the endotoxin tolerance model is an isolated cell type model. Direct comparison to the sepsis dataset, which is derived from mixed leucocyte population and has a number of significant covariates, might not be a valid approach. There are also additional, well documented problems when comparing expression data from different microarray platforms.

These problems can be partly overcome by using the coefficients for survival from the Bayesian hierarchical model, instead of the raw expression data. The biases due to cell type specificity, however, do remain.

With a 1.2 fold change in mean expression as threshold, and a Bayesian 95% credible interval not including zero, 240 probes from 215 genes were associated with survival in the Bayesian hierarchical, after correcting for important confounders. Only 13 of these genes are also significantly associated with endotoxin tolerance (Figure 14.12). The mean change in expression of these genes in the tolerance phenotype correlates well with their corresponding Bayesian coefficients (Pearson correlation coefficient -0.75, p = 0.003, Figure 14.13).

Of the 215 genes associated with sepsis mortality in the Bayesian analysis, only 155 are represented in the tolerance data set. The correlation between the mean expression differences for these genes is shown in Figure 14.14. The correlation is modest (Pearson correlation coefficient -0.28), but is statistically significant (p = 0.0003).

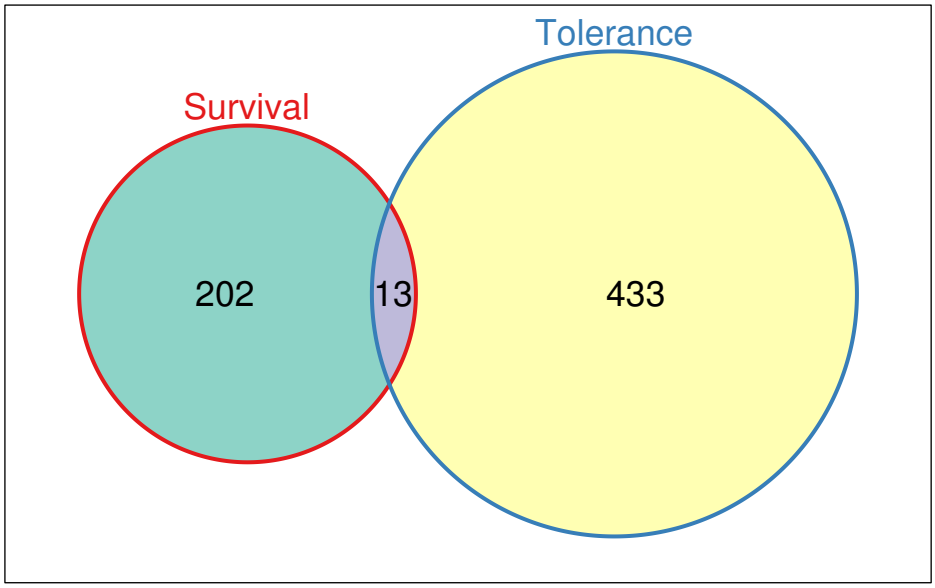


Figure 14.12: Overlap of the genes differentially expressed in sepsis outcomes in the Bayesian analysis, and in the endotoxin tolerance response.

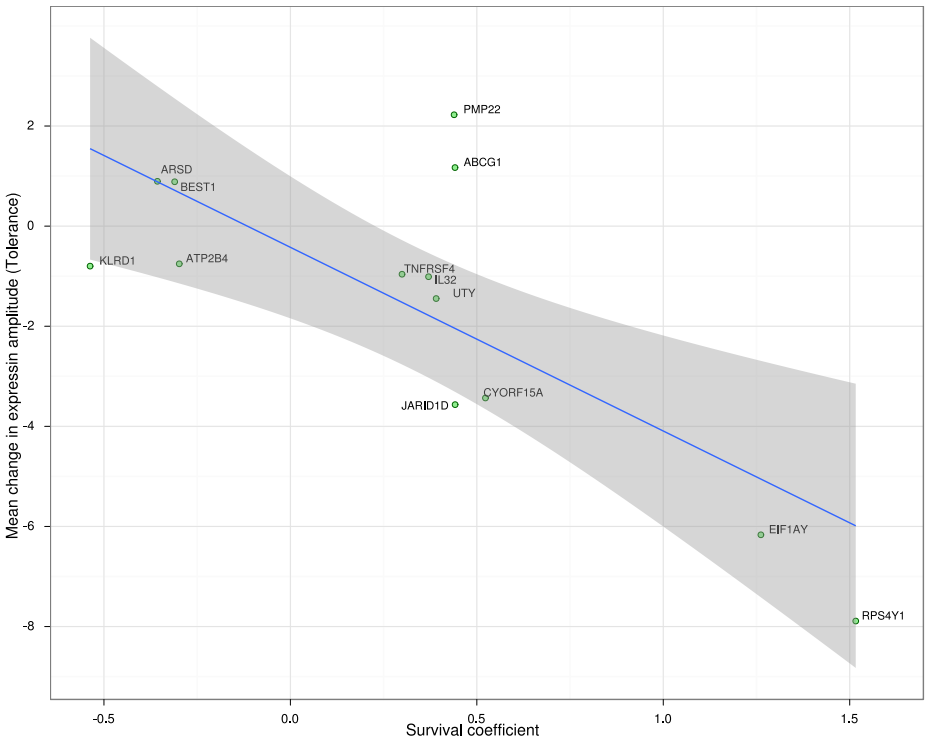


Figure 14.13: Correlation between the genes that are differentially expressed in sepsis outcomes in the Bayesian analysis, and in the endotoxin tolerance response.

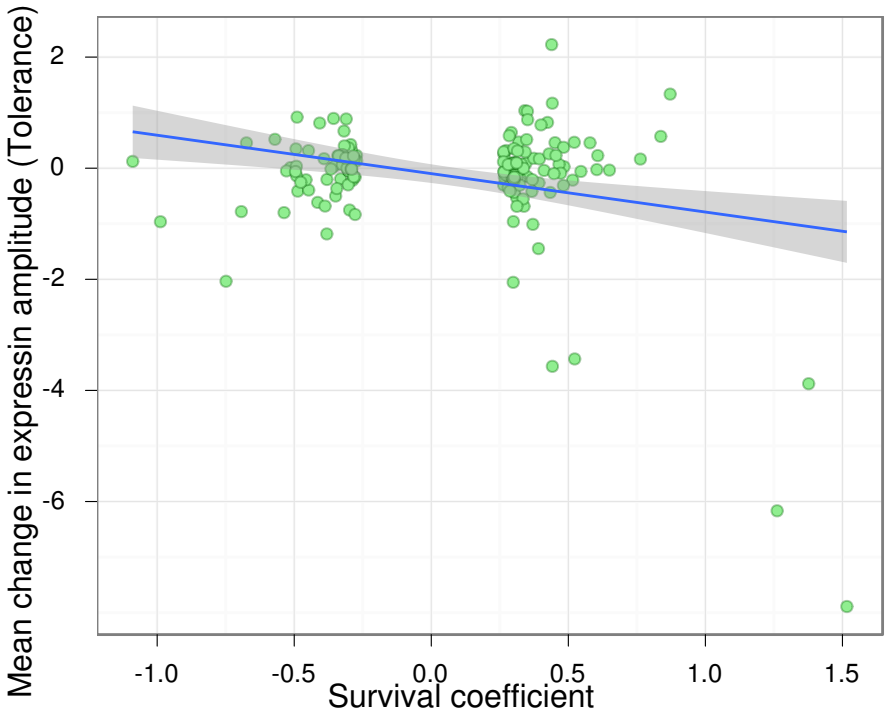


Figure 14.14: Correlation between the genes that are differentially expressed in sepsis outcomes in the Bayesian analysis

The evidence suggests that at least some aspects of the gene expression signature of the endotoxin tolerance phenotype might be related to survival in sepsis. The moderate correlations could be the result of lack of cell type specificity and the effect of confounders in the sepsis data set. However, because of the *a*-priori likelihood of gene overlap in related immune phenomena, the results could also be attributable to spurious associations.

Pathways

It is unlikely that one would be able to detect a precise genomic signature of endotoxin tolerance in patients with sepsis. However, aspects of the immuno-modulatory events related to endotoxin tolerance might still play a role in sepsis outcomes. The behaviour of pathways associated with tolerance in terms of sepsis outcomes is of considerable interest, especially when described in pathways derived from statistical tests that are robust to the effect of confounders. Computational limitations preclude the identifica-

tion of pathways based on the Bayesian coefficients, and this analysis is restricted to the pathways based on statistical tests not corrected for confounders.

Pathways significantly associated with the endotoxin tolerance phenotype were analysed. The leading edge genes contributing to the enrichment score of the gene set was extracted for the tolerance and survival contrasts. The significance of the pathway in relation to sepsis survival was estimated by counting the number of genes that are shared between both the leading edge gene lists, and correlating the mean expression difference for the shared genes.

Results of the analysis in CAP and FP survival are shown in Tables 14.8 and 14.9. Only results where there are at least two shared genes (minimum requirements for correlation tests) are shown. The pattern of association with tolerance related gene sets is different for CAP and FP. This result re-iterates the observation in earlier chapters about the unique gene expression patterns in CAP and FP.

Three gene sets show statistically significant correlations between CAP survival and endotoxin tolerance: KEGG graft vs host response, cellular defence response, and the anionic transmembrane transporter activity. Of these the number of shared genes between the two contrasts is low in the cellular defence response and anionic transmembrane transporter activity. The low number of shared genes, the modest correlations, and non-specific nature of the significant pathways decreases the likelihood of tolerance as a mechanism affecting survival in sepsis.

In FP survival, significant correlations are seen in steroid metabolic process and KEGG graft vs host disease. In addition there is a trend toward significance in three KEGG pathways: type 1 diabetes mellitus, autoimmune thyroid disease, and allograft rejection. Of these the steroid metabolic process gene set result is due to just 3 shared genes. The other gene sets show a much larger proportion of shared genes, thereby increasing confidence in the results. As with CAP, the associations are still non-specific, but the slightly larger number of associations with greater confidence suggests at least some parts of the FP survival outcome may be related to endotoxin tolerance.

The discrepancy between CAP and FP survival with respect to endotoxin tolerance could be explained by: different pathophysiological responses in CAP and FP, a temporal

Gene set	Tolerance	Survival	Shared	p	Estimate
Chr15q22	16	20	7	0.26	0.493
Biocarta ctl pathway	5	4	4	0.49	0.510
KEGG antigen processing and presentation	33	29	13	0.65	0.141
KEGG type I diabetes mellitus	16	21	12	0.21	-0.387
KEGG allograft rejection	16	22	12	0.21	-0.387
KEGG graft versus host disease	21	23	15	0.05	-0.519
KEGG autoimmune thyroid disease	14	21	11	0.17	-0.449
KEGG viral myocarditis	19	21	10	0.66	-0.158
Reactome dual incision reaction in TC NER	17	27	16	0.16	-0.366
Reactome immunoregulatory interactions	16	22	6	0.44	-0.392
Reactome MTOR signalling	9	7	3	0.53	-0.673
Cellular defence response	10	23	7	0.03	0.792
Proteinaceous extracellular matrix	12	11	5	0.46	-0.435
External side of plasma membrane	3	12	3	0.87	0.207
Membrane	278	277	61	0.40	0.110
Organelle outer membrane	6	14	4	0.77	0.234
Extracellular matrix	12	12	5	0.46	-0.435
Synapse	4	17	4	0.58	-0.425
Anion transmembrane transporter activity	7	10	3	0.01	1.000
Sulphuric ester hydrolase activity	8	12	8	0.23	-0.476

Table 14.8: Correlation of pathways between the endotoxin tolerance phenotype and CAP survival

bias where the early sampling in FP might correspond more closely to the time points in the tolerance experiment, and the use of a cluster robust statistic in CAP but not in FP.

The correlation of endotoxin tolerance enriched pathways to cluster specific mortality in CAP is shown in Supplementary Tables 22 to 24. The results show cluster specific differences in associations with endotoxin tolerance.

Mortality in cluster 3 has no associations with any of the endotoxin tolerance enriched gene sets. Only one gene set is association cluster 1 mortality. The cellular defence response gene set correlation is the result of 7 shared genes. Cluster 2 shows the most interesting associations with statistically significant correlations seen with chr15q22 genomic region, KEGG antigen processing and presentation pathway and the antigen presentation gene ontology gene set. The KEGG antigen presentation pathway

Gene set	Tolerance	Survival	Shared	p	Estimate
Chr15q22	16	21	8	0.13	-0.580
Biocarta CTL pathway	5	3	3	0.70	0.451
KEGG arachidonic acid metabolism	8	15	3	0.51	-0.696
KEGG antigen processing and presentation	33	28	14	0.71	-0.109
KEGG type i diabetes mellitus	16	20	10	0.06	-0.606
KEGG allograft rejection	16	18	10	0.06	-0.606
KEGG graft versus host disease	21	19	11	0.04	-0.623
KEGG autoimmune thyroid disease	14	18	10	0.06	-0.606
KEGG viral myocarditis	19	19	10	0.66	-0.161
Reactome dual incision reaction in tc ner	17	5	4	0.51	-0.490
Steroid metabolic process	13	10	3	0.05	0.997
Proteinaceous extracellular matrix	12	19	3	0.63	-0.548
Membrane	278	276	48	0.39	0.128
Extracellular matrix	12	8	3	0.62	-0.566

Table 14.9: Correlation of pathways between the endotoxin tolerance phenotype and FP survival

shows 62% correlation in 12 genes (30% of the tolerance leading edge gene set and 50% of survival leading edge gene set). This biological mechanism represented by this pathway is non-specific, and it is difficult to draw strong conclusions from this associations. However, the same pathway is not correlated in clusters 1 and 3 despite 13 and 10 shared genes respectively. This re-iterates the conclusions of unique genomic expression patterns in the clusters and different mechanisms of mortality. However, the relevance of endotoxin tolerance as a mechanism of poor outcome in severe sepsis is likely to be circumscribed, even in cluster 2.

14.7 Summary

Endotoxin tolerance is a well defined immune phenomenon that is consistently identified in human, animal, and cell line models of infection. The phenomenon can be defined on the basis of suppression of TNF protein production by immune cells previously exposed to an inflammatory stress. The genomic response of PBMCs derived from

healthy volunteers to different doses of LPS and to a second exposure to LPS was studied by hybridising mRNA transcripts to Illumina WGA6-v3 BeadArray platforms.

PBMCs exposed to LPS display a dose-dependent genomic response. However, the response to high dose LPS is not merely an exaggerated low-dose response, but appears to be a distinct entity with significant qualitative differences from the low-dose response. The phenotype of endotoxin tolerance is associated with the differential regulation of a relatively small panel of 446 genes, which relate to a number of pathways related to immune function, such as antigen processing and presentation, antigen binding and cell adhesion.

The relevance of the genomics of endotoxin tolerance in clinical sepsis is difficult to infer in this analysis, due to differences in cell types from which the mRNA transcripts are derived, and differences in the microarray platform to which the transcripts were hybridised. However, the gene expression determinants of survival, identified in the Bayesian analysis which partitions the observed variance across a number of covariates, show strong correlations with the determinants of endotoxin tolerance. Similarly, gene sets related to endotoxin tolerance show modest enrichments in sepsis outcomes with modest correlations between the leading edge genes in each of the phenotypes.

Part V

Conclusions and Discussion

15

Conclusions and Discussion

This work emphasizes the complexities of sepsis pathophysiology and the difficult bioinformatic issues involved in the analysis of complex data sets with multiple modulating influences.

15.1 Conclusions

Heterogeneity in Sepsis

Sepsis is treated as a homogeneous entity while considering clinical management and in most research related to clinical sepsis. Results from the previous literature and the work presented here demonstrate multiple sources of heterogeneity. In particular,

- Equivalency cannot be assumed either in terms of clinical behaviour or underlying cellular events among different diagnostic subgroups based on source of infection. The demographic data from this study, identifies significant differences between CAP and FP, both in outcomes and in phenotypic features. The genomic response also shows clear differences between the cohorts when compared to controls and in gene expression signatures of survival in each of the two cohorts.

- Even in the clinically homogeneous cohort of patients admitted to ICU with severe sepsis due to CAP, there is evidence of subclasses based on gene expression profiles. These profiles are associated with differences in clinical covariates and outcomes. These subclasses might be an universal phenomenon in sepsis and can be demonstrated in the FP cohort with lesser certainty.

A similar subclustering has been reported in paediatric sepsis based on a panel of nearly 7000 genes (Wong *et al.,* 2009b). In contrast, the subclasses identified in this work were derived from about 2000 genes which showed the greatest variance among a total of more than 4000 probes. Less than 1000 genes were differentially expressed between the clusters with the regression and Bayesian analyses. In contrast to the paediatric data set, the clusters were derived on the principal components rather than the raw expression data thus maximizing feature discoverability.

The clusters in the paediatric subset showed associations with steroid use, and Gram positive infections. The possibility that these clusters could be artefactual was not satisfactorily excluded, since the authors did not examine the influence of other covariates on gene expression. In contrast, in the work reported here, the effects of multiple covariates were examined in detail and significant associations were found with cell counts and PF ratio, but not to steroid use and the type of infection. The cluster structure was found to improve after robustly correcting for these covariates and was further validated in the same paediatric data set with a much smaller number of genes.

A pilot analysis (data not shown) of genes distinguishing the clusters, especially the class II MHC genes in B and T cells derived from healthy volunteers, reveals similar clusters in unstimulated cells. This is in agreement with a previously published observation of clusters in healthy subjects based on genes predicting prognosis in autoimmune disease (McKinney *et al.,* 2010). A number of inferences follow from these observations: There are distinct patterns of gene expression in genes related to inflammation and the immune response even in health. These could represent tendencies towards distinct gene expression patterns in response to an inflammatory stimulus, each having different associated risks and outcomes. The existence of these clusters in health is strong evidence for a genetic basis, and associations with specific SNPs or haplotypes is likely. The identification of these variants and an explanation of their evolutionary significance would be of considerable interest.

In conjunction with previous literature, the work presented in this thesis clearly demonstrates heterogeneity in sepsis relating to distinct patterns of gene expression. The distinct pathophysiological profiles that result from these differences could explain

the inconsistencies associated with clinical studies and has important implications in future study design, particularly for randomised controlled trials.

There are some interesting comparisons between this data set and the genome signatures associated with dengue shock syndrome (Devignot *et al.,* 2010) The dengue shock syndrome signature was derived from a smaller sample set of 48 patients of whom 19 were shocked, and the rest had non-shock variants of dengue. The mRNA transcripts were extracted from whole blood samples. The FDRs were derived from a non-Bayesian analysis. A relaxed FDR threshold of 0.1 was used without a mean expression threshold. The number of genes that constituted the predictive panel in dengue shock is comparable to the sepsis signatures when equivalent FDRs are used. The dengue signature was derived from data not corrected for differential cell counts and showed neutrophil specific effects (specifically the increased expression of *MMP8*, *CECAM6*, and *CECAM8*.). These genes and pathways were also seen in analogous comparisons in the sepsis data set. Intriguingly, some of these genes continue to show significant differences in expression after correction for neutrophil counts in the Bayesian hierarchical models. A host of metabolic pathways (specifically lipid related) were significantly enriched in dengue shock syndrome, which was also the case in the sepsis data set. There was also enrichment of the T cell activation pathway which is seen in the sepsis and the cardiac data sets, suggesting a common element of the host response to infection. The authors of this study did not attempt to create a predictive model, but given the small observed effect sizes, it is unlikely that the model would have had adequate predictive power.

The use of a large number of transcripts in the classification algorithms is a typical feature of genomic level classification of clinical phenotypes. The smallest successful genomic prediction panel involved nearly 400 transcripts and was used to predict human tuberculosis with high bacterial load. As with the viral predictors, this panel displayed considerable overlap with latent tuberculosis, significantly restricting its predictive power.

The Sepsis Response

The genomic events in sepsis were described by comparing sepsis samples to control patients who were scheduled to undergo cardiac surgery. This is in contrast to previously published studies involving adult patients in which patients admitted to ICU without sepsis were used as controls (Tang *et al.,* 2008, 2009). This could explain the low gene yield in these studies (less than 100) compared to the work presented here in which between 250 and 1000 genes were associated with the sepsis response. This is still lower than the paediatric studies of sepsis in which more than 2400 genes differentially expressed genes were identified (Wong *et al.,* 2007). The reason for the greater gene yield in the paediatric population is not clear, but could represent the relative lack of the cumulative environmental influences in the paediatric genome compared to the adult genome. This hypothesis would account for the "noisiness" of the adult sepsis dataset. Interestingly, if there is a strong environmental component to the sepsis response in adults, then there is a possibility that the response could be modulated towards favourable outcomes by directed interventions.

There is considerable overlap between the gene sets identified in previous paediatric and adult studies of sepsis and in this thesis. In particular, activation of the immune response, TLR signalling, cytokine responses and the ubiquitin response are consistently identified across data sets. Ubiquitin is now known to be a regulator of a number of immune responses with both intra-cellular and extra-cellular roles. Extra-cellular ubiquitin has been associated with decreased production of TNF as a consequence of reduced transcript level, and ubiquitin inhibition abolishes endotoxin tolerance in clinical sepsis (Majetschak *et al.,* 2003) Ubiquitin is also a key mediator of selective autophagy and influences proteasome mediated degradation (Kraft *et al.,* 2010).

The paediatric data sets identify complement cascade as an important component of the sepsis response. In this data set this association is only seen in cluster 3 of CAP and in FP. The complement cascade is also differentially regulated in sepsis when compared to the SIRS following cardiac surgery. The implications of these inconsistent associations are not clear.

In contrast to the inconsistent associations seen with canonical events of the immune response like cytokine release and the complement cascade, the analysis reported here also identifies a strong and consistent roles for ribosomal proteins and other regulatory events in sepsis. It might be more useful to define and investigate sepsis as perturbations in regulatory networks than the more conventional definitions based on clinical features and cytokine levels.

An intriguing finding, in this thesis, is the differential regulation of viral transcription related pathways in sepsis. Whether this association is the result of a true concomitant viral infection or the involvement of relevant genes in the absence of viral infection cannot be ascertained in this dataset. However, the association of viral pathways in only some of the contrasts and not in others, and the overall qualitative differences in the inflammatory patterns in these contrasts, suggest the likelihood of concomitant viral infections. The Bayesian analysis identified a pattern of gene expression specific to confirmed viral CAP but the definitive confirmation of these gene expression patterns in a suitable prospective experimental design is required. The effect of concomitant viral infections in clinical sepsis has not been examined in detail, but there is some evidence of increased risk of bacterial infections following a viral illness (Beadling *et al.,* 2004) and increased mortality in patients with co-existing viral infections (Lee *et al.,* 2010; Paddock *et al.,* 2012). Similarly, while there is no direct evidence of viral infections altering gene expression in sepsis, there is evidence for role of viral related genes in autoimmune diseases (Cadwell *et al.,* 2010).

A number of genes related to the heme-oxygenase system are differentially regulated in sepsis. The heme-oxygenase system has been related to production of NO, an important mediator of vasomotor stability in sepsis (Maines, 1997). The system also has immune functions related to oxidative injury (Wagener *et al.,* 2003). In addition, induction of heme-oxygenase reverses the pro-thrombotic state in mouse caecal ligation and puncture (CLP) and is associated with the increased production of activated protein C (Fei *et al.,* 2012).

Survival in Sepsis

There is a paucity of literature relating gene expression profiles to mortality in sepsis. The paediatric data sets identify genes encoding zinc binding proteins as showing consistent associations with outcome (Wong *et al.,* 2009a). This could not be confirmed in the work presented here. It is possible that this association is unique to paediatric sepsis.

A panel of 700 genes is associated with death after correcting for clustering in the data set reported here. Within each subclass there seem to be unique mechanisms associated with the risk of death. An interesting finding is the consistent association of metabolism related pathways with outcome risk in sepsis. Death was also associated with suppression of adaptive immunity. Interestingly, in paediatric sepsis this pathway was specifically associated with sepsis shock and not sepsis or SIRS (ibid.).

Covariates

Sepsis risk is modulated by age and sex. Age is known to affect various immune processes and is a predictor of mortality risk in the commonly used risk prediction scores. However, the relationship between age and outcome in the sepsis data set using the Bayesian hierarchical model is modest at best. This is at variance with previous literature where age is associated with immune cell dysfunction, and reduced production of cytokines resulting in a reduced inflammatory response (De Maio *et al.,* 2005). This could be because of the more inclusive model in the Bayesian analysis where age related factors are included explicitly as predictors. In particular, cell counts, cell type effects, and organ function effects are separately represented in the model.

In contrast, 31 genes show significant sex specific effects. There are only a few studies investigating the effect of sex on whole genome expression in sepsis. In a mouse model of *Coxiella burnetti* infection, 86% of the transcripts extracted from liver cells after 24 hours of infection showed sex specific changes (Textoris *et al.,* 2010). The number of genes associated with gender in the sepsis data set presented here is modest, but as with the mouse model shows associations with cell adhesion and signaling mechanisms.

Other covariates have a much larger effect on gene expression. The strongest effects are seen with relative cell counts and renal failure. There is a considerable overlap amongst the genes that are associated with relative cell counts. This is similar to findings in paediatric sepsis (Wong *et al.,* 2010a).

Gene expression profiling in adult sepsis has failed to identify signatures specific to the type of infection (Tang *et al.,* 2008, 2009). With a more robust analysis, the work presented here identifies significant differences between the infectious aetiologies. A panel of 31 genes differentiates Gram positive from Gram negative infections for example, while more striking changes are seen between bacterial and viral infections.

One interesting finding is the induction of oligoadenylate synthetases in atypical and bacterial infections with the suggestion of shared and independent regulatory mechanisms. Recently published work suggests that these synthetases are induced by Type 1 IFN, and there is differential regulation of these genes through the transcription factors STAT1 and STAT2 (Pulit-Penaloza *et al.,* 2012). The external validation of these observations is strong evidence for the robustness of the analysis and accuracy of the results.

Predictive models

A predictive model with 100 genes was developed to predict the subclasses identified in the paediatric data set. This model had a predictive accuracy of 91%. Using a different classifier algorithm, a panel of 20 genes was identified that could predict the CAP cluster membership with a less than 10% misclassification rate. Most of the misclassifications are from one cluster. A similar observation was found with the models that predict diagnosis, with poor prediction accuracy for FP.

Mortality predictions were generally disappointing in the present study. This is in concordance with the limited published literature that utilize genomic models to predict outcomes. One explanation could be the influence of variables that were not included in the model. For example, even in infections caused by the same bacterial species, the presence of a virulence factor can result in a 20% increase in absolute mortality (Nickerson *et al.,* 2009). Clinically useful predictive models such as APACHE

II and SAPS II are composite models including organ function markers as predictors. Genomic models either incorporate this information as uncorrected expression data for organ dysfunction or use a corrective model that makes it difficult to examine for factor interactions. A composite model where the genomic data adds to the current clinical models is the most accessible solution to this problem. However, clinical application of such a model will require the identification of a small panel of transcripts whose expression can be measured with ease at the bedside.

Limitations

An important limitation of the work presented here, and many other gene expression studies, is the inference of protein function from transcript levels, especially in the interaction related analyses using gene sets and pathways. This inference is unjustified because:

- Concordance between transcript and protein levels cannot be assumed. Post-translational modification of transcripts and transcript stability can alter protein levels.
- Even if there were no post-translational modifications, the relationship between transcript and protein level, and the subsequent protein kinetics is exponential. This means that the protein levels can not be inferred from transcript levels and that similar transcript levels can result in proteins whose effects are separated in time and space (Kholodenko *et al.,* 2010).
- Post-translational modifications can result in proteins that have different functional characteristics. For example alternate splicing can convert membrane based proteins to soluble forms, and can even change the effect of receptors from agonist to antagonist.
- Protein effects depend on cellular context. For example, DAP12 can function as an agonist and an antagnoist in the same cell type for the same external stimulus based on the presence of a high affinity ligand as compared to a low affinity ligand (Turnbull *et al.,* 2007).

- Protein effects depend on protein levels in a non-linear manner. For example, the degradation of IκBβ allows the transcription of NF-κB. However, paradoxically this effect is lost when IκBβ is completely inhibited (Rao *et al.,* 2010).

15.2 Future Work

Future work arising from the findings detailed in this thesis can be classified into three inclusive groups: the general framework of which this work forms a part, overcoming limitations, and examining hypotheses generated from this work.

Framework

As with other pathophysiologies, the observed cellular and phenotypic effects of sepsis are the cumulative result of a number of layers with complex interactions. These include 1. the genotypes, with the various structural variations and their interactions with packaging proteins, transcriptional and regulatory events that influence the production of RNA transcripts, 2. events at the transcript level including post-translational modifications, 3. events affecting transcript stability, 4. protein interactions including intracellular effectors such as enzymes, 5. membrane effects such as receptors, and organised membrane activity such as lipid rafts, and 6. extracellular effectors like cytokines and other modulatory and signalling proteins. Additional complexity is derived from cell level interactions such as the antigen processing, presentation and other immuno-modulatory events in the lymph nodes, tissue level interactions such as the interactions between leucocytes, endothelial cells, and the sub-endothelial matrix. In sepsis, especially in relation to mortality, there are additional organism level factors such as age related degradation in tissue function.

Clearly the genome level study of the sepsis response and the signatures of mortality in severe sepsis is only the first step in unravelling the complexity of this syndrome. This work is only a part of a larger investigative framework, examining the genetic basis of clinical sepsis.

Genome wide association study

The primary focus of the Genomic Advances in Sepsis (GAinS) study is to identify genotypes associated with risk of severe sepsis due to CAP and FP and mortality in severe sepsis. Analysis of the GWAS results is in progress; preliminary findings indicate modest associations with relevant clinical outcomes. Interpreting these associations in conjunction with the functional genomics data will help the identification of relevant candidate regions and SNPs for a more focussed interrogation.

Expression quantitative trait localisation

The results presented here can be further enhanced by conducting a genome wide expression Quantitative Trait Localisation (eQTL) analysis for the functional genomics data. Genome wide genotyping data is already available, and the overall success of this proposed project will depend on bioinformatics feasibility, selection of appropriate measures as traits, and the ability to place identified results in the larger clinical context.

A related work by our group has already identified cell type specific *c*is and *t*rans eQTLs (in publication), and a similar approach can be used with this clinical dataset. The selection of appropriate measures is slightly contentious, especially with a noisy dataset. Two solutions are possible; the use of principal components as the expression traits, and the use of "normalised" expression values. Both these solutions have been successfully used in this work; the former with the hierarchical agglomerative clustering of principal components, and the latter through robust Bayesian hierarchical modelling. The unique value of this work will be the integrative approach where any associated variation can be placed in the context of the GWAS results, gene expression data, clinical outcomes, and selected serum protein levels using stored plasma.

A pilot analysis has already been performed, albeit to the expression data instead of the more robust markers such as the principal components, with a small fraction of the data set. A representative result of the *cis* associations for *TLR4* is shown in Figure 15.1.

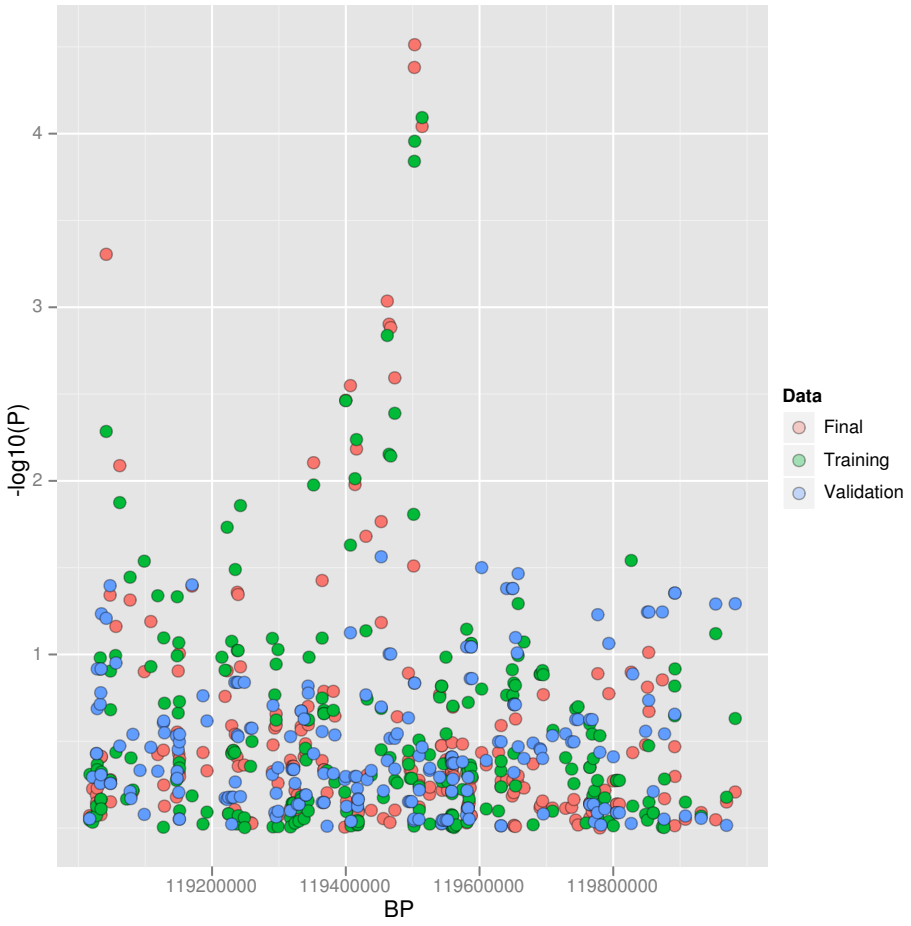


Figure 15.1: cis associations for *TLR4* expression. Genotyping was performed on Affymetrix GeneChip Human Mapping 500K array in 54 patients. Training = Association to expression data from the derivation data set (36 patients), Validation = Association to expression data from the validation data set (18 patients), Final = Association to combined expression data from both arrays with the array platform as a covariate.

Proteomics

The GAinS project resource includes the collection of plasma samples as detailed in an earlier chapter. This resource has not been systematically explored so far, but the value of a combined proteomic and genomic analysis is clear. Even the interrogation of a small number of specific protein targets is likely to be useful. For example, ELISA based quantification of TNF and IL6 in a small subset of the data showed minimal variation in TNF levels, but large variation in IL6 levels (unpublished data). Some of the variation is seen in the same patient, where more than 10 fold change in IL6 protein levels can be demonstrated on different sampling days.

Relating other immune phenomenon to sepsis

This work attempts to use data derived from cardiac surgery and data derived from a cell line model of endotoxin tolerance to form an integrative model of sepsis with varying success. Additional genome level expression data is available from work done by Narelle Magueri and Bert Mohr, where HapMap cell lines were exposed to heat shock. These chaperone proteins were differentially regulated in sepsis and endotoxin tolerance, and an integrative model including heat shock data will be useful. Data is also available from cells subjected to hypoxic stress; a model relevant to clinical sepsis especially in CAP. Dissecting the role of hypoxia would also be relevant, especially in the context of the associations of PF ratio with gene expression and CAP cluster structure identified in this work.

Overcoming limitations

As with any study, this work suffers from a number of limitations, some imposed by logistic necessities and others arising from the analysis. Future work could focus on overcoming these limitations.

Technological issues

The microarray technology to interrogate genome level transcription data is useful but suffers from a number of drawbacks. New technology like RNA sequencing can overcome some of the limitations and improve quantification and data accuracy at the cost of increased financial and possibly computational burden. Use of the new methods at least in a small set of samples could be used to validate the findings and test some hypotheses.

Validation of the CAP clusters

There was evidence of distinct patterns of host response in a relatively homogeneous cohort of patients with severe sepsis due to CAP. The strength of the evidence for these clusters is weakened by difficulty in estimating clusters in the validation cohort, poor predictability, and multiple samples from the same patient. However, there is some

evidence of some sub-structure in the FP cohort, even though the relationships to the CAP clusters in unclear. Validating these clusters in a larger cohort is therefore essential. Also identifying patients whose samples cross cluster boundaries with time, and their clinical behaviour will be useful in determining the relationships between the clusters.

Cell type specificity

Genomic responses to environmental changes exhibit cell type specificity. This work relates changes in gene expression to relative changes in the white cell counts. It will be useful to quantify the cell type specific genomic responses in sepsis, especially in serial samples, and to correlate these effects to clinical outcomes. It will also be useful to extend these to other cell types such as those derived from lymphoid tissue and endothelial cells. Animal models might be a feasible option for the latter.

Hypotheses and Directions

There are a number of hypothesis that can be explored from this data set, especially in prospective experimental designs. Some of these have been described earlier and are summarised below.

There is strong evidence for the presence of distinct subclusters in clinically homogeneous populations of sepsis based on gene expression profiles. Based on evidence presented in this thesis, there is support for the hypothesis that these distinct profiles are associated with sequence variants and haplotypic blocks especially in the class II MHC. It is also likely that these associations are modulated to environmental influences, possibly through epigenetic mechanisms.

The clinical relevance of these clusters should also be investigated. In addition to the characterization of the clusters in terms of clinical covariates, the implications of the clusters in selecting treatment strategies should be examined. It is likely that effectiveness of canonical treatment strategies such as steroids, and novel treatment strategies such as Vitamin D and anti-cytokine therapy would vary in each of the patient classes. Also interesting in clinical situations, particularly in the context of time and

route of feeding in patients with sepsis and tight glycaemic control, is the role of pathways involved in glucose and energy metabolism in sepsis.

Development of predictive models to identify inflammatory patterns and predict outcomes is of considerable potential clinical value. While gene expression patterns fail to provide strong predictive models on their own, a combined approach incorporating organ function effects and biochemical markers along with gene expression profiles may achieve the desired predictive accuracies.

Appendices

Supplementary Tables

Name	City
John Radcliffe Hospital	Oxford
Royal Berkshire Hospital	Reading
Bart's and the London Hospitals	London
Aberdeen Royal Infirmary	Aberdeen
Norfolk and Norwich University Hospitals	Norwich
Southend University Hospital	Southend
Royal Victoria Infirmary	Newcastle

Table 1: Recruiting ITUs

Id	Function
LOC100128056	RNA polymerase II, small phosphatase 2 pseudogene
LOC100129585	similar to hCG2011544
LOC100130904	similar to CD177 molecule
LOC100132528	similar to HNRPA1 protein
LOC100133329	similar to rCG63653
LOC100133390	hypothetical protein LOC100133390
LOC100133565	similar to hCG23738
LOC100133662	hypothetical protein LOC100133662
LOC100134273	similar to mCG7602
LOC136143	ribosomal protein L18 pseudogene
LOC387791	mitochondrial import receptor subunit TOM20 homolog
LOC387825	similar to 23 kD highly basic protein
LOC401676	60S ribosomal protein L13a-like
LOC402251	similar to Hypothetical protein MGC76216
LOC439949	uncharacterised LOC439949
LOC440063	peptidylprolyl isomerase A (cyclophilin A) pseudogene
LOC440341	similar to nuclear pore complex interacting protein
LOC441506	similar to laminin receptor 1
LOC441743	novel protein similar to RPL23A
LOC641768	similar to 40S ribosomal protein S26
LOC642113	Ig kappa chain V-III region HAH-like
LOC642817	uncharacterised LOC642817
LOC642989	similar to 40S ribosomal protein S25
LOC643357	similar to SMT3 suppressor of mif two 3 homolog 2
LOC644113	hypothetical protein LOC644113
LOC644936	actin, beta pseudogene
LOC644946	hypothetical protein LOC644946
LOC645138	similar to ribosomal protein S11
LOC645385	similar to Heterogeneous nuclear ribonucleoprotein A1
LOC645436	heterogeneous nuclear ribonucleoprotein A1 pseudogene
LOC645691	heterogeneous nuclear ribonucleoprotein A1 pseudogene
LOC647030	eukaryotic translation elongation factor 1 beta 2 pseudogene
LOC647307	phosphatidylethanolamine binding protein 1 pseudogene
LOC647450	similar to Ig kappa chain V-I region HK101 precursor
LOC647506	similar to Ig kappa chain V-I region HK101 precursor
LOC647856	similar to laminin receptor 1 (ribosomal protein SA)
LOC648210	similar to Heterogeneous nuclear ribonucleoprotein A1
LOC648249	similar to 40S ribosomal protein SA (p40)
LOC648771	60S ribosomal protein L12-like
LOC649143	similar to HLA class II, DRB1-9 beta chain precursor
LOC649548	similar to 60S ribosomal protein L32
LOC649821	similar to 60S ribosomal protein L14
LOC649839	similar to large subunit ribosomal protein L36a
LOC649923	similar to Ig gamma-2 chain C region
LOC651149	similar to 60S ribosomal protein L3 (L4)
LOC651202	similar to large subunit ribosomal protein L36a

Table 2: LOC genes referenced in the text, but do not have an official symbol or an ortholog

	Overall	Survivors	Non-Survivors	p
Patients	25	15	10	
Samples	56	41	15	
Day 1 samples	20	13	7	
Day 3 samples	20	14	6	
Day 5 samples	16	14	2	
Age (Median/IQR)	71 (16.5)	65.5 (15.3)	75.5 (14)	0.10
Sex	1:1	1.3:1	1:1.5	0.68

Table 3: Demographics of the CAP validation cohort. p values are only shown where relevant.

	Overall	Survivors	Non-Survivors	p
Temperature				
High	37.78 (0.76)	37.89 (0.77)	37.47 (0.68)	0.06
Low	36.21 (0.80)	36.27 (0.83)	36.02 (0.71)	0.29
SBP				
High	144 (20)	143 (18)	148 (26)	0.55
Low	97 (20)	100 (20)	88 (17)	0.05
MAP				
High	97 (12)	96 (11)	99 (16)	0.64
Low	65 (13)	67 (12)	58 (15)	0.05
HR				
High	120 (27)	118 (23)	127 (36)	0.37
Low	81 (15)	82 (16)	75 (13)	0.09

Table 4: Vital signs of the CAP validation cohort. SBP = systolic blood pressure, MAP = mean arterial pressure, HR = heart rate

	Overall	Survivors	Non-Survivors	p
WCC				
High	13.9 (8.0)	14.3 (6.7)	12.7 (11.5)	0.63
Low	11.9 (7.6)	12.1 (5.8)	11.5 (11.7)	0.85
DCC				
Neutrophils	85.3 (8.4)	84.7 (8.4)	87.3 (8.5)	0.35
Monocytes	6.0 (5.9)	6.3 (6.3)	5.3 (4.4)	0.51
Lymphocytes	8.6 (5.8)	9.0 (5.7)	7.4 (6.1)	0.42
Platelets	198 (102)	203 (99)	185 (115)	0.61
Haematocrit	17 (17)	15 (18)	20 (14)	0.38

Table 5: Cell counts in the CAP validation cohort. WCC = white cell count, DCC = differential cell count

	Overall	Survivors	Non-Survivors	p
Sodium				
High	140.8 (5.1)	141.7 (4.9)	139.1 (5.3)	0.31
Low	136.9 (4.9)	136.8 (4.2)	137.0 (6.3)	0.96
Potassium				
High	4.8 (1.0)	4.9 (1.1)	4.7 (0.6)	0.63
Low	3.8 (0.7)	3.6 (0.6)	4.0 (0.9)	0.35
Bilirubin	13.6 (10.9)	12.8 (9.5)	15.5 (14.0)	0.53
pH	7.28 (0.14)	7.29 (0.12)	7.27 (0.19)	0.81
Bicarbonate	23.9 (6.4)	25.3 (6.2)	21.6 (6.4)	0.15

Table 6: Biochemical markers in the CAP validation cohort

	Overall	Survivors	Non-Survivors	p
Urea	26.6 (4.3)	32.9 (5.4)	16.3 (7.4)	0.22
Urine Volume	1374 (1328)	1444 (1267)	1056 (874)	0.19
Creatinine				
High	198 (143)	206 (159)	173 (70)	0.30
Low	218 (151)	255 (178)	256 (59)	0.04
Respiratory Rate	33 (10)	33 (10)	32 (12)	0.73

Table 7: Markers of renal and respiratory function in the CAP validation cohort

	Overall	Survivors	Non-Survivors	p
SAPS II	66 (18.5)	63 (17)	70 (20)	0.49
Predicted Death Rate	78.5 (31.3)	73.6 (29.7)	83.8 (30.8)	0.82

Table 8: Outcome risk in the CAP validation cohort

Unique to Cluster 1						
ARL5B	ARRDC4	C14ORF100	C1ORF25	C2ORF76	C7ORF28B	CAPZA2
CEPT1	DNAJB14	ELMOD2	FAM160B1	FNDC3B	HERC4	HS.174794
KCNJ2	KIAA1600	LPCAT2	MIER1	MRPL33	MS4A6A	MYL6
POMP	PRKAA1	PSMA4	RAB10	RAP2C	RNASET2	RTN4
SELT	SLC2A3	ST8SIA4	SULT1B1	TMX1	TNPO1	UBE2A
USMG5	USP33	ZFP106				
Unique to Cluster 2						
GOLPH4	PVALB	SCGB1C1	XIST			
Unique to Cluster 3						
ATOX1	DAB2	LEMD3	PRICKLE1	RNF122		
Shared between Clusters 1 and 2						
ZNF518A						
Shared between Clusters 1 and 3						
FBXO11	GMFB	LMO4	NEBL	RAB6A	UBA6	
Shared between Clusters 2 and 3						
NGFRAP1						
Shared between all clusters						
AMFR	DCP2	HGD	HLA-DRB1	HLA-DRB5	HRK	KAT2B
MEG3	SCARNA9	SIGLEC14	WTAP			

Table 9: Shared and unique genes up-regulated in the cluster-wise response in CAP sepsis.

Unique to Cluster 1					
BACH2	BCL11B	BEX2	BTBD11	C16ORF30	C2ORF89
C5ORF39	CACNA1I	CCR7	CD2	CD27	CD40LG
CD5	DGKA	DKFZP761P0423	EEF1G	EPHX2	FAIM3
FAM102A	FAM62B	FBXL16	FCRL3	FLJ43681	FVT1
GIPC1	GRAP	GZMK	HAPLN3	HKDC1	HLA-DOA
HS.567469	HS.572649	IL7R	INPP5E	LBH	LDHB
LIPA	RPL10AP6	RPL19P12	RPS29P14	RPL21P125	RPS27P13
RPS27P23	RPL26P6	RPL21P87	RPS27P7	RPS2P9	RPS27P29
RPL26P33	RPS2P28	RPL12P32	LOC100133390	LOC100134273	RPL36P5
LOC136143	FAM153B	RPS12P23	RPL36P14	LOC387825	LOC388076
LOC388339	LOC388344	LOC388707	LOC389141	LOC389156	LOC390183
LOC391019	LOC391126	LOC391656	LOC391833	LOC401676	LOC439949
LOC440341	LOC441073	LOC441154	LOC441550	LOC441743	LOC642741
LOC642828	LOC642989	LOC643007	LOC644029	LOC644464	LOC644511
LOC644790	LOC644928	LOC644937	LOC645138	LOC645173	LOC645231
LOC645688	LOC646688	LOC646909	LOC647030	LOC647276	LOC647285
LOC647307	LOC647436	LOC647856	LOC648249	LOC648343	LOC648771
LOC649548	LOC649821	LOC649839	LOC651202	LOC651453	LOC652624
LOC653314	LOC653881	LOC727821	LOC727984	LOC728031	LOC728126
LOC728244	LOC728481	LOC728484	LOC728782	LOC728820	LOC729208
LOC729301	LOC729340	LOC729500	LOC730255	LOC730754	LOC731096
LRIG1	MAN1C1	MGC87895	MORC2	MPRIP	MYC
NFX1	P2RY10	PEBP1	PECI	PIK3IP1	PLCH2
PRKCH	PRKCQ	PRPF31	RGPD1	RORC	RPL10A
RPL12	RPL13A	RPL13L	RPL23	RPL23A	RPL3
RPL7A	RPL8	RPLP1	RPS18	RPS26P11	RPS3
RPS6	RPSA	SH3YL1	SPOCK2	STMN3	TCEA3
TCF7	TGIF1	TIMM9	TRAF5	TSPAN3	UBASH3A
ZBTB4	ZNF266	ZNF827	ZNF831		
Unique to Cluster 2					
Unique to Cluster 3					
BIN2	CDC42SE2	CDV3	LOC440063	LOC643668	LOC648210
LOC652458	MRPL45	NACAP1	RGPD8	SLC7A6	SMAP1
TCP1	TMED10	TMEM128	TRAT1	UFM1	YTHDC1
ZNF681					

Table 10: Unique genes down-regulated in the cluster-wise comparisons in CAP sepsis

Shared between Clusters 1 and 2					
LOC728823	RAPGEF6				
Shared between Clusters 1 and 3					
ABLIM1	ID3	IPO8	RPS7P10	NPM1P21	LOC100129585
RPS27P9	HNRNPA1P2	RPL36AP44	LOC100132528	RPL4P5	RPL37P6
LOC387791	LOC388654	LOC402251	LOC441506	LOC442270	LOC642817
LOC643873	LOC645385	LOC645436	LOC645691	LOC646785	LOC651149
LOC653773	LOC728060	LOC728643	LOC728732	LOC730029	MAL
MGC15763	NELL2	NKTR	RGPD2	RPS6P1	SF3A1
ST6GAL1	THOC3	TMEM194A	TNFRSF25	ZC3H14	
Shared between Clusters 2 and 3					
EIF1AY					
Shared between all clusters					
CCR6	CYORF15A	D4S234E	FAM134B	FBLN5	GPR128
HS.08889	HS.567464	IL23A	JARID1D	KLHL3	LEF1
LOC100128056	RPS26P54	LOC100133329	LOC100133662	LOC440991	LOC441377
LOC641768	LOC644191	LOC644934	LOC644936	LOC645979	LOC646966
LOC653820	LOC728572	MIR130A	PABPC3	PASK	PRKY
RPS26	RPS26L	RPS26P10	RPS4Y1	RPS4Y2	TBC1D4
TMEM204	TXK	XYLT1			

Table 11: Shared genes down-regulated in cluster-wise comparisons in CAP sepsis

Gene	Chr	Function	Coefficient	95% CI
CLEC7A	12	C-type lectin domain family 7, member A	1.56	0.60 – 2.51
HDC	15	histidine decarboxylase	1.14	0.10 – 2.14
HLA-DRB1	6	MHC, class II, DR beta 1	2.48	0.78 – 4.20
KYNU	2	kynureninase	1.12	0.27 – 1.96
C20orf108	20	chromosome 20 open reading frame 108		
STRADB	2	STE20-related kinase adaptor beta	1.26	0.20 – 2.34
SLC25A39	17	solute carrier family 25, member 39	1.47	0.41 – 2.54
AHSP	16	alpha haemoglobin stabilizing protein	1.50	0.01 – 2.98
FCER1A	1	Fc fragment of IgE	1.36	0.10 – 2.63
MYL4	17	myosin, light chain 4	1.26	0.12 – 2.43
HBM	16	haemoglobin, mu	2.15	0.64 – 3.65
GYPC	2	glycophorin C (Gerbich blood group)	1.19	0.28 – 2.10
C7orf28B	7	chromosome 7 open reading frame 28B		
TSPAN5	4	tetraspanin 5	1.02	0.09 – 1.98

Table 12: Genes discriminating atypical pneumonias from Gram positive infections. Genes induced in Gram positive infections. Coefficients are in log2 scale

Gene	Chr	Function	Coefficient	95% CI
SEPT4	17	septin 4	-1.58	-2.44 – -0.70
IFIT3	10	IFN protein with tetratricopeptide repeats 3	-2.04	-3.76 – -0.29
ERI1	8	exoribonuclease 1	-1.06	-1.64 – -0.44
EPSTI1	13	epithelial stromal interaction 1 (breast)	-1.87	-3.67 – -0.09
XAF1	17	XIAP associated factor 1	-1.62	-3.11 – -0.07
IFI44L	1	interferon-induced protein 44-like	-2.43	-4.69 – -0.14
IFI44	1	interferon-induced protein 44	-1.98	-3.63 – -0.23
GBP1	1	guanylate binding protein 1, interferon-inducible	-1.43	-2.78 – -0.08
HLA-C	6	MHC, class I, C	-1.31	-2.51 – -0.11
COL18A1	21	collagen, type XVIII, alpha 1	-1.22	-1.89 – -0.54
PARP12	7	poly (ADP-ribose) polymerase family, member 12	-1.02	-2.02 – -0.04
SERPING1	11	serpin peptidase inhibitor, (C1 inhibitor)	-2.10	-3.38 – -0.82
PARP14	3	poly (ADP-ribose) polymerase family, member 14	-1.01	-1.95 – -0.05
LY6E	8	lymphocyte antigen 6 complex, locus E	-2.32	-3.82 – -0.77
IRF7	11	interferon regulatory factor 7	-1.09	-1.86 – -0.33
RSAD2	2	radical S-adenosyl methionine domain	-1.84	-3.59 – -0.09
LAP3	4	leucine aminopeptidase 3	-1.40	-2.30 – -0.51
IFI6	1	interferon, alpha-inducible protein 6	-1.83	-3.47 – -0.15
BATF2	11	basic leucine zipper transcription factor	-1.09	-1.87 – -0.25
HERC5	4	hect domain and RLD 5	-1.83	-3.44 – -0.17
DHX58	17	DEXH (Asp-Glu-X-His) box polypeptide 58	-1.07	-2.08 – -0.06
OAS1	12	2′-5′-oligoadenylate synthetase 1	-1.65	-2.91 – -0.40
OAS3	12	2′-5′-oligoadenylate synthetase 3	-1.95	-3.63 – -0.33
OAS2	12	2′-5′-oligoadenylate synthetase 2	-1.81	-3.21 – -0.40
RTP4	3	receptor (chemosensory) transporter protein 4	-1.37	-2.45 – -0.32
HES4	1	hairy and enhancer of split 4 (Drosophila)	-1.23	-2.26 – -0.24
ISG15	1	ISG15 ubiquitin-like modifier	-2.52	-4.32 – -0.69

Table 13: Genes discriminating atypical pneumonias from Gram positive infections. Genes repressed in Gram positive infections. Coefficients are in log2 scale.

Gene	Chr	Function	Coefficient	95% CI
MME	3	membrane metallo-endopeptidase	1.42	0.04 – 2.81
USMG5	10	skeletal muscle growth 5 homolog	1.30	0.20 – 2.42
TPST1	7	tyrosylprotein sulfotransferase 1	1.33	0.10 – 2.62
SCGB1C1	11	secretoglobin	-1.34	-2.19 – -0.47
RBPMS2	15	RNA binding protein with multiple splicing 2	-1.50	-2.92 – -0.07
CCL5	17	chemokine (C-C motif) ligand 5	-1.19	-2.28 – -0.11
JUP	17	junction plakoglobin	-1.47	-2.81 – -0.21
RNY5	7	RNA, Ro-associated Y5	-2.01	-4.05 – -0.01
RNU11	1	RNA, U11 small nuclear	-1.24	-2.26 – -0.21

Table 14: Genes discriminating Gram negative pneumonias from Gram positive infections. Top panel lists the induced genes and the bottom panel repressed genes. Coefficients are in log2 scale.

Gene	Chr	Function	Coefficient	95% CI
IFIT2	10	IFN-induced protein with tetratricopeptide repeats	1.22	0.32 – 2.11
IFIT3	10	IFN-induced protein with tetratricopeptide repeats	1.90	0.55 – 3.22
IFIT1	10	IFN-induced protein with tetratricopeptide repeats	1.93	0.31 – 3.52
EPSTI1	13	epithelial stromal interaction 1 (breast)	1.70	0.42 – 3.02
CCR1	3	chemokine (C-C motif) receptor 1	1.11	0.54 – 1.66
XAF1	17	XIAP associated factor 1	1.46	0.49 – 2.48
IFI44L	1	interferon-induced protein 44-like	2.79	1.13 – 4.36
IFI44	1	interferon-induced protein 44	2.10	0.84 – 3.37
GBP1	1	guanylate binding protein 1, interferon-inducible	1.18	0.21 – 2.18
CFD	19	complement factor D (adipsin)	1.12	0.13 – 2.08
HIST1H4H	6	histone cluster 1, H4h	1.22	0.52 – 1.95
SERPING1	11	serpin peptidase inhibitor	1.25	0.36 – 2.19
SPATS2L	2	spermatogenesis associated, serine-rich 2-like	1.05	0.37 – 1.68
ZBP1	20	Z-DNA binding protein 1	1.05	0.32 – 1.74
LY6E	8	lymphocyte antigen 6 complex, locus E	1.64	0.55 – 2.74
DDX60	4	DEAD (Asp-Glu-Ala-Asp) box polypeptide 60	1.17	0.35 – 1.97
RSAD2	2	radical S-adenosyl methionine domain containing 2	1.77	0.49 – 3.13
SAMD9L	7	sterile alpha motif domain containing 9-like	1.12	0.39 – 1.90
IFI6	1	interferon, alpha-inducible protein 6	1.52	0.25 – 2.79
OTOF	2	otoferlin	2.33	1.57 – 3.08
HBM	16	haemoglobin, mu	1.14	0.04 – 2.24
TRIM22	11	tripartite motif containing 22	1.22	0.48 – 1.98
HERC5	4	hect domain and RLD 5	1.60	0.43 – 2.77
IFI27	14	interferon, alpha-inducible protein 27	3.73	2.16 – 5.24
OAS1	12	2′-5′-oligoadenylate synthetase 1, 40/46kDa	1.72	0.77 – 2.68
OAS3	12	2′-5′-oligoadenylate synthetase 3, 100kDa	1.84	0.66 – 3.05
OAS2	12	2′-5′-oligoadenylate synthetase 2, 69/71kDa	1.44	0.43 – 2.50
RTP4	3	receptor (chemosensory) transporter protein 4	1.10	0.26 – 1.89
HES4	1	hairy and enhancer of split 4 (Drosophila)	1.06	0.29 – 1.82
ISG15	1	ISG15 ubiquitin-like modifier	1.41	0.09 – 2.72

Table 15: Genes induced in Gram positive pneumonias when compared to viral infections. Coefficients are expressed in log2 scale.

Gene	Chr	Function	Coefficient	95% CI
MPO	17	myeloperoxidase	-1.40	-2.62 – -0.21
MS4A3	11	membrane-spanning (hematopoietic cell-specific)	-1.25	-2.41 – -0.10
DEFA4	8	defensin, alpha 4, corticostatin	-1.72	-3.17 – -0.22
HLA-DRB1	6	MHC, class II, DR beta 1	-1.87	-3.14 – -0.62
AZU1	19	azurocidin 1	-1.45	-2.61 – -0.25
PRTN3	19	proteinase 3	-1.27	-2.38 – -0.18
ELANE	19	elastase, neutrophil expressed	-1.89	-3.34 – -0.39
TREML4	6	triggering receptor expressed on myeloid cells	-1.04	-1.81 – -0.25
CD8A	2	CD8a molecule	-1.09	-2.00 – -0.24
CTSG	14	cathepsin G	-1.69	-3.12 – -0.34
TFF3	21	trefoil factor 3 (intestinal)	-1.19	-1.94 – -0.43
NFXL1	4	nuclear transcription factor	-1.07	-1.72 – -0.38

Table 16: Genes repressed in Gram positive pneumonias when compared to viral infections. Coefficients are expressed in log2 scale.

Gene	Chr	Function	Coefficient	95% CI
LAG3	12	lymphocyte-activation gene 3	1.40	0.11 – 2.68
KLRD1	12	killer cell lectin-like receptor subfamily	1.35	0.02 – 2.71
SYTL2	11	synaptotagmin-like 2	1.35	0.43 – 2.29
C20orf108	20	chromosome 20 open reading frame 108		
SESN3	11	sestrin 3	1.35	0.05 – 2.65
XRN2	20	5′-3′ exoribonuclease 2	1.33	0.08 – 2.59
ZNF683	1	zinc finger protein 683	1.34	0.25 – 2.40
VPREB3	22	pre-B lymphocyte 3	1.50	0.25 – 2.71
GZMK	5	granzyme K (granzyme 3; tryptase II)	1.48	0.04 – 2.88
SLC25A39	17	solute carrier family 25, member 39	1.85	0.18 – 3.53
AHSP	16	alpha haemoglobin stabilizing protein	2.47	0.17 – 4.83
ITGB3	17	integrin, beta 3 (platelet glycoprotein IIIa)	1.35	0.05 – 2.68
IL32	16	interleukin 32	1.48	0.03 – 2.94
RBPMS2	15	RNA binding protein with multiple splicing 2	2.04	0.39 – 3.62
HBM	16	haemoglobin, mu	3.20	0.87 – 5.47
JUP	17	junction plakoglobin	1.62	0.13 – 3.18
LPCAT2	16	lysophosphatidylcholine acyltransferase 2	1.97	0.22 – 3.66
SNCA	4	synuclein, alpha (component of amyloid precursor)	1.72	0.06 – 3.41
C7orf28B	7	chromosome 7 open reading frame 28B		
TSPAN5	4	tetraspanin 5	1.50	0.02 – 2.97

Table 17: Genes induced in atypical pneumonias when compared to Gram negative infections. Coefficients are in log2 scale.

Gene	Chr	Function	Coefficient	95% CI
IFIT3	10	IFN-induced protein with tetratricopeptide repeats 3	-2.85	-5.66 – -0.08
MAK	6	male germ cell-associated kinase	-1.98	-3.59 – -0.39
IFIH1	2	interferon induced with helicase C domain 1	-1.68	-3.15 – -0.09
IFITM3	11	interferon induced transmembrane protein 3	-1.61	-3.09 – -0.08
MIIP	1	migration and invasion inhibitory protein	-1.42	-2.53 – -0.35
COL18A1	21	collagen, type XVIII, alpha 1	-1.56	-2.62 – -0.51
BTN3A1	6	butyrophilin, subfamily 3, member A1	-1.39	-2.67 – -0.15
PARP12	7	poly (ADP-ribose) polymerase family, member 12	-1.59	-3.10 – -0.01
SERPING1	11	serpin peptidase inhibitor (C1 inhibitor)	-2.96	-4.93 – -1.02
TMEM140	7	transmembrane protein 140	-1.45	-2.32 – -0.53
XKR3	22	XK, Kell blood group complex subunit-related family	-1.34	-2.68 – -0.05
SCO2	22	SCO cytochrome oxidase deficient homolog 2	-1.37	-2.74 – -0.03
LY6E	8	lymphocyte antigen 6 complex, locus E	-2.56	-4.96 – -0.07
CARD17	11	caspase recruitment domain family, member 17	-1.49	-2.64 – -0.39
SLC22A16	6	organic cation/carnitine transporter	-1.44	-2.55 – -0.31
IRF7	11	interferon regulatory factor 7	-1.46	-2.76 – -0.22
RSAD2	2	radical S-adenosyl methionine domain containing 2	-2.74	-5.43 – -0.05
LAP3	4	leucine aminopeptidase 3	-1.77	-3.15 – -0.38
IFI6	1	interferon, alpha-inducible protein 6	-2.79	-5.41 – -0.15
WARS	14	tryptophanyl-tRNA synthetase	-1.38	-2.63 – -0.11
ITGB1BP1	2	integrin beta 1 binding protein 1	-1.43	-2.44 – -0.47
HERC5	4	hect domain and RLD 5	-2.72	-5.28 – -0.25
DHX58	17	DEXH (Asp-Glu-X-His) box polypeptide 58	-1.59	-3.13 – -0.00
TPST1	7	tyrosylprotein sulfotransferase 1	-1.67	-3.28 – -0.15
OAS1	12	2'-5'-oligoadenylate synthetase 1, 40/46kDa	-2.38	-4.32 – -0.40
OAS3	12	2'-5'-oligoadenylate synthetase 3, 100kDa	-2.70	-5.21 – -0.15
RTP4	3	receptor (chemosensory) transporter protein 4	-1.81	-3.46 – -0.10
HES4	1	hairy and enhancer of split 4 (Drosophila)	-1.64	-3.23 – -0.08
ISG15	1	ISG15 ubiquitin-like modifier	-3.54	-6.45 – -0.74

Table 18: Genes repressed in atypical pneumonias when compared to Gram negative infections. Coefficients are in log2 scale.

Gene	Chr	Function	Coefficient	95% CI
IFIT3	10	IFN-induced protein with tetratricopeptide repeats 3	2.72	0.12 – 5.10
IFIH1	2	IFN induced with helicase C domain 1	1.54	0.12 – 2.95
IFITM3	11	IFN induced transmembrane protein 3	1.79	0.44 – 3.16
IFI44L	1	IFN-induced protein 44-like	3.66	0.47 – 6.84
IFI44	1	IFN-induced protein 44	2.52	0.12 – 4.96
HIST1H4H	6	histone cluster 1, H4h	1.81	0.47 – 3.16
SERPING1	11	serpin peptidase inhibitor (C1 inhibitor)	2.11	0.39 – 3.87
HSPE1	2	heat shock 10kDa protein 1 (chaperonin 10)	1.39	0.42 – 2.33
SPATS2L	2	spermatogenesis associated, serine-rich 2-like	1.44	0.18 – 2.72
SCO2	22	SCO cytochrome oxidase deficient homolog 2 (yeast)	1.54	0.32 – 2.78
ZBP1	20	Z-DNA binding protein 1	1.43	0.03 – 2.79
CARD17	11	caspase recruitment domain family, member 17	1.43	0.44 – 2.43
HSH2D	19	hematopoietic SH2 domain containing	1.36	0.30 – 2.42
IRF7	11	interferon regulatory factor 7	1.43	0.28 – 2.52
DDX60	4	DEAD (Asp-Glu-Ala-Asp) box polypeptide 60	1.74	0.21 – 3.31
TMTC1	12	transmembrane and tetratricopeptide repeat containing	1.64	0.19 – 3.04
LGALS2	22	lectin, galactoside-binding, soluble, 2	1.65	0.23 – 3.13
RSAD2	2	radical S-adenosyl methionine domain containing 2	2.67	0.31 – 5.08
SAMD9L	7	sterile alpha motif domain containing 9-like	1.75	0.36 – 3.21
IFI6	1	interferon, alpha-inducible protein 6	2.48	0.06 – 4.84
OTOF	2	otoferlin	2.50	1.10 – 3.90
TRIM22	11	tripartite motif containing 22	1.75	0.31 – 3.16
HERC5	4	hect domain and RLD 5	2.50	0.17 – 4.76
DHX58	17	DEXH (Asp-Glu-X-His) box polypeptide 58	1.46	0.05 – 2.89
IFI27	14	interferon, alpha-inducible protein 27	3.55	0.53 – 6.45
MT2A	16	metallothionein 2A	1.54	0.26 – 2.82
OAS1	12	2′-5′-oligoadenylate synthetase 1, 40/46kDa	2.31	0.54 – 4.03
OAS3	12	2′-5′-oligoadenylate synthetase 3, 100kDa	2.60	0.29 – 4.87
RTP4	3	receptor (chemosensory) transporter protein 4	1.53	0.04 – 3.05
HES4	1	hairy and enhancer of split 4 (Drosophila)	1.47	0.06 – 2.89

Table 19: Genes induced in Gram negative pneumonias when compared to viral pneumonias. Coefficients are expressed in log2 scale.

Gene	Chr	Function	Coefficient	95% CI
KCNMA1	10	K^+ large conductance Ca^{2+}-activated channel	-1.58	-3.05 – -0.06
FAM102A	9	family with sequence similarity 102	-1.33	-2.17 – -0.49
KLRG1	12	killer cell lectin-like receptor subfamily G	-1.40	-2.68 – -0.03
HLA-DRB5	6	MHC, class II, DR beta 5	-3.08	-5.81 – -0.26
CD8A	2	CD8a molecule	-1.70	-3.31 – -0.02
GZMK	5	granzyme K (granzyme 3; tryptase II)	-1.51	-2.81 – -0.24
TDRD9	14	tudor domain containing 9	-1.45	-2.81 – -0.12
HOPX	4	HOP homeobox	-1.63	-3.04 – -0.25
IL32	16	interleukin 32	-1.63	-2.90 – -0.41
RBPMS2	15	RNA binding protein with multiple splicing	-1.97	-3.42 – -0.45
EOMES	3	eomesodermin	-1.64	-3.07 – -0.27
PROS1	3	protein S (alpha)	-1.73	-3.01 – -0.49
RNY5	7	RNA, Ro-associated Y5	-2.31	-4.49 – -0.17
RNU1-3	1	RNA, U1 small nuclear 3	-2.67	-4.87 – -0.55
RNU11	1	RNA, U11 small nuclear	-1.48	-2.56 – -0.37
RNU1-5	1	RNA, U1 small nuclear 5	-2.54	-4.68 – -0.40

Table 20: Genes repressed in Gram negative pneumonias when compared to viral pneumonias. Coefficients are expressed in log2 scale.

Gene	Chr	Function	Coefficient	95% CI
MS4A7	11	membrane-spanning 4-domains, subfamily A	1.33	0.27 – 2.37
CLEC7A	12	C-type lectin domain family 7, member A	1.91	0.81 – 3.01
CFD	19	complement factor D (adipsin)	2.09	0.57 – 3.57
KYNU	2	kynureninase	1.63	0.67 – 2.55
STRADB	2	STE20-related kinase adaptor beta	1.45	0.29 – 2.71
CA1	8	carbonic anhydrase I	2.36	0.38 – 4.41
SLC4A1	17	solute carrier family 4, anion exchanger	1.83	0.51 – 3.14
SLC25A39	17	solute carrier family 25, member 39	1.70	0.55 – 2.93
AHSP	16	alpha haemoglobin stabilizing protein	2.36	0.71 – 4.01
MYL4	17	myosin, light chain 4, alkali; atrial, embryonic	1.39	0.07 – 2.72
OTOF	2	otoferlin	1.90	0.78 – 3.07
HBM	16	haemoglobin, mu	3.29	1.56 – 5.03
ALAS2	X	aminolevulinate, delta-, synthase 2	1.84	0.10 – 3.54
HBD	11	haemoglobin, delta	1.77	0.00 – 3.51
GYPC	2	glycophorin C (Gerbich blood group)	1.36	0.32 – 2.38
C7orf28B	7	chromosome 7 open reading frame 28B		
IFI27	14	interferon, alpha-inducible protein 27	2.88	0.60 – 5.27
IRF5	7	interferon regulatory factor 5	1.35	0.31 – 2.39
SEPT4	17	septin 4	-1.34	-2.31 – -0.38
HLA-DRB4	6	MHC, class II, DR beta 4	-1.79	-3.34 – -0.22
NFXL1	4	nuclear transcription factor	-1.36	-2.35 – -0.31

Table 21: Genes discriminating atypical pneumonias from viral infections. Coefficients are in log2 scale.

Gene set	Tolerance	Survival	Shared	p	Estimate
Chr15q22	16	22	7	0.31	0.449
Biocarta CTL pathway	5	4	4	0.39	0.607
Biocarta UCALPAIN pathway	5	14	4	0.85	-0.154
KEGG antigen processing and presentation	33	27	13	0.53	0.193
KEGG type I diabetes mellitus	16	20	12	0.21	-0.389
KEGG allograft rejection	16	21	12	0.21	-0.389
KEGG graft versus host disease	21	19	12	0.19	-0.404
KEGG autoimmune thyroid disease	14	19	11	0.18	-0.433
KEGG viral myocarditis	19	14	9	0.76	-0.117
Reactome dual incision reaction in TC NER	17	28	17	0.27	-0.286
Reactome immunoregulatory interactions	16	20	6	0.43	-0.402
Reactome MTOR signalling	9	8	3	0.64	-0.541
Cellular defence response	10	16	7	0.03	0.805
Proteinaceous extracellular matrix	12	12	4	0.51	-0.491
Membrane	278	248	56	0.77	0.040
Extracellular matrix	12	13	4	0.51	-0.491
Synapse	4	17	4	0.34	-0.659
Anion transmembrane transporter activity	7	11	3	0.10	0.987

Table 22: Pathway correlation with tolerance in CAP cluster 1 survival. The numbers indicate the number of leading edge genes contributing to the enrichment of the gene set. Shared = number of genes that are shared in the leading edge gene subset for each of the contrasts, Estimate = Spearman correlation coefficient

Gene set	Tolerance	Survival	Shared	p	Estimate
Chr18q23	8	4	3	0.83	-0.266
Chr15q22	16	14	3	0.05	-0.997
Biocarta CTL pathway	5	8	3	0.59	-0.595
KEGG antigen processing and presentation	33	25	12	0.03	-0.624
KEGG type I diabetes mellitus	16	26	13	0.59	-0.164
KEGG allograft rejection	16	25	13	0.59	-0.164
KEGG graft versus host disease	21	26	15	0.06	-0.488
KEGG autoimmune thyroid disease	14	25	12	0.43	-0.253
KEGG arrhythmogenic RV cardiomyopathy ARVC	7	18	3	0.98	0.034
KEGG viral myocarditis	19	22	9	0.74	0.127
Reactome dual incision reaction in TC NER	17	10	4	0.24	-0.757
Reactome immunoregulatory interactions	16	23	4	0.44	-0.561
Steroid metabolic process	13	23	5	0.82	0.142
Maintenance of localization	8	7	3	0.73	-0.407
Cellular defence response	10	16	6	0.18	0.625
Membrane	278	256	55	0.64	0.064
Anion transmembrane transporter activity	7	14	3	0.22	-0.940
Antigen binding	7	7	3	0.06	0.996
Sulphuric ester hydrolase activity	8	8	5	0.63	0.296
Serine type endopeptidase inhibitor activity	13	6	4	0.58	0.424

Table 23: Pathway correlation with tolerance in CAP cluster 2 survival. The numbers indicate the number of leading edge genes contributing to the enrichment of the gene set. Shared = number of genes that are shared in the leading edge gene subset for each of the contrasts, Estimate = Spearman correlation coefficient

Gene set	Tolerance	Survival	Shared	p	Estimate
Chr15q22	16	23	8	0.49	-0.287
KEGG antigen processing and presentation	33	19	10	0.47	0.260
KEGG type i diabetes mellitus	16	20	10	0.63	-0.175
KEGG allograft rejection	16	20	10	0.63	-0.175
KEGG graft versus host disease	21	21	11	0.10	-0.525
KEGG autoimmune thyroid disease	14	21	9	0.39	-0.328
KEGG arrhythmogenic RV cardiomyopathy	7	13	3	0.21	-0.946
KEGG viral myocarditis	19	18	9	0.81	0.096
Reactome dual incision reaction in TC NER	17	8	5	0.81	-0.151
Steroid metabolic process	13	18	3	0.26	-0.916
Cell cell adhesion	8	21	3	0.68	-0.481
Cellular defence response	10	19	4	0.62	0.379
Membrane	278	270	60	0.10	0.216
Anion transmembrane transporter activity	7	11	3	0.32	0.875
Serine type endopeptidase inhibitor activity	13	4	3	0.81	0.292

Table 24: Pathway correlation with tolerance in CAP cluster 3 survival. The numbers indicate the number of leading edge genes contributing to the enrichment of the gene set. Shared = number of genes that are shared in the leading edge gene subset for each of the contrasts, Estimate = Spearman correlation coefficient

Supplementary Figures

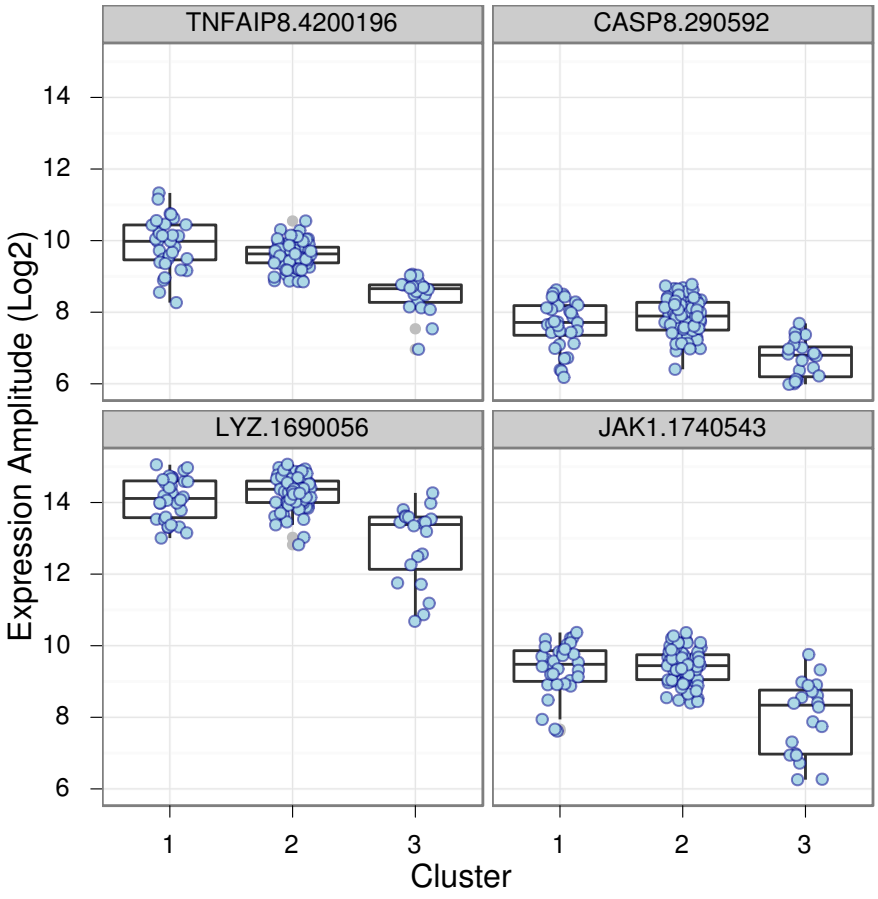


Figure 1: Distribution of some of the genes consistently repressed in cluster 3 compared to other clusters.*TNFAIP8* Tumour necrosis factor alpha induced protein 8, *LYZ* Lysozyme, *CASP8*Caspase 8, *JAK1* Janus kinase 1. Numbers following gene names are the Illumina probe identifiers.

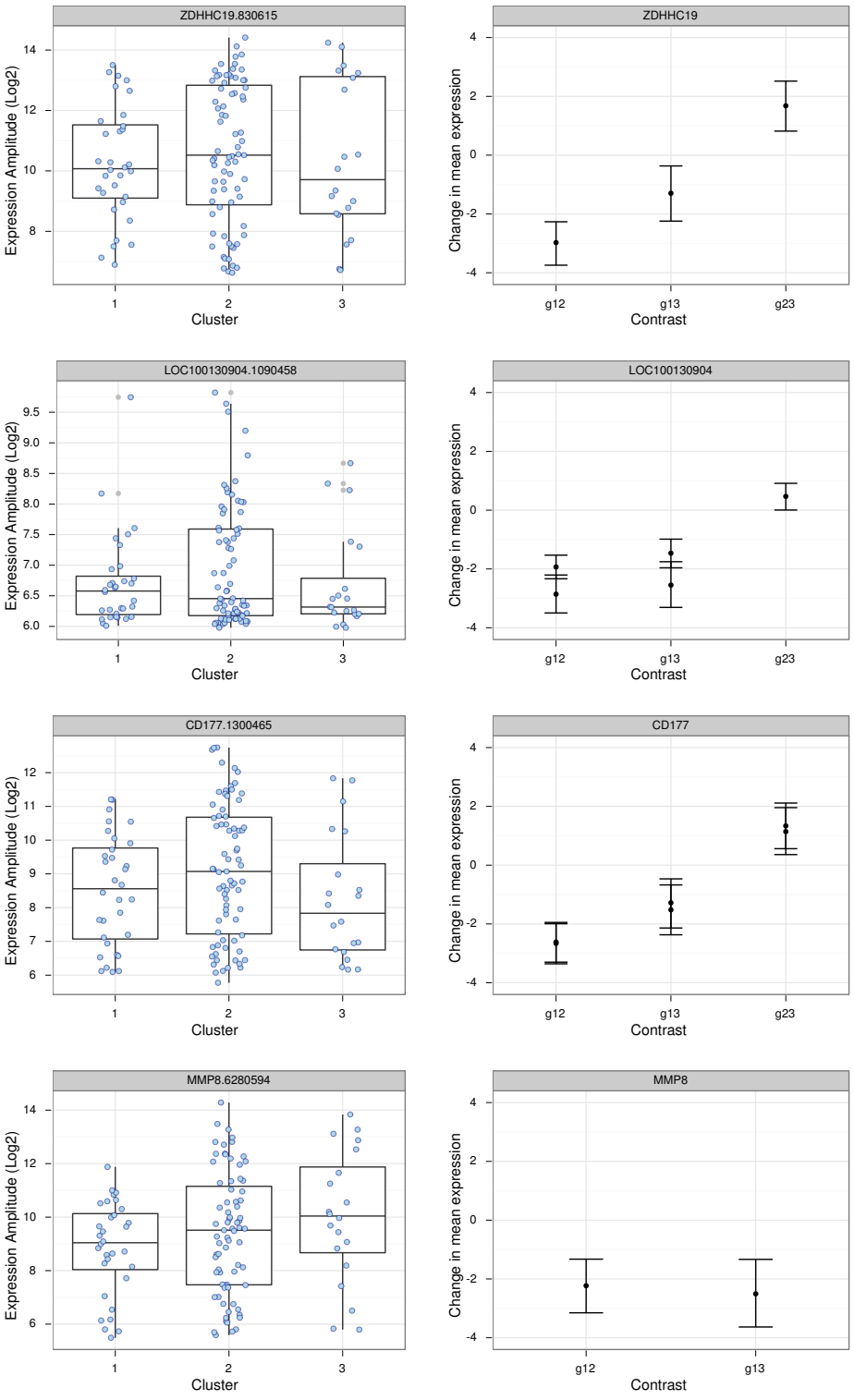


Figure 2: Genes with the biggest fold change differences between clusters 1 and 2. Panels on the left show the distribution of expression values. Panels on the right show the Bayesian estimate of mean differences and the 95% credible intervals. Multiple lines in a contrast are due to multiple probes for the same gene. Contrasts are labelled as "gab" which compares cluster b against cluster a. *ZDHHC19* Zinc finger domain containing protein, *LOC100130904* similar to CD177, *CD177* cell surface glycoprotein receptor, *MMP8* neutrophil matrix metallopeptidase 8

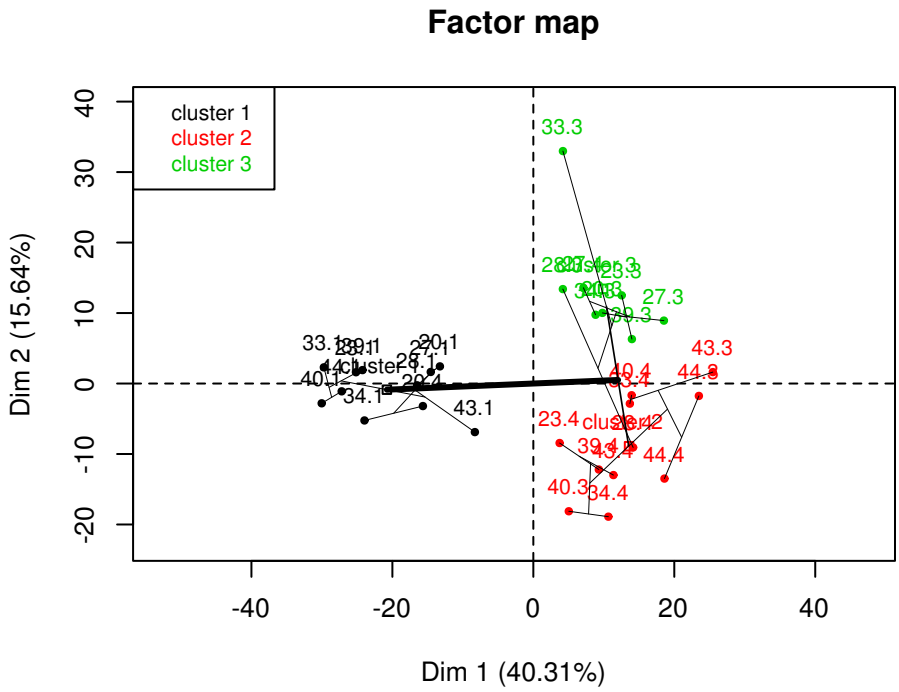


Figure 3: Hierarchical agglomerative clustering on the principal components for the validation cohort in the cardiac dataset. Each number represents one sample and is composed of the patient id preceding the decimal point and the time of sampling where 1 = Baseline, 3 = post op, and 4 = day 1 samples. There is some overlap between the two post surgical samples where the inflammatory response on day 1 is comparable to the immediate post op period.

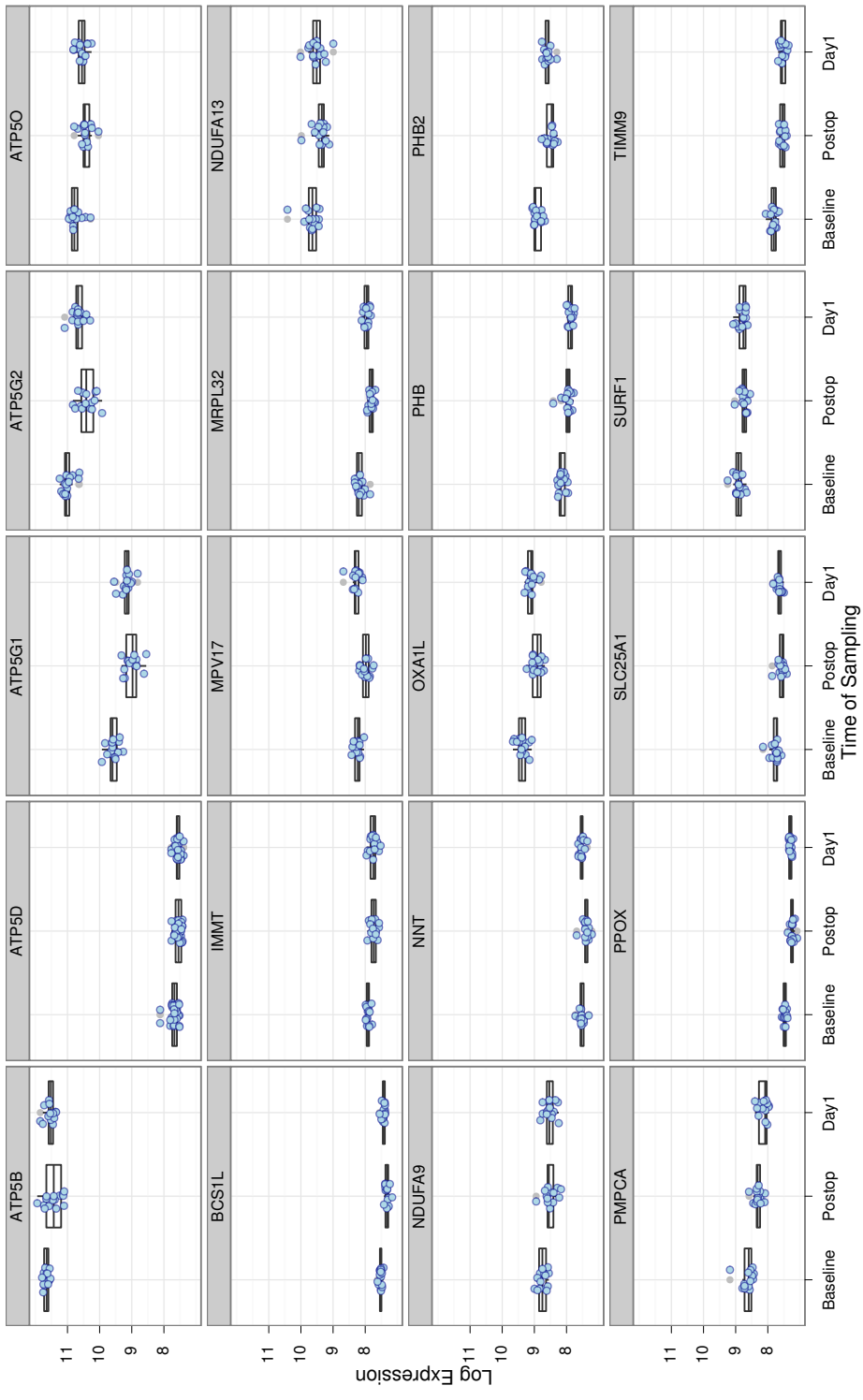


Figure 4: Leading edge genes that encode mitochondrial inner membrane proteins in cardiac surgery. All the genes have a low variance suggesting tight regulation in health and disease.

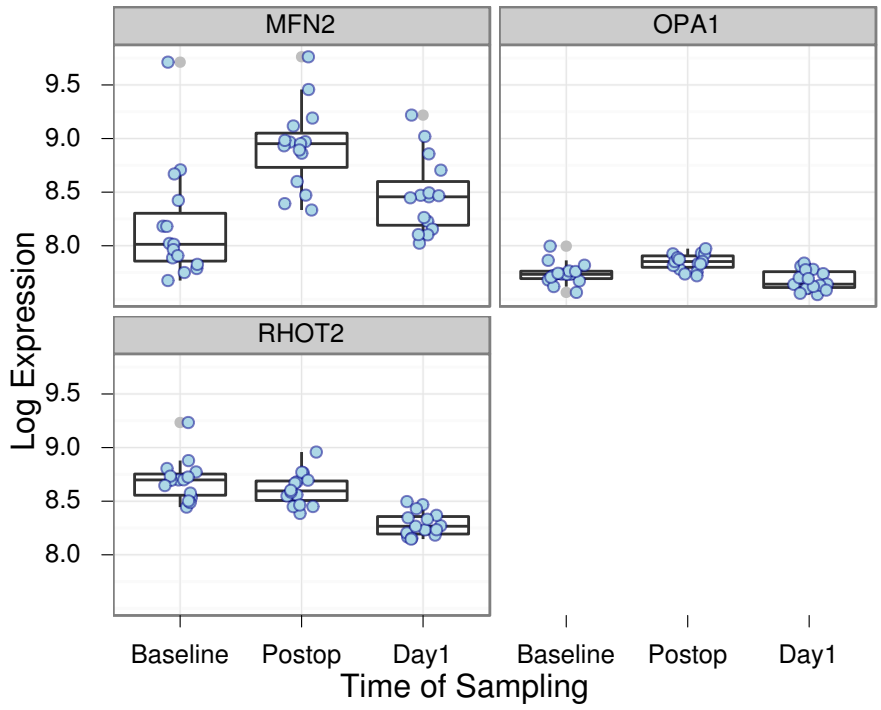


Figure 5: Leading edge genes that encode mitochondrial outer membrane proteins in cardiac surgery.

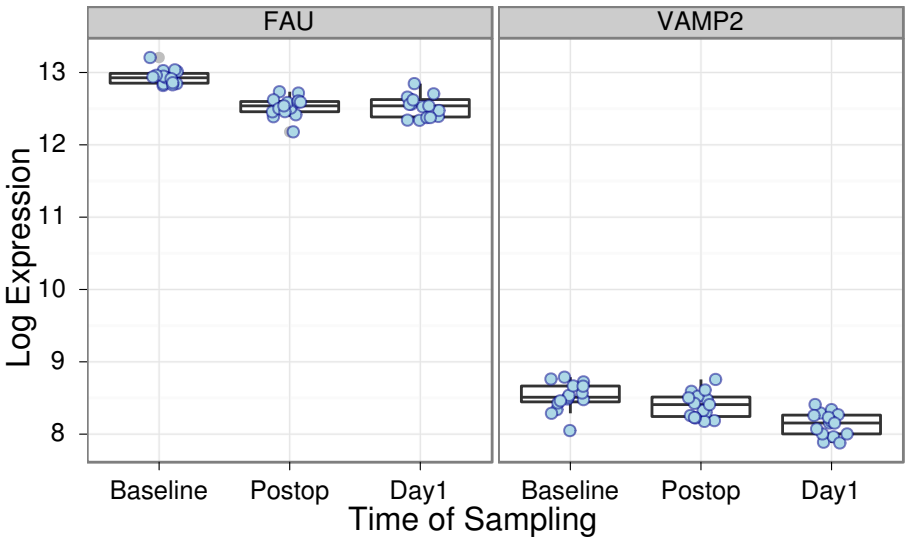


Figure 6: Leading edge genes of the Reactome insulin synthesis and secretion pathway in cardiac surgery

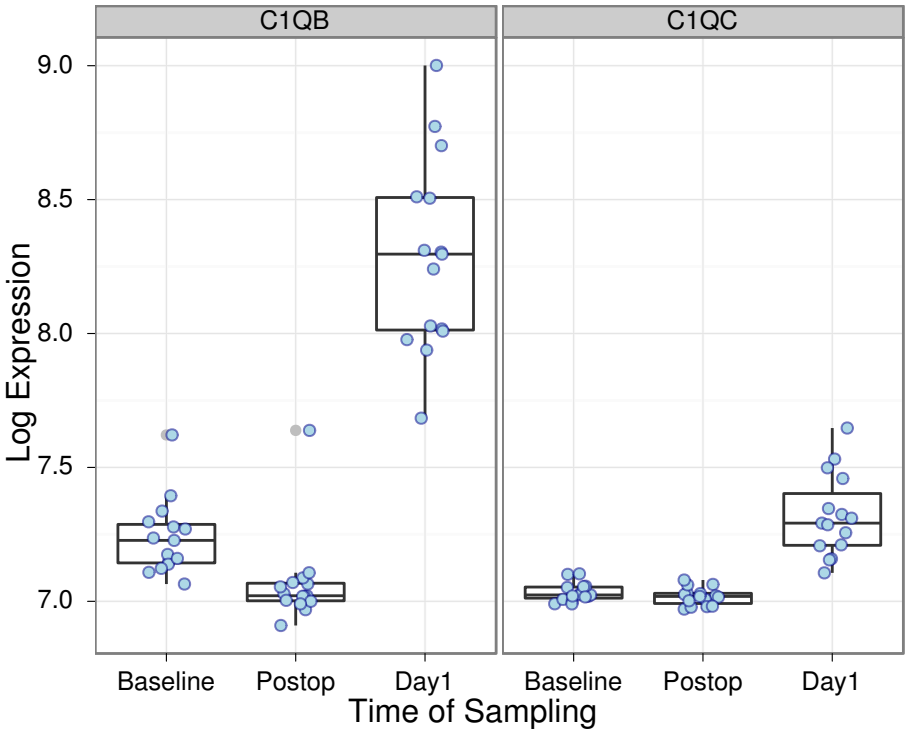


Figure 7: Leading edge genes related to the complement cascade in cardiac surgery

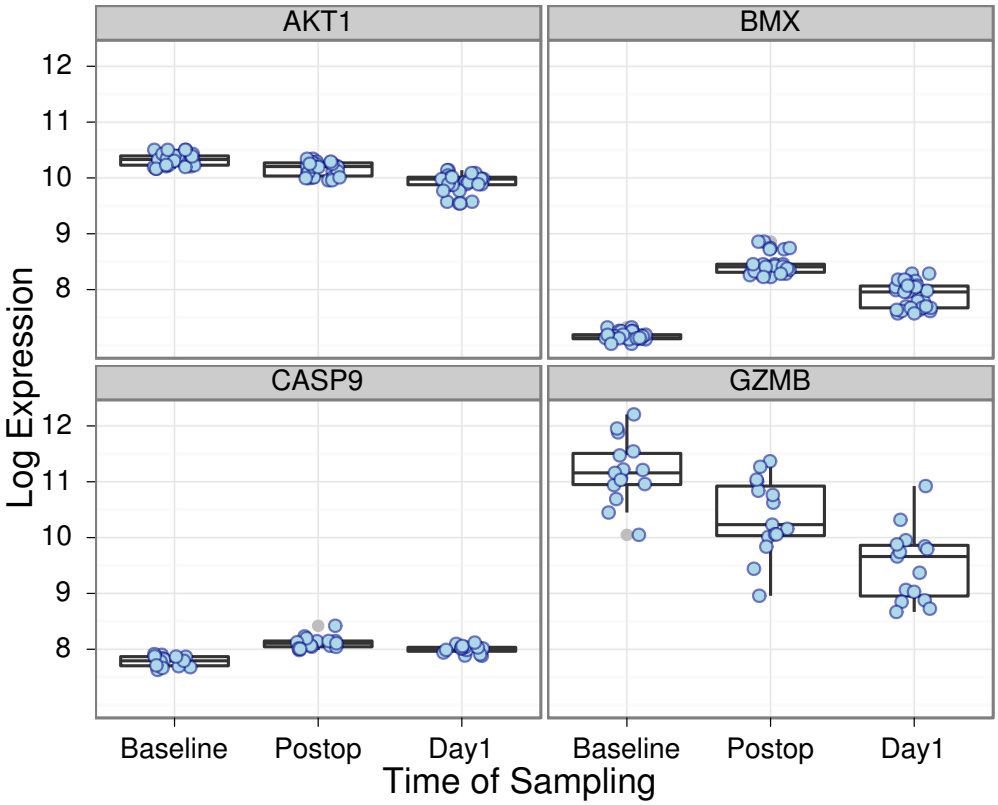


Figure 8: Leading edge genes related to the regulation of apoptosis in cardiac surgery

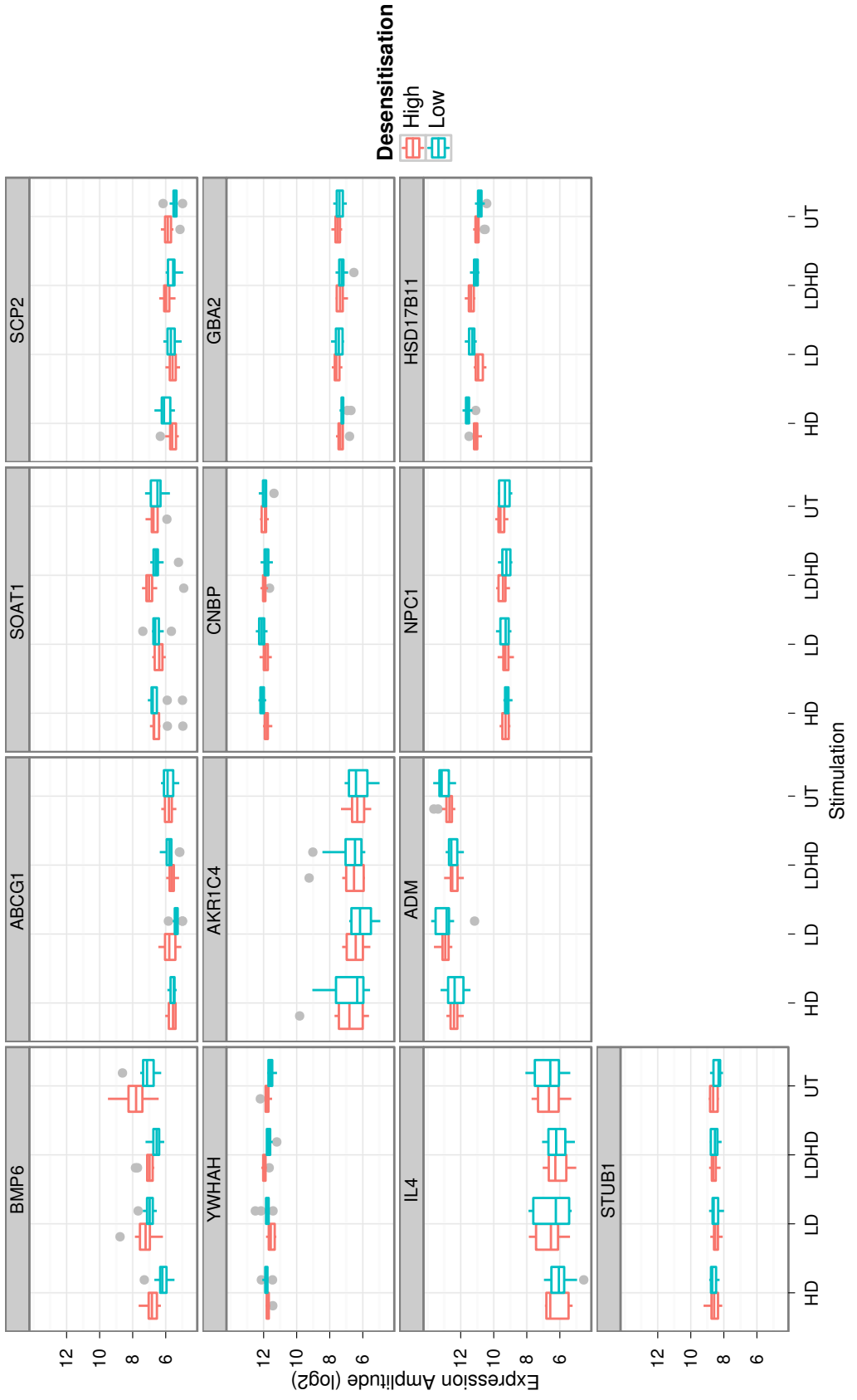


Figure 9: Leading edge genes in the steroid metabolic process for the endotoxin tolerance phenotype. Genes with high and modest effects contribute to the enrichment of the pathway.

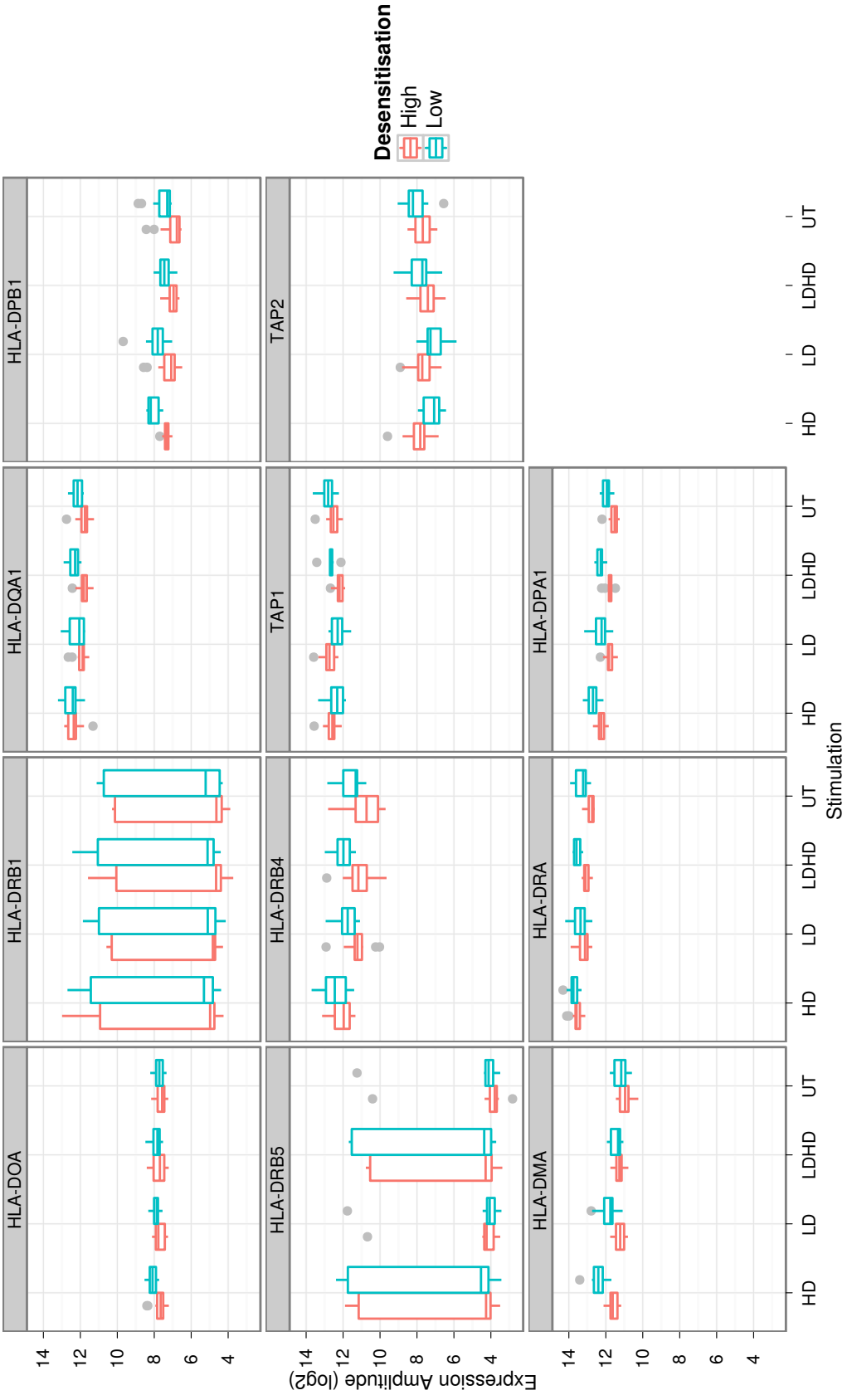


Figure 10: Leading edge genes in the antigen presentation pathway for the endotoxin tolerance phenotype. Genes with high and modest effects contribute to the enrichment of the pathway.

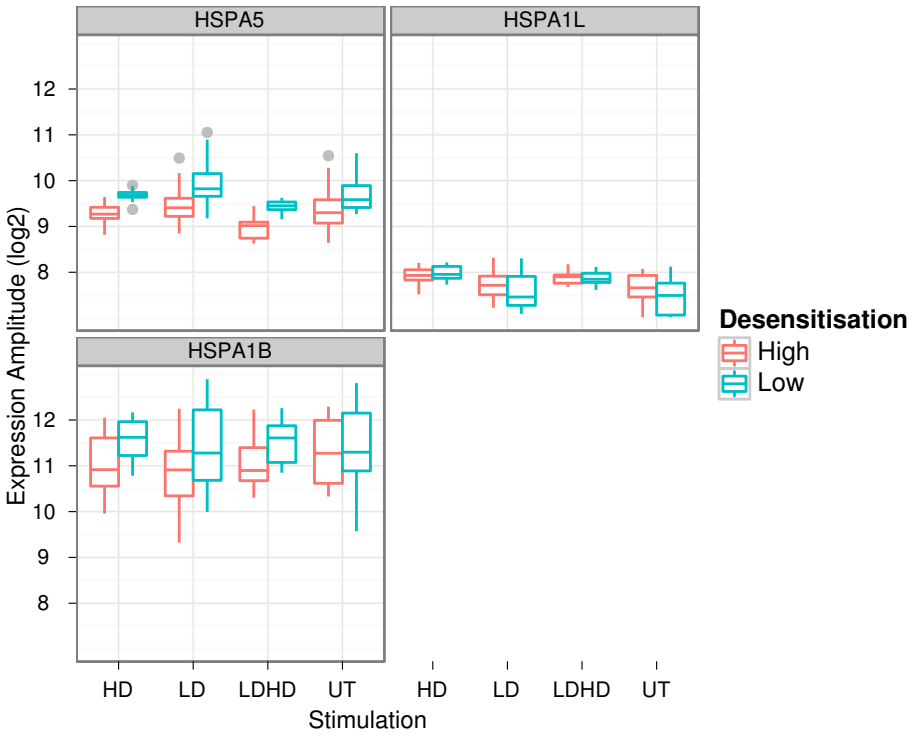


Figure 11: Leading edge genes related to heat shock proteins in the antigen presentation pathway for the endotoxin tolerance phenotype. Genes with high and modest effects contribute to the enrichment of the pathway.

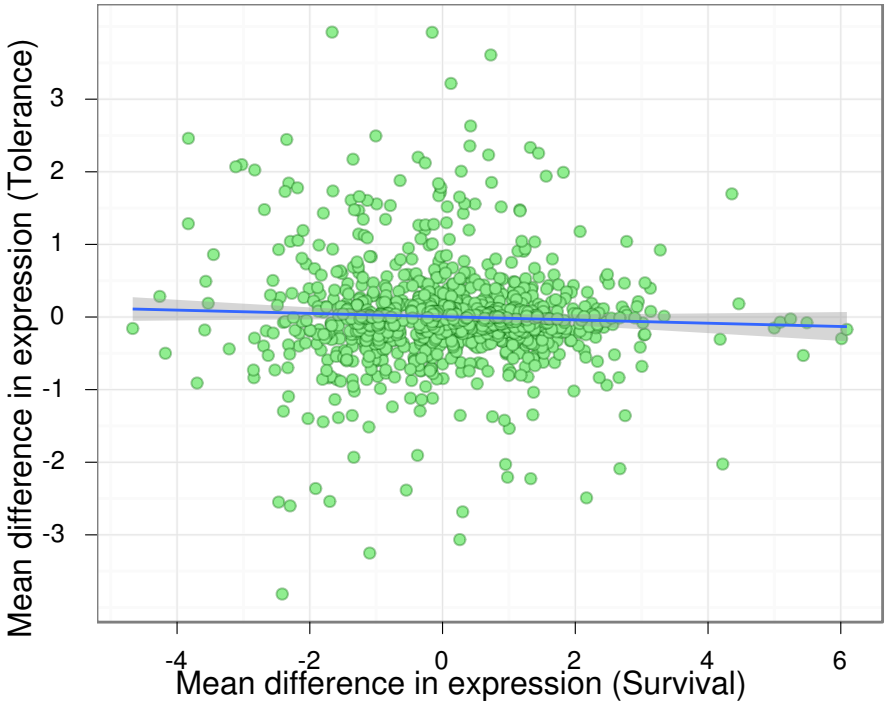


Figure 12: Correlation between the expression of the genes that are differentially expressed in either sepsis survival and the endotoxin tolerance response

Bibliography

1. Adrie, C., Azoulay, E., Timsit, J. & Vincent, J. in, 889–896 (Springer New York, 2007). ISBN: 978-0-387-49518-7.
2. Alves-Filho, J. C., Sônego, F., Souto, F. O., Freitas, A., Jr, W. A. V., Auxiliadora-Martins, M., Basile-Filho, A., McKenzie, A. N., Xu, D., Cunha, F. Q. & Liew, F. Y. Interleukin-33 attenuates sepsis by enhancing neutrophil influx to the site of infection. *Nature Medicine* **16,** 708–712. ISSN: 1078-8956 (May 2010).
3. Andersson, A., Rundgren, M., Kalman, S., Rooyackers, O., Brattstrom, O., Oldner, A., Eriksson, S. & Frithiof, R. Gut microcirculatory and mitochondrial effects of hyperdynamic endotoxaemic shock and norepinephrine treatment. *British Journal of Anaesthesia* **108.** PMID: 22157851, 254–261. ISSN: 1471-6771 (Feb. 2012).
4. Angele, M. K., Knöferl, M. W., Schwacha, M. G., Ayala, A., Cioffi, W. G., Bland, K. I. & Chaudry, I. H. Sex steroids regulate pro- and anti-inflammatory cytokine release by macrophages after trauma-hemorrhage. *American Journal of Physiology - Cell Physiology* **277,** C35–C42. ISSN: 0363-6143, 1522-1563 (July 1999).
5. Angus, D. C., Marrie, T. J., Obrosky, D. S., Clermont, G., Dremsizov, T. T., Coley, C., Fine, M. J., Singer, D. E. & Kapoor, W. N. Severe Community-acquired Pneumonia. *American Journal of Respiratory and Critical Care Medicine* **166,** 717–723. ISSN: 1073-449X, 1535-4970 (Sept. 2002).
6. Annane, D., Sébille, V., Troché, G., Raphaël, J., Gajdos, P. & Bellissant, E. A 3-Level Prognostic Classification in Septic Shock Based on Cortisol Levels and Cortisol Response to Corticotropin. *JAMA: The Journal of the American Medical Association* **283,** 1038–1045. ISSN: 0098-7484, 1538-3598 (Feb. 2000).
7. Apostolova, N., Garcia-Bou, R., Hernandez-Mijares, A., Herance, R., Rocha, M. & Victor, V. M. Mitochondrial antioxidants alleviate oxidative and nitrosative stress in a cellular model of sepsis. *Pharmaceutical Research* **28.** PMID: 21786065, 2910–2919. ISSN: 1573-904X (Nov. 2011).
8. Arbour, N. C., Lorenz, E., Schutte, B. C., Zabner, J., Kline, J. N., Jones, M., Frees, K., Watt, J. L. & Schwartz, D. A. TLR4 mutations are associated with endotoxin hyporesponsiveness in humans. *Nature Genetics* **25,** 187–191. ISSN: ERROR! NO ISSN (June 2000).
9. Arcaroli, J., Fessler, M. B. & Abraham, E. Genetic polymorphisms and sepsis. *Shock* **24,** 300–312. ISSN: 1073-2322 (Oct. 2005).

10. Astiz, M. E., Rackow, E. C., Still, J. G., Howell, S. T., Cato, A., Von Eschen, K. B., Ulrich, J. T., Rudbach, J. A., McMahon, G. & Vargas, R. Pretreatment of normal humans with monophosphoryl lipid A induces tolerance to endotoxin: a prospective, double-blind, randomized, controlled trial. *Critical Care Medicine* **23.** PMID: 8001393, 9–17. ISSN: 0090-3493 (Jan. 1995).

11. Bartlett, M. S. Properties of Sufficiency and Statistical Tests. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **160.** Article-Type: research-article / Full publication date: May 18, 1937 / Copyright © 1937 The Royal Society, 268–282. ISSN: 0080-4630 (May 1937).

12. Beadling, C. & Slifka, M. K. How do viral infections predispose patients to bacterial infections? *Current Opinion in Infectious Diseases* **17.** PMID: 15166819, 185–191. ISSN: 0951-7375 (June 2004).

13. Bécue-Bertaut, M. & Pagès, J. Multiple factor analysis and clustering of a mixture of quantitative, categorical and frequency data. *Comput. Stat. Data Anal.* **52,** 3255–3268. ISSN: 0167-9473 (Feb. 2008).

14. Beeson, P. Development of tolerance to typhoid bacterial pyrogen and its abolition by reticulo-endothelial blockade. *Proc. Soc. Exp. Biol. Med.* **61,** 248–250 (1946).

15. Bellamy, R., Ruwende, C., Corrah, T., McAdam, K. P., Whittle, H. C. & Hill, A. V. Variations in the NRAMP1 gene and susceptibility to tuberculosis in West Africans. *The New England Journal of Medicine* **338.** PMID: 9486992, 640–644. ISSN: 0028-4793 (Mar. 1998).

16. Bellili, N. M., Foucan, L., Fumeron, F., Mohammedi, K., Travert, F., Roussel, R., Balkau, B., Tichet, J. & Marre, M. Associations of the |[minus]|344 T>C and the 3097 G>A Polymorphisms of CYP11B2 Gene With Hypertension, Type 2 Diabetes, and Metabolic Syndrome in a French Population. *American Journal of Hypertension* **23,** 660–667. ISSN: 0895-7061 (Mar. 2010).

17. Berry, M. P. R. *et al.* An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* **466,** 973–977. ISSN: 0028-0836 (2010).

18. Berthier, M., Paradis, A., Tchernof, A., Bergeron, J., Prud'homme, D., Després, J. & Vohl, M. The interleukin 6 –174G/C Polymorphism is associated with indices of obesity in men. *Journal of Human Genetics* **48,** 0014–0019. ISSN: 1434-5161 (Jan. 2003).

19. Bhasin, J. M., Chakrabarti, E., Peng, D., Kulkarni, A., Chen, X. & Smith, J. D. Sex Specific Gene Regulation and Expression QTLs in Mouse Macrophages from a Strain Intercross. *PLoS ONE* **3.** PMID: 18197246 PMCID: 2174529. ISSN: 1932-6203 (Jan. 2008).

20. Boldrick, J. C., Alizadeh, A. A., Diehn, M., Dudoit, S., Liu, C. L., Belcher, C. E., Botstein, D., Staudt, L. M., Brown, P. O. & Relman, D. A. Stereotyped and specific gene expression programs in human innate immune responses to bacteria. *Proceedings of the National Academy of Sciences* **99,** 972–977 (Jan. 2002).

21. Bolstad, B. M., Irizarry, R. A., Åstrand, M. & Speed, T. P. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics* **19,** 185–193. ISSN: 1367-4803, 1460-2059 (Jan. 2003).

22. Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R. M. & Sibbald, W. J. 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 (June 1992).

23. Brouwer, M. C., de Gans, J., Heckenberg, S. G., Zwinderman, A. H., van der Poll, T. & van de Beek, D. Host genetic susceptibility to pneumococcal and meningococcal disease: a systematic review and meta-analysis. *The Lancet Infectious Diseases* **9,** 31–44. ISSN: 1473-3099 (Jan. 2009).

24. Brown, K., Brain, S., Pearson, J., Edgeworth, J., Lewis, S. & Treacher, D. Neutrophils in development of multiple organ failure in sepsis. *The Lancet* **368,** 157–169. ISSN: 0140-6736 (July 2006).

25. Bruhn, A., Verdant, C., Vercruysse, V., Su, F., Vray, B. & Vincent, J. Effects of dexamethasone on macrophage migration inhibitory factor production in sepsis. *Shock (Augusta, Ga.)* **26.** PMID: 16878025, 169–173. ISSN: 1073-2322 (Aug. 2006).

26. Bryant, C. & Fitzgerald, K. A. Molecular mechanisms involved in inflammasome activation. *Trends in Cell Biology* **19,** 455–464. ISSN: 0962-8924 (Sept. 2009).

27. Bucala, R., Model, P. & Cerami, A. Modification of DNA by reducing sugars: a possible mechanism for nucleic acid aging and age-related dysfunction in gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **81.** PMID: 6582469 PMCID: 344619, 105–109. ISSN: 0027-8424 (Jan. 1984).

28. Cadwell, K., Patel, K. K., Maloney, N. S., Liu, T., Ng, A. C., Storer, C. E., Head, R. D., Xavier, R., Stappenbeck, T. S. & Virgin, H. W. Virus-Plus-Susceptibility Gene Interaction Determines Crohn's Disease Gene Atg16L1 Phenotypes in Intestine. *Cell* **141,** 1135–1145. ISSN: 0092-8674 (June 2010).

29. Cai, L., Ji, A., de Beer, F. C., Tannock, L. R. & van der Westhuyzen, D. R. SR-BI protects against endotoxemia in mice through its roles in glucocorticoid production and hepatic clearance. *Journal of Clinical Investigation* **118,** 364–375. ISSN: 0021-9738 (Jan. 2008).

30. Calvano, J., Agnese, D., Um, J., Goshima, M., Singhal, R., Coyle, S., Reddell, M., Kumar, A., Calvano, S. & Lowry, S. Modulation of the lipopolysaccharide receptor complex (CD14, TLR4, MD-2) and toll-like receptor 2 in systemic inflammatory response syndrome-positive patients with and without infection: relationship to tolerance. *Shock* **20,** 415 (2003).

31. Cartin-Ceba, R., Kojicic, M., Li, G., Kor, D. J., Poulose, J., Herasevich, V., Kashyap, R., Trillo-Alvarez, C., Cabello-Garza, J., Hubmayr, R., Seferian, E. G. & Gajic, O. Epidemiology of Critical Care Syndromes, Organ Failures, and Life-Support Interventions in a Suburban US Community. *Chest* **140,** 1447–1455. ISSN: 0012-3692, 1931-3543 (Dec. 2011).

32. Caruana, R. & Niculescu-Mizil, A. in *Proceedings of the 23rd international conference on Machine learning* (ACM, Pittsburgh, Pennsylvania, 2006), 161–168. ISBN: 1-59593-383-2.

33. Castegren, M., Lipcsey, M., Söderberg, E., Skorup, P., Eriksson, M., Larsson, A. & Sjölin, J. Differences in Organ Dysfunction in Endotoxin Tolerant Pigs Under Intensive Care Exposed to a Second Hit of Endotoxin. *Shock (Augusta, Ga.)* PMID: 22266970. ISSN: 1540-0514 (Jan. 2012).

34. Chan, C., Li, L., McCall, C. E. & Yoza, B. K. Endotoxin Tolerance Disrupts Chromatin Remodeling and NF-B Transactivation at the IL-1 Promoter. *The Journal of Immunology* **175,** 461–468. ISSN: 0022-1767, 1550-6606 (July 2005).

35. Checchia, P., Anas, N., Thomas, N. J., Wong, H. R., Freishtat, R. J., Lin, R., Cvijanovich, N., Monaco, M., Shanley, T. P., Allen, G. L., Sakthivel, B. & Odoms, K. Validating the genomic signature of pediatric septic shock. *Physiological genomics* **34,** 127–34 (2008).

36. Chen, X., Yoza, B. K., El Gazzar, M., Hu, J. Y. Q., Cousart, S. L. & McCall, C. E. RelB Sustains IB Expression during Endotoxin Tolerance. *Clinical and Vaccine Immunology* **16,** 104–110. ISSN: 1556-6811, 1556-679X (Jan. 2009).

37. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452,** 429–435. ISSN: 0028-0836 (Mar. 2008).

38. Chong, D. L. W. & Sriskandan, S. Pro-inflammatory mechanisms in sepsis. *Contributions to Microbiology* **17.** PMID: 21659748, 86–107. ISSN: 1662-291X (2011).

39. Chopra, M., Golden, H. B., Mullapudi, S., Dowhan, W., Dostal, D. E. & Sharma, A. C. Modulation of Myocardial Mitochondrial Mechanisms during Severe Polymicrobial Sepsis in the Rat. *PLoS ONE* **6.** PMID: 21712982 PMCID: 3119671. ISSN: 1932-6203 (June 2011).

40. Clarke, J., Seo, P. & Clarke, B. Statistical expression deconvolution from mixed tissue samples. *Bioinformatics* **26,** 1043–1049. ISSN: 1367-4803, 1460-2059 (Apr. 2010).

41. Combes, A., Costa, M., Trouillet, J., Baudot, J., Mokhtari, M., Gibert, C. & Chastre, J. Morbidity, mortality, and quality-of-life outcomes of patients requiring ≥14 days of mechanical ventilation. *Critical Care Medicine* **31,** 1373–1381. ISSN: 0090-3493 (May 2003).

42. Cuthbertson, B. H., Roughton, S., Jenkinson, D., MacLennan, G. & Vale, L. Quality of life in the five years after intensive care: a cohort study. *Critical Care* **14,** R6. ISSN: 1364-8535 (Jan. 2010).

43. Davila, S. *et al.* Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. *Nature Genetics* **42.** PMID: 20694013, 772–776. ISSN: 1546-1718 (Sept. 2010).

44. De Magalhães, J. P., Curado, J. & Church, G. M. Meta-Analysis of Age-Related Gene Expression Profiles Identifies Common Signatures of Aging. *Bioinformatics* **25,** 875–881. ISSN: 1367-4803, 1460-2059 (Apr. 2009).

45. De Maio, A., Torres, M. B. & Reeves, R. H. Genetic determinants influencing the response to injury, inflammation, and sepsis. *Shock* **23,** 11–17. ISSN: 1073-2322 (Jan. 2005).

46. Della Valle, P., Pavani, G. & D'Angelo, A. The protein C pathway and sepsis. *Thrombosis Research* **129,** 296–300. ISSN: 0049-3848 (Mar. 2012).

47. Deng, L., Broaddus, R. R., McCampbell, A., Shipley, G. L., Loose, D. S., Stancel, G. M., Pickar, J. H. & Davies, P. J. A. Identification of a novel estrogen-regulated gene, EIG121, induced by hormone replacement therapy and differentially expressed in type I and type II endometrial cancer. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **11.** PMID: 16322283, 8258–8264. ISSN: 1078-0432 (Dec. 2005).

48. Devaraj, S., Dasu, M. R. & Jialal, I. Diabetes is a proinflammatory state: a translational perspective. *Expert review of endocrinology & metabolism* **5.** PMID: 20204165 PMCID: 2829992, 19–28. ISSN: 1744-6651 (Jan. 2010).

49. Devignot, S., Sapet, C., Duong, V., Bergon, A., Rihet, P., Ong, S., Lorn, P. T., Chroeung, N., Ngeav, S., Tolou, H. J., Buchy, P. & Couissinier-Paris, P. Genome-Wide Expression Profiling Deciphers Host Responses Altered during Dengue Shock Syndrome and Reveals the Role of Innate Immunity in Severe Dengue. *PLoS ONE* **5,** e11671 (July 2010).

50. Dhainaut, J., Laterre, P., Janes, J. M., Bernard, G. R., Artigas, A., Bakker, J., Riess, H., Basson, B. R., Charpentier, J., Utterback, B. G. & Vincent, J. Drotrecogin alfa (activated) in the treatment of severe sepsis patients with multiple-organ dysfunction: data from the PROWESS trial. *Intensive Care Medicine* **29,** 894–903. ISSN: 0342-4642 (June 2003).

51. Dimitriadou, E., Hornik, K., Leisch, F., Meyer, D. & Weingessel, A. *e1071: Misc Functions of the Department of Statistics (e1071), TU Wien* R package version 1.6 (2011).

52. Ding, X. Z., Feng, X. R., Borschel, R. H., Nikolich, M. P., Feng, J., Li, Y. S. & Hoover, D. L. HSP-70 mitigates LPS/SKI-induced cell damage by increasing sphingosine kinase 1 (SK1). *Prostaglandins & Other Lipid Mediators* **92,** 1–7. ISSN: 1098-8823 (June 2010).

53. Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G. M., Abecasis, G. R. & Cookson, W. O. C. A genome-wide association study of global gene expression. *Nat Genet* **39,** 1202–1207. ISSN: 1061-4036 (Oct. 2007).

54. Douville, R. N., Lissitsyn, Y., Hirschfeld, A. F., Becker, A. B., Kozyrskyj, A. L., Liem, J., Bastien, N., Li, Y., Victor, R. E., Sekhon, M., Turvey, S. E. & HayGlass, K. T. TLR4 Asp299Gly and Thr399Ile Polymorphisms: No Impact on Human Immune Responsiveness to LPS or Respiratory Syncytial Virus. *PLoS ONE* **5,** e12087 (2010).

55. Draisma, A., Dorresteijn, M. J., Bouw, M. P., van der Hoeven, J. G. & Pickkers, P. The Role of Cytokines and Inducible Nitric Oxide Synthase in Endotoxemia-Induced Endothelial Dysfunction. *Journal of Cardiovascular Pharmacology* **55,** 595–600. ISSN: 0160-2446 (June 2010).

56. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24,** 1547–1548 (July 2008).

57. Dunning, M. J., Barbosa-Morais, N. L., Lynch, A. G., Tavaré, S. & Ritchie, M. E. Statistical issues in the analysis of Illumina data. *BMC Bioinformatics* **9,** 85. ISSN: 1471-2105 (Feb. 2008).

58. Eachempati, S. R., Hydo, L. & Barie, P. S. Gender-Based Differences in Outcome in Patients With Sepsis. *Arch Surg* **134,** 1342–1347 (Dec. 1999).

59. Eisen, D. P., Dean, M. M., Boermeester, M. A., Fidler, K. J., Gordon, A. C., Kronborg, G., Kun, J. F. J., Lau, Y. L., Payeras, A., Valdimarsson, H., Brett, S. J., Ip, W. K. E., Mila, J., Peters, M. J., Saevarsdottir, S., van Till, J. W. O., Hinds, C. J. & McBryde, E. S. Low serum mannose-binding lectin level increases the risk of death due to pneumococcal infection. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America* **47.** PMID: 18611155, 510–516. ISSN: 1537-6591 (Aug. 2008).

60. Eisen, D. P., Dean, M. M., Thomas, P., Marshall, P., Gerns, N., Heatley, S., Quinn, J., Minchinton, R. M. & Lipman, J. Low mannosebinding lectin function is associated with sepsis in adult patients. *FEMS Immunology & Medical Microbiology* **48,** 274–282. ISSN: 1574-695X (Nov. 2006).

61. Ershler, W. B. & Keller, E. T. Age-Associated Increased Interleukin-6 Gene Expression, Late-Life Diseases, and Frailty. *Annual Review of Medicine* **51,** 245–270 (2000).

62. Ertel, W., Kremer, J. P., Kenney, J., Steckholzer, U., Jarrar, D., Trentz, O. & Schildberg, F. W. Downregulation of proinflammatory cytokine release in whole blood from septic patients. *Blood* **85,** 1341–1347. ISSN: 0006-4971, 1528-0020 (Mar. 1995).

63. Fahy, R. J., Exline, M. C., Gavrilin, M. A., Bhatt, N. Y., Besecker, B. Y., Sarkar, A., Hollyfield, J. L., Duncan, M. D., Nagaraja, H. N., Knatz, N. L., Hall, M. & Wewers, M. D. Inflammasome mRNA Expression in Human Monocytes during Early Septic Shock. *Am. J. Respir. Crit. Care Med.* **177,** 983–988 (May 2008).

64. Fairfax, B. P., Davenport, E. E., Makino, S., Hill, A. V., Vannberg, F. O. & Knight, J. C. A common haplotype of the tumour necrosis factor receptor 2 gene modulates endotoxin tolerance. *Journal of immunology (Baltimore, Md. : 1950)* **186.** PMID: 21282507 PMCID: 3211060, 3058–3065. ISSN: 0022-1767 (Mar. 2011).

65. Fairfax, B. P., Vannberg, F. O., Radhakrishnan, J., Hakonarson, H., Keating, B. J., Hill, A. V. & Knight, J. C. An integrated expression phenotype mapping approach defines common variants in LEP, ALOX15 and CAPNS1 associated with induction of IL-6. *Human Molecular Genetics* **19,** 720–730 (Feb. 2010).

66. Feezor, R. J., Baker, H. V., Mindrinos, M., Hayden, D., Tannahill, C. L., Brownstein, B. H., Fay, A., MacMillan, S., Laramie, J., Xiao, W., Moldawer, L. L., Cobb, J. P., Laudanski, K., Miller-Graziano, C. L., Maier, R. V., Schoenfeld, D., Davis, R. W. & Tompkins, R. G. Whole blood and leucocyte RNA isolation for gene expression analyses. *Physiological Genomics* **19,** 247–254. ISSN: 1094-8341, 1531-2267 (Nov. 2004).

67. Fei, D., Meng, X., Zhao, M., Kang, K., Tan, G., Pan, S., Luo, Y., Liu, W., Nan, C., Jiang, H., Krissansen, G. W., Zhao, M. & Sun, X. Enhanced induction of heme oxygenase-1 suppresses thrombus formation and affects the protein C system in sepsis. *Translational Research: The Journal of Laboratory and Clinical Medicine* **159.** PMID: 22243794, 99–109. ISSN: 1878-1810 (Feb. 2012).

68. Fingerle-Rowson, G., Koch, P., Bikoff, R., Lin, X., Metz, C. N., Dhabhar, F. S., Meinhardt, A. & Bucala, R. Regulation of Macrophage Migration Inhibitory Factor Expression by Glucocorticoids in Vivo. *The American Journal of Pathology* **162,** 47–56. ISSN: 0002-9440 (Jan. 2003).

69. Flohé, S., Dominguez Fernández, E., Ackermann, M., Hirsch, T., Börgermann, J. & Schade, F. Endotoxin tolerance in rats:Expression of TNF-, IL-6, IL-10, VCAM-1 and HSP 70 in lung and liver during endotoxin shock. *Cytokine* **11,** 796–804. ISSN: 1043-4666 (Oct. 1999).

70. Foster, S. L., Hargreaves, D. C. & Medzhitov, R. Gene-specific control of inflammation by TLR-induced chromatin modifications. *Nature* **447,** 972–978. ISSN: 0028-0836 (July 2007).

71. Garrabou, G., Morén, C., López, S., Tobías, E., Cardellach, F., Miró, Ò. & Casademont, J. The Effects of Sepsis on Mitochondria. *Journal of Infectious Diseases* **205,** 392–400. ISSN: 0022-1899, 1537-6613 (Feb. 2012).

72. Gazzar, M. E., Yoza, B. K., Hu, J. Y. .-., Cousart, S. L. & McCall, C. E. Epigenetic Silencing of Tumor Necrosis Factor during Endotoxin Tolerance. *Journal of Biological Chemistry* **282,** 26857–26864. ISSN: 0021-9258, 1083-351X (Sept. 2007).

73. Gentleman, R. C., Carey, V. J., Bates, D. M., *et al.* Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology* **5,** R80 (2004).

74. Glubb, D. M., Gearry, R. B., Barclay, M. L., Roberts, R. L., Pearson, J., Keenan, J. I., McKenzie, J. & Bentley, R. W. NOD2 and ATG16L1 polymorphisms affect monocyte responses in Crohn's disease. *World Journal of Gastroenterology : WJG* **17.** PMID: 21734790 PMCID: 3120942, 2829–2837. ISSN: 1007-9327 (June 2011).

75. Goldstein, B. A., Hubbard, A. E., Cutler, A. & Barcellos, L. F. An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. *BMC Genetics* **11,** 49. ISSN: 1471-2156 (June 2010).

76. Gon, Y., Hashimoto, S., Hayashi, S., Koura, T., Matsumoto, K. & Horie, T. Lower serum concentrations of cytokines in elderly patients with pneumonia and the impaired production of cytokines by peripheral blood monocytes in the elderly. *Clinical and Experimental Immunology* **106.** PMID: 8870709, 120–126. ISSN: 0009-9104 (Oct. 1996).

77. Goodin, J., Manrique, C., Dulohery, M., Sampson, J., Saettele, M. & Dabbagh, O. Effect of statins on the clinical outcomes of patients with sepsis. *Anaesthesia and Intensive Care* **39.** PMID: 22165357, 1051–1055. ISSN: 0310-057X (Nov. 2011).

78. Gordon, A. C., Waheed, U., Hansen, T. K., Hitman, G. A., Garrard, C. S., Turner, M. W., Klein, N. J., Brett, S. J. & Hinds, C. J. Mannose-binding lectin polymorphisms in severe sepsis: relationship to levels, incidence, and outcome. *Shock (Augusta, Ga.)* **25.** PMID: 16369192, 88–93. ISSN: 1073-2322 (Jan. 2006).

79. Gregg L., S. HIF-1, O2, and the 3 PHDs: How Animal Cells Signal Hypoxia to the Nucleus. *Cell* **107,** 1–3. ISSN: 0092-8674 (Oct. 2001).

80. Grundmann, U., Rensing, H., Adams, H. A., Falk, S., Wendler, O., Ebinger, N. & Bauer, M. Endotoxin desensitization of human mononuclear cells after cardiopulmonary bypass: role of humoral factors. *Anesthesiology* **93.** PMID: 10910482, 359–369. ISSN: 0003-3022 (Aug. 2000).

81. Guevara, R., Gianotti, M., Roca, P. & Oliver, J. Age and sex-related changes in rat brain mitochondrial function. *Cellular Physiology and Biochemistry: International Journal of Experimental Cellular Physiology, Biochemistry, and Pharmacology* **27.** PMID: 21471708, 201–206. ISSN: 1421-9778 (2011).

82. Guevara, R., Santandreu, F. M., Valle, A., Gianotti, M., Oliver, J. & Roca, P. Sex-dependent differences in aged rat brain mitochondrial function and oxidative stress. *Free Radical Biology & Medicine* **46.** PMID: 18992805, 169–175. ISSN: 1873-4596 (Jan. 2009).

83. Haralambous, E., Weiss, H. A., Radalowicz, A., Hibberd, M. L., Booy, R. & Levin, M. Sibling familial risk ratio of meningococcal disease in UK Caucasians. *Epidemiology and Infection* **130.** PMID: 12825725, 413–418. ISSN: 0950-2688 (June 2003).

84. Harrison, D. A., Brady, A. R. & Rowan, K. Case mix, outcome and length of stay for admissions to adult, general critical care units in England, Wales and Northern Ireland: the Intensive Care National Audit & Research Centre Case Mix Programme Database. *Critical Care* **9,** S1. ISSN: 1364-8535 (Feb. 2004).

85. Harrison, D. A., Welch, C. A. & Eddleston, J. M. The epidemiology of severe sepsis in England, Wales and Northern Ireland, 1996 to 2004: secondary analysis of a high quality clinical database, the ICNARC Case Mix Programme Database. *Critical Care* **10.** PMID: 16542492 PMCID: 1550902, R42. ISSN: 1364-8535 (2006).

86. Henckaerts, L., Nielsen, K. R., Steffensen, R., Van Steen, K., Mathieu, C., Giulietti, A., Wouters, P. J., Milants, I., Vanhorebeek, I., Langouche, L., Vermeire, S., Rutgeerts, P., Thiel, S., Wilmer, A., Hansen, T. K. & Van den Berghe, G. Polymorphisms in innate immunity genes predispose to bacteremia and death in the medical intensive care unit*. *Critical Care Medicine* **37,** 192–e3. ISSN: 0090-3493 (Jan. 2009).

87. Hill, A. V. S. Aspects of Genetic Susceptibility to Human Infectious Diseases. *Annual Review of Genetics* **40,** 469–486 (2006).

88. Hinshaw, L. B., Emerson, J. T E, Taylor, J. F B, Chang, A. C., Duerr, M., Peer, G. T., Flournoy, D. J., White, G. L., Kosanke, S. D. & Murray, C. K. Lethal Staphylococcus aureus-induced shock in primates: prevention of death with anti-TNF antibody. *The Journal of Trauma* **33.** PMID: 1433404, 568–573. ISSN: 0022-5282 (Oct. 1992).

89. Hinshaw, L. B., Tekamp-Olson, P., Chang, A. C., Lee, P. A., Taylor, J. F B, Murray, C. K., Peer, G. T., Emerson, J. T E, Passey, R. B. & Kuo, G. C. Survival of primates in LD100 septic shock following therapy with antibody to tumor necrosis factor (TNF alpha). *Circulatory Shock* **30.** PMID: 2178801, 279–292. ISSN: 0092-6213 (Mar. 1990).

90. Hu, H. & Li, X. Transcriptional regulation in eukaryotic ribosomal protein genes. *Genomics* **90,** 421–423. ISSN: 0888-7543 (Oct. 2007).

91. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics Enrichment Tools: Paths Toward the Comprehensive Functional Analysis of Large Gene Lists. *Nucleic Acids Research* **37,** 1–13. ISSN: 0305-1048, 1362-4962 (Jan. 2009).

92. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4.** PMID: 19131956, 44–57. ISSN: 1750-2799 (2009).

93. Huang, L., Chuang, I., Dong, H. & Yang, R. Hypoxia-inducible factor 1 regulates the expression of the mitochondrial ATPase inhibitor protein (IF1) in rat liver. *Shock (Augusta, Ga.)* **36.** PMID: 21412184, 90–96. ISSN: 1540-0514 (July 2011).

94. Huo, M., Xu, L., D, L., Liu, D., Liu, K., Chen, H., Han, W., Chen, Y., Y, W., Wang, J., Zhang, X., Zhao, X. & Huang, X. The effect of HLA disparity on clinical outcome after HLA-haploidentical blood and marrow transplantation. *Clinical Transplantation.* PMID: 21919963. ISSN: 1399-0012 (Sept. 2011).

95. Iba, T., Okamoto, K., Ohike, T., Tajirika, T., Aihara, K., Watanabe, S. & Kayhanian, H. Enoxaparin and fondaparinux attenuates endothelial damage in endotoxemic rats. *The Journal of Trauma and Acute Care Surgery* **72.** PMID: 22310127, 177–182. ISSN: 2163-0763 (Jan. 2012).

96. Idaghdour, Y., Storey, J. D., Jadallah, S. J. & Gibson, G. A Genome-Wide Gene Expression Signature of Environmental Geography in Leukocytes of Moroccan Amazighs. *PLoS Genet* **4,** e1000052 (Apr. 2008).

97. Imtiyaz, H. Z., Williams, E. P., Hickey, M. M., Patel, S. A., Durham, A. C., Yuan, L., Hammond, R., Gimotty, P. A., Keith, B. & Simon, M. C. Hypoxia-inducible factor 2alpha regulates macrophage function in mouse models of acute and tumor inflammation. *The Journal of Clinical Investigation* **120.** PMID: 20644254, 2699–2714. ISSN: 1558-8238 (Aug. 2010).

98. Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G. & Guedj, M. Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data: A Comparison of Variance Modeling Strategies. *PLoS ONE* **5,** e12336 (2010).

99. Jensen, U. S., Knudsen, J. D., Wehberg, S., Gregson, D. B. & Laupland, K. B. Risk factors for recurrence and death after bacteraemia: a population-based study. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* **17.** PMID: 21714830, 1148–1154. ISSN: 1469-0691 (Aug. 2011).

100. Jepson, A., Banya, W., Sisay-Joof, F., Hassan-King, M., Nunes, C., Bennett, S. & Whittle, H. Quantification of the relative contribution of major histocompatibility complex (MHC) and non-MHC genes to human immune responses to foreign antigens. *Infection and Immunity* **65.** PMID: 9038290 PMCID: 175062, 872–876. ISSN: 0019-9567 (Mar. 1997).

101. Jepson, A. P., Banya, W. A. S., Sisay-Joof, F., Hassan-King, M., Bennett, S. & Whittle, H. C. Genetic Regulation of Fever in Plasmodium falciparum Malaria in Gambian Twin Children. *The Journal of Infectious Diseases* **172.** ArticleType: research-article / Full publication date: Jul., 1995 / Copyright © 1995 Oxford University Press, 316–319. ISSN: 0022-1899 (July 1995).

102. Kanaan, Z., Gardner, S., Carruba, C., Mattingly, J., Druen, D. & Cheadle, W. G. Macrophage Genetic Reprogramming During Chronic Peritonitis is Augmented by LPS Pretreatment. *The Journal of Surgical Research.* PMID: 21696761. ISSN: 1095-8673 (May 2011).

103. Kaslow, R. A., McNicholl, J. & Hill, A. V. S. *Genetic susceptibility to infectious diseases* ISBN: 9780195174908 (Oxford University Press, 2008).

104. Kee, C., Cheong, K. Y., Pham, K., Waterer, G. W. & Temple, S. E. L. Genetic variation in heat shock protein 70 is associated with septic shock: narrowing the association to a specific haplotype. *International Journal of Immunogenetics* **35,** 465–473. ISSN: 1744-313X (Dec. 2008).

105. Kholodenko, B. N., Hancock, J. F. & Kolch, W. Signalling ballet in space and time. *Nature Reviews Molecular Cell Biology* **11,** 414–426. ISSN: 1471-0072 (June 2010).

106. Knaus, W. A., Draper, E. A., Wagner, D. P. & Zimmerman, J. E. APACHE II: a severity of disease classification system. *Critical Care Medicine* **13.** PMID: 3928249, 818–829. ISSN: 0090-3493 (Oct. 1985).

107. Koenig, W., Khuseyinova, N., Hoffmann, M. M., März, W., Fröhlich, M., Hoffmeister, A., Brenner, H. & Rothenbacher, D. CD14 C(-260)→T polymorphism, plasma levels of the soluble endotoxin receptor CD14, their association with chronic infections and risk of stable coronary artery disease. *Journal of the American College of Cardiology* **40,** 34–42. ISSN: 0735-1097 (July 2002).

108. Korosec Jagodic, H., Jagodic, K. & Podbregar, M. Long-term outcome and quality of life of patients treated in surgical intensive care: a comparison between sepsis and trauma. *Critical Care (London, England)* **10.** PMID: 16978417, R134. ISSN: 1466-609X (2006).

109. Kovarik, M., Muthny, T., Sispera, L. & Holecek, M. The dose-dependent effects of endotoxin on protein metabolism in two types of rat skeletal muscle. *Journal of Physiology and Biochemistry.* PMID: 22311459. ISSN: 1877-8755 (Feb. 2012).

110. Kraft, C., Peter, M. & Hofmann, K. Selective autophagy: ubiquitin-mediated recognition and beyond. *Nature Cell Biology* **12,** 836–841. ISSN: 1465-7392 (2010).

111. Kubaszek, A., Pihlajamäki, J., Punnonen, K., Karhapää, P., Vauhkonen, I. & Laakso, M. The C-174G Promoter Polymorphism of the IL-6 Gene Affects Energy Expenditure and Insulin Sensitivity. *Diabetes* **52,** 558–561. ISSN: 0012-1797, 1939-327X (Feb. 2003).

112. Kugelberg, E., Gollan, B. & Tang, C. M. Mechanisms in Neisseria meningitidis for resistance against complement-mediated killing. *Vaccine* **26.** PMID: 19388162 PMCID: 2686086, I34–I39. ISSN: 0264-410X (Dec. 2008).

113. Lagu, T., Lindenauer, P. K., Rothberg, M. B., Nathanson, B. H., Pekow, P. S., Steingrub, J. S. & Higgins, T. L. Development and validation of a model that uses enhanced administrative data to predict mortality in patients with sepsis. *Critical Care Medicine* **39.** PMID: 22005222, 2425–2430. ISSN: 1530-0293 (Nov. 2011).

114. Lall, H., Coughlan, K. & Sumbayev, V. V. HIF-1 protein is an essential factor for protection of myeloid cells against LPS-induced depletion of ATP and apoptosis that supports Toll-like receptor 4-mediated production of IL-6. *Molecular Immunology* **45,** 3045–3049. ISSN: 0161-5890 (June 2008).

115. Lapaquette, P., Bringer, M. & DarfeuilleMichaud, A. Defects in autophagy favour adherentinvasive Escherichia coli persistence within macrophages leading to increased proinflammatory response. *Cellular Microbiology.* ISSN: 1462-5822 (2012).

116. Laubach, V. E., Foley, P. L., Shockey, K. S., Tribble, C. G. & Kron, I. L. Protective roles of nitric oxide and testosterone in endotoxemia: evidence from NOS-2-deficient mice. *American Journal of Physiology - Heart and Circulatory Physiology* **275,** H2211–H2218. ISSN: 0363-6135, 1522-1539 (Dec. 1998).

117. Le Gall, J. R., Lemeshow, S. & Saulnier, F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. *JAMA: The Journal of the American Medical Association* **270.** PMID: 8254858, 2957–2963. ISSN: 0098-7484 (Dec. 1993).

118. Lê, S., Josse, J. & Husson, F. FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software* **25,** 1–18. ISSN: 1548-7660 (Jan. 2008).

119. Lee, N., Choi, K. W., Chan, P. K. S., Hui, D. S. C., Lui, G. C. Y., Wong, B. C. K., Wong, R. Y. K., Sin, W. Y., Hui, W. M., Ngai, K. L. K., Cockram, C. S., Lai, R. W. M. & Sung, J. J. Y. Outcomes of adults hospitalised with severe influenza. *Thorax* **65.** PMID: 20522848, 510–515. ISSN: 1468-3296 (June 2010).

120. Levy, M. M., Fink, M. P., Marshall, J. C., Abraham, E., Angus, D., Cook, D., Cohen, J., Opal, S. M., Vincent, J. & Ramsay, G. 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference. *Critical Care Medicine* **31.** PMID: 12682500, 1250–1256. ISSN: 0090-3493 (Apr. 2003).

121. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2,** 18–22 (2002).

122. Liu, F. Q., Liu, Y., Lui, V. C. H., Lamb, J. R., Tam, P. K. H. & Chen, Y. Hypoxia modulates lipopolysaccharide induced TNF-alpha expression in murine macrophages. *Experimental Cell Research* **314.** PMID: 18255061, 1327–1336. ISSN: 0014-4827 (Apr. 2008).

123. Liu, T. F., Yoza, B. K., El Gazzar, M., Vachharajani, V. T. & McCall, C. E. NAD+-Dependent SIRT1 Deacetylase Participates in Epigenetic Reprogramming During Endotoxin Tolerance. *Journal of Biological Chemistry* **286,** 9856–9864. ISSN: 0021-9258, 1083-351X (Mar. 2011).

124. Liu, X., Hao, L., Zhang, S., Ji, Y., Zhang, Y., Lu, X., Shi, B., Pei, H., Wang, Y., Chen, D., Guan, X. & Zheng, Y. Genetic repression of mouse VEGF expression regulates coagulation cascade. *IUBMB Life* **62.** PMID: 21086498, 819–824. ISSN: 1521-6551 (Nov. 2010).

125. Looney, M. R., Gropper, M. A. & Matthay, M. A. Transfusion-Related Acute Lung Injury*. *Chest* **126,** 249–258. ISSN: 0012-3692, 1931-3543 (July 2004).

126. Lorenz, E., Mira, J. P., Frees, K. L. & Schwartz, D. A. Relevance of Mutations in the TLR4 Receptor in Patients With Gram-Negative Septic Shock. *Arch Intern Med* **162,** 1028–1032 (May 2002).

127. Madách, K., Aladzsity, I., Szilágyi, Á., Fust, G., Gál, J., Pénzes, I. & Prohászka, Z. 4G/5G polymorphism of PAI-1 gene is associated with multiple organ dysfunction and septic shock in pneumonia induced severe sepsis: prospective, observational, genetic study. *Critical Care* **14.** PMID: 20429897 PMCID: 2887202, R79. ISSN: 1364-8535 (2010).

128. Maines, M. D. The heme oxygenase system: a regulator of second messenger gases. *Annual Review of Pharmacology and Toxicology* **37.** PMID: 9131263, 517–554. ISSN: 0362-1642 (1997).

129. Majetschak, M., Flohé, S., Obertacke, U., Schröder, J., Staubach, K., Nast-Kolb, D., Schade, F. U. & Stüber, F. Relation of a TNF gene polymorphism to severe sepsis in trauma patients. *Annals of Surgery* **230.** PMID: 10450735, 207–214. ISSN: 0003-4932 (Aug. 1999).

130. Majetschak, M., Krehmeier, U., Bardenheuer, M., Denz, C., Quintel, M., Voggenreiter, G. & Obertacke, U. Extracellular ubiquitin inhibits the TNF- response to endotoxin in peripheral blood mononuclear cells and regulates endotoxin hyporesponsiveness in critical illness. *Blood* **101,** 1882–1890. ISSN: 0006-4971, 1528-0020 (Mar. 2003).

131. Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R. & Bähler, J. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* **151.** PMID: 23101633, 671–683. ISSN: 1097-4172 (Oct. 2012).

132. Martin, C. M., Priestap, F., Fisher, H., Fowler, R. A., Heyland, D. K., Keenan, S. P., Longo, C. J., Morrison, T., Bentley, D. & Antman, N. A prospective, observational registry of patients with severe sepsis: the Canadian Sepsis Treatment and Response Registry. *Critical Care Medicine* **37.** PMID: 19050636, 81–88. ISSN: 1530-0293 (Jan. 2009).

133. Martin, G. S., Mannino, D. M., Eaton, S. & Moss, M. The epidemiology of sepsis in the United States from 1979 through 2000. *The New England Journal of Medicine* **348.** PMID: 12700374, 1546–1554. ISSN: 1533-4406 (Apr. 2003).

134. Mason, P. E., Al-Khafaji, A., Milbrandt, E. B., Suffoletto, B. P. & Huang, D. T. CORTICUS: The end of unconditional love for steroid use? *Critical Care* **13,** 309. ISSN: 1364-8535 (Aug. 2009).

135. Massberg, S., Grahl, L., Bruehl, M. v., Manukyan, D., Pfeiler, S., Goosmann, C., Brinkmann, V., Lorenz, M., Bidzhekov, K., Khandagale, A. B., Konrad, I., Kennerknecht, E., Reges, K., Holdenrieder, S., Braun, S., Reinhardt, C., Spannagl, M., Preissner, K. T. & Engelmann, B. Reciprocal coupling of coagulation and innate immunity via neutrophil serine proteases. *Nature Medicine* **16,** 887–896. ISSN: 1078-8956 (Aug. 2010).

136. Maugeri, N., Radhakrishnan, J. & Knight, J. C. Genetic determinants of HSP70 gene expression following heat shock. *Human Molecular Genetics* **19,** 4939–4947 (Dec. 2010).

137. Maung, A. A., Fujimi, S., Miller, M. L., MacConmara, M. P., Mannick, J. A. & Lederer, J. A. Enhanced TLR4 reactivity following injury is mediated by increased p38 activation. *Journal of Leukocyte Biology* **78,** 565–573. ISSN: 0741-5400, 1938-3673 (Aug. 2005).

138. McCall, C. E. & Yoza, B. K. Gene Silencing in Severe Systemic Inflammation. *Am. J. Respir. Crit. Care Med.* **175,** 763–767 (Apr. 2007).

139. McKinney, E. F., Lyons, P. A., Carr, E. J., Hollis, J. L., Jayne, D. R. W., Willcocks, L. C., Koukoulaki, M., Brazma, A., Jovanovic, V., Kemeny, D. M., Pollard, A. J., MacAry, P. A., Chaudhry, A. N. & Smith, K. G. C. A CD8+ T cell transcription signature predicts prognosis in autoimmune disease. *Nature Medicine* **16,** 586–591. ISSN: 1078-8956 (Apr. 2010).

140. Meduri, G. U. Clinical review: A paradigm shift: the bidirectional effect of inflammation on bacterial growth. Clinical implications for patients with acute respiratory distress syndrome. *Critical Care* **6,** 24. ISSN: 1364-8535 (Nov. 2001).

141. Medvedev, A. E., Kopydlowski, K. M. & Vogel, S. N. Inhibition of Lipopolysaccharide-Induced Signal Transduction in Endotoxin-Tolerized Mouse Macrophages: Dysregulation of Cytokine, Chemokine, and Toll-Like Receptor 2 and 4 Gene Expression. *The Journal of Immunology* **164,** 5564–5574. ISSN: 0022-1767, 1550-6606 (June 2000).

142. Melo, E. S., Barbeiro, D. F., Gorjão, R., Rios, E. C. S., Vasconcelos, D., Velasco, I. T., Szabo, C., Curi, R., de Lima-Salgado, T. M. & Soriano, F. G. Gene expression reprogramming protects macrophage from septic-induced cell death. *Molecular Immunology* **47,** 2587–2593. ISSN: 0161-5890 (Oct. 2010).

143. Metukuri, M. R., Reddy, C. M. T., Reddy, P. R. K. & Reddanna, P. Bacterial LPS Mediated Acute Inflammation-induced Spermatogenic Failure in Rats: Role of Stress Response Proteins and Mitochondrial Dysfunction. *Inflammation* **33,** 235–243. ISSN: 0360-3997, 1573-2576 (Jan. 2010).

144. Miller, R. A. The aging immune system: primer and prospectus. *Science (New York, N.Y.)* **273.** PMID: 8658199, 70–74. ISSN: 0036-8075 (July 1996).

145. Mira, J., Cariou, A., Grall, F., Delclaux, C., Losser, M., Heshmati, F., Cheval, C., Monchi, M., Teboul, J., Riché, F., Leleu, G., Arbibe, L., Mignon, A., Delpech, M. & Dhainaut, J. Association of TNF2, a TNF- Promoter Polymorphism, With Septic Shock Susceptibility and Mortality A Multicenter Study. *JAMA: The Journal of the American Medical Association* **282,** 561–568. ISSN: 0098-7484, 1538-3598 (Aug. 1999).

146. Morote-Garcia, J. C., Rosenberger, P., Kuhlicke, J. & Eltzschig, H. K. HIF-1-dependent repression of adenosine kinase attenuates hypoxia-induced vascular leak. *Blood* **111.** PMID: 18309031, 5571–5580. ISSN: 1528-0020 (June 2008).

147. Morris, J. Sidney M. Regulation of enzymes of the urea cycle and arginine metabolism. *Annual Review of Nutrition* **22.** PMID: 12055339, 87–105. ISSN: 0199-9885 (2002).

148. Murphy, T. J., Paterson, H. M., Kriynovich, S., Zang, Y., Kurt-Jones, E. A., Mannick, J. A. & Lederer, J. A. Linking the "two-hit" response following injury to enhanced TLR4 reactivity. *Journal of Leukocyte Biology* **77,** 16–23. ISSN: 0741-5400, 1938-3673 (Jan. 2005).

149. Nickerson, E. K., Wuthiekanun, V., Wongsuvan, G., Limmathurosakul, D., Srisamang, P., Mahavanakul, W., Thaipadungpanit, J., Shah, K. R., Arayawichanont, A., Amornchai, P., Thanwisai, A., Day, N. P. & Peacock, S. J. Factors Predicting and Reducing Mortality in Patients with Invasive Staphylococcus aureus Disease in a Developing Country. *PLoS ONE* **4,** e6512 (2009).

150. Nooh, M. M., El-Gengehi, N., Kansal, R., David, C. S. & Kotb, M. HLA Transgenic Mice Provide Evidence for a Direct and Dominant Role of HLA Class II Variation in Modulating the Severity of Streptococcal Sepsis. *The Journal of Immunology* **178,** 3076–3083. ISSN: 0022-1767, 1550-6606 (Mar. 2007).

151. Ogawa, S., Lozach, J., Benner, C., Pascual, G., Tangirala, R. K., Westin, S., Hoffmann, A., Subramaniam, S., David, M., Rosenfeld, M. G. & Glass, C. K. Molecular Determinants of Crosstalk between Nuclear Receptors and Toll-like Receptors. *Cell* **122,** 707–721. ISSN: 0092-8674 (Sept. 2005).

152. Paddock, C. D., Liu, L., Denison, A. M., Bartlett, J. H., Holman, R. C., Deleon-Carnes, M., Emery, S. L., Drew, C. P., Shieh, W., Uyeki, T. M. & Zaki, S. R. Myocardial injury and bacterial pneumonia contribute to the pathogenesis of fatal influenza B virus infection. *The Journal of Infectious Diseases* **205.** PMID: 22291193, 895–905. ISSN: 1537-6613 (Mar. 2012).

153. Papathanassoglou, E. D. E., Giannakopoulou, M. D. & Bozas, E. Genomic variations and susceptibility to sepsis. *AACN Advanced Critical Care* **17.** PMID: 17091041, 394–422. ISSN: 1559-7768 (Dec. 2006).

154. Parmigiani, G., Garrett-Mayer, E. S., Anbazhagan, R. & Gabrielson, E. A Cross-Study Comparison of Gene Expression Studies for the Molecular Classification of Lung Cancer. *Clinical Cancer Research* **10,** 2922–2927. ISSN: 1078-0432, 1557-3265 (May 2004).

155. Pastinen, T., Ge, B. & Hudson, T. J. Influence of human genome polymorphism on gene expression. *Review Literature And Arts Of The Americas* **15,** 9–16 (2006).

156. Paterson, H. M., Murphy, T. J., Purcell, E. J., Shelley, O., Kriynovich, S. J., Lien, E., Mannick, J. A. & Lederer, J. A. Injury Primes the Innate Immune System for Enhanced Toll-Like Receptor Reactivity. *The Journal of Immunology* **171,** 1473–1483. ISSN: 0022-1767, 1550-6606 (Aug. 2003).

157. Pathak, E., MacMillan-Crow, L. A. & Mayeux, P. R. Role of mitochondrial oxidants in an in vitro model of sepsis-induced renal injury. *The Journal of Pharmacology and Experimental Therapeutics* **340.** PMID: 22011433, 192–201. ISSN: 1521-0103 (Jan. 2012).

158. Peyssonnaux, C., Cejudo-Martin, P., Doedens, A., Zinkernagel, A. S., Johnson, R. S. & Nizet, V. Cutting Edge: Essential Role of Hypoxia Inducible Factor-1 in Development of Lipopolysaccharide-Induced Sepsis. *The Journal of Immunology* **178,** 7516–7519. ISSN: 0022-1767, 1550-6606 (June 2007).

159. Picard, C. *et al.* Pyogenic bacterial infections in humans with IRAK-4 deficiency. *Science (New York, N.Y.)* **299.** PMID: 12637671, 2076–2079. ISSN: 1095-9203 (Mar. 2003).

160. Pillay, J., Kamp, V. M., van Hoffen, E., Visser, T., Tak, T., Lammers, J., Ulfman, L. H., Leenen, L. P., Pickkers, P. & Koenderman, L. A subset of neutrophils in human systemic inflammation inhibits T cell responses through Mac-1. *The Journal of Clinical Investigation* **122.** PMID: 22156198, 327–336. ISSN: 1558-8238 (Jan. 2012).

161. Plummer, M. JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling (2003).

162. Polito, A. & Annane, D. Adrenocortical cell tolerance to lipopolysaccharide: A new mechanism for Critical Illness Related Corticosteroid Insufficiency*. *Critical Care Medicine* **39,** 597–598. ISSN: 0090-3493 (Mar. 2011).

163. Puel, A., Yang, K., Ku, C., von Bernuth, H., Bustamante, J., Santos, O. F., Lawrence, T., Chang, H., Al-Mousa, H., Picard, C. & Casanova, J. Heritable defects of the human TLR signalling pathways. *Journal of Endotoxin Research* **11,** 220–224. ISSN: 1753-4259, 1753-4267 (Aug. 2005).

164. Pulit-Penaloza, J. A., Scherbik, S. V. & Brinton, M. A. Activation of Oas1a gene expression by type I IFN requires both STAT1 and STAT2 while only STAT2 is required for Oas1b activation. *Virology* **425.** PMID: 22305621, 71–81. ISSN: 1096-0341 (Apr. 2012).

165. R Development Core Team. *R: A Language and Environment for Statistical Computing* ISBN 3-900051-07-0. R Foundation for Statistical Computing (Vienna, Austria, 2011).

166. Raju, D. *et al.* Vacuolating Cytotoxin and Variants in Atg16L1 That Disrupt Autophagy Promote Helicobacter pylori Infection in Humans. *Gastroenterology.* ISSN: 0016-5085 (2012).

167. Rao, P., Hayden, M. S., Long, M., Scott, M. L., West, A. P., Zhang, D., Oeckinghaus, A., Lynch, C., Hoffmann, A., Baltimore, D. & Ghosh, S. IB acts to both inhibit and activate gene expression at different stages of the inflammatory response. *Nature* **466.** PMID: 20740013 PMCID: 2946371, 1115–1119. ISSN: 0028-0836 (Aug. 2010).

168. Rayhane, N., Fitting, C. & Cavaillon, J. Dissociation of IFN- from IL-12 and IL-18 Production During Endotoxin Tolerance. *Journal of Endotoxin Research* **5,** 319–324. ISSN: 1753-4259, 1753-4267 (Oct. 1999).

169. Reinhart, K. & Karzai, W. Anti-tumor necrosis factor therapy in sepsis: Update on clinical trials and lessons learned. *Critical Care Medicine* **29.** ISSN: 0090-3493 (2001).

170. Remmers, D. E., Cioffi, W. G., Bland, K. I., Wang, P., Angele, M. K. & Chaudry, I. H. Testosterone: the crucial hormone responsible for depressing myocardial function in males after trauma-hemorrhage. *Annals of Surgery* **227.** PMID: 9637542 PMCID: 1191378, 790–799. ISSN: 0003-4932 (June 1998).

171. Rius, J., Guma, M., Schachtrup, C., Akassoglou, K., Zinkernagel, A. S., Nizet, V., Johnson, R. S., Haddad, G. G. & Karin, M. NF-kB links innate immunity to the hypoxic response through transcriptional regulation of HIF-1a. *Nature,* 1–6 (2008).

172. Robinson, K., Kruger, P., Prins, J. & Venkatesh, B. The metabolic syndrome in critically ill patients. *Best Practice & Research Clinical Endocrinology & Metabolism* **25,** 835–845. ISSN: 1521-690X (Oct. 2011).

173. Salomao, R., Brunialti, M. K. C., Gomes, N. E., Mendes, M. E., Diaz, R. S., Komninakis, S., Machado, F. R., da Silva, I. D. C. G. & Rigato, O. Toll-like receptor pathway signaling is differently regulated in neutrophils and peripheral mononuclear cells of patients with sepsis, severe sepsis, and septic shock*. *Critical Care Medicine* **37,** 132–139. ISSN: 0090-3493 (Jan. 2009).

174. Saracco, P., Vitale, P., Scolfaro, C., Pollio, B., Pagliarino, M. & Timeus, F. The coagulopathy in sepsis: significance and implications for treatment. *Pediatric Reports* **3.** PMID: 22355515 PMCID: 3283198. ISSN: 2036-749X (Dec. 2011).

175. Schmid, R., Baum, P., Ittrich, C., Fundel-Clemens, K., Huber, W., Brors, B., Eils, R., Weith, A., Mennerich, D. & Quast, K. Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3. *BMC Genomics* **11,** 349. ISSN: 1471-2164 (June 2010).

176. Schröder, J., Kahlke, V., Book, M. & Stüber, F. Gender differences in sepsis: genetically determined? *Shock (Augusta, Ga.)* **14.** PMID: 11028548, 307–310, 307–310. ISSN: 1073-2322 (Sept. 2000).

177. Shapiro, S. S. & Wilk, M. B. An Analysis of Variance Test for Normality (complete Samples). *Biometrika* **52,** 591–611. ISSN: 0006-3444, 1464-3510 (Dec. 1965).

178. Shi, D., Zhang, J., Jiang, H., Tong, C., Gu, G., Ji, Y., Summah, H. & Qu, J. LPS pretreatment ameliorates multiple organ injuries and improves survival in a murine model of polymicrobial sepsis. *Inflammation Research* **60,** 841–849. ISSN: 1023-3830 (2011).

179. Shnyra, A., Brewington, R., Alipio, A., Amura, C. & Morrison, D. C. Reprogramming of Lipopolysaccharide-Primed Macrophages Is Controlled by a Counterbalanced Production of IL-10 and IL-12. *The Journal of Immunology* **160,** 3729–3736. ISSN: 0022-1767, 1550-6606 (Apr. 1998).

180. Smyth, G. K. in *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (eds Gentleman, R., Carey, V., Dudoit, S., Irizarry, R. & Huber, W.) 397–420 (Springer, New York, 2005).

181. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3.** PMID: 16646809, Article3. ISSN: 1544-6115 (2004).

182. Soneson, C., Lilljebjörn, H., Fioretos, T. & Fontes, M. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC Bioinformatics* **11,** 191. ISSN: 1471-2105 (Apr. 2010).

183. Sørensen, T. I., Nielsen, G. G., Andersen, P. K. & Teasdale, T. W. Genetic and Environmental Influences on Premature Death in Adult Adoptees. *New England Journal of Medicine* **318,** 727–732 (1988).

184. Spooner, C. E., Markowitz, N. P. & Saravolatz, L. D. The role of tumor necrosis factor in sepsis. *Clinical Immunology and Immunopathology* **62,** S11–S17. ISSN: 0090-1229 (Jan. 1992).

185. Su, X., Sykes, J. B., Ao, L., Raeburn, C. D., Fullerton, D. A. & Meng, X. Extracellular heat shock cognate protein 70 induces cardiac functional tolerance to endotoxin: Differential effect on TNF- and ICAM-1 levels in heart tissue. *Cytokine* **51.** PMID: 20478717 PMCID: 2905144, 60–66. ISSN: 1043-4666 (July 2010).

186. Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. & Mesirov, J. P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *October* (2005).

187. Tait, S. W. G. & Green, D. R. Mitochondria and cell death: outer membrane permeabilization and beyond. *Nature Reviews Molecular Cell Biology* **11,** 621–632. ISSN: 1471-0072 (Aug. 2010).

188. Tang, B. M. P., McLean, A. S., Dawes, I. W., Huang, S. J., Cowley, M. J. & Lin, R. C. Y. Gene-expression profiling of gram-positive and gram-negative sepsis in critically ill patients. *Critical Care Medicine* **36.** PMID: 18379237, 1125–1128. ISSN: 1530-0293 (Apr. 2008).

189. Tang, B. M. P., McLean, A. S., Dawes, I. W., Huang, S. J. & Lin, R. C. Y. Gene-expression profiling of peripheral blood mononuclear cells in sepsis. *Critical Care Medicine* **37.** PMID: 19237892, 882–888. ISSN: 1530-0293 (Mar. 2009).

190. Tang, B. M. P., McLean, A. S., Dawes, I. W., Huang, S. J. & Lin, R. C. Y. The Use of Gene-Expression Profiling to Identify Candidate Genes in Human Sepsis. *American Journal of Respiratory and Critical Care Medicine* **176,** 676–684. ISSN: 1073-449X, 1535-4970 (Oct. 2007).

191. Textoris, J., Ban, L. H., Capo, C., Raoult, D., Leone, M. & Mege, J. Sex-Related Differences in Gene Expression Following Coxiella burnetii Infection in Mice: Potential Role of Circadian Rhythm. *PLoS ONE* **5,** e12190 (2010).

192. Thiel, M., Caldwell, C. C., Kreth, S., Kuboki, S., Chen, P., Smith, P., Ohta, A., Lentsch, A. B., Lukashev, D. & Sitkovsky, M. V. Targeted deletion of HIF-1alpha gene in T cells prevents their inhibition in hypoxic inflamed tissues and improves septic mice survival. *PloS One* **2.** PMID: 17786224, e853. ISSN: 1932-6203 (2007).

193. Tofil, N. M., Aronow, B. J., Thomas, N. J., Wong, H. R., Doctor, A., Lin, R., Cvijanovich, N., Monaco, M., Shanley, T. P., Kalyanaraman, M., Allen, G. L., Barnes, M., Sakthivel, B., Penfil, S. & Odoms, K. Genome-level longitudinal expression of signaling pathways and gene networks in pediatric septic shock. *Molecular medicine (Cambridge, Mass.)* **13,** 495–508 (2007).

194. Tressel, S. L., Kaneider, N. C., Kasuda, S., Foley, C., Koukos, G., Austin, K., Agarwal, A., Covic, L., Opal, S. M. & Kuliopulos, A. A matrix metalloprotease-PAR1 system regulates vascular integrity, systemic inflammation and death in sepsis. *EMBO Molecular Medicine* **3.** PMID: 21591259, 370–384. ISSN: 1757-4684 (July 2011).

195. Turnbull, I. R. & Colonna, M. Activating and inhibitory functions of DAP12. *Nature Reviews Immunology* **7,** 155–161. ISSN: 1474-1733 (Jan. 2007).

196. Vadim V., S. LPS-induced Toll-like receptor 4 signalling triggers cross-talk of apoptosis signal-regulating kinase 1 (ASK1) and HIF-1 protein. *FEBS Letters* **582,** 319–326. ISSN: 0014-5793 (Jan. 2008).

197. Van der Poll, T., Calvano, S. E., Kumar, A., Braxton, C. C., Coyle, S. M., Barbosa, K., Moldawer, L. L. & Lowry, S. F. Endotoxin induces downregulation of tumor necrosis factor receptors on circulating monocytes and granulocytes in humans. *Blood* **86,** 2754–2759. ISSN: 0006-4971, 1528-0020 (Oct. 1995).

198. Van der Poll, T. & Levi, M. Crosstalk between Inflammation and Coagulation: The Lessons of Sepsis. *Current Vascular Pharmacology.* PMID: 22272914. ISSN: 1875-6212 (Jan. 2012).

199. Vega, V. L., de Cabo, R. & De Maio, A. Age and caloric restriction diets are confounding factors that modify the response to lipopolysaccharide by peritoneal macrophages in C57BL/6 mice. *Shock* **22,** 248–253. ISSN: 1073-2322 (Sept. 2004).

200. Viemann, D., Barczyk, K., Vogl, T., Fischer, U., Sunderkötter, C., Schulze-Osthoff, K. & Roth, J. MRP8/MRP14 impairs endothelial integrity and induces a caspase-dependent and -independent cell death program. *Blood* **109,** 2453–2460. ISSN: 0006-4971, 1528-0020 (Mar. 2007).

201. Vincent, J. L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C. K., Suter, P. M. & Thijs, L. G. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure. On behalf of the Working Group on Sepsis-Related Problems of the European Society of Intensive Care Medicine. *Intensive Care Medicine* **22.** PMID: 8844239, 707–710. ISSN: 0342-4642 (July 1996).

202. Vogl, T., Tenbrock, K., Ludwig, S., Leukert, N., Ehrhardt, C., Zoelen, M. A. D. v., Nacken, W., Foell, D., Poll, T. v. d., Sorg, C. & Roth, J. Mrp8 and Mrp14 are endogenous activators of Toll-like receptor 4, promoting lethal, endotoxin-induced shock. *Nature Medicine* **13,** 1042–1049. ISSN: 1078-8956 (Sept. 2007).

203. Wagener, F. A. D. T. G., Volk, H., Willis, D., Abraham, N. G., Soares, M. P., Adema, G. J. & Figdor, C. G. Different Faces of the Heme-Heme Oxygenase System in Inflammation. *Pharmacological Reviews* **55,** 551–571. ISSN: , 1521-0081 (Online) (Sept. 2003).

204. Walmsley, S. R., Print, C., Farahi, N., Peyssonnaux, C., Johnson, R. S., Cramer, T., Sobolewski, A., Condliffe, A. M., Cowburn, A. S., Johnson, N. & Chilvers, E. R. Hypoxia-induced neutrophil survival is mediated by HIF-1–dependent NF-B activity. *The Journal of Experimental Medicine* **201,** 105–115. ISSN: 0022-1007, 1540-9538 (Jan. 2005).

205. Werth, N., Beerlage, C., Rosenberger, C., Yazdi, A. S., Edelmann, M., Amr, A., Bernhardt, W., von Eiff, C., Becker, K., Schäfer, A., Peschel, A. & Kempf, V. A. J. Activation of Hypoxia Inducible Factor 1 Is a General Phenomenon in Infections with Human Pathogens. *PLoS ONE* **5,** e11576 (July 2010).

206. Wichmann, M. W., Inthorn, D., Andress, H. .-. & Schildberg, F. W. Incidence and mortality of severe sepsis in surgical intensive care patients: the influence of patient gender on disease process and outcome. *Intensive Care Medicine* **26,** 167–172. ISSN: 0342-4642, 1432-1238 (Mar. 2000).

207. Williams, T. A., Dobb, G. J., Finn, J. C., Knuiman, M. W., Geelhoed, E., Lee, K. Y. & Webb, S. A. R. Determinants of long-term survival after intensive care. *Critical Care Medicine* **36.** PMID: 18434893, 1523–1530. ISSN: 1530-0293 (May 2008).

208. Witten, D. M. & Tibshirani, R. J. Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data. *Statistical Applications in Genetics and Molecular Biology* **8.** ISSN: 1544-6115 (June 2009).

209. Wolk, K., Kunz, S., Crompton, N. E. A., Volk, H. & Sabat, R. Multiple Mechanisms of Reduced Major Histocompatibility Complex Class II Expression in Endotoxin Tolerance. *Journal of Biological Chemistry* **278,** 18030–18036. ISSN: 0021-9258, 1083-351X (May 2003).

210. Wong, H. R., Cvijanovich, N., Allen, G. L., Lin, R., Anas, N., Meyer, K., Freishtat, R. J., Monaco, M., Odoms, K., Sakthivel, B. & Shanley, T. P. Genomic expression profiling across the pediatric systemic inflammatory response syndrome, sepsis, and septic shock spectrum. *Critical care medicine* **37.** PMID: 19325468 PMCID: 2747356, 1558–1566. ISSN: 0090-3493 (May 2009).

211. 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., Monaco, M., Odom, K. & Shanley, T. P. Identification of pediatric septic shock subclasses based on genome-wide expression profiling. *BMC Medicine* **7.** PMID: 19624809 PMCID: 2720987, 34. ISSN: 1741-7015 (July 2009).

212. Wong, H. R., Freishtat, R. J., Monaco, M., Odoms, K. & Shanley, T. P. Leukocyte subset-derived genome-wide expression profiles in pediatric septic shock. *Pediatric critical care medicine : a journal of the Society of Critical Care Medicine and the World Federation of Pediatric Intensive and Critical Care Societies* **11.** PMID: 20009785 PMCID: 2927840, 349–355. ISSN: 1529-7535 (May 2010).

213. 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., Penfil, S., Monaco, M., Tagavilla, M. A., Odoms, K., Dunsmore, K., Barnes, M. & Aronow, B. J. Genome level expression profiles in pediatric septic shock indicate a role for altered zinc homeostasis in poor outcome. *Physiological genomics* **30.** PMID: 17374846 PMCID: 2770262, 146–155. ISSN: 1094-8341 (July 2007).

214. Wong, H. R., Wheeler, D. S., Tegtmeyer, K., Poynter, S. E., Kaplan, J. M., Chima, R. S., Stalets, E., Basu, R. K. & Doughty, L. A. Toward a clinically feasible gene expression-based sub-classification strategy for septic shock: proof of concept. *Critical care medicine* **38.** PMID: 20639748 PMCID: 2943553, 1955–1961. ISSN: 0090-3493 (Oct. 2010).

215. Wray, G. M., Foster, S. J., Hinds, C. J. & Thiemermann, C. A cell wall component from pathogenic and non-pathogenic gram-positive bacteria (peptidoglycan) synergises with endotoxin to cause the release of tumour necrosis factor-alpha, nitric oxide production, shock, and multiple organ injury/dysfunction in the rat. *Shock (Augusta, Ga.)* **15.** PMID: 11220642, 135–142. ISSN: 1073-2322 (Feb. 2001).

216. Wysocka, M., Robertson, S., Riemann, H., Caamano, J., Hunter, C., Mackiewicz, A., Montaner, L. J., Trinchieri, G. & Karp, C. L. IL-12 Suppression During Experimental Endotoxin Tolerance: Dendritic Cell Loss and Macrophage Hyporesponsiveness. *The Journal of Immunology* **166,** 7504–7513. ISSN: 0022-1767, 1550-6606 (June 2001).

217. Xiao, W. *et al.* A Genomic Storm in Critically Injured Humans. *The Journal of Experimental Medicine.* ISSN: 0022-1007, 1540-9538 (Nov. 2011).

218. Xiong, Y. & Medvedev, A. E. Induction of endotoxin tolerance in vivo inhibits activation of IRAK4 and increases negative regulators IRAK-M, SHIP-1, and A20. *Journal of Leukocyte Biology* **90.** PMID: 21934070, 1141–1148. ISSN: 1938-3673 (Dec. 2011).

219. Yang, K. *et al.* Human TLR-7-, -8-, and -9-Mediated Induction of IFN-/ and - Is IRAK-4 Dependent and Redundant for Protective Immunity to Viruses. *Immunity* **23,** 465–478. ISSN: 1074-7613 (Nov. 2005).

220. Yang, Q., Calvano, S. E., Lowry, S. F. & Androulakis, I. P. A dual negative regulation model of Toll-like receptor 4 signaling for endotoxin preconditioning in human endotoxemia. *Mathematical Biosciences* **232,** 151–163. ISSN: 0025-5564 (Aug. 2011).

221. Yang, Q., Mattick, J. S. A., Orman, M. A., Nguyen, T. T., Ierapetritou, M. G., Berthiaume, F. & Androulakis, I. P. Dynamics of Hepatic Gene Expression Profile in a Rat Cecal Ligation and Puncture Model. *The Journal of Surgical Research.* PMID: 22381171. ISSN: 1095-8673 (Dec. 2011).

222. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. & Speed, T. P. Normalization for cDNA Microarray Data: A Robust Composite Method Addressing Single and Multiple Slide Systematic Variation. *Nucleic Acids Research* **30,** e15–e15. ISSN: 0305-1048, 1362-4962 (Feb. 2002).

223. Yang, Y., Yang, K. S., Hsann, Y. M., Lim, V. & Ong, B. C. The effect of comorbidity and age on hospital mortality and length of stay in patients with sepsis. *Journal of Critical Care* **25.** PMID: 19836195, 398–405. ISSN: 1557-8615 (Sept. 2010).

224. Zellweger, R., Wichmann, M. W., Ayala, A., Stein, S., DeMaso, C. M. & Chaudry, I. H. Females in proestrus state maintain splenic immune functions and tolerate sepsis better than males. *Critical Care Medicine* **25.** PMID: 8989185, 106–110. ISSN: 0090-3493 (Jan. 1997).

225. Zhang, Q., Raoof, M., Chen, Y., Sumi, Y., Sursal, T., Junger, W., Brohi, K., Itagaki, K. & Hauser, C. J. Circulating mitochondrial DAMPs cause inflammatory responses to injury. *Nature* **464,** 104–107. ISSN: 0028-0836 (Mar. 2010).

226. Zschiedrich, I., Hardeland, U., Krones-Herzig, A., Berriel Diaz, M., Vegiopoulos, A., Müggenburg, J., Sombroek, D., Hofmann, T. G., Zawatzky, R., Yu, X., Gretz, N., Christian, M., White, R., Parker, M. G. & Herzig, S. Coactivator function of RIP140 for NFB/RelA-dependent cytokine gene expression. *Blood* **112,** 264–276. ISSN: 0006-4971, 1528-0020 (July 2008).