THE UNIVERSITY
of EDINBURGH

Computational investigation of systemic pathway responses in severe pneumonia among the Gambian children and infants

James Jafali

A thesis presented for the degree of

Doctor of Philosophy,

THE UNIVERSITY of EDINBURGH

January 2018

Division of Infection & Pathway Medicine
College of Medicine & Veterinary Medicine
THE UNIVERSITY of EDINBURGH

Declaration

I hereby declare that this thesis and the work presented in it are my own.

I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University
- Where any part of this thesis has previous been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all the main sources of help
- Where the thesis is based on work done by myself and jointly with others, I have stated clearly what was done by others and what I have contributed myself

James Jafali

THE UNIVERSITY of EDINBURGH

Abstract

Pneumonia remains the leading cause of infectious mortality in under-five children, and the burden is highest in sub-Saharan Africa. To mitigate this burden, further knowledge is required to accelerate the development of innovative and cost-effective approaches. To gain a deeper insight into the pathogenesis of pneumonia, I investigated the central hypothesis that systemic pathway (cellular and molecular) responses underpin the development of severe pneumonia outcomes.

Mainly, I compared whole blood transcriptomes between severe pneumonia cases (clinically stratified as mild, severe and very severe) and non-pneumonia community controls (prospectively matched by age and sex). In total, 803 whole blood RNA samples were collected from Gambian children (aged 2-59 months) between 2007 and 2010, of which, 518 passed laboratory quality control criteria for the microarray analysis. After data cleaning, the final database reduced to 503 samples including the training (n=345) and independent validation (n=158) data sets.

To investigate the cellular responses, I applied computational deconvolution analysis to assess the variations of immune cell type proportions with pneumonia severity. To further enhance the computational performance, I applied a data fusion approach on 3,475 immune marker genes from different resources to derive an optimal and integrated blood marker list (IBML, m=277) for Neutrophils, Monocytes, NK, Dendritic, B and T cell types; which robustly performed better than the existing individual resources. Using the IBML resource, pneumonia severity was significantly associated with the depletion of B, T, Dendritic and NK cell types, and the elevation of Monocytes and neutrophil proportions (P-value<0.001).

At the molecular level, pneumonia severity was associated (false discovery rate<0.05) with a battery of systemic pathway (innate, adaptive and metabolic) responses in a range of biomedical databases. While the up-regulation of inflammatory innate responses was also observed in mild cases, severe pneumonia cases were predominantly associated with the co-inhibition of the cells of the adaptive immune response (B and T) and Natural killer cells, and the up-regulation of fatty acid and lipid metabolism. While most of these findings were anticipated, the involvement of NK cells was unexpected, and potentially presents a novel immune-modulation target for mitigating the burden of pneumonia. Together, the cellular and molecular pathways responses consistently support the central hypothesis that

systemic pathway responses contribute significantly to the development of severe pneumonia outcomes.

Clinically, the identification and appropriate treatment of patients at the higher risk of developing severe pneumonia outcomes remains the major challenge. To address that, I applied supervised machine-learning approaches on cellular pathway based transcriptomic features; and derived a 33-gene classifier (representing the NK, T, and neutrophils cell types), which accurately detected severe pneumonia cases in both the training (leave-one-out cross-validated accuracy=99%) and independent validation (accuracy=98%) datasets. Independently, similar performance (98% in each dataset) was associated with a subset (m=18) of the validated 52-gene neonatal sepsis classifier. Conversely, at least 75% of the cellular biomarkers were differentially expressed (false discovery rate<0.05) in bacterial neonatal sepsis. Further, very severe pneumonia cases were predominantly associated with antibacterial responses; and mild pneumonia cases with blood-culture-confirmed positivity were also associated with an increased frequency of differentially expressed genes. These findings suggest the significant contribution of bacterial septicaemia in the development of serious pneumonia outcomes. Together, this study highlights the future potential of host-derived systemic biomarkers for early identification and novel treatment modalities of high-risk cases presenting at a resource-constrained clinic with mild pneumonia. However, further validation studies are required.

Lay summary

Pneumonia (an infection of the lungs) is the leading cause of deaths in under-five children (more than HIV/AIDS, malaria and diarrhoea combined), especially in sub-Saharan Africa. To reduce this burden, innovative and cost-effective approaches are required. However, the development of severe pneumonia outcomes is not fully described.

This thesis investigated whether the development of pneumonia severity is associated with the immune responses that are detectable in the blood, which is clinically accessible. To address this, we applied a comprehensive approach called *transcriptomics* to compare whole blood samples between pneumonia cases (clinically classified as mild, severe and very severe pneumonia) and similar non-pneumonia community controls. The blood samples were collected from the Gambian children (2-59 months old) between 2007 and 2010. After data quality assurance, 345 samples were applied in the main analyses while 158 samples were kept for independent validations.

Whole blood is a complex mixture of a range of immune cell types and molecules, which vary in concentration between individuals with different conditions. In the main analyses, pneumonia severity was investigated at the cellular and molecular levels, and the findings were consistently associated with strong inflammatory responses. To investigate the cellular responses, I applied a powerful and cost-effective approach called *computational deconvolution analysis*. Firstly, I developed a data resource tool called IBML, which further enhanced the computational performance. Using this resource (IBML), pneumonia severity was simultaneously associated with an accumulation of pro-inflammatory cell types (neutrophils and monocytes), and the depletion of anti-inflammatory mediators (B, T, and Dendritic). At the molecular level, while the inflammatory responses were observed from mild to very severe pneumonia, severe pneumonia cases were predominantly associated with loss of regulatory control mechanisms. While these findings were anticipated, here pneumonia severity was unexpectedly associated with significant depletion and inhibition of natural killer (NK) cells. Potentially, this finding presents a novel intervention target for preventing serious outcomes in pneumonia.

Further, very severe pneumonia cases were predominantly associated with bacterial infections, which highlight the importance of early identification and appropriate treatment of cases at the higher risk of developing severe pneumonia

outcomes. To enhance the identification of high-risk pneumonia cases, here I developed a 33-gene molecular diagnostic tool, which accurately detected at least 98% of the serious pneumonia cases. Independently, these findings shared significant similarities with bacterial sepsis (blood infection), which strongly suggest the importance of blood-based (i.e. systemic) responses in the development of serious pneumonia outcomes. Together, this study highlights the potential of whole blood host-based signatures for future clinical identification and treatment of high-risk pneumonia cases especially in resource-limited settings where the burden is highest. However, further validation studies are required.

Acknowledgements

Firstly, I would like to thank my parents for the unique love, support, encouragement, and guidance. However, it takes a community to raise a child in Africa. Am therefore indebted to many people in my community including my grandparents, uncles, aunts, my many village friends, and mentors. Special thanks to my girlfriend and my best friend, *Maria*, for her encouragement, support, love, and patience throughout my PhD seclusion.

I am very grateful to my sponsors MRC Unit The Gambia (MRCG) and the Edinburgh Global Research Scholarship, and my supervisors for the opportunity to be part of this great project. I was privileged to have a strong team of unique supervisors, and I am indebted to their support, inspiration, suggestions, encouragement and the stimulating discussions throughout my studentship. Special thanks to my principal supervisor Professor Peter Ghazal and *Dr Thorsten Forster* for the weekly meetings and their unlimited accessibility; *Dr Stephen Howie* for all the clinical and epidemiological aspect of this project; *Dr David Jeffries* & Dr *Thorsten Forster* for the statistical and bioinformatics support; and *Dr Paul Dickinson* for the babysitting introduction to this project. Am also indebted to Professor *Harry Campbell* for his sacrifice as the thesis committee chair.

I take this opportunity to extend my gratitude to my previous sponsors (the Malawi government for the BSc scholarship, and the LSHTM's Tropical Epidemiological Group for the MSc fellowship); and my previous teachers for the invaluable knowledge, skills, morals, guidance, and inspirations. They won't be forgotten including my first teacher *Mrs. Mwangalika; Ustaz Amin Chiumbe, Mr. Kaliati, Mrs. Gondwe, Prof. Mercy Kazima, Prof. Tobias Chirwa and Prof. J.J. Namangale, and Prof. Shabar Jaffri and Dr. Christian Bottomley:* This list is too long….

During this PhD expedition, I found heaven in two vibrant and unique research communities: MRC unit, The Gambia (MRCG) and the DIPM "signalling pathway". I have enjoyed the experience and learned a lot from many people in both institutions. At MRCG, special thanks to the Research Training & Career Development team (Mariatou Sallah, *Adam Drammeh*, Ismaila Danso & Dr. Assan Jay) for all the administrative support. At the DIPM, many thanks to all the administrators, P.I.s and students: special thanks to Alan Ross and Marie Craigon for processing my RNA samples, and of course my brother in football Richard Perry for his invariable support. I also found the university system very helpful: thanks to all:)

Last but not least, I humbly thank all the children and infants who sacrificed their blood to participate in this study. Further, special credit to all the people who were involved in this project including the field workers, research clinicians (*Dr. Osaretin Omoroyi*, *Dr. Readon Ideh*, and *Dr. Bernard Ebruke*), scientific officers (*Eunice Machuka*), laboratory technicians, data managers, administrators and collaborators for their invaluable contribution to this project.

Above all, I thank GOD for everything and all the people around me.

Seek knowledge from the cradle to the grave

(Prophet Mohammed, PBUH)

Table of contents

List of figures

List of tables

List of abbreviations and acronyms

Abbreviation	Definition
ACSL1	Acyl-CoA synthetase long-chain family member 1
ACVR1B	Activin A receptor, type IB
ADCY3	Adenylate cyclase 3
ALPL	Alkaline phosphatase, liver/bone/kidney
ANKRD22	Ankyrin repeat domain 22
APC	Antigen presentaion cells
ARG1	Arginase 1
AUC	Area under the curve
BASP1	Brain abundant, membrane attached signal protein 1
BCL11B	B-cell CLL/lymphoma 11B (zinc finger protein)
BCL6	B-cell CLL/lymphoma 6
BEX2	Brain expressed X-linked 2
BP	Biological process
C19orf70	Chromosome 19 open reading frame 70
C1QB	Complement component 1, q subcomponent, B chain
CACNA2D3	Calcium channel, voltage-dependent, alpha 2/delta subunit 3
CCGs	Cell correlated genes
CD	Cluster of differentiation
CD177	CD177 molecule
CD247	CD247 molecule
CD5	CD5 molecule
CD7	CD7 molecule
CDHR3	Cadherin-related family member 3
CEACAM1	carcinoembryonic antigen-related cell adhesion molecule 1 (biliary glycoprotein)
CellSep	Aggregation of cellular pathway centric and validated sepsis biomarkers
CLC	Charcot-Leyden Crystal Galectin
CLC	Charcot-Leyden crystal galectin
CLEC4D	C-type lectin domain family 4, member D
CLEC5A	C-type lectin domain family 5, member A
CMTM2	CKLF-like MARVEL transmembrane domain containing 2
CNIH4	Cornichon family AMPA receptor auxiliary protein 4
CPPED1	calcineurin-like phosphoesterase domain containing 1
CSF3R	Colony stimulating factor 3 receptor (granulocyte)
CYP1B1	Cytochrome P450, family 1, subfamily B, polypeptide 1
DCGs	Differentially correlated genes
DEGs	Differentially expressed genes
DOCK5	Dedicator of cytokinesis 5
DSC2	Desmocollin 2
ECM	Extracellular matrix
ELOVL4	ELOVL fatty acid elongase 4
ENTPD7	Ectonucleoside triphosphate diphosphohydrolase 7
EXOC6	Exocyst complex component 6

FAM20A	Family with sequence similarity 20, member A
FC	Fold change
FCER1A	Fc fragment of IgE, high affinity I, receptor for; alpha polypeptide
FDR	False discovery rate
FGR	FGR proto-oncogene, Src family tyrosine kinase
FKBP5	FK506 binding protein 5
FLOT2	Flotillin 2
GALC	Galactosylceramidase
GCRMA	Guanine Cytosine Robust Multi-Array Analysis
GEO	Gene expression Ominbus
GO	Gene ontology
GPR114	G protein-coupled receptor 114
GPR160	G protein-coupled receptor 160
GRINA	Glutamate receptor, ionotropic, N-methyl D-aspartate-associated protein 1 (glutamate binding)
GSEA	Gene set expresion analsysis
GYG1	Glycogenin 1
HAZ	Height-for-age Z score
HIV	Human Immunodeficiency Virus
HP	Haptoglobin
IBML	Integrated Blood Marker List
ID3	Inhibitor of DNA binding 3, dominant negative helix-loop-helix protein
IGA	Immunoglobulin A
IL	Interleukine
IL18R1	Interleukin 18 receptor 1
IL18RAP	Interleukin 18 receptor accessory protein
IQR	Interquartile range
IRAK3	Interleukin-1 receptor-associated kinase 3
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIF1B	Kinesin family member 1B
KLF7	Kruppel-like factor 7 (ubiquitous)
KLRB1	Killer cell lectin-like receptor subfamily B, member 1
KNN	K-nearest neighbour
LAT	Linker for activation of T cells
LCN2	Lipocalin 2
LDA	Liner descrimanatory analsysis
LDLR	Low density lipoprotein receptor
LILRA4	Leukocyte immunoglobulin-like receptor, subfamily A (with TM domain), member 4
LINC01000	Long intergenic non-protein coding RNA 1000
LRRC70	Leucine rich repeat containing 70
LRRN1	Leucine rich repeat neuronal 1
LRRN3	Leucine-rich repeat neuronal protein 3
LRRN3	Leucine rich repeat neuronal 3
m	Number of genes
M	Number of tests

MAL	Mal, T-cell differentiation protein
MAPK	Mitogen-activated protein kinase
MAPK14	Mitogen-activated protein kinase 14
MAS	Microarray Analysis Suite
MBEI	Model-Based Expression Index
MBOAT7	Membrane bound O-acyltransferase domain containing 7
MCEMP1	Mast cell-expressed membrane protein 1
MF	Molecular function
MGAM	Maltase-glucoamylase (alpha-glucosidase)
MGR	Marker gene resources
MMP9	Matrix metallopeptidase 9
MMP9	Matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)
MRC	Medical Research Council
MSvS	Differentially expressed genes in mild, severe and very severe pneumonia
n	Number of samples
N	Population size
NCBI	National Center for Biotechnology Information
NECAB1	N-terminal EF-hand calcium binding protein 1
NF-kB	Nuclear factor kappa-light-chain-enhancer of activated B cells
NFIL3	Nuclear factor, interleukin 3 regulated
NK	Natural killer
NLRs	NOD-like receptors
NOD	Nucleotide-binding oligomerization domain
NUSE	Normalized Unscaled Standard Error
PCR	Polymerase chain reaction
PD1	Programmed death protein-1
PDZD4	PDZ domain containing 4
PECAM-1	Platelet endothelial cell adhesion molecule
PGAP3	Post-GPI attachment to proteins 3
PHF21A	PHD finger protein 21A
PPP1R3B	Protein phosphatase 1, regulatory subunit 3B
PRRs	Pathogen recognition receptors
PTGDR2	Prostaglandin D2 receptor 2
RAGE	Receptor for Advanced Glycation End products
RAPGEFL1	Rap guanine nucleotide exchange factor (GEF)-like 1
RF	Random forest
RLE	Relative Log Expression
RMA	Robust Multi-array Average
RNA	Ribonucleic acid
ROC	Receiver operation characteristics
ROCC	Receiver operation characteristics (ROC) based classifier
RVTH	The Royal Victoria Teaching Hospital
SCC	Scientific coordinating committee
SD	Standard deviation

SD	Standard deviation
SIRPG	Signal-regulatory protein gamma
SLC26A8	Solute carrier family 26 (anion exchanger), member 8
SLC2A14	Solute carrier family 2 (facilitated glucose transporter), member 14
SLC2A3	Solute carrier family 2 (facilitated glucose transporter), member 3
SLC2A4RG	SLC2A4 regulator
SLC37A3	Solute carrier family 37, member 3
SLC4A10	Solute carrier family 4, sodium bicarbonate transporter, member 10
SRPK1	SRSF protein kinase 1
ST20	Suppressor of tumorigenicity 20
STOM	Stomatin
SVM	Support vector machines
SvS	Differentially expressed genes in severe and very severe pneumonia
TCTN1	Tectonic family member 1
TECPR2	Tectonin beta-propeller repeat containing 2
TLRs	Toll-like receptors
TNF	Tumour necrosis factor
Under-five	Children younger than five years old
VNN1	Vanin 1
Vs	Differentially expressed genes in very severe pneumonia
VSN	Variance stabilisation normalisation
VSNRMA	Variance stabilisation normalisation (VSN) with median polish summarization as in the RMA method
WAZ	Weight-for-age Z score
WHO	World Health Organisation
WWOX	WW domain containing oxidoreductase

Chapter 1: Introduction

1.1 Introduction

This chapter presents an overview introduction of childhood pneumonia including the burden, pathogenesis, clinical practice challenges and host responses. Further, the implications of genome-wide profiling (particularly whole blood transcriptomics) in elucidating the pathway biology of pneumonia are highlighted. Finally, the central hypothesis, objectives and thesis outline are introduced.

1.1 The burden of childhood pneumonia

Child survival remains a major public health challenge worldwide especially in resource-limited countries such as the Sub-Saharan Africa and South-Eastern Asia [1-3]. Despite the scaled efforts to implement safe, effective and affordable interventions; infections still account for more than 50% of child fatality cases *[4, 5]*. In particular, pneumonia remains the single leading cause of mortality in children younger than five years old, with an estimated burden of one million deaths in 2013[2, 6].

Indeed, child mortality rates have fallen in the past decade [2, 5, 7-9]. However, the progress has largely depended on the country-specific wealth index thereby leaving resource-limited countries behind with a high burden of childhood pneumonia *[4]*. Therefore, robust and affordable approaches are urgently required to accelerate the reduction of childhood pneumonia burden in resource-limited settings like the sub Saharan Africa. However, such

approaches require a better understanding of the pathogenesis and host responses in childhood pneumonia infection and severity, which is not fully understood. In the next section, the risk factors for pneumonia are briefly discussed.

1.2 Risk factors for pneumonia

Pneumonia has several correlated risk factors both from the host and the environment. Firstly, exposure to potential pathogens is the main prerequisite risk factor for disease onset and a reservoir for transmission [10, 11]. Consequently, children living in the high burden regions like the sub-Saharan Africa have an increased risk. Notably, both pneumococcal diseases and asymptomatic carriage of pneumococcal strains are very high in this region including the Gambian children [5, 9, 10, 12, 13]

While exposure is an important risk factor, infectious pneumonia is often high among vulnerable people associated with compromised host responses. Consequently, underlying host factors such as extreme age (younger children or the elderly), low birth-weight, premature birth, malnutrition including micronutrients (i.e. zinc) deficiencies and suboptimal breast-feeding, and co-morbidities such as HIV AIDS, diarrhoea, malaria and asthma are among the host risk factors [7, 9, 11, 14, 15].

Further, environmental factors such living in crowded conditions, and exposure to in-door air pollution like biofuels and passive smoking compromise the epithelial host defense mechanisms[6, 16]. Furthermore, while pneumonia occurs throughout the year, seasonal variations in disease

incidence and aetiology have also been reported [17-20]. While it is difficult to quantify the magnitude of each risk factor (i.e. because they are highly correlated), the major themes for pneumonia risk factors are (i) underlying host factors and (ii) environmental exposures [16]. However, effective preventive measures against these risk factors are lacking especially in resource-limited settings (next section)

With regard to this thesis and subsequent studies, these risk factors present a potential source of confounding effects. To mitigate this, it is worth noting that this study applied a matched-case control study design to account for the potential confounding effects of age, sex and residential area (**Chapter 2**). However, it is not feasible to match for many factors at the study design level. Therefore, potential confounding factors were further investigated and accounted for during data analysis (**Chapter 3)**.

1.3 Prevention of pneumonia

"P*revention is better than cure" (*Desiderius Erasmus). Indeed, diagnosis and treatment of pneumonia remain the clinical challenges [1, 21-24] (next section). Therefore, prevention of pneumonia remains a public health priority [10, 25, 26]. In particular, innovative and integrated approaches targeting both pneumonia and the co-morbidities such as HIV, malaria and diarrhoea are required to eliminate potential pathogens and strengthen the host, and households, community and the health systems.

At the population level, socio-economic empowerment and community sensitization are vital to minimise the underlying risk factors such as poor

hygiene, malnutrition, over-crowding and in-door pollution. Further, strengthening of health systems (i.e. to enhance optimal diagnosis, and effective management of patients), especially in remote area where the burden is often high, is a corner stone for reducing the burden of pneumonia [27]. However, derivation and implementation of such approaches require evidence-based knowledge to influence policy change and stimulate the political will.

At the host level, vaccination remains the most successful protection against harmful diseases including pneumonia [2, 9, 16, 28]. However, pneumonia has a complex aetiology, and vaccines against many pathogens such as respiratory syncytial virus (RSV) are not available [29]. Further, while the coverage of pneumococcal conjugate vaccine (PCV) is very low in resource-limited countries, the increasing prevalence of non-vaccine serotypes (NVT) in high-coverage settings is becoming worrisome [10]. Thus, novel approaches are required to gain a deeper insight into the pathogenesis of pneumonia to facilitate the discovery of better vaccine targets and candidates.

1.4 Pathogenesis of pneumonia

Pneumonia is a disease of the lung parenchyma in the lower respiratory tract causing mild to very severe outcomes across all ages but more prevalent among the very young and very old age groups (under-five children and the elderly over 60 years old) and people with compromised immunity [30-32]. In the affected lungs, the characteristic features include consolidation of the

affected part and the alveolar air spaces are filled with exudate, inflammatory cells, and fibrin[33]. The main function of the respiratory tract system, which is divided into lower and upper tracts (**Figure1.1**), is to supply oxygen and remove carbon dioxide from the body. However, the air is often contaminated with allergens, toxic chemicals and potential invasive pathogens capable of causing serious diseases such as pneumonia (lungs), sepsis (blood) and meningitis (brain)[34].

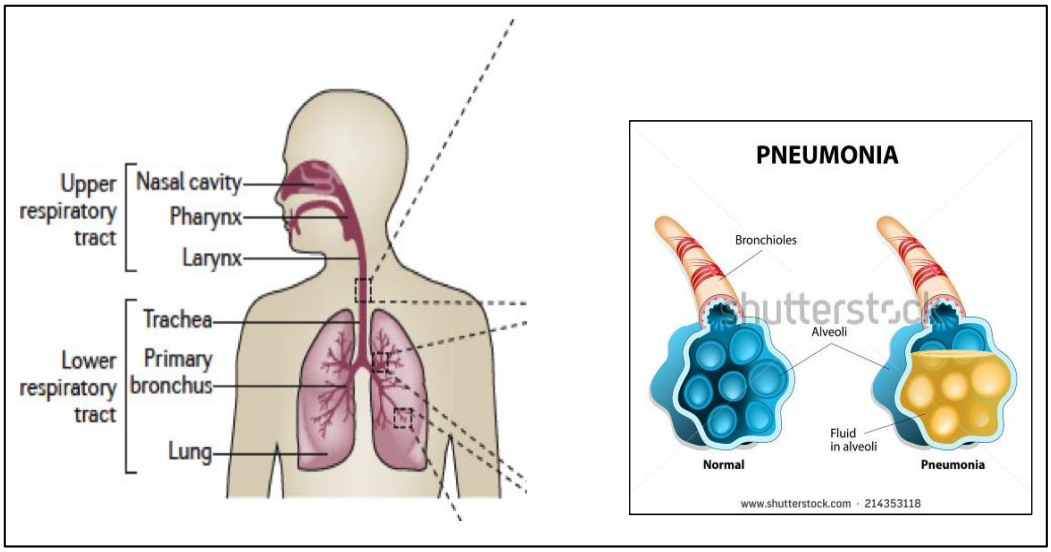


Figure 1.1: The respiratory tract system[31] showing pneumonia infected alveoli. The left and right figures were copied from Iwasaki (2016), page 3 (http://dx.doi.org/10.1038/nri.2016.117), and www.momjunction.com (respectively).

Normally, the upper respiratory system is colonised by commensal microorganisms while physical and chemical barriers protect the lower respiratory tract system [35, 36]. However, due to host and environmental risk factors as well as microbial virulence factors, these barriers can be breached [16, 37, 38]. When invasive bacteria are detected in the lower region, an inflammatory host response that includes the recruitment of neutrophils from the blood to the alveoli[39] are induced.

While these host responses are beneficial to eliminate the pathogen, deleterious inflammatory responses are often associated with disease severity causing tissue injury, pain and accumulation of debris from dead cells in the alveoli[38]. Consequently, these physiological changes compromise normal lung functions (gas exchange) and are often manifested in a range of clinical symptoms such as fever, difficulty in breathing or hypoxia even after the pathogen is cleared[40]. Therefore, a deeper understanding of the systems-level responses is vital to elucidate novel immuno-modulation factors responsible for excessive responses and severe outcomes. In the next section, the main causes of pneumonia are described.

1.5 Aetiology of pneumonia

Infectious pneumonia has a complex temporal-spatial aetiology including bacteria, viruses, fungi and atypical bacteria, which also cause meningitis (brain), sepsis (blood), media otitis (ear) and sinusitis (sinuses) across different geographical regions and seasons [28, 41, 42]. While no pathogen is identified in almost half of the clinical pneumonia cases[5, 16, 28], *Streptococcus pneumoniae* and *respiratory syncytial virus* (RSV) are the most common causes of bacterial and viral pneumonia, respectively[43]. In sub-Saharan Africa, bacterial pneumonia is more prevalent and is associated with more serious outcomes than viral pneumonia[9]. In particular, *Streptococcus pneumoniae* is the leading cause (33%) of pneumonia-related mortality in children younger than five years [5]. Other strains include *Haemophilus influenza, LegionellaChlamydia pneumoniae and mycoplasma pneumoniae* bacteria as well as *Influenza* and other virus.

Further, co-infection of bacterial with viral pathogens or other underlying morbidities such HIV, malaria and diarrhoea are frequently associated with severe outcomes. Similarly, the superinfection of viral and bacterial pathogens is common in the lungs (i.e. following an Influenza virus infection) [44-47].

Notably, the ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*,*Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* species) are frequently associated with severe nosocomial pneumonia and antibiotic resistance throughout the world [48]. Furthermore, *Pneumocystis jiroveci* is commonly isolated in HIV infected children *[27, 49, 50].* Due to this complex aetiology and co-morbidities, optimal aetiological stratification of pneumonia cases remains a major clinical challenge [28, 32, 40] (next section).

1.6 Clinical diagnosis of childhood pneumonia

1.6.1 Available approaches

Early and optimal aetiological stratification of cases presenting at the clinic with mild pneumonia is vital to prevent severe outcomes (associated with bacterial pneumonia) and mitigate the spread of antibiotic resistance due to unnecessary presumptive antibiotic treatment [24, 27]. However, existing diagnostic tools beyond physical examinations such as ***chest X-rays*** and ***culture-based*** *assays on blood or respiratory samples (induced sputum (IS),* nasopharyngeal airway (NPA), bronchoalveolar lavage (BAL)[51, 52]) are too expensive for resource-constrained health facilities, and have several

limitations [5, 40, 53, 54]. In particular, chest X-rays cannot establish disease aetiology, and frequent exposure to radiation is associated with long-term side effects [40]. Further, blood culture results are not rapid (24 to 48 hours) and have low sensitivity (<80%) [53, 55]. Consequently, empirical treatment (using clinical signs) of suspected bacterial cases is not uncommon.

HIV status	Signs	Treatment category	Treatment	
			First line	Second line
Any	Cough or cold	No pneumonia	Home care advice	Further diagnosis
Negative and less exposed children	Fast breathing or lower chest indrawing	Pneumonia	Oral amoxicillin (home therapy)	Referral
	Danger signs	Severe or very severe pneumonia	• Parenteral ampicillin (or penicillin) and gentamicin • Supportive therapy	Parenteral Ceftriaxone
HIV infected or exposed children	Chest-indrawing pneumonia or severe pneumonia	Severe or very severe pneumonia	• Parenteral ampicillin (or penicillin) and gentamicin • Supportive therapy	Parenteral Ceftriaxone
			Parenteral ampicillin (or penicillin) and Ceftriaxone	
	Suspected *Pneumocystis jirovecii*	Infants	Cotrimoxazole (additional)	

Table 1.1: Empirical treatment of childhood pneumonia. These are the revised guidelines (2014) by the World Health Organisation (WHO) [27] that were accessed at: http://www.who.int/maternal_child_adolescent/documents/child-pneumonia-treatment/en/

To minimise referrals, delayed treatment and severe outcomes, the World Health Organization (WHO) improvised highly sensitive guidelines (**Table1.1**) for enhancing treatment of suspected bacterial pneumonia at primary health facilities [27]. However, this criterion *is* consequently depleting antibiotic stocks and potentially exacerbating the spread of antibiotic resistance because bacterial pneumonia can have similar clinical presentations as viral pneumonia as well as other common infections such as malaria and

diarrhoea [40]. Moreover, the diagnosis of pneumonia by physical evaluation requires well-trained and experienced personnel who are rarely available in remote areas [56]. Therefore, new approaches are required to enhance clinical stratification of pneumonia. Potentially, gaining a deeper insight into the systemic pathway biology of pneumonia presents an innovative approach to derive robust diagnostic and prognostic biomarkers for optimal stratification of pneumonia cases (next section).

1.6.2 Potential approaches (Biomarkers)

Stratification of patients is a common challenge in many disease areas including cancer and infections; and the potential use of biomarkers present an attractive and cost-effective alternative solution [57]. Biomarkers are defined as *biological characteristics that can be objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention*[58, 59]. In pneumonia, diagnostic (to stratify bacterial and viral pneumonia cases) and prognostic (to predict mild cases at the higher risk of developing serious outcomes including mortality) biomarkers are urgently required.

In particular, blood-based biomarkers, which are clinically accessible, present a powerful and objective approach for enhancing clinical stratification and appropriate treatment of pneumonia cases. However, the application of single serum biomarkers such as blood counts, interleukins, C-reactive protein (CRP) and *procalcitonin* are lacking robustness and far from optimal [60-65]. Generally, the application of individual biomarkers does not capture

the complex pathway responses underpinning the pathogenesis of pneumonia [57, 58, 62, 65].

Recently, synergistic advances in genome-wide profiling experiments and data science approaches have spurred the investigation of molecular biomarkers, which are more robust and could potentially translate into point-of-care tests through technologies such PCR [66-68]. To date, molecular biomarkers include the occurrence of (i) genetic polymorphisms, gene mutations and methylation markers, and (ii) changes in RNA (gene expression) and microRNA (miRNA) abundance [69]. In particular, whole blood genome-wide profiling has become the mainstay of genomic research and future clinical practice. Importantly, whole blood is comprehensive and rich with biomarkers, and readily accessible tissue for clinical pathophysiological investigations. Therefore, systemic pathway-based biomarkers potentially present a powerful and accessible approach for enhancing clinical stratification of pneumonia cases.

However, successful translation of molecular biomarkers into clinical practice has several challenges including high cost, inadequate study design and statistical analyses, and limited accessibility of samples or impracticability (i.e. difficult implementation of protocols) [58, 70]. Firstly, many studies have applied retrospective and observational designs, which are often underpowered. In particular, observational studies are susceptible to potential bias and confounding effects, which undermines their generalizability. Further, many candidate biomarkers have lacked adequate

validations and follow–up studies either due to lack of similar studies or limited resources[58, 70]. Nevertheless, with the cost of genomic profiling going down, carefully designed studies investigating systemic pathway-based biomarkers such as whole blood transcriptomes (which are clinically accessible) have the potential to enhance the clinical stratification of pneumonia cases [71-73].

1.7 Treatment of pneumonia

Early prediction and treatment of mild pneumonia cases at the higher-risk of developing serious clinical outcomes is vital for mitigating the burden of hospital admissions and under-five mortality [27]. As highlighted in the previous section, treatment of pneumonia is often based on the World Health Organization (WHO) clinical algorithm (**Table1.1**). Mainly, treatment options include symptomatic treatment, oral or injectable antibiotics, and supportive therapy such as oxygen supplementation [27, 28, 74]. Special attention has to be given to immune-compromised patients and other risk groups as well as young children [27].

While upper respiratory tract infection (URTI) such as cough or common colds are only treated asymptomatically, lower respiratory tract infections (pneumonia) are usually treated by oral or injectable antibiotics [27]. However, not all mild cases will proceed to severe pneumonia if antibiotic treatment is withheld. Consequently, while rapid antibiotic treatment of bacterial cases is vital to prevent severe outcomes, empirical misclassification of non-bacterial (i.e. viral pneumonia) cases is depleting

antibiotic stocks and fueling the exacerbation of antibiotic resistance [40, 53, 56, 75-77]. Therefore, innovative approaches are required to enhance clinical stratification and treatment modalities for mild pneumonia cases, and derive alternative treatment options for antibiotic-resistant pathogens.

1.8 Host response in pneumonia

Despite being the most frequent infectious cause of child mortality, host responses in childhood pneumonia are not fully understood[31]. Generally, host response involves the interplay between the Innate and Adaptive pathways of the immune system (and recently metabolic pathways), which mainly differ on how they recognize pathogens [78, 79]. In pneumonia, these responses are further divided into local (within the lungs) and systemic (detected in the blood) [80, 81].

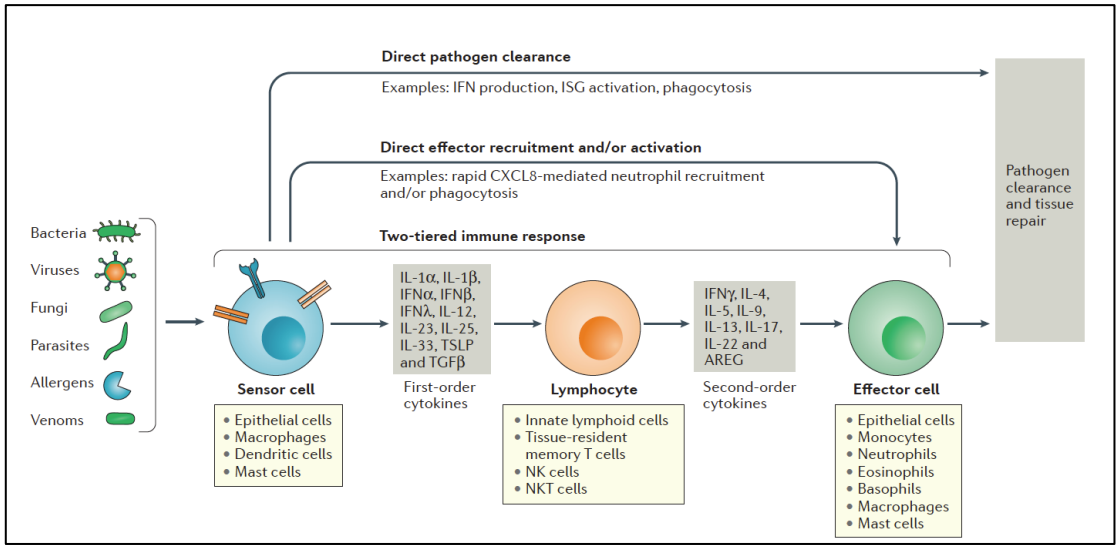


Figure 1.2: Host responses in the lungs following respiratory infections. The figure was accessed from Iwasaki (2016), page 2: (http://dx.doi.org/10.1038/nri.2016.117) [31]

1.8.1 Innate responses

The innate immunity provides the first line of defense (against the invading pathogens) beyond the physical and chemical barriers (i.e. cilia beating, tight junctions and mucus production) provided by the epithelial cells (ciliated,

club, goblet and basal cells) [79]. Mediators of the innate immunity include phagocytes (Neutrophils and Macrophages), antigen presentation cells (i.e. dendritic cells, macrophages), innate lymphoid cells (ILCs), and natural killer cells. While the innate response is rapid, pathogen recognition is less specific relying on pattern recognition receptors (PRRs) to detect pathogen-associated molecular patterns (PAMPs). The recognition of PAMPs by the PRRs initiates the cascade of cellular signaling pathways including

(i) The production of pro-inflammatory cytokines (i.e. TNF, IL-1, IL-6, IL-12), chemokines (i.e. CXCL8) and interferons (IFN) through the activation of transcription factors such AP-1, IRFs, NF-kB [82-84]

(ii) Phagocytosis of pathogens, degranulation (eosinophils, neutrophils and mast cells), and vasodilation of epithelial cells

(iii) Priming of the adaptive immunity through the antigen-presentation cells (i.e. dendritic cells) [39, 79, 85-89] [31].

Briefly, PAMPs are biochemical signatures, which are exclusively expressed in pathogens (not the host) and essential to their survival. They include a major family of biomolecules such as Lipopolysaccharide (LPS), Lipoprotein, Peptidoglycan, Lipoteichioc acids (LTAs) [90]. On the other hand, pathogen recognition receptors (PRRs) are germ-line encoded and evolutionary conserved molecules that are exclusive to the host. The major types of PRRs include (i) Toll-like receptors (TLRs), (ii) RIG-I-like receptors (RLRs) (iii) NOD-like receptors (NLRs), (iv) C-type lectin-like receptors (CLRs) [85, 87, 91-95].

Toll-like receptors (TLRs) are the most studied PRRs, which are located on the outer membrane or the endosome of the host cells (i.e. macrophages). On the outer membrane, TLR2 & TLR1 and TLR2 & TLR6 heterodimers recognize lipoproteins, LTA, PGN, lipoarabinomannan while TLR4 and TLR5 recognize LPS and flagellin respectively. In the endosome, TLR3, TLR7, TLR8 recognize viral or bacterial RNA while TLR9 recognizes viral or bacterial DNA, respectively. Activated TLRs often induce the production of pro-inflammatory cytokines such as TNF, IL-1, IL12, lL8 and lFNγ [85].

On the other hand, NOD (nucleotide-binding oligomerization domain)-like receptors (NLRs) and RIG-I-like recognize pathogens in the cytosol. NLRs comprise at least twenty families including NALP1 and NALP3, which form an inflammasome complex (with Caspase-1 and ASC) that mediates the production of inflammatory cytokines IL-1B and IL18 [87]. Further, the RLRs including RIG-I, MDA5, and LGP2 sense viral RNA and coordinate the production of type-I interferon, and the transcription of antiviral genes to eliminate the intracellular viral infection [91-93].

Other mediators of inflammation and antimicrobial activities are the circulating plasma protein complexes such as the complement system, C-reactive proteins [96] and antimicrobial peptides such as defensins and cathelicidins [97, 98]. In particular, the complement proteins are involved in *opsonophagocytosis*, inflammation (chemo-attraction of phagocytes) and the formation of microbial membrane attack complex (MAC) [99, 100]. The antimicrobial peptides are natural antibiotics, which also induce phagocytosis

and activate adaptive mediators such as CD4+ helper T cells[98]. On the other hand, danger-associated molecular signals (DAMPs) such as the high-mobility group box 1 (HMGB1) protein, reactive oxygen species (ROS) and nitric oxide (NO) promote phagocytic and inflammatory activities especially in the macrophages [31]. To elicit an effective immune response, the adaptive immunity is subsequently involved (next).

1.8.2 Adaptive responses

The adaptive immune response is delayed (days) but capable of recognizing a repertoire of more specific antigens than the innate immunity, and forms immunological memory for robust response upon re-infection with the same pathogen (a hallmark for vaccine development). Its activation is vital for an effective clearance of the pathogen, resolving the inflammation and wound healing. The main cell types of the adaptive immunity are the B and T cells.

In whole blood, the relative proportions of lymphocytes (B and T cells) approximately range between 20% to 40% [101]. Among them, the relative proportions for T and B cells can vary as follows: 61%- 85% and 7%–23%, respectively [102-105]. Functionally, the adaptive immunity is divided into $humoral$ and $cellular$ mediated responses [79, 86, 88, 106-108].

The *humoral-mediated* response is more rapid and mainly involves the production of antibodies (immunoglobulins) by the plasma B cells to neutralize and eliminate extracellular antigens. In particular, naïve B cells are produced and mature in the bone marrow, and circulate into the bloodstream

and other body tissues. Upon activation with an antigen, they differentiate into plasma (antibody producing B cells) and memory B cells. Generally, antibodies are involved in the opsonisation pathogens to (i) neutralize their virulence activities, and (ii) harness their phagocytosis by innate cells (i.e. macrophages) and membrane complex attack (MAC) through the activation of the classical complement pathway [79, 86, 88, 106-108] [79, 109-112].

Antibodies (immunoglobulins) are Y-shaped proteins, which are classified by their heavy chains including IgM (μ-chains), IgG (γ-chains), IgA (α-chains), IgD (δ-chains), and IgE (ε-chains). In particular, the IgG and IgA molecules are the most prevalent antibodies in the serum (systemic responses) and secretions (i.e. mucous, tears, saliva and milk), respectively. Notably, the IgM and IgG antibodies are more abundant for early (i.e. potential marker of acute disease) and long-term (i.e. potential marker for chronic disease) exposures, respectively. Further, the IgD and IgE antibodies are the least prevalent, and do not activate the complement pathway [79, 86, 106].

On the other hand, *cellular-mediated responses* are coordinated by the T-cells (mainly through the production of cytokines) to facilitate the elimination of intracellular antigens. Briefly, the T cells are mainly divided into the cytotoxic CD8 and CD4 T-cell pathways. The ratio of CD4+ T-cells to CD8+ T-cells varies from <1.0 to 2.0 [102-105]. The cytotoxic CD8 T cells recognize pathogen-infected host cells through the major histocompatibility class1 (MHC-I), and induce their apoptosis to contain the infection (i.e. virus) [79, 86, 88, 106-108].

The CD4 T-cells (on the other hand) recognize their antigens via the antigen-MCH-II class complex presented by the professional antigen presentations cells (APCs) especially the dendritic cells. Functionally, the CD4 T-cells are further divided into subclasses each producing a range of cytokines including **Th1** (IL12, IFNγ), **Th2** (IL-4, IL-2), and **Th17** (IL-6, IL-21, IL-23, TGF-β), **Th9** (TGF-β, lL-4), **iTreg** (TGF-β, lL-2), **Tfh** (IL6, IL-21), and **Tr1** (IL27, IL-10) [111]. Manly, these cytokines amplify and regulate the effector functions of other cells including (i) phagocytosis, (ii) antibody production by the plasma B cells and (iii) the cytotoxicity of CD8 T-cells and NK cells [79, 86, 88, 106-108]. Notably, regulatory T cells (Tregs) are vital for controlling excessive inflammation to maintain or restore a homeostatic environment[112]. However, they can be detrimental if activated prematurely or pathogen-modulated [113, 114].

1.8.3 Local and systemic responses

It is worth noting that host responses in pneumonia can be separated at the **local** and **systemic** levels [81, 115, 116]. The local responses are found within the lungs [31] while systemic responses are detectable in the circulating blood [115, 117]. Normally, the local responses are sufficient to clear the pathogen without causing serious clinical outcomes. However, due to host risk factors and pathogenic virulence factors, the local response boundaries are breached consequently inducing the systemic responses, which are often excessive and detrimental to the host and often associated with serious clinical outcomes [116, 118].

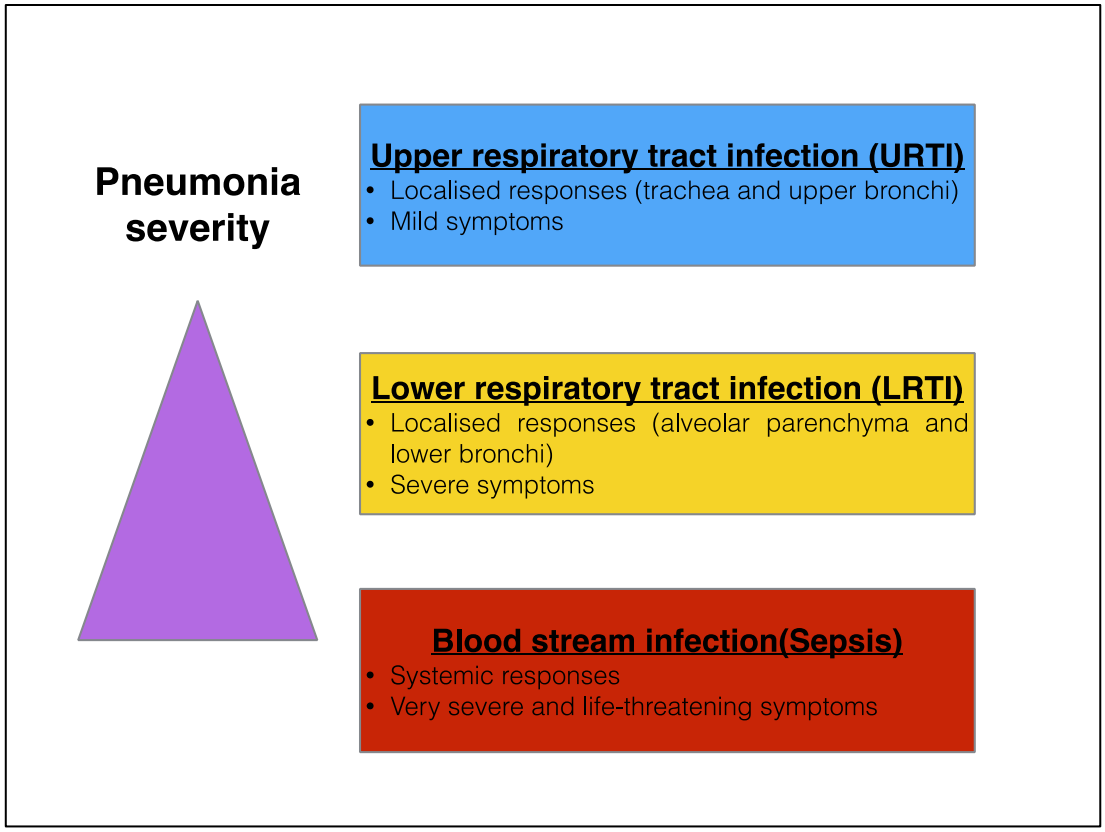


Figure 1.3: An illustration of local and systemic responses associated with pneumonia severity. Severity outcomes increase with loss of control, from the upper respiratory tract (top) to the blood stream (bottom).

Thus, while the *local responses* are vital to detect, contain and eliminate the invading bacteria within the lungs without causing serious clinical pathology, the involvement of the systemic responses potentially underpins the development of severe pneumonia outcomes [37, 116, 119-122]. Therefore, blood-based signatures present an opportunity for case stratification (i.e. biomarkers) and management of severe pneumonia cases. Moreover, changes in the blood reflect key pathophysiological changes for the entire body including the lungs [102]. Importantly, whole blood is a readily accessible tissue in clinical practice. Therefore, whole blood genome-wide profiling approaches (i.e. RNA Seq or microarray based transcriptomics), which have become a mainstay of genomic research and future translation

medicine [123], present a powerful and innovative approach for enhanced clinical stratification of pneumonia cases [65]. In this thesis, we applied whole blood transcriptomics (introduced in the next section) to gain deeper insights into the molecular and cellular pathway responses in severe pneumonia.

1.9 Whole blood transcriptomics

Proteins are the building blocks for cell structure and activities in disease and health. The human cell has about 20,000 protein-coding genes, which are activated at different time points depending on the state of the cell [124]. The genes are found on the DNA, which provides the template of protein-coding. In a human cell, the DNA is packed in twenty-three pairs of chromosomes inside the nucleus. However, proteins are coded indirectly through a transient intermediary molecule called messenger RNA (mRNA) that carries similar information as DNA and the protein [125]:

$$\boldsymbol{DNA} \xrightarrow{Transcription} \boldsymbol{mRNA} \xrightarrow{Translation} \boldsymbol{Protein}$$

While proteomics enables direct investigations of protein structures and physiological functions, protein analysis is more complicated than genomics (gene sequences) and transcriptomics (RNA abundance). Firstly, proteins functions are translated from the 4-nucleotide codes of DNA and mRNA into a much more complex code of 20 amino acids [126], which also depends on the specific structure they fold up into. Further, sample purification is also challenging [125, 127].

Transcriptomics is the study of the complete set of RNA transcripts that are produced within the cell (transcriptome) under certain conditions such as response to infections [123, 125, 128]. Here, the key assumption is that the RNA transcripts reflect the transcribed genes, and hence the state of the cell. In particular, advances in *microarray* technologies [129, 130] and recently *RNA-seq [131]* has spurred transcriptomic research investigating

pathogenesis pathways, biomarkers and therapeutic targets in a range of diseases including cancer [132, 133], infections [121, 134-136] and autoimmunity [137-139].

In pneumonia, several studies have applied whole blood transcriptomics to investigate the systemic responses and derived candidate biomarkers [66, 67, 140, 141]. However, such studies are rarely available in Sub-Saharan Africa, where the pneumonia burden is high. Here, this study applied the whole blood transcriptomics approach to gain a deeper insight into the systemic pathway responses that are associated with the severity of childhood pneumonia in resource-limited settings, focusing on the Gambian children and infants and using the microarray technology (next section).

1.9.1 The Microarray approach

Despite recent advances in RNA-seq, the microarray approach remains the method of choice for transcriptomic studies because it is cost-effective and more established[142] Briefly, this technology applies hybridization [129] to simultaneously assess the abundance of tens of thousands of RNA transcripts in multiple samples. For each sample, a raw data point called *CELFILE*, which contains expression signals for all the gene probes on the array, is generated [125, 129, 143].

1.9.2 Data analysis

The analyses of microarray data involve the application of various statistical and bioinformatics approaches to address the following objectives[144]:

(i) *Class discovery*: Identification of novel clusters in the data (unsupervised learning)

(ii) *Class comparison*: Identify candidate genes (i.e. differentially expressed genes) and signaling pathways (supervised analysis)

(iii) *Class prediction:* identification of candidate biomarkers

However, the raw data points are potentially confounded by non-biological variations [129, 145]. Therefore, data quality assurance is central to the main analysis process. To remove technical variations across the array, the raw database is subjected to statistical pre-processing algorithms for background correction, normalization, transformation, and summarization of probe-level data into the probe set [146, 147]. For studies with multiple sample batches, batch-effect normalization is also required in addition to raw data pre-processing [148-150]. Further, potential outliers are investigated before and after data pre-processing [151, 152]. Since not all transcripts are relevant to a particular disease, non-informative gene probes are often eliminated to minimize potential noise and false discoveries due to multiple testing [153]. To enable the contextualization of results, the gene probes are annotated to the universal gene IDs such as ENTREZIDs [71, 143].

1.10 Study aims and hypothesis

1.10.1 Rationale

Pneumonia remains the leading infectious cause of under-five mortality especially in resource-limited settings including the Gambia [2, 4]. Therefore, innovative approaches are required to gain deeper insights into the pathogenesis of pneumonia that could facilitate the discovery of cost-effective vaccines as well as diagnostic and prognostic tools applicable for resource-constrained settings where the burden is highest [154].

Here, this thesis has applied a range of data science approaches using a genome-wide whole blood transcriptome to gain a deeper insight into the systemic pathway responses associated with the clinical pneumonia severity among the Gambian children and infants. Importantly, whole blood is a readily accessible clinical tissue and whole blood transcriptomics has become the mainstay of genomic research and future translation medicine [71, 102]. Further, the methodology and the cost of molecular profiling are improving [155]. Therefore, this approach presents an innovative, clinically accessible and powerful resource for elucidating the pathway biology, and enhancing the clinical practice of severe pneumonia.

1.10.2 The central hypothesis

This study has investigated the following central hypothesis:

- ***Systemic pathway responses underpin the development of severe pneumonia outcomes*.**

This hypothesis implies that while the local responses (compartmentalised in the lungs) are the priority, systemic responses are crucially involved in severe pneumonia outcomes. To address this hypothesis, the following objectives were pursued.

1.10.3 Study objectives

The aim of this study is to gain a deeper insight into the systemic pathway responses associated with pneumonia severity. In particular, the following specific objectives were addressed:

1) To investigate the cellular pathway responses associated with pneumonia severity
2) To investigate the molecular pathway responses associated with pneumonia severity
3) To identify candidate biomarkers for early detection of mild pneumonia cases at the higher risk of developing severe pneumonia outcomes.

1.11 Thesis outline

This thesis has seven chapters. The study methodology is described in **Chapter 2**, and extended into **Chpater 3** where data curation, characterization and quality assurance findings are presented. To investigate the cellular pathways, **Chapter 4** applied computational deconvolution analysis approaches to assess the cellularity of whole blood in pneumonia severity. To investigate the molecular pathway responses to severe pneumonia, **Chapter 5** applied a computational pathway analysis approach

using a range of biochemical pathway databases. To address the final objective, in **Chapter 6** I coupled cellular pathway biology with machine-learning approaches to investigate the candidate biomarkers for severe pneumonia. Finally, **Chapter 7** presents an overall summary and discussion.

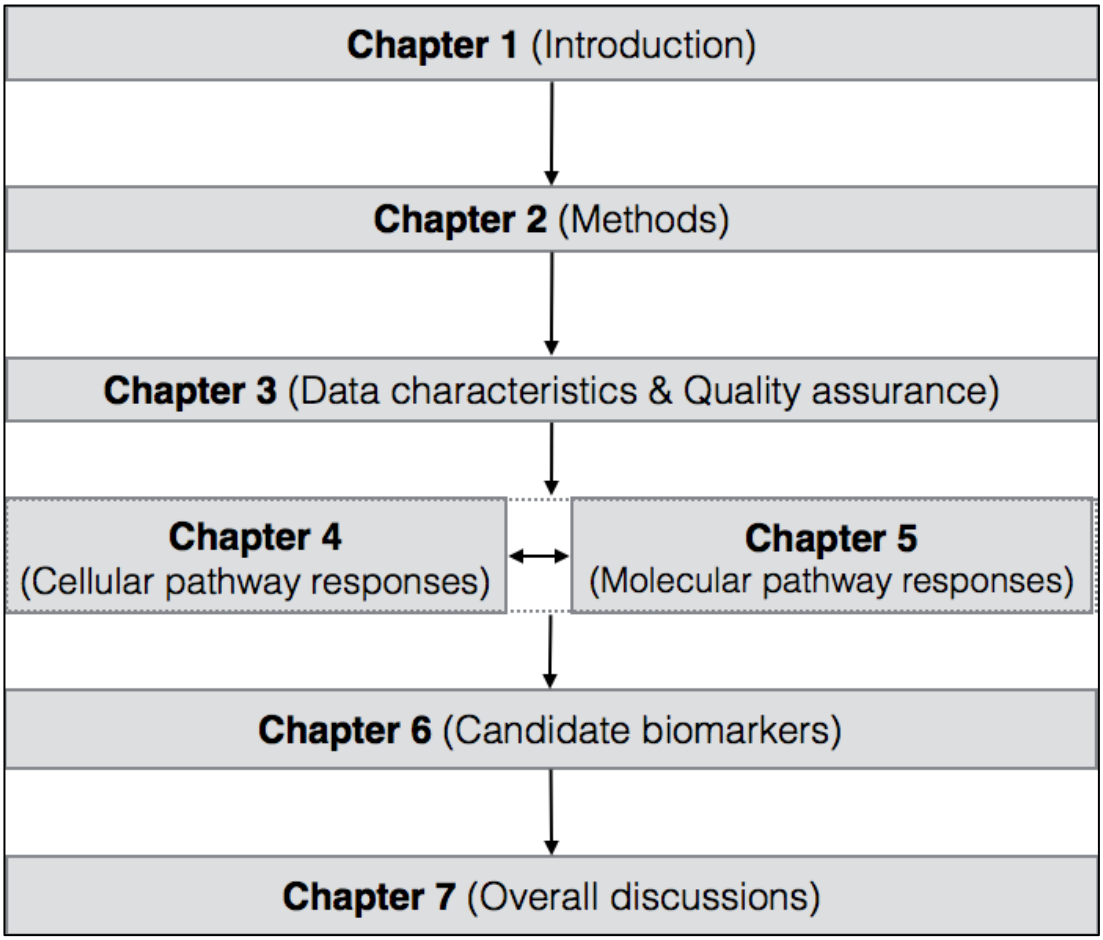


Figure 1.4: Systematic diagram of thesis chapters.

Chapter 2: Material and methods

2.1 Introduction

This chapter introduces the study materials and methodology including the study design, the central data resources and statistical analyses.

2.2 Materials

This thesis mainly analyzed a whole blood transcriptome and the corresponding metadata records (clinical, demographic and microbial databases), which were complemented by a range of publicly available data resources including gene expression data from the Gene Expression Omnibus (GEO) at the National Center for Biotechnology Information (NCBI) [156], and biochemical pathway databases such as KEGG, REACTOME and Gene Ontology (GO) [157]. Computationally, most of the analyses were conducted in R statistical programming language [158] using a range of packages especially from the Bioconductor repository [159]. Further, the CellMix package [160, 161], which is not on the Bioconductor, was also extensively applied. Metadata cleaning and descriptive analysis tables were conducted in Stata 12 (StataCorp, Texas) [162].

2.3 Study setting

In this study, participants (children aged 2-59 months) were recruited from The Gambia in West Africa. The Gambia is a small (area=11,295 km^2) and resource-limited country, which is mostly surrounded by Senegal. Eligible participants were recruited from the coastal semi-urban in the Greater Banjul area (training dataset) or the rural Basse area (Validation data set) between June 2007 and September 2010 (**Figure2.1**) [15].

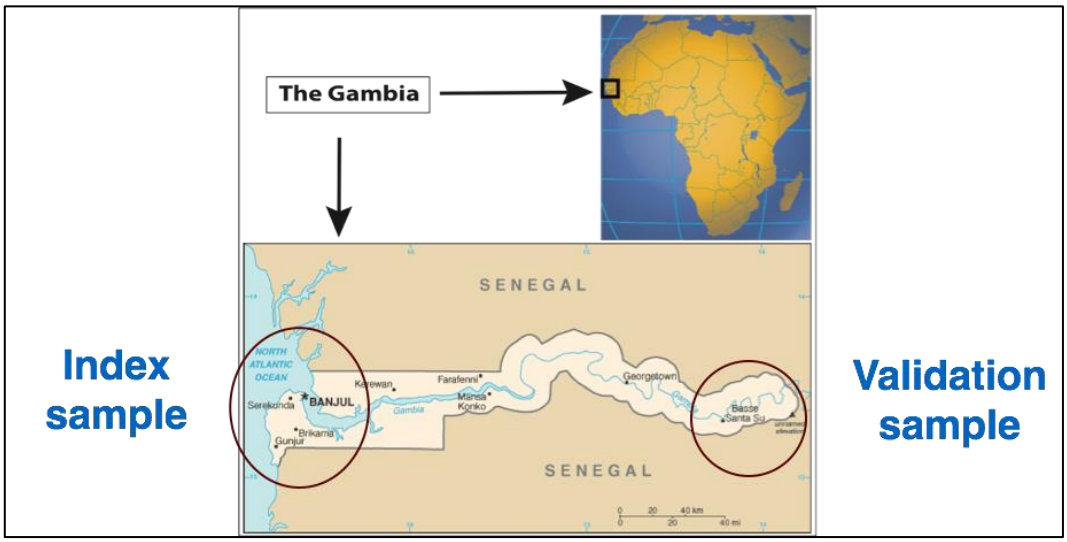


Figure 2.1: Map of Gambia showing major towns and study sites.
(http://www.accessgambia.com/information/map.html)

2.4 Study population

The Gambia has an estimated population of 2 million people (density=176.3 people/square km) (https://data.unicef.org/country/gmb). Notably, the proportion of under-five children is the highest in this population (**Figure2.2**).

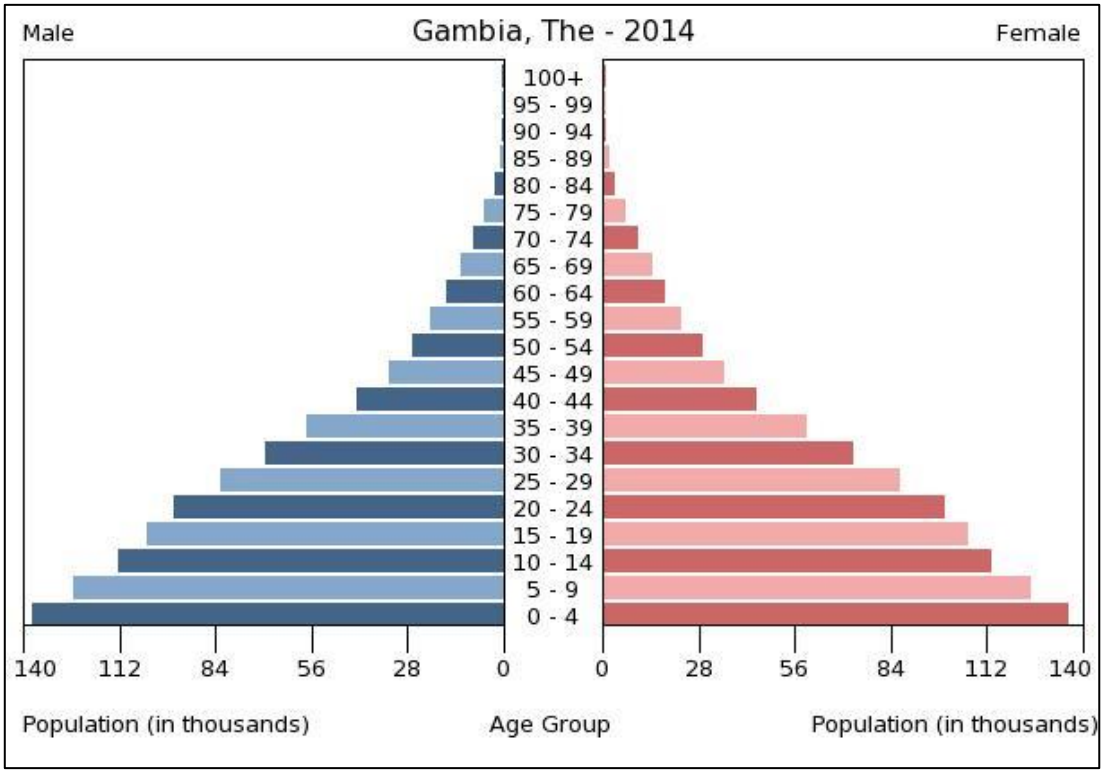


Figure 2.2: The population structure of the Gambia.
(http://www.allcountries.org/world_fact_book_2016/gambia_the/gambia_the_people.html)

Life expectancy is estimated at 58 years for men and 69 years for women but HIV (Human Immunodeficiency Virus) prevalence is very low (less than 2%). The under-five mortality rate is estimated at 69 per 1000 live births [163], which is above the United Nations (UN)'s 2015 Millennium Development Goal number 4 (MDG4). In 2010, malaria (20%), pneumonia (15%) and prematurity (14%) were the leading causes of under-five mortality (http://www.commonwealthhealth.org). Aetiologically, *Streptococcus pneumonia* is the leading cause of pneumonia in this population [41]. At the time of blood sample collection, the coverage for conjugate *Haemophilus influenzae type b* (Hib) vaccine was high but there was no routine usage of pneumococcal conjugate vaccine (PCV), which is the case currently in many African countries[41].

2.5 Ethics approval

Written informed consents for participation in the study were obtained from the parents or legal guardians of all participants. The study was approved by the Gambian Government-Medical Research Council Joint Ethics Committee in Banjul (SCC/EC1062).

2.6 Microarray experiment

Briefly, a typical microarray experiment involves the following steps:

1) Study design and sample collection
2) RNA sample preparation,
3) Reverse transcribe and label the mRNA,
4) Hybridization of the labeled target to the microarray plate

5) Scan the microarray to quantify the expression signals
6) Image analysis
7) Statistical and bioinformatics analysis

2.7 Study design and recruitment of participants

Study participants were recruited using a prospectively matched case-control study design. The eligible cases were children and infants (aged 2-59 months) who were clinically classified as mild, severe or very severe pneumonia at a designated study clinic. To define pneumonia severity levels, the modified World Health Organisation (WHO) criteria (**Table2.1**) were applied. It is worth noting that oxygen saturation was measured in all the participants.

Pneumonia criteria		**Clinical phenotype**
Age, sex, season and community matched non-pneumonia children aged 2-59 months old		Controls
Children aged 2-59 months with cough OR difficulty in breathing	≥50 breaths/min in children 2-11 months old ≥40 breaths in children 1-5 years old	Mild
	Lower chest in-drawing / head nodding / Nasal flaring / Grunting	Severe
	Oxygen saturation<90%	Very severe

Table 2.1: Classification of pneumonia cases into severity groups. This criteria was adapted from the World health organization (WHO)'s guidelines for case management at health facility [1]. In particular, this criterion includes oxygen saturation data.

In the Greater Banjul region, the pneumonia cases were recruited at the Medical Research Council (MRC) hospital in Fajara, the Royal Victoria Teaching Hospital (RVT), or the major health centres at Fajikunda, Serekunda or Brikama. Together, these samples formed the training database (here called Fajara) and were applied for all primary analyses. In

the validation sample (from the rural Basse area), all the pneumonia cases were recruited at the Basse health centre. In all the sites, children with a cough of ≥2 weeks, or severe anemia (hemoglobin level <6 g/dL) or confirmed wheeze were excluded.

To account for potential confounding, severe and non-severe pneumonia cases were frequency-matched by age, sex, location and season. Further, all the pneumonia cases were also matched to non-pneumonia community controls. In particular, community controls were selected from a compound located at least 50 paces (in a randomly selected direction) from the compound with a pneumonia case [15]. It is worth noting that, the validation dataset was kept independent from all the primary analyses in order to validate the performance of candidate biomarkers in **Chapter 6**.

2.8 Collection and processing of whole blood samples

All the eligible cases and controls (n=1527) were bled on the day of recruitment. Whole blood samples were collected for blood culture/PCR based bacterial detection and into PAXgene tubes (https://www.preanalytix.com) for RNA isolation. Additional lung aspirate samples were taken from some individuals for bacterial culture. Sufficient whole blood RNA samples (n=803) were extracted and prepared at Medical Research Council Unit, The Gambia (MRCG) laboratories and shipped for transcriptomics analyses to Peter Ghazal's group at the Division of Infection and Pathway Medicine (DIPM), University of Edinburgh Medical School, United Kingdom.

2.9 RNA quality control and microarray experiment

In Edinburgh, whole blood RNA samples were subjected to further quality control analysis to assess sample volume (sufficient quantity) and purity using the Nanodrop ND1000 spectrophotometer (https://www.thermofisher.com), and RNA integrity using the RNA 6000 Nano Chip run on an Agilent Bioanalyser 02936A (Agilent Technologies: http://www.agilent.com). Eligible RNA samples (n=518) were hybridized and sent for genome-wide microarray profiling using the Affymetrix HGU219 platform at AROS (http://arosab.com).

It is worth noting that the microarray profiling was conducted in two batches: Batch1 (n=447) in 2013 and Batch2 (n=71) in November 2014. The second Batch was added to minimise the demographic data imbalances (potentially confounding variations) between the study groups, and improve the sample size of very severe pneumonia (i.e. after sample size re-assessment. The final database was reduced to 503 samples (**Table2.2**) after removing outliers (described in section 2.1.4 and 3.3.1.1).

Site	Controls	Mild	Severe	Very severe	Total
(a) Number of recruited participants					
Total	714	321	443	49	1527
Fajara	402	175	232	24	833
Basse	312	146	211	25	694
(b) Whole blood microarray experiment samples					
Total	175	137	162	44	518
Fajara	128	91	121	20	360
Basse	47	46	41	24	158
(i) Whole blood samples batch1					
Total	144	124	153	26	447
Fajara	116	90	118	16	340
Basse	28	34	35	10	107
(ii) Whole blood samples batch2					
Total	31	13	9	18	71
Fajara	12	1	3	4	20
Basse	19	12	6	14	51
(c) Final analyses data (After data cleaning)					
Total	167	136	158	42	503
Fajara	120	90	117	18	345
Basse	47	46	41	24	158
(i) Whole blood samples batch1					
Total	139	123	149	24	435
Basse	28	34	35	10	107
Fajara	111	89	114	14	328
(ii) Whole blood samples batch2					
Total	28	13	9	18	68
Basse	19	12	6	14	51
Fajara	9	1	3	4	17

Table 2.2: The distribution of eligible samples. The table shows the number of samples at recruitment (a), microarray profiling (b) and final analyses (c). Pneumonia cases were matched to non-pneumonia community controls by age, sex and location.

2.9.1 The Affymetrix HGU219 microarray platform

To investigate the gene expression profiles associated with pneumonia severity, this thesis applied the Affymetrix GeneChip technology particularly using the HGU219 microarray platform. Generally, Affymetrix GeneChips apply 16-25 pairs of oligonucleotide probes (collectively called a probe set) to investigate a gene. To improve robustness, multiple probe sets are often

applied to investigate a single gene. Each probe pair comprises a perfect match (PM) and a mismatch (MM) probe to assess sensitivity and specificity, respectively. In particular, the MM probes have a single sequence mismatch at the middle (i.e. 13th position) to estimate background noise due to non-specific binding [147]

The Affymetrix HGU219 array is a *single-channel* (*one-color)* microarray platform, which provides intensity data for each probe or probe set indicating a relative level of hybridization with the labeled target. For each sample, the intensity data represent relative RNA abundance when compared to other samples or conditions processed in the same experiment. This array platform was designed using sequences selected from the UniGene database 219 (build date March 30, 2009), RefSeq version 36 (13 July 2009) and full-length human mRNA's from GenBank® (downloaded May 12, 2009). For each RNA sample, the HGU219 platform stores the raw data for the analysis into a CELFILE, which contains expression signals for 54613 gene-probes representing 47,000 transcripts and their variants [164-166]. However, it is should be noted that the HGU219 array platform was designed without the mismatch (MM) probe data. (https://www.thermofisher.com/order/catalog/product/901595).

2.10 Sample size re-assessment

While the blood samples were collected from the Gambian population, the original study design applied the variability estimates from a neonatal study that was conducted at the Royal infirmary of Edinburgh in the United Kingdom[167]. In this thesis (**Chapter 3**), another sample size analysis was conducted to re-assess the statistical power of the study groups. In the

current sample size analysis, variability estimates were estimated from a whole blood transcriptome of Gambian children (GSE20436) who participated in the Trachoma study as healthy controls [168].

Mainly, this analysis estimated the number of samples that were statistically powered (90%) to detect at least a two–fold change in gene expression at 5% false discovery rate (FDR) in at least 90% of the genome.

Therefore, assuming constant variance between the groups, the following input parameters were used:

- Effect size =log2(Fold change)=1
- Type I error (α)=5% or 1%,
- Type II error (β)=10% (i.e. power=90%)
- Variability: A vector of between-sample (n=20) standard deviation (SD) values estimated from each gene probe using the GSE20436 data.

In this analysis, I applied the ssize function that is implemented in the *ssize R* package [169]. To account for multiple testing, this function applies the Bonferroni multiple testing correction method [170], which is very stringent. In this method, the multiple-testing-adjusted P-value is defined as $\mathbf{q_i = M * p_i}$; where $\mathbf{p_i}$ is the corresponding raw P-value and **M** represents the total number hypotheses tested. Here, a less stringent approach called rough false discovery rate (RFDR) was applied such that $\mathbf{q_i = p_i * 2M/(M+1)}$.

2.11 Data cleaning (metadata records)

To ensure data quality and completeness, the demographic, clinical and microbial databases were subjected to an intensive data cleaning to identify relevant variables, suspicious values, and missing data. The data queries were resolved using the hard copy reference database (SCC1062) secured in the archive department at the MRC Unit, in the Gambia.

2.12 Quality assurance of the microarray database

To ensure data quality, the microarray databases (training and validation sets, respectively) were subjected to raw data pre-processing, batch-effect correction, outlier detection and gender analysis (**Chapter 3**).

2.12.1 Pre-processing of the raw data

2.12.1.1 Introduction

Microarray gene expression [130] analysis seeks to investigate meaningful biological variations in the abundance of mRNA transcripts that are associated with different phenotypes of interest such as infections and disease severity [71, 168, 171]. However, these interesting variations are often obscured by unwanted non-biological variations within and between the arrays [130, 145, 147]. Therefore, appropriate data pre-processing [145, 172] is mandatory to effectively

(i) Remove background noise due to non-specific binding and spatial heterogeneity [173]
(ii) Normalize technical variations between the arrays due to sample handling (i.e. hybridization) [174-178]
(iii) Stabilize the variance (i.e. data transformation) and
(iv) Summarize probe-specific signals into a probe set (gene) level data [145, 146, 179, 180].

However, it should be noted that the HGU219 GeneChip is optimised without the mismatch (MM) probes (as explained in section 2.9.1). Consequently, raw data pre-processing algorithms [181] that require mismatch probe data such as Average difference [182], Li Wong [183], MAS.5 [181], PLIER [181] and GCRMA [184] were not applicable. To identify an appropriate pre-processing algorithm, here we assessed the performance of RMA [147] and VSN [146] algorithms, which do not require the MM data (more details below).

2.12.1.2 Robust Multi-Array Average (RMA)

The RMA algorithm was developed by Irizarry et. al (2003) [147], and is implemented in R Bioconductor $affy$ package using the rma function [185]. Usually, pre-processing algorithms subtract MM data from PM to adjust for background noise, which often generates negative values (if MM>PM). This observation suggests that the MM data capture more than background noise, and potentially introduces bias when adjusted for background correction. Therefore, the RMA method completely ignores the MM data and the following step are applied on the PM data: (i) background correction, (ii) quantile normalization [186], (iii) log2 transformations, and (iv) summarization of probe level data into a probe-set (i.e. gene) level data using robust $median\ polish$ [187].

2.12.1.2.1 RMA background correction

Formerly, RMA background correction involves the deconvolution of the observed perfect match (PM) signal for array i, probe j on probe set (i.e.

gene) k into background (BG) noise and real signal (S) components [147, 188].

$$PM_{ijk} = BG_{ijk} + S_{ijk}$$

$$where\ BG_{ijk} \sim N(\mu_i, \sigma_i^2), and\ S_{ijk} \sim Exp(\lambda_{ijk})$$

To get the background corrected signals, a transformation $B(.)$ is applied such that

$$B(PM_{ijk}) = E(S_{ijk}|PM_{ijk})$$

$$= PM_{ijk} - \mu - \lambda\sigma^2 + \frac{\varphi((PM_{ijk} - \mu_i - \lambda_{ijk}\sigma_i^2)/\sigma_i)) - \varphi((PM_{ijk} + \lambda_{ijk}\sigma_i^2)/\sigma_i))}{\phi((PM_{ijk} - \mu_i - \lambda_{ijk}\sigma_i^2)/\sigma_i)) - \phi(PM_{ijk} + \lambda_{ijk}\sigma_i^2)/\sigma_i)) - 1}$$

Where $\phi(.)\ and\ \varphi(.)$ represent the Gaussian (N(0,1)) cumulative distribution and probability density functions, respectively. The parameters ($\mu, \sigma\ and\ \lambda$) are estimated separately within each array using the observed PM data. Finally, $E(S_{ijk}|PM_{ijk})$ is the background corrected PM data [147, 185].

2.12.1.2.2 Quantile normalization

To remove unwanted technical variations between the arrays, the RMA algorithm applies $quantile\ normalization$ on the background corrected data ($E(S_{ijk}|PM_{ijk})$) [172, 173]. Briefly, quantile normalization seeks to project the array-specific n vectors onto a diagonal of an n-dimensional quantile plot (where n=number of arrays). In particular, quantile normalization replaces the $r^{th} - ranked$ value in each array by the rank-specific arithmetic mean calculated from all the $r^{th} - ranked$ values across the arrays. For example, consider 15, 12, 10, 11 and 16 as the highest ranked values in each array (n=5). Then, each value will be replaced by the following arithmetic mean: (15+12+10+ 11+16)/5=64/5=12.8 after normalization. The same applies to other ranks. While this approach is robust and non-parametric, it makes

arrays similar by forcing extreme values to be identical potentially generating force negative results in differential expression analyses (http://bmbolstad.com/stuff/qnorm.pdf) [189].

2.12.1.2.3 Data transformation (log2)

To stabilize the variance across the arrays, improve linearity and minimise the effects of outliers, the RMA algorithm applies the log2 transformation on the background-corrected and quantile normalized data [147].

2.12.1.2.4 Summarization of probe level data

This is the final step of the RMA algorithm where probe level data are summarized into a probe set (gene). The goal is to estimate the true $log2(expression\ level)$ for gene (probe set) k on array i (μ_{ik}) using the following model:

$$Y_{ijk} = \mu_{ik} + \alpha_{jk} + \epsilon_{ijk}$$

Where Y_{ijk} represent the observed background-adjusted, quantile-normalized, and log-transformed PM intensity for array i, on probe j for probe set (gene) k, and $\alpha_{jk} =$ the affinity effect of probe j on gene (probe set) k such that $\sum_{j=1}^{J} \alpha_{jk}$=0.

To estimate μ_{ik}, the robust $median\ polish$ algorithm [187, 188] is applied within each gene (probe set) k [147]. Let $\boldsymbol{A_{0k}}$ be the observed $n \times J$ $(n\ arrays\ and\ J\ probes)$ matrix for the background-corrected, quantile-normalised and log2-transformed PM intensities for gene (probe set) k across all the arrays. Then, the $median\ polish\ algorithm$ alternatively subtracts the column and row medians from the column and row vectors (respectively) of the matrix $\boldsymbol{A_{0k}}$, until the column and row medians of $\boldsymbol{A_{0k}}$

converge to zero vectors. Further, let $\boldsymbol{A_{1k}}$ be the resulting matrix of residuals after the $median\ polish$ operation. Then, μ_{ik} values are estimated by the row means (i.e. within array mean across the probes) of the following matrix $\boldsymbol{D_k} = (\boldsymbol{A_{0k}} - \boldsymbol{A_{1k}})$. Thus, $\hat{\mu}_{ik}$ are called robust multi-array average (RMA) estimates for gene (probe set) k sample i [147, 188].

2.12.1.3 Variance stabilization normalization (VSN)

The VSN pre-processing algorithm was developed by Huber et. al (2002) [146]. For Affymetrix GeneChip data, VSN involves (i) background correction (ii) between-array normalisation and (iii) variance stabilization transformation. To summarise the probe-level data into a gene (probe set), the VSN authors recommend using the $median\ polish$ as explained in the RMA method [190]. The VSN pre-processing is implemented using the $justvsn$ function in the R Bioconductor package called vsn [190]. In this thesis, we combined the VSN algorithm with $median\ polish$ summarisation using the $vsnrma$ function in the $affy$ package [147, 185].

Briefly, VSN is a model-based algorithm that works in two steps: (i) an affine data transformation (cantering and scaling) to calibrate systematic technical experimental factors including sample handling variations and background noise (ii) a general log2 ($glog_2$) transformation to stabilise the dependency of variance on mean [190]. To calibrate the data (step1), the following model is applied:

$$y_{ki}^{*} = \lambda_{si} y_{ki} + \alpha_{si}$$

Where y_{ki} is the observed raw value for probe k in array i, y_{ki}^{*} is the calibrated value for y_{ki} through the scaling and shifting parameters λ_{si} and

α_{si}, respectively. The s index denotes the strata of the probes within each array [191, 192]. However, in this analysis we that assumed all the probes on each array were subjected to the same systematic effects (one stratum), and therefore applied an array-wide calibration using the following reduced model (without the s index):

$$y_{ki}^{*} = \lambda_i y_{ki} + \alpha_i \qquad (1)$$

To stabilize the variance of the calibrated signals, the following inverse hyperbolic transformation (h) is applied:

$$h_{ki} = arsinh(\lambda_0 y^{*} + \alpha_0) \qquad (2)$$

$$= \log\left(\lambda_0 y^{*} + \alpha_0 + \sqrt{(\lambda_0 y^{*} + \alpha_0)^2 + 1}\right)$$

Combining the two steps (1) and (2)), VSN formerly seeks to solve the following equation [190, 191]:

$$h_{ki} = arsinh(e^{b_i} * y_{ik} + a_i)$$

Where $a_i = \alpha_i + \lambda_0 \alpha_i$ and $b_i = \log(\lambda_0 \lambda_i)$ are the combined calibration and transformation parameter for features from array i, which are estimated using maximum likelihood and a robust procedure similar to least *trimmed sum of squares regression* [191, 192].

2.12.1.3.1 The choice of the transformation function

Let y_{ki} be the observed perfect match (PM) intensity value for probe k on array i, which is deconvoluted into a noise parameter α_{ki} and the true expression signal x_{ki} multiplied by the proportionality factor β_{ki} as shown below:

$$y_{ki} = \alpha_{ki} + \beta_{ki} x_{ki} \qquad (1)$$

$$\Rightarrow \frac{(Y_{ki} - a_i)}{\beta_i} = m_{ki} e^{\eta_{ki}} + v_{ki} \qquad (2)$$

Where $\eta_{ki} \sim N(0, \sigma_\eta^2)$ and $v_{ki} \sim N(0, \sigma_v^2)$; $\beta_{ki} = \beta_i \gamma_k e^{\eta_{ki}}$; $\alpha_{ki} = a_i + \bar{v}$; $m_{ki} = \gamma_k x_{ki}$ and $v_k = \bar{v}/\beta_i$. Using (2), it can be shown that the variance of Y_{ki} has a quadratic relationship with its mean (3), where $c^2 = Var(e^\eta)E^2(e^\eta)$.

$$\Rightarrow Var(Y_{ki}) = c^2(E(Y_{ki}) - a_i)^2 + \beta_i^2 \sigma_v^2 \qquad (3)$$

Therefore, VSN seeks to find a transformation $\boldsymbol{h}(.)$ that keeps $Var(Y_{ki})$ constant (i.e SD not dependent on the mean) using the intuition of the $Delta\ method$ approximation of sample variance (i.e. based on Taylor series expansion). For a family of random variables Y_u with E(Y_u)=u and Var(Y_u)=v(u) and a differentiable function h defined on the range of Y_u, the Taylor series approximation of the transformation $h(Y_u)$ is:

$$h(Y_u) \approx h(u) + h'(u)Y_u$$

$$\Rightarrow Var\big(h(Y_u)\big) \approx h'(u)^2 v(u)$$

Intuitively, integrating $h'(u) = v^{-\frac{1}{2}}(u)$ would generate a variance-stabilizing transformation $h(y) = \int \frac{1}{\sqrt{v(u)}} du$. Notably, $arsinh(x) = \int \frac{1}{\sqrt{x^2+1}}$. Therefore, rearranging $v(u) = Var(Y_{ki})$ as expressed in equation (3) in the form of $(x^2 + 1)$, approximately yields:

$$h(y_{ki}) = arcsinh\left(\frac{Y_{ki} - a_i}{b_i}\right) = \mu_{ki} + \varepsilon_{ki}$$

Where $b_i = \beta_i \sigma_v / c$; $\mu_{ki} = E(\text{arcsinh}(\frac{c}{\sigma_v}(m_{ki}e^{\eta_{ki}} + v_{ki})) \approx E(\text{arcsinh}(\frac{c}{\sigma_v} m_{ki})$ is the transformed true abundance of probe k on array i; and $\varepsilon_{ki} \sim N(0, c^2)$ [191]. To generate probe set (i.e. gene) level data, the $median\ polish\ summarisation$ was applied to on the VSN-transformed probe-level data (i.e. $\hat{\mu}_{ki}$ values) [191]

2.12.1.4 Comparison of MRA and VSNRMA

To compare the performance (i.e. stability) of the two algorithms (RMA and VSNRMA), three graphical criteria were applied:

(i) ***Standard deviation (SD) against rank of mean*:** This criterion assessed the stability of signal-to-noise ratio using the plot of standard deviation (SD) against the mean. In particular, the *meanSdPlot* function implemented in the *vsn* package [146] was applied to assess the distribution of $\boldsymbol{SD}(\boldsymbol{X_k})$ against $\boldsymbol{RANK}(\overline{\boldsymbol{X}}_{\boldsymbol{k}})$**;** where $\boldsymbol{X_k}$the gene expression vector for gene probe k and $\overline{\boldsymbol{X}}_{\boldsymbol{k}}$ is the mean value across the samples. Ideally, a constant distribution (horizontal line) implied good performance. Otherwise, false negative or positive discoveries would be expected if the mean and SD values are positively or negatively related, respectively.

(ii) ***Correlation (r_{XY}) against standard deviation (SD):*** This criterion assessed the correlations of randomly selected pairs (5000 pairs) of gene probes (i.e. from independent signaling pathways). Ideally, no significant correlation is expected between each random pair of gene probes **X** and **Y** (i.e. r=0). Therefore, a good algorithm should have a stable relationship (horizontal line, y=0) when the correlations values ($\mathbf{r_{XY}}$) were plotted against the mean standard deviation (**mean*(SD(X), SD(Y))***. Here, two functions *CorrSample* and *plot.corr.sample that are* implemented in the *maCorrPlot R* package were applied [193] to estimate and plot the values, respectively.

(iii) ***Distribution of absolute rank deviation (ARD)****[175]:* This criterion compares the stability (between-sample standard deviations) of gene expression values sharing the same rank across the gene probes. Ideally, low ARD values indicate better performance. Here, a density plot was applied to compare the distributions of the ARD vectors associated with the RMA and VSNRMA methods.

2.12.2 Batch effect variations

As described above (**Table2.2**), the RNA samples were profiled in two batches (n=447 in 2013 and n=71 in 2014). However, raw data pre-processing methods are not effective against batch-effect variations [71, 149]. To investigate the presence of batch-effect variations, principal component analysis (PCA) visualization was applied on the pre-processed transcriptomes. In particular, this analysis was restricted to the gene expression profiles for the negative control probes (i.e. Nonspecific probes), which are designed to remain constant under different biological conditions. To get the principal component (PC) scores, the *prcomp* function in the stats R package[194] was applied. To visualize the batch-effect sample clusters, PC2 (y-axis) was plotted against PC1 (x-axis)

To eliminate the unwanted batch-effect variations, an empirical Bayes normalization called **ComBat**, which is implemented in the *sva* R Bioconductor package *[195],* was applied. Briefly, **ComBat** solves the following location/scale (L/S) equation:

$$y_{ijg} = \alpha_g + X\beta_g + \gamma_{ig} + \delta_{ig}\varepsilon_{ijg};$$ where:

- y_{ijg} =the expression value for a gene ***g*** in sample ***j*** from batch ***i***
- α_g is the overall gene expression,
- ***X*** =a design matrix for sample conditions (other confounders), and
- β_g= a vector of regression coefficients corresponding to ***X***.
- ε_{ijg} = an error term assumed to follow a Normal distribution N(0, σ^2).
- γ_{ig} and δ_{ig} represent the additive and multiplicative batch effects for batch ***i*** on gene ***g***, respectively.

In particular, Combat estimates the L/S model parameters by pooling information across genes in each batch to shrink the batch-effect parameter estimates towards the overall mean of the batch-effect estimates (across genes) in three steps [196]:

Step 1: Data standardization using least square estimates

Step 2: Empirical estimation of batch effect parameters using the standardized expression data matrix (step1) and parametric priors

Step 3: Batch effects adjustment using the Empirical Bayes (EB) parameter estimates (step2) [150, 195, 196].

2.12.3 Detection of outlier samples in the microarray databases

Outliers cause deleterious effects in statistical analyses including microarray gene expression data. To identify the potential outliers, the *ArrayQualityMetrics* Bioconductor package [151] was applied before and after data pre-processing. For the raw database, six criteria were applied and any sample that was flagged by at least four criteria was eliminated as an

outlier. Of them, three criteria were also applicable to the pre-processed databases. Here, samples that were flagged by at least two of the three criteria methods were also eliminated. The six methods are described in **Table2.3** below.

Database applicability	Method	Criteria
Raw database only	**Hoeffding's statistic D_a on the joint distribution of A and M values for each array**; where; M = log2(L_1)-log2(L_2), A = 1/2 (log2(L_1)+log2(L_2)), L_1 is the intensity of the array studied, and L_2 is the intensity of a "pseudo"-array that consists of the median across arrays	D_a>0.15
	Distance between arrays Metric= S_a : Sum of L_1 distances (D_{ab}) between arrays ($S_a = \Sigma_b D_{ab}$); Where $D_{ab=}$ mean ($\|M_{ai} - M_{bi}\|$) and M_{ai} is the value of the *i*-th probe on the *a*-th array.	Boxplot of **S_a**
	Relative distribution of intensity values (M_{ai}) Metric= K_a : Kolmogorov-Smirnov statistic between each array's distribution and the distribution of the pooled data	Boxplot of **K_a**
Raw and pre-processed databases	**Relative Log Expression (RLE)** Metric= R_a: Kolmogorov-Smirnov statistic R_a between each array's RLE values and the pooled, overall distribution of RLE values	Boxplot of **R_a**
	Normalized Unscaled Standard Error (NUSE) Metric= N_a: 75% quantile of each array's NUSE values	Boxplot of **N_a**
	Spatial distribution of M values. Metric=F_a :The sum of the absolutes value of low frequency Fourier coefficients	Boxplot of **F_a**

Table 2.3: Description of outlier detection methods for the microarray database. The table shows the outlier detection methods that were implemented using the arrayQualityMetrics R Bioconductor package [151]. **Box plot**: For each metric (i.e. S_a), samples that lie beyond the extremes of the whiskers of a boxplot (p25-1.5*IQR; p75+1.5*IQR) were considered as potential outliers: p25=25th percentile, p75=75th percentile, and IQR=interquartile range.

2.12.4 Gender analysis

The Y chromosome-specific genes present a powerful molecular signature for distinguishing sex phenotypes (male or female) [197]. To validate the gender variable in the demographic database, gene expression profiles for Y-linked genes (m=65) were subjected to principal component analysis (PCA) to predict sex phenotypes. To visualize the sample clusters, a scatter plot for PC2 against PC1 scores was applied (as described in batch-effect section). For Y- linked genes with multiple probes, a gene probe with the

maximum median value across the samples was selected. Suspicious samples were verified using the reference database at MRC unit, in The Gambia (SCC1062).

2.12.5 Non-specific feature filtering

While genome-wide approaches provide comprehensive data reassures, not all the gene features are relevant to a particular disease[125]. To minimise the dimensionality of the data and reduce the potential of false discoveries (i.e. in differential expression analyses), a range of non-specific filtering approaches were applied to eliminate the irrelevant gene probes. In particular, gene probes that failed any of the following criteria were eliminated.

1. **Annotation:** Gene probes without standard annotations (i.e. Gene SYMBOL or ENTREZID) were filtered out.
2. **Signal intensity:** gene probes with lower intensity values as compared to the negative controls were also eliminated. For each sample, the following threshold was applied: $A_i = M_i + 2 * MAD_i$; where
    - **M**=median value between the gene negative control gene probes.
    - $\mathbf{MAD_i}$= median absolute deviation across the negative control probes.

Here, a gene probe was eliminated if at least 10% of the signal intensities were less than $\mathbf{A_i}$.

3. ***Between-sample variability***: This filter eliminated the gene probes with low coefficient of variation (CV) between the samples (<10%).

2.12.6 Consolidation of gene probes

In the microarray assay, a gene is investigated using multiple probes[198]. To remove the redundant gene probes, the filtered database was further

subjected to a **maximum mean** filter [199]. In particular, a gene probe with the maximum mean across the samples was selected for each gene.

2.13 Investigation of cellular pathway responses to pneumonia (Chapter 4)

Whole blood is a complex tissue comprising multiple immune cell types in varying proportions between the samples of different phenotypes [102]. To investigate the cellular pathway responses (objective one), I assess the association between pneumonia severity and the proportions of immune cell types. To estimate the sample proportions of immune cell types, I applied computational deconvolution analysis approach on the training data whole blood transcriptome. In particular, I sought to estimate **P** (i.e. the sample proportions of immune cell types) using the following linear model equation:

$$\boldsymbol{W} = \boldsymbol{E} * \boldsymbol{P}; \text{ where}$$

- **W** represent the observed data matrix (**m** genes by **n** samples) for the heterogeneous whole blood transcriptome
- **E** is a data matrix (**m** genes by **k** cell types) for the cell type-specific expression signals
- **P** is a data matrix (**k** cell types by **n** samples) for the sample proportions of the **k** cell types in **E**.

In a typical transcriptomic experiment like this study, the microarray assay or *RNA-seq* are often applied to measure **W** while **E** and **P** are often unknown. When **E** is known, *partial deconvolution* algorithms are often applied [139, 200]. While optimised cell type-specific expression signatures (i.e. **E**) exist for *partial deconvolution [137, 139, 200],* their application is limited by

platform-specific differences. Alternatively, *semi-supervised deconvolution approaches, which* apply marker genes to estimate **P** without the knowledge of **E** [201], are more applicable because they are robust to array platform differences. However, the existing marker genes resources have little overlap and vary in performance. Therefore, I further sought to enhance computation performance by applying a data fusion approach to derive a unified marker gene list called IBML (Integrated blood marker list) was derived (next section).

2.13.1 Derivation of IBML

Briefly, IBML stands for an Integrated Blood Marker List. This list contains cell type specific marker genes for six human immune cell types: B, T, NK, Dendritic, Monocytes and Neutrophils. To derive IBML the following steps were applied:

1. ***Selection of eligible marker genes from the CellMix R package[160].*** This package has compiled comprehensive resources for computational deconvolution analyses including marker genes. To select eligible markers, the following criteria were applied:
    a. Human genome
    b. A marker gene associated with any of the six cell types above
    c. Valid identification ID corresponding to the ENTREZID.

2. ***Aggregation of all eligible markers regardless of cell type.*** To facilitate subsequent analyses, all the eligible markers were annotated to ENTREZIDs.

3. ***Selection of cell type-specific markers:*** To select the cell type-specific markers, AUC values from the ROC analysis [202, 203] were applied. To calculate the AUC values, the **GSE22886** data set [204] was applied. This data set was originally applied to derive the IRIS marker gene list [204], and here it was preferred because of better coverage of cell types and sample sizes. For each marker gene, AUC values were estimated for each type. To calculate the AUC values, expression values for each cell type were compared against the average values from the other cell types (one versus other). Thus, a dummy variable (gold standard) was generated for each cell type **j** such that sample **i=1** if it represent cell type **j,** otherwise **i=0**:

$$\boldsymbol{y_{ji}} = \begin{cases} 1, if\ i = j \\ 0, if\ i \neq j \end{cases}$$

An AUC value ranges between 0 (perfect negative discrimination), 0.5 (no discrimination) and 1 (perfect positive discrimination). Here, a marker gene **g** was assigned to cell **k** if $\mathbf{AUC_{gk}}$=1.

4. ***Validation of selected markers****.* The same approach (AUC=1) was applied on the **GSE1133** [205] and **GSE28490** [206] data sets. To derive the final list of markers in the IBML, markers that were consistently associated with AUCgk=1 in all the three data sets were selected.
5. ***Performance assessment of IBML****.* Here, IBML was applied to deconvolute whole blood transcriptomes with existing (laboratory-measured) proportions of immune cell types (**GSE20300, GSE3649, GSE87301, GSE25504, GSE64385**) using the *ssFrobenius* algorithm

[160, 201]. To assess the agreement between the measured and predicted proportions, Pearson correlation coefficient (R) values were applied. The performance of IBML was compared to the original marker gene resources in the CellMix package. **Table2.5** shows the datasets that were applied in this analysis.

Application	Data	Types	Reference
Selection of marker genes	GSE22886	Purified samples	Abbas(2005)[204]
Validation selected marker genes (ROC analysis)	GSE1133	Purified samples	Su (2004) [205]
	GSE28490	Purified samples	Allantaz(2012)[206]
Performance assessment of IBML	GSE20300	Whole blood	Shen-Orr (2010)
	GSE64385	Whole blood	Becht(2016)
	Pneumonia (Unpublished)	Whole blood	Ghazal et. al
	GSE87301	Whole blood	Shannon(2012)
	GSE25504	Whole blood	Smith(2014)

Table 2.4: Gene expression data applied to derive the IBML marker gene resource. The data were downloaded NCBI's gene expression omnibus [156]

2.13.2 Application of IBML on the pneumonia database

To investigate the cellular pathway responses in pneumonia, the IBML resource was applied to deconvolute sample proportion of immune cell types using the pneumonia whole blood transcriptome. To further validate the performance of IBML, the predicted proportions of neutrophils and lymphocytes were directly compared (Pearson's r) to the corresponding values existing in the clinical database. Subsequently, the deconvoluted proportions were compared between the pneumonia severity groups. In particular, linear regression approach was applied to adjust for potential confounders (age, nutrition status and antibiotic usage). Then, fitted values were compared graphically using the whisker and boxplots.

2.14 Unsupervised machine learning

2.14.1 Introduction

This section introduces the unsupervised machine learning algorithms that were applied to independently assess the inherent structures of the central data resources for this thesis. In particular, principal component analysis (PCA) and T-Distributed Stochastic Neighbourhood Embedding (T-SNE) algorithms were applied for dimensional reduction and visualisation. Further, clustering algorithms such as K-means and hierarchical agglomerative clustering (with heatmaps) were applied to identify inherent sample clusters.

2.14.2 Principal component analysis

Principal component analysis (PCA) is a linear dimensional reduction technique (by Pearson K. (1901) [207] and Hotellling H. (1933) [208] for visualizing high dimensional data [209-211]. Given a quantitative data set $\boldsymbol{X}$ with n samples and p variables $(\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_p)$, PCA transforms the original variables into linear combinations of new uncorrelated variables $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_p)$ called principal components (PCs). In particular, the PCs $(\boldsymbol{\xi}_j)$ are derived in a decreasing order of importance (variance) such that more data variability is captured by the leading $k \leq p$ principal components, which enables data reduction. While PCA is not a clustering algorithm, scatter plots of the leading components often reveal important data clusters. However, PCA is sensitive to data scale, and data standardization is mandatory if the original variables have different scale units [210].

Formerly, principal component analysis (PCA) solves for matrix $\boldsymbol{A}$ using the following definition:

$$\boldsymbol{\xi}_j = \sum_{i=1}^{p} a_{ij}\boldsymbol{x}_i \Leftrightarrow \boldsymbol{\xi} = \boldsymbol{A}^T\boldsymbol{X}, \qquad where\ cor(\xi_j, \xi_k) \approx 0 \ \ for\ i \neq j$$

Where $\boldsymbol{x}_i$ are the original variables in $\boldsymbol{X}$, and each column j of matrix **A** contains the coefficients (a_{ij}) for component $\boldsymbol{\xi}_j$. In particular, the columns of **A,** $\boldsymbol{a}_j = (a_{j1}, a_{j2}, \ldots a_{jp})$, contain the eigenvectors for the covariance matrix ($\boldsymbol{\Sigma}$) of the original data X (or correlation matrix of the standardized data). Further, the corresponding eigenvalue for $\boldsymbol{a}_j$ (λ_j) represent the variance of PC_j such that $\lambda_1 > \lambda_2 > \cdots > \lambda_p$. Therefore, an $eigendecomposition$ of X is often applied to derive the principal components as follows [210]:

1. Standardize the original variables in x_j to mean 0 and variance 1. This step is mandatory if the original variables have different scale units.
2. Extract the correlation matrix of the standardized data $\widehat{\boldsymbol{\Sigma}}$
3. Perform an $eigendecomposition$ of $\widehat{\boldsymbol{\Sigma}}$ to get the eigenvectors, which are the coefficients (a_{ij}) of $\hat{\boldsymbol{\xi}}_j = \sum_{i=1}^{p} a_{ij}\boldsymbol{x}_i$. This algorithm is implemented in R using the $princomp$ function (stats package)

However, in this thesis we applied the singular value decomposition (SVD) approach to derive the principal components ($\hat{\boldsymbol{\xi}}_j$) using the $prcomp$ function (stats package) in R. Notably, the SVD approach has better numerical accuracy than $eigendecomposition$ because the variances are computed using unbiased divisor $(N-1)$ (https://stat.ethz.ch/R-manual/R-patched/library/stats/html/prcomp.html).

To visualize the transformed data, samples were projected into a two-dimensional scatter plot using the first (x-axis) and second (y-axis) principal components.

Similarly, the SVD approach solves for the principal components coefficients (a_j) (i.e. eigenvectors of $\boldsymbol{\Sigma}$) but without estimating the sample covariance/correlation matrix $\boldsymbol{\Sigma}$ as in the $eigendecompsiotion$. In particular, this approach is based on the following property: *the right singular vectors of a matrix* $\boldsymbol{Z}$ *are the eigenvectors of* $\mathbf{Z}^{\mathrm{T}}\mathbf{Z}$*.* Briefly, the SVD for matrix $\boldsymbol{Z}_{m\times n}$ with rank r is defined as follows:

$$\boldsymbol{Z} = \boldsymbol{UDV^T} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{r}} \sigma_{\boldsymbol{i}} \boldsymbol{u_i} \boldsymbol{v_i^T}$$

Where $\boldsymbol{D}$ is the diagonal matrix of the singular values $\sigma_{\boldsymbol{i}}$, and the column vectors of matrices $\boldsymbol{U}_{(m\times r)}$ and $\boldsymbol{V}_{(n\times r)}$ represent the left and right singular vectors (respectively) for **Z** such that $\boldsymbol{U^T U} = \boldsymbol{V^T V} = \boldsymbol{I_r}$**.**

Therefore (using the property above), the column vectors of $\boldsymbol{V}_{(n\times r)}$ contain the coefficients (a_j) for the principal components $(\hat{\boldsymbol{\xi}}_j)$ if $\boldsymbol{Z} = \frac{\mathbf{1}}{\sqrt{\boldsymbol{n-1}}}(\boldsymbol{X} - \mathbf{1}\boldsymbol{m}^T)$ such that $\boldsymbol{Z^T Z} = \boldsymbol{\Sigma}$ (i.e. covariance matrix of **X**). Here**,** **X** represents the standardised input data matrix $(n \times p)$, **m** is a p-dimensional vector of sample means and **1** is an n-dimensional column vector of ones. In other ways, PCA involves deconvoluting the column right singular vectors $(\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots \boldsymbol{v_r})$ of matrix **Z**, where the singular values (σ_i) are the standard deviations $\left(\sqrt{eigenvalues}\right)$ for the resulting principal components $\left(\boldsymbol{\xi_j}\right)$ [210].

2.14.3 T-Distributed Stochastic Neighbourhood Embedding (T-SNE)

T-SNE is a nonlinear unsupervised machine-learning algorithm (developed by Laurens van der Maaten and Geoffrey Hinton, 2008), to enhance dimensionality reduction and visualization of high-dimensional data [212]. Notably, while T-SNE and PCA have a common goal, PCA is a linear

algorithm that preserves the global structure of the data. On the other hand, T-SNE seeks a nonlinear projection of high dimensional data into a low dimensional space while preserving the local and global structures [213] . For each sample i, distributions of the similarity scores calculated using the low ($\boldsymbol{Q}_i$) and high ($\boldsymbol{P}_i$) dimensional data. In particular, the similarity of two data points is defined by the conditional probability such that point x_i would pickas its neighbour if neighbours were picked in proportion to the Gaussian probability density cantered at However, it should be noted that T-SNE was adapted from the SNE (Stochastic Neighbourhood Embedding) as described below [214].

Formerly, let **X** and **Y** be the data points in the high and low dimensional spaces (respectively).Then, the corresponding similarity scores between sample i and j in the high ($p_{j|i}$) and low ($q_{j|i}$) dimensional spaces are defined as follows:

$$p_{j|i} = \frac{\exp\left(-\frac{\left\|\mathrm{x_i} - x_j\right\|^2}{2{\sigma_i}^2}\right)}{\sum_{k \neq l} \exp\left(-\frac{\|\mathrm{x_k} - x_l\|^2}{2{\sigma_i}^2}\right)}; p_{j|i} = 0 \; if \; i = j$$

$$q_{j|i} = \frac{\exp\left(-\left\|\mathrm{y_i} - y_j\right\|^2\right)}{\sum_{k \neq l} \exp(-\|\mathrm{y_k} - y_l\|^2)}; where \; q_{j|i} = 0 \; if \; i = j$$

Notably, $\sigma = 1/\sqrt{2}$ in the low dimensional space. This parameter is associated with the model *perplexity* (i.e. the number of close neighbours for each point). To find the optimal projection of the data into low the dimensional space, SNE minimises the following Kullback-Leibler (KL) divergence cost function (C) using gradient decent algorithm:

$$C = \sum_{i=1}^{n} KL(P_i|Q_i) = \sum_{i=1}^{n}\sum_{j=1}^{n} p_{j|i}\log\left(\frac{p_{j|i}}{q_{j|i}}\right)$$

$$\Rightarrow \frac{\delta C}{\delta y} = 2\sum_{j}^{n}(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

$$\Rightarrow y_{(t)} = y^{(t-1)} + \eta\frac{\delta C}{\delta y} + \alpha(t)(y^{(t-1)} - y^{(t-2)}$$

Where, $\alpha(t) =$ momentum term, $y(t)$=solution at time t, and η=the learning rate. Notably, the KL function is asymmetric such that the penalty (C) is higher if the distance between two points increases after the projection into the low dimensional space ($p_{j|i} > q_{j|i}$). However, the SNE algorithm has two major challenges (i) crowding of data points (ii) the cost function is difficult to optimise.

To reduce the overcrowding problem, T-SNE algorithm applies a heavy tailed Cauchy distribution (t-distribution with one degree of freedom) to project the data into the low dimensional space. Notably, the t-distribution is robust to outliers than the Gaussian distribution. Further, T-SNE applies symmetric similarity scores such that $p_{ij} = p_{ji}$ and $q_{ij} = q_{ji}$, which reduces the complexity of the cost function (C) thereby improving computational efficiency. Formerly, the similarity scores for high (p_{ij}) and low (q_{ij}) dimensional spaces, the optimisation problem at iteration t ($y_{(t)}$) and the T-SNE algorithm are presented below.

$$p_{ji} = p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n};\ \ where\ p_{ij} = 0\ if\ i = j$$

$$q_{ji} = q_{ij} = \frac{\left(1 + \left\|\mathrm{y_i} - x_j\right\|^2\right)^{-1}}{\sum_{k\neq l}(1 + \|\mathrm{y_k} - x_l\|^2)^{-1}};\ \ where\ q_{ij} = 0\ if\ i = j$$

$$\Rightarrow \frac{\delta C}{\delta y} = 4\sum_{j}^{n}(p_{ij} - q_{ij})(y_i - y_j)\left(1 + \|y_i - y_j\|^2\right)^{-1}$$

$$\Rightarrow y_{(t)} = y^{(t-1)} + \eta\frac{\delta C}{\delta y} + \alpha(t)(y^{(t-1)} - y^{(t-2)}$$

Where n=total number of data points, $\alpha(t) =$ momentum term, $y(t)$=solution at time t, and η=the learning rate. The T-SNE algorithms is briefly outline below [212]:

The t-SNE Algorithm [215]

1. Compute pairwise affinities $p_{i|j}$ with perplexity $Perp$
2. Set $p_{ij} = \frac{p_{i|j}+p_{j|i}}{2n}, where\ n = number\ of\ samples$
3. Sample initial solution $Y^{(O)} = y_1, y_2, \dots \dots y_n\ \ from\ N(0, 10^{-4}\boldsymbol{I})$

for t=1**,** T **do** {

1. Compute the low-dimensional affinities q_{ij}
2. Compute the gradient $\frac{\delta C}{\delta y}$
3. Set $y_{(t)} = y^{(t-1)} + \eta\frac{\delta C}{\delta y} + \alpha(t)(y^{(t-1)} - y^{(t-2)}$

}

}

Nevertheless, T-SNE has some limitations including the difficult choice of the perplexity parameter and interpretation of results. For example, T-SNE captures more local or global variations if the perplexity parameter is too small or too large (respectively) consequently leading to different structures. Further, cluster size and distances between clusters has no clear interpretation, and random noise can lead into false positive structures [216].

2.14.4 Hierarchical agglomerative clustering (HAC)

HAC is a family of unsupervised machine learning algorithms for exploring high dimensional data clusters in a two-dimensional space (typically genes on the rows and samples on the columns), often using $dendrograms$ and $HeatMaps$ [210, 217]. Briefly, HAC algorithms start with each sample as a

cluster and iteratively merge nearest clusters until all the samples belong to one cluster (i.e. bottom to top approach). At each step, the nearest clusters are merged using different linkage algorithms including $single$, $complete$, $average$, $centroid$, and $Ward's$ methods [210, 218] (**Table 2.5)**.

Linkage method	Linkage approach	Equation
Single (Single nearest)	Merges two clusters with the minimum of the minimum distances.	$D(C_1, C_2) = \min_{x_1 \in C_1, x_2 \in C_2} D(x_1, x_2)$
Complete (Complete furthest)	Merges two clusters with the minimum of the maximum distances.	$D(C_1, C_2) = \max_{x_1 \in C_1, x_2 \in C_2} D(x_1, x_2)$
Group average (Unweighted pair group mean averaging)	Merges two clusters with the minimum average pairwise distances between the sample points	$D(C_1, C_2) = \frac{1}{\lvert c_1 \rvert}\frac{1}{\lvert c_2 \rvert}\sum_{x_1 \in C_1}\sum_{x_2 \in C_2} D(x_1, x_2)$
Centroid	Merges two clusters with the minimum distance between their centroids (means)	$D(C_1, C_2) = D\left(\left(\frac{1}{\lvert c_1 \rvert}\sum_{\boldsymbol{x} \in C_1} \vec{x}\right), \quad \left(\frac{1}{\lvert c_2 \rvert}\sum_{\boldsymbol{x} \in C_2} \vec{x}\right)\right)$
Ward's method	Merges two clusters with the smallest change (**∇**) in the total distance to the centroid.	$\nabla = TD - (D_{C1} + D_{C2})$ Where • $TD = \sum_{x \in C_1 \cup C_2} D(\boldsymbol{x}, \mu_{C_1 \cup C_2})$ • $D_{C1} = \sum_{x_1 \in C_1} D(x_1, \mu_{c1})$ • $D_{C2} = \sum_{x_2 \in C_2} D(x_2, \mu_{c2})]$

Table 2.5: Linkage algorithms for hierarchical agglomerative clustering (HAC) analysis . D(.)=Distance function [210, 217].

To identify the nearest clusters, various distance metrics such as the $Euclidian\ distance\left(\sqrt{\sum_{j=1}^{P}(x_{ij} - x_{kj})^2}\right)$, $Manhattan$ or $City - block\left(\sum_{j=1}^{P}|x_{ij} - x_{kj}|\right)$, $Minkowski$ distance with order D $\left(\sqrt[D]{\sum_{j=1}^{P}|x_{ij} - x_{kj}|^D}\right)$, $Chebyshev\ (max|x_{ij} - x_{kj}|)$ and $Canberra$ distance $\left(\sum_{j=1}^{P}\frac{|x_{ij} - x_{kj}|}{|x_{ij} + x_{kj}|}\right)$ are applied. Here, x_i and x_k represent a P-dimensional data points in clusters i and k (respectively), and P is the number of features in the input data [125, 210, 219].

In this analysis, we applied the *Euclidian distance* and the $complete\ linkage\ clustering$ algorithm to visualise the high-dimensional data using $HeatMaps$ and$dendrograms$. Notably, the choice of the linkage algorithm depends on the analysis objective and could reveal different structures. Here, the complete linkage algorithm was preferred in order to generate compact clusters (i.e. spherical clusters with consistent diameters), and avoid the $chaining\ problem$ associated the single linkage. Further, the single and complete linkage algorithms are computationally efficient but could be susceptible to extreme outliers [125, 210, 219], which were eliminated during quality control analysis (**Chapter 3**). In particular, the $heatmap.3$ function was applied: (https://www.rdocumentation.org/packages/GMD/versions/0.3.3/topics/heatmap.3).

2.14.5 K-means clustering

In **Chapter3**, we applied the K-means clustering algorithm [220] to explore the data clusters associated with the pneumonia transcriptome. For an input data matrix $X_{n\times p}$, the k-means algorithm partitions the n samples $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots . \boldsymbol{x}_n, where\ x_i \in \mathbb{R}^{\boldsymbol{p}})$ into K pre-defined clusters such that the within-group sum of squared deviations from the centroid are minimized. The distortion function is defined as follows:

$$J(K) = \sum_{k=1}^{K}\sum_{i=1}^{n}(\|\boldsymbol{x}_{ik} - \boldsymbol{c}_k\|)^2$$

Where $\boldsymbol{c}_k$ =centroid for cluster C_k, $k = 1,2..K; and\ \boldsymbol{x}_{ik} \in C_k$. In particular, the K-means algorithm works as follows [210]:

1. Initialize the algorithm with random centroids $(\boldsymbol{c}_1, \boldsymbol{c}_2, \ldots, \boldsymbol{c}_K)$. This step is often repeated to avoid local minima.

2. Assign each sample ($\boldsymbol{x}_i \in \mathbb{R}^{\boldsymbol{p}}$) to the nearest cluster (C_k) based on the minimum distance (i.e using the Euclidian distance metric) to the centroid $(\boldsymbol{c}_k)$.
3. Within each cluster, calculate the new centroid $\boldsymbol{c}_k$ using the mean of the samples in that cluster (C_k) (i.e. $\boldsymbol{c}_k = \sum_{x \in C_k} \boldsymbol{x}$).
4. Re-assign the samples to the nearest clusters based on the new centroids.
5. Repeat step (2) to (4) until the algorithm converges (i.e. $J(K)$ remains constant).

2.14.5.1 Cluster stability using the Jaccard coefficient with bootstrap samples

While K-means is a popular clustering algorithm, choosing the number of clusters (K) is often difficult. To select stable number of the K-means clusters (K), in this thesis we applied the *Jaccard coefficient [221]* using a sequence of 1000 bootstrap samples. Briefly, the *Jaccard coefficient* is defined as the proportion of elements shared by two sets A and B as defined below [222-224]:

$$Jaccard\ coefficient = \frac{|A \cap B|}{|A \cup B|}$$

To estimate the stability (S_k) of cluster C_k (where k=1,..K), 1000 Jaccard coefficients comparing C_k (i.e. based on the original data) to each bootstrap sample-based cluster $(C_{bk}, where\ b = 1,2, \ldots 1000)$ were averaged using an arithmetic mean. Thus, the stability of each cluster was defined as the mean of its $Jaccard\ coefficients$ as compared to the clustering based on all bootstrap iterations [225].

$$S_k = \frac{1}{1000} \sum_{b=1}^{b=1000} \frac{|C_k \cap C_{bk}|}{|C_k \cup C_{bk}|}$$

To identify the optimal number of clusters (K), the cluster-specific stability estimates (S_k) were averaged based on the pre-defined choice of K as follows [225]:

$$S_K = \frac{1}{K} \sum_{k=1}^{K} S_k$$

Formerly, let $\boldsymbol{D}_{(\boldsymbol{n}\times\boldsymbol{p})} = (\boldsymbol{x}_1, x_2, \ldots . \boldsymbol{x}_n, where\ x_i \in \mathbb{R}^{\boldsymbol{p}})$ be the original matrix. Then, this thesis applied the following algorithm to assess the stability of clusters ranging from K=2 to K=10 using the $clusterboot$ function in the fpc R package [225]:

1) Apply the K-means clustering using the original data $\boldsymbol{D}_{(\boldsymbol{n}\times\boldsymbol{p})}$
2) Draw a bootstrap sample $\boldsymbol{D}^{(\boldsymbol{b})}_{(\boldsymbol{n}\times\boldsymbol{p})}$ with replacement from the original data $\boldsymbol{D}_{(\boldsymbol{n}\times\boldsymbol{p})}$, where b=1, 2,..1000).
3) Apply the K-means clustering on $\boldsymbol{D}^{(\boldsymbol{b})}_{(\boldsymbol{n}\times\boldsymbol{p})}$
4) For each cluster $\boldsymbol{C}_k$, where k=1, 2,…,K; compare the agreemen between the original cluster and the bootstrap sample-based cluster ($\boldsymbol{C}_{kb}$) using the Jaccard coefficient.
5) Repeat steps (2) to (4) many times (i.e. B=1000).
6) Estimate the stability (S_k) of each cluster $\boldsymbol{C}_k$ by the arithmetic mean of its $Jaccard\ coefficients$ over all the B bootstrap iterations
7) Calculate the overall stability (S_K) associated with the choice of K partitions of the data using the arithmetic mean of the cluster-specific stability scores (S_k).
8) Repeat steps (1) to (7) for different values of K and select the partition with the highest mean stability across the K clusters (S_K).

2.15 Investigation of systemic molecular responses

In this thesis, molecular responses were investigated at the gene and pathway analytic levels. At the gene analytic level, differentially expressed genes (DEGs) were characterized. Subsequently, the DEGs were applied to investigate pathway responses using a range of biochemical pathway databases.

2.15.1 Identification of differentially expressed genes (DEGs) using empirical Bayes moderated t-test.

Microarray gene expression studies often seek to identify differentially expressed genes (DEGs) for further investigations such as pathway and biomarker analyses [71, 168, 171]. To identify the DEGs associated with pneumonia severity, this thesis applied an empirical Bayes moderated t-test by Smyth et. al. (2004) [226]. In particular, the moderated t-test seeks to overcome the $denominator\ challenge$ (dependence of the t-statistic on sample variance), which is a potential problem in microarray studies especially with small sample sizes [227-234].

Here, we introduce the ordinary t-test followed by the moderated test in the context of two independent samples (X and Y). Let $X_g = (x_{g1}, x_{g2}, .x_n) \sim iid\ N(\mu_{gx},\ \sigma_g^2)$ and $Y_g = (y_{g1}, y_{g2}, \ldots y_{gn}) \sim iid\ N(\mu_{gy},\ \sigma_g^2)$ be the expression values for gene g in group X and Y (respectively), where μ_{gx} and μ_{gy} are the population means and σ_g^2 is a shared variance (i.e. assuming homoscedasticity: $\sigma_{gx}^2 \approx \sigma_{gy}^2 \approx \sigma_g^2$). Then, an ordinary t-test seeks to test the null hypothesis that $\boldsymbol{H_0}: \boldsymbol{\mu_{gx} = \mu_{gy}}$ against the alternative hypothesis that $\boldsymbol{H_0}: \boldsymbol{\mu_{gx} \neq \mu_{gy}}$ using the following statistic:

$$t_g = \frac{(\bar{X}_g - \bar{Y}_g) - (\boldsymbol{\mu_{gx}} - \boldsymbol{\mu_{gy}})}{S_{g(p)}\sqrt{v_g}} = \frac{\bar{X}_g - \bar{Y}_g}{S_{g(p)}\sqrt{v_g}} \sim t_{n+m-2}$$

Where $\bar{X}_g = \frac{1}{n}\sum_{i=1}^{n} x_{gi}$ and $\bar{Y}_g = \frac{1}{m}\sum_{i=1}^{m} y_{gi}$ are the group-specific sample means, $v_g = \left(\frac{1}{n} + \frac{1}{m}\right)$ and $S_{g(p)} = \sqrt{\frac{(n-1)S_{gx}^2 + (m-1)S_{gy}^2}{n+m-2}}$ is the pooled sample standard deviation where $S_{gx}^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}\left(x_{gi} - \bar{X}_g\right)^2$ and $S_{gy}^2 = \frac{1}{(m-1)}\sum_{i=1}^{m}\left(y_{gi} - \bar{Y}_g\right)^2$ are the group-specific sample variances [235].

Notably, the goal of differential gene expression analysis is to select top-ranked genes often using the p-value from the t-statistic. However, the ordinary t-statistic (t_g) is inversely related to the pooled variance ($S_{g(p)}$), which varies across the genes. Thus, genes with low variance (i.e. due to signal intensities) tend to have inflated values of t_g, and therefore more likely to be incorrectly declared as significant (false positive discoveries). Consequently, this *denominator challenge* potentially generates misleading lists of candidate genes for further investigations [228, 230, 231, 234].

To mitigate the unwanted dependence of t_g on $S_{g(p)}$, an empirical Bayes [227, 229] moderated t-test (implemented in the $limma\ R\ package$ [236]) is often applied in differential gene expression analyses [145, 171, 178, 226]. Basically, this approach applies a hierarchical Bayesian model to shrink the gene-level pooled variance S_{gp} towards the pooled estimate (i.e. borrowing strength from the distribution of all the genes). In particular, the moderated t-test statistic for gene g ($\widetilde{t_g}$) has the following closed form:

$$\widetilde{t_g} = \frac{\overline{X_g} - \overline{Y_g}}{\tilde{S}_{g(p)}\sqrt{v_g}} \sim t_{d_g + d_0}$$

Where d_g is the observed degrees of freedom (df), and $\tilde{S}^2_{g(p)} = \frac{d_g S^2_{g(p)} + d_0 S^2_0}{d_g + d_0}$ is the posterior mean of the population variance (σ^2_g), which is estimated using the prior variance (S_0) and degrees of freedom (d_0) given the observed variance(S_{gp}). Thus, the observed variance (S_{gp}) is shrunk towards the pooled estimate ($\tilde{S}^2_{g(p)}$) by the prior variance (S_0) and degrees of freedom (d_0) such that $\tilde{S}^2_{g(p)} = S^2_{g(p)}\ if\ d_0 = 0$. The prior parameters ($S_0$ and d_0) are empirically estimated from the observed data using a series of closed form equations (more details are in Smyth et. al (2004)) [226]. In summary, the moderated t-test is a hybrid test applying a Bayesian estimate (pooled variance) into a classic statistic framework.

In this thesis, the limma $Bioconductor\ R$ package[226] was applied to conduct the moderated t-test. In particular, this analysis identified differentially expressed genes (DEGs) between the non-pneumonia controls and each severity state (mild, severe and very severe), respectively. To adjust for potential confounding, the following covariates were included in the design matrix of the linear model: (i) age, (ii) nutrition status and (iii) antibiotic usage in the previous week. The nutrition status covariate was estimated from the principal component analysis (PC1) of weight-for-age (under-weight), height-for-age (stunting) and weight-for height (wasting) Z-scores. To guard against false discoveries due to multiple testing, the Benjamini and Hochberg FDR correction method was applied ([237, [238]].

For each severity state, the *TopTable* function was applied to rank and select the DEGs using the following criteria:

1. B>0, where B is the log odds for a gene being differentially expressed i.e. if p=probability of a gene being differentially expressed, then B=log(p/1-p).
2. FDR (i.e. Adjusted P-value) <0.05
3. |Fold change| ≥2 were of much interest (Otherwise, stated)

2.16 Adjusting for false discoveries due to multiple testing

2.16.1 Introduction

In classic statistics, hypothesis testing is associated with type-I (α) and type-II (β) errors. Type-I error occurs when a true null hypothesis is incorrectly rejected (false positive discovery (i.e. FP in **Table2.6**)) while type-II error occurs when there is no sufficient evidence to reject a false null hypothesis (false negative discovery; scenario FN in **Table2.6**) [235, 238].

Null hypothesis (H_0) true?	**Test significant (reject H_0)?**		**Total**
	Yes	**No**	
Yes	**FP**	TN	m0
No	TP	**FN**	m1
Total	R	A	m

Table 2.6: Classification of multiple hypothesis tests FP=false positive, FN=false negatives, TP=true positives, TN=True negatives, R=total number of rejected hypotheses, m= total number of hypotheses.

While both errors are critical and often kept at minimal rates (i.e. α≤5% and β≤10%) by study design, type-I error is often considered more serious than type-II error. In this section we discuss the potential impact of multiple hypothesis testing on false positive discoveries. Potentially, the number of false positive discoveries (FP) is directly related to the number of hypotheses tested (m) where α (type-I error) is proportionality constant.

$$FP = \alpha * m$$

Consequently, more false positive discoveries (FP) are expected for large number tests (m) even if α is fixed constant; which is more serious for high-dimensional data such as microarray transcriptomes [239]. To account for false discoveries, a range of multiple testing correction procedures[240] such as the Bonferroni method [241] and Benjamini & Hochbergh's (BH) procedure [238] are often applied (discussed below).

2.16.2 The Bonferroni multiple testing correction

The Bonferroni correction method seeks to control for the Family-Wise Error Rate (FWER): the probability of incorrectly rejecting at least one true null hypothesis ($Prob(FP > 0)$, **Table2.6**) [240]. For a given α and m independent tests, the FWER is defined as follows:

$$FWER = 1 - (1 - \alpha)^m \geq \alpha;\ \ Since\ (1 - \alpha)^m \leq (1 - \alpha)\ \forall\ m \geq 1$$

Notably, the Sidak correction method rejects any null hypothesis (H_i) if the corresponding P-value (P_i) is less than $1 - (1 - \alpha)^m$ ($i.e. Reject\ H_i\ if\ p_i < p_{crit} = 1 - (1 - \alpha)^m$). However, this approach is more conservative for large number of dependent tests [242]. On the other hand, the Bonferroni approach rejects any null hypothesis (H_i) if the corresponding P-value (P_i) is less than $\frac{\alpha}{m}$ [240]. Equivalently, the raw p-values (P_i) are multiplied by m such that test i is declared significant if and only if $m * P_i < \alpha$.This approach is motivated by the following probabilistic property (i.e. Boole's inequality), which is also valid for dependent tests:

$$\text{FWER} = \bigcup_{i=1}^{m} \alpha_i \leq \sum_{i}^{m} \alpha_i = m\alpha,$$

Where m = number tests conducted, and α_i = type-I error for test i. Ideally, fixing $\alpha_i = \frac{\alpha}{m} = \alpha^*$ keeps the FWER≤ *α* (the desired type-I error):

$$\mathrm{FWER} = \bigcup_{1}^{m} \alpha^* \leq \sum_{i}^{m} \alpha^* = m\alpha^* = m\left(\frac{\alpha}{m}\right) = \alpha,$$

While the Bonferroni correction is simple and justifiable in some extreme circumstances (i.e. comparing the efficacy of multiple competing drugs to increase the confidence on the chosen drug [238]), controlling for FWER is very stringent consequently reducing the statistical power (i.e. more false negatives) [238, 243, 244]. Hence, it is not ideal for high dimensional data.

2.16.3 Benjamini & Hochbergh (BH) procedure

In this thesis, we applied the Benjamini & Hochbergh (BH) procedure to control for multiple testing across the gene features [238]. The BH procedure seeks to control False Discovery rates (FDR): the expected proportions of false discoveries among the rejected hypotheses [238]. Generally, FDR correction corrections are less stringent than the FWER correction procedures (i.e. Bonferroni method) and suitable for genome-wide analysis. Notably, the BH procedure enables more statistical power but at the expense of higher type-I error rate than the Bonferroni method [229, 239, 245]. Formally, false discovery rate (FDR) is defined as follows:

$$FDR = prob(R > 0) * E\left(\frac{FP}{R}\middle| R > 0\right)$$

Where $E(.)$ denote the expected value, FP=false positives and R=rejected hypotheses (**Table 2.6)** [238]. Therefore, FDR=FWER if all the rejected null hypotheses are true (i.e. FP=R), and no action is taken if one hypothesis is rejected (R=1).

2.16.3.1 Implementation of the BH procedure

For a given α (type-I error), the BH procedure applies the following algorithm to control for FDR among m independent tests:

1. Rank the corresponding p-values (P_k) for each hypothesis test (H_k) in ascending order such that $P_1 < P_2 < \cdots < P_{m-1} < P_m$
2. Find the largest k^* such that $P_k \leq \frac{k^*}{m}\alpha$, where k=1, 2,… k^*.
3. Reject all the null hypotheses (H_k) for $k = 1,2,..,k^*$.

2.16.3.2 Geometric interpretation (implementation) of the BH procedure

Geometrically, the BH procedure is equivalent to the following algorithm:

1. Plot the raw P-values (P_i) on the y-axis against the rank k=1,2,..m (x-axis)
2. Superimpose a straight line $P_k = \frac{\alpha}{m} * k$ such that $\frac{\alpha}{m} = slope$ and y-intercept=0
3. Reject all the null hypotheses (H_k) associated with the points below the straight the line in (2).

2.16.3.3 Assumption

It is worth noting that the BH procedure is valid for m independent tests. To relax this assumption, the Benjamini–Hochberg–Yekutieli procedure [244] introduces $c(m)$ into the denominator of the HM inequality such that $P_k \leq \frac{\alpha}{m*c(m)} * k^*$, where

- c(m)=1, for independent or positively-correlated tests
- $c(m) = \sum_{i=1}^{m} \frac{1}{i}$, for arbitrary dependency

- $c(m) = \sum_{i=1}^{m} \frac{1}{i} \approx \ln(m) + \gamma + \frac{1}{2m}$, for negatively-correlated tests, where γ is a Euler–Mascheroni constant.

Other approaches for controlling the FDR include the positive false discovery rate ($pFDR = E\left(\frac{FP}{R} \middle| R > 0\right)$) by Storey et. al (2002) [239, 246, 247]. Further, bootstrap and permutation procedures have also been proposed [239, 243, 248].

2.17 Identification of significant pathways (Chapter 5)

Pathway analysis presents a powerful system-level approach (as compared to single gene analysis, which ignores the proteins interactions), for investigating candidate vaccines, therapeutic targets and the pathogenesis of diseases [249]. In particular, this approach incorporates validated biochemical pathway databases such as KEGG [250] to facilitate the interpretation of long lists of candidate genes. To identify significant pathways, the following approaches are often applied:

1) Over-representations analysis (ORA) of candidate gene lists using the Fisher's exact test [249, 251-253].
2) Score-based gene set enrichment analysis (GSEA) [157, 254, 255], or
3) More complex analyses that account for the pathway topology (structure) [256-265].

In this thesis, we investigated the pathways associated with the development of pneumonia severity using the following hierarchical candidate lists of differentially (FDR<0.05, |FC|≥2) expressed genes (DEGs):

1. **MSvS:** DEGs that were jointly associated with mild, severe and very severe pneumonia
2. **SvS:** DEGs that were jointly associated with severe and very severe pneumonia but not mild pneumonia.
3. **Vs:** DEGs that were uniquely associated with very severe pneumonia

To investigate these pre-defined genes lists (**MSvS, SvS** and **vS**), we applied the Fisher's exact test-based approach (ORA) using the following biochemical databases: (i) KEGG [250] (ii) REACTOME [266] (iii) Gene ontology (GO) [267] and (iv) HALLMARK. The databases were downloaded from the MSigDB website (http://software.broadinstitute.org/gsea/msigdb/index.jsp) [157].

Notably, gene ontology (GO) terms are classified into three key categories: (i) Cellular Component (CC) where gene products are active, (ii) Molecular Function (MF), which represent the biological function of gene or gene product and (iii) Biological Process (BP), which represent pathways or larger processes that multiple gene products are involved in [268]. However, it is worth noting that the primary goal of gene ontology (GO) terms is protein functional annotation. Therefore, not all the GO terms and their relations represent valid functional protein associations such as protein-protein interactions or mRNA co-expression [268, 269].

2.17.1 Fisher's exact test

Briefly, over-representation analysis (ORA) involves testing the null hypothesis that two lists of genes are independent using a 2×2 contingency table. For example, the columns could represent the list of genes in a

specific biochemical pathway (i.e. TLR4 pathway), whilst the rows could represent the candidate genes of interest such as differentially expressed genes (DEGs) in severe pneumonia (**Table 2.7**) [270].

Differentially expressed gene (DEGs)?	Gene in the pathway?		Total
	Yes	No	
Yes	a	b	(a+b)
No	c	d	(c+d)
Total	(a+c)	(b+d)	(a+b+c+d)=m

Table 2.7: An illustration of possible outcomes in over-representation analysis (ORA) The Fisher's exact test (instead of the Chi-square test) is applied to assess the association between the column and row variables if the expected value of the, b, c, or d is less than 5.

In particular, the following hypotheses are investigated:

H_0: The row and column outcomes are independent (i.e. a/(a+b)=c/(c+d) or the Odds ratio= ad/bc=1)

H_1: There is an association between the row and column outcomes (a/(a+b)≠c/(c+d) or the Odds ratio= ad/bc≠1)

To test the null hypothesis, the Chi-square test for association of two categorical variables is often applied. However, the Chi-square test depends on asymptotic probabilistic properties (i.e. Chi-square distribution), and is not valid when the expected cell counts are less than 5 [235, 271]. Inevitably, this problem is common when investigating small pathways or small candidate gene lists.

Alternatively, the Fisher's exact test is applied. Notably, this approach is valid for any sample size because the P-values are calculated from "exact" probabilities of the observed data and more extreme scenarios. However, the Fisher's exact test is more conservative than the Chi-square test, and

computationally expensive for large sample sizes. With reference to **Table2.7**, the exact probabilities are calculated using the $Hypergeometric$ distribution as follows:

$$Prob(X = a) = \frac{\binom{(a+b)}{a}\binom{(c+d)}{c}}{\binom{m}{(a+c)}} = \frac{\binom{(a+b)}{b}\binom{(c+d)}{d}}{\binom{m}{(b+d)}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!m!}.$$

To calculate the 2-sided p-value, the exact probabilities of the observed data and more extreme scenarios (while fixing the marginal totals) are added as shown in the following algorithm [235, 271]:

1. Calculate the exact probability of the observed data: $P_0 = Prob(X = a)$
2. Reshuffle the table and calculate the $P_i = Prob(X = a_i)$ for all the possible values of a (while fixing the marginal totals).
3. Add all the probabilities that are less than or equal to P_0

   $i.e. Pvalue = \sum_{Pi \leq P0} P_i$.

Manually, the 2-sided P-value is conveniently estimated by doubling the one-sided p-value (assuming a symmetric distribution) as follows:

$Pvalue = 2 * Prob(X \leq a)$; Where a is an observed cell count.

In this thesis, the Fisher's exact test was conducted in R using the $fisher.test$ function (stats package). To adjust for false discovery rate (FDR) due to multiple testing, the $Benjamini\ and\ Hochbergh$ (BH) procedure [238] was applied (within each pathway database) using the $p.adjust\ function$ (stats package).

2.17.2 Limitations of over-representation analysis (ORA)

Notably, the ORA approach is simple, flexible and computationally efficient, and was ideal for our analysis because we investigated pre-defined lists of candidate genes using a comprehensive range of biochemical databases.

However, this black-box approach ignores important information such as pathway structure and gene strength (i.e. expression levels) potentially leading to loss of power [249]. Alternatively, more powerful approaches that account for the pathway structure and the expression intensities could reveal more insights into the pathogenesis of pneumonia [256, 272]. For example, $Sanguinetti\ (2006)$ [259, 260] and $Ocone\ (2011)$ [261] proposed probabilistic models to infer the regulatory activities of transcriptional factors [258-261, 273]. Other pathway structure-based approaches include DAEP (Differential Expression Analysis for Pathways) by Haynes (2013) [274], TAEP (topology-based pathway enrichment analysis) by Yang (2017) [275] and structural equations modeling (SEM) based approaches [276-279].

2.17.3 Network analysis

To compensate for the ORA approach, the **STRING** database was applied to identify validated functional protein-protein network interactions that were associated with pneumonia severity (https://string-db.org/) [280]. Further, we applied the **Pathview** tool (https://pathview.uncc.edu/) [281] to visualize significant KEGG pathways. For each pathway map, up and down regulated genes were highlighted in red and green (respectively) colours while the non-significant and un-annotated genes were represented in grey and white colours, respectively (**Appendix**).

2.18 Investigation of candidate biomarkers for severe pneumonia (Chapter 6)

In this analysis, I coupled cellular pathway biology with machine-learning approaches to derive candidate biomarkers for the detection of mild pneumonia cases at the higher risk of developing severe pneumonia

outcomes. In particular, this analysis involved the following steps (**Figure2.4**):

(i) Cellular pathway-based feature selection (training data)

(ii) Internal performance assessment (training data)

(iii) Independent performance validation (validation dataset)

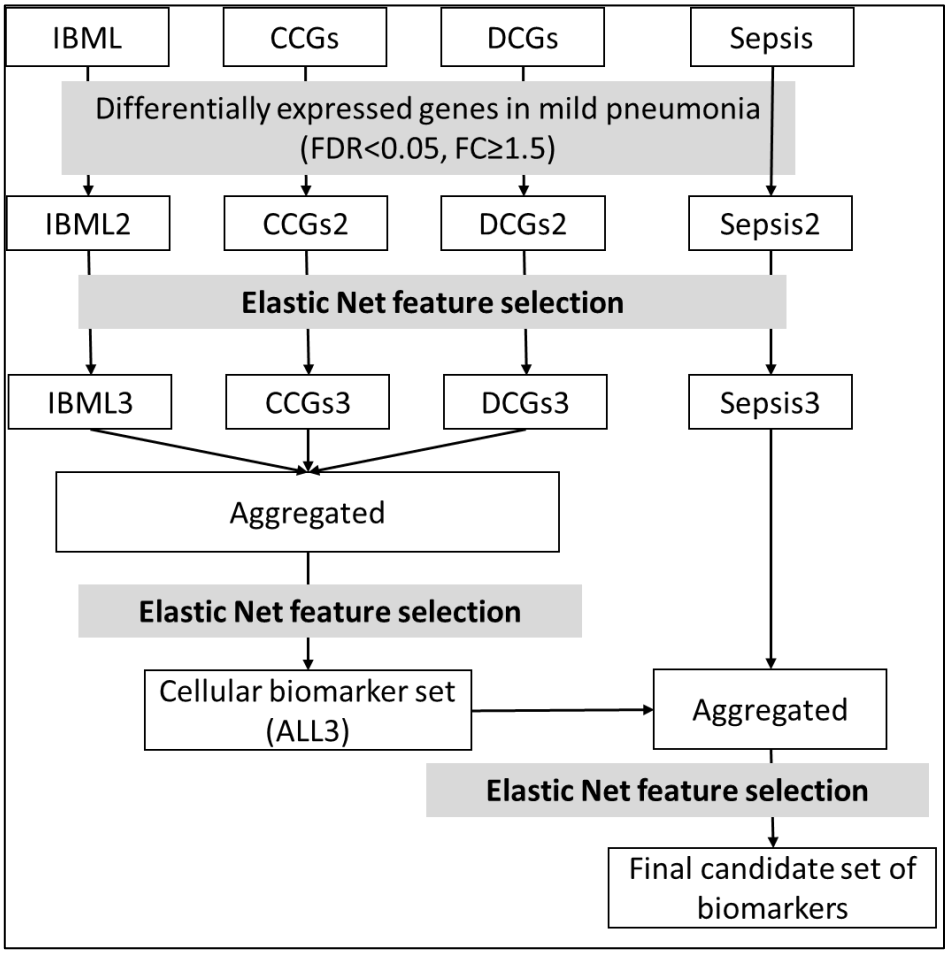


Figure 2.3: An illustration of feature selection for candidate biomarkers of severe pneumonia. Elastic net feature selection was repeated 100 times and markers that were selected all the times were retained. **Abbreviations**: FDR=false discovery rate, FC=fold change, IBML=Integrated Blood Marker List, CCG=Cell proportions Correlated Genes, DCG=Differentially Cell proportion Correlated Genes.

2.18.1 Feature selection

Mainly, feature selection combined machine-learning and cellular pathway centric approach. Further, sepsis markers were also independently assessed and aggregated into the final candidate biomarker set (illustrated in **Figure2.3**). To select the cellular-based features the following criteria were applied:

1. **IBML:** This marker list was derived to enhance computational deconvolution analysis in **Chapter 4.**

2. **Cell correlated genes (CCGs):** These are the genes that were positively associated (FDR<0.05) with the deconvoluted (using IBML) proportions of immune cell types. Here, empirical Bayes linear regression (*limma* R package) [226] was applied to identify the CCGs while adjusting for the potential confounders and false discoveries (BH method)[238]. To remove duplicates, genes that were associated with multiple cell types were assigned to the cell type with the highest positive correlation across all the samples.

3. **Differentially correlated genes (DCGs):** These were the genes with significant statistical interaction (FDR<0.05) between pneumonia severity and the deconvoluted (using IBML) proportions of immune cell type. To derive the DCGs list, empirical Bayes linear regression (*limma* R package) was applied to test for the interaction terms while adjusting for the potential confounders. To remove duplicates, genes that were associated with multiple cell types were assigned to the cell type with the highest positive correlation among the pneumonia cases.

4. **Sepsis markers:** Here, a 52-gene validated neonatal sepsis classifier was applied [136].

For each list, eligible markers were subjected to Elastic Net feature selection (glmm package) [282]. In particular, two criteria were applied to define marker eligibility: (i) differentially expressed in mild pneumonia (ii) showing trend in fold change (increasing/decreasing) with pneumonia severity. To select robust biomarkers, the Elastic net feature selection was repeated 100 times, and markers that were selected together all the time were retained. To identify the optimal values of the model parameters, the cross-validated (here using leave-one-out) *cv.glmnet* function in the *glmnet* R package [282] was applied.

It is worth noting that while the mild pneumonia cases were applied to select eligible gene features, they were excluded from the subsequent analyses (Elastic net feature selection and classification). In particular, the Elastic Net algorithm involved the application of a regularized logistic regression comparing non-pneumonia samples to severe outcomes (severe and very severe pneumonia cases). In particular, the outcome variable for the logistic regression model was coded as follows:

$$y_i = \begin{cases} 1, if\ i = Severe\ or\ very\ severe\ pneumonia \\ 0, if\ i = Nonpnemonia\ control \end{cases}$$

For each cellular list (IBMLs, CCGs, DCGs), candidate biomarkers were selected at the cellular level (i.e. neutrophils, NK, T) followed by an aggregation of the cell type-specific biomarkers. To derive the unified set of cellular biomarkers (ALL3), cellular based biomarkers (i.e. IBML3, CCGs3 and DCGs3, **Figure2.4**) were also aggregated. Finally, an aggregation of cellular-based (ALL3) and sepsis (Sepsis3) biomarkers were also

investigated. At each level, the eligible markers were subjected to the Elastic Net feature selection to select an optimal and robust biomarker set (**Figure2.3**).

2.18.2 Internal performance assessment

To assess the performance of the selected biomarkers (at each level), five classification algorithms were applied: (i) Support vector machine (SVM) (ii) K-nearest neighbour (KNN), (iii) Random forest, (iv) Linear Discriminant Analysis (LDA) and (v) the ROC analysis-based classifier (ROCC) (**Table2.8**). To minimise the prediction bias, each algorithm was coupled with the leave-one-out cross-validation (LOOCV) approach. For each model, the following out of sample performance indices were applied: accuracy, sensitivity, specificity, balanced accuracy (mean of sensitivity and specificity), negative predictive value (NPV) and positive predictive value (PPV).

Algorithm (reference)	**Description**	**R package (Function)**	**Hyperparameters**
SVM[283]	Support vector machines	e1071(svm)	Default: Kernel =radial basis
RF[284]	Random forest	randomForest (randomForest)	Default: ntree=500
KNN[285]	K-nearest neighbour	Class (knn.cv)	K=5, l=0
LDA[286]	Linear discriminatory analysis	MASS (lda)	Default
ROCC[287]	Receiver operation characteristic (ROC) analyses based classifier	rocc (o.rocc)	xgenes=all the selected genes

Table 2.8: Classification algorithms applied to assess the performance of candidate biomarkers for severe pneumonia.

2.18.3 Independent validation of candidate biomarkers

To independently validate the candidate biomarker sets, the training data classifiers were applied to predict severe pneumonia cases in the Basse data set, which was kept independent from all primary analyses. To derive the classifiers, the SVM algorithm was applied. Briefly, support vector machine (SVM) classifier seeks to identify the optimal separating hyperplane, which maximizes the margin of the training data[283]. Here, this algorithm was associated with the best performance in all the training data classification.

2.19 Supervised machine learning algorithms

This section provides a description of the supervised machine learning algorithms that were applied in this thesis (**Chapter 6**) to investigate cellular-based candidate biomarkers for severe pneumonia including (i) Elastic Net feature selection, and (ii) ROCC, KNN, SVM and Random Forest classifiers. In particular, multiple classification algorithms were applied to assess the robustness of the candidate classifiers.

2.19.1 Feature selection using the Elastic Net logistic regression

In this analysis, Elastic Net logistic regression was applied to select optimal subsets of transcriptomic classification features for severe pneumonia. Briefly, the Elastic Net feature selection combines the regularization penalties for LASSO ($L_1 - norm$) and Ridge ($L_2 - norm$) regressions, which enables sparsity and grouped feature selection while stabilizing the variance of regression coefficients especially for correlated variables. Here, the ordinary logistic regression, and the Ridge, LASSO and Elastic Net regularization penalties are introduced.

2.19.1.1 Ordinary logistic regression

Let **x** be an $n \times p$ input data matrix (n samples and p gene features) and $y_i \in \{0,1\}$ be the class label for sample i=1,2,..n. Then, logistic regression models the probability of Y=1 given **x,** (P(Y=1|**x**), using the $logit$ function (i.e. log odds):

$$logit(p) = \log\left(\frac{p_\theta}{1-p_\theta}\right) = \beta_0 + \boldsymbol{\beta^T x} = \boldsymbol{\theta^T x},$$

$$\Rightarrow p_\theta = P(Y=1|x,\boldsymbol{\theta}) = \frac{e^{\boldsymbol{\theta^T x}}}{1+e^{\boldsymbol{\theta^T x}}} = \frac{1}{1+e^{-\boldsymbol{\theta^T x}}}$$

Where $\boldsymbol{\theta}$ are the regression coefficients. Notably, the $logit$ transformation extends the limits of the predicted probabilities $(\hat{p}_\theta)$ from (0,1) to $(-\infty,+\infty)$ [288-292]. To predict the class of a new sample j, an optimal threshold of $\hat{p}_\theta$ is applied (i.e. $y_j = 1$ if $\hat{p}_\theta > 0.5$). The regression coefficients $(\widehat{\boldsymbol{\theta}})$ are estimated using maximum likelihood estimation (MLE), which seeks to to minimize $\frac{-l(\boldsymbol{\theta}|\boldsymbol{x})}{n}$ of the Binomial distribution [235, 289-291, 293] where:

$$P(Y = y_i|\boldsymbol{x},\boldsymbol{\theta}) = {p_\theta}^{y_i}(1-p_\theta)^{(1-y_i)}$$

$$\Rightarrow L(\boldsymbol{\theta}|\boldsymbol{x}) = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} {p_\theta}^{y_i}(1-p_\theta)^{(1-y_i)}$$

$$\Rightarrow l(\theta|\boldsymbol{x}) = log\big(L(\boldsymbol{\theta}|\boldsymbol{x})\big) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} y_i log(p_\theta) + (1-y_i)\log(1-p_\theta)$$

However, ordinary logistic regression is liable to over-fitting and unstable for large number of input variables. In particular, the variance of the estimated coefficients $(\widehat{\boldsymbol{\theta}})$ is often high for correlated input variables. To overcome that, Ridge regularization is often applied.

2.19.1.2 Ridge regularization

Ridge regression applies an $L_2\ norm$ penalty to control the variance of regression coefficients for correlated variables [294-296]. Briefly, Ridge regularization seeks to minimize the ordinary logistic regression cost function $C(y, x, \boldsymbol{\beta})$, subject to $\|\boldsymbol{\beta}\|_2^2 < c^2$ constraint. Specifically, the following cost function is applied to shrink the regression coefficients:

$$\min(C(y, x, \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_2^2)$$

Where λ is a Lagrange multiplier and c is a constant. Geometrically, the $\|\boldsymbol{\beta}\|_2^2 < c^2$ constraint is equivalent to a circle with radius c (**Figure2.4**). However, ridge regression lacks sparsity (no coefficient is set to zero), and therefore not ideal for feature selection. To achieve sparsity, the LASSO penalty is often applied.

2.19.1.3 LASSO regularization

Least Absolute Shrinkage and Selection Operator (LASSO) regression is a popular feature selection algorithm (by Tibshirani, 1996), which applies an $\boldsymbol{L_1}$ norm penalty to shrink the estimated coefficients (some to zero)[297] . Formerly, LASSO seeks to minimize the ordinary regression cost function $C(y, x, \boldsymbol{\beta})$ subject to $\|\boldsymbol{\beta}\|_1 < k$, which reduces to the following form:

$$\min(C(y, x, \boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1)$$

Where λ is a Lagrange multiplier and k is a constant. Geometrically, the $\|\boldsymbol{\beta}\|_1 < k$ constraint is equivalent to a diamond (**Figure2.4**) such that the cost function $C(y, x, \boldsymbol{\beta})$ can only touch the edges thereby forcing some coefficients to zero (hence a sparse model). However, LASSO regression cannot select more variables than the training examples (p≤n); and is not ideal for grouped variable selection because it randomly selects one variable from a group of

correlated variables and ignores the rest. To overcome these limitations, Elastic Net feature selection is often applied.

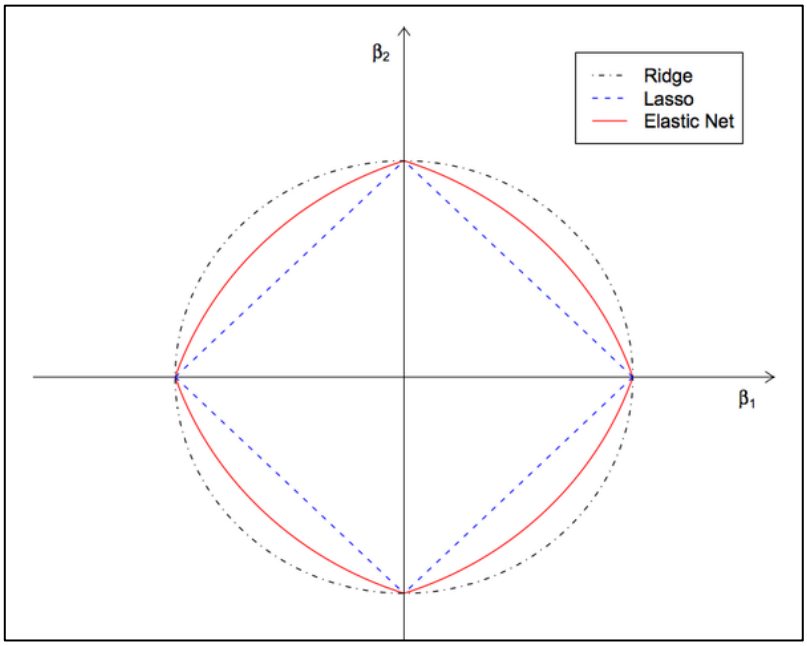


Figure 2.4: A two-dimensional representation of the regularization penalties. Ridge (black circle), LASSO (blue diamond) and Elastic Net (red share). The figure is courtesy of Hastie & Zou (2004) (https://web.stanford.edu/~hastie/TALKS/enet_talk.pdf) [298].

2.19.1.4 Elastic Net regularization

Briefly, Elastic Net (EN) regularization linearly combines the $\boldsymbol{L_2}$ norm (applied in Ridge regularisations) and $\boldsymbol{L_1}$ norm (applied in LASSO regularisations) penalties [299-301]. Notably, this combination of regularization penalties enables the EN algorithm to achieve model sparsity and grouped feature selection beyond the number of training examples (p≥n). Formerly, the EN algorithm seeks to minimize the logistic regression cost function $C(y, x, \boldsymbol{\beta})$ subject to $J(\boldsymbol{\beta}) = \alpha\|\boldsymbol{\beta}\|_1 + (1-\alpha)\|\boldsymbol{\beta}\|_2^2$, where $\alpha \in (0,1)$. Notably, Elastic Net is equivalent to LASSO if $\alpha = 1$ or ridge regularisation if $\alpha = 0$. In particular, the cost function for an Elastic Net regularised logistic regression has the following form where λ is a Lagrange multiplier:

$$\min_{(\beta_0,\boldsymbol{\beta})\in\mathbb{R}^{p+1}} -\left(\frac{1}{n}\sum_{i=1}^{n} y_i(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta}) - \log\left(1 + e^{(\beta_0 + \boldsymbol{x}_i^T\boldsymbol{\beta})}\right)\right) + \left(\lambda\left((1-\alpha)\|\boldsymbol{\beta}\|_2^2\right)/2 \; + \alpha\|\boldsymbol{\beta}\|_1\right)$$

For $\alpha \in (0,1)$, the L_1 part ($\alpha\|\boldsymbol{\beta}\|_1$) generates a sparse model while the quadratic L_2 part ($\frac{(1-\alpha)\|\boldsymbol{\beta}\|_2^2}{2}$) removes the limitation on the number of selected variables, encourages grouping effect, and stabilizes the L_1 regularization path. Therefore, the ElasticNet is a more powerful and flexible hybrid algorithm combining the strengths of the ridge and LASSO regularizations [298].

In this thesis, we applied the Elastic Net feature selection using the R Bioconductor package $Glmnet$ [302]. In particular, leave-one-out cross validation (LOOCV) was applied (using the $cv.glmnet$ function) to estimate the optimal value for λ while fixing $\alpha = 0.8$. While the classification performance was high, an optimal combination of both hyperparameters ($\lambda\ and\ \alpha$) using cross-validation (which requires more computational time) would have achieved more optimal results.

2.19.2 The ROC analysis based Classifier (ROCC)

The ROCC algorithm (by *Lauss et. al (2010)*) is a parameter-free binary classier, which is mainly based on the receiver operation characteristic (ROC) analysis [287]. Firstly, an area under the ROC curve (ROCAUC) filter is applied to select a predefined number of high discriminatory features. However, the feature selection step was not required in this analysis because the classifier features were pre-selected using the elastic net regression. To derive a classification rule, the selected features are collapsed into a univariate $metagene$ using a within-sample arithmetic mean (across the selected features). Finally, the $metagene$ is subjected to ROC analysis to determine an optimal cut-off threshold (associated with the

highest accuracy using the training data) for predicting the class of new samples. To account for platform-specific differences, the $metagene$ is calculated on standardised input features such that the within-sample mean=0 and standard deviation=1. The ROCC algorithm is outlined below.

Let $\{(x_1, y_1), (x_2, y_2), .., (x_n, y_n)\}$ be the training data such that **x** is a standardized p-dimensional input feature matrix and $y_i \in \{0,1\}$ represent the class labels such that:

$$y_i = \begin{cases} 1 & if\ sample\ i\ is\ a\ case \\ 0 & if\ sample\ i\ is\ a\ control \end{cases}$$

Then, the ROCC algorithm works as follows:

1. Decide the optimal number of features (k≤p) to include in the classier (i.e. k=10)
2. Calculate the area under the ROC curve (ROCAUC) for all the *p* features with respect to y.
3. Select the top *k* features with the highest max(AUC, 1-AUC)
4. Among the selected k features, negate (i.e. multiply by -1) all the features that are inversely related with y (i.e. AUC<0.5).
5. Generate a univariate $metagene$ (n-by-1 vector) using the within-sample arithmetic means across the k features.
6. Rank the $metagene$ values in ascending order, and identify the optimal cut-off (i.e. the mean value between two samples) associated with the highest accuracy for predicting y_i in the training data.
7. Apply the cut-off threshold in (6) to predict y_l for new sample l using the $metagene$ value calculated from input vector x_l.

2.19.3 K-nearest neighbours (K-NN) classier

The K-NN is one of the simplest and non-parametric classification algorithm, in which a new sample is classified based on the majority vote of its nearest neighbours [125, 210, 303-305]. To identify the nearest neighbours for new

sample j, distance metrics such as the Euclidian distance $\left((d(x_j, x_i) = \sqrt{\sum_{k=1}^{p}(x_{ki} - x_{kj})^2}\right)$ are applied. To avoid ties, an odd number of neighbours (k=1,3,5,..) is often selected. Alternatively, ties could be resolved by comparing the sums of distances between the classes of the selected k neighbours. In this analysis, we applied the default R package value of k=5. However, this arbitrary choice could generate suboptimal results compared to a cross-validation approach (described in **section 2.20**).

2.19.4 Linear discriminant analysis (LDA)

The LDA classifier seeks an optimal linear combination (similar to PCA analysis) of the input features ($x_i \in \mathbb{R}^p$) such that the training samples are projected into a direction that maximises the separation of the class labels ($y_i \in \{-1,1\}$) [125, 210, 306]. Formerly, the LDA algorithm solves for the optimal values of the weight vector $\boldsymbol{w}$ and threshold b such that

$$\boldsymbol{w}^T\boldsymbol{x} + b > 0 \;\; if \; y_i = 1 \quad AND \quad \boldsymbol{w}^T\boldsymbol{x} + b < 0 \;\; if \; y_i = -1$$

$$\Longrightarrow y_i(w^T\boldsymbol{x} + b) > 0$$

To derive the classification rule, the following Fisher's criterion (J_F) is often maximized with respect to the direction vector **w:**

$$J_F = \frac{|\boldsymbol{w}^T(\boldsymbol{m}_1 - \boldsymbol{m}_2|^2}{\boldsymbol{w}^T S_w \boldsymbol{w}}$$

Where $\boldsymbol{m}_1$ and $\boldsymbol{m_2}$ are the group-specific sample means for y=1 and y=-1 (respectively), and $\boldsymbol{S}_w$ is the pooled covariance matrix given by:

$$S_w = \frac{1}{(n_1 + n_2 - 2)}(n_1\hat{\Sigma}_1 + n_2\hat{\Sigma}_2)$$

$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{1}}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{2}}$ are the maximum likelihood estimates of the group-specific covariance matrices for class1 (y=1, n=n1) and class2 (y=-1, n=n2) respectively. Without loss of generality (i.e. applying the unit proportionality constant), the Fisher's criterion gives the following closed solutions after solving for $\frac{d\boldsymbol{J_F}}{d\boldsymbol{w}} = 0$:

$$\boldsymbol{w} \propto S_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

$$\Longrightarrow \boldsymbol{w} = S_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2)$$

$$\Longrightarrow b = -\frac{1}{2}(\boldsymbol{m}_1 - \boldsymbol{m}_2)^T \boldsymbol{S}_w^{-1}(\boldsymbol{m}_1 - \boldsymbol{m}_2) - \log\left(\frac{p_2)}{p_1)}\right),$$

Where $p_1 = \frac{n_1}{(n_1+n_2)}$ and $p_2 = \frac{n_2}{(n_1+n_2)}$ are the proportions of samples in class 1 and class 2 respectively. However, it should be noted that the Fisher's criterion is optimal if the input features follow the Gaussian distribution and the class-specific covariance matrices are similar (Webb, 2002, p:127-129) [210].

2.19.5 Random forest

2.19.5.1 Introduction

Random Forest is an ensemble of many classification or regression trees (CART) that are grown (trained) on bootstrap samples using random subsets of the input features [284]. Briefly, ensemble classifiers seek to improve the synergetic performance of weak classifiers [307]. Here, we focus on the random forest for classification trees.

2.19.5.2 Classification trees

Classification trees are very intuitive classifiers and can be applied to almost any type of data scale [210]. As illustrated in **Figure2.5**, a classification tree involves a sequence of binary splits of the training data from the **root node** through the **internal nodes (blue box)** to the **leaf nodes**, where samples

are finally classified [308, 309]. In particular, (i) a root node has outgoing arrows only (grey box), (i) internal nodes have both incoming and outgoing arrows (blue box), and (iii) leaf nodes have incoming arrows only (gold boxes).

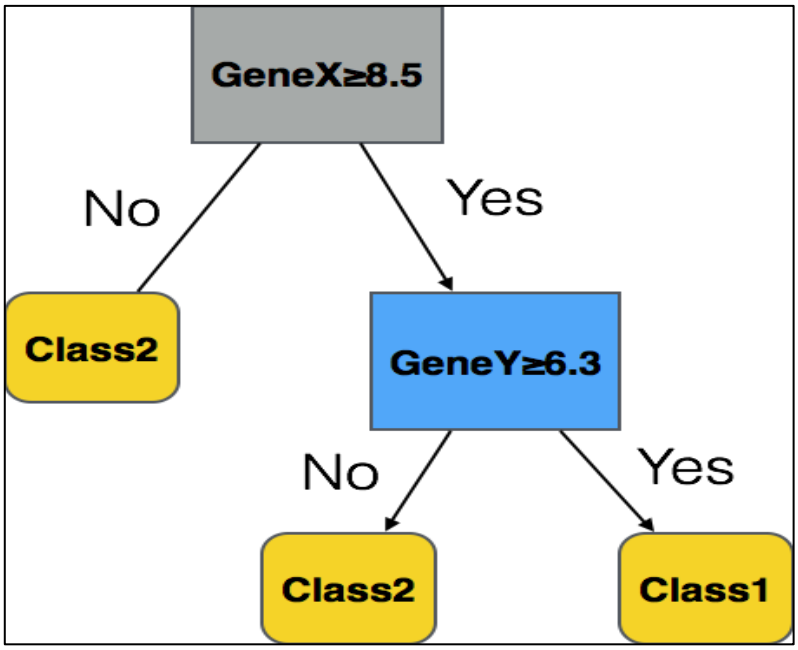


Figure 2.5: A classification tree diagram. The figure shows an illustration of a classification tree algorithm for binary outcomes (Class1 and Class2) [308, 309].

At each node, the best gene feature (and the optimal cut-off value) is applied to split the data into more homogenous groups called child nodes. If no better split is attainable, the current node is maintained as a leaf node. While various metrics including $information\ gain$ and $variance\ reduction$ exist, the $Gini\ impurity\ index$ is often applied to select the optimal split of the training data into child nodes [284, 308, 310]. In each child node c, the Gini index is calculated as follows:

$$Gini(c) = 1 - \sum_{k=1}^{2} p_k^2$$

Where p_k is the proportion of samples in class $k = 1,2$ within that child node such that $p_1 + p_2 = 1$. Notably, the $Gini(c)$ measures the misclassification

rate of samples in each child node. To get the overall $Gini\ impurity\ index$ associated with each split, the weighted (proportional to the size of child node c) $Gini(c)$ values are added.

$$Gini_{(split)} = \sum_{c=1}^{2} p_c * Gini(c)$$

Where p_c =the proportion of samples in child node $c = 1,2$ with respect to the parent node such that $p_1 + p_2 = 1$, and $Gini(c)$ is the corresponding $Gini\ index$. However, individual classification trees are weak classifiers that liable to over-fitting (high variance) and lack robustness [210, 284].

2.19.5.3 Random forest algorithm

To improve the performance (variance, accuracy and robustness) of individual classification trees, the random forest algorithm (as illustrated in **Figure2.6**) applies an ensemble technique called BAGGING (**B**ootstrap and **AGG**regat**ING**)[311-313]. In particular, the decision rule is based on the majority vote of many classification trees that are trained using bootstrap samples from the training data. Notably, the bootstrapping approach enables the internal validation of the random forest classier using an average of Out-of-Bag (OOB) errors estimated by predicting the class of training examples that are not included in the bootstrap sample of a particular tree [314].

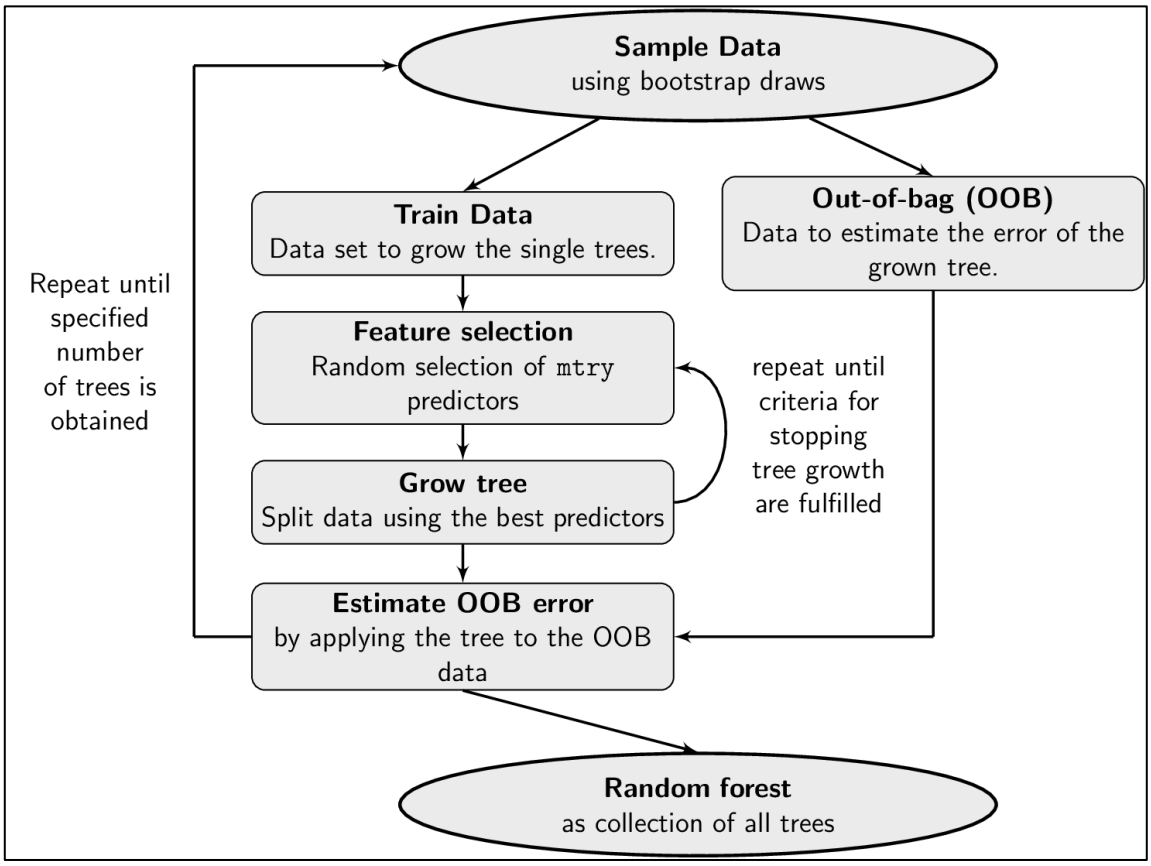


Figure 2.6: An illustration of a Random Forest algorithm. The figure was taken from Boulesteix et. al (2011)[314].

Formerly, Let $\boldsymbol{D_{n\times p}} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x_2}, y_2), \ldots\ldots (\boldsymbol{x}_n, y_n)\}$ be the training data set, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is a p-dimensional vector of gene features in sample i and $y_i \in \{-1, 1\}$ is the corresponding class label. Then, the $RandomForest$ involves the following main steps:

1. Draw B bootstrap samples (with replacement) from the original input data **D**.
2. Build a tree classifier T_b using each $\boldsymbol{D}^{(b)}_{\boldsymbol{n\times p}}$ bootstrap sample (b=1, 2, .B).
3. Assign new sample $\boldsymbol{x}_k$ to the class y_k based on majority vote of the B classification trees (T_b).

To build the random forest classifier, this thesis applied the $RandomForest$ R package [315, 316] using the $Gini\ impurity\ index$ and default hyperparameter (i.e. number of trees and number of features per split (mtry))

values. While the classification performance was high, cross-validated $hyperparameters$ (described in **section 2.20**) would have generated more optimal results.

2.19.6 Support vector machines (svm)

2.19.6.1 Introduction

Support vector machine (SVM) is the most powerful and successful linear classification algorithm based on the idea of margin and kernel tricks[283]. Briefly, SVM seeks the best separating line ($hyperplane$) with the "widest margin" between the classes of the training data. While a hard margin (**Figure2.7a**) is sufficient for linearly separable classification problems, $soft\ margins$ (**Figure2.7b**) and $kernel\ tricks$ are often applied in more complex nonlinear situations [317-319]. Here, the hard-margin SVM (linearly separable data) is introduced followed by an extension to the nonlinear situations.

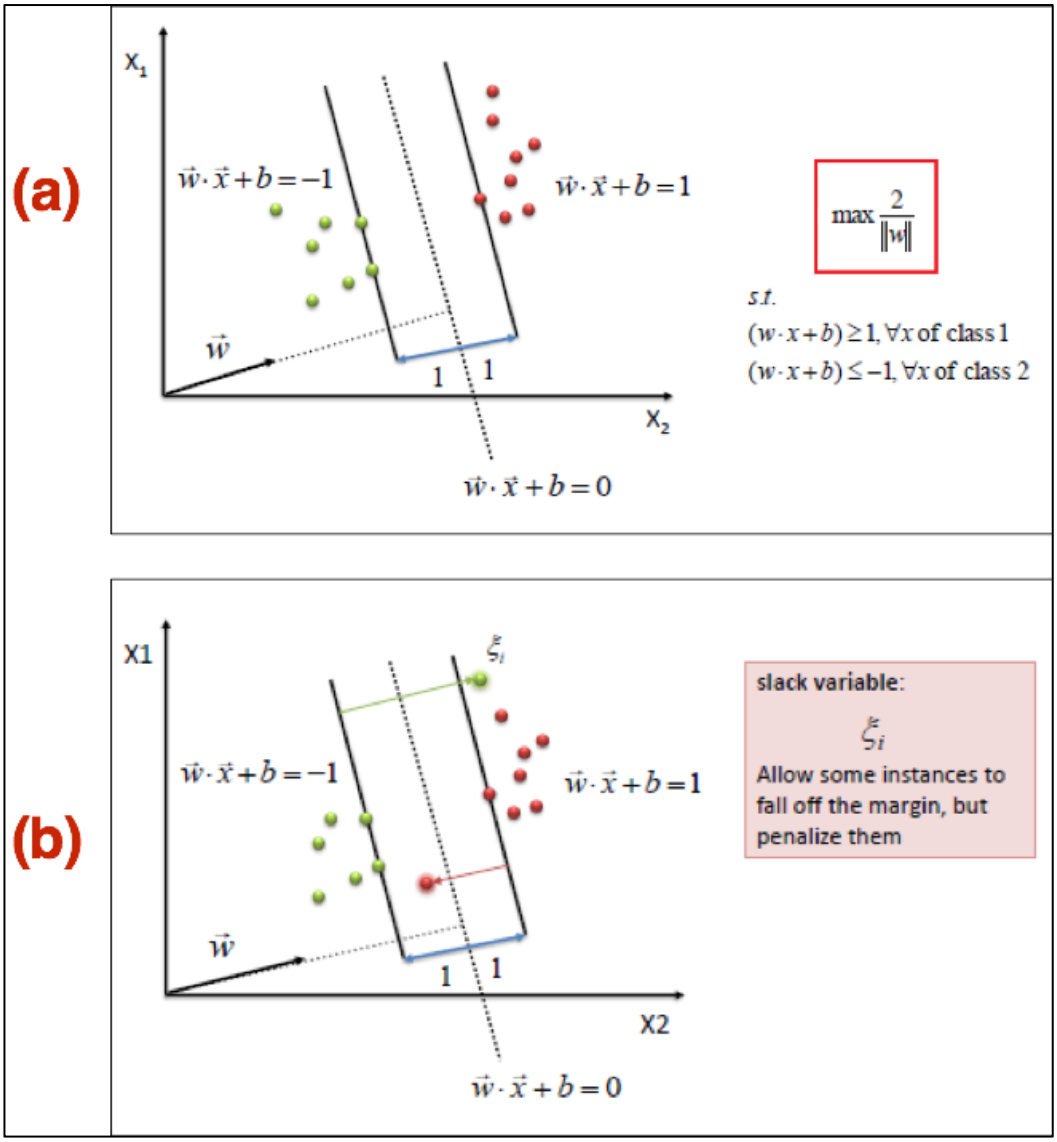


Figure 2.7: A two-dimensional illustration of a support vector machines (SVM) classifier : (a) hard margin SVM, (b) soft margin SVM where x1 and x2 are the classification features. The two classes are represented by red (y=+1) and green (y=-1) dots. The images are courtesy of Dr. Saed Sayad(2018)[320] (https://www.saedsayad.com/support_vector_machine.htm)

2.19.6.2 The hard margin SVM

Let $\boldsymbol{D} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x_2}, y_2), \ldots\ldots (\boldsymbol{x}_n, y_n)\}$ be the training data set, where $\boldsymbol{x}_i \in \mathbb{R}^p$ is a p-dimensional vector of gene features in sample i and $y_i \in \{-1, 1\}$ is the corresponding class label. To classify a new sample k, SVM applies the following decision criteria:

$$y_k = \begin{cases} +1 \; if \; \boldsymbol{w}^T\boldsymbol{x} + b \geq 1 \\ -1 \; if \; \boldsymbol{w}^T\boldsymbol{x} + b \leq 1 \end{cases} \Longrightarrow y_k(\boldsymbol{w}^T\boldsymbol{x} + b) \geq 1$$

Here, $\boldsymbol{w}^T\boldsymbol{x} + b = 0$ is the best separating line (hyperplane) between the classes, where $(\boldsymbol{w}^T\boldsymbol{x} + b \geq 1)$ and $(\boldsymbol{w}^T\boldsymbol{x} + b \geq 1)$ are lower and upper boundaries of its margin (respectively). Notably, the training examples touching (i.e. supporting) the margin (i.e. $\boldsymbol{w}^T\boldsymbol{x}^- + b = -1$ or $\boldsymbol{w}^T\boldsymbol{x}^+ + b = 1$) are called the "$support\ vectors$". Therefore, the margin (M) is the shortest distance between two support vectors on the lower (x^-) and (x^+) upper boundary of the hyperplane ($\boldsymbol{w}^T\boldsymbol{x} + b = 0$) such that:

$$M = d(x^+, x^-) = \|x^+ - x^-\| = \|(\boldsymbol{w}^T x^+ + b) - (\boldsymbol{w}^T x^- + b)\| = 2$$

$$\Longrightarrow \|\boldsymbol{w}^T(x^+ - x^-)\| = 2$$

$$\Longrightarrow \|\boldsymbol{w}^T\|\|(x^+ - x^-)\| = 2$$

$$\Longrightarrow \|(x^+ - x^-)\| = \frac{2}{\|\boldsymbol{w}\|}$$

$$\Longrightarrow\ M = \frac{2}{\|\boldsymbol{w}\|}$$

Therefore, the SVM algorithm seeks to maximise the margin $\frac{2}{\|\boldsymbol{w}\|}$ while correctly classifying the samples (i.e. $y_i(\boldsymbol{w}^T\boldsymbol{x} + b) \geq 1$). In practice, SVM minimises the following constrained quadratic problem:

$$\arg min\ \left(\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}\right)\ \ subject\ to\ \ \ \ y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1$$

$$\Longrightarrow L(\boldsymbol{w}, b, \propto) = \frac{1}{2}\boldsymbol{w}^T w - \sum_{i=1}^{n} \propto_i (y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) - 1)$$

Where $y_i \in \{-1,1\}, i = 1,.2,..n$ are the class labels, $\alpha_i \geq 0$ are the Lagrange multipliers according to the *Karush–Kuhn–Tucker (KKT) conditions [210, 283, 321].* Therefore differentiating $L(\boldsymbol{w}, b, \propto)$ with respect to $\boldsymbol{w}$ gives

$$\boldsymbol{w} = \sum_{i=1}^{n} \propto_i y_i x_i$$

$$\Longrightarrow L(\propto) = \sum_{i=1}^{n} \propto_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \propto_i \propto_j x_i^T x_j$$

Where $i \neq j$ are the pairs of the training examples such that $\sum_{i=1}^{n} \propto_i y_i = 0$ and $\propto_i (y_i(\boldsymbol{w}^T \boldsymbol{x_i} + b) - 1) = 0$ $(KKT\ condition)$. Thus, $\propto_i \approx 0$ for all the non-support vector examples $(y_i(\boldsymbol{w}^T \boldsymbol{x_i} + b) \neq 1)$ such that $\boldsymbol{w}$ is efficiently estimated using the support vectors (SV) only.

$$\boldsymbol{w} = \sum_{i \in SV} \propto_i y_i x_i$$

Where SV is the set of support vectors (SV). In particular, the $hard\ margin$ SVM applies the following algorithm:

1. Minimise $L(\propto) = \sum_{i=1}^{n} \propto_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \propto_i \propto_j x_i^T x_j$ with respect to **α** subject to $\propto_i \geq 0$ and $\sum_{i=1}^{n} \propto_i y_i = 0$
2. Solve for $\boldsymbol{w} = \sum_{i \in SV} \propto_i y_i x_i$ using the $\propto_i \geq 0$ values for the support vectors (SV)
3. Solve for b using $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1$ for any support vector i
4. Predict the class of new sample h as follows: $y_h = \begin{cases} +1 & if\ \boldsymbol{w}^T \boldsymbol{x_h} + b > 0 \\ -1 & if\ \boldsymbol{w}^T \boldsymbol{x_h} + b < 0 \end{cases}$

2.19.6.3 Soft margin extension

The $hard\ margin$ approach assumes that the training examples are linearly separable by one unit away $(y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1)$ from the best separating hyperplane $(\boldsymbol{w}^T \boldsymbol{x}_i + b = 0)$. However, this rigid approach is liable to over-fitting due to outliers and nonlinearity. Notably, the generalization of SVM classifiers depend on the number of support vectors $(y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1)$ such that

$$E(\epsilon_{out}) = \frac{E(number\ of\ support\ vectors)}{n - 1}$$

Where $\epsilon_{out} = \text{out of sample error}$and E(.) denotes the expected value [322, 323]. To accommodate the errors and maintain a wider margin, the $soft\ margin$ approach introduces a slack parameter ξ thereby allowing some training samples to cross the margin boundaries (i.e. misclassified). Formerly, the soft margin SVM solves for

$$\arg\min\left(\frac{1}{2}\boldsymbol{w}^T w + C\sum_{i=1}^{n}\xi_i\ \right) subject\ to\quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i$$

Here, C is the regularization parameter capturing the importance of the slack parameter ξ_i with respect to the margin $\left(\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}\right)$ [210, 283]. Notably, the objective function reduces to a $hard\ margin$ SVM if ξ_i=0, and the penalty increases for large values of ξ_i [324]. However, the cost function $\boldsymbol{L}(\propto)$ remains unchanged but $\propto_i$ is upper-bounded by the regularization parameter C such that SVM seeks to minimize

$$L(\propto) = \sum_{i=1}^{n} \propto_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \propto_i \propto_j x_i^T x_j$$

with respect to **α** subject to $0 \leq \propto_i \leq C$ and $\sum_{i=1}^{n} \propto_i y_i = 0$ for all i=1, 2, ..n.

2.19.6.4 The kernel trick

The $kernel\ trick$ enables the SVM to learn nonlinear problems with linear machinery. Briefly, a kernel function is applied to map the training data (**x**) into an infinitely high dimensional space (**Z**) where the data is potentially linearly separable ($\mathbf{x} \rightarrow \mathbf{Z} \in \mathbb{R}^{\infty}$) [324-327]. For a kernel function Z, the $soft\ margin$ SVM seeks to minimise:

$$L(\propto) = \sum_{i=1}^{n} \propto_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \propto_i \propto_j z_i^T z_j$$

With respect to $\propto$ subject to $0 \leq \propto_i \leq C$ and $\sum_{i=1}^{n} \propto_i y_i = 0$ for all i=1, 2, ..n where C is a regularisation parameter. Notably, the $kernel\ trick$ mainly

involves calculating the dot product $z_i^T z_j$ without explicit mapping into the high-dimensional feature space (Z). Further, the complexity of the kernel transformation is not directly related to over-fitting because the error rate depends on the number of support vectors. Therefore, the SVM classifier has the computational feasibility to learn non-linear problems in infinitely high-dimensional space while using the machinery of linear algorithms [324, 325].

According to Mercer's theorem, a kernel function K(x,y) is a continuous function involving a scalar (dot) product of input features in a particular feature space such that the following conditions are satisfied [210, 324, 325]

(i) $K(x_i, x_j) = K(x_i, x_j)$ (Symmetry)

(ii) $\sum_{i=1}^{n}\sum_{j=1}^{n} K(x_i x_j)c_i c_j \geq 0$ (Positive semi-definite).

In particular, the following kernel functions are often applied in SVM classification [328]:

- **Linear**: $\mathrm{K}_{Lin}(\mathrm{x},\mathrm{y}) = \mathrm{x}^T\mathrm{y}$
- **Polynomial**: $\mathrm{K}_{Poly}(\mathrm{x},\mathrm{y}) = (\mathrm{x}^T\mathrm{y}+1)^D \; where \; D \in \mathbb{R}$
- **Gaussian Radial basis function (RBF)**:

$$\mathrm{K}_{RBF}(\mathbf{x},\mathbf{y}) = \exp\left(\frac{-\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) \Leftrightarrow \exp(-\gamma\|\mathrm{x}-\mathrm{y}\|^2), where\ \gamma > 0.$$

In this thesis, we applied the soft-margin SVM with the radial basis function (RBF) kernel using the **e1071** R package [329]. In particular, the RBF is a general kernel function without assuming prior knowledge about the data [328]. It is worth noting that here we applied the default value (in R) for the regularization parameter C. While the classification performance was high, a

cross-validated value (described in **section 2.20**) for C [322] would produce optimal results.

2.20 Internal validation of classifiers

A classier is a function that maps unlabelled instances to a phenotypic class label using internal data structures [330]. To select the best classifier, internal validation is required to estimate the out-of-bag (OOB) error (or accuracy). These include Holdout, Bootstrapping and K-folds cross-validation [330-332]. In this thesis, we applied leave-one-out cross-validation (LOOCV), a special case for the K-folds cross-validation where k=n (i.e. training data sample size).

2.20.1 Cross-validation

Briefly, Cross-validation involves splitting the training data into K mutually exclusive folds, where each k^{th} fold is applied (once) as a testing data set for the model trained on the other k-1 folds combined [330-333]. Notably, Cross-validation generates a distribution of estimates, which is desirable to estimate model variance and robustness [330, 332, 333]. Further, it is associated with less bias than the Bootstrap validation approach [330].

2.20.2 Nested cross validation

While the classification performance was high (**Chapter 6**), it is worth noting that this thesis applied default values for the hyperparameters (**Table 2.8** and **Table 6.1**) potentially generating suboptimal results. To get optimal results, nested cross-validation is recommended [332, 333]. Briefly, a two-level nested cross-validation incorporates an inner cross-validation loop (using the K-1 training folds) to identify the best combination of hyperparameters for the

k^{th} iteration of the model selection cross-validation loop [331, 332]. The algorithm is outlined below:

2.20.2.1 Nested cross validation algorithm

1. Split the data into K folds
2. For each k=1,2,…,K {
3. Keep the k^{th} fold for model testing
4. Further split the data for the remaining K-1 folds (combined) into J folds
5. For each j=1,2, ..J {
    a. Keep the data for the j^{th} fold for model testing
    b. Use the J-1 folds data to train the model using each possible combination of the hyperparameters
    c. Assess the performance of model hyperparameters using the unused data in fold j.

    }
6. Choose the model with the best combination of hyperparameters across all the J folds.
7. Apply the unused data in the k^{th} fold (1) to assess the performance of the classifier trained in (6)

    }
8. Choose the best model across all the K folds in (1).

Chapter 3: Data characteristics and quality assurance

3.1 Introduction

The aim of this chapter is to evaluate (characteristics and quality) and curate existing data resources for their use in subsequent chapters of this thesis.

3.2 Background

Advances in genome-wide profiling such as whole blood transcriptomics have spurred biomedical research for elucidating the pathogenesis, biomarkers and therapeutic targets for a wide range of diseases including cancer, infections and autoimmunity; and they present an innovative approach for future translation of personalized medicine [70, 123, 334]. In this thesis, I have analyzed a microarray-based whole blood transcriptome (and the corresponding demographic, clinical and microbiology databases) for Gambian children aged 2-59 months to gain a deeper insight into the systemic pathway responses to severe pneumonia. However, making meaningful inferences from high-throughput data has several challenges including confounding non-biological variations, high dimensionality, limited study design, inadequate sample sizes and limited phenotypic data[71]. Therefore, data quality assurance is mandatory[181]. To assess the validity and quality of the central data resources for this thesis, this chapter has applied a range of statistical quality control approaches on the microarray database and the corresponding metadata records. These findings will highlight the strengths and limitations of the available data resources, and will thereby guide subsequent analyses for addressing the primary objectives of this thesis.

To evaluate the characteristics and ensure the quality of the existing data resources for this thesis; this chapter has addressed the following objectives:

1) To assess the characteristics and validity of the existing data resources
2) To assess the data quality and limitations (i.e. adequacy, completeness, imbalances) in accordance with the primary objectives
3) To enhance data quality (i.e. data cleaning, pre-processing) and identify key covariates (potential confounders) for subsequent analyses

3.3 Results

3.1.1 Demographic and clinical characteristics of study participants

In this study, eligible participants (children aged 2-59 months old) were recruited from two geographical regions in the Gambia, West Africa. The training data were collected from the semi-urban coastal area (here called **Fajara**), and the validation sample was collected from the rural upper region called **Basse**. In total, 1527 children who were clinically classified as mild, severe and very severe pneumonia, and their prospectively matched (by age, sex, location) non-pneumonia community controls were recruited and bled. Of them, sufficient whole blood RNA samples (n=803) were isolated for transcriptomics analysis in Edinburgh. After laboratory quality control analysis, 518 RNA samples were subjected to the microarray assay. However, the final database reduced to n=503 after data cleaning (i.e. 15 outliers were excluded). Of them, 69%(n=345) and 31%(n=158) represent the training and validation populations, respectively. It is worth noting that the

validation dataset (**Basse**, n=158) was kept independent from all the primary analyses for validation of candidate biomarkers of severe pneumonia in **Chapter6**.

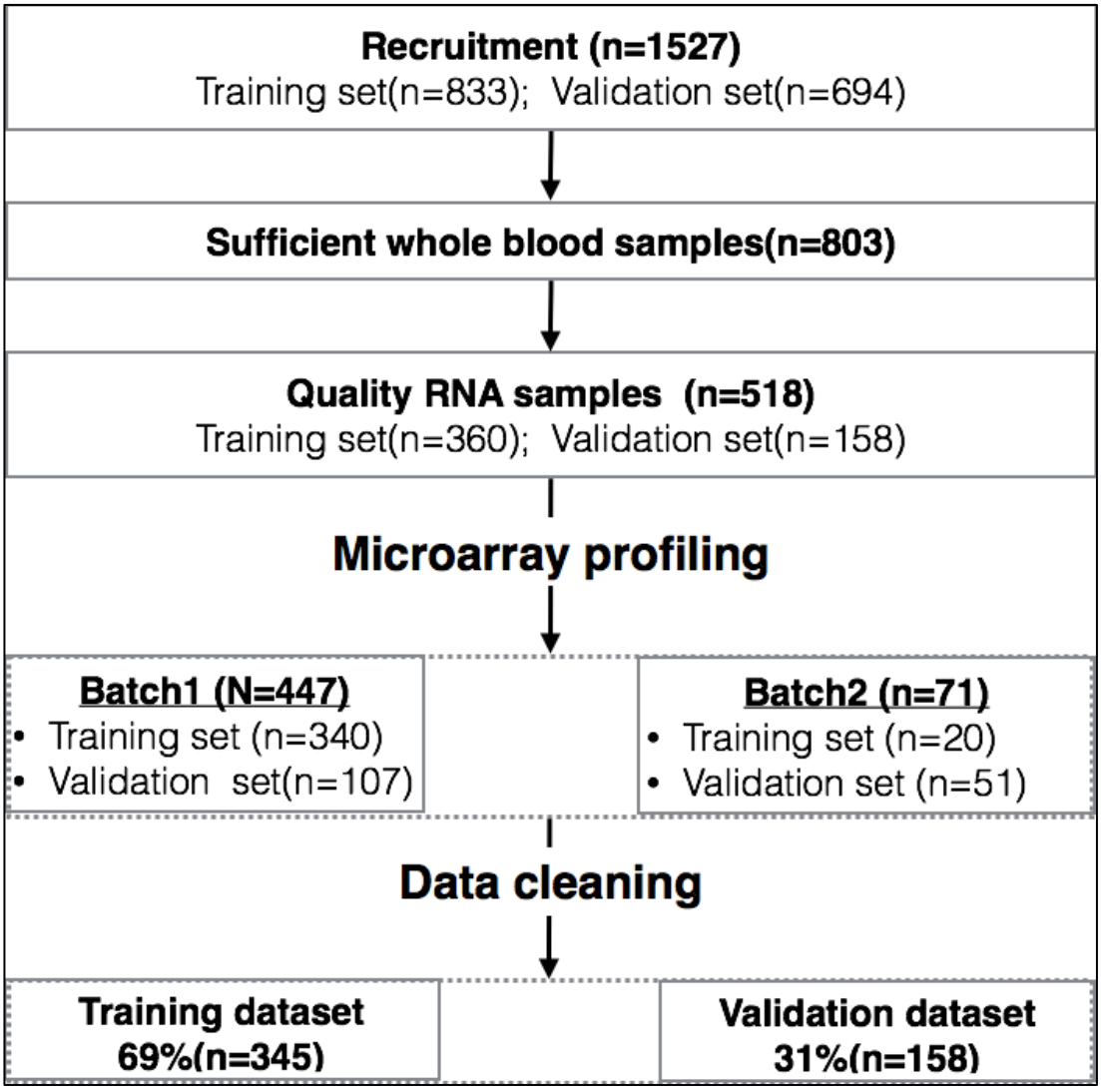


Figure 3.1: Sample recruitment and processing

As shown **Table3.1**, the group sample sizes ranged between n=18 (very severe) and n=120 (non-pneumonia controls) in the training data (where most primary analyses will be conducted); and between n=24(very severe) and n=47 (non-pneumonia controls) in the validation dataset. While both datasets have more samples in the first Batch, the second Batch had more validation samples (n=51) than the training population (n=17). Further, while

gender and seasonality were similar between the training and validation datasets, it is worth noting that the validation set (the rural sample) was relatively younger and associated with worse clinical outcomes than in the training set. In particular, the rural population was associated with the higher prevalence of severe pneumonia outcomes, malnutrition (stunting and under-weight) and iron intake than the training set; and the vice versa for vitamin A supplementation ($p<0.05$, respectively). Potentially, this data imbalance may impact the performance of candidate biomarkers in subsequent analysis (**Chapter6**). On the other hand, this variability will partially enable to assess the robustness of the candidate biomarkers prior to subsequent validations (**Figure3.1**). To ensure data quality and completeness, the metadata records were subjected to data cleaning, and the prevalence of missing data was investigated (next section).

Factor	Basse (Validation set)	Fajara (Training set)	P-value
Total	**158**	**345**	
Pneumonia severity			**<0.001**
Control	47 (29.7%)	120 (34.8%)	
Mild	46 (29.1%)	90 (26.1%)	
Severe	41 (25.9%)	117 (33.9%)	
Very Severe	24 (15.2%)	18 (5.2%)	
Microarray sample batches			**<0.001**
Batch1	107 (67.7%)	328 (95.1%)	
Batch2	51 (32.3%)	17 (4.9%)	
Demographics			
Age in months			
Median (IQR)	11.0 (5.3, 23.3)	14.2 (7.9, 22.6)	0.034
Age groups (months)			**<0.001**
2-5	46 (29.1%)	54 (15.7%)	
6-11	36 (22.8%)	83 (24.1%)	
12-23	38 (24.1%)	134 (38.8%)	
24-59	38 (24.1%)	74 (21.4%)	
Gender			**0.49**
Female	73 (46.2%)	148 (42.9%)	
Male	85 (53.8%)	197 (57.1%)	
Season			0.065
Dry	68 (43.0%)	179 (51.9%)	
Wet	90 (57.0%)	166 (48.1%)	
Nutrition status			
Under weight (Weight-for-Age)			
WAZ score, mean (SD)	-1.4 (1.3)	-1.1 (1.3)	0.062
Moderate underweight (WAZ<-2)	45 (28.7%)	90 (26.1%)	0.55
Severe underweight (WAZ<-3)	19 (12.1%)	19 (5.5%)	0.010
Stunting (Height-for-Age)			
HAZ score, mean (SD)	-1.0 (2.4)	-0.7 (1.4)	0.027
Moderate (HAZ<-2)	27 (17.3%)	44 (12.9%)	0.19
Severe (HAZ<-3)	14 (9.0%)	10 (2.9%)	0.003
Wasting (Weight-for-Height)			
WHZ score, mean (SD)	-1.2 (1.3)	-1.1 (1.4)	0.38
Moderate (WHZ<-2)	37 (23.9%)	69 (20.2%)	0.35
Severe (WAZ<-3)	8 (5.2%)	26 (7.6%)	0.32
Clinical values			
Iron supplementation	32 (53.3%)	28 (27.5%)	<0.001
Vitamin A (within 6 months)	51 (34.0%)	170 (52.5%)	<0.001
Antibiotic usage (within 2 weeks)	9 (6.8%)	57 (17.1%)	0.004

Table 3.1: Demographic and clinical characteristics of the participants. Reported P-values are comparing the distributions of the training and validations data sets using the Chi-square / Fisher's exact test (categorical variables) or student t-test/Mann-Whitney test (continuous variables). WAZ=weight-for-Age Z score, WHZ=weight –for-height Z score HAZ=height-for-age Z score

3.1.2 Data completeness (missing data)

Missing data cause serious problems in statistical analyses, which undermine the statistical power to detect study effects (i.e. elimination of samples with missing values) and generate biased results (i.e. missing data correlating with study outcomes)[335]. It is worth noting that the microarray database had no missing values. To ensure the quality and completeness of the metadata records, relevant variables were extracted and screened for missing or suspicious values, and validated using the reference database (**SCC1062**), which is securely stored at the Medical Research Council (MRC) unit, The Gambia.

As shown in **Figure 3.2**, the prevalence of missing values (yellow colour) was very minimal especially among the key variables for the primary analyses of this thesis. Comparatively, the prevalence of missing data (i.e cell blood counts) is lower in the training data where more primary analyses are required. Together, these findings highlight the quality of the existing central resources for this thesis. In the next section, I investigated the quality of the microarray database.

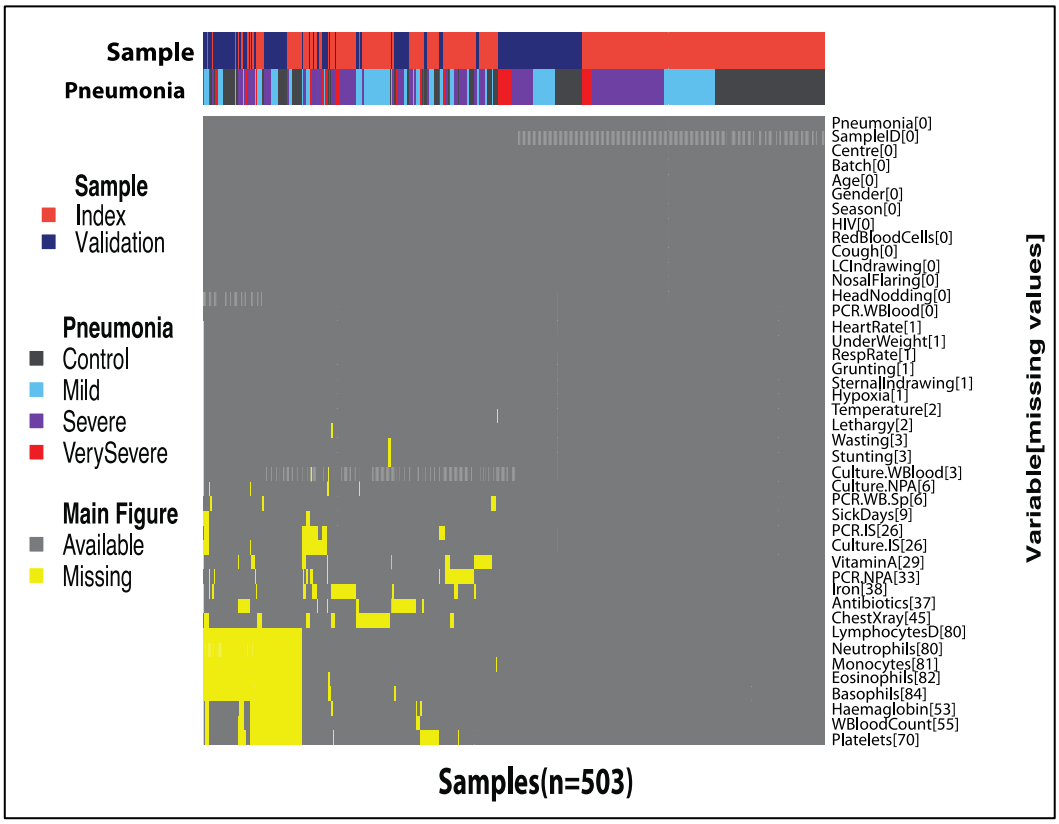


Figure 3.2: The heatmap showing the unsupervised clustering of samples base on the distribution of missing values. The main figure shows missing (yellow colour) and observed (grey colour) values data in the demographic and clinical valuables (x-axis). The samples (columns) are annotated by site (training or validation sample) and pneumonia severity (legend).

3.1.3 Quality assurance of the microarray database

The quality of microarray data depends on several factors at different stages of the study [129]. Therefore data quality assurance is an integral part of the main analysis to eliminate the confounding non-biological variations and mitigate the challenge of multiple testing [151, 153, 336]. To ensure the quality of the microarray transcriptome database, I applied a range of statistical methods to remove the unwanted variations (i.e. hybridization and batch-effect variations), outliers and non-informative gene probes. Here, the training and validation data sets were analysed separately. Prior to that, sample size analysis was done to re-assess the adequacy (i.e. statistical power) of the existing resources (next section).

3.1.3.1 Re-assessment of sample sizes and statistical power

Meaningful statistical inference requires adequately-powered studies, and sample size analyses (or power calculations) are vital [337]. For comparative studies (like this thesis), sample size analysis estimates the minimum numbers of required samples (per group) that are statistically powered to detect meaningful effect sizes (i.e. mean difference, rate or ratio). In particular, the following input parameters are required: (i) desired statistical power (type II error), (ii) significance level (type I error), (iii) minimum effect size (i.e. mean difference, ratio or rate) and (iv) variability estimates for each study group. Mathematically, sample size estimates are positively related to the statistical power and population variability, and the vice versa for effect size and significance level. While investigators decide reasonable effect size, statistical power and significance level, estimation of population variability is more challenging and often relies on previous studies or pilot data.

For the multidimensional transcriptomic data, sample size and power analysis further account for false discoveries due to multiple testing [167, 169, 338]. In particular, sample size estimates are statistically powered to detect multiple-testing-adjusted meaningful effect sizes (i.e. two-fold changes in gene expressions between two groups) in the desired percentage (i.e. 90%) of the gene probes on the array platform. Thus, variability estimates are required for each gene probe.

In this thesis, the original study design for the microarray database (Gambian children) was powered using variability estimates from the Neonatal Study (n=56) conducted at the Royal Infirmary of Edinburgh (United Kingdom)

[167]. Approximately, 100 samples (per group) were 90% statistically powered to detect a two-fold change in differential expression in at least 90% of the gene probes on the array at a significance level of *alpha = 0.001* (corrected for multiple testing by the Bonferroni method [170]). However, these estimates were based on the variability estimates from a different population, an older profiling technology (potentially with more technical variable) and a very stringent approach for multiple testing corrections (Bonferroni correction [170]).

To guide subsequent analyses, this analysis re-assessed the statistical power of the training database (where primary analyses will be conducted) using variability estimates from the same population and less stringent parameters applicable to this thesis (**Figure3.3**). In particular, the following question was addressed: *How many samples are statistically powered (90%) to detect at least 2-fold change in at least 90% of the gene probes in microarray database while controlling for false discovery rate (FDR) at 5%?* To estimate the population variability, I applied the whole blood transcriptome for Gambian children who participated as healthy controls (n=20) in the Trachoma study (**GSE29463**) by *Natividad et al.(2010)* [168]. In this analysis, constant variability was assumed between the study groups.

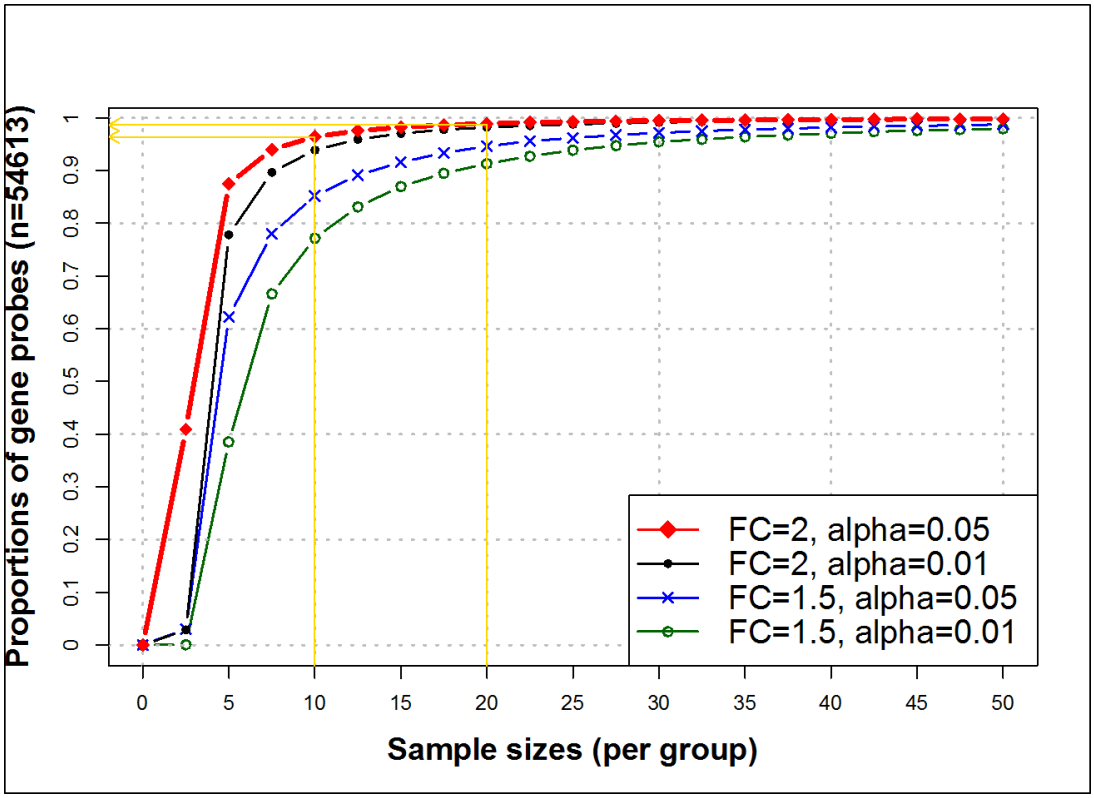


Figure 3.3: Sample size estimates for the microarray database. The curved lines indicate the number of minimum required samples (x-axis) statistically powered (90%) to detect fold changes (legend) in a particular proportion of the gene probes (y-axis). Variability was estimated from healthy controls in the GSE29463 study [168]. FC=fold change (i.e. size effect), alpha= False discovery rate (FDR) adjusted Type-I error (α).

In **Figure3.3**, the sample size estimates for fold-change=2 and 1.5, and FDR=0.01 and 0.05 are presented. According to the current analysis, at least 10 samples are sufficient (90% statistically powered) to detect at least 2-fold change in gene expression of at least 96% of the gene probes on the microarray; and the proportion of detectable gene probes increases to 98% with 20 samples (**Figure3.3, red curve**). Similarly, 20 samples are sufficient to detect more stringent effects (fold- change=1.5, FDR=0.01) in at 90% of the gene probes on the array.

In the training data (where most primary analyses were conducted), the sample sizes ranged between n=18 (very severe pneumonia) and n=120

(non-pneumonia controls) cases (**Table3.1**). According to the current sample size analysis (**Figure3.3)**, 18 samples are statistically powered (90%) to detect at least 1.5-fold change or 2-fold change (FDR<0.05) in at least 90% or 97% (respectively) of the gene probes on the microarray. Together, this analysis suggests that the available database resources are adequately-powered to address the primary objectives of this thesis. In the next section, technical variations in the raw databases were investigated.

3.1.3.2 Pre-processing of raw data

3.1.3.2.1 Why raw data pre-processing is required?

Microarray raw data points are not informative due to non-biological (i.e hybridization) variations across the array [129]. As illustrated in **Figure3.4**, the variability of raw data expression profiles is higher even within the same study group (very severe pneumonia samples in the training sample). To normalise the unwanted variations, raw data pre-processing [147, 339] is often mandatory (next section).

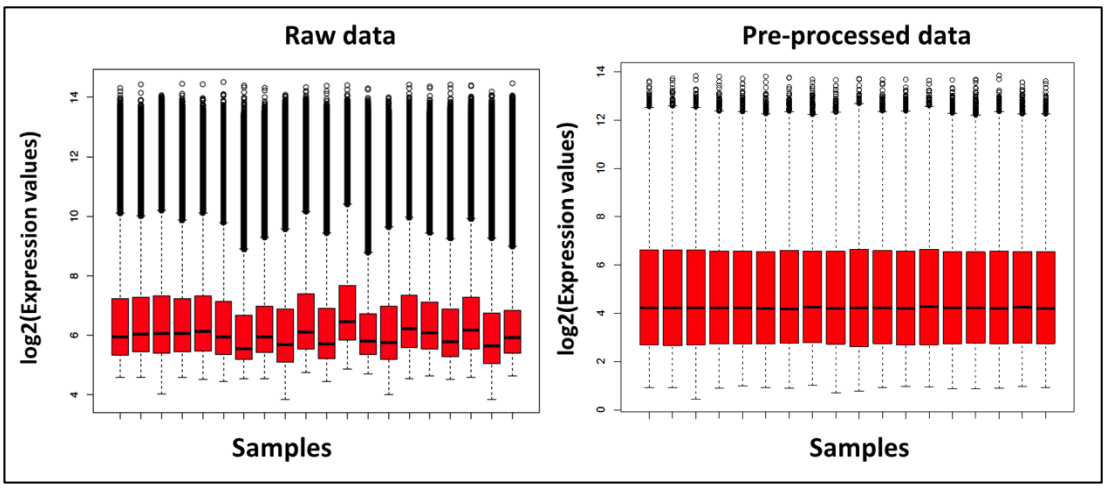


Figure 3.4: An illustration of sample variability before and after data pre-processing. This plot is based on gene expression signals for samples with very severe pneumonia in the training database (n=20).

3.1.3.2.2 Selection of appropriate algorithm for raw data pre-processing

Raw microarray data points are confounded by technical variations, which

require *pre-processing* (background correction, normalisation, transformation and summarization of probe-specific data) [182]. While many statistical algorithms exist, careful selection of an appropriate algorithm is equally important. Notably, many algorithms require expression data for mismatch probes, which are not included in the design of the microarray platform applied in this thesis (Affymetrix HGU219)[166].

To select the most appropriate pre-processing method, two widely applied algorithms that do not require mismatch probes were compared graphically (**Figure3.5**): (i) **RMA**[1] by Irizarry, et al.(2003)[147] and **VSNRMA**[2] by Huber, et al.(2002)[146]. While both methods apply *median polish summarization*, the key steps (background, normalisation and transformation) are different. Moreover, *Irizarry et al. (2006)* noted that the accuracy/precision (bias/variance) trade-off is driven mostly by background correction[181].

In all the three graphical methods, good performance is measured by the stability of the curves. Using these criteria, a better pre-processing algorithm should have:

(i) Constant variance at different mean values (**Figure3.5a**),

(ii) No correlation (i.e. r≈0) between pairs of randomly selected gene features (i.e. probe sets represeting genes from functionally independent pathways) regardless of their variance (**Figure3.5b**),

(iii) Lower absolute rank deviation (ARD) (**Figure3.5b**). Briefly, ARD is

[1] **RMA**=Robust Multi-array Average
[2] **VSNRMA**=Variance stabilisation normalisation (VSN) with median polish summarization as in the RMA method

the between-sample standard deviation (SD) for the gene expression values sharing the same rank across gene probes[175].

As shown in **Figure3.5**, the RMA algorithm was consistently associated with more stable results than the VSNRMA algorithm. In particular, while the distributions of variance against mean (**Figure3.5a**) were similar, the RMA method was associated with more stable correlation and ARD values(**Figure3.5b-c**). Similar results were observed in the validation data set, and the **RMA** algorithm was selected. In the next section, I investigated the presence of batch-effect variations beyond raw data pre-processing.

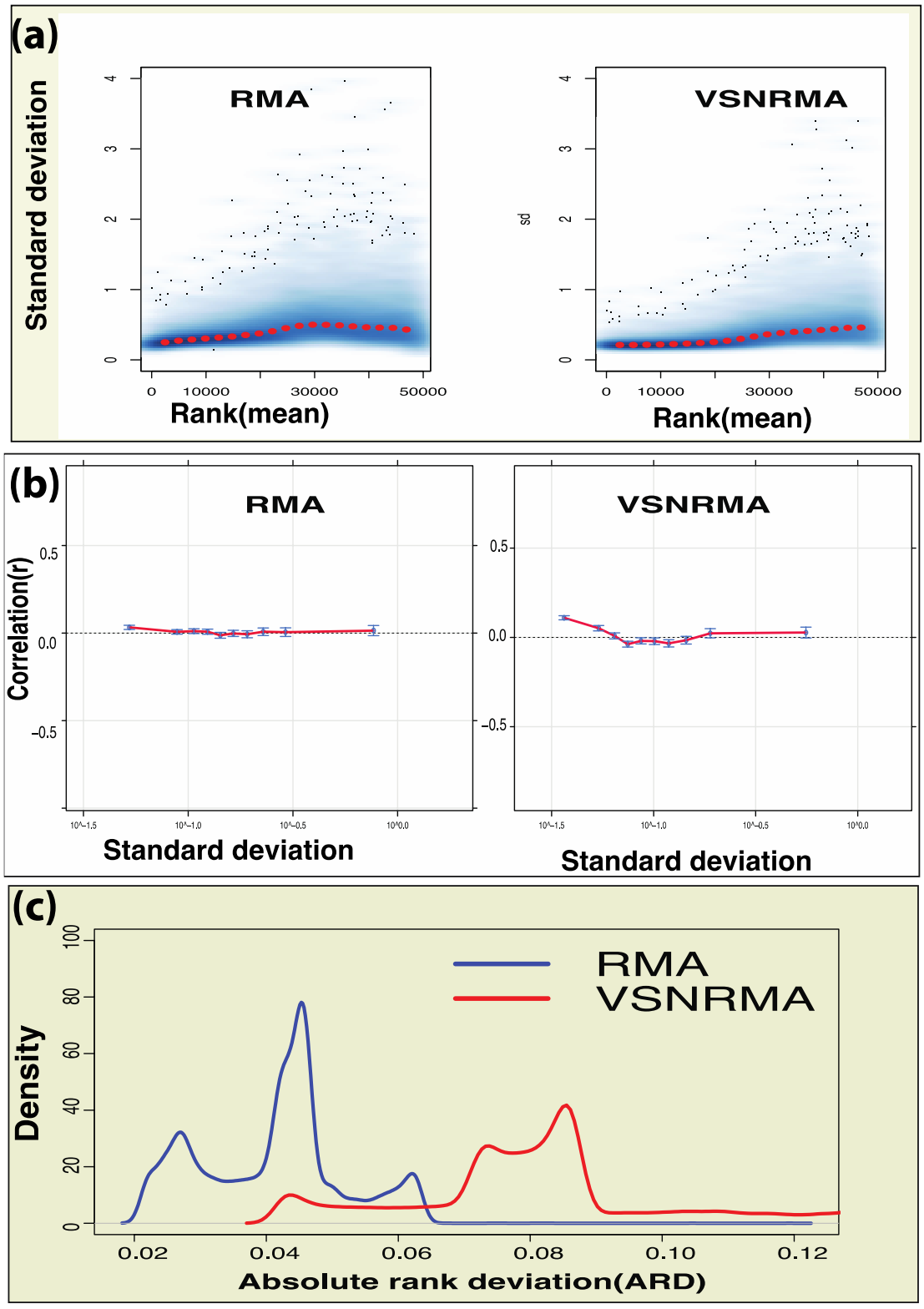


Figure 3.5: Performance assessment of raw data pre-processing methods (RMA and VSNRMA) in the training data. **(a)** Per gene probe standard deviation (y-axis) against ranked mean values (x-axis). **(b)** Correlation between randomly selected pair of gene probes (y-axis) against the mean of their standard deviations. **(c)** Absolute rank deviation (ARD), which is between the samples standard deviations of gene expression values sharing the same rank across the gene features[175].

3.3.1.1 Assessment and correction of batch effects variations

Batch-effect variations come from several data handling sources including differences between personnel, laboratory conditions, array platforms and time of the experiment [195]. However, raw data pre-processing algorithms are often not adequate to remove batch-effect variations [150, 195, 340]. In this study, the microarray experiment was conducted in two batches, 447 samples in 2013 and 71 samples in 2014, which required investigation. The second batch was particularly added to minimise demographic data imbalances between the study groups.

To assess the batch-effect variations, principal component analysis (PCA) was applied to identify unsupervised samples clusters. To account for the biological variations, gene expression profiles for the negative control probes (which are designed to remain constant under different biological conditions) were applied in this analysis. As shown in **Figure3.6**, batch-effect variations were observed in both the training (**Figure3.6a)** and validation (**Figure3.6c)** datasets beyond raw data pre-processing.

To remove the unwanted batch-effect variations, **ComBat** normalisation algorithm was applied [195, 196]. Briefly, **ComBat** is a Bayesian data standardisation algorithm, which empirically estimates parameters for the location-scale (l-s) model using normal and inverse gamma distribution priors. While several batch-effect correction algorithms exist[341], **ComBat** is robust to small sample batches and already implemented in the *sva* Bioconductor package[195]. Here, the **ComBat** algorithm successfully

removed the batch-effect variations in both the training (**Figure3.6b)** and validation **(Figure3.6d**) data sets.

To assess whether **Combat** was not overcorrecting, a sensitivity analysis was conducted using randomly simulated batches (within Batch1 of the training data). Unlike with the real batches, sample clustering for the randomly generated batches remained unchanged before (**Figure3.6e**) and after (**Figure3.6f**) **Combat** adjustment, which reassured its effectiveness. Together, this analysis highlights the challenge of batch-effect variations, which should be avoided by study design or at least investigated beyond data pre-processing. In the next sections, outliers were investigated.

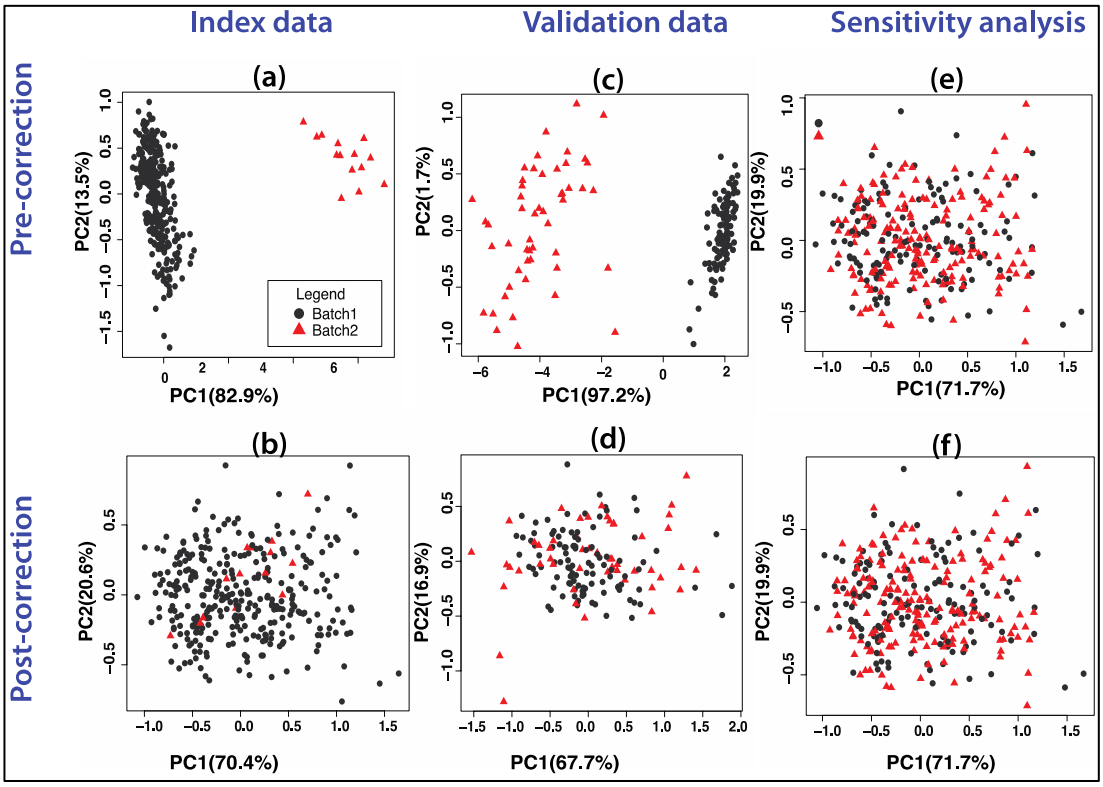


Figure 3.6: Identification and normalisation of batch effect variations. The figure panels show sample clustering before (row1) and after (row2) batch effect correction in the training **(a-b)** and validation **(c-d)** data. Sensitivity analysis used randomly simulated batches within the Batch1 of the training data (e-f).

3.3.1.1 Investigation of outliers

Outliers are samples that deviate from the global (abnormal samples) or group-specific (misclassified samples) distributions. These can cause deleterious effects on statistical inference and machine learning approaches such as (i) reduced statistical power, (ii) biased estimates, (iii) violation of normality assumptions or (iv) over-fitting of models[152, 342].

To detect the suspected global outliers, the *arrayQualityMetrics* algorithm in R Bioconductor [151, 343] was applied before and after data pre-processing (**Figure3.7**). At the time of this analysis, six metrics (more details in Chapter 2) involving the relative distribution of expression signals, the distance between the arrays, and the absolute quality of each sample were implemented. While all the metrics were applicable to the raw database, only three were applicable to the pre-processed database. Here, samples that were detected by at least two-third of the applicable methods in the raw (4/6) or pre-processed (2/3) databases were eliminated (respectvely) as outliers (**Figure3.7**).

As shown in **Figure3.7**, 15 outliers were detected in the training data. Of them, 8 were detected in the raw database, and 7 after data pre-processing. However, using the same criteria, no outlier was detected in the validation data set. Subsequently, 345 and 158 samples were analysed in the training and validation data sets, respectively (**Figure3.7)**. In the next section, I applied the Y-linked genes to validate the sex variable (gender analysis).

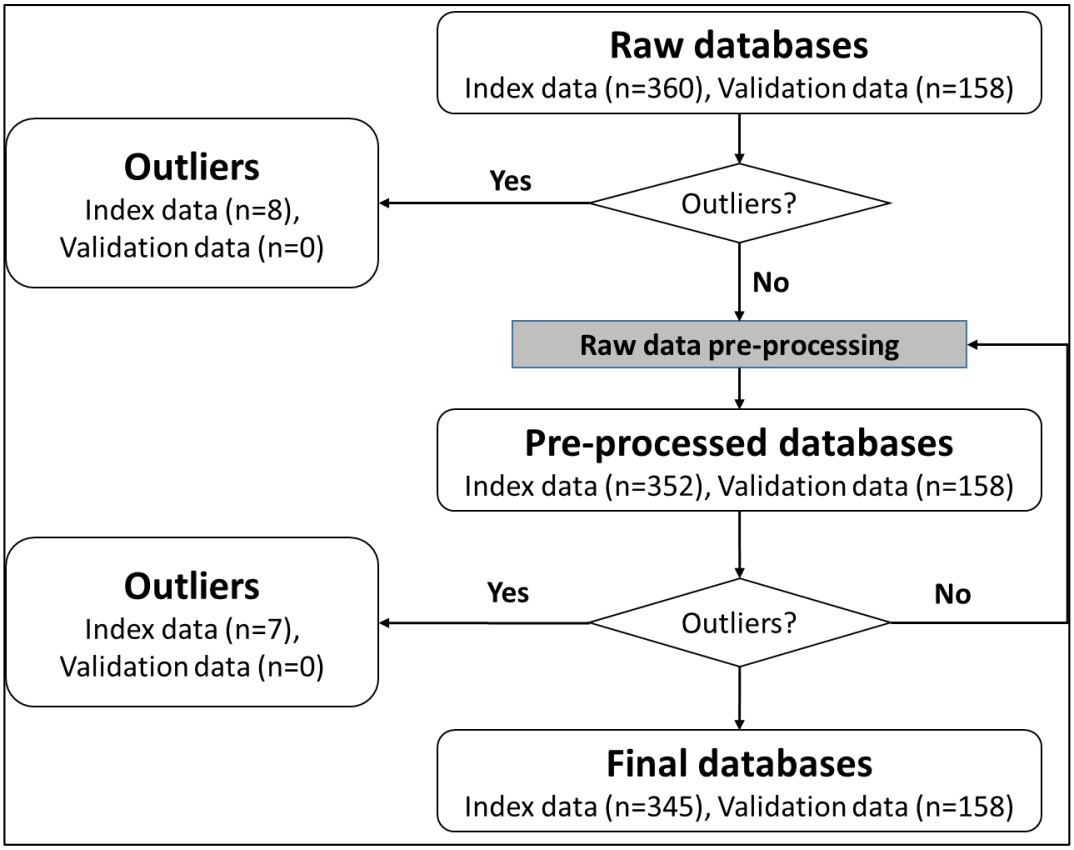


Figure 3.7: Detection of outliers in the training and validation data sets, respectively. Outliers were detected sequentially in the raw and pre-processed databases using the ***arrayQualityMetrics*** package[14, 23]

3.3.1.1 Gender analysis: Molecular identification of potentially sex-misclassified samples

In vertebrates, the Y chromosome is a sex-determining region of the DNA [197], and a powerful molecular signature for classifying sex phenotypes (male or female). To further assess the quality of the existing data resources, this property was applied to validate the sex variable labels (gender analysis) in the demographic database. In particular, an expression signature of the Y-linked genes (n=65) was subjected to principal component analysis to identify suspicious samples (**Figure3.8**).

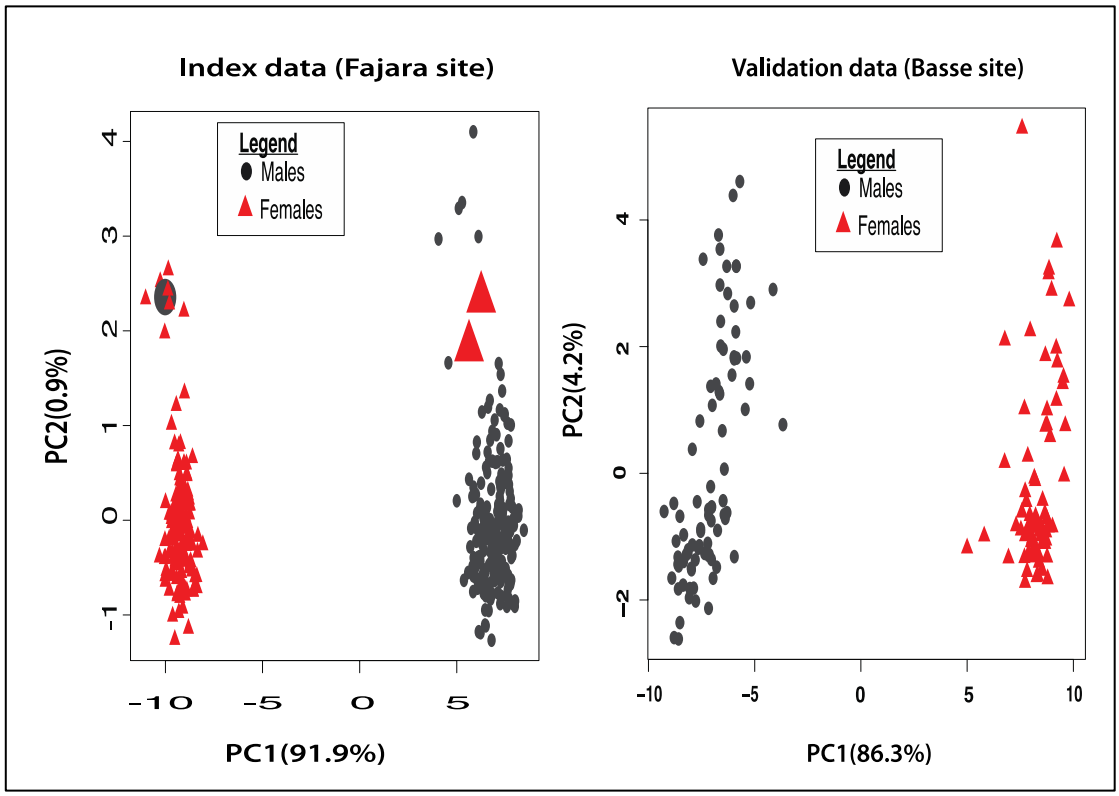


Figure 3.8: Gender analysis. Principal component analysis visualisation of samples using 65 Y-chromosome specific genes

As shown in **Figure3.8**, the Y-linked signature clearly distinguished between the male and female samples in both data sets. Notably, only 3(<1%) samples (one male and two females) were potentially misclassified in the training data, which suggest the quality of the current database. Further, these sample labels were consistent with the reference database in The Gambia, suggesting a data collection error (not data entry), or other unexplained factors. Unlike in the previous section, the suspicious samples were retained for further investigation since sex is neither the main outcome nor a serious confounder (**Table3.3**). To minimise false discoveries due to multiple testing, non-informative or redundant gene probes were eliminated (next section).

3.3.1.1 Filtering for non-informative or redundant gene probes

While a whole blood transcriptome provides a comprehensive approach for investigating the systemic pathway responses and candidate biomarkers[71, 123], not all the genes are relevant to every disease[125]. Further, the analysis of high-throughput data often suffers from the "*curse of dimensionality"* (i.e. analysing more variables than sample sizes) including false discoveries due to multiple testing[71], and feature selection challenges in machine learning[344-346]. In this thesis, the HGU219 array platform has analysed 49386 variables (gene probes) using 345 samples in the training database.

To mitigate the potential for false discoveries due to multiple testing, non-informative gene probes were eliminated prior to differential gene expression analyses[153]. Of the 49386 gene probes, 32677 (66%) were eliminated using the following non-specific joint criteria (**Figure3.9**):

(i) Lack of annotation (i.e. ENTREZID),

(ii) Low signal intensity: In each sample **i**, the threshold $C_i = Median_i + 2 * MAD_i$ was applied; where *Median* and *MAD* are the sample median and "median absolute deviation" values (respectively) estimated from the expression values across the negative control gene probes. Thus, gene probes with expression values less than the $\mathbf{C_i}$ threshold in at least 5% of the samples were eliminated.

(iii) Low variability: coefficient of variation (CV)<10%.

However, the microarray technology assay often applies multiple gene probe sets to investigate a single gene. To remove the redundant gene probes among the filtered gene probes (m=16709), a maximum mean filter was applied. In particular, a gene probe with the maximum mean across all the samples was retained for each gene. Together, 11037 gene probes representing unique genes were selected for subsequent differential expression analyses (**Chapter 5**).

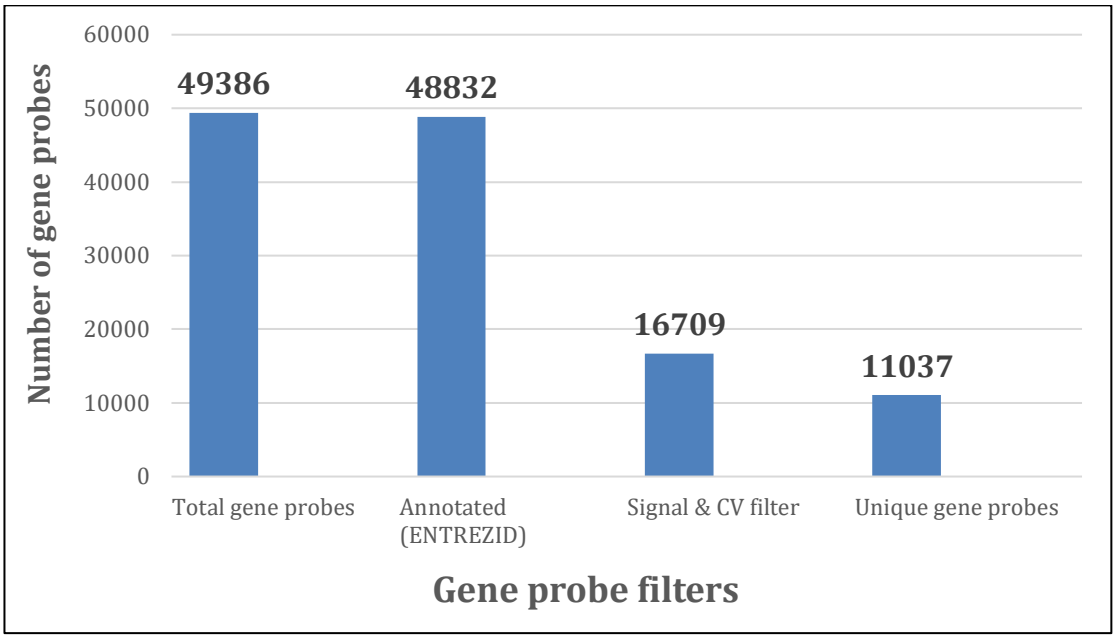


Figure 3.9: Non-specific filtering of gene probes in the training data prior to differential gene expression.

3.1.4 Molecular phenotyping of samples

This section investigates the association between molecular signatures and clinical phenotypes.

3.3.1.1 Prediction of samples with suspected bacterial septicaemia using the neonatal sepsis classifier [136]

This thesis investigates the hypothesis that systemic pathway responses underpin the development of severe pneumonia outcomes. Systemic molecular biomarkers are powerful resources for predicting disease outcomes [347]. To assess the association between pneumonia severity and bacterial septecaemia (blood infection), I applied a validated 52-gene neonatal sepsis classifier by Smith, et al., (2014), [136]

Briefly, the sepsis classifier (above) applies a transcriptomic signature of pathway biology (innate, adaptive and metabolic) derived genes (m=52) to identify an optimal threshold for predicting the class (positive or negative) of new samples using a ROC analysis-based classification algorithm [348]. Using this classifier on the pneumonia database (**Table3.2**), it was predicted that 53.6%(185) of the training sample (n=345) had bacterial sepsis. Notably, the prevalence of septicaemia increased with pneumonia severity from 67.8% in mild pneumonia cases to 100% in very severe pneumonia cases (p-value<0.001), and differed significantly between the (i) non-pneumonia controls and mild pneumonia cases (p-value<0.001), (ii) mild and severe pneumonia cases (p<0.00) and (iii) mild and very severe pneumonia cases (p-value<0.003), but not between the severe and very severe pneumonia cases (p=0.215).

	Control		Mild		Severe		Very Severe		P-value
Diagnostic tools	N	n(%)	N	n(%)	N	n(%)	N	n(%)	
N	120		90		117		18		
Sepsis classifier	120	2 (1.7%)	90	61 (67.8%)	117	104 (88.9%)	18	18 (100%)	<0.001
Chest x-ray	120	0 (0.0%)	57	50 (87.7%)	113	102 (90.3%)	18	17 (94.4%)	<0.001
Blood culture results (all)	119	19 (16.0%)	90	12 (13.3%)	117	17 (14.5%)	17	2 (11.8%)	0.94
Blood culture results (No contaminants)	108	8 (7.4%)	88	10 (11.4%)	114	14 (12.3%)	17	2 (11.8%)	0.66
PCR results	120	38 (31.7%)	90	51 (56.7%)	117	66 (56.4%)	18	12 (66.7%)	<0.001

Table 3.2: Molecular and clinical phenotypes in the training data. The table shows stratified (by pneumonia severity) proportions of samples with (i) suspected bacterial septicaemia (based on **Sepsis classifier**), (ii) significant chest-x-ray pathology (**Chest x-ray**) and (iii) **blood culture** confirmed results with (all) and without (no contaminants) samples labelled as contaminants. N=Total number of samples analysed with a pneumonia study group (denominator); n=number of samples with positive outcome. P-values were generated from Fisher's exact test for associations.

On the other hand, while pneumonia severity was significantly associated with significant chest x-ray pathology and PCR positivity (P-value<0.001, no significant differences were observed between the mild and severe or very

severe pneumonia cases (p-value>0.5). Further, BloodCulture-confirmed results lacked sensitivity, and there was no significant association with pneumonia severity (p-value>0.5). In particular, only 52 samples (14.58%) had blood-culture confirmed positive results including Streptococcus *pneumonia (n*=14; 4.08%), *Staphylococcus aureus (n*=4; 1.17%), contaminants (n=16;4.66%), and other organisms (n=16; 4.66%) such as Bacillus species (n=5), *Micrococci species* (n=6), *non-typeable Haemophilus influenzae* (n=1), *Streptococcus viridan* (n=3) and *Streptococcus species* (n=1). Nevertheless, it is worth noting that the higher prevalence of *Streptococcus pneumonia isolates* is consistent with several aetiology studies worldwide [41, 50, 349, 350].

In summary, these findings highlight the (i) limitations of the existing standard diagnostic tools [55], (ii) the important contribution of bacterial septicaemia in the development of serious pneumonia outcomes and (iii) the potential of systemic molecular signatures for clinical stratification of pneumonia cases (investigated further in **Chapter 6**). To gain an overview of the training whole blood transcriptome, unsupervised clustering approaches were applied (next section).

3.3.1.2 Identification of inherent sample clusters (unsupervised clustering)

Class discovery is among the main objectives of microarray analyses, where unsupervised approaches are applied. In the previous sections, principal component analysis (PCA) has revealed sex (**Figure3.8**) and batch-effect (**Figure3.6**) related sample clusters, respectively. Here, a similar approach

(unsupervised learning) was applied to assess the overview structure of the training data transcriptome. In particular, this analysis assessed whether sample clustering reflected study outcomes (pneumonia severity) or unaccounted confounders.

To identify the dominant and stable clusters, k-means clustering algorithm [351] coupled with bootstrap resampling[225] was applied on the most variable genes (m=76) with at least 30% coefficient of variation (CV) [352] across all the samples. To estimate the cluster stability, the *Jaccard coefficient*[225] was applied. Briefly, this coefficient estimates the proportion of bootstrap samples in which the original clustering is reproduced, and mathematically *ranges between 0% (no stability) and 100% (perfect stability).* To visualise the sample clusters, the T-SNE (t-distributed stochastic neighbor embedding) dimensionality reduction algorithm [212] was applied (**Figure3.10**)**.**

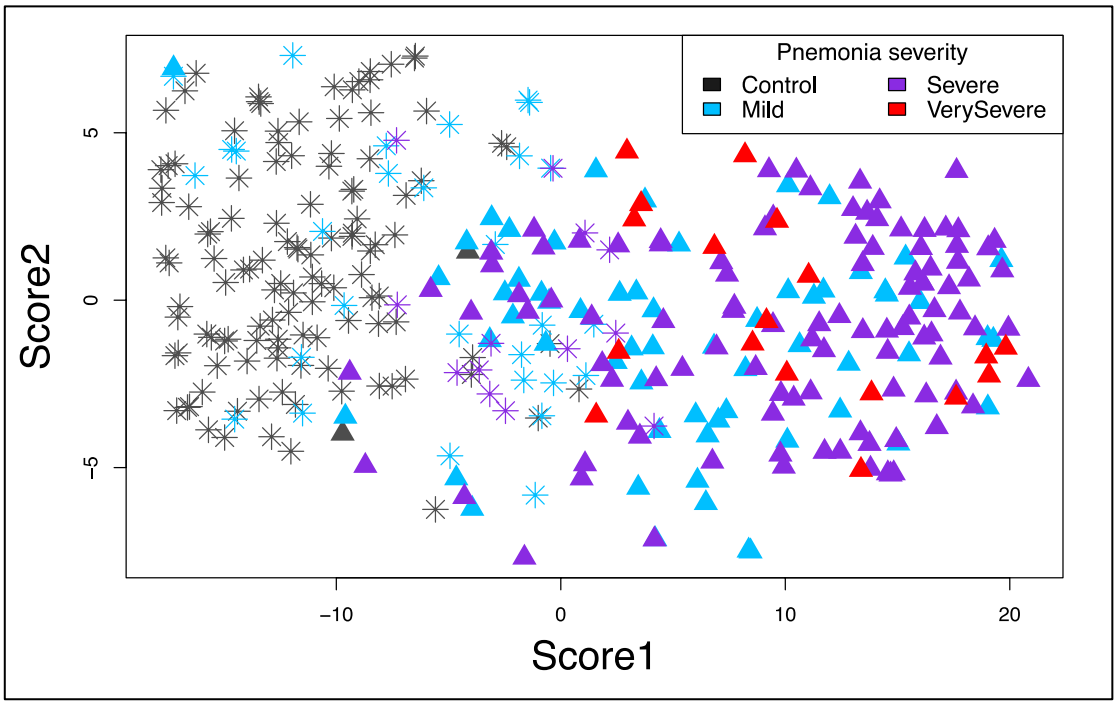


Figure 3.10: Unsupervised clustering of samples in training data (n=345). Data visualisation is based on the scores derived from the T-SNE algorithm, which is non-linear and more powerful than the principal component analysis (PCA) [212] (details in chapter2). Each dot represents a sample annotated by pneumonia severity state (legend) and predicted septicaemia (Triangles=Positive, stars=negative).

Here, I assessed the stability of different numbers of clusters between two and ten. Among them, two k-mean clusters (222 samples in *cluster1* and 123 in *cluster2*) were associated with the best stability (*Jaccard coefficient* =99.3% and 0.98.7% respectively). As shown in **Figure3.10**, this data partitioning was significantly associated with pneumonia severity and the bacterial septicaemia (using the sepsis classifier), respectively (P-value<0.001). In particular, the *cluster1* was associated with better clinical outcomes representing 100%(120), 64%(58/90), 33%(39/117) and 28%(5/18) of the non-pneumonia controls, mild, severe and very severe pneumonia cases (respectively). Similarly, 100%(160) and 34%(62/185) of the samples with negative and positive septicaemia predictions (respectively) were also associated with *cluster1*. Firstly, these findings highlight the quality of the central data resources since the clustering reflects the study

hypothesis (i.e. not major unaccounted confounding factors). Further, the association between the unsupervised clusters and bacterial septicaemia suggest the potential involvement of systemic responses in the development of severe pneumonia outcomes, and highlights the potential of whole blood transcriptomes in the clinical stratification of pneumonia cases. In the next section, potential epidemiological challenges were investigated.

3.1.5 Epidemiological considerations

3.3.1.3 Identification of potential confounders in the training data

A confounder is nuisance factor that is associated with both the exposure and outcome variable [353, 354]. If not accounted for, the imbalances in these factors often generate misleading conclusions. In this study, we can loosely define pneumonia severity as exposure, and cellular and molecular responses as the outcomes. While the study groups were sufficiently powered (**Figure3.3**, **Table3.1**) and matched by study design, residual confounding is inevitable especially in observational studies [355]. Therefore, potential confounders were investigated in the training data where most of the cellular and molecular pathway responses (primary objectives) were investigated to identify key covariates for subsequent analyses. To identify the potential confounders, this section investigated the associations between pneumonia severity (exposure), and the clinical and demographic variables

As shown in **Table3.3,** the distributions of sex, season and Vitamin A supplementation were similar between the study groups (P-value>0.05). However, age, nutrition status, sample batch and antibiotic usage significantly (P-value<0.05) differed between the pneumonia severity groups

(i.e. potential confounders). In particular, younger children, poor nutrition status and antibiotic usage were associated with worse clinical outcomes. Together, this analysis has identified age, nutrition status, and antibiotic usage as potential covariates for subsequent analyses. To investigate further, the potential confounders were assessed molecularly (next section).

Factor	Control	Mild	Severe	Very Severe	P-value
N	120	90	117	18	
Demographics					
Age in months					
Median (IQR)	15.5 (7.8, 23.2)	14.8 (10.0, 24.1)	14.1 (7.7, 23.2)	7.3 (5.6, 12.2)	0.007
Age groups (months)					0.050
<6	18 (15.0%)	7 (7.8%)	22 (18.8%)	7 (38.9%)	
6-11	29 (24.2%)	23 (25.6%)	25 (21.4%)	6 (33.3%)	
12-23	47 (39.2%)	37 (41.1%)	45 (38.5%)	5 (27.8%)	
24-59	26 (21.7%)	23 (25.6%)	25 (21.4%)	0 (0.0%)	
Gender					0.66
Female	49 (40.8%)	37 (41.1%)	52 (44.4%)	10 (55.6%)	
Male	71 (59.2%)	53 (58.9%)	65 (55.6%)	8 (44.4%)	
Season					0.44
Dry	68 (56.7%)	48 (53.3%)	55 (47.0%)	8 (44.4%)	
Wet	52 (43.3%)	42 (46.7%)	62 (53.0%)	10 (55.6%)	
Sample batches					<0.001
Batch1	111 (92.5%)	89 (98.9%)	114 (97.4%)	14 (77.8%)	
Batch2	9 (7.5%)	1 (1.1%)	3 (2.6%)	4 (22.2%)	
Nutrition status					
Under weight (Weight-for-Age)					
WAZ score, mean (SD)	-0.8 (1.1)	-1.2 (1.2)	-1.5 (1.4)	-0.9 (1.5)	<0.001
Moderate underweight (WAZ<-2)	13 (10.8%)	24 (26.7%)	48 (41.0%)	5 (27.8%)	<0.001
Severe underweight (WAZ<-2)	1 (0.8%)	5 (5.6%)	12 (10.3%)	1 (5.6%)	0.018
Stunting (Height-for-Age)					
HAZ score, mean (SD)	-0.6 (1.2)	-0.7 (1.1)	-0.8 (1.8)	0.2 (1.4)	0.042
Moderate stunting (HAZ<-2)	11 (9.3%)	9 (10.1%)	23 (19.7%)	1 (5.6%)	0.055
Severe stunting (HAZ<-2)	1 (0.8%)	2 (2.2%)	7 (6.0%)	0 (0.0%)	0.096
Wasting (Weight-for-Height)					
WHZ score, mean (SD)	-0.6 (1.0)	-1.1 (1.3)	-1.4 (1.7)	-1.3 (1.8)	<0.001
Moderate wasting (WHZ<-2)	10 (8.5%)	19 (21.3%)	34 (29.1%)	6 (33.3%)	<0.001
Severe wasting (WAZ<-3)	0 (0.0%)	7 (7.9%)	16 (13.7%)	3 (16.7%)	<0.001
Clinical values					
HIV positive			4 (6%)	1 (11%)	0.59
Iron supplementation			25 (28%)	3 (23%)	0.71
Vitamin A (within 6 months)	59 (51.3%)	50 (56.8%)	51 (49.0%)	10 (58.8%)	0.68
Antibiotic usage (within 2 weeks)	2 (1.7%)	16 (18.0%)	32 (29.6%)	7 (41.2%)	<0.001

Table 3.3: Demographic and clinical characteristics of the training sample (Fajara). Reported P-values are for Chi-square / Fisher's exact test (categorical variables) or student t-test/Mann-Whitney test (continuous variables). WAZ=weight-for-Age Z score, WHZ=weight-for-height Z score, HAZ=height-for-age Z score, IQR=interquartile range, SD=standard deviation. Potential confounders are highlighted in RED colour.

3.3.1.4 Identification of potentially confounded genes

In **Table3.3**, age, nutrition status, antibiotic usage and batch-effect were associated with pneumonia severity (exposure). To further investigate these potential confounders and identify the key covariates for subsequent analyses, this section assessed the associations with the study outcomes (gene expression). For each factor, the numbers of potentially confounded genes were estimated (**Figure3.11b**).

Firstly, I estimated the numbers of differentially expressed genes (DEGs) before (pre) and after (post) adjusting for each potential confounder. In each analysis, the empirical Bayes moderated t-test (using the *limma* package [226]) was applied to identify the DEGs between the non-pneumonia controls and each pneumonia severity group, respectively. As illustrated in **Figure3.11a**, potentially confounded genes were exclusively significant (FDR<0.05, |FC|≥2) before (positively confounded) or after (negatively confounded) adjusting for a particular covariate (i.e. age).

In overall, negative confounding (i.e. masked genes, blue colour) was more predominant than false positive (red colour) discoveries (**Figure3.11b**). Comparatively, age (n=216) was the strongest confounder followed by antibiotic usage (n=89) and nutrition status especially stunting (n=82). Potentially, these confounders may undermine the systemic pathway responses in very severe pneumonia where more participants were younger, malnourished and associated with more antibiotic usage (**Table3.3**). Therefore, it is important to adjust for these variables in subsequent analyses.

However, while all the nutrition status variables (stunting, under-weight and wasting) were identified as potential confounders, these indices are correlated, and not ideal to be adjusted in the same model (i.e. to avoid *multicollinearity*). Instead, principal component analysis (PCA) was applied to transform the nutrition status indices into uncorrelated principal component (PC) scores. Here, the first principal component, which captured 66% of the variability in data, was selected as surrogate covariates for nutrition status. Together, this analysis has identified age, antibiotic usage and nutrition status as key covariates for subsequent analyses. *In the next section, I investigated the presence of effect-modification.*

(a)

Genes	Gene significant?		Gene confounded?
	Before adjusting for covariate (Pre)	After adjusting for covariate (Pre)	
Gene1	✔	✔	No
Gene2	✖	✔	Yes (Negative)
Gene3	✔	✖	Yes (Positive)
Gene4	✖	✖	No

(b)

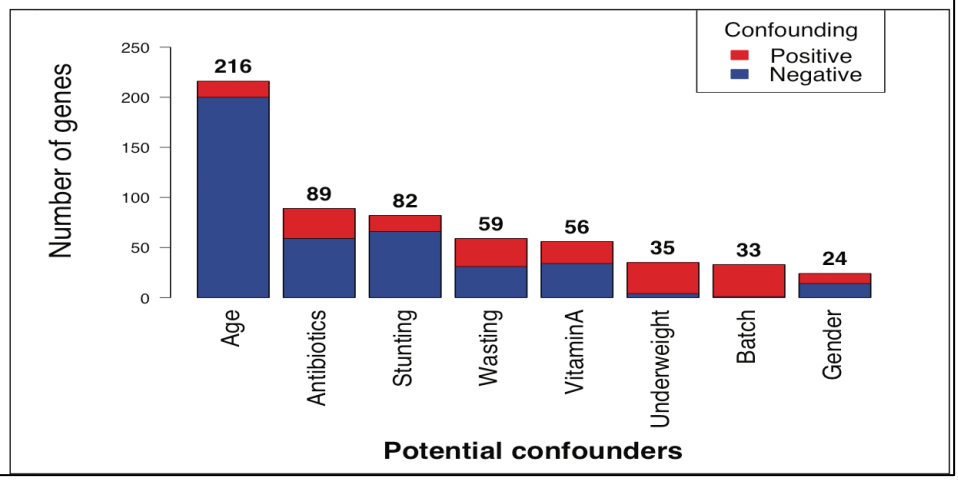


Figure 3.11: Numbers of potentially confounded genes in the training data. **(a)** An illustration of confounded genes: **Gene1** and **Gene2** are not confounded because their significance does not depend on the confounder. On the other hand, **Gene3** and **Gene4** were confounded negatively (masked by the confounder) and positively (false positive discovery driven by the confounder) respectively.**(b)** Number of genes of positively (red) and negatively (blue) confounded genes (y-axis) by each potential confounder (x-axis).

3.3.1.1 Identification of effect-modified genes

Effect modification is an epidemiological term for stratification, which occurs when an exposure (i.e. pneumonia severity) has different (strength or direction) outcomes (i.e. gene expression) across the strata of a third variable (i.e. age groups) [356]. For example, a gene can be up regulated in males but down regulated in females or strongly up-regulated in older children but not significant in infants. While confounding is always a nuisance factor creating false discoveries, effect modification provides important insights into subgroup variations (i.e. magnitude and direction of association). To further characterize the existing data resources, this section investigated the number of effect-modified genes (EMGs) across the strata of clinical and demographic variables including the potential confounders (**Figure3.12)**.

To identify potentially modified genes, empirical Bayes moderated F-test (limma package [226]) was applied to test for significant interaction between each potential effect modifier (i.e. age) and pneumonia severity. For genes with significant interaction terms (P-value<0.05), subgroup-specific contrasts were tested (empirical Bayes moderated t-tests) to identify differentially expressed genes (FDR<0.05, FC≥2) between non-pneumonia controls and severe pneumonia groups (respectively). As illustrated in **Figure3.12a**, effect-modified genes (i.e. Gene4 and Gene5) had significant interaction (column2) and different conclusions across the strata (columns 3 and 4).

(a)

Genes	Interaction test significant?	Subgroup effect significant?		Significantly effect-modified?
		Group1	Group2	
Gene1	✗			No
Gene2	✓	✗	✗	No
Gene3	✓	✓	✓	No
Gene4	✓	✓	✗	Yes
Gene5	✓	✗	✓	Yes

(b)

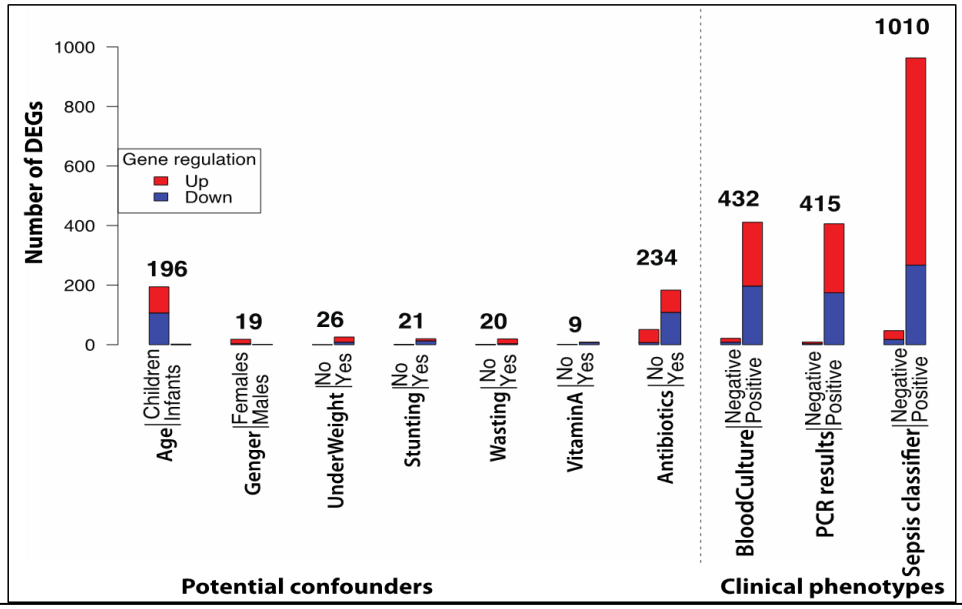


Figure 3.12: Numbers of effect-modified genes in the training data. An illustration of effect modified genes: **Gene1** did not pass the interaction test (no subgroup effects were tested) while the effects of **Gene2** and **Gene3** did not differ between the subgroups, hence not significantly modified. On the other hand, **Gene4** and **Gene5** passed the interaction test and were exclusively significant in one subgroup, hence significantly modified. Total number of effect-modified genes across the clinical and demographic strata when severe pneumonia groups were compared to non-pneumonia controls, respectively. For each variable (x-axis), the total numbers of effect-modified genes are shown on top of the stratum-specific bars: **DEGs**=differentially expressed genes (FDR<0.05, |FC|≥2)

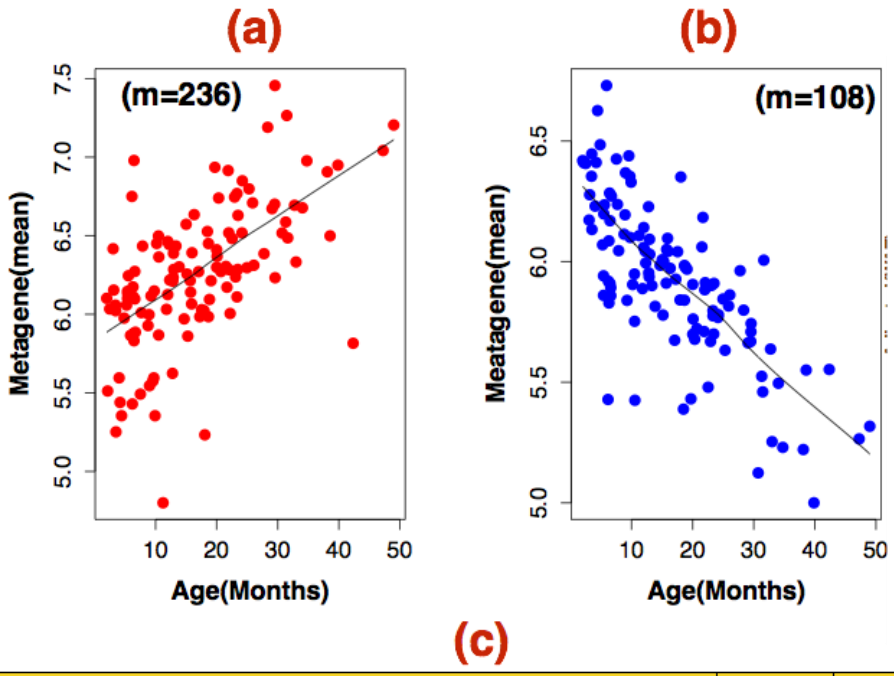
In overall, clinical phenotypes were associated with more effect-modified genes than the potential confounders (**Figure3.12b**). Notably, septicaemia (predicted by the sepsis classifier [136], **Table3.2**) had the highest number of effect-modified genes (m=1010). At the gene analytic level, pneumonia cases with suspected bacterial septicaemia (blood culture, PCR and sepsis classifier) were associated with stronger systemic molecular responses. These findings support the central hypothesis that systemic responses underpin the development of severe pneumonia outcomes and further

suggest the importance of bacterial aetiology in serious pneumonia outcomes (discussed more in **Chapter6**).

Consistent with the confounder analysis, age (m=196) and antibiotic usage (m=234) were associated with the highest numbers of effect-modified genes among the potential confounders. Notably, order children and antibiotic usage (especially down-regulated genes) were associated with stronger systemic molecular responses. Together, these findings provide an insight into the gradient of the molecular responses in severe pneumonia, further highlighting the importance of adjusting for the potential confounders in subsequent analyses. To gain more insight into the age dependency, age analysis was conducted *(next section).*

3.3.1.2 Characterization of age-dependent genes among the non-pneumonia controls (age analysis)

In the previous sections, age has emerged as a strong confounder (**Figure3.11**) and effect-modifier (**Figure3.12**). At the gene analytic level, older children were associated with enhanced systemic responses in pneumonia than the infants. To gain more insights into the ontogeny of systemic pathway responses, age-dependent genes were characterised. To account for the potential confounding effects, this analysis was restricted to the non-pneumonia controls, and adjusted for nutrition status. In particular, empirical Bayes moderated linear regression analysis approach (*limma* package) was applied to identify the genes that were associated (FDR<0.05) with age (continuous scale) while adjusting for the potential confounding effects of nutrition status (**Figure3.13**).



(c)

KEGG Pathway	Hits	FDR
TNF signalling pathway	9	0.000965
Cytokine-cytokine receptor interaction	13	0.00131
Chagas disease (American trypanosomiasis)	7	0.0107
Amoebiasis	7	0.0132
Staphylococcus aureus infection	5	0.0158
NF-kappa B signaling pathway	6	0.0184
Osteoclast differentiation	7	0.0184
Hematopoietic cell lineage	6	0.0184
Epithelial cell signaling in Helicobacter pylori infection	5	0.0272

Figure 3.13: Characterisation of age-dependent genes among the non-pneumonia controls. (a)-(b) Scatter plots showing the association between the up (a) or down (b) regulated genes with increased age (x-axis). Each dot represents a sample; and the y-axis represents the sample-specific mean expression value across the up (a) or down (b) regulated genes: m=number of genes. (c) Enriched KEGG pathways associated with the up-regulated gene in (a). Hits: The numbers of up-regulated genes that were enriched on each pathway. FDR=Multiple testing adjusted p-values (Benjamini–Hochberg (BH) procedure[238]) from the hypergeometric test.

In total, 344 genes (up-regulated=236, down-regulated=108) were significantly associated with age (FDR<0.05) among the non-pneumonia controls (**Figure3.13a-b**). To assess the molecular functions associated with the age-dependent gene sets, the STRING database for protein-protein-network analysis [280] was applied. While the down-regulated genes (m=108) were not associated with significant pathways, the up-regulated

genes (m=236) were predominantly associated with elevated basal levels of the pro-inflammatory systemic innate responses according to the KEGG pathway database (**Figure3.13c**). Similarly, Burl et al. (2011) also observed age-dependent maturation of the pro-inflammatory cytokine responses among Gambian infants (n=120) [357]. While natural developments of the host system partially explain this observation, other contributing factors may include (i) underlying asymptomatic diseases and (ii) sequelae of prolonged exposure to hazardous chemicals or infections. Together, these findings suggest the potential of exploring age-dependent systemic responses towards the implementation of personalized clinical management of pneumonia cases in resource-limited settings.

3.4 Discussion

This chapter has introduced, evaluated and curated the central data resources for this thesis to ensure data quality and facilitate subsequent analyses. Briefly, the central data resources include the microarray whole blood transcriptome and the corresponding phenotypic databases (demographic, clinical, microbiology). Whole blood is rich and readily accessible tissue for clinical investigations, and its application in genome-wide investigations (i.e. transcriptomics) has become a mainstay of comprehensive genomic research and future translation medicine for a wide range of diseases including cancer, infections and autoimmunity [71, 102, 123, 204, 358]. Therefore, these data resources provide a powerful and innovative approach for gaining deeper insights into the pathogenesis of pneumonia, and present an opportunity for future clinical stratification and

treatment modalities of pneumonia cases. However, inferring from transcriptomic data has several limitations, which require careful considerations from study design to the final data analysis [129, 147, 182, 359]. Here, the strengths and limitations of the existing resources are highlighted.

3.1.6 Strengths

To detect meaningful biological insights, an appropriate study design and sufficient sample sizes are fundamental. In this study, a matched case-control study design was implemented to account for the potential confounding effects of age, sex, season and location. Further, while many genomic studies are underpowered (i.e. due to financial or ethical constraints, or lack of appropriate sample size estimates[68, 71]), the original study design was sufficiently powered using conservative approaches (i.e. the Bonferroni correction[170]). Further, sample size reassessment suggested that the existing data resources are sufficiently powered to address the primary objectives; and to detect meaningful biological effects even in subgroup analyses.

Further, another limitation with high-throughput data is lack of reliable phenotypic data [57, 68, 70, 71]. Here, the whole blood transcriptome has a comprehensive database for metadata records including clinical, demographic and laboratory phenotypes, which was subjected to intensive data-cleaning to ensure data quality. Notably, missing data were very minimal especially among the key variables for addressing primary objectives of this thesis. Further, gender analysis identified minimal

suspicious samples (<1%), which reassured the quality and completeness of the available data resources. Furthermore, the application of the sepsis classifier [360] to molecularly predict samples with bacterial septicaemia provides a powerful approach for data cleaning and addressing the primary objectives.

To enhance the quality of the whole blood transcriptome and minimise the potential confounding effects, several statistical approaches were applied to eliminate the non-biological variations in the data. Firstly, to account for the technical (i.e. hybridization) variations across the array[145], an appropriate algorithm for the pre-processing of raw data was carefully selected; and potential outliers were eliminated. It is worth noting that the design of the array platform for this database (HGU219) does not include the expression data for the mismatch probes. Therefore, while several pre-processing algorithms such as MAS, GCRMA, MBEI (or Li & Wong) exist[339], here the Robust Multi-Array (RMA)[147] and variance stabilizing normalization (VSN) algorithms were applicable. In particular, the RMA [147] algorithm empirically outperformed VSN in both the training and validation data, and successfully normalised the unwanted sample variations.

However, the current whole blood transcriptome was processed in two sample batches; and raw data pre-processing algorithms are not optimised to eliminate batch-effect variations [340]. While this problem is better prevented at the study design stage, here a computational solution (***comBat*** algorithm) was applied and successfully resolved the unwanted batch-effect

variations[195]. While several batch-effect correction algorithms such as DWD weighted discrimination (DWD), surrogate variable analysis (SVA), Mean-centering (PAMR) and Geometric ratio-based method (Ratio_G) exist, ***ComBat*** remains the most successful algorithm [150]. Importantly, it is robust to small sample sizes and readily available in an R Bioconductor environment [195].

While appropriate study design is vital for minimizing confounding effects [353, 354, 356], it is equally important to investigate and account for residual confounding during analyses. On one hand, it is very challenging to account for several confounders during sample collection. On the other hand, residual confounding is almost inevitable especially in observational studies [361]. Here, potential confounders and effect-modifiers were comprehensively investigated to identify key covariates for subsequent analyses (age, nutrition status and antibiotic usage). Notably, age-dependencies were consistently observed in confounder, effect-modification and age analyses, and the findings were consistent with previous observations in the same population [357]. While age-dependencies present a confounding challenge in the investigations of systemic responses, these findings present an opportunity for future personalized clinical interventions in pneumonia.

Further, effect-modification analysis revealed systemic response differences across the demographic and clinical strata. Notably, children with bacterial septicaemia (i.e. blood culture, PCR and the sepsis classifier) were consistently associated with stronger systemic responses in severe

pneumonia. These findings suggest that the data structure reflect the intended study objectives (i.e. not unaccounted confounders), and they support the central hypothesis that systemic pathway responses underpin the development of severe pneumonia states. Clinically, the highlighted importance of bacterial aetiology presents an opportunity for host-based biomarkers and treatment modalities in pneumonia cases (**Chapter 6**).

3.1.7 Limitations

Despite the highlighted strengths above, these data resources have some limitations. Firstly, it is worth noting that this is an observational study design, which is susceptible to potential confounders and has limited interpretations [355]. At the individual level, the samples were collected at a single time point (i.e. cross-sectional study design). Consequently, this database has lacked vital follow up data such as patient outcomes. Preferably, a longitudinal study design would enable the proper investigations of causality and prognostic biomarkers. Further, while antibiotic usage was identified as a key confounder, this data was based on the reported testimony. Potentially, this approach is susceptible to recall-bias (i.e. due to loss of memory) [362-364] and could be misleading (i.e. paracetamol). Furthermore, while batch-effect variations were normalized computationally, it is important to process all the samples in a single experiment.

3.1.8 Conclusion

In summary, this chapter has identified the strengths and limitations, and enhanced the quality of the available data resources to facilitate subsequent analyses. In summary this thesis has adequate and high-quality data resources for primary analyses and independent validations.

Chapter 4: Computational deconvolution analysis of cellular responses using whole blood transcriptomes

4.1 Introduction

This chapter investigates systemic cellular pathway responses associated with pneumonia severity. Whole blood is a complex mixture comprising a wide range of immune cell types, which vary in proportions between samples of different phenotypes. To quantify the variations in the proportions of immune cell types (cellular responses) in severe pneumonia, here I applied a computational approach called *computational deconvolution analysis* (**Figure4.1**). I further sought to enhance the computational performance by applying a data fusion approach to derive an optimal and Integrated Blood Marker List (here on called IBML). IBML provides a single unified marker gene resource for enhanced computational deconvolution of whole blood transcriptomes; and was extensively applied in subsequent analyses.

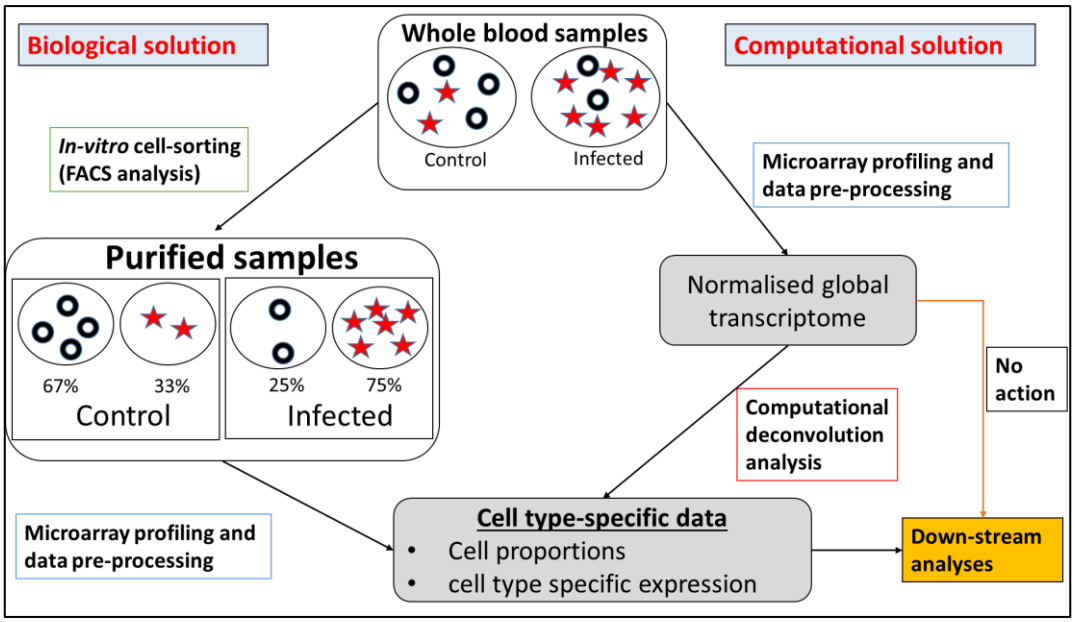


Figure 4.1: Quantifying of cell type-specific information from heterogeneous whole blood samples in transcriptomic analyses. The biological solution (anti-clockwise) involves an intermediate biophysical cell-sorting step before gene expression profiling while the computational solution (clockwise) estimates cell type-specific information directly from

the whole blood transcriptome. However, many studies are oblivious (No action) of the cellular context of whole blood transcriptomes.

4.2 Background

Whole blood is rich and readily available tissue for pathophysiological investigations in biomedical research and clinical practice [71, 102]. Further, the application of whole blood samples in genome-wide profiling studies such as transcriptomics has become a mainstay for discovering key biological pathways, biomarkers and therapeutic targets for a wide range of diseases including infections, cancer and autoimmunity[71, 72]. However, whole blood samples have a complex cellularity including myeloid (i.e. neutrophils, monocytes) and lymphocytes (i.e. T, B cells) immune cell subpopulations, which usually correlate with clinical phenotypes. For example, pneumonia is associated with vigorous recruitment of neutrophils to the lungs [39] and a decrease in the lymphocytes subpopulations[365], thereby changing their proportions in the blood stream[366-370].

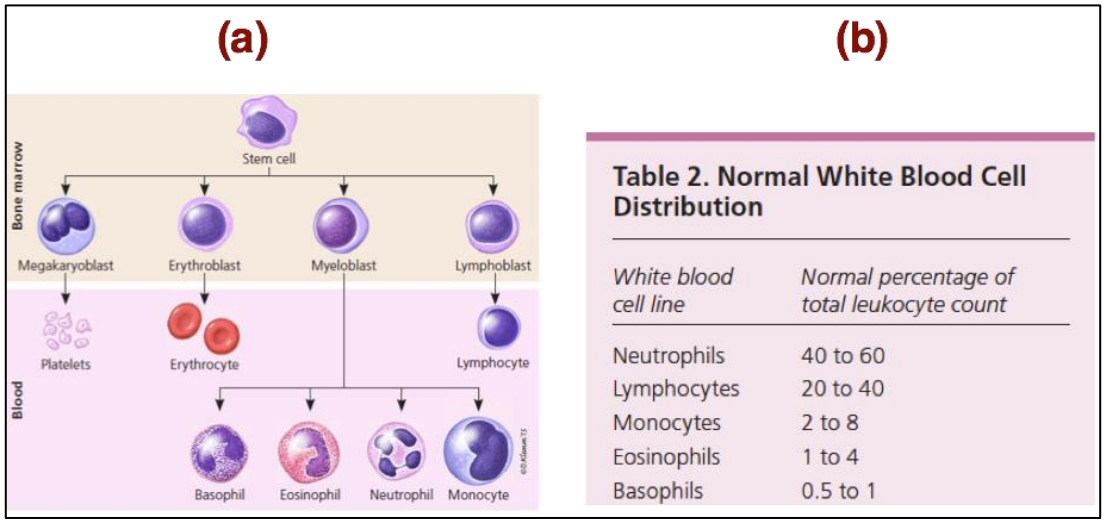


Figure 4.2: White blood cell maturation (a) and normal proportions ranges (b) The figure panels were copied for Riley et. al (2015) : www.aafp.org/afp, *Volume 92, Number 11 ,pages: 1005-1006* [101].

Generally, neutrophils are the most abundant white cells (**Figure4.2**), which increase or decrease in disease [39, 102, 367]. Among the peripheral blood

mononuclear cells (PBMC), the relative proportion of monocytes ranges between 2% and 10%. Within the lymphocytes subpopulations, the relative proportions for T and B cells are 61- 85% and 7–23%, respectively. Further, the CD4+ T-cells to CD8+ T-cells ratio can vary from <1.0 to 2.0 [102-105].

While whole blood transcriptomics (and similar technologies) presents a powerful approach for elucidating systemic host response pathways, ignoring the variations in sample proportions of immune cell types is both a challenge and a missed opportunity[73, 139, 200]. On one hand, these variations potentially confound downstream molecular analyses such that strong signals from less abundant cell types (i.e. NKs, basophils) are diluted or masked by signals from more abundant cell types (i.e. neutrophils). On the other hand, changes in cellular proportions provide an overview state of the immune system such as cell proliferation, differentiation or apoptosis[73]. Therefore, knowledge of the cellular context of whole blood transcriptomes is vital for streamlined analyses and accounting for potential confounding.

Standard biophysical cell-sorting techniques such as magnetic bead sorting, Fluorescence Activated Cell Sorting (FACS) have several limitations [139, 200, 371]. Logistically, these methods require more resources and are timing-consuming. Biologically, cell purification neglects the systemic view of the data and potentially interferes with the gene expression signals consequently introducing another confounding layer [139, 200, 371]. Further, while this thesis has laboratory measurements for cell blood counts of neutrophils and lymphocytes, more detailed data especially for the

lymphocyte compartment (i.e. NK, B and T) were desirable but unfortunately not available.

Computational deconvolution analysis has proved to be a powerful and cost-effective approach for enumerating cell proportions directly from the heterogeneous whole blood transcriptomes[72, 200]. Notably, Abbas et al. (2009) [137], Shannon (2014) [139] and Shen-Orr et al. (2010) [372] successfully deconvoluted whole blood transcriptomes in the contexts of systemic lupus erythematosus, acute kidney allograft rejection and post-kidney transplant, respectively. Further, the *CellMix* toolbox (R package) has compiled a comprehensive open resource comprising algorithms, expression signatures and marker gene lists for different immune cell types, which has facilitated the application of computational deconvolution analysis in whole blood transcriptomics [160].

While expression signatures are often used in *partial deconvolution analyses* [200], marker gene lists for a given cell type are of more general use because they are robust to platform-specific differences [201]. However, the overlap between the existing marker genes lists for a given immune cell type is poor, presenting the end users with a selection challenge. This level of heterogeneity develops inconsistency and variable performance that affects reliability. Further, in this thesis an aggregation of all eligible markers was associated with reduced performance, which suggested the presence of non-specific or noisy markers.

Data science integration approaches provide an opportunity for a step change in the assessment and optimization of these large heterogeneous data resources. To further enhance the computational deconvolution of whole blood transcriptomes, here I applied a comprehensive and unbiased data fusion approach to derive an optimal and integrated blood marker gene lists (IBML). Briefly, IBML provides a unified and optimised single application resource comprising highly specific immune markers from multiple marker gene resources (MGR), which robustly enhanced the prediction of cell type proportions from independent whole blood gene expression data sets.

Subsequently, the IBML resource was applied to deconvolute the pneumonia database, which enabled the investigation of cellular pathway resources in severe pneumonia. In particular, this chapter addressed the following specific objectives:

1. To derive an optimal integrated blood marker gene list (IBML) for enhanced deconvolution of whole blood gene expression data
2. To deconvolute and characterize the cellularity (variations of sample proportions of immune cell types) f whole blood in pneumonia severity
3. To identify cell type-specific molecular differences associated with pneumonia severity

4.3 Results

4.1.1 Optimisation of an integrated blood marker gene list (IBML)

Briefly, IBML was derived to provide a unified, reduced and optimised single marker gene resource (MGR) for enhanced computational deconvolution analysis of human whole blood transcriptomes. To achieve that, three key

steps were involved: (i) selection of eligible markers from eligible marker gene resources (ii) data-driven filtering of eligible markers, and (iii) independent performance assessment (illustrated in **Figure4.3**). Details for each step are outlined in the subsequent subsections.

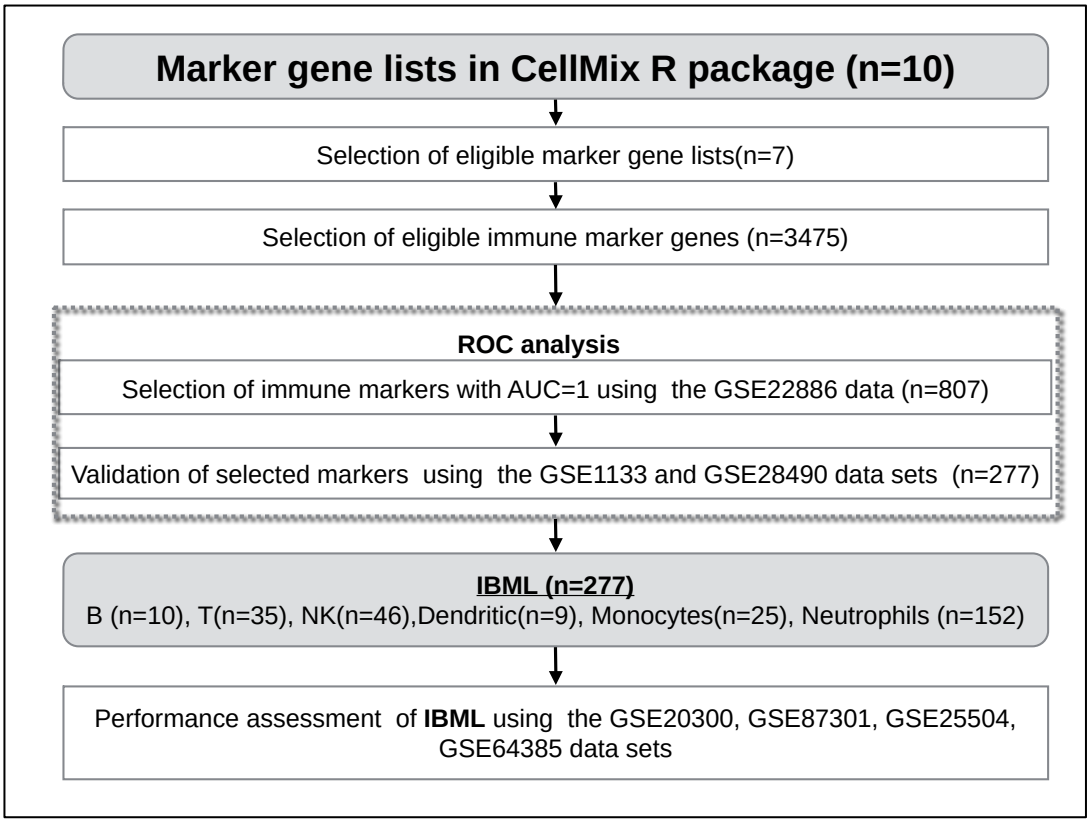


Figure 4.3: Optimisation of an optimal integrated blood marker gene list (IBML)

4.3.1.1 Selection of eligible markers

In this analysis, eligible markers marker genes were selected from the CellMix R package [160], which has compiled comprehensive and accessible resources for computational deconvolution analysis. Here, we focused on valid human marker genes (with corresponding ENTREZIDs) for neutrophils, monocytes, dendritic, NK, T or B cell types. At the time of this analysis (in 2014), the CellMix toolbox [160] had ten MGRs comprising thousands of marker genes for different tissues, organs and immune cell types for different species including human and rat[160, 161, 200]. Here, seven eligible MGRs

(**Table4.1)** comprising 3475 immune marker genes that were associated with the six cell types above were selected**.** To select highly specific markers for each cell type, ROC analysis optimisation was applied (next section).

4.3.1.2 ROC analysis optimisation

The area under the receiver-operating characteristic (ROC) curve (AUC) is a robust measure for predictive performance of quantitative variables on binary outcomes [202, 203]. Briefly, AUC values range between 0 (perfect negative predictor) and 1 (perfect positive predictor) where AUC=0.5 means not better than random discrimination. To select a reduced list of highly specific marker genes (i.e. IBML), AUC values associated with each cell type were calculated for all the eligible markers (n=**3475**). For each marker gene, the expression values of each cell type were compared against the combined average of the other cell types (one-versus-other comparison).

To identify highly specific and robust markers, IBML markers were selected in two steps (selection then validation) using three independent cell-sorted transcriptome databases. In the first step, a total of 807 highly specific cell type markers (with AUC=1) were selected using the **GSE22886** data set [156, 204]. This data was preferred because it is more comprehensive (i.e. has information for more cell types) and has better sample sizes than the other original MGRs. To independently validate the selected markers (n=801), the same approach (AUC=1) was applied using the **GSE1133** and **GSE28490** data sets [156, 205, 206] thereby reducing the final list in IBML to 277 markers for human Neutrophils (152), Monocytes (25), Dendritic (9), NK (46), T (35) and B (10) immune cell types (**Figure4.3 & Table4.1**). To assess

the performance of IBML, independent whole blood transcriptomes were deconvoluted (next section).

Marker gene resources (reference)	Immune cell types					
	B	**T**	**NK**	**Dendritic**	**Monocytes**	**Neutrophils**
Abbas[137]	9	10	3	31	27	16
IRIS[204]	83	74	15	67	72	41
HaemAtlas[358]	175	46	66	Null	185	695
Palmer[102]	231	151	Null	Null	Null	263
Grigoryev[138]	5	30	5	Null	4	2
VeryGene[373]	10	5	17	7	6	Null
CDBlood[160]	1	1	Null	Null	1	2
IBML	**10**	**35**	**46**	**9**	**25**	**152**

Table 4.1: Refined marker genes resources (MGRs). The table shows the distributions of cell type-specific marker genes that were extracted from the CellMix package[160]. The IBML marker genes are presented in the **Appendix A,** and as a supplementary Excel file **(IBMLgenes)**.

4.3.1.3 Performance assessment of IBML

In the previous section, I applied ROC analysis on the purified transcriptomes to derive IBML. To assess the performance of IBML, here I deconvoluted independent whole blood transcriptomes (**GSE20300**[137], **GSE87301**[374]**, GSE25504** [136] and **GSE64385** [375]), which are heterogeneous. In particular, these data sets were chosen because they have existing laboratory-measured proportions of immune cell types to enable direct comparison between the predicted and reference values. Firstly, five algorithms (*DSA [371]*, *ssKL*, *meanProfile* and *ssFrobenius*) [160, 201] were assessed, and the ***ssFrobenius*** algorithm was associated with the highest performance. Using that algorithm (*ssFrobenius*), the predicted sample proportions of immune cell types were directly compared with the existing standard values (laboratory measured) using the Pearson correlation coefficients (r). To identify the optimal list between the IBML and the original MGRs (K=7), the cell type-specific r-values were compared across the MGRs (**Figure4.4**).

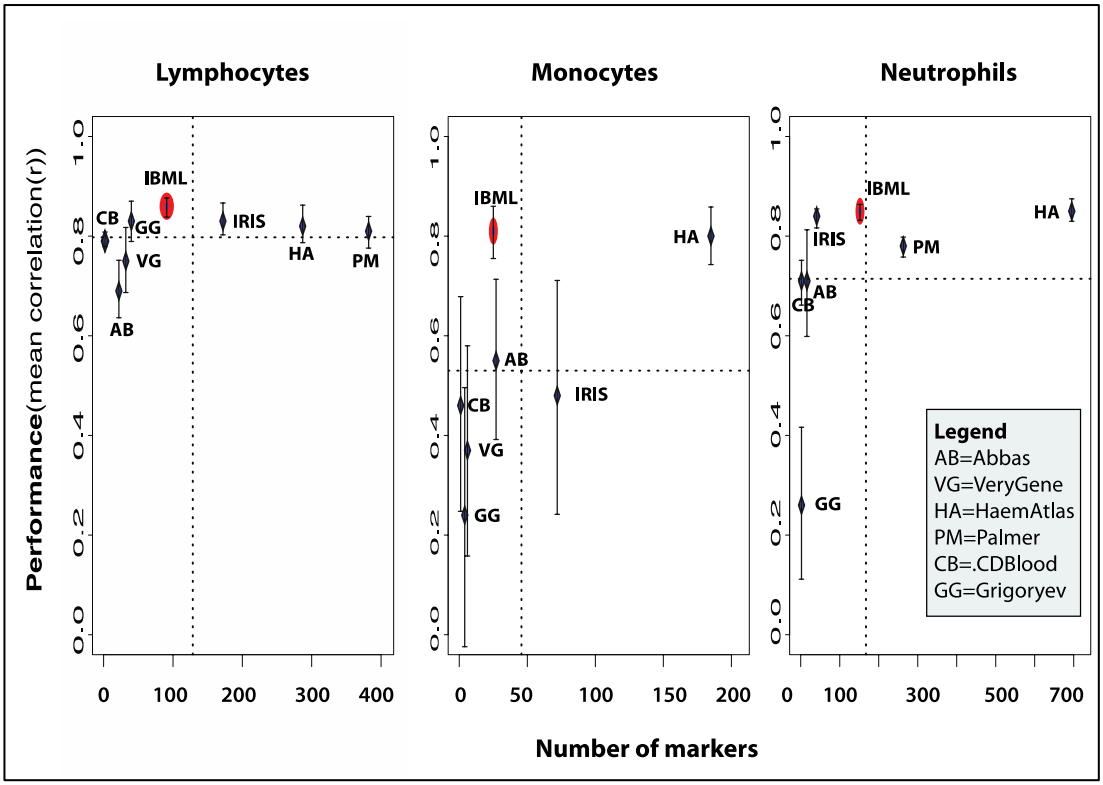


Figure 4.4: Comparative performance assessment of IBML. Each dot (and error bar) represents a mean (and standard errors) value for Pearson correlations between laboratory-measured and computationally estimated proportions of cell types across the benchmark data sets. The dotted lines represent the average values across the MGRs (i.e. vertical=the mean number of markers; horizontal=the mean performance).

Ideally, an optimal MGR should have a minimal size but associated with high performance (top left quadrant of **Figure4.4**). While the *IRIS* and *HaemAtlas* resources were associated with the highest performance among the original MGRs, the later had the highest number of markers (>1000) and the former was inferior in monocytes (hence not optimal). On the other hand, IBML has minimal number of markers but robustly associated with high performance (R≥0.8) in all the cell types (top left quadrant). Together, these findings suggest that the IBML provides a unified, optimal and robust candidate marker gene resource for enhanced computational deconvolution analysis of whole blood transcriptomes. Subsequently, the IBML resource was applied on the pneumonia database (next section).

4.1.2 Implementation of IBML in the pneumonia database

As already mentioned, this chapter has sought to apply a computational approach to investigate cellular pathway responses in pneumonia severity. So far, I have applied independent public data resources (marker gene lists, purified expression data and whole blood expression data) to derive an optimal marker gene resource (IBML) for enhanced computational deconvolution analysis of whole blood transcriptomes. To validate the performance of IBML in the pneumonia database, the same approach (correlation analysis) was applied (next section).

4.3.1.4 Performance validation of IBML in the pneumonia database

Firstly, IBML was applied to deconvolute sample proportions of immune cell types (T, B, NK, dendritic, monocytes and neutrophils) from the pneumonia whole blood transcriptome. It is worth noting that this database has existing laboratory-measured cell proportions (complete blood counts) for neutrophils and lymphocytes. To validate the performance of IBML in this database, the deconvoluted proportions were directly compared to the corresponding laboratory-measured values using the Pearson's correlation coefficient (r). Interestingly, the performance of IBML remained high (**r≥0.83**) in both cell types (**Figure4.5**). This finding further suggests the robustness and applicability of IBML in the pneumonia database. Subsequently, IBML and the deconvoluted proportions of immune cell types were applied to investigate the cellular pathway responses (next sections).

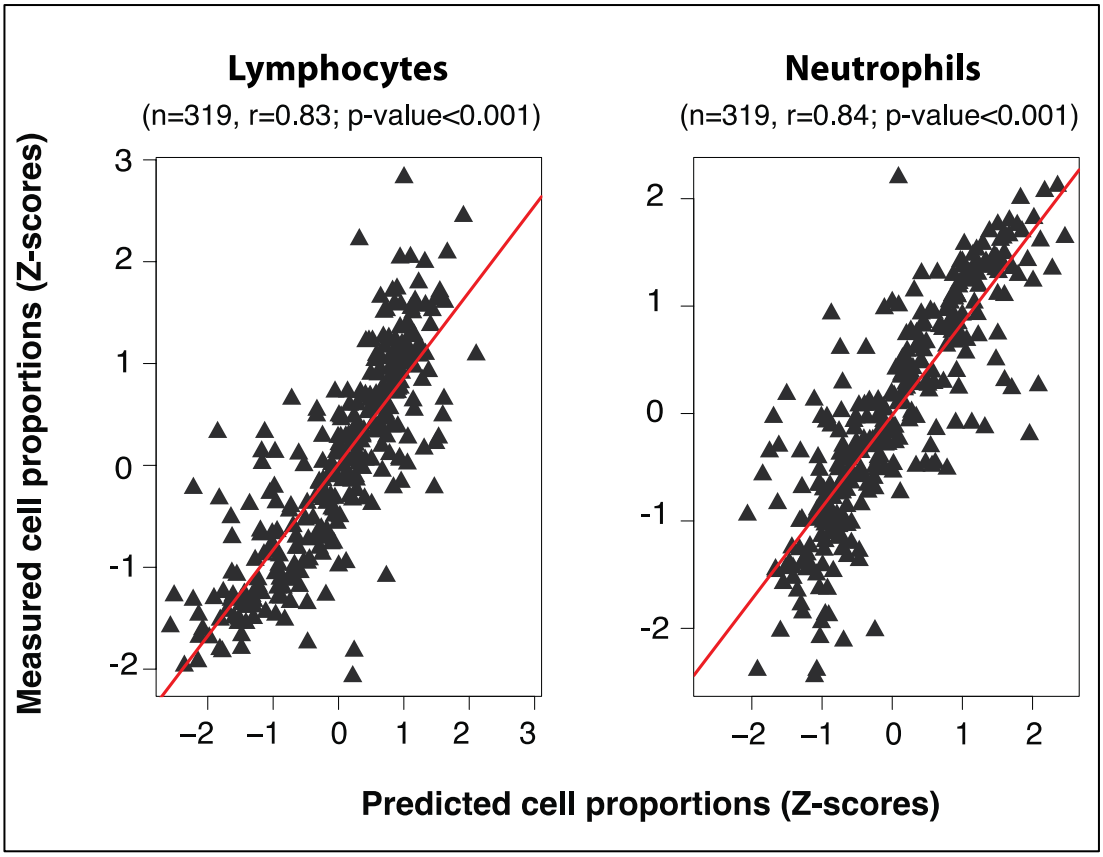


Figure 4.5: Independent validation of IBML using the pneumonia database: The figure shows scatter plots between the deconvoluted (x-axis) and laboratory-measured (y-axis) sample proportions of lymphocyte (left) and neutrophil immune cell types. To achieve the same scale between the x and y axes, the proportion values were standardized into Z-scores (i.e. each value was subtracted the mean and divided by the standard deviation). Each triangle in the main figures represent a sample, and n=sample size, r=Pearson's correlation coefficient; P-value =correlation test (H_0:r=0).

4.1.2.1 Age-dependent variations in the proportions of immune cell types

In **Chapter 3**, age was molecularly associated with elevated baseline status of the pro-inflammatory innate responses among the non-pneumonia controls. To investigate the corresponding cellular response levels, here I assessed whether the deconvoluted proportions of the immune cell types varied significantly with age. To account for the confounding effects of pneumonia severity and malnutrition, this analysis was restricted to the non-pneumonia controls (n=120), and linear regression analysis approach was applied to adjust for nutrition status (**Figure4.6)**

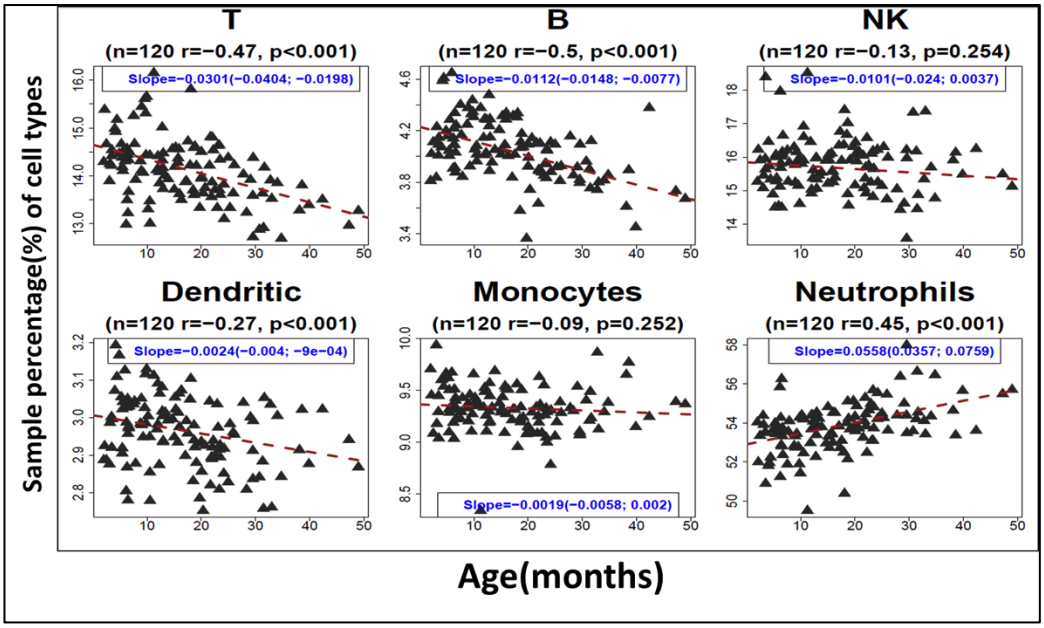


Figure 4.6: The associations between age (x-axis) and sample proportions of immune cell types (y-axis). In each scatter plot, the dots represent samples, r= partial correlation coefficient (Pearson's) adjusted for nutrition status; slope= linear regression coefficient and its 95% confidence interval, n=sample size and p=P-value for the significance for the slope.

While the data points show high variability, age was indeed associated with significant (P-value<0.001) variations in the proportions of B, T, Dendritic and neutrophils (**Figure4.6**). In particular, the proportions of neutrophils (myeloid) increased with age (r=0.45) and vice-versa for the adaptive response mediators: B(r=-0.50), T (r=-0.47) and dendritic cells (r=-27). These findings are consistent with the molecular findings in **Chapter 3** and similar previous studies. Notably, Burr et. al (2011) reported age-dependent maturation of Toll-like receptor (TLR)-mediated cytokine responses in healthy Gambian Infants[357]. Further, Mandala et al. (2010) also observed a significant age-dependent reduction in lymphocytes but not NK cells among healthy Malawian children (n= 539) [376]. While these results were anticipated, and due to natural ontogeny and environmental factors. These findings highlight the importance of adjusting for age differences in the primary analyses of this thesis. In the subsequent sections, cellular responses associated with pneumonia severity were investigated.

4.3.1.5 Associations between sample proportions of immune cell types and pneumonia severity

Clinical phenotypes such as pneumonia severity reflect the underlying host responses from different immune cell types [53]. To investigate the systemic cellular pathway responses in severe pneumonia, this section assessed the variations in the deconvoluted proportions of immune cell types across the pneumonia severity groups. For each cell type, the linear regression analysis approach was applied to quantify the association between pneumonia severity and the deconvoluted proportions (F-test P-values) while adjusting for the potential confounders (age, nutrition status and antibiotic usage).

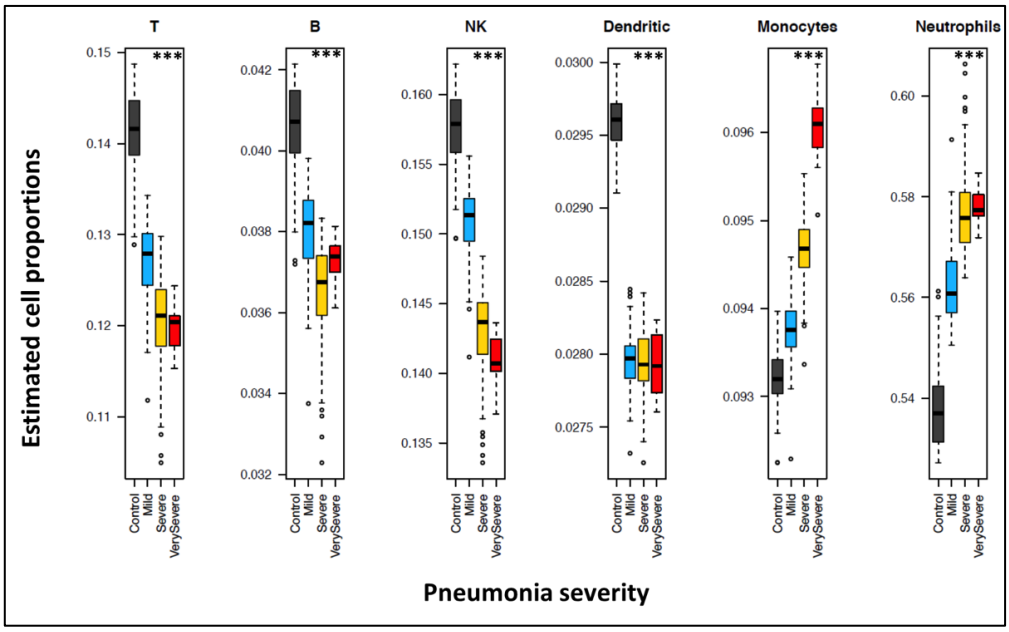


Figure 4.7: Deconvoluted sample proportions of immune cell types. Linear regression analysis was applied to assess the association between the sample proportions of each cell type (outcome variable) and pneumonia severity while adjusting for potential confounders (age, nutrition status and antibiotic usage). The box and whisker plots show the distribution of the adjusted proportions across the pneumonia severity groups (x-axis). The indicated P-values are for the adjusted F-tests. Abbreviations: NK=natural killer cells, *** = P-value<0.0001

In average, neutrophils and dendritic cells represented the highest and lowest proportions, respectively (**Figure4.7**). Comparatively, pneumonia severity was associated with significant variations in the sample proportions of all the six cell types (P-value<0.0001). In particular, the depletion of

lymphocytes (B, T and NK) and dendritic cell types; and the elevation of neutrophils and monocytes levels were associated with increased pneumonia severity (respectively). Notably, while the elevation of myeloid (neutrophils and monocytes) and depletion of adaptive lymphoid (B and T cells) cell subpopulations (respectively) are frequently reported [367, 368, 377, 378], the unexpected potential involvement of human natural killer (NK) cells in the pathogenesis of severe pneumonia remains elusive [379-382]. Potentially, this finding presents a novel therapeutic target for immune-modulation and management of severe pneumonia cases. Further, it is also worth noting that the cellular proportions distinguished the pneumonia severity groups suggesting the potential of cellular pathway-based biomarkers in the stratification of pneumonia cases (**Chapter 6**). Together, these findings highlight the potential of computational deconvolution analysis, and support the hypothesis that cellular pathway responses underpin the development of severe pneumonia states. However, further studies are required to gain a deeper insight into involvement of NK cells in pneumonia. In the next sections, cell type-specific molecular responses were investigated.

4.1.2.2 *Identification of differentially expressed marker genes in severe pneumonia*

Ideally, an investigation of cell type-specific molecular responses can provide a deeper insight into the pathogenesis of severe pneumonia. It is expected that a marker gene for a given cell type is exclusively expressed in that particular cell type [161, 371]. Therefore, immune marker genes resources like the **IBML** provide a reliable estimate of cell type-specific expression profiles. To partially understand the differential gene expression profiles associated with pneumonia severity at cellular level, immune markers in the IBML resource (n=277) were subjected to empirical Bayes moderated t-test to identify differentially expressed genes (FDR<0.05, |FC|≥1.5) between the non-pneumonia controls and severe pneumonia groups (**Figure4.8**).

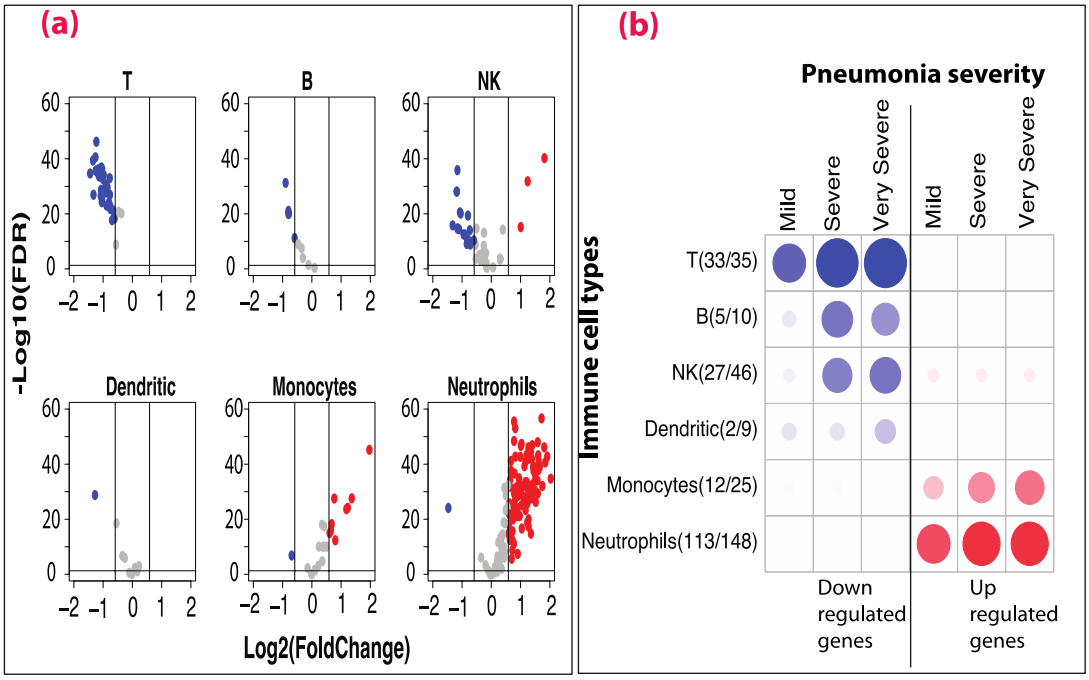
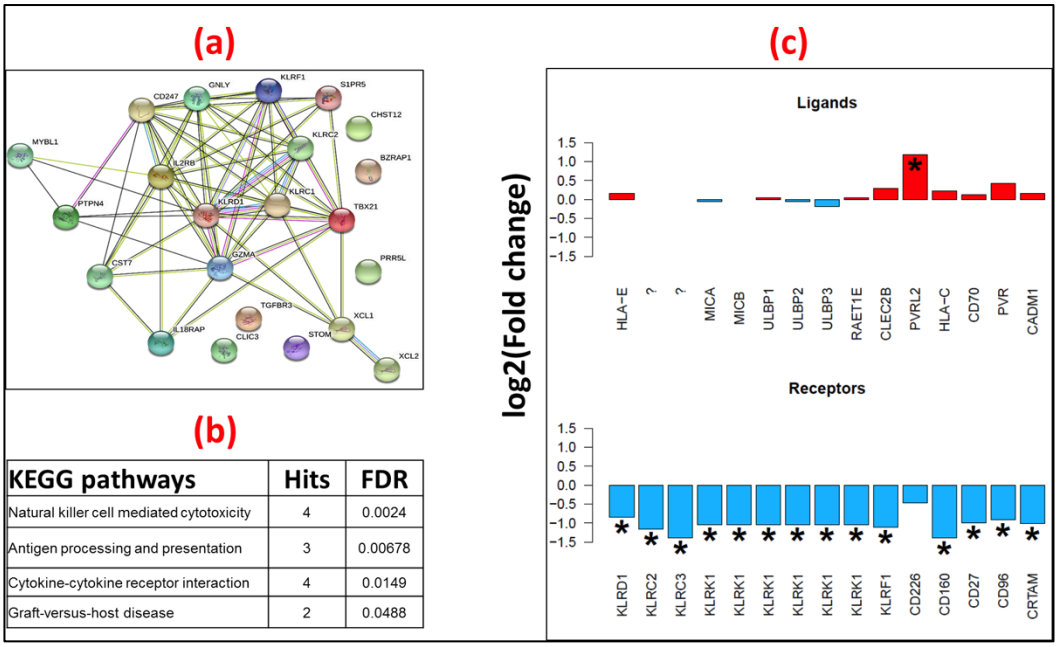


Figure 4.8: Differentially expressed immune marker genes in pneumonia severity. (a) Volcano plots showing up (red) or down (blue) regulated (FDR<0.05, |FC|≥1.5) cell type-specific marker genes (in IBML) between non-pneumonia controls (n=120) and pneumonia cases (n=225). **(b)** Correlogram summary showing the proportions of up (red) and down (blue) regulated genes between non-pneumonia controls (n=120) and each severity group, respectively (mild (n=90), severe (n=117), very severe (n=18)). The fractions of differentially expressed makers in at least one severity state are annotated on the x-axis (i.e. 33 of the 35 T cell markers were down-regulated at least in any of the severity groups)

Interestingly, the molecular responses are consistent with the distribution of the cellular proportions. In overall, the number of cell type-specific differentially expressed genes increased with pneumonia severity (**Figure4.8b**). In particular, pneumonia severity correlated with the up-regulation of monocytes and neutrophils markers, and the down-regulation of markers for NK, dendritic, B and T cells. Notably, 59%(m=27) of the NK markers were potentially involved in pneumonia severity. Specifically, while three markers (CST7, IL18RAP and STOM) were up regulated (**Figure4.8**), 24 (52%) markers were consistently down-regulated with pneumonia severity. The down regulation of NK markers with pneumonia further suggests a potential protective role of natural killer cells in pneumonia. To investigate further, the next section assessed the molecular functions associated with the NK cell markers

4.1.3 Functional analysis of natural killer (NK) cell markers associated with pneumonia severity

It is worth noting that the current study design is not adequate to gain a deeper understanding of the role of NK cells in pneumonia severity. Nevertheless, here I partially assessed whether pneumonia severity is associated with NK-specific molecular functions. Firstly, I applied the STRINGs database [280] to assess whether the significant NK markers (n=27) were associated with any protein-protein functional network (**Figure4.9a-b**). Interestingly, these genes were principally associated with the down regulation of the natural killer cytotoxicity network (**Figure4.9a-b**).



KEGG pathways	Hits	FDR
Natural killer cell mediated cytotoxicity	4	0.0024
Antigen processing and presentation	3	0.00678
Cytokine-cytokine receptor interaction	4	0.0149
Graft-versus-host disease	2	0.0488

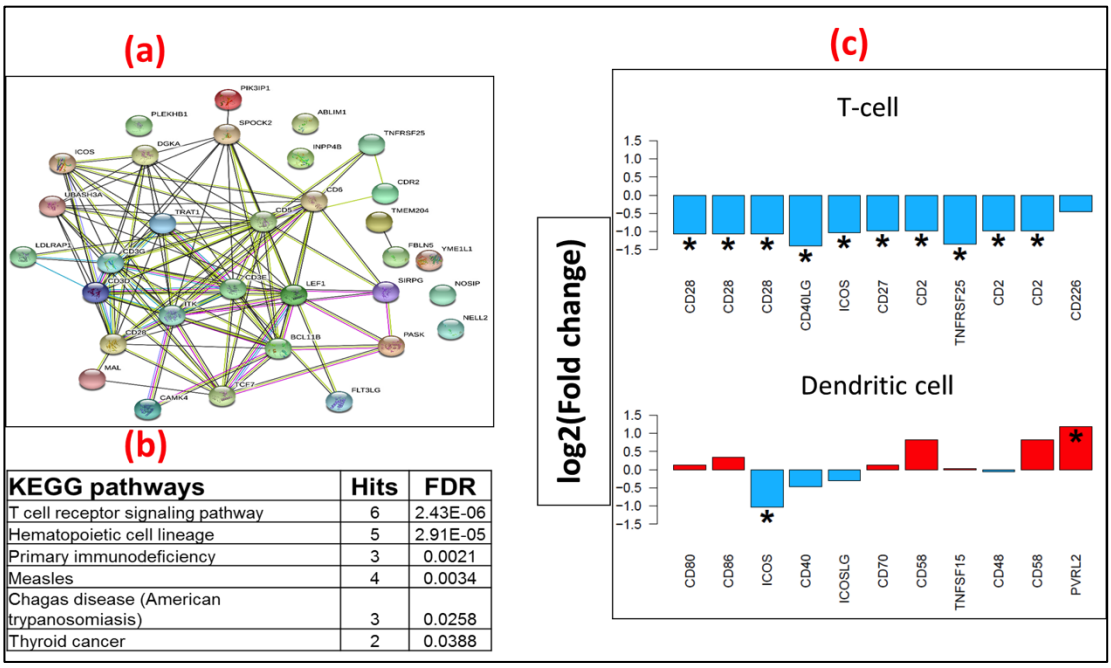
Figure 4.9: Molecular functions associated with differentially expressed NK markers in pneumonia severity. **(a)** Protein-protein interaction network (STRING database [280]) for differentially expressed NK markers (n=27) between non-pneumonia controls and severe pneumonia groups. **(b)** Enriched pathways (FDR<0.05) in KEGG biomedical database: FDR=False discovery rate **(c)** Gene expression fold changes between no-pneumonia controls and all pneumonia cases for activating receptors on NK cells and their corresponding ligands: *= differentially expressed gene (FDR<0.05, |FC|≥2).

Further, to assess the co-stimulatory functions of NK cells in pneumonia, I assessed the expression profiles of the validated NK receptors and their corresponding ligands [383, 384]. Notably, significant down regulation (FDR<0.05, |FC|>2) of the co-stimulatory receptors on the NK surface were associated with non-significant gene expressions on the corresponding ligands (**Figure4.9c**). On the other hand, non-significant down-regulation of the CD266 receptors was associated with the significant up-regulation (FDR<0.05, |FC|>2) of its corresponding ligand (PVRL2) [383, 384]. Together, these findings further suggest the previously unknown involvement of NK cells in pneumonia. Nevertheless, these predictions require future experimental validation studies. In the next section, the same approach was applied on the T cells.

4.1.4 Investigating the cross-talk between dendritic and T cells in pneumonia

Dendritic cells are professional antigen presentation cells (APCs) bridging the innate and adaptive immune arms through co-stimulation and co-inhibition of T cells [385, 386]. Here, the depletion of both cell types and down-regulation of their markers were significantly associated with pneumonia severity. To assess whether pneumonia severity was associated with T-cell specific protein-protein associated functional networks, differentially expressed T cell markers (n=33) were subjected to functional analysis using the STRING database [280] (**Figure4.10a-b**). Principally, pneumonia severity was associated with a connected functional network of T cell receptor signalling pathway (FDR<0.001).

To investigate the co-stimulatory cross-talk between dendritic and T cells associated with pneumonia severity, I assessed the differential gene expression profiles for the corresponding receptors and ligands between the two cell types (**Figuer4.10c**). Interestingly, non-significant change of expression or significant down-regulation (FDR<0.05, |FC|>2) of the activating receptors on dendritic cell were associated with significant down regulation of the corresponding ligands on T-cells, highlighting the positive regulatory role of dendritic cells on T cell activities in the pathogenesis of pneumonia. Together, these analyses provide insights into cellular pathway responses in severe pneumonia; and support the central hypothesis that systemic pathway (cellular) responses underpin the development of severe pneumonia outcomes.



KEGG pathways	Hits	FDR
T cell receptor signaling pathway	6	2.43E-06
Hematopoietic cell lineage	5	2.91E-05
Primary immunodeficiency	3	0.0021
Measles	4	0.0034
Chagas disease (American trypanosomiasis)	3	0.0258
Thyroid cancer	2	0.0388

Figure 4.10: Molecular functions associated with differentially expressed T markers in pneumonia severity. **(a)** Protein-protein interaction network (STRING database [280]) for down-regulated T markers (n=33) between non-pneumonia controls and severe pneumonia groups. **(b)** Enriched pathways (FDR<0.05) in KEGG biomedical database: FDR=False discovery rate **(c)** Gene expression fold changes between no-pneumonia controls and all pneumonia cases for co-stimulatory receptors (dendritic cell) and their corresponding ligands (T cell): *= differentially expressed gene (FDR<0.005, |FC|≥2).

4.4 Discussion

This chapter has applied whole a blood transcriptome to investigate the systemic cellular pathway responses in severe in pneumonia. While whole blood is readily accessible and widely applied for genomic research and routine clinical practice [123], careful considerations are required to quantify and account for its complex cellularity that often correlate with clinical phenotypes[102]. However, the quantification of cell type specific information remains a challenge [200, 371]. Standard biophysical cell purification approaches are very expensive, time-consuming and have limited biological interpretation due to potential confounding and lack of the systemic perspective [73, 102, 139, 200, 371]. Further, this study has limited information on the cellularity of the whole blood samples.

To investigate the cellular pathway responses in pneumonia, here I applied a powerful and cost-effective computational solution called computational deconvolution analysis. Notably, this chapter has (i) generated an optimised marker gene resource (IBML) for enhanced deconvolution of whole blood transcriptomes and (ii) provided insights into the unexpected potential involvement of NK cells in the pathogenesis of severe pneumonia, and could be a potential target for future intervention in severe pneumonia.

4.1.5 Strengths and limitations of computational deconvolution analysis

While computational deconvolution analysis is a powerful and attractive alternative to the standard laboratory methods, it has some limitations. Generally, standard deconvolution methods require knowledge of cellular proportions to estimate cell type-specific expression profiles (*direct partial*

deconvolution) or the vice-versa (*reverse partial deconvolution*), which are rarely available and platform-specific[200]. Alternatively, *semi-supervised deconvolution algorithms are more attractive because they only require* marker genes, which are robust to platform-specific differences [201, 371]. However, existing marker gene resources have little overlap for a given cell type, vary in performance, and are potentially contaminated by noisy markers, which is another challenge for end-users.

Here, I investigated the hypothesis that a unified, optimal and reliable marker genes resource would enhance computational deconvolution analysis of whole blood transcriptomes [200, 204]. To address this, I applied a data fusion approach to derive a unified and optimized marker gene lists called IMBL (Integrated blood marker list). Relative to its reduced size, IBML was associated with robust and enhanced performance than the individual MGRs. Importantly, this analysis provides the first comprehensive comparative performance assessment of the existing marker gene resources (MGRs) applicable to computational deconvolution analysis of whole blood. While the IRIS marker list [204] is often applied[371, 387], here *HaemAtlas*[358] was associated with the best performance among original MGRs. Nevertheless, IBML provides a simplified and unified single application resource for enhanced computational deconvolution of whole blood transcriptomes.

However, to maximize the potential of computational deconvolution, marker genes resources (MGRs) with more cell type coverage and granularity

comprising markers for different activation states are required. In particular, deep deconvolution of specific cell subcomponents such as regulatory T cells (Tregs); and less abundant cell types such as basophils, eosinophils and mast cells would provide comprehensive insights into the pathogenesis of diseases. However, lack of comprehensive benchmark data sets (i.e. gene expression profiles for purified cell types) remains the challenge[200]. Further, while ROC analysis was applied to mitigate the challenge of limited sample sizes, derivation of IBML would have benefited from more sufficiently-powered studies. Nevertheless (according to this analysis), IBML provides the optimal choice among the existing marker genes resources.

4.1.6 Biological insights from computational deconvolution analysis

To assess whether pneumonia severity is associated with systemic cellular pathway responses, IBML was applied to deconvolute the whole blood transcriptome. While adjusting for potential confounders including age, pneumonia severity was significantly associated with the elevation of markers for myeloid cells (neutrophils and monocytes) and the depletion of B, T, Dendritic and NK cells respectively. Further, the elevation and depletion of immune cell types were consistently associated with the up or down-regulation of their molecular markers (respectively). Together, these findings consistently support the central hypothesis that systemic pathway responses underpin the development of severe pneumonia.

While the elevation of myeloid (monocytes and neutrophils) and depletion of adaptive (B and T cells) cells are frequently associated with inflammatory responses[367, 370, 378], the depletion of NK and dendritic cells were not

anticipated and somehow controversial[379, 388]. Dendritic cells are professional antigen presentation cells (APCs) mainly responsible for priming the T cells [385, 386]. Depletion of dendritic cells may partly contribute to the depletion or malfunction of T cells in the blood. Alternatively, T cell depletion may be due to other factors such as apoptosis, necrosis or migration of T cells to the lungs [116].

This analysis has identified a novel involvement of NK cells in pneumonia severity. Natural killer (NK) cells are the innate lymphoid cells well known for their cytotoxicity against tumours and virally infected cells [383]. However, recent studies have revealed both pro and anti-inflammatory regulatory roles such as (i) sending negative signals to primed macrophages[389], (ii) killing or promoting the maturation of dendritic cells[390], (iii) suppressing auto-reactive B cells[391], and (iv) promoting differentiation of CD4+ T cells or killing primed T cells [383, 392-396]. In that regard, this finding suggests the novel central role of NK cells in the pathogenesis of pneumonia including positive regulation of dendritic cells [392, 395, 397], and negative regulations of inflammation[383]. In mice models, depletion of NKs is also associated with severe outcomes [398-400] further suggesting the involvement of NK cells in pneumonia. The number of NK cells was not age dependent in the non-pneumonia controls suggesting their specific involvement in pneumonia (**Figure4.6**). These findings require further validation preferably in a longitudinal study.

4.5 Conclusion

In summary, the results presented in chapter highlights the potential and the challenges of a computational deconvolution analysis for streamlining analysis of whole blood transcriptomes. In conclusion, while further investigations are required to elucidate the novel role of NK cells in the pathogenesis of pneumonia, these findings support the central hypothesis that systematic cellular pathway responses underpin the pathogenesis of pneumonia. Potentially, the NK findings present a novel target for immune-modulation and clinical management of pneumonia cases.

Chapter 5: Computational investigation of systemic molecular pathway responses in severe pneumonia

5.1 Introduction

This thesis has investigated systemic pathway responses associated with pneumonia severity, both at the cellular (**Chapter 4**) and molecular (**Chapter 5**) analytic levels In particular, this chapter has applied a range of validated biomedical pathway databases to investigate the molecular pathway responses associated with pneumonia severity.

5.2 Background

Pneumonia is an inflammatory disease of the lung parenchyma (lower respiratory tract system) causing significant morbidity and mortality worldwide [2]. The respiratory tract system is crucial for gas exchange between the body and outside environment. However, this task brings constant exposure to potential pathogens in the air [31]. To ensure normal gas exchange, colonization of the upper respiratory tract (i.e. by commensals) is often tolerated[10, 35] while the lower tract is strictly guarded by physical and chemical barriers. When these barriers are evaded, immune responses are induced to eliminate the invading pathogen [78].

It is worth noting that host responses in pneumonia are investigated at either *"local"* (within the lung) or *"systemic"* (in the blood stream) levels[80, 81]. In non-severe cases, local responses are mainly important, and are tightly regulated and compartmentalized within the lungs to prevent systemic responses, which can be associated with deleterious inflammatory outcomes

[31]. However, due to host (virulence) and pathogen (susceptibility) factors, these barriers can be breached consequently inducing the systemic responses in the circulating blood [31, 39, 115, 116]. Thus, while the local responses are sufficient in mild cases, systemic responses are vital for outcomes of severe pneumonia outcomes [115]. Therefore, blood-based immune signatures should mainly reflect the key components of the pathway responses in severe pneumonia. Moreover, changes in the blood reflect key components of host responses for not only targeted diseases, but also the entire body including the lungs [71]. Importantly, whole blood is clinically accessible for pathophysiological investigations [102].

While whole blood genome-wide profiling provides a comprehensive and powerful approach for investigating systemic responses, single-gene analysis approaches have failed to realize the potential of such multi-dimensional and highly correlated data [255, 256]. Host responses often involve multiple immune mediators inducing a cascade of signalling events, which are manifested in clinical phenotypes such severe pneumonia [36, 37, 99, 122, 401]. However, these univariate approaches often generate long lists of candidate genes, which lack the systems-level perspective of the immune system and are very challenging to interpret. Further, such lists are potentially confounded by false discoveries due to multiple testing, and consequently lack stability (little overlap) between similar studies*[254, 402]*.

To gain more insight into the pathogenesis of severe pneumonia, here I have applied a pathway analysis approach where the quantum of analysis is a set

of genes involved in a particular biomedical pathway such as Toll-like receptor signalling pathway [157, 249, 255, 403]. This approach incorporates existing biological knowledge to condense and contextualize a long list of candidate genes into a short list of meaningful biochemical processes. Notable biomedical pathway databases include the Gene Ontology (GO), KEGG and REACTOME [143, 404]. Moreover, pathway-based approaches present an opportunity for robust biomarkers, which capture the disease pathogenesis [65, 274, 405].

To identify molecular pathway responses associated with pneumonia severity, the following specific objectives were investigated using the training transcriptome (n=345) and a range of biomedical pathway databases:

1. To identify molecular pathway responses that were uniquely associated with severe pneumonia outcomes
2. To identify potentially prognostic pathways that were associated pneumonia from mild to very severe outcomes.
3. To assess whether very severe pneumonia cases were associated with unique molecular pathway responses

5.3 Approach

To address the objectives above, differentially expressed genes (DEGs) between non-pneumonia controls and severe pneumonia states (mild, severe and very severe, respectively) were classified into the following unique subsets corresponding to each specific objective (**Figure5.1**):

1. **MSvS**: DEGs that were jointly associated with all the pneumonia states (objective two)
2. **SvS**: DEGs that were associated with the development of severe and very severe pneumonia states (objective one)
3. **Vs**: DEGS that were uniquely associated with the development of very severe pneumonia (objective three)

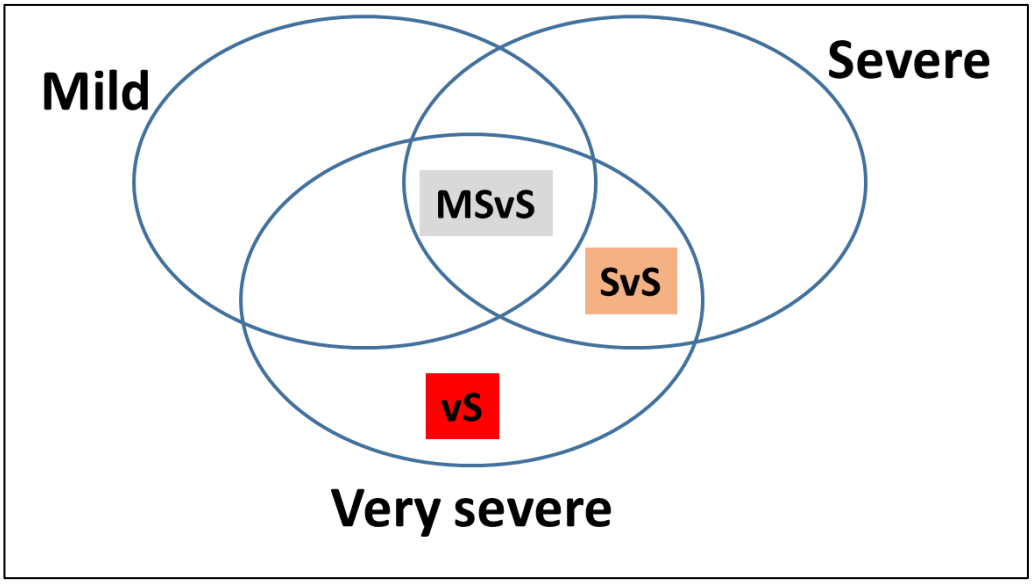


Figure 5.1: Subsets of differentially expressed genes (DEGs) that were applied to assess molecular pathways associated pneumonia severity. Each circle represents the set of DEGs between non-pneumonia controls and each severity state. MSvS=DEGs shared by mild, severe, very severe pneumonia states; SvS= DEGs shared by severe and very severe pneumonia states, and Vs= DEGs unique to very severe pneumonia.

Subsequently, up and down-regulated genes in each subset (**MSvS**, **SvS** and **vS**) were applied to identify enriched pathways (FDR<0.05) using the following biochemical pathway databases: KEGG, REACTOME, HALLMARK

and Gene Ontology (GO). The databases were downloaded from the Molecular Signature DataBase *(MSigDB)* repository (http://software.broadinstitute.org/gsea/msigdb) *[254, 402].* Here, multiple pathway databases were applied to gain a comprehensive view of the molecular pathway responses in pneumonia severity, and to account for the heterogeneity of the existing biochemical pathway resources. To identify significant pathways, Fisher's exact test for association was applied (using the *stats* package in R [158]). To guard against false discoveries, the raw P-values were adjusted for multiple testing across the pathways in each database using the Benjamini-Hochberg method [238].

5.4 Results

5.4.1 Single-gene analysis: Identification of differentially expressed genes in pneumonia

To investigate the molecular pathway responses in severe pneumonia, a single-gene analysis was done to identify the candidate gene sets (illustrated in **Figure5.1**) for subsequent pathway analyses. Firstly, an, empirical Bayes moderated t-test was applied (*limma* Bioconductor package [226]) to (i) identify differentially expressed genes between non-pneumonia controls and severe pneumonia states (mild, severe and very severe, respectively) while adjusting for potential confounders (age and nutrition status and antibiotic usage) and false discoveries due to multiple testing (BH method[238]).

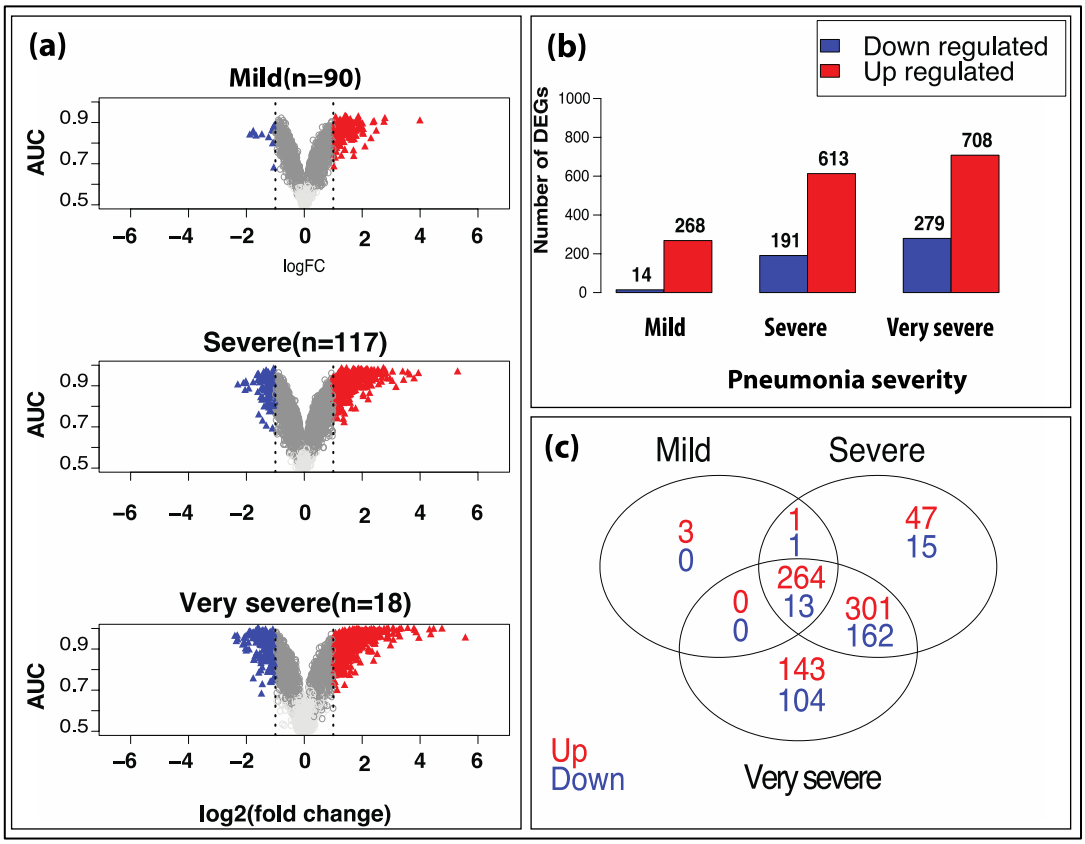


Figure 5.2: Differential gene expression profiles between severe pneumonia states and non-pneumonia controls (n=120). **(a)** Volcano plots showing the fold changes for up (red) and down (blue) regulated genes (|fold change|≥2, FDR<0.05): AUC=area under the ROC curve. **(b)** A bar plot showing the trend (number) of up and down regulated genes. **(c)** A Venn diagram showing the overlaps of differentially expressed genes between severe pneumonia states. The lists of differentially expressed genes with pneumonia severity are presented as a supplementary Excel file (DEGsPneumonia).

Of the 11037 genes that passed the non-specific filtering criteria (**Chapter 3**), 295 and 759 genes were significantly (FDR<0.05 and |FC|≥2) down and up regulated (respectively) in at least one pneumoina severity state (**Figure 5.2**). Notably, the fold changes (**Figure 5.2a**) and the number of diferentially expressed genes (**Figure 5.2b**) increased with pneumoina severity. At the gene the level, this finding supports the current hypothesis that sytemic molecular responses are associated with pneumonia severity.

Based on the contrasts defined in **Figure5.1**, 277 (**MSvS** subset), 463 (**SvS** subset) and 247 (**Vs** subset) were eligible for the subsequent pathway

analyses (**Figure5.2c**). In particular, 277 genes (13 down-regulated and 266 up-regulated) were associated with all the severe pneumonia states from mild to very severe (**MSvS** subset). Notably, very few genes (m=13) were down regulated in mild pneumonia potentially suggesting the importance of inhibitory responses in the development of severe pneumonia outcomes. Further, 463 genes (162 down-regulated and 301 up-regulated) were uniquely associated with severe pneumonia outcomes (both severe and very severe pneumonia). Finally, 247 genes (104 down-regulated and 143 up-regulated) were uniquely associated with very severe pneumonia state (**Vs** subset). In the next sections, these gene sets (**MSvS, SvS** and **vS**) were applied for pathway analyses to investigate the systemic molecular pathway responses in pneumonia severity.

5.4.2 Systemic molecular pathway responses in mild pneumonia

In this section, I investigated the pathways that were associated with the up (m=264) and down (m=13) regulated genes in all the severe pneumonia states from mild to very severe pneumonia (**MSvS** subset). While the original study design does not allow a formal analysis for prognostic biomarkers (i.e. samples were collected at a single time point), this contrast presents an opportunity to identify potentially prognostic pathways for early identification and monitoring of high-risk pneumonia cases. However, the down-regulated genes (m=13) were too small for a proper pathway analysis. Further, using the STRINGS database [280], this list was not associated with any functionally connected protein-protein networks. Instead, the individual genes were characterized (next section).

5.4.2.1 Investigation of down-regulated genes in mild pneumonia (m=13)

Table5.1 presents the potential molecular functions for the down-regulated genes in mild pneumonia (m=13). Mainly, this subset was associated with the down-regulation of cell adhesion molecules, T cell functions (CCR3, CLC, and LRRN3) and allergic responses. While the interpretation of a single-gene analysis is limited, the inhibition of T cells is consistent with the cellular pathway responses (**Chapter 4**). In the next section, the up-regulated pathways were investigated.

SYMBOL	Description	Functions
CCR3	C-C Motif Chemokine Receptor 3	Epithelial CCR3 mediates the release of IL8 through LPS-induced lung inflammation[406] Expressed by T lymphocytes co-localizing with eosinophils in allergic inflammation[407]
CDHR3	Cadherin Related Family Member 3	Cadherins are calcium-dependent cell adhesion proteins
CLC	Charcot-Leyden crystal protein	Regulates immune responses through the recognition of cell-surface glycans. Essential for the energy and suppressive function of CD25-positive regulatory T-cells (Treg)
CTGF	Connective tissue growth factor	Major connective tissue mitoattractant secreted by vascular endothelial cells. Promotes proliferation and differentiation of chondrocytes. Mediates heparin- and divalent cation-dependent cell adhesion in many cell types including fibroblasts, myofibroblasts, endothelial and epithelial cells. Enhances fibroblast growth factor-induced DNA synthesis
DPH6	Diphthamine Biosynthesis 6	Amidase that catalyzes the last step of diphthamide biosynthesis using ammonium and ATP. Diphthamide biosynthesis consists in the conversion of an L-histidine residue in the translation elongation factor (EEF2) to diphthamide (By similarity) [124, 408].
FCER1A	Fc fragment of IgE, high affinity I	Binds to the Fc region of immunoglobulins epsilon. High affinity receptor. Responsible for initiating the allergic response. Binding of allergen to receptor-bound IgE leads to cell activation and the release of mediators (such as histamine) responsible for the manifestations of allergy. The same receptor also induces the secretion of important lymphokines
LRRN3	Leucine rich repeat neuronal 3	Potential role in initiation of the primary immune response through mediation of interaction between T cells and dendritic cells.
NRCAM	Neuronal cell adhesion molecule;	Cell adhesion protein that is required for normal responses to cell-cell contacts in brain and in the peripheral nervous system.
OLIG2	Oligodendrocyte lineage transcription factor 2	Involved in a chromosomal translocation t(14;21)(q11.2;q22) associated with T-cell acute lymphoblastic leukaemia[409].
PRSS33	Protease, serine, 33	Serine protease that has amidolytic activity, cleaving its substrates before Arg residues
SIGLEC8	Sialic acid binding Ig-like lectin 8	Putative adhesion molecule that mediates sialic-acid dependent binding to cells.
STMN3	Stathmin-like 3	Exhibits microtubule-destabilizing activity, which is antagonized by STAT3
TRABD2A	TraB domain	Involved in Wnt-protein signalling

	containing 2A	

Table 5.1: Potential functions of genes that were jointly down-regulated in all the pneumonia severity states. The molecular functions are according to the STRING database [280], otherwise a citation is provided.

5.4.2.2 Investigation of up-regulated pathways in pneumonia

In this analysis, I investigated the pathways that were associated the up-regulated genes (m=264) in all the severe pneumonia states (**MSvS** subset) from mild to very severe pneumonia state. While this gene set (m=264) was associated (FDR<0.05) with a range of pathway responses (i.e. innate, adaptive and metabolic functions), the activation of pro-inflammatory innate responses was more predominant (**Table5.2**).

In particular, mild pneumonia was associated with the activation of pathogen recognition receptors (PRRs) pathways including the Toll-like receptors (TLRs), intracellular NOD-like receptors (Inflammasome) and the multi-ligand receptor for Advanced Glycation End products (RAGE). Notably, these are central players in inflammatory responses [85, 410-412]. Consistently, several pro-inflammatory pathways were also activated including the (i) complement system, (ii) MAPK signalling, (iii) NFKB transcription factor, (iv) production of pro-inflammatory cytokines (TNFa, IL1, IL6, IL8, growth factors and INFN gamma) (v) chemotaxis of myeloid cell types (i.e. neutrophils, macrophages and dendritic cells). Notably, the activation of interferon gamma (IFNg) cytokine signals (mostly from the T and NK lymphocytes) is particularly important for enhanced bacterial phagocytosis in macrophages [413, 414], which suggest the important contribution of bacterial infection in pneumonia severity.

Further, mild pneumonia was also associated with the activation of antimicrobial, stress (responses to reactive oxygen species, wounding and heat), adaptive (B and T cell activation and differentiation) and metabolic pathway responses. Notably, the biosynthesis of pro-inflammatory compounds such as oxygen reactive species, nitric oxide and triglyceride as well as xenobiotic metabolism [415-417] and cholesterol haemostasis [418, 419] corresponded with the catabolism of an anti-inflammatory *pyrimidine-ribonucleoside* [420, 421].

To assess whether this gene set (m=264) was associated with functional molecular networks, the STRING database for protein-protein interactions [280] was applied. As shown in **Figure5.3**, pro-inflammatory networks involving intracellular (NLRs) and extracellular (TLRs) pathogen recognition receptors (PRRs), inflammatory cytokines, the complement system (through an IL8 receptor (CXCR2)), coagulation (via MMP9), as well as activation and regulation of transcription factors (through MAPK14) were interconnected.

Together, these findings suggest the involvement of blood-based (systemic) responses, and highlight the interplay between the innate and metabolic pro-inflammatory systemic pathway responses in mild pneumonia. Potentially, blood-based signatures could be applied to detect high-risk cases among the patients presenting at the clinic with mild pneumonia. Notably, these findings are consistent with the cellular responses in **Chapter 4**, and support the central hypothesis that systemic pathway responses underpin the development of severe pneumonia.

Table 5.2: Pathways associated with up-regulated genes in all disease states (Mild, severe and very severe)

SOURCE	ID	Total	Hits	FDR
(a)Innate (Receptors)				
GO:0038187(BP):	Pattern Recognition Receptor Signalling Pathway	109	8	0.0076
GO:0008329(MF)	Signalling Pattern Recognition Receptor Activity	17	3	0.0304
KEGG(hsa04620)	Toll Like Receptor Signalling Pathway	102	8	0.0012
REACTOME	Toll Receptor Cascades	118	9	0.0036
GO:0002224(BP)	Toll Like Receptor Signalling Pathway	85	7	0.0092
GO:0034121 (BP)	Regulation Of Toll Like Receptor Signalling Pathway	49	5	0.0168
GO:0034142 (BP)	Toll Like Receptor 4 Signalling Pathway	18	3	0.0407
REACTOME	Activated Tlr4 Signalling	93	7	0.0099
GO:0002755 (BP)	Myd88 Dependent Toll Like Receptor Signalling Pathway	32	6	<0.001
REACTOME	Myd88 Mal Cascade Initiated On Plasma Membrane	83	6	0.0157
KEGG	NOD Like Receptor Signalling Pathway	62	5	0.0178
GO:0001653 (MF)	Peptide Receptor Activity	133	9	0.0069
REACTOME	Peptide Ligand Binding Receptors	188	8	0.0314
GO:0098543 (BP)	Detection Of Other Organism	19	3	0.045
GO:0050786 (MF)	Rage Receptor Binding	11	4	0.0026
(b)Innate (Transcription Factors)				
REACTOME	NFKB And Map Kinases Activation Mediated By Tlr4 Signalling Repertoire	72	5	0.0314
GO:0051092 (BP)	Positive Regulation Of NF KAPPAB Transcription Factor Activity	132	11	<0.001
GO:0042346 (BP)	Positive Regulation Of NF KAPPAB Import Into Nucleus	27	5	0.0034
GO:0042348 (BP)	Regulation Of NF KAPPAB Import Into Nucleus	48	5	0.0179
GO:0042993 (BP)	Positive Regulation Of Transcription Factor Import Into Nucleus	51	5	0.0214
GO:0000187 (BP)	Activation Of MAPK Activity	137	8	0.0192
(c)Innate (Cytokines)				
HALLMARK	IL6 JAK Stat3 Signalling	87	9	<0.001
GO:0050707 (BP)	Regulation Of Cytokine Secretion	149	13	<0.001
GO:0032677 (BP)	Regulation Of Interleukin 8 Production	61	9	<0.001
GO:0032757 (BP)	Positive Regulation Of Interleukin 8 Production	45	8	<0.001
GO:0050715 (BP)	Positive Regulation Of Cytokine Secretion	96	9	<0.001
GO:0032612 (BP)	Interleukin 1 Production	15	4	0.0044
GO:0032715 (BP)	Negative Regulation Of Interleukin 6 Production	35	5	0.0065
GO:0050663 (BP)	Cytokine Secretion	38	5	0.0092
REACTOME	IL1 Signalling	39	5	0.0099
GO:0001816 (BP)	Cytokine Production	120	8	0.0105
GO:0004896 (MF)	Cytokine Receptor Activity	89	7	0.0107
GO:0032729 (BP)	Positive Regulation Of Interferon Gamma Production	65	6	0.012
REACTOME	Signalling By ILs	107	7	0.0145
REACTOME	G Alpha I Signalling Events	195	9	0.0167
GO:0010575 (BP)	Positive Regulation Of Vascular Endothelial Growth Factor Production	26	4	0.0168
GO:0032675 (BP)	Regulation Of Interleukin 6 Production	104	7	0.0177
HALLMARK	Il2 Stat5 Signalling	200	8	0.0225
HALLMARK	Interferon Gamma Response	200	8	0.0225
GO:0010574(BP)	Regulation Of Vascular Endothelial Growth Factor Production	31	4	0.0241

GO:0005149(MF)	Interleukin 1 Receptor Binding	16	3	0.0298
GO:0019956 (MF)	Chemokine Binding	21	3	0.0304
GO:0032732 (BP)	Positive Regulation Of Interleukin 1 Production	36	4	0.0335
GO:0032649 (BP)	Regulation Of Interferon Gamma Production	97	6	0.0412
GO:1903555(BP)	Regulation Of Tumor Necrosis Factor Superfamily Cytokine Production	101	6	0.0444
REACTOME	Chemokine Receptors Bind Chemokines	57	4	0.047
HALLMARK	TNFa Signalling Via NFKB	200	21	<0.001
GO:0070851(MF)	Growth Factor Receptor Binding	129	8	0.0145
(d) Innate (Chemotaxis)				
GO:0030595 (BP)	Leukocyte Chemotaxis	117	11	<0.001
GO:0060326 (BP)	Cell Chemotaxis	162	12	<0.001
GO:0097529 (BP)	Myeloid Leukocyte Migration	99	7	0.0167
GO:0050920 (BP)	Regulation Of Chemotaxis	180	9	0.0221
GO:0097530 (BP)	Granulocyte Migration	75	7	0.0054
GO:1902622 (BP)	Regulation Of Neutrophil Migration	32	4	0.0227
GO:0002407 (BP)	Dendritic Cell Chemotaxis	16	3	0.0312
GO:0036336 (BP)	Dendritic Cell Migration	21	4	0.0105
GO:0002274 (BP)	Myeloid Leukocyte Activation	98	8	0.0044
GO:0042116 (BP)	Macrophage Activation	31	4	0.0241
(e) Innate (Complement system)				
HALLMARK	Complement	200	13	<0.001
KEGG	Complement And Coagulation Cascades	69	7	<0.001
GO:0001848 (MF)	Complement Binding	19	3	0.0394
(f)Innate (Inflammation or Stress)				
GO:0031663 (BP)	Lipopolysaccharide Mediated Signalling Pathway	31	5	0.0054
GO:0050829 (BP)	Defence Response To Gram Negative Bacterium	43	5	0.0141
GO:0001530 (MF)	Lipopolysaccharide Binding	21	4	0.0107
GO:0006953 (BP)	Acute Phase Response	43	5	0.0141
HALLMARK	Inflammatory Response	200	21	<0.001
GO:0050729 (BP)	Positive Regulation Of Inflammatory Response	113	7	0.0243
GO:0002526 (BP)	Acute Inflammatory Response	73	6	0.0177
GO:0071260(BP)	Cellular Response To Mechanical Stimulus	80	6	0.0226
GO:0071216 (BP)	Cellular Response To Biotic Stimulus	163	9	0.0141
KEGG	LEISHMANIA Infection	72	8	<0.001
KEGG	Systemic Lupus Erythematosus	140	8	0.0071
GO:0000302 (BP)	Response To Reactive Oxygen Species	191	10	0.012
GO:0031960 (BP)	Response To Corticosteroid	176	10	0.0088
GO:0009266 (BP)	Response To Temperature Stimulus	148	10	0.0034
GO:0009408 (BP)	Response To Heat	89	7	0.0105
GO:1904018 (BP)	Positive Regulation Of Vasculature Development	133	8	0.0168
GO:1902883 (BP)	Negative Regulation Of Response To Oxidative Stress	35	4	0.0288
GO:0090083 (BP)	Regulation Of Inclusion Body Assembly	16	3	0.0312
GO:1903036 (BP)	Positive Regulation Of Response To Wounding	162	8	0.0342
HALLMARK	Coagulation	138	6	0.0383
REACTOME	Amyloids	83	5	0.0444
(g) Adaptive				
GO:0002285 (BP)	Lymphocyte Activation Involved In Immune Response	98	6	0.0428
GO:0002292 (BP)	T Cell Differentiation Involved In Immune Response	29	4	0.0214
GO:0002286 (BP)	T Cell Activation Involved In Immune Response	60	5	0.0312
GO:0019864 (MF)	IgG Binding	12	3	0.0149
GO:0019865 (MF)	Immunoglobulin Binding	23	4	0.0124

GO:0002460 (BP)	Adaptive Immune Response Based On Somatic Recombination Of Immune Receptors Built From Immunoglobulin Superfamily Domains	154	11	<0.001
GO(BP)	Negative Regulation Of Adaptive Immune Response	37	4	0.0359
(h)Metabolism				
HALLMARK	Cholesterol Homeostasis	74	6	0.0029
HALLMARK	Xenobiotic Metabolism	200	9	0.0082
GO:0005536 (MF)	Glucose Binding	12	3	0.0149
REACTOME	Organic Cation Anion Zwitterion Transport	13	3	0.0154
GO:0046133 (BP)	Pyrimidine RIBONUCLEOSIDE Catabolic Process	11	3	0.0168
GO:0016810 (MF)	Hydrolase Activity Acting On Carbon Nitrogen But Not Peptide Bonds	143	8	0.0173
GO:0044262 (BP)	Cellular Carbohydrate Metabolic Process	144	8	0.0214
GO:0072529 (BP)	Pyrimidine Containing Compound Catabolic Process	32	4	0.0241
GO:0006768 (BP)	Biotin Metabolic Process	14	3	0.0243
GO:0051770 (BP)	Positive Regulation Of Nitric Oxide Synthase Biosynthetic Process	14	3	0.0243
REACTOME	Triglyceride Biosynthesis	38	4	0.0243
GO:0044275 (BP)	Cellular Carbohydrate Catabolic Process	33	4	0.0246
GO:2000377 (BP)	Regulation Of Reactive Oxygen Species Metabolic Process	152	8	0.0246
GO:0006638 (BP)	Neutral Lipid Metabolic Process	85	6	0.0259
GO:2000379 (BP)	Positive Regulation Of Reactive Oxygen Species Metabolic Process	86	6	0.0259
GO:0006767 (BP)	Water Soluble Vitamin Metabolic Process	88	6	0.0259
GO:0051769 (BP)	Regulation Of Nitric Oxide Synthase Biosynthetic Process	19	3	0.0407
KEGG	GLYCEROLIPID Metabolism	49	4	0.0489
GO:0016813 (MF)	Hydrolase Activity Acting On Carbon Nitrogen But Not Peptide Bonds In Linear AMIDINES	11	3	0.0145
(i) Cell Cycle				
KEGG	Hematopoietic Cell Lineage	88	8	<0.001
REACTOME	Meiotic Synapsis	73	6	0.0139
REACTOME	Packaging Of Telomere Ends	48	5	0.014
REACTOME	Chromosome Maintenance	122	7	0.0167
REACTOME	RNA Pol I Promoter Opening	62	5	0.0182
REACTOME	Telomere Maintenance	75	5	0.0399
REACTOME	Meiosis	116	6	0.0402
REACTOME	Meiotic Recombination	86	5	0.047
REACTOME	RNA Pol I Transcription	89	5	0.0495
REACTOME	Deposition Of New CENPA Containing Nucleosomes At The Centromere	64	6	0.0099
KEY: BP=Gene ontology biological process. MF=Gene ontology molecular function. KEGG= Kyoto Encyclopedia of Genes and Genomes [250]				

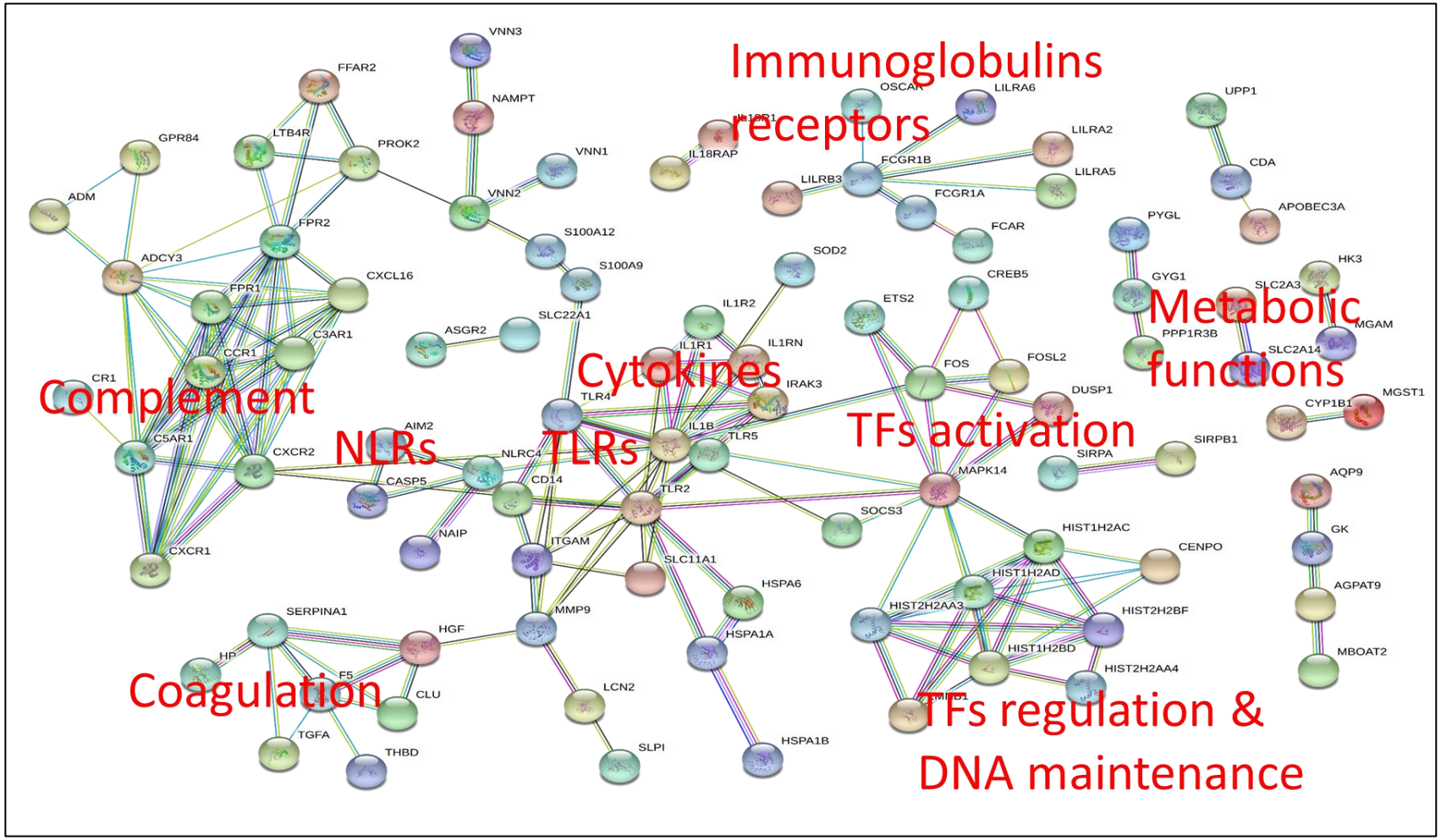


Figure 5.3: Activated protein-protein networks associated with mild pneumonia. Up-regulated genes in all the severe pneumonia states (MSvS, m=264) were analysed using the STRING database[280]. TFs=Transcription factors, NLRs=NOD like receptors, TLRs=Toll-like receptors.

5.4.3 Systemic molecular pathway responses in severe pneumonia

In the previous section, systemic pathway responses that are jointly associated with all the pneumonia severity levels were investigated. To gain a deeper insight into the development of severe pneumonia, here I investigated the pathways that were associated with the up (m=301) and down (m=162) regulated genes in severe and very severe pneumonia but not in mild pneumonia (**SvS** gene set). The down and up-regulated pathways are presented below (respectively).

5.4.3.1 Investigation of down-regulated pathways in severe pneumonia

Briefly, 162 genes were exclusively down-regulated from severe to very severe pneumonia. Principally, this gene set was associated with the down-regulation of ant-inflammatory responses especially in T-cell signalling (**Table5.3**).

As shown in **Table 5.3**, severe pneumonia cases were associated with significant inhibitions of T-cell functions including CD28 dependent co-stimulation (which is vital for priming the adaptive responses [385]), T cell selection, differentiation, proliferation and receptor signaling. Further, severe and very severe pneumonia cases were jointly associated with the down-regulation of the anti-inflammation pathways including *regulation of IL4 and IL10* [116], *regulation of phagocytosis*, *production of immunoglobulin A (IGA) antibody in B cells* [422-425] and the *WNT beta catenin signalling pathway [426]. Furthermore, the* PECAM-1 (CD31) adhesive and signaling pathway, which is associated with the regulation of T-cell homeostasis, effector function and trafficking [427] as well as regulatory [428, 429] and protective [430] roles in inflammation was also down regulated.. . Notably, severe pneumonia was also associated with the down-regulation of the Natural *killer (NK) cell-mediated*

cytotoxicity pathway (innate immunity). This finding is consistent with the cellular responses observed in **Chapter 4** (i.e. pneumonia severity correlated with the depletion of NK cells) further suggesting the potential involvement of NK cells in the pathogenesis of severe pneumonia, and presents a novel target for potential immune-modulation.

To assess whether the down-regulated genes in severe pneumonia states (n=162) were associated with functionally interconnected molecular networks, the STRING database for protein-protein interactions [280] was applied (**Figure5.4**). Mainly, this gene set was associated with an interconnected protein-protein network involving T cell and natural killer (NK) cell functions, and the *Wnt signalling pathway*. Together, these findings suggest an important interplay between the dendritic, T and NK cells in the pathogenesis of severe pneumonia states. To investigate further, the next section assessed on the up-regulated pathways.

Table 5.3: Pathways associated with the down-regulated genes for severe and very severe pneumonia states (m=162)

SOURCE	ID	Total	Hits	FDR
(a) Innate				
KEGG	Natural Killer Cell Mediated Cytotoxicity	137	11	<0.001
GO:0001910(BP)	Regulation Of Leukocyte Mediated Cytotoxicity	53	4	0.0305
GO:0002717(BP)	Positive Regulation Of Natural Killer Cell Mediated Immunity	21	3	0.0247
KEGG	Cell Adhesion Molecules (CAMs)	134	5	0.0401
REACTOME	PECAM1 Interactions	10	2	0.0293
HALLMARK	Complement	200	6	0.0346
HALLMARK	WNT Beta Catenin Signalling	42	3	0.0346
KEGG	WNT Signalling Pathway	151	5	0.0453
KEGG	Antigen Processing and Presentation	89	4	0.0432
(b) Cytokines				
GO:0032673(BP)	Regulation Of Interleukin 4 Production	31	4	0.0086
GO:0032753(BP)	Positive Regulation Of Interleukin 4 Production	22	3	0.0275
GO:0071353(BP)	Cellular Response To Interleukin 4	26	3	0.0354
GO:0032653(BP)	Regulation Of Interleukin 10 Production	46	4	0.0218
(c) Adaptive(T Cells)				
REACTOME	Generation Of Second Messenger Molecules	27	6	<0.001
GO:0031343(BP)	Positive Regulation Of Cell Killing	39	4	0.0136
GO:0002699(BP)	Positive Regulation Of Immune Effector Process	156	7	0.0136
KEGG	T Cell Receptor Signalling Pathway	108	10	<0.001
REACTOME	TCR Signalling	54	6	<0.001
REACTOME	Translocation Of ZAP70 To Immunological Synapse	14	4	<0.001
REACTOME	Costimulation By The CD28 Family	63	6	<0.001
GO:0050852(BP)	T Cell Receptor Signalling Pathway	146	9	<0.001
GO:0030217(BP)	T Cell Differentiation	123	8	0.0019
GO:0042102(BP)	Positive Regulation Of T Cell Proliferation	95	7	0.0025
GO:0046641(BP)	Positive Regulation Of Alpha Beta T Cell Proliferation	19	4	0.0026
GO:0042129(BP)	Regulation Of T Cell Proliferation	147	8	0.0034
GO:0046632(BP)	Alpha Beta T Cell Differentiation	45	5	0.0035
GO:0046640(BP)	Regulation Of Alpha Beta T Cell Proliferation	23	4	0.0035
REACTOME	Phosphorylation Of CD3 And TCR Zeta Chains	16	3	0.0051
REACTOME	PD1 signalling	18	3	0.0065
GO:0046631(BP)	Alpha Beta T Cell Activation	54	5	0.0067
REACTOME	CD28 Dependent Pi3k AKT Signalling	22	3	0.0083
GO:0043383(BP)	Negative T Cell Selection	13	3	0.0092
GO:0043368(BP)	T Cell Selection	36	4	0.0121
GO:0045061(BP)	Thymic T Cell Selection	19	3	0.0208
REACTOME	CD28 Costimulation	32	3	0.0241
GO:0046635(BP)	Positive Regulation Of Alpha Beta T Cell Activation	51	4	0.0281
REACTOME	Downstream TCR Signalling	37	3	0.0293
REACTOME	CD28 Dependent VAV1 Pathway	11	2	0.0331
GO:0005070(MF)	SH3 -SH2 Adaptor Activity	52	5	0.0117
GO:0031294(BP)	Lymphocyte Costimulation	78	11	<0.001
(d) Adaptive (other)				
GO:0042288(MF)	MHC Class I Protein Binding	19	3	0.0222
KEGG	Intestinal Immune Network For IGA Production	48	3	0.0453
GO:0070665(BP)	Positive Regulation Of Leukocyte Proliferation	136	9	<0.001
KEGG	Hematopoietic Cell Lineage	88	6	0.0013
GO:0002705(BP)	Positive Regulation Of Leukocyte Mediated Immunity	85	6	0.0064

GO:0006968(BP)	Cellular Defence Response	60	5	0.0089
GO:0002708(BP)	Positive Regulation Of Lymphocyte Mediated Immunity	69	5	0.0136
GO:0002312(BP)	Cell Activation Involved In Immune Response	139	6	0.0305
GO:1901623(BP)	Regulation Of Lymphocyte Chemotaxis	20	3	0.0221
GO:0050851(BP)	Antigen Receptor Mediated signalling Pathway	195	9	0.0026
(e) Other				
KEGG	Allograft Rejection	38	3	0.0397
HALLMARK	Allograft Rejection	200	11	<0.001
KEGG	Graft Versus Host Disease	42	3	0.0401
KEGG	Primary Immunodeficiency	35	6	<0.001
HALLMARK	Apical Surface	44	3	0.0346
REACTOME	The Role Of NEF In HIV1 Replication And Disease Pathogenesis	28	4	0.0014
REACTOME	NEF Mediates Down Modulation Of Cell Surface Receptors By Recruiting Them To Clathrin Adapters	21	3	0.0079
GO:0002230(BP)	Regulation Of Defence Response To Virus By Virus	29	4	0.0071
GO:0016444(BP)	Somatic Cell DNA Recombination	33	4	0.0092
GO:0033151(BP)	V D J Recombination	16	3	0.0136
GO:0002200(BP)	Somatic Diversification Of Immune Receptors	42	4	0.0143
GO:0071594(BP)	Thymocyte Aggregation	45	4	0.0208
GO:0015026(MF)	Coreceptor Activity	38	4	0.021
GO:0035591(MF)	Signalling Adaptor Activity	74	5	0.021
KEGG	Colorectal Cancer	62	4	0.0227
KEGG	Autoimmune Thyroid Disease	53	3	0.0453
KEGG	Basal Cell Carcinoma	55	3	0.0453
KEGG	Endometrial Cancer	52	3	0.0453
KEGG	Melanogenesis	102	4	0.0453
KEGG	Vibrio Cholerae Infection	56	3	0.0453

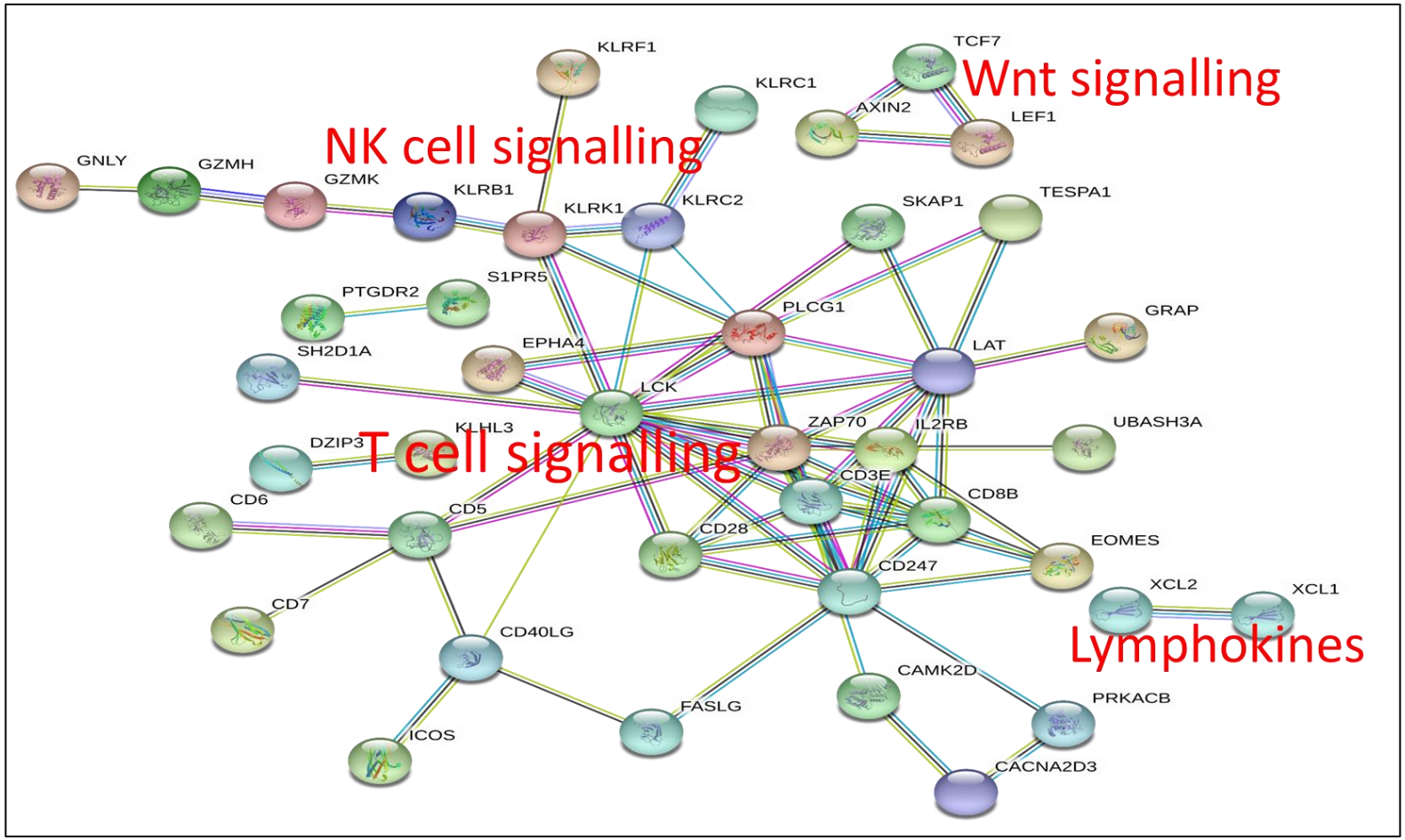


Figure 5.4: Down-regulated protein-protein networks associated with severe pneumonia. Down regulated genes in severe and very severe pneumonia were investigated using the STRING database [280]. NK=Natural killer, Wnt=Wingless-type MMTV integration site family member..

5.4.3.2 *Investigation of up-regulated pathways in severe pneumonia*

At the gene analytic level, severe and very severe pneumonia cases were uniquely associated with the up-regulation of 301 genes. At pathway analytic level, this gene set was principally associated with the activation of innate and metabolic pathway responses (**Table5.4**). Severe pneumonia outcomes were associated with further amplification of pro-inflammatory innate responses such as complement system, cell chemotaxis, platelet degranulation and IL6 production. Notably, severe pneumonia cases were further associated with the activation of fatty acid and lipid metabolism pathways including Sphingolipid, Lysophospholipid, unsaturated fatty acids and lipopolysaccharide metabolism (**Table5.4**).

In summary, severe pneumonia outcomes are predominantly associated with the inhibition of adaptive and NK cell responses, and the activation of fatty acid and lipid metabolism pathways consequently promoting the deleterious pro-inflammatory responses. These finding further highlight the interplay between the innate, adaptive and metabolic pathways in pneumonia and support the central hypothesis that systemic pathway responses underpin the pathogenesis of severe pneumonia. *To gain more insights into the pathogenesis of pneumonia severity, the next section investigated the molecular pathway responses to very severe pneumonia.*

Table 5.4: Pathways associated with the up-regulated genes for severe and very severe pneumonia (m=301)

SOURCE	ID	TOTAL	HITS	FDR
(a) Innate				
Hallmark	Complement	200	14	<0.001
Hallmark	IL6 JAK Stat3 Signalling	87	9	<0.001
Hallmark	Inflammatory Response	200	13	<0.001
GO:0060326(BP)	Cell Chemotaxis	162	9	0.034

GO:0004896(MF)	Cytokine Receptor Activity	89	8	0.0068
GO:0032675(BP)	Regulation Of Interleukin 6 Production	104	8	0.0173
GO:0032755(BP)	Positive Regulation Of Interleukin 6 Production	68	6	0.0292
GO:0002699(BP)	Positive Regulation Of Immune Effector Process	156	9	0.0292
GO:0002698(BP)	Negative Regulation Of Immune Effector Process	102	7	0.0363
GO:0043300(BP)	Regulation Of Leukocyte Degranulation	41	5	0.0267
GO:0002576(BP)	Platelet Degranulation	107	7	0.0433
KEGG	Hematopoietic Cell Lineage	88	8	0.0045
GO:0034113(BP)	Heterotypic Cell Cell Adhesion	27	4	0.0329
Hallmark	Epithelial Mesenchymal Transition	200	10	0.0082
(b) Metabolism				
GO:1901568(BP)	Fatty Acid Derivative Metabolic Process	96	9	0.0043
GO:0033559(BP)	Unsaturated Fatty Acid Metabolic Process	109	9	0.0068
GO:1901570(BP)	Fatty Acid Derivative Biosynthetic Process	46	6	0.0071
GO:0006636(BP)	Unsaturated Fatty Acid Biosynthetic Process	58	6	0.0185
GO:0006633(BP)	Fatty Acid Biosynthetic Process	114	8	0.0215
GO:0072330(BP)	Monocarboxylic Acid Biosynthetic Process	172	11	0.0068
GO:0030148(BP)	Sphingolipid Biosynthetic Process	77	8	0.0053
KEGG	Sphingolipid Metabolism	40	5	0.0133
GO:0046467(BP)	Membrane Lipid Biosynthetic Process	114	10	0.0043
GO:0006643(BP)	Membrane Lipid Metabolic Process	184	11	0.0135
GO:0071617(MF)	Lysophospholipid Acyltransferase Activity	19	4	0.0147
GO:0009247(BP)	Glycolipid Biosynthetic Process	62	6	0.0258
GO:0006665(BP)	Sphingolipid Metabolic Process	138	9	0.0215
GO:0006688(BP)	Glycosphingolipid Biosynthetic Process	25	4	0.0292
GO:0006687(BP)	Glycosphingolipid Metabolic Process	69	6	0.0292
GO:0032369(BP)	Negative Regulation Of Lipid Transport	26	4	0.0292
GO:0050996(BP)	Positive Regulation Of Lipid Catabolic Process	25	4	0.0292
GO:0032368(BP)	Regulation Of Lipid Transport	95	7	0.0292
GO:0019370(BP)	Leukotriene Biosynthetic Process	23	6	<0.001
GO:0006691(BP)	Leukotriene Metabolic Process	33	7	<0.001
GO:0006672(BP)	Ceramide Metabolic Process	73	8	0.0043
GO:0046513(BP)	Ceramide Biosynthetic Process	43	6	0.0068
GO:0046173(BP)	Polyol Biosynthetic Process	26	5	0.0068
GO:0016755(MF)	Transferase Activity Transferring Amino Acyl Groups	26	5	0.0068
GO:0035091(MF)	Phosphatidylinositol Binding	200	11	0.0147
GO:0001574(BP)	Ganglioside Biosynthetic Process	18	4	0.0167
GO:0001573(BP)	Ganglioside Metabolic Process	26	4	0.0292
GO:1903175(BP)	Alcohol Biosynthetic Process	111	8	0.0215
GO:0097503(BP)	Sialylation	21	4	0.0215
GO:1903509(BP)	Liposaccharide Metabolic Process	114	8	0.0246
GO:0048268(BP)	CLATHRIN Coat Assembly	12	3	0.0292
GO:0034311(BP)	Diol Metabolic Process	11	3	0.0292
GO:0033690(BP)	Positive Regulation Of Osteoblast Proliferation	11	3	0.0292
GO:1903307(BP)	Positive Regulation Of Regulated Secretory Pathway	49	5	0.0343
GO:0046519(BP)	Sphingoid Metabolic Process	13	3	0.0385

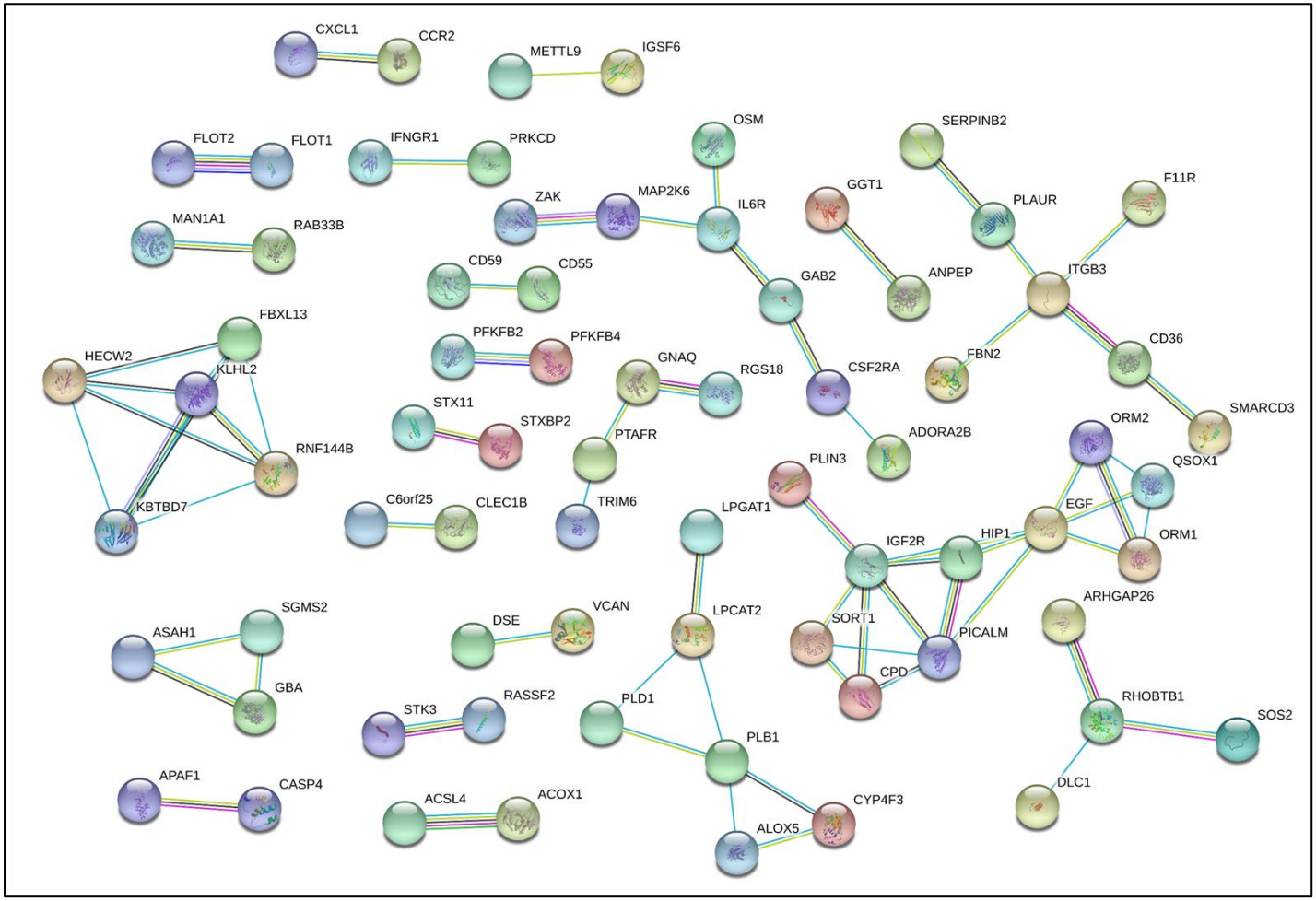


Figure 5.5: Up-regulated protein-protein networks associated with severe and very pneumonia. The STRING database [280] was applied to investigate the protein-protein interactions networks that were associated with the up-regulated genes in severe and very severe pneumonia.

5.4.4 Systemic molecular pathway responses in very severe pneumonia

Clinically, the distinction between severe and very severe pneumonia is subtle and subjective, and are often treated equally (inpatients)[1]. Here, I assessed whether very severe pneumonia cases have a unique set of systemic molecular pathway responses (objective three). At the gene level, 247 genes (down=104, up=143) were uniquely associated with the clinical definition of very severe pneumonia (**vS** subset). Here, the corresponding pathways (up and down-regulated) were investigated.

5.4.4.1 Investigation of the down-regulated pathways in very severe pneumonia

Briefly, the development of very severe pneumonia was principally associated with the down-regulation of adaptive pathway responses especially in T cells (**Table5.5**). While most of these pathways were also observed in severe pneumonia, very severe pneumonia cases were predominantly associated with the inhibition of regulatory effector responses including the production of regulatory cytokines (IL2 and IL17), and the regulation of cellular responses (i.e. phagocytosis and B cell mediated immunity). To investigate further, the next section investigated the up-regulated pathways in very severe pneumonia.

Table 5.5: Pathways associated with the down-regulated genes in very severe pneumonia only (m=104)

SOURCE	ID	Total	Hits	FDR
GO:0050851(BP)	Antigen Receptor Mediated Signalling Pathway	195	6	0.0155
GO:0002460 (BP)	Adaptive Immune Response Based On Somatic Recombination Of Immune Receptors Built From Immunoglobulin Superfamily Domains	154	5	0.0255
GO:0002312(BP)	Cell Activation Involved In Immune Response	139	6	0.0055
GO:0002443(BP)	Leukocyte Mediated Immunity	189	5	0.0492
GO:0050715(BP)	Positive Regulation Of Cytokine Secretion	96	4	0.0448
GO:0032660 (BP)	**Regulation Of Interleukin 17 Production**	**22**	**4**	**0.0011**
GO:0032740 (BP)	**Positive Regulation Of Interleukin 17 Production**	**13**	**3**	**0.0032**
HALLMARK	**IL2 STAT5 Signalling**	200	5	0.0443
GO:0002819 (BP)	Regulation Of Adaptive Immune Response	123	6	0.0032
GO:0002821(BP)	Positive Regulation Of Adaptive Immune Response	73	5	0.0032
GO:0002285(BP)	Lymphocyte Activation Involved In Immune Response	98	5	0.0082
GO:0002699(BP)	Positive Regulation Of Immune Effector Process	156	6	0.0083
GO:0002712 (BP)	Regulation Of B Cell Mediated Immunity	41	3	0.0402
GO:0002705 (BP)	Positive Regulation Of Leukocyte Mediated Immunity	85	6	0.0011
GO:0002703 (BP)	Regulation Of Leukocyte Mediated Immunity	156	7	0.0021
GO:1903363 (BP)	Negative Regulation Of Cellular Protein Catabolic Process	64	4	0.0124
GO:0031343 (BP)	Positive Regulation Of Cell Killing	39	3	0.036
GO:0002708(BP)	Positive Regulation Of Lymphocyte Mediated Immunity	69	6	<0.001
GO:0002706(BP)	Regulation Of Lymphocyte Mediated Immunity	114	7	<0.001
HALLMARK	Allograft Rejection	200	6	0.0141
GO:0050728(BP)	Negative Regulation Of Inflammatory Response	100	4	0.0448
GO:1900425(BP)	Negative Regulation Of Defence Response	144	5	0.0268
GO:0002286(BP)	T Cell Activation Involved In Immune Response	60	5	0.0021
GO:0002292 (BP)	T Cell Differentiation Involved In Immune Response	29	4	0.0021
GO:2000514(BP)	Regulation Of CD4 Positive Alpha Beta T Cell Activation	38	4	0.0032
GO:0043379(BP)	T Cell Differentiation	123	6	0.0032
GO:0046632(BP)	Alpha Beta T Cell Differentiation	45	4	0.0055
GO:0050852(BP)	T Cell Receptor Signalling Pathway	146	6	0.0058
GO:0046631(BP)	Alpha Beta T Cell Activation	54	4	0.0087
GO:2000516(BP)	Positive Regulation Of CD4 Positive Alpha Beta T Cell Activation	27	3	0.0155
GO:0046634(BP)	Regulation Of Alpha Beta T Cell Activation	68	4	0.0174
GO:0035710(BP)	CD4 Positive Alpha Beta T Cell Activation	34	3	0.0268
GO:0045058(BP)	T Cell Selection	36	3	0.0296
GO:0046637(BP)	Regulation Of Alpha Beta T Cell Differentiation	46	3	0.0492

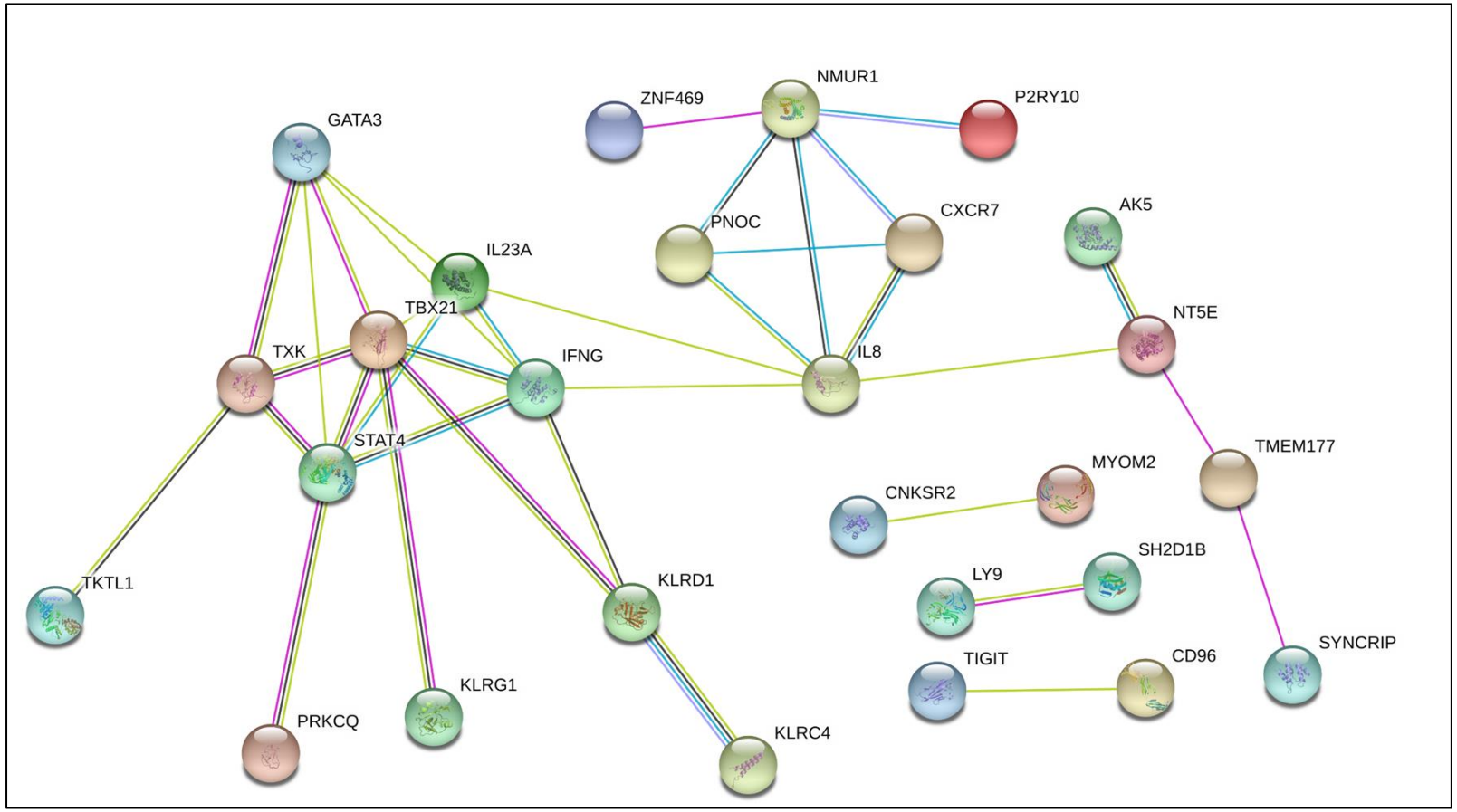


Figure 5.6: Down-regulated protein-protein networks associated with very severe pneumonia. The STRING database [280] was applied to investigate the protein-protein interactions networks that were associated with the down-regulated genes in very severe pneumonia

5.4.4.2 *Investigation of up-regulated pathways in very severe pneumonia*

Here, the up-regulated genes in very severe pneumonia were predominantly associated with the activation of stress, antimicrobial and wound healing processes (**Table5.6 & Figure5.7**). Firstly, the predominant activation of stress and wound healing processes such as extracellular matrix (ECM) organization [431], epithelial mesenchymal transition organization [432], collagen formation[433] and coagulation[434] and apoptosis suggest that deleterious systemic responses underpin the development of very severe pneumonia states. Notably, the predominant activation of antimicrobial activities such as *defensin* antimicrobial peptides[435], and responses to bacterium suggests the important contribution of bacterial infections in very severe pneumonia outcomes. Further, this finding is similar to host responses in sepsis [136] suggesting the involvement of bacterial septicaemia in the development of serious pneumonia outcomes, and present an opportunity to investigate blood-based biomarkers for clinical stratification and treatment modalities of high-risk pneumonia cases (**Chapter 6**). Together, these findings support the central hypothesis that systemically suppressed and activated molecular pathway responses underpin the development severe pneumonia states

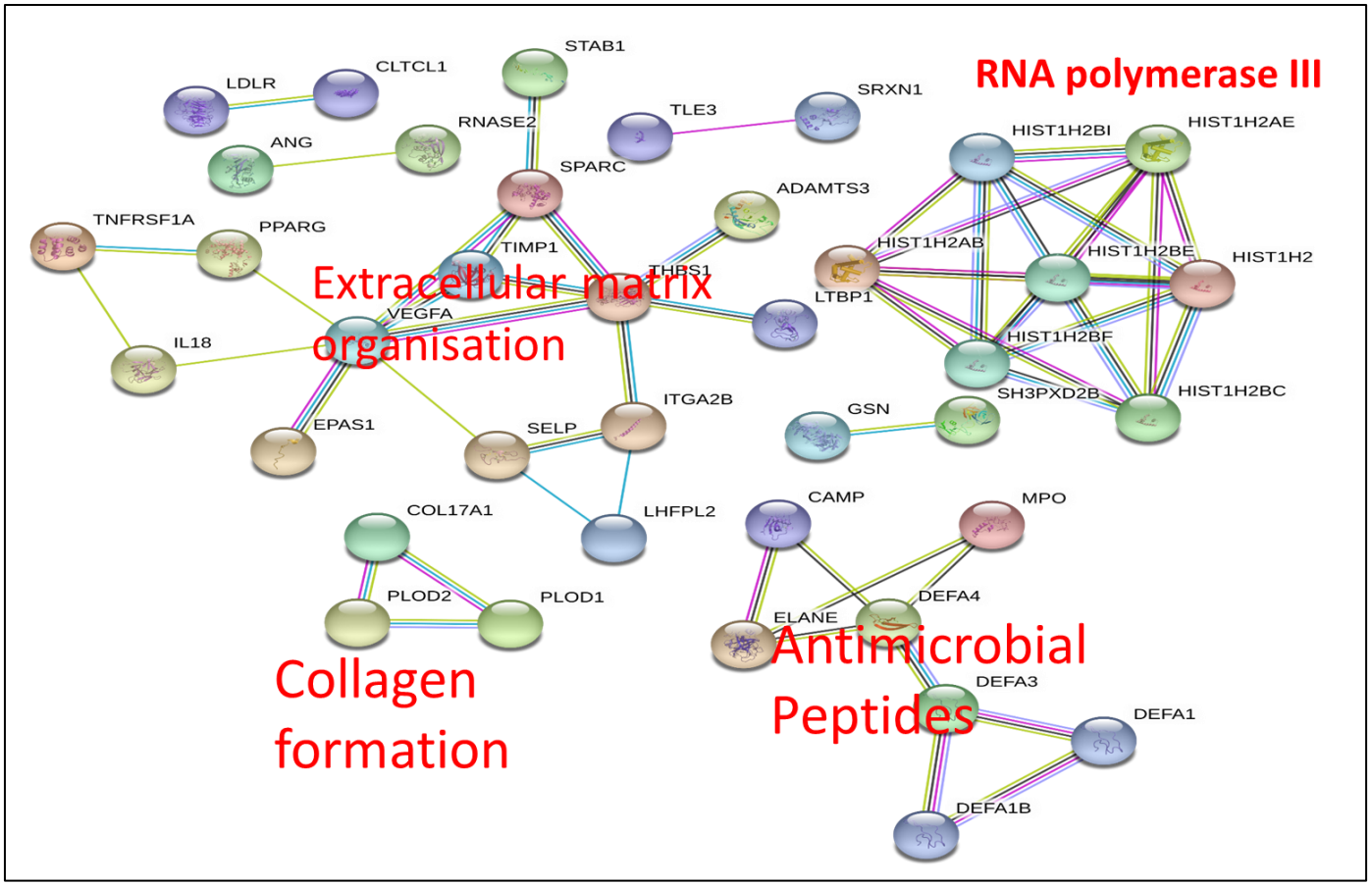


Figure 5.7: UP-regulated protein-protein networks associated with very severe pneumonia. The STRING database [280] was applied to investigate the protein-protein interactions networks that were associated with the up-regulated genes in very severe pneumonia.

Table 5.6: Pathways associated with the up-regulated genes in very severe pneumonia (m=143)

SOURCE	ID	Total	Hits	FDR
(a) Innate, antimicrobial, stress or wound healing				
GO:0001906(BP)	Cell Killing	56	6	<0.001
REACTOME	Defensins	51	4	0.0032
GO:0002227(BP)	Innate Immune Response In Mucosa	23	9	<0.001
GO:0019730(BP)	Antimicrobial Humoral Response	52	11	<0.001
GO:0050830(BP)	Defence Response To Gram Positive Bacterium	72	12	<0.001
GO:0050829(BP)	Defence Response To Gram Negative Bacterium	43	4	0.0161
GO:0030228(MF)	Lipoprotein Particle Receptor Activity	16	3	0.013
GO:0030169(MF)	Low Density Lipoprotein Particle Binding	15	3	0.013
GO:0071814(MF)	Protein Lipid Complex Binding	24	3	0.027
GO:0002251(BP)	Organ or Tissue Specific Immune Response	33	9	<0.001
GO:0006959(BP)	Humoral Immune Response	187	12	<0.001
GO:0001878(BP)	Response to Yeast	13	5	<0.001
KEGG	Systemic Lupus Erythematosus	140	9	<0.001
GO:0035821(BP)	Modification Of Morphology Or Physiology Of Other Organism	100	8	<0.001
GO:0044144(BP)	Modulation Of Growth Of Symbiont Involved In Interaction With Host	16	3	0.0131
GO:0051702(BP)	Interaction With Symbiont	52	4	0.0301
GO:0071936(MF)	Co-receptor Activity	66	4	0.0305
GO:0043901(BP)	Negative Regulation Of Multi Organism Process	151	6	0.0339
GO:0075157(BP)	Positive Regulation Of G Protein Coupled Receptor Protein Signalling Pathway	25	3	0.032
GO:0050832(BP)	Response To Fungus	52	9	<0.001
GO:0050832(BP)	Defence Response To Fungus	39	8	<0.001
GO:0044364(BP)	Disruption Of Cells Of Other Organism	26	6	<0.001
REACTOME	Response To Elevated Platelet Cytosolic Ca2	89	6	<0.001
GO:0002576(BP)	Platelet Degranulation	107	6	0.0081
HALLMARK	UV Response Down	144	7	0.0011
HALLMARK	Reactive Oxygen Species Pathway	49	3	0.0435
REACTOME	Extracellular Matrix Organization	87	6	<0.001
GO:0050840(MF)	Extracellular Matrix Binding	51	4	0.0246
HALLMARK	Epithelial Mesenchymal Transition	200	8	0.0011
HALLMARK	Coagulation	138	7	0.0011
REACTOME	Collagen Formation	58	4	0.005
GO:0001936(BP)	Regulation Of Endothelial Cell Proliferation	98	6	0.0057
GO:0030520(BP)	Intracellular Estrogen Receptor Signalling Pathway	20	3	0.0232
GO:0008201(MF)	Heparin Binding	157	7	0.013
REACTOME	Amyloids	83	8	<0.001
(B) Cell Cycle				
REACTOME	Packaging Of Telomere Ends	48	7	<0.001
REACTOME	Deposition Of New CENPA Containing Nucleosomes At The Centromere	64	7	<0.001
REACTOME	RNA Pol I Promoter Opening	62	7	<0.001
REACTOME	Meiotic Synapsis	73	7	<0.001
REACTOME	Telomere Maintenance	75	7	<0.001
REACTOME	Meiotic Recombination	86	7	<0.001
REACTOME	RNA Pol I Transcription	89	7	<0.001
REACTOME	Meiosis	116	7	<0.001
REACTOME	Chromosome Maintenance	122	7	<0.001
REACTOME	RNA Pol I RNA Pol III And Mitochondrial Transcription	122	7	<0.001

GO:0030041(BP)	Actin Filament Polymerization	23	4	0.0012
GO:0008154(BP)	Actin Polymerization Or Depolymerization	37	4	0.0081
HALLMARK	Apoptosis	161	5	0.0435

5.5 Discussion

This chapter investigated the systemic molecular pathway responses that are associated with pneumonia severity (mild, severe and very severe) using the training set of the whole blood transcriptome (n=345). At the gene analytic level, pneumonia severity was significantly associated with a battery of molecular response signatures. The absolute fold changes and the number of differentially expressed genes (DEGs) increased with pneumonia severity (**Figure5.2**). Using a range of biochemical pathway databases, this analysis has revealed significant systemic pathway (innate, adaptive, metabolic and cell-cycle) responses in pneumonia, which support the central hypothesis that systemic pathway responses underpin the development of severe pneumonia outcomes. These findings show high agreement between the biochemical pathways databases, and are consistent with the cellular pathway responses in **Chapter 4**.

5.5.1 Agreements between the biochemical pathway databases

To enhance the interpretation of high-throughput data, a range of pathway analysis algorithms and biochemical databases has been applied [249, 256].. To gain a comprehensive view, here I investigated a range of biochemical pathway resources that are extensively applied in pathway analyses (KEGG, REACTOME, GO, HALLMARK) *[254, 402].* Despite the lack of standardized nomenclature, the findings of this analysis showed a high level of agreement between the pathway responses that were independently enriched across the biochemical pathway databases. For example, pathogen recognition receptors (PRRs), complement system, natural killer cell (NK), adaptive (B and T cells) and lipid metabolism pathway responses were consistently

enriched in all the resources. These agreements have increased the confidence in these results, which in turn support the central hypothesis suggesting the significant contribution of systemic pathway responses in the development of severe pneumonia outcomes.

5.5.2 Systemic molecular pathway responses in severe pneumonia

As hypothesized, the development of severe pneumonia states was associated with a battery of significant systemic pathway responses involved in the innate, adaptive and metabolic signalling pathways. In particular, while the up-regulation of innate responses and cholesterol metabolism were associated with the development of all severe pneumonia states (i.e. mild to very severe); the development of severe pneumonia states were predominantly associated with the activation of fatty acid and lipid metabolism pathways, and the inhibition of adaptive effector functions as well as NK cell signalling. While many of these findings were anticipated[36, 37, 39, 80, 112, 436, 437], the potential involvement of NK cells in the pathogenesis of pneumonia presents a novel finding on the potential target for immune-modulation and case management. Finally, very severe pneumonia cases were predominantly associated with antimicrobial and wound-healing responses.

Together, these findings underscore the importance of systemic pathway responses in development of severe pneumonia outcomes. Consistently, Fernandez-Botran et al. (2014) also observed that severe pneumonia cases were associated with compromised local responses but enhanced systemic responses; and the vice versa for the non-severe cases[436]. Further, these

findings also share several similarities with host responses in sepsis[120, 136, 381, 438-440], which further suggest the important contribution of systemic pathway responses in the development of severe pneumonia outcomes. Thus, while immune responses in pneumonia are often compartmentalized within the alveoli (local responses)[441], severe pneumonia outcomes are potentially caused by the de-compartmentalization (leakage) of the local responses into the circulating blood (i.e. involvement the systemic responses)[116]. Furthermore, the predominant involvement of antibacterial responses in very severe pneumonia outcomes particularly suggests the important contribution of bacterial septicaemia in the development of serious pneumonia outcomes. Clinically, this finding presents an innovative approach for blood-based biomarkers to enhance rapid identification and appropriate treatment of high-risk mild pneumonia cases (**Chapter 6**).

5.5.3 Agreement between cellular and molecular pathway responses

In this thesis, I sought to investigate the central hypothesis that systemic pathway (cellular and molecular) responses underpin the development of severe pneumonia states. In **Chapter 4**, pneumonia severity was associated with the depletion of T, B, NK and Dendritic cells, and the elevation of monocyte and neutrophil cell proportions, which are consistent with the current molecular pathway responses (**Chapter5**). In particular, the depletion of Dendritic, B, T and NK proportions (**Chapter 4**) molecularly corresponded with the inhibition of antigen presentation (i.e. dendritic and T-cells), adaptive responses (B and T cells) and natural killer (NK) cell cytotoxicity. On the other hand, elevation of monocytes and neutrophils subpopulations

molecularly corresponded with the activation of pro-inflammatory innate and stress responses including phagocytosis, neutrophil chemotaxis and granulation, and the production of reactive oxidative species and nitric acid. Thus, the molecular findings support the functional involvement of the cellular response mediators in the pathogenesis of pneumonia severity. Together, these findings (cellular and molecular pathway responses) jointly support the central hypothesis for this thesis that systemic pathway (cellular and molecular) responses underpin the development of severe pneumonia states. Nevertheless, the potential involvement of NK cells in severe pneumonia requires further investigations.

5.5.4 Limitations

However, this analysis had some limitations. Firstly, these findings reflect a cross-section view of the immune response since whole blood samples were collected at a single time point. Preferably, a longitudinal study design would provide more insights into the pathogenesis of pneumonia. Secondly, these findings largely depend on the current state of the existing biochemical pathway databases, which remains an area of active research. Potentially, these data and current results may yield yet knew findings if new data resources come online.

Further, the over representation analysis (ORA) pathway analysis approach (i.e. using the Fisher's exact test) ignored the pathway structure, and the strength the gene features and their interactions. Though computationally expensive, pathway analysis approaches that account for the pathway structure (as suggested in **Chapter 2, section 2.4.4**, page 70) [258-261,

273, 276-279, 442] would have revealed more insights into the pathogenesis of childhood pneumonia. Furthermore, differentially expressed genes (DEGs) based on arbitrary cut-offs values (FDR<0.05 and |FC|≥2) were applied. While these cut-offs are biologically and statistically reasonable, a sensitivity analysis on a range of cut-offs would potentially generate more robust results. Alternatively, parameter free methods such as gene set enrichment analysis (GSEA)[157] would be applied. However, such methods require more computational time (i.e. permutation tests) and are potentially confounded by irrelevant genes and pathways [256]. Nonetheless, the agreements between different pathway resources, and between molecular and cellular pathway responses suggest reliability and validity.

5.5.5 Conclusion

In summary, in this chapter it was shown that pneumonia severity is associated with a battery of significant systemic molecular pathway responses. These findings highlight the interplay between the innate, adaptive and metabolic systemic pathway responses in severe pneumonia. Importantly, the molecular (**Chapter 5**) and cellular (**Chapter 4**) pathway responses consistently support the central hypothesis that systemic pathway responses underpin the development of severe pneumonia states. Notably, the potential involvement of NK cells presents a novel target for immuno-modulation, and requires further investigations.

Chapter 6: Systemic cellular pathway-based candidate biomarkers of severe pneumonia

6.1 Introduction

Early identification of mild pneumonia patients at the higher risk of developing poor outcomes remains a major public health challenge. This chapter investigates host-based systemic (whole blood) biomarkers for severe pneumonia, to facilitate early detection of high-risk cases presenting at the resource-constrained clinic with mild pneumonia. Based on the results from the previous chapters, here I have coupled cellular pathway biology with machine-learning approaches to select (using the Fajara training data, n=345) and validate (using the Basse validation data set, n=345) the performance of systemic candidate biomarkers (transcriptomic classifier features) for early detection of severe pneumonia cases.

6.2 Background

Pneumonia is caused by a range of pathogens including viruses and bacteria; and remains the leading infectious cause of mortality in under-five children worldwide [2, 7, 9, 50, 443]. Clinically, the detection of microbiological aetiology in patients presenting with symptoms of pneumonia remains a major challenge especially in resource-limited where the burden is highly concentrated [5, 28]. While pneumonia has a complex aetiology, bacterial cases are often associated with more serious outcomes [5]. Therefore, early identification and appropriate treatment of bacterial cases is a cornerstone for mitigating the burden of childhood pneumonia and under-five mortality[27].

However, standard diagnostic tools (i.e. Chest x-rays and blood culture) are both too expensive for remote healthcare facilities and also not optimal in their sensitivity and specificity. Firstly, chest x-rays are not aetiology-specific; and unnecessary frequent exposure to radiation is a potential risk factor for serious conditions like cancer[28]. Further, blood culture lacks sensitivity and the turnaround is too long (typically 24-48 hours or longer) to delay empirical therapy being started [53, 55]. To mitigate serious bacterial outcomes, the World Health Organisation (WHO) criteria[27] prioritises sensitivity over specificity. Although this helps to avoid undertreating serious cases, it also results broadly in over-treatment with antibiotics, which has the consequences of heavier financial costs and the potential increase in antibiotic resistance [350, 444, 445].

Alternatively, host-based biomarkers present a potential paradigm shift in the clinical management of pneumonia towards practical personalized treatment [154]. While single serum biomarkers have shown potential[56, 96, 446-449], whole blood genome-wide profiling presents a potentially more robust and innovative approach to explore systemic pathway-based candidate biomarkers[65, 121, 136, 450-452] to enhance the stratification and management of pneumonia cases. Importantly, whole blood is a rich and clinically accessible tissue for pathophysiological investigations, and molecular profiling has become a mainstay for future translational medicine[123].

Chapter 6: Candidate biomarkers

The aim of this chapter is to derive candidate biomarkers for early detection of mild pneumonia cases at the higher risk of developing severe outcomes (i.e. bacterial cases). However, it is worth noting that this analysis lacked proper gold standard data on disease etiology (i.e. viral or bacterial pneumonia). Therefore, this analysis is based on the main assumption that bacterial infections underpin the pathogenesis of severe pneumonia outcomes. In this regard, the strategy is to derive a classifier that accurately distinguishes severe pneumonia cases from non-pneumonia controls (i.e. extreme cases). Subsequently, this classifier would be applied to stratify mild pneumonia cases into low-risk and high-risk treatment groups (**Figure6.1**).

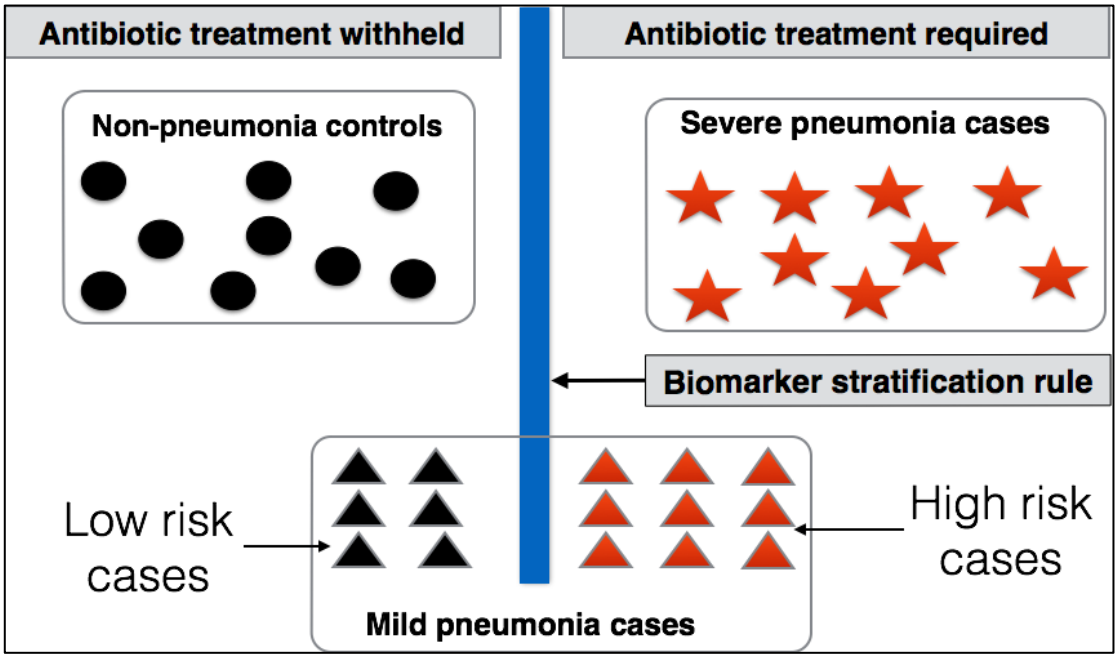


Figure 6.1: An Illustration of the potential clinical application of the current candidate biomarkers for stratification and treatment of mild pneumonia cases. Using a biomarker stratification rule that distinguishes extreme cases (severe pneumonia cases and non-pneumonia controls), mild pneumonia cases with suspected bacterial infection (red triangles) would benefit from early antibiotic treatment while withholding treatment for the low risk cases (black triangles).

6.3 Approach

In particular, I have applied machine-learning approaches to assess the performance of cellular pathway-based biomarkers (range: 18-37). Feature selection and internal performance assessment using leave-one-out cross-validation (LOOCV) were implemented in the training data (**Fajara**) whole blood transcriptome (n=345); followed by an independent validation in the **Basse** data set (n=158), which was kept independent from primary analyses. Throughout this chapter, **n** and **m** represent the number of subjects and genes, respectively.

6.3.1 Feature selection and performance assessment

As illustrated in **Figure6.2**, the investigation of candidate biomarkers for severe pneumonia involved three main steps: (i) feature selection (ii) internal performance assessment using the training data and (iii) independent validation using the Basse data set. Briefly, feature selection coupled cellular pathway biology and machine learning approaches. In particular, machine-learning approaches were applied to select and investigate the performance of cellular pathway based transcriptomic features. The cellular pathway-based features and the machine learning approaches are described below.

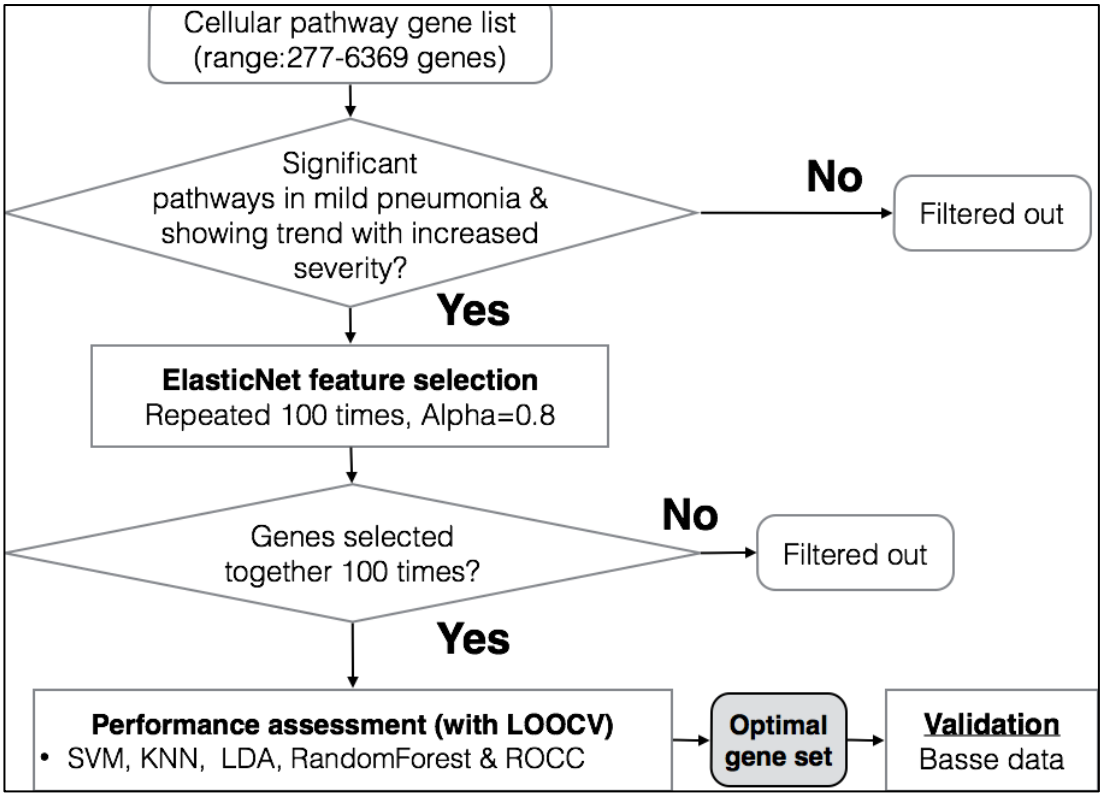


Figure 6.2: Feature selection of candidate biomarkers for severe pneumonia. The cellular pathway gene lists ranged between 277 markers in the IBML and 6369 cell-correlated genes. The Elastic Net (EN) feature selection was applied to the gene features that were differentially expressed in mild pneumonia (compared to non-pneumonia controls).

6.3.1.1 Selection of cellular pathway-based transcriptomic features

Briefly, the cellular pathway-based features were independently selected from the following lists:

i. **IBML:** an integrated blood marker lists (m=277), derived in **Chapter 4**

ii. **CCGs**: Cell correlated genes (m=6369)

iii. **DCGs**: Differentially correlated genes (m=720).

The specific details for each cellular list are provided in subsequent sections. For each cellular list, candidate biomarkers were selected at the cellular level (i.e. B, NK, or neutrophils, respectively) followed by an aggregation. To derive a unified cellular classifier, candidate biomarkers from the three lists were also aggregated.

At each level, eligible features were selected using the following two criteria: (i) differentially expressed genes (FDR<0.05 & |FC|≥1.5) in mild pneumonia, and (ii) genes showing trended response (i.e. increased or decreased fold change) with increased pneumonia severity. Independently, the same approach was applied to select eligible markers from the validated 52-gene neonatal sepsis classifier [136]. Subsequently, each eligible feature set was subjected to machine learning feature selection to identify the optimal number of candidate biomarker set (next section).

6.3.1.2 *Machine-learning feature selection*

For each eligible set of biomarkers, the Elastic Net (EN) feature selection algorithm (implemented in the *glmnet R package*[282]) was applied to select an optimal combination of synergetic features for distinguishing severe pneumonia cases from the non-pneumonia controls. It is worth noting that mild pneumonia cases were excluded from machine learning analyses. To select the optimal parameter values, the *cv.glmnet function* was implemented using the leave-one-out cross validation (LOOCV). Further, the EN algorithm was repeated hundred times to enable the selection of robust and stable combination of features. Using this approach, candidate biomarkers that were selected together all the time (100 times) were selected and assessed for classification performance.

To assess the classification of the candidate biomarker sets, five classification algorithms coupled with leave-one-out cross validation (LOOCV) were applied (**Table6.1**).

Algorithm (reference)	Description	R package (Function)	Parameters
SVM [283]	Support vector machines	e1071(svm)	Default: Kernel =radial basis
RF [284]	Random forest	randomForest (randomForest)	Default: ntree=500
KNN [285]	K-nearest neighbour	Class (knn.cv)	K=5, l=0
LDA [286]	Linear discriminant analysis	MASS (lda)	Default
ROCC [287]	Receiver operation characteristic (ROC) analyses based classifier	Rocc (o.rocc)	xgenes=all selected genes

Table 6.1: Classification algorithms applied to assess the performance of candidate biomarkers for severe pneumonia. The table shows the R package (name and specific function) and parameter settings applied in this analysis (more details in **Chapter 2**).

6.4 Results

6.4.1 Instigating the association between bacterial septicaemia and systemic responses in pneumonia

In this chapter, the main assumption is that systemic responses to bacterial infections are associated with more serious outcomes. To partially assess that, here I investigated whether mild pneumonia cases with confirmed or suspected bacterial aetiology (i.e. high-risk cases) were associated with stronger responses at the gene analytic level. In particular, I applied the empirical Bayes moderated t-test [226] to identify differentially expressed genes (adjusting for potential confounders) in mild pneumonia stratified by bacterial infection phenotypes (**Figure6.3**).

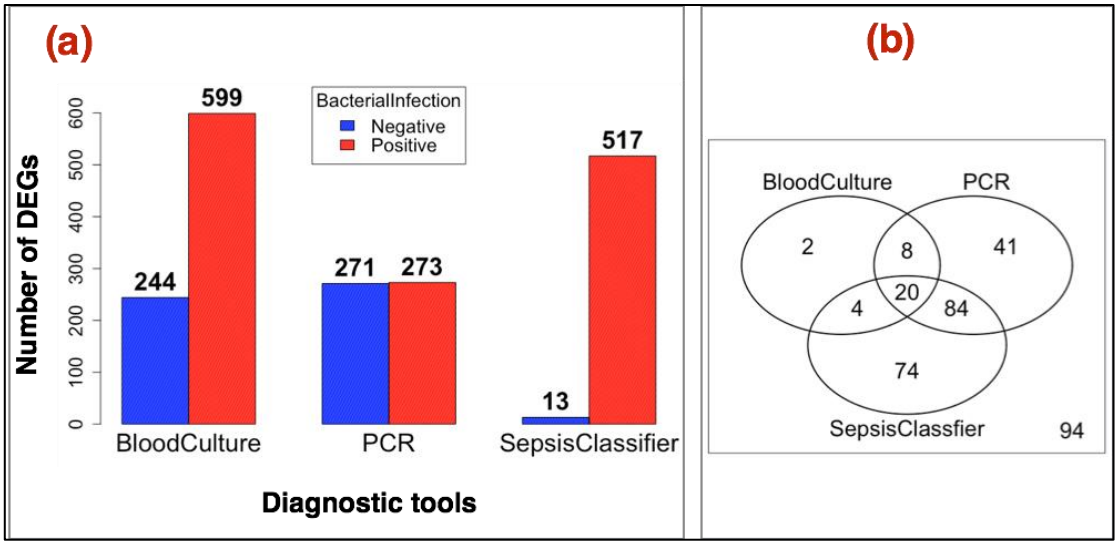


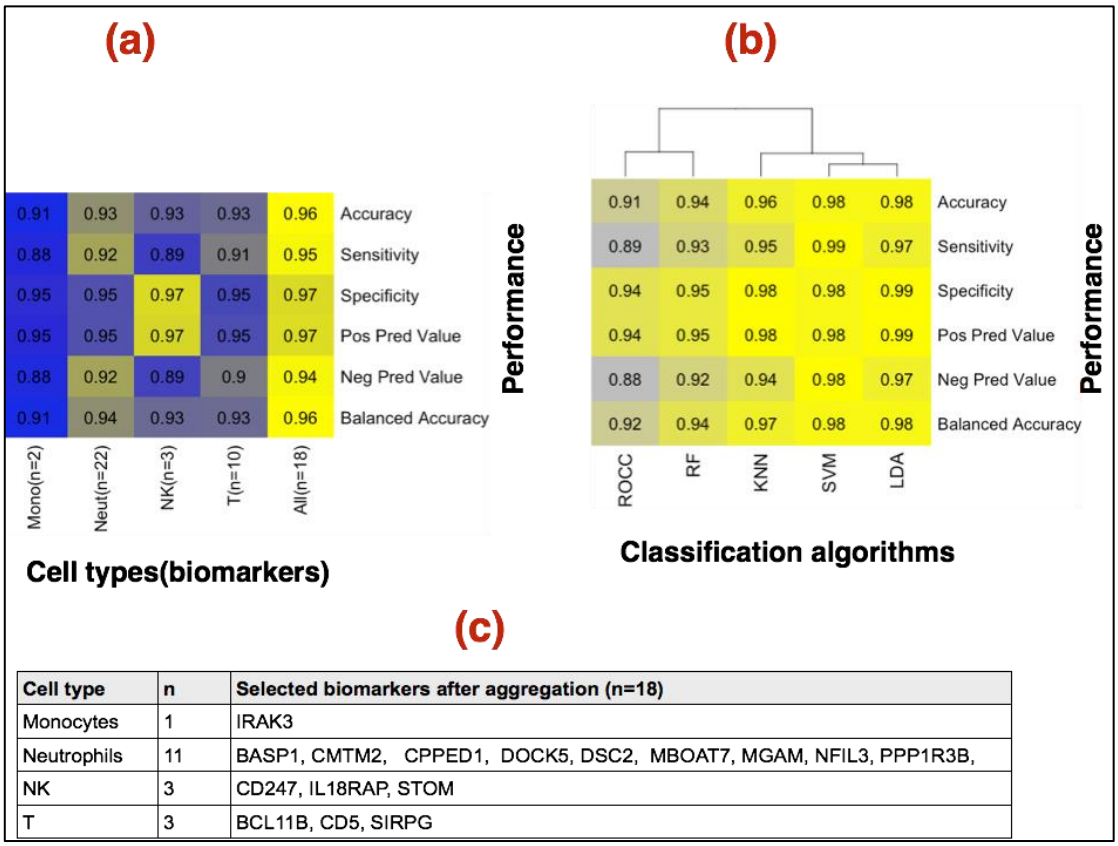
Figure 6.3: Systemic responses in mild pneumonia stratified by bacterial infection. **(a)** Number of differentially expressed genes between mild pneumonia cases and non-pneumonia controls (y-axis) stratified by bacterial infection (x-axis). **(b)** An overlap of suspected bacterial cases in the training data between the between diagnostic tools.

As shown in **Figure6.3**, bacterial infection was indeed associated with stronger responses. In particular, while no differences were observed with the PCR stratification (which is very sensitive), mild pneumonia cases with blood culture-confirmed or suspected septicaemia (using a 52-gene neonatal sepsis classifier [136]) were associated with stronger qualitative systemic

responses. Notably, while Blood-Culture and PCR results lacked sensitivity and specificity (respectively), the sepsis classifier was associated the strongest difference between the negative and positive cases, which highlights the accuracy and potential of systemic pathway-based biomarkers. Together, these findings support the hypothesis that bacterial septicaemia importantly contributes to the development of severe pneumonia in this population. In the next sections, the composition and performance of the candidate biomarkers for severe pneumonia are investigated.

6.4.2 Feature selection from the IBML list

IBML is an optimized list of marker genes (m=277) for immune cell types, which was derived in **Chapter 4** of this thesis to enhance computational deconvolution of whole blood transcriptomes. It constitutes cell type-specific marker genes for B (m=10), T (m=35), NK (m=46), Dendritic (m=9), Monocytes (m=25) and Neutrophils (m=152). To assess the classification performance of these markers in severe pneumonia, the feature selection approach described above (**Figure6.2**) was applied.



(c)

Cell type	n	Selected biomarkers after aggregation (n=18)
Monocytes	1	IRAK3
Neutrophils	11	BASP1, CMTM2, CPPED1, DOCK5, DSC2, MBOAT7, MGAM, NFIL3, PPP1R3B,
NK	3	CD247, IL18RAP, STOM
T	3	BCL11B, CD5, SIRPG

Figure 6.4: Performance assessment of candidate biomarkers selected from the IBML. **(a)** Average performances (across the five algorithms) for cellular level and aggregated (All) candidate biomarkers: Neut=neutrophils, Mono=monocytes, NK=natural killer cells and All=aggregated. **(b)** Algorithm-specific performance of the aggregated biomarkers (n=18), which are displayed in part **(c)**.

At the cellular level, more neutrophils markers (m=22) were selected than in the monocytes (m=2), NK (m=3) and T (m=10) cell types. However, the average performances across the classification algorithms (**Table6.1**) were similar (accuracy: 93%) especially with the T and NK cells (**Figure6.4a**). On the other hand, the performance improved to 96% accuracy after the aggregation (**All**) despite the reduction in the number of selected features (m=18). In particular, this set combines markers from the B (5%), T (17%), NK (17%) and 61% neutrophils compartments (**Figure6.4c**). At the algorithm specific level (**Figure6.4b**), the Support Vector Machine (SVM) and Linear Discriminant Analysis (LDA) algorithms were associated with the highest performance (accuracy=98%, respectively) on the aggregated biomarker set

(n=18). Together, this finding highlights the synergetic interplay between the adaptive and innate cell types, and the potential of cellular-based biomarkers in pneumonia. To expand the search domain for cellular candidate biomarkers, the next section applied the IBML resource (m=277) to investigate the performance of cell-correlated genes (**CCGs**).

6.4.3 Feature selection from the cell-correlated genes (CCGs, m=6369)

Cell-correlated genes (CCGs) are genes that were positively correlated with the deconvoluted (using IBML) proportions of a particular immune cell type, regardless of pneumonia status (**Figure6.5**). To derive the **CCGs** list, the IBML resource (m=277) was applied to deconvolute the sample proportions of immune cell types from the training data (n=345). For each cell type, empirical Bayes regression (limma package[226]) was applied to identify significantly correlated genes (FDR<0.05) with the deconvoluted proportions (regardless of pneumonia severity); while adjusting for the potential confounders (age, nutrition status and antibiotic usage) and multiple testing (BH method[238]) as follows:

$$Gene_{ig} = \alpha + B1 * CellProp_{ic} + Confounders_i$$

Where;

- *Gene*$_{ig}$ represents the expression value for gene **g** in sample **i**;
- **CellProp**$_{ic}$ is the proportion for cell type **c** in sample **i**,
- **B1** is the regression coefficient for **CellProp** variable,
- **Confounders**$_i$ is a vector of values for the potential confounders in sample **i**.
- α =the intercept

In particular, **Gene**$_g$ was assigned to cell type **c** if **B1**>0 and FDR <0.05 (i.e. significant positive association). To avoid duplicates, genes that were associated with multiple cell types were assigned to the cell type with highest

positive Pearson's correlation value (r). In total, the CCGs (m=6369) list contains unique genes for T (m=2063), B (m=590), NK (m=734), Monocytes (m=125) and neutrophils (m=2982). On average, the correlations ranged between **r**=0.83 in monocytes to **r**=0.98 in neutrophils (**Figure 6.5**). However, monocytes-associated genes were not eligible for feature selection (i.e. not differentially expressed in mild pneumonia).

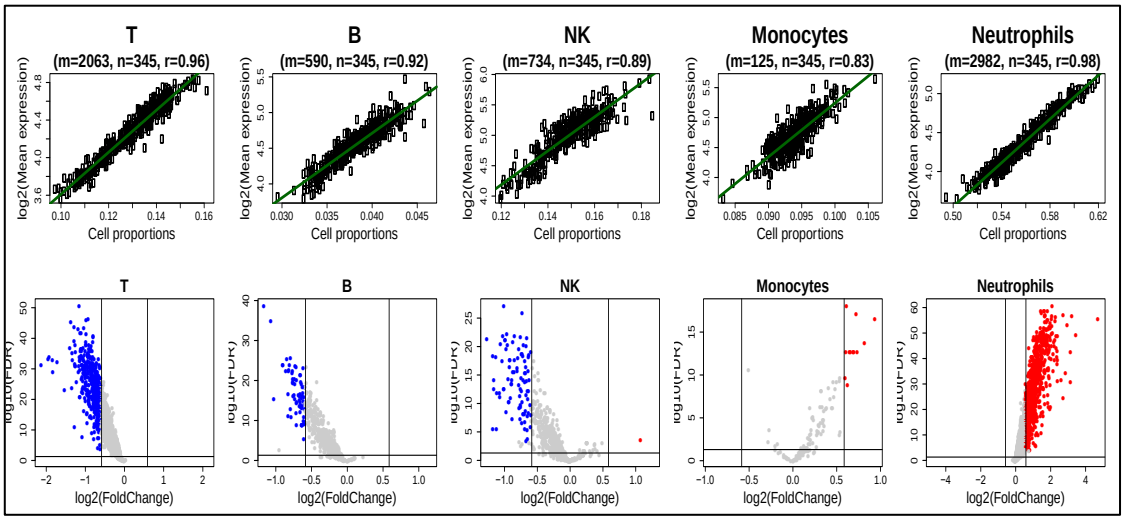
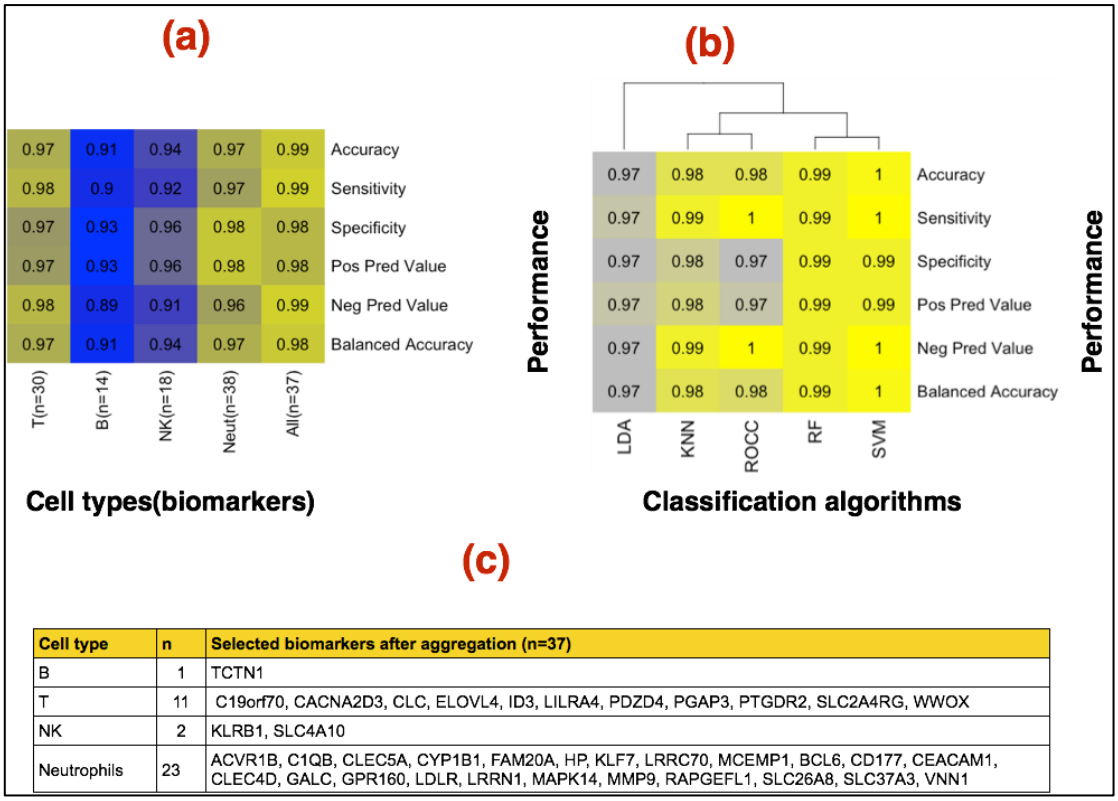


Figure 6.5: The distribution of cell-correlated genes (CCGs). The first row (scatter plots) shows the correlation between the deconvoluted proportions of each cell type (x-axis) and the mean expression values for the CCGs (y-axis). Each dot represents a sample. The second row (volcano plots) shows the differentially expressed CCGs between non-pneumonia controls and all pneumonia cases combined. Each dot represents a gene.

After feature selection, 30 (T), 14 (B), 18 (NK) and 38 (Neutrophils) biomarkers were selected at the cellular level. On average (across the classification algorithms), T cell and neutrophils based features were associated with the highest performance (accuracy=97%). Similarly, the performance improved to 99% accuracy (**Figure6.6a**) after the aggregation (m=37) comprising 62%(23) neutrophils, 30% T and 5.4%(2) NK cell markers (**Figur6c**). This finding further suggests a synergetic interplay between the neutrophil, NK, T cell-based features for the classification of severe

pneumonia cases. Again, while all the algorithms performed well (accuracy>97%), the SVM algorithm was consistently associated with the highest performance (accuracy≥100%) on this list (**Figure6.6b**), which suggests robustness. To account for pneumonia severity in the correlation structure, the next section assessed the performance of different-correlated genes (DCGs).



Cell type	n	Selected biomarkers after aggregation (n=37)
B	1	TCTN1
T	11	C19orf70, CACNA2D3, CLC, ELOVL4, ID3, LILRA4, PDZD4, PGAP3, PTGDR2, SLC2A4RG, WWOX
NK	2	KLRB1, SLC4A10
Neutrophils	23	ACVR1B, C1QB, CLEC5A, CYP1B1, FAM20A, HP, KLF7, LRRC70, MCEMP1, BCL6, CD177, CEACAM1, CLEC4D, GALC, GPR160, LDLR, LRRN1, MAPK14, MMP9, RAPGEFL1, SLC26A8, SLC37A3, VNN1

Figure 6.6: Feature selection and performance assessment of cell-correlated genes (CCGs) candidate biomarkers. **(a)** Average performances (across the five algorithms) for cellular level and aggregated (All) candidate biomarkers: Neut=neutrophils, NK=natural killer cells and All=aggregated **(b)** Algorithm-specific performance of the aggregated biomarkers (n=37), which are displayed in part **(c)**.

6.4.4 Feature selection from differentially-correlated genes (DCGs)

Differentially correlated genes (DCGs) are the genes with different (strength or direction) correlation structures (with the deconvoluted proportions of immune cell types) between the pneumonia severity states (**Figure6.7**). For example, the average correlation coefficient with the proportions of NK cells

varied between **r**=0.16 (among the non-pneumonia controls) and r=87 (among the pneumonia cases). Statistically, these are the genes with significant interaction terms between pneumonia severity and sample proportions of immune cell types (effect-modified genes).

To derive the **DCGs** list, the IBML resource (m=277) was applied to deconvolute the sample proportions of immune cell types from the training data (n=345). For each cell type, empirical Bayes regression (limma package[226]) was applied to identify genes with significant interactions (FDR<0.05) between pneumonia severity and the deconvoluted (using IBML) proportions as follows:

$$Gene_{ig} = \alpha + B1 * CellProp_{ic} + B2 * Pneum_i + B3 * Pneum\_CellProp_{ic} + Confounders_i$$

Where

- *Gene$_{ig}$* represents the expression value for gene **g** in sample **i**;
- **CellProp$_{ic}$** is the proportion for cell type **c** in sample **i**,
- **Pneum$_i$**=0 if sample i is a non-pneumonia control, otherwise=1
- **Pneum_CellProp$_c$=**the interaction term between pneumonia status and the deconvoluted proportions of cell type **c**.
- **B3** is the regression coefficient for the interaction term (Pneum_CellProp)
- **Confounders$_i$** is a vector of values for the potential confounders in sample **i**.
- α =the intercept

In particular, **Gene$_g$** was assigned to cell type **c** if **B3**>0 and FDR <0.05 (i.e. significant interaction or effect modification). To avoid duplicates, genes that

were associated with multiple cell types were assigned to the cell type with highest positive correlation (Pearson's r) among the pneumonia cases (**Pneum**$_i$=1). In total, the DCGs list contains 720 genes for T (m=172), NK (m=133) and Neutrophils (m=415) cell types (**Figure6.8**).

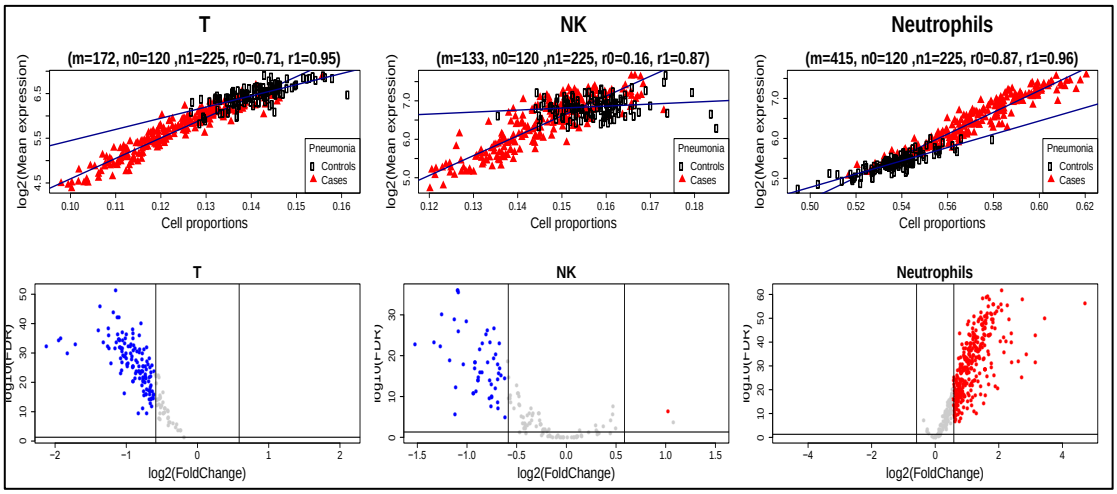
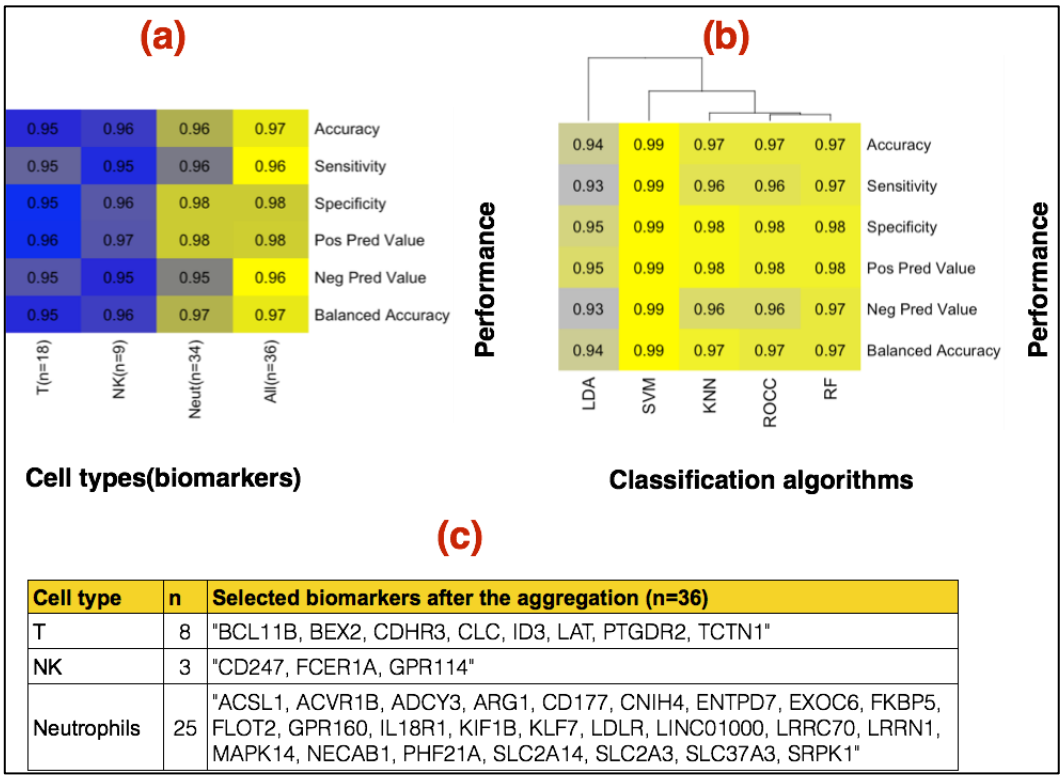


Figure 6.7: The distribution of differentially correlated genes (DCGs). The first row (scatter plots) shows the correlation between the deconvoluted proportions of each cell type (x-axis) and average profile of its selected genes (y-axis) stratified pneumonia. Each dot represents a sample. The second row (volcano plots) shows the differential gene expression of DCGs between non-pneumonia controls and all pneumonia cases combined. Each dot represents a gene.

At the cellular level, more neutrophils biomarkers (m=34) were selected than the T (m=18) and NK (m=9) cellular compartments. However, the average performances (accuracy 95%-96%) across the classifiers were similar (**Figure6.8a**). Similarly, the performance slightly improved to accuracy=97% after the aggregation (m=36) representing 22%(m=8) T, 8%(m=3) NK and 70%(m=25) neutrophils cell types (**Figure6.8c**). Again, the SVM algorithm was associated with the highest performance (accuracy=99%) in this list (**Figure6.8b**) further suggesting robustness. Together, these findings further highlight the potential of cellular-based biomarkers in pneumonia, and the robustness of the SVM algorithm. In the next section, the cellular candidate biomarker sets (IBM, CCGs and DCGs) were aggregated.



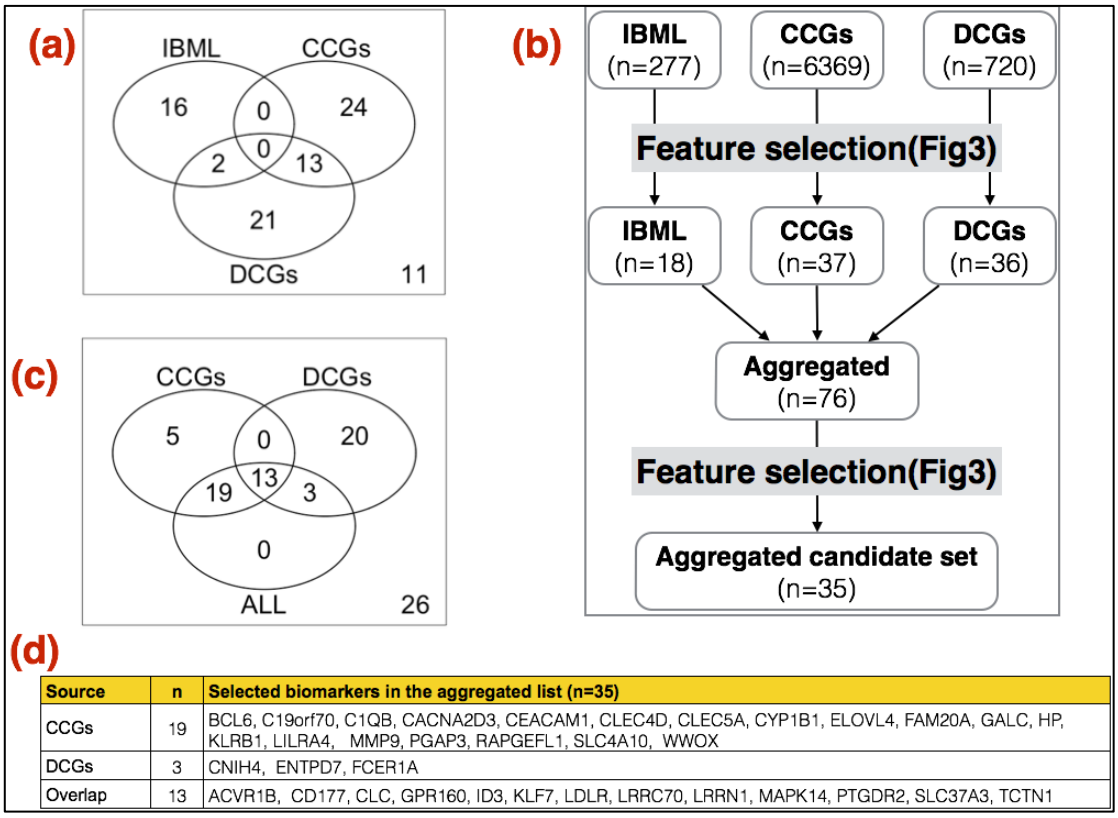
(c)

Cell type	n	Selected biomarkers after the aggregation (n=36)
T	8	"BCL11B, BEX2, CDHR3, CLC, ID3, LAT, PTGDR2, TCTN1"
NK	3	"CD247, FCER1A, GPR114"
Neutrophils	25	"ACSL1, ACVR1B, ADCY3, ARG1, CD177, CNIH4, ENTPD7, EXOC6, FKBP5, FLOT2, GPR160, IL18R1, KIF1B, KLF7, LDLR, LINC01000, LRRC70, LRRN1, MAPK14, NECAB1, PHF21A, SLC2A14, SLC2A3, SLC37A3, SRPK1"

Figure 6.8: Feature selection and performance assessment of differentially correlated genes (DCGs) candidate biomarkers: **(a)** Average performances (across the five algorithms) for the cellular level and aggregated (All) candidate biomarkers: Neut=neutrophils, and NK=natural killer cells. **(b)** Algorithm-specific performance of the aggregated biomarkers (n=36), which are displayed in part **(c)**.

6.4.5 Aggregation of the cellular candidate biomarker sets

So far this chapter has derived three cellular pathway-based candidate biomarker sets representing IBML (m=18), CCGs (m=37) and DCGs (m=36). However, 80% of these biomarkers were unique to a particular set (**Figure6.9a**). To derive a unified candidate set, these markers (m=76) were aggregated and subjected to the same feature selection approach (**Figure6.2**) as illustrated in **Figure6.9b**.



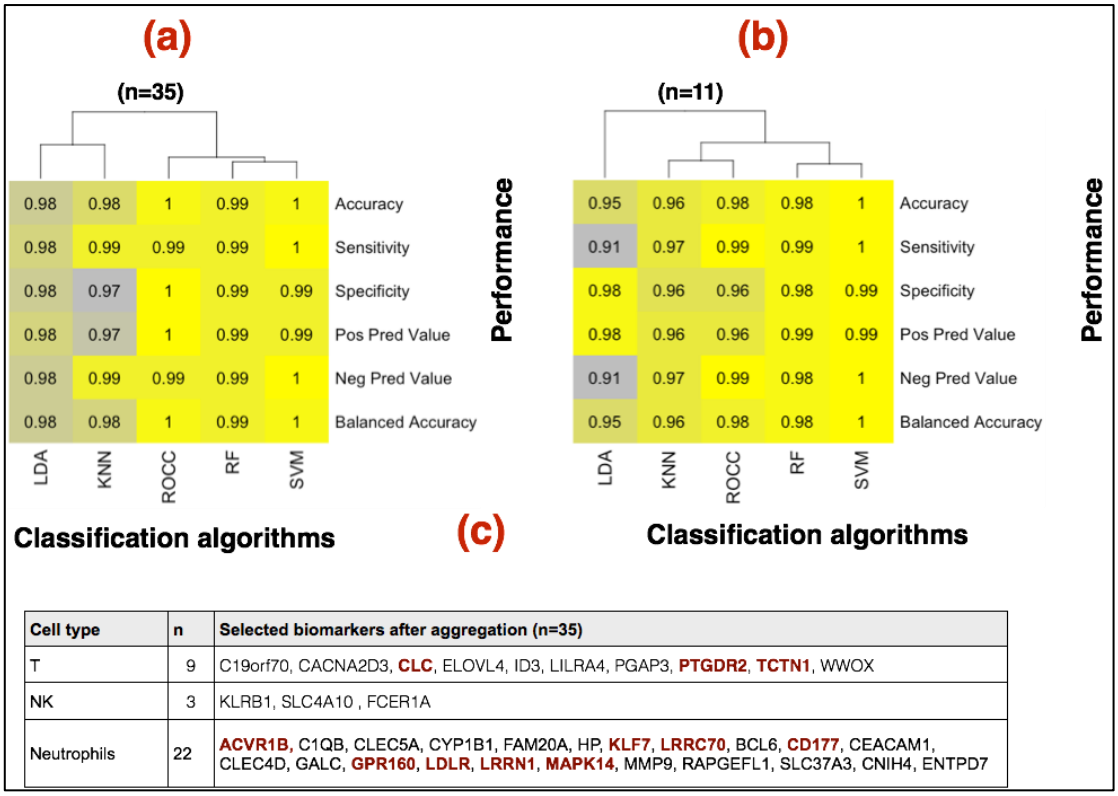
(d)

Source	n	Selected biomarkers in the aggregated list (n=35)
CCGs	19	BCL6, C19orf70, C1QB, CACNA2D3, CEACAM1, CLEC4D, CLEC5A, CYP1B1, ELOVL4, FAM20A, GALC, HP, KLRB1, LILRA4, MMP9, PGAP3, RAPGEFL1, SLC4A10, WWOX
DCGs	3	CNIH4, ENTPD7, FCER1A
Overlap	13	ACVR1B, CD177, CLC, GPR160, ID3, KLF7, LDLR, LRRC70, LRRN1, MAPK14, PTGDR2, SLC37A3, TCTN1

Figure 6.9: Aggregation of cellular-based biomarkers: **(a)** The overlap of cellular candidate biomarker sets (IBML, CCGs and DCGs). **(b)** An illustration of biomarker aggregation using the feature selection in Figure6.2. **(c)-(d)** the distribution of the aggregated list (no IBML markers were selected). ALL=the aggregated list, IBML=an integrated blood marker list (derived in Chapter 4), CCGs=Cellular correlated genes; DCGs=differentially correlated genes.

In total, 35 biomarkers were selected from the CCGs (m=32) and DCGs (n=16) sets but not IBML (m=0). Among them, 13 markers were common in both lists (**Figure 6.9c-d)**. At the cellular level, this candidate list (n=35) represents 63%(22) neutrophils, 26%(9) T and 9%(3) NK cell types (**Figure6.10c**). Notably, this list was consistently associated with high

performance across all the classification algorithms (accuracy: 98%-100%), which suggest robustness (**Figure6.10a**). Again, the performance of the SVM algorithm remained the highest.



Cell type	n	Selected biomarkers after aggregation (n=35)
T	9	C19orf70, CACNA2D3, **CLC**, ELOVL4, ID3, LILRA4, PGAP3, **PTGDR2**, **TCTN1**, WWOX
NK	3	KLRB1, SLC4A10 , FCER1A
Neutrophils	22	**ACVR1B**, C1QB, CLEC5A, CYP1B1, FAM20A, HP, **KLF7**, **LRRC70**, BCL6, **CD177**, CEACAM1, CLEC4D, GALC, **GPR160**, **LDLR**, **LRRN1**, **MAPK14**, MMP9, RAPGEFL1, SLC37A3, CNIH4, ENTPD7

Figure 6.10: Performance assessment of the aggregated cellular biomarkers. (a) Performance of an aggregation of all cellular biomarker sets (n=35). **(b)** Performance of a reduced model selected the overlapping markers between CCGs and DCGs models (m=11). **(c)** Cellular distribution of the selected markers (m=35). Overlapping markers in the reduced model (m=11) are highlighted in red. CCGs=Cellular correlated genes; DCGs=differentially correlated genes.

To assess whether a reduced model would replicate the performance, I assessed the performance of the overlapping features (m=13). After feature selection (**Figure6.2**) 11 markers were selected. However, this set was associated with a reduced performance in the LDA, KNN and ROCC algorithms (**Figure6.10b**) potentially suggesting lack of robustness. Nevertheless, based on the SVM and RF-based classifiers (which remained unchanged), this list potentially presents a reduced model for the cellular-

based biomarkers (highlighted in red in **Figure6.10c**). Together, these findings have consistently highlighted the potential of cellular pathway-based biomarkers in severe pneumonia; the synergetic interplay between the neutrophils, T and NK cell-based features; and the robustness of the SVM algorithm. In the next section, I assessed the performance of sepsis markers in severe pneumonia.

6.4.6 Feature selection from the 52-gene sepsis classifier

In **Figure6.3**, septicaemia was associated with an increased frequency of differentially expressed genes in mild pneumonia suggesting its importance in the development of severe pneumonia outcomes. To directly assess whether sepsis biomarkers can distinguish severe pneumonia cases, I applied the same feature selection (**Figure6.2**) on the validated 52-gene neonatal sepsis classifier by Smith et al. (2014) [136] (**Figure6.11**).

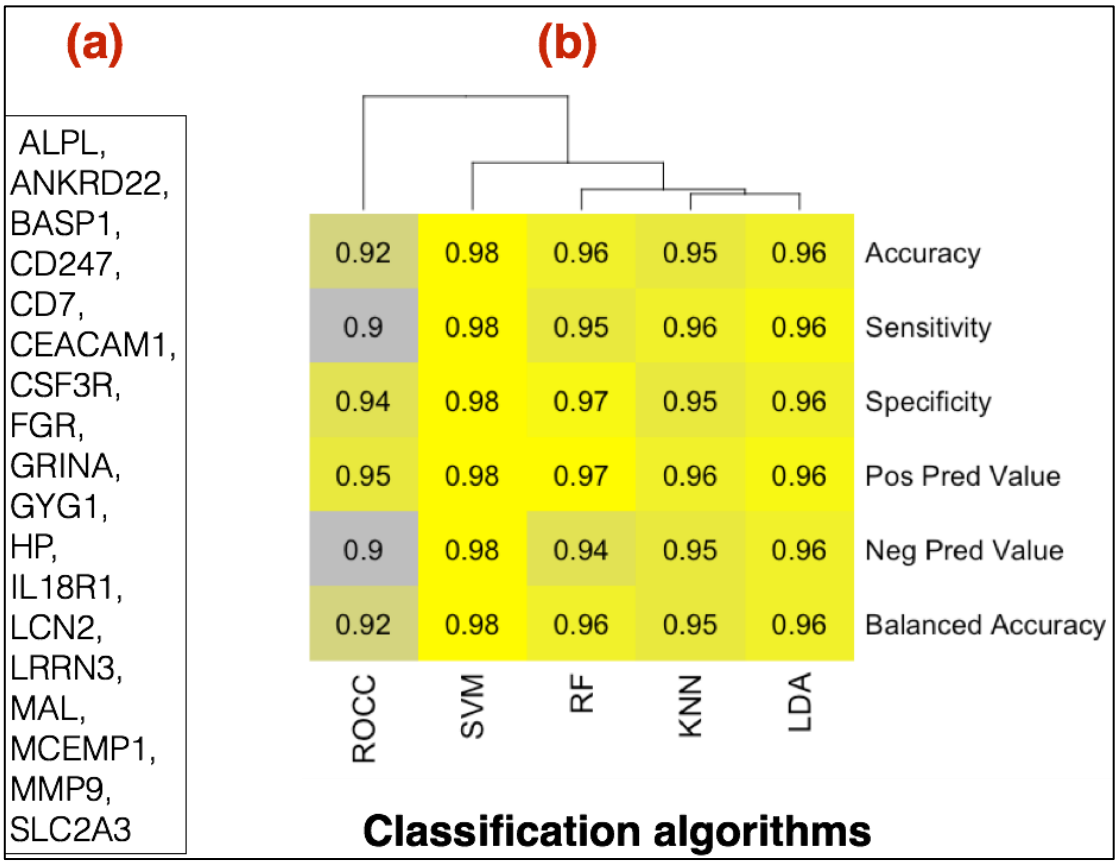


Figure 6.11: Performance assessment of sepsis markers in pneumonia. **(a)** Optimal selected marker (m=18). **(b)** Performance assessment using five algorithms (x-axis).

Of them (m=52), 18 markers were selected (**Figure6.11a**). Interestingly, the performance was similar to the cellular biomarker sets especially with the SVM algorithm (Accuracy=98%). To further investigate this agreement, I assessed the expression profiles of the cellular pathway-based candidate biomarkers using the Edinburgh neonatal sepsis database [136]. Interestingly, at least 70% (**Figure6.12**) of the cellular-based biomarkers were differentially expressed (FDR<0.05) in sepsis; suggesting a stronger agreement between the cellular and sepsis markers in pneumonia severity. Thus, these findings further support the important contribution of bacterial septicaemia in the development of serious pneumonia outcomes. To further improve the performance of the candidate biomarkers, the next section, aggregated the cellular-based and sepsis markers.

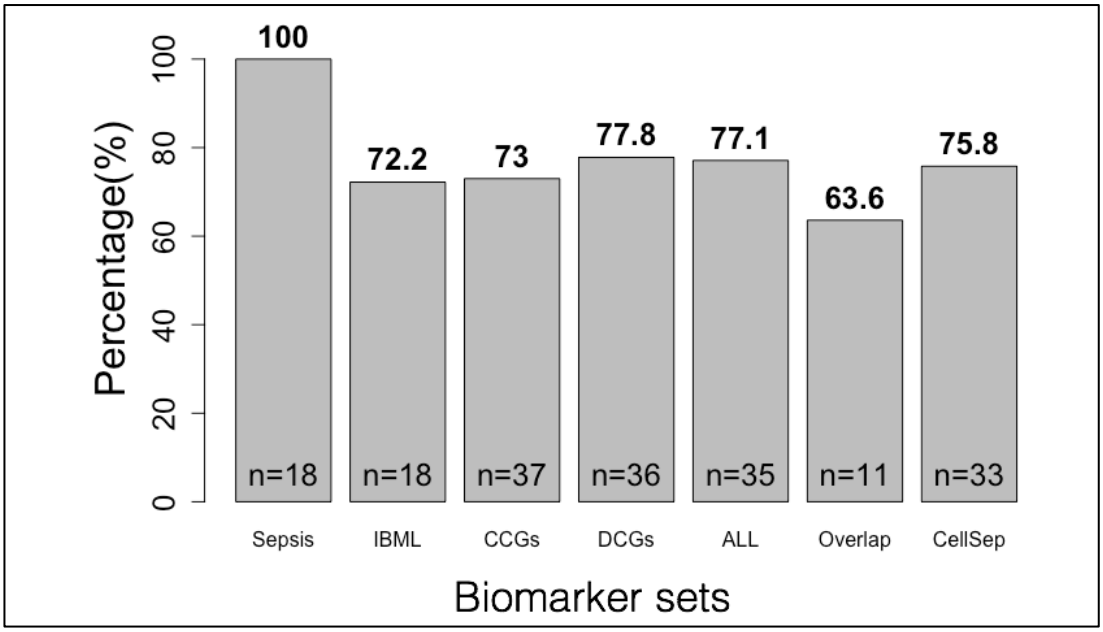


Figure 6.12: High agreement between sepsis and severe pneumonia. Each bar represents the proportion (y-axis) of candidate biomarkers of severe pneumonia (x-axis) that were differentially expressed (FDR<0.05) in neonatal sepsis using the Edinburgh neonatal sepsis database [136]. IBML=Markers selected from the IBML resource derived in Chapter 4 (section 6.4.2). **CCGs**=Cellular correlated genes (section 6.4.3). **DCGs**=differentially correlated genes (section 6.4.4). ALL=A combination cellular-based biomarkers (section 6.4.5). **Sepsis**=Biomarkers selected from the validated bacterial sepsis classifier (section 6.4.6). **Overlap**=an intersection of **CCG** and **DCG** biomarkers. **CellSep**=A combination of ALL and sepsis biomarker sets.

6.4.7 Aggregation of cellular-based and sepsis biomarkers

So far I have independently assessed the performance of cellular-based and validated sepsis biomarkers. To derive a unified candidate biomarker set, cellular-based (aggregated set, m=35) and sepsis (m=18) biomarkers were aggregated. After feature selection (**Figure6.2**), 33 markers were selected; and this biomarker set is called **CellSep**. Except for the LDA algorithm, this aggregation was associated with an improved performance especially on the ROCC, KNN, Random forest (RF) algorithms whilst the SVM remained high (**Figure6.13a**). Interestingly, the selected biomarkers (**Figure6.13c)** clearly distinguished severe pneumonia cases from the non-pneumonia controls using the unsupervised principal component analysis (**Figure6.13b**), which suggest their applicability with less complicated algorithms. Together, this

finding highlights the important contribution of bacterial septicaemia in severe pneumonia; which also support the central hypothesis that systemic pathway responses underpin the development of severe pneumonia outcomes. To summarise the feature selection process, the classification performance of all the candidate biomarkers were directly compared (next section).

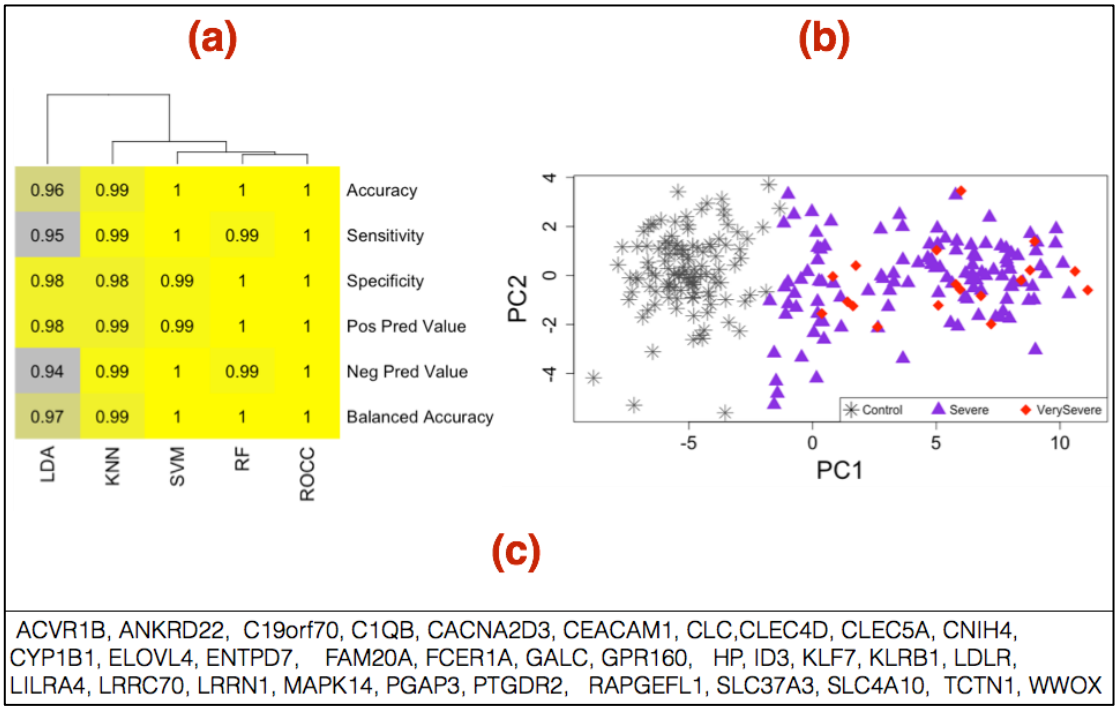


Figure 6.13: Aggregation of cellular-based and validated sepsis biomarkers. Supervised **(a)** and unsupervised (principal component analysis) **(b)** performance assessment of the selected biomarkers **(c)**.

6.4.8 Perfomance summary of candidate biomarker sets

This section provides a summary and direct performance comparison of the candidate biomarker sets across all the algorithms. In **Figure6.14**, the candidate biomarker sets are ordered (ascending) by the average accuracy (big circle) across the classification algorithms. Firstly, the KNN, RF and SVM algorithms were consistently associated with higher performance than the LDA and ROCC classifiers. Notably, the SVM (triangular symbols) had the highest performance in all the candidate biomarkers sets.

While all the biomarker sets were associated with high performance (accuracy>95%), the aggregation of cellular-based and sepsis biomarkers (**CellSep**, m=33) were associated with the highest performance (**Figure6.14**). Therefore, according to this analysis, **CellSep** is the final candidate biomarker set. To validate the candidate biomarker sets, the SVM-based classifiers were applied to predict severe pneumonia cases in the **Basse** dataset (n=158), which was kept independently from the primary analyses of this thesis (next section).

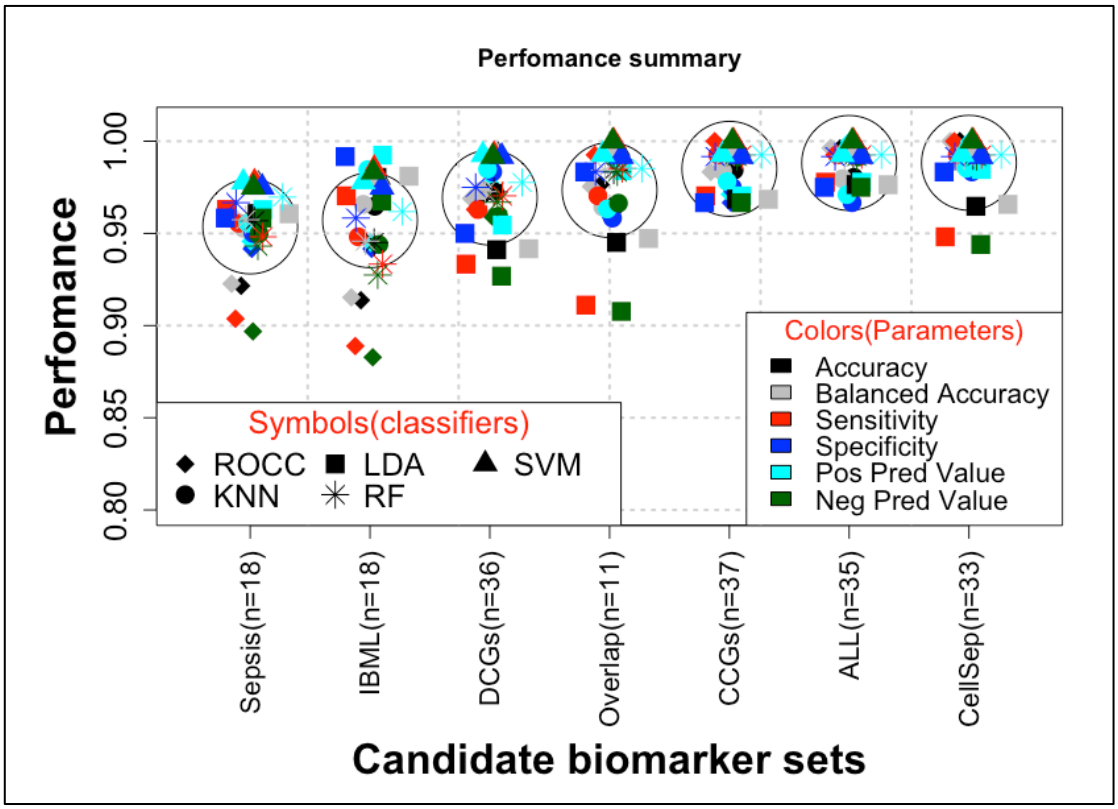


Figure 6.14: An training data performance summary of candidate biomarkers. For each candidate biomarker set (x-axis), each dot represents an algorithm-specific (symbol legend) performance (colour legend). The large circles represent the average accuracy across all the algorithms. Sepsis=Markers from the 52-validated neonatal sepsis classifier [136]; Overlap=an overlap between cell-correlated genes (CCGs) and differentially correlated genes (DCGs). ALL=an aggregation between IBML, CCGs and DCGs. CellSep=an aggregation between ALL and Sepsis biomarker sets

6.4.9 Independent validation of candidate biomarkers using the Basse dataset

In this thesis, data were collected from two different geographical regions within The Gambia. By study design, the training data set was collected from a semi-urban coastal area (**Fajara**) while the validation data set represents the rural population in the upper region (**Basse**). So far, I have applied the training data set to investigate cellular and molecular pathway responses; as well as train and test candidate biomarkers for severe pneumonia. To validate the performances of the candidate biomarkers, the SVM-based classifiers were applied to predict severe pneumonia cases in the validation data set (**Basse**, n=158).

Prior to that, the validation database was subjected to data cleaning to identify potentially mislabelled samples (**Figure6.15**). In particular, principal component analysis (PCA) and the Neonatal Sepsis classifier (m=52)[136] identified eight suspicious samples (three non-pneumonia controls and five severe pneumonia cases). Subsequently, the suspicious samples were eliminated. However, to assess the negative impact of mislabelled samples in biomarker analysis, I conducted sensitivity analysis to validate the performance of the candidate biomarkers with and without the potentially mislabelled samples (**Figure6.16**).

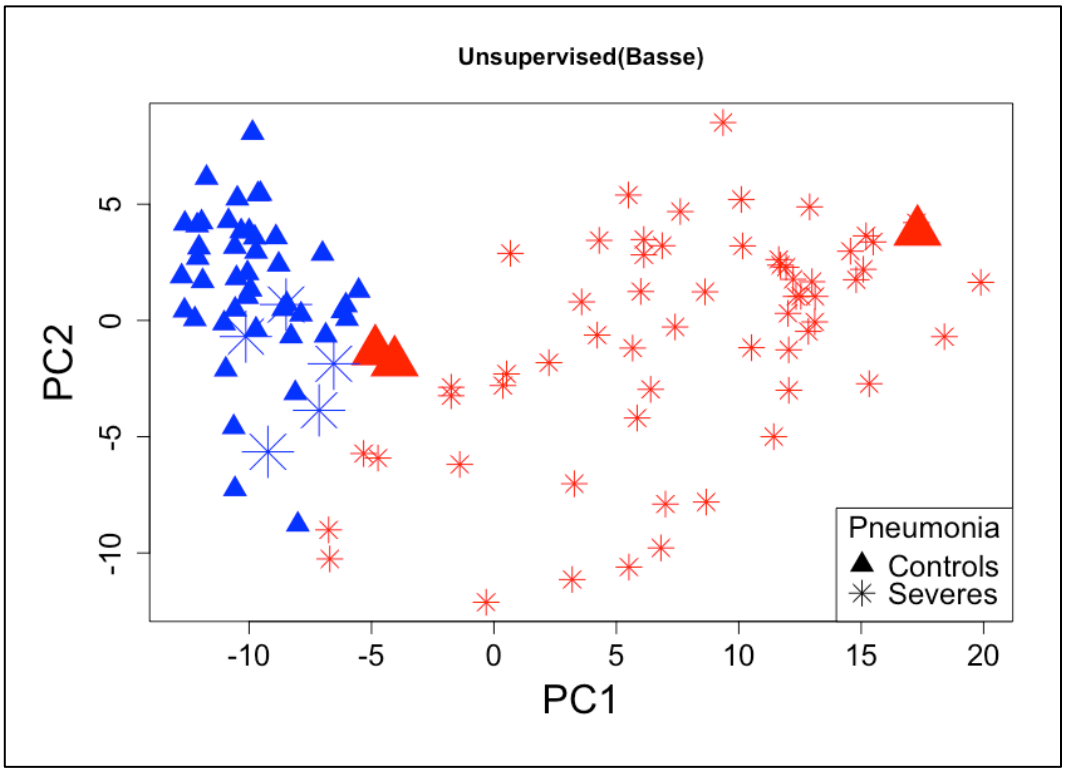


Figure 6.15: Unsupervised description of samples in the validation data set. The plot shows a principal component analysis plot using the most variable genes (m=100) selected using the coefficient of variation (CV) statistic across all the samples. The symbols represent pneumonia status (legend) and colours represent suspected bacterial infection (red=positive, blue=negative) predicted using the neonatal sepsis classifier [136].

In general, the performance was high, and further improved after the data cleaning across all the candidate biomarker sets. Similar to the training data, an aggregation of sepsis and cellular-based biomarkers (**CellSep**) was also associated with the highest performance in the validation data set (accuracy=98%). Notably, the same performance (accuracy=98%) was replicated by the Sepsis (m=18) and DCGs (m=36) biomarker sets. Together, these findings independently highlight the potential systemic biomarkers in pneumonia, and the important contribution of bacterial septicaemia in the development of serious pneumonia outcomes.

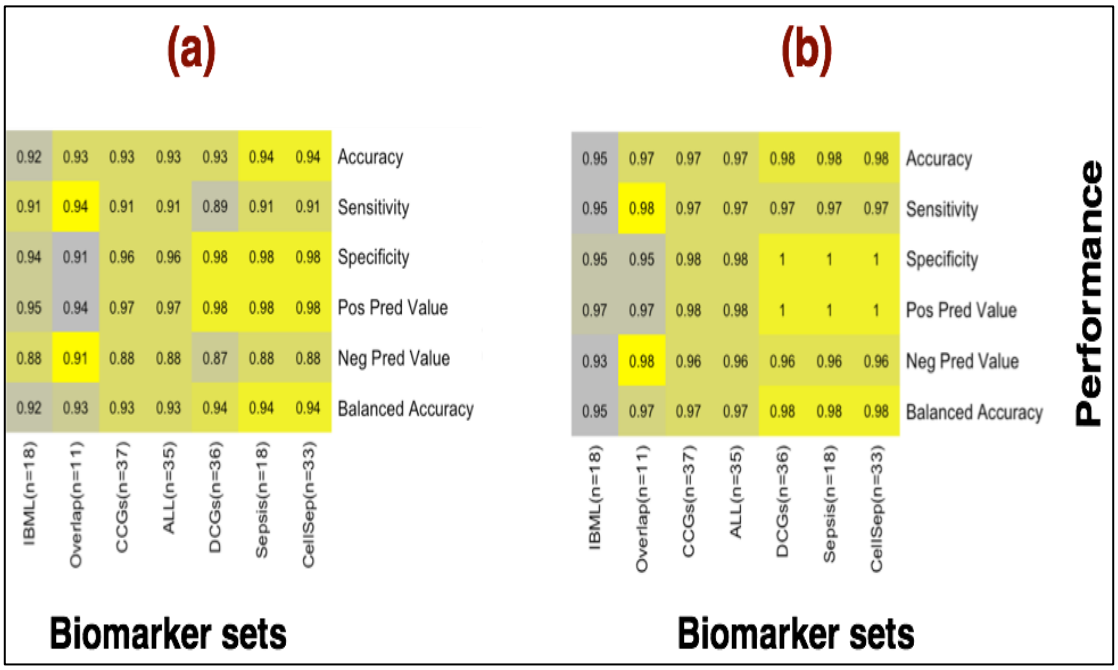


Figure 6.16: Independent validation of the candidate biomarker sets using the Basse data set. The figure shows the classification performance before (a) and after (b) data cleaning. In each figure, the candidate biomarker sets (x-axis) are ordered from lowest to highest performer. Classification was based on the support vector machine (SVM) algorithm.

6.4.10 Molecular stratification of mild pneumonia cases into high and low risk groups

To assess the potential applicability of the candidate biomarkers (i.e. as a proof of concept), the 33-gene SVM classier (**CellSep**) was applied to predict mild pneumonia cases that were at the higher risk of developing severe pneumonia in both the training and validation data sets. In particular, 71 and 22 cases were predicted as high-risk cases in the training and validation datasets, respectively (**Table6.2, Table6.3**).

In both data sets, the demographic characteristics were similar between the low-risk and high-risk. On the other hand, the high-risk cases were associated with poor clinical outcomes such as elevated neutrophils counts, depletion of lymphocytes (P-value<0.001), lower oxygen saturation, and higher heart rate, respiratory rate and body temperature. Further, the high-risk mild cases were associated with higher prevalence of chest x-ray

pathology, and bacterial septicaemia (blood culture isolates, PCR and the sepsis classifier). Notably, this stratification shows strong association with the original neonatal sepsis classifier [136], which further suggests the importance of bacterial septicaemia in the development of serious pneumonia outcomes. Together, this finding highlights the potential of host-based systemic biomarkers in the clinical stratification and treatment modalities of patients presenting at the clinic with mild pneumonia.

Training data (Fajara)			
Factor	**High-Risk (n=71)**	**Low-Risk (n=19)**	**P-value**
Demographics characteristics			
Age in months, median (IQR)	15.7 (10.8, 24.1)	13.3 (7.0, 19.7)	0.13
Sex			0.67
Female	30 (42%)	7 (37%)	
Male	41 (58%)	12 (63%)	
Season			0.94
Dry	38 (54%)	10 (53%)	
Wet	33 (46%)	9 (47%)	
Under-weight (WAZ), mean (SD)	-1.2 (1.2)	-1.2 (1.0)	0.90
Stunting (HAZ), mean (SD)	-0.7 (1.1)	-0.6 (1.1)	0.74
Wasting (WHZ), mean (SD)	-1.1 (1.3)	-1.3 (1.4)	0.64
Clinical phenotypes			
Haemoglobin, mean (SD)	9.8 (1.8)	10.4 (1.7)	0.25
Neutrophils, median (IQR)	53.7 (46.0, 66.6)	42.0 (33.3, 47.9)	<0.001
Lymphocytes, median (IQR)	38.2 (26.0, 46.4)	51.0 (45.2, 57.9)	<0.001
Platelets, mean (SD)	380.0 (161.6)	433.5 (126.3)	0.23
Temperature, median (IQR)	38.2 (37.7, 39.1)	37.6 (36.7, 38.3)	<0.001
Cough	69 (97%)	12 (63%)	<0.001
Heart_rate, mean (SD)	151.8 (15.0)	149.0 (17.8)	0.48
Respiratory rate, mean (SD)	56.0 (8.7)	52.2 (8.6)	0.092
Oxygen saturation, mean (SD)	**96.7 (1.7)**	**98.2 (1.6)**	**0.001**
Chest X-ray (positive)	42 (91%)	8 (73%)	0.092
PCR (Positive)	40 (56%)	11 (58%)	0.90
Sepsis classifier (positive)	58 (82%)	3 (16%)	<0.001
BloodCulture			0.80
Contaminants	2 (3%)	0 (0%)	
No growth	61 (86%)	17 (89%)	
S.aureus	2 (3%)	1 (5%)	
S.pneumoniae	6 (8%)	1 (5%)	

Table 6.2: Demographic and clinical characteristics of mild pneumonia cases in the training data. The samples were molecularly classified into low and high-risk groups using the SVM based 33-gene classifier representing the cellular and sepsis biomarkers (CellSep).

(b) Validation data (Basse)			
Factor	**High-Risk (n=22)**	**Low-Risk (n=24)**	**P-value**
Demographic characteristics			
Age in months, median (IQR)	8.2 (5.0, 20.6)	10.9 (4.5, 17.4)	0.88
Sex			0.55
Female	10 (45%)	13 (54%)	
Male	12 (55%)	11 (46%)	
Season			0.31
Dry	6 (27%)	10 (42%)	
Wet	16 (73%)	14 (58%)	
Under-weight (WAZ), mean (SD)	-1.5 (1.1)	-1.1 (1.2)	0.21
Stunting (HAZ), mean (SD)	-1.0 (1.3)	-0.5 (1.3)	0.18
Wasting (WHZ), mean (SD)	-1.3 (1.2)	-1.1 (1.1)	0.64
Clinical phenotypes			
Haemoglobin, mean (SD)	**8.8 (2.4)**	**10.2 (1.7)**	**0.048**
WBC_Total2, median (IQR)	16.4 (11.6, 28.9)	9.6 (7.1, 14.6)	0.006
Neutrophils, median (IQR)	**58.0 (49.3, 68.6)**	**42.5 (31.5, 52.0)**	**0.005**
Lymphocytes, median (IQR)	**34.7 (25.0, 43.6)**	**52.9 (41.4, 61.6)**	**0.004**
Platelets, mean (SD)	**289.2 (143.9)**	**366.2 (107.6)**	**0.074**
Temperature, median (IQR)	39.0 (38.5, 39.4)	38.2 (37.2, 38.7)	0.007
Cough	20 (91%)	15 (62%)	0.024
Heart_rate, mean (SD)	159.2 (14.5)	149.7 (17.3)	0.050
Respiratory rate, mean (SD)	**62.3 (9.8)**	**60.3 (12.2)**	**0.56**
Oxygen saturation, mean (SD)	**95.9 (2.1)**	**97.2 (1.7)**	**0.025**
Chest X-ray(positive)	17 (81%)	10 (59%)	0.13
PCR (Positive)	5 (23%)	3 (12%)	0.36
Sepsis classifier (positive)	19 (86%)	4 (17%)	<0.001
BloodCulture			0.51
Contaminants	3 (14%)	4 (17%)	
No growth	17 (77%)	20 (83%)	
S.aureus	1 (5%)	0 (0%)	
S.pneumoniae	1 (5%)	0 (0%)	

Table 6.3: Demographic and clinical characteristics of mild pneumonia cases in the validation data. The samples were molecularly classified into low and high-risk groups using the SVM based 33-gene classifier representing the cellular and sepsis biomarkers (**CellSep**).

6.5 Discussion

In this chapter, I have investigated systemic candidate biomarkers for severe pneumonia under the hypothesis that bacterial septicaemia underpins the development of severe pneumonia outcomes. In particular, the main objective was to derive a classifier for potential stratification of patients presenting at the clinic with mild pneumonia into low-risk and high-risk treatment groups.

This analysis had several motivations. Firstly, the burden of pneumonia remains unacceptably high, and mainly due to lack of optimal and affordable diagnostic tools [1]. Evidently, this analysis lacked complete data on pneumonia aetiology. To mitigate the burden of childhood pneumonia, innovative approaches such as biomarkers are required to enhance the clinical stratification and appropriate treatment modalities for mild pneumonia cases. In particular, systemic response (host-based) biomarkers potentially present an opportunity for a paradigm shift in the clinical practice of pneumonia [154]. Here, the observation that cases with bacterial septicaemia were associated with stronger systemic responses highlighted the importance of bacterial septicaemia in severe pneumonia outcomes. Whole blood is readily accessible tissue in clinical practice and molecular profiling has become a mainstay of genomic research and future translation medicine [71].

6.5.1 Systemic biomarkers in severe pneumonia

According to the World Health Organisation (WHO), *a biomarker is a chemical, its metabolite, or the product of an interaction between a chemical and some target molecule or cell that is measured in the human body [27]*.

Ideally, biomarkers should be accessible, accurate (sensitive and specific), robust, reproducible, and reflect the disease pathogenesis [58-62, 446]. To derive robust and biologically meaningful biomarkers for severe pneumonia, here I coupled cellular pathway biology and machine-learning approaches to derive systemic transcriptomic classifier features.

Using that approach, I have derived a 33-gene cellular pathway-centric classifier comprising markers from neutrophils, T and NK cells. Notably, this classifier was consistently associated with high performance in the training and validation data sets. In particular, a support vector machine based classifier accurately distinguished severe pneumonia cases in the training data (accuracy=100%) and independently validated (accuracy=98%) in the Basse data set, which was kept independent from primary analyses. On the other hand, all the misclassified samples were either associated with poor RNA quality, bacterial infection (positive controls and negative cases) or antibiotic usage. Together, these results highlight the accuracy and robustness of systemic cellular pathway-centric transcriptomic biomarkers in pneumonia (**Supplementary Table6A1**).

6.5.2 Strengths

This approach had several advantages. Firstly, the cellular and molecular pathway responses consistently supported the central hypothesis that systemic pathway responses underpin the development of severe pneumonia outcomes (**Chapter 4** and **Chapter 5**). Potentially, these biomarkers reflect the pathogenesis of severe pneumonia, and therefore robust. Interestingly, the interplay between neutrophils, T and NK cell-based

features was consistently associated with high performance across a range of classification algorithms, which suggest robustness.

To further enhance the robustness, a powerful and regularized machine-learning feature selection (Elastic Net[282]) approach was applied to select compatible and non-redundant biomarkers. Further, this approach was repeated 100 times to select a robust combination of biomarkers that were consistently selected together all the time (100 times). Unlike the filter methods that independently focus on the strength of individual gene features [345, 346], here the focus was to derive an optimal and synergetic combination of features that reflect the pathway biology of pneumonia. Furthermore, several classification algorithms were applied to assess the performance of the candidate features. Among them, the support vector machine (SVM) had the highest performance, which is consistent with several comparative studies [319, 453, 454].

6.5.3 The agreement between cellular centric and sepsis biomarkers

Clinically, early identification of bacterial pneumonia cases is very important to prevent serious outcomes but the standard tools have several limitations [28]. While this analysis lacked complete microbial benchmark data, there was a significant agreement between the cellular-based and bacterial sepsis markers. Firstly, mild pneumonia cases with suspected septicemia (using the sepsis classifier [136]) or blood culture confirmed positive results were associated with stronger systemic responses, which suggested the important contribution of bacterial septicaemia in the development of severe pneumonia outcomes. Independently, validated sepsis markers (n=18) were

associated with high classification performance in severe pneumonia (accuracy=98% in both the training and validation data sets). On the other hand, at least 75% of the cellular pathway-centric biomarkers were also significantly associated with neonatal sepsis in the Edinburgh database [136]. This agreement highlights the importance of bacterial septicaemia in the development of severe pneumonia outcomes, and is consistent with epidemiological findings that bacterial aetiology is more associated with serious outcomes including mortality [5]. Clinically, these candidate biomarkers present a powerful and accessible potential for enhanced stratification and treatment modalities for patients presenting at the clinic with mild pneumonia.

6.5.4 Limitations

While the classification performance (based on Leave-One-Out Cross Validation (LOOCV)) was high, it is worth noting that the classifiers were trained using default values for the hyperparameters, which potentially generated suboptimal results. To improve the performance, nested cross-validation (as described in **Chapter 2**, section 2.6, pages 95-96) is recommended to identify the best combination of the hyperparameters.

Further, the analysis lacked complete aetiological data. Ideally, it would be straightforward to derive candidate biomarkers for treatment modalities of mild pneumonia if viral and bacterial cases were known. However, the existing standard diagnostic tools are suboptimal to guarantee “gold standard” data. Secondly, it is worth noting that the current study design was not adequate for a proper investigation of prognostic biomarkers mainly due

to lack of follow-up data including survival outcomes. Preferably, a longitudinal study design would be ideal to investigate prognostic biomarkers for predicting mild cases that would progress to severe states.

Furthermore, these candidate biomarkers lacked similar studies for further independent validations. Generally, biomarkers require rigorous validations to establish robustness and generalizability prior to routine clinical application. Therefore, more validation work will follow (more details in **Chapter7**). Nonetheless, robust measures were applied to address the objectives of this chapter within the realm of the available resources.

6.5.5 Conclusion

Despite the limitations in this chapter, the identification of a highly accurate and robust 33-gene classifier was presented representing the systemic cellular pathway responses involving the neutrophils, T and NK cell immune compartments. Importantly, the findings of this chapter suggest the hypothesis that bacterial septicaemia underpins the development of severe pneumonia outcomes, which is vital for treatment modalities. In conclusion, these findings present a novel and powerful approach for the early identification of mild pneumonia cases at the higher risk of developing severe outcomes. However, further validations are required.

Chapter 7: Discussion

7.1 Introduction

This chapter presents a summary of key findings, discussion of the strengths and limitations, and future outlook of this thesis.

7.2 Motivation

Despite the scaled efforts to improve child survival [4], infections-attributable mortality rates remain high in children younger than five years old (under-five) [2]. In particular, pneumonia remains the leading infectious cause of under-five mortality especially in resource-limited countries like the sub-Saharan Africa [2, 3, 5, 350]. Pneumonia has a complex aetiology including viral and bacterial infections, and disease pathogenesis is not fully understood. Consequently, prevention, diagnosis, and treatment of pneumonia remain public health challenges. In particular, the derivation of optimal vaccines [29] and early identification of bacterial pneumonia cases remain the major public health challenges for promoting child survival [24, 55, 154].

Further, the existing vaccines [25, 26] and the standard diagnostic tools (i.e. Chest x-ray, blood culture) are rarely available in remote settings where the burden is highly concentrated [1]. In these settings, the diagnosis of pneumonia is further complicated by the presence of co-morbidities with overlapping clinical presentations such as malaria and diarrhoea, which potentially lead to misclassifications of patients [5]. Thus, while effective antibacterial therapies exist, the delayed treatment of bacterial pneumonia cases is associated with the development of more serious outcomes

including mortality [5, 24]. On the other hand, unnecessarily presumptive antibiotic treatment is not cost-effective and potentially exacerbating the spread of antibiotic resistance [24, 40]. Therefore, innovative approaches are required to enhance the stratification and treatment modalities especially for patients presenting at a resource-constrained clinic with mild pneumonia. Potentially, gaining a deeper understanding of the systemic pathway responses in pneumonia would unravel key immuno-modulation candidates for novel vaccine candidates, robust biomarkers and therapeutic targets [154].

7.3 Approach and data resources

In this thesis, it was hypothesised that systemic pathway responses are associated with the development of severe pneumonia outcomes. In other words, while the compartmentalised local immune responses (within the lungs) are often crucial for the detection and clearance of the invading pathogens[455], the involvement of systemic (blood-based) pathway responses contributes significantly to the development of serious pneumonia outcomes including mortality [80, 116, 436].

Importantly, whole blood potentially contains a large number of biomarkers and clinically accessible tissue for pathophysiological investigations. Further, whole blood genome-wide profiling has become a mainstay of genomic research and future translation medicine in range of diseases including cancer, infections and autoimmunity[123]. Potentially, whole blood transcriptomics presents a powerful and innovative solution to enhance the clinical practice of childhood pneumonia in resource-limited settings. In

particular, this approach presents a comprehensive opportunity to gain a deeper insight into the pathogenesis of severe pneumonia and explore the potential of systemic response-based biomarkers (i.e. a paradigm shift from pathogen-based to host-based factors [154]).

Mainly, this thesis has sought to address two main objectives: (i) investigation of systemic pathway (cellular and molecular) responses associated with the clinical severe pneumonia states and (ii) identification of candidate biomarkers for high-risk pneumonia cases among the patients presenting at the clinic with mild pneumonia. To address that, I have analysed a whole blood transcriptome comprising the training (n=345) and validation (n=158) data sets. Whole blood samples were collected from a matched-case control study involving Gambian children and infants aged 2-59 months old. The cases were clinically classified as mild, severe and very severe pneumonia, and prospectively matched (by age, sex and location) to non-pneumonia community controls to mitigate the potential effect of confounding (**Chapter 2**).

Firstly, sample size re-assessment revealed that the study groups were statistically powered to address the primary objectives (**Chapter 3**). To ensure data quality and completeness, the central data resources (the transcriptome and corresponding metadata) were subjected to a range of quality control measures including intensive data cleaning, pre-processing[147] and batch-effect correction[148, 150, 195] (**Chapter 3**). Further, potential confounding effects such as age and nutrition status

differences were further investigated and accounted for during data analysis. To guard against false discoveries, all analyses were adjusted for multiple testing using the *Benjamini & Hochberg's* false discovery rate (FDR) control procedure[238], which is less stringent than the traditional *Bonferroni* procedure[170]. Together, this thesis has adequate and high-quality data resources for primary and validation analyses.

7.4 The significant involvement of systemic pathway responses in severe pneumonia

To gain a deeper insight into pathogenesis of severe pneumonia, the systemic pathways responses were investigated at the cellular and molecular analytic levels.

7.1.1 Cellular pathway responses

Whole blood is complex tissue with heterogeneous cellularity including myeloid and lymphocytes cell types, which vary in proportions within and between samples and often correlates with clinical phenotypes such as pneumonia severity[102, 200, 358, 456, 457]. To investigate the cellular pathway responses, here I applied a powerful and yet cost-effective computational solution called computational deconvolution analysis [72, 73, 200] (**Chapter 4**). Briefly, this approach estimates cell type-specific information (i.e. proportions and cell type-specific gene expression signatures) directly from the whole blood transcriptomes without incurring intermediate costs for cell-sorting techniques such as FACs analysis, which are further limited by the availability of cell surface markers[139, 200, 456]. However, computational deconvolution analysis remains an area of active research, which requires more reliable input resources (i.e. marker genes lists, expression signatures and algorithms) to facilitate its application in the

mainstream analyses of high-throughput data such as whole blood transcriptomes[160, 200, 371].

To further enhance the computational deconvolution analysis of whole blood transcriptomes and related datasets, this thesis has applied a data fusion approach to derive an optimal and integrated blood marker list called **IBML**. This analysis (derivation of IBML) had several motivations. Firstly, marker genes (*semi-supervised deconvolution*) are more applicable than gene expression signatures (*partial deconvolution*) because they are robust to array platform-specific differences [160, 201]. Further, IBML provides a single unified application resource because the existing marker gene resources (MGR) for a given immune cell type were found to be molecularly distinct (i.e. have little overlap) consequently presenting the end-user with a selection challenge. Furthermore, a data filtering approach was applied because an aggregation of all the eligible markers (m=3,475) was associated with a reduced performance suggesting the presence of non-specific and noisy markers. Briefly, IBML contains highly specific marker genes (m=277) for T (m= 35), B(m=10), NK(m= 46), Dendritic (m=9) , Monocytes (m=25) and neutrophils (m=152) cell types, and was associated with enhanced and robust performance in a range of independent benchmark whole blood transcriptomes. Together, IBML presents a unified and optimal application resource for enhanced computational deconvolution analysis of whole blood.

Subsequently, the IBML resource was applied to deconvolute the training whole blood transcriptome (n=345). As anticipated [367, 368, 370, 378, 385],

pneumonia severity was associated with significant depletion of adaptive response mediators (B, T, Dendritic), and elevation of pro-inflammatory innate mediators (Monocytes and neutrophils). Unexpectedly, this analysis further revealed the depletion of natural killer (NK) cells in severe pneumonia. While this finding is consistent with the observation that NK-depleted mice are susceptible to lung infections[394, 398], the role of NK cells in human pneumonia remains elusive[458] and controversial[388], and therefore requires further investigations. Nevertheless, this potential protective role of NK cells presents a novel immuno-modulation target for mitigating the burden of pneumonia worldwide. Together, these findings highlight the potential of computational deconvolution analysis, and support the central hypothesis that systemic cellular pathway responses underpin the development of severe pneumonia states.

7.1.2 Molecular pathway responses

Systemic molecular responses were investigated at the gene and pathway analytic levels. At the gene analytical level, absolute fold changes and the number of differentially expressed genes (DEGs) increased significantly with pneumonia severity. To gain a comprehensive insight into the systemic pathway responses in severe pneumonia, the differentially expressed genes (DEGs) were investigated using a range of biochemical pathway database resources (KEGG, REACTOME, GO and HALLMARK). At the pathway analytic level, pneumonia severity was associated with a significant interplay between the innate, adaptive and metabolic pathways, which support the central hypothesis that systemic pathway responses underpin the development of severe pneumonia states (**Chapter 5**).

In particular, while pro-inflammatory innate and cholesterol metabolism pathway responses were associated with all the severe pneumonia states (from mild to very severe pneumonia); the development of severe and very severe pneumonia outcomes were predominantly associated with the co-inhibition of NK cell signalling (innate) and the adaptive effector responses especially in T cells; and the activation of fatty acid and lipid metabolism. Notably, these findings are consistent with the cellular pathway responses (**Chapter 4**) including the potential involvement of NK cells in the pathogenesis of severe pneumonia. Further, it was also observed that very severe pneumonia cases were predominantly associated with antibacterial responses. This finding particularly suggests the importance of bacterial septicaemia in the development of serious pneumonia outcomes. Clinically, this presents an opportunity for the application of systemic response-based candidate biomarkers for early detection and treatment modalities of severe pneumonia cases. Together, these findings consistently support the central hypothesis that systemic pathway (cellular and molecular) responses underpin the development of severe pneumonia states.

7.5 The potential of systemic pathway response-based candidate biomarkers of severe pneumonia

Pneumonia has a complex aetiology but mostly dominated by bacteria and viruses[28]. While viral pneumonia cases are more prevalent, delayed appropriate treatment of the bacterial pneumonia cases is associated with more serious outcomes including mortality [5]. However, the aetiological stratification of pneumonia cases remains a clinical challenge especially in

resource-constrained settings where the burden of childhood pneumonia is highest [40]. Therefore, innovative and cost-effective approaches for early identification (and treatment modalities) of high-risk pneumonia cases are required to mitigate the global burden of childhood pneumonia and promote child survival[1, 24]. Potentially, host-based systemic pathway response derived biomarkers presents a direct and robust approach for enhancing clinical stratification and treatment modalities of pneumonia cases[56, 60-62, 65, 70, 140, 154, 446, 459].

To assess the potential of systemic biomarkers, I coupled cellular pathway biology with machine learning approaches to derive a whole blood-based transcriptomic classifier for the detection of severe pneumonia cases. Using that approach, I have derived a 33-gene transcriptomic classifier comprising candidate biomarkers from the NK, T and neutrophils cellular pathways. This signature (m=33) was robustly associated with high performance across a range of classification algorithms in both the training (accuracy=99%), validation (accuracy=98%) datasets. These findings highlight the potential of systemic pathway response-based transcriptomic biomarkers in pneumonia.

It is worth noting that due to lack of complete “gold standard” aetiology data, this analysis compared extreme clinical pneumonia severity phenotype labels (non-pneumonia versus severe and very severe pneumonia) to derive the candidate biomarkers. Ideally, these biomarkers are intended to for the prediction of high-risk pneumonia cases among the patients presenting at the clinic with mild pneumonia. Firstly, this approach was motivated by the

predominance of cellular centric pathway responses in severe pneumonia (**Chapter 4** and **Chapter 5**). For treatment modalities of bacterial pneumonia cases, this feature selection approach was driven by the assumption that bacterial septicaemia contributes significantly to the development of serious pneumonia outcomes.

In this thesis, this assumption was supported by several observations. Firstly, pneumonia severe was significantly associated septicaemia (**Table3.2**, **Chapter 3**). Further, effect-modification analysis revealed that pneumonia cases with BloodCulture-confirmed bacterial infection or suspected septicaemia (using the validated sepsis classifier) were associated with an increased frequency of differentially expressed genes (**Chapter 3**) including the mild cases (**Chapter6**). In **Chapter 5**, very severe pneumonia cases were also associated with predominant antibacterial pathway responses. Furthermore, there was a strong agreement between the cellular-based and sepsis biomarkers (**Chapter6**). In particular, a subset (m=18) of the 52-gene validated neonatal sepsis classifier [136] was associated with similar classification performance (99% and 98% accuracy in the training and validation data, respectively). Conversely, at least 75% of the cellular based biomarkers were also differentially expressed in the neonatal sepsis database. Moreover, epidemiology studies consistently associate the bacterial aetiology especially *Streptococcus pneumonia* with more serious outcomes[5, 7, 9, 22, 41, 350]. Together, these candidate biomarkers present a novel approach for early identification and treatment modalities of high-risk patients presenting at a clinic with mild pneumonia. In the next

sections, I highlight the overall strengths and limitations of this thesis followed by suggested future recommendations.

7.6 Study strengths

This section highlights the study strengths. To our knowledge, this is the first comprehensive and adequately powered study in the sub-Saharan African region to investigate the systemic pathway responses and candidate biomarkers for childhood severe pneumonia. In particular, the application of whole blood transcriptome presents a comprehensive and innovative approach for elucidating the pathogenesis of pneumonia and investigating robust candidate biomarkers for treatment modalities. Importantly, whole blood is a clinically accessible and acceptable tissue for pathophysiological investigations [102]. Further, whole blood genome-wide profiling has become the mainstay of biomedical research and future translation medicine[71, 121, 123, 154, 450, 451]. Therefore, these findings have a potential to improve the clinical practice of pneumonia especially in resource-limited settings where the burden is very high [2, 3].

This thesis has sufficient and high quality data resources. The study groups were adequately powered to enable meaningful primary analyses (n=345) and independent validations (n=158). Secondly, the whole blood transcriptome was annotated with high quality metadata records (clinical, demographic databases), and the databases were carefully curated to ensure data quality and completeness. To mitigate biased results, a prospective matched-case-control study design was implemented to account

for the potential confounding effects of age, sex, season and location. To further enhance the data quality, several statistical approaches were applied to account for non-biological variations in the data including raw data pre-processing, batch-effect correction and confounder analysis to identify key covariates for subsequent analyses (**Chapter 3**). To guard against false discoveries, all the analyses were adjusted for multiple testing.

Beyond data quality assurance, this thesis has benefited from a range of data science approaches including data fusion, computational deconvolution analysis, computational pathway analysis, and pathway biology-coupled machine learning analyses. Interestingly, these approaches complimented each other and generated robust results, which were consistent with the central hypothesis. In particular, computational deconvolution analysis derived the **IBML** marker gene resource (m=277), which further enhanced the investigation of cellular pathways (**Chapter 4**). This resource presents a single unified application resource for streamlined analysis and interpretations of future whole blood transcriptomes. Further, the cellular pathway responses (**Chapter 4**) were consistent with the molecular findings from the computational pathway analysis (**Chapter 5**). Notably, these findings revealed a novel involvement of NK cells in pneumonia severity, which presents a potential target for immune-modulation and case management. Finally, the application of machine learning approaches on the cellular pathway-centric transcriptomic features derived a robust and highly accurate candidate classifier (**CellSep**, m=33) for the detection of severe pneumonia cases (**Chapter6**). This signature presents a novel and

accessible approach for early identification and potential treatment modalities for patients at the high-risk of developing severe pneumonia among the patients presenting at a resource-constrained clinic with mild pneumonia. Inevitably, this study has some limitations (**next section**).

7.7 Study limitations

In this section, the limitations of this study are highlighted. Firstly, the original study design had some limitations. While the study groups were statistically powered and matched for the potential confounders, it is worth noting that this is an observational study design. Thus, while observational studies reflect the real conditions in clinical practice (hence more generalizable)[460], they are often susceptible to potential bias[68, 355, 460, 461]. Further, whole blood samples were collected at a single time point without follow-up data (i.e. cross-sectional study). Therefore, these findings provide a limited cross-sectional view of the systemic pathway responses at the individual level, and lack information on patient survival, which is vital for deriving long-term interventions. Preferably, a longitudinal study design may provide more insights into the individual-level trajectories, and enable proper investigation of prognostic biomarkers.

As already mentioned, viruses and bacteria are the predominant causes of pneumonia; but the aetiological stratifications of pneumonia cases remains a clinical challenge[28]. While **Chapter6** investigated candidate biomarkers for severe pneumonia, lack of complete aetiology data was a setback. Ideally, knowledge of samples with viral, bacterial and co-infections would have enabled the formal investigations of aetiology-specific pathways and

classifier signatures. However, the existing gold standard diagnostic tools have several limitations [55].

Similarly, these findings were potentially limited by the shortcomings of the existing data science resources. Firstly, while processing and analysis of microarray data has improved, generalisation of results is often limited by the technical variations across the array platforms [148-150, 341, 462]. Further, while genomic profiling is becoming affordable [155, 463]; the cost is still too high for routine clinical application especially in resource-constrained settings.

Further, the derivation of the IBML resource (**Chapter 4**) lacked more detailed resources to enable *deep deconvolution* of minor cell subcomponents (i.e. Tregs cells [464]) at different activation stages. Furthermore, the findings from the molecular pathway analysis reflect the current state of the existing knowledge archived in the biochemical pathway databases (**Chapter 5**). Potentially, these resources may provide different insights if more data come online. Similarly, while **Chapter6** has derived and independently validated candidate biomarkers for severe pneumonia, further independent validations are required. However, such data are not available especially in the sub-Saharan African region.

Nonetheless, the current study has adequate and high quality resources, and applied robust approaches to address the central hypothesis and study objectives. Moreover, the molecular pathway responses were independently replicated across a range of biochemical resources, and were consistent with

the cellular pathway responses. Importantly, these findings support the central hypothesis that systemic pathway responses underpin the development of severe pneumonia outcomes. To address some of the limitations, the next section provides suggested recommendations for future work.

7.8 Suggested recommendations and future outlook

Mainly, the findings of this thesis would benefit from further validations. In the short term, further analyses using publicly available data resources are required to enhance and validate the IBML resources and the candidate biomarkers. For IBML, the priority is to improve the cell type coverage and granularity, and conduct further performance validations using whole blood transcriptomes from a range of diseases. Similarly, the candidate biomarkers require further validation at least in data sets with known pneumonia aetiology. Further, both resources could benefit from alternative and computationally intensive approaches such as cross-validated (i.e. bootstrapping) and ensemble feature selection approaches [465]. Potentially, those approaches may generate more robust and cost-effective (reduced) biomarker sets for routine clinical practice.

In the long-term perspective, longitudinal studies are required to validate the current results. To validate the computational deconvolution analysis results, standard approaches such as FACS [466-468] could be applied to re-assess the cellular pathway responses especially the potential involvement of NK in pneumonia severity. Simultaneously, the cell-sorted blood samples could be

applied to molecularly validate the cellular responses at the cellular level (i.e. cell type-specific transcriptome analysis). Further, PCR-based methods [469-471] could be applied to validate the candidate biomarkers of severe pneumonia (**Chapter6**). Ideally, the 33-gene classifier can be applied to stratify mild pneumonia cases (at recruitment) into low-risk (treatment withheld) and high-risk (antibiotic treatment administered immediately) groups, and monitor temporal changes in clinical symptoms. In expectation, both groups should resolve their symptoms with time. Alternatively, another low-risk group can be randomised to an antibiotic treatment as the current treatment standard. In expectation, the antibiotic intervention should have no significant differences between the low-risk groups.

7.9 Conclusions

Together, the findings of this thesis consistently support the central hypothesis that systemic pathway (cellular and molecular) responses underpin the development of severe pneumonia outcomes. Notably, the potential involvement of NK cells in the pathogenesis of pneumonia present a novel immune-modulation target for mitigating the burden of pneumonia. Further, the discovery of the 33-gene cellular pathway-centric classifier (supported by the observed strong association between bacterial septicaemia and pneumonia severity) potentially presents a novel and accessible approach for early identification and treatment modalities of high-risk mild pneumonia cases. In conclusion, these findings present a strong foundation for innovative future studies aimed at mitigating the burden of childhood pneumonia especially in resource-limited settings (i.e. the sub-Saharan Africa) where the burden is highly concentrated.

Chapter 8: References

1. WHO: **Revised WHO classification and treatment of childhood pneumonia at health facilities**. In. Edited by Maternal n, child and adolescent health. Geneva: World Health organisation; 2014.

2. **Liu L, Oza S, Hogan D, Perin J, Rudan I, Lawn JE, Cousens S, Mathers C, Black RE:** Global, regional, and national causes of child mortality in 2000-13, with projections to inform post-2015 priorities: an updated systematic analysis. *Lancet* 2015, **385**(9966):430-440.

3. **You D, Hug L, Ejdemyr S:** Global, regional, and national levels and trends in under-5 mortality between 1990 and 2015, with scenario-based projections to 2030: a systematic analysis by the UN Inter-agency Group for Child Mortality Estimation (vol 386, pg 2275, 2015). *Lancet* 2015, **386**(10010):2256-2256.

4. UN: **The Millennium Development Goals Report 2015**. In.: United Nations; 2015.

5. **Walker CL, Rudan I, Liu L, Nair H, Theodoratou E, Bhutta ZA, O'Brien KL, Campbell H, Black RE:** Global burden of childhood pneumonia and diarrhoea. *Lancet* 2013, **381**(9875):1405-1416.

6. **Zar HJ, Madhi SA, Aston SJ, Gordon SB:** Pneumonia in low and middle income countries: progress and challenges. *Thorax* 2013, **68**(11):1052-1056.

7. **Rudan I, Boschi-Pinto C, Biloglav Z, Mulholland K, Campbell H:** Epidemiology and etiology of childhood pneumonia. *Bulletin of the World Health Organization* 2008, **86**(5):408-416.

8. **Liu L, Johnson HL, Cousens S *et al*:** Global, regional, and national causes of child mortality: an updated systematic analysis for 2010 with time trends since 2000. *Lancet* 2012, **379**(9832):2151-2161.

9. **Rudan I, O'Brien KL, Nair H *et al*:** Epidemiology and etiology of childhood pneumonia in 2010: estimates of incidence, severe morbidity, mortality, underlying risk factors and causative pathogens for 192 countries. *Journal of global health* 2013, **3**(1):010401.

10. **Usuf E, Bottomley C, Adegbola R, Hall A:** Pneumococcal Carriage in Sub-Saharan Africa—A Systematic Review. . *PLoS ONE* 2014, **9**(1).

11. **Wonodi CB, Deloria-Knoll M, Feikin DR *et al*:** Evaluation of risk factors for severe pneumonia in children: the Pneumonia Etiology Research for Child Health study. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2012, **54 Suppl 2**:S124-131.

12. **Kwambana BA, Barer MR, Bottomley C, Adegbola RA, Antonio M:** Early acquisition and high nasopharyngeal co-colonisation by Streptococcus pneumoniae and three respiratory pathogens amongst Gambian new-borns and infants. *BMC infectious diseases* 2011, **11**:175.

13. **Reddy EA, Shaw AV, Crump JA:** Community-acquired bloodstream infections in Africa: a systematic review and meta-analysis. *The Lancet infectious diseases* 2010, **10**(6):417-432.

14. **Fonseca Lima EJ, Mello MJ, Albuquerque MF, Lopes MI, Serra GH, Lima DE, Correia JB:** Risk factors for community-acquired pneumonia in children under five years of age in the post-pneumococcal conjugate vaccine era in Brazil: a case control study. *BMC pediatrics* 2016, **16**(1):157.

15. **Howie SR, Schellenberg J, Chimah O *et al*:** Childhood pneumonia and crowding, bed-sharing and nutrition: a case-control study from The Gambia. *The international journal of tuberculosis and lung disease : the official journal of the International Union against Tuberculosis and Lung Disease* 2016, **20**(10):1405-1415.

16. **Rudan I, O'Brien KL, Nair H *et al*:** Epidemiology and etiology of childhood pneumonia in 2010: estimates of incidence, severe morbidity, mortality, underlying risk factors and causative pathogens for 192 countries. *Journal of global health* 2013, **3**(1):010401.

17. **Cilloniz C, Ewig S, Gabarrus A, Ferrer M, Puig de la Bella Casa J, Mensa J, Torres A:** Seasonality of pathogens causing community-acquired pneumonia. *Respirology* 2017, **22**(4):778-785.

18. **Bojang A, Jafali J, Egere UE, Hill PC, Antonio M, Jeffries D, Greenwood BM, Roca A:** Seasonality of Pneumococcal Nasopharyngeal Carriage in Rural Gambia Determined within the Context of a Cluster Randomized Pneumococcal Vaccine Trial. *PLoS One* 2015, **10**(7):e0129649.

19. **Brewster DR, Greenwood BM:** Seasonal variation of paediatric diseases in The Gambia, west Africa. *Annals of tropical paediatrics* 1993, **13**(2):133-146.

20. **Enwere G, Cheung YB, Zaman SM *et al*:** Epidemiology and clinical features of pneumonia according to radiographic findings in Gambian children. *Tropical medicine & international health : TM & IH* 2007, **12**(11):1377-1385.

21. **Feikin DR, Feldman C, Schuchat A, Janoff EN:** Global strategies to prevent bacterial pneumonia in adults with HIV disease. *The Lancet infectious diseases* 2004, **4**(7):445-455.

22. **Pletz MW, Welte T, Ott SR:** Advances in the prevention, management, and treatment of community-acquired pneumonia. *F1000 medicine reports* 2010, **2**:53.

References

23. **Abubakar I, Gautret P, Brunette GW, Blumberg L, Johnson D, Poumerol G, Memish ZA, Barbeschi M, Khan AS:** Global perspectives for prevention of infectious diseases associated with mass gatherings. *The Lancet infectious diseases* 2012, **12**(1):66-74.

24. **Pletz M, Rohde G, Welte T, Kolditz M, Ott S**: Advances in the prevention, management, and treatment of community-acquired pneumonia [version 1; referees: 2 approved], vol. 5; 2016.

25. **Roca A, Bottomley C, Hill PC, Bojang A, Egere U, Antonio M, Darboe O, Greenwood BM, Adegbola RA:** Effect of age and vaccination with a pneumococcal conjugate vaccine on the density of pneumococcal nasopharyngeal carriage. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2012, **55**(6):816-824.

26. **Loo JD, Conklin L, Fleming-Dutra KE, Knoll MD, Park DE, Kirk J, Goldblatt D, O'Brien KL, Whitney CG:** Systematic review of the indirect effect of pneumococcal conjugate vaccine dosing schedules on pneumococcal disease and colonization. *The Pediatric infectious disease journal* 2014, **33 Suppl 2**:S161-171.

27. WHO: **Principles of evaluating health riks in children associated with exposure to chemicals** In. Edited by Environment WHODoPotH, Chemicals I-OPftSMo. Geneva: World Health Organisation; 2006.

28. **Pletz MW, Rohde GG, Welte T, Kolditz M, Ott S:** Advances in the prevention, management, and treatment of community-acquired pneumonia. *F1000Research* 2016, **5**.

29. **Giersing BK, Karron RA, Vekemans J, Kaslow DC, Moorthy VS:** Meeting report: WHO consultation on Respiratory Syncytial Virus (RSV) vaccine development, Geneva, 25-26 April 2016. *Vaccine* 2017.

30. **WHO:**Fact sheet Series 331: Pneumonia [http://www.who.int/mediacentre/factsheets/fs331/en/]

31. **Iwasaki A, Foxman EF, Molony RD:** Early local immune defences in the respiratory tract. *Nature reviews Immunology* 2017, **17**(1):7-20.

32. **Scott JA, Brooks WA, Peiris JS, Holtzman D, Mulholland EK:** Pneumonia research to reduce childhood mortality in the developing world. *The Journal of clinical investigation* 2008, **118**(4):1291-1300.

33. **Gamache J, Harrington A:**Bacterial Pneumonia: Practice Essentials, Background, Pathophysiology [https://emedicine.medscape.com/article/300157-overview#a2]

34. **Hindmarsh PC, Matthews DR, Brain C, Pringle PJ, Brook CG:** The application of deconvolution analysis to elucidate the pulsatile nature of

growth hormone secretion using a variable half-life of growth hormone. *Clinical endocrinology* 1990, **32**(6):739-747.

35. **Bogaert D, de Groot R, Hermans PWM:** Streptococcus pneumoniae colonisation: the key to pneumococcal disease. *The Lancet infectious diseases* 2004, **4**(3):144-154.

36. **Henriques-Normark B, Tuomanen EI:** The pneumococcus: epidemiology, microbiology, and pathogenesis. *Cold Spring Harbor perspectives in medicine* 2013, **3**(7).

37. **Dockrell DH, Whyte MKB, Mitchell TJ:** Pneumococcal pneumonia: mechanisms of infection and resolution. *Chest* 2012, **142**(2):482-491.

38. **Prina E, Ranzani OT, Torres A:** Community-acquired pneumonia. *Lancet* 2015, **386**(9998):1097-1108.

39. **Craig A, Mai J, Cai S, Jeyaseelan S:** Neutrophil recruitment to the lungs during bacterial pneumonia. *Infection and immunity* 2009, **77**(2):568-575.

40. **Rambaud-Althaus C, Althaus F, Genton B, D'Acremont V:** Clinical features for diagnosis of pneumonia in children younger than 5 years: a systematic review and meta-analysis. *The Lancet infectious diseases* 2015, **15**(4):439-450.

41. **Howie SR, Morris GA, Tokarz R *et al*:** Etiology of severe childhood pneumonia in the Gambia, West Africa, determined by conventional and molecular microbiological analyses of lung and pleural aspirate samples. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2014, **59**(5):682-685.

42. **Levine OS, O'Brien KL, Deloria-Knoll M *et al*:** The Pneumonia Etiology Research for Child Health Project: a 21st century childhood pneumonia etiology study. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2012, **54 Suppl 2**:S93-101.

43. **Gilani Z, Kwong YD, Levine OS, Deloria-Knoll M, Scott JA, O'Brien KL, Feikin DR:** A literature review and survey of childhood pneumonia etiology studies: 2000-2010. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2012, **54 Suppl 2**:S102-108.

44. **McCullers JA:** The co-pathogenesis of influenza viruses with bacteria in the lung. *Nature reviews Microbiology* 2014, **12**(4):252-262.

45. **Hament JM, Kimpen JL, Fleer A, Wolfs TF:** Respiratory viral infection predisposing for bacterial disease: a concise review. *FEMS immunology and medical microbiology* 1999, **26**(3-4):189-195.

46. **Jakab GJ:** Mechanisms of virus-induced bacterial superinfections of the lung. *Clinics in chest medicine* 1981, **2**(1):59-66.

47. **Rynda-Apple A, Robinson KM, Alcorn JF:** Influenza and Bacterial Superinfection: Illuminating the Immunologic Mechanisms of Disease. *Infection and immunity* 2015, **83**(10):3764-3770.

48. **Santajit S, Indrawattana N:** Mechanisms of Antimicrobial Resistance in ESKAPE Pathogens. *BioMed research international* 2016, **2016**:2475067.

49. **Akter S, Shamsuzzaman SM, Jahan F:** Community acquired bacterial pneumonia: aetiology, laboratory detection and antibiotic susceptibility pattern. *The Malaysian journal of pathology* 2014, **36**(2):97-103.

50. **Caglayan Serin D, Pullukcu H, Cicek C, Sipahi OR, Tasbakan S, Atalay S, Pneumonia Study G:** Bacterial and viral etiology in hospitalized community acquired pneumonia with molecular methods and clinical evaluation. *Journal of infection in developing countries* 2014, **8**(4):510-518.

51. **Rock MJ:** The diagnostic utility of bronchoalveolar lavage in immunocompetent children with unexplained infiltrates on chest radiograph. *Pediatrics* 1995, **95**(3):373-377.

52. **Radha S, Afroz T, Prasad S, Ravindra N:** Diagnostic utility of bronchoalveolar lavage. *Journal of cytology* 2014, **31**(3):136-138.

53. **Wang K, Bhandari V, Chepustanova S, Huber G, O'Hara S, O'Hern CS, Shattuck MD, Kirby M:** Which Biomarkers Reveal Neonatal Sepsis? *Plos One* 2013, **8**(12).

54. **Deloria-Knoll M, Feikin DR, Scott JA, O'Brien KL, DeLuca AN, Driscoll AJ, Levine OS, Pneumonia Methods Working G:** Identification and selection of cases and controls in the Pneumonia Etiology Research for Child Health project. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2012, **54 Suppl 2**:S117-123.

55. **Iroh Tam PY, Bernstein E, Ma X, Ferrieri P:** Blood Culture in Evaluation of Pediatric Community-Acquired Pneumonia: A Systematic Review and Meta-analysis. *Hospital pediatrics* 2015, **5**(6):324-336.

56. **Huang H, Ideh RC, Gitau E *et al*:** Discovery and validation of biomarkers to guide clinical management of pneumonia in African children. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2014, **58**(12):1707-1715.

57. **Hakansson L:** Development of Biomarkers: What Are the Scientific Hurdles? *Ejc Suppl* 2007, **5**(9):10-11.

58. **Mayeux R:** Biomarkers: potential uses and limitations. *NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics* 2004, **1**(2):182-188.

59. **Strimbu K, Tavel JA:** What are biomarkers? *Current opinion in HIV and AIDS* 2010, **5**(6):463-466.

60. **Summah H, Qu JM:** Biomarkers: a definite plus in pneumonia. *Mediators of inflammation* 2009, **2009**:675753.

61. **Blasi F, Stolz D, Piffer F:** Biomarkers in lower respiratory tract infections. *Pulmonary pharmacology & therapeutics* 2010, **23**(6):501-507.

62. **Christ-Crain M, Opal SM:** Clinical review: the role of biomarkers in the diagnosis and management of community-acquired pneumonia. *Critical care* 2010, **14**(1):203.

63. **Scott JA, Wonodi C, Moisi JC *et al*:** The definition of pneumonia, the assessment of severity, and clinical standardization in the Pneumonia Etiology Research for Child Health study. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2012, **54 Suppl 2**:S109-116.

64. **King C, McCollum ED, Mankhambo L *et al*:** Can We Predict Oral Antibiotic Treatment Failure in Children with Fast-Breathing Pneumonia Managed at the Community Level? A Prospective Cohort Study in Malawi. *PLoS One* 2015, **10**(8):e0136839.

65. **Ackermann M, Strimmer K:** A general modular framework for gene set enrichment analysis. *BMC bioinformatics* 2009, **10**:47.

66. **Scicluna BP, Klein Klouwenberg PM, van Vught LA *et al*:** A molecular biomarker to diagnose community-acquired pneumonia on intensive care unit admission. *American journal of respiratory and critical care medicine* 2015, **192**(7):826-835.

67. **Nakaya HI, Wrammert J, Lee EK *et al*:** Systems biology of vaccination for seasonal influenza in humans. *Nature immunology* 2011, **12**(8):786-795.

68. **Mehta S, Shelling A, Muthukaruppan A, Lasham A, Blenkiron C, Laking G, Print C:** Predictive and prognostic molecular markers for cancer medicine. *Therapeutic advances in medical oncology* 2010, **2**(2):125-148.

69. **Nalejska E, Maczynska E, Lewandowska MA:** Prognostic and predictive biomarkers: tools in personalized oncology. *Molecular diagnosis & therapy* 2014, **18**(3):273-284.

70. **Khoury MJ, Gwinn M, Yoon PW, Dowling N, Moore CA, Bradley L:** The continuum of translation research in genomic medicine: how can we accelerate the appropriate integration of human genome discoveries into health care and disease prevention? *Genet Med* 2007, **9**(10):665-674.

71. **Chaussabel D, Pascual V, Banchereau J:** Assessing the human immune system through blood transcriptomics. *Bmc Biol* 2010, **8**.

72. **Repsilber D, Kern S, Telaar A, Walzl G, Black GF, Selbig J, Parida SK, Kaufmann SH, Jacobsen M:** Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach. *BMC bioinformatics* 2010, **11**:27.

73. **Zhao Y, Simon R:** Gene expression deconvolution in clinical samples. *Genome medicine* 2010, **2**(12):93.

74. **Grant GB, Campbell H, Dowell SF, Graham SM, Klugman KP, Mulholland EK, Steinhoff M, Weber MW, Qazi S:** Recommendations for treatment of childhood non-severe pneumonia. *The Lancet infectious diseases* 2009, **9**(3):185-196.

75. **Makoka MH, Miller WC, Hoffman IF *et al*:** Bacterial infections in Lilongwe, Malawi: aetiology and antibiotic resistance. *BMC infectious diseases* 2012, **12**:67.

76. **Laxminarayan R, Duse A, Wattal C *et al*:** Antibiotic resistance—the need for global solutions. *The Lancet infectious diseases* 2013, **13**(12):1057-1098.

77. **Ashu EE, Jarju S, Dione M, Mackenzie G, Ikumapayi UN, Manjang A, Azuine R, Antonio M:** Population structure, epidemiology and antibiotic resistance patterns of Streptococcus pneumoniae serotype 5: prior to PCV-13 vaccine introduction in Eastern Gambia. *BMC infectious diseases* 2016, **16**(1):33.

78. **Eddens T, Kolls JK:** Host defenses against bacterial lower respiratory tract infection. *Current opinion in immunology* 2012, **24**(4):424-430.

79. **Abbas AK, Lichtman AH**: Basic immunology : functions and disorders of the immune system, 3rd edn. Philadelphia, PA: Saunders/Elsevier; 2009.

80. **Paats MS, Bergen, I. M., Hanselaar WE, van Zoelen EC, Hoogsteden HC, Hendriks RW, van der Eerden MM:** Local and systemic cytokine profiles in non-severe and severe community-acquired pneumonia. *ERJ Express* 2012.

81. **Paats MS, Bergen IM, Hanselaar WE, van Zoelen EC, Verbrugh HA, Hoogsteden HC, van den Blink B, Hendriks RW, van der Eerden MM:** T helper 17 cells are involved in the local and systemic inflammatory response in community-acquired pneumonia. *Thorax* 2013, **68**(5):468-474.

82. **Tak PP, Firestein GS:** NF-kappaB: a key role in inflammatory diseases. *The Journal of clinical investigation* 2001, **107**(1):7-11.

83. **Lawrence T:** The nuclear factor NF-kappaB pathway in inflammation. *Cold Spring Harbor perspectives in biology* 2009, **1**(6):a001651.

84. **Sheller JR, Polosukhin VV, Mitchell D, Cheng DS, Peebles RS, Blackwell TS:** Nuclear factor kappa B induction in airway epithelium

increases lung inflammation in allergen-challenged mice. *Experimental lung research* 2009, **35**(10):883-895.

85. **Kawai T, Akira S:** Toll-like Receptors and Their Crosstalk with Other Innate Receptors in Infection and Immunity. *Immunity* 2011, **34**(5):637-650.

86. **Janeway C, Travers P, Walport M, Shlomchik M**: Immunobiology: The Immune System in Health and Disease., 5th edition. edn. New York:: Garland Science.; 2001.

87. **Franchi L, Warner N, Viani K, Nunez G:** Function of Nod-like receptors in microbial recognition and host defense. *Immunological reviews* 2009, **227**(1):106-128.

88. **Abbas A, Lichtman A, S. P**: Cellular and Molecular Immunology, 3rd edition. edn. Philadelphia:: Elsevier Saunders; 2011.

89. **Kadioglu A, Andrew PW:** The innate immune response to pneumococcal lung infection: the untold story. *Trends Immunol* 2004, **25**(3):143-149.

90. **Mahla RS, Reddy MC, Prasad DV, Kumar H:** Sweeten PAMPs: Role of Sugar Complexed PAMPs in Innate Immunity and Vaccine Biology. *Frontiers in immunology* 2013, **4**:248.

91. **Yoneyama M, Fujita T:** Function of RIG-I-like receptors in antiviral innate immunity. *The Journal of biological chemistry* 2007, **282**(21):15315-15318.

92. **Loo YM, Gale M, Jr.:** Immune signaling by RIG-I-like receptors. *Immunity* 2011, **34**(5):680-692.

93. **Uzri D, Greenberg HB:** Characterization of rotavirus RNAs that activate innate immune signaling through the RIG-I-like receptors. *PLoS One* 2013, **8**(7):e69825.

94. **Wakefield D, Gray P, Chang J, Di Girolamo N, McCluskey P:** The role of PAMPs and DAMPs in the pathogenesis of acute and recurrent anterior uveitis. *The British journal of ophthalmology* 2010, **94**(3):271-274.

95. **Varki A:** Since there are PAMPs and DAMPs, there must be SAMPs? Glycan "self-associated molecular patterns" dampen innate immunity, but pathogens can mimic them. *Glycobiology* 2011, **21**(9):1121-1124.

96. **Zhydkov A, Christ-Crain M, Thomann R, Hoess C, Henzen C, Werner Z, Mueller B, Schuetz P, Pro HSG:** Utility of procalcitonin, C-reactive protein and white blood cells alone and in combination for the prediction of clinical outcomes in community-acquired pneumonia. *Clinical chemistry and laboratory medicine* 2015, **53**(4):559-566.

97. **Ganz T:** Defensins: Antimicrobial peptides of innate immunity. *Nature Reviews Immunology* 2003, **3**(9):710-720.

98. **Oppenheim JJ, Biragyn A, Kwak LW, Yang D:** Roles of antimicrobial peptides such as defensins in innate and adaptive immunity. *Annals of the rheumatic diseases* 2003, **62 Suppl 2**:ii17-21.

99. **Paterson GK, Mitchell TJ:** Innate immunity and the pneumococcus. *Microbiol-Sgm* 2006, **152**:285-293.

100. **Kerr AR, Paterson GK, Riboldi-Tunnicliffe A, Mitchell TJ:** Innate immune defense against pneumococcal pneumonia requires pulmonary complement component C3. *Infection and immunity* 2005, **73**(7):4245-4252.

101. **Riley LK, Rupert J:** Evaluation of Patients with Leukocytosis. *American family physician* 2015, **92**(11):1004-1011.

102. **Palmer C, Diehn M, Alizadeh AA, Brown PO:** Cell-type specific gene expression profiles of leukocytes in human peripheral blood. *BMC genomics* 2006, **7**:115.

103. **Reichert T, DeBruyere M, Deneys V *et al*:** Lymphocyte subset reference ranges in adult Caucasians. *Clinical immunology and immunopathology* 1991, **60**(2):190-208.

104. **Roman S, Moldovan I, Calugaru A, Regalia T, Sulica A:** Lymphocyte subset reference ranges in Romanian adult Caucasians. *Romanian journal of internal medicine = Revue roumaine de medecine interne* 1995, **33**(1-2):27-36.

105. **Shahabuddin S:** Quantitative differences in CD8+ lymphocytes, CD4/CD8 ratio, NK cells, and HLA-DR(+)-activated T cells of racially different male populations. *Clinical immunology and immunopathology* 1995, **75**(2):168-170.

106. **Solomon A, Weiss DT:** Structural and functional properties of human lambda-light-chain variable-region subgroups. *Clinical and diagnostic laboratory immunology* 1995, **2**(4):387-394.

107. **Abbas AK, Lichtman AH**: Basic immunology : functions and disorders of the immune system, 2nd edn. Philadelphia, PA: Elsevier Saunders; 2006.

108. **Abbas AK, Lichtman AH**: Basic Immunology., 8th edn. Philadelphia:: Elsevier Saunders; 2015.

109. **Zhu J, Yamane H, Paul WE:** Differentiation of effector CD4 T cell populations (*). *Annual review of immunology* 2010, **28**:445-489.

110. **Okoye IS, Wilson MS:** CD4+ T helper 2 cells--microbial triggers, differentiation requirements and effector functions. *Immunology* 2011, **134**(4):368-377.

111. **Luckheeram RV, Zhou R, Verma AD, Xia B:** CD4(+)T cells: differentiation and functions. *Clinical & developmental immunology* 2012, **2012**:925135.

112. **Chen K, Kolls JK:** T cell-mediated host immune defenses in the lung. *Annual review of immunology* 2013, **31**:605-633.

113. **Martinez NE, Karlsson F, Sato F, Kawai E, Omura S, Minagar A, Grisham MB, Tsunoda I:** Protective and detrimental roles for regulatory T cells in a viral model for multiple sclerosis. *Brain pathology* 2014, **24**(5):436-451.

114. **Simonetta F, Bourgeois C:** CD4+FOXP3+ Regulatory T-Cell Subsets in Human Immunodeficiency Virus Infection. *Frontiers in immunology* 2013, **4**:215.

115. **Fernandez-Serrano S, Dorca J, Coromines M, Carratala J, Gudiol F, Manresa F:** Molecular inflammatory responses measured in blood of patients with severe community-acquired pneumonia. *Clinical and diagnostic laboratory immunology* 2003, **10**(5):813-820.

116. **Bordon J, Aliberti S, Fernandez-Botran R *et al*:** Understanding the roles of cytokines and neutrophil activity and neutrophil apoptosis in the protective versus deleterious inflammatory response in pneumonia. *International journal of infectious diseases : IJID : official publication of the International Society for Infectious Diseases* 2013, **17**(2):e76-83.

117. **Pennings JL, Schuurhof A, Hodemaekers HM *et al*:** Systemic signature of the lung response to respiratory syncytial virus infection. *PLoS One* 2011, **6**(6):e21461.

118. **Deng JC, Standiford TJ:** The systemic response to lung infection. *Clinics in chest medicine* 2005, **26**(1):1-9.

119. **Multz AS, Cohen R:** Systemic response to pneumonia in the critically ill patient. *Seminars in respiratory infections* 2003, **18**(2):68-71.

120. **Kellum JA, Kong L, Fink MP *et al*:** Understanding the inflammatory cytokine response in pneumonia and sepsis: results of the Genetic and Inflammatory Markers of Sepsis (GenIMS) Study. *Archives of internal medicine* 2007, **167**(15):1655-1663.

121. **Mejias A, Ramilo O:** Transcriptional profiling in infectious diseases: ready for prime time? *The Journal of infection* 2014, **68 Suppl 1**:S94-99.

122. **Wilson R, Cohen JM, Jose RJ, de Vogel C, Baxendale H, Brown JS:** Protection against Streptococcus pneumoniae lung infection after

nasopharyngeal colonization requires both humoral and cellular immune responses. *Mucosal immunology* 2015, **8**(3):627-639.

123. **Chaussabel D:** Assessment of immune status using blood transcriptomics and potential implications for global health. *Seminars in immunology* 2015, **27**(1):58-66.

124. **Apweiler R, Bairoch A, Wu CH *et al*:** UniProt: the universal protein knowledgebase. *Nucleic acids research* 2004, **32**(suppl_1):D115-D119.

125. **Amaratunga D, Cabrera J**: Exploration and Analsysis of DNA Microarray and Protein Array Data. New Jersey: John Wiley & Sons; 2004.

126. **Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, Ferrari R:** Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in bioinformatics* 2016, **19**(2):286-302.

127. **Gräslund S, Nordlund P, Weigelt J *et al*:** Protein production and purification. *Nature methods* 2008, **5**(2):135.

128. **Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T:** Transcriptomics technologies. *PLoS computational biology* 2017, **13**(5):e1005457.

129. **Jaksik R, Iwanaszko M, Rzeszowska-Wolny J, Kimmel M:** Microarray experiments and factors which affect their reliability. *Biology direct* 2015, **10**:46.

130. **Bumgarner R:** Overview of DNA microarrays: types, applications, and their future. *Current protocols in molecular biology / edited by Frederick M Ausubel [et al]* 2013, **Chapter 22**:Unit 22 21.

131. **Wang Z, Gerstein M, Snyder M:** RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews Genetics* 2009, **10**(1):57-63.

132. **Golub TR, Slonim DK, Tamayo P *et al*:** Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999, **286**(5439):531-537.

133. **Tibshirani R, Hastie T, Narasimhan B, Chu G:** Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(10):6567-6572.

134. **Mans JJ, Lamont RJ, Handfield M:** Microarray analysis of human epithelial cell responses to bacterial interaction. *Infectious disorders drug targets* 2006, **6**(3):299-309.

135. **Mejias A, Suarez NM, Ramilo O:** Detecting specific infections in children through host responses: a paradigm shift. *Current opinion in infectious diseases* 2014, **27**(3):228-235.

136. **Smith CL, Dickinson P, Forster T *et al*:** Identification of a human neonatal immune-metabolic network associated with bacterial infection. *Nature communications* 2014, **5**:4649.

137. **Abbas AR, Wolslegel K, Seshasayee D, Modrusan Z, Clark HF:** Deconvolution of blood microarray data identifies cellular activation patterns in systemic lupus erythematosus. *PLoS One* 2009, **4**(7):e6098.

138. **Grigoryev YA, Kurian SM, Avnur Z *et al*:** Deconvoluting post-transplant immunity: cell subset-specific mapping reveals pathways for activation and expansion of memory T, monocytes and B cells. *PLoS One* 2010, **5**(10):e13358.

139. **Shannon CP, Balshaw R, Ng RT *et al*:** Two-stage, in silico deconvolution of the lymphocyte compartment of the peripheral whole blood transcriptome in the context of acute kidney allograft rejection. *PLoS One* 2014, **9**(4):e95224.

140. **Parnell GP, McLean AS, Booth DR *et al*:** A distinct influenza infection signature in the blood transcriptome of patients with severe community-acquired pneumonia. *Critical care* 2012, **16**(4):R157.

141. **Damron FH, Oglesby-Sherrouse AG, Wilks A, Barbier M:** Dual-seq transcriptomics reveals the battle for iron during Pseudomonas aeruginosa acute murine pneumonia. *Scientific reports* 2016, **6**:39172.

142. **Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X:** Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One* 2014, **9**(1):e78644.

143. **Zhang Y, Szustakowski J, Schinke M:** Bioinformatics analysis of microarray data. *Methods in molecular biology* 2009, **573**:259-284.

144. **Tarca AL, Romero R, Draghici S:** Analysis of microarray experiments of gene expression profiling. *American journal of obstetrics and gynecology* 2006, **195**(2):373-388.

145. **Reimers M:** Making informed choices about microarray data analysis. *PLoS computational biology* 2010, **6**(5):e1000786.

146. **Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M:** Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 2002, **18 Suppl 1**:S96-104.

147. **Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP:** Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 2003, **4**(2):249-264.

148. **Johnson WE, Li C, Rabinovic A:** Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007, **8**(1):118-127.

149. **Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA:** Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature reviews Genetics* 2010, **11**(10):733-739.

150. **Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C:** Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* 2011, **6**(2):e17238.

151. **Kauffmann A, Gentleman R, Huber W:** arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. *Bioinformatics* 2009, **25**(3):415-416.

152. **Shieh AD, Hung YS:** Detecting outlier samples in microarray data. *Statistical applications in genetics and molecular biology* 2009, **8**:Article 13.

153. Gentleman R, Carey VJ, Huber W, Hahne F: **genefilter: genefilter: methods for filtering genes from high-throughput experiments. R package version 1.52.1**. In.

154. **Mejias A, Suarez NM, Ramilo O:** Detecting specific infections in children through host responses: a paradigm shift. *Current opinion in infectious diseases* 2014, **27**(3):228-235.

155. **Christensen KD, Dukhovny D, Siebert U, Green RC:** Assessing the Costs and Cost-Effectiveness of Genomic Sequencing. *Journal of personalized medicine* 2015, **5**(4):470-486.

156. **Barrett T, Wilhite SE, Ledoux P *et al*:** NCBI GEO: archive for functional genomics data sets--update. *Nucleic acids research* 2013, **41**(Database issue):D991-995.

157. **Subramanian A, Kuehn H, Gould J, Tamayo P, Mesirov JP:** GSEA-P: a desktop application for Gene Set Enrichment Analysis. *Bioinformatics* 2007, **23**(23):3251-3253.

158. R Development Core Team: **R: A language and environment for statistical computing. Version 3.2** In., Vienna, Austria.: R Foundation for Statistical Computing 2014.

References

159. **Gentleman RC, Carey VJ, Bates DM *et al*:** Bioconductor: open software development for computational biology and bioinformatics. *Genome biology* 2004, **5**(10):R80.

160. **Gaujoux R, Seoighe C:** CellMix: a comprehensive toolbox for gene expression deconvolution. *Bioinformatics* 2013, **29**(17):2211-2212.

161. Renaud G, Seoighe C: **A Comprehensive Toolbox for Gene Expression Deconvolution CRAN.** . In., 1.6.2 edn; 2013: R package.

162. StataCorp: **Stata Statistical Software: Release 12. College Station, TX: StataCorp LP.** In.; 2011.

163. **Alkema L, Chao F, You D, Pedersen J, Sawyer CC:** National, regional, and global sex ratios of infant, child, and under-5 mortality and identification of countries with outlying ratios: a systematic assessment. *The Lancet Global health* 2014, **2**(9):e521-e530.

164. **Lockhart DJ, Dong H, Byrne MC *et al*:** Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology* 1996, **14**(13):1675-1680.

165. **Dalma-Weiszhausz DD, Warrington J, Tanimoto EY, Miyada CG:** The affymetrix GeneChip platform: an overview. *Methods in enzymology* 2006, **410**:3-28.

166. AFFYMETRIX: **Affymetrix Human Genome U219 Array Strip User Guide**. In.: Affymetrix.; 2010.

167. **Smith CL, Dickinson P, Forster T *et al*:** Quantitative assessment of human whole blood RNA as a potential biomarker for infectious disease. *The Analyst* 2007, **132**(12):1200-1209.

168. **Natividad A, Freeman TC, Jeffries D, Burton MJ, Mabey DC, Bailey RL, Holland MJ:** Human conjunctival transcriptome analysis reveals the prominence of innate defense in Chlamydia trachomatis infection. *Infection and immunity* 2010, **78**(11):4895-4911.

169. **Warnes GR, Liu P, Li F:** R package ssize: Estimate Microarray Sample Size. 2012.

170. **Aickin M, Gensler H:** Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *Am J Public Health* 1996, **86**(5):726-728.

171. **Demmer RT, Pavlidis P, Papapanou PN:** Bioinformatics techniques in microarray research: applied microarray data analysis using R and SAS software. *Methods in molecular biology* 2010, **666**:395-417.

172. **Wu Z, Irizarry RA:** Preprocessing of oligonucleotide array data. *Nature biotechnology* 2004, **22**(6):656-658; author reply 658.

173. **Ritchie ME, Silver J, Oshlack A, Holmes M, Diyagama D, Holloway A, Smyth GK:** A comparison of background correction methods for two-colour microarrays. *Bioinformatics* 2007, **23**(20):2700-2707.

174. **Hoffmann R, Seidl T, Dugas M:** Profound effect of normalization on detection of differentially expressed genes in oligonucleotide microarray data analysis. *Genome biology* 2002, **3**(7):RESEARCH0033.

175. **Kroll TC, Wolfl S:** Ranking: a closer look on globalisation methods for normalisation of gene expression arrays. *Nucleic acids research* 2002, **30**(11):e50.

176. **Smyth GK, Speed T:** Normalization of cDNA microarray data. *Methods* 2003, **31**(4):265-273.

177. **Wu W, Xing EP, Myers C, Mian IS, Bissell MJ:** Evaluation of normalization methods for cDNA microarray data by k-NN classification. *BMC bioinformatics* 2005, **6**:191.

178. **Steinhoff C, Vingron M:** Normalization and quantification of differential expression in gene expression microarrays. *Briefings in bioinformatics* 2006, **7**(2):166-177.

179. **Schmidt MT, Handschuh L, Zyprych J, Szabelska A, Olejnik-Schmidt AK, Siatkowski I, Figlerowicz M:** Impact of DNA microarray data transformation on gene expression analysis - comparison of two normalization methods. *Acta biochimica Polonica* 2011, **58**(4):573-580.

180. **Venet D, Detours V, Bersini H:** A measure of the signal-to-noise ratio of microarray samples and studies using gene correlations. *PLoS One* 2012, **7**(12):e51013.

181. **Irizarry RA, Wu Z, Jaffee HA:** Comparison of Affymetrix GeneChip expression measures. *Bioinformatics* 2006, **22**(7):789-794.

182. **Quackenbush J:** Microarray data normalization and transformation. *Nature genetics* 2002, **32 Suppl**:496-501.

183. **Li C, Wong WH:** Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(1):31-36.

184. **Z. W, R. I, R. G, M. MF, A. SF:** Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association* 2004, **99**::909–917. do.

References

185. **Gautier L, Cope L, Bolstad BM, Irizarry RA:** affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 2004, **20**(3):307-315.

186. **Bolstad BM, Irizarry RA, Astrand M, Speed TP:** A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 2003, **19**(2):185-193.

187. **Klawonn F, Jayaram B, Crull K, Kukita A, Pessler F:** Analysis of contingency tables based on generalised median polish with power transformations and non-additive models. *Health information science and systems* 2013, **1**:11.

188. Steven JR: **Introduction to Preprocessing: RMA (Robust Multi-Array Average)**. In.: Utah State University; 2014.

189. Bolstad BM: **Probe Level Quantile Normalization of High Density Oligonucleotide Array Data**. In.; 2001.

190. Huber W: **Introduction to robust calibration and variance stabilisation with VSN**. In.; 2014.

191. **Huber W, von Heydebreck A, Sueltmann H, Poustka A, Vingron M:** Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical applications in genetics and molecular biology* 2003, **2**:Article3.

192. **Durbin BP, Rocke DM:** Variance-stabilizing transformations for two-color microarrays. *Bioinformatics* 2004, **20**(5):660-667.

193. **Ploner A, Miller LD, Hall P, Bergh J, Pawitan Y:** Correlation test to assess low-level processing of high-density oligonucleotide microarray data. *BMC bioinformatics* 2005, **6**:80.

194. R Development Core Team: **R: A language and environment for statistical computing.** . In., 3.2 edn., Vienna, Austria.: R Foundation for Statistical Computing 2014.

195. **Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD:** The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 2012, **28**(6):882-883.

196. **Johnson WE, Li C, Rabinovic A:** Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007, **8**(1):118-127.

197. **Piprek RP:** Genetic mechanisms underlying male sex determination in mammals. *Journal of applied genetics* 2009, **50**(4):347-360.

References

198. **Schneider S, Smith T, Hansen U:** SCOREM: statistical consolidation of redundant expression measures. *Nucleic acids research* 2012, **40**(6):e46.

199. **Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, Horvath S:** Strategies for aggregating gene expression data: the collapseRows R function. *BMC bioinformatics* 2011, **12**:322.

200. **Shen-Orr SS, Gaujoux R:** Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current opinion in immunology* 2013, **25**(5):571-578.

201. **Gaujoux R, Seoighe C:** Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. *Infection, genetics and evolution : journal of molecular epidemiology and evolutionary genetics in infectious diseases* 2012, **12**(5):913-921.

202. **Eng J:** Receiver operating characteristic analysis: utility, reality, covariates, and the future. *Academic radiology* 2013, **20**(7):795-797.

203. **Hajian-Tilaki K:** Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian journal of internal medicine* 2013, **4**(2):627-635.

204. **Abbas AR, Baldwin D, Ma Y *et al*:** Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes and immunity* 2005, **6**(4):319-331.

205. **Su AI, Wiltshire T, Batalov S *et al*:** A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(16):6062-6067.

206. **Allantaz F, Cheng DT, Bergauer T *et al*:** Expression profiling of human immune cell subsets identifies miRNA-mRNA regulatory relationships correlated with cell type specific expression. *PLoS One* 2012, **7**(1):e29979.

207. **Pearson K:** On lines and planes of closest fit to systems of points in space. *Philos Mag* 1901, **2**:559–572.

208. **Hotelling H:** Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 1933, **24**:417–441, 498–520.

209. **Jolliffe IT**: Principal component analysis, 2nd edn edn. New York: Springer,; 2002.

210. **Webb AR**: Statistical Pattern Recognition, 2nd edition edn. Sussex: John Wiley & Sons LTD; 2002.

211. **Jackson JE**: A user's guide to principal components. . New York: Wiley; 1991.

References

212. **van der Maaten L, Hinton G:** Visualizing Data using t-SNE. *J Mach Learn Res* 2008, **9**:2579-2605.

213. Saurabh J: **Comprehensive Guide on t-SNE algorithm with implementation in R & Python**. In: *Analytics Vidhya.* https://http://www.analyticsvidhya.com/blog/2017/01/t-sne-implementation-r-python/; 2017.

214. **Hinton GE, Roweis ST**: Stochastic neighbor embedding. In: *Advances in neural information processing systems: 2003*. 857-864.

215. Kangeyan D: **t-SNE tutorial**. In.; 2017.

216. **Wattenberg M, Viégas F, Johnson I:** How to use t-sne effectively. *Distill* 2016, **1**(10):e2.

217. **Langfelder P, Zhang B, Horvath S:** Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* 2008, **24**(5):719-720.

218. **Yim O, Ramdeen KT:** 2015. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. *Quant Methods Psychol*
2015, **11**:8–21.

219. Lavrenko V: **Hierarchical Clustering 3: single-link vs. complete-link**. In: *Science & Technology.* 2014.

220. **Jain AK:** Data clustering: 50 years beyond K-means. *Pattern Recogn Lett* 2010, **vol. 31**:651-666.

221. **Jaccard P:** Distribution de la florine alpine dans la Bassin de Dranses et dans quelques regiones voisines. . *Bulletin de la Societe Vaudoise des Sciences Naturelles* 1901, **37**:241–272.

222. **Prokopenko D, Hecker J, Silverman EK, Pagano M, Nothen MM, Dina C, Lange C, Fier HL:** Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 Genomes Project. *Bioinformatics* 2016, **32**(9):1366-1372.

223. **Zhang S, Wu X, You Z:** Jaccard distance based weighted sparse representation for coarse-to-fine plant species recognition. *PLoS One* 2017, **12**(6):e0178317.

224. **Mammone N, Ieracitano C, Adeli H, Bramanti A, Morabito FC:** Permutation Jaccard Distance-Based Hierarchical Clustering to Estimate EEG Network Density Modifications in MCI Subjects. *IEEE transactions on neural networks and learning systems* 2018.

225. **Hennig C:** Cluster-wise assessment of cluster stability. *Computational statistics & data analysis* 2007, **52**(1):258-271.

226. **Smyth GK:** Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology* 2004, **3**:Article3.

227. **Efron B, Tibshirani R, Storey JD, Tusher V:** Empirical Bayes analysis of a microarray experiment. . *Journal of the American Statistical Association* 2001, **96**:1151-1160.

228. **Tusher VG, Tibshirani R, Chu G:** Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* 2001, **98**(9):5116-5121.

229. **Efron B, Tibshirani R:** Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology* 2002, **23**(1):70-86.

230. **L□onnstedt I, Speed TP:** Replicated microarray data. *Statistica Sinica* 2002, **12**:31-46.

231. **Smyth GK, Yang YH, Speed T:** Statistical issues in cDNA microarray data analysis. *Methods in molecular biology* 2003, **224**:111-136.

232. **Kim SY, Lee JW, Sohn IS:** Comparison of various statistical methods for identifying differential gene expression in replicated microarray data. *Statistical methods in medical research* 2006, **15**(1):3-20.

233. **Astrand M, Mostad P, Rudemo M:** Empirical Bayes models for multiple probe type microarrays at the probe level. *BMC bioinformatics* 2008, **9**:156.

234. **Parmigiani G, Garrett ES, Irizarry RA, Zeger SL**: The Analysis of Gene Expression Data: Methods and Software. . New York: Springer; 2003.

235. **Kirkwood B, Sterne J:** Medical statistics. 2. *Malden: Blackwell Science* 2003.

236. Smyth GK, Thorne, N. P., and Wettenhall, J. : **Limma: Linear Models for Microarray Data User's Guide.** . In.; 2003.

237. **Benjamini Y, Drai D, Elmer G, Kafkafi N, Golani I:** Controlling the false discovery rate in behavior genetics research. *Behavioural brain research* 2001, **125**(1-2):279-284.

238. **Benjamini Y, Hochberg Y:** Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* 1995, **57**(1):289-300.

239. **Storey JD, Tibshirani R:** Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(16):9440-9445.

240. **Caldas de Castro M, Singer BH:** Controlling the false discovery rate: a new application to account for multiple and dependent tests in local statistics of spatial association. *Geographical Analysis* 2006, **38**(2):180-208.

241. **Hommel G:** A Comparison of Two Modified Bonferroni Procedures. *Biometrika* 1989, **76**(3):624–625.

242. **Abdi H:** Bonferroni and Šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics* 2007, **3**:103-107.

243. **Yekutieli D, Benjamini Y:** Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference* 1999, **82**(1-2):171-196.

244. **Benjamini Y, Yekutieli D:** The control of the false discovery rate in multiple testing under dependency. *Annals of statistics* 2001:1165-1188.

245. **Birkner MD, Pollard KS, van der Laan MJ, Dudoit S:** Multiple testing procedures and applications to genomics. 2005.

246. **Storey JD:** A Direct Approach to False Discovery Rates. *Journal of the Royal Statistical Society B* 2002, **64**(3):479–498.

247. **Storey JD:** 'The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value. *The Annals of Statistics* 2003, **31**(6):2013–2035.

248. **Pollard KS, Dudoit S, van der Laan MJ**: Multiple testing procedures: the multtest package and applications to genomics. In: *Bioinformatics and computational biology solutions using R and bioconductor.* Springer; 2005: 249-271.

249. **Khatri P, Sirota M, Butte AJ:** Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology* 2012, **8**(2):e1002375.

250. **Kanehisa M, Goto S:** KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* 2000, **28**(1):27-30.

251. **Goeman JJ, Buhlmann P:** Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007, **23**(8):980-987.

252. **Huang da W, Sherman BT, Lempicki RA:** Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* 2009, **37**(1):1-13.

253. **Khatri P, Draghici S:** Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 2005, **21**(18):3587-3595.

254. **Subramanian A, Tamayo P, Mootha VK *et al*:** Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(43):15545-15550.

255. **Irizarry RA, Wang C, Zhou Y, Speed TP:** Gene set enrichment analysis made simple. *Statistical methods in medical research* 2009, **18**(6):565-575.

256. **Rahmatallah Y, Emmert-Streib F, Glazko G:** Gene Sets Net Correlations Analysis (GSNCA): a multivariate differential coexpression test for gene sets. *Bioinformatics* 2014, **30**(3):360-368.

257. **Du J, Yuan Z, Ma Z, Song J, Xie X, Chen Y:** KEGG-PATH: Kyoto encyclopedia of genes and genomes-based pathway analysis using a path analysis model. *Molecular bioSystems* 2014, **10**(9):2441-2447.

258. **Asif HM, Rolfe MD, Green J, Lawrence ND, Rattray M, Sanguinetti G:** TFInfer: a tool for probabilistic inference of transcription factor activities. *Bioinformatics* 2010, **26**(20):2635-2636.

259. **Sanguinetti G, Lawrence ND, Rattray M:** Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics* 2006, **22**(22):2775-2781.

260. **Sanguinetti G, Rattray M, Lawrence ND:** A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. *Bioinformatics* 2006, **22**(14):1753-1759.

261. **Ocone A, Sanguinetti G:** Reconstructing transcription factor activities in hierarchical transcription network motifs. *Bioinformatics* 2011, **27**(20):2873-2879.

262. **Geistlinger L, Csaba G, Kuffner R, Mulder N, Zimmer R:** From sets to graphs: towards a realistic enrichment analysis of transcriptomic systems. *Bioinformatics* 2011, **27**(13):i366-373.

263. **Glaab E, Baudot A, Krasnogor N, Valencia A:** TopoGSA: network topological gene set analysis. *Bioinformatics* 2010, **26**(9):1271-1272.

264. **Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R:** A systems biology approach for pathway level analysis. *Genome research* 2007, **17**(10):1537-1545.

265. **Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, Kim CJ, Kusanovic JP, Romero R:** A novel signaling pathway impact analysis. *Bioinformatics* 2009, **25**(1):75-82.

References

266. **Croft D, O'kelly G, Wu G *et al*:** Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research* 2010, **39**(suppl_1):D691-D697.

267. **Ashburner M, Ball CA, Blake JA *et al*:** Gene Ontology: tool for the unification of biology. *Nature genetics* 2000, **25**(1):25.

268. **Peng J, Wang H, Lu J, Hui W, Wang Y, Shang X:** Identifying term relations cross different gene ontology categories. *BMC bioinformatics* 2017, **18**(Suppl 16):573.

269. **Chagoyen M, Pazos F:** Quantifying the biological significance of gene ontology biological processes--implications for the analysis of systems-wide data. *Bioinformatics* 2010, **26**(3):378-384.

270. **Rivals I, Personnaz L, Taing L, Potier MC:** Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* 2007, **23**(4):401-407.

271. Freeman JV, Campbell MJ: **THE ANALYSIS OF CATEGORICAL DATA: FISHER'S EXACT TEST**. In.

272. **Glazko G, Rahmatallah Y, Zybailov B, Emmert-Streib F:** Extracting the Strongest Signals from Omics Data: Differentially Expressed Pathways and Beyond. *Methods in molecular biology* 2017, **1613**:125-159.

273. **Rolfe MD, Ocone A, Stapleton MR, Hall S, Trotter EW, Poole RK, Sanguinetti G, Green J, Sys MOSC:** Systems analysis of transcription factor activities in environments with stable and dynamic oxygen concentrations. *Open biology* 2012, **2**(7):120091.

274. **Haynes WA, Higdon R, Stanberry L, Collins D, Kolker E:** Differential expression analysis for pathways. *PLoS computational biology* 2013, **9**(3):e1002967.

275. **Yang Q, Wang S, Dai E, Zhou S, Liu D, Liu H, Meng Q, Jiang B, Jiang W:** Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. *Briefings in bioinformatics* 2017.

276. **Liu B, de la Fuente A, Hoeschele I:** Gene network inference via structural equation modeling in genetical genomics experiments. *Genetics* 2008, **178**(3):1763-1776.

277. **Wu X, Sharpe K, Zhang T, Chen H, Zhu W, Li E, Taghavi S, Van Der Lelie D**: Comparative genetic pathway analysis using structural equation Modeling. In: *ICCABS: 2011*. 190-195.

278. **Pepe D, Grassi M:** Investigating perturbed pathway modules from gene expression data via structural equation models. *BMC bioinformatics* 2014, **15**:132.

279. **Martinez SA, Beebe LA, Thompson DM, Wagener TL, Terrell DR, Campbell JE:** A structural equation modeling approach to understanding pathways that connect socioeconomic status and smoking. *PLoS One* 2018, **13**(2):e0192451.

280. **Szklarczyk D, Morris JH, Cook H *et al*:** The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research* 2017, **45**(D1):D362-D368.

281. **Luo W, Pant G, Bhavnasi YK, Blanchard SG, Jr., Brouwer C:** Pathview Web: user friendly pathway visualization and data integration. *Nucleic acids research* 2017.

282. **Friedman J, Hastie T, Tibshirani R:** Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* 2010, **33**(1):1-22.

283. **Cortes C, Vapnik V:** Support-Vector Networks. *Mach Learn* 1995, **20**(3):273-297.

284. **Breiman L:** Random forests. *Mach Learn* 2001, **45**(1):5-32.

285. **Altman NS:** An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *American Statistician* 1992, **46**(3):175-185.

286. **Hall P, Hallen BA, Selander H:** Linear Discriminatory Analysis - Patient Classifying Method for Research and Production Control. *Methods of information in medicine* 1971, **10**(2):96-&.

287. **Lauss M, Frigyesi A, Ryden T, Hoglund M:** Robust assignment of cancer subtypes from expression data using a uni-variate gene expression average as classifier. *BMC cancer* 2010, **10**:532.

288. **Helfenstein U, Steiner M:** The use of logistic discrimination and receiver operating characteristics (ROC) analysis in dentistry. *Community dental health* 1994, **11**(3):142-146.

289. **Ng AY, Jordan MI**: On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In: *Advances in neural information processing systems: 2002*. 841-848.

290. **Allison PD**: Logistic regression using SAS: Theory and application: SAS Institute; 2012.

291. **Tripepi G, Jager KJ, Dekker FW, Zoccali C:** Linear and logistic regression analysis. *Kidney international* 2008, **73**(7):806-810.

292. **Menard S**: Applied logistic regression analysis, vol. 106: Sage; 2002.

References

293. **Hosmer DW, Lemeshow S, Sturdivant RX**: Applied logistic regression, vol. 398: John Wiley & Sons; 2013.

294. **Hoerl AE, Kennard RW:** Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences* 1970, **12**(1):55-67.

295. **Le Cessie S, Van Houwelingen JC:** Ridge estimators in logistic regression. *Applied statistics* 1992:191-201.

296. **Marquaridt DW:** Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences* 1970, **12**(3):591-612.

297. **Tibshirani R:** Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B (Methodological)* 1996:267-288.

298. Zou H, Hastie T: **Regularization and Variable Selection via the Elastic Net**
. In.; 2004.

299. **Zou H, Hastie T:** Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2005, **67**(2):301-320.

300. **Zou H, Hastie T:** Regression shrinkage and selection via the elastic net, with applications to microarrays. *JR Stat Soc Ser B* 2003, **67**:301-320.

301. **Ogutu JO, Schulz-Streeck T, Piepho H-P**: Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions. In: *BMC proceedings: 2012*. BioMed Central: S10.

302. Trevor H, Junyang Q: **Glmnet Vignette**. In.; 2014.

303. **Li L, Weinberg CR, Darden TA, Pedersen LG:** Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics* 2001, **17**(12):1131-1142.

304. **Li L, Zhang Y, Zhao Y:** k-Nearest Neighbors for automated classification of celestial objects. *Science in China Series G: Physics, Mechanics and Astronomy* 2008, **51**(7):916-922.

305. **Altman NS:** An introduction to kernel and nearest-neighbor nonparametric regression. *The American statistician* 1992, **46**(3):175-185.

306. **Martínez AM, Kak AC:** Pca versus lda. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 2001(2):228-233.

References

307. **Oza NC, Russell S**: Online ensemble learning: University of California, Berkeley; 2001.

308. **Strobl C, Malley J, Tutz G:** An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological methods* 2009, **14**(4):323.

309. **Breiman L**: Classification and regression trees: Routledge; 2017.

310. **Goldstein BA, Polley EC, Briggs FB:** Random forests for genetic association studies. *Statistical applications in genetics and molecular biology* 2011, **10**(1):32.

311. **Quinlan JR**: Bagging, boosting, and C4. 5. In: *AAAI/IAAI, Vol 1: 1996*. 725-730.

312. **Lemmens A, Croux C:** Bagging and boosting classification trees to predict churn. *Journal of Marketing Research* 2006, **43**(2):276-286.

313. **Maclin R, Opitz D:** An empirical evaluation of bagging and boosting. *AAAI/IAAI* 1997, **1997**:546-551.

314. Boulesteix AL, Janitza S, Kruppa J, König IR: **Overview of Random Forest Methodology and Practical Guidance with Emphasis on Computational Biology and Bioinformatics**. In.: Department of Statistics, University of Munich; 2012.

315. Breiman L: **Manual On Setting Up, Using, And Understanding Random Forests V3.1**. In.; 2002.

316. Liaw A: **randomForest v4.6-14: Classification And Regression With Random Forest**. In.; 2018.

317. **Burges CJ:** A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery* 1998, **2**(2):121-167.

318. **Hsu C-W, Chang C-C, Lin C-J:** A practical guide to support vector classification. 2003.

319. **Statnikov A, Wang L, Aliferis CF:** A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC bioinformatics* 2008, **9**.

320. Sayad S: **Support Vector Machine - Classification (SVM)**. In: *An Introduction to Data Science.* vol. 2018; 2010.

321. **Gordon G, Tibshirani R:** Karush-kuhn-tucker conditions. *Optimization* 2012, **10**(725/36):725.

References

322. Abu-Mostafa YS: **Support Vector Machines**. In: *Lecture 14* caltech; 2012.

323. OpenCourseWare M: **Support Vector Machines**. In: *Learning 16* 2014.

324. Abu-Mostafa YS: **Kernel Methods**. In: *Lecture 15.* caltech; 2012.

325. Pavel L, Blaine N: **Theory of Kernel Functions**. In: *Advanced Topics in Machine Learning.* 2012.

326. **Aronszajn N:** Theory of reproducing kernels. *Transactions of the American mathematical society* 1950, **68**(3):337-404.

327. **Hille E:** Introduction to general theory of reproducing kernels. *The Rocky Mountain Journal of Mathematics* 1972, **2**(3):321-368.

328. Team DF: **Kernel Functions-Introduction to SVM Kernel & Examples** In: *Machine Learning Tutorials.* vol. 2018. https://data-flair.training/blogs/svm-kernel-functions/; 2017.

329. **Karatzoglou A, Meyer D, Hornik K:** Support vector machines in R. 2005.

330. **Kohavi R**: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Ijcai: 1995*. Montreal, Canada: 1137-1145.

331. Folta W: **Machine Learning Tip: Nested Cross Validation – When (Simple) Cross Validation Isn't Enough**. In: *Mach Learn.* vol. 2018. https://http://www.predictiveanalyticsworld.com/patimes/nested-cross-validation-simple-cross-validation-isnt-enough/8952/: Predictive Analytics Times; 2017.

332. Albon C: **Nested Cross Validation**. In: *Mach Learn.* vol. 2018. https://chrisalbon.com/machine_learning/model_evaluation/nested_cross_validation/; 2017.

333. **Cawley GC, Talbot NL:** On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res* 2010, **11**(Jul):2079-2107.

334. **Biomarkers Definitions Working G:** Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical pharmacology and therapeutics* 2001, **69**(3):89-95.

335. SPSS: **Missing data: the hidden problem (white paper)**. In.

336. **Jaksik R, Iwanaszko M, Rzeszowska-Wolny J, Kimmel M:** Microarray experiments and factors which affect their reliability. *Biology direct* 2015, **10**(1):46.

337. **Jones SR, Carley S, Harrison M:** An introduction to power and sample size estimation. *Emergency medicine journal : EMJ* 2003, **20**(5):453-458.

338. **Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A:** False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* 2005, **21**(13):3017-3024.

339. **Harr B, Schlotterer C:** Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons. *Nucleic acids research* 2006, **34**(2):e8.

340. **Walker WL, Liao IH, Gilbert DL, Wong B, Pollard KS, McCulloch CE, Lit L, Sharp FR:** Empirical Bayes accomodation of batch-effects in microarray data using identical replicate reference samples: application to RNA expression profiling of blood from Duchenne muscular dystrophy patients. *BMC genomics* 2008, **9**:494.

341. **Muller C, Schillert A, Rothemeier C *et al*:** Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. *PLoS One* 2016, **11**(6):e0156594.

342. **Rasmussen JL:** Evaluating Outlier Identification Tests: Mahalanobis D Squared and Comrey Dk. *Multivariate behavioral research* 1988, **23**(2):189-202.

343. **Kauffmann A, Huber W:** Microarray data quality control improves the detection of differentially expressed genes. *Genomics* 2010, **95**(3):138-142.

344. **Wang J, Bo TH, Jonassen I, Myklebost O, Hovig E:** Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. *BMC bioinformatics* 2003, **4**:60.

345. **Jirapech-Umpai T, Aitken S:** Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. *BMC bioinformatics* 2005, **6**:148.

346. **Saeys Y, Inza I, Larranaga P:** A review of feature selection techniques in bioinformatics. *Bioinformatics* 2007, **23**(19):2507-2517.

347. **Banchereau R, Hong S, Cantarel B *et al*:** Personalized Immunomonitoring Uncovers Molecular Networks that Stratify Lupus Patients. *Cell* 2016, **165**(3):551-565.

348. **Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D:** Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics* 2005, **21**(20):3896-3904.

349. **Palmu AA, Saukkoriipi A, Snellman M *et al*:** Incidence and etiology of community-acquired pneumonia in the elderly in a prospective population-

based study. *Scandinavian journal of infectious diseases* 2014, **46**(4):250-259.

350. **Mathew JL, Singhi S, Ray P *et al*:** Etiology of community acquired pneumonia among children in India: prospective, cohort study. *Journal of global health* 2015, **5**(2):050418.

351. **Jain AK, Murty MN, Flynn PJ:** Data clustering: A review. *Acm Comput Surv* 1999, **31**(3):264-323.

352. **Yang MG, Xiao Z, Shi Q *et al*:** Synthesis of 3-phenylsulfonylmethyl cyclohexylaminobenzamide-derived antagonists of CC chemokine receptor 2 (CCR2). *Bioorganic & medicinal chemistry letters* 2012, **22**(3):1384-1387.

353. **Vander Weele TJ:** Confounding and effect modification: distribution and measure. *Epidemiologic methods* 2012, **1**(1):55-82.

354. **Lee PH, Burstyn I:** Identification of confounder in epidemiologic data contaminated by measurement error in covariates. *BMC medical research methodology* 2016, **16**:54.

355. **Pocock SJ, Elbourne DR:** Randomized trials or observational tribulations? *The New England journal of medicine* 2000, **342**(25):1907-1909.

356. **Shrier I, Pang M:** Confounding, effect modification, and the odds ratio: common misinterpretations. *Journal of clinical epidemiology* 2015, **68**(4):470-474.

357. **Burl S, Townend J, Njie-Jobe J *et al*:** Age-dependent maturation of Toll-like receptor-mediated cytokine responses in Gambian infants. *PLoS One* 2011, **6**(4):e18185.

358. **Watkins NA, Gusnanto A, de Bono B *et al*:** A HaemAtlas: characterizing gene expression in differentiated human blood cells. *Blood* 2009, **113**(19):e1-9.

359. **Geller SC, Gregg JP, Hagerman P, Rocke DM:** Transformation and normalization of oligonucleotide microarray data. *Bioinformatics* 2003, **19**(14):1817-1823.

360. **Smith CL, Dickinson P, Forster T *et al*:** Identification of a human neonatal immune-metabolic network associated with bacterial infection. *Nature communications* 2014, **5**.

361. **VanderWeele TJ, Shpitser I:** A new criterion for confounder selection. *Biometrics* 2011, **67**(4):1406-1413.

362. **Vrijheid M, Deltour I, Krewski D, Sanchez M, Cardis E:** The effects of recall errors and of selection bias in epidemiologic studies of mobile phone

use and cancer risk. *Journal of exposure science & environmental epidemiology* 2006, **16**(4):371-384.

363. **Coughlin SS:** Recall bias in epidemiologic studies. *Journal of clinical epidemiology* 1990, **43**(1):87-91.

364. **Walter SD:** Recall bias in epidemiologic studies. *Journal of clinical epidemiology* 1990, **43**(12):1431-1432.

365. **Fantin B, Joly V, Elbim C, Golmard JL, Gougerot-Pocidalo MA, Yeni P, Carbon C:** Lymphocyte subset counts during the course of community-acquired pneumonia: evolution according to age, human immunodeficiency virus status, and etiologic microorganisms. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 1996, **22**(6):1096-1098.

366. **de Jager CP, van Wijk PT, Mathoera RB, de Jongh-Leuvenink J, van der Poll T, Wever PC:** Lymphocytopenia and neutrophil-lymphocyte count ratio predict bacteremia better than conventional infection markers in an emergency care unit. *Critical care* 2010, **14**(5):R192.

367. **de Jager CP, Wever PC, Gemen EF, Kusters R, van Gageldonk-Lafeber AB, van der Poll T, Laheij RJ:** The neutrophil-lymphocyte count ratio in patients with community-acquired pneumonia. *PLoS One* 2012, **7**(10):e46561.

368. **Koh YW, Kang HJ, Park C *et al*:** The ratio of the absolute lymphocyte count to the absolute monocyte count is associated with prognosis in Hodgkin's lymphoma: correlation with tumor-associated macrophages. *The oncologist* 2012, **17**(6):871-880.

369. **Terradas R, Grau S, Blanch J, Riu M, Saballs P, Castells X, Horcajada JP, Knobel H:** Eosinophil count and neutrophil-lymphocyte count ratio as prognostic markers in patients with bacteremia: a retrospective cohort study. *PLoS One* 2012, **7**(8):e42860.

370. **Yoon NB, Son C, Um SJ:** Role of the neutrophil-lymphocyte count ratio in the differential diagnosis between pulmonary tuberculosis and bacterial community-acquired pneumonia. *Annals of laboratory medicine* 2013, **33**(2):105-110.

371. **Zhong Y, Wan YW, Pang K, Chow LM, Liu Z:** Digital sorting of complex tissues for cell type-specific gene expression profiles. *BMC bioinformatics* 2013, **14**:89.

372. **Shen-Orr SS, Tibshirani R, Khatri P *et al*:** Cell type-specific gene expression differences in complex tissues. *Nature methods* 2010, **7**(4):287-289.

373. **Yang X, Ye Y, Wang G, Huang H, Yu D, Liang S:** VeryGene: linking tissue-specific genes to diseases, drugs, and beyond for knowledge discovery. *Physiological genomics* 2011, **43**(8):457-460.

374. **Shannon CP, Hollander Z, Wilson-McManus J, Balshaw R, Ng RT, McMaster R, McManus BM, Keown PA, Tebbutt SJ:** White blood cell differentials enrich whole blood expression data in the context of acute cardiac allograft rejection. *Bioinformatics and biology insights* 2012, **6**:49-61.

375. **Becht E, Giraldo NA, Lacroix L *et al*:** Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome biology* 2016, **17**(1):218.

376. **Mandala WL, MacLennan JM, Gondwe EN, Ward SA, Molyneux ME, MacLennan CA:** Lymphocyte subsets in healthy Malawians: implications for immunologic assessment of HIV infection in Africa. *The Journal of allergy and clinical immunology* 2010, **125**(1):203-208.

377. **Jones RO, Brittan M, Anderson NH, Conway Morris A, Murchison JT, Walker WS, Simpson AJ:** Serial characterisation of monocyte and neutrophil function after lung resection. *BMJ open respiratory research* 2014, **1**(1):e000045.

378. **Curbelo J, Luquero Bueno S, Galvan-Roman JM *et al*:** Inflammation biomarkers in blood as mortality predictors in community-acquired pneumonia admitted patients: Importance of comparison with neutrophil count percentage or neutrophil-lymphocyte ratio. *PLoS One* 2017, **12**(3):e0173947.

379. **Souza-Fonseca-Guimaraes F, Adib-Conquy M, Cavaillon JM:** Natural killer (NK) cells in antibacterial innate immunity: angels or devils? *Molecular medicine* 2012, **18**:270-285.

380. **Hall LJ, Murphy CT, Hurley G, Quinlan A, Shanahan F, Nally K, Melgar S:** Natural killer cells protect against mucosal and systemic infection with the enteric pathogen Citrobacter rodentium. *Infection and immunity* 2013, **81**(2):460-469.

381. **Hotchkiss RS, Tinsley KW, Swanson PE, Grayson MH, Osborne DF, Wagner TH, Cobb JP, Coopersmith C, Karl IE:** Depletion of dendritic cells, but not macrophages, in patients with sepsis. *Journal of immunology* 2002, **168**(5):2493-2500.

382. **Rosendahl A, Bergmann S, Hammerschmidt S, Goldmann O, Medina E:** Lung dendritic cells facilitate extrapulmonary bacterial dissemination during pneumococcal pneumonia. *Frontiers in cellular and infection microbiology* 2013, **3**:21.

383. **Vivier E, Tomasello E, Baratin M, Walzer T, Ugolini S:** Functions of natural killer cells. *Nature immunology* 2008, **9**(5):503-510.

384. **Pegram HJ, Andrews DM, Smyth MJ, Darcy PK, Kershaw MH:** Activating and inhibitory receptors of natural killer cells. *Immunology and cell biology* 2011, **89**(2):216-224.

385. **Howard CJ, Charleston B, Stephens SA, Sopp P, Hope JC:** The role of dendritic cells in shaping the immune response. *Animal health research reviews* 2004, **5**(2):191-195.

386. **McDermott DS, Weiss KA, Knudson CJ, Varga SM:** Central role of dendritic cells in shaping the adaptive immune response during respiratory syncytial virus infection. *Future virology* 2011, **6**(8):963-973.

387. **Chikina M, Zaslavsky E, Sealfon SC:** CellCODE: a robust latent variable approach to differential expression analysis for heterogeneous cell populations. *Bioinformatics* 2015, **31**(10):1584-1591.

388. **Kerr AR, Kirkham LA, Kadioglu A, Andrew PW, Garside P, Thompson H, Mitchell TJ:** Identification of a detrimental role for NK cells in pneumococcal pneumonia and sepsis in immunocompromised hosts. *Microbes and infection / Institut Pasteur* 2005, **7**(5-6):845-852.

389. **Lapaque N, Walzer T, Meresse S, Vivier E, Trowsdale J:** Interactions between Human NK Cells and Macrophages in Response to Salmonella Infection. *Journal of immunology* 2009, **182**(7):4339-4348.

390. **Harizi H:** Reciprocal crosstalk between dendritic cells and natural killer cells under the effects of PGE2 in immunity and immunopathology. *Cellular & molecular immunology* 2013, **10**(3):213-221.

391. **Che SL, Huston DP:** Natural-Killer-Cell Suppression of Igm Production. *Natural immunity* 1994, **13**(5):258-269.

392. **Walzer T, Dalod M, Vivier E, Zitvogel L:** Natural killer cell-dendritic cell crosstalk in the initiation of immune responses. *Expert opinion on biological therapy* 2005, **5 Suppl 1**:S49-59.

393. **Niemeyer M, Darmoise A, Mollenkopf HJ, Hahnke K, Hurwitz R, Besra GS, Schaible UE, Kaufmann SH:** Natural killer T-cell characterization through gene expression profiling: an account of versatility bridging T helper type 1 (Th1), Th2 and Th17 immune responses. *Immunology* 2008, **123**(1):45-56.

394. **Small CL, McCormick S, Gill N, Kugathasan K, Santosuosso M, Donaldson N, Heinrichs DE, Ashkar A, Xing Z:** NK cells play a critical protective role in host defense against acute extracellular Staphylococcus aureus bacterial infection in the lung. *Journal of immunology* 2008, **180**(8):5558-5568.

395. **Yoshida O, Akbar F, Miyake T, Abe M, Matsuura B, Hiasa Y, Onji M:** Impaired dendritic cell functions because of depletion of natural killer cells

disrupt antigen-specific immune responses in mice: restoration of adaptive immunity in natural killer-depleted mice by antigen-pulsed dendritic cell. *Clinical and experimental immunology* 2008, **152**(1):174-181.

396. **Crouse J, Xu HC, Lang PA, Oxenius A:** NK cells regulating T cell responses: mechanisms and outcome. *Trends in Immunology* 2015, **36**(1):49-58.

397. **Joyee AG, Qiu H, Fan Y, Wang S, Yang X:** Natural killer T cells are critical for dendritic cells to induce immunity in Chlamydial pneumonia. *American journal of respiratory and critical care medicine* 2008, **178**(7):745-756.

398. **Broquet A, Roquilly A, Jacqueline C, Potel G, Caillon J, Asehnoune K:** Depletion of natural killer cells increases mice susceptibility in a Pseudomonas aeruginosa pneumonia model. *Critical care medicine* 2014, **42**(6):e441-450.

399. **Christaki E, Diza E, Giamarellos-Bourboulis EJ *et al*:** NK and NKT Cell Depletion Alters the Outcome of Experimental Pneumococcal Pneumonia: Relationship with Regulation of Interferon-gamma Production. *Journal of immunology research* 2015, **2015**:532717.

400. **Ebbo M, Gerard L, Carpentier S *et al*:** Low Circulating Natural Killer Cell Counts are Associated With Severe Disease in Patients With Common Variable Immunodeficiency. *EBioMedicine* 2016, **6**:222-230.

401. **Koppe U, Suttorp N, Opitz B:** Recognition of Streptococcus pneumoniae by the innate immune system. *Cellular microbiology* 2012, **14**(4):460-466.

402. **Mootha VK, Lindgren CM, Eriksson KF *et al*:** PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics* 2003, **34**(3):267-273.

403. **Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ:** GAGE: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics* 2009, **10**:161.

404. **Eddy JA, Hood L, Price ND, Geman D:** Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC). *PLoS computational biology* 2010, **6**(5):e1000792.

405. **Kim S, Kon M, DeLisi C:** Pathway-based classification of cancer subtypes. *Biology direct* 2012, **7**:21.

406. **Li B, Dong C, Wang G, Zheng H, Wang X, Bai C:** Pulmonary epithelial CCR3 promotes LPS-induced lung inflammation by mediating release of IL-8. *Journal of cellular physiology* 2011, **226**(9):2398-2405.

407. **Gerber BO, Zanni MP, Uguccioni M, Loetscher M, Mackay CR, Pichler WJ, Yawalkar N, Baggiolini M, Moser B:** Functional expression of the

eotaxin receptor CCR3 in T lymphocytes co-localizing with eosinophils. *Current biology : CB* 1997, **7**(11):836-843.

408. **Su X, Lin Z, Chen W, Jiang H, Zhang S, Lin H:** Chemogenomic approach identified yeast YLR143W as diphthamide synthetase. *Proceedings of the National Academy of Sciences* 2012, **109**(49):19983-19987.

409. **Lin YW, Deveney R, Barbara M, Iscove NN, Nimer SD, Slape C, Aplan PD:** OLIG2 (BHLHB1), a bHLH transcription factor, contributes to leukemogenesis in concert with LMO1. *Cancer research* 2005, **65**(16):7151-7158.

410. **Chavakis T, Bierhaus A, Nawroth PP:** RAGE (receptor for advanced glycation end products): a central player in the inflammatory response. *Microbes and infection / Institut Pasteur* 2004, **6**(13):1219-1225.

411. **Chuah YK, Basir R, Talib H, Tie TH, Nordin N:** Receptor for advanced glycation end products and its involvement in inflammatory diseases. *International journal of inflammation* 2013, **2013**:403460.

412. **Kawai T, Akira S:** The roles of TLRs, RLRs and NLRs in pathogen recognition. *International immunology* 2009, **21**(4):317-337.

413. **Shtrichman R, Samuel CE:** The role of gamma interferon in antimicrobial immunity. *Current opinion in microbiology* 2001, **4**(3):251-259.

414. **Arango Duque G, Descoteaux A:** Macrophage cytokines: involvement in immunity and infectious diseases. *Frontiers in immunology* 2014, **5**:491.

415. **Jimenez-Garza O, Guo L, Byun HM, Carrieri M, Bartolucci GB, Zhong J, Baccarelli AA:** Promoter methylation status in genes related with inflammation, nitrosative stress and xenobiotic metabolism in low-level benzene exposure: Searching for biomarkers of oncogenesis. *Food and chemical toxicology : an international journal published for the British Industrial Biological Research Association* 2017.

416. **Wahlang B, Prough RA, Falkner KC *et al*:** Polychlorinated Biphenyl-Xenobiotic Nuclear Receptor Interactions Regulate Energy Metabolism, Behavior, and Inflammation in Non-alcoholic-Steatohepatitis. *Toxicological sciences : an official journal of the Society of Toxicology* 2016, **149**(2):396-410.

417. **Zhou C, Tabb MM, Nelson EL *et al*:** Mutual repression between steroid and xenobiotic receptor and NF-kappaB signaling pathways links xenobiotic metabolism and inflammation. *The Journal of clinical investigation* 2006, **116**(8):2280-2289.

418. **Tall AR, Yvan-Charvet L:** Cholesterol, inflammation and innate immunity. *Nature reviews Immunology* 2015, **15**(2):104-116.

419. **Ghosh S:** Macrophage cholesterol homeostasis and metabolic diseases: critical role of cholesteryl ester mobilization. *Expert review of cardiovascular therapy* 2011, **9**(3):329-340.

420. **Karoui A, Allouche F, Deghrigue M, Agrebi A, Bouraoui A, Chabchoub F:** Synthesis and pharmacological evaluation of pyrazolopyrimidopyrimidine derivatives: anti-inflammatory agents with gastroprotective effect in rats. *Medicinal chemistry research : an international journal for rapid communications on design and mechanisms of action of biologically active agents* 2014, **23**:1591-1598.

421. **Fairbanks LD, Bofill M, Ruckemann K, Simmonds HA:** Importance of ribonucleotide availability to proliferating T-lymphocytes from healthy humans. Disproportionate expansion of pyrimidine pools and contrasting effects of de novo synthesis inhibitors. *The Journal of biological chemistry* 1995, **270**(50):29682-29689.

422. **Mkaddem SB, Christou I, Rossato E, Berthelot L, Lehuen A, Monteiro RC:** IgA, IgA receptors, and their anti-inflammatory properties. *Current topics in microbiology and immunology* 2014, **382**:221-235.

423. **Monteiro RC:** Role of IgA and IgA fc receptors in inflammation. *Journal of clinical immunology* 2010, **30**(1):1-9.

424. **Watanabe T, Kanamaru Y, Liu C, Suzuki Y, Tada N, Okumura K, Horikoshi S, Tomino Y:** Negative regulation of inflammatory responses by immunoglobulin A receptor (FcalphaRI) inhibits the development of Toll-like receptor-9 signalling-accelerated glomerulonephritis. *Clinical and experimental immunology* 2011, **166**(2):235-250.

425. **Ben Mkaddem S, Rossato E, Heming N, Monteiro RC:** Anti-inflammatory role of the IgA Fc receptor (CD89): from autoimmunity to therapeutic perspectives. *Autoimmunity reviews* 2013, **12**(6):666-669.

426. **Ma B, Hottiger MO:** Crosstalk between wnt/beta-Catenin and NF-kappa B Signaling Pathway during Inflammation. *Frontiers in immunology* 2016, **7**.

427. **Marelli-Berg FM, Clement M, Mauro C, Caligiuri G:** An immunologist's guide to CD31 function in T-cells. *Journal of cell science* 2013, **126**(Pt 11):2343-2352.

428. **Privratsky JR, Newman DK, Newman PJ:** PECAM-1: conflicts of interest in inflammation. *Life sciences* 2010, **87**(3-4):69-82.

429. **Privratsky JR, Tourdot BE, Newman DK, Newman PJ:** The anti-inflammatory actions of platelet endothelial cell adhesion molecule-1 do not involve regulation of endothelial cell NF-kappa B. *Journal of immunology* 2010, **184**(6):3157-3163.

430. **Woodfin A, Voisin MB, Nourshargh S:** PECAM-1: a multi-functional molecule in inflammation and vascular biology. *Arteriosclerosis, thrombosis, and vascular biology* 2007, **27**(12):2514-2523.

431. **Sorokin L:** The impact of the extracellular matrix on inflammation. *Nature reviews Immunology* 2010, **10**(10):712-723.

432. **Lopez-Novoa JM, Nieto MA:** Inflammation and EMT: an alliance towards organ fibrosis and cancer progression. *EMBO molecular medicine* 2009, **1**(6-7):303-314.

433. **Marshall BG, Wangoo A, Cook HT, Shaw RJ:** Increased inflammatory cytokines and new collagen formation in cutaneous tuberculosis and sarcoidosis. *Thorax* 1996, **51**(12):1253-1261.

434. **Levi M, van der Poll T:** Inflammation and coagulation. *Critical care medicine* 2010, **38**(2 Suppl):S26-34.

435. **Ganz T:** Defensins: antimicrobial peptides of innate immunity. *Nature reviews Immunology* 2003, **3**(9):710-720.

436. **Fernandez-Botran R, Uriarte SM, Arnold FW *et al*:** Contrasting inflammatory responses in severe and non-severe community-acquired pneumonia. *Inflammation* 2014, **37**(4):1158-1166.

437. **Quinton LJ, Jones MR, Robson BE, Mizgerd JP:** Mechanisms of the hepatic acute-phase response during bacterial pneumonia. *Infection and immunity* 2009, **77**(6):2417-2426.

438. **Hotchkiss RS, Osmon SB, Chang KC, Wagner TH, Coopersmith CM, Karl IE:** Accelerated lymphocyte death in sepsis occurs by both the death receptor and mitochondrial pathways. *Journal of immunology* 2005, **174**(8):5110-5118.

439. **van der Poll T, Opal SM:** Host–pathogen interactions in sepsis. *The Lancet infectious diseases* 2008, **8**(1):32-43.

440. **Hotchkiss RS, Monneret G, Payen D:** Immunosuppression in sepsis: a novel understanding of the disorder and a new therapeutic approach. *The Lancet infectious diseases* 2013, **13**(3):260-268.

441. **Monton C, Torres A:** Lung inflammatory response in pneumonia. *Monaldi archives for chest disease = Archivio Monaldi per le malattie del torace* 1998, **53**(1):56-63.

442. **Fowler D, Hodgekins J, Garety P, Freeman D, Kuipers E, Dunn G, Smith B, Bebbington PE:** Negative cognition, depressed mood, and paranoia: a longitudinal pathway analysis using structural equation modeling. *Schizophrenia bulletin* 2012, **38**(5):1063-1073.

443. **Lazzerini M, Seward N, Lufesi N *et al*:** Mortality and its risk factors in Malawian children admitted to hospital with clinical pneumonia, 2001-12: a retrospective observational study. *The Lancet Global health* 2016, **4**(1):e57-68.

444. **Prina E, Ranzani OT, Polverino E *et al*:** Risk factors associated with potentially antibiotic-resistant pathogens in community-acquired pneumonia. *Annals of the American Thoracic Society* 2015, **12**(2):153-160.

445. **Moran GJ, Krishnadasan A, Gorwitz RJ, Fosheim GE, Albrecht V, Limbago B, Talan DA, Group EMINS:** Prevalence of methicillin-resistant staphylococcus aureus as an etiology of community-acquired pneumonia. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2012, **54**(8):1126-1133.

446. **Seligman R, Ramos-Lima LF, Oliveira Vdo A, Sanvicente C, Pacheco EF, Dalla Rosa K:** Biomarkers in community-acquired pneumonia: a state-of-the-art review. *Clinics* 2012, **67**(11):1321-1325.

447. **Wacker C, Prkno A, Brunkhorst FM, Schlattmann P:** Procalcitonin as a diagnostic marker for sepsis: a systematic review and meta-analysis. *Lancet Infectious Diseases* 2013, **13**(5):426-435.

448. **Porfyridis I, Georgiadis G, Vogazianos P, Mitis G, Georgiou A:** C-reactive protein, procalcitonin, clinical pulmonary infection score, and pneumonia severity scores in nursing home acquired pneumonia. *Respiratory care* 2014, **59**(4):574-581.

449. **Agnello L, Bellia C, Di Gangi M, Lo Sasso B, Calvaruso L, Bivona G, Scazzone C, Dones P, Ciaccio M:** Utility of serum procalcitonin and C-reactive protein in severity assessment of community-acquired pneumonia in children. *Clinical biochemistry* 2016, **49**(1):47-50.

450. **Oshansky CM, Zhang W, Moore E, Tripp RA:** The host response and molecular pathogenesis associated with respiratory syncytial virus infection. *Future microbiology* 2009, **4**(3):279-297.

451. **Blankley S, Berry MP, Graham CM, Bloom CI, Lipman M, O'Garra A:** The application of transcriptional blood signatures to enhance our understanding of the host response to infection: the example of tuberculosis. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2014, **369**(1645):20130427.

452. **Kaforou M, Wright VJ, Levin M:** Host RNA signatures for diagnostics: an example from paediatric tuberculosis in Africa. *The Journal of infection* 2014, **69 Suppl 1**:S28-31.

453. **Khondoker M, Dobson R, Skirrow C, Simmons A, Stahl D:** A comparison of machine learning methods for classification using simulation with multiple

real data examples from mental health studies. *Statistical methods in medical research* 2016, **25**(5):1804-1823.

454. **Byvatov E, Fechner U, Sadowski J, Schneider G:** Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of chemical information and computer sciences* 2003, **43**(6):1882-1889.

455. **Nagaoka Y, Nosaka N, Yamada M, Yashiro M, Washio Y, Baba K, Morishima T, Tsukahara H:** Local and Systemic Immune Responses to Influenza A Virus Infection in Pneumonia and Encephalitis Mouse Models. *Disease markers* 2017, **2017**:2594231.

456. **Venet D, Pecasse F, Maenhaut C, Bersini H:** Separation of samples into their constituents using gene expression data. *Bioinformatics* 2001, **17 Suppl 1**:S279-287.

457. **Whitney AR, Diehn M, Popper SJ, Alizadeh AA, Boldrick JC, Relman DA, Brown PO:** Individuality and variation in gene expression patterns in human blood. *Proceedings of the National Academy of Sciences of the United States of America* 2003, **100**(4):1896-1901.

458. **Culley FJ:** Natural killer cells in infection and inflammation of the lung. *Immunology* 2009, **128**(2):151-163.

459. **Christ-Crain M, Muller B:** Biomarkers in respiratory tract infections: diagnostic guides to antibiotic prescription, prognostic markers and mediators. *The European respiratory journal* 2007, **30**(3):556-573.

460. **Onder G:** [The advantages and limitations of observational studies]. *Giornale italiano di cardiologia* 2013, **14**(3 Suppl 1):35-39.

461. **Wiedermann CJ:** The limitations of observational studies on the treatment of severe sepsis. *Critical care* 2002, **6**(6):546-547; author reply 548.

462. **Fasold M, Binder H:** Variation of RNA Quality and Quantity Are Major Sources of Batch Effects in Microarray Expression Data. *Microarrays* 2014, **3**(4):322-339.

463. **Roh SW, Abell GC, Kim KH, Nam YD, Bae JW:** Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends in biotechnology* 2010, **28**(6):291-299.

464. **Neill DR, Fernandes VE, Wisby L *et al*:** T regulatory cells control susceptibility to invasive pneumococcal pneumonia in mice. *PLoS pathogens* 2012, **8**(4):e1002660.

465. **Liu H, Liu L, Zhang H:** Ensemble gene selection by grouping for microarray data classification. *Journal of biomedical informatics* 2010, **43**(1):81-87.

466. **Hardy RR, Hayakawa K, Haaijman J, Herzenberg LA:** B-cell subpopulations identifiable by two-color fluorescence analysis using a dual-laser FACS. *Annals of the New York Academy of Sciences* 1982, **399**:112-121.

467. **Carter J, Newport A, Keeler KD, Dresser DW:** FACS analysis of changes in T and B lymphocyte populations in the blood, spleen and lymph nodes of pregnant mice. *Immunology* 1983, **48**(4):791-797.

468. **von Recum-Knepper J, Sadewasser A, Weinheimer VK, Wolff T:** FACS-based analysis reveals an asymmetric induction of interferon stimulated genes in response to seasonal influenza a virus. *Journal of virology* 2015.

469. **Hazelton BJ, Thomas LC, Unver T, Iredell JR:** Rapid identification of Gram-positive pathogens and their resistance genes from positive blood culture broth using a multiplex tandem RT-PCR assay. *Journal of medical microbiology* 2013, **62**(Pt 2):223-231.

470. **Ho V, Yeo SY, Kunasegaran K, De Silva D, Tarulli GA, Voorhoeve PM, Pietersen AM:** Expression analysis of rare cellular subsets: direct RT-PCR on limited cell numbers obtained by FACS or soft agar assays. *BioTechniques* 2013, **54**(4):208-212.

471. **Foongladda S, Mongkol N, Petlum P, Chayakulkeeree M:** Multi-probe real-time PCR identification of four common Candida species in blood culture broth. *Mycopathologia* 2014, **177**(5-6):251-261.

Chapter 9: Appendices

9.1 Appendix A (Chapter 4): An optimal Integrated Blood Marker List (IBML).

Table 9.1: Immune cell type-specific marker genes compiled in IBML. The table shows the distribution of marker genes that are associated with six immune cell types (B cells, T cells, NK cells, dendritic cells, Monocytes and Neutrophils). IBML was derived in **Chapter 4**.

CELLTYPE	SYMBOL	ENTREZID	GENENAME
B	CD19	930	CD19 molecule
B	CD79A	973	CD79a molecule, immunoglobulin-associated alpha
B	CD79B	974	CD79b molecule, immunoglobulin-associated beta
B	FCRL2	79368	Fc receptor-like 2
B	IGLJ3	28831	immunoglobulin lambda joining 3
B	KIAA0125	9834	KIAA0125
B	OSBPL10	114884	oxysterol binding protein-like 10
B	P2RX5	5026	purinergic receptor P2X, ligand-gated ion channel, 5
B	POU2AF1	5450	POU class 2 associating factor 1
B	TPD52	7163	tumor protein D52
T	ABLIM1	3983	actin binding LIM protein 1
T	BCL11B	64919	B-cell CLL/lymphoma 11B (zinc finger protein)
T	CAMK4	814	calcium/calmodulin-dependent protein kinase IV
T	CD28	940	CD28 molecule
T	CD3D	915	CD3d molecule, delta (CD3-TCR complex)
T	CD3E	916	CD3e molecule, epsilon (CD3-TCR complex)
T	CD3G	917	CD3g molecule, gamma (CD3-TCR complex)
T	CD5	921	CD5 molecule
T	CD6	923	CD6 molecule
T	CDR2	1039	cerebellar degeneration-related protein 2, 62kDa
T	DGKA	1606	diacylglycerol kinase, alpha 80kDa
T	FBLN5	10516	fibulin 5
T	FLT3LG	2323	fms-related tyrosine kinase 3 ligand
T	ICOS	29851	inducible T-cell co-stimulator
T	IL7R	3575	interleukin 7 receptor
T	INPP4B	8821	inositol polyphosphate-4-phosphatase, type II, 105kDa
T	ITK	3702	IL2-inducible T-cell kinase
T	ITPKB	3707	inositol-trisphosphate 3-kinase B
T	LDLRAP1	26119	low density lipoprotein receptor adaptor protein 1
T	LEF1	51176	lymphoid enhancer-binding factor 1
T	LEPROTL1	23484	leptin receptor overlapping transcript-like 1

T	MAL	4118	mal, T-cell differentiation protein
T	NELL2	4753	NEL-like 2 (chicken)
T	NOSIP	51070	nitric oxide synthase interacting protein
T	PASK	23178	PAS domain containing serine/threonine kinase
T	PIK3IP1	113791	phosphoinositide-3-kinase interacting protein 1
T	PLEKHB1	58473	pleckstrin homology domain containing, family B (evectins) member 1
T	SIRPG	55423	signal-regulatory protein gamma
T	SPOCK2	9806	sparc/osteonectin, cwcv and kazal-like domains proteoglycan (testican) 2
T	TCF7	6932	transcription factor 7 (T-cell specific, HMG-box)
T	TMEM204	79652	transmembrane protein 204
T	TNFRSF25	8718	tumor necrosis factor receptor superfamily, member 25
T	TRAT1	50852	T cell receptor associated transmembrane adaptor 1
T	UBASH3A	53347	ubiquitin associated and SH3 domain containing A
T	YME1L1	10730	YME1-like 1 ATPase
NK	ARAP2	116984	ArfGAP with RhoGAP domain, ankyrin repeat and PH domain 2
NK	ASCL2	430	achaete-scute family bHLH transcription factor 2
NK	AUTS2	26053	autism susceptibility candidate 2
NK	BZRAP1	9256	benzodiazepine receptor (peripheral) associated protein 1
NK	CD247	919	CD247 molecule
NK	CHST12	55501	carbohydrate (chondroitin 4) sulfotransferase 12
NK	CLIC3	9022	chloride intracellular channel 3
NK	CST7	8530	cystatin F (leukocystatin)
NK	F2R	2149	coagulation factor II (thrombin) receptor
NK	GNLY	10578	granulysin
NK	GNPTAB	79158	N-acetylglucosamine-1-phosphate transferase, alpha and beta subunits
NK	GZMA	3001	granzyme A (granzyme 1, cytotoxic T-lymphocyte-associated serine esterase 3)
NK	GZMB	3002	granzyme B (granzyme 2, cytotoxic T-lymphocyte-associated serine esterase 1)
NK	HEG1	57493	heart development protein with EGF-like domains 1
NK	IL12RB2	3595	interleukin 12 receptor, beta 2
NK	IL18RAP	8807	interleukin 18 receptor accessory protein
NK	IL2RB	3560	interleukin 2 receptor, beta
NK	JAK1	3716	Janus kinase 1
NK	KIR2DL2	3803	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 2
NK	KIR2DL4	3805	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 4
NK	KIR2DL5A	57292	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 5A
NK	KIR2DL5B	553128	killer cell immunoglobulin-like receptor, two domains, long cytoplasmic tail, 5B
NK	KIR2DS1	3806	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 1

NK	KIR2DS4	3809	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 4
NK	KIR3DL3	115653	killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 3
NK	KIR3DL3	100133046	killer cell immunoglobulin-like receptor three domains long cytoplasmic tail 3
NK	KIR3DS1	3813	killer cell immunoglobulin-like receptor, three domains, short cytoplasmic tail, 1
NK	KLRC1	3821	killer cell lectin-like receptor subfamily C, member 1
NK	KLRC2	3822	killer cell lectin-like receptor subfamily C, member 2
NK	KLRD1	3824	killer cell lectin-like receptor subfamily D, member 1
NK	KLRF1	51348	killer cell lectin-like receptor subfamily F, member 1
NK	MACF1	23499	microtubule-actin crosslinking factor 1
NK	MYBL1	4603	v-myb avian myeloblastosis viral oncogene homolog-like 1
NK	NCAM1	4684	neural cell adhesion molecule 1
NK	PDGFRB	5159	platelet-derived growth factor receptor, beta polypeptide
NK	PRF1	5551	perforin 1 (pore forming protein)
NK	PRR5L	79899	proline rich 5 like
NK	PTPN4	5775	protein tyrosine phosphatase, non-receptor type 4 (megakaryocyte)
NK	RGS3	5998	regulator of G-protein signaling 3
NK	S1PR5	53637	sphingosine-1-phosphate receptor 5
NK	STOM	2040	stomatin
NK	TBX21	30009	T-box 21
NK	TFDP2	7029	transcription factor Dp-2 (E2F dimerization partner 2)
NK	TGFBR3	7049	transforming growth factor, beta receptor III
NK	XCL1	6375	chemokine (C motif) ligand 1
NK	XCL2	6846	chemokine (C motif) ligand 2
NK	YPEL1	29799	yippee-like 1 (Drosophila)
Dendritic	ALCAM	214	activated leukocyte cell adhesion molecule
Dendritic	ATP1B1	481	ATPase, Na+/K+ transporting, beta 1 polypeptide
Dendritic	CCDC88A	55704	coiled-coil domain containing 88A
Dendritic	CD1E	913	CD1e molecule
Dendritic	CLEC10A	10462	C-type lectin domain family 10, member A
Dendritic	MRC1	4360	mannose receptor, C type 1
Dendritic	PON2	5445	paraoxonase 2
Dendritic	SPINT2	10653	serine peptidase inhibitor, Kunitz type, 2
Dendritic	UBE2A	7319	ubiquitin-conjugating enzyme E2A
Monocytes	ANXA1	301	annexin A1
Monocytes	AP1S2	8905	adaptor-related protein complex 1, sigma 2 subunit
Monocytes	ARHGEF10L	55160	Rho guanine nucleotide exchange factor (GEF) 10-like
Monocytes	ASGR1	432	asialoglycoprotein receptor 1
Monocytes	ASGR2	433	asialoglycoprotein receptor 2
Monocytes	CD14	929	CD14 molecule
Monocytes	CSTA	1475	cystatin A (stefin A)

Monocytes	CYBB	1536	cytochrome b-245, beta polypeptide
Monocytes	DIAPH2	1730	diaphanous-related formin 2
Monocytes	DUSP6	1848	dual specificity phosphatase 6
Monocytes	FXYD6	53826	FXYD domain containing ion transport regulator 6
Monocytes	IRAK3	11213	interleukin-1 receptor-associated kinase 3
Monocytes	METTL9	51108	methyltransferase like 9
Monocytes	MS4A6A	64231	membrane-spanning 4-domains, subfamily A, member 6A
Monocytes	NLRP3	114548	NLR family, pyrin domain containing 3
Monocytes	P2RY2	5029	purinergic receptor P2Y, G-protein coupled, 2
Monocytes	PID1	55022	phosphotyrosine interaction domain containing 1
Monocytes	PSAP	5660	prosaposin
Monocytes	SLC7A7	9056	solute carrier family 7 (amino acid transporter light chain, y+L system), member 7
Monocytes	SLX1A-SULT1A3	100526830	SLX1A-SULT1A3 readthrough (NMD candidate)
Monocytes	SLX1B-SULT1A4	100526831	SLX1B-SULT1A4 readthrough (NMD candidate)
Monocytes	SULT1A3	6818	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 3
Monocytes	SULT1A4	445329	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 4
Monocytes	TPPP3	51673	tubulin polymerization-promoting protein family member 3
Monocytes	VCAN	1462	versican
Neutrophils	AATK	9625	apoptosis-associated tyrosine kinase
Neutrophils	ABHD5	51099	abhydrolase domain containing 5
Neutrophils	ABTB1	80325	ankyrin repeat and BTB (POZ) domain containing 1
Neutrophils	ACOX1	51	acyl-CoA oxidase 1, palmitoyl
Neutrophils	ARAP3	64411	ArfGAP with RhoGAP domain, ankyrin repeat and PH domain 3
Neutrophils	ARHGEF40	55701	Rho guanine nucleotide exchange factor (GEF) 40
Neutrophils	B3GNT8	374907	UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 8
Neutrophils	BASP1	10409	brain abundant, membrane attached signal protein 1
Neutrophils	BEST1	7439	bestrophin 1
Neutrophils	BID	637	BH3 interacting domain death agonist
Neutrophils	BTNL8	79908	butyrophilin-like 8
Neutrophils	C5AR1	728	complement component 5a receptor 1
Neutrophils	CAMK2G	818	calcium/calmodulin-dependent protein kinase II gamma
Neutrophils	CANT1	124583	calcium activated nucleotidase 1
Neutrophils	CCNJL	79616	cyclin J-like
Neutrophils	CCR3	1232	chemokine (C-C motif) receptor 3
Neutrophils	CD46	4179	CD46 molecule, complement regulatory protein
Neutrophils	CEACAM3	1084	carcinoembryonic antigen-related cell adhesion molecule 3
Neutrophils	CENPBD1P1	65996	CENPBD1 pseudogene 1
Neutrophils	CEP19	84984	centrosomal protein 19kDa

Neutrophils	CFLAR	8837	CASP8 and FADD-like apoptosis regulator
Neutrophils	CHI3L1	1116	chitinase 3-like 1 (cartilage glycoprotein-39)
Neutrophils	CIR1	9541	corepressor interacting with RBPJ, 1
Neutrophils	CKLF	51192	chemokine-like factor
Neutrophils	CKLF-CMTM1	100529251	CKLF-CMTM1 readthrough
Neutrophils	CMTM2	146225	CKLF-like MARVEL transmembrane domain containing 2
Neutrophils	CPD	1362	carboxypeptidase D
Neutrophils	CPPED1	55313	calcineurin-like phosphoesterase domain containing 1
Neutrophils	CREB5	9586	cAMP responsive element binding protein 5
Neutrophils	CREBRF	153222	CREB3 regulatory factor
Neutrophils	CTBS	1486	chitobiase, di-N-acetyl-
Neutrophils	CXCR1	3577	chemokine (C-X-C motif) receptor 1
Neutrophils	CYP4F3	4051	cytochrome P450, family 4, subfamily F, polypeptide 3
Neutrophils	DAPK2	23604	death-associated protein kinase 2
Neutrophils	DGAT2	84649	diacylglycerol O-acyltransferase 2
Neutrophils	DOCK5	80005	dedicator of cytokinesis 5
Neutrophils	DSC2	1824	desmocollin 2
Neutrophils	EGLN1	54583	egl-9 family hypoxia-inducible factor 1
Neutrophils	EMR3	84658	egf-like module containing, mucin-like, hormone receptor-like 3
Neutrophils	EPHB1	2047	EPH receptor B1
Neutrophils	EPOR	2057	erythropoietin receptor
Neutrophils	F11R	50848	F11 receptor
Neutrophils	FAM129A	116496	family with sequence similarity 129, member A
Neutrophils	FAM212B	55924	family with sequence similarity 212, member B
Neutrophils	FAM53C	51307	family with sequence similarity 53, member C
Neutrophils	FAS	355	Fas cell surface death receptor
Neutrophils	FFAR2	2867	free fatty acid receptor 2
Neutrophils	FPR1	2357	formyl peptide receptor 1
Neutrophils	FRAT2	23401	frequently rearranged in advanced T-cell lymphomas 2
Neutrophils	GMFG	9535	glia maturation factor, gamma
Neutrophils	GPR97	222487	G protein-coupled receptor 97
Neutrophils	HIST1H2AC	8334	histone cluster 1, H2ac
Neutrophils	HIST1H2BC	8347	histone cluster 1, H2bc
Neutrophils	HIST1H2BE	8344	histone cluster 1, H2be
Neutrophils	HIST1H2BF	8343	histone cluster 1, H2bf
Neutrophils	HIST1H2BG	8339	histone cluster 1, H2bg
Neutrophils	HIST1H2BI	8346	histone cluster 1, H2bi
Neutrophils	HOTAIRM1	100506311	HOXA transcript antisense RNA, myeloid-specific 1
Neutrophils	HSPA6	3310	heat shock 70kDa protein 6 (HSP70B')
Neutrophils	IDS	3423	iduronate 2-sulfatase
Neutrophils	IFRD1	3475	interferon-related developmental regulator 1
Neutrophils	IKBIP	121457	IKBKB interacting protein

Neutrophils	INAFM1	255783	InaF-motif containing 1
Neutrophils	ITPRIP	85450	inositol 1,4,5-trisphosphate receptor interacting protein
Neutrophils	KATNBL1	79768	katanin p80 subunit B-like 1
Neutrophils	KCNJ15	3772	potassium inwardly-rectifying channel, subfamily J, member 15
Neutrophils	KDM6B	23135	lysine (K)-specific demethylase 6B
Neutrophils	KIAA1324	57535	KIAA1324
Neutrophils	KIF13A	63971	kinesin family member 13A
Neutrophils	LGALSL	29094	lectin, galactoside-binding-like
Neutrophils	LITAF	9516	lipopolysaccharide-induced TNF factor
Neutrophils	LOC643072	643072	uncharacterized LOC643072
Neutrophils	LRG1	116844	leucine-rich alpha-2-glycoprotein 1
Neutrophils	LRP10	26020	low density lipoprotein receptor-related protein 10
Neutrophils	LRRC4	64101	leucine rich repeat containing 4
Neutrophils	LST1	7940	leukocyte specific transcript 1
Neutrophils	LYN	4067	LYN proto-oncogene, Src family tyrosine kinase
Neutrophils	MAP2K4	6416	mitogen-activated protein kinase kinase 4
Neutrophils	MBOAT2	129642	membrane bound O-acyltransferase domain containing 2
Neutrophils	MBOAT7	79143	membrane bound O-acyltransferase domain containing 7
Neutrophils	MGAM	8972	maltase-glucoamylase (alpha-glucosidase)
Neutrophils	MME	4311	membrane metallo-endopeptidase
Neutrophils	MMP25	64386	matrix metallopeptidase 25
Neutrophils	MNDA	4332	myeloid cell nuclear differentiation antigen
Neutrophils	MPZL3	196264	myelin protein zero-like 3
Neutrophils	MRVI1	10335	murine retrovirus integration site 1 homolog
Neutrophils	MSL1	339287	male-specific lethal 1 homolog (Drosophila)
Neutrophils	MSRB1	51734	methionine sulfoxide reductase B1
Neutrophils	MTHFS	10588	5,10-methenyltetrahydrofolate synthetase (5-formyltetrahydrofolate cyclo-ligase)
Neutrophils	MXD1	4084	MAX dimerization protein 1
Neutrophils	NATD1	256302	N-acetyltransferase domain containing 1
Neutrophils	NCF4	4689	neutrophil cytosolic factor 4, 40kDa
Neutrophils	NCOA1	8648	nuclear receptor coactivator 1
Neutrophils	NDEL1	81565	nudE neurodevelopment protein 1-like 1
Neutrophils	NFIL3	4783	nuclear factor, interleukin 3 regulated
Neutrophils	OAZ2	4947	ornithine decarboxylase antizyme 2
Neutrophils	PGS1	9489	phosphatidylglycerophosphate synthase 1
Neutrophils	PHC2	1912	polyhomeotic homolog 2 (Drosophila)
Neutrophils	PHF20L1	51105	PHD finger protein 20-like 1
Neutrophils	PIGB	9488	phosphatidylinositol glycan anchor biosynthesis, class B
Neutrophils	PIGX	54965	phosphatidylinositol glycan anchor biosynthesis, class X
Neutrophils	POLB	5423	polymerase (DNA directed), beta
Neutrophils	PPP1R3B	79660	protein phosphatase 1, regulatory subunit 3B

Neutrophils	PPP4R1	9989	protein phosphatase 4, regulatory subunit 1
Neutrophils	PROK2	60675	prokineticin 2
Neutrophils	R3HDM4	91300	R3H domain containing 4
Neutrophils	RAF1	5894	Raf-1 proto-oncogene, serine/threonine kinase
Neutrophils	RALB	5899	v-ral simian leukemia viral oncogene homolog B
Neutrophils	REM2	161253	RAS (RAD and GEM)-like GTP binding 2
Neutrophils	REPS2	9185	RALBP1 associated Eps domain containing 2
Neutrophils	RGL4	266747	ral guanine nucleotide dissociation stimulator-like 4
Neutrophils	RGS18	64407	regulator of G-protein signaling 18
Neutrophils	RNASET2	8635	ribonuclease T2
Neutrophils	RNF149	284996	ring finger protein 149
Neutrophils	ROPN1L	83853	rhophilin associated tail protein 1-like
Neutrophils	S100P	6286	S100 calcium binding protein P
Neutrophils	S1PR4	8698	sphingosine-1-phosphate receptor 4
Neutrophils	SEC14L1	6397	SEC14-like 1 (S. cerevisiae)
Neutrophils	SLC22A4	6583	solute carrier family 22 (organic cation/zwitterion transporter), member 4
Neutrophils	SLC25A37	51312	solute carrier family 25 (mitochondrial iron transporter), member 37
Neutrophils	SLC45A4	57210	solute carrier family 45, member 4
Neutrophils	SLPI	6590	secretory leukocyte peptidase inhibitor
Neutrophils	SRGN	5552	serglycin
Neutrophils	ST20	400410	suppressor of tumorigenicity 20
Neutrophils	ST20-MTHFS	100528021	ST20-MTHFS readthrough
Neutrophils	ST6GALNAC2	10610	ST6 (alpha-N-acetyl-neuraminyl-2,3-beta-galactosyl-1,3)-N-acetylgalactosaminide alpha-2,6-sialyltransferase 2
Neutrophils	STEAP4	79689	STEAP family member 4
Neutrophils	STK40	83931	serine/threonine kinase 40
Neutrophils	STX3	6809	syntaxin 3
Neutrophils	TBXAS1	6916	thromboxane A synthase 1 (platelet)
Neutrophils	TECPR2	9895	tectonin beta-propeller repeat containing 2
Neutrophils	TIGD3	220359	tigger transposable element derived 3
Neutrophils	TLE3	7090	transducin-like enhancer of split 3
Neutrophils	TLR6	10333	toll-like receptor 6
Neutrophils	TMCC1	23023	transmembrane and coiled-coil domain family 1
Neutrophils	TMCC3	57458	transmembrane and coiled-coil domain family 3
Neutrophils	TMEM154	201799	transmembrane protein 154
Neutrophils	TMEM71	137835	transmembrane protein 71
Neutrophils	TNFRSF10C	8794	tumor necrosis factor receptor superfamily, member 10c, decoy without an intracellular domain
Neutrophils	TOPORS-AS1	100129250	TOPORS antisense RNA 1
Neutrophils	TREM1	54210	triggering receptor expressed on myeloid cells 1
Neutrophils	TRPM6	140803	transient receptor potential cation channel, subfamily M, member 6

Neutrophils	TSEN34	79042	TSEN34 tRNA splicing endonuclease subunit
Neutrophils	UBE2B	7320	ubiquitin-conjugating enzyme E2B
Neutrophils	UBE2R2	54926	ubiquitin-conjugating enzyme E2R 2
Neutrophils	UBXN2B	137886	UBX domain protein 2B
Neutrophils	USP15	9958	ubiquitin specific peptidase 15
Neutrophils	VNN2	8875	vanin 2
Neutrophils	VNN3	55350	vanin 3
Neutrophils	XPO6	23214	exportin 6
Neutrophils	ZDHHC18	84243	zinc finger, DHHC-type containing 18
Neutrophils	ZNF117	51351	zinc finger protein 117

.

9.2 Appendix B (Chapter 5): Annotation of differentially expressed genes on KEGG pathways.

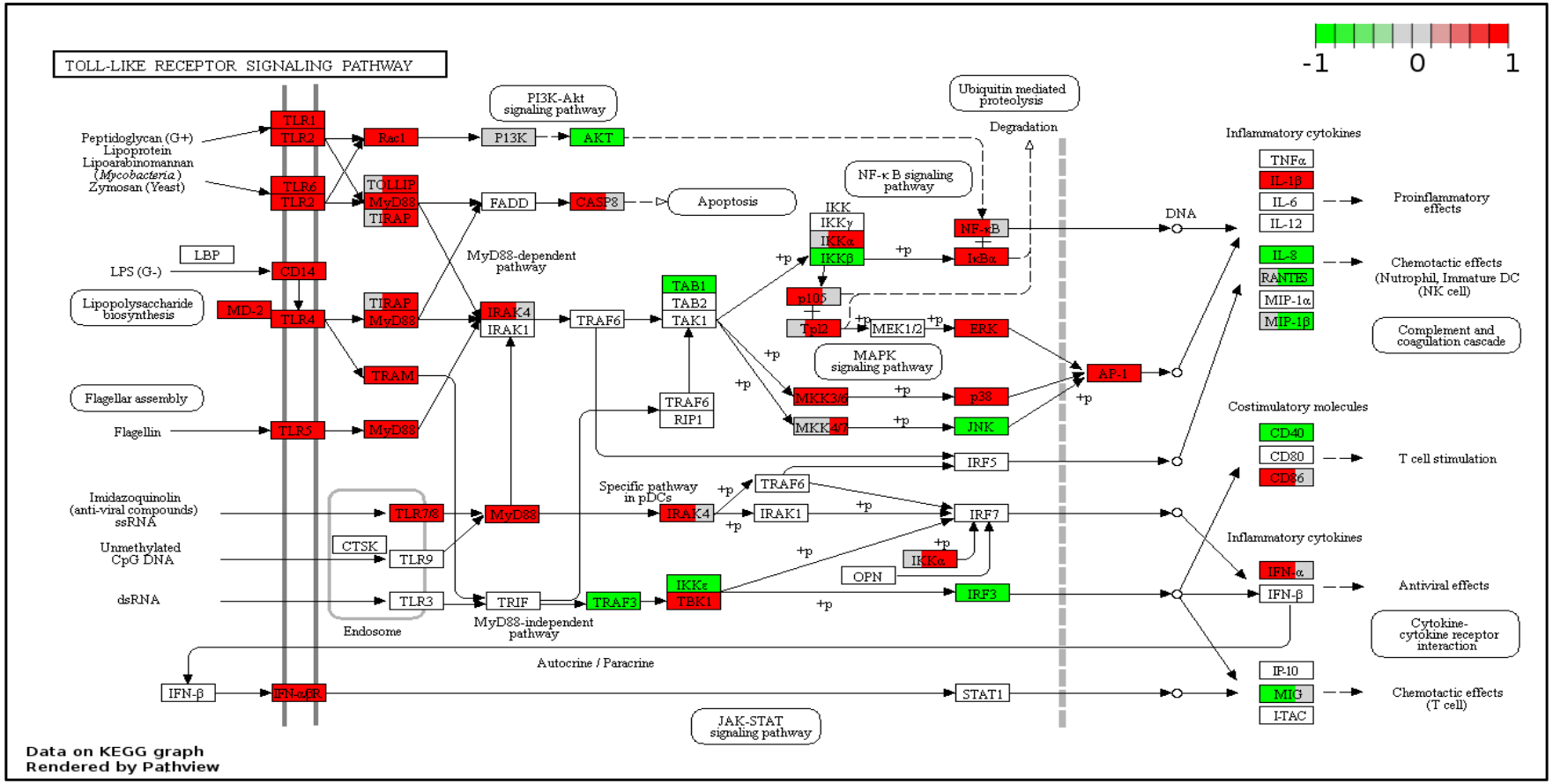


Figure 9.1: The Toll-like receptor KEGG pathway map (hsa04620) showing differentially expressed genes in pneumonia (FDR<0.05). Each coloured box is divided into three sections representing mild (left), severe (middle) and very severe (right) pneumonia states. Colours: (i) White=Gene not analysed, (ii) Grey=Gene not significant (FDR>0.05), (iii) Green=Down-regulated genes and (iv) red=up-regulated gene. The pathway map was produced using the PathView Web (https://pathview.uncc.edu/).

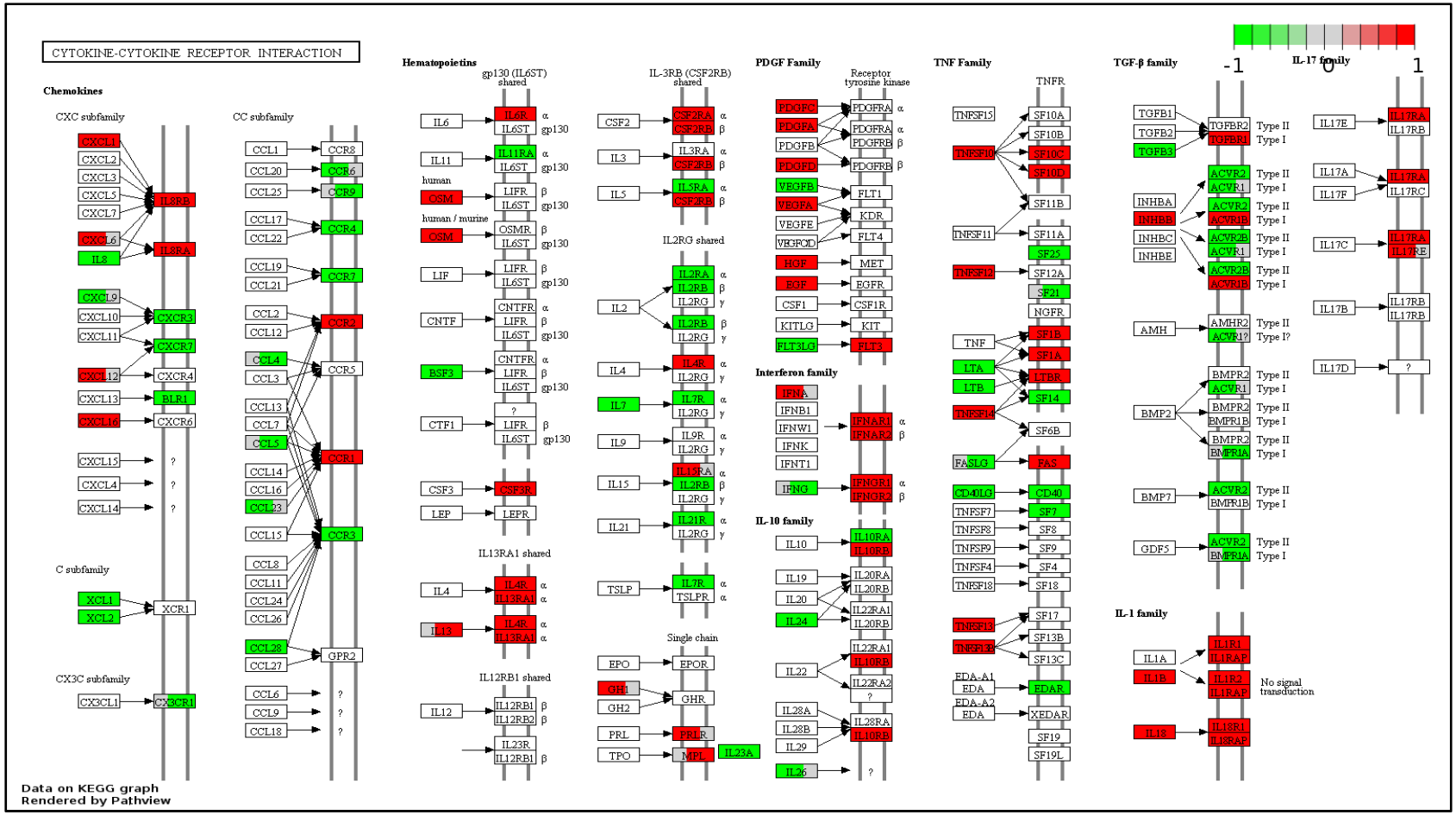


Figure 9.2: The Cytokine-cytokine receptor interaction KEGG map (hsa04060) showing differentially expressed genes in pneumonia (FDR<0.05). Each coloured box is divided into three sections representing mild (left), severe (middle) and very severe (right) pneumonia states. Colours: (i) White=Gene not analysed, (ii) Grey=Gene not significant (FDR>0.05), (iii) Green=Down-regulated genes and (iv) red=up-regulated gene. The pathway map was produced using the PathView Web (https://pathview.uncc.edu/).

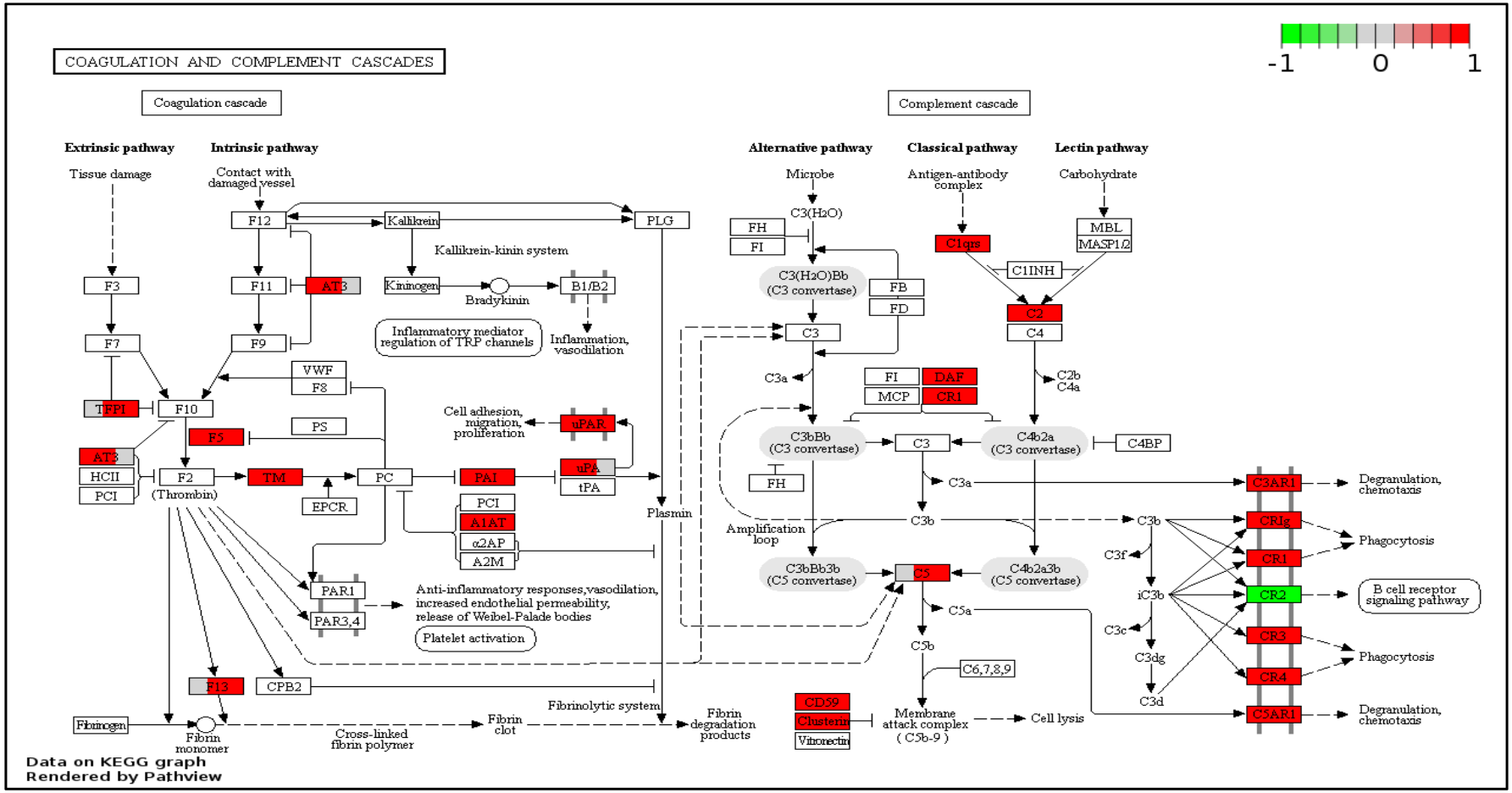


Figure 9.3: The Complement and coagulation cascades KEGG map (hsa04610) showing differentially expressed genes in pneumonia (FDR<0.05). Each coloured box is divided into three sections representing mild (left), severe (middle) and very severe (right) pneumonia states. Colours: (i) White=Gene not analysed, (ii) Grey=Gene not significant (FDR>0.05), (iii) Green=Down-regulated genes and (iv) red=up-regulated gene. The pathway map was produced using the PathView Web (https://pathview.uncc.edu/).

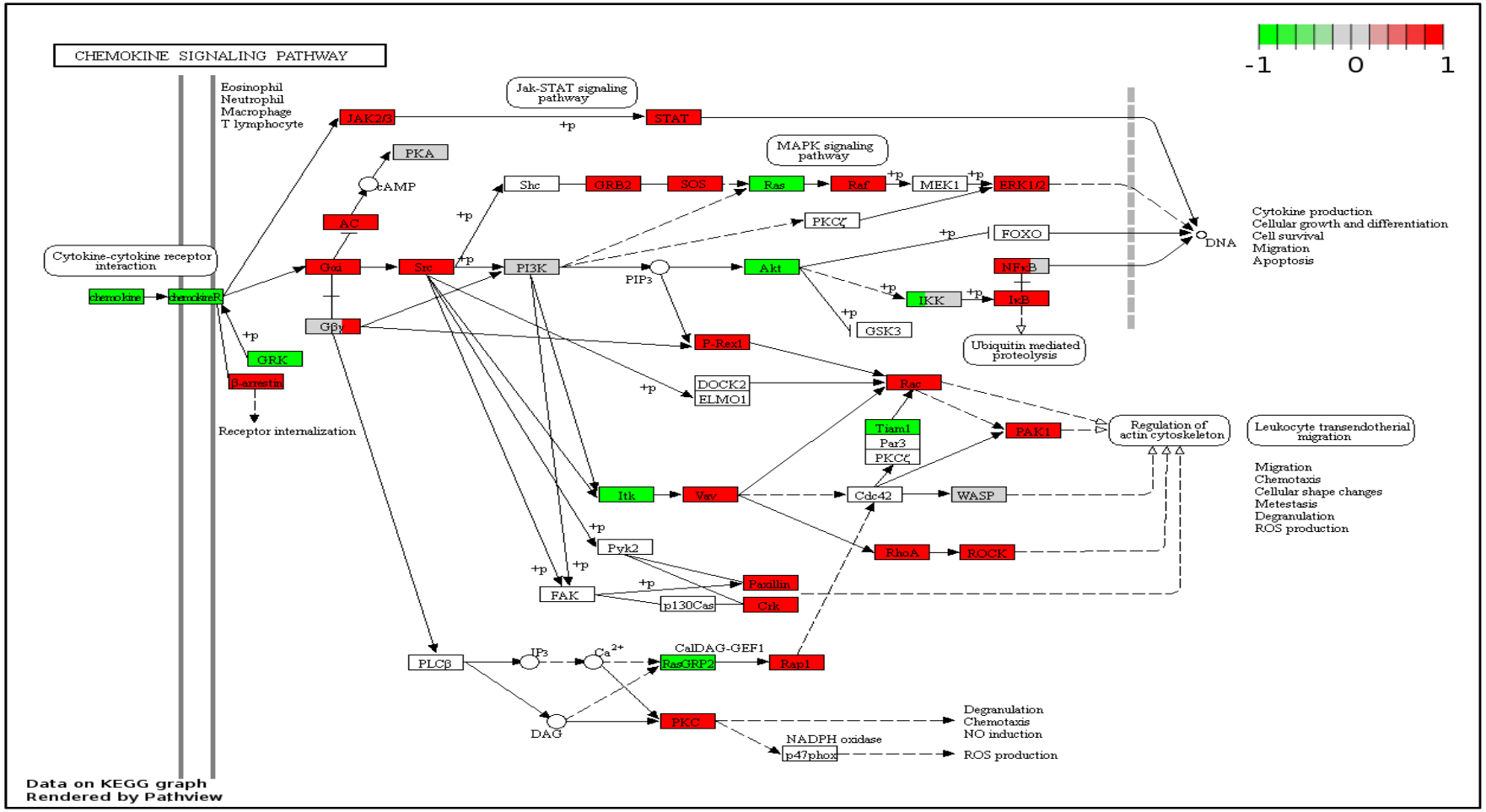


Figure 9.4: The Chemokine signalling pathway KEGG map (hsa04062-) showing differentially expressed genes in pneumonia (FDR<0.05). Each coloured box is divided into three sections representing mild (left), severe (middle) and very severe (right) pneumonia states. Colours: (i) White=Gene not analysed, (ii) Grey=Gene not significant (FDR>0.05), (iii) Green=Down-regulated genes and (iv) red=up-regulated gene. The pathway map was produced using the PathView Web (https://pathview.uncc.edu/).

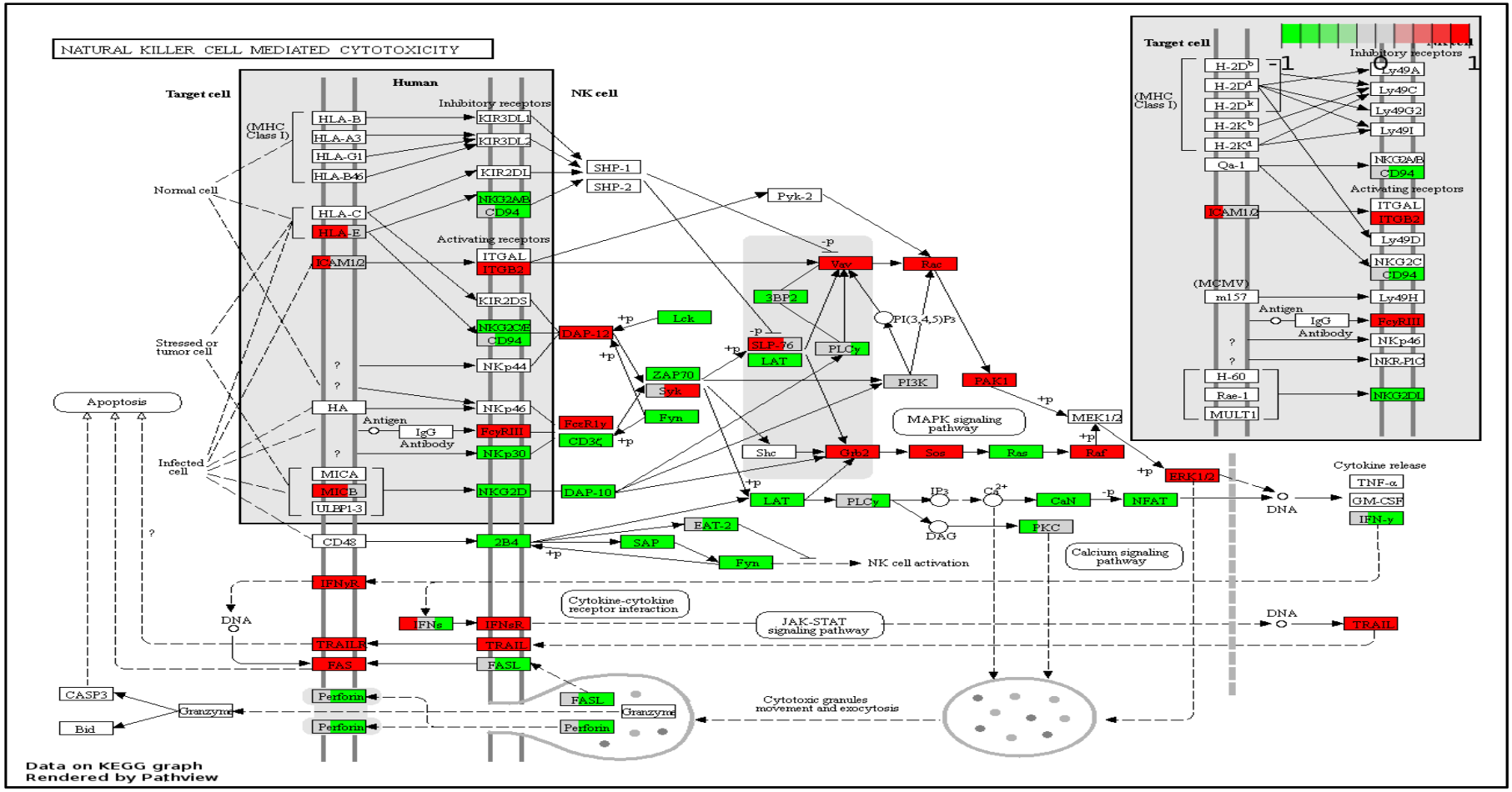


Figure 9.5: The Natural killer cell mediated cytotoxicity KEGG map (hsa04650) showing differentially expressed genes in pneumonia (FDR<0.05). Each coloured box is divided into three sections representing mild (left), severe (middle) and very severe (right) pneumonia states. Colours: (i) White=Gene not analysed, (ii) Grey=Gene not significant (FDR>0.05), (iii) Green=Down-regulated genes and (iv) Red=up-regulated gene. The pathway map was produced using the PathView Web (https://pathview.uncc.edu/).

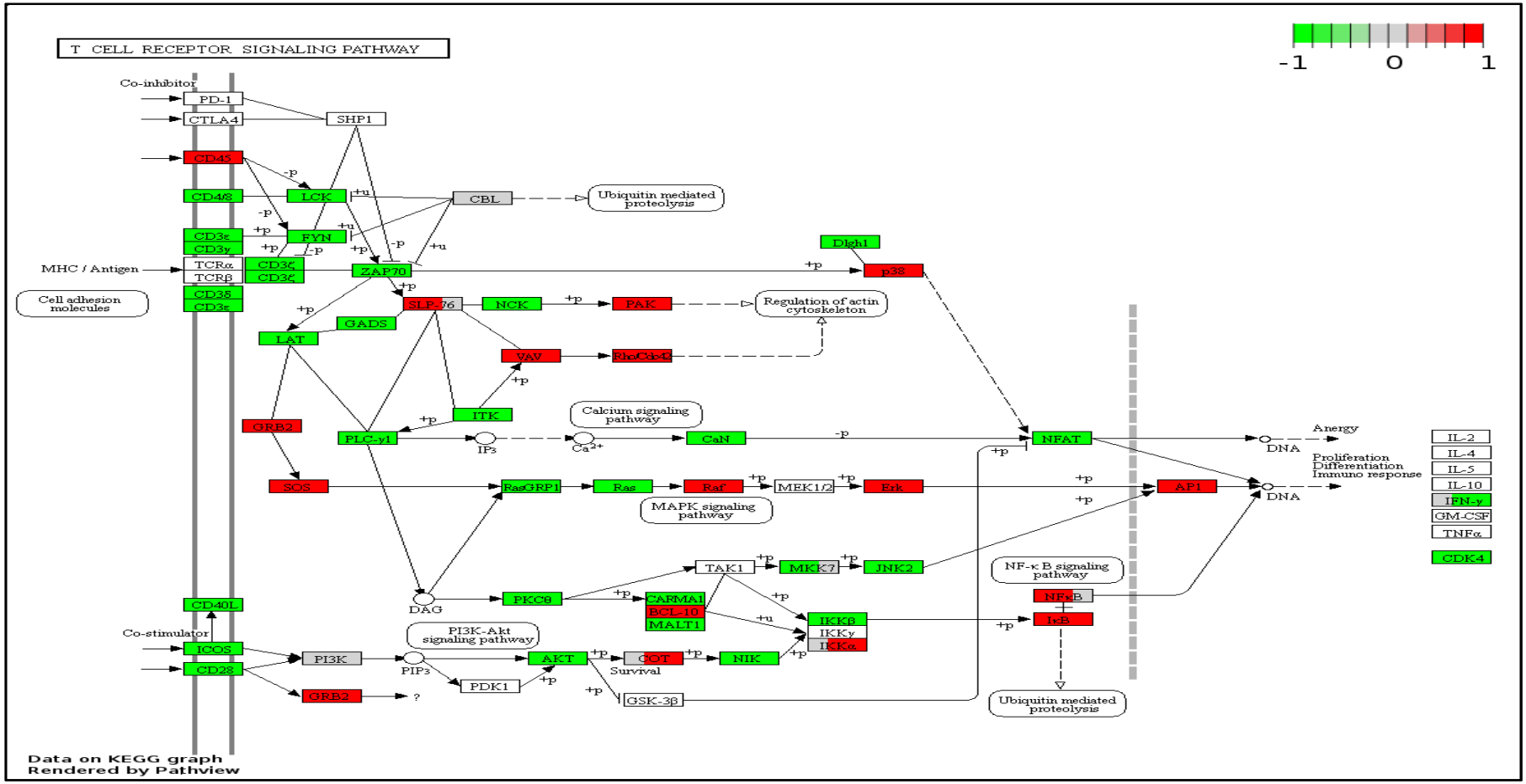
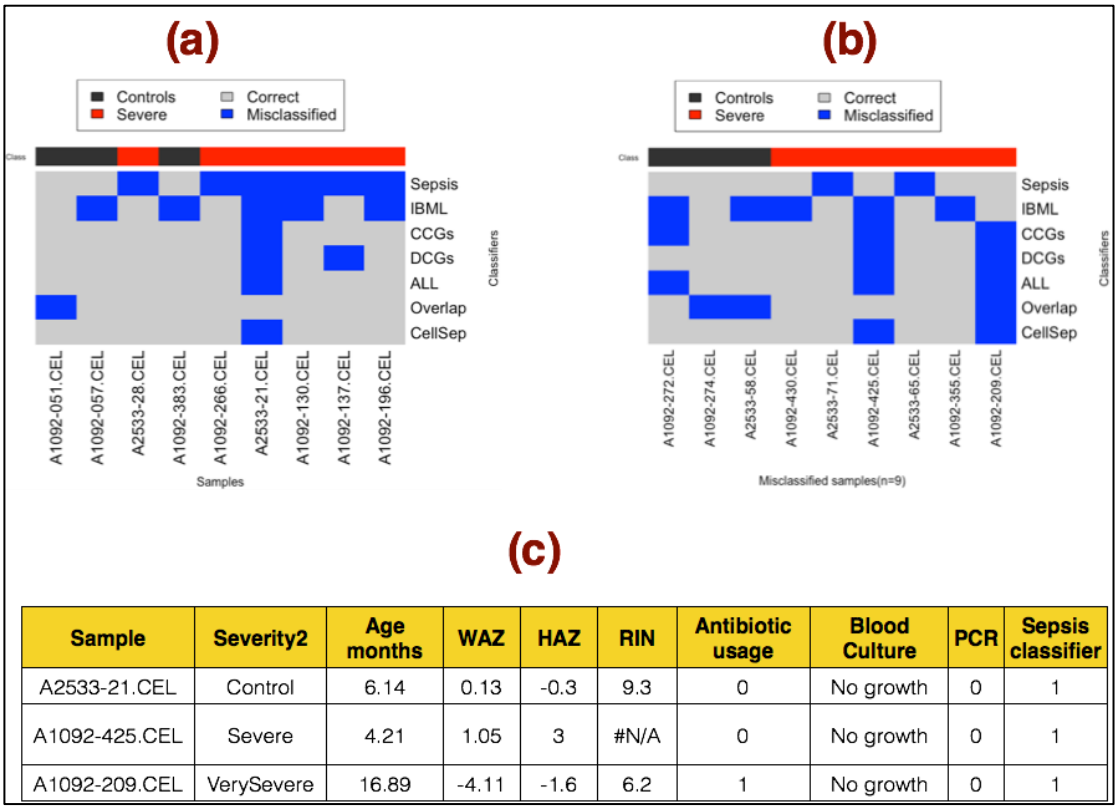


Figure 9.6:The T cell receptor signaling pathway KEGG map (hsa04660) showing differentially expressed genes in pneumonia (FDR<0.05). Each coloured box is divided into three sections representing mild (left), severe (middle) and very severe (right) pneumonia states. Colours: (i) White=Gene not analysed, (ii) Grey=Gene not significant (FDR>0.05), (iii) Green=Down-regulated genes and (iv) Red=up-regulated gene. The pathway map was produced using the PathView Web (https://pathview.uncc.edu/)

9.3 Appendix C (Chapter6): Misclassified samples in biomarker analysis.



(c)

Sample	Severity2	Age months	WAZ	HAZ	RIN	Antibiotic usage	Blood Culture	PCR	Sepsis classifier
A2533-21.CEL	Control	6.14	0.13	-0.3	9.3	0	No growth	0	1
A1092-425.CEL	Severe	4.21	1.05	3	#N/A	0	No growth	0	1
A1092-209.CEL	VerySevere	16.89	-4.11	-1.6	6.2	1	No growth	0	1

Figure 9.7: Misclassified samples in the training (a) and the validation (b) datasets across the biomarker sets. In total, three samples were misclassified by the final biomarker set (**CellSep**) in the training (n=1) and validation (n=2) data sets. In the training data, a non-pneumonia control sample (A2533-21.CEL), which was associated with bacterial septicaemia, was misclassified (**c**). In the validation data set, the misclassified samples (A1092-425.CEL and A1092-209.CEL) were associated with overall poor sample quality as measured by the RNA integrity number (RIN) was associated with one and two misclassified samples