An integrated metagenomic approach to investigating disease heterogeneity in sepsis due to community acquired pneumonia

Cyndi Goh

Linacre College

University of Oxford

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

Michaelmas Term, 2019

Abstract

An integrated metagenomic approach to investigating disease heterogeneity in sepsis due to community acquired pneumonia

Cyndi Goh, Linacre College, Michaelmas Term, 2019

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

Sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection. It is an increasing global burden associated with high mortality, long-term disability and shortened life expectancy. Clinical management of sepsis remains supportive rather than curative and progress in sepsis research has been severely constrained by a heterogeneous disease phenotype, limiting the interpretation of clinical trials and the development of effective therapeutic interventions. One source of heterogeneity is the pathogen but the frequent failure of clinical microbiology to identify the infecting organism in sepsis has limited efforts to understand the effect of disease heterogeneity involving the pathogen. Community-acquired pneumonia (CAP) is the most common cause of sepsis and clinical microbiology is unable to provide a diagnosis in approximately 60% of cases, suggesting that alternative methods such as clinical metagenomics are required for improved diagnostics.

Clinical metagenomics involves the application of next-generation sequencing technologies to characterise all the DNA and/or RNA present in a sample, enabling analysis of the entire microbiome as well as the human host genome or transcriptome from patient samples. This thesis presents the development and validation of *Castanet*, a method for targeted metagenomic sequencing using probe-based enrichment. Clinical metagenomic data is presented for 573 patients admitted to intensive care with sepsis due to CAP, including 447 patients for whom clinical microbiology did not identify a pathogen. In addition, droplet digital PCR data is presented for the most frequently identified bacteria (*Streptococcus pneumoniae*) and virus (Epstein-Barr virus) in the metagenomic cohort. Finally, this thesis explores how improved resolution of microbiology in the sepsis cohort can be applied to transcriptomic and genomic-based approaches to understand the host response in sepsis. This includes exploration of Epstein-Barr virus reactivation, differential gene expression analysis for different pathogens, and analysis of the association between specific HLA alleles and susceptibility to different pathogens.

This thesis demonstrates the usefulness of integrating metagenomic data with other omic approaches to enable improved understanding of the heterogeneous host response in sepsis, with opportunities for a precision medicine approach.

Acknowledgements

Firstly, I would like to thank my two supervisors Julian Knight and Ellie Barnes. Julian has made the Knight group a very happy place to spend four years of my life and I am grateful to him for his patience and guidance, and always having time for me despite the busyness of running a large group. During a period of illness, I particularly appreciated Ellie's wisdom and support.

Members of the Knight group have also provided me with immense support. Special thanks must go to Katie Burnham who has been endlessly willing to share scripts, troubleshoot problems big and small and point me in the right direction.

I have also been enormously grateful for the care and support I received from Dr Rose Freeman, Dr Evie Kemp, Dr Desi Choi, and Dr Rebecca Knowles Bevis. This DPhil would not have been possible without you.

Finally, thanks must go to my cycling buddies Becca Kearney and Mimi Harrison who have helped me maintain my sanity on two-wheeled trips involving cake. My family, Mum, Wendi and Jeremy have been an endless source of love and support. And of course, Oscar who never fails to cheer me up with licks and cuddles.

Declarations

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

Study recruitment: The Genomic Advances in Sepsis (GAinS) study began recruiting patients in 2005 from 34 intensive care units across the UK. From October 2015, I have been responsible for liaising with research nurses, providing supplies, maintaining sample records, and collecting and processing samples.

Pathogen probes: Probe design was carried out by Azim Ansari. Curation of meningitis pathogen sequences from RefSeq was performed by Ivo Elliott. Curation of additional pneumonia pathogen sequences was performed by myself.

Metagenomics: I carried out the nucleic acid extractions (supervised by Anthony Brown) and library preparation (supervised by Amy Trebes) for the initial two optimisation experiments and 32 metagenomic samples. Nucleic acid extractions for the subsequent samples were carried out by Anthony Brown and George Macintyre. Library preparation for the remainining samples were carried out by Hubert Slawinski and Mariateresa de Cesare. Data processing and analysis was performed in collaboration with Tanya Golubchik.

Cardiac surgery patients: The cardiac surgery patients were recruited by Eduardo Svoren and Bart's and the London NHS Trust.

Axiom microbiome array: Library preparation was performed by Hannah Matten.

Gene expression: RNA sample processing was performed by Yuxin Mi, Alice Allcock, Katie Burnham, Emma Davenport, Jayachandran Radhakrishnan, Narelle Magueri, and Ashley Thorpe. Microarray gene expression data was generated by the Wellcome Sanger Institute (WSI) and Wellcome Centre for Human Genetics (WHG) Core Genomics facilities.

HLA work: probe design was carried out by Azim Ansari. Library preparation for HLA enrichment experiments were performed by Mariateresa de Cesare. DNA samples were processed by myself, Yuxin Mi, Andrew Kwok, Alice Allcock, Katie Burnham, Jayachandran Radhakrishnan, Emma Davenport, Anna Rautanen, and Tara Mills. Genotyping data was generated by the WHG and WSI Core Genomics facilities. Genotyping QC was performed by Emma Davenport and Katie Burnham. SNP2HLA imputation was performed by Justin Whalley.

Submitted Abstracts

Quest for the one true test: enrichment-based metagenomic sequencing in paediatric meningitis and adult sepsis
Oral: Applied Bioinformatics and Public Health Microbiology (2019)
T Golubchik, **C Goh**, MA Ansari, M de Cesare, A Trebes, D Bonsall, A Brown, M Sadarangani, P Piazza, K Jolley, I Elliott, C Ip, H Slawinski, A Coxon, G Meddaugh, P Hutton, CJ Hinds, E Barnes, AJ Pollard, JC Knight, R Bowden

Using *in vivo* eQTL interactions to identify the regulatory drivers of variation in the transcriptomic response to sepsis
Poster: Immunogenomics of Disease: Accelerating to Patient Benefit (2019)
EE Davenport, KL Burnham, **C Goh**, J Radhakrishnan, P Hutton, TC Mills, A Rautanen, AC Gordon, N Soranzo, AVS Hill, CJ Hinds, S Raychaudhuri, JC Knight

Using *in vivo* eQTL interactions to identify the genetic drivers of the transcriptomic response to sepsis
Poster: American Society of Human Genetics (2018)
EE Davenport, KL Burnham, **C Goh**, J Radhakrishnan, P Hutton, TC Mills, A Rautanen, AC Gordon, N Soranzo, AVS Hill, CJ Hinds, S Raychaudhuri, JC Knight

Associated Publications

Enhanced understanding of the host-pathogen interaction in sepsis: new opportunities for omic approaches
The Lancet Respiratory Medicine 2017; 5: 212-23
C Goh, JC Knight

Targeted metagenomic sequencing enhances the identification of pathogens associated with acute infection
Nature Microbiology (under review), bioRxiv (2019) /textithttps://doi.org/10.1101/716902
C Goh, T Golubchik, MA Ansari, M de Cesare, A Trebes, I Elliott, D Bonsall, P Piazza, A Brown, H Slawinski, N Martin, S Defres, MJ Griffiths, JE Bray, MC Maiden, P Hutton, CJ Hinds, T Solomon, E Barnes, AJ Pollard, M Sadarangani, JC Knight, R Bowden

Epstein-Barr virus reactivation in sepsis is associated with an immunosuppressed host transcriptomic endotype
Critical Care (under review)
C Goh, KL Burnham, MA Ansari, M de Cesare, T Golubchik, P Hutton, LE Overend, EE Davenport, CJ Hinds, R Bowden, JC Knight

Contents

List of Figures

List of Tables

Abbreviations

APACHE-II	Acute physiology and chronic health evaluation II
AST	Antimicrobial stewardship team
AUC	Area under the curve
bp	Base pairs
BWA	Burrows-Wheeler Aligner
CAP	Community-acquired pneumonia
cDNA	Complementary deoxyribonucleic acid
CF	Combined with fragmentation
ChiMES	Childhood meningitis and encephalitis study
CI	Confidence interval
CLiMax	Confidence likelihood maximisation
CMV	Cytomegalovirus
CnoF	Combined with no fragmentation
CSF	Cerebrospinal fluid
ddPCR	Digital droplet PCR
DNA	Deoxyribonucleic acid
DNase	Deoxyribonuclease
dsDNA	Double-stranded deoxyribonucleic acid
dsRNA	Double-stranded ribonucleic acid
EBV	Epstein-Barr virus
eCRF	Electronic case record form
EDTA	Ethylenediaminetetraacetic acid
ELISA	Enzyme-linked immunosorbent assay
EPIC	Aetiology of pneumonia in the community
eQTL	Expression quantitative trait loci
ERCC	External ribonucleic acid controls consortium
ESI-MS	Electrospray ionisation mass spectrometry
FDR	False discovery rate
FER	Fes/Fps related tyrosine kinase
GAinS	Genomic advances in sepsis
GWAS	Genome-wide association study
GWLS	Genome-wide linkage study
HBV	Hepatitis B virus
HCV	Hepatitis C virus
HHV6	Human herpesvirus 6
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HR	Hazard ratio
HSV	Herpes simplexvirus
HWE	Hardy-Weinberg equilibrium
ICU	Intensive care unit
IPA	Ingenuity pathway analysis
IRF	Interferon regulatory factor
MAF	Minor allele frequency

MAFFT	Multiple alignment using fast Fourier transform
MALDI-TOF	Matrix-associated laser desorption/ionisation-time of flight
MARS	Molecular diagnosis and risk stratification of sepsis
MHC	Major histocompatibility complex
MiDAS	Microbial detection analysis software
miRNA	Micro ribonucleic acid
MOSAIC	Mechanisms of Severe Acute Influenza Consortium
MR	Misclassification rate
NCBI	National centre for biotechnology information
NGS	Next-generation sequencing
NIBSC	National institute for biological standards and control
NS1	Non-structural protein 1
nt	Nucleotide
PBWT	Positional Burrows-Wheeler Transform
PC	Principal component
PCA	Principal component analysis
PCR	Polymerase chain reaction
pHBV	Hepatitis B virus plasmid
pQTL	Protein quantitative trait loci
PRR	Pattern recognition receptor
REALISM	Reanimation low immune status markers
RF	Random forest
rMLST	Ribosomal multilocus sequence typing
RNA	Ribonucleic acid
ROC	Receiver operating curve
rps	Ribosomal protein subunit
rRNA	Ribosomal ribonucleic acid
RSV	Respiratory syncytial virus
SIRS	Systemic inflammatory response syndrome
SNP	Single nucleotide polymorphism
SOFA	Sequential organ failure assessment
SRS	Sepsis response signature
ssRNA	Single stranded RNA
STOP-HCV	Stratified medicine to optimise the treatment of patients with hepatitis C virus infection
SURPI	Sequence-based ultra-rapid pathogen identification
T1DGC	Type 1 diabetes genetics consortium
TTV	Torque teno virus
UK	United Kingdom
USA	United States of America
VANISH	Vasopressin vs norepinephrine as initial therapy in septic shock
VCA	Viral capsid antigen
VMR	Viral multiplex reference
VPS13A	Vacuolar protein sorting 13 homolog A
vsn	Variance stabilisation and normalisation
VZV	Varicella zoster virus
WHG	Wellcome Centre for Human Genetics

WSI	Wellcome Sanger Institute

1

Introduction

This chapter presents the aims of this thesis, and provides an overview of pre-existing knowledge relevant to these goals

The overall objective of this thesis is to understand the role of microbiological aetiology in contributing to the observed heterogeneous host response in sepsis due to community acquired pneumonia (CAP). Sepsis is defined as life-threatening organ dysfunction caused by a dysregulated host response to infection (Singer et al. 2016). It is a clinically heterogeneous syndrome with the pathogen as one source of heterogeneity. Clinical microbiology fails to provide a diagnosis in approximately 47% of cases (Gupta et al. 2016), suggesting that alternative methods such as clinical metagenomics are required for improved diagnostics. This work is conducted as part of the UK Genomic Advances in Sepsis study (GAinS, http://ukccggains.com).

In this thesis, I will describe the development of a library preparation method suitable for the metagenomic sequencing of both RNA- and DNA-based pathogens from plasma. I will integrate metagenomic, PCR-based and clinical microbiology data to improve microbiological phenotyping in the GAinS CAP

sepsis cohort. Finally, I will apply this improved microbiological phenotyping to transcriptomic and genomic datasets to better characterise disease heterogeneity.

In this chapter, I will summarise the existing literature as it relates to the work to be described in this thesis. First, I will discuss the sources of sepsis heterogeneity and how this is limiting progression in sepsis research. Next, I will describe how better understanding of the host-pathogen interaction can substantially enhance, and in turn benefit from, current and future application of omics-based approaches to understand the host response in sepsis. I will discuss how clinical metagenomics provides new opportunities in clinical microbiology and describe the applications of clinical metagenomics. Finally, I will detail the specific aims of this thesis.

1.1 Sepsis heterogeneity

The increasing global burden of sepsis is associated with an unacceptably high mortality rate (Vincent et al. 2014), long-term disability (Barnato et al. 2011) and shortened life expectancy (Cuthbertson et al. 2013). Although the importance of early recognition and prompt treatment cannot be overemphasised (Seymour et al. 2017), there remains a significant cohort of sepsis patients who progress to develop a severely dysregulated host response, with failure of vital organs, for whom supportive care in an intensive care unit (ICU) cannot prevent death. For such patients, it has been postulated that some form of adjunctive therapy designed, for example, to modulate the immune response, might improve outcomes. Disappointingly however, multiple clinical trials of such therapies have been uniformly negative (Marshall 2014). Contributory factors include the limitations of animal models (Buras et al. 2005), suboptimal clinical trials, the poorly understood pathophysiology of this ill-defined clinical syndrome and, perhaps most importantly, the substantial variation in the sepsis response both

within and between individuals.

1.1.1 Factors influencing heterogeneity

Heterogeneity in sepsis is driven by a combination of factors (Figure 1.1 A) and may manifest as individual differences in the clinical and molecular response to infection (Figure 1.1 B).

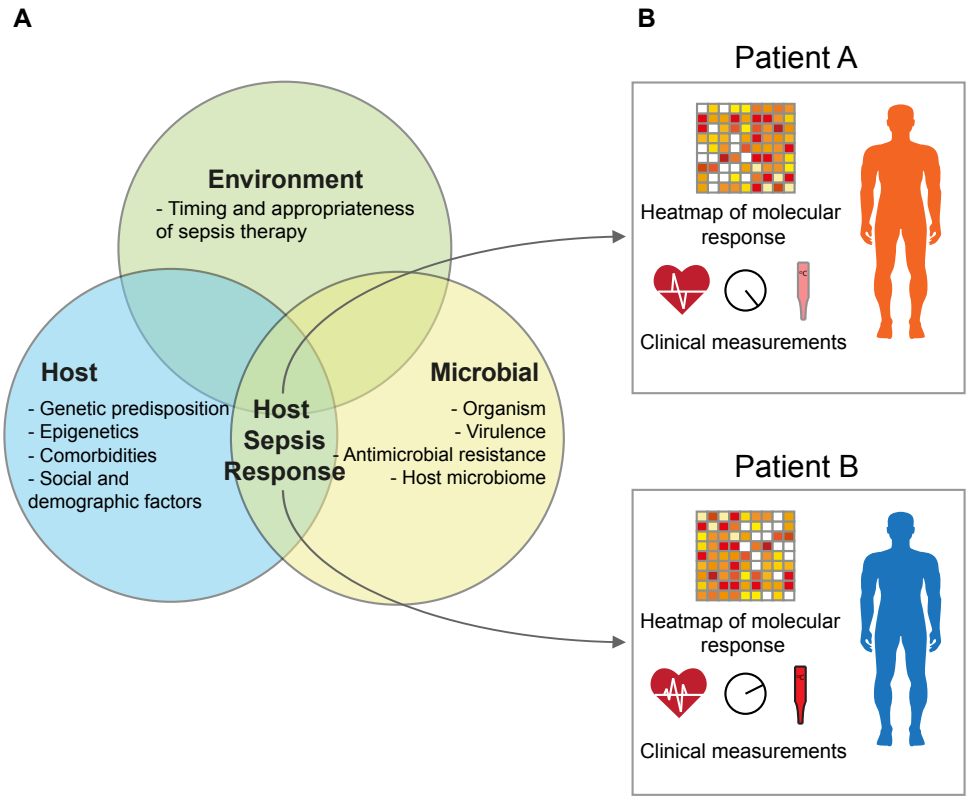


Figure 1.1: Heterogeneity in sepsis. Heterogeneity in the host response to sepsis can be resolved into clinically relevant endotypes. **A.** An interaction between host, environmental, and microbial factors leads to heterogeneity in the individual host response. **B.** This heterogeneity may be observed in the molecular and/or physiological response to sepsis, and may be resolved into patient subgroups or endotypes with clinically relevant and distinct disease phenotypes.

1.1.2 Failure of clinical trials

There has been overwhelming failure of over 100 Phase II and Phase III randomised clinical trials of therapies targeting the dysregulated inflammatory response in sepsis (Marshall 2014). These include drugs which:

1. Non-selectively suppress inflammation (e.g. ibuprofen (Bernard et al. 1997))
2. Neutralize microbial products (e.g. human anti-endotoxin monoclonal antibody (McCloskey et al. 1994))
3. Neutralize host inflammatory mediators (e.g. tumour necrosis factor receptor antagonist (Cohen and Carlet 1996))
4. Non-selectively target inflammatory mediators (e.g. intravenous immunoglobulins (Werdan et al. 2007))
5. Stimulate immune function (e.g. granulocyte colony-stimulating factor (Marti-Carvajal et al. 2012))
6. Have anticoagulant function (e.g. activated protein C (Root et al. 2003))

Given the substantial heterogeneity seen in sepsis at both molecular and physiological levels, it is unsurprising that these trials have failed. All have used simple physiological parameters to identify the at-risk population and studied a reduction in mortality rate as a primary outcome.

1.1.3 Precision medicine approaches

"Precision medicine" has been defined as *"treatments targeted to the needs of individual patients on the basis of genetic, biomarker, phenotypic or psychosocial charaacteristics that distinguish a given patient from other patients with similar clinical presentations"* (Jameson and Longo 2015).

In the context of sepsis, several examples indicate how a precision medicine approach may hold potential for benefit in clinical trials of sepsis therapy. In the example of drotrecogin alfa (recombinant human activated protein C), underlying host genetics may influence treatment outcome. A genome-wide association study (GWAS) of treatment response in sepsis patients receiving drotrecogin alfa identified several single nucleotide polymorphism (SNP) combinations in biologically relevant genes which were associated with a positive treatment outcome (Man et al. 2013). In the top combination of three SNPs (observed in 26% of the cohort), an large absolute risk reduction in 28-day mortality of 41.7% was observed (compared with 6.1% in the whole cohort). This result needs to be treated cautiously given the absence of a replication cohort. Nevertheless. the underlying principle is an important one; subgroups of individuals with certain genotypes might benefit from therapies that have been evaluated as non-efficacious in a heterogeneous cohort.

A further example comes in the form of a post-hoc analysis of the Vasopressin vs Norepinephrine as Initial Therapy in Septic Shock (VANISH) trial (Antcliffe et al. 2019). In this study, patients were randomised to receive norepinephrine or vasopressin followed by hydrocortisone or placebo. Using a gene set defined *a priori* (Davenport et al. 2016), patients were also categorized into two groups based on their peripheral blood leukocyte gene expression pattern, SRS1 (sepsis response signature 1) and SRS2. The authors observed an interaction between assignment to hydrocortisone or placebo, and SRS endotype (p=0.02) whereby hydrocortisone administration was associated with increased mortality in those with an SRS2 phenotype (odds ratio = 7.9; 95% CI 1.6-39.9). This result is biologically plausible as individuals with the SRS2 phenotype were relatively immunocompetent compared to the SRS1 phenotype; abolishing this immunocompetent phenotype through the administration of the corticosteroid is likely to have been harmful. Again, this result illustrates how a precision

medicine approach may be beneficial in trials of sepsis therapy by using transcriptomic signatures to select individuals who might benefit from a particular treatment.

1.2 Omics-based approaches and the importance of the pathogen

Omics-based methods have applied high-throughput techniques to enable understanding of the molecular mechanisms of sepsis from gene level to biological phenotype. Genomics, transcriptomics, epigenomics, proteomics, metabolomics, and metagenomics have all contributed to our understanding of sepsis pathophysiology and disease heterogeneity.

In this section, I will illustrate how omics-based approaches as applied to sepsis will remain limited while studies using such approaches focus on the host to the exclusion of the pathogen. Examples from sepsis and other infectious diseases will be used to illustrate where efforts to account for the degree of heterogeneity involving the pathogen have led to valuable insights into host biology.

1.2.1 Genomics

Three decades ago, a seminal epidemiological study (Sorensen et al. 1988) documented the surprising observation that premature mortality from severe infection is strongly heritable. Since then, a number of genomic approaches have been applied to studying the association between host genetics and sepsis susceptibility or disease outcome. Many of the results have been conflicting or underwhelming; a systematic review (Clark and Baudouin 2006) of 76 candidate gene studies in sepsis assessed the majority of studies to be of low to moderate quality. Using Bayesian statistics, the authors calculated that 30 of the 32 SNPs

associated with sepsis had at least a 50% probability of being a false-positive finding if the prior probability of a true association was set at a liberal cutoff of 0.01.

Currently, there are only two reported GWAS of sepsis in the literature (Rautanen et al. 2015) (Scherag et al. 2016); each of the studies identified different associations with 28-day survival. In the first study (Rautanen et al. 2015), the only genome-wide significant association involved a SNP (rs4957796) located in an intron of the Fps/Fes related tyrosine kinase *(FER)* gene on chromosome 5. Interestingly, this association reached significance only when cause of sepsis and microbiology was accounted for. Statistical significance was reached only when patients with sepsis from pneumonia were analysed independently from those with intra-abdominal infection (p combined=$5.6x10^{-8}$), with effect size increasing when bacterial cases of pneumonia were analysed separately (odds ratio reduced from 0.56 to 0.40, indicating increased effect size). *FER* is implicated in sepsis-relevant pathways involving neutrophil chemotaxis and endothelial permeability; electroporation-mediated gene therapy of a plasmid containing the protective SNP improves survival in murine models of traumatic lung contusion and pneumonia (Dolgachev et al. 2016). However, two independent studies (a second genome-wide association study (Scherag et al. 2016) and a targeted validation study (Schoneweck et al. 2015)) have failed to replicate this association involving the *FER* gene.

The second GWAS (Scherag et al. 2016) analysed a wide range of infection types (lung, abdominal, urogenital, bone, soft tissue, and wound infections) and identified a different association (rs117983287) localising to the vacuolar protein sorting 13 homolog A *(VPS13A)* gene on chromosome 9 (p=$8.16x10^{-8}$). Failure of replication between these genome-wide association studies, despite similar methods and outcome of interest, is potentially attributable to the different causes of sepsis investigated in each study.

1.2.2 Transcriptomics

Transcriptomic studies in sepsis have aimed to define disease subphenotypes with distinct host response features. These include the two GAinS Sepsis Response Signature (SRS) endotypes (Davenport et al. 2016) and the four Molecular diagnosis and risk stratification of sepsis (MARS) Mars endotypes (Scicluna et al. 2017) (see Section 5.1.2). However, challenges in dissecting the various sources of disease heterogeneity have limited efforts to apply these studies to sepsis biomarker development.

Studies comparing host transcriptomic signatures between different classes of microbiology have yielded varying results. One study comparing neutrophil gene expression in ICU patients with Gram-positive and Gram-negative sepsis failed to identify significant differences (Tang et al. 2008). This result is unsurprising given the small patient numbers (18 Gram-positive patients, 25 Gram-negative patients) and the diverse range of sepsis causes studied. For example, the Gram-negative group included many infection types (respiratory, intra-abdominal, urinary, and central nervous system) from ten different bacterial species. A unique transcriptome profile may have been observed if the authors had refined their analysis to a particular infection (e.g. respiratory) or pathogen type in a larger cohort of patients.

A similar approach applied genome-wide gene expression profiling to peripheral whole blood in sepsis caused by acute respiratory illness. Three gene sets were defined that classified patients into bacterial, viral, co-infected, and non-infected causes with a classification accuracy superior to that of procalcitonin (87% vs 78%; $p<0.03$) (Tsalik et al. 2016).

Cheng and colleagues (Cheng et al. 2016) studied critically ill patients with sepsis who had proven *Escherichia coli* bacteraemia and candida fungaemia. Analysis of peripheral blood gene expression microarray data showed that the

majority of genes (5977) with significant differential expression from healthy volunteers (adjusted $p<0.05$) were common to both pathologies. Genes encoding proteins within pathways relevant to glycolysis and oxidative metabolism were upregulated in both groups. However, a unique transcriptional response specific to each pathogen was also seen (*E. coli* bacteraemia, 1718 genes; candida fungaemia 830 genes).

Multi-cohort analysis of publicly available transcriptomic data from patients with acute infection provides further evidence for distinct gene expression patterns depending on the pathogen (Sweeney et al. 2016). Analysis of datasets from eight cohorts of patients enabled derivation of a seven-gene set, which discriminated bacterial from viral infection in 30 independent cohorts. In combination with a separate 11-gene set that discriminates infection from no infection (Sweeney et al. 2015), the authors derived an antibiotics decision model that had 94% sensitivity and 59.8% specificity for bacterial infection.

All these studies highlight the importance of accounting for underlying microbiological heterogeneity in studies of host gene expression in sepsis, such that any differences between study groups can be clearly distinguished from the signature attributable to the pathogen.

1.2.3 Regulation of gene expression

Regulation of gene expression is dynamic, involving highly coordinated processes from transcription to post-translational stages. An extensive review of published GWAS identified that the majority of trait-associated and disease-associated SNPs were located in non-protein coding regions of the genome, with 88% situated in intergenic or intronic regions (Hindorff et al. 2009). These genetic variants are believed to contribute to the heritability of inter-individual variation in gene expression. In fact, SNPs associated with complex traits are

more likely to be expression quantitative loci (eQTL) than frequency-matched SNPs (Nicolae et al. 2010). Given that eQTL show strong heritability (Wright et al. 2014), these regulatory genetic variants could reflect the major selective pressure that pathogens have exerted on human genetics.

Quantitative trait loci mapping approaches define the association of a SNP with an intermediate phenotype (e.g. levels of a transcript (eQTL), protein (pQTL), or metabolite). These approaches have convincingly shown the importance of cellular context and environment on the activity of a regulatory variant. For example, in-vitro studies of human monocytes exposed to biological stimuli relevant to bacterial (lipopolysaccharide) and mycobacterial or viral (interferon-γ) infections have shown that exposure to these stimuli is necessary to show the association of SNPs with the expression of approximately 50% of innate immune system-related genes (Fairfax et al. 2014). In some examples, the direction of effect on gene expression (i.e. an increase or decrease in gene expression) can even differ between different conditions. A similar approach applied to human primary dendritic cells exposed to mycobacterium tuberculosis identified 198 loci associated with gene expression that were not observed in unstimulated cells, reflecting the effect of the host-pathogen interaction (Barreiro et al. 2012).

These observations are highly relevant to studies of sepsis patients, where we must consider the biological context set by a particular infecting pathogen and its influence on underlying host genomic modulation of gene expression. In a study of GAinS ICU patients with sepsis from CAP (Davenport et al. 2016) there was evidence of eQTL in genes associated with viral respiratory infection. This was observed even though only 9% of patients within this cohort had CAP of confirmed viral cause; this analysis would have been enhanced by more detailed microbiological phenotyping of the cohort.

The relevance of non-coding RNAs in the regulation of gene expression has become increasingly recognised. MicroRNAs (miRNAs) have been of particular

interest as potential biomarkers in sepsis. However, conflicting results pose a challenge for the interpretation and application of these studies. For example, significantly reduced miR-223 levels were observed in one cohort of patients with sepsis, compared with patients with systemic inflammatory reponse syndrome and healthy individuals (Wang et al. 2010), while elevated miR-223 levels were observed in another cohort of patients with sepsis compared with healthy individuals (Wang et al. 2012). In a third study (Benz et al. 2015), no differences in miR-223 levels were found between individuals with and without sepsis. The authors of the third study attribute these conflicting results to important differences in methods of normalisation. However, differences in cause of infection deserve consideration. The first of these three studies included only ICU surgical patients with intra-abdominal and trauma-related infection, whereas the other two recruited patients with respiratory infection and a broader range of infective pathologies. Even within the context of different strains of the same virus, significant differences in miR-223 expression can be observed. In mice infected with a reconstructed pandemic 1918 H1N1 influenza A virus, distinct patterns of lung tissue miRNA expression were observed compared with those infected with a seasonal H1N1 influenza A virus (strain A/Texas/36/91) (Li et al. 2010). Specifically, levels of miR-223 were nearly three-times higher with the pandemic strain than with the seasonal strain.

1.2.4 Proteomics and metabolomics

Advances in mass-spectrometry-based methods enables large-scale analysis of the protein and metabolite composition of biological samples in sepsis research. There is the potential for unique functional insights beyond those afforded by genomics and transcriptomics since there is no direct association between mRNA expression and protein or metabolite levels.

Langley and colleagues (Langley et al. 2013) studied patients presenting to

the emergency department with suspected community-acquired sepsis. Unique plasma proteomic and metabolomic profiles were seen in the three groups of patients: sepsis survivors, sepsis non-survivors, and individuals with a non-infective systemic inflammatory response syndrome. Non-survivors had profiles indicative of impaired mitochondrial fatty acid β oxdiation. The authors of this study developed a prognostic logistic regression model based on seven parameters (including four carnitine esters, lactate, age, and haematocrit). This model was validated in two separate cohorts and successfully classified sepsis survivors and non-survivors with an accuracy of 85%, superior to that of using lactate, Sequential Organ Failure Assessment, or Acute Physiology and Chronic Health Evaluation II scores. Interestingly, no substantial differences were seen between the plasma metabolome and proteome of patients with sepsis due to *Streptococcus pneumoniae*, *Staphylococcus aureus* and *Escherichia coli*. The authors postulated that this might have been because of heterogeneity with respect to infection site, and the possibility that subtle differences were overwhelmed by a generalised septic response. Indeed, in a different cohort of patients with Gram-positive and Gram-negative sepsis, ELISA-based quantification of a more limited panel of 11 plasma cytokines showed higher interleukin-1β, interleukin-6 and interleukin-18 concentrations in the Gram-positive group compared with the Gram-negative group (Feezor et al. 2003). In another study (Huang et al. 2014), recruiting a more homegeneous cohort of Gambian children with pneumonia, a mass-spectrometry-based proteomic analysis identified 42 proteins that differentiated severe from non-severe pneumonia and non-severe pneumonia from controls (Huang et al. 2014). One of these proteins was neutrophil gelatinase-associated lipocalin (lipocalin-2), which discriminates pneumonia of probable bacterial cause from viral cause. Children with plasma concentrations of lipocalin-2 more than 163 ng/ml were nine times more likely to have a positive blood culture with a clinically significant isolate. Proteomic and metabolomic studies of sepsis represent a rich, untapped source of potential

disease biomarkers for the clinical setting.

1.2.5 Epigenetics

From the perspective of evolutionary biology, the co-existence of pathogen and host is a major selective pressure on the genetic diversity of both organisms (Sironi et al. 2015). While random point mutations in DNA (SNPs) are a key source of diversity, the role of epigenetic changes in contributing to more rapid changes to organism phenotype is being increasingly recognised (Rando and Verstrepen 2007). Both viruses (Paschos and Allday 2010) and bacteria (Hamon and Cossart 2008) have the potential to induce epigenetic changes in humans, modulating the biological interaction between pathogen and host with significant consequences to disease course.

For example, the H3N2 influenza A virus carries a histone-like sequence in its non-structural protein 1 (NS1) tail. This allows it to interact with the host epigenome; binding of the NS1 protein to the human polymerase associated factor 1 transcription elongation complex allows the virus to target sites of actively transcribed antiviral genes, suppressing the antiviral response (Marazzi et al. 2012). Interestingly, the H1N1 influenza A virus does not possess this histone-like NS1 tail, which might explain the different disease phenotypes seen between various strains.

There are few epigenetics studies of sepsis. However, there is substantial corroborative evidence to support epigenetics as a promising approach to advance our understanding of the disease. In an animal model of acute lung injury and sepsis, anaesthetised mice were exposed to a dual pulmonary insult of aspiration of a *Staphylococcus aureus* culture and mechanical ventilation (Bomsztyk et al. 2015). After only 6 hours, there was decreased expression of the angiogenic genes *ANGPT1*, *TEK*, and *KDR* in the lung, kidney, and

liver. Chromatin immunoprecipitation assays showed a decrease in relative RNA polymerase II abundance and histone deacetylation at these genes, providing evidence for the role of epigenetic changes in contributing to sepsis-induced endothelial dysfunction.

In the previously mentioned study of gene expression in adult ICU patients with sepsis due to CAP (Davenport et al. 2016), the locations of the observed sepsis eQTL showed substantial overlap with epigenetic marks observed in monocytes stimulated with bacterial lipopolysaccharide. This overlap included deoxyribonuclease (DNase) I hypersensitive sites (i.e regions of chromatin where DNase I activity results in the DNA being accessible to transcription factor binding) and histone marks associated with enhancer and promoter regions (i.e. specific covalent modifications such as H3K27ac, H3K4me1, H3K4me3).

In summary, results from studies in animal models and in humans with sepsis suggest a strong role for epigenetic changes in sepsis pathophysiology. However, to translate these changes to patient benefit (e.g. identification of novel therapeutic targets), these epigenetic mechanisms will need to be dissected further. Studies that take account of specific infection types will be an important way to achieve this aim.

1.2.6 Metagenomics

Host and pathogen factors interact to result in the heterogeneous sepsis immune response. However, this host-pathogen interaction does not occur in isolation but within a rich microbial context. The term pathobiome illustrates this concept - that a microorganism's pathogenicity depends not just on its specific virulence factors, but also on host factors, environmental factors, and the microbial community of which it is a part (Vayssier-Taussat et al. 2014).

Advances in metagenomics has been facilitated by the increasing affordability

of high-throughput sequencing. Of relevance to sepsis are the mechanisms by which one microorganism (or a community of microorganisms) can influence the interaction of another microorganism with the host. This synergism between multiple pathogens is well-recognised in influenza virus infection, where adherence of *Streptococcus pneumoniae* to the respiratory epithelium is facilitated by viral neuraminidase (Peltola et al. 2005). Metagenomics now enables the consideration of synergism beyond a handful of organisms, by revealing new mechanisms through which the whole microbiome influences host susceptibility to infection.

Mucosal surfaces represent a critical interface for the microbiome-host interaction; microorganisms previously considered to be pure commensals act both directly and indirectly at sites such as the intestinal and respiratory epithelium to protect the host from potentially pathogenic organisms. Directly, the intact microbiome provides a barrier effect by limiting the supply of essential resources at these sites. Indirectly, the microbiome interacts with immune cells at mucosal surfaces, leading to the modulation of key innate and adaptive immune pathways (Thaiss et al. 2016). In murine models, the bacterial component of the gut microbiome plays an essential part in enabling persistent norovirus infection by limiting the efficacy of interferon-γ mediated innate immunity (Baldridge et al. 2015). Adaptive immunity is also mediated by the microbiome, with the response to influenza virus shown to depend on a healthy lung bacterial microbiome. Mice receiving a 4-week course of oral antibiotics developed marked dysbiosis of the lung microbiome, leading to defective CD4 T-cell, CD8 T-cell, and B-cell mediated immunity to subsequent intranasal influenza virus challenge (Ichinohe et al. 2011). This suggests that antibiotics might be not just unnecessary, but actually harmful in cases of viral respiratory infection with potentially serious implications in patients with chronic lung conditions (e.g. chronic obstructive pulmonary disease or cystic fibrosis), who frequently receive

antibiotics as prophylaxis or for treatment of exacerbations.

Large-scale endeavours such as the National Institutes of Health Human Microbiome Project (Peterson et al. 2009) have revealed the extent to which the microbiome differs between healthy individuals. This diversity is often lost in critical illnesses such as sepsis, in which both disrupted host pathophysiology and clinical interventions to manage the underlying condition result in striking changes to the microbiome (Dickson 2016). There is convincing evidence to suggest that host genetic variants interact with the microbiome to result in the development of immune-related disease; for example, K/BxN transgenic mice do not develop inflammatory arthritis in a germ-free environment since segmented filamentous bacteria in the gut are essential to T-helper-17 cell activation and subsequent joint inflammation (Wu et al. 2010). This could have implications in sepsis, in which the relevance of host genetics to disease might be better understood in the context of inter-individual differences in the microbiome.

There is potential for metagenomic analysis of the microbiome to yield prognostic information in sepsis: in conditions as diverse as bronchiectasis (Rogers et al. 2014) and metabolic syndrome (Le Chatelier et al. 2013), the microbiome has been shown to correlate with disease phenotype and outcome. Metagenome-wide association studies (Zhang et al. 2015b) that identify associations between microbial genes and disease traits have found that restoration of dysbiosis in the dental microbiome is associated with good response to disease-modifying anti-rheumatic drugs in rheumatoid arthritis. It is not inconceivable that microbiome-based tests could similarly be identified as biomarkers for early diagnosis, patient stratification, and evaluation of treatment response in sepsis. Additionally, the microbiome represents an attractive therapeutic target in sepsis, with murine models showing a reduction in circulating six hour *in vivo* aged neutrophils following antibiotic depletion of the gut-microbiota, with corresponding improvements in survival from endotoxin-induced septic shock

(Zhang et al. 2015a). As an important source of heterogeneity in sepsis, characterisation of the microbiome and integration of this second genome (Grice and Segre 2012) into other omics-based approaches should be a key priority in further sepsis research.

1.3 New opportunities in clinical microbiology

A USA-based epidemiological study (Gupta et al. 2016) of nearly 7 million sepsis patients hospitalised between 2001 and 2010 estimated the incidence of culture-negative sepsis at 47%, with culture negativity identified as an independent predictor of mortality. Making a microbiological diagnosis is a key clinical priority, enabling effective antimicrobial therapy within the framework of responsible stewardship. Emerging new technologies bring the potential of more rapid and detailed resolution of microbiology in sepsis, and can be classified into one of three main approaches: PCR-based techniques, mass-spectrometry-based methods, and nucleic acid sequencing (next-generation or high-throughput sequencing) methods.

1.3.1 Multiplex PCR

Multiplex PCR platforms enable the simultaneous detection of multiple pathogens with high sensitivity and specificity. Antimicrobial susceptibility data can be provided and there is potential for automation. Point-of-care platforms are being increasingly commonplace in the clinical setting with the advantage of rapid turnaround times and user-friendly sample processing.

The BioFire FilmArray (bioMerieux, Marcy l'Etoile, France) respiratory (Xu et al. 2013) and meningitis/encephalitis (Lee et al. 2019) panels are two examples of disease specific multiplex PCR platforms. The respiratory panel (Xu et al. 2013)

tests for 15 viral pathogens in respiratory specimens and was implemented in a regional paediatric hospital in the USA with a median turnaround time of 1.4 hours. Over 2,500 specimens were tested by FilmArray and a direct fluorescence assay in parallel. The FilmArray panel detected rhinovirus in 20% of samples and coronavirus in 6% of samples, both organisms were not part of the direct fluorescence assay panel. Whilst the respiratory panel only tests for viruses, the meningitis/encephalitis panel tests for 14 pathogens, including 6 bacteria, 7 viruses, and 1 fungal species from cerebrospinal fluid (Lee et al. 2019). Forty-two individuals presenting to the emergency department with relevant symptoms were tested using the FilmArray platform; six positive samples were detected with an 88% agreement rate with conventional microbiology testing.

There are two main disadvantages with these platforms. Firstly, detection is limited to pathogens in the pre-specified probe panel, which are limited in their scope. Therefore, the platforms are limited in geographical locations where there is a high prevalence of infection caused by microorganisms not included in the panel. Secondly, false-negative results may arise where variation among pathogen genomes leads to failure to detect the organism. For example, a variant strain of *Chlamydia trachomatis* in Sweden led to false negative testing by PCR (Ripa and Nilsson 2006) because the variant strain had a deletion of 377bp in the cryptic plasmid, the region targeted by PCR testing.

1.3.2 Mass spectrometry

There are generally two methods of mass spectrometry applied to microbiological testing: matrix-assisted laser desorption ionisation-time of flight mass spectrometry (MALDI-TOF) and electrospray ionisation-mass spectrometry (ESI-MS) (Buchan and Ledeboer 2014). The difference in the two techniques lies in the generation of the ions. In MALDI-TOF, the analyte is allowed to dry before being overlaid by a weak acid matrix material. It is then exposed to a laser

which ionises and desorbs it from the sample plate. The created ions are then accelerated through a vacuum by the application of an electrostatic field. In contrast, analytes need to be in the liquid phase for ESI-MS. The solute is passed through a heated capillary and exposed to a high voltage to generate an aerosol of ions.

Huang and colleagues compared MALDI-TOF and antimicrobial stewardship team (AST) intervention with routine clinical care in individuals with bloodstream infection (Huang et al. 2013). Routine clinical care included analysis of blood culture results using the VITEK-2 system (biochemical testing for antimicrobial susceptibility information) without AST intervention. In the study group, the AST intervention consisted of provision of evidence-based antimicrobial recommendations after receiving a positive result. The intervention group showed significantly decreased time to organism identification, decreased time to effective antimicrobial therapy, decreased 30-day mortality and decreased ICU length of stay, compared with the pre-intervention control group.

ESI-MS has also been studied in individuals with bloodstream infection (Vincent et al. 2015). In 616 individuals with bloodstream infection, ESI-MS identified a pathogen in 37% of individuals compared with a 11% diagnostic rate for blood cultures. An expert panel was enlisted to provide independent clinical analysis and concluded that ESI-MS could potentially have altered treatment in up to 57% of patients.

1.3.3 Nucleic acid sequencing

See next section (Section 1.4) for a description of short-read nucleic acid sequencing.

1.3.4 Nanopore sequencing

Nanopore sequencing is an emerging NGS technology that performs long-read sequencing with real-time sequence analysis (Jain et al. 2016). Commercially available platforms include the MinION (1 flow cell), GridION (5 flow cell capacity), and PromethION (48 flow cell capacity). The MinION has been successfully applied to outbreak surveillance, e.g. in epidemiological outbreak surveillance of Ebola virus in west Africa (Quick et al. 2016) and Salmonella in a UK-based inpatient setting (Quick et al. 2015).

Examples of application to diagnostics in clinical syndromes remains limited due to inherent challenges; MinION sequencing has been applied to urinary tract infections (Schmidt et al. 2017), prosthetic joint infections (Sanderson et al. 2018), and lower respiratory infections (Charalampous et al. 2019). Schmidt and colleagues (Schmidt et al. 2017) analysed 10 heavily infected urine specimens (>10^7 cfu/ml) and 5 specimens of healthy urine spiked with multi-drug resistant *Escherichia coli*. With a turnaround time of 4 hours, there was 100% agreement between MinION sequencing result, Illlumina sequencing, and clinical microbiology. In addition, 51/55 resistance genes detected by Illumina sequencing were found by minION sequencing. However, the authors noted a number of limitations with their study, including the fact that only heavily infected urine was analysed, the cost of sequencing (only one sample was sequenced per flow cell) and the poor identification of allelic variants and resistance-conferring mutations due to poor base-calling accuracy. Compared to Illumina sequencing, the potential for scalability in terms of number of samples and data volumes remains relatively limited in Nanopore sequencing.

1.4 Clinical metagenomics

Clinical metagenomic next-generation sequencing has been described as the characterisation of *"all DNA and/or RNA present in a sample, enabling analysis of the entire microbiome as well as the human host genome or transcriptome in patient samples"* (Chiu and Miller 2019).

1.4.1 Applications

Infectious disease diagnostics. NGS approaches can be untargeted or targeted in their approach. Untargeted approaches involve the sequencing of the entire DNA and/or RNA component of a sample. Typically, >99% of the reads obtained are human in origin, so sensitivity for a pathogen can be a critical issue. However, advantages include the unbiased nature of the approach, enabling comprehensive analysis of all pathogens in a single assay.

Targeted approaches include the following three approaches: (i) targeted PCR amplification of a conserved region (e.g. 16S ribosomal RNA (rRNA) gene amplification for bacteria (Watanabe et al. 2018), and 18S rRNA and internal transcribed spacer gene amplification for fungi (Wagner et al. 2018)), (ii) targeted PCR amplification of a whole genome using tiled primers (Quick et al. 2017), and (iii) hybrid-capture based techniques whereby metagenomic libraries are subjected to hybridisation using capture "bait" probes (Bonsall et al. 2015). Targeted approaches are able to increase the proportion of pathogen reads in the sequence data, thereby increasing sensitivity with potential reductions in cost of sequencing because less total sequencing reads are required to achieve a given depth of pathogen coverage.

Clinical microbiome analysis. There is increasing awareness of the role of the microbiome in the pathogenesis of both acute and chronic illnesses.

One group charcterised the microbiome of blood in sepsis (n=62) patients and healthy volunteers (n=23) (Gosiewski et al. 2017). The authors observed striking differences between the taxonomic composition of the two groups; the microbiome of healthy volunteers was composed of mainly anaerobic bacteria whereas the microbiome of septic patients was composed of mainly aerobic and microaerophilic microorganisms. There was decreased representation of the Actinobacteria phyla in sepsis patients and decreased representation of the Proteobacteria phyla in healthy volunteers.

Human transcriptomic analysis. Although clinical metagenomics focuses on the microbial reads recovered from a sample, analysis of human gene expression is also possible if RNA libraries have been constructed from the clinical samples. This incidental RNA-seq data has the potential to inform analysis of pathogen reads. For example, infective and non-infective pathologies can be differentiated by combined host and pathogen read data analysis. Clinical metagenomic NGS of tracheal aspirate samples enabled differentiation of individuals with acute respiratory failure from lower respiratory tract infections versus those with non-infectious causes based on pathogen, microbiome diversity, and host gene expression metrics (Langelier et al. 2018b). Also, dual transcriptome analysis enables improved understand of host-pathogen interactions. Analysis of Gambian children infected with *Plasmodium falciparum* revealed distinct human and parasite gene expression profiles associated with severe malaria phenotypes (Lee et al. 2018). Up to 99% of human differential gene expression was driven by differences in parasite load and coexpression analyses revealed interactions between host and parasite, with marked co-regulation of translation genes in severe malaria. Furthermore, RNA-seq can be especially useful in cases of infection where the causative pathogen is only transiently present in the host, e.g. Lyme disease (Marques 2015) or Zika virus infection (Landry and St George 2017), since distinct pathogen-specific human transcriptomic signatures may be

identified.

Applications in oncology. Sequencing of tumour samples can be used to identify viruses associated with cancer (e.g. herpesviruses, papillomaviruses, and polyomaviruses) and also to identify virus-host interactions. For example, metagenomic NGS enabled the identification of a previously unknown polyomavirus, Merkel cell polyomavirus, in skin tissue analysis of Merkel cell carcinoma (a rare but aggressive dermatological cancer) (Feng et al. 2008). In addition, targeted metagenomic sequencing enabled eight whole genomes of Epstein-Barr virus to be sequenced from primary nasopharyngeal carcinoma biopsy specimens (Kwok et al. 2014). Clinical metagenomics may also inform the treatment of cancers associated with viral infection. For example, Kanwal and colleagues (Kanwal et al. 2017) showed that in hepatitis C virus patients treated with direct-acting antiviral agents, those with sustained virological response showed a reduction in the risk of hepatocellular carcinoma.

1.4.2 Hybrid-capture based techniques

Of the three approaches to targeted metagenomics described above, the hybrid-capture based technique is the most versatile as it does not depend on the specificity of highly conserved primers for PCR amplification. The capture procedure is typically applied after nucleic acid extraction and library preparation. Here, RNA baits are preferable to DNA baits as RNA:DNA duplexes hybridise with greater efficiency and stability than DNA:DNA hybrids (Lesnik and Freier 1995). The hybridisation can be carried out on a solid support (i.e. array-based) (Schuenemann et al. 2018), or more commonly, in-solution (Briese et al. 2015). For the in-solution technique , biotinylated baits bind the target of interest (Gnirke et al. 2009). Then, Streptavidin coated magnetic beads are added to the solution, which bind to the biotinylated baits. Subsequent washing steps lead to non-specific unbound molecules being washed away before the sample is

sent for NGS.

Hybrid-capture based techniques have been applied to clinical metagenomic NGS of bacteria (Allicock et al. 2018), viruses (Depledge et al. 2011) (Briese et al. 2015) (Wylie et al. 2015), fungi (Amorim-Vaz et al. 2015), and parasites (Bright et al. 2012).

For bacterial sequencing, the BacCapSeq platform includes a probe set comprised of 4.2 million oligonucleotide probes which enable detection and characterisation of bacteria, virulence determinants and antimicrobial resistance genes (Allicock et al. 2018). The capture resulted in up to a 1000-fold increase in sensitivity with blood samples.

For viral sequencing, the first example of hybrid-capture based NGS involved a study which targeted three full length herpesvirus genomes (Varicella-Zoster Virus, Epstein-Barr virus and Kaposi's sarcoma-associated Herpesvirus) (Depledge et al. 2011). A range of clinical sample types were tested across 13 samples and full length herpesvirus genomes reconstructed at high read depth. Other examples of hybrid-capture based NGS applied to viruses include ViroCap (Wylie et al. 2015) and VirCapSeq-VERT (Briese et al. 2015). Virocap includes a panel of probes designed to enrich for nucleic acid from 34 families of DNA and RNA viruses (190 viral genera and 337 species) that infect vertebrate hosts, excluding human endogenous retroviruses (Wylie et al. 2015). The authors showed that the probes only required a minimum of 58% probe-target homology for enrichment and sequencing. VirCapSeq-VERT includes 2 million probes covering 207 viral taxa that infect vertebrates (Briese et al. 2015). The authors demonstrated that the capture resulted in a 100- to 10,000-fold increase in viral reads from blood and tissue homogenates compared to conventional Illumina sequencing using established virus enrichment procedures (filtration, nuclease treatments, and RiboZero rRNA subtraction). Novel viruses were also detected where their genomes were approximately 60% similar to probes included in the

capture library.

For fungal sequencing, Amorim-Vaz and colleagues (Amorim-Vaz et al. 2015) designed a set of 55,342 probes covering 6,094 *Candida albicans* open reading frames. Results showed approximately 1000-fold enrichment of *C. albicans* reads in biological samples and a detection of more than 86% of its genes.

For parasite sequencing, *Plasmodium vivax* enrichment enabled an increase in sequencing yield from 0.5% to a median of 55%, with 5X coverage across 93% of the *P. vivax* genome (Bright et al. 2012).

1.4.3 Application to sepsis

Apart from the work described in this thesis (Goh et al. 2019), there are several recent examples in the literature of clinical metagenomics as applied to adult sepsis patients. None of these examples include the simultaneous sequencing and analysis of bacterial, DNA viral, and RNA viral reads.

Blauwkamp and colleagues (Blauwkamp et al. 2019) performed microbial cell-free DNA sequencing from the plasma of 348 adult patients presenting to the Emergency Department who met sepsis alert criteria. Next-generation sequencing (NGS) yielded a positive diagnosis in 49% of individuals, compared to an 18% diagnostic yield from blood cultures and 38% diagnostic yield from all conventional microbiological testing combined. The method enabled identification of bacteria, DNA viruses, fungi, and eukaryotic parasites. NGS was particularly helpful in identifying a cause of sepsis in patients who had received preceding antibiotics when compared to conventional microbiological testing. The authors also describe their experience of NGS processing of the first 2000 plasma samples submitted to their Clinical Laboratory Improvement Amendments certified and College of American Pathologists accredited laboratory. They achieved a 98% rate of sample reporting within one day after

sample receipt.

In a smaller study, Grumaz and colleagues (Grumaz et al. 2019) also applied cell-free DNA sequencing to plasma from septic shock patients (n=48). NGS yielded a diagnosis in 72% of individuals compared to a 33% diagnostic rate from blood cultures. A clinical expert panel reviewed the NGS results; 96% of NGS results were deemed plausible and were judged to have led to a change to more adequate therapy in 53% of cases.

In six Vietnamese hospitals, NGS was applied to 492 clinical samples from 386 patients with community acquired sepsis (Anh et al. 2019). A range of sample types were sequenced including serum, nasal/throat swabs, stool and cerebrospinal fluid. Only sequencing reads aligning to viral sequences were analysed. The authors confirmed each positive NGS result with PCR and considered only those validated by PCR as positive. 21 viral species known to be infectious in humans were identified in 13.4% of patients.

Finally, a study of Chinese ICU patients with sepsis (n=78) described the preparation of DNA libraries from plasma and Ion Torrent sequencing (Long et al. 2016). NGS yielded a 31% diagnostic rate versus 13% from blood cultures.

1.4.4 Challenges

There are a number of challenges associated with clinical metagenomics, which will be described in this section.

Sensitivity. As discussed above, typically >99% of metagenomic NGS reads generated originate from the host. This is because the majority of nucleic acid in clinical specimens is human in origin, and also because human genomes are far larger than that of bacteria and viruses. Strategies to increase sensitivity for microbial reads include targeted metagenomic approaches (described above) as well as other purification or enrichment procedures to increase pathogen

yield. For bacterial pathogens, the most common first step is culture, which can favour contaminating, culturable or fast-growing organisms. For viral pathogens and some bacteria, yield-maximising methods include filtration to remove host cells (Allander et al. 2001), sample treatment with nucleases to digest nucleic acid not protected within cells or virions (Allander et al. 2001) (Charalampous et al. 2019), and high-speed gradient centrifugation to concentrate virus particles (Breitbart and Rohwer 2005). Each of these procedures reduces throughput and may bring about bias.

Interpretation of positive results. In the context of a metagenomic positive result, it can be challenging to differentiate infection from colonisation or contamination. In certain samples (e.g. respiratory specimens), the normal microbiota can be rich and complicate interpretation of results. Systematic PCR testing of nasopharyngeal or bronchoalveolar lavage specimens from patients in intensive care, admitted with pneumonia, identified respiratory viruses in over a third of patients, the relevance of which remains uncertain (Choi et al. 2012). One potential solution is the use of spike-in based calibration to enable quantitative analysis through conversion of percentage reads to colony-forming units per milliliter (Stammler et al. 2016). Another approach is to assess the host immune response by simultaneous analysis of the human transcriptome. Langelier and colleagues evaluated a composite metric of immunity genes in haematopoietic cellular transplant patients as a biomarker of active infection, using this to differentiate colonisation from active infection (Langelier et al. 2018a).

The use of controls (healthy individuals as well as no-template controls) can also be a useful strategy, especially for differentiating causative pathogens from extraneous sources of DNA ubiquitous in commonly used reagents used in the extraction and library preparation procedures (Salter et al. 2014). For example, Grumaz and colleagues (Grumaz et al. 2016) assigned a sepsis indicating

quantifier score to each identified microbe in a sample, to indicate the probability of it being a true positive finding (on the basis of the number of reads mapping to the microbe's reference genome in the patient sample, relative to healthy individuals). A further approach is to evaluate the proportion of the genome covered by the reads mapping to an organism; where reads are localised to limited regions of the genome, this is more likely to represent contamination when compared to reads spanning larger regions of the genome.

Turnaround times. Time from sample receipt to data generation can vary, and has been reported to be anywhere from 6 hours to 7 days (average 48 hours) (Simner et al. 2018). These long turnaround times arise because of laborious library preparation stages essential to making a clinical sample suitable for sequencing as well as time taken for the generation of sequence data. However, automated processes are becoming increasingly commonplace and new technologies, such as Oxford Nanopore Technologies, require less pre-processing and provide real-time data analysis (Quick et al. 2016).

Data analysis. A sizeable gap exists between the generation of NGS data and the analysis and processing of this data to generate clinically relevant information. Although bioinformatic pipelines have been implemented in clinical settings (e.g. the sequence-based ultrarapid pathogen identification SURPI pipeline (Wilson et al. 2019)), there is still lack of standardisation of data processing and analysis methods. In addition, bioinformatics packages and pipelines currently require a substantial degree of bioinformatics expertise that is not routinely available in clinical settings.

The absence of well-curated databases also poses a challenge. Draft and partial genomes on NCBI (National Centre for Biotechnology Information) may lead to erroneous results. For example, false positive results may arise from mapping to low complexity regions in the reference sequence or to contaminants from database entries that contain reads to human DNA or sequencing adapters.

Also, false negative results may arise due to incomplete or missing taxonomic representation in databases.

1.5 Specific aims and objectives

The aims of this thesis are to:

- **Develop a library preparation method for metagenomic sequencing of plasma samples (Ch. 3)**
  Commercially available library preparation methods are available for either RNA or DNA, but not both within a single streamlined method. Sensitivity is likely to be a limiting factor in metagenomic sequencing from plasma; hybrid-capture based techniques provide an opportunity to improve sensitivity. I aim to:
  1. Use probe-based enrichment to increase sensitivity for sequencing organisms relevant to sepsis due to CAP from plasma
  2. Optimise a library preparation method suitable for sequencing both DNA- and RNA-based organisms from plasma
  3. Evaluate performance of the library preparation method against a known positive control reference set
- **Improve definition of microbiological aetiology using targeted metagenomics, PCR-based pathogen analysis, and clinical data (Ch. 4)**
  Various methods can be used to improve the microbiological diagnosis rate in GAinS CAP sepsis patients. I aim to improve microbiological classification of GAinS CAP sepsis patients through:
  1. Application of targeted metagenomics to plasma samples
  2. Use of droplet digital PCR to assay *Streptococcus pneumoniae* and Epstein-

Barr virus from plasma samples

3. Use of the Axiom Microbiome Array

- **Improve characterisation of disease heterogeneity in patients with no previous microbiological diagnosis through integrated application of metagenomic, transcriptomic and genomic techniques (Ch. 5)**

  Integrated -omic approaches have the potential to enhance our understanding of the heterogeneous host response in sepsis. Specifically, the most commonly identified bacterial (*Streptococcus pneumoniae*) and viral (influenza) infections will be studied as well as the most commonly reactivated virus (Epstein-Barr Virus). These will be related to Sepsis Response Signature endotype, total leukocyte gene expression and underlying host genotype (HLA type). In this chapter, I aim to:

  1. Characterise the extent and implications of EBV reactivation and integrate this with host transcriptomic data
  2. Investigate the association between *Streptococcus pneumoniae* bacterial load and sepsis endotype (SRS status)
  3. Describe host transcriptomic signatures of viral infection, influenza infection, and *Streptococcus pneumoniae* infection and identify predictive gene sets for these infections
  4. Investigate the association between host genotype (HLA type) and susceptibility to different classes of infection

2

Materials and Methods

This chapter describes the patient cohorts, laboratory methods, and bioinformatic methods used in this thesis

2.1 Genomic Advances in Sepsis

The UK Genomic Advances in Sepsis (GAinS) study (http://ukccggains.com) is a multicentre prospective study initiated in 2005 by the UK Critical Care Genomics group with the original aim of characterising genetic variants that affect susceptibility to and outcomes from sepsis. A bioresource arising from this study includes biological samples and phenotypic information from over 2,000 individuals with sepsis from community acquired pneumonia (CAP) or faecal peritonitis (FP) admitted to intensive care units (ICUs). This thesis focuses

only on the subset of individuals with sepsis from CAP. Recruitment was initially carried out in 34 ICUs and remains ongoing in 4 ICUs across the UK.

2.1.1 GAinS patient recruitment and exclusion criteria

Sepsis patients were recruited through the GAinS study from 34 participating ICUs between 2005 and 2019. Patients were recruited if they met the diagnostic criteria for severe sepsis in use at the time of study initiation (Sepsis-2, 2001 American College of Chest Physicians/Society of Critical Care Medicine consensus definition (Levy et al. 2003)). Sepsis was defined as infection with signs of systemic inflammation 2.1 and classified as severe when associated with organ dysfunction. CAP was defined as febrile illness associated with cough, sputum production, breathlessness, leukocytosis and radiological features of pneumonia acquired prior to or within 48 hours of hospital admission (Lim et al. 2009). Exclusion criteria included immunocompromise, admission for palliative care only, and pregnancy.

Infection, documented or suspected, and some of the following:	
General variables	Fever (core temperature >38.3°C) Hypothermia (core temperature <36°C) Heart rate >90/min or >2 SD above the normal value for age Tachypnoea Altered mental status Significant oedema or positive fluid balance (>20 ml/kg for 24 hours) Hyperglycaemia (plasma glucose >7.7 mol/L) in the absence of diabetes
Inflammatory variables	Leukocytosis (WBC count >12,000/uL) Leukopenia (WBC count <4,000/uL) Normal WBC count with >10% immature forms Plasma C-reactive protein >2 SD above the normal value Plasma procalcitonin >2 SD above the normal value
Haemodynamic variables	Arterial hypotension (SBP <90 mmHg, MAP <70, or an SBP decrease >40 mm Hg in adults or >2 SD below normal for age) SvO_2 >70% Cardiac index >3.5 L/min/M^2
Organ dysfunction variables	Arterial hyperaemia (PaO_2/FiO_2 <300 mm Hg) Acute oliguria (urine output <0.5 ml/kg/h or 45 mmol/L for at least 2 hours) Creatinine increase (>0.5 mg/dL) Coagulation abnormalities (INR >1.5 or APTT >60 seconds) Ileus (absent bowel sounds) Thrombocytopenia (platelet count <100,000/uL) Hyperbilirubinaemia (plasma total bilirubin >4 mg/dL or 70 mmol/L)
Tissue perfusion variables	Hyperlactataemia (>1 mmol/L) Decreased capillary refill or mottling

Table 2.1: Diagnostic criteria for sepsis. WBC=white blood cell; SBP=systolic blood pressure; MAP=mean arterial blood pressure; SvO2=mixed venous oxygen saturation; INR=international normalised ratio; APTT=activated partial thromboplastin time

2.2 Additional cohorts

Two other cohorts were studied: (a) patients with hepatitis C virus infection, and (b) patients undergoing cardiac surgery.

2.2.1 Hepatitis C virus infection patient recruitment and exclusion criteria

The hepatitis C virus (HCV) infection cohort was recruited through the NIHR Oxford Biomedical Research Centre Prospective Cohort Study in Hepatitis C virus infection. This aim of this study was phenotypic and genotypic characterisation of a cohort of patients with current and resolved HCV infection. Adult patients were recruited if HCV infection was confirmed by detectable viral RNA or if there was evidence of spontaneously resolved infection confirmed by the presence of HCV antibodies in the absence of detectable HCV RNA. Exclusion criteria included clinician judgment that the patient was unlikely to participate in a long-term study.

2.2.2 Cardiac surgery patient recruitment and exclusion criteria

Patients undergoing elective cardiac surgery requiring cardiopulmonary bypass (coronary artery bypass grafting, valve replacement, or valve repair) were recruited to the Genomic Advances in Cardiac Surgery study by Dr Eduardo Svoren and Professor Charles Hinds (Bart's and the London NHS Trust). This study aimed to investigate the host inflammatory response induced by elective cardiac surgery involving cardiopulmonary bypass. They were included in this thesis as uninfected negative controls for sepsis. Patients were excluded if they were immunocompromised, undergoing an emergency operation, had malignancy, or were unable to provide informed consent.

2.3 Metagenomics

2.3.1 Nucleic acid extraction

Total nucleic acid extraction was performed using the NucliSENS easyMag platform (Biomerieux). Typically, 500 μl of extracted plasma was eluted in 25 μl of buffer. Postextraction quality control was performed using the Agilent 2100 Bioanalyzer platform and/or the Qubit dsDNA High Sensitivity Assay (Thermo Fisher Scientific) in a subset of samples.

2.3.2 Library preparation methods

Four library prepration methods were evaluated.

1. RNA: We used the NEBNext Ultra Directional RNA Library Preparation Kit for Illumina (New England Biolabs) with several modifications to the manufacturers guidelines including: fragmentation for 4 minutes at 94°C, omission of Actinomycin D at first-strand reverse transcription, library amplification for 15 PCR cycles using custom indexed primers and post-PCR clean-up with 0.85x volume Ampure XP (Beckman Coulter).

2: DNA: The Nextera DNA Library Preparation Kit (Illumina) was used according to the manufacturer's guidelines.

3: Combined with fragmentation (CF): This involved the RNA protocol (NEBNext Ultra Directional RNA Library Preparation) (1) followed by the DNA protocol (Nextera DNA Library Preparation)(2).

4. Combined with no fragmentation (CnoF): This involved the RNA protocol (1), with omission of fragmentation, end repair, and adaptor ligation steps, followed by the DNA protocol (2).

2.3.3 Spike-ins

RNA: We used the Ambion ERCC RNA Spike-In Mix 1 (Thermo Fisher Scientific) consisting of 92 synthetic transcripts between 250-2000 nucleotides in length, at a range of pre-specified concentrations (External RNA Controls Consortium 2005).

DNA: Multiple restriction enzyme digest of three synthetic plasmids was performed according to manufacturer instructions (New England BioLabs) (Table 2.2).

Plasmid	Original size (bp)	Restriction enzymes	Fragment sizes (bp)
pHBV	6820	AccI, AlwNI, HindIII, NdeI	800, 1178, 1652, 3190
p1990	4808	AccI, AlwNI, HindIII	401, 588, 1796, 2023
p2022	3356	AccI, AlwNI, NdeI	379, 1099, 1878

Table 2.2: DNA spike-in controls. Plasmids, restriction enzymes, and resulting fragment sizes.

The three plasmids were pooled in equal mass ratios and spiked-in at 3% sample DNA concentration by mass.

2.3.4 Probe-based enrichment

In collaboration with a paediatric meningitis study, a custom probe panel covering bacterial and viral pathogens relevant to meningitis and pneumonia was designed using the Agilent SureDesign service. This included probes complementary to three ERCC's (External RNA Controls Consortium; ERCC14, ERCC25, ERCC116) and the pHBV plasmid fragment. The probe set included 52,101 120nt RNA oligonucleotide probes (5.87×10^6 bp).

1μg of each indexed pooled library was enriched using the Agilent SureSelect^{XT} Target Enrichment System for Illumina Paired-End Multiplexed Sequencing Library protocol with one major modification to the recommended protocol.

This involved capture on a post-PCR indexed pool with use of oligonucleotide blockers complementary to adapter sequences.

2.3.5 Data processing

De-multiplexed sequence read-pairs were trimmed of adapter sequences using Trimmomatic v0.36 (Bolger et al. 2014). Fastq files containing the trimmed reads were then classified with the metagenomic classifier Kraken v1 (Wood and Salzberg 2014) using a custom database comprised of human, bacterial, viral and fungal genomes. Unclassified reads as well as reads classified as bacterial or viral were aligned using bwa v0.7.12(Li and Durbin 2009) to a multi-fasta reference comprised of consensus sequences corresponding to the enrichment probe set, sequencing contaminants (e.g. *Alteromonas* species) and potential clinical sample contaminants including non-pathogenic *Streptococcus* species.

2.4 Digital droplet PCR

Digital droplet PCR (ddPCR) was performed for targets from several microorganisms: (a) *S. pneumoniae*, (b) influenza, (c) Epstein-Barr virus (EBV), and (d) cytomegalovirus (CMV). The assay was performed on samples following nucleic acid extraction as above. For the RNA-based pathogen (influenza) nucleic acid extraction was followed by first strand cDNA synthesis (SuperScript III First-Strand Synthesis System, Invitrogen).

Sample processing was performed in triplicate (1.5μl per replicate) following the recommended workflow (QX200 ddPCR system, Bio-Rad). Custom-designed PrimeTime (IDT) primer/probe sets targeting the *S. pneumoniae* capsular polysaccharide biosynthesis (*cpsA*), influenza A matrix (M), EBV Epstein-Barr nuclear antigen 1 (*EBNA-1*), and CMV envelope glycoprotein B (*UL55*) genes

were designed based on published sequence data (Park et al. 2010) (Shu et al. 2011) (Ryan et al. 2004) (Sedlak et al. 2014) (Table A.1). A total of 20μl of each reaction mixture was loaded onto a DG8 cartridge (Bio-Rad) with 70μl of droplet generation oil (Bio-Rad) and placed in the QX100 Droplet Generator (Bio-Rad). Droplets were transferred to a 96-well PCR plate and PCR amplification performed on a C1000 Touch Thermal Cycler (Bio-Rad). Following amplification, the plate was loaded onto the QX100 Droplet Reader (Bio-Rad). Data was analysed with the QuantaSoft analysis software.

2.5 Epstein Barr Virus Serology

Enzyme linked immunosorbent assay (ELISA) was used to test for the presence of IgG and IgM antibodies against the EBV viral capsid antigen (VCA) using proprietary kits (Abcam). The manufacturer's instructions were followed. Plasma samples (10μl) were diluted to the recommended 1:100 concentration and run in duplicate. Absorbance was measured at 450nm using the CLARIOstar plate reader (BMG Labtech). Samples were considered to be positive if the absorbance value was greater than 10% over the cut-off control absorbance value.

2.6 Axiom Microbiome Array

Plasma samples from ten individuals were tested on the Axiom Microbiome Array platform (Affymetrix). For each sample, total nucleic acid was extracted from 500μl plasma and 21 out of the 25μl eluted product (equivalent to 420μl plasma) was processed according to the Axiom 2.0 Assay Protocol. Each array includes 1,277,846 target probes and 60,152 random negative control probes. The target probes represent 135,555 sequences from 12,513 microbial species from five domains: archaea, bacteria, fungi, protozoa and viruses. Sample

processing involved parallel processing of samples on a microarray plate with isothermal whole-genome amplification, hybridisation to 35-mer oligonucleotide probes, washing and scanning on the GeneTitan Multi-Channel instrument. Data analysis was performed using the Axiom Microbial Detection Analysis Software (MiDAS) based on a Composite Likelihood Maximisation Method (CLiMax) algorithm. Probes were considered positive if signal intensity exceeded the 99th percentile of the random control probe intensities and if more than 20% of target-specific probes were detected.

2.7 Transcriptomics

2.7.1 Sample collection

Serial samples for RNA were obtained by collecting 5ml blood into Vacuette EDTA tubes (Becton Dickinson). Using a vacutainer system, blood was passed across a LeukoLOCK leukocyte enrichment filter (Ambion), isolating the total blood leukocyte population. The filtered leukocytes were stabilised with RNA*later*. Filters were stored and transported at -80°C. GAinS samples were collected on days 1, 3, and/or 5 of ICU admission. Cardiac samples were collected prior to induction of anaesthesia, immediately post-operative and 24 hours post-operative.

2.7.2 RNA extraction

RNA extraction was performed using the Total RNA Isolation Protocol (Ambion). Purified RNA, depleted of globin mRNA, was extracted from the LeukoLOCK filters. Filter contents were lysed and eluted with a guanidine thiocyanate-based solution. Degradation of cellular proteins and DNA was carried out using Proteinase K and DNase I respectively. Magnetic bead technology was used to

purify the RNA. Spectrophotometry (Nanodrop 2000; Thermo Scientific) was used to quantify the RNA yield and quality of a small subset was evaluated using on-chip electrophoresis (Bioanalyzer; Agilent).

2.7.3 Microarray data processing and analysis

Genome-wide gene expression data was generated on 1000ng RNA using the Illumina Human-HT-12 v4 Expression BeadChip gene expression platform comprising 47,231 probes. The report for analysis was generated by Illumina's Genomestudio software. The four microarray datasets were generated by Dr Jayachandran Radhakrishnan (Radhakrishnan 2010), Dr Emma Davenport (Davenport 2011) and Dr Katie Burnham (Burnham 2014 and Burnham 2016). The first two datasets were generated at the Wellcome Sanger Institute (WSI, Cambridge) whilst the final two datasets were generated at the Wellcome Centre for Human Genetics (WHG, Oxford).

For each dataset, data backgrounds were subtracted and probes with a detection value of <0.95 in >95% of samples were filtered out. The raw data was transformed and normalised using the Variance Stabilisation and Normalisation (vsn) R package (Huber et al. 2002). QC checks including Principal Component Analysis (PCA) was carried out in R to identify batch and array effects. The four individual datasets were then combined and probe filtering repeated using the parameters described above, followed by normalisation using vsn. The ComBat function from the R package sva (Leek et al. 2012) was used to directly estimate and remove the known batch effects.

Differential gene expression was analysed using the R package limma (Ritchie et al. 2015), which fits a generalised linear model to the expression of each gene and uses an empirical Bayes approach to account for overall variance in the dataset. Genes with an FDR <0.05 and ≥1.5 were considered to be

differentially expressed. Pathway enrichment analysis was carried out with XGR (Fang et al. 2016) using the xEnricherGenes function and taking all genes tested for differential expression as the background. Predictive gene signatures were derived using the elastic net method (Zou 2005) (Herberg et al. 2016). This variable selection algorithm combines the lasso and ridge regression methods of shrinkage by minimising the number of variables included (lasso) whilst also making the model less dependent on any single variable (ridge).

2.8 Genomics

2.8.1 DNA extraction

DNA was extracted from buffy coat or whole blood using one of three protocols: (a) Qiagen DNA extraction protocol, (b) Maxwell 16 Blood purification kit (Promega), or (c) QIAamp Blood Midi kit protocol (Qiagen). In the Qiagen protocol, cell lysis is followed by sequential removal of unwanted cell components (e.g. protein and RNA). The Maxwell automated protocol uses paramagnetic particles to purify DNA following cell lysis while the QIAamp kit uses a spin column. DNA yield was determined by spectrophotometry (Nanodrop) or fluoresecence using the Quant-iT PicoGreen kit (Invitrogen).

2.8.2 Genotyping and data processing

Three genome-wide genotyping datasets were generated. The first dataset had previously been generated at the WHG for 295 CAP patients and 63 cardiac surgery patients and 730,525 SNPs using the Illumina HumanOmniExpress BeadChip (Davenport 2014). The second dataset was generated for 655 patients at the WSI using the Infinium CoreExome BeadChip (Illumina; 551,839 SNPs) and the Illuminus genotype calling algorithm. The third dataset was generated

for 307 patients at the WHG using the Infinium Global Screening Array BeadChip (Illumina; 654,027 SNPs).

PLINK was used for the genotyping QC (Anderson et al. 2010). Samples were excluded on the basis of discordant sex information, proportion of missing genotypes >0.02, heterozygosity rate, identity by descent. SNPs were excluded if they had missing data proportion >0.05, minor allele frequency (MAF) <0.01, and Hardy-Weinburg equilibrium (HWE) p $<1x10^{-5}$.

2.8.3 Imputation

Each of the genotyping datasets was imputed independently against the Haplotype Reference Consortium (HRC) release 1.1 panel using the Sanger Imputation Service (McCarthy et al. 2016). Pre-imputation checks were carried out using the HRC Imputation Tool (Will Rayner, WHG; www.well.ox.ac.uk/ wraymer/tools). Genotypes were phased using Eagle 2 (Loh et al. 2016) and imputed using PBWT (Durbin 2014). SNPs with an info score <0.9 were removed.

2.8.4 HLA enrichment

127 libraries previously prepared for metagenomic sequencing using the combined no fragmentation protocol described above were enriched for sequencing using a custom designed set of HLA probes designed by Dr Azim Ansari. The enrichment protocol and sequencing are as described for the metagenomic samples.

2.8.5 HLA assignment

For patients with genotyping data, HLA alleles were imputed using the SNP2HLA package (Jia et al. 2013). Two-digit and four-digit alleles were imputed for the HLA-A, -C, -B, -DRB1, -DQA1, and -DQB1 gene loci within the MHC region on chromosome 6. SNP2HLApackagev1.0.2, Beagle.3.0.4, linkage2beagle2.0 and Plink1.07 were used following recommended parameters with 10 iterations and a marker window size of 1000. The pre-built Type 1 Diabetes Genetics Consortium (T1DGC) reference panel of 5225 European individuals and 8961 binary markers was downloaded along with the SNP2HLA tool and used as a training set for the HLA imputation. As well as the imputed HLA alleles, imputation posterior probabilities were also determined to inform the accuracy of the imputed alleles.

2.9 Statistical analysis

Statistical analysis was carried out in R. Demographic data were compared between groups by χ^2 test for categorical data, t-test for continuous parametric data, Mann-Whitney U-test for continuous non-parametric data, and log rank test for survival. ROC curves were plotted using the pROC R package (Robin et al. 2011).

3

Development of Metagenomic Sequencing Workflow

This chapter describes the development of an enrichment-based metagenomic sequencing workflow for application to plasma samples from individuals with sepsis due to community acquired pneumonia.

3.1 Introduction

We anticipated two main challenges associated with metagenomic sequencing from sepsis plasma samples. Firstly, sensitivity for detecting microbial nucleic acids would be low since human nucleic acids would be present in greater excess, especially with antimicrobial use prior to sample collection. Secondly, the design of a single library preparation method suitable for sequencing both DNA and RNA when almost all examples of library preparation methods in the literature target only one type of nucleic acid. These challenges were addressed through the application of probe-based enrichment and a combined DNA and RNA library preparation workflow, respectively and we termed this probe-based targeted enrichment method *Castanet*.

3.1.1 Probe-based enrichment

Our starting point for targeted metagenomics using oligonucleotide enrichment probes was work done (Bonsall et al. 2015) by the Stratified Medicine to Optimise the Treatment of Patients with Hepatitis C Virus Infection (STOP-HCV) consortium (led by Professor Ellie Barnes, University of Oxford). The genome of Hepatitis C virus (HCV) is highly diverse with strains subdivided into seven genotypes which differ at approximately 30-35% positions across the 9650 nt genome. This makes work on HCV highly applicable its genomic diversity is analogous to the range of diverse bacteria and viruses we would need to target in CAP.

Bonsall and colleagues observed greater than 10^3 fold enrichment in mid-range viral load samples (10^4-10^5 IU/ml). This degree of enrichment enabled an increased depth of sequencing to be achieved, enabling more affordable sequencing through multiplexing of larger numbers of samples for the same amount of sequencing capacity.

In addition, the STOP-HCV group were able to exploit a phenomenon they observed, namely that a 20% divergence between probe and target was tolerated with minimal reduction in enrichment efficiency. They started with a probe panel covering 4 genotypes and augmented it to improve coverage for the four already-included subtypes and six additional subtypes. To do this, a consensus sequence for each HCV subtype was generated. Then, for each whole genome sequence in a reference set, genomic regions with less than 80% identity to a starting panel of probes were identified. For each of these regions, the subtype consensus sequence was considered as a reference if there was $\geq$ 80% similarity. If not, a probe was added to cover that genomic region. In this way, a cost-effective, non-redundant set of probes was generated.

3.1.2 Ribosomal multilocus sequence typing

The ribosomal multilocus sequence typing (rMLST) scheme (Jolley et al. 2012) for combined taxonomy and typing in bacteria is a development from 16S rRNA gene approaches (Woese 1987). This approach indexes variation of the 53 bacterial ribosome protein subunit (*rps*) genes to enable resolution into groups at all taxonomic and most typing levels. This is possible as the *rps* genes are conserved enough to enable taxonomic organisation and yet sufficiently diverse to enable species and type characterisation. As of 23 July 2019, the rMLST database (http://pubmlst.org/rmlst) contained 304,876 genomes and 1,861,484 alleles.

The rMLST scheme can be applied to the generation of an efficient panel of probes which enable species/type characterisation without coverage of the whole bacterial genome. In addition, the targeting of conserved regions means that probes are able to capture bacterial organisms despite intra-species genomic variation.

3.1.3 Library preparation methods for metagenomics

There are few examples in the literature of a single combined library preparation workflow suitable for the metagenomic sequencing of both DNA and RNA. A 2018 review (Forbes et al. 2018) details 65 peer-reviewed studies of diagnostic clinical metagenomics. The majority of studies describe sequencing-based techniques applied only to DNA-based or RNA-based genomes with a few examples describing parallel RNA and DNA workflows (Langelier et al. 2018a), (Salzberg et al. 2016). Only one small study (n=6) (Doan et al. 2016) successfully sequenced reads from DNA and RNA viruses (as well as fungi and parasites) from ocular fluid using a single protocol. This involved nucleic acid extraction, cDNA synthesis, and subsequent processing using the Illumina Nextera DNA Library Prep kit.

Other examples of combined library preparation methods include that described in the VirCapSeq-VERT method (Briese et al. 2015) which involved nucleic acid extraction, cDNA synthesis, fragmentation using ultrasonication, and subsequent processing using the KAPA library prepration kit. However, the authors did not trial this method on actual clinical specimens.

3.1.4 Viral multiplex reference

We evaluated our workflow using a Viral Multiplex Reference (VMR) control (11/242) available through the UK National Institute for Biological Standards and Control (NIBSC). The reagent contains 25 viruses covering a range of genome types (dsDNA, dsRNA, ssRNA+, ssRNA-), sizes (6.8-233.7 kb), envelope types and pre-assayed concentrations (Table B.3). The viruses represent a range of common hazard group 2 human viruses (UK Advisory Committee on Dangerous Pathogens classification) and were propagated in cell culture or by egg passage, or isolated from clinical specimens.

Mee and colleagues (Mee et al. 2016) document a study involving 15 laboratories who were invited to process the VMR control using their own wet-lab and informatics methods. In this study, 6/25 target viruses were detected by all laboratories and two laboratories detected all 25 viruses. We will compare the performance of *Castanet* against these 15 laboratories.

3.1.5 Aims

1. To use probe-based enrichment to increase sensitivity for sequencing organisms relevant to sepsis from CAP
2. To optimise a library preparation method suitable for sequencing both DNA and RNA-based organisms from plasma

3. To evaluate performance of the library preparation with enrichment against a known positive control reference set

3.2 Results

3.2.1 Probe panel development

Development of the probe panel was performed in collaboration with the Childhood Meningitis and Encephalitis Study (ChiMES) group. We compiled a list of viral and bacterial pathogens relevant to paediatric meningitis and adult sepsis from CAP in the UK (Table B.1; Table B.2). We also included several pathogens of current interest during the time of probe set development (Zika virus, Chikungunya virus). Considering the number of distinct entries on our list (116, from 17 virus families and 35 bacterial species) and the criteria for inclusion, we inferred that any omissions of a priori less likely organisms, including relevant fungal or parasite pathogens, would comprise rare (<1% frequency) or novel and therefore unsuspected causes of meningitis or pneumonia and sepsis. We also included probes to 4 spike-in control sequences for methodological evaluation of *Castanet*.

We targeted similar lengths of genomic sequence for each pathogen to achieve a comparable assay sensitivity, optimising the breadth of pathogens we could target and avoiding bias in favour of larger genomes. For each of the viruses, we downloaded from NCBI RefSeq the full set of complete genomes available at 1st August 2015. We constructed genome alignments using MAFFT from which to design the probes. For each of the included herpesviruses, whose genomes exceed 100 kbp, this involved a low-diversity region of 20kb whilst for all other viruses we used the whole genome. For bacterial species, we took advantage of the ribosomal multilocus sequence typing (rMLST) scheme, which targets 53

genes encoding ribosomal proteins present in all bacteria and resolves bacteria to a sub-species level, extracting relevant sequences from the rMLST database on 11 December 2015.

Probe design was carried out by Dr Azim Ansari. In previous work with HCV, it had been observed that sequence capture efficiency is preserved when probe and target sequences differ at up to 20% of positions, and that exploiting sequence similarity to avoid redundancy can make probe design substantially more efficient without sacrificing performance. Accordingly, for each sequence alignment we constructed a tree using pairwise distances, within which we identified clusters such that all sequences were less than 5% divergent from one another. The 5.86 x 10^6 bases of cluster consensus sequences were used to design a panel of 52,101 Agilent SureSelect, 120 nucleotide RNA probes on the complementary strand.

3.2.2 Evaluation of four library prepration methods

We chose to perform this initial library preparation (library prep) development on plasma from patients infected with Hepatitis C Virus due to a previously established workflow with this sample type within the STOP-HCV consortium.

Four different library prep methods were compared in five HCV samples: (i) RNA; (ii) DNA; (iii) Combined with Fragmentation (CF. RNA method followed by DNA method); and (iv) Combined with no Fragmentation (CnoF; DNA method preceded by reverse transcription with random primers). To assess the suitability of each method for DNA and RNA-based pathogens, samples were spiked with an RNA (External RNA Controls Consortium Spike-In Mix, ERCC) and DNA positive control (dsDNA plasmid fragments). Following total nucleic acid extraction, the DNA and RNA content of each sample was assayed (Agilent 2100 Bioanalyser platform). This enabled us to spike in the plasmid DNA and

ERCC RNA and 3% and 1% concentration by mass respectively. There was no enrichment stage in these experiments.

Sequencing yielded a mean total read count of $1.5x10^6$ (range $0.9\text{-}2.6x10^6$) across the five HCV samples and four library preps with the majority of sequences aligning to the human reference genome (mean 95.6%; range 91.5-99.1%). Libraries following RNA prep were significantly lower in cDNA concentration, requiring volumes for equimolar pooling in excess of the other three methods by an average of 16-fold. Unsurprisingly, following the DNA prep method, no reads aligning to RNA sequences (HCV or ERCC) were identified in any sample (Figure 3.1). These observations indicate that neither the RNA nor DNA prep in isolation would be suitable for the sequencing of sepsis samples. Thus, the remainder of this chapter concentrates on comparing the two combined library preps.

HCV yield was highest in the CnoF prep, superior even to the standard RNA library prep method used by the STOP-HCV group (Figure 3.1c and d). The CnoF prep also yielded a higher percentage of reads mapping to ERCC than the other three methods (Figure 3.1a). However, the CF prep was superior to the CnoF prep in yield of DNA plasmid reads (Figure 3.1b).

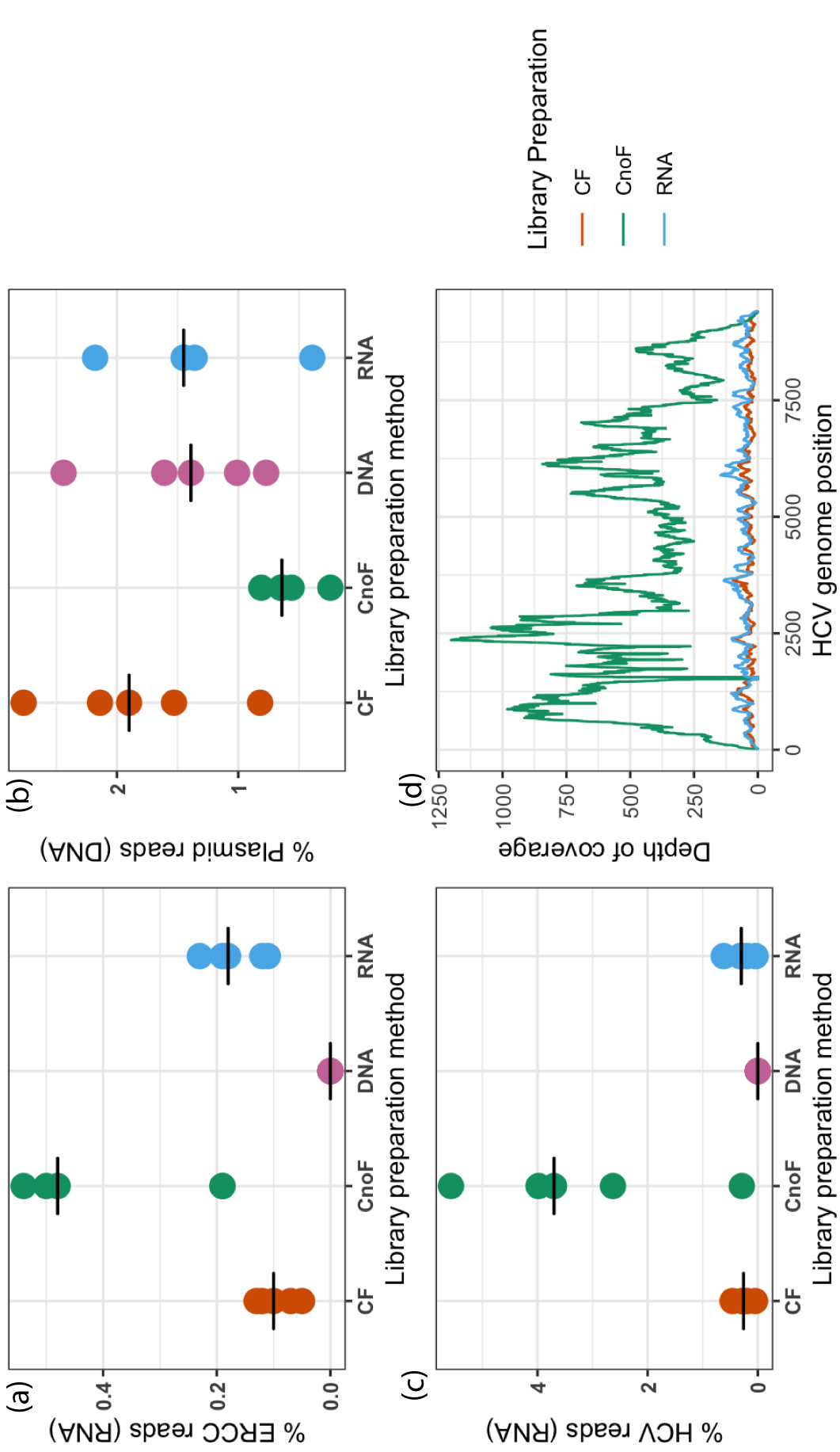


Figure 3.1: Comparison of four library preparation methods Plasma samples from five patients with HCV infection were processed by four library preparation methods in parallel. Performance of the different library preparation methods was evaluated with regards to: (a) ERCC RNA yield; (b) Plasmid DNA yield; (c) HCV RNA yield; (d) HCV genome coverage. (ERCC=External RNA Controls Consortium Spike-in Mix; CF=Combined with Fragmentation; CnoF=Combined with no Fragmentation; HCV=Hepatitis C Virus

For each sample, we calculated the input DNA:RNA spike-in ratio and compared this against the ratio of DNA:RNA (plasmid:ERCC) reads recovered (Table 3.1). We observed that the yield of DNA:RNA reads closely matched the input spike-in concentrations of DNA:RNA with the CnoF prep, indicating that this method was similarly efficient for recovery of both DNA and RNA when compared to the CF prep.

		HCV1	HCV2	HCV3	HCV4	HCV5
Input (Plasmid:ERCC mass)	Both	1.5	5.9	1.4	1.8	1.8
Output (Plasmid:ERCC reads)	CF	12.0	53.6	11.0	21.3	16.4
	CnoF	1.1	4.2	1.3	1.3	1.2

Table 3.1: Sequencing yield of plasmid:ERCC reads relative to plasmid:ERCC spike-in mass. The combined with no fragmentation (CnoF) and combined with fragmentation (CF) library preps are compared in plasma samples from five patients with HCV infection (HCV1 to HCV5)

3.2.3 Evaluation of the Combined no Fragmentation protocol

A subsequent experiment was performed to further evaluate the CnoF library prep protocol (Figure 3.2), with the following specific aims: (i) to compare relative yield of reads originating from RNA and DNA; (ii) to evaluate the relationship of sequencing yield with read length and concentration for RNA and DNA; and (iii) to trial the CnoF protocol in sepsis patients. The sepsis samples were obtained from GAinS patients with sepsis due to CAP.

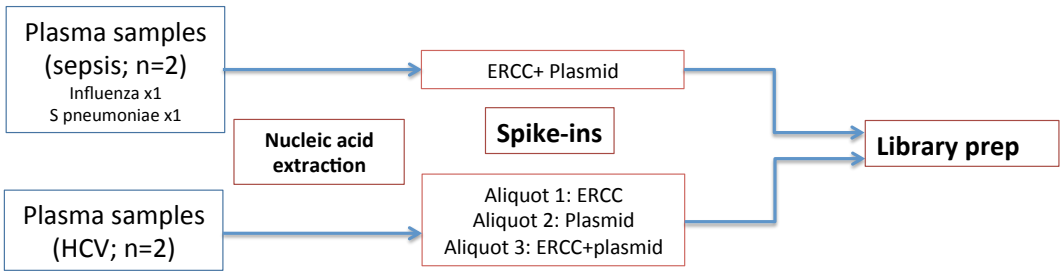


Figure 3.2: Workflow for CnoF Evaluation Experiment. Each HCV sample was divided into three aliquots following nucleic acid extraction and spiked with (i) ERCC only, (ii) plasmid only, or (iii) both ERCC and plasmid. Each sepsis sample was spiked with both ERCC and plasmid.

Sequencing yielded a mean total read count of 24.8×10^6 with minimal variation between HCV and sepsis samples (range $22.7\text{-}28.3 \times 10^6$) (Table 3.2).

The proportion of reads mapping to the ERCC and plasmid references remained consistent whether the controls were spiked in individually or in combination, confirming that RNA was not impacting recovery of DNA or vice versa (Figure 3.3).

The plasmid and ERCC controls were spiked-in at 1% by mass of the initial sample DNA and RNA concentrations respectively. Thus, the relative mass ratio of plasmid to ERCC spike-ins varied between the samples. However, the ratio of plasmid to ERCC reads yielded reflected the input mass ratio fairly consistently between samples, indicating minimal bias towards RNA or DNA (Table 3.3).

	HCV2	HCV4	Sepsis1	Sepsis2
Total reads ($\times 10^6$)	23.9	22.7	25.7	25.1
% Human	95.2	96.4	96.4	99.1
% HCV	1.98	0.44	0	0
% ERCC	0.23	0.27	0.20	0.18
% Plasmid	0.38	0.50	0.23	0.19

Table 3.2: Percentage of reads aligning to each of the relevant reference genomes. Results are for samples with dual ERCC and plasmid spike-ins. In this experiment, plasma samples from two patients with HCV infection (HCV2 and HCV4) and two patients with sepsis due to CAP (Sepsis1 and Sepsis2) was evaluated.

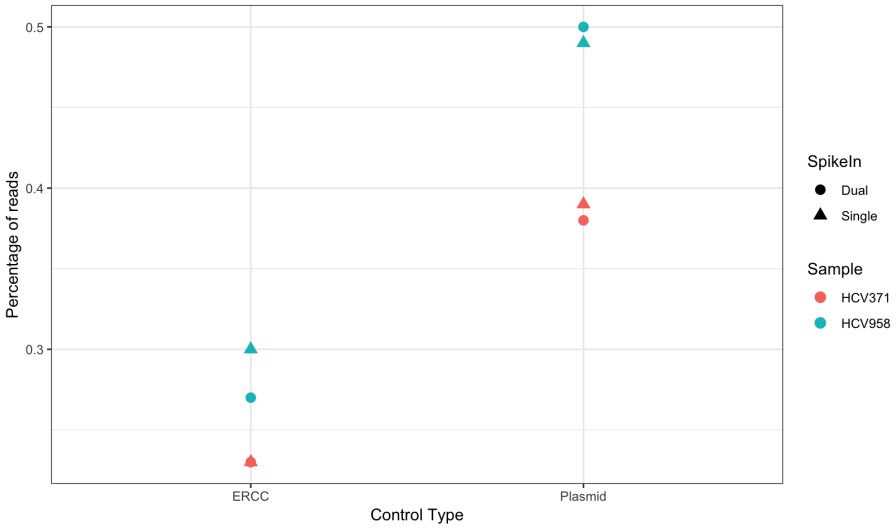


Figure 3.3: Sequencing yield of controls Yield of ERCC and plasmid compared between aliquots receiving single and dual spike-ins

Sample	Input ratio (mass)	Output ratio (reads)
HCV2	1.94	1.65
HCV4	2.40	1.86
Sepsis1	1.12	1.15
Sepsis2	0.70	1.03

Table 3.3: Sequencing yield of plasmid:ERCC (reads) relative to spike-in input (mass). The relative performance of the library preparation method for sequencing RNA and DNA was evaluated by comparing input to output ratios for plasmid:ERCC (DNA:RNA). Plasma samples from two patients with HCV infection (HCV2 and HCV4) and two patients with sepsis due to CAP (Sepsis1 and Sepsis2 was evaluated.

There was no association between sequencing yield and fragment size for either plasmid or ERCC (Figure 3.4), indicating that the CnoF protocol performs consistently across the range of fragment sizes studied (plasmid 379-3190bp; ERCC 250-2000nt).

For the ERCCs, sequencing yield was proportional to input concentration (Figure 3.5). This association was less clear for the plasmid spike-ins. Although sequencing yield was highest for the p1990 plasmid (which was spiked-in at the highest concentration), the yield of p2022 was higher than that of pHBV despite a higher spike-in concentration of pHBV relative to p2022 (Figure 3.5). The differences observed between ERCC and plasmid probably reflect inaccuracies in plasmid nucleic acid quantification rather than differences between RNA and DNA processing in the library prep method.

Finally, we did not observe significant differences in total read count or recovery of spike-ins between plasma from sepsis patients compared to HCV patients.

3.2.4 Evaluation of performance using a Viral Multiplex Reference control

We combined four 1ml aliquots of reagent and made two replicates of a series of five dilutions (neat, 1:10, 1:100, 1:500, 1:1000) in phosphate-buffered saline solution, forming 500ul aliquots for extraction and library preparation.

We used dilutions of a commercially available mixture of viruses (NIBSC Viral Multiplex Reference 11/242) to assess the quantitative range of detection of our method. For two undiluted VMR replicates, *Castanet* sequencing yielded 9.1 and 10.9×10^7 reads. We detected all 21 viruses for which we had enrichment probes, with at least 8.65 reads per million in enriched samples, as well as two viruses not included (Sapovirus and Astrovirus; Figure 3.6). Norovirus GI and GII were not detected; we did not have probes to capture this virus. A likely

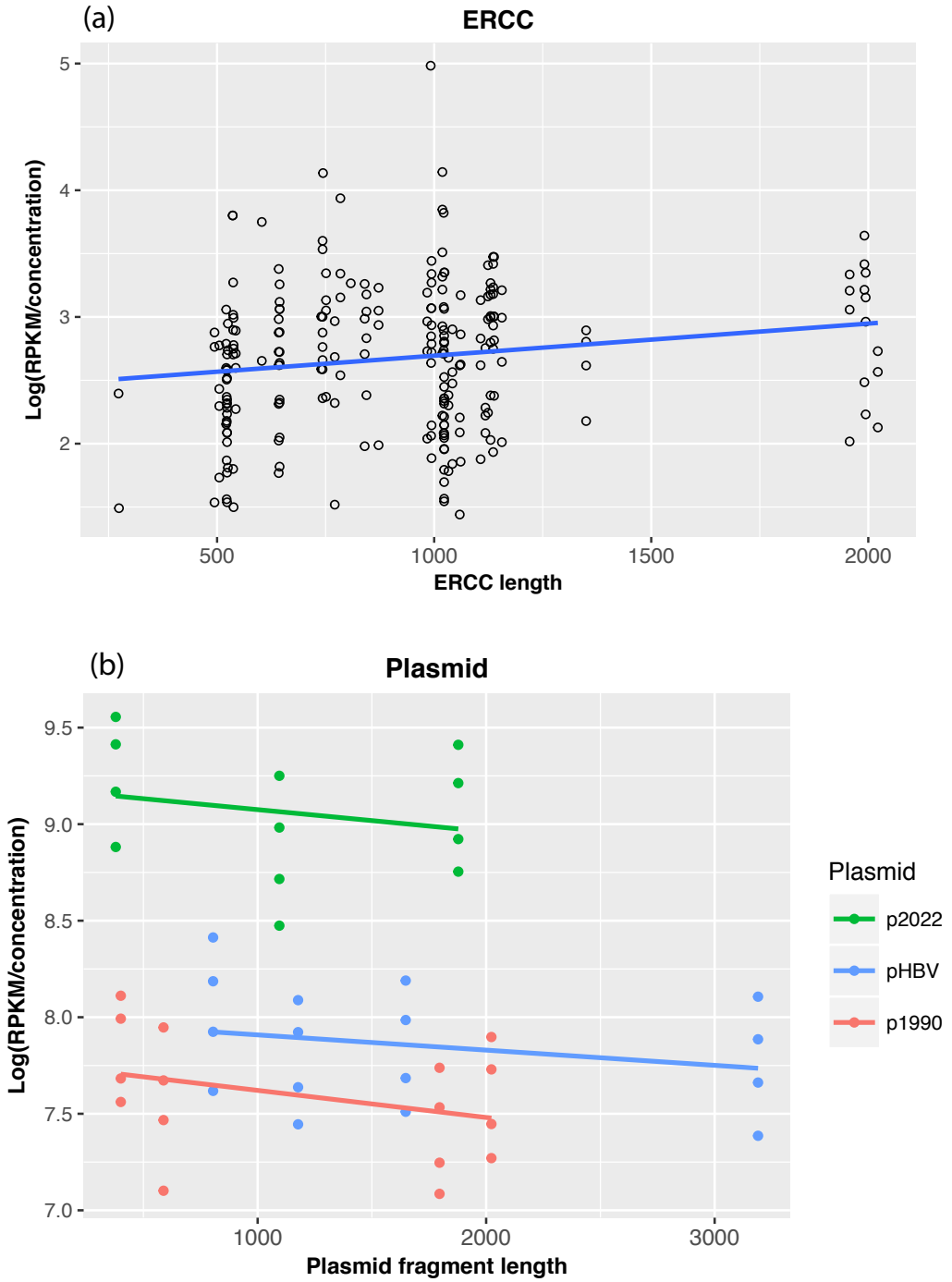


Figure 3.4: Relationship between read yield and fragment length. (a) ERCC, (b) Plasmid. The y-axis displays read count normalised for fragment length, total read count per sample, and fragment concentration. RPKM = reads per kilobase of fragment per million reads.

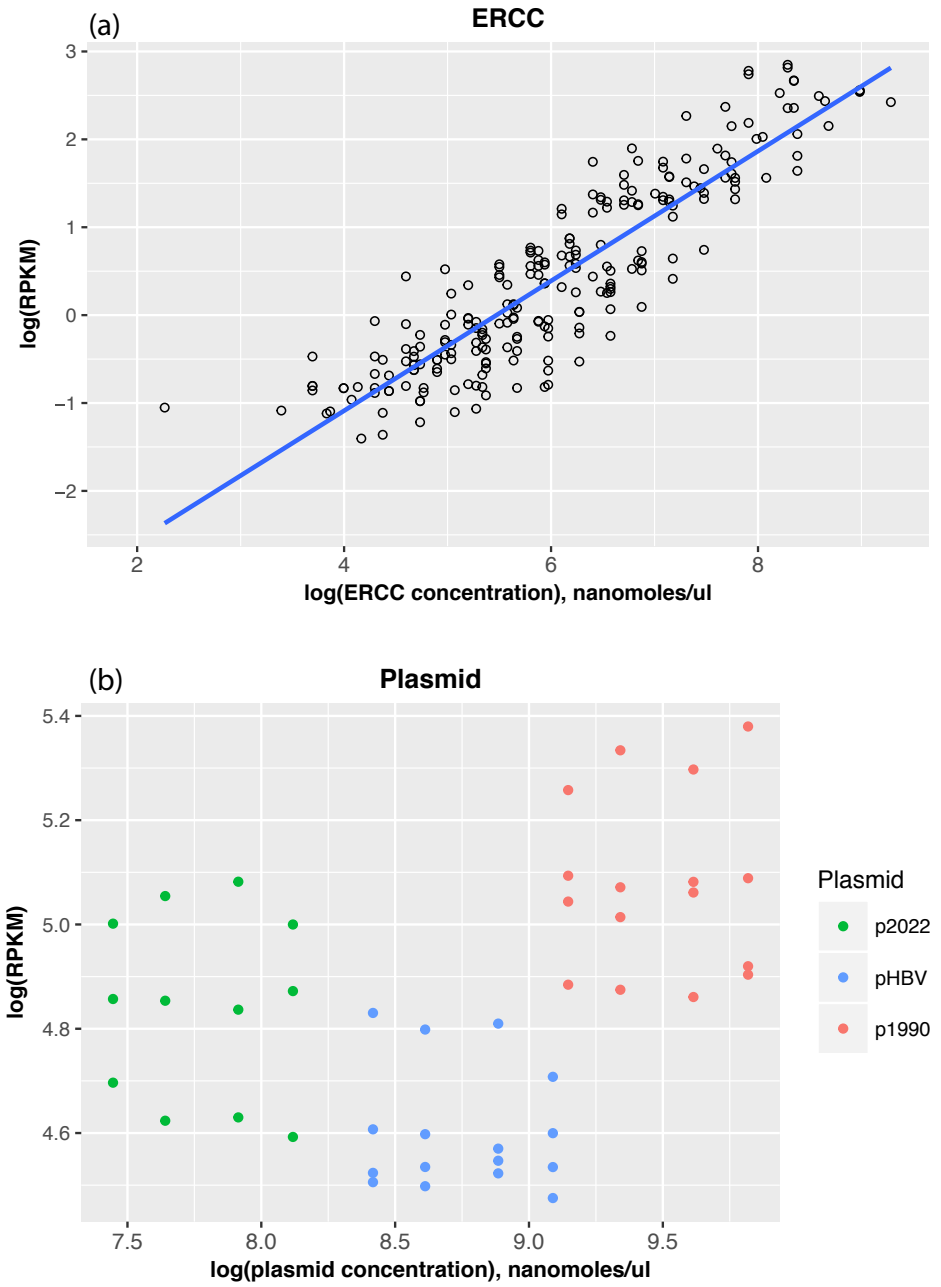


Figure 3.5: Relationship between read yield and input concentration (molarity). (a) ERCC, (b) Plasmid. RPKM = reads per kilobase of fragment per million reads.

reason for this difference is that Sapovirus and Astrovirus were present in high enough concentrations to be sequenced without enrichment whilst Norovirus was present in lower concentrations. This is consistent with the 80% failure rate in sequencing one or both Norovirus species in an evaluation of 15 laboratories with the same VMR control.

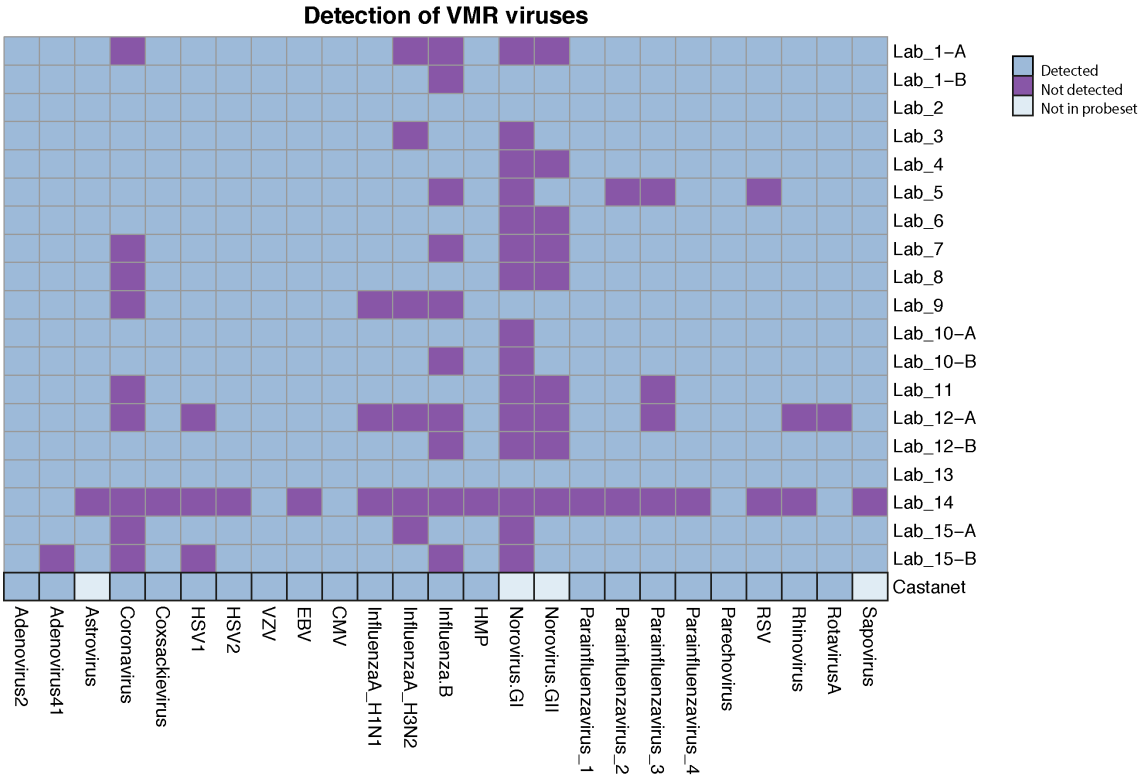


Figure 3.6: Comparison of Castanet with 15 other laboratories for sequencing of viruses in the NIBSC Viral Multiplex Reference 11/242. Mee and colleagues document the performance of 15 laboratories in sequencing the 25 viruses in the VMR. HSV=herpes simplex virus; VZV=varicella zoster virus; EBV=Epstein-Barr virus; CMV=cytomegalovirus; HMP=human metapneumovirus; RSV=respiratory syncytial virus

For individual microorganisms, we observed a linear relationship between organism load and sequencing yield. The VMR included five viruses where viral load had been quantified by the NIBSC using qPCR. For each virus, the number of deduplicated reads was proportional to input concentration across the dilution series (neat, 1 in 10, 1 in 100, 1 in 500, 1 in 1000) (Figure 3.7). However, the relationship between input viral load and yield of deduplicated reads differed between viruses. We observed a 10^2-10^3-fold enrichment for the five quantified

viruses (Figure 3.8).

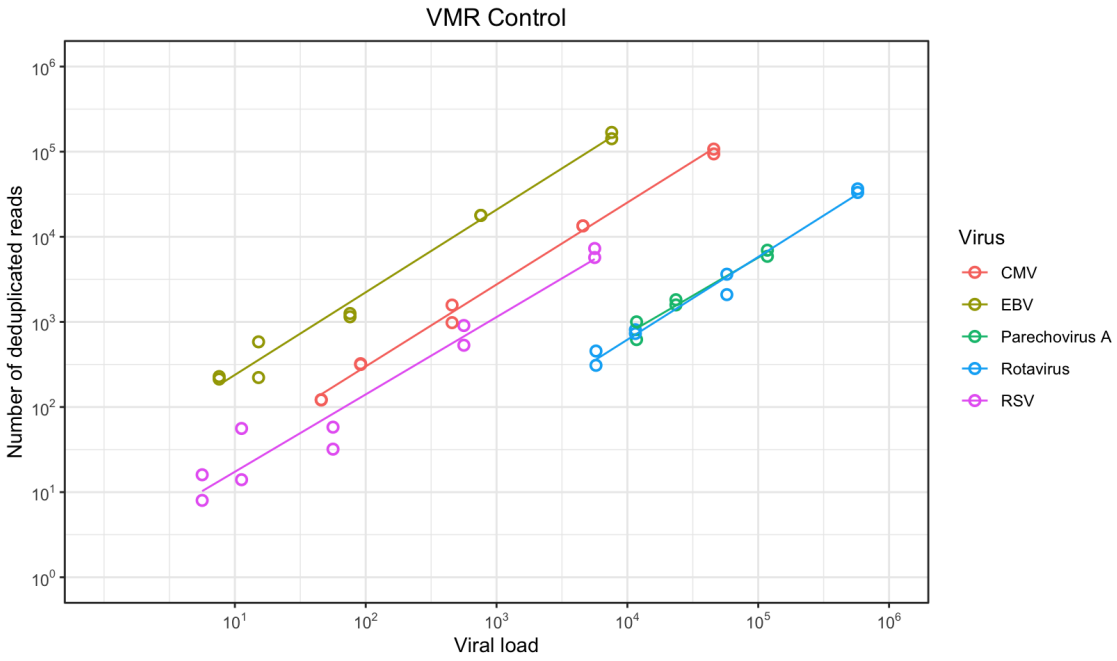


Figure 3.7: Relationship between viral load and sequencing yield in Viral Multiplex Reference (VMR) samples. The VMR was sequenced at a range of dilutions in two replicates. For the five viruses in the VMR that had been quantified by the NIBSC using qPCR, the relationship between viral load and sequencing yield is plotted. (CMV=cytomegalovirus; EBV=Epstein-Barr Virus; RSV=Respiratory Syncytial Virus)

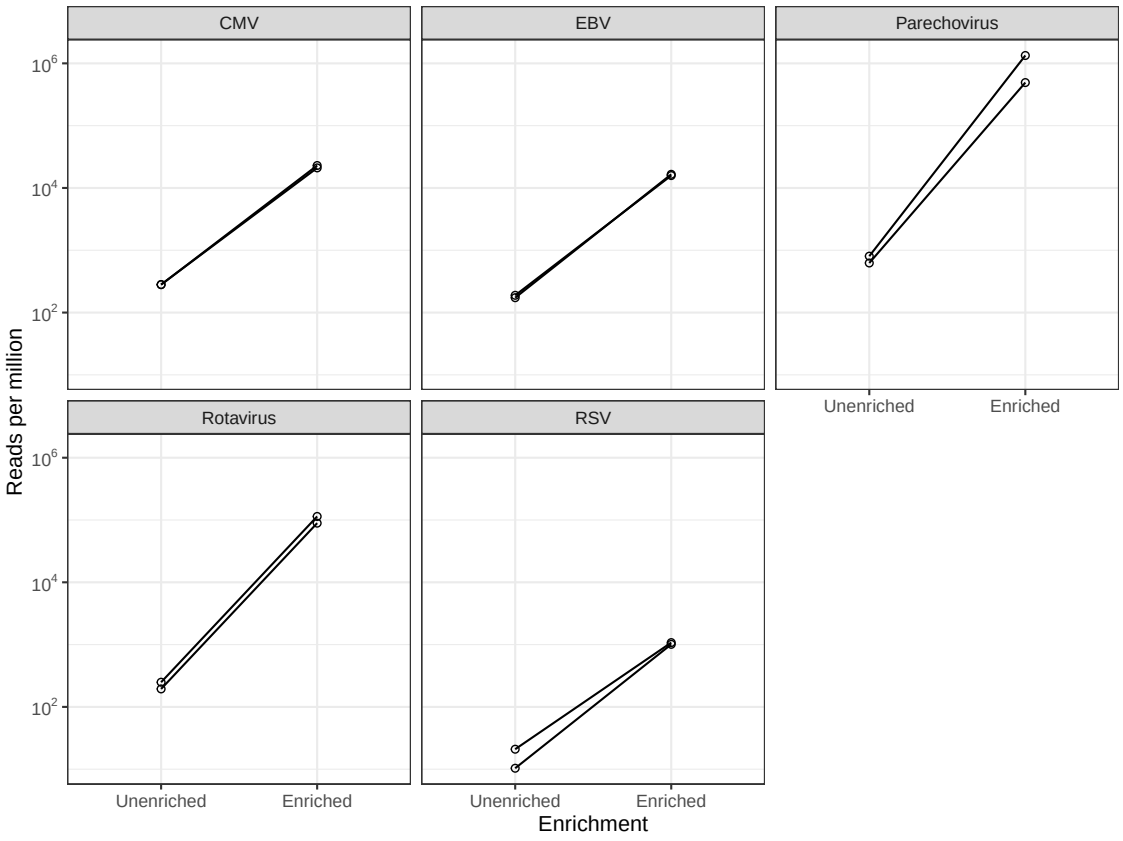


Figure 3.8: Increase in sequencing yield with enrichment The VMR was sequenced in two replicates at undiluted concentration. The sequencing yield (reads per million total reads) is plotted for each of the five viruses quantified by the NIBSC using qPCR. (CMV=cytomegalovirus; EBV=Epstein-Barr Virus; RSV=Respiratory Syncytial Virus)

3.2.5 Data processing

A data processing pipeline (Figure 3.9) was developed in collaboration with Dr Tanya Golubchik from the ChiMES project.

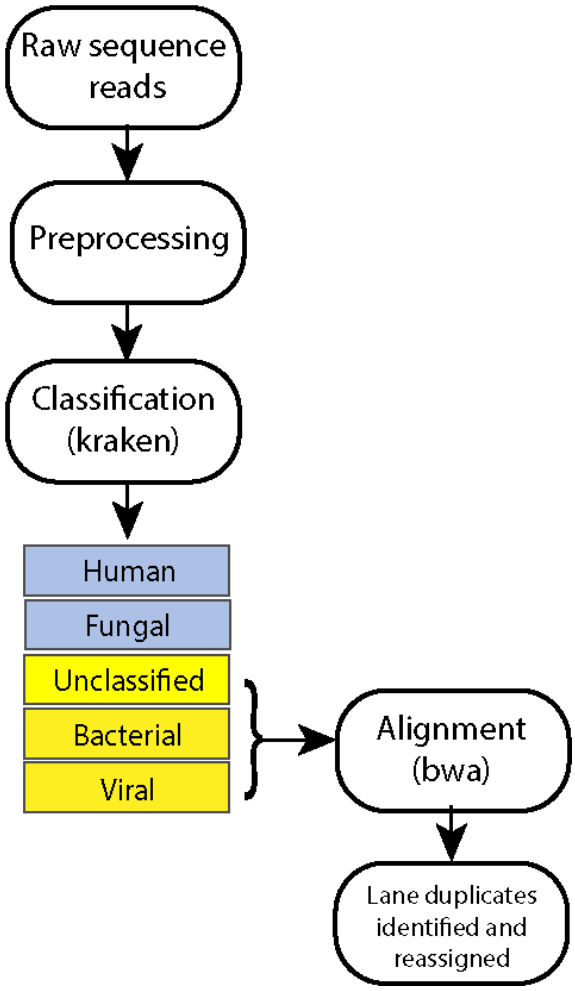


Figure 3.9: Data processing pipeline

De-multiplexed sequence read-pairs were trimmed of adapter sequences using Trimmomatic v0.36, with the ILLUMINACLIP options set to 2:10:7:1:true MINLEN:50, using the set of Illumina adapters supplied with the software (Bolger et al. 2014). The trimmed reads were then classified using Kraken v1 (Wood and Salzberg 2014) using a custom database containing the human genome (GRCh38 build), all RefSeq bacterial and viral genomes, and a selection of fungal genomes that were most likely to be associated with cases of meningitis (Cuomo 2017). These were: *Aspergillus fumigatus*, *Candida* spp., *Coccidioides*

spp., *Cryptococcus* spp., *Histoplasma capsulatum*, *Paracoccidioides brasiliensis*, and *Pneumocystis* spp. This custom database including both human and microbial reads was created to optimise the accuracy of classification, minimising the likelihood of human reads being classified as microbial or vice versa.

Subsequently, reads identified by Kraken as bacterial or viral were aligned using BWA v0.7.1243 with default settings to a multi-fasta reference of consensus sequences corresponding to the enrichment probe targets, augmented with sequences of known or suspected contaminants. These included (i) reagent contaminants (*Alteromonas* and *Achromobacter* spp.), (ii) genomes of two viruses known to have been sequenced on the same flow cell: MVMPCG spike-in control and Echovirus 7; and (iii) the rMLST sequences of commensal Streptococcus species (*S. mitis*, *S. oralis*, *S. pseudopneumoniae*) that were thought to be likely contaminants in clinical samples. This two step process of Kraken classification followed by BWA alignment has not previously been described; it was used in the *Castanet* data processing pipeline as the initial Kraken step efficiently and accurately separates human/fungal from bacterial/viral reads, enabling the BWA mapping to occur in a computationally efficient manner, accurately aligning reads to the probe sequences.

Following BWA alignment, we corrected our sequencing results for index misassignment (index hopping). This well-recognised phenomenon occurs when a small proportion of reads belonging to a sample gets misassigned to a different index in the multiplexed pool. For each sequencing pool, we identified duplicated reads and reassigned all reads in each duplicate cluster to the sample with the highest number of reads in that cluster. This data processing step is unique to *Castanet* and decreases the number of false positives arising from index hopping.

After duplicate reassignment, we calculated a set of descriptive statistics for each sample and target organism. These included sequencing depth with and without

deduplication, and coverage of target sequences at various depth thresholds. The collected statistics were combined with available laboratory data and ddPCR results where available and the resulting data frame used to train a random forest model (Section 4.2.1).

3.3 Discussion

In this chapter, I have described *Castanet*, a versatile probe-based enrichment sequencing method that combines the analysis of RNA and DNA from the same starting material in a single protocol and enriches for pathogens of interest using a modestly sized panel of probes.

3.3.1 Combined library preparation method

We developed a single library preparation workflow which combines separate RNA and DNA workflows in a single streamlined protocol, enabling the successful sequencing of both spike-in controls (ERCC and plasmid) as well as the VMR control. The experiments involving the spike-in controls were performed without enrichment whilst those involving the VMR control were performed with enrichment.

The CnoF protocol enabled sequencing of targets originating from RNA and DNA without bias, as evidenced by the yield of DNA:RNA closely matching input ratios of DNA:RNA for the plasmid (DNA) and ERCC (RNA) spike-in controls. This lack of bias is particularly important for detecting cases of co-infection in CAP, e.g. co-infection with DNA-based *Streptococcus pneumoniae* and RNA-based influenza A virus.

We also observed that there was no association between sequencing yield and fragment size for either plasmid or ERCC. This is important as it demonstrates

a lack of bias towards particular fragment sizes, which could in theory preferentially favour the sequencing of one organism over another. However, the analysis is limited to the range of fragment sizes studied (plasmid 379-3190bp; ERCC 250-2000nt).

For the five quantified viruses in the VMR, we observed a linear relationship between input concentration and sequencing yield. However, we noted that this relationship differed between viruses, presumably because of differences in the enriched (genomic) sequence length, the efficiency of sequencing library formation and capture and, perhaps the calibration of qPCR assays. Nevertheless, our results indicate that for a particular organism, deduplicated read counts can be compared between samples to provide information about relative organism loads.

3.3.2 Enrichment for targets of interest

We developed a probe panel covering 116 organisms from 17 virus families and 35 bacterial species. Considering the extent of coverage, the size of the probe panel was modest at 5.86 x 10^6 bases, with minimal redundancy and associated cost savings.

To my knowledge, this is the first published example of a probe panel that includes both bacteria and viruses as well as one targeting specific diseases. Other examples of probe panels include the separate bacterial BacCapSeq (Allicock et al. 2018) and viral VirCapSeq-Vert (Briese et al. 2015) panels which target all human pathogenic bacteria and vertebrate viruses respectively.

As our probe panel was designed based on conserved *rps* genes of the bacterial genome, there are several limitations. Firstly, we are unable to distinguish between different species of the Enterobacteriaceae family because of sequence homology within this region. Secondly, we did not enrich for regions encoding

for virulence genes and antimicrobial resistance genes. Although we did not set out with the latter aim, future iterations of the probe panel would benefit from extending beyond the rMLST system.

We observed between a 10^2 to 10^3-fold enrichment for the five quantified VMR viruses. The two higher viral load viruses (Parechovirus, Rotavirus) showed higher fold-change enrichment (10^3-fold vs 10^2-fold) compared to the lower viral load viruses (CMV, EBV, RSV). This is in keeping with the observations of (Bonsall et al. 2015) who noted a 10^3 fold enrichment for mid-range viral load samples but lower fold-change enrichment for lower viral loads.

3.3.3 Data processing pipeline

Our aim with the data processing was to implement a computationally efficient pipeline that would enable accurate alignment of microbial sequences. One challenge was the low signal to noise ratio, with the majority of samples containing >95% human reads. We dealt with this by using kraken for classification prior to bwa alignment. By building a custom database with human and microbial reference sequences, we could identify the human sequences and discard them before BWA alignment, making the latter stage substantially faster.

Another challenge was contaminants, including those from kit reagents, patient skin/mucosa, and simultaneously sequenced samples. We dealt with this by adding these sequences to the multi-fasta file for alignment so that the contaminants would map to the correct references rather than mis-mapping to a closely related pathogenic organism.

One area for future work is the diagnosis of fungal infections. Currently, we do not enrich for fungal reads and any sequences classified as fungal by kraken are discarded prior to BWA alignment. Fungal causes of CAP are currently underappreciated (Chen et al. 2001) and may represent a proportion of the cases

which remain diagnosed after routine bacteriology/virology.

3.4 Conclusions

In this chapter, I have described *Castanet*, a targeted metagenomic approach using enrichment probes for the sequencing of DNA and RNA-based bacteria and viruses from patient samples. Here, I have evaluated *Castanet* in terms of its performance for sequencing positive controls (plasmid and ERCC spike-ins; VMR control). In the next chapter, *Castanet* is applied to a cohort of patient samples and further evaluated.

4

Improved classification of microbiological aetiology in sepsis

This chapter explores the diagnosis of microbiological aetiology using various methods

4.1 Introduction

4.1.1 Clinical microbiology of the GAinS CAP cohort

Of the 1222 patients with sepsis due to CAP in the GAinS cohort, a majority of 729 patients (60%), were recorded as having no positive microbiology diagnosis (Table 4.1). This figure compares similarly to that of the Etiology of Pneumonia in the Community (EPIC) study (Jain et al. 2015) which evaluated CAP requiring hospitalisation among adult patients in the USA and identified 62% as lacking a microbiology diagnosis.

Interestingly, the GAinS and EPIC cohorts differed in that bacterial infections were predominant in the GAinS cohort (33%) whereas viral infections were predominant in the EPIC cohort (23%). The most likely reason for the high viral diagnosis rate in the EPIC cohort is that all patients were systematically tested

Diagnosis	Number of patients (%)	
	GAinS (n=1222)	EPIC (n=2259)
Unknown	729 (60)	1406 (62)
Bacterial	399 (33)	247 (11)
Viral	80 (7)	530 (23)
Mixed bacterial/viral	12 (1)	59 (3)
Fungal	2 (0.2)	17 (1)

Table 4.1: Microbiological classification for the GAinS (n=1222) and EPIC (n=2259) cohorts. GAinS data based on curated electronic Case Record Form data. EPIC data obtained from the Etiology of Pneumonia in the Community Study (Jain et al. 2015).

for viral infections by PCR on nasopharyngeal swabs whereas the GAinS cohort were tested according to clinician judgement. In addition, the higher prevalence of bacterial infections in the GAinS cohort may have been due to a higher severity of illness; GAinS recruited sepsis patients whilst EPIC recruited hospitalised CAP patients.

4.1.2 GAinS electronic Case Record Form data

The CAP microbiology section of the GAinS electronic Case Record Form (eCRF) is divided into three main sections (Figure C.1). In this web-based form, research nurses are presented with a series of tick boxes and free text sections relating to:

1. The name or category of any identified micro-organism.
2. The source (sample type) of any identified micro-organism.
3. Miscellaneous data (e.g. complicating factors such as presence of a pleural effusion, and vaccination status for pneumococcus/influenza).

The eCRF aims to be simple in its format, unfortunately however, it lacks sufficient resolution to enable detailed microbiological phenotyping. For example, the eCRF format doesn't allow for a negative test result to be recorded, this is particularly relevant in the case of molecular testing for viral infections which are frequently not tested for. In addition, the eCRF does not allow for an

identified organism to be matched to the source of the positive microbiological test, this is problematic if more than one organism is identified. Furthermore, one tick box option under source of positive microbiological test is "culture of lung secretions" which does not allow differentiation between an expectorated sputum sample and a directed broncho-alveolar lavage sample, samples with very different diagnostic values. The free text boxes of the eCRF often contained valuable clinical microbiology information. Thus, manual curation of the eCRF data was performed.

4.1.3 Childhood Meningitis and Encephalitis Study (ChiMES) cohort

The targeted metagenomics work was performed in collaboration with the Childhood Meningitis and Encephalitis Study (ChiMES) investigators. ChiMES (http://www.encephuk.org/studies/ukchimes) is a UK-based multi-centre clinical study of over 3000 children with suspected meningitis and encephalitis. A subset of patients (n=243) were included in the targeted metagenomics aspect of the study, consisting of patients with a positive microbiological diagnosis (n=108) and patients without a microbiological diagnosis (n=135) (Table 4.2). A further control group of meningitis negative individuals (n=22, patients requiring a lumbar puncture for reasons other than meningitis) was studied.

Diagnosis	**Number of patients (%)**
Unknown	121 (50)
Enterovirus	45 (19)
Human parechovirus	14 (6)
Streptococcus pneumoniae	15 (6)
Neisseria meningitidis	13 (5)
Other	35 (15)
Total	**243**

Table 4.2: Microbiological classification for the ChiMES (n=243) cohort. These diagnoses were made from routine clinical microbiological testing of CSF and/or blood.

4.1.4 Digital droplet PCR (ddPCR)

Digital droplet PCR (ddPCR) is a method which utilises water-oil emulsion droplet technology. In this technique, a droplet generation step is performed which partitions a 20μl sample into approximately 20,000 nanolitre-sized droplets. The principle is that each droplet contains one or no copies of the target molecule and the PCR reaction occurs simultaneously in each droplet containing the target molecule. Advantages of ddPCR include absolute quantification without the need for a standard curve (Poisson statistics are employed), increased precision due to the degree of sample partitioning, and increased signal-to-noise ratio where the relative concentration of target to background is low.

ddPCR was chosen as a method to be applied to the GAinS samples for several reasons. Firstly, it provided an independent method for evaluating the performance of *Castanet*. This is particularly applicable because a clinical microbiological diagnosis of a particular infection type would not necessarily mean the organism was present in plasma, both because the original diagnosis could have arisen from a different specimen type and also because the plasma sampled for metagenomics was collected after antibiotic administration. Secondly, ddPCR provided a quantitative method of assessing pathogen load in plasma.

4.1.5 Random forests

Random forests (Breiman 2001) are supervised machine learning algorithms which are capable of performing both regression and classification tasks. A forest is created which consists of a number of decision trees. Each decision tree sees a subset of the training data (bootstrapping) and at each decision node, a random sample of m predictors is chosen from the full set of p predictors. For classification tasks, each tree votes for a class and the forest chooses the class

which has the most votes by the majority of trees in the forest. Advantages of random forests leading to their suitability for classification tasks based on (meta)genomic data include their ability to handle missing values, their robustness to overfitting, and their ability to handle large datasets with high dimensionality.

4.1.6 Axiom Microbiome Array

The Axiom Microbiome Array (Affymetrix) is a commercially available platform that enables detection of all organisms in a sample. Organisms are identified at species- or strain-level resolution within a single reaction. The platform is based on microarray technology, with 1,277,846 target probes and 60,152 random negative control probes. The target probes represent 135,555 sequences from 12,513 microbial species from five domains: archaea, bacteria, fungi, protozoa and viruses.

4.1.7 Aims

To improve microbiological classification of GAinS sepsis patients with CAP through:

1. Application of targeted metagenomics to plasma samples
2. Use of droplet digital PCR (ddPCR) to assay *Streptococcus pneumoniae* and Epstein-Barr Virus from plasma samples
3. Use of the Axiom Microbiome Array

4.2 Results

4.2.1 Targeted metagenomics

Large scale analysis of clinical samples. There was successful sequencing of a total of 854 individuals, derived from 243 ChiMES meningitis cases, 27 non-meningitis negative-control CSF samples, 573 GAinS sepsis cases, and 11 negative-control plasma samples (Figure 4.1). The 243 meningitis included 122 patients for which a pathogen had been identified by clinical microbiology (108 from CSF only, plus 14 from a blood sample +/- CSF) and 121 where no pathogen had been identified before sequencing. The sepsis cases comprised 126 for which a pathogen had been identified and 447 chosen for the study because no pathogen had been identified in any relevant sample. The clinical characteristics for the 573 GAinS CAP sepsis cases are summarised in Table 4.3.

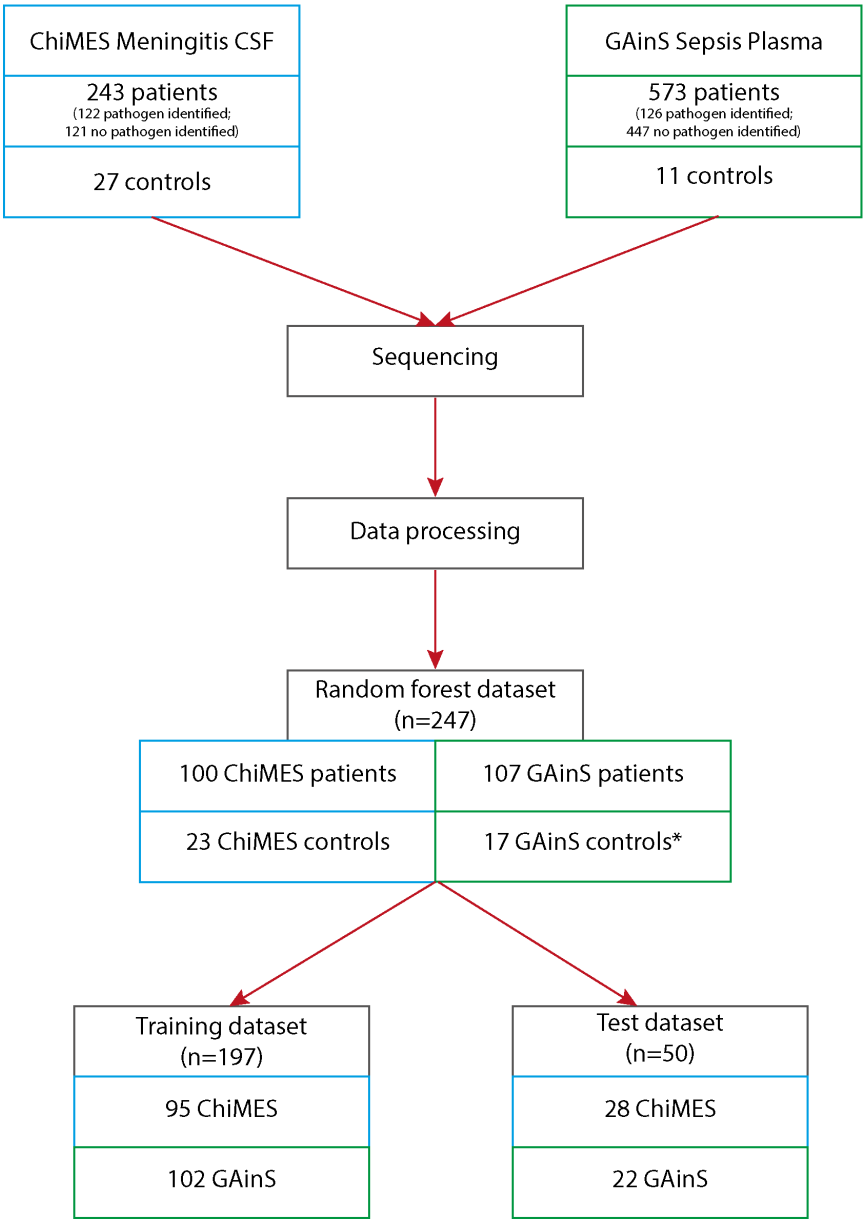


Figure 4.1: Flowchart of samples analysed. Samples undergoing targeted metagenomic sequencing are detailed in the flowchart. Control samples for ChiMES are CSF samples obtained from individuals without meningitis/encephalitis. Control samples from GAinS are plasma samples obtained from non-infected individuals undergoing elective cardiac surgery. CSF samples are outlined in blue and plasma samples are outlined in green. Samples used to train and test the random forest algorithm are detailed in the flowchart. *The number of GAinS control samples in the random forest dataset (17) exceeds the number of patients detailed initially (11) because several patient samples were split into several aliquots and sequenced.

Development of random forest algorithm. In order to evaluate the ability of *Castanet* to identify pathogens in real clinical samples, a 'truth dataset' of samples was compiled whose status for particular pathogens was known with confidence. For meningitis cases, the pathogen identified by clinical microbiology testing was accepted as the truth state for microbiology-positive samples. For the 126 pathogen-positive CAP sepsis cases, the pathogen identification had often been made in a sample other than plasma and most of the plasma samples in the collection had been obtained after administration of antibiotics, two situations in which plasma levels of pathogens might have been undetectable by any method. Accordingly, a positive result by ddPCR for *S. pneumoniae* or Epstein-Barr virus (EBV) was used to define a sample subset with which to learn the characteristics of pathogen-true-positive sequencing data.

Another key issue in interpreting metagenomic sequencing data, especially in large pools of samples, is to distinguish low-level positives from true-negative samples. Here, the *S. pneumoniae* ddPCR assays were used to estimate the threshold for considering a sample sequence-positive. Samples with fewer than either 72 total (specificity 0.94, sensitivity 0.83) or 4 de-duplicated (specificity 0.84, sensitivity 0.85) sequence reads from any single pathogen were excluded from consideration as sequencing-positive (i.e. they were not considered by the random forest algorithm) (Figure 4.2).

Characteristic	GAinS cohort (n=573)
Age (years)	61
Male sex	324 (57%)
Charlson comorbidity index	1.1
Mortality (28-day)	121 (21%)
ICU length of stay (days)	9.9
SOFA score (day 1)	6.2
SOFA score (maximum)	7.1
Mechanical ventilation	460 (80%)
Vasopressors	297 (52%)
Earliest ICU day of sampling	Day 1: 302 (53%); Day 3: 192 (34%); Day 5: 79 (14%)

Table 4.3: Clinical characteristics of GAinS metagenomic cohort (n=573) Mean or count data is summarised for the various clinical characteristics. ICU length of stay excludes patients who died in ICU. Earliest day of sampling refers to the earliest timepoint at which a plasma sample was collected for metagenomic analysis.

Combining the above criteria, 100 ChiMES meningitis CSF and 107 GAinS sepsis plasma samples were identified for inclusion as samples with known pathogen status. In addition, negative control samples were included in this cohort (CSF from non-meningitis patients, n=23; plasma from sepsis-negative patients, n=17) to provide instances of pathogen-negative data. Plasma and CSF samples that were microbiology-positive for a particular pathogen were deemed negative for other pathogens. Reads aligning to viruses known to reactivate in sepsis (herpes simplex virus, cytomegalovirus, human herpes virus 6, JC virus) were excluded from the analysis, apart from those EBV samples where ddPCR data was available.

The 247 samples defined above were randomly allocated to training and test datasets in an 80:20 ratio. The training dataset (197 samples: 95 CSF; 102 plasma) were used to train a random forest classifier that used a set of variables derived from the sequencing data to derive a score between 0 and 1 to indicate whether it was positive for each organism with reads in a sample. Most samples contained reads from multiple organisms and the random forest returns a score for each one of these organisms.

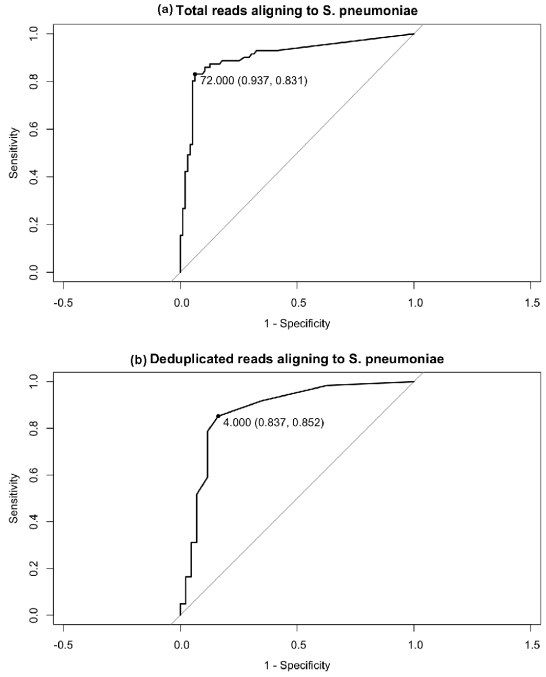


Figure 4.2: *S. pneumoniae* detection thresholds in GAinS samples. Sample read statistics were assessed as predictors of ddPCR-positive status for *S. pneumoniae* detection. ROC curves of (a) total reads and (b) de-duplicated reads as predictors are presented here. Youden's J statistic was used to select the respective thresholds of 72 total reads (specificity 0.94, sensitivity 0.83) and 4 de-duplicated reads (specificity 0.84, sensitivity 0.85) for calling a sample positive.

The test dataset comprised 50 samples (28 CSF; 22 plasma). A cut-off random forest score of 0.465 was selected for classification of the test set, to appropriately weight specificity over sensitivity (Figure 4.3). At this threshold, there were five false negatives and one false positive in the ChiMES test dataset and one false negative and three false positives in the GAinS test dataset. In the combined set of test samples (Figure 4.4), the sensitivity was 86.7% (39 of 45 true positives) and the specificity was 98.6% (283 of 287 true negatives). Among the most informative sequencing data metrics for prediction were the numbers of total and de-duplicated reads matching a pathogen, taken as the respective proportions of reads aligning to all pathogens in the probeset, and whether a high proportion of the targeted region (regions in the probeset) for that pathogen were covered by reads.

Since excluding pathogens from a diagnosis can also be clinically useful, the performance of the method in predicting negative status was assessed. With a random forest score threshold of 0.015, 59.2% of true negatives were correctly identified, with a specificity of 97.8%, implying that for many samples it is possible to exclude many possible pathogens without erroneously ruling out true positives.

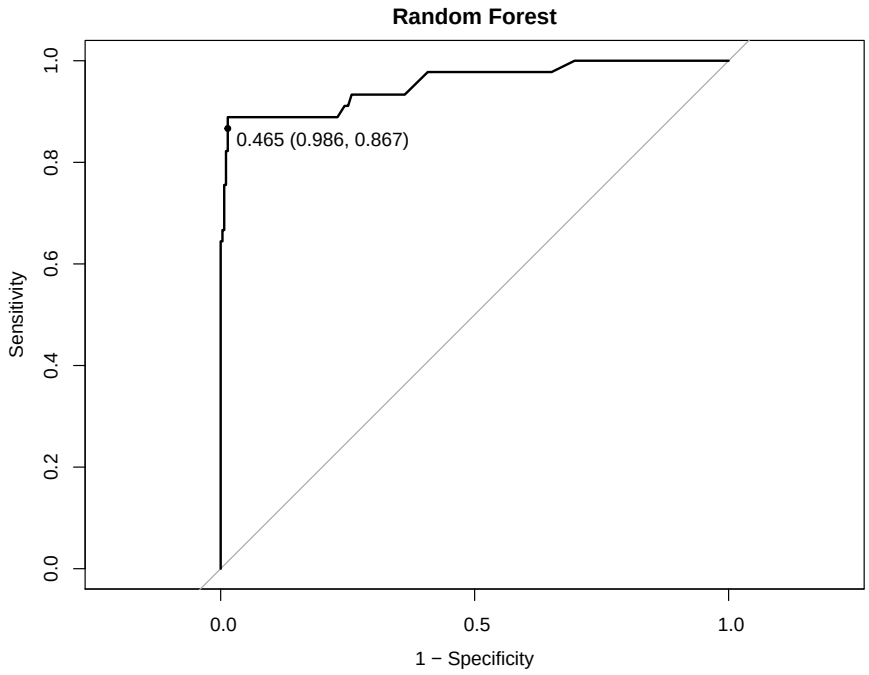


Figure 4.3: Random forest ROC curve. ROC curve derived from random forest training dataset (n=197; ChiMES n=95, GAinS n=102) to choose random forest score threshold for predicting positive samples. A threshold of 0.465 was selected to call samples as positive for a particular pathogen. At this threshold, specificity was 0.986 and sensitivity 0.867.

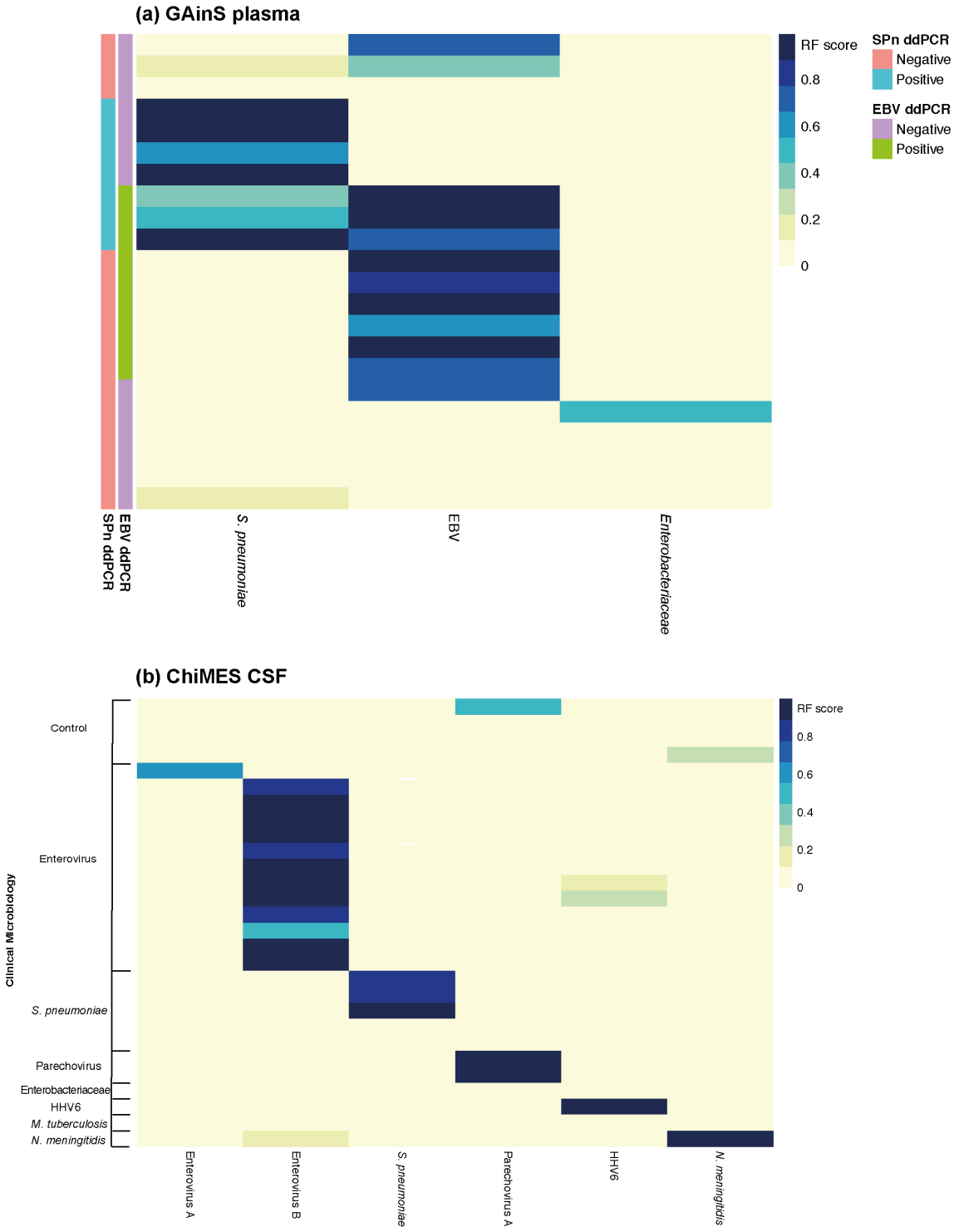


Figure 4.4: GAinS/ChiMES test dataset. Performance of *Castanet* in clinical samples with a positive microbiology diagnosis. The test dataset included 50 samples: (a) 22 plasma samples (20 GAinS, 2 sepsis-negative controls); (b) 28 CSF samples (24 ChiMES, 4 meningitis-negative controls). The combined overall test dataset specificity and sensitivity was 0.986 and 0.867 respectively. Only organisms detected in each sample set are included as columns in each panel. RF=random forest; HHV6=human herpes virus 6; SPn=*Streptococcus pneumoniae*; EBV=Epstein-Barr Virus; TB=*Mycobacterium tuberculosis*; ddPCR=droplet digital PCR.

Application to samples with no previous microbiology diagnosis. Of the 729 GAinS patients with sepsis due to CAP that had no known microbiological diagnosis (Table **??**), 447 had plasma available for analysis. In this group of samples, in which no causative pathogen had been identified by routine clinical microbiology, *Castanet* identified one or more pathogens in 37% of samples (n=165), including both bacteria and viruses that in many cases were likely to have been causative (Figure 4.5). Among such pathogens, instances of EBV, human herpes virus 6, herpes simplex virus, JC virus and cytomegalovirus in sepsis may represent viral reactivation in the context of critical illness, while *Burkholderia cerpacia* and *Nocardia asteroides* sequences have been previously noted to represent contamination of samples (Salter et al. 2014). Excluding these likely reactivations and contaminants, *Castanet* made 50 new detections in sepsis patients, comprising 11% of previously unresolved cases (Table 4.4).

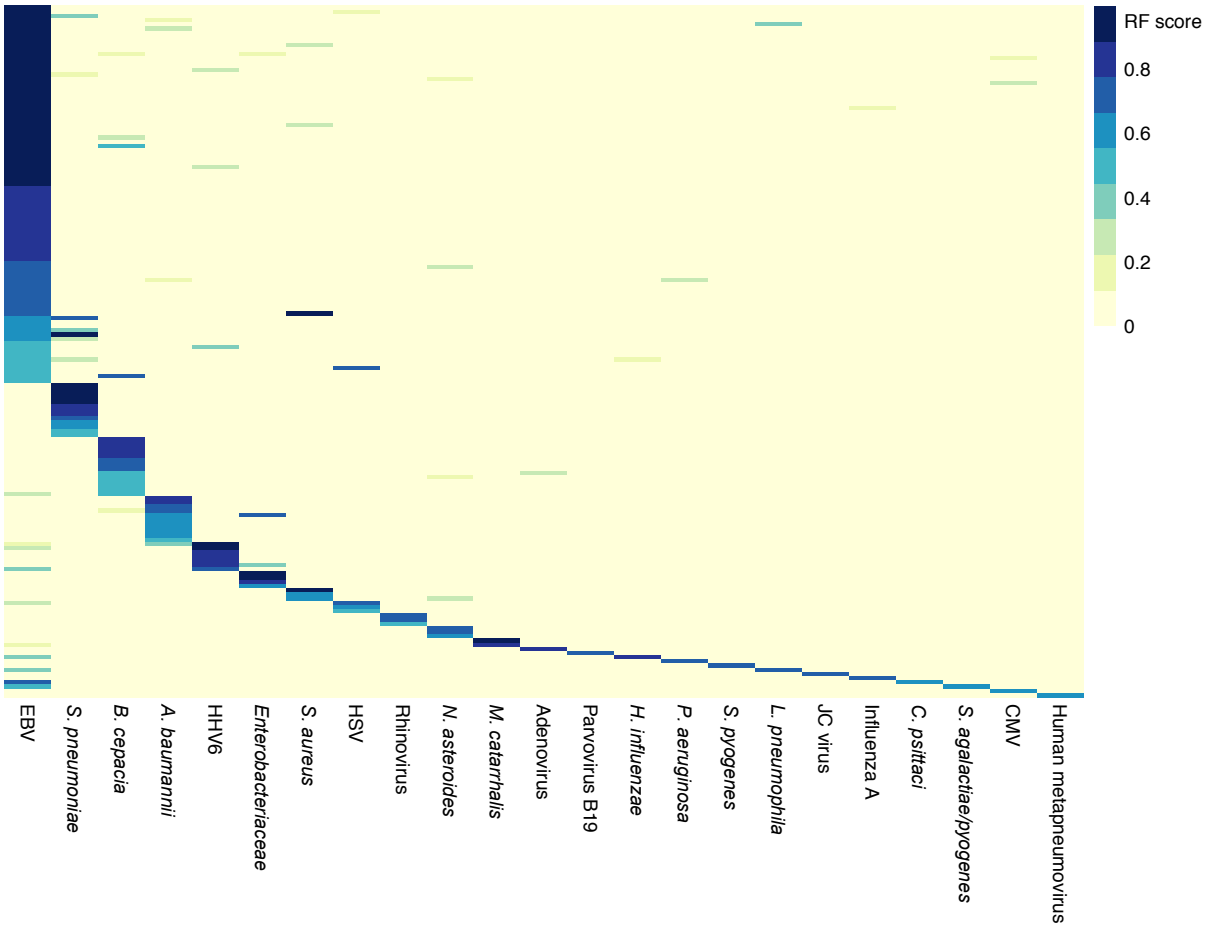


Figure 4.5: Performance of *Castanet* in GAinS CAP sepsis cases with no clinical microbiology diagnosis. The heatmap depicts samples from patients with no microbiological diagnosis where at least one organism was detected at a random forest (RF) score >0.465 (n=165/447). Only organisms detected in each sample set are included as columns in the heatmap.

4.2.2 Digital droplet PCR in GAinS samples

ddPCR was performed on GAinS samples with sepsis due to CAP as an independent method for evaluating the performance of *Castanet* and also to enable quantitative evaluation of pathogen load in the samples.

Streptococcus pneumoniae. ddPCR for the *S. pneumoniae cpsA* gene (conserved in nearly all 90 known serotypes of *S. pneumoniae* (**Morona2004**)) was performed on the following samples:

1. 139 samples from 92 patients with *S. pneumoniae* infection diagnosed by clinical microbiology.
2. 15 samples from 15 patients with no microbiology diagnosis.
3. 4 samples from 4 patients with influenza infection diagnosed by clinical microbiology.

ddPCR was performed on the latter two groups of samples for cross-validation because *S. pneumoniae* reads had been unexpectedly identified on metagenomic sequencing. This was prior to the development of the *Castanet* random forest algorithm, thus the remainder of this section will focus only on the ddPCR result from the 92 patients with *S. pneumoniae* infection diagnosed by clinical microbiology.

Of the 139 samples tested, 47 (33.8%) were positive for *S. pneumoniae*. ddPCR results were compared with *Castanet* sequencing (Table 4.5). If ddPCR is considered as the gold standard result, this would mean Castanet has a specificity of 100% and a sensitivity of 83% for *S. pneumoniae*. There was no significant association between ddPCR positivity and blood culture positivity for *S. pneumoniae* (χ^2=0.13, d.f.=1, p=0.72).

Epstein-Barr virus. ddPCR was performed on 619 samples from 565 patients for Epstein-Barr virus (EBV). See Section 5.2.3.

Organism	Number of patients
Streptococcus pneumoniae	16
Acinetobacter baumannii	11
Enterobacteriaceae	5
Staphylococcus aureus	4
Moraxella catarrhalis	2
Haemophilus influenza	1
Pseudomonas aeruginosa	1
Streptococcus pyogenes	1
Legionella pneumophila	1
Chlamydia psittaci	1
Rhinovirus	3
Adenovirus	1
Parvovirus B19	1
Influenza A	1
Human metapneumovirus	1
Total	**50**

Table 4.4: New pathogen identifications made by *Castanet* among sepsis plasma samples.

		Castanet		Total
		Negative	Positive	
ddPCR	Negative	92	0	92
	Positive	8	39	47
Total		100	39	139

Table 4.5: ddPCR and Castanet results for *S. pneumoniae in GAinS CAP sepsis samples*. Comparison of results in samples with both ddPCR and *Castanet* data (n=139 samples from n=92 patients).

Influenza virus. ddPCR was performed for the influenza A matrix gene on 14 plasma samples from 14 patients. All patients had been diagnosed with influenza infection by PCR of a nasopharyngeal swab or respiratory specimen. All samples were negative for influenza via *Castanet* and also tested ddPCR negative.

Relationship between organism load and sequencing yield Similar to the finding for the five quantified VMR viruses (Figure 3.7), a linear relationship between input pathogen load and sequencing yield was observed in a subset of sepsis plasma samples for which bacterial and viral load of *S. pneumoniae* (n=102) and EBV (n=199) respectively had been quantified by ddPCR (Figure 4.6). Samples where both the ddPCR result and sequencing reads were zero were excluded from the analysis. These findings indicate that the number of de-duplicated reads obtained from targeted enrichment provides data on quantitative yield.

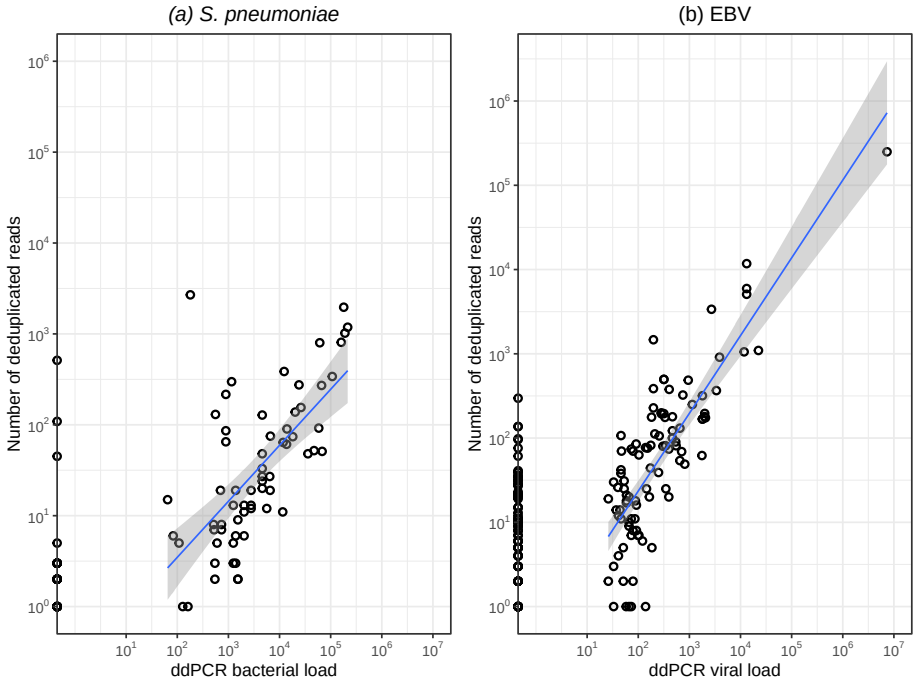


Figure 4.6: Organism load and sequencing yield in sepsis samples. De-duplicated read yield was plotted against pathogen load, estimated by ddPCR in samples from a subset of cases in which (a) *Streptococcus pneumoniae* (n=102) or (b) Epstein-Barr virus (EBV) (n=199) was detected by sequencing. A linear relationship between pathogen load and sequencing yield was observed for each organism (*S. pneumoniae*: Pearson's $R^2 = 0.449$, $p = 8.8 \times 10^{-5}$; EBV: Pearson's $R^2 = 0.702$, $p = 2.9 \times 10^{-16}$.)

4.2.3 Axiom Microbiome Array

We evaluated plasma from ten patients on the Axiom Microbiome Array platform (Affymetrix). This included four patients with *S. pneumoniae* CAP sepsis, three patients with influenza CAP sepsis, and three uninfected cardiac surgery controls. The Axiom Microbiome Array was evaluated only in terms of detection for DNA-based pathogens as limited sample availability meant that RNA in the samples was not processed (this would have included performing a second assay in parallel, with cDNA synthesis).

Of the 10 samples, all 4 *S. pneumoniae* sepsis samples were positive for *S. pneumoniae* by *Castanet* sequencing (Figure 4.7). In contrast, the Axiom Microbiome Array platform only detected *S. pneumoniae* in 2 of the 4 samples. When compared with *Castanet* sequencing, the Axiom Microbiome Array platform had a 50% sensitivity and 100% specificity for *S. pneumoniae*, although the number of samples was very small.

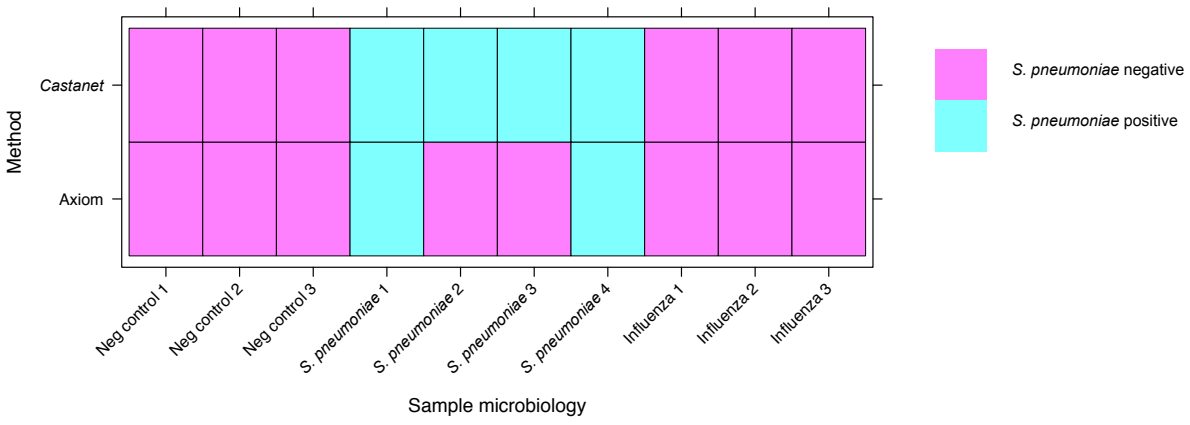


Figure 4.7: Axiom Microbiome Array results for *Streptococcus pneumoniae* compared with *Castanet*. Clinical microbiology status and detection for *S. pneumoniae* is displayed in the heatmap.

In total, 10 organisms were detected across the 10 samples (Figure 4.8). The majority of organisms were of little clinical consequence, e.g. Torque teno virus and *Propionibacterium acnes*. However, *Escherichia coli* was detected in 9 out of 10 samples, presumably reflecting sample or reagent contamination. Quantitative data is not a feature available via the Axiom Microbial Detection Analysis Software (MiDAS).

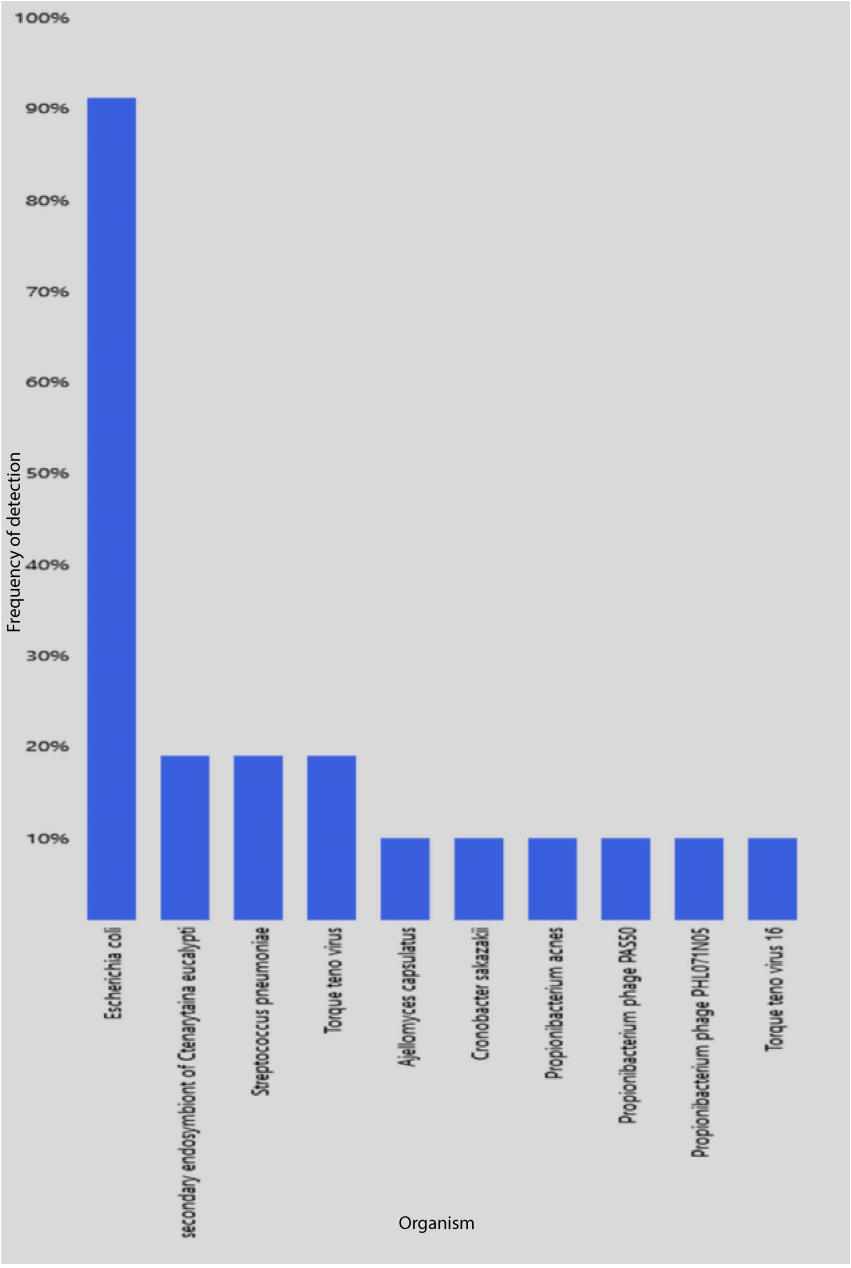


Figure 4.8: Summary of Axiom Microbiome Array results. All organisms detected in any of the 10 samples are depicted in the bar chart with the frequency of detection displayed on the y-axis.

4.2.4 Application to the GAinS cohort

Of the three methods explored in this chapter (*Castanet*, ddPCR, Axiom Microbiome Array), improved microbiological phenotyping for GAinS CAP sepsis patients was obtained primarily through *Castanet* targeted metagenomic sequencing.

In the GAinS CAP sepsis patients undergoing metagenomic sequencing (n=573), *Castanet* identified a likely causative pathogen in 176/573 (31%) of cases. This was a 9% improvement on the 126/573 (22%) of cases with a microbiological diagnosis arising from clinical microbiology alone (Table 4.6).

Integrating this into the GAinS CAP sepsis cohort as a whole (n=1222) (Table 4.1), the application of targeted metagenomics reduced the number of individuals with no microbiology diagnosis from 729 (60%) to 693 (57%) (Table 4.7).

Organism	Clinical microbiology	Clinical microbiology + *Castanet*	Prevalence in overall GAinS cohort
Unknown	447 (78%)	397 (69%)	60%
S. pneumoniae	92 (16%)	108 (19%)	15%
Other bacterial	6 (1%)	33 (6%)	18%
Influenza	13 (2%)	14 (2%)	5%
Other viral	15 (3%)	21 (4%)	3%
Total	**573**	**573**	100%

Table 4.6: Summary of microbiology in GAinS patients with sepsis due to CAP undergoing *Castanet* sequencing (n=573). The prevalence of each organism in the overall GAinS cohort (n=1222) based on clinical microbiology alone is displayed in the last column to show how the GAinS metagenomic cohort (n=573) compares in microbiology diagnosis to the overall cohort.

Diagnosis	Number of patients (%)	
	With metagenomics	Without metagenomics
Unknown	693 (57)	729 (60)
Bacterial	426 (35)	399 (33)
Viral	86 (7)	80 (7)
Mixed bacterial/viral	15 (1)	12 (1)
Fungal	2 (0.2)	2 (0.2)

Table 4.7: Summary of microbiology in entire cohort of GAinS patients with sepsis due to CAP (n=1222). Diagnoses made on the basis of metagenomics and clinical microbiology and compared with those made on the basis of clinical microbiology only.

4.3 Discussion

4.3.1 Targeted metagenomics and droplet digital PCR

The *Castanet* workflow enabled successful sequencing of a range of bacteria and viruses (both RNA-based and DNA-based) from a subset of sepsis plasma samples. In spite of difficulties describing a 'truth dataset' of known positives and known negatives, the characteristics of pathogen-positive and pathogen-negative samples were defined using a random forest model that combines data across pathogens and diseases. In the test dataset, a specificity of 98.6% and sensitivity of 86.7% was observed. Similarly, high specificity (100%) and sensitivity (83%) was also observed when *Castanet* was evaluated against a gold-standard independent method of detection (ddPCR) for *S. pneumoniae*.

Despite this high sensitivity of detection, a causative organism was only identified in 11% of plasma samples from patients who had no positive microbiology result from routine clinical microbiological testing. This was lower than the figure observed in the ChiMES cohort, where 32% of similar samples yielded a new diagnosis with *Castanet*. There are several likely explanations for this. Firstly, timing of sample collection was suboptimal for diagnostic metagenomics; the earliest time at which a plasma sample was obtained was on the first day of ICU admission after the consenting process. By this time, the

diagnosis of sepsis would have been made and antimicrobial drugs administered. Secondly, the nature of the sample type would have contributed to the low diagnosis rate. In the setting of sepsis secondary to pneumonia, plasma may contain little or no pathogen material from agents such as influenza A virus where viraemia is not typically a feature of disease.

Future work to enable more comprehensive evaluation of *Castanet* sequencing would include recruitment and sampling of patients at earlier time points prior to the administration of antimicrobials, with consent procedures such as emergency waiver of consent in place. In addition, sampling of other relevant specimen types for CAP sepsis (e.g. nasopharyngeal swabs for viral infections) is likely to increase the rate of pathogen detection.

Other opportunities for future work include improving probe design to enable better resolution of pathogenic species within clusters of closely related bacteria. Currently, *Castanet* is unable to differentiate between species within the *Enterobacteriaceae* family or between *Streptococcus pyogenes* and *Streptococcus agalactiae*. This work could also focus on bacterial sequences of particular clinical interest such as factors associated with virulence and anti-microbial resistance.

4.3.2 Axiom Microbiome Array

Compared with *Castanet*, the Axiom Microbiome Array had low sensitivity (50%) for *S. pneumoniae* detection. This may have been because the manufacturers recommend a quantity of 50-100ng DNA per sample which was unachievable in this experiment, given the low DNA quantities (2-20ng) usually obtained following nucleic acid extraction of the typical 500μl aliquot of plasma received from the recruiting centres. Other disadvantages include the absence of quantitative data, meaning it was difficult to determine whether the presence of *E. Coli* was as a contaminant or pathogenic. Also, separate workflows (and

thus increased sample volumes) are required for RNA and DNA pathogens. These disadvantages outweighed the main benefit of the Axiom Microbiome Array which included the straightforward sample processing, data generation and analysis steps.

For analysis of sample types like plasma from sepsis patients where sensitivity is a major challenge and specimen volumes are limited, the Axiom Microbiome Array is not a suitable method for diagnostics.

4.4 Conclusions

In the GAinS cohort of sepsis CAP patients, the majority of individuals (60%) lack a microbiological diagnosis despite extensive laboratory testing. In this chapter, I have shown that targeted metagenomics enables new diagnoses to be made in this challenging cohort using the *Castanet* workflow. This has important clinical and scientific implications, enabling more accurate prescribing of antimicrobial therapy and better understanding of disease heterogeneity.

5

Integration of Microbiology with the Host Response

This chapter explores the integration of metagenomic data with host transcriptomic and genomic data in order to understand the host response in sepsis.

5.1 Introduction

Application of omics-based methodologies is advancing understanding of the dysregulated host immune response to infection in sepsis. However, the frequently elusive nature of the infecting organism has limited efforts to understand the effect of disease heterogeneity involving the pathogen. This chapter explores how a better understanding of microbiology in the GAinS cohort can be applied to transcriptomic and genomic-based approaches to understand the host response in sepsis.

5.1.1 Immunosuppression in sepsis

Over the last two decades, there has been a significant shift in our conceptual understanding of immune dysfunction in sepsis. Earlier thinking and sepsis definitions centred around the Systemic Inflammatory Response Syndrome (SIRS) and the corresponding hyperinflammatory and immune activating processes. Now, increasing evidence suggests that immunosuppression is a key disease feature, co-occuring with immune activation, and an important contributor to patient morbidity and mortality (Daviaud et al. 2015).

In 2001, Munford and Pugin (Munford and Pugin 2001) published one of the first studies demonstrating that both pro-inflammatory and anti-inflammatory processes occur rapidly after the onset of sepsis. Later, Boomer and colleagues (Boomer et al. 2011) published further evidence in the form of histological, biochemical and flow cytometric studies documenting defects in immunity in the tissues of patients dying from sepsis. T cell exhaustion, the functional impairment of effector T cells associated with decrease cytokine production, loss of proliferative capacity and decreased cytotoxicity, was described as a key mechanism in sepsis-induced immmunosuppression.

More recently, transcriptomic studies describing sepsis endotypes have highlighted features of immunosuppression as important in distinguishing endotypes associated with worse outcome (Davenport et al. 2016) (Scicluna et al. 2017).

5.1.2 Sepsis Endotypes

Previous work in the Knight laboratory (Davenport et al. 2016) describes two sepsis endotypes in the GAinS cohort from genome-wide gene expression analysis of peripheral blood leukocytes. Unsupervised hierarchical cluster

analysis of the top 10% most variable probes (n=2619) was performed on a discovery cohort of 265 GAinS CAP patients. Two groups were identified, Sepsis Response Signature 1 and 2 (SRS1 and SRS2). Individuals with an SRS1 endotype were characterised by features of immunosuppression, including endotoxin tolerance, T-cell exhaustion, and downregulation of HLA class II. Importantly, this group demonstrated higher 14-day mortality (discovery cohort hazard ratio 2.4, 95% CI 1.3-4.5, p-0.005; validation cohort HR 2.8, 95% CI 1.5-5.1, p=0.0007), supporting the observation that patients dying of sepsis show marked immunosuppression.

Similar findings were made independently by the MARS consortium (Scicluna et al. 2017). Sepsis patients admitted to ICUs in the Netherlands were studied and four molecular endotypes (Mars1-4) were described using unsupervised consensus clustering in a discovery cohort of 306 patients from whole blood genome-wide gene expression profiling. Mars1 patients had the worst outcome, with increased 28-day mortality (HR vs all other endotypes 1.86, 95% CI 1.21-2.86, p=0.0045). Similar to SRS1, this high-risk endotype was characterised immunosuppression, with decreased expression of genes involved in key innate and adaptive immune cell functions.

5.1.3 Viral reactivation

Studies describing a high frequency of viral reactivation in sepsis patients further support the concept that previously immunocompetent individuals can develop a varying degree of functional immunosuppression in response to severe infection. Walton and colleagues observed that 42.7% of sepsis patients displayed viraemia with more than one reactivated virus (Walton et al. 2014); the MARS consortium describe a 68% frequency of herpesvirus viraemia amongst individuals with septic shock (Ong et al. 2017).

Reactivated viruses include the herpesviruses Epstein-Barr virus (EBV), cytomegalovirus (CMV), herpes simplex virus (HSV), and human herpesvirus 6 (HHV-6), torque teno virus (TTV), and the polyomaviruses BK and JC virus. In ICU sepsis patients, EBV is the most commonly observed reactivated virus at a frequency of 32-48% in plasma (Walton et al. 2014) (Ong et al. 2017). Neither study observed an association with mortality for EBV reactivation alone although the MARS consortium identified a 3.17 hazard ratio (HR) (95% CI 1.41-7.13) for mortality with concurrent CMV and EBV reactivation (Ong et al. 2017). In addition, Walton and colleagues observed an increased incidence of fungal infections, mean Sequential Orgain Failure Assessment (SOFA) score and ICU length of stay with EBV reactivation (Walton et al. 2014). CMV is another commonly reactivated virus, observed at a frequency of approximately 18% in the plasma of previously immunocompetent sepsis patients (Ong et al. 2017). CMV reactivation has been observed to be associated with increased ICU and hospital length of stay as well as prolonged mechanical ventilation in critically ill sepsis patients (Heininger et al. 2011).

5.1.4 Epstein-Barr virus

Like other herpesviruses, EBV is characterised by its ability to remain dormant in human cells after primary infection. Approximately 90% of adults worldwide are EBV seropositive (Cohen 2000), the highest rate of any herpesvirus.

Following B lymphocyte infection, latency is established with persistent infection arising due to a dynamic balance between viral evasion strategies and host immune responses. Of the 100 genes encoded by the 172Kbp EBV genome, ten are expressed during latency, establishing and maintaining the ”immortalized” state.

Whilst the viral genes and products characterising latency have been well-

studied, triggers for the shift from latency to lytic replication and reactivation are not clearly defined. In general, reactivation disease is not believed to be a marked issue with EBV (unlike other herpesviruses, e.g. CMV) apart from in the context of post-transplant lymphoproliferative disorders.

5.1.5 Transcriptomic signatures of viral infection

GAinS dataset. Previous work done in the Knight laboratory by Dr Emma Davenport included the definition of gene expression signatures for different infection types in the GAinS CAP cohort /parenciteDavenport2014. Four contrasts were made with differentially expressed probes defined as those with a FDR <0.05 and a modest fold change >1.2. The contrasts made are summarised in Table 5.1.

Of interest is the viral vs no viral infection analysis, which identified 66 differentially expressed probes. These mapped to 54 genes in Ingenuity Pathway Analysis (IPA) with enrichment of pathways involving the role of pattern recognition receptors (PRRs) in recognition of viruses and the activation of interferon-regulatory factors (IRF) by cytosolic PRRs.

A prediction model for viral infection was defined using the GeneRave R package (CSIRO Bioinformatics R package version 3.0.8) (Kiiveri 2008). To reduce the number of possible predictors, only probes that were differentially expressed at

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

Table 5.1: Previous work: differential gene expression analysis for various classes of infection. Genome-wide gene expression data from microarray was analysed in patients with sepsis due to CAP. Differential expression analysis was performed for different microbiological classes of infection. Differentially expressed probes are defined as those with FDR <0.05 and fold change >1.2.

FDR <0.05 and fold change >1.2, mapped to genes in IPA and were moderate to highly expressed (expression >6.5) in a proportion of samples equating to the smallest group of the comparison were used. Six genes were selected for the prediction model: *IFI27*, *TGIF2*, *LY6E*, *CCNY*, *DYNLL2*, and *LAMP3*. Using leave-one-out cross-validation, a ROC curve (AUC 0.89) and MR (9.1) was determined for the model.

Limitations of this analysis include the lack of accurate microbiological phenotyping in the comparator ”non-viral” group. The electronic case record forms (eCRF) only provides information where testing yields a positive result, thus we are unable to confirm cases where testing has yielded a negative result. As viral testing is typically only performed during influenza outbreaks, there is a high likelihood the ”non-viral” group included cases of undiagnosed viral infection. This may explain why only 66 differentially expressed probes were identified at a modest fold change of 1.2.

Other datasets. A number of published studies have described gene signatures differentiating viral from bacterial infection. Two examples include the seven-gene set described by Sweeney *et al* (Sweeney et al. 2016) and the two-gene disease risk score (DRS) described by Herberg *et al* (Herberg et al. 2016).

Sweeney and colleagues (Sweeney et al. 2016) combined 8 datasets from adults and children with CAP, febrile illness and/or sepsis to form a discovery dataset of 426 individuals (n=142 viral; n=284 bacterial). Gene expression was analysed from whole blood or peripheral blood mononuclear cells. The authors performed differential expression analysis and identified 72 genes with a FC ¿2 and FDR ¡0.01. A greedy forward search method was then utilised to identify a set of seven genes which discriminated viral from bacterial infection. This included three ”viral” genes (/textitIFI27, JUP, LAX1) and four ”bacterial” genes (*HK3*, *TNIP1*, *GPAA1*, *CTSB*). ROC analysis was performed with an AUC value of 97% (95% CI: 89%-99%) in the discovery cohort. The gene set was subsequently validated in

a combined validation cohort consisting six datasets of 341 individuals (n=203 viral; n=138 bacterial). Here, it performed with an AUC value of 91% (95% CI 82%-96%).

Herberg and colleagues (Herberg et al. 2016) prospectively recruited febrile children presenting to hospital. Genome-wide gene expression analysis was performed on RNA extracted from whole blood. Differential gene expression analysis identified 285 probes differentially expressed at a FC ¿2 and FDR ¡0.05. The elastic net method (Zou 2005) was used to identify 38 probes from initial gene set. Elastic net is a variable selection algorithm that is an alternative to standard linear regression with least squares fitting, particularly suited to cases where the number of predictors greatly exceeds the number of observations. Elastic net combines the lasso and ridge regression methods of shrinkage, minimising the number of variables included (lasso) whilst also making the model less dependent on any one variable (ridge).

Subsequently, forward selection-partial least squares was used to eliminate highly correlated transcripts and a two transcript signature was identified. This involved a ratio of *IFI44L* expression to *FAM89A* expression, which the authors termed a disease risk score (DRS). ROC analysis was performed, yielding an AUC of 96.3% (95% CI 87.4%-100%) in the test dataset of 29 individuals and an AUC of 97.4% (95% CI 91.2%-100%) in the validation dataset of 51 individuals.

5.1.6 HLA

The extreme variability seen in the Major Histocompatibility Complex (MHC) region is believed to have arisen due to human-viral co-evolution. It is unsurprising therefore, that associations are seen between specific HLA alleles and infectious disease susceptibility, progression and outcome.

Notable examples include human immunodeficiency virus 1 (HIV-1), hepatitis

B virus (HBV) and hepatitis C virus (HCV) infection. In HIV-1 infection, Carrington and colleagues (Carrington et al. 1999) tested 63 alleles and found six (A*29, B*27, B*35, B*41, Cw*04 and Cw*12) to be significantly associated with disease progression in Caucasians. In addition, elevated *HLA-A* expression levels were found to be negatively correlated with control of HIV infection (Ramsuran et al. 2018). In HBV infection, Nishida and colleagues (Nishida et al. 2016) tested 144 alleles and found DQB1*06:01 to have the strongest association with susceptibility to chronic infection in Japanese individuals. Finally, in HCV infection, B*27 (Neumann-Haefelin et al. 2006) and B*57 (Kim et al. 2011) have both been found to be associated with a higher rate of viral clearance.

In addition, genome-wide association studies (GWAS) and genome-wide linkage studies (GWLS) have identified polymorphisms in the HLA regions to be associated with viral (HIV, HBV, HCV), bacterial (leprosy, tuberculosis), and parasitic (malaria, leishmaniasis, and schistosomiasis) infections (Blackwell et al. 2009). For example, the International HIV Controllers Study (Pereyra et al. 2010) identified over 300 genome-wide significant SNPs within the MHC to be associated with control of HIV-1 whilst the first GWAS of leprosy susceptibility reported associations with SNPs in six genetic loci, including the HLA-DR region (Zhang et al. 2009).

In summary, these examples in the literature suggest it would be of particular interest to investigate the association between specific HLA alleles and susceptibility to different microbiological classes of sepsis.

5.1.7 Aims

The overall aim of this chapter is to explore how improved resolution of microbiology in the sepsis cohort can be applied to transcriptomic and genomic-based approaches to understand the host response in sepsis. Specifically, the most

commonly identified bacterial (*Streptococcus pneumoniae*) and viral (influenza) infections will be studied as well as the most commonly reactivated virus (Epstein Barr virus). These will be related to Sepsis Response Signature endotype, total leukocyte gene expression and underlying host genotype (HLA type). Specific aims are as follows.

1. To characterise the extent and implications of EBV reactivation and integrate this with host transcriptomic data.
2. To investigate the association between *Streptococcus pneumoniae* bacterial load and sepsis endotype (SRS status).
3. To describe host transcriptomic signatures of viral infection, influenza infection, and *Streptococcus pneumoniae* infection and identify predictive gene sets for these infections.
4. To investigate the association between host genotype (HLA type) and susceptibility to different classes of infection.

5.2 Results

5.2.1 QC, normalisation and combination of microarray datasets

In this chapter, I analyse microarray gene expression data from the GAinS study processed in four batches over a six year period. These include samples from patients with sepsis due to CAP or faecal peritonitis (FP) enrolled through the GAinS study, recruited from 34 participating ICUs between 2005 and 2016. Serial samples were obtained on the first, and/or third, and/or fifth day of ICU admission. The total peripheral blood leukocyte population was isolated from whole blood for RNA extraction using the LeukoLOCK system (Invitrogen).

Dataset	Samples post-QC	CAP	FP	Cardiac
Radhakrishnan 2010	236	124 (39/45/40)	94 (37/34/23)	18 (6/6/6)
Davenport 2011	339	262 (130/86/46)	0	77 (39/0/38)
Burnham 2014	159	106 (42/42/22)	53 (25/15/13)	0
Burnham 2016	143	72 (24/24/24)	71 (23/24/24)	0
Total	**877**	**564**	**218**	**95**

Table 5.2: Summary of microarray datasets. Numbers in brackets refer to samples at each timepoint. The three timepoints for the sepsis patients are day 1, day 3 and day 5 after ICU admission. The three timepoints for the cardiac surgery patients are pre-operative, immediately post-operative, 24 hours post-operative.

All four of these datasets were generated by former DPhil students in the Knight lab: Radhakrishnan 2010, Dr Jayachandran Radhakrishnan; Davenport 2011, Dr Emma Davenport; Burnham 2014, Dr Katie Burnham; Burnham 2016, Dr Katie Burnham. Samples from cardiac surgery patients (sepsis control patients) were taken prior to and after their operation, and included in the two earlier cohorts. These samples were processed by Dr Eduardo Svoren (Barts and the London).

Genome-wide gene expression data was generated for all four cohorts using the Illumina Human-HT-12 v4 Expression BeadChip (47,231 probes). The data QC and normalisation was performed for each dataset in parallel. This was performed by myself for the Radhakrishnan 2010 and Davenport 2011 datasets, and by Dr Katie Burnham for the Burnham 2014 and Burnham 2016 datasets. Table 5.2 summarises the samples included in each dataset following QC. Following QC of each dataset, the four datasets were combined.

Radhakrishnan 2010 The Radhakrishnan 2010 dataset (n=264) (Radhakrishnan 2012) included 234 samples from 138 sepsis patients and 30 samples from 10 cardiac surgery patients. Twenty-two samples were excluded prior to data normalisation (3 samples from a patient subsequently discovered as failing to meet study inclusion criteria; 2 mislabelled samples; 5 technical replicates; 12 samples from cardiac surgery patients with missing consent forms).

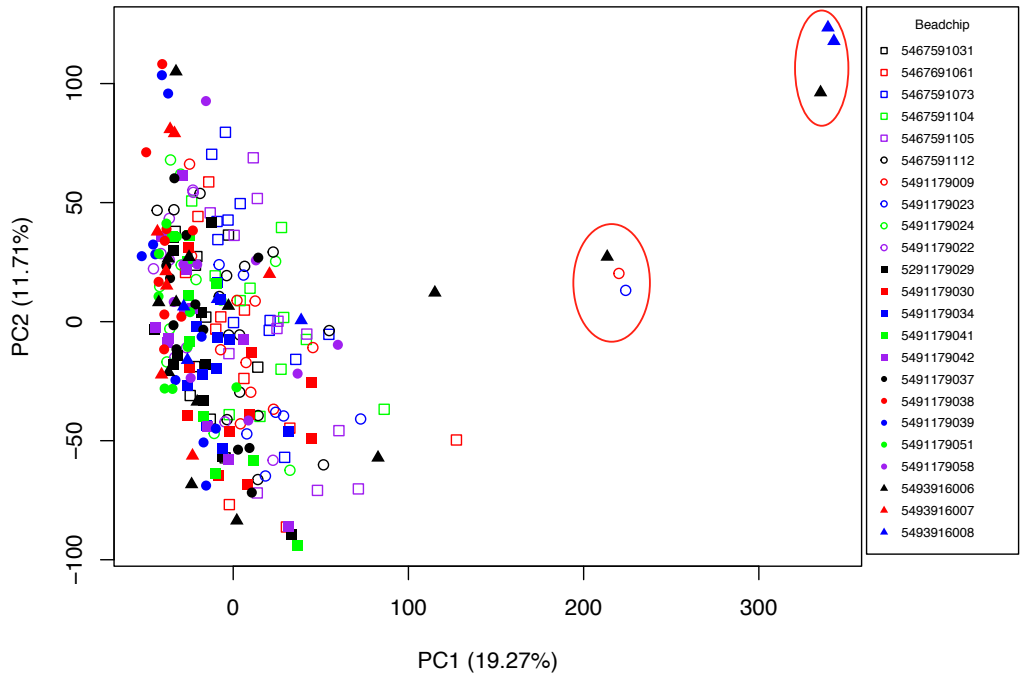
Six outlying samples were identified through Principal Component Analysis

(PCA and other QC measures (Figure 5.1) and excluded. Probes that did not have a detection p-value <0.05 in at least 5% of samples were removed (19,978 probes). Following normalisation and QC, expression data was available for 27,253 probes in 236 samples (124 CAP, 94 FP, 18 cardiac surgery) from 143 patients (73 CAP, 64 FP, 6 cardiac surgery).

Davenport 2011 The Davenport 2011 dataset has been published as the discovery cohort in Davenport et al. 2016. This initial dataset (n=432) included 306 samples from 306 CAP patients and 126 samples from 63 cardiac surgery patients. Eighty-five samples were excluded prior to data normalisation (48 samples from 4 chips with failed hybridisation; 34 samples from cardiac surgery patients with missing consent forms; 1 patient subsequently discovered as failing to meet study inclusion criteria; 1 patient who withdrew consent; 1 patient with suspected active leukaemia).

Eight outlying samples were identified through PCA and other QC measures (Figure 5.2) and excluded. Probes that did not have a detection p-value <0.05 in at least 5% of samples were removed (20,805 probes). Following normalisation and QC, expression data was available for 26,426 probes in 339 samples (262 CAP, 77 cardiac surgery) from 301 patients (262 CAP, 39 cardiac surgery).

(a) Principal Component Analysis of all samples



(b) Principal Component Analysis of 236 samples following outlier removal

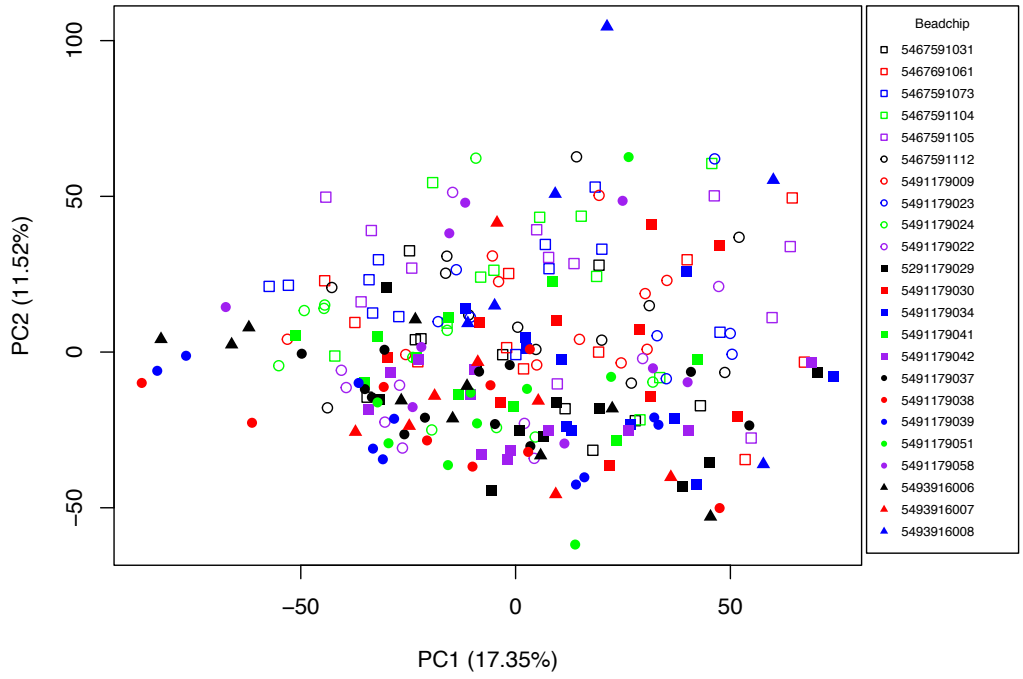


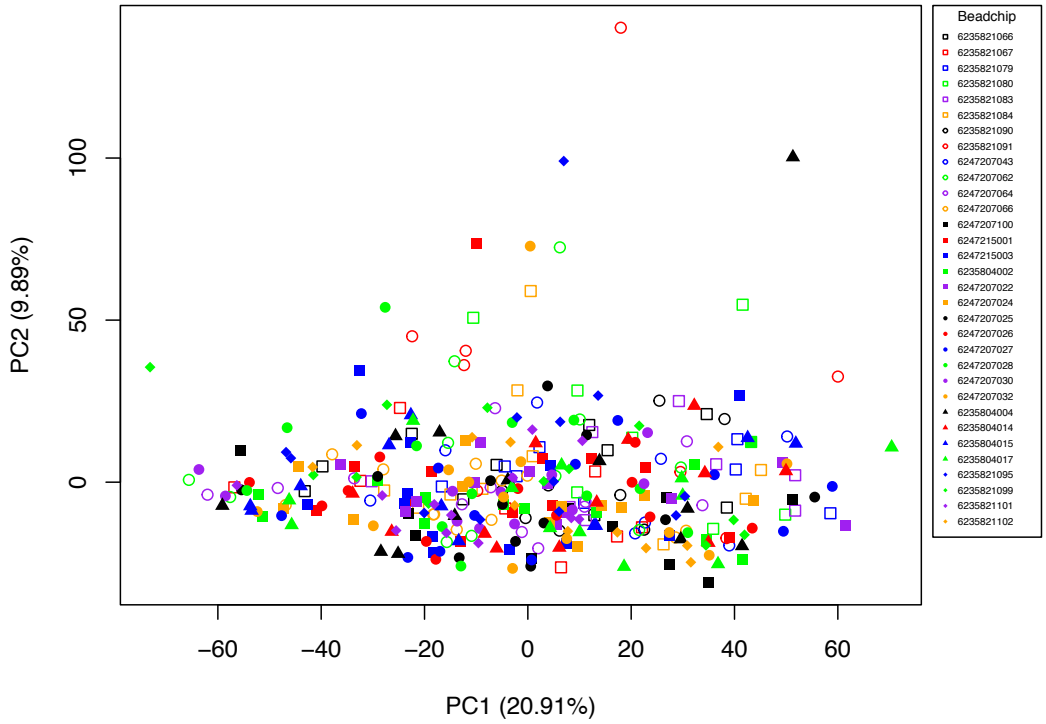
Figure 5.1: Principal Component Analysis: Radhakrishnan 2010 The first two principal components of the gene expression data are plotted, with the amount of variance explained by each noted; (a) when all samples are included in the analysis, and (b) after the removal of six outliers. Samples are coloured according to beadchip and outlying samples are circled in red.

Burnham 2014 Data QC and normalisation was carried out by Dr Katie Burnham (Burnham 2017). Post-QC, this dataset was comprised of 159 samples (106 CAP, 53 FP) from the same number of patients.

Burnham 2016 Data QC and normalisation was carried out by Dr Katie Burnham (Burnham 2017). Post-QC, this dataset was comprised of 143 samples (72 CAP, 71 FP) from 48 patients (24 CAP, 24 FP).

Combination of datasets Following separate QC and normalisation, the four datasets were combined. There were 92 additional probes in the two more recent datasets, which were removed. In addition, there were 61 samples that were repeated in the Davenport 2011 dataset following initial analysis in the Radhakrishnan dataset. Given that the SRS endotypes were defined and published for the Davenport 2011 dataset, the samples in this dataset were retained and duplicates in the Radhakrishnan 2010 dataset removed. Probes that did not have a detection p-value <0.05 in at least 5% of samples were removed (19,003 probes). Data was then normalised before the ComBat function from the R package sva applied to remove known batch effects (Figure 5.3). In this combined dataset, expression data was available for 28,228 probes for 816 samples (509 CAP, 218 FP, 89 cardiac surgery) from 591 patients (408 CAP, 141 FP, 42 cardiac surgery).

(a) Principal Component Analysis of all samples



(b) Principal Component Analysis of 339 samples following outlier removal

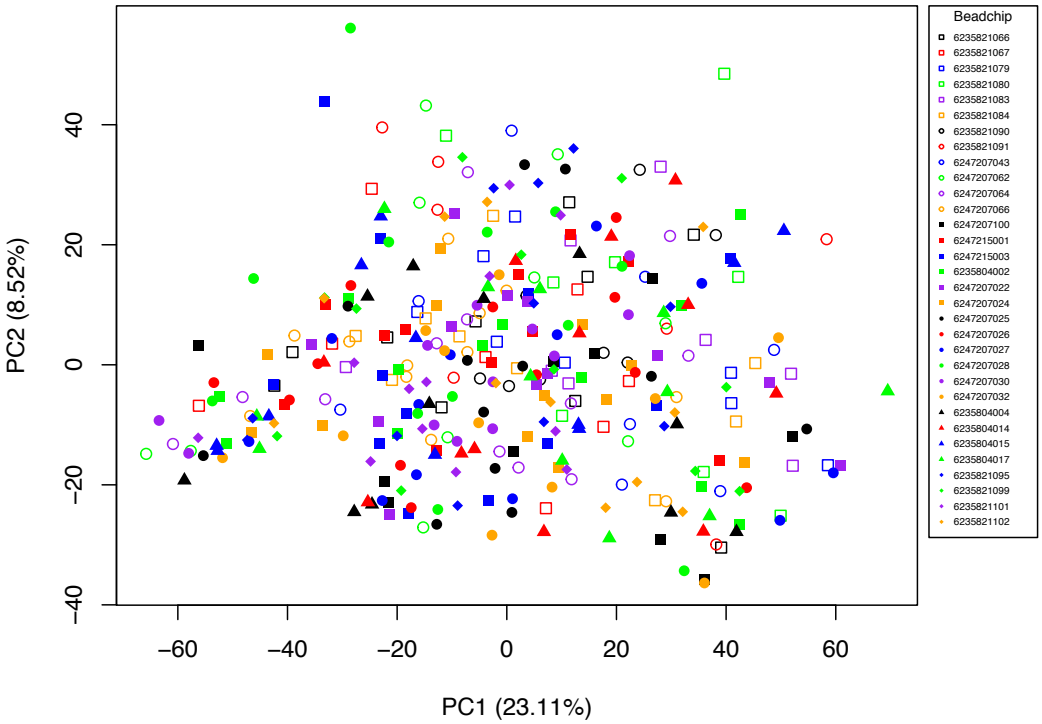


Figure 5.2: Principal Component Analysis: Davenport 2011 The first two principal components of the gene expression data are plotted, with the amount of variance explained by each noted; (a) when all samples are included in the analysis, and (b) after the removal of eight outliers. Samples are coloured according to beadchip and outlying samples are circled in red.

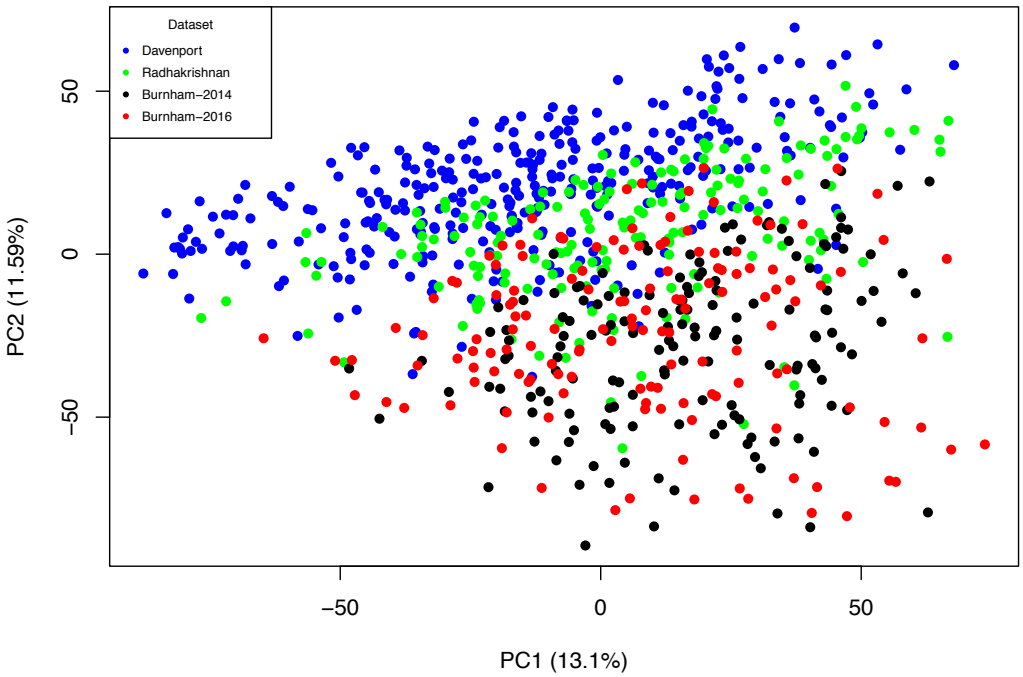
5.2.2 Evidence for immunosuppression: viral reactivation

Immunosuppression is a key pathophysiological process in sepsis, with viral reactivation seen as a consequence. Amongst 757 samples from 573 GAinS patients with CAP undergoing metagenomic sequencing (defined here as the metagenomic cohort), viral reactivation was observed in 24% of individuals, with EBV the most commonly observed virus (Table 5.3). Notably, this was observed despite there being an enrichment for earlier time points from ICU admission (Day 1=302; Day 3=284; Day 5=171). Six individuals showed simultaneous reactivation of EBV and a second virus.

Other reactivated virus	None	HSV	CMV	HHV-6	JC virus
EBV-negative (n=438)	422 (74%)	5 (0.9%)	1 (0.2%)	8 (1.4%)	2 (0.3%)
EBV-positive (n=135)	129 (23%)	4 (0.7%)	1 (0.2%)	0 (0%)	1 (0.2%)

Table 5.3: Incidence of viral reactivation in the metagenomic cohort (n=573). Six patients demonstrated reactivation with EBV and a second virus.

(a) Principal Component Analysis of all samples



(b) Principal Component Analysis following ComBat to remove batch effects

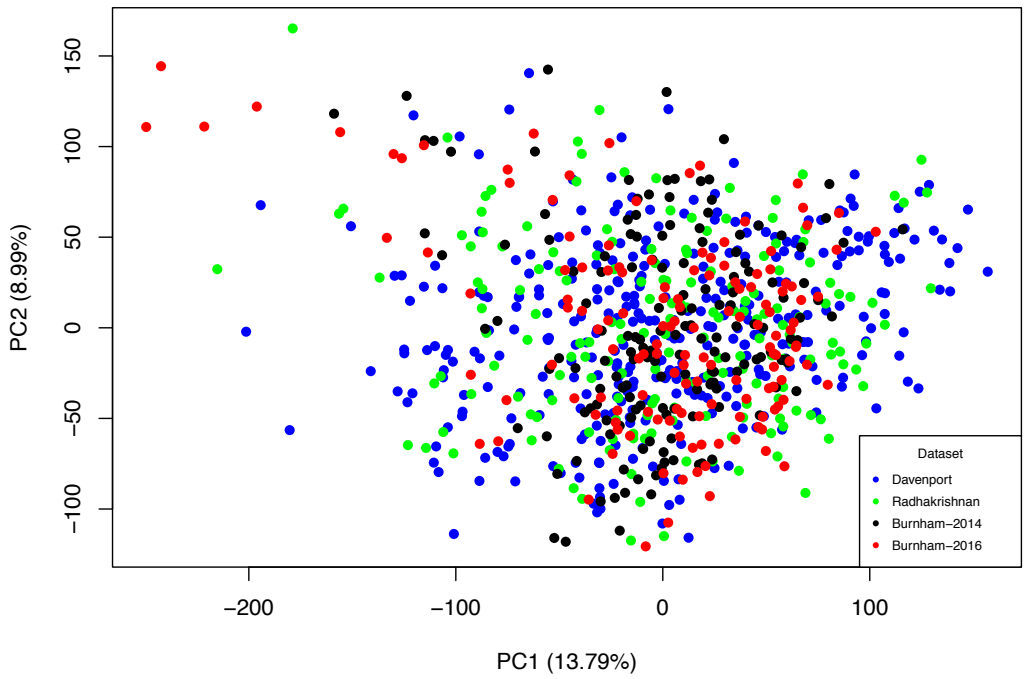


Figure 5.3: Principal Component Analysis: Combined dataset of 816 samples from GAinS CAP and FP patients. The first two principal components of the gene expression data are plotted, with the amount of variance explained by each noted; (a) prior to batch effects being removed (b) after removal of batch effects by ComBat. Samples are coloured according to the four datasets: Radhakrishnan (n=175), Davenport (n=339), Burnham-2014 (n=159), Burnham-2016 (n=143).

As the incidence of CMV reactivation was unexpectedly low, droplet digital PCR was carried out on a subset of samples overlapping with the metagenomic cohort (n=98). Samples were selected based on: (a) host transcriptomic data availability (overlap with Davenport 2011 cohort); (b) CMV positivity from metagenomics (n=1); and (c) sample availability. Of the 98 samples assayed, only two were positive for CMV including the one positive sample identified by metagenomics. The second sample (missed by metagenomics) was positive for CMV at a low level (86 copies/ml).

5.2.3 Evidence for immunosuppression: EBV

Cohort description. EBV was assayed from plasma samples taken at one or more time points (day 1 and/or day 3 and/or day 5 of ICU admission) by two methods, targeted metagenomics (757 samples from 573 patients) and digital droplet PCR (ddPCR) (619 samples from 565 patients), forming the two cohorts for this section. Targeted metagenomics enabled characterisation of multiple reactivated viruses in plasma, whilst ddPCR enabled a quantitative approach focused on EBV, enabling assessment of individuals from the metagenomic cohort at later time points as well as additional individuals. There was an overlap of 409 patients in the application of the two methods, with a total of 731 patients evaluated overall.

Serology. Plasma from 40 EBV ddPCR-positive patients, selected to represent the full range of viral loads observed, were tested for IgG and IgM antibodies against EBV Viral Capsid Antigen (VCA). All individuals tested (n=40; 100%) were positive for the IgG VCA antibody and negative for IgM VCA antibody, indicating that sepsis had coincided with viral reactivation and not primary infection.

Incidence of reactivation with time after ICU admission. Combining both

cohorts, a total of 1042 unique samples (731 patients) from days 1,3 and 5 after ICU admission were evaluated by either targeted metagenomics, ddPCR, or both. Where the ddPCR and metagenomic result differed (15% of samples), the positive result was taken forward for further analysis. The overall incidence of EBV reactivation at any time point was 37% (271/731). The incidence of EBV reactivation increased significantly over time (χ^2=24.6; d.f.=2; p=4.6 x 10^{-6}) (Table 5.4).

Clinical outcomes. Clinical outcome data was compared between EBV-negative (n=460) and EBV-positive (n=271) individuals (Table 5.5); patients were considered EBV-positive if at least one sample at any time point was positive by ddPCR or targeted metagenomics. Twenty-eight-day mortality was higher (27% vs 20%; p=0.04) and ICU length of stay was longer (12.9 days vs 9.2 days; p=0.004) in the EBV-positive group. In addition, the EBV-positive group had more organ failures with higher day 1 Sequential Organ Failure Assessment (SOFA) scores (6.9 vs 5.9; p=0.00011) and maximum SOFA scores (7.9 vs 6.7; p=3.6 x 10^{-6}).

Day of sampling	EBV negative	EBV positive	Total
1	245 (75%)	82 (25%)	327
3	247 (70%)	106 (30%)	353
5	209 (58%)	153 (42%)	362

Table 5.4: EBV status by day of sampling after ICU admission. GAinS patients with sepsis due to CAP had their EBV status evaluated by targeted metagenomic analysis or ddPCR of plasma samples obtained on days 1, 3 and/or 5 of ICU admission.

Characteristic	EBV-negative (n=460)	EBV-positive (n=271)	p-value
Age	60.5 (17-92)	62.4 (19-91)	0.11
Male sex^	275 (60%)	141 (52%)	0.049
Mortality (28-day)‘	92 (20%)	72 (27%)	0.040
ICU length of stay‘	9.2	12.9	0.0040
SOFA score (day 1)*	5.9	6.9	0.00011
SOFA score (maximum)*	6.7	7.9	3.6×10^{-6}

Table 5.5: Clinical characteristics and outcome data compared between EBV-negative and EBV-positive GAinS patients with sepsis due to CAP (total n=731). ICU length of stay analysis only includes patients surviving to ICU discharge. T-test performed except where indicated (^Chi-squared test; *Mann-Whitney U-test; ‘Log-rank test).

Association with SRS endotype. SRS group membership (determined at the first available time point after ICU admission) was evaluated in the context of EBV positivity over the first 5 days of ICU admission. Considering SRS endotype as a categorical trait, there was a greater proportion of SRS1 patients within the EBV-positive group (44% vs 35%; p=0.097) but this was not statistically significant (Table 5.6). However, this binary classification is a simplification of what is a continuous spectrum of gene expression patterns, with some individuals at either extreme and others with more moderate SRS signatures in the centre. Therefore, we reasoned that considering SRS endotype as a continuous trait might increase statistical power to detect an association with EBV-positivity.

Characteristic	EBV-negative (n=239)	EBV-positive (n=152)	p-value
Age	63.5 (19-91)	63.5 (17-91)	1.00
Male sexˆ	149 (62%)	88 (58%)	0.44
Sepsis Response Signature 1ˆ	84 (35%)	67 (44%)	0.097
Mortality (28-day)'	69 (29%)	53 (35%)	0.20
ICU length of stay'	10.7	13.4	0.070
SOFA score (day 1)*	6.2	7.0	0.019
SOFA score (maximum)*	7.1	8.1	0.0066

Table 5.6: Clinical characteristics and outcome data compared between EBV-negative and EBV-positive GAinS patients with sepsis due to CAP with gene expression data available.ICU length of stay analysis only includes patients surviving to ICU discharge. T-test performed except where indicated (ˆChi-squared test; *Mann-Whitney U-test; 'Log-rank test).

The difference in gene expression between SRS1 and SRS2 individuals can be observed in the first principal component (PC1) of principal component analysis (PCA) of the top 10% most variable genes (Figure 5.4). Thus, we compared PC1 between EBV-positive and EBV-negative individuals and found EBV-positive individuals had lower PC1 values, i.e. a more SRS1-like gene expression (Figure 5.5) (mean PC1 score -2.8 vs 1.8; p=0.014; t-test).

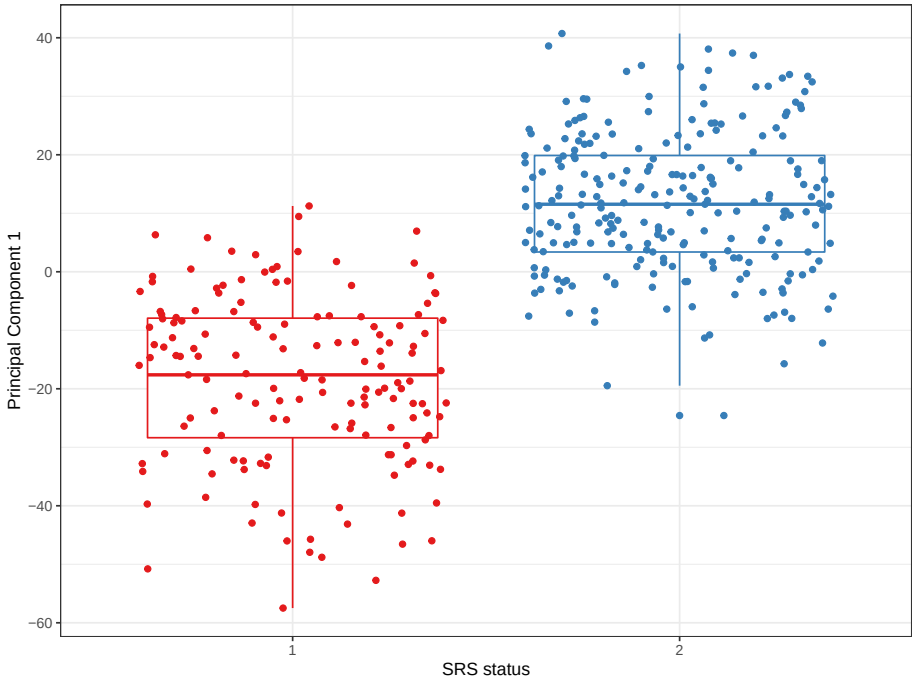


Figure 5.4: Principal Component 1 Principal component analysis was performed on the top 10% most variable genes expressed by peripheral blood leukocytes in GAinS patients with sepsis due to CAP. Individuals with both gene expression data and known EBV status (n=391) are included here with the first principal component (PC1) plotted against SRS endotype.

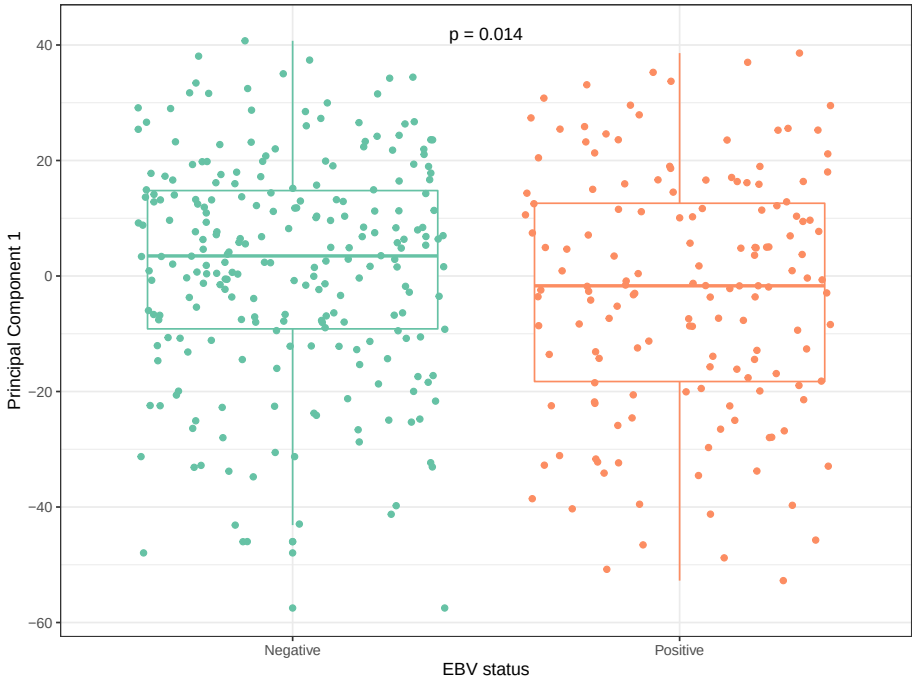


Figure 5.5: Boxplot of first available SRS as a continuous trait against EBV status over the first five days of ICU admission. GAinS patients with sepsis due to CAP for whom gene expression data and EBV status are avaiable are included here (n=391). The values for principal component 1 (PC1) provide a continuous measure of SRS endotype with higher values representing an SRS2 endotype and lower values representing an SRS1 endotype.

Levels of EBV viraemia. Digital droplet PCR was used to assay EBV in plasma samples (619 samples; 565 patients) over the first five days of admission (days 1,3, and/or 5). The maximum EBV load was related to SRS endotype determined from the first available timepoint after ICU admission (Figure 5.6). The median EBV load was higher in the SRS1 compared to the SRS2 endotype patients (211 vs 106 copies/ml; p=0.025; Mann-Whitney U-Test). Thus, among those in whom the virus was detected, the two SRS endotypes differed in the amount of EBV measured.

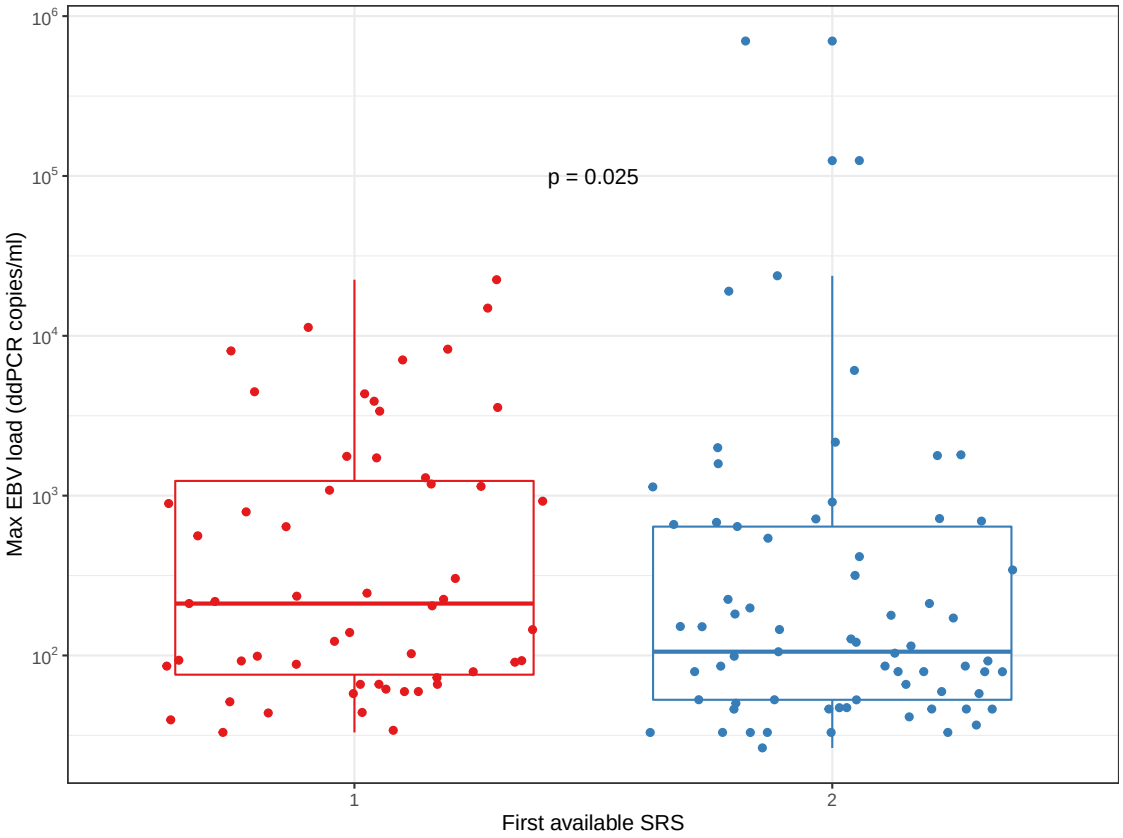


Figure 5.6: Boxplot of maximum EBV load over the first 5 days of ICU admission (ddPCR) against first available SRS status after ICU admission. GAinS patients with sepsis due to CAP were analysed with respect to EBV viral load by ddPCR and SRS endotype by genome-wide gene expression microarray analysis from peripheral blood leukocytes.

Gene expression signature. Global gene expression from the peripheral blood total leukocyte population was compared between EBV-positive and EBV-negative individuals (Figure 5.7). The differential expression analysis included

only gene expression from patient samples at the first available time point; 28228 probes were tested. Nine genes were differentially expressed at a fold-change >1.5 and FDR <0.05 (Table D.1). These include *CACNA2D3* (FDR 0.00521, fold change 1.55, downregulated in EBV-positive patients) which is a tumour suppressor gene downregulated in primary nasopharyngeal cancer and nasopharyngeal cell lines compared with non-tumorigenic cells (Wong et al. 2013) and *KIAA0101* (FDR 0.0128, fold change 1.51, upregulated in EBV-positive patients), an Epstein Barr Virus Nuclear Antigen 1 target gene (Satoh JK 2013).

Since SRS status is associated with EBV status and is therefore a confounding factor, we repeated the differential expression analysis, this time including SRS as a covariate in the linear model (Figure 5.8). Twelve genes (Table D.2) were found to be differentially expressed; there was an overlap of seven genes with the previous analysis.

Of the seven overlapping genes, there were several with a known role in EBV pathophysiology. These include *CDC20* (FDR 0.00451, fold change 1.53, upregulated in EBV-positive patients) which binds to EBV encoded proteins, activating the mitotic checkpoint and facilitating lytic EBV replication (Li et al. 2015) and *PRTN3* (FDR 0.0468, fold change 1.59, upregulated in EBV-positive patients) which has been observed to be co-expressed with the basigin gene, overexpressed in nasopharyngeal cancer (Gao et al. 2017).

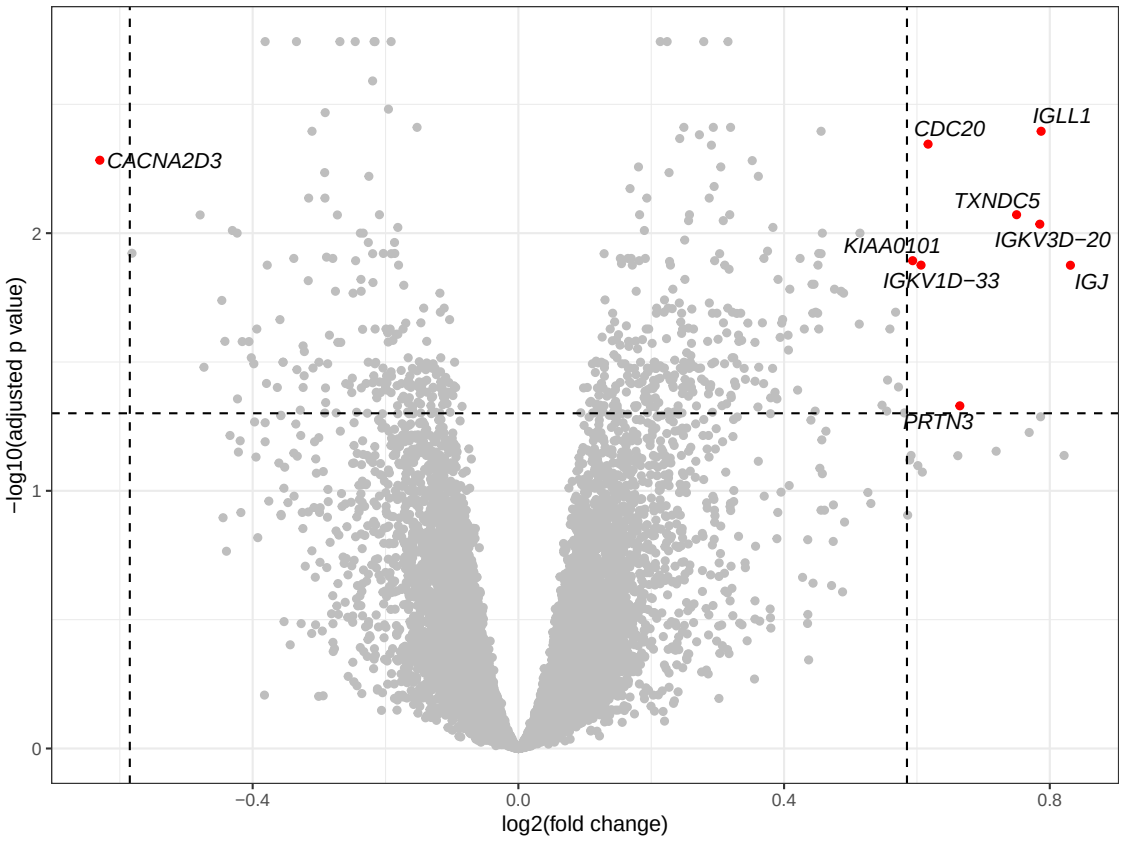


Figure 5.7: Volcano plot showing differentially expressed probes between EBV-positive and EBV-negative GAinS patients with sepsis due to CAP. Probes in red (labelled) are differentially expressed at a fold-change of 1.5 and p-value of 0.05. Positive fold change corresponds to upregulation in Epstein-Barr Virus-positive individuals.

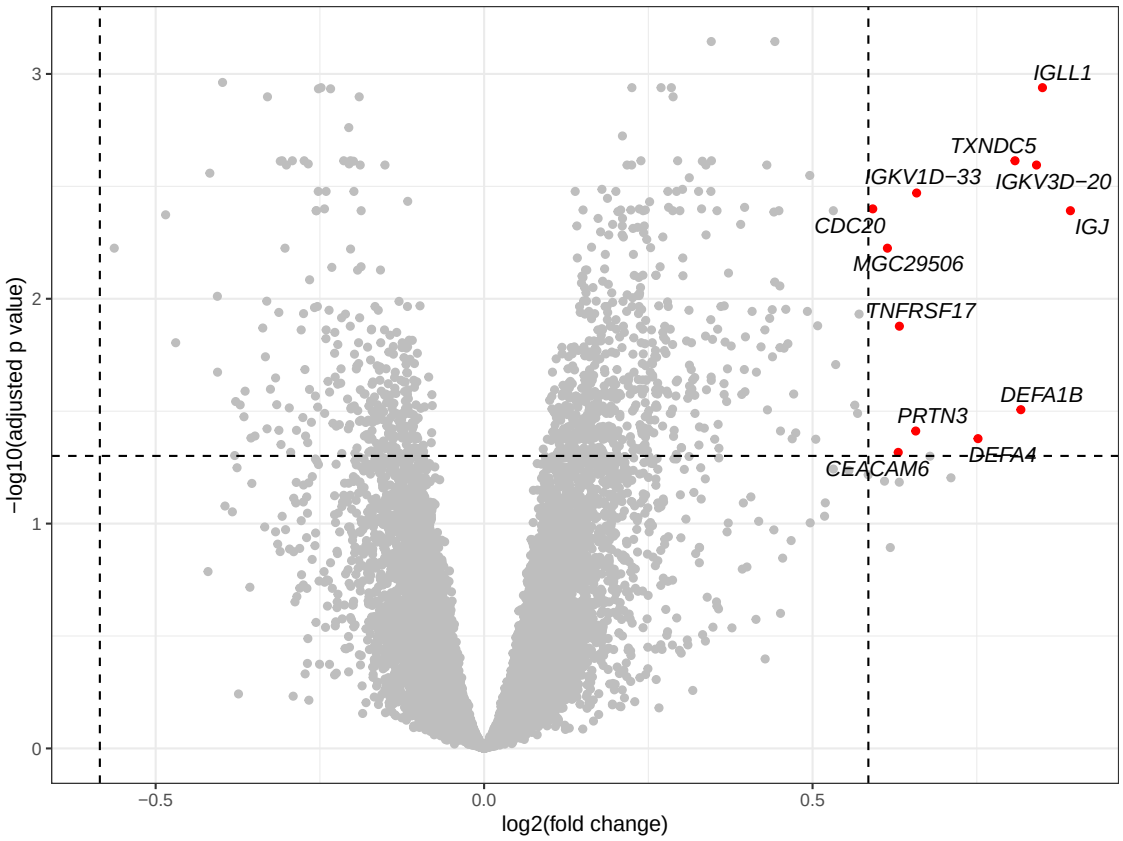


Figure 5.8: Volcano plot showing differentially expressed probes between EBV-positive and EBV-negative GAinS patients with sepsis due to CAP, with SRS endotype included in the linear model. Probes in red (labelled) are differentially expressed at a fold-change of 1.5 and p-value of 0.05. Positive fold change corresponds to upregulation in Epstein-Barr Virus-positive individuals.

5.2.4 Evidence for immunosuppression: *Streptococcus pneumoniae*

Of the 573 individuals in the metagenomic cohort, 109 had evidence of *Streptococcus pneumoniae* infection. This was comprised of diagnoses made by clinical microbiology (n=92) and new diagnoses made by *Castanet* (n=17). *S. pneumoniae* ddPCR data was available on 99/109 individuals. Of these 99 individuals, 38 were ddPCR positive for *S. pneumoniae*. Clinical characteristics were compared between the ddPCR positive (n=38) and ddPCR negative (n=61) groups (Table 5.7).

Characteristic	ddPCR-negative (n=61)	ddPCR-positive (n=38)	p-value
Age	58.1 (19-84)	57.8 (18-89)	0.94
Male sexˆ	33 (54%)	14 (37%)	0.14
SRS1ˆ	15/27 (56%)	12/21 (57%)	1.00
Charlson comorbidity index*	1.11	0.6	0.028
Mortality (28-day)‘	12 (20%)	6 (16%)	0.6
ICU length of stay‘	10.7	13.2	0.3
SOFA score (day 1)*	7.0	7.2	0.39
SOFA score (maximum)*	8.0	8.1	0.42
Mechanical ventilationˆ	13 (21%)	3 (7.9%)	0.14
Vasopressorsˆ	20 (33%)	10 (26%)	0.65

Table 5.7: Clinical characteristics and outcome data compared between ddPCR *S. pneumoniae*-negative and *S. pneumoniae*-positive GAinS patients with sepsis due to CAP. ICU length of stay analysis only includes patients surviving to ICU discharge. T-test performed except where indicated (ˆChi-squared test; *Mann-Whitney U-test; ‘Log-rank test).

There was no difference in outcome status of both groups in terms of 28-day mortality, ICU length of stay, SOFA score, mechanical ventilation and vasopressor use. Interestingly however, there was a difference in the Charlson comorbidity index between the two groups with *S. pneumoniae* positive patients showing lower levels of premorbid disease.

In addition, individuals with a high *S. pneumoniae* bacterial load ($\geq 10^3$ copies/ml) were compared with individuals with a bacterial load below this threshold (Table 5.8). There was no difference in the key outcomes of 28-day mortality, mechanical ventilation and vasopressor use.

Characteristic	Low/negative *S. pneumoniae* load (n=89)	High *S. pneumoniae* load (n=10)	p-value
Mortality (28-day)‘	15 (16.9%)	3 (30%)	0.3
Mechanical ventilationˆ	0 (0%)	10 (100%)	0.31
Vasopressorsˆ	62 (70%)	7 (70%)	1

Table 5.8: Clinical characteristics and outcome data compared between ddPCR *S. pneumoniae*-low/negative and ddPCR *S. pneumoniae*-high GAinS patients with sepsis due to CAP. Individuals with a bacterial load of $\geq 10^3$ were classified as having a high *S. pneumoniae* bacterial load. (ˆChi-squared test; ‘Log-rank test).

Given the findings with EBV, it was possible that levels of *Streptococcus pneumoniae* bacteraemia might also differ between the two SRS1 endotypes. For patients with *S. pneumoniae* infection diagnosed by clinical microbiology (n=92), ddPCR was used to assay *S. pneumoniae* in plasma samples over the first five days of admission (days 1,3, and/or 5) and the earliest value noted. For individuals with detectable bacterial load in plasma, this was evaluated against SRS endotype from the same day where expression data was available (n=18) (Figure 5.9). There was a statistically significant difference between the two groups, with SRS1 endotype patients showing a higher median *S. pneumoniae* load compared to SRS2 endotype patients (12293 vs 234 copies/ml; p=0.0022; Mann-Whitney U Test).

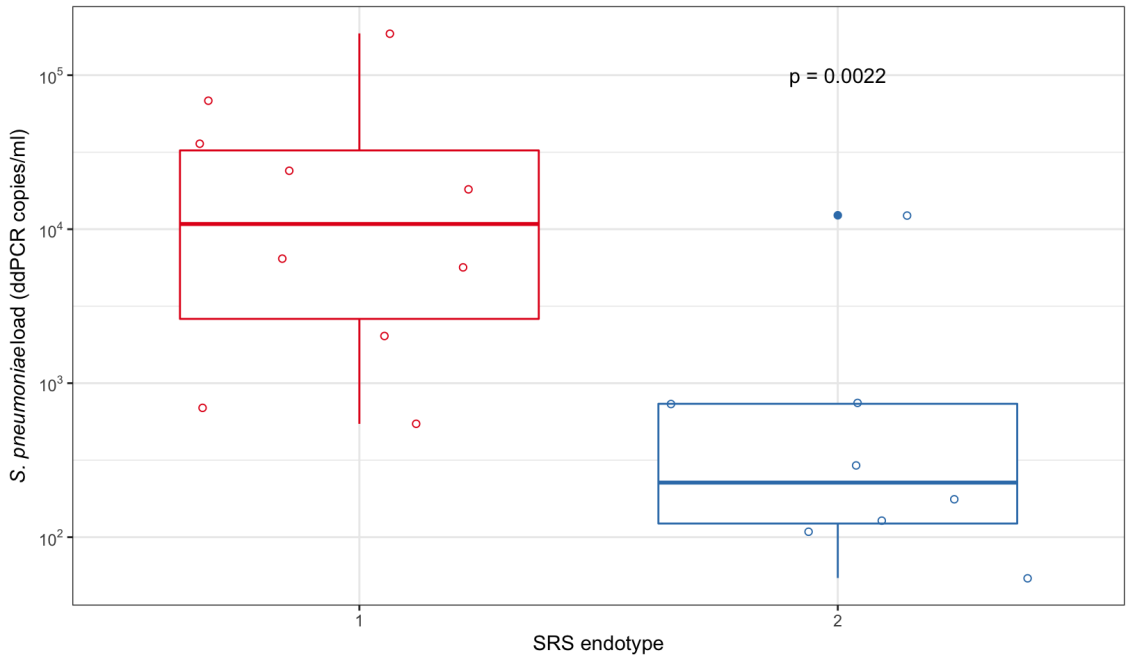


Figure 5.9: Boxplot of *Streptococcus pneumoniae* load and SRS status in GAinS patients with sepsis due to CAP. Bacterial load is assessed by ddPCR and SRS status is determined from genome-wide gene expression analysis of peripheral blood leukocytes by microarray. The highest bacterial load over the first five days of ICU admission is plotted against the SRS endotype from the first available time point after ICU admission.

Characteristic	Viral (n=30)	Bacterial (n=136)	p-value
Age	57.8 (25-83)	63.0 (18-92)	0.11
Male sexˆ	14 (47%)	78 (57%)	0.29
SRS1ˆ	9 (30%)	72 (53%)	0.038
Charlson comorbidity index*	0.8	1.1	0.52
Mortality (28-day)‘	9 (30%)	38 (28%)	0.3
ICU length of stay‘	17.7	13.4	0.3
SOFA score (day 1)*	6.1	7.1	0.12
SOFA score (maximum)*	7.3	8.2	0.33

Table 5.9: Viral vs bacterial infection: clinical characteristics and outcome data in GAinS patients with sepsis due to CAP. ICU length of stay analysis only includes patients surviving to ICU discharge. T-test performed except where indicated (ˆChi-squared test; *Mann-Whitney U-test; ‘Log-rank test).

5.2.5 Transcriptomic signature of viral infection

Of the 408 CAP patients with expression data following QC, there were 166 individuals with a diagnosis of either bacterial (n=136) or viral (n=30) infection. Microbiological phenotyping was based on clinical information from the eCRF and metagenomic data. Individuals with a mixed bacterial/viral infection, fungal infection or no positive microbiology were excluded from the comparison since they could not be accurately allocated to either the bacterial or viral category.

Clinical characteristics and outcome. Table 5.9 details the clinical characteristics and outcome data for the bacterial and viral groups. There was no significant difference in age, sex, baseline comorbidities, mortality, length of stay or SOFA score between the two groups. However, there were significantly more patients with an SRS2 endotype in the viral group compared with the bacterial group. PC1 was plotted against infection type (Figure 5.10), showing that patients with bacterial infection were evenly distributed across the spectrum of SRS1/SRS2 whilst patients with viral infection displayed a more SRS2-like endotype.

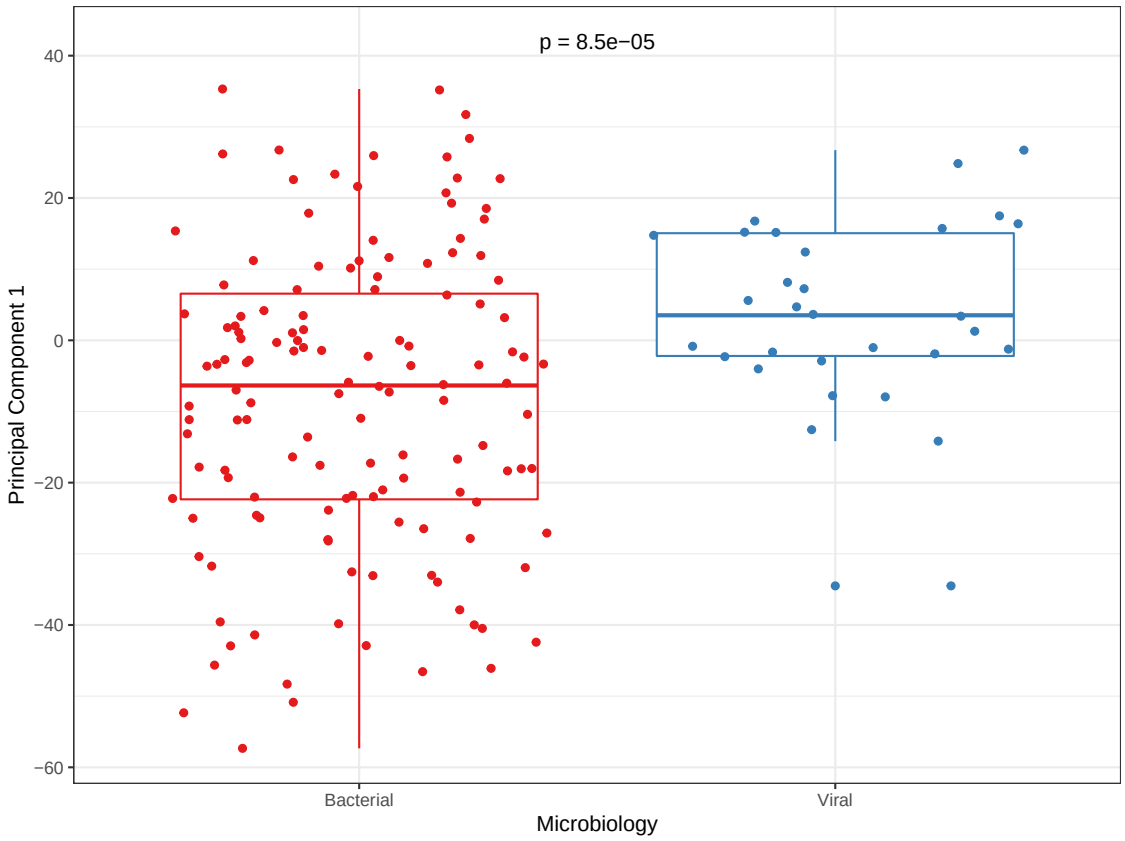


Figure 5.10: Boxplot of first available SRS as a continuous trait against microbiology in GAinS patients with sepsis due to CAP. SRS endotype is determined from genome-wide gene expression analysis of peripheral blood leukocytes by microarray. The values for principal component 1 (PC1) are derived from PCA of the top 10% most variable genes and provide a continuous measure of SRS endotype with higher values representing an SRS2 endotype and lower values representing an SRS1 endotype.

Differential gene expression. The gene expression of individuals with viral (n=30) and bacterial (n=136) infection was contrasted. This identified 206 differentially expressed probes (FDR <0.05 and fold change >1.5) with the majority upregulated in individuals with viral infection (Figure 5.23) (Table D.3).

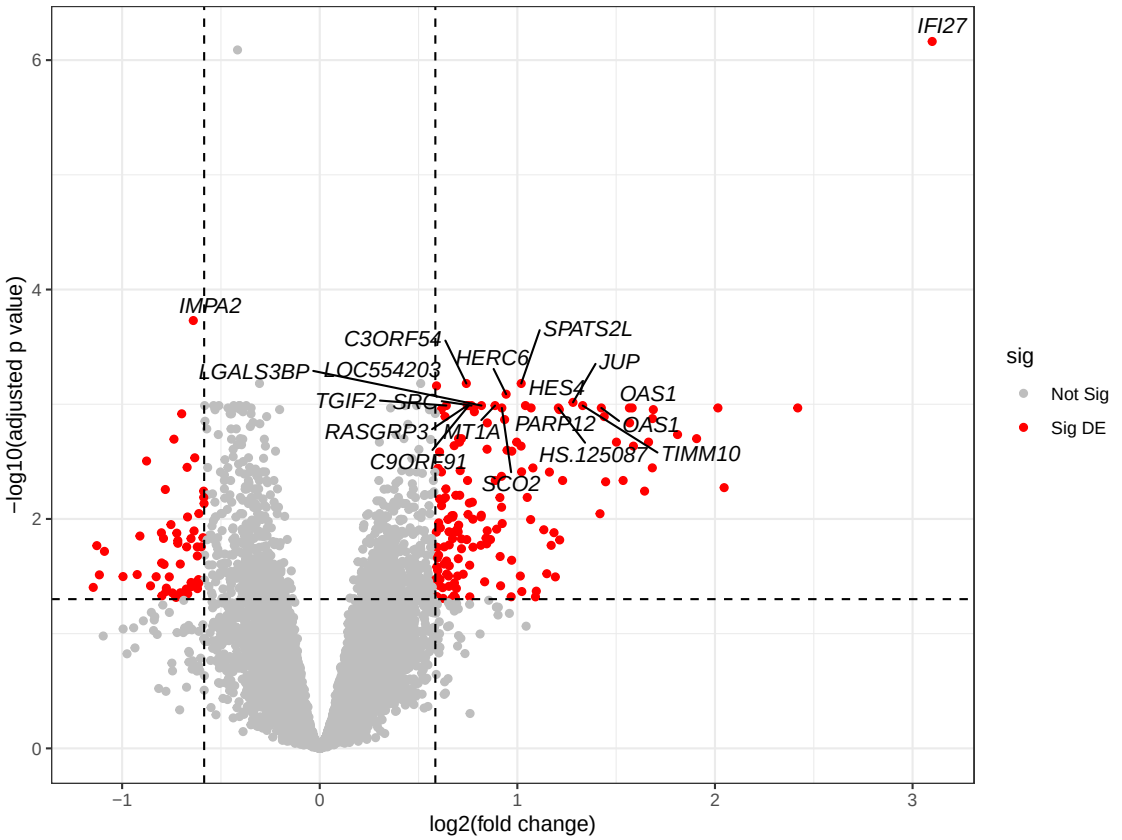


Figure 5.11: Volcano plot of differentially expressed probes in viral vs bacterial infection in GAinS patients with sepsis due to CAP. Probes in red are differentially expressed at a FDR <0.05 and fold change >1.5. Positive fold change corresponds to upregulation in individuals with viral infection. Top 20 significantly differentially expressed probes are labelled.

Pathway enrichment. Enrichment analysis was performed on the 206 probes identified as being differentially expressed using the R package XGR (Fang et al. 2016) and the Gene Ontology database (Consortium 2019) (Ashburner et al. 2000). A number of pathways specific to the immune response to viral infection were identified (Figure 5.12), including the type 1 interferon signalling pathway and the regulation of viral genome replication.

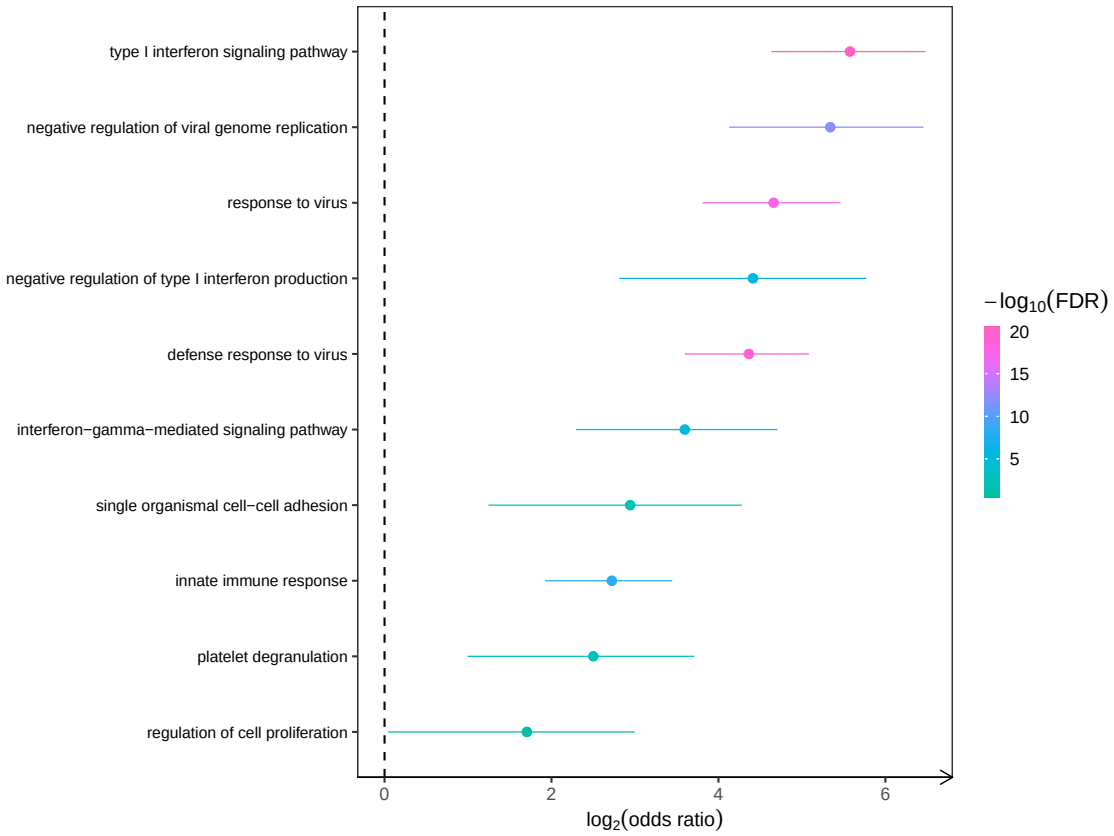


Figure 5.12: Pathway enrichment for viral infection in GAinS patients with sepsis due to CAP. The differentially expressed probes between patients with viral and bacterial infection were used to determine pathway enrichment with XGR. The top ten most enriched pathways using the Gene Ontology database are shown.

Predictive gene signature. Differentiating viral from bacterial CAP remains a challenge in the clinical setting. Therefore, a subset of genes was selected for a prediction model from the 206 probes differentially expressed between individuals with viral and bacterial infection.

The 166 individuals with bacterial (n=136) or viral (n=30) infection formed the training cohort for the model. Logistic regression with variable selection was applied to the 206 differentially expressed probes (FDR <0.05, fold change >1.5) using the elastic net method (Zou 2005) (Herberg et al. 2016).

Ten probes, corresponding to ten genes were selected by the model: *BTBD11*, *C3ORF54*, *IFI27*, *IMPA2*, *MT1A*, *RNASE1*, *SIGLEC10*, *SRC*, *TIMM10*, *TSPAN13*. In the training dataset, the signature had an AUC of 88.6% (95% CI 80.7%-96.5%) (Figure 5.14). Youden's method was used to select a threshold (0.2017) for discriminating bacterial from viral infection. At this threshold, 114/136 bacterial infections were correctly predicted (specificity 83.8% [95% CI 77.2%-89.7%]) whilst 25/30 viral infections were correctly predicted (sensitivity 83.3% [95% CI 70.0%-96.7%]) (Figure 5.13). This equated to a misclassification rate of 16.3%.

Due to limitations in data availability, it was not possible to select a single validation cohort which included both bacterial and viral infections. Therefore two validation cohorts were used, MOSAIC and VANISH. The Mechanisms of Severe Acute Influenza Consortium (MOSAIC) cohort included 109 adults with confirmed influenza while the Vasopressin vs Norepinephrine as Initial Therapy in Septic Shock (VANISH) cohort included 24 adults with predominantly bacterial sepsis requiring vasopressors.

For the combined validation cohort, the signature had an AUC of 97.1% (95% CI 94.6%-99.5%) (Figure 5.14). Using the threshold selected for the training cohort, 20/23 bacterial infections were correctly predicted (specificity 87.0% [95% CI 73.8%-100.0%]) whilst 100/110 viral infections were correctly predicted (sensitivity 90.1% [95% CI 85.5%-96.4%]) (Figure 5.13). This equated to a

misclassification rate of 9.8%.

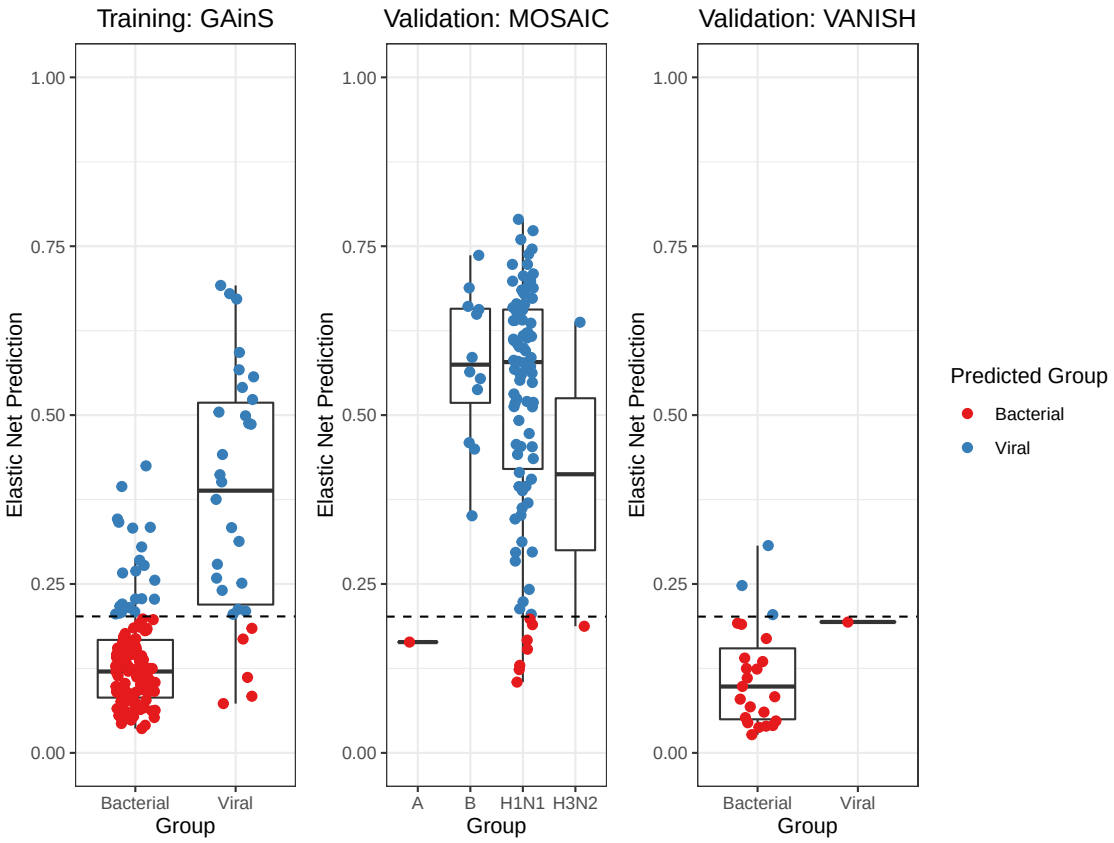


Figure 5.13: Boxplot of predictions from viral vs bacterial gene signature in GAinS patients with sepsis due to CAP. Each point corresponds to a patient. Actual microbiological groups are plotted on the x-axis whilst predicted classifications are denoted by point colour. The MOSAIC groupings correspond to influenza virus types. Threshold for discriminating bacterial from viral infection is denoted by the dashed line. The elastic net prediction value (y-axis) can range from 0 (indicating bacterial infection) to 1 (indicating viral infection).

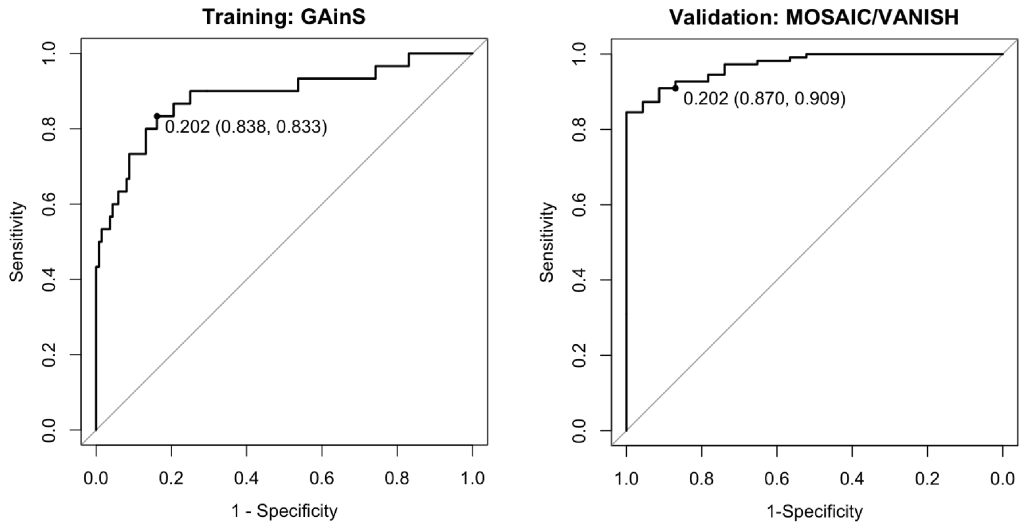


Figure 5.14: ROC analysis for viral vs bacterial signature in GAinS patients with sepsis due to CAP. The threshold for discriminating bacterial from viral infection is noted (0.202), together with the specificity and sensitivity at that threshold in brackets. AUC for the training cohort was 88.6% (95% CI 80.7%-96.5%) whilst AUC for the combined validation cohort was 97.1% (95% CI 94.6%-99.5%). MR was 16.3% for the training cohort and 9.8% for the validation cohort.

5.2.6 Previously derived transcriptomic signatures of viral infection

Sweeney seven-gene set. The previously described seven-gene set derived by Sweeney and colleagues (Sweeney et al. 2016) was applied to 166 GAinS patients (n=136 with bacterial infection; n=30 with viral infection). This involved calculating a composite score for each individual derived from the geometric mean of viral genes minus the geometric mean of bacterial genes, multiplied by the ratio of viral genes to bacterial genes (3/4). ROC analysis was performed, with an AUC of 81% (95% CI 71%-90%). Youden's method was used to select a threshold for discriminating viral from bacterial infection (Figure 5.15). At the selected threshold of -1.76, 130/136 bacterial infections were correctly predicted (specificity 95.6% [95% CI 91.9%-98.5%]) whilst 18/30 viral infections were correctly predicted (sensitivity 60.0% [95% CI 43.3%-76.7%) (Figure 5.16). This equated to a misclassification rate of 10.8% .

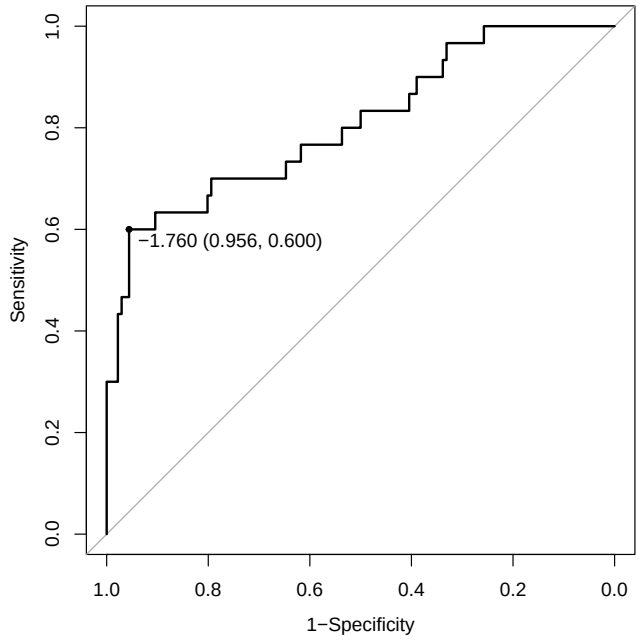


Figure 5.15: ROC analysis for Sweeney seven-gene set. The threshold for discriminating bacterial from viral infection is noted (-1.76), together with the specificity and sensitivity at that threshold in brackets. AUC for the GAinS cohort was 81% (95% CI 71%-90%).

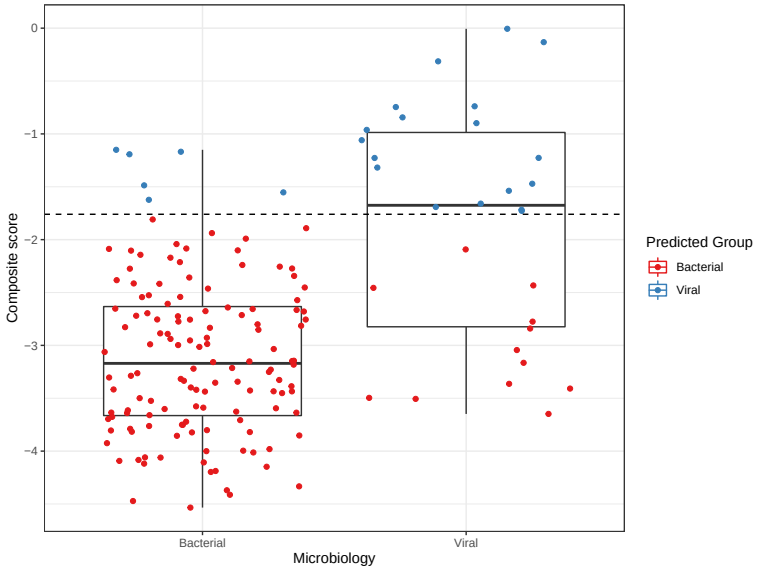


Figure 5.16: Boxplot of predictions from Sweeney seven-gene set. Each point corresponds to a patient. Actual microbiological groups are plotted on the x-axis whilst predicted classifications are denoted by point colour. Threshold for discriminating bacterial from viral infection (-1.76) is denoted by the dashed line. The composite score (y-axis) is derived from calculating the geometric mean of viral genes minus the gemoetric mean of bacterial genes, multiplied by the ratio of viral genes to bacterial genes (3/4).

Herberg two-transcript disease risk score. This previously described two-transcript ratio of *IFI44L* to *FAM89A* expression was applied to the same group of 166 GAinS patients (n=136 with bacterial infection; n=30 with viral infection). ROC analysis was performed with an AUC of 79.0% (95% CI 70.0%-88.0%). Youden's method was used to select a threshold for discriminating viral from bacterial infection (Figure 5.17). At the selected threshold of -4.49, 117/136 bacterial infections were correctly predicted (specificity 86.0% [95% CI 80.2%-91.2%]) whilst 20/30 viral infections were correctly predicted (sensitivity 66.7% [95% CI 50.0%-83.3%]) (Figure 5.18). This equated to a misclassification rate of 17.5%.

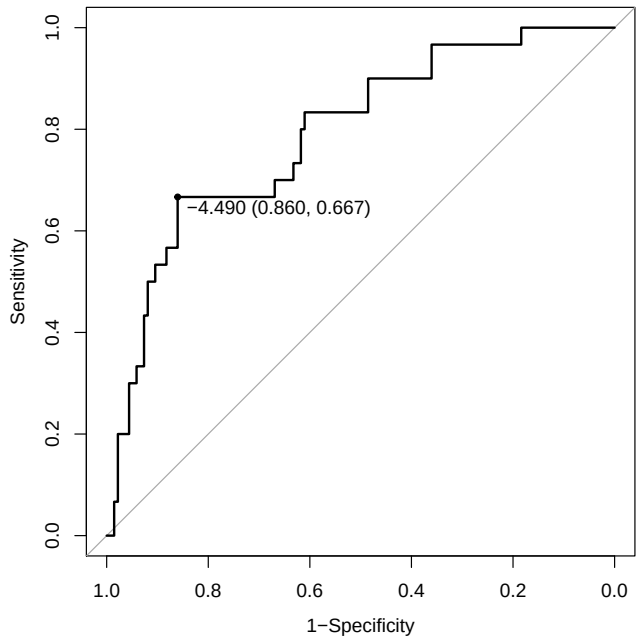


Figure 5.17: ROC analysis for Herberg disease risk score. The threshold for discriminating bacterial from viral infection is noted (-4.49), together with the specificity and sensitivity at that threshold in brackets. AUC for the GAinS cohort was 79% (95% CI 70%-88%).

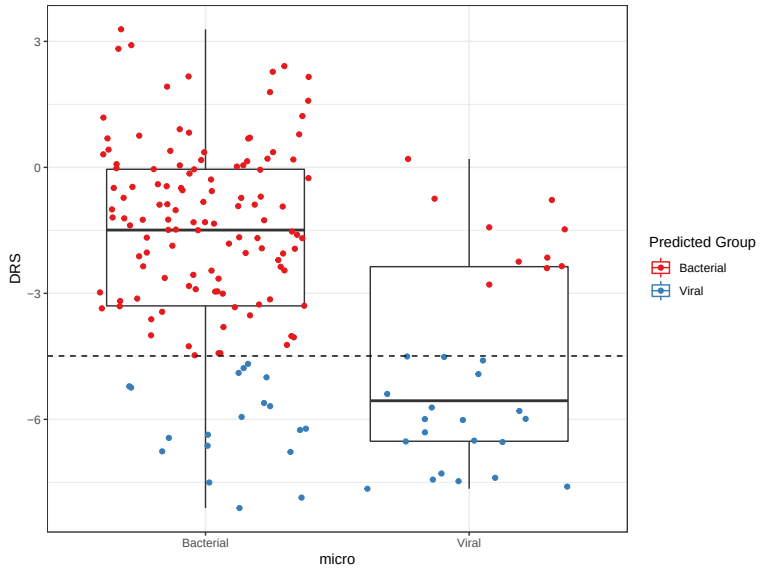


Figure 5.18: Boxplot of predictions from Herberg disease risk score. Each point corresponds to a patient. Actual microbiological groups are plotted on the x-axis whilst predicted classifications are denoted by point colour. Threshold for discriminating bacterial from viral infection (-4.49) is denoted by the dashed line. The composite score (y-axis) is derived from the disease risk score of *FAM89A* expression minus *IFI44L* expression. DRS=disease risk score

5.2.7 Transcriptomic signature of influenza infection

Differential gene expression. Two thirds of the viral cohort in the preceding section had influenza infection. The gene expression of these individuals (n=20) was contrasted with that of individuals with bacterial infection (n=136). This identified 139 differentially expressed probes (FDR <0.05 and fold change >1.2) (Figure 5.19) (Table D.4).

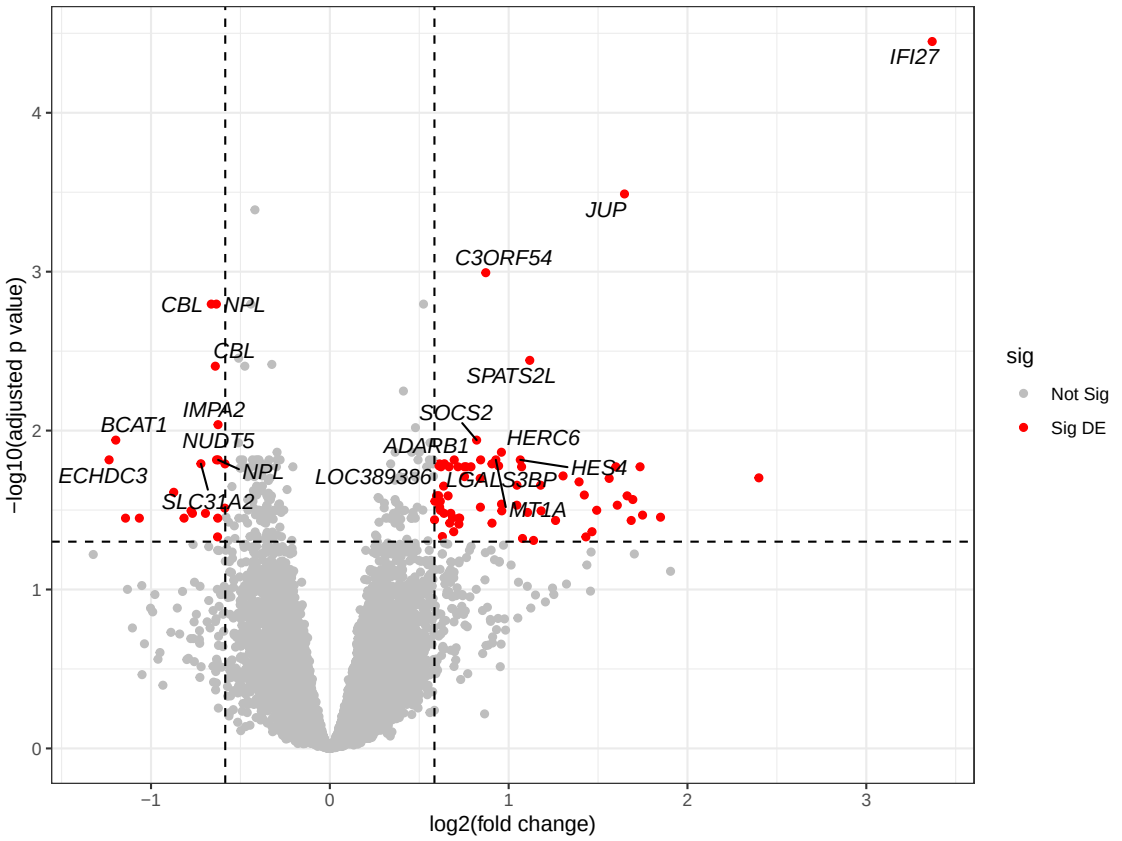


Figure 5.19: Volcano plot of differentially expressed probes in influenza vs bacterial infection for GAinS patients with sepsis due to CAP. Probes in red are differentially expressed at a FDR <0.05 and fold change >1.5. Positive fold change corresponds to upregulation in influenza-positive individuals. Top 20 significantly differentially expressed probes are labelled.

Pathway enrichment. Enrichment analysis was performed on the 139 probes identified as being differentially expressed using the R package XGR (Fang et al. 2016) and the Gene Ontology database (Consortium 2019) (Ashburner et al. 2000). Similar pathways identified in the viral vs bacterial analysis were identified again here (Figure 5.20), including the type 1 interferon signalling pathway, the regulation of viral genome replication, and viral response.

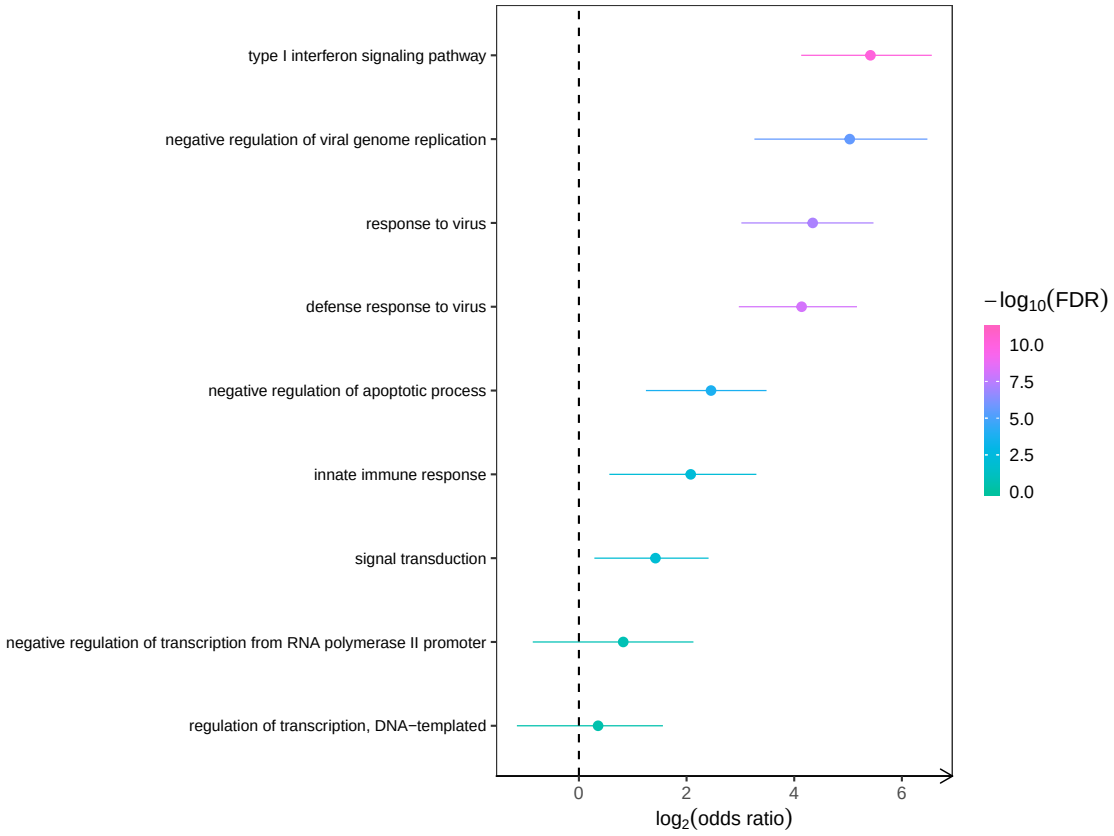


Figure 5.20: Pathway enrichment for influenza infection in GAinS patients with sepsis due to CAP. The differentially expressed probes between patients with influenza and bacterial infection were used to determine pathway enrichment with XGR. The top ten most enriched pathways using the Gene Ontology database are shown.

Predictive gene signature. The elastic net method was used again here to select a subset of predictive genes from the 139 differentially expressed probes. The training dataset comprised 156 individuals from the GAinS cohort, 20 with influenza infection and 136 with bacterial infection.

Seven probes corresponding to seven genes were selected by the elastic net model: *IFI27*, *JUP*, *C3ORF54*, *NPL*, *CBL*, *UBQLNL*, *LOC401845*. In the training

dataset, the signature had an AUC of 90.1% (95% CI 80.4%-99.8%) (Figure 5.22). Youden's method was used to select a threshold (0.182) for discriminating bacterial from influenza infection. At this threshold, 123/136 bacterial infections were correctly predicted (specificity 90.4% [95% CI 85.3%-94.9%]) whilst 16/20 viral infections were correctly predicted (sensitivity 80% [95% CI 60%-95%]) (Figure 5.21). This equated to a misclassification rate of 10.9%.

The two validation cohorts (MOSAIC and VANISH) were used again here with the MOSAIC cohort comprising 109 individuals with influenza infection and the VANISH cohort comprising 23 patients with bacterial sepsis and 1 patient with influenza sepsis.

For the combined validation cohort, the signature had an AUC of 92.9% (95% CI 88.5%-97.4%) (Figure 5.22). Using the threshold identified in the training cohort, 20/23 bacterial infections were correctly predicted (specificity 87% [95% CI 73.9%-100%]) whilst 89/110 viral infections were correctly predicted (sensitivity 80.9% [95% CI 73.6%-88.2%]) (Figure 5.21). This equated to a misclassification rate of 18%.

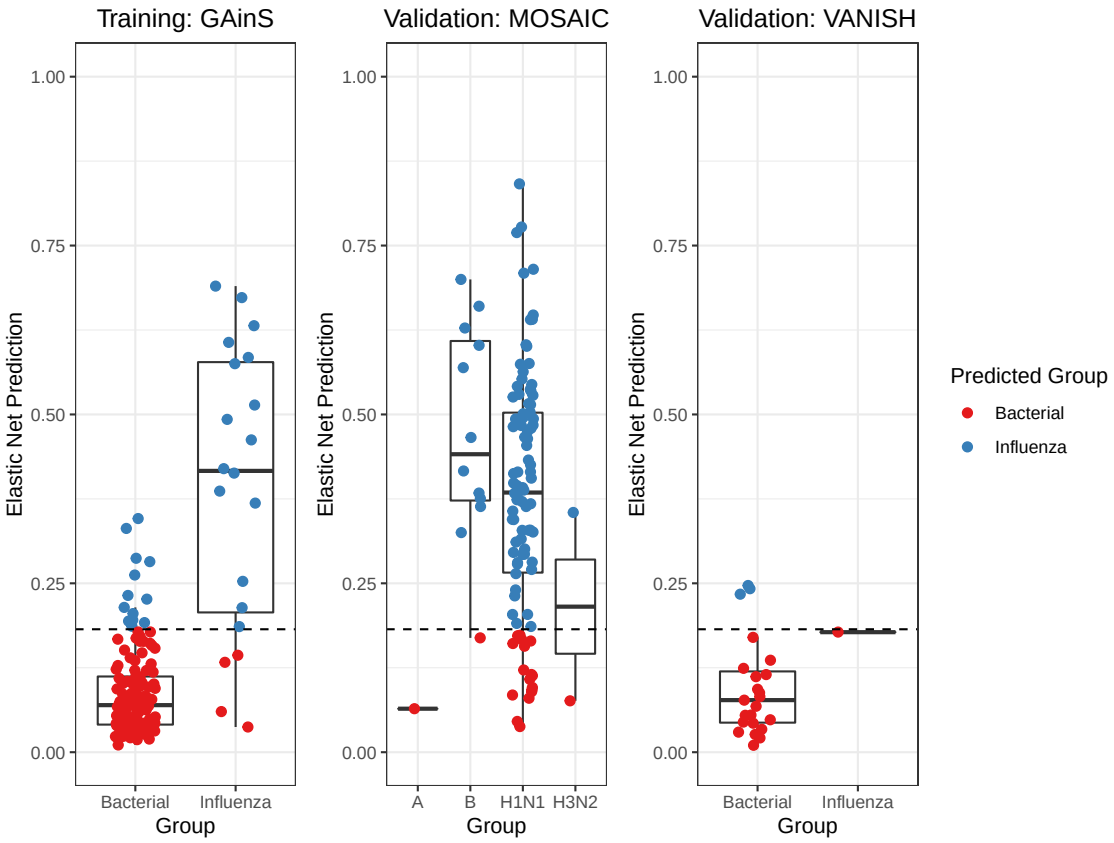


Figure 5.21: Boxplot of predictions from influenza vs bacterial gene signature in GAinS patients with sepsis due to CAP. Each point corresponds to a patient. Actual microbiological groups are plotted on the x-axis whilst predicted classifications are denoted by point colour. The MOSAIC groupings correspond to influenza virus types. Threshold for discriminating bacterial from influenza infection is denoted by the dashed line. The elastic net prediction value (y-axis) can range from 0 (indicating bacterial infection) to 1 (indicating influenza infection).

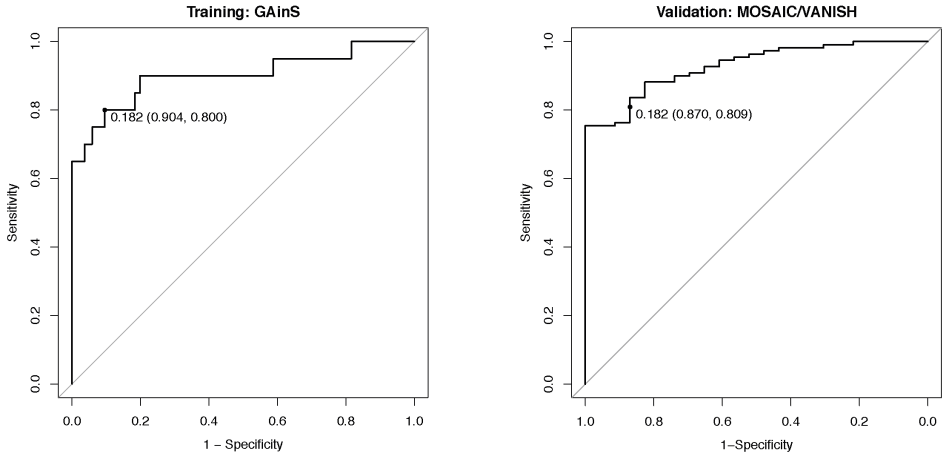


Figure 5.22: ROC analysis for influenza vs bacterial signature. The threshold for discriminating bacterial from influenza infection is noted (0.182), together with the specificity and sensitivity at that threshold in brackets. AUC for the training cohort was 90.1% (95% CI 80.4%-99.8%) whilst AUC for the combined validation cohort was 92.9% (95% CI 88.5%-97.4%). MR was 10.9% for the training cohort and 18% for the validation cohort.

5.2.8 Transcriptomic signature of influenza vs viral infection

Gene expression of individuals with influenza infection (n=20) was contrasted with that of individuals with non-influenza viral infection (n=10). There were no differentially expressed probes at FDR <0.05 and fold change >1.5.

5.2.9 Transcriptomic signature of *Streptococcus pneumoniae* infection

Total leukocyte gene expression was compared between individuals with *S. pneumoniae* infection (n=50) diagnosed by either clinical microbiology or metagenomics and individuals without *S. pneumoniae* infection (n=113). The comparator group included individuals with viral infections and non-pneumococcal bacterial infections. Three patients co-infected with *S. pneumoniae* and another organism were excluded from the analysis.

Differential expression analysis revealed no differentially expressed genes

between the two groups (FDR <0.05 and fold change >1.5).

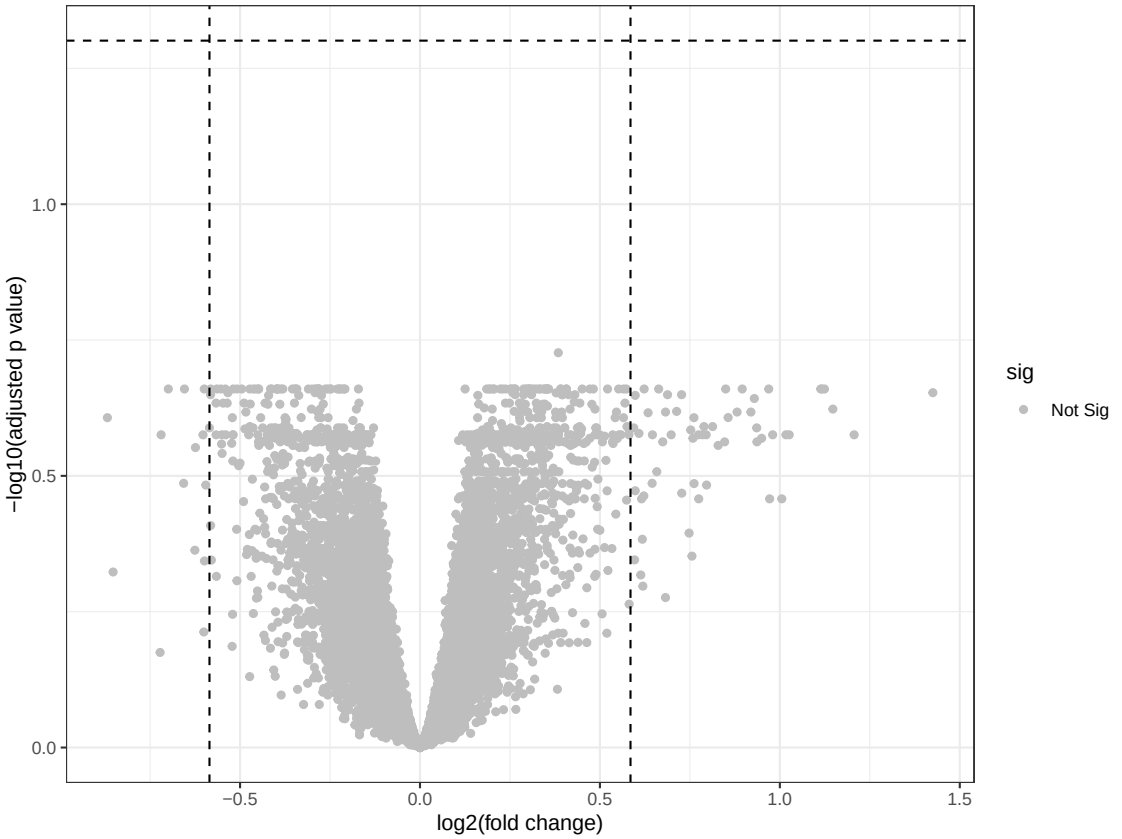


Figure 5.23: Volcano plot of differentially expressed probes in *S. pneumoniae* vs non-*S. pneumoniae* infection for GAinS patients with sepsis due to CAP. There are no differentially expressed probes at an FDR <0.05 and fold change >1.5.

5.2.10 Summary of differential expression analyses

The differential gene expression analyses described in the previous sections are summarised in the following table (Table 5.10).

5.2.11 Genomics: association with HLA

For this section, the training dataset included 613 GAinS patients with sepsis due to CAP with HLA imputed from genotyping data or HLA typing from sequencing data whilst the validation dataset included 274 individuals genotyped separately from the training cohort, with HLA imputation from the genotyping data. Alleles

Comparison (no. of patients)	Patients	Differentially expressed probes
Viral (30) vs bacterial (136)	166	206
Influenza (20) vs bacterial (136)	156	139
Influenza (20) vs viral (10)	30	0
S. pneumoniae (50) vs non-S. pneumoniae (113)	163	0

Table 5.10: Summary of differential expression analysis. Differentially expressed probes defined as those with FDR <0.05 and fold change >1.5.

were tested for association with the phenotype of interest if they were present in the training cohort at a prevalence of ≥2%.

Influenza. In the training cohort (n=613), there were 30 individuals with and 583 individuals without a diagnosis of influenza sepsis. Sixty-two HLA class I (HLA-A, HLA-B, HLA-C) and class II (HLA-DQA1, HLA-DQB1, HLA-DRB1) alleles were tested for association at 2-digit resolution. After correction for multiple testing, two alleles (B*35 and C*12) had a statistically significant association with influenza sepsis (Chi-squared test; adjusted p=0.00152 and adjusted p=0.00214 respectively).

The two alleles were subsquently tested for association with influenza sepsis in the validation cohort (n=274) where there were 33 individuals with and 241 individuals without a diagnosis of influenza sepsis. Here, there was a statistically significant association for B*35 (p=0.0256) but not for C*12 (p=0.352 in the opposite direction) with the presence of influenza sepsis.

HLA-B*35 is present in the England Leeds cohort (n=5024) at a prevalence of 12.7% (*Allele Frequencies in Worldwide Populations*). This was similar to the prevalence of 12.6% observed in the GAinS training cohort (n=76/611 heterozygous, 1/611 homozygous). There was an increased prevalence of the allele in the influenza positive individuals in the training cohort (n=6/30 heterozygous, 1/30 homozygous; 23.3%) as well as the validation cohort (n=4/33 heterozygous, 1/33 homozygous; 15.2%) (Figure 5.24).

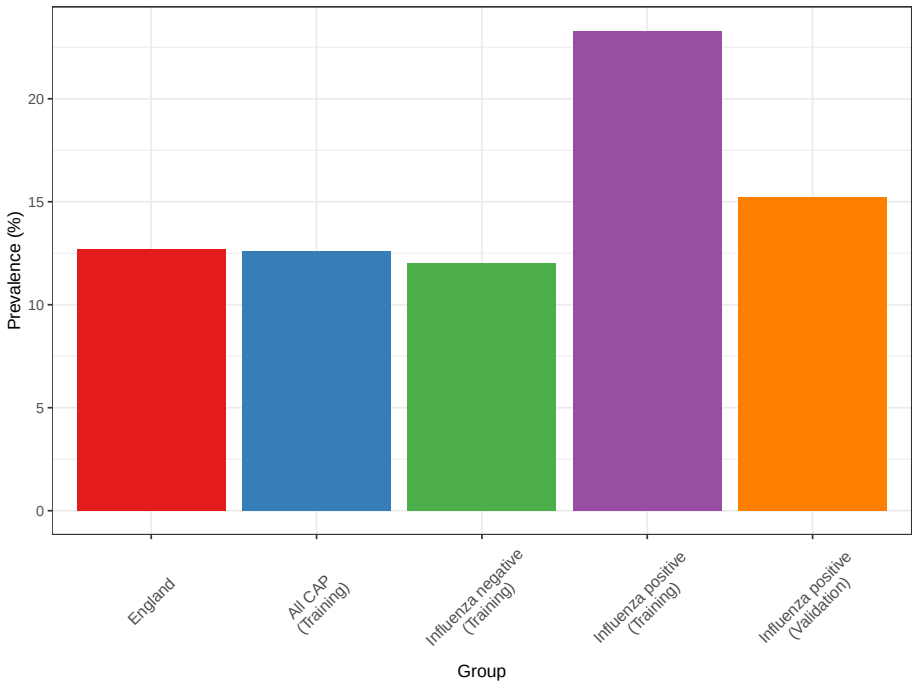


Figure 5.24: Bar graph of HLA-B*35 prevalence in GAinS patients with sepsis due to CAP. The percentage of individuals with at least one HLA-B*35 allele is plotted on the y-axis.

It was not possible to contrast the prevalence of HLA-B*35 between influenza positive and negative individuals at four digit resolution due to small patient numbers.

Streptococcus pneumoniae. In the training cohort (n=613), there were 113 individuals with and 500 individuals without a diagnosis of *S. pneumoniae* sepsis. Sixty-two HLA class I (HLA-A, HLA-B, HLA-C) and class II (HLA-DQA1, HLA-DQB1, HLA-DRB1) alleles were tested. After correction for multiple testing, there were no statistically significant associations with *S. pneumoniae* sepsis.

EBV reactivation. Of the 613 individuals in the training cohort, there was EBV reactivation data on 547 individuals (n=197 EBV positive; n=350 EBV negative). Sixty-three HLA class I (HLA-A, HLA-B, HLA-C) and class II (HLA-DQA1, HLA-DQB1, HLA-DRB1) alleles were tested for association. After correction for multiple testing, only one allele (B*07) approached statistical significance (Chi-squared test; adjusted-p=0.0733).

There were 166 individuals in the validation cohort for whom EBV reactivation data was available (n=68 EBV positive; n=98 EBV negative). HLA-B*07 was tested for association with EBV status and again the result did not reach statistical significance (p=0.102).

5.3 Discussion

5.3.1 EBV reactivation and immunosuppression

Viral reactivation in sepsis has not previously been analysed in the literature within the context of host gene expression. EBV reactivation in sepsis is common, increases over time and is associated with longer ICU stays and increased mortality. Consistent with the concept that viral reactivation in sepsis is a consequence of immune compromise, EBV reactivation was more frequent in those with the SRS1 immunocompromised sepsis transcriptomic endotype. Differences in the molecular response to EBV reactivation are also seen, in terms of total leukocyte gene expression profiles, with a small number of differentially expressed genes (n=13) between the EBV-positive and EBV-negative individuals. Although a number of these genes are of unknown function, *CACNAD3*, *CDC20*, *KIAA0101* and *PRTN3* are all implicated in EBV pathophysiology. In clinical practice, viral reactivation in the critically ill is generally considered to be an epiphenomenon and a marker of illness severity that is of little clinical significance. It remains unclear whether reactivated viruses contribute to the pathophysiology of the host response to sepsis and whether specific treatment is indicated.

In previously published work, the SRS endotype have been described as a categorical trait (Davenport et al. 2016). Here, it is described for the first time as a continuous phenotype. By using the values of principal component 1,

which separates patients with SRS1 from SRS2 in principal component analysis, it was shown that individuals with a more pronounced SRS1 phenotype are more likely to demonstrate reactivation of EBV, most likely as a consequence of decreased immune competence. One area for future work should therefore be to use immunophenotyping to determine whether higher levels of viraemia are associated with increased markers of immunosuppression.

The observed association between EBV reactivation and both morbidity and mortality highlights the clinical importance of this phenomenon. There was an increase in 28-day mortality in individuals with EBV reactivation compared to those without (27% vs 20%; p=0.04). This is in keeping with one small study of critically ill sepsis and non-sepsis patients (Libert et al. 2015) although two larger studies focusing on sepsis patients (Walton et al. 2014) (Ong et al. 2017) did not find an association between EBV reactivation and mortality. EBV-positive patients also had increased levels of organ dysfunction compared with EBV-negative patients. This is in keeping with the findings of Walton and colleagues (Walton et al. 2014) who describe a similar association with SOFA score.

In the overall cohort of 573 patients, an overall 37% incidence of EBV reactivation was observed. This finding is consistent with other studies in which the incidence of EBV reactivation in plasma has been estimated at 32-48% (Walton et al. 2014) (Ong et al. 2017). In contrast, a lower rate of EBV and other viral reactivation was observed in the metagenomic cohort. This is most likely because sample selection for metagenomics prioritised the sequencing of earlier time point samples to enable microbiological diagnosis of CAP with 40% of samples obtained on the first day of ICU admission. Walton and colleagues (Walton et al. 2014) observed that the median time to reactivation for EBV was 5 days, so it is highly probably that samples EBV-negative on day one would become EBV-positive at later timepoints.

These findings have some potential implications for the management of sepsis.

The switch from latency to lytic replication requires the virus to escape from host immune surveillance. In EBV-positive patients, an SRS1 phenotype is associated with higher levels of viraemia, with 17 SRS1 individuals having EBV loads of ¿10^3 copies/ml (vs 11 in the SRS2 group). These levels of viraemia are considerable and prompt the question as to whether such levels are detrimental, perhaps because they compromise host immunity and warrant specific treatment. For example, it is known that EBV reactivation is associated with increased expression of proteins such as an IL-10 homologue which inhibits monocyte/macrophage function (Moore et al. 2001) and several proteins which impair interferon alpha and gamma release (Morrison et al. 2001) (Cohen and Lekstrom 1999). Importantly however, a trial of specific antiviral therapy in CMV reactivation had to be terminated early due to increased mortality in the treatment arm (Cowley et al. 2017).

Finally, immune therapies for sepsis have been and continue to undergo evaluation and it is likely that an individualised approach will be required. We propose that serial qPCR of EBV viral loads in the clinical setting be considered as part of a biomarker panel to comprehensively assess a patient's immune status, such as that being used in the REAnimation Low Immune Status Markers (REALISM) project (Rol et al. 2017).

5.3.2 *Streptococcus pneumoniae* infection and sepsis endotype

For patients with *S. pneumoniae* infection where there was detectable bacterial load in plasma by ddPCR, there was a difference in bacterial load depending on SRS endotype, with SRS1 patients displaying higher levels of *S. pneumoniae* bacteraemia. This is in keeping with the SRS1 endotype being associated with a relatively immunosuppressed transcriptomic phenotype, which could potentially reduce the ability of the host immune system to control levels of infection.

Rello and colleagues (Rello et al. 2009) showed that *S. pneumoniae* bacterial load was independently associated with morbidity and mortality in patients with pneumococcal CAP. Individuals with a *S. pneumoniae* bacterial load of $\geq 10^3$ copies/ml had an increased risk for septic shock (OR 8.00), the need for mechanical ventilation (OR 10.50), and hospital mortality (OR 5.43).

In the GAinS cohort, there was no association seen between ddPCR *S. pneumoniae* positivity and vasopressor use, mechanical ventilation or 28-day mortality. This remained the case even when individuals with a bacterial load of $\geq 10^3$ copies/ml were compared to those with a bacterial load below this threshold.

The disparity between these results and those of Rello and colleagues may be due to two reasons. Firstly, the GAinS patients received antibiotics prior to sample collection so ddPCR-measured bacterial load almost certainly did not reflect peak bacteraemia levels. Secondly, the analysis was limited by only 10 patients in the high bacterial load group, so there might not have been sufficient power to detect a statistically significant difference between the groups.

However, Rello and colleagues' results are still relevant in that they indicate an association between bacterial load and disease severity. This supports the finding that individuals in the GAinS cohort with the higher mortality SRS1 endotype have higher bacterial loads; these are individuals with greater sepsis severity, reflected in degree of immunosuppression and corresponding *S. pneumoniae* bacterial load.

5.3.3 Transcriptomic signatures of infection

The four contrasts made are summarised in Table 5.10.

Previous work had identified a six gene signature for differentiating viral from non-viral infection which performed with a reasonable MR (9.1%). However, there were serious limitations to this analysis in that the "non-viral" comparator

group was poorly defined, with approximately half lacking a microbiological diagnosis.

This work was improved upon by better microbiological phenotyping (primarily through metagenomics and curation of clinical data) and an increase in cohort size with the addition of three further gene expression datasets. This resulted in the viral vs bacterial comparison yielding 206 differentially expressed genes at FDR <0.05 and fold change >1.5 (previous work had identified only 2 differentially expressed probes at FDR <0.05 and fold change >1.2).

Two predictive gene expression signatures were identified using the elastic net method. The viral vs bacterial signature performed reasonably well in the validation cohort (MR 9.8%) whilst the influenza vs bacterial signature performed less well, with a higher MR of 18% in the validation cohort.

Of the 10 genes identified in the viral vs bacterial signature, a number are of particular relevance to viral infection, increasing our confidence in the predictive gene set. *IFI27*, encodes an interferon alpha-inducible protein which has been described as a biomarker for differentiating influenza from bacterial respiratory infection (Tang et al. 2017). *SRC* encodes a non-receptor tyrosine kinase protein that interacts with viral proteins mediating key host-pathogen interactions (Pagano et al. 2013).

There were no differentially expressed genes identified in the influenza vs viral analysis or the *S. pneumoniae* vs non-*S. pneumoniae* analysis. This parallels the results from previous work (Section 5.1.5) which identified only 2 differentially expressed genes in influenza vs viral infection and none in Gram-positive vs Gram-negative infections. It is possible that our modest patient numbers meant there was insufficient power to detect a difference between the different classes of infection. Alternatively, these results suggest that the transcriptomic response in sepsis may be attributable to the class of infection (i.e. viral or bacterial) rather than specific to individual organisms. This is supported by published work

describing a lack of genes differentiating Gram-positive from Gram-negative sepsis (Tang et al. 2008).

5.3.4 Genomics: association with HLA

An association between HLA-B*35 positivity and influenza sepsis was detected in both training and validation cohorts. This finding has not previously been described in sepsis patients or in a Caucasian population. However, a case-control study (n=138 cases vs n=225 controls) of the influenza A H1N1/09 pandemic found that B*35:01:01-C*07:02:01 was one of two haplotypes observed at higher frequency in H1N1/09 influenza patients from a Mexican population (Falfan-Valencia et al. 2018). Similarly, a case-control study of influenza A H1N1/09 pandemic patients in north-east India observed the frequency of HLA-B*35 to be higher in cases (n=35) versus controls (n=35). This preliminary finding in the GAinS cohort should be treated with caution and future work would include validating the finding in a cohort with larger numbers of influenza cases, restricting the analysis to Caucasian individuals and consideration of haplotype structure.

A possible association between HLA-B*07 positivity and EBV reactivation was detected in the GAinS cohort. Although this did not reach statistical significance, evidence in the literature is suggestive of this being a genuine finding. In multiple sclerosis, where EBV plays a critical role in disease development, HLA-B*07 is found at higher prevalence in a patients compared with controls (Jilek et al. 2012). In the same Swiss cohort, ex vivo studies showed that the HLA-B*07 restricted EBV-specific CD8+ T cell response was dysregulated in multiple sclerosis patients (Jilek et al. 2012). In another study (Agostini et al. 2018), EBV viral load was higher in multiple sclerosis patients positive for HLA-B*07 compared to those negative for the allele.

No association between HLA class I and II alleles and *S. pneumoniae* infection was detected. This is consistent with the literature, where no positive associations have been described. A possible reason for the positive findings in viral infection but negative finding in bacterial infection may be that an effective host response against viruses relies more heavily on cell-mediated immunity through class I and class II HLA molecules.

5.4 Conclusions

Improved microbiological phenotyping has the potential to enhance our understanding of the host response in sepsis. In this chapter, I have explored the interaction between microbiology and host transcriptomic sepsis endotypes, performed differential gene expression analysis for different classes of infection, and investigated the effect of HLA alleles on susceptibility to different microbiological classes of infection. The results have shown integrating metagenomic data with other -omic datasets to be a promising approach.

6

General Discussion

This chapter outlines the broader conclusions and future directions suggested by the work described in this thesis

General intro

6.1 A section

6.2 Limitations and future work

6.3 Conclusion

Conclusions.

A

Materials and Methods

Organism	Amplicon	Primer/probe	Nucleotide sequence
S. pneumoniae	67 bp	Taqman probe	5'-/5HEX/AAT GTT ACG/ZEN/CAA CTG ACG AG/3IABkFQ/-3'
		Forward primer	5'-GCT GTT TTA GCA GAT AGT GAG ATC GA-3'
		Reverse primer	5'-TCC CAG TCG GTG CTG TCA-3'
Influenza A	185 bp	Taqman probe	5'-/56-FAM/TGC AGT CCT/ZEN/CGC TCA CTG GGC ACG /3IABkFQ/-3'
		Forward primer	5'-AGG GCA TTY TGG ACA AAK CGT CTA-3'
		Reverse primer	5'-GAC CRA TCC TGT CAC CTC TGA C-3'
Epstein-Barr virus	75 bp	Taqman probe	5'-/56-FAM/AGG GAG ACA/ZEN/CAT CTG GAC CAG AAG GC/3IABkFQ/-3'
		Forward primer	5'-TCT TTG AGG TCC ACT GCC G-3'
		Reverse primer	5'-TAC AGG ACC TGG AAA TGG CC-3'
Cytomegalovirus	66 bp	Taqman probe	5'-/56-FAM/TG GGC AAC C/ZEN/A CCG CAC TGA GG/3IABkFQ/-3'
		Forward primer	5'-TGG GCG AGG ACA ACG AA-3'
		Reverse primer	5'-TGA GGC TGG GAA GCT GAC AT-3'

Table A.1: Primer/probe sets used for the digital droplet PCR experiments

B

Application of Metagenomic Sequencing to Sepsis Samples

Virus family	Virus species
Adenoviridae	Human adenovirus
Arenaviridae	Lassa virus Lymphocytic choriomeningitis virus
Coronaviridae	Human coronavirus HKU1, NL63, OC43, 229E Middle East respiratory syndrome coronavirus Severe acute respiratory syndrome coronavirus
Flaviviridae	Dengue virus Japanese encephalitis virus Murray Valley encephalitis virus St Louis encephalitis virus Tick-borne encephalitis virus West Nile virus Yellow fever virus Zika virus
Herpesviridae	Human herpesvirus 3 (Varicella zoster virus) Human herpesvirus 4 (Epstein Barrvirus) Human herpesvirus 5 (Cytomegalovirus) Human herpesvirus 6-7 (Roseolovirus) Human herpesvirus 8 (Kaposi's sarcoma-associated herpesvirus) Herpes simplex virus 1-2
Orthomyxoviridae	Influenza virus A-C
Paramyxoviridae	Hendra henipavirus Human metapneumovirus Humanparainfluenzavirus1-5 Measles morbilivirus Mumps rubulavirus Nipah henipavirus Respiratory syncytial virus Sosugavirus
Parvoviridae	Human bocavirus Human parvovirus B19 Human parvovirus 4 Primate erythroparvovirus 1 Primate tetraparvovirus 1
Peribunyaviridae	California encephalitis virus
Phenuiviridae	Rift valley fever virus Sandfly fever Naples virus Sandfly fever Sicilian virus
Picornaviridae	Cardiovirus A-B Coxsackie A virus ECHO virus Enterovirus A, B, D Hepatitis A virus Parechovirus A-B Rhinovirus A-C Rosavirus A Salivivirus
Polyomaviridae	BK virus JC polyomavirus
Reoviridae	Rotavirus A-C
Rhabdovirus	Australian bat lyssavirus Duvenhage lyssavirus European bat lyssavirus 1-2 Lagos bat lyssavirus Mokola lyssavirus Rabies lyssavirus
Togaviridae	Chikungunya virus Eastern equine encephalitis virus Rubella virus Venezuelan equine encephalitis virus Western equine encephalitis virus

Table B.1: Viruses included in enrichment probe set

Bacterial genus	Bacterial species
Acinetobacter	*baumanii* *calcoaceticus*
Bartonella	*henselae*
Bordetella	*pertussis*
Borrelia	*burgdorferi*
Brucella	
Burkholderia	*cepacia*
Chlamydophila	*pneumoniae* *psittaci*
Coxiella	*burnetii*
Enterobacter	*aerogenes* *cloacae*
Escherichia	*coli*
Haemophilus	*influenzae* *parainfluenzae*
Klebsiella	*pneumoniae* *oxytoca*
Legionella	*pneumophila*
Leptospira	
Listeria	*monocytogenes*
Moraxella	*catarrhalis*
Mycobacterium	*avium* *intracellulare* *tuberculosis*
Mycoplasma	*pneumoniae*
Neisseria	*meningitidis*
Nocardia	
Pseudomonas	*aeruginosa*
Serratia	*marcescens*
Staphylococcus	*aureus*
Stenotrophomonas	*maltophilia*
Streptococcus	*agalactiae* *pneumoniae* *pyogenes*
Treponema	*pallidum*

Table B.2: Bacteria included in enrichment probe set

Table B.3: Viral Multiplex Reference reagent 11/242 (UK NIBSC). This reference set included 25 viruses of various nucleic acid types, envelope types and genome sizes. 21/25 viruses had corresponding enrichment probes in our probe panel. (Adapted from Mee et al.)

Group	Family	Species/serotype	Envelope	Genome size	Included in probeset	Concentration (log10 copies/ml)
dsDNA	Adenoviridae	Adenovirus 2	No	35.9	Yes	NA
		Adenovirus 41		34.2	Yes	NA
	Herpesviridae	Human herpesvirus 1	Yes	151.2	Yes	NA
		Human herpesvirus 2		154.7	Yes	NA
		Human herpesvirus 3 (VZV)		124.8	Yes	NA
		Human herpesvirus 4 (EBV)		171.7	Yes	3.88
		Human herpesvirus 5 (CMV)		233.7	Yes	4.66
dsRNA	Reoviridae	Rotavirus A	No	18.5	Yes	6.76
ssRNA (+)	Astroviridae	Astrovirus	No	6.8	No	NA
	Caliciviridae	Norovirus GI	No	7.6	No	NA
		Norovirus GII		7.5	No	NA
		Sapovirus C12		7.5	No	NA
	Coronaviridae	Coronavirus 229E	Yes	27.2	Yes	NA
	Picornaviridae	Coxsackievirus B4	No	7.4	Yes	NA
		Rhinovirus A39		7.1	Yes	NA
		Parechovirus 3		7.2	Yes	7.07
ssRNA (-)	Orthomyxoviridae	Influenza A virus H1N1	Yes	13.2	Yes	NA
		Influenza A virus H3N2		13.6	Yes	NA
		Influenza B virus		14.2	Yes	NA
	Paramyxoviridae	Metapneumovirus A	Yes	13.3	Yes	NA
		Parainfluenzavirus 1		15.5	Yes	NA
		Parainfluenzavirus 2		15.7	Yes	NA
		Parainfluenzavirus 3		15.4	Yes	NA
		Parainfluenzavirus 4		17.4	Yes	NA
		Respiratory syncytial virus A2		15.2	Yes	3.75

C

Improved Classification of Microbiological Aetiology in Sepsis

MICROBIOLOGICAL DATA: COMMUNITY ACQUIRED PNEUMONIA

Streptococcus pneumoniae	☐ [CCacpneumo1 / 0 1]	Chlamydia pneumoniae	☐ [ccpneumo5 / 0 1]
Haemophilus influenzae	☐ [ccpneumo2 / 0 1]	Mycoplasma pneumonia	☐ [ccpneumo3 / 0 1]
Staphyloccocus aureus	☐ [ccpneumo6 / 0 1]	Pseudomonas aeruginosa	☐ [ccpneumo7 / 0 1]
Legionella spp	☐ [ccpneumo4 / 0 1]	Mixed organisms	☐ [ccpneumo8 / 0 1]
Not known	☐ [ccpneumo9 / 0 1]		
Viral	☐ [ccpneumo10 / 0 1]	State type if known	[speviral] 2
Other	☐ [ccpneumo11 / 0 1]	Specify	[ccpneumo12] 3

Lung organism identification based on :

Culture of lung secretions	☐ [LungOrg1 / 0 1]	Serology	☐ [LungOrg3 / 0 1]
Culture of pleural effusion	☐ [CultPleuEffus / 0 1]	Blood culture	☐ [LungOrg2 / 0 1]
Other technique	☐ [LungOrg4 / 0 1]	Specify	[otherTechnSpecify]
Complicating factors	☐ Yes [ComplFactor/1] ☐ No [ComplFactor/2]		
Pleural effusion	☐ [compliF1 / 0 1]	Domiciliary oxygen	☐ [compliF4 / 0 1]
Empyema	☐ [compliF2 / 0 1]	Domiciliary ventilation	☐ [compliF5 / 0 1]
Recent discharge from an acute care facility	☐ [compliF3 / 0 1]	Cystic fibrosis	☐ [compliF6 / 0 1]

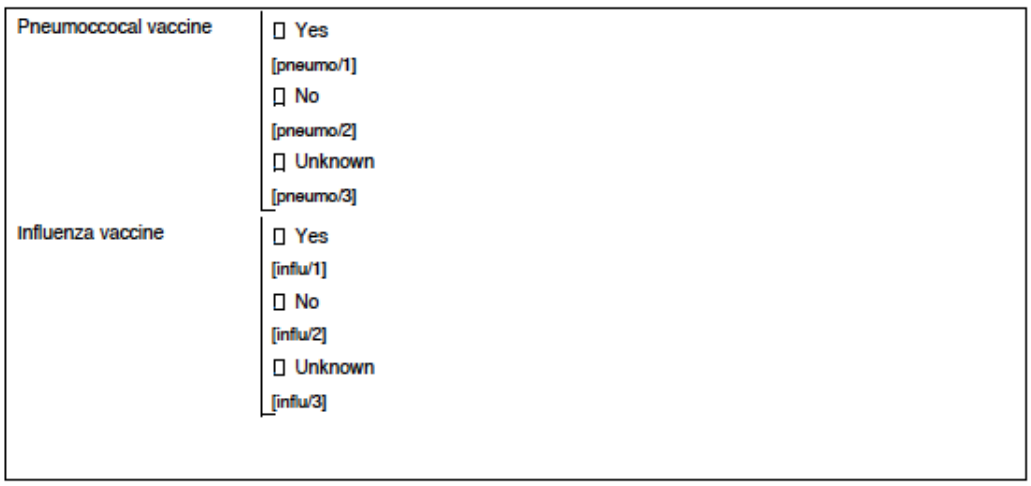


Figure C.1: GAinS electronic case record form: microbiology section

D

Integration of Microbiology with the Host Response

Gene	log2(fold change)	FDR
IGLL1	0.79	4.02E-03
CDC20	0.62	4.51E-03
CACNA2D3	-0.63	5.21E-03
TXNDC5	0.75	8.48E-03
IGKV3D-20	0.78	9.23E-03
KIAA0101	0.59	1.28E-02
IGKV1D-33	0.61	1.33E-02
IGJ	0.83	1.33E-02
PRTN3	0.66	4.68E-02

Table D.1: Summary of differentially expressed genes in EBV-positive vs EBV-negative individuals. Nine genes were significant at fold-change >1.5 and FDR <0.05.

Gene	log2(fold change)	FDR
IGLL1	0.85	1.15E-03
TXNDC5	0.81	2.43E-03
IGKV3D-20	0.84	2.54E-03
IGKV1D-33	0.66	3.38E-03
CDC20	0.59	3.98E-03
IGJ	0.89	4.06E-03
MGC29506	0.61	5.96E-03
TNFRSF17	0.63	1.32E-02
DEFA1B	0.82	3.11E-02
PRTN3	0.66	3.87E-02
DEFA4	0.75	4.19E-02
CEACAM6	0.63	4.82E-02

Table D.2: Summary of differentially expressed genes in EBV-positive vs EBV-negative individuals with SRS included as a covariate in the linear model. Twelve genes were significant at fold-change >1.5 and FDR <0.05.

Gene	log2(fold change)	FDR
IFI27	3.10	6.88E-07
IMPA2	-0.64	1.86E-04
C3ORF54	0.74	6.59E-04
SPATS2L	1.02	6.59E-04
LOC554203	0.59	6.91E-04
HERC6	0.94	8.17E-04
JUP	1.28	9.63E-04
TIMM10	1.33	1.03E-03
SRC	0.77	1.03E-03
HES4	1.04	1.03E-03
LGALS3BP	0.82	1.03E-03
RASGRP3	0.75	1.03E-03
TGIF2	0.64	1.03E-03
MT1A	0.89	1.03E-03
C9ORF91	0.75	1.03E-03
OAS1	1.42	1.08E-03
SCO2	0.92	1.08E-03
HS.125087	1.21	1.08E-03
PARP12	1.07	1.08E-03
EPSTI1	2.02	1.08E-03
OAS2	1.57	1.08E-03
IFI44L	2.42	1.08E-03
SP140	0.78	1.08E-03
LY6E	1.69	1.11E-03
CXCL10	0.78	1.16E-03
TMCO3	-0.70	1.21E-03
HS.72010	0.63	1.28E-03
XAF1	1.44	1.28E-03
IFI44	1.68	1.33E-03
RNASE1	0.94	1.36E-03
GALM	0.85	1.45E-03
IFIT3	1.57	1.45E-03
RSAD2	1.81	1.84E-03
LAMP3	0.72	1.99E-03
HS.386275	0.70	1.99E-03
PLB1	-0.74	2.02E-03
MT2A	1.00	2.14E-03
SERPING1	1.50	2.14E-03
OAS3	1.66	2.14E-03
CDKN1A	0.71	2.14E-03
LHFPL2	0.68	2.30E-03
BTN3A3	0.85	2.47E-03
PARP14	0.95	2.52E-03
IFIT5	0.97	2.57E-03
FHL2	0.61	2.60E-03
CXCL16	-0.63	2.93E-03
SIGLEC10	-0.88	3.13E-03
NLRC4	-0.67	3.55E-03
ISG15	1.68	3.59E-03
ZNF684	0.60	3.63E-03

Table D.3: Summary of differentially expressed genes in viral vs bacterial infection. The top 50 most significantly differentially expressed genes are listed here.

Gene	log2(fold change)	FDR
IFI27	3.52	1.3E-06
JUP	1.68	2.0E-04
C3ORF54	0.90	4.7E-04
NPL	-0.66	7.9E-04
SPATS2L	1.17	1.1E-03
CBL	-0.68	1.4E-03
IMPA2	-0.67	1.7E-03
HERC6	1.01	4.8E-03
LGALS3BP	0.89	5.1E-03
HES4	1.12	6.0E-03
MT1A	0.98	6.2E-03
BCAT1	-1.20	7.0E-03
RASGRP3	0.80	7.0E-03
ECHDC3	-1.25	7.2E-03
CDC123	-0.62	7.2E-03
MTE	0.64	7.2E-03
SCO2	1.00	7.2E-03
SP140	0.83	7.2E-03
LOC389386	0.66	7.2E-03
GALM	0.95	7.2E-03
UBQLNL	0.60	7.2E-03
SOCS2	0.82	7.2E-03
LY6E	1.83	7.2E-03
HS.125087	1.25	7.2E-03
OAS2	1.69	7.2E-03
NUDT5	-0.65	7.5E-03
MT2A	1.12	8.0E-03
IFI44L	2.55	8.0E-03
OAS1	1.48	8.0E-03
ADARB1	0.70	8.0E-03
LAMP3	0.79	8.0E-03
LOC401845	0.62	8.0E-03
PARP12	1.11	8.0E-03
CDKN1A	0.80	8.0E-03
SLC31A2	-0.73	8.0E-03
CXCL10	0.78	8.0E-03
FHL2	0.70	8.0E-03
XAF1	1.51	8.0E-03
HIST1H4E	0.59	8.0E-03
DUSP5	0.66	8.1E-03
TGIF2	0.65	8.6E-03
PDE9A	0.73	8.9E-03
HS.72010	0.66	8.9E-03
IFI44	1.77	9.4E-03
TIMM10	1.28	9.4E-03
LOC652694	1.34	9.8E-03
RAB11FIP3	0.60	9.8E-03
CD69	0.86	1.0E-02
SRC	0.74	1.0E-02
OAS3	1.79	1.0E-02

Table D.4: Summary of differentially expressed genes in influenza vs bacterial infection. The top 50 most significantly differentially expressed genes are listed here.

References

Agostini, S., Mancuso, R., Guerini, F. R., et al. (2018). "HLA alleles modulate EBV viral load in multiple sclerosis." eng. *Journal of translational medicine* 16 (1): p. 80.

Allander, T., Emerson, S. U., Engle, R. E., et al. (2001). "A virus discovery method incorporating DNase treatment and its application to the identification of two bovine parvovirus species". *Proc Natl Acad Sci U S A* 98: pp. 11609–14.

Allicock, O. M., Guo, C., Uhlemann, A.-C., et al. (2018). "BacCapSeq: a Platform for Diagnosis and Characterization of Bacterial Infections." eng. *mBio* 9 (5).

Amorim-Vaz, S., Tran, V. D. T., Pradervand, S., et al. (2015). "RNA Enrichment Method for Quantitative Transcriptional Analysis of Pathogens In Vivo Applied to the Fungus Candida albicans." eng. *mBio* 6 (5): e00942–15.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., et al. (2010). "Data quality control in genetic case-control association studies." eng. *Nature protocols* 5 (9): pp. 1564–73.

Anh, N. T., Hong, N. T. T., Nhu, L. N. T., et al. (2019). "Viruses in Vietnamese Patients Presenting with Community-Acquired Sepsis of Unknown Cause." eng. *Journal of clinical microbiology* 57 (9).

Antcliffe, D. B., Burnham, K. L., Al-Beidh, F., et al. (2019). "Transcriptomic Signatures in Sepsis and a Differential Response to Steroids. From the VANISH Randomized Trial." eng. *American journal of respiratory and critical care medicine* 199 (8): pp. 980–986.

Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." eng. *Nature genetics* 25 (1): pp. 25–9.

Baldridge, M. T., Nice, T. J., McCune, B. T., et al. (2015). "Commensal microbes and interferon-lambda determine persistence of enteric murine norovirus infection." eng. *Science (New York, N.Y.)* 347 (6219): pp. 266–9.

Barnato, A. E., Albert, S. M., Angus, D. C., et al. (2011). "Disability among elderly survivors of mechanical ventilation." eng. *American journal of respiratory and critical care medicine* 183 (8): pp. 1037–42.

Barreiro, L. B., Tailleux, L., Pai, A. A., et al. (2012). "Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection." eng. *Proceedings of the National Academy of Sciences of the United States of America* 109 (4): pp. 1204–9.

Benz, F., Tacke, F., Luedde, M., et al. (2015). "Circulating microRNA-223 serum levels do not predict sepsis or survival in patients with critical illness." eng. *Disease markers* 2015: p. 384208.

Bernard, G. R., Wheeler, A. P., Russell, J. A., et al. (1997). "The effects of ibuprofen on the physiology and survival of patients with sepsis. The Ibuprofen in Sepsis Study Group." eng. *The New England journal of medicine* 336 (13): pp. 912–8.

Blackwell, J. M., Jamieson, S. E., and Burgner, D. (2009). "HLA and infectious diseases." eng. *Clinical microbiology reviews* 22 (2): 370–85, Table of Contents.

Blauwkamp, T. A., Thair, S., Rosen, M. J., et al. (2019). "Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease." eng. *Nature microbiology* 4 (4): pp. 663–674.

Bolger, A. M., Lohse, M., and Usadel, B. (2014). "Trimmomatic: a flexible trimmer for Illumina sequence data". *Bioinformatics* 30: pp. 2114–20.

Bomsztyk, K., Mar, D., An, D., et al. (2015). "Experimental acute lung injury induces multi-organ epigenetic modifications in key angiogenic genes implicated in sepsis-associated endothelial dysfunction." eng. *Critical care (London, England)* 19: p. 225.

Bonsall, D., Ansari, M. A., Ip, C., et al. (2015). "ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens". *F1000Res* 4: p. 1062.

Boomer, J. S., To, K., Chang, K. C., et al. (2011). "Immunosuppression in patients who die of sepsis and multiple organ failure." eng. *JAMA* 306 (23): pp. 2594–605.

Breiman, L (2001). "Random Forests". *Machine Learning* 45: pp. 4–32.

Breitbart, M. and Rohwer, F. (2005). "Method for discovering novel DNA viruses in blood using viral particle selection and shotgun sequencing". *Biotechniques* 39: pp. 729–36.

Briese, T., Kapoor, A., Mishra, N., et al. (2015). "Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis." eng. *mBio* 6 (5): e01491–15.

Bright, A. T., Tewhey, R., Abeles, S., et al. (2012). "Whole genome sequencing analysis of Plasmodium vivax using whole genome capture." eng. *BMC genomics* 13: p. 262.

Buchan, B. W. and Ledeboer, N. A. (2014). "Emerging technologies for the clinical microbiology laboratory". *Clin Microbiol Rev* 27: pp. 783–822.

Buras, J. A., Holzmann, B., and Sitkovsky, M. (2005). "Animal models of sepsis: setting the stage." eng. *Nature reviews. Drug discovery* 4 (10): pp. 854–65.

Burnham, K. L. (2017). "Functional Genomics of the Sepsis Response". PhD thesis. University of Oxford.

Carrington, M., Nelson, G. W., Martin, M. P., et al. (1999). "HLA and HIV-1: heterozygote advantage and B*35-Cw*04 disadvantage." eng. *Science (New York, N.Y.)* 283 (5408): pp. 1748–52.

Charalampous, T., Kay, G. L., Richardson, H., et al. (2019). "Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection." eng. *Nature biotechnology* 37 (7): pp. 783–792.

Chen, K. Y., Ko, S. C., Hsueh, P. R., et al. (2001). "Pulmonary fungal infection: emphasis on microbiological spectra, patient outcome, and prognostic factors." eng. *Chest* 120 (1): pp. 177–84.

Cheng, S.-C., Scicluna, B. P., Arts, R. J. W., et al. (2016). "Broad defects in the energy metabolism of leukocytes underlie immunoparalysis in sepsis." eng. *Nature immunology* 17 (4): pp. 406–13.

Chiu, C. Y. and Miller, S. A. (2019). "Clinical metagenomics." eng. *Nature reviews. Genetics* 20 (6): pp. 341–355.

Choi, S.-H., Hong, S.-B., Ko, G.-B., et al. (2012). "Viral infection in patients with severe pneumonia requiring intensive care unit admission." eng. *American journal of respiratory and critical care medicine* 186 (4): pp. 325–32.

Clark, M. F. and Baudouin, S. V. (2006). "A systematic review of the quality of genetic association studies in human sepsis." eng. *Intensive care medicine* 32 (11): pp. 1706–12.

Cohen, J. and Carlet, J. (1996). "INTERSEPT: an international, multicenter, placebo-controlled trial of monoclonal antibody to human tumor necrosis factor-alpha in patients with sepsis. International Sepsis Trial Study Group." eng. *Critical care medicine* 24 (9): pp. 1431–40.

Cohen, J. I. (2000). "Epstein-Barr virus infection." eng. *The New England journal of medicine* 343 (7): pp. 481–92.

Cohen, J. I. and Lekstrom, K. (1999). "Epstein-Barr virus BARF1 protein is dispensable for B-cell transformation and inhibits alpha interferon secretion from mononuclear cells." eng. *Journal of virology* 73 (9): pp. 7627–32.

Consortium, T. G. O. (2019). "The Gene Ontology Resource: 20 years and still GOing strong." eng. *Nucleic acids research* 47 (D1): pp. D330–D338.

Cowley, N. J., Owen, A., Shiels, S. C., et al. (2017). "Safety and Efficacy of Antiviral Therapy for Prevention of Cytomegalovirus Reactivation in Immunocompetent Critically Ill Patients: A Randomized Clinical Trial." eng. *JAMA internal medicine* 177 (6): pp. 774–783.

Cuomo, C. A. (2017). "Harnessing Whole Genome Sequencing in Medical Mycology". *Curr Fungal Infect Rep* 11: pp. 52–59.

Cuthbertson, B. H., Elders, A., Hall, S., et al. (2013). "Mortality and quality of life in the five years after severe sepsis." eng. *Critical care (London, England)* 17 (2): R70.

Davenport, E. E. (2014). "Functional genomics of variation in response to infection: insights into severe sepsis and common variable immune deficiency disorders". PhD thesis. University of Oxford.

Davenport, E. E., Burnham, K. L., Radhakrishnan, J., et al. (2016). "Genomic landscape of the individual host response and outcomes in sepsis: a prospective cohort study." eng. *The Lancet. Respiratory medicine* 4 (4): pp. 259–71.

Daviaud, F., Grimaldi, D., Dechartres, A., et al. (2015). "Timing and causes of death in septic shock." eng. *Annals of intensive care* 5 (1): p. 16.

Depledge, D. P., Palser, A. L., Watson, S. J., et al. (2011). "Specific capture and whole-genome sequencing of viruses from clinical samples". *PLoS One* 6: e27805.

Dickson, R. P. (2016). "The microbiome and critical illness." eng. *The Lancet. Respiratory medicine* 4 (1): pp. 59–72.

Doan, T., Wilson, M. R., Crawford, E. D., et al. (2016). "Illuminating uveitis: metagenomic deep sequencing identifies common and rare pathogens". *Genome Med* 8: p. 90.

Dolgachev, V. A., Goldberg, R., Suresh, M. V., et al. (2016). "Electroporation-mediated delivery of the FER gene in the resolution of trauma-related fatal pneumonia." eng. *Gene therapy* 23 (11): pp. 785–796.

Durbin, R. (2014). "Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT)." eng. *Bioinformatics (Oxford, England)* 30 (9): pp. 1266–72.

Fairfax, B. P., Humburg, P., Makino, S., et al. (2014). "Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression." eng. *Science (New York, N.Y.)* 343 (6175): p. 1246949.

Falfan-Valencia, R., Narayanankutty, A., Resendiz-Hernandez, J. M., et al. (2018). "An Increased Frequency in HLA Class I Alleles and Haplotypes Suggests Genetic Susceptibility to Influenza A (H1N1) 2009 Pandemic: A Case-Control Study." eng. *Journal of immunology research* 2018: p. 3174868.

Fang, H., Knezevic, B., Burnham, K. L., et al. (2016). "XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits." eng. *Genome medicine* 8 (1): p. 129.

Feezor, R. J., Oberholzer, C., Baker, H. V., et al. (2003). "Molecular characterization of the acute inflammatory response to infections with gram-negative versus gram-positive bacteria." eng. *Infection and immunity* 71 (10): pp. 5803–13.

Feng, H., Shuda, M., Chang, Y., et al. (2008). "Clonal integration of a polyomavirus in human Merkel cell carcinoma." eng. *Science (New York, N.Y.)* 319 (5866): pp. 1096–100.

Forbes, J. D., Knox, N. C., Peterson, C.-L., et al. (2018). "Highlighting Clinical Metagenomics for Enhanced Diagnostic Decision-making: A Step Towards Wider Implementation." eng. *Computational and structural biotechnology journal* 16: pp. 108–120.

Gao, L., Zhong, J.-C., Huang, W.-T., et al. (2017). "Integrative analysis of BSG expression in NPC through immunohistochemistry and public high-throughput gene expression data." eng. *American journal of translational research* 9 (10): pp. 4574–4592.

Gnirke, A., Melnikov, A., Maguire, J., et al. (2009). "Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing." eng. *Nature biotechnology* 27 (2): pp. 182–9.

Goh, C., Golubchik, T., Ansari, M., et al. (2019). "Targeted metagenomic sequencing enhances the identification of pathogens associated with acute infection". *BioRxiv*.

Gosiewski, T., Ludwig-Galezowska, A. H., Huminska, K., et al. (2017). "Comprehensive detection and identification of bacterial DNA in the blood of patients with sepsis and healthy volunteers using next-generation sequencing method - the observation of DNAemia." eng. *European journal of clinical microbiology & infectious diseases : official publication of the European Society of Clinical Microbiology* 36 (2): pp. 329–336.

Grice, E. A. and Segre, J. A. (2012). "The human microbiome: our second genome." eng. *Annual review of genomics and human genetics* 13: pp. 151–70.

Grumaz, S., Stevens, P., Grumaz, C., et al. (2016). "Next-generation sequencing diagnostics of bacteremia in septic patients". *Genome Med* 8: p. 73.

Grumaz, S., Grumaz, C., Vainshtein, Y., et al. (2019). "Enhanced Performance of Next-Generation Sequencing Diagnostics Compared With Standard of Care Microbiological Diagnostics in Patients Suffering From Septic Shock." eng. *Critical care medicine* 47 (5): e394–e402.

Gupta, S., Sakhuja, A., Kumar, G., et al. (2016). "Culture-Negative Severe Sepsis: Nationwide Trends and Outcomes". *Chest* 150: pp. 1251–1259.

Hamon, M. A. and Cossart, P. (2008). "Histone modifications and chromatin remodeling during bacterial infections." eng. *Cell host & microbe* 4 (2): pp. 100–9.

Heininger, A., Haeberle, H., Fischer, I., et al. (2011). "Cytomegalovirus reactivation and associated outcome of critically ill patients with severe sepsis." eng. *Critical care (London, England)* 15 (2): R77.

Herberg, J. A., Kaforou, M., Wright, V. J., et al. (2016). "Diagnostic Test Accuracy of a 2-Transcript Host RNA Signature for Discriminating Bacterial vs Viral Infection in Febrile Children." eng. *JAMA* 316 (8): pp. 835–45.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., et al. (2009). "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." eng. *Proceedings of the National Academy of Sciences of the United States of America* 106 (23): pp. 9362–7.

Huang, A. M., Newton, D., Kunapuli, A., et al. (2013). "Impact of rapid organism identification via matrix-assisted laser desorption/ionization time-of-flight combined with antimicrobial stewardship team intervention in adult patients with bacteremia and candidemia." eng. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 57 (9): pp. 1237–45.

Huang, H., Ideh, R. C., Gitau, E., et al. (2014). "Discovery and validation of biomarkers to guide clinical management of pneumonia in African children." eng. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 58 (12): pp. 1707–15.

Huber, W., Heydebreck, A. von, Sultmann, H., et al. (2002). "Variance stabilization applied to microarray data calibration and to the quantification of differential expression". *Bioinformatics* 18 Suppl 1: S96–104.

Ichinohe, T., Pang, I. K., Kumamoto, Y., et al. (2011). "Microbiota regulates immune defense against respiratory tract influenza A virus infection." eng. *Proceedings of the National Academy of Sciences of the United States of America* 108 (13): pp. 5354–9.

Jain, M., Olsen, H. E., Paten, B., et al. (2016). "The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community." eng. *Genome biology* 17 (1): p. 239.

Jain, S., Self, W. H., Wunderink, R. G., et al. (2015). "Community-Acquired Pneumonia Requiring Hospitalization among U.S. Adults." eng. *The New England journal of medicine* 373 (5): pp. 415–27.

Jameson, J. L. and Longo, D. L. (2015). "Precision medicine–personalized, problematic, and promising." eng. *The New England journal of medicine* 372 (23): pp. 2229–34.

Jia, X., Han, B., Onengut-Gumuscu, S., et al. (2013). "Imputing amino acid polymorphisms in human leukocyte antigens." eng. *PloS one* 8 (6): e64683.

Jilek, S., Schluep, M., Harari, A., et al. (2012). "HLA-B7-restricted EBV-specific CD8+ T cells are dysregulated in multiple sclerosis." eng. *Journal of immunology (Baltimore, Md. : 1950)* 188 (9): pp. 4671–80.

Jolley, K. A., Bliss, C. M., Bennett, J. S., et al. (2012). "Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain". *Microbiology* 158: pp. 1005–15.

Kanwal, F., Kramer, J., Asch, S. M., et al. (2017). "Risk of Hepatocellular Cancer in HCV Patients Treated With Direct-Acting Antiviral Agents." eng. *Gastroenterology* 153 (4): 996–1005.e1.

Kiiveri, H. T. (2008). "A general approach to simultaneous model fitting and variable elimination in response models for biological data with many more variables than observations." eng. *BMC bioinformatics* 9: p. 195.

Kim, A. Y., Kuntzen, T., Timm, J., et al. (2011). "Spontaneous control of HCV is associated with expression of HLA-B 57 and preservation of targeted epitopes." eng. *Gastroenterology* 140 (2): 686–696.e1.

Kwok, H., Wu, C. W., Palser, A. L., et al. (2014). "Genomic diversity of Epstein-Barr virus genomes isolated from primary nasopharyngeal carcinoma biopsy samples." eng. *Journal of virology* 88 (18): pp. 10662–72.

Landry, M. L. and St George, K. (2017). "Laboratory Diagnosis of Zika Virus Infection." eng. *Archives of pathology & laboratory medicine* 141 (1): pp. 60–67.

Langelier, C., Zinter, M. S., Kalantar, K., et al. (2018a). "Metagenomic Sequencing Detects Respiratory Pathogens in Hematopoietic Cellular Transplant Patients". *Am J Respir Crit Care Med* 197: pp. 524–528.

Langelier, C., Kalantar, K. L., Moazed, F., et al. (2018b). "Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults." eng. *Proceedings of the National Academy of Sciences of the United States of America* 115 (52): E12353–E12362.

Langley, R. J., Tsalik, E. L., Velkinburgh, J. C. van, et al. (2013). "An integrated clinico-metabolomic model improves prediction of death in sepsis." eng. *Science translational medicine* 5 (195): 195ra95.

Le Chatelier, E., Nielsen, T., Qin, J., et al. (2013). "Richness of human gut microbiome correlates with metabolic markers." eng. *Nature* 500 (7464): pp. 541–6.

Lee, H. J., Georgiadou, A., Walther, M., et al. (2018). "Integrated pathogen load and dual transcriptome analysis of systemic host-pathogen interactions in severe malaria." eng. *Science translational medicine* 10 (447).

Lee, S. H., Chen, S.-Y., Chien, J.-Y., et al. (2019). "Usefulness of the FilmArray meningitis/encephalitis (M/E) panel for the diagnosis of infectious meningitis and encephalitis in Taiwan." eng. *Journal of microbiology, immunology, and infection = Wei mian yu gan ran za zhi*.

Leek, J. T., Johnson, W. E., Parker, H. S., et al. (2012). "The sva package for removing batch effects and other unwanted variation in high-throughput experiments." eng. *Bioinformatics (Oxford, England)* 28 (6): pp. 882–3.

Lesnik, E. A. and Freier, S. M. (1995). "Relative thermodynamic stability of DNA, RNA, and DNA:RNA hybrid duplexes: relationship with base composition and structure." eng. *Biochemistry* 34 (34): pp. 10807–15.

Levy, M. M., Fink, M. P., Marshall, J. C., et al. (2003). "2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference." eng. *Critical care medicine* 31 (4): pp. 1250–6.

Li, H. and Durbin, R. (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform". *Bioinformatics* 25: pp. 1754–60.

Li, L., Deng, X., Mee, E. T., et al. (2015). "Comparing viral metagenomics methods using a highly multiplexed human viral pathogens reagent". *J Virol Methods* 213: pp. 139–46.

Li, Y., Chan, E. Y., Li, J., et al. (2010). "MicroRNA expression and virulence in pandemic influenza virus-infected mice." eng. *Journal of virology* 84 (6): pp. 3023–32.

Libert, N., Bigaillon, C., Chargari, C., et al. (2015). "Epstein-Barr virus reactivation in critically ill immunocompetent patients." eng. *Biomedical journal* 38 (1): pp. 70–6.

Lim, W. S., Baudouin, S. V., George, R. C., et al. (2009). "BTS guidelines for the management of community acquired pneumonia in adults: update 2009". *Thorax* 64 Suppl 3: pp. iii1–55.

Loh, P.-R., Danecek, P., Palamara, P. F., et al. (2016). "Reference-based phasing using the Haplotype Reference Consortium panel." eng. *Nature genetics* 48 (11): pp. 1443–1448.

Long, Y., Zhang, Y., Gong, Y., et al. (2016). "Diagnosis of Sepsis with Cell-free DNA by Next-Generation Sequencing Technology in ICU Patients". *Arch Med Res* 47: pp. 365–371.

Man, M., Close, S. L., Shaw, A. D., et al. (2013). "Beyond single-marker analyses: mining whole genome scans for insights into treatment responses in severe sepsis." eng. *The pharmacogenomics journal* 13 (3): pp. 218–26.

Marazzi, I., Ho, J. S. Y., Kim, J., et al. (2012). "Suppression of the antiviral response by an influenza histone mimic." eng. *Nature* 483 (7390): pp. 428–33.

Marques, A. R. (2015). "Laboratory diagnosis of Lyme disease: advances and challenges." eng. *Infectious disease clinics of North America* 29 (2): pp. 295–307.

Marshall, J. C. (2014). "Why have clinical trials in sepsis failed?" eng. *Trends in molecular medicine* 20 (4): pp. 195–203.

Marti-Carvajal, A. J., Sola, I., Lathyris, D., et al. (2012). "Human recombinant activated protein C for severe sepsis." eng. *The Cochrane database of systematic reviews* (3): p. CD004388.

McCarthy, S., Das, S., Kretzschmar, W., et al. (2016). "A reference panel of 64,976 haplotypes for genotype imputation." eng. *Nature genetics* 48 (10): pp. 1279–83.

McCloskey, R. V., Straube, R. C., Sanders, C., et al. (1994). "Treatment of septic shock with human monoclonal antibody HA-1A. A randomized, double-blind, placebo-controlled trial. CHESS Trial Study Group." eng. *Annals of internal medicine* 121 (1): pp. 1–5.

Mee, E. T., Preston, M. D., Minor, P. D., et al. (2016). "Development of a candidate reference material for adventitious virus detection in vaccine and biologicals manufacturing by deep sequencing". *Vaccine* 34: pp. 2035–2043.

Moore, K. W., Waal Malefyt, R. de, Coffman, R. L., et al. (2001). "Interleukin-10 and the interleukin-10 receptor." eng. *Annual review of immunology* 19: pp. 683–765.

Morrison, T. E., Mauser, A., Wong, A., et al. (2001). "Inhibition of IFN-gamma signaling by an Epstein-Barr virus immediate-early protein." eng. *Immunity* 15 (5): pp. 787–99.

Munford, R. S. and Pugin, J. (2001). "Normal responses to injury prevent systemic inflammation and can be immunosuppressive." eng. *American journal of respiratory and critical care medicine* 163 (2): pp. 316–21.

Neumann-Haefelin, C., McKiernan, S., Ward, S., et al. (2006). "Dominant influence of an HLA-B27 restricted CD8+ T cell response in mediating HCV clearance and evolution." eng. *Hepatology (Baltimore, Md.)* 43 (3): pp. 563–72.

Nicolae, D. L., Gamazon, E., Zhang, W., et al. (2010). "Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS." eng. *PLoS genetics* 6 (4): e1000888.

Nishida, N., Ohashi, J., Khor, S.-S., et al. (2016). "Understanding of HLA-conferred susceptibility to chronic hepatitis B infection requires HLA genotyping-based association analysis." eng. *Scientific reports* 6: p. 24767.

Ong, D. S. Y., Bonten, M. J. M., Spitoni, C., et al. (2017). "Epidemiology of Multiple Herpes Viremia in Previously Immunocompetent Patients With Septic Shock." eng. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 64 (9): pp. 1204–1210.

Pagano, M. A., Tibaldi, E., Palu, G., et al. (2013). "Viral proteins and Src family kinases: Mechanisms of pathogenicity from a "liaison dangereuse"." eng. *World journal of virology* 2 (2): pp. 71–8.

Park, H. K., Lee, H. J., and Kim, W. (2010). "Real-time PCR assays for the detection and quantification of Streptococcus pneumoniae". *FEMS Microbiol Lett* 310: pp. 48–53.

Paschos, K. and Allday, M. J. (2010). "Epigenetic reprogramming of host genes in viral and microbial pathogenesis." eng. *Trends in microbiology* 18 (10): pp. 439–47.

Peltola, V. T., Murti, K. G., and McCullers, J. A. (2005). "Influenza virus neuraminidase contributes to secondary bacterial pneumonia." eng. *The Journal of infectious diseases* 192 (2): pp. 249–57.

Pereyra, F., Jia, X., McLaren, P. J., et al. (2010). "The major genetic determinants of HIV-1 control affect HLA class I peptide presentation." eng. *Science (New York, N.Y.)* 330 (6010): pp. 1551–7.

Peterson, J., Garges, S., Giovanni, M., et al. (2009). "The NIH Human Microbiome Project." eng. *Genome research* 19 (12): pp. 2317–23.

Quick, J., Ashton, P., Calus, S., et al. (2015). "Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella." eng. *Genome biology* 16: p. 114.

Quick, J., Loman, N. J., Duraffour, S., et al. (2016). "Real-time, portable genome sequencing for Ebola surveillance." eng. *Nature* 530 (7589): pp. 228–232.

Quick, J., Grubaugh, N. D., Pullan, S. T., et al. (2017). "Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples." eng. *Nature protocols* 12 (6): pp. 1261–1276.

Radhakrishnan, J. (2012). "Functional Genomics of Severe Sepsis and Septic Shock". PhD thesis. University of Oxford.

Ramsuran, V., Naranbhai, V., Horowitz, A., et al. (2018). "Elevated HLA-A expression impairs HIV control through inhibition of NKG2A-expressing cells." eng. *Science (New York, N.Y.)* 359 (6371): pp. 86–90.

Rando, O. J. and Verstrepen, K. J. (2007). "Timescales of genetic and epigenetic inheritance." eng. *Cell* 128 (4): pp. 655–68.

Rautanen, A., Mills, T. C., Gordon, A. C., et al. (2015). "Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study." eng. *The Lancet. Respiratory medicine* 3 (1): pp. 53–60.

Rello, J., Lisboa, T., Lujan, M., et al. (2009). "Severity of pneumococcal pneumonia associated with genomic bacterial load." eng. *Chest* 136 (3): pp. 832–840.

Ripa, T. and Nilsson, P. (2006). "A variant of Chlamydia trachomatis with deletion in cryptic plasmid: implications for use of PCR diagnostic tests." eng. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin* 11 (11): E061109.2.

Ritchie, M. E., Phipson, B., Wu, D., et al. (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." eng. *Nucleic acids research* 43 (7): e47.

Robin, X., Turck, N., Hainard, A., et al. (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves." eng. *BMC bioinformatics* 12: p. 77.

Rogers, G. B., Zain, N. M. M., Bruce, K. D., et al. (2014). "A novel microbiota stratification system predicts future exacerbations in bronchiectasis." eng. *Annals of the American Thoracic Society* 11 (4): pp. 496–503.

Rol, M.-L., Venet, F., Rimmele, T., et al. (2017). "The REAnimation Low Immune Status Markers (REALISM) project: a protocol for broad characterisation and follow-up of injury-induced immunosuppression in intensive care unit (ICU) critically ill patients." eng. *BMJ open* 7 (6): e015734.

Root, R. K., Lodato, R. F., Patrick, W., et al. (2003). "Multicenter, double-blind, placebo-controlled study of the use of filgrastim in patients hospitalized with pneumonia and severe sepsis." eng. *Critical care medicine* 31 (2): pp. 367–73.

Ryan, J. L., Fan, H., Glaser, S. L., et al. (2004). "Epstein-Barr virus quantitation by real-time PCR targeting multiple gene segments: a novel approach to screen for the virus in paraffin-embedded tissue and plasma". *J Mol Diagn* 6: pp. 378–85.

Salter, S. J., Cox, M. J., Turek, E. M., et al. (2014). "Reagent and laboratory contamination can critically impact sequence-based microbiome analyses". *BMC Biol* 12: p. 87.

Salzberg, S. L., Breitwieser, F. P., Kumar, A., et al. (2016). "Next-generation sequencing in neuropathologic diagnosis of infections of the nervous system". *Neurol Neuroimmunol Neuroinflamm* 3: e251.

Sanderson, N. D., Street, T. L., Foster, D., et al. (2018). "Real-time analysis of nanopore-based metagenomic sequencing from infected orthopaedic devices." eng. *BMC genomics* 19 (1): p. 714.

Satoh JK N; Yamamoto, Y. (2013). "Molecular network of chromatin immunopreciptation followed by deep sequencing-based (ChIP-Seq) Epstein-Barr virus nuclear antigen 1-target cellular genes supports biological implications of Epstein-Barr virus persistence in multiple sclerosis." *Clinical and Experimental Neuroimmunology*: pp. 181–192.

Scherag, A., Schoneweck, F., Kesselmeier, M., et al. (2016). "Genetic Factors of the Disease Course after Sepsis: A Genome-Wide Study for 28Day Mortality." eng. *EBioMedicine* 12: pp. 239–246.

Schmidt, K., Mwaigwisya, S., Crossman, L. C., et al. (2017). "Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing". *J Antimicrob Chemother* 72: pp. 104–114.

Schoneweck, F., Kuhnt, E., Scholz, M., et al. (2015). *Common genomic variation in the FER gene: useful to stratify patients with sepsis due to pneumonia?* eng. United States.

Schuenemann, V. J., Kumar Lankapalli, A., Barquera, R., et al. (2018). "Historic Treponema pallidum genomes from Colonial Mexico retrieved from archaeological remains." eng. *PLoS neglected tropical diseases* 12 (6): e0006447.

Scicluna, B. P., Vught, L. A. van, Zwinderman, A. H., et al. (2017). "Classification of patients with sepsis according to blood genomic endotype: a prospective cohort study." eng. *The Lancet. Respiratory medicine* 5 (10): pp. 816–826.

Sedlak, R. H., Cook, L., Cheng, A., et al. (2014). "Clinical utility of droplet digital PCR for human cytomegalovirus". *J Clin Microbiol* 52: pp. 2844–8.

Seymour, C. W., Gesten, F., Prescott, H. C., et al. (2017). "Time to Treatment and Mortality during Mandated Emergency Care for Sepsis." eng. *The New England journal of medicine* 376 (23): pp. 2235–2244.

Shu, B., Wu, K. H., Emery, S., et al. (2011). "Design and performance of the CDC real-time reverse transcriptase PCR swine flu panel for detection of 2009 A (H1N1) pandemic influenza virus". *J Clin Microbiol* 49: pp. 2614–9.

Simner, P. J., Miller, S., and Carroll, K. C. (2018). "Understanding the Promises and Hurdles of Metagenomic Next-Generation Sequencing as a Diagnostic Tool for Infectious Diseases." eng. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 66 (5): pp. 778–788.

Singer, M., Deutschman, C. S., Seymour, C. W., et al. (2016). "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)." eng. *JAMA* 315 (8): pp. 801–10.

Sironi, M., Cagliani, R., Forni, D., et al. (2015). "Evolutionary insights into host-pathogen interactions from mammalian sequence data." eng. *Nature reviews. Genetics* 16 (4): pp. 224–36.

Sorensen, T. I., Nielsen, G. G., Andersen, P. K., et al. (1988). "Genetic and environmental influences on premature death in adult adoptees." eng. *The New England journal of medicine* 318 (12): pp. 727–32.

Stammler, F., Glasner, J., Hiergeist, A., et al. (2016). "Adjusting microbiome profiles for differences in microbial load by spike-in bacteria." eng. *Microbiome* 4 (1): p. 28.

Sweeney, T. E., Shidham, A., Wong, H. R., et al. (2015). "A comprehensive time-course-based multicohort analysis of sepsis and sterile inflammation reveals a robust diagnostic gene set." eng. *Science translational medicine* 7 (287): 287ra71.

Sweeney, T. E., Wong, H. R., and Khatri, P. (2016). "Robust classification of bacterial and viral infections via integrated host gene expression diagnostics." eng. *Science translational medicine* 8 (346): 346ra91.

Tang, B. M., Shojaei, M., Parnell, G. P., et al. (2017). "A novel immune biomarker IFI27 discriminates between influenza and bacteria in patients with suspected respiratory infection." eng. *The European respiratory journal* 49 (6).

Tang, B. M. P., McLean, A. S., Dawes, I. W., et al. (2008). "Gene-expression profiling of gram-positive and gram-negative sepsis in critically ill patients." eng. *Critical care medicine* 36 (4): pp. 1125–8.

Thaiss, C. A., Zmora, N., Levy, M., et al. (2016). "The microbiome and innate immunity." eng. *Nature* 535 (7610): pp. 65–74.

Tsalik, E. L., Henao, R., Nichols, M., et al. (2016). "Host gene expression classifiers diagnose acute respiratory illness etiology." eng. *Science translational medicine* 8 (322): 322ra11.

Vayssier-Taussat, M., Albina, E., Citti, C., et al. (2014). "Shifting the paradigm from pathogens to pathobiome: new concepts in the light of meta-omics." eng. *Frontiers in cellular and infection microbiology* 4: p. 29.

Vincent, J. L., Brealey, D., Libert, N., et al. (2015). "Rapid Diagnosis of Infection in the Critically Ill, a Multicenter Study of Molecular Detection in Bloodstream Infections, Pneumonia, and Sterile Site Infections". *Crit Care Med* 43: pp. 2283–91.

Vincent, J.-L., Marshall, J. C., Namendys-Silva, S. A., et al. (2014). "Assessment of the worldwide burden of critical illness: the intensive care over nations (ICON) audit." eng. *The Lancet. Respiratory medicine* 2 (5): pp. 380–6.

Wagner, K., Springer, B., Pires, V. P., et al. (2018). "Molecular detection of fungal pathogens in clinical specimens by 18S rDNA high-throughput screening in comparison to ITS PCR and culture." eng. *Scientific reports* 8 (1): p. 6964.

Walton, A. H., Muenzer, J. T., Rasche, D., et al. (2014). "Reactivation of multiple viruses in patients with sepsis." eng. *PloS one* 9 (2): e98819.

Wang, H.-j., Zhang, P.-j., Chen, W.-j., et al. (2012). "Four serum microRNAs identified as diagnostic biomarkers of sepsis." eng. *The journal of trauma and acute care surgery* 73 (4): pp. 850–4.

Wang, J.-f., Yu, M.-l., Yu, G., et al. (2010). "Serum miR-146a and miR-223 as potential new biomarkers for sepsis." eng. *Biochemical and biophysical research communications* 394 (1): pp. 184–8.

Watanabe, N., Kryukov, K., Nakagawa, S., et al. (2018). "Detection of pathogenic bacteria in the blood from sepsis patients using 16S rRNA gene amplicon sequencing analysis." eng. *PloS one* 13 (8): e0202049.

Werdan, K., Pilz, G., Bujdoso, O., et al. (2007). "Score-based immunoglobulin G therapy of patients with sepsis: the SBITS study." eng. *Critical care medicine* 35 (12): pp. 2693–2701.

Wilson, M. R., Sample, H. A., Zorn, K. C., et al. (2019). "Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis." eng. *The New England journal of medicine* 380 (24): pp. 2327–2340.

Woese, C. R. (1987). "Bacterial evolution." eng. *Microbiological reviews* 51 (2): pp. 221–71.

Wong, A. M. G., Kong, K. L., Chen, L., et al. (2013). "Characterization of CACNA2D3 as a putative tumor suppressor gene in the development and progression of nasopharyngeal carcinoma." eng. *International journal of cancer* 133 (10): pp. 2284–95.

Wood, D. E. and Salzberg, S. L. (2014). "Kraken: ultrafast metagenomic sequence classification using exact alignments". *Genome Biol* 15: R46.

Worldwide Populations, A. F. in. *Allele Frequencies in Worldwide Populations*. Accessed 13 August 2019. URL: http://allelefrequencies.net.

Wright, F. A., Sullivan, P. F., Brooks, A. I., et al. (2014). "Heritability and genomics of gene expression in peripheral blood." eng. *Nature genetics* 46 (5): pp. 430–7.

Wu, H.-J., Ivanov, I. I., Darce, J., et al. (2010). "Gut-residing segmented filamentous bacteria drive autoimmune arthritis via T helper 17 cells." eng. *Immunity* 32 (6): pp. 815–27.

Wylie, T. N., Wylie, K. M., Herter, B. N., et al. (2015). "Enhanced virome sequencing using targeted sequence capture." eng. *Genome research* 25 (12): pp. 1910–20.

Xu, M., Qin, X., Astion, M. L., et al. (2013). "Implementation of filmarray respiratory viral panel in a core laboratory improves testing turnaround time and patient care". *Am J Clin Pathol* 139: pp. 118–23.

Zhang, D., Chen, G., Manwani, D., et al. (2015a). "Neutrophil ageing is regulated by the microbiome." eng. *Nature* 525 (7570): pp. 528–32.

Zhang, F.-R., Huang, W., Chen, S.-M., et al. (2009). "Genomewide association study of leprosy." eng. *The New England journal of medicine* 361 (27): pp. 2609–18.

Zhang, X., Zhang, D., Jia, H., et al. (2015b). "The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment." eng. *Nature medicine* 21 (8): pp. 895–905.

Zou H; Hastie, T (2005). "Regularization and variable selection via the elastic net". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67: pp. 301–320.