Rare Genetic Variants and Susceptibility to Severe Bacterial Diseases



Anne Wanjiru Ndungu

Somerville College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2015

Rare Genetic Variants and Susceptibility to Severe Bacterial Diseases

Anne Wanjiru Ndungu

Somerville College, University of Oxford, D.Phil

Michaelmas 2015

Abstract

Infectious diseases are a major cause of morbidity and mortality worldwide. *Streptococcus pneumoniae* and *Neisseria meningitidis* are major causes of severe bacterial disease which can manifest as invasive disease such as bacteraemia and meningitis. Exposure to these pathogens is relatively widespread, yet only a minority of individuals develop invasive disease. A host genetic component to infectious disease susceptibility has been implied from twin and adoptee studies. A role for rare large effect genetic variants in predisposition to infection has been demonstrated through the study of individuals with primary immunodeficiencies. However, a majority of these studies have been undertaken in individuals with a history of recurrent disease or in multi-case families. The relative role of rare genetic variants of moderate to large effect at the population level has not been widely explored. This thesis presents effort made using next generation sequencing methods to identify rare genetic variants that lead to increased susceptibility to bacterial disease focussing on meningococcal disease, pleural infection(empyema), pneumococcal disease and sepsis phenotypes.

Using an exome sequencing approach in 13 cases with invasive meningococcal disease, a novel mutation leading to a complement deficiency and increased risk of meningococcal infection was identified and functionally validated in one individual. This mutation in the *CFP* gene was demonstrated as leading to impaired properdin secretion. Further analysis implicated loss of function mutations in *CD4* and *ZAP70* as novel loci for meningococcal disease susceptibility.

A case control association analysis for sepsis susceptibility highlighted the possible role for small Rho GTPases in sepsis pathology. By aggregating all rare predicted deleterious mutations in a gene, four genes in this pathway, (*ROCK2*, *ARHGAP18*, *FYN* and *CDC42BPG*) were implicated as having an excess of rare deleterious variants in sepsis samples compared to population controls.

A similar approach identified low frequency genetic variants in the *CD109* gene as predisposing to empyema susceptibility in children.

Finally, preliminary evidence from adult individuals with invasive pneumococcal disease points to a potential role of the *RNASE7* gene in invasive pneumococcal disease susceptibility. This association was primarily due to a predicted deleterious missense mutation present in cases and absent in controls.

Taken together, these results have identified a number of potential loci with rare variants associated with susceptibility to severe phenotypes of bacterial diseases.

Acknowledgements

Firstly, I would like to thank my supervisors Dr Stephen Chapman and Dr Katherine Elliott. The patience, guidance and support you have provided over the last three years has been remarkable and I will be forever grateful.

Secondly I would like to thank the members of Hill genetics group and colleagues at the Wellcome Trust Centre for Human Genetics. To be surrounded by such brilliant minds has been a great pleasure and I would not have wished for better colleagues. Your enthusiasm for science has been contagious and I am a much better scientist because of each one of you. To Prof. Adrian Hill, you have been a great mentor. John Broxholme, I would have given up on programming a long time ago were it not for your amazing patience and understanding through all my incessant questions and coding challenges right from the beginning, thank you!

I wish to thank great mentors I have encountered Prof. Kevin Marsh and Dr Evelyn Gitau who believed in me and encouraged me to set out on this path, Prof. Katsuyuki Yui who taught me all I know on molecular cloning techniques used in this project.

To the great friendships I have formed during the course of this research, thank you for keeping me grounded and sane. Nyambura Irungu, this would have been a much more difficult journey were it not for the great travels, laughs and tears we have had together. I thank Sally Njoki and Lucy Njeri for the constant support and encouragement. Despite being miles away, you have both taught me the true meaning of perseverance. To the great friends I have made in the UK, thank you for making it a home away from home for me.

And to my family, you have been a great source of strength that I could tap into whenever I ran low. My mum, for being my biggest cheerleader throughout this journey and to my dad for setting the foundations that led me into science. To my siblings, my two brothers Moses and Kamau for your constant push for me to give my all, my sister Irene, the one person I could call at any time day or night, you have provided such amazing and constant support. This has been as much my project as it has been yours, you have shared in all the highs and lows. Words cannot do justice to how grateful I am. Thank you all.

Attributions

The sample collections presented in this thesis are due to the great efforts made by individuals involved in the GenOSept, GAinS, Oxfordshire Pneumococcal surveillance programme and the National Programme for Enhanced Pneumococcal Surveillance teams. Tara Mills performed all the DNA extractions and quantification of the empyema samples.

The Wellcome Trust Centre for Human Genetics, Oxford Genomics Centre conducted all the sequencing for all samples used in these studies. Lorna Witty, Lorne Lornie and Simon Engledow managed these projects during sequencing and John Broxholme performed the quality control for raw reads as well as mapping of the empyema and IPD samples. Sanger sequencing was carried out by Lynne Richardson at the Zoology Department, University of Oxford

I would also like to acknowledge the contributions made by members of the Simon Draper research group at the Old Road Campus Research Building. Jing Jin routinely maintained in culture the HEK293E cells used in this project and Daniel Alanine for use of his molecular biology reagents. Keith Morris at the Wellcome Trust Centre for Human Genetics Cellular Imaging Core assisted in flourescent microscopy.

All research work presented in this thesis is original and conducted by the author at the Wellcome Trust Centre for Human Genetics under the supervision of Dr Stephen Chapman and Dr Katherine Elliott. Any part of this work not conducted by the author is cited in the text. No part of this thesis has been submitted for any degree, diploma or certificate or other qualification at the University of Oxford or elsewhere.

Contents

List of Figures viii

List of Tables ix

List of Abbreviations x

1 Introduction 1
- 1.1 The global burden of bacterial diseases . . . 1
- 1.2 Sepsis . . . 1
- 1.3 Invasive meningococcal disease . . . 4
- 1.4 Invasive pneumococcal disease . . . 8
- 1.5 Pleural infection (empyema) . . . 11
- 1.6 A genetic basis for human disease . . . 12
- 1.7 Current methods in human genetics of complex traits . . . 14
- 1.8 Rare variant association analysis . . . 16
- 1.9 Statistical methods in rare variant analysis . . . 18
- 1.10 Variant prioritisation in rare variant analysis . . . 19
- 1.11 Rare host genetic variants in infectious diseases . . . 20
- 1.12 Overview of this thesis . . . 23

2 Materials and Methods 25
- 2.1 Introduction . . . 25
- 2.2 Disease cohorts . . . 25
  - 2.2.1 Meningococcal disease . . . 25
  - 2.2.2 Community acquired pneumonia . . . 26
  - 2.2.3 Oxfordshire invasive pneumococcal disease . . . 26
  - 2.2.4 Pleural infection (empyema) . . . 27
- 2.3 Control cohorts . . . 27
  - 2.3.1 Avon Longitudinal Survey of Parents and Children (ALSPAC) 28
  - 2.3.2 Scottish Family Health Study obesity controls (UK10K-OBESITY-GS) . . . 28

2.3.3 The 1958 Birth Cohort (58BC) . . . 28
2.4 Laboratory methods . . . 29
2.4.1 DNA extraction and quantification . . . 29
2.4.2 Quality control . . . 29
2.4.3 Whole genome amplification . . . 29
2.4.4 Exome capture and sequencing . . . 30
2.5 Bioinformatics pipeline for the analysis of sequence data . . . 30
2.5.1 Mapping . . . 30
2.5.2 Post mapping processing . . . 31
2.5.3 Marking duplicates . . . 31
2.5.4 Local realignment around indels . . . 31
2.5.5 Recalibrate base quality score . . . 32
2.5.6 Variant calling . . . 32
2.5.7 Quality controls of variants . . . 33
2.6 Analytical methods . . . 35
2.6.1 Population stratification . . . 35
2.6.2 Multi-Dimensional Scaling (MDS) . . . 36
2.6.3 Principal Component Analysis (PCA) . . . 36
2.7 Statistical analysis . . . 37
2.7.1 Single variant association analysis . . . 37
2.7.2 Gene-based methods of analysis . . . 38
2.7.3 Strategies for aggregation of variants . . . 38
2.7.4 Statistical methods of gene-based analysis . . . 39
2.7.5 Genomics analysis of biological networks . . . 39
2.8 Molecular biology methods used in validation of genetic findings . . . 40
2.8.1 Sanger sequencing . . . 40
2.8.2 Molecular cloning . . . 41

3 Rare genetic variation and susceptibility to meningococcal disease 46
3.1 Introduction . . . 46
3.2 Results . . . 47
3.2.1 Sequence metrics . . . 47
3.2.2 Novel variants in meningococcal disease . . . 48
3.2.3 Rare deleterious mutations in meningococcal disease . . . 61
3.3 Discussion . . . 69

4 Genetic susceptibility to sepsis **73**
4.1 Introduction . . . 73
4.2 Results . . . 74
4.2.1 Sequencing metrics . . . 74
4.2.2 Analysis with ALPSAC controls . . . 76
4.2.3 Analysis with Generation Scotland: Scottish Family Health Study (UK10K-OBESITY-GS) controls . . . 79
4.2.4 Association analysis using 1958 Birth Cohort (58BC) controls 87
4.3 Discussion . . . 98

5 Genetic susceptibility to paediatric pleural infection **102**
5.1 Introduction . . . 102
5.2 Results . . . 102
5.2.1 Pilot study . . . 102
5.2.2 Population structure . . . 107
5.2.3 Single variant analysis . . . 107
5.2.4 Gene burden analysis . . . 121
5.2.5 Pathway enrichment analysis . . . 124
5.3 Discussion . . . 127

6 Genetic susceptibility to invasive pneumococcal disease **130**
6.1 Introduction . . . 130
6.2 Results . . . 131
6.2.1 Pilot study . . . 131
6.2.2 Population structure . . . 135
6.2.3 Single variant analysis . . . 135
6.2.4 Gene burden analysis . . . 147
6.2.5 Pathway enrichment analysis . . . 151
6.3 Discussion . . . 152

7 Discussion **155**

8 Exome sequencing **161**
8.1 Library preparation . . . 161
8.2 Cluster generation . . . 161
8.3 Sequencing . . . 162

9 KEGG pathway analysis for genes with case only loss of function mutations in meningococcal cases **163**

10 Inclusion criteria for the GenOSept study **170**
10.1 Inclusion criteria 170
10.2 Exclusion criteria 172

11 IPD research protocol for ethics application **173**

12 Whole genome amplification in exome sequencing **185**
12.1 Results 185
12.2 Discussion 186

Bibliography **188**

List of Figures

3.1 Sequencing coverage in meningococcal samples 49
3.2 Confirmation of R346C mutation 51
3.3 Vector map for 1990_pENTR4LPTOS 53
3.4 GeneArt® Strings™ (Invitrogen) 346Arg DNA fragment 54
3.5 GeneArt® Strings™ (Invitrogen) 346Cys DNA fragment 55
3.6 Properdin expression assays 56
3.7 HEK293 intracellular staining 57
3.8 Secondary structure for the trimeric form of properdin 58
3.9 Interaction networks for genes with case only LoF 64
3.10 Meningococcal disease gene burden analysis 66
3.11 Filtration and prioritisation of novel variants in meningococcal case . 67
3.12 Case-control variant prioritisation 68

4.1 CAP sequencing coverage 75
4.2 ALSPAC-CAP population structure analysis 77
4.3 ALSPAC-CAP single variant analysis 78
4.4 UK10K-CAP MDS analysis 80
4.5 UK10K-GS-CAP Q-Q plots 85
4.6 58BC-CAP MDS analysis 88
4.7 Protein network analysis for Rho GTPase genes with mutations . . . 97

5.1 Coverage in all empyema samples 106
5.2 58BC-empyema MDS analysis 109
5.3 Manhattan plots for single variant analysis 119
5.4 Q-Q plots for single variant analysis 120

6.1 All IPD samples sequence coverage 134
6.2 58BC-IPD MDS analysis 136
6.3 Manhattan plot for IPD single variant analysis 145
6.4 Plots for IPD single variant analysis 146

List of Tables

3.1 Sequence metrics summary for meningococcal disease cases 48
3.2 KEGG pathways enriched for novel mutations in meningococcal cases 60
3.3 Genes with LoFs in meningococcal cases only 63

4.1 Summary of CAP sequence metrics 74
4.2 UK10K-GS-CAP single variant analysis 83
4.3 UK10K/GS-CAP gene burden analysis 86
4.4 58BC-CAP single variant analysis 90
4.5 KEGG pathways enrichment for top variants in single variant analysis 92
4.6 58BC-CAP gene burden analysis 95
4.7 58BC-CAP pathway analysis 96

5.1 Sequence metrics for 12 empyema pilot samples 104
5.2 Individual metrics for pilot samples 105
5.3 Empyema single variant analysis 110
5.4 Pneumococcal empyema single variant analysis 115
5.5 Empyema gene burden analysis 123
5.6 Pneumococcal empyema gene burden analysis 125

6.1 Summary sequence metrics for 12 IPD pilot samples 132
6.2 Individual metrics for pilot samples 133
6.3 IPD single variant analysis 137
6.4 IPD gene burden analysis 149

9.1 KEGG pathways with case only mutations 164

12.1 Exome sequencing of whole genome amplified DNA 187

List of Abbreviations

ACCP . . . . . American college of chest physicians

ALSPAC . . . Avon longitdinal survey of parents and children

BMI . . . . . Body mass index

C3 . . . . . Complement component 3

CAP . . . . . Community acquired pneumonia

CAST . . . . . Cohort allelic sums test

CCDS . . . . . Consensus coding sequence

CMC . . . . . Combined multivariate and collapsing

CMV . . . . . Cytomegalovirus

COPD . . . . . Chronic obstructive pulmonary disease

CSF . . . . . Cerebrospinal fluid

CVID . . . . . Common variable immunodeficiency

DNA . . . . . Deoxyribonucleic acid

EBV . . . . . Epstein barr virus

EDTA . . . . . Ethylenediaminetetraacetic acid

EMMAX . . . Efficient mixed model association eXpedited

EPACTS . . . Efficient and parallelizable association container toolbox

ExAC . . . . . Exome aggregation consortium

fHbp . . . . . Factor H binding protein

FS . . . . . Fisher's strand

GAinS . . . . . Genomic Advances in Sepsis

GATK . . . . . Genome analysis tool-kit

GenOSept . . Genetics Of Sepsis and Septic Shock

GFP . . . . . Green flourescent protein

GO . . . . . Gene ontology

GQ . . . . . Genotype quality

GWAS . . . . . Genome-wide association study

HEK293E . . . Human embryonic kidney cells expressing EBV EBNA1 antigen

HLA . . . . . Human leukocyte antigen

HRUN . . . . Homopolymer run

IBS . . . . . Identity by state

ICU . . . . . Intensive care unit

IFN . . . . . Interferon

IMD . . . . . Invasive meningococcal disease

IPD . . . . . Invasive pneumococcal disease

KEGG . . . . . Kyoto encyclopedia of genes and genomes

LB . . . . . Luria Bertan

LD . . . . . Linkage disequillibrium

LDL . . . . . Low density lipoproteins

LMM . . . . . Linear mixed model

LoF . . . . . Loss of function

LOS . . . . . Lipooligosaccharides

MAC . . . . . Membrane attack complex

MAF . . . . . Minor allele frequency

MDS . . . . . Multi-dimensional scaling

MHC . . . . . Major histocompatibility complex

MLST . . . . . Multilocus sequence typing

mRNA . . . . Messenger ribonucleic acid

NadA . . . . . neisserial adhesin A

NCBI . . . . . National centre for biotechnological information

NF-κB . . . . . Nuclear Factor-κB

NHBA . . . . *Neisseria* heparin-binding antigen

NHLBI . . . . National heart, lung and blood institute

OXREC . . . . Oxford research ethics committee

PCA . . . . . Principal components analysis

PCR . . . . . Polymerase chain reaction

PCV . . . . . Pneumococcal conjugate vaccine

PEI . . . . . Polyethylenimine

PID . . . . . Primary immunodeficiency

PPSV . . . . . Pneumococcal polysaccharide vaccine

Pra . . . . . Protein altering

SCCM . . . . . Society of critical care medicine

SCID . . . . . Severe combined immunodeficiency

SIFT . . . . . Sorting intolerant from tolerant

SIRS . . . . . Systemic inflammatory response syndrome

SKAT . . . . . Sequence kernel association test

SNP . . . . . Single nucleotide polymorphism

SOC . . . . . Super optimal broth with catabolite repression

TCA . . . . . Tri-carboxylic acid

TDT . . . . . Transmission disequlibrium test

TLR . . . . . Toll-like receptor

UCSC . . . . . University of California at Santa Cruz

UK . . . . . United Kingdom

UTR . . . . . Untranslated region

VCF . . . . . Variant call format

VEP . . . . . Variant effect predictor

WES . . . . . Whole exome sequencing

WGS . . . . . Whole genome sequencing

Chapter 1
Introduction

1.1 The global burden of bacterial diseases

Bacterial diseases remain a major public health challenge in both the developing and developed world. Despite the advent of antibiotics, development of modern health care facilities and an improved understanding of the pathogenic mechanisms underlying disease, bacterial diseases are still associated with high rates of morbidity and mortality worldwide. In 2013, 20% of global deaths in neonates were due to sepsis and pneumonia [1]. Lower respiratory tract infections (LRTI) which include pneumonia and bronchitis were the fourth leading cause of death globally in 2012, leading to 3.1 million deaths. In low income countries, LRTIs accounted for 91 in every 100,000 deaths in 2012, resulting in more deaths than HIV/AIDS, malaria or tuberculosis. Together, these four diseases account for over 30% of all deaths in low income countries [2]. The high burden of infectious diseases in both developing and the developed world reflects a gap in our knowledge on invasive diseases for adequate interventions both preventative and therapeutic.

1.2 Sepsis

The term sepsis can be traced to Hippocrates (ca.460-370BC), derived from the Greek word “sípsi” which translates to “make rotten, decay or putrefaction”. Prior to the 19th century, the pathophysiology remained unknown with theories such as contact of a wound with bad air or oxygen suggested as possible causes. Ignaz Semmelweis, widely

credited with reducing maternal deaths after childbirth by introducing simple hand washing measures, deduced that septic fever was due to external matter that gained access to the bloodstream [3]. It was not until Louis Pasteur's advent of the germ theory that Joseph Lister linked microbes to disease and introduced sterilisation practices during surgery [4].

An initial consensus definition for sepsis was published in 1992 following an American College of Chest Physicians/Society of Critical Care Medicine (ACCP/SCCM) Consensus Conference held in 1991 [5]. Sepsis was defined as a systemic inflammatory response to infection, differentiating it from systemic inflammatory response syndrome (SIRS) which can occur in the absence of an infection. The following criteria were formally outlined for use in sepsis diagnosis:

1. A body temperature greater than 38°C or less than 36°C
2. A heart rate greater than 90 beats per minute
3. A respiratory rate greater than 20 breaths per minute or hyperventilation as indicated by a $PaCO_2$ of less than 32mmHg
4. A white blood cell count greater than 12,000/mm^3, less than 4,000/mm^3 or the presence of more than 10% immature neutrophils

Furthermore, formal definitions for the terms bacteraemia, septicaemia, severe sepsis and septic shock were resolved. Bacteraemia was defined as the presence of viable bacteria in the blood. Severe sepsis was defined as sepsis associated with organ dysfunction, hypotension or hypoperfusion, while septic shock was defined as sepsis complicated by hypoperfusion despite adequate fluid resuscitation [5,6]. These criteria were expanded in 2001 to reflect varied signs and symptoms as observed by physicians during patient care. Moreover, guidelines on classification of patients were provided and termed "PIRO". These allowed the stratification of patients based on **P**redisposing conditions, the nature and extent of **I**nfection, the nature and magnitude of the host **R**esponse, and the degree of concomitant **O**rgan damage [6]. Whereas sepsis has

historically been defined with regards to inflammatory processes, it is now widely accepted that a range of pathways other than the immune system are dysregulated in sepsis [7]. At the time of writing, the third international consensus definitions for sepsis and septic shock have been outlined to reflect an improved understanding of the pathophysiological processes that occur [8]. Sepsis is now defined as a dysregulated response to an infection leading to organ dysfunction [8, 9]. The Sequential Organ Failure Assessment (SOFA) scoring system is now recommended to define the degree of organ failure [9]. A SOFA score greater than 2 in addition to an infection is indicative of sepsis. A baseline SOFA score of 0 is refers to an infection with no organ damage, a score of 2 or greater is associated with a 2-25 fold increase in mortality rate in individuals with an infection [8, 9].

Initially, the pathophysiology of sepsis was attributed to an over-stimulated immune response following an infection [7, 10]. Based on animal studies implicating increased cytokine levels with both increased organ damage and correlation with death [11], clinical trials were conducted aimed at modulating the host immune response [12–16]. While an activated pro-inflammatory state occurs in the initial stages of sepsis, it has now emerged that there follows a period of prolonged immunosuppression [17, 18]. This has been attributed to a shift in the adaptive immune response from a primarily pro-inflammatory type 1 helper T cell (Th1)-driven response to a type 2 helper T cell (Th2)-driven immunosuppressed state [7]. Moreover, there is increased apoptosis of lymphocytes and dendritic cells resulting in enhanced immunosuppression [19, 20].

The incidence of sepsis has continued to rise in recent years despite significant advances in intensive care medicine; moreover, sepsis-associated mortality rates remain high with different studies reporting mortality rates of 20- 50% [21–25]. Mortality rates are highly dependent on disease syndrome with one study reporting 16%, 20% and 46% for sepsis, severe sepsis and septic shock respectively [26]. Furthermore, sepsis survivors suffer from high rates of morbidity and mortality for up to five years after the septic episode, in addition to considerably reduced quality of life [27, 28]. Recently, the most common causes of hospital readmission in severe sepsis survivors were reported as reoccurrence of sepsis, heart failure, chronic obstructive pulmonary disease (COPD),

pneumonia and urinary tract infections; 12% of re-admissions involved a primary diagnosis of infection [29]. The precise mechanisms leading to increased morbidity and mortality are unclear; however prolonged post-sepsis immunosuppression has been suggested as a possible mechanism [30].

Between 1979 and 1987, Gram negative bacteria were the major cause of sepsis while Gram positive bacteria were the most identified group of pathogens from 1987 to 2000 [23]. More recently, Gram negative bacteria were isolated from 62% of all septic patients in a large multi-centre study [31]. Community acquired pneumonia (CAP) remains the leading cause of sepsis, severe sepsis and septic shock [25,32]. *Streptococcus pneumoniae* (pneumococcus) is the most commonly identified pathogen underlying CAP, although over 37% of all CAP cases lack an identified causal organism [25, 33].

Risk factors for sepsis are varied and include both host and environmental factors. Sepsis is more likely to occur in older adults over 65 years of age and especially residents of long-term care facilities [34]. Moreover, sepsis incidences are higher in individuals of lower socioeconomic background [35]. Chronic conditions such as diabetes, pulmonary diseases such as COPD and chronic kidney disease have also been shown to associate with sepsis risk [34].

1.3 Invasive meningococcal disease

Meningococcal disease is a major cause of meningitis and sepsis worldwide and follows invasive infection with the Gram negative bacterium *Neisseria meningitidis* which exclusively infects humans [36]. The first clinical description of meningococcal disease is attributed to Gaspard Vieusseux following an outbreak of the disease in Geneva in 1805 [37]. It wasn't until 1887 however, that the causative pathogen *Neisseria meningitidis* (initially referred to as *Diplococcus intracellularis meningitidis*) was isolated from the cerebrospinal fluid (CSF) of a patient by Anton Weichselbaum [38]. Before the early 1900s, mortality from meningococcal meningitis ranged above 70% [39], however, in 1913 Simon Flexner reported significant reduction in mortality rates to 31% in the first human trial using serum from immunised horses [39]. The subsequent

introduction of sulphonamides further reduced mortality to 10-20%, replacing serum therapy as the treatment of choice [40]. This development also led to sulphonamides being administered prophylactically, particularly in military settings to eliminate asymptomatic carriage [41]. However, widespread sulphonamide resistance led to increased reliance on antibiotics for the treatment of meningococcal diseases [42].

Asymptomatic nasopharyngeal carriage of *N. meningitidis* is relatively common in the general population, affecting approximately 5-10% of adults in non-epidemic settings with a peak of 23% in adolescents [43, 44]. In addition to age, carriage rates increase in crowded conditions such as in military barracks and the annual Hajj pilgrimage as well as due to increased social contact [45, 46]. In a small minority of individuals however, the pathogen is able to cross mucosal surfaces and invade the bloodstream, leading to invasive disease. Invasive meningococcal disease (IMD) presents with symptoms ranging from fever and rash to fulminant infection and death occurring within hours in otherwise healthy individuals [43, 47]. Clinical presentations of IMD can be:

1. Bacteraemia without sepsis, typically presenting as a low-grade fever in children [48]

2. Meningococcal sepsis (meningococcemia) without meningitis [48, 49]

3. Meningitis with or without meningococcemia [48, 49]

Meningitis and meningococcemia are the most common and severe types of IMD [48,49]. Meningitis occurs in about 30-60% of all cases and is characterised by fever, altered level of consciousness, photosensitivity, headaches and neck stiffness [48–50]. Mortality rates range from 5-18% [49]. Meningococcemia represents 20-30% of all cases of IMD [49]. A characteristic symptom of meningococcal sepsis is the appearance of a purpuric rash (purpura fulminans) usually in the lower extremities of the body [49]. The rash is indicative of thrombocytopenia and a dysregulated coagulation cascade [48].

There are 13 serogroups of *Neisseria meningitidis*, characterised by distinct capsular polysaccharides [51]. However, due to exchange of genetic material, meningococci are

genetically diverse and can be further subclassified into sequence types and clonal complexes [52]. Multilocus sequence typing (MLST) is now the gold standard for identifying meningococcal clonal groups [53, 54]. Housekeeping genes are sequenced and used to generate a profile for each sequence type [53, 55]. The sequence types are then grouped into clonal complexes [53, 55]. Sequence types and clonal complexes are deposited in the Neisseria MLST database (http://pubmlst.org/neisseria/) [56]. Five serogroups: A, B, C, W135 and Y are the common causes of IMD with 90% of incidences due to A, B and C [51]. However, incidences of IMD due to serogroup X have been reported in Niger [57]. In Europe, serogroup B is now the major cause of disease after the introduction of the serogroup C vaccine, while serogroup A is a common cause of epidemics in the meningitis belt of sub-Saharan Africa though outbreaks due to C, W-135, Y and X also occur [51].

The transition from asymptomatic carriage to invasive disease is dependent on a complex interplay between host susceptibility, microbial virulence factors and environmental factors, although the exact mechanism of invasion remains unclear [49]. *N. meningitidis* possesses several structural features that promote virulence. The pathogen expresses type IV filamentous pili which promote adhesion to host cells. Pili-mediated adhesion subsequently stimulates release of methyltransferase which then alters pili adhesion promoting movement and colonisation of new regions as well as invasion [58]. A common feature amongst invasive meningococcal isolates is the expression of specific opacity protein variants. Opacity proteins, specifically Opa and Opc, play a role in attachment of microbe to host cells by binding to CD66 and heparan sulphate proteoglycan receptors on host cells [50, 59, 60]. Invasive isolates of *N. meningitidis* are encapsulated, indicating a crucial role of the hydrophilic capsule in pathogenesis [61]. Indeed the capsule acts as a barrier to macrophage phagocytosis as well as resistance to complement mediated killing in the bloodstream [61]. The outer membrane of the meningococcus is composed of lipooligosaccharides (LOS), an endotoxin that resembles lipopolysaccharides (LPS) on other Gram negative pathogens but bears some distinct structural differences [36]. The bacterium releases copious amounts of LOS in vesicle-like structures known as blebs [62]. Endotoxin is a potent

activator of both the innate and the adaptive immune systems. Circulating levels of LOS directly correlate with increased levels of pro-inflammatory cytokines [63]. Moreover, LOS levels are associated with increased disease severity and poor prognosis [49,64]. The highest levels of circulating LOS are detected in individuals with fulminant meningococcal sepsis, a severe clinical presentation with mortality rates of up to 50% [64]. A key component in meningococcal invasion and survival in the bloodstream is the factor H binding protein (fHbp) expressed on the surface of the pathogen which recruits and binds human complement factor H to its surface. Complement factor H negatively regulates the alternative arm of the complement system through degradation of complement component 3 (C3). By binding factor H on their cell surface, meningococci are thus able to evade complement mediated killing in the bloodstream [65–67]. Factor H binding protein is widely expressed across meningococcal strains and is highly immunogenic, stimulating production of anti-fHbp serum bactericidal antibodies [68–70]. Due to this, fHbp, in addition to *Neisseria* heparin-binding antigen (NHBA) and the neisserial adhesin A (NadA) are key components in the newly licenced meningococcal serogroup B vaccine (4CMenB, Bexsero®) [71, 72].

The incidence of meningococcal disease is inversely correlated with levels of serum bactericidal antibodies [73, 74]. Infants have the least levels of bactericidal antibodies and the highest force of infection, defined as the rate at which susceptible individuals acquire either asymptomatic carriage or infectious disease [75]. Infants also have a higher carriage to disease ratio estimated at 0.07-0.12, this, however reduces by 10-50 fold within the first 10 years of life as immunity develops [75–77].

Current meningococcal vaccines are either composed of purified capsular polysaccharide antigens or polysaccharide antigens conjugated to a protein for enhanced immunogenicity. There are three licenced polysaccharide vaccines: a bivalent (A and C), a tetravalent (A, C and W-135) and the quadrivalent (A.C.W-135 and Y) [78]. Due to poor immunogenicity in children, polysaccharide vaccines are recommended for use in individuals over 2 years [79]. Licenced meningococcal conjugate vaccines are either monovalent (A, B, C) or quadrivalent (A, B, C, W-135) or in combination

with *Haemophilus influenzae* type b (HibMenC) [78]. The protein conjugate is either a diphtheria or tetanus toxoid [78].

1.4 Invasive pneumococcal disease

Streptococcus pneumoniae is a fastidious Gram positive bacteria that usually exists as pairs of cocci (diplococci) but may also be observed singly or in short chains; it does not form spores and is non-motile. The pneumococcus is alpha-haemolytic, whereby only partial breakdown of haemoglobin is observed when the microbe is cultured in blood agar. This is evident as a green discolouration surrounding the bacterial colony, in contrast to beta-haemolytic streptococci such as *Streptococcus pyogenes* where complete haemolysis results in an almost transparent region around the colony. Due to a lack of the catalase enzyme, the organism requires an external catalase source such as blood for culture.

There are over 90 serotypes of the pneumococcus that colonise the respiratory tract of mammals, and these are characterised by unique polysaccharides forming the capsule [80]. Serotyping pneumococcal isolates has traditionally depended on the Quellung reaction, whereby a suspension of the microbe is mixed with serotype-specific antisera and observed under a microscope; a positive reaction is indicated by swelling of the capsule [81–83]. Recently however, there has been increased use of high throughput methods such as microarray, multiplex PCR and sequetyping [82].

The nasopharynx forms the primary ecological niche for pneumococcus. The pathogen is well adapted to survive on respiratory tract surfaces. The capsular polysaccharides are negatively charged thus repelling the also negatively charged mucopolysaccharides on the surface of epithelium, preventing entrapment by mucus. The pneumococcal capsule inhibits attachment to endothelial cells during invasion; consequently, the pneumococcus has the ability to undergo phase variation. The capsule changes thickness, varying from opaque to transparent, the transparent forms have less capsule and more easily attach to host cells [84]. Moreover, the bacteria express adhesins such as phosphorylcholine that facilitate attachment of

host cell surfaces. Pneumococci express 10-15 choline binding proteins on their cell wall which also facilitate attachment to endothelial cells and modulate host immune responses during pathogenesis [85].

Asymptomatic colonisation by the pneumococcus is relatively widespread: estimated at 50% for 2-3 year old children in developed countries and as high as 60-90% for children in developing countries [86, 87]. In a minority of individuals however, the pathogen is able to cross mucosal barrier and cause disease in both immunocompetent and immunocompromised hosts. Pneumococcal infection has varied clinical manifestations ranging from mild otitis media and sinusitis to serious invasive disease. Invasive pneumococcal disease (IPD) is defined as the isolation of the pathogen from a normally sterile site, most commonly blood (diagnostic of bacteraemia) or CSF (diagnostic of meningitis) [88–90]. Clinical manifestations and the age at which disease occurs have been observed to correlate with certain pneumococcal serotypes. Serotypes 6, 10 and 23 are more commonly isolated in CSF while serotypes 1, 4 and 14 are isolated from bacteraemic patients [91]. Moreover, serotypes 1 and 3 dominate in children while 6B, 14 and 23 dominate in adults [91].

A major determinant of microbial virulence is the presence of a capsule. During colonisation of mucosal surfaces, the negatively charged capsule repels mucopolysaccharides on host cells preventing entrapment by mucus [85]. Furthermore, encapsulated serotypes have been shown to be resistant to complement-mediated phagocytosis and are the most frequently isolated serotypes in clinical isolates [85, 92]. In addition, the pathogen produces pneumolysin, a potent cytotoxin which inactivates immune cells through lysis of cell membranes [85]. Pneumolysin inhibits respiratory burst as well as cytokine production in immune cells [85, 93].In addition pneumolysin and neuraminidases A and B have been implicated as key determinants of invasion in mouse models of pneumococcal meningitis [94]. Another major virulence factor is resistance to immunoglobulin A1 (IgA1)-mediated killing [95]. IgA1 is the most abundant immunoglobulin on mucosal surfaces. Cleaving of IgA1 produces pneumococcal antigens bound to the Fab fragments and releasing the Fc fragment of the cleaved

antibody [85, 95]. Due to a lack of an Fc region, immune response to these antigens is not initiated [85].

Nasopharyngeal colonisation is a prerequisite for invasive disease, however, in addition to microbial virulence factors, host risk factors have also been identified that predispose to pneumococcal infections. Asplenia, whether congenital or arising from surgical splenectomy, remains a major risk factor pneumococcal infections despite widespread vaccination [96, 97]. Furthermore individuals less than 2 years of age or over 65 years are at an increased risk of infection [98]. Chronic comorbidities such as such as pulmonary, cardiovascular,liver, kidney, HIV and diabetes mellitus have all been identified as leading to an increased risk in pneumococcal disease. Moreover smoking, nutrition status and socio-economic background also lead to increased risk of IPD [98–101].

Mortality from IPD in developed countries ranges from 11-30% despite advances in medical treatments [102]. The mortality associated with invasive bacterial disease in developing countries is also high: bacteraemia is associated with a mortality rate of 28% among hospitalised Kenyan children, compared with 5.7% among other admissions [103]. Since the introduction of the *Haemophilus influenzae* type b vaccine in immunization programmes, the pneumococcus now accounts for the largest proportion of bacterial meningitis [37, 104]. In Europe, over half of all bacterial meningitis cases are attributable to pneumococcus, resulting in mortality rates as high as 30% and neurological sequelae in 30% of survivors [37, 104, 105].

Currently, there are two pneumococcal polysaccharide vaccines (PPSV14 and PPSV23) and three pneumococcal conjugate vaccines; 7-valent, 10-valent and 13-valent conjugate vaccines (PCV7, PCV10 and PCV13 respectively). The polysaccharide vaccines are poorly immunogenic in children younger than two years as they mount a T-cell independent driven response yet B cell compartments in infants are immature [106]. Polysaccharide vaccines are currently recommended for use in adults >65 years in the UK [106]. Following the introduction of PCV7 in routine immunisation more than a decade ago, there was a 94% decrease in IPD due to vaccine serotypes as well as 45% decrease in overall cases of IPD due to herd immunity [107]. However,

there was increased incidence of disease due to non-vaccine serotypes such as 19A, leading to licensure of PCV10 and PCV13 to include additional serotypes [107–109]. Recently, results have been published from a large scale impact study on the effect of PCV13 three years after introduction in routine immunisation in the United States [108]. PCV13 significantly reduced IPD cases in children less than 2 years although the case fatality rates remained unchanged in both children and adults [108]. Interestingly, the proportion of IPD cases presenting with bacteraemia significantly increased in children less than 18 years while a significant reduction was observed in adults over 18 years [108].

1.5 Pleural infection (empyema)

Parapneumonic effusions are characterised by fluid accumulation in the pleural space and are a common complication of pneumonia, occurring in up to 44% of adults with bacterial pneumonia [110]. The majority of effusions are uncomplicated and resolve quickly; however in a small proportion of individuals, the pleural space becomes infected resulting in complicated parapneumonic effusions or empyema. The incidence rate of empyema complicating pneumonia has increased fivefold in the last 20 years with the greatest burden in children [111–114].

Risk factors reported to associate with progression to empyema include ibuprofen use prior to admission and the invading pathogen species [111, 115]. Pneumonia from *Streptococcus pyogenes* and *Staphylococcus aureus* are more likely to develop pleural effusions than pneumococcal pneumonia; 83% of pneumonia cases due to *S. pyogenes* compared to 32% of pneumococcal pneumonia were reported to develop empyema [116]. However, due to the large number pneumococcal CAPs, *Streptococcus pneumoniae* remains a major cause of empyema despite widespread use of pneumococcal conjugate vaccines [111]. The rates of pneumococcal empyema in children less than 16 years of age remained unchanged before and after introduction of PCV7 with 75% of cases attributable to serotypes that were then included in PCV13 [112]. In the UK, the introduction of PCV13 into routine immunisation in 2010 has led to a

dramatic 51% reduction in pneumococcal empyema in children, notably in those aged less than two years [117]. Similar reduction in rates have been observed in the United States with a further 46% and 37% reduction in children aged 2–4 years and 5–17 years, respectively [117, 118].

The pathophysiology of empyema progresses in three stages. In the initial exudative stage, interstitial fluid seeps into the pleural space due to increased vascular permeability [119, 120]. In this stage, the fluid is sterile with normal glucose and pH levels as well as a low neutrophil count [120]. In most cases, this stage resolves with antibiotics without the need for drainage [119]. Without treatment however, the effusion progresses into the second stage, referred to as the fibrinopurulent stage. The pathogen invades the pleural space resulting in an infiltration of neutrophils to fight the infection. The accumulation of neutrophils and dead bacteria leads to accumulation of pus in the pleural space and the exudate becomes suppurative [119]. Moreover phagocytic killing and release of bacterial products results in further inflammation [121] and there is deposition of fibrin resulting in formation of fluid pockets (loculation) [120]. During this process, the pleural fluid glucose concentration drops followed by an accumulation of lactate dehydrogenase and a decrease of pH [122]. The third stage is the organisational stage with a marked increase in fibroblast proliferation and subsequent formation of an inelastic 'collagenous peel' along pleural membranes [121]. The pleural peel restricts lung re-expansion impairing lung function and promoting further fluid accumulation and infection due to decreased pressure [120]. Like most complex diseases, progression from uncomplicated pneumonia to empyema is likely a reflection of complex interactions between host susceptibility, pathogen and environmental factors. The precise mechanism of progression from pneumonia to empyema remains poorly understood although a recently developed murine model holds promise [123].

1.6 A genetic basis for human disease

A heritable basis for human diseases has long been known, with the maternal transmission of haemophilia described in Jewish writings from the 2$^{\text{nd}}$century [124].

The foundations of genetics were laid by Gregor Mendel while working on peas in 1866. He noted the presence of fundamental factors that were passed on unchanged from parent to offspring [125]. His work was however forgotten until the early 1900s [126]. The re-discovery of Mendel's work allowed Archibald Garrod to describe the recessive inheritance pattern of alkaptonuria after noting high incidence rates among offspring from consanguineous unions [127]. The hereditary material identified by Mendel was later identified as deoxyribonucleic acid (DNA) organised into genes that determine the protein that is then expressed in cells. The history of sickle cell disease provides a classic example of the identification of a genetic alteration leading to a human disease. Sickle cell disease was first described by James B. Herrick in 1910 who observed peculiar shaped red cells in a young dental student from Grenada [128]. The patient suffered recurrent respiratory infections and finally died from pneumonia [129]. It wasn't until 1949 that Linus Pauling discovered the molecular basis for sickling through observations of the difference in physical and chemical properties of haemoglobin from normal and sickled red cells [130]. This was followed shortly after by the identification of the amino acid residue change from a glutamic acid residue to valine in the haemoglobin polypeptide chain [131].

Genetic traits can be classified as either monogenic or complex. Monogenic traits follow the classic pattern of inheritance described by Mendel where a single gene is responsible for manifestation of the trait (phenotype). Complex traits on the other hand are dependent on interactions between genetics and the environment for the phenotype to manifest. In diseases with a monogenic pattern of inheritance (Mendelian diseases), a disruption in the gene will certainly lead to manifestation of the disease. However, in complex diseases, genetic determinants of disease vary in their magnitude of effect; this is termed penetrance. Genetic variants with a high penetrance will result in a large proportion of individuals who carry the genetic variant manifesting the disease.

Historically, mapping of genes associating with disease employed either linkage analysis in pedigrees or candidate gene analysis in population based studies [132, 133]. Linkage analysis aimed to identify loci that co-segregated with disease more often than would be expected by chance in pedigrees [132, 133]. This approach has been highly

successful in mapping loci in Mendelian traits [132]. However, linkage analysis is underpowered for use in the study of complex traits such as diabetes and cardiovascular diseases [132,134]. In contrast to Mendelian diseases where genetic factors are necessary and sufficient to cause disease, the genetic component in complex diseases is never fully penetrant and a large number of pedigrees would be required for detection.

In 1996, Risch and Merikangas proposed that for complex traits, where loci associating with the trait will more often than not tend to have modest effect sizes, association studies were more powerful than linkage studies [134]. Candidate gene case-control studies as well as parent offspring trios were initially widely used in the analysis of complex trait with the candidate genes selected based on biological plausibility. In trio studies the Transmission Disequilibrium test (TDT) assessed alleles passed on to affected offspring and compared to what was expected under Mendelian segregation [135]. In case-control studies, the differential frequency of a genetic variant was then determined between unrelated cases and controls. However a great majority of these studies failed to replicate in independent studies. Indeed from a review of 166 candidate gene studies, where each association had at least three publications, only six of these replicated in independent studies [136]. The most common reasons for the poor reproducibility were a lack of statistical power due to small sample sizes, lack of statistical correction for multiple comparisons and poor ethnicity matching in the study samples as the frequency of alleles show marked differences based on ancestry [136, 137]. However, there were notable successes from robust candidate gene studies for example the association between Factor V Leiden and deep vein thrombosis, *APOE* and Alzheimer's disease and a 32 base pair deletion in the chemokine receptor *CCR5* and HIV-1 infection [136].

1.7 Current methods in human genetics of complex traits

In 1990, the Human Genome Project was launched and the first draft sequences of the human genome were published a decade later in 2001 [138]. This facilitated large

scale studies into understanding human genetic variation, notably the subsequent establishment of the International HapMap Project in 2003 [139]. This was a collaboration aimed at providing a dense characterisation of human genetic diversity at single nucleotides in different populations of African, Asian and European ancestry [139]. This project initially aimed to genotype one common single nucleotide polymorphism (SNP) every five kilobases, whereby a common SNP was defined as having a minor allele frequency (MAF) greater than 5% [139, 140]. A characteristic feature of genetic variation is the non-random association of SNPs whereby alleles at different loci tend to be inherited together, referred to as linkage disequilibrium (LD). In addition to providing maps of common SNPs, the project also identified the underlying genome-wide correlation of SNPs with their neighbours into haplotype blocks, providing a genome-wide map of LD structure [140]. This remarkable resource, coupled to great technological advances accelerated rapid progress to large scale genome wide association studies (GWAS). In GWAS, SNPs are selected to 'tag' other SNPS in high LD. Rapid development of array genotyping enabled high throughput genotyping of hundreds of thousands of SNPs, allowing unbiased genome-wide association analysis. The Wellcome Trust Case Control Consortium reported a landmark study in 2007, providing the earliest demonstration on the utility of GWAS in complex trait genetics [141]. Moreover, methods that enable accurate estimation of genotypes that have not been directly genotyped have been developed [142].

GWAS are motivated by the "common disease- common variant" hypothesis which proposes that a significant proportion of complex diseases such as diabetes, asthma and cardiovascular diseases are due to common variants that modify disease risk [143, 144]. GWAS have greatly advanced our understanding of the role of common polymorphisms in common multifactorial diseases. Novel loci have been implicated in a myriad of traits such as diabetes, lipid levels, schizophrenia, cancer and cardiovascular diseases [145]. Though the underlying biology or mechanism of action for most GWAS loci remains unclear, notable successes include uncovering the role of the complement pathway in age related macular degeneration as well as IL23 signalling in Crohn's disease and psoriasis [146].

However, it is now clear that common variants have only a modest effect on common traits -with effect sizes typically less than two- and only account for a small proportion of the heritability observed from familial aggregation [145–147]. As noted in studies of loci underlying variability in height, statistically significant GWAS loci account for only 5% of the estimated 80% height heritability, while 71 loci implicated in Crohn's disease account for only 23.2% of heritability [148, 149]. The source of the remaining "missing heritability" of common disease has been subject to much debate [147, 150, 151].

1.8 Rare variant association analysis

It has been widely proposed that rare genetic variants could account for this "missing heritability" [147, 150]. The "common disease-rare variant" hypothesis proposes that for a common disease, the genetic risk is attributable to multiple variants of high penetrance that are individually rare but collectively account for a significant proportion of the disease [143, 144, 152]. However, a strict consensus on the definition of "rare" in this context has not been agreed. While it is generally accepted that variants <1% are rare, various studies have used either <5%, <1% or <0.5% to delimit rare from common [144, 152, 153]. Generally, variants <1% are considered "rare" while variants of 1-5% are considered "low frequency" [154].

Evolutionary genetics teaches that a variant leading to a fitness cost will tend to be maintained at a low frequency or eliminated from a population, although exceptions occur when the variant is under balancing selection due to conferring a selective advantage. A key example is the rs334 mutation in the *HBB* locus, where mutant homozygotes suffer from sickle cell anaemia while heterozygote carriers are protected from *Plasmodium falciparum* malaria. This balancing selection has led to the *HBB* allele occurring at a high frequency in malaria endemic regions despite severe morbidity and mortality resulting from the sickle cell disease [155, 156]. Such examples however are rare in the human genome and are unlikely to account for a significant proportion of common diseases. Further support for a significant role for rare variants in complex diseases arises from observations that a majority of deleterious variants will tend to have

an allele frequency of less than 1% [152, 154, 157, 158]. Moreover, early candidate gene resequencing studies provided empirical evidence for rare variants in complex diseases. Notable examples from these early studies include multiple rare variants in *IFIH1* associating with Type 1 diabetes [159], rare and low frequency variants in multiple genes associating with lipoprotein levels and risk of cardiovascular diseases [160–162], and rare variants associating with obesity [163]. In addition, re-sequencing of GWAS loci identified functional rare variants associating with inflammatory bowel disease that were independent of the GWAS signal [164].

Until recently, genotyping arrays have been designed to capture common polymorphisms (MAF >5%). However, there is a poor correlation (LD) between common and rare variants and therefore inference of rare variants from common polymorphisms with a high degree of certainty is challenging [165]. Sequencing provides the ideal way to discover rare variants, ideally through sequencing the whole genome (WGS) as this would allow detection of the full spectrum of mutations, whether single nucleotide or large structural changes as well as variation in both coding and non-coding regions of the genome. However, although the cost of WGS has dramatically reduced, it is still prohibitive for large sample sizes. Sequencing of protein coding regions of the genome (the exome) provides a suitable cost effective alternative [166]. Details of the exome sequencing procedure are in appendix chapter A. The human exome consists of approximately 20,000 protein-coding genes with over 180,000 exons [167, 168]. The consensus coding sequence project (CCDS) which curates consistent protein annotations between the University of California at Santa Cruz (UCSC) genome browser [169], the National Centre for Biotechnological Information (NCBI) RefSeq database [170] and Ensembl [171], lists a conservative estimate of 18,826 protein coding genes as of November 2015. The RefSeq database lists 26,546 genes including non-coding genes while Ensembl lists 20,296 protein coding genes. Despite accounting for only 1% of the human genome, the exome harbours 85% of all genetic variants that lead to Mendelian disorders [172, 173]. Whole exome sequencing (WES) has proved a robust tool in the identification of causal variants underlying Mendelian

diseases [173–176]. In complex traits, exome sequencing has been used to uncover rare variants in population based studies [154, 177–188]

1.9 Statistical methods in rare variant analysis

Association analysis methods in GWAS are robust and well developed. However, as the allele frequency is a major determinant of power in such study designs, such methods would be underpowered for rare variant analysis in even modest sample sizes [150, 165, 185, 189–191]. For example for an allele with 2% frequency and odds ratio of three, study power is 70% at GWAS significance threshold of 5.00×10^{-8} in 1,000 case-control pairs for a disease with 1% prevalence [191]. For a variant occurring in 1 in 2,000 individuals, a 40-fold increase in risk would be needed to obtain the same level of power [191]. Exceptions might be the so-called "Goldilocks' alleles" which are low frequency alleles of 1-5% [151, 192–194]. These alleles might have modest effects on a trait and are common enough to detect through single variant analysis but are excluded from GWAS due to <5% allele frequency threshold in most GWAS.

Currently, methods of association analysis in rare variants aggregate multiple rare variants in a functional unit, primarily a gene [189, 195–199]. This method has been shown to be more powerful than single variant association analysis and reduces the multiple testing burden [165]. Earlier methods collapse variants in a gene to an indicator variable to represent presence or absence of a rare variant [165]. The Combined Multivariate and Collapsing (CMC) test divides rare variants into subgroups based on allele frequency and assesses the joint effect of both common and rare variants [165]. The Cohort Allelic Sums Test (CAST) collapses the variants by summing up the rare alleles in a gene [200] while in the Weighted Sums Test (WST) each variant is assigned a weight as a function of frequency [201]. The underlying assumption in these tests is that all the aggregated variants are causal and have the same direction of effect on the phenotype [202]. However, empirical data has demonstrated that some genes can harbour variants with different direction of effect. A good example is in the *PCSK9* gene where missense mutations are associated

with severe hypercholesterolemia while loss of function mutations associated with low levels of LDL cholesterol [160–162, 203–206]. In such a scenario, collapsing tests suffer from weakened signal if variants of opposite effect or non-causal variants are collapsed [207]. Subsequently, tests have been developed that are robust to inclusion of both non-causal variants and variants with different direction of effect [202, 208]. The C-alpha test compares distribution of counts for each variant to the expected under a normal distribution and then detects a pattern across all variants in the region being tested, however this test does not allow inclusion of covariates or analysis in quantitative traits [202]. The Sequence Kernel Association Test (SKAT) aggregates the score statistic for each individual variant across the region being tested; each score is assigned a weight based on direction of effect. This test is similar to C-alpha but allows inclusion of covariates [208–210]. Moreover, this test has been shown to be most robust to inclusion of non-causal variants [189].

1.10 Variant prioritisation in rare variant analysis

A major challenge in rare variant analysis is the identification of causal variants amongst the multitude of mostly neutral variation identified through sequencing. Aggregation based methods for variant analyses are known to suffer from significant loss of power due to inclusion of non-causal variants. Since a majority of deleterious variants will tend to be rare, allele frequency thresholds are implemented to exclude common variants from aggregation [152, 154, 157, 158]. Due to a redundancy in the genetic code, not all base pair changes lead to a change in amino acid encoded. Missense mutations leading to an amino acid change are intrinsically more likely to impact protein function than mutations in which the same amino acid is retained (synonymous). Therefore, synonymous mutations are generally excluded in aggregation analysis. Furthermore, not all missense mutations will have a significant impact on protein functionality; bioinformatics tools have been developed that incorporate both computational and empirical methods to predict the impact of missense mutations on protein function. Two widely used prediction tools are Polymorphism Phenotyping-2 (PolyPhen-2) [211]

and Sorting Intolerant From Tolerant (SIFT) [212]. Both tools use a phylogenetic approach, comparing evolutionary conservation across multiple homologous proteins to predict whether the substitution falls within the spectrum of tolerated changes. Mutations arising in regions of the genome that display marked conservation across species are most likely to have an impact on protein function [211–213]. In addition, PolyPhen-2 relies on sequence based characteristics of the substitutions such as whether it occurs in a biologically active domain or whether the substitution leads to changes in the 3-dimensional structure of the protein [211]. Incorporation of prediction information while prioritising variants for analysis has proven useful in improving power for rare variant analysis [188–190, 195].

However, due to a tendency for functional prediction tools to overcall pathogenic variants, they lack in sensitivity and specificity necessary for clinical genomic studies [214]. For instance, in a recent study of 33 missense mutations in 23 immune related genes, only 20% of variants predicted as deleterious by PolyPhen-2 had the expected phenotypes in homozygous mice [215]. However all predicted benign variants were functionally neutral. Moreover an in-depth analysis of all possible missense mutations in the *TP53* gene revealed 42% of predicted deleterious variants to be of little or no functional impact but only 4% of those predicted as benign were false negatives [215]. These results highlight the high number of false positive pathogenic variants but a low number of false negatives. However, these tools still prove useful in variant prioritisation for association analysis though further validation of causality would be necessary [153, 215–217].

1.11 Rare host genetic variants in infectious diseases

Humans display a striking variability in susceptibility and outcome to infectious diseases. Early examples of a genetic influence on infectious diseases are provided by haemoglobinopathies. A possible role for genetics influencing disease susceptibility was first suggested by J.B.S Haldane in 1949, when he hypothesised that the high

incidence of thalassemia in Greeks and Italians was due to a heterozygous advantage providing protection against malaria which also happened to be prevalent in these regions [218]. This was subsequently followed by Allison's observations in 1954 that carriers of the sickle trait were also protected against malaria (described in section 1.8) [155, 156]. A significant host genetic component to infection was further implied in a landmark study on Danish adoptees from the 1980s which reported a 5.8-fold increase in the risk of death from infection in adoptees if a biological parent died prematurely from infection [219].

Infectious diseases have been a major cause of premature death throughout human history and hence exert enormous selection pressure on the human genome. It would therefore follow that genetic variants leading to compromised resistance to infectious agents would be rare or even private in the population [220–222]. The converse is also true: variants conferring resistance to an infection will tend to be maintained at a higher frequency in the population where the infectious risk exists. This is exemplified by the aforementioned sickle cell trait and malaria, where the frequency of the rs334 variant ranges from 0 to 40% varying with malaria endemicity [155, 156, 222].

Great insight into the nature of rare variants in relation to infection has been gained from the field of primary immunodeficiencies. Traditionally, primary immunodeficiencies have been defined as inborn errors of immunity resulting from a fully penetrant disruption in a single gene, characterised by a predisposition to a wide spectrum of recurrent infections usually in childhood. While this definition holds true for classic syndromes such as X-linked agammaglobulinemia [223], congenital neutropenia [224] and combined severe combined immunodeficiency (SCID) [225], it is now increasingly recognised that primary immunodeficiencies can also result from genetic variants of incomplete penetrance that predispose to a narrow range of pathogens in otherwise healthy individuals [226, 227]. Empirical evidence includes rare mutations in the innate immune Toll-like receptor (TLR)-Nuclear Factor-κB (NF- κB) signalling pathway that have been described in association with increased susceptibility to pyogenic bacterial infections, in particular IPD [226, 228–232]. Moreover, rare mutations in the IL-12/Interferon-gamma (IFN-γ) pathway have been implicated

in increased predisposition to infection with intracellular pathogens such as non-tuberculous mycobacteria and *Salmonella* [226, 233–236]. Furthermore, rare mutations leading to deficiencies in early and late components of the complement system have been implicated in increased susceptibility to *Neisseria meningitidis* infections [237–242].

These observations have led to a change in the definition of primary immunodeficiency and an increasing recognition that such selective immunodeficiency states might be more common than previously supposed. Recently identified immunodeficiencies are characterised by a predisposition to a narrow range of pathogens and varied age of onset. Picard et al. identified three children with IRAK-4 deficiency and recurrent infections with *Streptococcus pneumoniae* and *Staphylococcus aureus* whose frequency reduced with increasing age [230]. Reduction of episodes with age has been attributed to maturation of the adaptive immunity that complements the selectively impaired innate immunity in some of these individuals [227]. In contrast, some immunodeficiencies such as common variable immunodeficiency (CVID) typically present in adulthood [227]. Moreover, newly defined primary immunodeficiencies display a degree of locus heterogeneity with mutations in different genes conferring the same phenotype. Mutations in *IRAK-4* and *NEMO* have been demonstrated to lead to similar phenotypes, for example recurrent IPD [228]. Additionally, allelic heterogeneity- where different mutations in the same gene result in the same phenotype- is recognised. Different mutations in complement genes leading to a complement deficiency and increased predisposition to meningococcal disease have been reported [238].

Due to increasing identification of primary immunodeficiencies in otherwise healthy individuals, together with notable allelic and locus heterogeneity observed with selective immunodeficiency, a significant proportion of infectious diseases in individuals with no other identifiable risk factors could be due to rare genetic variants of variable penetrance. This was implied in a population based study of meningococcal disease susceptibility in which resequencing the *TLR4* gene in meningococcal cases and controls identified an excess of rare heterozygous mutations in individuals with meningococcal sepsis [243]. In another example, exome sequencing identified rare variants in *DCTN4* gene that were significantly associated with early *Pseudomonas aeruginosa* airway infection in

adults with cystic fibrosis [180]. There have been very few reported population-based studies to date on rare genetic variants in infectious disease phenotypes.

1.12 Overview of this thesis

This thesis outlines efforts made in characterising the role of moderate to large effect genetic variants in susceptibility to severe bacterial disease using next generation sequencing technology in population based cohorts. The challenges of using shared controls in the analysis of rare variants are outlined, as well as progress and consensus on rare variant association analysis in complex diseases. This thesis demonstrates the potential for next generation sequencing to identify novel selective primary immunodeficiencies as well as highlighting novel loci and mechanisms predisposing to invasive bacterial diseases. Chapter 3 describes the analysis of rare large effect genetic variants in susceptibility to invasive meningococcal disease. Described is a systematic method of analysis for rare variants in an extreme phenotype design, outlining the variant prioritisation strategies that uncover a novel causal variant. Also described in this chapter are results from the subsequent characterisation of the molecular consequences of the variant on protein function. Chapter 4 describes efforts made to identify genetic variants underlying sepsis arising from community acquired pneumonia. This chapter highlights the challenge of appropriate controls in sequencing studies as well as analytical issues regarding stratification of rare variants in closely related populations. Furthermore, results implicating a novel pathway in sepsis pathology are presented. In chapter 5, the role of rare genetic variants underlying childhood empyema is investigated. This chapter describes a systematic analysis of rare and low frequency mutations, highlighting the role of "Goldilocks's alleles" of 1-5% allele frequency in infectious diseases susceptibility. Moreover, this chapter identifies a genetic basis for predisposition in pathways previously implicated through immunological studies in the pathogenesis of empyema. Chapter 6 describes efforts made to characterise large effect genetic variants from a population based study of invasive pneumococcal disease. Results from this analysis highlight the role of the innate immune system as a first line of

defence against invasive disease and the implication of novel mechanisms of protection that are impaired during pneumococcal invasion. Finally chapter 7 summarises and discusses these findings as well as the contributions they make in furthering our understanding of genetic predisposition to infectious diseases. Recommendations for further investigation of the novel genes and pathways implicated in the phenotypes studied are also presented, with the hope that this work will lead to further insight into the mechanisms leading to invasive bacterial disease.

Chapter 2
Materials and Methods

2.1 Introduction

This chapter describes the sample collections studied, the bioinformatics pipeline and tools used in the processing of next generation sequencing data, the statistical analyses, and finally the molecular biological techniques used to functionally validate genetic findings.

2.2 Disease cohorts

2.2.1 Meningococcal disease

Individuals in this collection were recruited as part of the Genetics of Sepsis in Europe (GenOSept) project [25, 244]. The overall aim of the GenOSept project was to define host genetic predisposition to morbidity and mortality due to sepsis. Diagnosis of sepsis was based on the International Consensus Criteria published in 2003 (appendix chapter C) [6]. Patients admitted to Intensive Care Unit (ICU) with sepsis due to community acquired pneumonia (CAP), faecal peritonitis, pancreatitis or meningococcal disease were recruited from 143 centres across 16 countries in Europe between September 2005 and October 2009 [25, 244]. For this study, only critically ill adult individuals admitted to an ICU in the UK with microbiologically proven meningococcal disease were used. The clinical definition of meningococcal disease was meningitis and/or meningococcemia and/or meningococcal sepsis proven to be due to or suspected to be due to *Neisseria meningitidis*. The causative pathogen was determined either through

CSF gram stain, CSF culture or CSF PCR. There were 15 individuals who fitted these criteria. The mean age was 43 years while 53% were male. One individual was excluded from further study due to ethnicity and a second individual was excluded due to a previously identified complement component 7 deficiency. Written informed consent was obtained from all patients or a legal representative, and ethical approval for the study was provided by the Multi-Centre Research Ethics Committee for Scotland.

2.2.2 Community acquired pneumonia

Individuals were recruited as part of Genomic Advances in Sepsis (GAinS) study, the UK collaborative arm of the GenOSept project. GAinS began in 2005 as part of GenOSept and continued to actively recruit patients with CAP and faecal peritonitis after GenOSept had completed using the same protocol. For this study, 270 individuals with sepsis due to CAP from this sample collection were included. The average age for these individuals was 42 years. Pneumococcal infections accounted for 22% of cases while *Haemophilus influenzae*, *Staphylococcus aureus* and viral infections accounted for 5% each; 48% of cases did not any identified microbiological cause. At the time of writing, this sample collection was still ongoing.

2.2.3 Oxfordshire invasive pneumococcal disease

The invasive pneumococcal disease (IPD) sample collection was established as part of an enhanced surveillance of IPD in three Oxfordshire hospitals (John Radcliffe, Horton General and Wycombe General) between 1995 and 1999, coordinated by Professor Derrick Crook at the John Radcliffe Hospital Department of Microbiology. IPD was defined as isolation of *Streptococcus pneumoniae* from a normally sterile site, most commonly blood (bacteraemia) or cerebrospinal fluid (meningitis). Samples were collected on diagnosis from all hospitalised adult patients with microbiologically proven IPD. Clinical details, including age, gender, diagnosis and co-morbidities, were recorded. The mean age was 58 years and 50% were male. Frequencies of clinical presentation were as follows: pneumonia 69%, isolated bacteraemia 15%, meningitis 11% and others 5%. Pneumococcal serotypes were identified and the distribution of serotypes was very

similar to that of previous UK studies, with serotype 14 being the commonest [90]. From this collection, 105 samples passed quality control and were used in this study. Samples in this study are existing holdings of DNA collected between 1995 and 1999 as per the regulations prior to the Human Tissue Act (2004) [245]. All samples were obtained from excess blood taken for routine diagnostics from living patients. Consent for high throughput genetic analysis was not required at the time of collection. In view of this, an ethics application for genome-wide high throughput genetic analysis was submitted to the Oxford Research Ethics Committee (OXREC) and was subsequently granted under reference MSD-IDREC-C1-2014-112 (appendix chapter D).

2.2.4 Pleural infection (empyema)

The empyema sample collection consists of over 600 samples collected from children 0-16 years with pleural infection from multiple sites across the UK recruited as part of the National Programme for Enhanced Pneumococcal Surveillance. All individuals recruited had a sample of pleural fluid removed as part of their routine management. Empyema was diagnosed as pyothorax with fistula or pyothorax without fistula as per the International Classification of Disease 10 codes J86.0 and J86.9. Saliva collection was favoured over blood collection because it is a less invasive method of sampling in very young children. A total of 101 samples were selected for sequencing based on European ethnicity, quality of DNA, and availability of microbiological diagnosis. Furthermore, individuals selected had no relevant co-morbidities or established primary immunodeficiencies. The mean age was 4.2 years and 64% were male. The causative pathogen was identified as *Streptococcus pneumoniae* (68%), *Streptococcus pyogenes* (Group A) (21%), *Staphylococcus aureus* (5%) and other *Streptococci* species (5%).

2.3 Control cohorts

The following population-based controls were used:

2.3.1 Avon Longitudinal Survey of Parents and Children (ALSPAC)

The ALSPAC study is a longitudinal birth cohort of over 14,000 women recruited while pregnant and due to give birth between April 1991 and December 1992 from the Bristol area and its surroundings [246]. This project is part of the 'Children of the 90s' long term health research project and the participants have been followed up for over two decades [247, 248]. Whole genome sequencing was carried out on 1,927 individuals as part of the UK10K project, a large scale genetics project whose aim was to sequence 10,000 individuals from the UK in order to characterise the role of rare and low frequency genetic variation in complex human diseases (http://www.uk10k.org). The median coverage for this dataset was 6X whole genome coverage.

2.3.2 Scottish Family Health Study obesity controls (UK10K-OBESITY-GS)

The Generation Scotland: Scottish Family Health Study (UK10K-GS: SFHS) is a family-based cohort study for the genetic basis of complex diseases and responses to treatment [249]. It involves over 24,000 volunteers from across Scotland recruited between 2006 and 2011. From this cohort, whole exome sequence data for 369 unrelated individuals was available. These individuals were selected as part of a study on the genetics of obesity and all had a BMI >40.

2.3.3 The 1958 Birth Cohort (58BC)

Individuals in this sample collection were recruited as part of the National Child Developmental Study and include all individuals from England, Scotland and Wales born during one week in 1958 [250]. Samples have been genome-wide genotyped and used successfully as shared controls in GWAS [141]. Whole exome sequence data from 1,000 individuals in the 58BC was generated by Professor Nazneen Rahman's team in the Division of Genetics and Epidemiology at The Institute of Cancer Research, London and made freely available to the scientific community.

2.4 Laboratory methods

2.4.1 DNA extraction and quantification

2.4.1.1 Extraction from blood

Blood was collected in EDTA tubes, transported at room temperature and then stored at -80°C until extraction. DNA was then extracted using the QIAamp DNA Blood Midi Kit (Qiagen) as per manufacturer's instructions. This method was used for extraction of meningococcal disease and CAP samples.

2.4.1.2 Extraction from saliva

Saliva was collected in Oragene DNA OG-250 kits (DNA Genotek) and transported at room temperature. Samples were all stored at room temperature until extraction. DNA was extracted by Dr Tara Mills using standard procedures from Oragene. All DNA samples were quantified using Quant-iT^{TM} PicoGreen® dsDNA Assay Kit (Invitrogen) as per manufacturer's instruction. The PicoGreen reagent fluoresces when bound to double strand DNA. Fluorescence was measured using the TECAN XFluor^{TM} software and DNA concentration estimated from a standard curve.

2.4.2 Quality control

For all samples, quality of DNA was assessed by gel electrophoresis using 1% agarose gel (Sigma-Aldrich, UK) dissolved in 1X Tris-Acetate EDTA (TAE) buffer (Sigma-Aldrich, UK). Ethidium bromide was used in visualisation of DNA migration on gel. Highly denatured genomic DNA appears as a smear on agarose gel upon electrophoresis. Samples showing significant smearing on electrophoresis were either re-extracted or excluded from sequencing.

2.4.3 Whole genome amplification

Whole genome amplification was carried out using the Genomiphi V2 DNA Amplification Kit (GE Healthcare). The assay is based on whole genome amplification by isothermal replacement. Hexamer primers bind randomly to DNA and the strand is

isothermally synthesised using Phi29 polymerase. This can yield 4-7µg of whole genome amplified DNA from 10ng starting DNA in a 20µl reaction over a 1.5 hour period.

2.4.4 Exome capture and sequencing

2.4.4.1 Sample preparation

2.4.4.1.1 Meningococcal disease and CAP For these two disease cohorts, samples were diluted to 50ng/µl of DNA using the appropriate volume of low Tris-EDTA. For each sample, 60µl (3µg DNA) was then submitted for sequencing at the Oxford Genomics Centre, Wellcome Trust Centre for Human Genetics.

2.4.4.1.2 Empyema and IPD Available samples from these two cohorts were found to have varied amounts of DNA: samples with greater than 500ng of genomic DNA were selected for sequencing in this project.

2.4.4.2 Exome enrichment and sequencing

For meningococcal disease samples, targeted capture of exonic regions was done using the TruSeq^{TM} Exome Enrichment Kit (Illumina). For the CAP, empyema and IPD samples, targeted capture were carried out using SeqCap EZ Human Exome Library v3.0. All samples were sequenced on the HiSeq 2000 platform (Illumina).

2.5 Bioinformatics pipeline for the analysis of sequence data

2.5.1 Mapping

The Stampy algorithm was used for mapping [251]. Stampy is a hash-based algorithm that creates a hash table of selected 15-mers of the reference genome. For each 15-mer overlapping region in the read, the algorithm searches the hash table to produce possible mapping locations which are subsequently filtered and remapped based on sequence similarity. Stampy was used in hybrid mode in which Burrows-Wheeler Aligner (BWA) - a fast compression-based algorithm - mapped the reads first and then Stampy remapped poorly mapped reads [252]. This enabled a combination of speed from BWA

with increased sensitivity from Stampy leading to superior mapping compared to other alignment algorithms [251]. The reference genome GRCh37, specifically the hs37d5, was used for alignment and all subsequent processing. In addition to the GRCh37 sequences, the hs37d5 genome also contains "decoy" sequences for reads that do not map to the human genome. Such reads are most often derived from viruses such the Epstein Barr virus whose DNA sequences are often captured together with human DNA.

2.5.2 Post mapping processing

The analysis pipeline used was modelled on the Genome Analysis ToolKit (GATK) best practice guidelines [253] described below.

2.5.3 Marking duplicates

Duplicate reads can arise due to library prep procedures and as PCR artefact. Such reads have the same start and end positions as well as having an identical sequence. They can lead to erroneous estimates of depth and compounding of errors. For all sequence data, duplicates were marked rather than removed using the MarkDuplicates algorithm implemented in Picard tools. Subsequently, reads flagged as duplicates are not considered during downstream processing and variant calling.

2.5.4 Local realignment around indels

Short insertions and deletions (indels) can usually lead to mismatches in surrounding bases due to incorrect mapping. These mapping artefacts are easily mistaken for polymorphisms and have implications for downstream quality score recalibration. Even though Stampy is more sensitive to mapping around indels than other alignment algorithms [251], the program considers each read independently during the mapping process. In order to provide consistent mapping across all reads spanning the indels, the Genome Analysis ToolKit (GATK) algorithms 'RealignerTargetCreator' and 'IndelRealigner' were used for local realignment. The first tool identifies regions for realignment, followed by local de novo re-assembly of reads that do not perfectly match the reference sequence in that region [254].

2.5.5 Recalibrate base quality score

The GATK algorithms 'BaseRecalibrator' and 'PrintReads' were used to calculate a well-calibrated and more accurate quality score for each base [253, 254]. Each base in a read has an associated Phred-scaled quality score -defined as the probability that the base was called incorrectly- assigned by the base-calling algorithm in the sequencing machine; however variation in this assigned quality score can arise with machine cycle and the sequence context of each base. Base quality recalibration makes the quality score more accurate by taking into account other covariates not considered by the base-calling algorithm that may affect the quality of a base. The base quality recalibration assigns a more accurate empirically recalculated base score that is derived from analysing not only all the reads from the same sequencing run (read groups) but also covariates associated with each base such as its positon on the read and surrounding bases. The process involved grouping reads based on covariates such as read group, quality score, base context and cycle. By providing a list of validated polymorphisms such as those from 1000genomes project, the recalibrator then assumes that all other mismatches not in the provided list are errors. This information together with covariate groups is then used to calculate an error rate used to correct each score resulting in a well-calibrated empirical base score crucial for downstream variant calling [254, 255].

2.5.6 Variant calling

Two variant callers from GATK were used:

2.5.6.1 UnifiedGenotyper

For the initial variant calling of meningococcal samples, UnifiedGenotyper was used [253, 256]. This caller takes a reference file in FASTA format and indexed BAM files to output variant calls as Variant Call Format (VCF) files. UnifiedGenotyper walks along each base on the reference genome and analyses all reads in that region. By using a genotype likelihood model, the algorithm uses a prior probability, P(G),

of the genotype and uses the read data to calculate a posterior probability for the genotype given the data, P(G|D).

$$P(G|D) = P(G)P(D|G)/(\sum_{a} xP(G_i)P(D|G_i)) \tag{2.1}$$

The conditional probability of the data given the genotype $P(D|G)$ is derived from

$$P(D|G) = \prod j(P(D_j H_1)/2 + P(D_j H_2)/2) \tag{2.2}$$

The diploid genotype G is given by H_1H_2. The value of $P(D_jH_n)$ which is the conditional probability of the data given the allele across all reads, is determined by a pairwise hidden Markov model. The genotype is then given as the one with the highest posterior probability. UnifiedGenotyper is an aggressive caller: the output file contains many false positives and stringent downstream filtration is recommended [253, 256].

2.5.6.2 HaplotypeCaller

During the course of my research project, HaplotypeCaller replaced UnifiedGenotyper as the recommended tool for variant calling due to superior sensitivity in calling both SNPs and short indels [253, 255, 257]. As opposed to UnifiedGenotyper which analyses each base at a time, HaplotypeCaller identifies an active region with a suspected variant. The tool performs local de novo assembly of reads in the active region to resolve misaligned reads. Subsequently, HaplotypeCaller is much more computationally intensive and much slower than UnifiedGenotyper for calling greater than 100 samples [253, 255, 257]. For CAP, IPD, empyema and reanalysis of meningococcal disease with generation Scotland controls, HaplotypeCaller was used for variant calling.

2.5.7 Quality controls of variants

2.5.7.1 Variant quality score recalibration (VQSR)

For this analysis, the GATK tool 'VariantRecalibrator' was used. The tool uses variant associated metrics such as quality by depth (QD), variant position on the read (ReadPosRankSum) and mapping quality (MQ) metrics from a set of true variants - either the International HapMap project [139, 140] or 1000 Genomes [258] -

to generate an adaptive error model. The model is then applied to variants in the analysis dataset to derive a probability of each variant being real. Each variant is then annotated with a VQSLOD score - a log odds ratio reflecting the probability of the variant being true versus the probability of the variant being false under the adaptive error model generated. The filter flag for each variant is then adjusted accordingly, with high VQLSOD variants annotated as 'PASS' while low VQSLOD variants have a 'LOW_VQSLOD' flag in the filter field. This annotation can then be used for downstream filtration where only variants with 'PASS' filter are used for analysis.

2.5.7.2 Hard filters

It is recommended that VQSR is run on more than 30 exome samples [253, 259]. For this reason, hard filters were used for initial analysis of 13 meningococcal exomes. The following filters were applied to each variant using the GATK tool 'VariantFiltration':

1. Quality Depth (QD >4): since variant confidence increases with depth, only variants with greater than four reads were retained.

2. Fisher's exact p-value for strand bias (FS >60): if a variant showed bias for either forward or reverse reads, it was excluded from analysis.

3. Mapping Quality (MQ >40): variants falling in low quality reads across all samples were excluded using this filter.

4. Length of Homopolymer Run (HRUN >5): variants falling in long homopolymer runs were excluded by this filter. Such variants are most likely artefacts, arising from the base-calling algorithm or from mismapping.

Each filter field was then annotated as 'PASS' if the variant passed all thresholds or 'HARD_TO_VALIDATE' if the variant failed any of the filters.

2.5.7.3 Genotype level filters

The metrics described above do not filter on individual genotypes. The genotype quality metric (GQ) is a Phred scaled score representing the confidence in the called genotype, with a higher score indicating that the called genotype is likely the true genotype at that site [260]. A genotype quality threshold of 20 (GQ >20) was used for all samples. This indicates a 99.9% confidence that the called genotype is the true genotype. The algorithm VCFtools was used to implement this filter; all genotypes failing this threshold were re-coded as missing [260]. A variant with greater than 15% of all samples re-coded to missing was excluded from further analysis.

2.5.7.4 Variant annotation

Variant effect predictor (VEP) was used for annotation [261]. For each variant, information on the gene and location of the variant in the gene (whether coding, intronic, intergenic or upstream) was added. Furthermore, information on the consequence of the variant on protein sequence such as missense, synonymous or altering a stop codon or splice site was added. Functional predictions were derived from PolyPhen-2 [211] and Sorting Intolerant From Tolerant (SIFT) [212] for missense variants while the VEP plugin LOFTEE was used for functional prediction of stop gain or stop loss, splice site disrupting and frameshift variants [261]. Additionally, allele frequency annotations from 1000 genomes and Exome Aggregation Consortium (ExAC) were added [262].

2.6 Analytical methods

2.6.1 Population stratification

Allele frequencies between individuals may differ due to genetic ancestry differences. If cases and controls are ascertained from different populations, spurious associations due to ancestry stratification rather than the phenotype of interest will occur [263,264]. Confounding due to either poor case control matching of ethnicity leading to population stratification, genotyping errors or batch effects in the cases and controls can be

detected empirically based on genome-wide inflation of test statistic popularly referred to as genomic control [265, 266]. Robust methods that infer genetic ancestry from the thousands of markers in GWAS have been developed to correct for stratification. Multidimensional scaling (MDS) and Principal Components Analysis (PCA) were used in this project.

2.6.2 Multi-Dimensional Scaling (MDS)

In order to characterise the extent of population stratification between cases and controls, MDS as implemented in the program PLINK v1.06 was used [267]. Identity by state (IBS) describes the amount of genetic information shared between individuals. PLINK estimates the IBS distance for all pairs of individuals with the resulting pairwise IBS then used in a clustering process to create an IBS matrix. In this procedure, each individual starts as a cluster of 1, and the two closest clusters are then merged. Clusters to be merged are determined by comparing similarities of members whereby clusters with too many dissimilar members are excluded [267]. In MDS, each point in a high dimension space, as expressed by the IBS matrix, is projected to a low dimension space while preserving the distance between the two points. The variation of genetic data is reduced to user specified dimensions allowing visualisation of each pair of data in a two dimensional space. MDS was performed with reference samples from the HapMap Project [139, 140] in order to identify and subsequently remove population outliers in all analysis.

2.6.3 Principal Component Analysis (PCA)

Principal components analysis is another widely used method of inferring and correcting stratification using ancestry informative markers [268]. Genetic variation is quantitatively expressed as continuous axes, with the major axes representing the greatest amount of variation. The derived components are then used as fixed covariates in a linear regression model to test for association between genotype and phenotype [263]. The first ten ancestry informative principal components derived from

common markers using the program EIGENSTRAT [268] were used as covariates in a linear mixed model association analysis presented in Chapter 4.

2.7 Statistical analysis

2.7.1 Single variant association analysis

Two tests were used for single variant association analysis, Fisher's exact and linear mixed model (LMM) association tests. In addition to case-control association analysis, results from these analyses were used to assess any inflation of test statistic reflecting poor quality control or population stratification.

2.7.1.1 Fisher's exact test

This test was used for analysis of association with susceptibility to meningococcal disease, presented in Chapter 3. The test is used to determine a non-random distribution of categorical variables in a contingency table and is especially useful if any cells in the contingency table have values less than five. The difference between observed and expected values under the assumption of independence is calculated to give an exact probability value for the relationship between the variables.

2.7.1.2 Linear mixed models

Linear mixed models (LMM) have the advantage over MDS and PCA in that they can simultaneously correct for multiple levels of relatedness [269]. LMM assess the pairwise relationship across all individuals, enabling detection of not only geographic stratification and relatedness but also modelling of polygenic effects. Efficient Mixed Model Association eXpedited (EMMAX) as implemented in the package Efficient and Parallelizable Association Container Toolbox (EPACTS v3.2.6) was used for all mixed model association analysis [270].

2.7.2 Gene-based methods of analysis

Due to the fact that only a small proportion of individuals will carry a rare variant in addition to the need for adjusting for a large number of tests, methods have been developed that aggregate the signal from multiple rare variants across a functional unit which can either be a gene or a biological pathway [165]. Methods aggregating by both gene and pathway were explored.

2.7.3 Strategies for aggregation of variants

Different variant aggregation scenarios were considered. A minor allele frequency threshold of 1% was used to define rare variants while low frequency variants were defined as alleles of 1-5% allele frequency [144,151–153,192–194]. A novel mutation was defined as one that was not present in the Single Nucleotide Polymorphism database (dbSNP build 137), 1000 Genomes database, the National Heart, Lung and Blood Institute (NHLBI) database of 6500 exomes, or the ExAC database of over 60,000 exomes. The following subgroups were analysed [188]:

2.7.3.1 Loss of Function (LoF)

1. All LOFs, which included only splice site variants, stop codon gain or loss, frameshift variants and inframe deletions and insertions.

2. Novel LOFs analysis was used in the study of meningococcal disease susceptibility presented in Chapter 3. Burden of novel loss of function mutations was compared in cases versus controls and p-values derived from Fisher's exact test.

2.7.3.2 Protein altering mutations

1. All rare and low frequency missense variants predicted as deleterious by either PolyPhen-2 or SIFT and all rare and low frequency LoF mutations.

2.7.4 Statistical methods of gene-based analysis

2.7.4.1 Burden test

Burden testing as implemented in PLINK/SEQ [271] was used. The raw count of alleles less than 1% was compared in cases against controls with permutation to calculate an empirical p-value. This method was used in Chapter 3 for analysis of meningococcal cases.

2.7.4.2 Sequence Kernel Association Testing (SKAT)

SKAT is a variance component test that allows for variants in a gene to have bi-directional effect [210]. A score statistic is derived for each variant weighed by minor allele frequency. SKAT uses multiple regression based methods to model the effect of variants on phenotype as random effects and allows for inclusion of covariates as fixed effects in the model. This method - as opposed to burden based methods - is robust to inclusion of variants of no effect or variants of opposite effect in the region being tested, and so provides improved power over burden tests if there are both risk and protective variants in the same gene [210]. Notably, if the majority of the variants are causal, burden tests are more powerful. Recently, an optimised method appropriately named SKAT-O has been developed that boosts the power of the test with increasing number of causal variants, allowing combination of burden and variance component based test [208, 272]. Additionally, this method allows for small-sample adjusted p-value calculations. SKAT-O with small sample size adjustment as implemented in EPACTS was used in the analysis of CAP, IPD and empyema in Chapters 4, 5 and 6 respectively.

2.7.5 Genomics analysis of biological networks

Due to genetic heterogeneity underlying susceptibility to complex diseases, protein-protein interactions can be a useful way to dissect whether products of genes contributing to a similar phenotype interact in any way. In this project, STRING [273] database was used to identify underlying biological networks in defined gene sets. The program allows analysis of statistical enrichment of gene ontology categories, KEGG pathways

or protein-protein interaction networks in user-defined gene lists. The database was used in the systematic application of the following approaches to pre-defined gene sets:

1. Enrichment of Gene Ontology (GO) [274] terms in pre-defined gene sets.
2. Enrichment of genes in pathways as defined in the Kyoto Encyclopaedia of Genes and Genomes (KEGG) [275].
3. Enrichment of protein-protein interactions in the gene sets.

For all approaches, all terms, pathways or interactions associated with at least one gene in the gene-set being tested were compared against a genomic background. A hypergeometric test was used to derive a p-value corrected for multiple testing using the Benjamini and Hochberg method [273].

2.8 Molecular biology methods used in validation of genetic findings

2.8.1 Sanger sequencing

Each PCR reaction consisted of: 1.2x of PCR buffer (Qiagen), 240µM each dNTP (Qiagen), 1.2mM $MgCl_2$ (Qiagen), 240µM each primer, 0.3 units of HotStar Taq DNA polymerase (Qiagen) and 20ng of DNA. Deionised water was used to bring the reaction volume to 15µl. The cycling conditions were: 95°C for 15 minute and 45 cycles of 94°C for 30 seconds, 69°C for 30 seconds, and 72°C for 2 minutes. This was followed by a final extension at 72°C for 10 minutes. Successful amplification was determined by gel electrophoresis using 2% (w/v) agarose gel with ethidium bromide. To prepare the PCR products for sequencing, the ExoSAP-IT kit was used to remove unconsumed dNTPs and excess primers. The following reaction mixture was used: 4µl PCR product to be cleaned, 10 units Antarctic Phosphatase (5000U/ml, New England Biolabs), 4 units Antarctic phosphate buffer (New England Biolabs), 2 units Exonuclease l (New England Biolabs) and 4µl water for a final reaction volume of 10µl. For the sequencing reaction each reaction mix consisted of 2µl cleaned PCR product,

0.25µl Big Dye Terminator Sequencing Mix v3.1 (ABI), 0.94x BigDye Dilution Buffer, and 0.5µM sequencing primer and 5.4µl water for a reaction volume of 10µl. The cycling conditions were 95°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes for 25 cycles. The DNA was then precipitated using ethanol and Sanger sequencing was carried out at the Zoology Department, University of Oxford.

2.8.2 Molecular cloning

2.8.2.1 GeneArt® gene fragment synthesis

The open reading frame for Complement Factor Properdin (CFP; described further in Chapter 3) was synthesised using DNA GeneArt® Strings^{TM} (Invitrogen). Two forms of the gene were designed, the wild type form and a mutant form where the sequence was edited to reflect a single base pair change at position 249 of the sequence from a G to A, resulting in an amino acid change from arginine to cysteine. The sequence was then optimised as follows. The signal peptide was confirmed by loading the sequence into signalP 4.0 Server [276]. A Kozak sequence was inserted immediately upstream of the first methionine. Two restriction sites were then inserted, a *Hind*III restriction site at the 5' end and a *Not*I restriction site at the 3' end of the sequence. Codons were optimised for efficient translation in HEK293E cells. Four random bases were then added at each end to anchor the restriction sites. Illustrations of these fragments are presented in Chapter 3 (Figures 3.3 and 3.4).

2.8.2.2 Ligation and transformation of *E.coli*

The plasmid 1990_pENTR4LPTOS (Chapter 3, Figure 3.3) was used in this experiment, kindly provided by Dr. Simon Draper, Jenner Institute, University of Oxford. The DNA fragments and the plasmid DNA were digested with *Hind*III (New England BioLabs) and *Not*I (New England BioLabs) to create compatible overhangs. Restriction was carried out in 20µl reactions of 2 units NEB buffer, 10 units of each of the enzymes, 1µg of DNA fragments and 1µg of the plasmid. The volume in each tube was then adjusted to 20µl by adding water. Digestion was carried out at 37°C for 2 hours. For a vector:insert ratio of 1:1, the following reaction mix was prepared:

1 unit of DNA ligase, 1X ligation buffer 5µl each of plasmid and appropriate insert DNA (either wild type, mutant or just plasmid from digestion reaction), and deionised water for a final reaction volume of 20µl. Ligation was carried out overnight at room temperature. High efficiency DH5α cells (New England BioLabs) were thawed on ice and three aliquots of 30µl each were made. To each aliquot 3µl of ligation reaction was added to respective tubes and the mix incubated on ice for 30 minutes. The reaction was then heat shocked for 30 seconds at 42°C and immediately placed on ice. After 5 minutes, 250µl of Super Optimal broth with Catabolite repression (SOC) broth (New England BioLabs Inc.) was then added to each tube and the reaction incubated at 37°C for 1 hour with vigorous shaking. Each reaction was then plated on pre-warmed LB agar plates with Kanamycin and incubated overnight at 37°C.

2.8.2.3 Colony screening

To identify successful transformations, PCR screening was carried out on six randomly selected colonies from each plate. Each reaction mix consisted of 10µl of ReddyMix (ThermoScientific), 240µM of each primer and one colony picked with a sterile pipette tip. The PCR cycling conditions were as follows. An initial heating step of 98°C for 10 minutes was followed by 10 cycles of: 98°C for 30 seconds, 68°C for 2 minutes and 72°C for 3 minutes. This was then followed by 25 cycles of 98°C for 10 seconds, 68°C for 30 seconds, 72°C for 2 minutes and the reaction was held at 10°C. Each PCR reaction was then run on 1% agarose gel to confirm successful ligations.

2.8.2.4 Plasmid preparation

For each colony that showed successful transformation, a plasmid preparation of 6ml Luria Bertani (LB) broth supplemented with 150µg Kanamycin was made. Each culture was then seeded with the respective successfully ligated colony. The culture was then incubated overnight at 37°C with vigorous shaking. DNA was extracted from an aliquot of the culture using the QIAprep Spin Miniprep Kit (Qiagen) as per manufacturer's instructions. Sanger sequencing was the carried out to confirm correct sequence of the insert. After Sanger sequencing confirmation, a maxi-prep

culture of 150mls of LB broth supplemented with 3.5mg Kanamycin was seeded with an aliquot of the previous mini-prep culture. DNA was extracted using the QIAGEN Plasmid Plus Maxi Kit as per manufacturer's instructions.

2.8.2.5 Transfection of HEK293E cells

Human Embryonic Kidney cells expressing EBV EBNA1 antigen (HEK293E) were used for this experiment due to their ease of transfection and culture and their potential to produce up to 1000µg of secreted protein. Furthermore, these cells lack endogenous expression of CFP in the native state which would otherwise have confounded the experiment [277]. In addition, a mammalian expression system as opposed to bacterial systems was chosen as it allowed for correct folding and post translational modifications of proteins. HEK293 cells were routinely maintained in culture at the Old Road Campus Research building by Dr. Jing Jin. On the day before transfection (day one), cells were split to yield 100mls of HEK293E cells at a density of 1.0 x10^6 cells/ml. Cell density was determined as follows. Firstly, 20µl of cells were mixed with 0.4% of trypan blue. A 20µl aliquot of this mixture was then loaded to a disposable counting chamber. When observed under a light microscope, dead cells absorb the dye while viable cells exclude dye. Viable cells in one 4x4 chamber were counted and multiplied by 2.0 x10^4 to yield the density of cells/ml of culture. Split cells at the correct density were then incubated at 37°C in 5% CO_2 overnight for next day transfection. Every experiment was carried out in triplicate: a separate culture for the wild type, mutant and a mock negative control. On the day of transfection (day two), cells were divided into three flasks of 30 ml each. The time the cells were out of the incubator was kept to a minimal. For each experiment 1µg of DNA was needed for each ml of culture i.e. 30µg DNA for a 30ml culture. Of the total DNA, 5% had to comprise a GFP expressing plasmid in order to check for transfection efficiency. The total DNA was pre-incubated for 15 minutes with the transfection reagent Polyethylenimine (PEI) in Invitrogen Freestyle 293 expression medium. This was then added drop wise to the respective flask while swirling and the culture returned to the incubator. On day four, the cells were counted again. Transfection efficiency

was determined by counting cells that had green fluorescence under blue light. A successful transfection had to generate greater than 40% GFP-positive cells. On day five, cells were pelleted by centrifugation and a protease inhibitor cocktail added to the supernatant. A 2ml aliquot of the supernatant was added to 18mls of ice-cold ethanol for overnight precipitation at -20°C. The rest was immediately stored at -80°C. The cells were retained and processed immediately.

2.8.2.6 Western blot

Precipitated supernatant proteins were separated on Criterion™ TGX™ precast gel (Bio-Rad). Separated proteins were then transferred to a Nitrocellulose membrane using the Trans-Blot® Turbo Transfer System (Bio-Rad). The membrane was blocked with 3% BSA dissolved in PBS. The blocked membrane was probed with unconjugated Properdin/CFP Rabbit anti-Human Polyclonal antibody (LifeSpan Biosciences, Inc.) followed by incubation with alkaline phosphatase conjugated Donkey anti-rabbit secondary antibody (Jackson ImmunoResearch Inc.). Pooled human serum was used as positive control. The results were confirmed using anti-properdin mouse monoclonal primary antibody (VWR International).

2.8.2.7 Intracellular staining

To enable microscopic analysis of the cell culture, Poly-L-Lysine was used to attach cells to a glass slide. Poly–L-Lysine - a cationic homopolymer of the amino acid lysine - is routinely used in cell culture as a very efficient attachment factor to promote adherence of cells in culture. An aliquot of the suspension culture was added to a Poly-L-Lysine coated slide (ThermoScientific) and incubated for 45 minutes at 37°C. Non-adherent cells were washed off with ice-cold PBS. The cells were then fixed with 2% paraformaldehyde and the slide blocked using 1% PBS-BSA. Permeabilisation was achieved using 0.5% PBS-Triton X-100. The cells were then incubated with unconjugated Properdin/CFP Rabbit anti-Human polyclonal antibody (LifeSpan Biosciences, Inc.). Excess antibody was washed off and the cells incubated with Alexa Fluor 647 goat anti-rabbit IgG (H+L) secondary antibody (Life Technologies). Cells

were then mounted with VECTASHIELD® Mounting Media with DAPI. Microscopy was carried out using a Zeiss 510 MetaHead confocal fluorescence microscope and image capture and analysis was done using Zeiss LSM Image Browser Version 4.2.0.141.

Chapter 3

Rare genetic variation and susceptibility to meningococcal disease

3.1 Introduction

Asymptomatic nasopharyngeal carriage of *N. meningitidis* is relatively common, averaging at 10-35% in the general population with highest rates of 23-55% reported for adolescents [43, 44, 55]. Despite this, disease prevalence is relatively rare with incidence rates of 1:100,000 in European populations. Therefore, despite relatively common carriage, the incidences of disease remains rare with the rate of invasive disease amongst carriers (case:carrier ratio) an average 1:10,000, varying highly with age [75–77, 278]. It is increasingly recognised that host genetic variation may in part explain why some individuals are susceptible to invasive meningococcal disease [238, 243, 279, 280]. Indeed a third of the 30-fold increase in risk for siblings of cases is attributable to genetic factors [279]. Recently, common polymorphisms in the gene region encoding complement factor H (*CFH*) and CFH-related protein 3 (*CFH3*) have been described that associate at genome-wide level with a small increase in risk of developing meningococcal disease in childhood [281]. Furthermore, studies on selected individuals with recurrent episodes or multi-case families have uncovered rare, highly penetrant host genetic defects in both early and late complement components that predispose to meningococcal disease [12, 237, 239, 282, 283]. Individuals deficient in late complement components are associated with a 7,000-10,000 fold increase in risk of invasive disease over the general population [278].

The relative contribution of rare mutations to meningococcal disease susceptibility at the population level remains unknown however. This study uses a whole exome sequencing approach to study the potential role of rare large effect genetic variants in the aetiology of meningococcal sepsis in a cohort of critically ill unrelated adults presenting to ICUs in the UK with meningococcal disease. Fifteen individuals with microbiologically proven meningococcal disease were recruited as part of the Genetics of Sepsis and Septic Shock in Europe (GenOSept) consortium, a multi-centre study involving 143 centres across 16 countries in Europe. Whole exome data from 369 unrelated Scottish individuals from the UK10K-OBESITY-GS study and 1,000 individuals from the 1958 birth cohort were used as population controls [249, 250]. Two individuals were excluded from this study due to ethnicity and an already identified primary immunodeficiency. Further details on these cohorts are described in Materials and Methods section 2.2.

3.2 Results

3.2.1 Sequence metrics

A summary of the sequencing and mapping metrics for each individual with meningococcal disease are presented in Table 3.1. The TruSeq Exome enrichment kit probes are designed to capture 201,032 exons in 20,794 genes for a target region totalling 62MB. An average of 90% of the target regions was covered at greater than 10X while 83% were covered at 20X across all samples. This is in line with the recommended target of 80-95% of target regions covered at 20X in whole exome sequencing studies [199]. All samples had greater than 99% of all raw reads mapping to the reference human genome GRCh37 (hs37d5). There were 173,146 variants detected across 13 samples. Of these, 125,324 variants passed quality control with 28.4% occurring in coding regions. Loss of function (LoF) mutations - i.e. splice site variants, stop codon gain or loss, frameshift variants and inframe deletions or insertions - accounted for 3.24% of all variants passing quality control. Moreover, 12,535 (10.8%) of variants in coding

Table 3.1: Summary of sequence metrics. A summary of whole exome sequence metrics across 13 individuals with menigococcal disease. The TruSeq™ Target Enrichment kit (Illumina Inc. San Diego, CA, USA) was used for exome capture.

Sequencing metrics	**Metric value**
Size of target region	62,085,295
Mean depth of target region	66.2
Percentage of target bases covered >= 10x (%)	89.9
Percentage of target bases covered >= 20x (%)	82.9
Percentage of target bases covered >= 30x (%)	73.3
Average duplication rate (%)	5.2

regions were rare in the Exome Aggregation Consortium (ExAC) and 1000 Genomes online databases [258, 262]. LoFs accounted for only 1.8% of variants in coding regions.

3.2.2 Novel variants in meningococcal disease

As large effect risk variants are likely to be rare or private in the general population, initial analysis was focused on all predicted functional novel variants missing from ExAC and 1000 Genomes. The prediction algorithms PolyPhen-2 and SIFT were used in prediction of the functional impact of missense variants on protein function [211,212]. A lenient filtration strategy was adopted, whereby a missense variant predicted as deleterious by any one of the two algorithms was included in analysis (Figure 3.11). There were 1,896 mutations that were present in meningococcal disease cases but not present in the ExAC and 1000 Genomes databases, 87.5% were in non-coding regions. Excluding synonymous variants, there were 168 variants in coding regions. These included 25 LoFs and 133 missense variants of which 99 were predicted as deleterious by at least one of the prediction algorithms used. There were no novel mutations shared across individuals, moreover, there were no genes that had a novel mutation in more

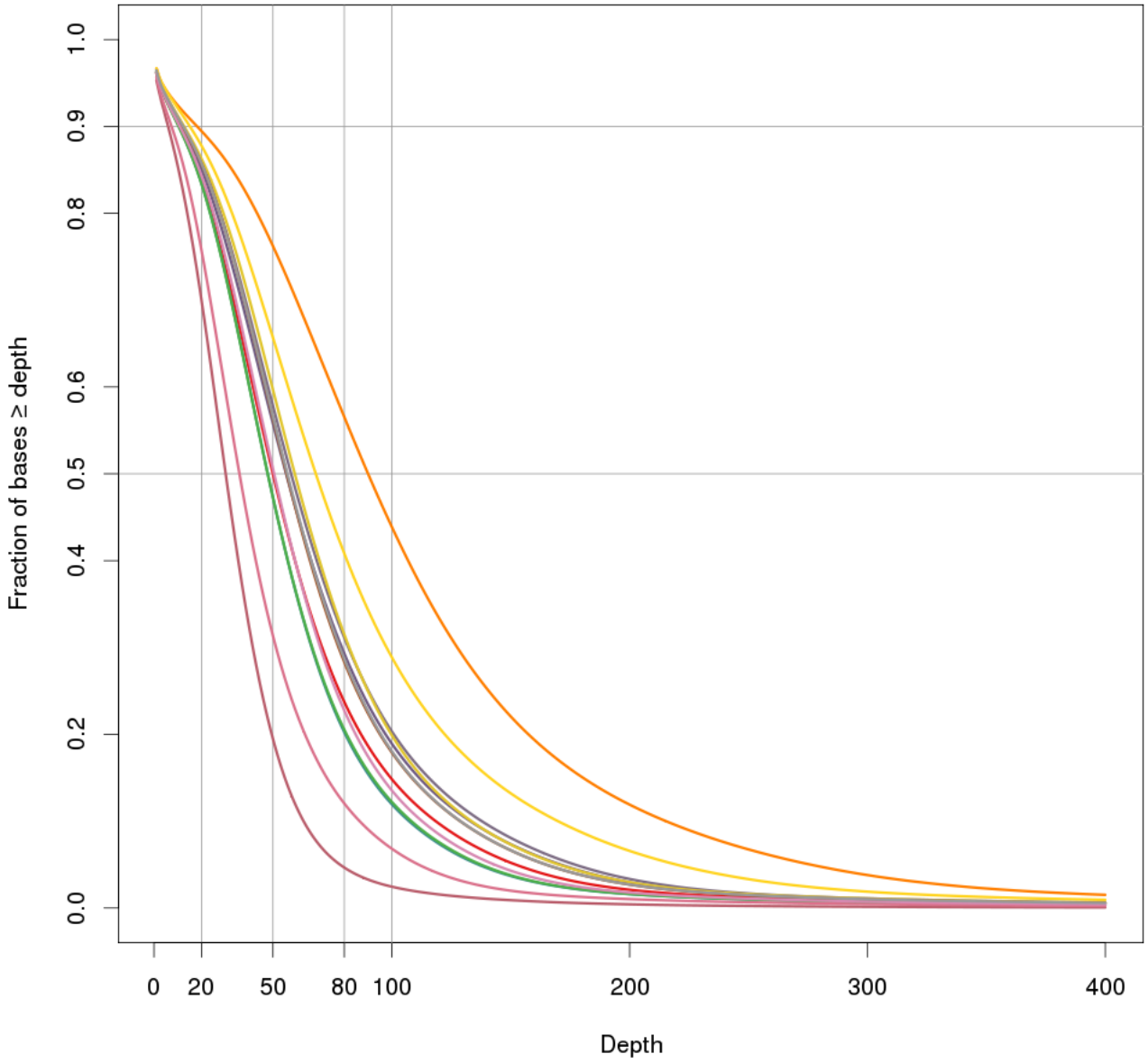
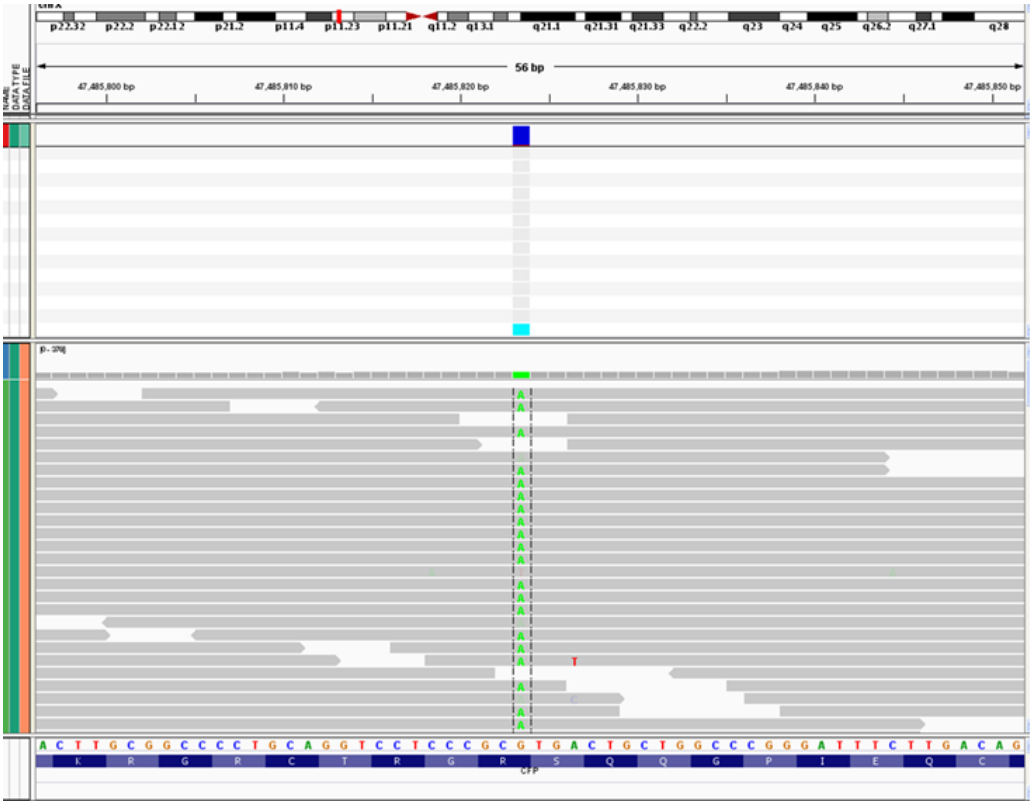
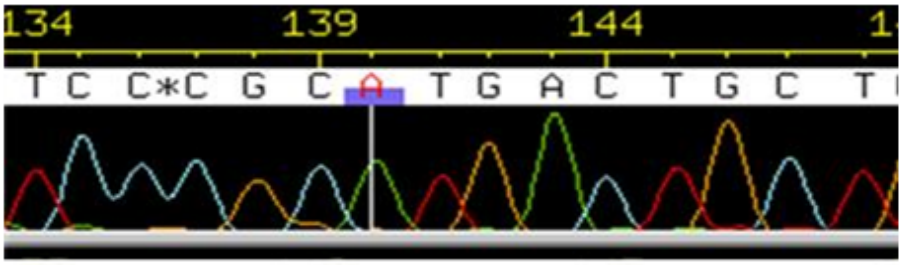


Figure 3.1: Coverage across meningococcal samples. Exome sequencing coverage in 13 individuals with meningococcal disease showing the proportion of the 62MB targeted region covered at various depths. Each line represents metrics for one individual. An average of 90% of regions targeted for capture were covered at greater than 10X in line with recommended depth for exome sequencing.

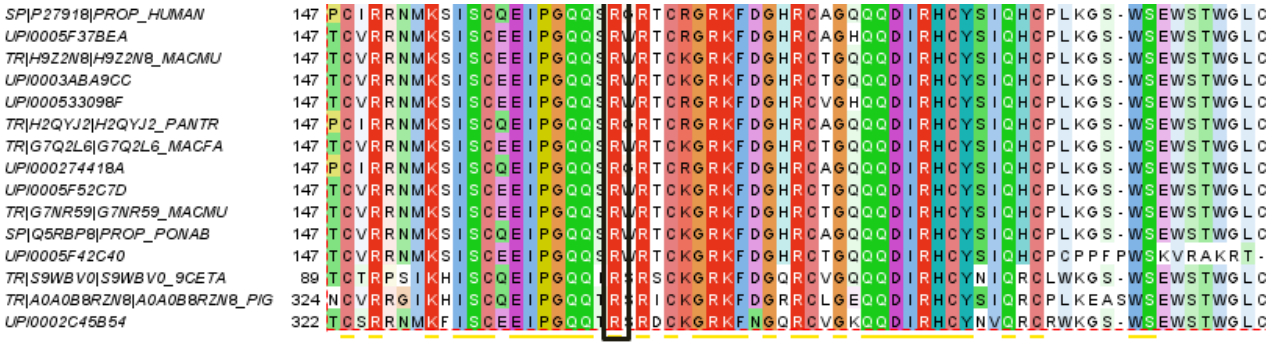
than one individual. Two genes, *LZTR1* and *CFP*, had homozygous/hemizygous novel mutations. The *LZTR1* gene encoding the leucine-zipper-like transcriptional regulator 1 protein had a novel homozygous splice site variant in one individual. Mutations in this gene predispose to schwannomatosis, a predisposition to tumours in various parts of the body [284]. No immune-related phenotypes have been described for this gene, including any animal model; knockout mice exhibit craniofacial defects and mortality at 3-4 weeks after birth [285]. A novel hemizygous missense mutation in one male individual was identified in the X-chromosomal *CFP* gene encoding the complement factor properdin, a positive regulator of the alternative complement pathway. This variant - a G to A base pair change – led to an amino acid substitution from an arginine to a cysteine at residue 346 in the polypeptide chain. It occurred at a highly conserved site and was predicted as deleterious by all prediction algorithms (Figure 3.2a, 3.2c). The mutation was subsequently confirmed by Sanger sequencing (Figure 3.2b).



(a)



(b)



(c)

Figure 3.2: The R346C mutation in the *CFP* gene.(a) Integrated Genome Viewer snapshot showing pile-up of reads. (b) Sanger sequencing trace confirming the variant and (c) conservation of the residue (in rectangle) across species

3.2.2.1 Functional impact of Arg346Cys mutation

Mutations in the *CFP* gene have been previously associated with an increased risk of meningococcal infections in multi-case families and individuals with recurrent episodes of meningococcal disease [240, 241, 286–288]. In order to characterise the functional impact of the novel missense mutation on protein expression and function, two GeneArt® Strings™ (Invitrogen) DNA fragments corresponding to the open reading frames of wild type and the mutant form of the *CFP* gene were synthesised; the design of each fragment is described in Materials and Methods section 2.8.2.1. Codons were optimised for expression and translation in HEK293E and the polypeptide fragments differed at amino acid 346. The plasmid 1990_pENTR4LPTOS used in transfection is shown in Figure 3.3, DNA fragments cloned are shown in Figures 3.4 and 3.5. The cell culture supernatant of HEK293E cells transiently transfected with either the mutant 346Cys or wild type 346Arg form of *CFP* were assayed for secreted properdin by western blot. Indirect detection using a polyclonal primary antibody targeting amino acids 203-233 revealed an absence of secreted protein in cells transfected with the mutated gene compared to cells transfected with wild type. This result was confirmed using a monoclonal antibody (Figure 3.6a). A quantitative polymerase chain reaction revealed presence of mRNA in both wild type and mutant cells (Figure 3.6b). Subsequent intracellular staining of HEK293E cells demonstrated detectable intracellular protein in both wild type and mutant cells (Figure 3.7). Taken together these findings indicate intracellular production but defective extracellular secretion in the *CFP*-346Cys mutant form.

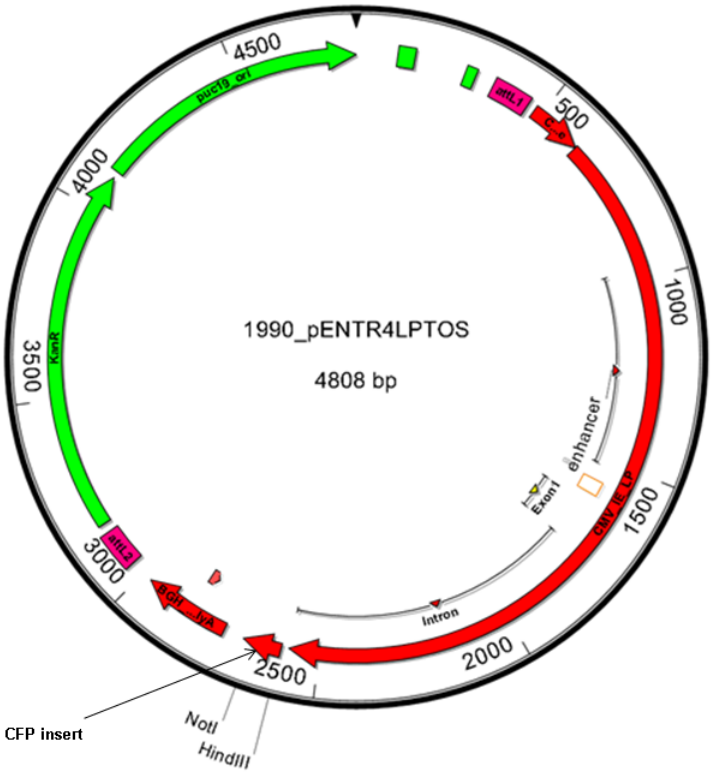


Figure 3.3: Vector map for 1990_pENTR4LPTOS. This plasmid was used for transfection of wild type and mutant fragments in HEK293E cells. The plasmid had a Kanamycin resistance gene and a strong CMV promoter to drive over-expression of the protein.

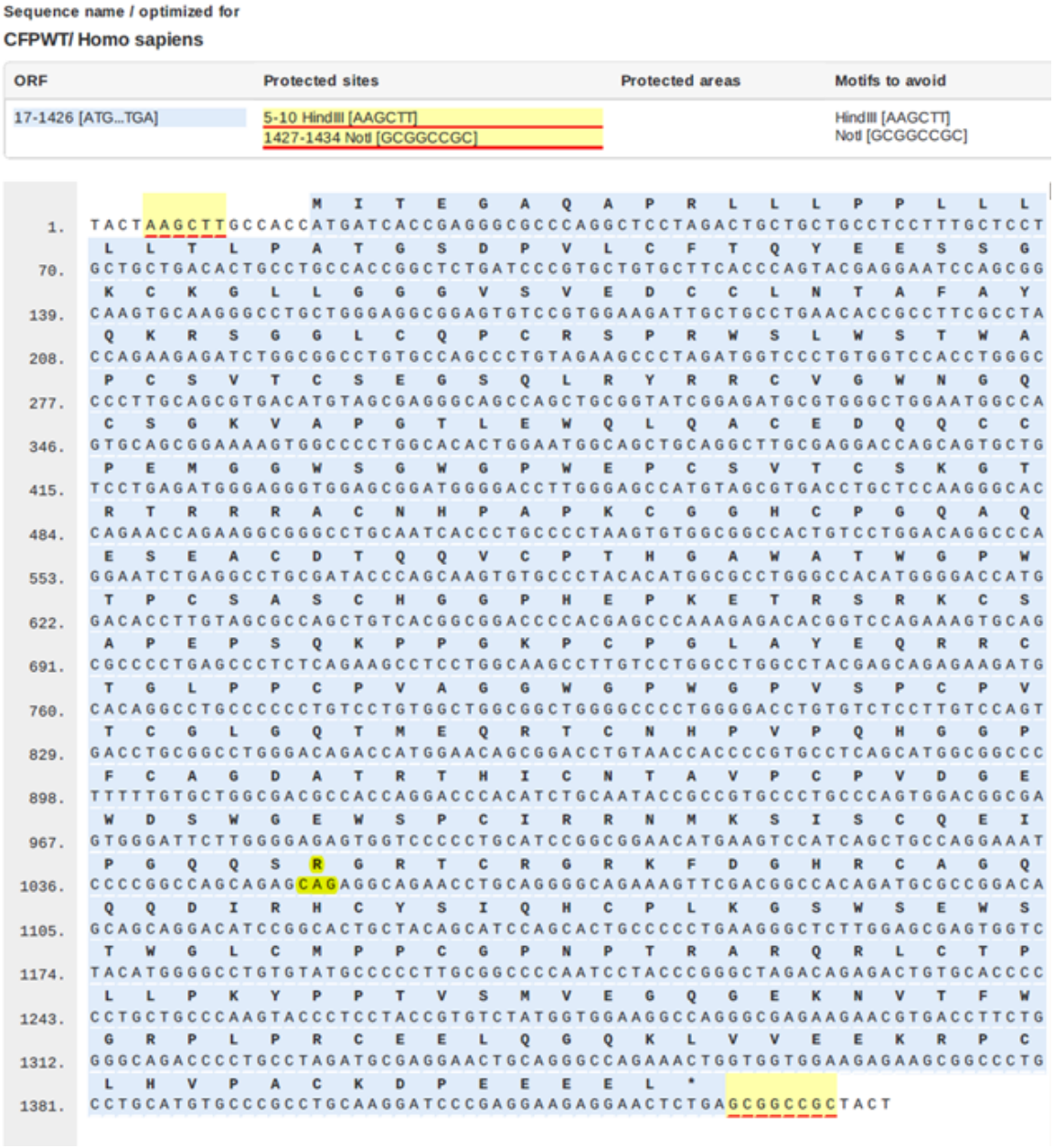


Figure 3.4: GeneArt® Strings™ (Invitrogen) 346Arg DNA fragment. Wild type *CFP* DNA fragment encoding 469 amino acid with arginine(highlighted) at position 346.

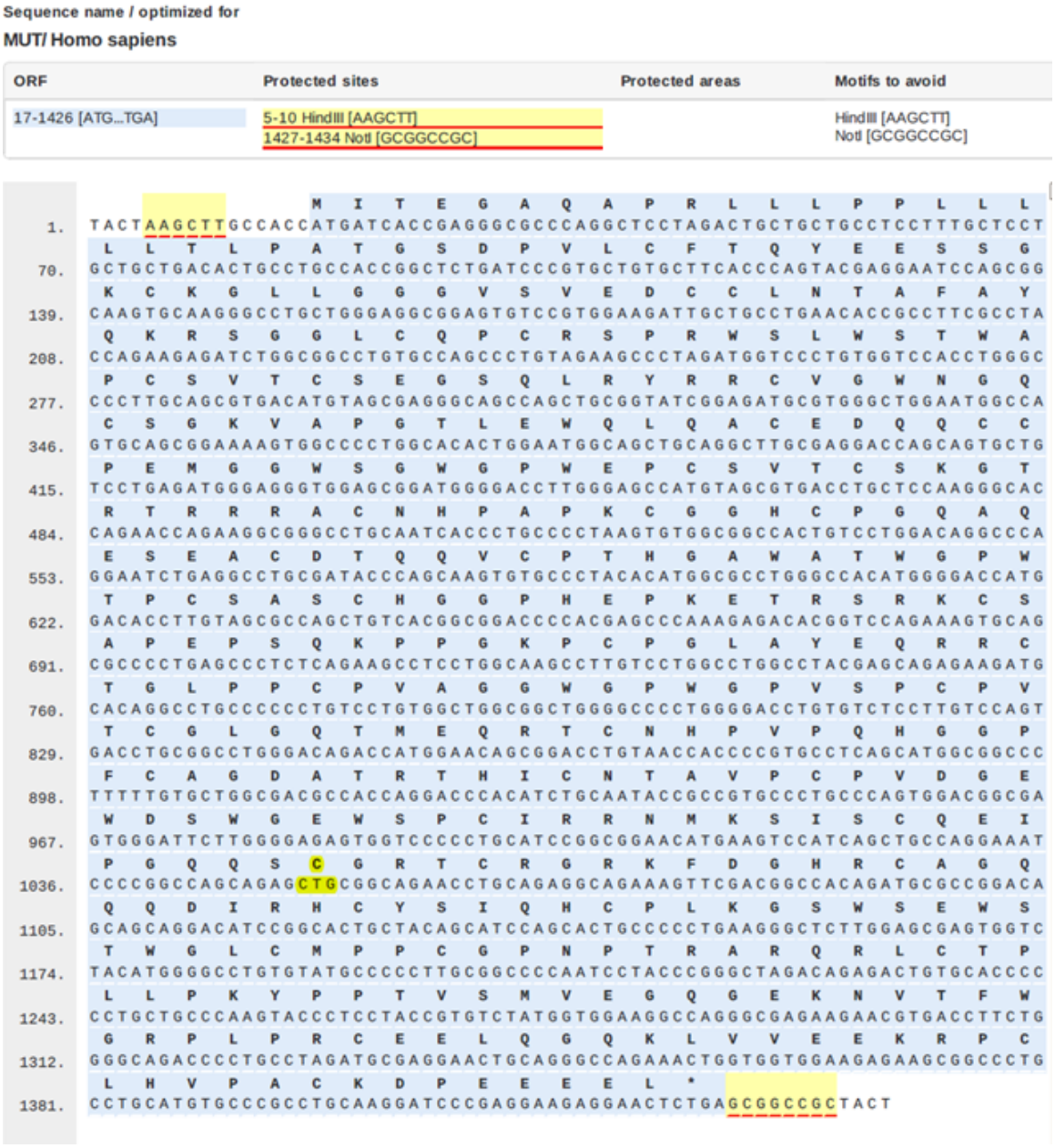
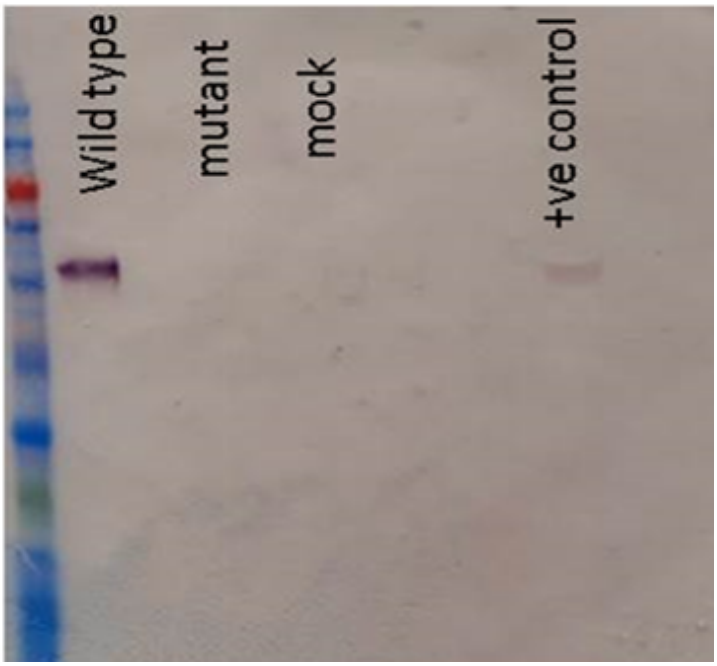
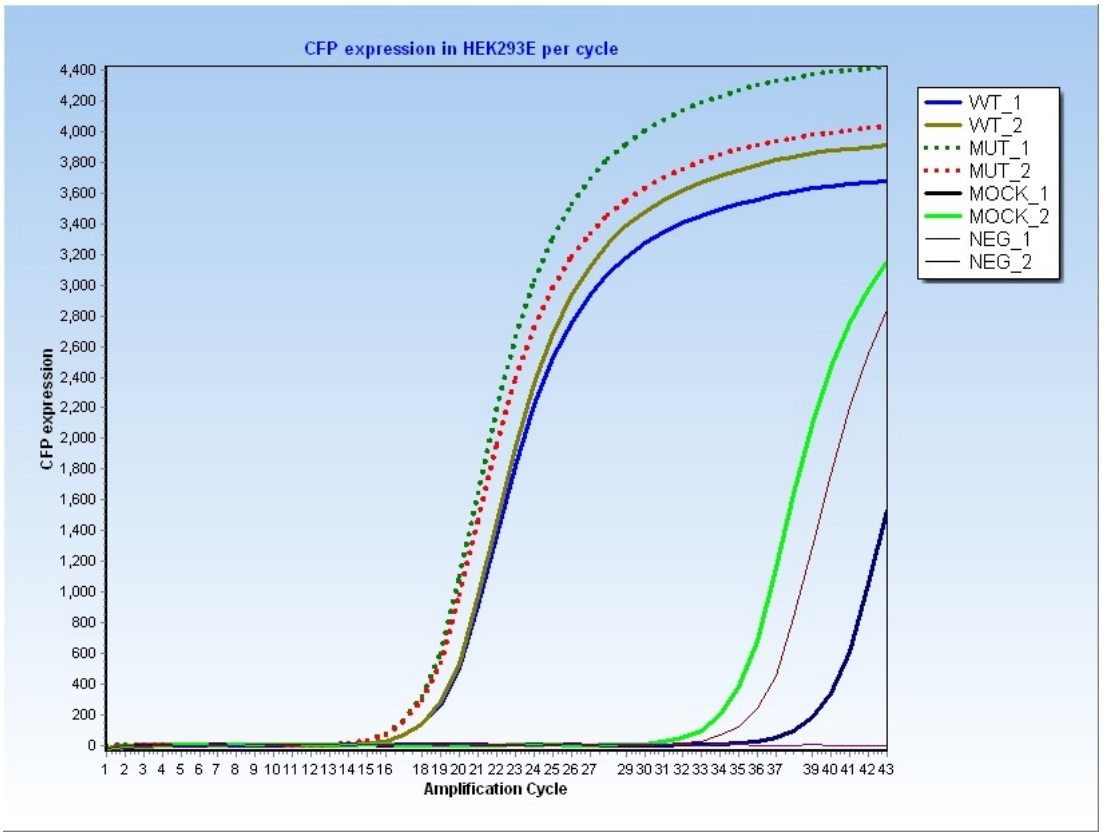


Figure 3.5: GeneArt® Strings™ (Invitrogen) 346Cys DNA fragment. Mutant *CFP* DNA fragment encoding cysteine (highlighted) at position 346.



(a) Western blot assay



(b) qPCR analysis

Figure 3.6: Properdin protein expression assays. Expression of properdin in HEK293 was assayed by (a) western blot of culture supernatant (b) quantification of mRNA expression using qPCR.

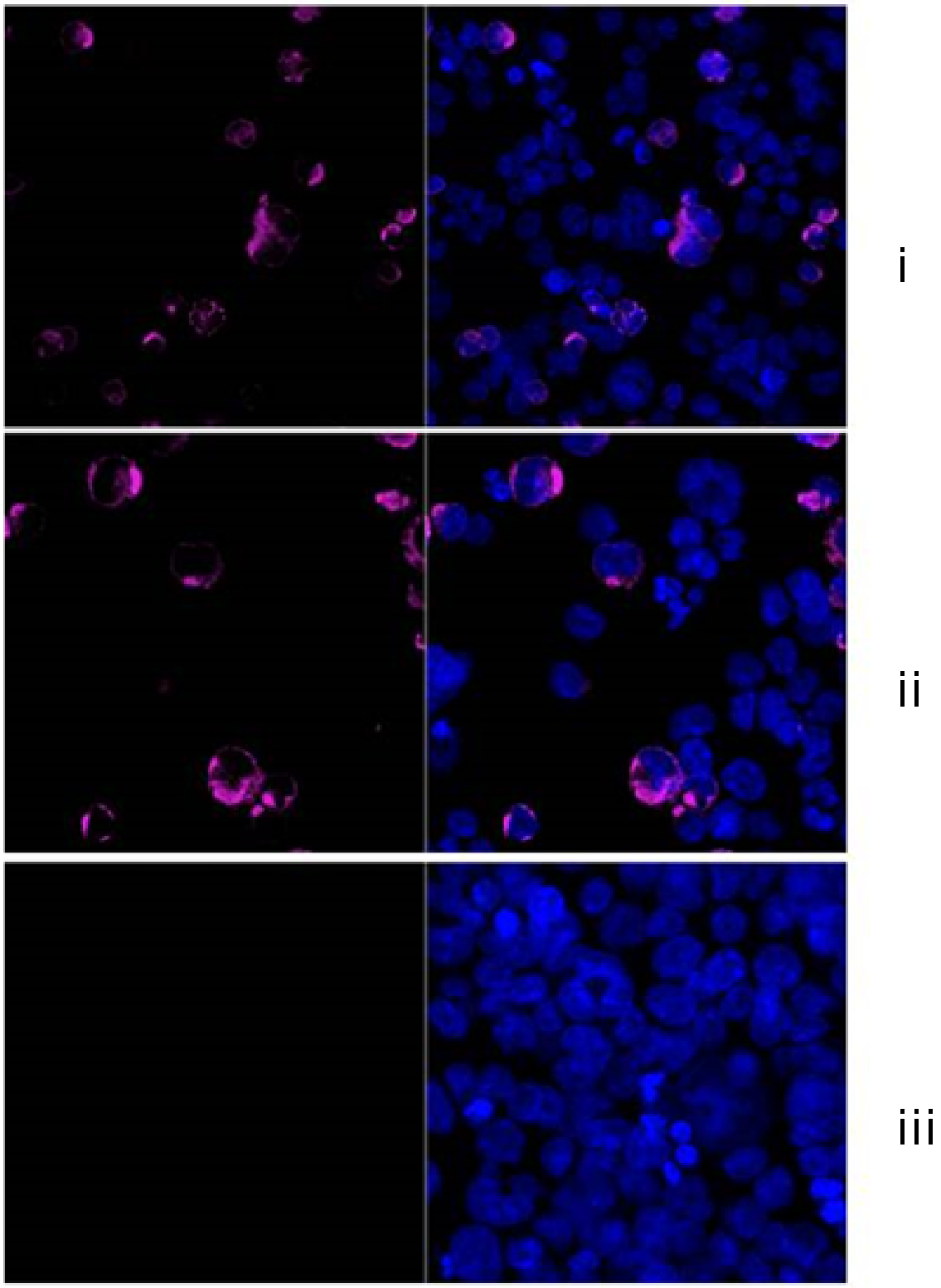


Figure 3.7: HEK293 intracellular staining. Intracellular staining of HEK293 cells transfected with the wild type (i), mutant (ii) and a mock culture (iii). Indirect staining with unlabelled primary antibody was followed by detection with Alexa Fluor® 647 conjugated secondary antibody.

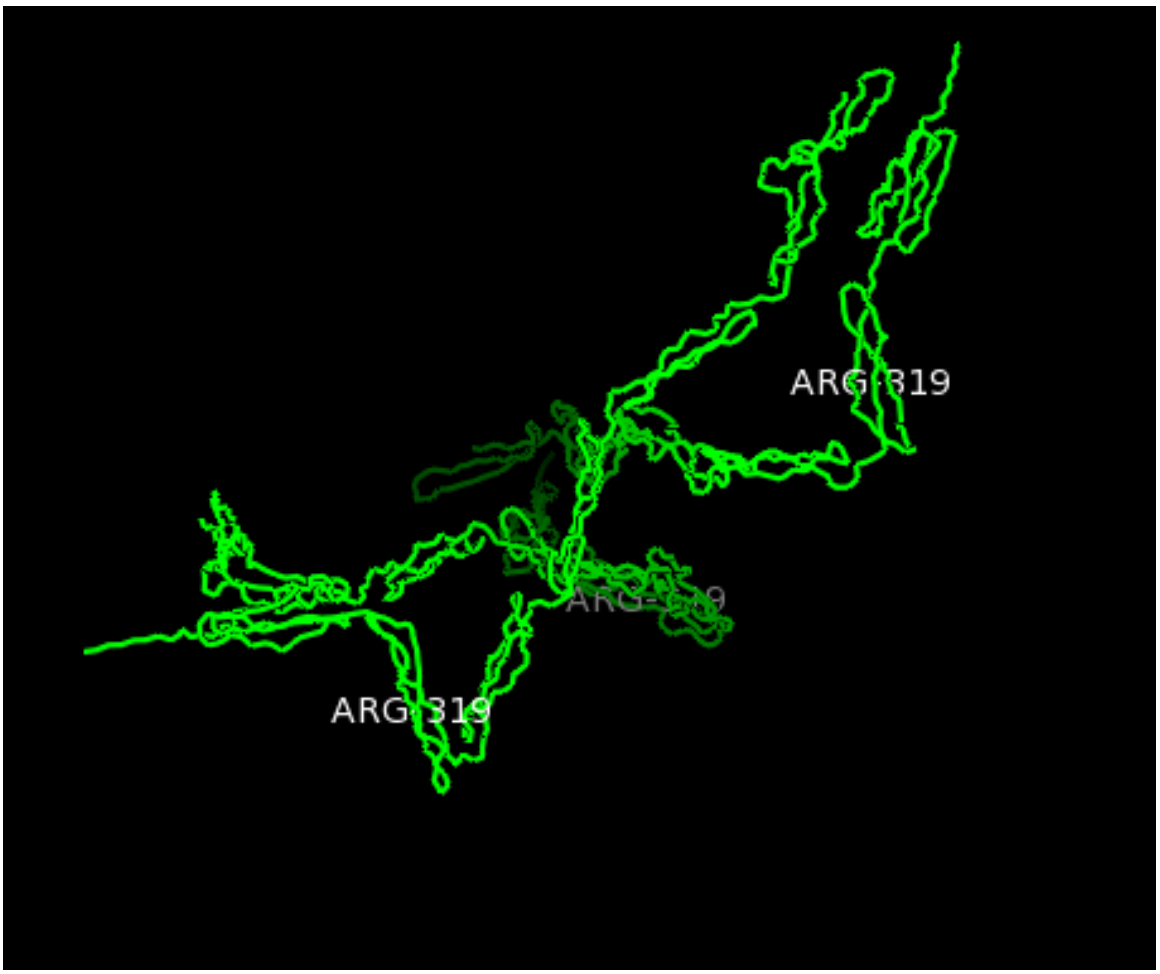


Figure 3.8: Secondary structure for the trimeric form of properdin. Highligted is the arginine residue mutated to cysteine in each monomer. The arginine residue is one of two highly conserved residues which together with three disulphide bonds maintain structural integrity of the monomer and facilitate formation of higher order oligomers for secretion.

3.2.2.2 Biological networks in all genes with novel mutations

Following initial analysis of individual novel variants, several functional genomic approaches were next used to test for enrichment of biological patterns relevant to disease. Ninety nine genes had novel missense mutations predicted as deleterious by either PolyPhen-2 or SIFT while 25 genes had a novel LoF for a total of 124 genes with a novel, potentially deleterious mutation. The program STRING [273] (Materials and Methods, section 2.7.4) was used to investigate over-represented GO categories, KEGG pathways or protein-protein interaction networks in the 124 genes with novel mutations. There were 39 protein-protein interactions among these genes, however, this did not represent a statistically significant enrichment of interactions above what was expected by chance ($P_{\text{corrected}}$=0.56). No GO terms were significantly overrepresented in these genes. There were 63 biological processes associated with at least one gene in the list. Of these, five had a nominal p-value of less than 0.05; however, none was significant after correction for multiple testing. The 124 genes were mapped to 114 KEGG pathways, 79 had only one gene on the list mapped to the pathway. Nine pathways (Table 3.2) achieved nominal significance though none were significant after correction for multiple testing.

Table 3.2: KEGG pathways with nominal enrichment of novel mutations. For 124 genes with novel predicted deleterious mutations, nine KEGG pathways had $P_{uncorrected}<0.05$.

GO-ID	KEGG pathway	No.ofGenes	$P_{uncorrected}$	$P_{corrected}$
5410	Hypertrophic cardiomyopathy (HCM)	4	$1.47x10^{-3}$	0.27
5414	Dilated cardiomyopathy	4	$1.91x10^{-3}$	0.2.7
5168	Herpes simplex infection	5	$3.91x10^{-3}$	0.37
5412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	3	$9.50x10^{-3}$	0.68
52	Galactose metabolism	2	$1.38x10^{-2}$	0.69
4022	cGMP-PKG signalling pathway	4	$1.46x10^{-2}$	0.69
2010	ABC transporters	2	$2.37x10^{-2}$	0.97
970	Aminoacyl-tRNA biosynthesis	2	$2.72x10^{-2}$	0.97
5416	Viral myocarditis	2	$4.27x10^{-2}$	1.00

3.2.3 Rare deleterious mutations in meningococcal disease

Analysis of rare mutations was next performed, based on the hypothesis that due to a fitness cost, large effect susceptibility variants for early onset and life-threatening infectious disease are unlikely to be identified at high frequency in the general population [144, 158, 289]. A comprehensive analysis of all rare coding variation defined as alleles of less than 1% allele frequency or missing from 1000genomes and ExAC databases, was undertaken. HaplotypeCaller was used for the joint calling of genetic variants from the 13 meningococcal cases together with 369 unrelated controls from the Generation Scotland study (Figure 3.12). A two-pronged analysis strategy was employed. In the first instance, a separate analysis of all highly disruptive LoF mutations was performed. These included all mutations likely to lead to a truncated protein or alter exon splicing. The second approach involved analysis of all protein-altering mutations which included all missense variants predicted as deleterious by at least one prediction algorithm as well as LoF mutations.

3.2.3.1 Loss of function mutations

Analysis of LoF mutations was carried out separately due to their higher probability of having a severe impact on protein function. There were 8,078 LoF mutations in 13 cases and 369 unrelated controls. Of these, 3,810 were rare while 670 were novel (25 of these in cases). A total of 3,127 genes had a rare or novel LoF with 365 genes having a qualifying variant in cases. Seventy four genes had a LoF mutation in cases only, all in single case individuals (Table 3.3). Three genes, *ASMT*, *GRIPAP1* and *BMX* had hemizygous mutations while one autosomal gene *LZTR1* (described in section 3.2.2) had a novel homozygous mutation. Functional genomics analysis revealed no GO terms or protein-protein interactions that were significantly enriched after correction for multiple testing. For pathway analysis, 94 non-exclusive KEGG pathways were associated with at least one gene (Appendix B). The most significant KEGG term was 'primary immunodeficiency' with a nominal p-value of 6.74 $x10^{-3}$ primarily driven by *CD4* and *ZAP70*; this however, did not withstand correction for multiple testing.

Among the 94 pathways, 24 related to the immune system and infectious diseases (viral and bacterial) functional hierarchies. Additionally, 13 pathways were involved in signal transduction including innate immune signalling pathways (appendix Table B.1). This apparent enrichment of rare LoF mutations in immune-related genes in cases prompted further analysis of the 74 genes in a further 1,000 population controls from the 1958 birth cohort. Of these, 30 genes did not have a rare or novel LoF in a further 1,000 controls (Figure 3.8), all mutations were singletons and 10 were novel. Protein-protein interaction analysis revealed a statistically significant enrichment of interactions in these 30 genes with $P_{\text{corrected}}$=0.04 (Figure 3.8). There were five direct interactions observed. One interaction (*ZAP70*, *CD4*) involved key immune related genes, two interactions (*GABRD*, *GABRA2*, *GLB1*, *ASPA*) involved metabolic genes, and one interaction (*TRPC3*, *FKBP4*) involved oestrogen receptor genes while the fifth interaction (*BMPR2*, *RUNX2*) involved a transcription growth factor gene and transcription factor regulating differentiation of hematopoietic stem cells.

Table 3.3: Genes with case only LoF. Seventy four genes had rare or novel loss of function mutations identified in meningococcal cases but not present in 369 controls.

Gene	Variant	Gene	Variant
SYTL1	splice_site variant	*C10orf47*	splice_site variant
SNRNP40	splice_site variant	*ILK*	splice_site variant
EXTL2	splice_site variant	*CD5*	splice_site variant
CACNA1E	splice_site variant	*SHANK2*	splice_site variant
GABRD	splice_site variant	*DYNC2H1*	splice_site variant
CYP4Z1	splice_site variant	*TMEM126A*	splice_site variant
SMC6	splice_site variant	*CCDC82*	splice_site variant
QPCT	splice_site variant	*POLE*	splice_site variant
TMEM127	splice_site variant	*NDUFA9*	splice_site variant
ERBB4	splice_site variant	*RERG*	splice_site variant
ABCA12	splice_site variant	*PPFIA2*	splice_site variant
EIF2AK3	splice_site variant	*ATP2A2*	splice_site variant
ZAP70	splice_site variant	*FKBP4*	splice_site variant
BMPR2	splice_site variant	*CD4*	splice_site variant
KALRN	splice_site variant	*SLC46A3*	splice_site variant
ROPN1B	splice_site variant	*UPF3A*	inframe deletion
GLB1	splice_site variant	*NID2*	splice_site variant
C3orf14	splice_site variant	*ATG2B*	splice_site variant
RFC1	stop lost	*SYT16*	splice_site variant
AGPAT9	splice_site variant	*NEO1*	splice_site variant
TLR2	stop gained	*GPR176*	splice_site variant
GABRA2	splice_site variant	*SEZ6L2*	splice_site variant
TRPC3	splice_site variant	*HSD17B2*	splice_site variant
MDN1	splice_site variant	*VPS4A*	splice_site variant
AIG1	splice_site variant	*TRPV3*	splice_site variant
TTBK1	splice_site variant	*ALOXE3*	splice_site variant
RUNX2	splice_site variant	*SLC26A11*	splice_site variant
CRISP3	splice_site variant	*ASPA*	splice_site variant
ABCB4	splice_site variant	*EIF3G*	splice_site variant
TMEM168	splice_site variant	*CSNK1G2*	splice_site variant
FAM71F1	splice_site variant	*LZTR1*	splice_site variant
C7orf49	inframe insertion	*SREBF2*	splice_site variant
SDR16C5	inframe deletion	*TTLL8*	splice_site variant
E2F5	splice_site variant	*BAIAP2L2*	splice_site variant
DCSTAMP	splice_site variant	*GRIPAP1*	splice_site variant
SHARPIN	splice_site variant	*BMX*	inframe deletion
COL15A1	splice_site variant	*ASMT*	splice_site variant

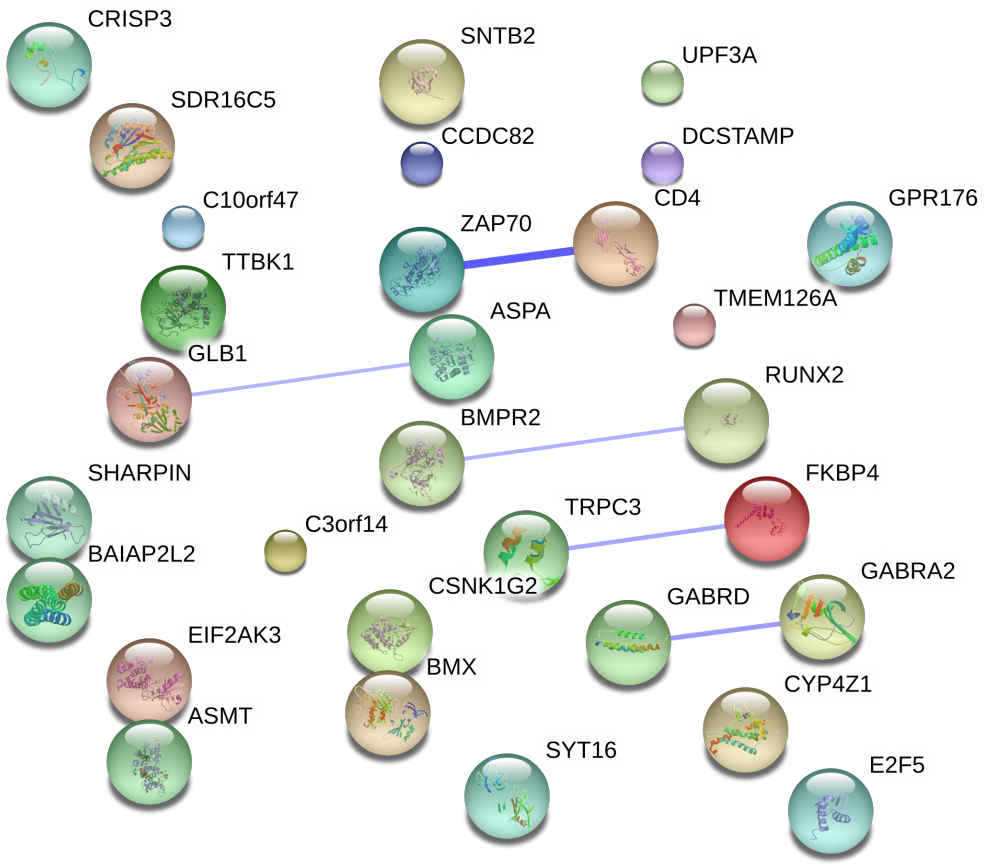


Figure 3.9: Case only LoF direct protein-protein interaction network. Five interaction networks for among 30 genes with rare and novel LoF mutations identified in cases only. Genes are represented by nodes while blue lines show direct interactions between the two genes. The thickness is scaled to illustrate level of confidence.

3.2.3.2 Rare protein altering mutations

This analysis included all predicted functional coding mutations, either LoF or missense mutations predicted as deleterious by at least one prediction algorithm, with an allele frequency of less than 1% in ExAC and 1000 Genomes. There were a total of 23,459 qualifying variants in 10,814 genes. Two strategies were adopted for analysis. In the first instance, testing for a burden of rare deleterious mutations in cases against controls, a one-sided burden test for each of the over 10,828 genes with a qualifying variant was performed using the program PLINK/SEQ (described in Materials and Methods section 2.7.3.1). Unsurprisingly, due to the small number of cases, burden analysis did not reveal any gene with a significant excess of rare deleterious alleles after Bonferroni correction. The most highly associated gene *MFF*, encoding a mitochondrial fission factor, had a permutation derived p-value of 1.24×10^{-3} (Figure 3.8).

Due to pre-existing evidence (including results from this project described in section 3.2.2.1) highlighting the role of complement genes in susceptibility to meningococcal disease, a comprehensive analysis of the burden of mutations in genes associated with the complement system was next performed. Of 200 complement-associated genes, controls had mutations in 26 genes. Only four genes had a mutation in at least one case individual, the novel hemizygous mutation in *CFP*, heterozygous mutations in *GPLD1*, *KRT1* and *C1QB*. Together, these did not represent an excess over that expected from 10 million permutations of case-control status ($P_{corrected} = 0.51$). Individually, the *CFP* mutation had been functionally validated (section 3.2.2.1). The *GPLD1* and *KRT1* genes were noted to contain the same rare mutations in controls as well. The deleterious mutation in *C1QB* has been reported to show no difference in allele frequency in 122 meningococcal meningitis cases compared to 302 controls [290].

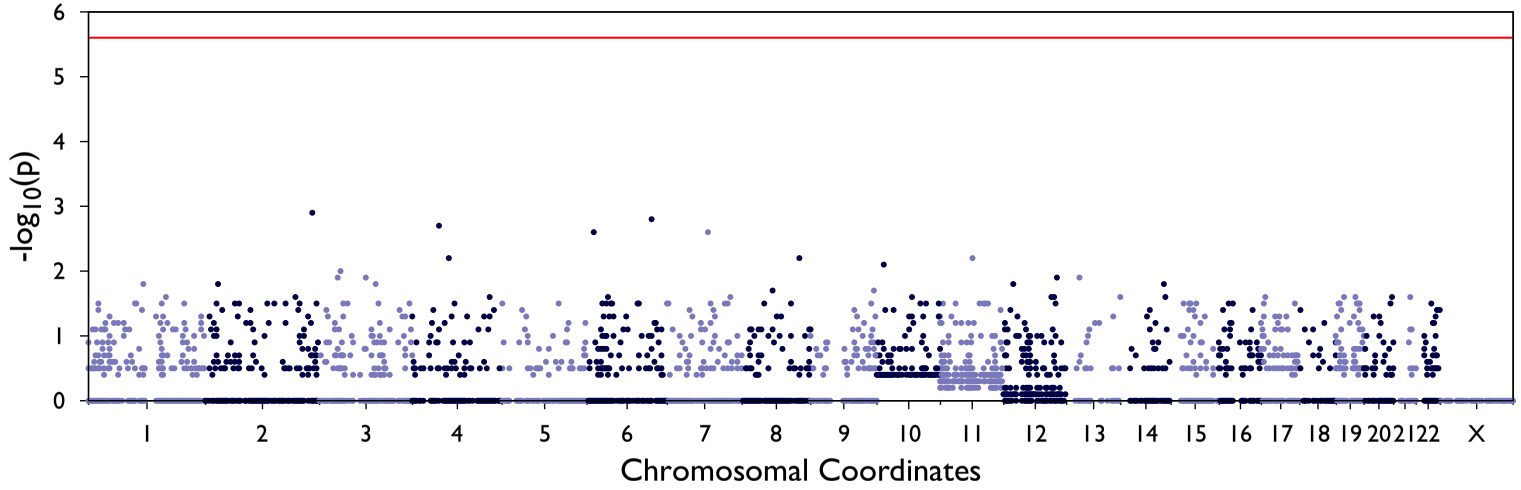


Figure 3.10: Manhattan plot for gene burden analysis. Association analysis for an excess of deleterious mutations in meningococcal cases. The *MFF* gene had strongest signal P=1.24x10^{-3}. The red line indicates exome-wide significance threshold of 2.5x10^{-6}.

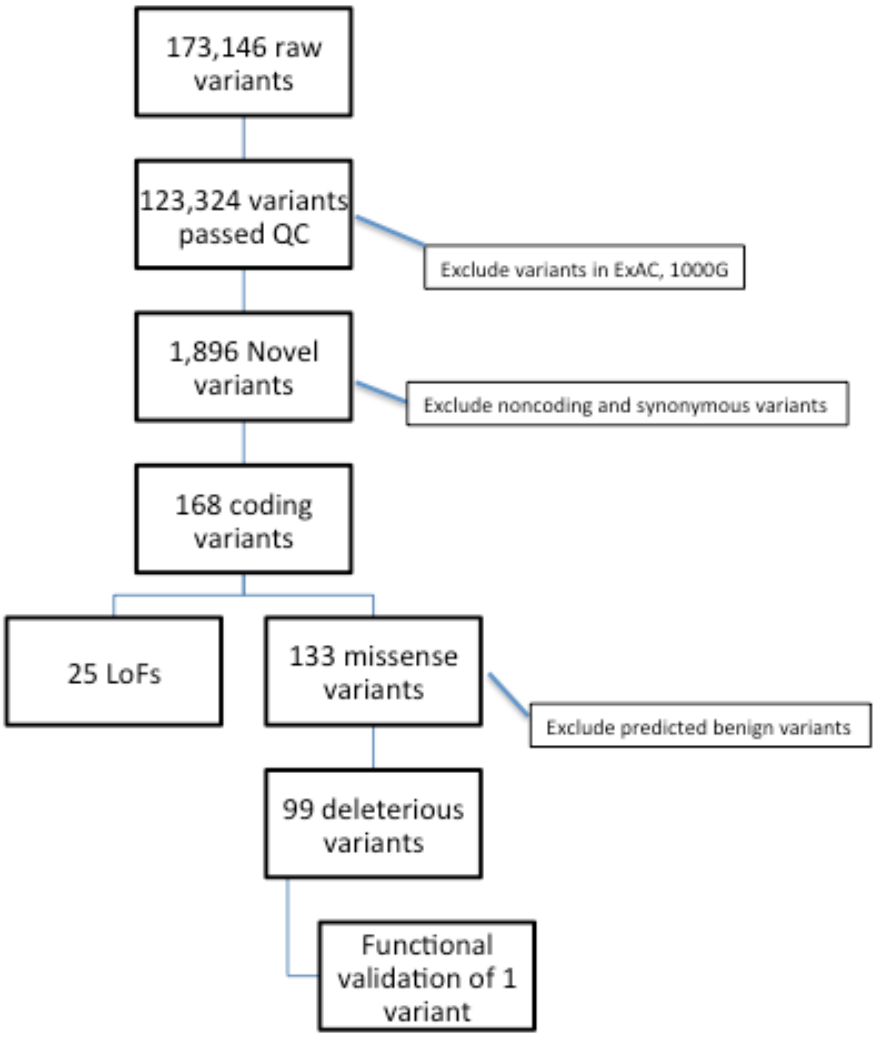


Figure 3.11: Schematic of filtration and prioritisation of novel variants in meningococcal case. This process prioritised variants in meningococcal cases only based on novelty and predicted biological function. Functional follow-up of one missense variant was based on biologicall plausability due to previously reported cases

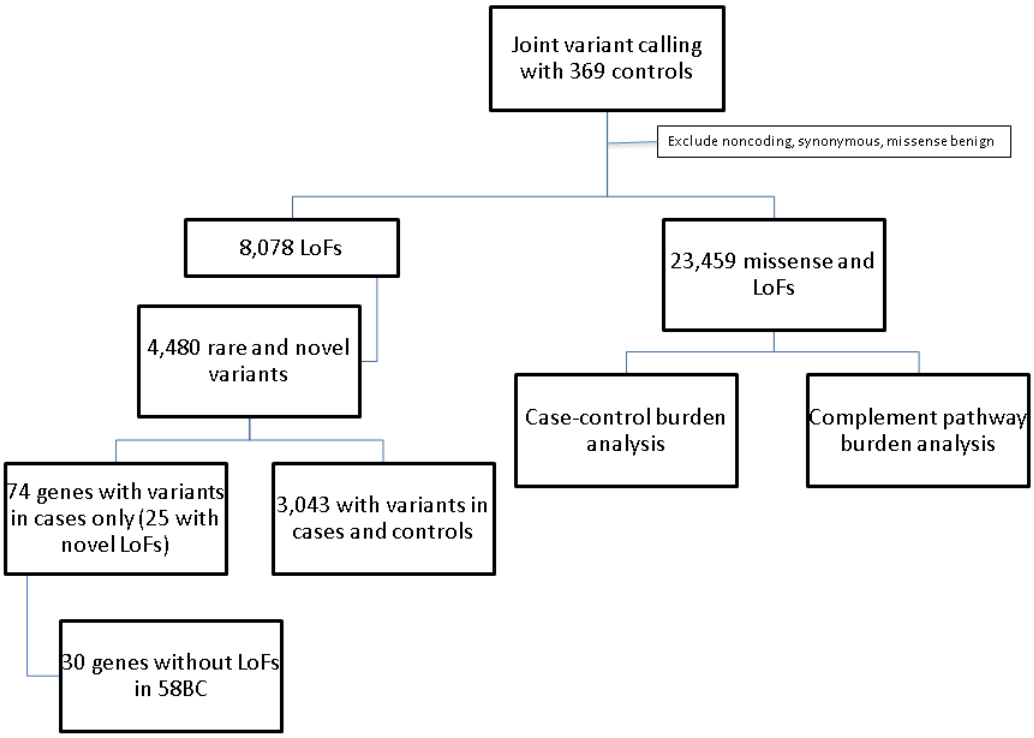


Figure 3.12: Case control variant filtration. The burden of deleterious mutations in all genes and in complement pathway genes was compared between cases and controls. Candidate genes with loss of function mutations were further prioritised by comparison with exomes in the UK10K project and 1958 birth cohort

3.3 Discussion

These experiments aimed to identify the role of rare large effect genetic variants predisposing to meningococcal disease using next generation sequencing methods in a cohort of adult individuals presenting to UK intensive care units with meningococcal disease.

This study identified a novel, functional mutation in the *CFP* gene encoding the complement factor properdin as a potentially causal variant for meningococcal disease. Properdin is the only known positive regulator of the complement pathway. It serves to stabilise the alternative pathway C3-convertase enzyme allowing deposition of the potent opsonin C3b on microbial cell surfaces, formation of the membrane attack complex (MAC) as well as amplification of the complement response to the pathogen via formation of more C3b molecules [291]. Three types of X-linked properdin deficiency have been described: type I leads to complete absence of properdin in serum, type II results in partial deficiency where there is detectable properdin in serum but at only 10% of normal concentrations, and type III is characterised by normal levels of dysfunctional properdin in serum [286, 292]. Mutations leading to any of the three types of deficiencies have been described that lead to a 250-fold increase in risk of meningococcal disease [286]. In this project, a novel missense mutation was identified in a male individual and subsequently shown to cause a type I properdin deficiency, with preserved intracellular production but apparently defective extracellular secretion. Meningococcal disease in properdin-deficient individuals is typically characterised by a late age-of-onset and infection with uncommon serogroups [293]. This particular individual was 19 years old at the time of this episode of infection; information on the meningococcal serogroup was not available. The lack of secreted properdin and presence of intracellular protein in the mutant form is most likely explained by protein misfolding leading to defective secretion. The structure of the properdin molecule has two highly conserved arginine residues, six highly conserved cysteine residues and three highly conserved tryptophan residues [294]. In serum, properdin exists as dimers, trimers and tetramers in varying ratios. The highly conserved residues as well as the disulphide bonds are essential in the formation of these secondary structures. The

R346C mutation replaces one of the conserved arginine resides with a cysteine. This likely results in formation of other disulphide bonds within the higher order oligomers and in is highly likely to lead to a compromised structure during protein folding as has been observed with CFH [294–296]. This phenomenon has been demonstrated for other single gene disorders such as cystic fibrosis due to the *CFTR* F508del variant, α1-antitrypsin deficiency, and partial IFN-γR2 deficiency [297, 298].

The role of mutations with predicted loss of protein function was also investigated in the context of meningococcal disease susceptibility. Loss of function mutations are generally maintained at low frequency by natural selection. The apparent enrichment of rare LoFs in immune-related genes in meningococcal disease presents an exciting opportunity for further follow-up. This however should be interpreted cautiously due to the small number of individuals studied and the striking depletion of genuine loss of function mutations in the human genome [299]. Immunological genes have recently been reported to be highly tolerant to functional genetic variation [300], however, lack of LoFs in 1,369 controls from Generation Scotland and 58BC represents evidence of constraint against highly deleterious genetic variation in these 30 genes. Two genes are of particular interest. Firstly, *ZAP70*, encoding a cytoplasmic kinase involved in antigen response in T cells. Homozygous and compound heterozygous mutations in this gene lead to severe combined immunodeficiency (SCID), with parents and unaffected siblings being heterozygous [301]. One individual with meningococcal disease was found to be heterozygous for a rare splice region variant. The role of this gene in meningococcal disease has not been investigated before. While SCID patients are susceptible to a wide range of bacterial and viral infections, literature on meningococcal infections in SCID patients is lacking. Secondly *CD4*, encoding a glycoprotein expressed on lymphocytes acts a co-receptor for T-cell binding to antigen presenting cells such as MHC class II. Individuals deficient in CD4-expressing T-cells are susceptible to a wide range of infections as is evident in the context of CD4 T cell depletion in HIV infections [302]. Partial defects in these two genes could result in impaired cellular responses and an increased susceptibility to bacterial pathogens. However further functional work is needed for robust conclusions.

This study did not identify an excessive burden of rare deleterious mutations in any gene or more specifically in the complement pathway in meningococcal cases compared to controls. Due to the small numbers of individuals studied, this is perhaps unsurprising as there would be limited power to detect such an association. More importantly, although the complement pathway as a whole was not enriched for deleterious variants, proof of principle for a role of individual mutations of large effect underlying severe infectious disease is provided by the identification and validation of a single mutation in a single individual predisposing to disease.

While compelling evidence is presented in this chapter, potential confounders exist. Data on other risk factors such as the pathogenicity of the infecting meningococcus strain, the context of the infection-whether during endemic or epidemic seasons- or any previous and subsequent episodes of disease are unavailable and could confound analyses.

In conclusion, this research demonstrates the potential of exome sequencing to identify novel, large effect variants in severe phenotypes of infectious disease, even in the setting of small sample sizes. Significantly larger sample sizes would be required however to fully define the role of other potentially pathogenic mutations, such as the *CD4* and *ZAP70* mutations. Complement deficiency is usually only suspected in multi-case families or a history of recurrent episodes of meningococcal disease. Indeed, one individual was excluded from whole-exome sequencing because of a history of recurrent meningococcal disease and an established diagnosis of complement component C7 deficiency. In this study of adults with meningococcal sepsis, 2 out of 14 individuals therefore have evidence of deleterious mutations in complement pathway components. Recently, a four component meningococcal serogroup B vaccine (4CMenB) has been introduced into routine immunisation in the UK, a key component in this vaccine is the meningococcal factor H binding protein, used by the pathogen to evade complement mediated killing by subverting complement factor H. In light of this, coupled to previous evidence on complement deficiencies and findings from this study, the key role of complement in host defence against meningococci continues to be highlighted,

raising the possibility that a significant fraction of adults presenting with invasive meningococcal disease may harbour an underlying selective primary immunodeficiency.

Chapter 4
Genetic susceptibility to sepsis

4.1 Introduction

There is strong evidence implicating genetic factors in the susceptibility and outcome of sepsis [219]. The risk of death from infection was observed to increase over five-fold for adoptees whose biological parent died prematurely from infectious disease [219]. The human genome epidemiology (HuGE) navigator database lists 308 genes associated with sepsis [303]. Currently, the majority of studies have primarily employed a candidate gene approach, whereby genes are selected based on a known or suspected role in the pathogenesis of sepsis [21, 304]. Candidate gene studies on cytokines such as *IL17A*, *TNF-α* and *IL-6* as well as components of the innate immune system such as *MBL*, *CD14* and *TLR4* have been reported [305–312] . However most of these studies have produced inconsistent results with poor replication in independent studies [304,313]. This has been partly attributed to the heterogeneous nature of sepsis syndromes and poorly defined patient phenotypes [304]. New, high throughput genetic methods have the ability to greatly improve our understanding of sepsis. Recently a GWAS on sepsis identified *FER* as a novel locus with common polymorphisms influencing sepsis outcomes [244]. However, population based exome- or genome-wide studies on the role of large effect rare variants have not been reported. This study aims to identify host genetic factors influencing susceptibility to sepsis. The case cohort comprised of 270 individuals aged over 18 years recruited from ICU's across the UK with sepsis arising from community acquired pneumonia. Definitions of

sepsis were based on the International Consensus Criteria [5]. Over the course of this study, three cohorts ALSPAC (whole genome sequencing data) [246], UK10K-OBESITY-GS [249] and 58BC [141, 250] were used as controls representative of the UK population for rare variant analysis. Further details on these cohorts are described in Materials and Methods section 2.2.

4.2 Results

4.2.1 Sequencing metrics

Exome capture for the 270 samples was achieved using the Nimblegen SeqCap EZ v3.0 (Roche Life Sciences) capture kit. Of the 64MB region targeted, 92% were captured at 10X or more while 87% of all targeted regions were covered at greater than 20X (Table 4.1). The average duplication rate across all samples was 15%. Overall, the sequencing achieved high coverage with the average depth of sequencing across all samples being 65X (Table 4.1 and Figure 4.1).

Table 4.1: Summary of sequence metrics Whole exome sequencing metrics for 270 sepsis samples. The proportions of the 64MB targeted region covered at various depths are shown

Sequencing metrics	**CAP270**
Size of target region	64,190,747
Mean depth of target region	65.4X
Percentage of target bases covered >= 10x (%)	91.6
Percentage of target bases covered >= 20x (%)	86.5
Percentage of target bases covered >= 30x (%)	77.9
Average duplication rate (%)	14.9

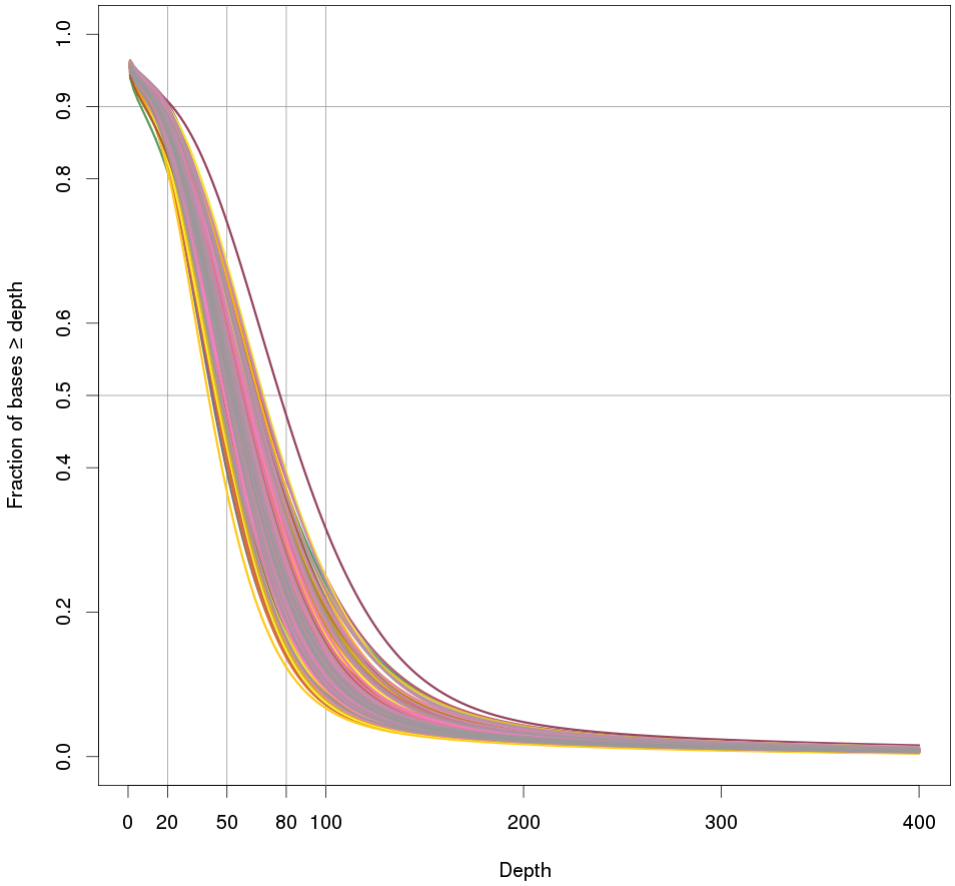


Figure 4.1: CAP sequencing coverage. Exome sequencing coverage in 270 CAP samples. The lines show the proportion of the 64MB targeted region covered at various depths for each individual. All samples had a mean coverage greater than 49X with an average of 65X across all samples.

4.2.2 Analysis with ALPSAC controls

4.2.2.1 Population structure

Multi-dimensional scaling (MDS) as implemented in PLINK was used to assess for underlying substructure in the 270 sepsis cases and 1,927 ALSPAC controls. For estimates of genetic ancestry, clustering of the top dimensions was performed incorporating dimensions from individuals in the HapMap3 dataset. Sixty individuals in ALSPAC were identified as having a genetic ancestry distinct from other European populations on the two most informative MDS components and were subsequently excluded from all further analysis as population outliers (Figure 4.2).

4.2.2.2 Association analysis

4.2.2.2.1 Single variant analysis In order to account for underlying cryptic relatedness not captured in MDS, mixed model association analysis was carried out using EMMAX as implemented in the statistical toolkit EPACTS. A kinship matrix calculated from common polymorphisms was applied in the mixed model. Significant deflation of test statistic with λ of 0.56 (Figure 4.3a) was observed. To exclude overcorrection due to the kinship matrix, reanalysis was performed using a Fisher's exact test. By contrast, significant inflation of test statistic (λ=1.16) was observed, indicating underlying confounding in the analysis (Figure 4.3b). Moreover, discrepancies were observed in allele frequencies in ALPSAC genotypes and European populations in 1000 Genomes. This could potentially have been due to the much lower (6X) sequencing depth in the controls. Due to this finding, in addition to unexplained deflation of test statistic in the mixed model analysis, ALSPAC whole genome data were deemed unsuitable for this analysis.

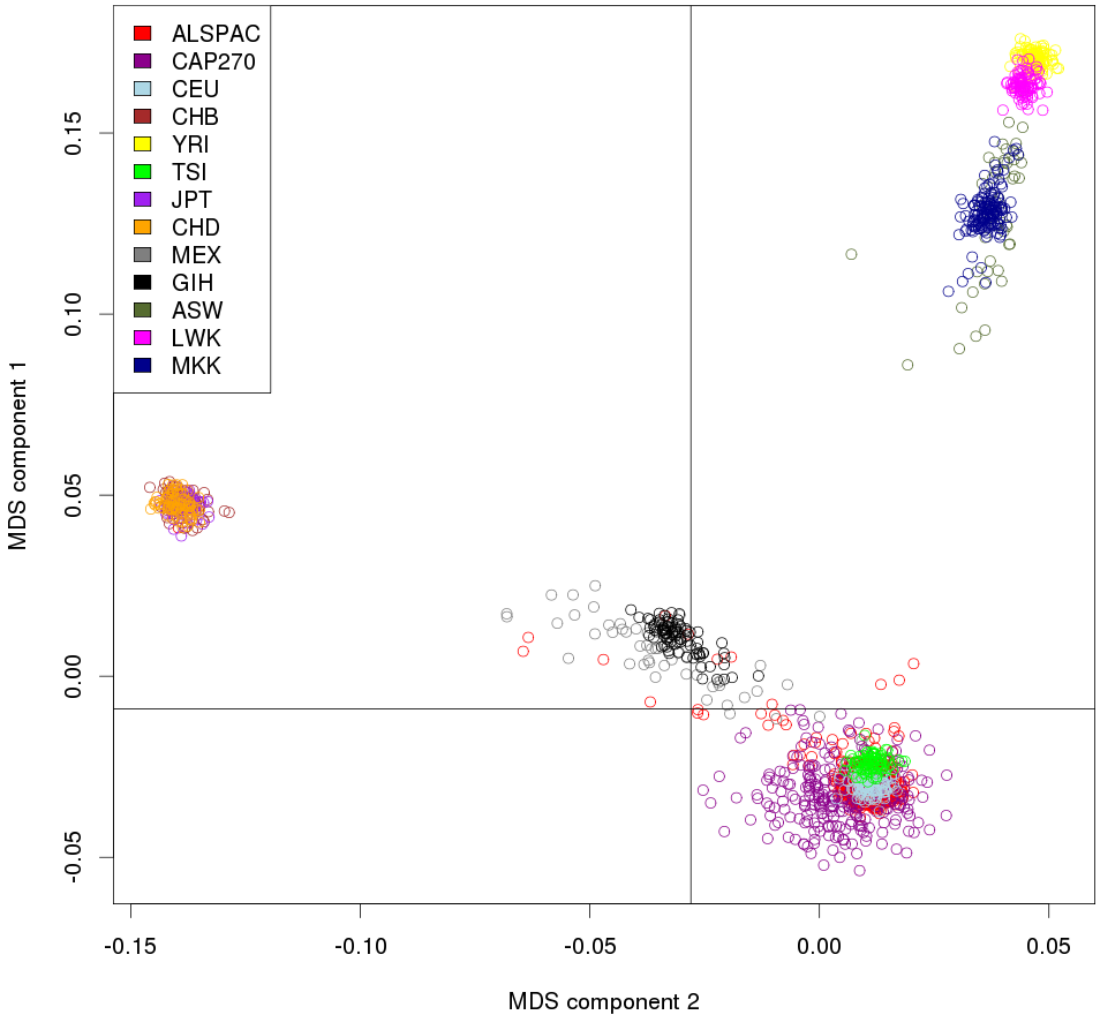
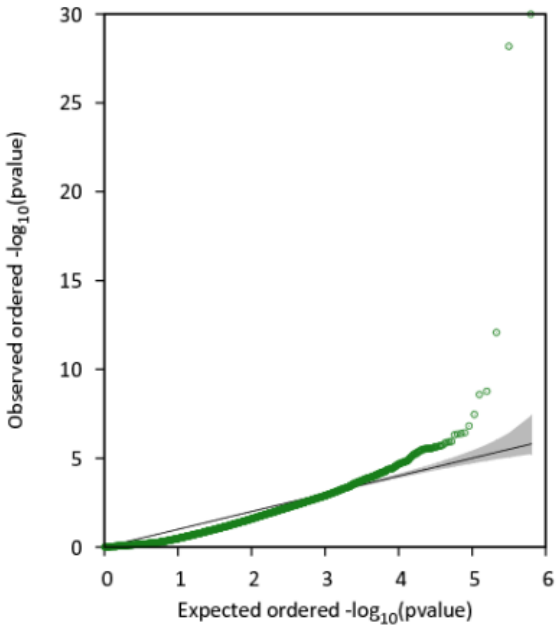
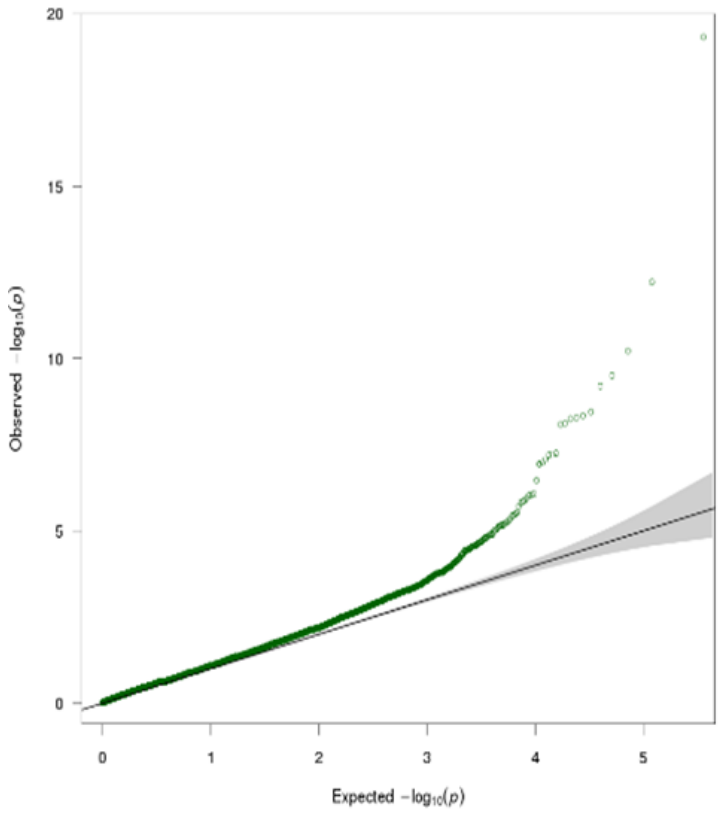


Figure 4.2: ALSPAC-CAP population structure analysis. Multi-Dimensional Scaling(MDS) of the first two components comparing MDS components for populations in HapMap3 with study samples. The two axes represent genetic similarity within and between human populations. Individuals show the greatest genetic similarity within continental populations. Sixty ALSPAC controls did not cluster with the European populations in HapMap3. **ALSPAC** ALSPAC population controls, **CAP270** CAP sepsis cases, **CEU** Utah residents with Northern and Western European ancestry from the CEPH collection, **CHB** Han Chinese from Beijing, China, **YRI** Yoruba in Ibadan, Nigeria, **TSI** Tuscani in Italia, **JPT** Japanese in Tokyo, Japan, **CHD** Chinese in Metropolitan Denver, **MXL** Mexican ancestry in Los Angeles, **GIH** Gujarati Indians in Houston, Texas, **ASW** African ancestry from SouthWest USA, Colorado, **LWK** Luhya in Webuye, Kenya, California, **MKK** Maasai in Kinyawa, Kenya



(a) EMMAX (λ =0.56)



(b) Fisher's exact(λ =1.16)

Figure 4.3: Single variant analysis. Association analysis using (a) mixed model and (b) Fisher's exact . Deflation of test statistic was observed with the mixed model while significant inflation was evident using the Fisher's exact test.

4.2.3 Analysis with Generation Scotland: Scottish Family Health Study (UK10K-OBESITY-GS) controls

4.2.3.1 Population structure

MDS analysis was performed in PLINK for 270 sepsis cases and 369 unrelated individuals from the Generation Scotland (GS) study on genetics of obesity [249]. MDS components were clustered together with HapMap samples for genetic ancestry estimation. Two cases and 18 controls were identified as not clustering with the primary European populations in HapMap and were excluded from further analysis (Figure 4.4).

4.2.3.2 Fine scale population structure

Due to the controls being of primarily Scottish ancestry, analysis was undertaken to determine whether confounding due to fine scale structure was adequately addressed by a kinship matrix in the linear mixed model. Single variant association analysis was performed using EMMAX for the 486,812 variants that passed quality control. Overall, genomic control had no evidence of inflation with λ of 1.02. Of note however, the *LCT* gene encoding the enzyme lactase dehydrogenase had two common variants with p-values of 8.5 x10^{-3} and 9.6 x10^{-3} respectively (Figure 4.5a). The *LCT* locus has been identified as a target of recent selection; large allele frequency differences at this locus exist even with geographically close populations [141, 314]. Although not reaching formal statistical significance, the presence of fine scale structure not adequately corrected by the mixed model was of concern as this may lead to spurious association signal [315]. To further correct for residual substructure, the top 10 principal components calculated using the program EIGENSTRAT were included as covariates in the EMMAX test [315]. Inclusion of principal components in addition to the kinship matrix had no effect on lambda which remained at 1.02, however, there was no residual signal at the *LCT* locus, with the most significant variant having a p-value of 0.03 (Figure 4.5b).

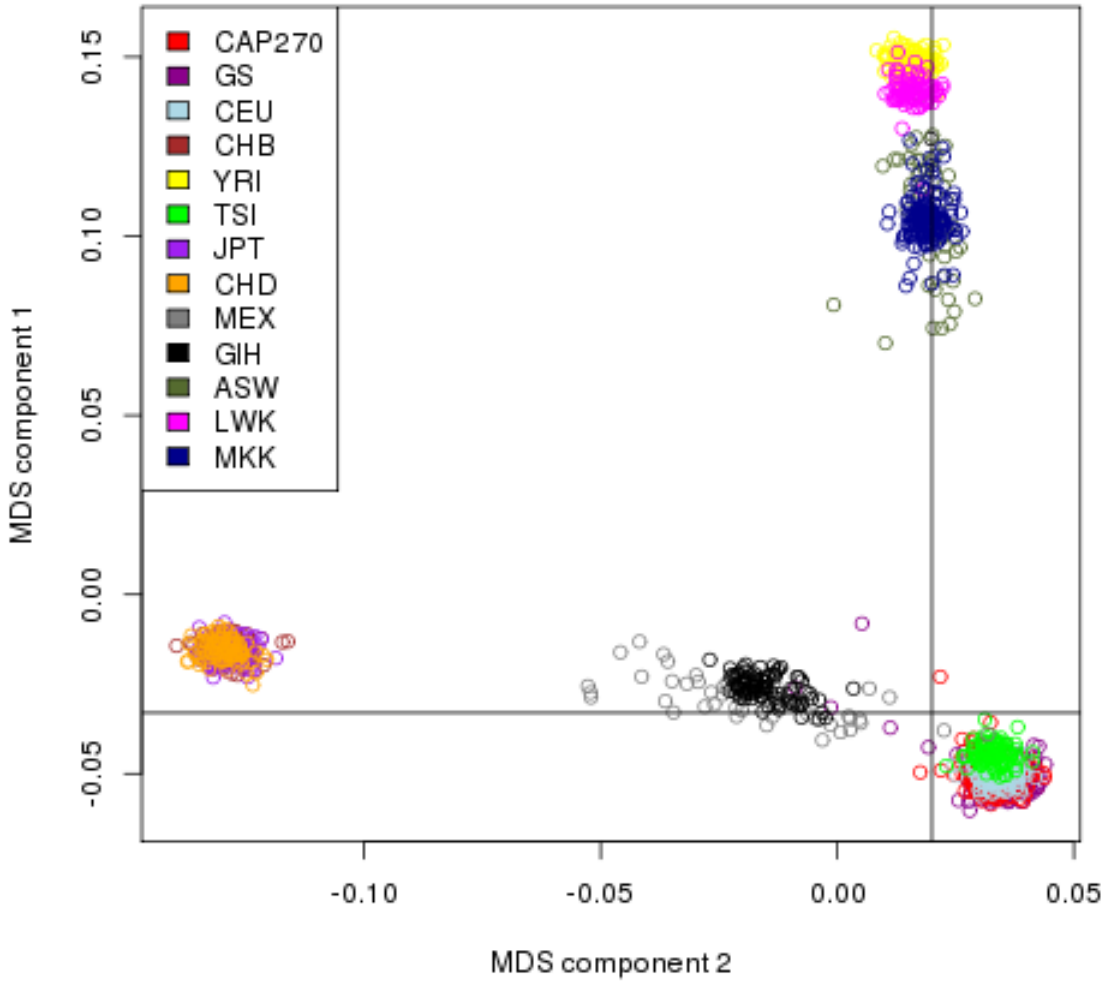


Figure 4.4: Population structure analysis. Multi-Dimensional Scaling(MDS) of the first two components comparing MDS components for populations in HapMap3 with study samples. The two axes represent genetic similarity within and between human populations. Individuals show the greatest genetic similarity within continental populations Two cases and 18 controls did not cluster with the European populations in HapMap3. **GS** Generation Scotland obesity controls, **CAP270** CAP sepsis cases, **CEU** Utah residents with Northern and Western European ancestry from the CEPH collection, **CHB** Han Chinese from Beijing, China, **YRI** Yoruba in Ibadan, Nigeria, **TSI** Tuscani in Italia, **JPT** Japanese in Tokyo, Japan, **CHD** Chinese in Metropolitan Denver, **MXL** Mexican ancestry in Los Angeles, **GIH** Gujarati Indians in Houston, Texas, **ASW** African ancestry from SouthWest USA, Colorado, **LWK** Luhya in Webuye, Kenya, California, **MKK** Maasai in Kinyawa, Kenya

4.2.3.3 Single variant association analysis

Subsequent analysis with the Scottish controls included the top 10 principal components in the mixed model. In the analysis of all single variants for an association with sepsis susceptibility, there were no variants that achieved statistical significance for the 486,812 variants tested. However, 42 variants had suggestive evidence of an association with sepsis susceptibility ($P<1.0\text{x}10^{-4}$, Table 4.2). The most significantly associated variant was a common missense polymorphism in the *DFNB31* gene with $P=1.52\text{x}10^{-7}$. The p.Met613Thr polymorphism was predicted as benign by both PolyPhen-2 and SIFT. A second signal - although not reaching statistical significance - was in the *OR9G1* gene with $P=1.81\text{x}10^{-7}$. However comparison of allele frequencies at this locus with non-Finnish Europeans in the ExAC database [262] revealed a discrepancy. The minor allele was rare in the Scottish controls, with an allele frequency of 2% compared to 50% in the ExAC database, an indication of possible genotyping errors at this variant in Scottish controls.

4.2.3.4 Burden analysis

Burden analysis aggregated classes of potentially functional variants within each gene and tested for an excess of deleterious variants in cases compared to controls using SKAT test [190, 208–210]. Two classes of variants were tested: highly disruptive loss of function mutations (LoFs; stop gain or loss, splice site, frameshift and inframe insertions and deletions) and all protein altering mutations (missense mutations predicted as deleterious by either PolyPhen-2 or SIFT and all LoFs). Two frequency thresholds were used: 1% to delimit rare variants and 5% for joint analysis of rare and low frequency variants [154, 195–197]. Collapsing rare protein altering mutations alone, the strongest association was observed in the *NECAP2* gene with borderline significance of $P=2.59\text{x}10^{-6}$ (Table 4.3). This gene had one rare deleterious missense mutation present in five cases but absent in controls. In the joint collapsing of rare and low frequency variants (<5%), the *SPATC1L* gene had the strongest association with sepsis susceptibility though not to statistical significance ($P=2.14\text{x}10^{-5}$). In

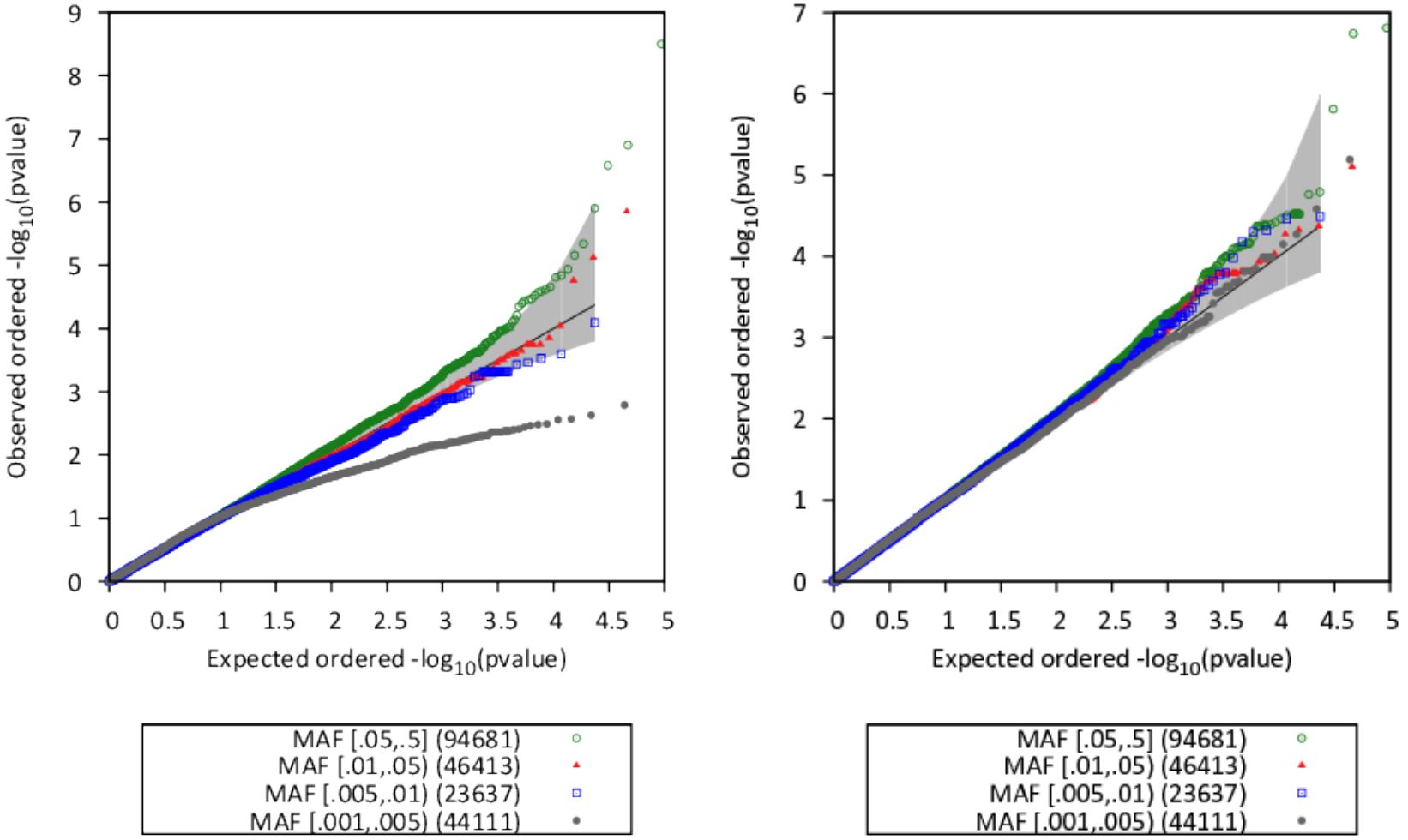
total, there were 22 genes with a suggestive association for either rare variants alone or jointly aggregating rare and low frequency variants with sepsis susceptibility (P<1.0x10^{-3}, Table 4.3). The program STRING was used to explore underlying biological interactions within this set of genes. There were no protein-protein interactions amongst the gene products; additionally there was no over-representation of genes from any KEGG pathways or enrichment of genes with the same GO annotation. A separate analysis was undertaken to investigate the role of highly disruptive mutations in sepsis susceptibility by aggregating LoF mutations only. There were no genes that achieved exome-wide significance. Seven genes however had suggestive evidence with P<1.0x10^{-4} (Table 4.4). In the joint analysis of rare and low frequency variants, the same seven genes as for rare variants had the strongest evidence of association. None achieved statistical significance.

Table 4.2: Single variant analysis. Association statistics for sepsis susceptibility using Scottish controls from the Generation Scotland study. The top 10 principal components were included in this analysis to control for residual fine scale structure. Association statistics in this table are single nucleotide variants with a P_{EMMAX}<1.00x10^{-4}. Also shown are the allelic counts of cases and controls (Case count and Control count respectively). There were no variants that significantly associated with sepsis susceptibility (P_{EMMAX}<1.02x10^{-7})

Gene	**Pos**	**Allele**	**Variant type**	P_{EMMAX}	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
DFNB31	9:117169033	A/G	Missense	1.53x10^{-7}	68/140/43	67/172/104
OR9G1	11:56468554	A/C	Missense	1.81x10^{-7}	212/50/2	329/13/1
EPB41L5	2:120889173	G/C	Intron	1.54x10^{-6}	16/152/53	31/128/190
MOV10	1:113243309	G/A	Utr3	6.41x10^{-6}	265/3/0	354/0/0
DAB2IP	9:124522313	C/T	Synonymous	7.79x10^{-6}	238/29/0	337/17/0
IGFL1	19:46733916	G/A	Intron	1.62x10^{-5}	21/118/127	48/180/116
RAB35	12:120546145	C/G	Intron	1.72x10^{-5}	197/60/3	309/45/0
ARHGEF2	1:155932492	C/T	Synonymous	1.99x10^{-5}	266/2/0	354/0/0
KIAA0182	16:85706073	C/G	Missense	1.99x10^{-5}	266/2/0	354/0/0
NECAP2	1:16785378	A/T	Missense	2.61x10^{-5}	263/5/0	354/0/0
RP11-191L9.6	22:48023221	G/A	Non-coding transcrip	2.64x10^{-5}	266/2/0	354/0/0
ARCN1	11:118471423	T/C	Synonymous	2.99x10^{-5}	194/66/8	284/69/1
PEAR1	1:156875037	T/G	Intron	3.02x10^{-5}	170/54/0	192/108/15
IGFL1	19:46733912	A/G	Intron	3.11x10^{-5}	21/118/127	48/180/118
ADHFE1	8:67365821	G/C	Intron	3.16x10^{-5}	249/7/0	351/2/0
IFT46	11:118432279	A/G	Intron	3.44x10^{-5}	194/67/7	284/69/1
RRS1	8:67342257	G/A	Synonymous	3.46x10^{-5}	259/8/0	352/2/0
ARCN1	11:118444525	G/A	Intron	3.73x10^{-5}	192/66/8	285/67/2
TMEM25	11:118405068	A/G	Missense	4.04x10^{-5}	175/66/7	277/72/2
IFT46	11:118416253	C/T	Intron	4.09x10^{-5}	190/68/8	280/72/2
ARCN1	11:118451901	T/C	Intron	4.16x10^{-5}	192/66/8	284/68/2
SMG7	1:183486859	G/A	Synonymous	4.17x10^{-5}	1/34/233	6/90/258
SLC15A2	3:121659202	G/A	Intron	4.20x10^{-5}	249/18/1	344/9/1
RAB35	12:120546136	G/A	Intron	4.20x10^{-5}	201/58/3	310/44/0
RQCD1	2:219447912	A/G	Intron	4.72x10^{-5}	0/16/186	0/0/353
DOT1L	19:2210778	G/A	Synonymous	4.78x10^{-5}	268/0/0	346/8/0
PIP5K1C	19:3667450	G/A	Intron	4.92x10^{-5}	234/6/0	342/1/0
DFNA5	7:24756765	C/T	Intron	5.27x10^{-5}	264/4/0	339/0/0
VPRBP	3:51472033	A/G	Intron	5.32x10^{-5}	267/1/0	301/14/0
IFT46	11:118430519	C/T	Missense	5.67x10^{-5}	193/66/9	280/72/2
TLE1	9:84199010	G/A	Utr3	5.93x10^{-5}	265/2/0	354/0/0
PLD1	3:171417632	C/G	Intron	6.48x10^{-5}	268/0/0	343/11/0

Continued on next page...

Gene	**Pos**	**Allele**	**Variant type**	P_{EMMAX}	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
IFT46	11:118425063	A/G	Intron	6.82×10^{-5}	191/67/9	277/71/3
IFT46	11:118422868	C/T	Intron	6.84×10^{-5}	193/67/8	280/72/2
PDIA4	7:148709346	T/C	Missense	7.05×10^{-5}	265/3/0	354/0/0
IFT46	11:118416073	A/C	Intron	7.15×10^{-5}	191/67/7	280/72/2
IFT46	11:118425419	G/T	Intron	7.35×10^{-5}	191/67/8	279/72/2
ALDH3B2	11:67431038	G/A	Intron	7.42×10^{-5}	0/38/229	3/72/247
AKR1C4	10:5254820	A/C	Intron	7.62×10^{-5}	204/55/6	167/101/4
DFNB31	9:117166246	A/G	Missense	7.92×10^{-5}	80/135/42	89/183/82
DENND5B	12:31542191	A/G	Intron	7.96×10^{-5}	51/108/42	140/171/43
PYCR1	17:79893270	C/T	Synonymous	9.20×10^{-5}	265/2/0	354/0/0
FETUB	3:186370454	C/T	Utr3	9.31×10^{-5}	223/17/1	342/12/0
HJURP	2:234749151	C/T	Intron	9.31×10^{-5}	49/134/78	88/187/78
IFT46	11:118425051	G/A	Intron	9.35×10^{-5}	192/68/7	269/69/2
FRAS1	4:79366891	G/A	Missense	9.63×10^{-5}	266/2/0	354/0/0
TMEM25	11:118405377	C/A	Utr3	9.78×10^{-5}	190/69/6	270/70/1
BMP2K	4:79754932	C/T	Intron	9.81×10^{-5}	266/2/0	352/0/0
SPTLC2	14:78036658	C/G	Intron	9.89×10^{-5}	162/67/3	296/56/2



(a) EMMAX without 10 PCs

(b) EMMAX with 10 PCs as covariates

Figure 4.5: Q-Q plots for fine scale population structure analysis. Single variant analysis using a mixed model (a) without additional principal components and (b) with the top 10 principal components as covariates. Residual signal at LCT locus corrected in addition to correction of deflated statistics for rare variants

Table 4.3: Sepsis gene-based analysis. Genes with suggestive evidence of an association with sepsis (P_{SKAT} <1.0x10^{-3}), comparing sepsis cases against Scottish controls. Rare variants(<1% MAF[1]) were analysed separately as well as jointly with low frequency variants (1-5%)

		Rare (less than 1%)		Rare & low frequency (less than 5%)	
	Gene	**No. of variants**	**P_{SKAT}**	**No. of variants**	***P_{SKAT}***
Pra[2]	*NECAP2*	1	2.59x10^{-6}	2	5.61x10^{-3}
	RNF112	8	9.54x10^{-6}	8	5.63x10^{-5}
	MBD6	8	6.01x10^{-5}	8	2.48x10^{-5}
	RPS6KB2	6	6.21x10^{-5}	6	8.92x10^{-5}
	KIAA1033	9	7.68x10^{-5}	11	9.50x10^{-3}
	SPATC1L	1	7.75x10^{-5}	1	2.14x10^{-5}
	MYRFL	3	2.86x10^{-4}	3	2.40x10^{-4}
	OR11H4	5	4.12x10^{-4}	5	3.34x10^{-4}
	COL14A1	18	5.12x10^{-4}	20	1.54x10^{-3}
	ANXA11	8	5.46x10^{-4}	8	7.36x10^{-4}
	TCP11L1	1	5.54x10^{-4}	1	7.00x10^{-4}
	CDK1	1	5.90x10^{-4}	1	6.48x10^{-4}
	HS3ST2	1	5.99x10^{-4}	1	6.18x10^{-4}
	SLC23A3	1	5.99x10^{-4}	1	6.38x10^{-4}
	CANX	1	6.38x10^{-4}	2	0.14
	SECISBP2L	6	6.53x10^{-4}	6	9.12x10^{-4}
	MT1B	3	7.01x10^{-4}	3	6.07x10^{-4}
	HSPB3	1	7.10x10^{-4}	1	6.28x10^{-4}
	VCAM1	4	8.28x10^{-4}	4	7.67x10^{-4}
	PUSL1	2	8.83x10^{-4}	2	8.88x10^{-4}
	PEX2	2	9.55x10^{-4}	3	0.39
	E2F7	2	0.76	3	7.41x10^{-4}
LoFs[3]	*MYRFL*	1	1.22x10^{-4}	1	6.10E-5
	ADAM15	1	1.74x10^{-4}	1	2.52x10^{-4}
	ADH5	1	6.18x10^{-4}	1	6.28x10^{-4}
	TRMT2B	1	6.18x10^{-4}	1	7.21x10^{-4}
	SLC23A3	1	6.68x10^{-4}	1	5.99x10^{-4}
	NT5DC1	1	7.00x10^{-4}	1	5.81x10^{-4}
	ECE2	4	9.32x10^{-4}	4	6.66x10^{-4}
	GPR98	9	5.12x10^{-3}	10	5.84x10^{-4}
	NYNRIN	2	1.08x10^{-3}	2	9.42x10^{-4}

[1] MAF:minor allele frequency

[2] Pra: Protein altering i.e. missense mutation predicted as deleterious by either PolyPhen-2 or SIFT, stop-gain, stop-loss, splice site or frameshift mutations

[3] LoFs: Loss of Function mutations i.e. stop-gain,stop-loss, splice site or frameshift mutations

4.2.4 Association analysis using 1958 Birth Cohort (58BC) controls

In the analysis of Scottish controls, using additional principal components mitigated the effect of fine scale structure on association analysis. However, the Scottish controls were recruited as part of a study on the genetics of obesity; all individuals had a BMI >40, presenting a challenge in distinguishing whether significant loci associated with risk of sepsis or reduced risk of obesity. Due to the efforts of Professor Nazneen Rahman's team in the Division of Genetics and Epidemiology at The Institute of Cancer Research, London, whole exome sequence data from 1,000 individuals in the 58BC was made freely available for use in academic research during the course of the study [250]. This resource provided more suitable, matched population controls for the cases in this study as the controls were recruited from various regions in the UK and have been widely used as controls in many GWAS [141].

4.2.4.1 Population structure

Population outliers were detected using MDS analysis in PLINK. Individuals of non-European ancestry were identified through clustering of study samples with populations in the HapMap3 data. One individual in the 58BC did not cluster with European populations or with the study samples and was excluded from subsequent analysis (Figure 4.6).

4.2.4.2 Single variant analysis

Single variants were tested for an association with sepsis using linear mixed model analysis. There were 310,837 variants that passed quality control metrics as described in Materials and Methods section 2.5.7 requiring a $P_{\text{EMMAX}}<1.61\text{x}10^{-7}$ for statistical significance in the single variant analysis. Of these, 300,990 were coding changes (missense and LoFs), 81% of less than 0.05 minor allele frequency. There were seven statistically significant associations (Table 4.4). Two significant variants were in genes belonging to the keratin-associated family of proteins; these common missense polymorphisms in *KRTAP4-7* and *KRTAP5-8* genes had $P_{\text{EMMAX}} =1.01\text{x}10^{-18}$ and $P_{\text{EMMAX}}=2.58\text{X}10^{-8}$

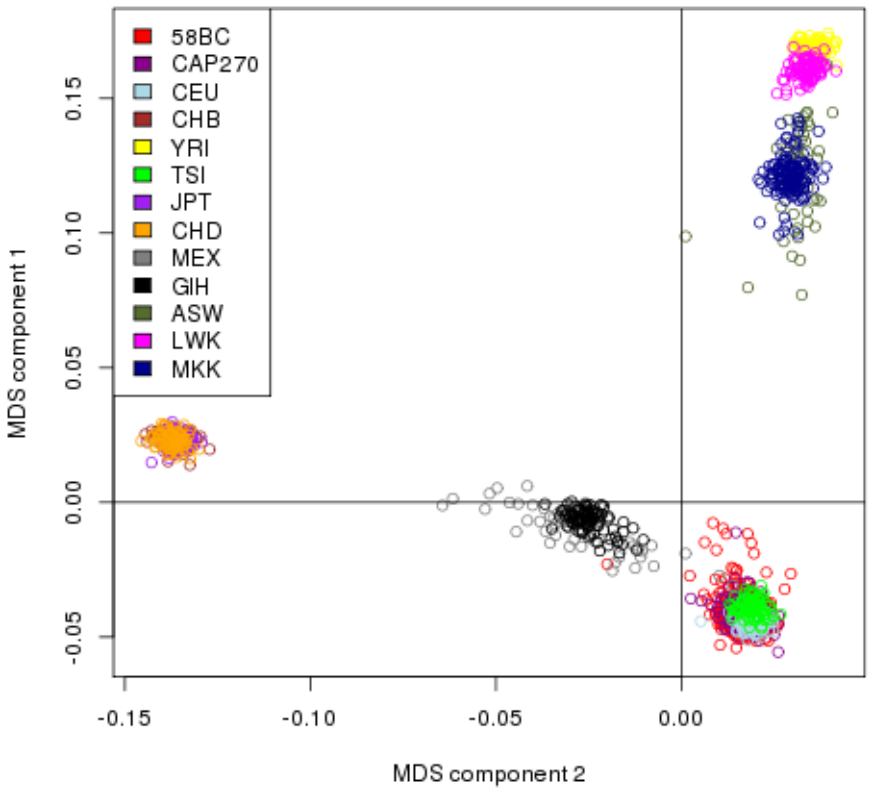


Figure 4.6: Population structure analysis. Multi-Dimensional Scaling(MDS) of the first two components comparing MDS components for populations in HapMap3 with study samples. The two axes represent genetic similarity within and between human populations. Individuals show the greatest genetic similarity within continental populations. One control individual did not cluster with the European populations in HapMap3. **58BC** 1958 birth cohort controls, **CAP270** CAP sepsis cases, **CEU** Utah residents with Northern and Western European ancestry from the CEPH collection, **CHB** Han Chinese from Beijing, China, **YRI** Yoruba in Ibadan, Nigeria, **TSI** Tuscani in Italia, **JPT** Japanese in Tokyo, Japan, **CHD** Chinese in Metropolitan Denver, **MXL** Mexican ancestry in Los Angeles, **GIH** Gujarati Indians in Houston, Texas, **ASW** African ancestry from SouthWest USA, Colorado, **LWK** Luhya in Webuye, Kenya, California, **MKK** Maasai in Kinyawa, Kenya

respectively. Both these variants were predicted as benign by PolyPhen-2 and SIFT. In addition, there were three significant variants in the extended HLA region: two synonymous variants in *HLA-DQA2* and *HLA-DQB2* and one intronic variant in *PGBD1* with P_{EMMAX}=4.78x10^{-13}, P_{EMMAX}=1.47X10^{-11} and P_{EMMAX}=8.01x10^{-13} respectively. Furthermore, an additional 53 single nucleotide variants had suggestive evidence of an association with sepsis susceptibility (P_{EMMAX}<1.00x10^{-4}). An analysis of underlying biological patterns in these genes did not identify enrichment of protein-protein interactions or GO terms in these genes. To test for an enrichment of genes involved in the same pathway, 11 pathways in the KEGG database had statistically significant over-representation ($P_{\text{corrected}}$<0.05) amongst the variants with 60 top variants. Enrichment was particularly evident in antigen presentation pathways and autoimmune disorders involving *HLA* and *CD40* gene products (Table 4.5).

Table 4.4: Single variant analysis. Association statistics for all single variants with P_{EMMAX} <1.00x10^{-4} in the analysis of sepsis susceptibility using 58BC controls. Also shown are the allelic counts of cases and controls (Case count and Control count respectively). There were seven variants that significantly associated with sepsis susceptibility (P_{EMMAX}<1.61x 10^{-7})

Gene	**Pos**	**Allele**	**Variant type**	$\boldsymbol{P_{EMMAX}}$	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
KRTAP4-7	17:39240511	A/T	Missense	1.01x10^{-18}	37/229/1	129/496/358
HLA-DQA2	6:32712979	C/T	Synonymous	4.78x10^{-13}	232/21/0	645/267/0
PGBD1	6:28265640	A/G	Intron	8.01x10^{-13}	244/14/1	879/0/0
HLA-DQB2	6:32726838	C/T	Synonymous	1.47x10^{-11}	18/165/81	47/373/493
GALNT9	12:132862976	G/A	Synonymous	4.98x10^{-10}	18/108/19	510/385/80
KRTAP5-8	11:71249183	T/C	Missense	2.58x10^{-8}	116/20/4	490/400/95
AC011467.1	19:22868965	T/A	Missense	1.37x10^{-7}	233/3/0	844/125/8
CAND2	3:12861506	G/A	Intron	2.09x10^{-7}	260/9/0	890/1/0
HLA-DQB2	6:32726803	C/T	Missense	3.01x10^{-7}	68/166/27	175/545/224
NAGPA	16:5075246	C/G	Utr3	5.70x10^{-7}	0/7/177	0/1/961
CADPS2	7:122033473	G/A	Intron	8.19x10^{-7}	263/6/0	986/0/0
COL6A2	21:47546382	C/A	Intron	1.56x10^{-6}	235/16/1	744/174/15
CEP120	5:122737877	C/T	Intron	1.93x10^{-6}	8/103/40	318/476/186
COL19A1	6:70881768	G/A	Intron	4.08x10^{-6}	256/13/0	963/9/0
KRT81	12:52681080	G/A	Synonymous	4.86x10^{-6}	32/96/18	489/408/83
DPY19L3	19:32974913	A/G	Utr3	5.69x10^{-6}	122/80/14	705/258/20
OR51A4	11:4968198	C/A	Missense	6.03x10^{-6}	18/77/174	16/225/752
OR51A4	11:4967532	G/C	Missense	8.47x10^{-6}	174/77/17	750/222/15
KL	13:33627858	C/G	Intron	9.27x10^{-6}	207/56/4	522/279/24
MYLK4	6:2675398	C/T	Intron	1.02x10^{-5}	264/5/0	972/0/0
SNORD113	14:101393745	G/T	Non-coding transcript	1.07x10^{-5}	260/8/0	977/3/0
TPTE	21:10941967	T/C	Missense	1.18x10^{-5}	262/5/0	991/0/0
CD40	20:44746982	T/C	Utr5	1.20x10^{-5}	17/86/70	49/342/582
HLA-DRB6	6:32521155	G/C	Splice-site	1.55x10^{-5}	244/15/0	969/12/0
SUV420H2	19:55857777	G/T	Intron	1.76x10^{-5}	130/108/29	322/457/153
KANK4	1:62740446	T/C	Synonymous	1.87x10^{-5}	35/143/90	84/418/477
TCF7L1	2:85530986	G/A	Intron	1.89x10^{-5}	264/5/0	935/0/0

Continued on next page...

Gene	**Pos**	**Allele**	**Variant type**	P_{EMMAX}	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
BTBD2	19:1990825	G/A	Splice-site	1.97×10^{-5}	134/60/4	777/170/5
DDX11	12:31249225	T/C	Intron	1.99×10^{-5}	115/101/21	353/424/172
KANK4	1:62740449	T/C	Synonymous	1.99×10^{-5}	35/143/90	84/419/476
STX8	17:9408381	C/T	Missense	2.32×10^{-5}	255/13/0	972/9/0
CXXC1	18:47809854	C/T	Intron	2.50×10^{-5}	262/6/0	982/1/0
SKA1	18:47906416	A/G	Intron	3.06×10^{-5}	38/117/97	75/402/486
YARS2	12:32906869	C/T	Synonymous	3.10×10^{-5}	265/4/0	968/0/0
AC010536.1	16:87733941	G/A	Utr3	3.37×10^{-5}	255/12/1	965/11/0
SYNE2	14:64498016	G/A	Missense	3.48×10^{-5}	259/8/0	953/3/0
NSUN2	5:6632909	G/A	Intron	3.76×10^{-5}	4/107/56	220/465/246
ATP5E	20:57607363	C/T	Utr5	4.17×10^{-5}	116/7/1	948/9/0
DFNB31	9:117169033	A/G	Missense	4.37×10^{-5}	68/138/43	189/498/288
HNF1A	12:121435342	C/T	Synonymous	4.55×10^{-5}	61/91/28	481/409/77
GGCX	2:85788175	T/A	Intron	4.75×10^{-5}	90/133/39	195/457/195
HLA-A	6:29911154	C/A	Missense	4.81×10^{-5}	22/79/40	324/467/170
ALOX12	17:6913652	G/A	Synonymous	5.01×10^{-5}	242/25/0	786/183/14
KIAA0182	16:85688481	C/T	Synonymous	6.15×10^{-5}	229/14/0	920/13/0
PHF2	9:96440808	C/A	Utr3	6.24×10^{-5}	39/94/40	383/427/143
MAD1L1	7:1855531	G/A	Utr3	6.36×10^{-5}	109/70/7	681/218/15
HIST2H2AC	1:149858542	A/G	Synonymous	6.46×10^{-5}	225/4/0	906/0/0
PEAR1	1:156875037	T/G	Intron	7.52×10^{-5}	165/46/1	606/306/28
HLA-DQA1	6:32609312	A/C	Missense	7.83×10^{-5}	59/104/67	130/441/327
POMC	2:25383887	G/A	Utr3	7.87×10^{-5}	61/71/7	634/300/36
ACACB	12:109670681	C/G	Intron	7.97×10^{-5}	264/4/0	910/0/0
ITGA7	12:56081922	A/G	Intron	8.19×10^{-5}	259/8/1	973/5/0
FAM40B	7:129120749	A/G	Intron	8.47×10^{-5}	265/4/0	975/0/0
CYP4B1	1:47283505	G/T	Intron	8.74×10^{-5}	215/47/2	608/254/13
TMPRSS9	19:2415801	G/A	Synonymous	9.29×10^{-5}	87/23/4	879/95/0
KHSRP	19:6417855	G/A	Splice-site	9.59×10^{-5}	12/89/10	450/435/93
C6orf201	6:4079858	C/G	Missense	9.63×10^{-5}	144/25/1	651/256/34
NTRK1	1:156837847	T/C	Intron	9.82×10^{-5}	215/47/1	863/96/0
PPIG	2:170493103	T/A	Missense	9.85×10^{-5}	65/139/59	178/469/342
DNAJC13	3:132172579	A/G	Intron	9.87×10^{-5}	250/5/0	967/1/0

Table 4.5: KEGG pathways over-represented amongst top variants Eleven KEGG pathways were significantly over-represented amongst single variants with P_{EMMAX} <1.00x10^{-4}.

GO-id	Term	NumberOfGenes	$P_{uncorrected}$	$P_{corrected}$(FDR)	Genes
5330	Allograft rejection	4	1.90x10^{-6}	5.44x10^{-4}	*CD40,HLA-DQA1,HLA-DQA2,HLA-A*
5320	Autoimmune thyroid disease	4	8.41x10^{-6}	1.21x10^{-3}	*HLA-DQA2,HLA-A,CD40,HLA-DQA1*
5416	Viral myocarditis	4	1.34x10^{-5}	1.28x10^{-3}	*HLA-DQA2,HLA-A,CD40,HLA-DQA1*
5310	Asthma	3	5.44x10^{-5}	3.90x10^{-3}	*CD40,HLA-DQA1,HLA-DQA2*
5322	Systemic lupus erythematosus	4	1.11x10^{-4}	5.58x10^{-3}	*HLA-DQA2,CD40,HLA-DQA1,HIST2H2AC*
5332	Graft-versus-host disease	3	1.17x10^{-4}	5.58x10^{-3}	*HLA-DQA1,HLA-DQA2,HLA-A*
4940	Type I diabetes mellitus	3	1.60x10^{-4}	6.57x10^{-3}	*HLA-DQA1,HLA-DQA2,HLA-A*
4672	Intestinal immune network for IgA production	3	2.13x10^{-4}	7.65x10^{-3}	*CD40,HLA-DQA1,HLA-DQA2*
4514	Cell adhesion molecules (CAMs)	4	4.57x10^{-4}	1.46x10^{-2}	*HLA-A,HLA-DQA1,HLA-DQA2,CD40*
4612	Antigen processing and presentation	3	7.07x10^{-4}	2.03x10^{-2}	*HLA-DQA2,HLA-A,HLA-DQA1*
5169	Epstein-Barr virus infection	4	1.56x10^{-3}	4.06x10^{-2}	*HLA-A,HLA-DQA1,HLA-DQA2,CD40,HIST2H2AC*

4.2.4.3 Gene burden analysis

In the gene burden analysis of protein altering mutations, there were no genes with a statistically significant excess of rare (<1%) protein altering mutations. The most significant genes with suggestive evidence of association were *CCDC159* with P=5.35x10^{-5} and *ARHGAP18* with P=7.51x10^{-5} (Table 4.6). The *CCDC159* (Coiled-coil domain containing 159) is ubiquitously expressed and has no known function. The gene had four singleton mutations in four different cases and none in controls, one nonsense mutation and three deleterious missense mutations. The *ARHGAP18* gene encodes a Rho GTPase activating protein whose role is regulation of signalling pathways in cytoskeleton remodelling. There were four mutations in this gene, all were singleton deleterious missense mutations present in four different cases and absent in controls. In total, 14 genes had suggestive evidence of an association of rare variants with sepsis susceptibility (P<1.0x10^{-3}, Table 4.6). Using the program STRING, there were no significantly enriched biological networks in these genes. There were no direct protein-protein interactions amongst the proteins encoded, nor were there any GO terms or KEGG pathways that were significantly overrepresented.

Further analysis was performed testing the joint contribution of rare (<1%) and low frequency (1-5%) variants to sepsis susceptibility. There were no statistically significant genes while jointly testing rare and low frequency variants (Table 4.6). The top gene was *ARHGAP18* with P=7.51x10^{-5}. The *CCDC159* gene was no longer significant in this analysis (P=0.22). In total, 12 genes had suggestive evidence of association (P<1.00x10^{-4}) in the joint analysis. There were no direct protein-protein interactions amongst the gene products or enrichment in GO terms or KEGG pathway in this set of genes.

The second class of variants tested was LoF mutations. Allele frequency thresholds were implemented as for protein altering mutations. There we no genes that achieved exome-wide significance for an association of rare and low frequency LoFs with sepsis susceptibility. Six genes had suggestive evidence of an association with sepsis, *ATP13A1*, *TOM1L2*, *CDC42BPG*, *LACE1*, *NSMCE4A* and *PUM2* (Table 4.6). The

ATP13A1 gene had three singleton splice variants in three different cases but none in controls. There were two splice site mutations in *TOM1L2* gene, one mutation was present in only one case and absent in controls while the second mutation was present in four cases and absent in controls. The *CDC42BPG* gene had three singleton splice variants in cases only while the *LACE1* gene had a single splice site variant present in four cases and absent in controls. There were two splice region variants in *NSMCE4A*, one rare variant present in five cases (1.8%) and one control(0.1%) and one low frequency variant present in eight cases (2.9%) and 22(2.2%) controls. In *PUM2* gene, one splice site variant was present in seven cases(2.7%) and three controls (0.1%).

4.2.4.4 Pathway analysis

Burden analysis was expanded to assess whether genes of similar function were perturbed in sepsis. Deleterious mutations were aggregated as per canonical pathways defined in the KEGG database or in genes with the same gene ontology annotation. At the time of this analysis, there were 299 canonical pathways manually curated in the KEGG database requiring $P<1.25\text{x}10^{-4}$ for statistical significance. Though not to statistical significance, the "Citrate cycle (TCA cycle)" pathway had the strongest association while aggregating just rare variants or jointly aggregating rare and low frequency variants ($P=9.92\text{x}10^{-4}$ and $P=2.63\text{x}10^{-3}$ respectively). Of the 30 genes mapped to this pathway in KEGG, 26 genes had at least one deleterious mutation in our study samples. Canonical pathways in KEGG with less than nominal p-value ($P=0.05$) are shown in Table 4.7. Using a parallel approach, genes were classified by their biological process and molecular function from the GO database [274, 316, 317]. There were a total of 14,698 GO biological processes and molecular function annotations requiring a conservative $P<3.40\text{x}10^{-6}$ after Bonferroni correction. There were no GO terms significantly enriched for rare or low frequency deleterious mutation. The ontology "enzyme-directed rRNA 2'-O-methylation" (GO:0000453) had the strongest association in both rare and joint analysis with $P=1.24\text{x}10^{-4}$ and $P=1.07 \text{ x } 10^{-4}$ respectively. Three genes (*MRM1*, *FTSJ2* and *FTSJ3*) are annotated with this biological process which involves methylation of ribosomal RNA.

Table 4.6: Gene based analysis. Genes with suggestive evidence of an association with sepsis (P_{SKAT} <1.0x10^{-3}), comparing sepsis cases against 58BC controls. Rare variants(<1% MAF[1]) were analysed separately as well as jointly with low frequency variants (1-5%)

		Rare (<1%)		Rare & low frequency (<5%)	
	Gene	**No. of variants**	**P_{SKAT}**	**No. of variants**	**P_{SKAT}**
Pra[2]	*CCDC159*	4	5.35x10^{-5}	5	0.22
	ARHGAP18	4	7.51x10^{-5}	4	7.51x10^{-5}
	ANO9	3	1.57x10^{-4}	3	3.10x10^{-4}
	PNN	4	1.58x10^{-4}	4	2.39x10^{-4}
	IDH3B	3	3.88x10^{-4}	3	1.10x10^{-3}
	MRM1	3	4.55x10^{-4}	3	4.55x10^{-4}
	PUM2	5	4.81x10^{-4}	5	2.34x10^{-4}
	OR10K1	3	4.83x10^{-4}	3	4.83x10^{-4}
	ROCK2	3	4.98x10^{-4}	3	4.98x10^{-4}
	FYN	3	5.45x10^{-4}	4	0.49
	TMEM184A	2	7.32x10^{-4}	2	1.08x10^{-3}
	FTSJ3	3	7.64x10^{-4}	3	4.70x10^{-4}
	HTRA4	3	8.15x10^{-4}	3	5.33x10^{-4}
	LACE1	2	9.53x10^{-4}	2	8.08x10^{-4}
	DENND1B	3	1.30x10^{-3}	3	5.04x10^{-4}
	PCNXL2	11	0.39	13	8.08x10^{-4}
	HLA-DRB6	0	NA	1	1.97x10^{-5}
LoFs[3]	*ATP13A1*	3	1.25x10^{-4}	3	1.25x10^{-4}
	TOM1L2	2	3.44x10^{-4}	2	3.30x10^{-4}
	CDC42BPG	3	3.83x10^{-4}	3	3.83x10^{-4}
	LACE1	1	4.62x10^{-4}	1	4.17x10^{-4}
	NSMCE4A	1	5.49x10^{-4}	2	5.00x10^{-2}
	PUM2	1	5.69x10^{-4}	1	3.84x10^{-4}
	HLA-DRB6	0	NA	1	8.36x10^{-6}

[1] MAF:minor allele frequency

[2] Pra: Protein altering i.e. missense mutation predicted as deleterious by either PolyPhen-2 or SIFT, stop-gain, stop-loss, splice site or frameshift mutations

[3] LoFs: Loss of Function mutations i.e. stop-gain,.stop-loss, splice site or frameshift mutations

Table 4.7: Pathway analysis KEGG pathways with P<0.05 in the analysis aggregating deleterious variant across pathways in KEGG database.

Rare (<1%)		**Rare & low frequency (<5%)**	
Pathway	**PVALUE**	**Pathway**	**PVALUE**
Citrate cycle (TCA cycle)	9.92×10^{-4}	Citrate cycle (TCA cycle)	2.63×10^{-3}
Ribosome	1.30×10^{-3}	Mismatch repair	8.81×10^{-3}
mRNA surveillance pathway	3.26×10^{-3}	Other glycan degradation	1.30×10^{-2}
Phagosome	4.37×10^{-3}	NOD-like receptor signalling pathway	1.70×10^{-2}
Carbon metabolism	1.03×10^{-2}	mRNA surveillance pathway	1.74×10^{-2}
Regulation of autophagy	1.69×10^{-2}	Homologous recombination	2.53×10^{-2}
2-Oxocarboxylic acid metabolism	1.77×10^{-2}	Notch signalling pathway	2.61×10^{-2}
Alanine, aspartate and glutamate metabolism	2.41×10^{-2}	Small cell lung cancer	3.69×10^{-2}
Legionellosis	2.68×10^{-2}	Insulin secretion	4.18×10^{-2}
Biosynthesis of antibiotics	2.70×10^{-2}	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	4.22×10^{-2}
Biosynthesis of amino acids	3.33×10^{-2}	Base excision repair	4.53×10^{-2}
Cardiac muscle contraction	3.52×10^{-2}	Cyanoamino acid metabolism	4.63×10^{-2}
RNA transport	3.85×10^{-2}	Sphingolipid metabolism	4.64×10^{-2}
Nicotine addiction	4.13×10^{-2}	ECM-receptor interaction	4.75×10^{-2}
Chronic myeloid leukemia	4.45×10^{-2}	Bacterial invasion of epithelial cells	4.91×10^{-2}
Taste transduction	4.45×10^{-2}		
Galactose metabolism	4.50×10^{-2}		

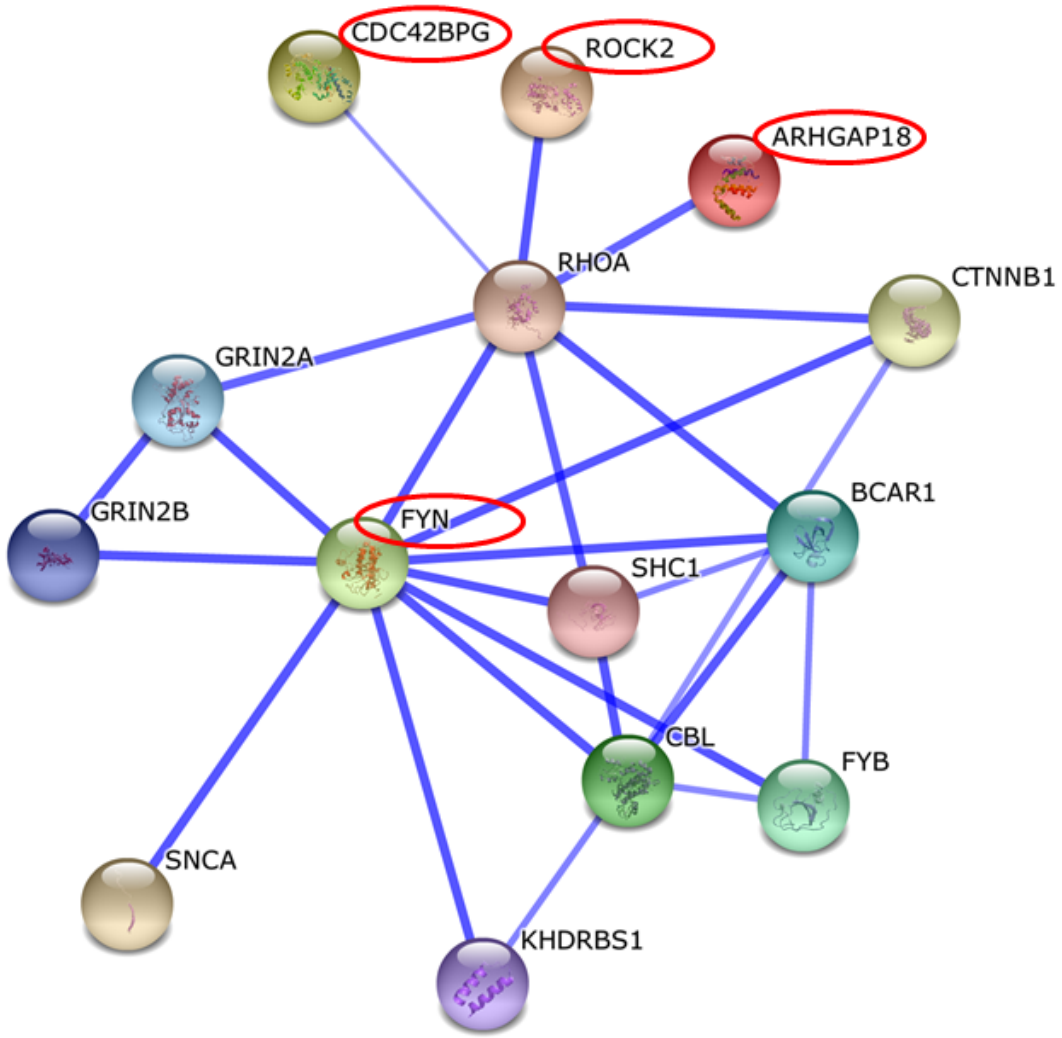


Figure 4.7: Protein-protein interaction network. A network showing the Fyn-RhoA-ROCK protein interaction network. Diagram generated in STRING using CDC42BPG, ARHGAP18, FYN and ROCK2 as input. Genes with deleterious mutations in sepsis samples are highlighted.

4.3 Discussion

Despite apparently adequate treatment of the underlying infection and supportive therapy, sepsis remains a leading cause of death in intensive care units. The pathophysiology of sepsis is complex and poorly understood but is characterised by dysregulation of the innate and adaptive immune system, the coagulation cascade as well as the endocrine system [7, 18]. This chapter describes efforts made to identify novel genes and pathways underlying sepsis susceptibility.

For this study, preliminary analyses using ALSPAC whole genome data and Scottish whole exome data as controls are presented for completeness, as key quality control criteria in the data processing pipeline were developed from these analyses. Furthermore, fundamental questions relating to use of genetic data processed on different pipelines as controls as well as adequate ethnicity matching in rare variant analyses were raised. Both these questions have not been adequately addressed in reported studies in regards to rare variant analyses [318, 319]. The associations observed with Scottish controls in both single variant analysis as well as gene burden analysis might represent an association with obesity as only three single variants had also suggestive evidence with 58BC controls. One common missense polymorphism in the *DFNB31* gene which encodes the whirlin protein, a component of sensory cells in the inner ear, had suggestive evidence in both analyses. Mutations in this gene are linked to autosomal recessive deafness as well as Usher syndrome type 2 [320, 321]. An intronic polymorphism in the *PEAR1* gene encoding the platelet endothelial aggregation receptor 1 also had suggestive significance in both analyses. This gene presents an interesting candidate due to the dysregulated coagulation observed in sepsis; however, replication in independent studies is necessary.

As the 58BC controls provided the best matched controls for this study, results from these analyses are discussed in more detail below. In the single variant analyses, seven single variants in keratin family genes as well the extended *HLA* locus achieved statistical significance. The *HLA* region contains highly homologous sequences and mapping of short reads to this region is prone to error despite strict quality

control. *HLA* genes have a key role in antigen presentation and *HLA* alleles are linked to various infectious diseases and autoimmune diseases [322, 323]. However, extensive fine mapping of this region would be needed to identify specific *HLA* alleles associating with sepsis. Similarly, the keratin-associated proteins are highly polymorphic [324, 325]. An association with infectious disease for members of this gene family has not been reported.

The *ARHGAP18* gene consistently had a strong signal in both rare and joint analysis. This gene had four singleton missense mutations predicted as deleterious in four different sepsis cases but none in controls. *ARHGAP18* encodes a Rho A GTPase-activating protein that modulates changes in the actin cytoskeleton in response to external signals [326]. Moreover, the ARHGAP18 protein has been demonstrated to regulate cell shape, spreading and migration of cells through its conversion of Rho A GTPase to an inactive state [326]. Rho A GTPase activity is executed by means of effector kinases termed Rho-associated kinases 1 and 2 (ROCK1 and ROCK2 respectively) [327]. Interestingly, the *ROCK2* gene also had suggestive evidence of an association with sepsis (P=4.98x10^{-4}). This gene had three singleton missense mutations predicted as deleterious present in three cases and none in controls. These sepsis cases were different individuals to those harbouring deleterious mutations in the *ARHGAP18* gene. ROCK proteins are crucial in maintaining endothelial integrity. Activation of ROCK leads to disruption of endothelial cell junctions increasing permeability and promoting leukocyte trans-endothelial migration during inflammation [328, 329]. Rho kinases have been implicated in autoimmune diseases such as Behcet disease as well as cardiovascular and pulmonary diseases including asthma [327, 330].

The *FYN* gene encoding the FYN tyrosine kinase also had suggestive evidence of association primarily through rare variants (P=5.45x10^{-4}). The gene had two splice site variants and one missense variants all present in single cases. In the human lung, FYN kinase has been demonstrated as a crucial component in TNFα induced vascular permeability [331]. TNFα is a key inflammatory mediator in sepsis and promotes remodelling of epithelial cells via FYN kinase mediated phosphorylation of adherens

junction proteins leading to altered permeability in the lungs [238, 307, 331–333]. The *FER* gene -a recently reported locus associating with 28-day mortality from sepsis- encodes a tyrosine kinase of similar function to the FYN-kinase [244,334]. In addition to the lung epithelia, FYN kinase is one of the major kinases expressed on T lymphocytes and is crucial in T cell activation [335]. Interestingly, FYN kinase is also involved in Rho A GTPase signalling and remodelling of the cytoskeleton [331, 332, 335–340]. In addition, FYN activation has been shown to stabilise the integrity of the epithelial barrier and protect against LPS induced sepsis in mice [336].

While aggregating rare LoF variants alone, *CDC42BPG* gene had suggestive evidence of association (P=3.83x10^{-4}). The gene had three splice site variants in three different cases and none in controls. Similar to ROCK proteins, CDC42BPG serves as one of the downstream effectors for CDC42 - another member of the Rho GTPases - and also binds weakly to Rho A [341, 342].

In summary, four genes *ARHGAP18*, *ROCK2*, *CDC42BPG* and *FYN* had suggestive evidence of rare variants associating with risk of sepsis. Strikingly, these four genes are each involved in signalling through small GTPases of the Rho subfamily (Figure 4.7) [341]. There were 22 of the 270 (8.1%) cases and 22 of the 999 controls (2.2%) contained rare deleterious mutations across all four genes. Due to the incomplete curation of current pathway databases, this interaction was not highlighted by pathway analysis with *ARHGAP18* and *CDC42BPG* not annotated to any pathway in KEGG.

There is increasing evidence supporting the role of Rho GTPase signalling in infectious diseases, as the activity of these GTPases results in a wide range of biological processes such as endocytosis during antigen presentation, phagocytosis and cell proliferation [343, 344]. Indeed, the Wiskott-Aldrich syndrome, a primary immunodeficiency characterised by recurrent bacterial, viral and fungal infections, results from rare highly penetrant mutations in the *WASP* gene which encodes another effector for CDC42 GTPase [345]. Furthermore, various pathogens such as *Salmonella typhimurium* and *Neisseria meningitidis* have been shown to utilise the Rho GTPase pathway to subvert the immune system [346–349]. In addition, Rho GTPase signalling has been shown to mediate NF-κB signalling in response to LPS [350–352].

The heterogeneity of sepsis presents a major challenge in genetic studies. Potential confounders in this analysis such as age and comorbidities were addressed by selecting only cases under 60 years of age and excluding all individuals with any reported comorbidities. However, although the pneumococcus was the major causal pathogen for CAP in this study, accounting for 22% of all cases, other bacterial and viral species jointly accounted for 15% of cases. A significant proportion of cases (48%) lacked a microbiology proven causal pathogen. This pathogen heterogeneity may result in different pathophysiological processes leading to sepsis as each pathogen may elicit different cellular responses. To disentangle this, extensive clinical data such as vaccination status and immunological phenotyping in addition to genetic data would be required. For this study, the few number of individuals, less than 10 in each category with the exception of pneumococcal cases, precluded a subgroup analysis based on the pathogen.

These preliminary results highlight four different genes in the Rho GTPase signalling pathway as having a potential impact on sepsis susceptibility. These novel genes have not been previously implicated in sepsis and at the time of writing had not been deposited in GeneMatcher -a database for curating genetic findings with a clinical interest [353]. Results in this single analysis suggest preliminary evidence of association, and replication of these results through targeted re-sequencing of all genes in this pathway in independent cohorts is required to provide confirmatory evidence of association.

Chapter 5

Genetic susceptibility to paediatric pleural infection

5.1 Introduction

Thoracic empyema is a suppurative infection of the pleural cavity usually occurring as a complication of pneumonia. The incidence rate of paediatric pleural empyema has continued to rise despite improved healthcare. The role of host genetics in paediatric empyema has remained largely unexplored. Candidate gene studies have implicated *NFKBIA*, *PTPN22* and *CYP1A1* in susceptibility to empyema in adults [88, 89, 354]. However, unbiased studies using recent genome-wide methods have not been reported. This study aims to use a whole exome sequencing approach to identify novel genetic loci associating with paediatric empyema. All selected empyema cases were previously well children below the age of 18 years admitted to hospitals across the UK. Population controls were 1,000 individuals from the 1958 birth cohort who had whole exome sequence data freely available to the scientific community. Further details on these cohorts are in Materials and Methods section 2.2.

5.2 Results

5.2.1 Pilot study

In order to establish the quality of DNA in the empyema samples, a pilot study of 12 samples was undertaken to assess quality of exome sequencing data. The empyema

samples in this study were saliva samples and in many cases were collected from very young children. Saliva volumes were relatively low, resulting in low DNA yield. Comparing sequence metrics from 270 high quality DNA samples from the CAP study (Chapter 4) sequenced on the same platform and processed with the same pipeline, high rates of RNA contamination were observed in the saliva extractions. This was attributed to lack of RNAse treatment in the DNA extraction protocol; however, this did not appear to have an effect on library preparation as all 12 pilot samples passed quality control. In addition, the rate of duplicate reads was observed to be higher for the saliva extractions, an average of 20.8%, compared to 14% in the CAP samples (Table 5.1). In coverage analysis, the proportion of targeted regions covered at greater than 20X was 75% in the pilot samples compared to 86 % in the CAPs, however, the mean coverage across all samples was greater than 30X (Figure 5.1 and Table 5.2) sufficient for calling heterozygous variants with greater than 90% confidence [172]. In view of this, sample quality was deemed sufficient and an additional 88 samples were sequenced for a total of 100 empyema cases in this study.

Table 5.1: Summary of sequence metrics Comparing pilot empyema samples with exome data from 270 samples of good quality DNA. Both datasets had 64MB of the genome captured using the Nimblegen SeqCap EZ V3.0 kit

Sequencing metrics	**CAP270**	**EMP12**
Size of target region	64,190,747	64,190,747
Mean depth of target region	65.4x	54.3x
Percentage of target bases covered >= 10x (%)	91.6	87.1
Percentage of target bases covered >= 20x (%)	86.5	75.1
Percentage of target bases covered >= 30x (%)	77.9	61.6
Average duplication rate (%)	14.9	20.8

Table 5.2: Sequence metrics for individual samples in the pilot study. Reads are shown in millions before and after removal of duplicates as well as the proportion of the 64MB targeted region covered at 10x, 20x and 30x coverage.

SampleID	RawReads	MappedReads	Duplicates(%)	Reads afterDedup	MeanCovofTarget	10xTarget	20xTarget	30xTarget
sample_1	34.1M	33.4M	20.2	27.2M	31.2	79	56.9	37.5
sample_2	44.6M	44.3M	19.7	35.8M	40.3	82.9	66.2	49.9
sample_3	45.5M	45.2M	21	36.0M	40.6	85	69.4	52.1
sample_4	47.6M	47.2M	20.8	37.7M	44	86.2	71.9	55.6
sample_5	53.7M	53.0M	19.8	43.0M	48.9	87.3	74.8	60.2
sample_6	57.7M	57.3M	20.3	46.0M	51.6	87	75.6	62.5
sample_7	60.9M	60.5M	19.6	49.0M	55.6	89.7	80.6	68.2
sample_8	63.4M	63.0M	21.3	49.9M	57.3	88.8	79.3	67.3
sample_9	64.6M	63.9M	20.1	51.6M	58	88.6	78.4	66.4
sample_10	64.8M	64.5M	21.3	51.0M	59.1	88.2	78.4	66.9
sample_11	64.6M	64.1M	22	50.3M	59.9	89.1	80.2	68.8
sample_12	110.9M	110.0M	23.5	84.8M	105.4	93	89.3	84.1

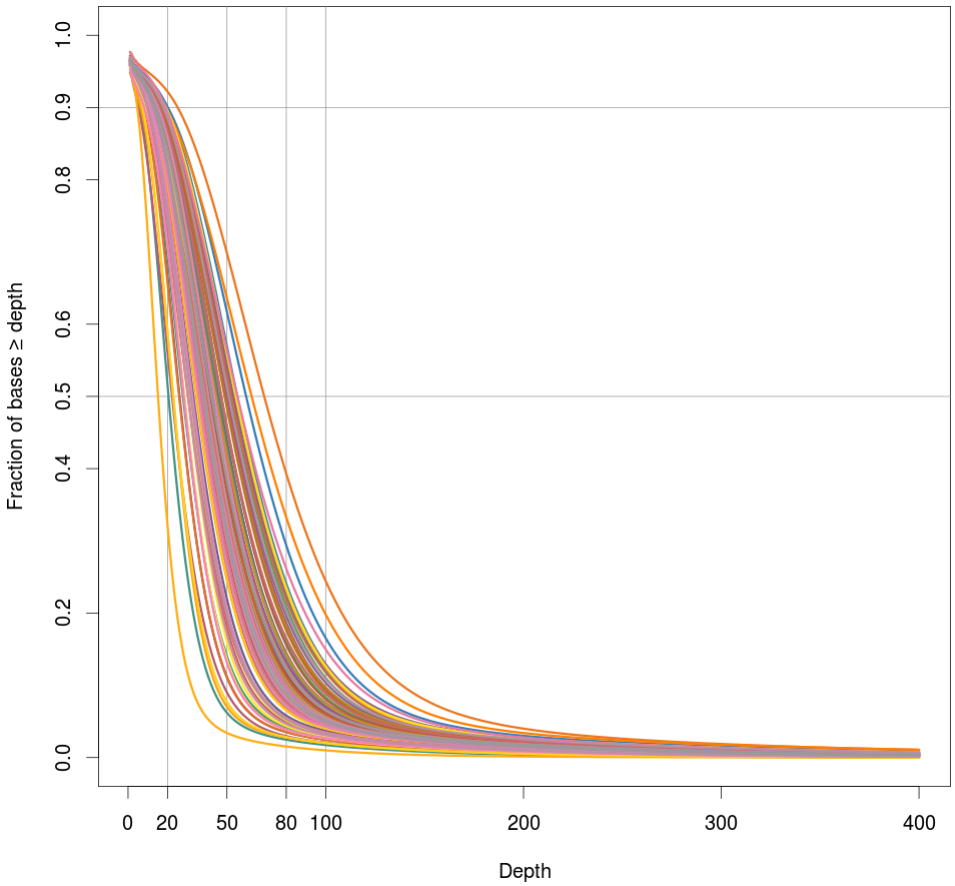


Figure 5.1: Empyema coverage. Exome sequencing coverage in 88 samples extracted from saliva. The lines show the proportion of the 64MB targeted region covered at various depths for each individual. All but one sample had an average depth of greater than 20X. Over 95% of all regions targeted were covered at least once.

5.2.2 Population structure

To minimise confounding of association analysis due to differences in genetic ancestry and underlying sample structure, Multi-Dimensional Scaling (MDS) as implemented in PLINK was used to detect population outliers [267,319,355]. Diverse populations from HapMap3 were used to identify continent-specific clustering of the study samples. On visualisation in a two-dimensional space, two empyema cases and three controls were observed to cluster with non-European populations and were excluded as population outliers from all subsequent analysis (Figure 5.2).

5.2.3 Single variant analysis

Linear mixed model association analysis (EMMAX) was performed to compare the distribution of genotypes between cases and controls [269, 270, 356]. A kinship matrix calculated from common variants was used to control for cryptic relatedness not captured by major axes of MDS. Of the 98 empyema cases, 63 had microbiologically proven *Streptococcus pneumoniae* as the causative pathogen. Subgroup analyses to identify pneumococcal specific genetic risk factors were performed concurrently with the main empyema analysis.

5.2.3.1 Empyema

There were a total 181,367 variants that passed quality control as described in Materials and Methods section 2.5.7. Of these, 8,379 were disruptive LoFs (frameshifts, stop, splice site and inframe mutations) of which 72% were of less than 0.05 allele frequency. Single variant analysis for association with risk of empyema was carried out for all variants and a $P<2.76\text{x}10^{-7}$ was therefore required for statistical significance after Bonferroni correction for all the variants tested. Two exonic variants achieved statistical significance (Table 5.3, Figures 5.3a and 5.4a). The strongest association was in a missense variant in the *CD109* gene (p.Gly828Glu, $P_{\text{EMMAX}}=7.58\text{x}10^{-8}$). This low frequency variant was predicted as deleterious by both PolyPhen-2 and SIFT and was identified in 10% of cases and 2% of controls. A rare splice site variant in the 12th

exon of *ACOX2* gene was also significantly associated with empyema (rs201210623, P_{EMMAX}=2.50 x 10^{-7}). The variant was present in 2% of cases and 0.1% of controls. A third significant association was a rare synonymous variant in *PPP2R2C* gene (rs369542799, P_{EMMAX}=2.81x10^{-7}) present in three cases but absent in controls. An additional 76 variants had suggestive evidence of association (P_{EMMAX}<1.0X10^{-4}, Table 5.3). To interrogate all the top variants for an enrichment of pathways or interaction networks, the program STRING was used [273]. This integrated approach revealed 53 protein-protein interactions in the top signals; however this did not represent a significant enrichment above the null, $P_{\text{corrected}}$= 0.45. Additionally, there were no canonical pathways in KEGG [275] or Gene Ontology (GO) [274] terms significantly enriched in these genes.

5.2.3.2 Pneumococcal empyema

Single variant analysis was also undertaken for the 63 cases with empyema due to *Streptococcus pneumoniae*. There were 113,437 variants that passed quality control requiring P<4.40x10^{-7} for a statistically significant association. Twelve rare single nucleotide variants appeared to be significantly associated with pneumococcal empyema (Table 5.4, Figures 5.3b and 5.4b). All variants, with the exception of the splice variant in *ACOX2* gene, were intronic, synonymous or benign missense. An additional 130 variants had suggestive evidence of association (Table 5.4) interrogation of these variants for biological patterns using the program STRING did not reveal statistically significant enrichment in KEGG pathways, protein-protein interactions or over-representation of GO terms.

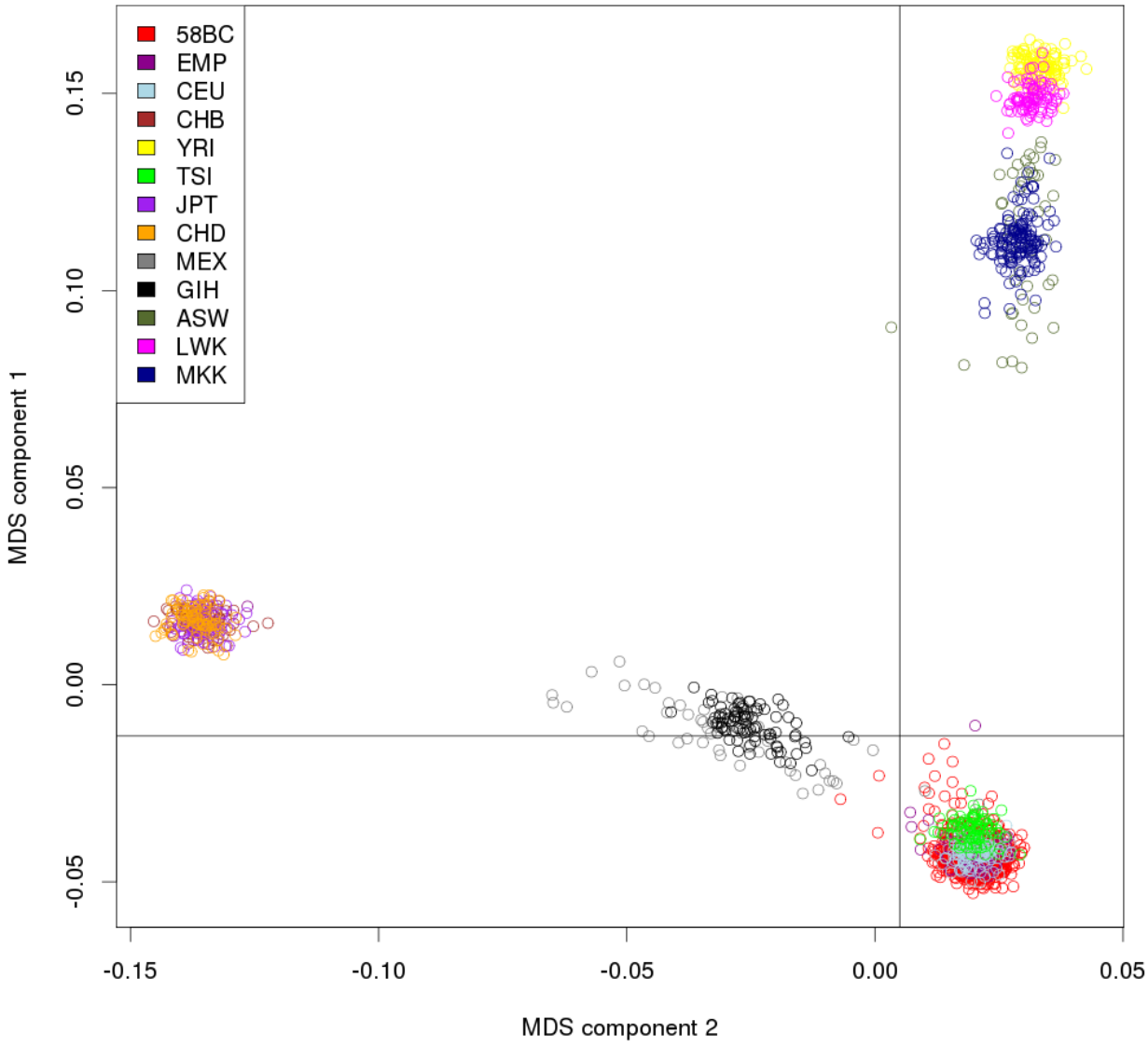


Figure 5.2: Population structure analysis. Multi-Dimensional Scaling(MDS) of the first two components comparing MDS components for populations in HapMap3 with study samples. The two axes represent genetic similarity within and between human populations. Individuals show the greatest genetic similarity within continental populations. Two empyema cases and three controls did not cluster with the European populations in HapMap3. **58BC** 1958 birth cohort population controls, **EMP** empyema cases, **CEU** Utah residents with Northern and Western European ancestry from the CEPH collection, **CHB** Han Chinese from Beijing, China, **YRI** Yoruba in Ibadan, Nigeria, **TSI** Tuscani in Italia, **JPT** Japanese in Tokyo, Japan, **CHD** Chinese in Metropolitan Denver, **MXL** Mexican ancestry in Los Angeles, **GIH** Gujarati Indians in Houston, Texas, **ASW** African ancestry from SouthWest USA, Colorado, **LWK** Luhya in Webuye, Kenya, California, **MKK** Maasai in Kinyawa, Kenya

Table 5.3: Single variant analysis. Association statistics for empyema showing all single nucleotide variants with a P_{EMMAX} <1.0x10^{-4}. Also shown are the allelic counts of cases and controls (Case count and Control count respectively). Top two variants were significantly associated with empyema (P_{EMMAX}<2.76 x 10^{-7}) while 158 variants had suggestive evidence of association (P_{EMMAX}<1.0x10^{-4})

Gene	**Pos**	**Allele**	**Variant type**	$\boldsymbol{P_{EMMAX}}$	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
CD109	6:74497102	G/A	Missense	7.58x10^{-8}	81/17/0	917/41/0
ACOX2	3:58508217	G/A	Normal	2.50x10^{-7}	91/4/0	863/2/0
PPP2R2C	4:6331036	G/A	Synonymous	2.81x10^{-7}	95/3/0	962/0/0
FSTL5	4:162577683	T/G	Intron	1.04x10^{-6}	92/4/0	888/2/0
MED24	17:38178627	A/G	Intron	1.11x10^{-6}	42/17/1	295/462/134
POLR3A	10:79769544	G/T	Intron	1.38x10^{-6}	88/9/1	939/20/0
PTP4A2	1:32375603	T/C	Intron	1.42x10^{-6}	94/2/0	956/0/0
C5orf42	5:37169169	T/C	Synonymous	1.51x10^{-6}	92/6/0	971/6/0
TMEM169	2:216965360	C/T	Utr3	1.70x10^{-6}	94/4/0	972/2/0
ACSF3	16:89167404	T/C	Synonymous	1.79x10^{-6}	10/47/38	36/313/624
ASH2L	8:37985897	A/G	Synonymous	1.81x10^{-6}	0/10/88	0/21/955
USP33	1:78180445	T/C	Missense	2.00x10^{-6}	95/3/0	969/1/0
DEPDC1B	5:59982996	C/T	Missense	2.12x10^{-6}	95/3/0	939/1/0
GSDMC	8:130764188	C/G	Intron	2.15x10^{-6}	95/3/0	966/1/0
LDHAL6B	15:59500255	C/T	Synonymous	2.16x10^{-6}	57/28/0	866/114/5
DOPEY2	21:37600028	C/T	Synonymous	2.18x10^{-6}	95/3/0	960/1/0
ITGA8	10:15647805	C/T	Intron	2.38x10^{-6}	87/8/0	959/13/0
HK3	5:176308894	A/G	Intron	2.43x10^{-6}	77/13/0	921/35/0
PTPRC	1:198725059	G/A	Missense	2.71x10^{-6}	95/3/0	912/1/0
LGI3	8:22009053	G/A	Synonymous	2.74x10^{-6}	93/2/0	972/0/0
VIT	2:37000915	C/A	Intron	2.74x10^{-6}	89/2/0	957/0/0
HERC2	15:28413640	C/T	Missense	2.76x10^{-6}	71/2/0	964/0/0
PPFIA2	12:81738422	T/C	Intron	2.98x10^{-6}	95/2/0	962/0/0
MUC2	11:1086051	C/T	Missense	3.15x10^{-6}	70/3/0	953/1/0
YARS	1:33248032	C/T	Missense	3.18x10^{-6}	96/2/0	952/0/0
ZNHIT1	7:100861166	T/C	Utr5	3.38x10^{-6}	96/2/0	961/0/0
AURKB	17:8110585	C/T	Missense	3.43x10^{-6}	91/3/0	959/1/0
FAM76A	1:28053937	T/G	Intron	3.47x10^{-6}	94/2/0	957/0/0
CD97	19:14492272	T/C	Utr5	3.59x10^{-6}	48/28/0	336/476/146
OR10X1	1:158549429	G/A	Synonymous	3.82x10^{-6}	95/3/0	983/1/0
FNDC3A	13:49782000	C/T	Utr3	4.00x10^{-6}	96/2/0	982/0/0
NA	1:85009837	G/T	Intergenic	4.18x10^{-6}	94/2/0	937/0/0
OCA2	15:28096573	C/T	Missense	4.20x10^{-6}	90/2/0	955/0/0
OCA2	15:28116336	C/T	Synonymous	4.27x10^{-6}	91/2/0	958/0/0

Continued on next page...

Gene	Pos	Allele	Variant type	P_{EMMAX}	Case count (ref/het/mut)	Control count (ref/het/mut)
SYNE2	14:64587686	A/G	Synonymous	$4.47x10^{-6}$	96/2/0	979/0/0
STRA6	15:74483660	A/C	Intron	$4.48x10^{-6}$	91/2/0	966/0/0
HERC2	15:28424100	C/T	Synonymous	$4.52x10^{-6}$	94/2/0	961/0/0
ABCC4	13:95815804	A/G	Intron	$4.56x10^{-6}$	95/3/0	958/1/0
VWA8	13:42393336	G/C	Intron	$4.58x10^{-6}$	95/2/0	859/0/0
HERC2	15:28443682	A/G	Intron	$4.59x10^{-6}$	96/2/0	980/0/0
HERC2	15:28446638	A/G	Synonymous	$4.61x10^{-6}$	96/2/0	977/0/0
PARPBP	12:102547741	G/A	Missense	$4.87x10^{-6}$	94/3/0	967/1/0
PARPBP	12:102558369	C/A	Missense	$4.98x10^{-6}$	95/3/0	975/1/0
IPO5	13:98670831	A/C	Synonymous	$5.10x10^{-6}$	96/2/0	971/0/0
DOK1	2:74782639	C/T	Intron	$5.14x10^{-6}$	95/2/0	872/0/0
UBD	6:29523787	G/A	Missense	$5.25x10^{-6}$	95/2/0	986/0/0
DDX21	10:70725147	A/G	Synonymous	$5.26x10^{-6}$	95/3/0	973/1/0
CDC42EP4	17:71282516	G/T	Synonymous	$5.51x10^{-6}$	55/2/0	965/0/0
ATP5F1	1:111999312	A/G	Missense	$5.51x10^{-6}$	96/2/0	973/0/0
MCM4	8:48883214	C/G	Synonymous	$5.72x10^{-6}$	96/2/0	957/0/0
41883	7:55874726	T/G	Intron	$5.93x10^{-6}$	96/2/0	855/0/0
PCGF1	2:74732932	T/C	Intron	$6.14x10^{-6}$	93/3/0	972/1/0
INADL	1:62267321	C/T	Missense	$6.16x10^{-6}$	96/2/0	974/0/0
TOMM34	20:43585156	G/A	Intron	$6.21x10^{-6}$	95/3/0	972/1/0
RP1	8:55541005	C/T	Synonymous	$6.22x10^{-6}$	96/2/0	976/0/0
SOX7	8:10683623	G/A	Intron	$6.50x10^{-6}$	55/38/4	320/496/112
SVIL	10:29812509	C/T	Missense	$6.67x10^{-6}$	96/2/0	944/0/0
CNIH4	1:224563471	G/A	Intron	$6.71x10^{-6}$	96/2/0	954/0/0
CCNA1	13:37014255	A/G	Missense	$6.72x10^{-6}$	96/2/0	973/0/0
DPP10	2:116539979	A/G	Normal	$7.16x10^{-6}$	92/6/0	970/8/0
CDC42BPG	11:64609200	G/A	Synonymous	$7.17x10^{-6}$	94/2/0	971/0/0
DHTKD1	10:12133642	C/T	Missense	$7.24x10^{-6}$	96/2/0	976/0/0
COL7A1	3:48627933	G/T	Missense	$7.76x10^{-6}$	91/2/0	961/0/0
FLRT1	11:63883990	C/T	Missense	$7.90x10^{-6}$	96/2/0	940/0/0
NA	7:98881262	G/A	Intergenic	$8.16x10^{-6}$	95/2/0	969/0/0
PLA2G4D	15:42379807	G/A	Intron	$8.24x10^{-6}$	88/2/0	896/0/0
XRCC2	7:152357883	A/G	Intron	$8.39x10^{-6}$	48/3/0	967/2/0
CDH13	16:83065664	G/C	Synonymous	$8.43x10^{-6}$	95/2/0	973/0/0
TMEM135	11:87029156	C/T	Intron	$8.54x10^{-6}$	84/2/0	951/0/0
LOXHD1	18:44181425	T/G	Missense	$8.54x10^{-6}$	96/2/0	912/0/0
HEXIM1	17:43226665	G/C	Synonymous	$8.57x10^{-6}$	91/5/0	965/5/0
ZRANB2	1:71530223	T/C	Utr3	$8.57x10^{-6}$	94/2/0	973/0/0
ZNF470	19:57089370	G/A	Missense	$8.61x10^{-6}$	96/2/0	986/0/0
CXCR6	3:45987980	G/A	Missense	$8.80x10^{-6}$	93/2/0	945/0/0

Continued on next page...

Gene	Pos	Allele	Variant type	P_{EMMAX}	Case count (ref/het/mut)	Control count (ref/het/mut)
GRIN2A	16:9923264	A/G	Intron	9.26×10^{-6}	93/2/0	938/0/0
LRRC48	17:17907651	C/T	Intron	9.35×10^{-6}	92/6/0	872/7/0
SLC25A15	13:41382553	T/C	Intron	9.48×10^{-6}	94/2/0	963/0/0
AP4B1	1:114444057	A/G	Intron	9.65×10^{-6}	92/2/0	934/0/0
BMP1	8:22059342	G/A	Missense	9.96×10^{-6}	91/2/0	938/0/0
ASZ1	7:117023243	T/C	Intron	9.97×10^{-6}	94/2/0	873/0/0
NPC1	18:21118528	G/C	Missense	1.01×10^{-5}	94/2/0	967/0/0
MUC16	19:9006766	G/A	Missense	1.05×10^{-5}	96/2/0	970/0/0
SALL3	18:76754962	G/T	Missense	1.06×10^{-5}	89/2/0	918/0/0
SACS	13:23908169	T/C	Synonymous	1.07×10^{-5}	93/4/0	972/4/0
MED16	19:873601	C/T	Intron	1.08×10^{-5}	84/6/0	936/9/0
ZNF221	19:44470964	G/A	Missense	1.08×10^{-5}	96/2/0	978/0/0
ODZ4	11:78440539	C/T	Synonymous	1.09×10^{-5}	95/2/0	966/0/0
KCNC4	1:110766392	C/T	Synonymous	1.11×10^{-5}	96/2/0	976/0/0
BCAS3	17:58786681	A/G	Missense	1.12×10^{-5}	96/2/0	977/0/0
PSMC3IP	17:40725490	G/A	Normal	1.15×10^{-5}	95/2/0	982/0/0
C12orf50	12:88381777	T/C	Intron	1.17×10^{-5}	96/2/0	965/0/0
SACS	13:23945189	T/C	Intron	1.17×10^{-5}	94/4/0	956/4/0
LBP	20:36979250	G/A	Synonymous	1.17×10^{-5}	84/3/0	955/1/0
NUAK1	12:106457619	G/A	Utr3	1.24×10^{-5}	96/2/0	981/0/0
CARS2	13:111293895	A/AT	Frameshift	1.35×10^{-5}	93/5/0	967/6/0
CREB3	9:35736017	C/T	Intron	1.41×10^{-5}	89/5/0	955/7/0
DMBT1	10:124358420	C/A	Synonymous	1.42×10^{-5}	96/2/0	983/0/0
DMBT1	10:124333306	A/G	Intron	1.43×10^{-5}	96/2/0	975/0/0
ITK	5:156649836	T/A	Intron	1.44×10^{-5}	92/2/0	898/0/0
PLCG2	16:81969847	C/T	Synonymous	1.49×10^{-5}	96/2/0	962/0/0
DNAJB13	11:73677180	T/C	Normal	1.54×10^{-5}	93/2/0	914/0/0
L3MBTL4	18:6301890	G/A	Intron	1.82×10^{-5}	96/2/0	964/0/0
CIZ1	9:130943027	C/T	Missense	1.83×10^{-5}	96/2/0	952/0/0
SLC9A8	20:48446596	T/C	Intron	1.84×10^{-5}	71/25/1	863/120/3
NA	1:43352995	A/T	Intergenic	1.86×10^{-5}	60/29/7	755/213/13
P2RX3	11:57135466	C/G	Intron	1.91×10^{-5}	94/2/0	946/0/0
NA	1:43352494	A/AG	Intergenic	1.95×10^{-5}	60/31/6	754/214/12
POU2F2	19:42603792	C/T	Intron	1.98×10^{-5}	88/2/0	974/0/0
NAT9	17:72771699	G/A	Intron	1.99×10^{-5}	89/2/0	877/0/0
SENP5	3:196626551	G/A	Missense	2.00×10^{-5}	96/2/0	981/0/0
XPO4	13:21429842	G/A	Missense	2.03×10^{-5}	96/2/0	971/0/0
CMPK1	1:47840778	C/T	Intron	2.10×10^{-5}	92/6/0	964/9/0
LUC7L2	7:139102335	A/C	Missense	2.15×10^{-5}	89/8/0	944/16/0
PITX2	4:111538827	G/A	Utr3	2.17×10^{-5}	96/2/0	989/0/0

Continued on next page...

Gene	Pos	Allele	Variant type	P_{EMMAX}	Case count (ref/het/mut)	Control count (ref/het/mut)
AACSP1	5:178224519	T/C	Intron	2.17x10^{-5}	29/49/20	431/437/70
DGKH	13:42784942	G/T	Intron	2.25x10^{-5}	95/2/0	898/0/0
MYL5	4:674328	C/T	Missense	2.26x10^{-5}	96/2/0	961/0/0
CNTNAP4	16:76495842	C/A	Intron	2.27x10^{-5}	42/43/12	203/465/188
CAMTA1	1:7528032	C/T	Intron	2.36x10^{-5}	86/2/0	943/0/0
MTBP	8:121500554	A/G	Missense	2.36x10^{-5}	96/2/0	932/0/0
AC022498.1	3:187896894	G/A	Utr5	2.39x10^{-5}	86/6/0	921/7/1
ARHGEF2	1:155936308	G/A	Intron	2.64x10^{-5}	93/4/0	963/4/0
SMPDL3A	6:123127486	G/A	Missense	2.65x10^{-5}	96/2/0	972/0/0
TRPM8	2:234869565	C/T	Missense	2.95x10^{-5}	96/2/0	971/0/0
NFE2	12:54686097	G/A	Utr3	2.99x10^{-5}	90/7/0	963/12/0
MTSS1	8:125575041	C/A	Synonymous	3.03x10^{-5}	92/2/0	900/0/0
EVI5L	19:7918234	C/G	Intron	3.07x10^{-5}	89/3/0	951/2/0
THOP1	19:2796197	G/A	Intron	3.24x10^{-5}	88/7/0	942/14/0
KCNG1	20:49620909	G/T	Synonymous	3.29x10^{-5}	88/3/0	917/2/0
NA	3:49828873	C/T	Essential	3.38x10^{-5}	95/2/0	979/0/0
SLC26A6	3:48671025	C/G	Missense	3.41x10^{-5}	95/2/0	968/0/0
KCNMB3	3:178977772	C/A	Intron	3.78x10^{-5}	96/2/0	987/0/0
OR10V1	11:59480759	C/T	Missense	3.88x10^{-5}	89/3/0	980/2/0
SLC22A23	6:3324142	G/A	Synonymous	4.06x10^{-5}	92/2/0	939/0/0
COL6A2	21:47546382	C/A	Intron	4.42x10^{-5}	84/2/0	693/170/15
KIAA0430	16:15688608	T/G	Utr3	4.55x10^{-5}	95/3/0	978/2/0
OR7G3	19:9236967	T/C	Synonymous	4.61x10^{-5}	95/3/0	976/2/0
LGR5	12:71898516	T/C	Intron	5.25x10^{-5}	75/20/2	859/107/0
MAST2	1:46494589	T/C	Synonymous	5.31x10^{-5}	94/3/0	937/2/0
MINK1	17:4797910	G/A	Missense	5.31x10^{-5}	77/14/0	924/44/1
PABPC1L	20:43541529	C/T	Intron	5.63x10^{-5}	94/3/0	863/2/0
GTSF1	12:54849917	C/G	Utr3	5.78x10^{-5}	84/13/0	922/41/0
LATS2	13:21555544	C/G	Intron	5.99x10^{-5}	29/54/11	153/470/243
BTN2A2	6:26390912	C/T	Normal	6.09x10^{-5}	95/3/0	984/1/0
LRP2	2:169999317	G/A	Intron	6.11x10^{-5}	95/3/0	931/2/0
CMPK1	1:47838652	A/G	Missense	6.29x10^{-5}	92/6/0	953/10/0
GOLGA3	12:133374911	C/T	Intron	6.29x10^{-5}	87/3/0	953/2/0
SLC11A2	12:51389573	A/G	Normal	6.33x10^{-5}	93/5/0	960/7/0
UNC79	14:94173247	A/G	Synonymous	6.55x10^{-5}	94/3/0	953/2/0
INPPL1	11:71941033	G/C	Missense	7.21x10^{-5}	86/3/0	967/2/0
SRCAP	16:30740333	A/G	Missense	7.23x10^{-5}	92/6/0	878/10/0
FAM60A	12:31440645	G/A	Synonymous	7.92x10^{-5}	94/2/0	969/1/0
LINGO4	1:151774712	T/C	Missense	7.96x10^{-5}	92/3/0	967/2/0
C1orf123	1:53681725	C/T	Missense	7.97x10^{-5}	90/5/0	967/6/0

Continued on next page...

Gene	**Pos**	**Allele**	**Variant type**	P_{EMMAX}	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
CDADC1	13:49833462	T/C	Intron	8.01×10^{-5}	93/3/0	958/2/0
WDFY4	10:50013402	T/C	Missense	8.03×10^{-5}	26/50/18	464/419/87
ATP6V1B1	2:71162996	A/G	Upstream	8.36×10^{-5}	17/39/7	125/448/358
LCE1C	1:152777901	G/C	Synonymous	8.85×10^{-5}	95/3/0	964/2/0
DNAH8	6:38816574	T/C	Synonymous	9.46×10^{-5}	95/3/0	970/2/0
GIMAP7	7:150217031	G/A	Utr5	9.61×10^{-5}	93/5/0	948/7/0

Table 5.4: Single variant analysis. Association statistics for pneumococcal empyema showing all single nucleotide variants with a P_{EMMAX} <1.0x10^{-4}. Also shown are the allelic counts of cases and controls (Case count and Control count respectively). Twelve variants were significantly associated with pneumococcal empyema (P_{EMMAX}<4.40x10^{-7}) while 130 variants had suggestive evidence of association (P_{EMMAX}<1.0x10^{-4})

Gene	**Pos**	**Allele**	**Variant type**	$\boldsymbol{P_{EMMAX}}$	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
PPP2R2C	4:6331036	G/A	Synonymous	6.54x10^{-11}	63/3/0	962/0/0
FSTL5	4:162577683	T/G	Intron	6.51x10^{-10}	60/4/0	888/2/0
DOPEY2	21:37600028	C/T	Synonymous	8.44x10^{-9}	63/3/0	960/1/0
PTPRC	1:198725059	G/A	Missense	1.54x10^{-8}	63/3/0	912/1/0
ASH2L	8:37985897	A/G	Synonymous	1.89x10^{-8}	0/9/57	0/21/955
SACS	13:23908169	T/C	Synonymous	3.65x10^{-8}	61/4/0	972/4/0
SACS	13:23945189	T/C	Intron	4.11x10^{-8}	62/4/0	956/4/0
ARHGEF2	1:155936308	G/A	Intron	1.35x10^{-7}	61/4/0	963/4/0
THOP1	19:2796197	G/A	Intron	1.40x10^{-7}	56/7/0	942/14/0
UNC79	14:94173247	A/G	Synonymous	3.74x10^{-7}	62/3/0	953/2/0
ACOX2	3:58508217	G/A	Splice_site	3.95x10^{-7}	60/3/0	863/2/0
KIAA0430	16:15688608	T/G	Utr3	3.98x10^{-7}	63/3/0	978/2/0
OR7G3	19:9236967	T/C	Synonymous	4.54x10^{-7}	63/3/0	976/2/0
PKHD1L1	8:110499005	C/T	Missense	7.75x10^{-7}	62/4/0	959/4/0
DPP10	2:116539979	A/G	Splice_site	8.54x10^{-7}	61/5/0	970/8/0
FBN1	15:48779352	C/T	Missense	8.72x10^{-7}	63/3/0	954/2/0
LRP2	2:169999317	G/A	Intron	9.42x10^{-7}	63/3/0	931/2/0
GOLGA3	12:133374911	C/T	Intron	9.80x10^{-7}	57/3/0	953/2/0
FAM174B	15:93173653	G/A	Intron	2.64x10^{-6}	58/3/0	936/3/0
UBQLN3	11:5530705	C/T	Synonymous	2.73x10^{-6}	64/2/0	962/1/0
PLXNA1	3:126753686	C/T	Utr3	2.73x10^{-6}	60/5/0	933/9/0
GAS6	13:114531587	C/T	Missense	3.65x10^{-6}	62/3/0	963/3/0
HEXIM1	17:43226665	G/C	Synonymous	4.17x10^{-6}	61/4/0	965/5/0
CD109	6:74497102	G/A	Missense	5.75x10^{-6}	55/11/0	917/41/0
ACO2	22:41903903	C/T	Synonymous	5.88x10^{-6}	62/2/0	922/1/0
TAB2	6:149731791	G/T	Utr3	5.91x10^{-6}	58/8/0	963/22/0
TAB2	6:149731794	A/T	Utr3	5.91x10^{-6}	58/8/0	963/22/0
DNAJC8	1:28527704	C/T	Utr3	6.15x10^{-6}	62/3/0	977/3/0
PLXNA1	3:126755336	C/T	Utr3	6.16x10^{-6}	53/2/0	974/1/0
SLC6A5	11:20668480	C/T	Synonymous	6.21x10^{-6}	63/3/0	969/3/0
TTC39A	1:51756228	C/G	Missense	6.23x10^{-6}	64/2/0	859/1/0
ZMYM2	13:20633757	A/G	Intron	6.80x10^{-6}	64/2/0	963/1/0
ZMYM2	13:20633540	A/G	Intron	6.93x10^{-6}	63/2/0	936/1/0
NECAP2	1:16778370	G/A	Missense	7.72x10^{-6}	61/3/0	963/3/0

Continued on next page...

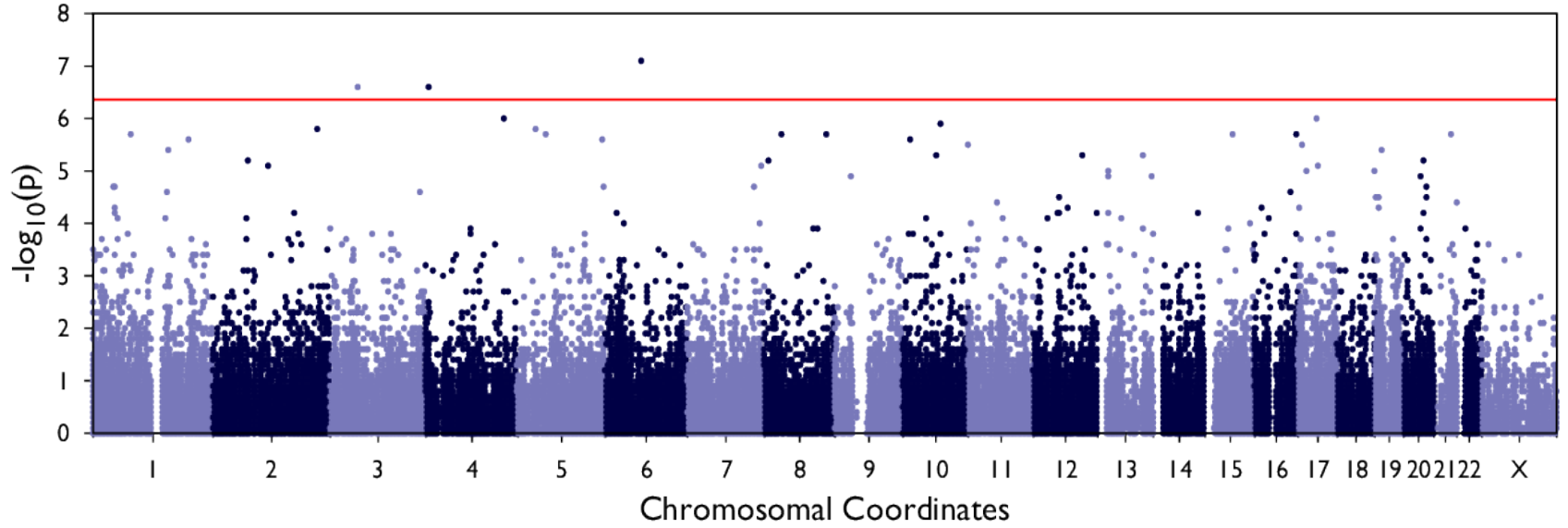
Gene	Pos	Allele	Variant type	P_{EMMAX}	Case count (ref/het/mut)	Control count (ref/het/mut)
TMEM159	16:21190895	G/A	Utr3	$8.43x10^{-6}$	62/2/0	963/1/0
PTBP3	9:114980707	C/T	Downstream	$8.63x10^{-6}$	60/3/0	933/3/0
CYP2C8	10:96818106	T/A	Missense	$8.76x10^{-6}$	63/3/0	958/3/0
KIAA0355	19:34844863	G/T	Utr3	$8.82x10^{-6}$	64/2/0	978/1/0
LUC7L2	7:139060918	A/G	Intron	$8.86x10^{-6}$	64/2/0	976/1/0
LRRC41	1:46763947	G/C	Intron	$9.12x10^{-6}$	63/3/0	942/3/0
ACSF3	16:89167404	T/C	Synonymous	$9.90x10^{-6}$	7/33/24	36/313/624
FCHSD1	5:141023893	G/A	Synonymous	$1.00x10^{-5}$	63/3/0	945/3/0
GZMB	14:25101305	C/T	Missense	$1.05x10^{-5}$	59/2/0	923/1/0
SPG11	15:44898257	A/G	Synonymous	$1.06x10^{-5}$	64/2/0	977/1/0
PARP12	7:139724356	G/T	Utr3	$1.08x10^{-5}$	60/2/0	958/1/0
MYCBP2	13:77748559	T/C	Synonymous	$1.12x10^{-5}$	64/2/0	973/1/0
AURKB	17:8110585	C/T	Missense	$1.12x10^{-5}$	60/2/0	959/1/0
CDH23	10:73556957	G/A	Missense	$1.13x10^{-5}$	62/2/0	964/1/0
FAM227A	22:38993440	A/G	Intron	$1.14x10^{-5}$	62/2/0	929/1/0
VWDE	7:12376614	G/A	Missense	$1.15x10^{-5}$	63/2/0	960/1/0
PTPRB	12:70949014	G/A	Missense	$1.15x10^{-5}$	58/7/0	946/19/0
SCIN	7:12617777	T/A	Missense	$1.15x10^{-5}$	64/2/0	976/1/0
ANK3	10:61833499	G/A	Synonymous	$1.26x10^{-5}$	64/2/0	981/1/0
DGKH	13:42748315	G/A	Intron	$1.27x10^{-5}$	59/6/0	957/15/0
ALPK2	18:56247152	C/A	Missense	$1.29x10^{-5}$	63/3/0	975/3/0
KDSR	18:60998846	G/A	Utr3	$1.32x10^{-5}$	64/2/0	976/1/0
ABCC4	13:95815804	A/G	Intron	$1.40x10^{-5}$	64/2/0	958/1/0
KLHL3	5:137056315	T/C	Intron	$1.41x10^{-5}$	63/2/0	971/1/0
KIAA0586	14:58955600	A/G	Splice_site	$1.43x10^{-5}$	62/3/0	854/3/0
PARPBP	12:102547741	G/A	Missense	$1.43x10^{-5}$	63/2/0	967/1/0
CD34	1:208072455	C/T	Missense	$1.44x10^{-5}$	64/2/0	976/1/0
MUC2	11:1086051	C/T	Missense	$1.44x10^{-5}$	45/2/0	953/1/0
PARPBP	12:102558369	C/A	Missense	$1.47x10^{-5}$	64/2/0	975/1/0
C20orf197	20:58644991	C/T	Utr5	$1.49x10^{-5}$	52/2/0	962/1/0
F13B	1:197030087	C/T	Synonymous	$1.54x10^{-5}$	60/5/0	963/10/0
NGFR	17:47579544	C/T	Synonymous	$1.56x10^{-5}$	52/2/0	938/1/0
NSUN2	5:6623378	T/C	Synonymous	$1.59x10^{-5}$	64/2/0	966/1/0
DDX21	10:70725147	A/G	Synonymous	$1.60x10^{-5}$	64/2/0	973/1/0
CACHD1	1:65124409	C/T	Synonymous	$1.61x10^{-5}$	64/2/0	937/1/0
CDH11	16:65022148	G/A	Missense	$1.68x10^{-5}$	64/2/0	971/1/0
PPIP5K2	5:102513557	T/C	Splice_site	$1.69x10^{-5}$	64/2/0	977/1/0
C5orf42	5:37169169	T/C	Synonymous	$1.76x10^{-5}$	62/4/0	971/6/0
DYNC1I1	7:95614242	C/T	Synonymous	$1.76x10^{-5}$	64/2/0	969/1/0
NA	1:120378772	G/A	Intergenic	$1.77x10^{-5}$	64/2/0	978/1/0

Continued on next page...

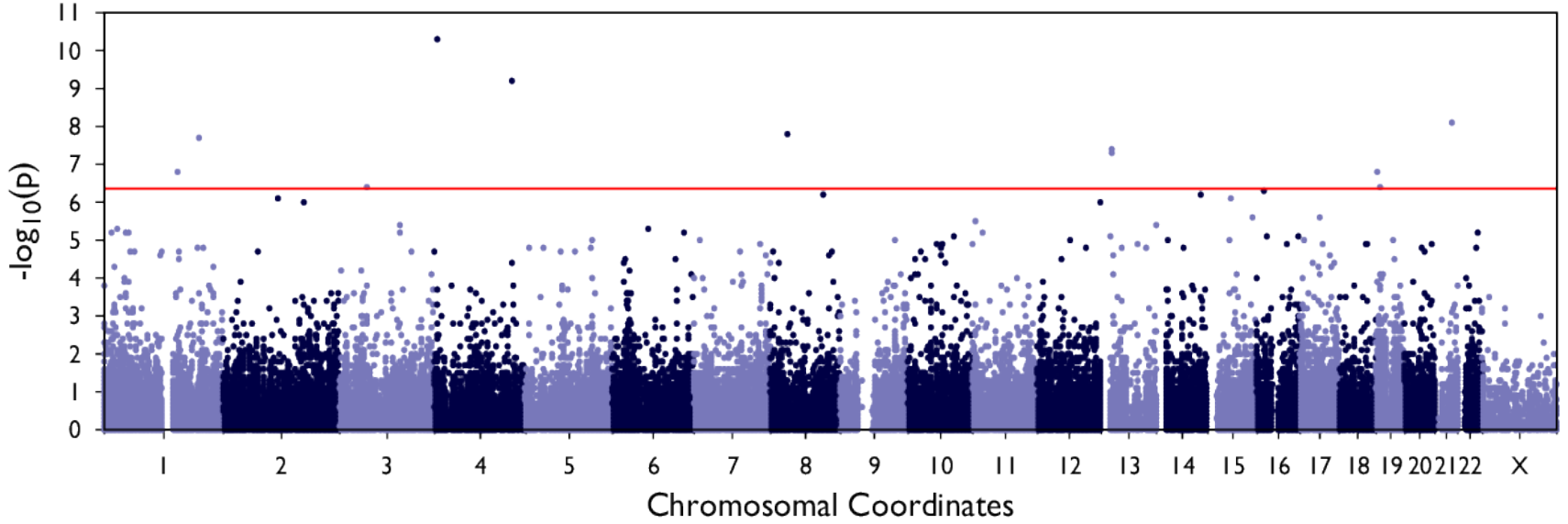
Gene	Pos	Allele	Variant type	P_{EMMAX}	Case count (ref/het/mut)	Control count (ref/het/mut)
LBP	20:36979250	G/A	Synonymous	2.02×10^{-5}	53/2/0	955/1/0
HERC4	10:69692476	C/T	Missense	2.06×10^{-5}	64/2/0	974/1/0
OR10X1	1:158549429	G/A	Synonymous	2.16×10^{-5}	64/2/0	983/1/0
DEFB1	8:6735409	A/C	Utr5	2.19×10^{-5}	62/2/0	966/1/0
PABPC1L	20:43585156	G/A	Intron	2.20×10^{-5}	64/2/0	972/1/0
UVSSA	4:1348888	G/A	Intron	2.23×10^{-5}	44/3/0	889/4/0
ANKRD26	10:27371802	T/C	Intron	2.27×10^{-5}	64/2/0	969/1/0
TSC22D2	3:150127116	C/G	Utr5	2.30×10^{-5}	60/2/0	942/1/0
PCGF1	2:74732932	T/C	Intron	2.31×10^{-5}	63/2/0	972/1/0
GSDMC	8:130764188	C/G	Intron	2.32×10^{-5}	64/2/0	966/1/0
TRIM45	1:117661003	C/T	Missense	2.32×10^{-5}	58/2/0	973/1/0
HSPB11	1:54405745	A/G	Missense	2.32×10^{-5}	61/2/0	971/1/0
HK1	10:71048437	C/G	Utr5	2.33×10^{-5}	64/2/0	961/1/0
GIMAP7	7:150217031	G/A	Utr5	2.45×10^{-5}	62/4/0	948/7/0
PLEKHM1P	17:62782689	T/C	Exon	2.47×10^{-5}	0/2/64	0/1/965
ATP12A	13:25267051	A/G	Intron	2.84×10^{-5}	63/2/0	982/1/0
EYA4	6:133852531	C/T	Utr3	2.96×10^{-5}	62/3/0	976/4/0
SLC11A2	12:51389573	A/G	Splice_site	2.98×10^{-5}	62/4/0	960/7/0
ANKRD31	5:74464817	T/C	Missense	3.03×10^{-5}	63/3/0	971/3/0
ITGA8	10:15646357	A/G	Splice_site	3.10×10^{-5}	64/2/0	969/1/0
FER1L6	8:125061950	G/T	Missense	3.16×10^{-5}	64/2/0	962/1/0
ZNF37A	10:38406392	G/A	Missense	3.23×10^{-5}	64/2/0	981/1/0
OR2H1	6:29430213	G/A	Missense	3.24×10^{-5}	53/10/0	943/40/0
SLC39A11	17:71027833	A/G	Synonymous	3.31×10^{-5}	62/4/0	951/8/0
ARHGEF11	1:156946728	T/C	Intron	3.43×10^{-5}	63/2/0	923/1/0
SAMD4B	19:39869140	C/A	Splice_site	3.70×10^{-5}	61/3/0	924/4/0
SLC39A5	12:56625045	T/C	Utr5	3.83×10^{-5}	0/1/64	0/0/978
POLR3A	10:79769544	G/T	Intron	3.99×10^{-5}	60/5/1	939/20/0
NAF1	4:164067063	T/C	Intron	4.00×10^{-5}	64/2/0	956/1/0
MED24	17:38178627	A/G	Intron	4.10×10^{-5}	28/08/2001	295/462/134
NKD2	5:1038526	A/G	Synonymous	4.26×10^{-5}	0/1/16	0/0/951
ABCA8	17:66871666	A/C	Intron	4.63×10^{-5}	59/6/0	885/16/0
BTN2A2	6:26390912	C/T	Splice_site	4.77×10^{-5}	64/2/0	984/1/0
WDR60	7:158705761	C/T	Missense	4.84×10^{-5}	64/2/0	974/1/0
SLC6A4	17:28524936	T/C	Utr3	4.85×10^{-5}	63/2/0	982/1/0
SLC18A1	8:20005154	G/A	Synonymous	5.13×10^{-5}	60/4/0	968/7/0
NUP133	1:229622124	G/A	Synonymous	5.17×10^{-5}	61/3/0	956/4/0
NBPF3	1:21795388	A/G	Missense	5.46×10^{-5}	24/36/6	191/505/279
ITPR1	3:4887956	G/C	Utr3	5.85×10^{-5}	62/3/0	974/4/0
ITPR1	3:4887957	T/G	Utr3	5.85×10^{-5}	62/3/0	974/4/0

Continued on next page...

Gene	**Pos**	**Allele**	**Variant type**	P_{EMMAX}	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
C3orf49	3:63821914	C/G	Intron	6.12x10^{-5}	0/1/65	0/0/962
FYCO1	3:45965190	G/A	Missense	6.42x10^{-5}	61/5/0	938/12/0
ARHGDIA	17:79825897	A/G	Utr3	6.47x10^{-5}	0/1/46	0/0/971
MIPEP	13:24330786	C/T	Intron	6.76x10^{-5}	55/3/0	958/5/0
TMEM104	17:72834071	A/G	Utr3	6.86x10^{-5}	0/1/13	0/0/957
SFT2D1	6:166733660	T/G	Utr3	6.94x10^{-5}	62/3/0	966/4/0
C10orf113	10:21435340	ACACT/A	Frameshift	6.95x10^{-5}	32/30/4	681/276/18
TDRD9	14:104429453	A/G	Synonymous	7.04x10^{-5}	0/1/65	0/0/983
THOC1	18:252559	G/A	Synonymous	7.54x10^{-5}	0/1/65	0/0/975
PYCR1	17:79891147	T/C	Synonymous	7.61x10^{-5}	0/1/56	0/0/970
CACNA1A	19:13441051	G/A	Splice_site	8.01x10^{-5}	63/2/0	962/2/0
XRCC2	7:152357883	A/G	Intron	8.08x10^{-5}	32/2/0	967/2/0
BNIP2	15:59970318	T/C	Intron	8.22x10^{-5}	59/4/0	938/8/0
SIRT7	17:79872241	A/G	Synonymous	8.30x10^{-5}	0/1/61	0/0/963
MUC16	19:8973988	C/T	Missense	8.41x10^{-5}	63/3/0	956/4/0
USP28	11:113670052	T/A	Synonymous	8.43x10^{-5}	0/1/65	0/0/971
NA	7:98895579	G/A	Intergenic	8.62x10^{-5}	62/3/0	976/4/0
TMEM217	6:37180544	T/C	Utr3	8.92x10^{-5}	62/4/0	974/7/0
ITGA8	10:15647805	C/T	Intron	9.36x10^{-5}	58/5/0	959/13/0
NLRP12	19:54314585	G/A	Intron	9.40x10^{-5}	0/1/65	0/0/967
NA	1:43352494	A/AG	Intergenic	9.50x10^{-5}	39/22/4	754/214/12
ATP13A3	3:194147859	C/T	Missense	9.69x10^{-5}	62/4/0	973/8/0
PRKD2	19:47177913	A/G	Missense	9.70x10^{-5}	0/1/43	0/0/979
PCDHB14	5:140604621	C/T	Missense	9.72x10^{-5}	52/6/0	967/18/0
RNF10	12:120998481	G/A	Intron	9.72x10^{-5}	0/1/65	0/0/931
HGF	7:81358953	C/T	Synonymous	9.74x10^{-5}	61/4/0	946/9/0
SLA/TG	8:134114759	C/T	Intron	9.80x10^{-5}	58/4/0	960/9/0
PINX1	8:10683623	G/A	Intron	9.98x10^{-5}	37/26/2	320/496/112

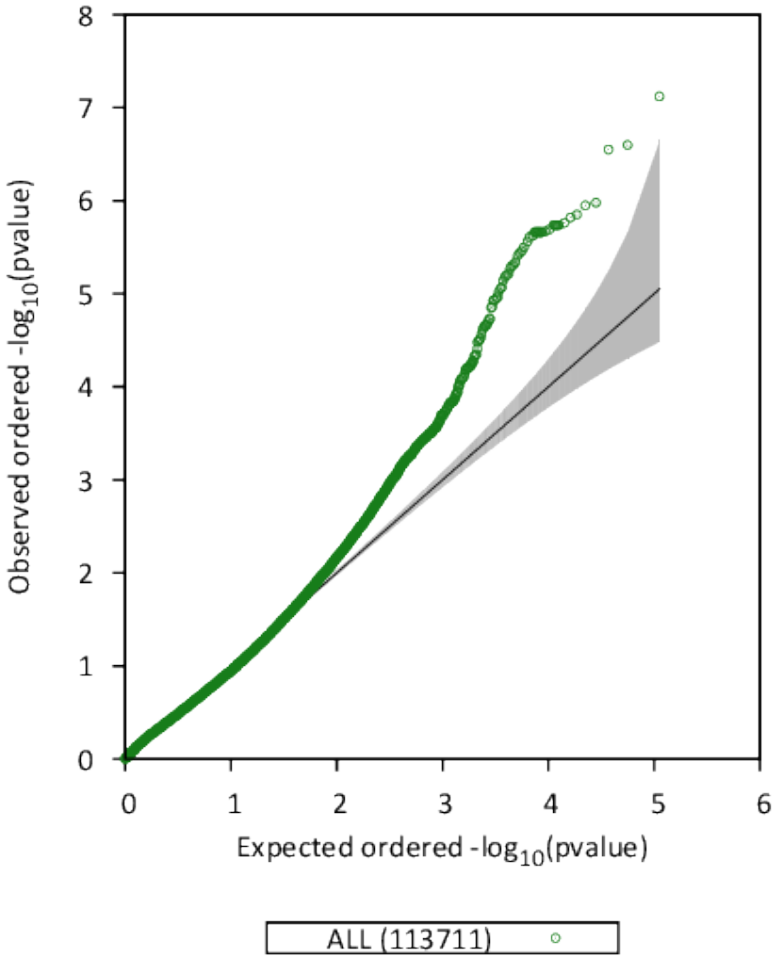


(a) Empyema Manhattan plot

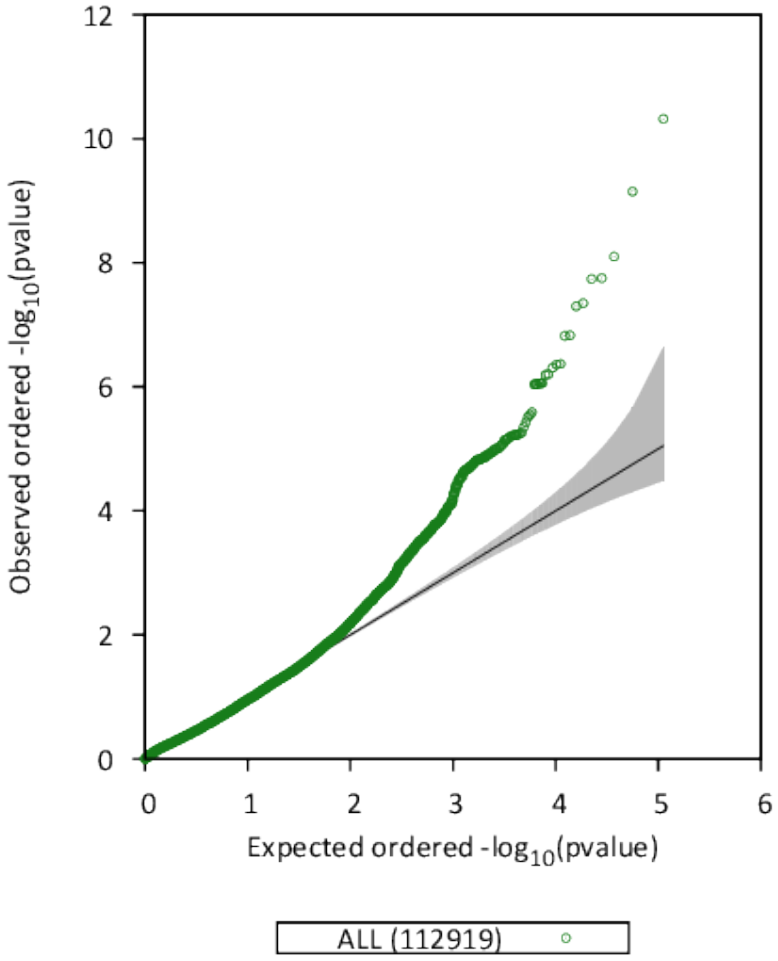


(b) Pneumococcal empyema Manhattan plot

Figure 5.3: Single variant analysis. Manhattan plots showing single variant association analysis for empyema (a) and pneumococcal empyema (b).P-values are from EMMAX mixed model test using a kinship matrix calculated from common variants.



(a) Empyema analysis Q-Q plot



(b) Pneumococcal empyema analysis Q-Q plot

Figure 5.4: Single Variant analysis. Q-Q plots showing single variant association analysis for empyema (a) and pneumococcal empyema subgroup (b). There was no evidence of residual inflation for empyema, with λ=1.04. However, borderline inflation of association statistic was observed in the pneumococcal empyema subgroup analysis, λ=1.06.

5.2.4 Gene burden analysis

5.2.4.1 Empyema

Cumulative effect of rare (<1%) and low frequency (1-5%) variants was assessed by collapsing deleterious variants in a gene and association determined using SKAT [210]. Variants to be aggregated were selected as described in Materials and Methods section 2.7.2. Initial analysis tested for a genic burden of protein altering mutations; these included rare and low frequency nonsense and splice site altering mutations (LoFs) as well as missense mutations predicted as deleterious by either Polyphen-2 or SIFT. There were no genes with an exome-wide significant burden of rare or low frequency variants. Eleven genes had suggestive evidence (P<1.00x10^{-3}) of an association (Table 5.5). Joint analysis of rare and low frequency mutations revealed three additional genes *CD109*, *MINK1* and *RBM44* with suggestive excess of deleterious mutations associating with empyema risk (Table 5.5). The strongest evidence was in the *CD109* gene, P=6.24x10^{-5}. This gene had 10 rare and low frequency variants, comprising four LoFs and six deleterious missense mutations with 31% of cases having at least one deleterious mutation compared to 18% of controls. In total, 14 genes had suggestive evidence for an excess of rare and low frequency protein altering mutations. Analysis of enriched biological patterns in this subset of genes using the program STRING [273] did not reveal enrichment for any GO terms or KEGG pathways. Three gene products had direct protein–protein interactions: *KIF18B*, *AURKB* and *CCNA1*. However, this did not represent a statistically significant enrichment of interactions after genome-wide permutation (P=0.067).

A separate analysis was undertaken to assess for a burden of highly disruptive LoF mutations; this were defined as all rare and low frequency nonsense and splice site altering mutations. There were no genes with a statistically significant excess of LoFs. Four genes *ACOX2*, *CARS2*, *BTN2A2* and *DPP10* had suggestive evidence of an association with empyema. *ACOX2*, *CARS2* and *BTN2A2* each had a single LoF mutation driving the association. The strongest evidence was in *ACOX2* gene P=2.5x10^{-4}, this was driven by a splice site mutation present in four cases and two

controls. The *CARS2* gene had a P=4.34x10^{-4} due to a frameshift variant present in five cases and six controls. The *BTN2A2* association signal was driven by a splice site variant (rs377607368) present in three cases and one control while the *DPP10* association was due to three splice site variants. Two *DPP10* variants were exclusively in controls while one variant (rs141057562) was present in six cases (6%) and only eight controls (0.8%).

5.2.4.2 Pneumococcal empyema

There were no genes that achieved an exome-wide significant association with pneumococcal empyema when testing rare variants alone or jointly testing rare and low frequency mutations in a gene, however, a trending association was evident in the 11 genes listed in Table 5.6. The strongest signal was in the olfactory receptor *OR2H1* gene, P=1.26x10^{-4} with three deleterious missense mutations in the joint analysis of rare and low frequency variants. The primary variant driving the association was a low frequency missense mutation present in 16% of pneumococcal empyema cases and 4% of controls. The association for this gene was not significant when considering only the two rare variants (P=0.97). The missense variant, p.Val223Met, was predicted as deleterious by all algorithms. Further association signals were in the cadherin gene *CDHR4* (P=5.03x10^{-4}) with four rare deleterious variants in three cases (3%) and three (0.3%) controls and in the ubiquitin gene, *UBD* (P=6.18x10^{-4}). The *UBD* association was primarily driven by a deleterious missense mutation present in two cases (2%) but absent in controls, however, the association was less significant with inclusion of low frequency variants with P=0.037 (Table 5.6).

To define the contribution of highly disruptive mutations to pneumococcal empyema, LoF mutations were analysed separately. No genes achieved exome-wide significance, however four genes- *ACOX2*, *PSMC3IP*, *DPP10* and *CDHR4*- showed a trending association with pneumococcal empyema (Table 5.6). The mutations in *ACOX2* and *DPP10* were also significant in empyema and have been described in section 5.2.3.1 above. The *PSMC3IP* gene had one splice site variant present in two cases and absent

Table 5.5: Empyema gene-based analysis. Genes with suggestive evidence of association with empyema (P_{SKAT} <1.0x10^{-3}) in the gene-based analysis. Rare variants(< 1% MAF [1]) were analysed separately as well as jointly with low frequency variants (1-5%)

		Rare (less than 1%)		Rare & low frequency (less than 5%)	
	Gene	**No. of variants**	**P_{SKAT}**	**No. of variants**	**P_{SKAT}**
Pra[2]	*CD109*	5	2.60x10^{-1}	10	6.24x10^{-5}
	MINK1	3	6.00x10^{-1}	4	2.09x10^{-4}
	C1orf123	1	5.11x10^{-4}	1	3.17x10^{-4}
	ACOX2	2	2.80x10^{-4}	2	3.90x10^{-4}
	CARS2	1	6.13x10^{-4}	1	5.92x10^{-4}
	CCNA1	4	7.48x10^{-4}	4	6.01x10^{-4}
	DPP10	3	5.13x10^{-4}	3	6.38x10^{-4}
	RBM44	1	NA	1	6.50x10^{-4}
	AURKB	1	8.28x10^{-4}	1	8.01x10^{-4}
	SALL3	4	6.84x10^{-4}	4	8.82x10^{-4}
	KIF18B	7	6.99x10^{-4}	7	9.11x10^{-4}
	METTL22	4	5.56x10^{-4}	4	9.33x10^{-4}
	BTN2A2	2	6.64x10^{-4}	4	2.00x10^{-2}
	PPP1R15B	4	9.41x10^{-4}	4	1.00x10^{-3}
LoFs[3]	*ACOX2*	1	2.50x10^{-4}	1	2.00x10^{-4}
	CARS2	1	4.34x10^{-4}	1	4.71x10^{-4}
	BTN2A2	1	7.34x10^{-4}	1	4.98x10^{-4}
	DPP10	3	8.22x10^{-4}	3	6.21x10^{-4}

[1]MAF:minor allele frequency
[2]Protein altering: missense mutation predicted as deleterious by either PolyPhen-2 or SIFT, stop-gain, stop-loss, splice site or frameshift mutations
[3]LoFs: Loss of Function mutations i.e stop-gain,stop-loss, splice site or frameshift mutations

in controls while *CDHR4* had two splice site variants, one in two cases and absent in controls and a second identified in only one control.

5.2.5 Pathway enrichment analysis

Perturbations in functionally related genes can lead to the same phenotype. To test this in paediatric empyema susceptibility, the cumulative effect of deleterious variants in functionally related genes was undertaken. In the first instance, a burden of mutations in 299 pathways defined in the KEGG database was tested using SKAT [208, 210]. For empyema, the top canonical pathway for rare variants was "arachidonic acid metabolism" with P=3.19x10^{-3}. There are 58 genes mapped to this pathway in KEGG of which 41 had rare deleterious mutations across cases and controls. For joint analysis of rare and low frequency variants, "mineral absorption" had the strongest signal with P=8.39x10^{-3}, with 34 of the 51 genes mapped to this pathway having a deleterious variant. For pneumococcal empyema, the "arachidonic acid metabolism" pathway had the strongest association for rare variants with P=6.16x10^{-4}. In joint analysis of rare and low frequency variants, "adherens junction" pathway had the strongest association with P=8.24x10^{-3}.

A complementary approach assessed whether a burden of disruptive mutations significantly associated with any particular GO terms [274]. There were 14,698 GO terms requiring a conservative P<3.40x10^{-6}. Genes with deleterious variants were associated with 12,533 biological processes and molecular function GO terms, none achieved statistical significance. In the empyema analysis, “transcription coactivator activity “(GO:0003713) had the strongest association with P=1.70x10^{-4} for variants <1% MAF. In joint analysis with low frequency variants, "negative regulation of keratinocyte proliferation" (GO:0010839) had the strongest association with P=4.02x10^{-5} with eight of the 14 genes with this GO annotation having deleterious variants.

In pneumococcal empyema GO annotation analysis, ‘omega-hydroxylase P450 pathway’ (GO:0097267) had the strongest association for both rare variants alone (P=1.07x10^{-4}) and joint analysis with low frequency variants (P=2.62x10^{-4}). Of the

Table 5.6: Pneumococcal empyema gene-based analysis. Genes with suggestive evidence of association with pneumococcal empyema subgroup (P_{SKAT} <1.0x10^{-3}) in the gene-based analysis. There were 63 pneumococcal empyema individuals for this analysis. Rare variants(< 1% MAF [1]) were analysed separately as well as jointly with low frequency variants (1-5%)

		Rare (less than 1%)		Rare & low frequency (less than 5%)	
	Gene	**No. of variants**	$\boldsymbol{P_{SKAT}}$	**No. of variants**	$\boldsymbol{P_{SKAT}}$
Pra[2]	*OR2H1*	2	9.74x10^{-1}	3	1.26x10^{-4}
	CDHR4	4	8.13x10^{-4}	4	5.03x10^{-4}
	ACOX2	2	9.36x10^{-4}	2	6.26x10^{-4}
	PSMC3IP	1	7.04x10^{-4}	1	6.43x10^{-4}
	RELL1	1	NA	1	8.69x10^{-4}
	DPP10	3	9.00x10^{-4}	3	1.22x10^{-3}
	UBD	1	6.18x10^{-4}	2	3.74x10^{-2}
LoFs[3]	*ACOX2*	1	7.16x10^{-4}	1	5.44x10^{-4}
	PSMC3IP	1	7.47x10^{-4}	1	6.42x10^{-4}
	DPP10	3	9.21x10^{-4}	3	5.69x10^{-4}
	CDHR4	2	9.40x10^{-4}	2	7.45x10^{-4}

[1]MAF:minor allele frequency

[2]Pra:Protein altering i.e missense mutation predicted as deleterious by either PolyPhen-2 or SIFT,stop-gain,stop-loss, splice site or frameshift mutations

[3]LoFs: Loss of Function mutations i.e stop-gain,stop-loss, splice site or frameshift mutations

nine genes associated with this ontology term, eight genes had deleterious mutations across the study samples.

5.3 Discussion

This study provides a systematic analysis of the potential role of rare and low frequency mutations in susceptibility to paediatric empyema. From this study, there is preliminary evidence to suggest that predicted functional deleterious variants in the *CD109* gene may predispose to empyema. In addition to a single deleterious missense mutation significantly over-represented in cases, there is suggestive evidence of a burden of deleterious variants in this gene in empyema cases compared to population controls. This evidence was strongest for all empyema cases regardless of causative pathogen suggesting a phenotype rather than pathogen-specific association. The *CD109* gene encodes a glycosylphosphatidylinositol-anchored protein (GPI) expressed on platelets, activated T cells and endothelial cells. On platelets, *CD109* encodes the human platelet antigen 15 (HPA-15) which has been linked to neonatal alloimmune thrombocytopenia [357]. CD109 protein acts as a negative regulator of TGF-β signalling by promoting internalisation of TGF-β receptors in various cell types, however, its role on activated T cells is yet unknown [358,359]. Of note, the TGF-β signalling pathway has previously been implicated in empyema pathogenesis. Through induction of vascular endothelial growth factor (VEGF), TGF-β promotes permeability and vasodilation of micro-vessels resulting in large pleural effusions [360, 361]. Moreover, due to its critical role in fibrosis, TGF-β is a key mediator in the progression from uncomplicated pleural effusion to loculated pleural empyema [362], a severe complication often requiring prolonged treatment and invasive interventions.

A multifaceted approach was also used to identify pathways and biological functions in genes of similar function disrupted in empyema. From pathway analysis, a potential role for arachidonic acid metabolism was implied; this suggestive association was stronger in the pneumococcal empyema subgroup analysis. Interestingly, when a complementary approach grouping genes by their ontology was applied, cytochrome p450 genes involved in omega-hydroxylation of arachidonic acid (GO:0097267) had the strongest suggestive association with pneumococcal empyema suggesting a possible role for this subset of genes. The cytochrome p450 family of enzymes are involved in

the oxidation of xenobiotics such as drugs as well as endogenous compounds such as omega-6 fatty acids [363]. Cytochrome p450 genes have a critical role in inflammation and polymorphisms in the cytochrome p450 gene *CYP1A1* have previously been reported in association with susceptibility to community acquired pneumonia and related complications including empyema [354,364]. Though this was not replicated in our study (P=0.38), three genes in this pathway *CYP2C8*, *CYP2C9* and *CYP11B2* had statistical evidence just above the suggestive threshold of $P<1.0\text{x}10^{-3}$: *CYP2BC* ($P=5.37\text{x}10^{-3}$), *CYP2C9* ($P=2.08\text{x}10^{-3}$) and *CYP11B2* ($P=1.46\text{x}10^{-3}$) in the genic burden analysis. The role of these genes in empyema needs further investigation.

In the pneumococcal subgroup gene burden analysis, two genes *ACOX2* and *DPP10* had trending association with both empyema overall and pneumococcal empyema specifically. The *ACOX2* gene encodes the branched chain acyl-CoA oxidase which breaks down long chain fatty acids. Expression of *ACOX2* is regulated by *STAT3*, and it has been hypothesised to have a role in surfactant phospholipid homeostasis in pneumocytes [365]. Furthermore, mutations in *STAT3* have been implicated in hyper-immunoglobulin-E syndrome, a primary immunodeficiency characterised by recurrent skin and lung infections [366]. Reduced expression of ACOX2 in STAT3 deficient mice has been linked to bacterial lipopolysaccharide mediated lung injury [365, 367]. The *DPP10* gene on the other hand encodes a dipeptidyl peptidase whose function is to modulate voltage-gated potassium channels. Variants in this gene have been liked to asthma and COPD [368, 369]. The role of these two genes in pneumonia and empyema pathogenesis has not been investigated before.

Finally, despite plausible biological pathways and processes being implicated in pathway analysis, a major shortcoming of this analysis is the incomplete mapping of genes to pathways in existing databases. The *CD109* gene displayed the strongest evidence for both single variant and gene burden analysis for empyema, however, the gene is not mapped to any canonical pathway in KEGG despite extensive evidence supporting a key role for *CD109* in TGF-β signalling [358,359,370–381]. Furthermore, epidemiological evidence supports acute viral infections and prior use of antibiotic and non-steroidal anti-inflammatory drugs as risk factors for empyema [111, 115, 382, 383].

For this study, this information was not available at the time of analysis and may present potential confounders. Moreover, due to heterogenity in causal pathogens for empyema, differences in pathogenicty of the organism would be of concern. However, the *CD109* gene remained significant despite the pneumococcal subgroup analysis, suggestive of the gene mediating disease in a common mechanistic manner rather than a pathogen-specific pathophysiological process.

In summary, the *CD109* gene was implicated as a novel susceptibility locus for paediatic empyema. Furthermore, the *ACOX2* gene was had suggestive association with pneumococcal empyema specifically. Although the pathway analysis had severe shortcomings due to incompletely curated databases, the cytochrome p450 gene family could play a role in empyema pathogenesis. These preliminary results implicate interesting biologically plausible candidates for further follow-up. Replication in independent cohorts or robust functional analysis is necessary.

Chapter 6

Genetic susceptibility to invasive pneumococcal disease

6.1 Introduction

The relative contribution of common and rare genetic variants to IPD susceptibility has been primarily explored through candidate gene studies, mainly involving genes in the Toll-like receptor (TLR)-Nuclear Factor-κB (NF-κB) signalling pathway [304, 322]. Common polymorphisms in various loci such as *MBL2*, *PTPN22*, *NFKBIA*, *NFKBIE*, *NFKBIZ* and *TIRAP* have been identified as having low to modest effects on risk [88, 89, 322, 384]. A role of rare, highly penetrant genetic variants has been highlighted through studies on single-gene inborn errors of immunity predisposing to pyogenic infections in individuals with primary immunodeficiency. Highly penetrant mutations in *IRAK4* and *MYD88* have been described that predispose primarily to pneumococcal infections [228–230]. Moreover, mutations in *NEMO* and *NFKBIA* genes cause the anhidrotic ectodermal dysplasia with immunodeficiency (EDA-ID) syndrome and predispose to infections from a much broader range of pathogens [228, 229, 231, 385]. The majority of these studies have been limited to small numbers of individuals, usually with recurrent infective episodes. Population based studies on the contribution of rare deleterious variants to IPD have not been widely reported. However, a recent population based resequencing study of *IRAK4*, *MYD88* and *IKBKG* genes in adults with IPD did not identify any rare deleterious mutations associating with risk of disease

in these individuals [386]. In order to identify novel genes and mutations predisposing to IPD, an unbiased exome- or genome-wide population based study is necessary.

This chapter describes efforts made to identify novel genes and mutations predisposing to IPD using an exome-wide case control strategy. The IPD cohort included UK Caucasian adults with microbiologically proven IPD while population controls were individuals from the 1958 birth cohort (58BC). Further details on these cohorts including clinical IPD phenotypes and pneumococcal serotype distribution in IPD cases are described in Materials and Methods section 2.2.

6.2 Results

6.2.1 Pilot study

DNA from the IPD cases had been extracted more than 15 years ago and stored at -80°C with intermittent thawing. To evaluate the quality of sequencing data from these samples, a pilot study of 12 samples was undertaken. Of concern was degradation of DNA after long term storage and the minimum amount of DNA required for successful sequencing. The quality of the data was assessed by comparison to sequence data from 270 good quality samples from the sepsis study (Chapter 4) processed in-house using the same pipeline. The results are summarised in Tables 6.1 and 6.2. IPD pilot samples were observed to have lower rates of duplication, an average of 8% compared to 14% in the CAPs. Additionally, the fraction of targeted bases covered at greater than 20X was lower, 76% in IPD compared to 87% in CAP. However, the average depth across all sites was greater than 30X, sufficient to detect heterozygous calls with 95% confidence [172]. From this pilot, it was determined that a minimum of 500ng of DNA with minimal degradation as determined by gel electrophoresis was sufficient for sequencing. A further 93 samples met these criteria and were subsequently sequenced for a total of 105 IPD cases in this study (Figure 6.1).

Table 6.1: Summary of sequence metrics. Comparing pilot IPD samples with exome data from 270 samples of good quality DNA. Both datasets had 64MB of the genome captured using the Nimblegen SeqCap EZ V3.0 kit

Sequencing metrics	**CAP270**	**IPD-12**
Size of target region	64,190,747	64,190,747
Mean depth of target region	65.4X	46.7X
Percentage of target bases covered >= 10x (%)	91.6	88.4
Percentage of target bases covered >= 20x (%)	86.5	75.9
Percentage of target bases covered >= 30x (%)	77.9	60.7
Average duplication rate (%)	14.9	8.2

Table 6.2: Sequence metrics for individual samples in the IPD pilot study. Reads are shown in millions before and after removal of duplicates as well as the proportion of the 64MB targeted region covered at 10X, 20X and 30X coverage.

SampleID	RawReads	MappedReads	Duplicates(%)	Reads afterDedup	MeanCovofTarget	10xTarget	20xTarget	30xTarget
Sample1	53.6M	53.2M	8.4	49.1M	44.5	88.5	75.6	59.5
Sample2	42.7M	42.4M	7.3	39.5M	34.9	85.4	67.1	46.8
Sample3	72.9M	72.5M	8.1	66.9M	57.4	90.6	82.4	71
Sample4	63.5M	63.1M	8	58.5M	51.6	89.6	80	67
Sample5	58.4M	58.1M	8.4	53.5M	48.9	89.2	78.7	64.5
Sample6	45.3M	45.0M	7.7	41.8M	37.7	86	69.4	50.8
Sample7	58.9M	58.6M	8.4	54.0M	48.9	89.3	78.4	64.1
Sample8	62.6M	62.3M	8	57.5M	51.4	89.2	78.7	65.4
Sample9	52.2M	51.9M	8.5	47.7M	43.9	88.2	75	58.7
Sample10	37.8M	37.3M	8.5	34.6M	32.1	84.1	63.4	42
Sample11	63.5M	63.1M	8.4	58.2M	53.2	90.1	80.5	67.6
Sample12	67.7M	67.4M	8.6	61.9M	57	90.4	82.1	70.5

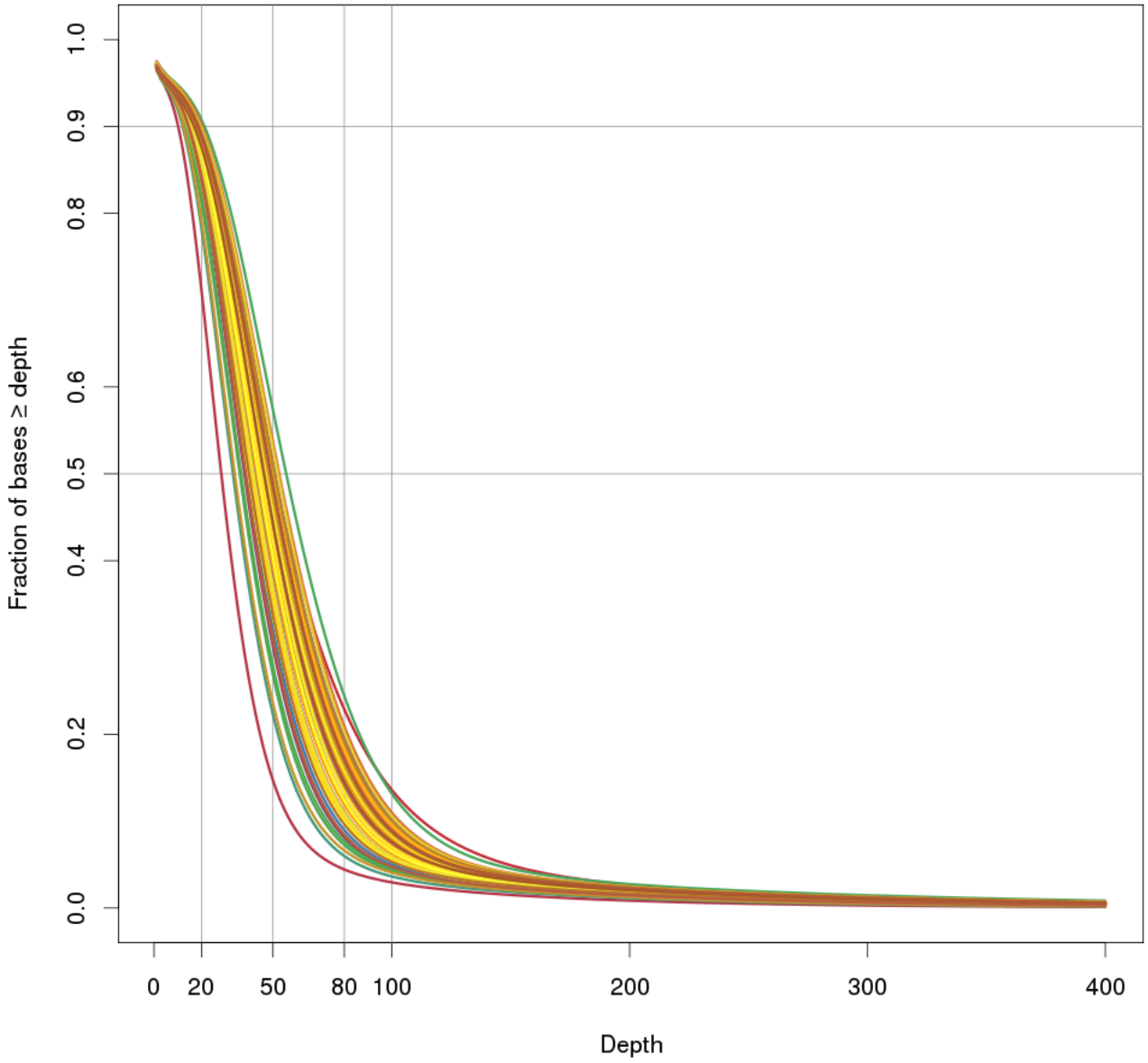


Figure 6.1: IPD coverage. Exome sequencing coverage in 93 IPD samples. The lines show the proportion of the 64MB targeted region covered at various depths per sample. All samples had an average depth of greater than 30X. Over 95% of all regions targeted were covered at least once.

6.2.2 Population structure

To minimise confounding due differences in genetic ancestry, population outliers were detected using MDS implemented in PLINK. Both IPD cases and 58BC controls clustered primarily with other European populations in the HapMap3 dataset (Figure 6.2). Three cases and three controls were identified as either clustering towards populations of African ancestry or of Indian and Mexican ancestry. They were subsequently excluded as outliers leaving 103 cases and 997 controls for all subsequent analysis.

6.2.3 Single variant analysis

Single variant case-control analysis was performed using a linear mixed model(EMMAX) for all variants that passed quality control. There were 221,290 variants in the analysis requiring a $P_{\text{EMMAX}}<2.26 \text{ x}10^{-7}$ for statistical significance. Of these variants, 72,802 were coding changes - either missense or loss of function mutations - of which 59,406 were of less than 5% allele frequency. Sixteen variants appeared to be significantly associated with IPD susceptibility (Table 6.3, Figures 6.3 and 6.4a and b). Nine of these were in non-coding regions; 6 were missense variants while 2 were in UTR regions. Of the 6 significant missense variants, variants in *C1QTNF9B*, *HLA-DQA1*, *MOCOS* and *MMS19* were predicted as benign by PolyPhen-2 and SIFT while the missense mutations in *GJB5* and *SLC11A2* were predicted as deleterious by both PolyPhen2 and SIFT. In *SLC11A2*, the p.Leu464Ile mutation was present in 5 cases and 2 controls ($P_{\text{EMMAX}}=9.68\text{x}10^{-8}$); the gene encodes a divalent ion transporter with a key role in metabolism of iron in macrophages. The *GJB5* encodes a gap-junction protein that facilitates cell to cell trafficking of small molecules; the p.Arg22Cys mutation was present in three cases and absent in controls ($P_{\text{EMMAX}}=1.01\text{x}10^{-7}$). There were 259 additional variants with suggestive association with IPD susceptibility ($P_{\text{EMMAX}}<1.00\text{x}10^{-4}$). Among the variants with significant and suggestive evidence of association, integrated analysis of biological networks using the program STRING [273] did not reveal a statistically significant enrichment of KEGG pathways or GO terms after correction for multiple testing in this list of genes.

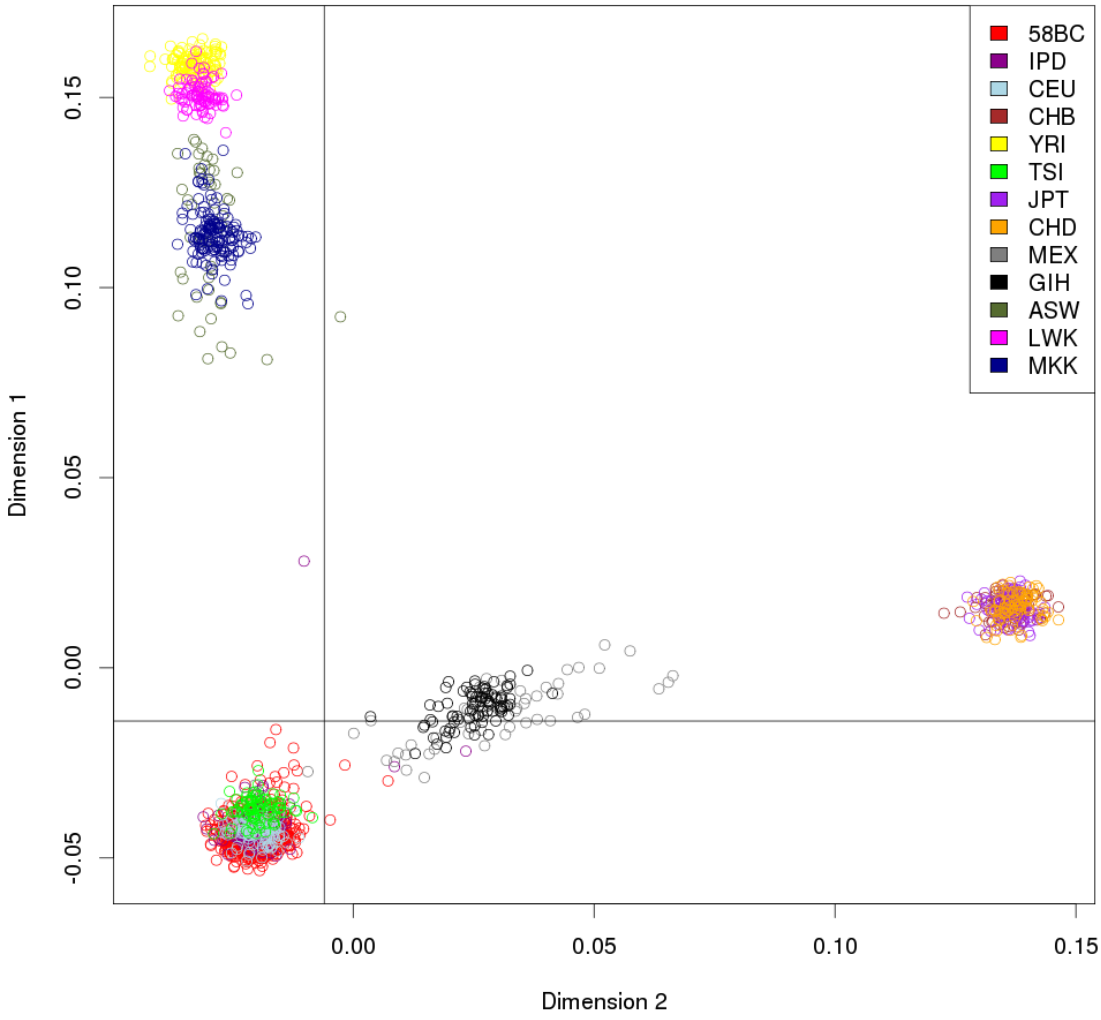


Figure 6.2: Population structure analysis. Multi-Dimensional Scaling(MDS) of the first two components comparing MDS components for populations in HapMap3 with study samples. The two axes represent genetic similarity within and between human populations. Individuals show the greatest genetic similarity within continental populations. Three IPD cases and 3 controls did not cluster with the European populations in HapMap3. **58BC** 1958 birth cohort population controls, **IPD** Invasive pneumococcal disease cases, **CEU** Utah residents with Northern and Western European ancestry from the CEPH collection, **CHB** Han Chinese from Beijing, China, **YRI** Yoruba in Ibadan, Nigeria,**TSI** Tuscani in Italia, **JPT** Japanese in Tokyo, Japan, **CHD** Chinese in Metropolitan Denver, **MXL** Mexican ancestry in Los Angeles, **GIH** Gujarati Indians in Houston, Texas, **ASW** African ancestry from SouthWest USA, Colorado, **LWK** Luhya in Webuye, Kenya, California, **MKK** Maasai in Kinyawa, Kenya

Table 6.3: Single variant analysis. Association statistics for IPD showing all single nucleotide variants with a P_{EMMAX} <1.00 x 10^{-4}. Also shown are the allelic counts of cases and controls (Case count and Control count respectively). Sixteen variants were significantly associated with IPD (P_{EMMAX}<2.26x10^{-7})

Gene	**Pos**	**Allele**	**Variant type**	$\boldsymbol{P_{EMMAX}}$	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
NSMAF	8:59501025	T/C	Intron	2.49x10^{-10}	85/6/2	470/386/115
AC018865.8	2:130807780	G/C	Non-coding transcript	2.03x10^{-9}	46/31/6	235/319/334
MMS19	10:99223674	T/C	Missense	3.31x10^{-9}	98/4/0	979/1/0
CDK11A	1:1654064	A/T	Utr5	3.87x10^{-9}	0/16/76	26/372/449
AP3B2	15:83332052	G/A	Intron	4.03x10^{-9}	99/3/0	865/0/0
CNTN6	3:1371421	C/G	Intron	5.49x10^{-9}	99/3/0	938/0/0
SEP14	7:55874726	T/G	Intron	1.03x10^{-8}	99/3/0	960/0/0
PTPRG	3:62254674	T/C	Intron	1.37x10^{-8}	98/4/0	978/0/0
C1QTNF9B	13:24468329	A/G	Missense	5.20x10^{-8}	77/10/5	413/388/132
SLC11A2	12:51386017	G/T	Missense	9.68x10^{-8}	97/5/0	985/2/0
GJB5	1:35222995	C/T	Missense	1.01x10^{-7}	99/3/0	978/0/0
ZNRD1-AS1	6:29975047	G/T	Non-coding transcript	1.07x10^{-7}	12/38/15	439/379/65
HLA-DQA1	6:32609312	A/C	Missense	1.07x10^{-7}	29/36/11	130/441/328
SMEK2	2:55804414	A/G	Intron	1.35x10^{-7}	99/3/0	982/0/0
SMEK2	2:55812141	C/G	Intron	1.44x10^{-7}	99/3/0	966/0/0
MOCOS	18:33775295	T/C	Missense	1.56x10^{-7}	99/3/0	982/1/0
CLCNKA	1:16360444	C/T	Utr3	2.35x10^{-7}	65/23/0	403/279/178
ATP2C2	16:84441926	T/C	Intron	3.14x10^{-7}	95/7/0	891/7/0
CHUK	10:101949433	T/C	Intron	6.01x10^{-7}	99/3/0	916/1/0
AP3B1	5:77396882	G/A	Intron	6.26x10^{-7}	100/2/0	976/0/0
PHYHIPL	10:61004952	T/C	Synonymous	6.51x10^{-7}	100/2/0	980/0/0
ABI3	17:47299919	A/G	Missense	9.29x10^{-7}	95/6/0	921/7/0
PHF1	6:33381944	A/G	Intron	9.52x10^{-7}	98/4/0	958/1/0
ATRN	20:3584936	C/T	Intron	9.53x10^{-7}	99/3/0	952/0/0
MTCH1	6:36940467	G/A	Synonymous	1.00x10^{-6}	97/5/0	974/1/0
PNPLA5	22:44287073	G/A	Stop-gain	1.02x10^{-6}	89/2/0	904/0/0
PARVB	22:44528849	G/A	Missense	1.03x10^{-6}	100/2/0	974/0/0
GPR144	9:127239197	G/A	Intron	1.07x10^{-6}	99/3/0	941/1/0
NOS2	17:26095966	T/C	Intron	1.12x10^{-6}	95/2/0	964/0/0
C9	5:39311336	A/T	Missense	1.14x10^{-6}	98/4/0	984/2/0
TRPV1	17:3495465	C/T	Missense	1.22x10^{-6}	99/3/0	957/0/0
KIAA1239	4:37448552	C/T	Missense	1.24x10^{-6}	100/2/0	985/0/0

Continued on next page...

Gene	**Pos**	**Allele**	**Variant type**	P_{EMMAX}	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
CELA2B	1:15813836	T/C	Synonymous	1.25×10^{-6}	97/5/0	960/3/0
DNAH10	12:124397851	A/G	Synonymous	1.32×10^{-6}	100/2/0	948/0/0
SCN3A	2:165971916	C/T	Intron	1.33×10^{-6}	99/3/0	943/0/0
TRAPPC9	8:141461103	T/C	Missense	1.34×10^{-6}	100/2/0	977/0/0
PADI3	1:17609560	A/C	Missense	1.36×10^{-6}	98/3/0	962/1/0
BCORL1	X:129148812	C/T	Synonymous	1.48×10^{-6}	94/4/0	969/2/0
ARFGAP1	20:61909483	C/T	Missense	1.59×10^{-6}	97/3/0	973/0/0
RNASE7	14:21511461	G/A	Missense	1.69×10^{-6}	98/4/0	978/0/0
HDAC2	6:114265517	T/G	Missense	1.78×10^{-6}	99/3/0	985/0/0
DNAH8	6:38810444	G/A	Missense	1.82×10^{-6}	100/2/0	975/0/0
EHD1	11:64621005	G/C	Utr3	1.90×10^{-6}	98/2/0	989/0/0
LLGL1	17:18133277	A/G	Missense	1.96×10^{-6}	98/4/0	968/1/0
GDAP1L1	20:42908245	A/G	Utr3	1.99×10^{-6}	100/2/0	989/0/0
PIK3C2B	1:204401314	G/C	Intron	2.05×10^{-6}	100/2/0	950/0/0
CTR9	11:10800279	A/G	Missense	2.11×10^{-6}	100/2/0	980/0/0
FAM186A	12:50748169	T/G	Missense	2.39×10^{-6}	100/2/0	982/0/0
RNF217	6:125379202	A/G	Missense	2.46×10^{-6}	98/4/0	978/1/0
CSNK1D	17:80210443	G/A	Synonymous	2.50×10^{-6}	100/2/0	979/0/0
ID4	6:19840545	C/T	Utr3	2.53×10^{-6}	95/7/0	979/8/0
PER3	1:7880231	A/C	Missense	2.58×10^{-6}	99/3/0	988/0/0
EBNA1BP2	1:43632801	C/T	Intron	2.59×10^{-6}	100/2/0	980/0/0
COL5A1	9:137709578	C/T	Intron	2.60×10^{-6}	97/2/0	938/0/0
ESYT2	7:158528253	C/T	Missense	2.61×10^{-6}	99/3/0	983/0/0
ARNT2	15:80889548	G/A	Utr3	2.64×10^{-6}	99/3/0	981/0/0
PPP1R18	6:30645012	A/G	Utr3	2.65×10^{-6}	99/3/0	986/0/0
ARHGAP17	16:24963698	T/C	Splice_site	2.73×10^{-6}	99/3/0	984/0/0
ZSCAN23	6:28402103	G/A	Utr3	2.85×10^{-6}	99/3/0	987/0/0
MEOX2	7:15651692	T/C	Utr3	2.90×10^{-6}	100/2/0	988/0/0
MEOX2	7:15652068	T/G	Missense	2.97×10^{-6}	100/2/0	981/0/0
RNLS	10:90332611	A/G	Intron	3.05×10^{-6}	100/2/0	978/0/0
SGIP1	1:67209391	G/T	Utr3	3.09×10^{-6}	89/13/0	958/24/0
RIMBP2	12:130882300	G/A	Utr3	3.10×10^{-6}	100/2/0	982/0/0
ACRBP	12:6752681	C/T	Intron	3.11×10^{-6}	100/2/0	964/0/0
C1S	12:7170171	T/G	Intron	3.17×10^{-6}	96/6/0	974/5/0
HTR1A	5:63256231	A/T	Utr3	3.19×10^{-6}	100/2/0	960/0/0
CARS2	13:111299567	G/A	Synonymous	3.24×10^{-6}	99/2/0	963/0/0
UPB1	22:24919586	G/A	Splice_site	3.46×10^{-6}	100/2/0	970/0/0
TUBG2	17:40811997	G/T	Intron	3.51×10^{-6}	28/28/3	195/425/303
DCK	4:71895835	A/G	Utr3	3.53×10^{-6}	97/2/0	980/0/0
C17orf103	17:21144220	C/T	Utr3	3.54×10^{-6}	100/2/0	975/0/0

Continued on next page...

Gene	**Pos**	**Allele**	**Variant type**	P_{EMMAX}	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
SYCP2L	6:10931639	G/A	Intron	3.57x10^{-6}	100/2/0	979/0/0
MYO5A	15:52622505	C/G	Intron	3.61x10^{-6}	100/2/0	929/0/0
AKAP11	13:42876833	G/A	Synonymous	3.82x10^{-6}	100/2/0	991/0/0
P2RX3	11:57135466	C/G	Intron	3.84x10^{-6}	100/2/0	975/0/0
CEACAM3	19:42300572	C/G	Utr5	4.03x10^{-6}	94/3/0	976/1/0
HIST4H4	12:14924004	A/G	Synonymous	4.09x10^{-6}	99/2/0	984/0/0
NEB	2:152492772	G/A	Missense	4.12x10^{-6}	97/5/0	979/5/0
PLD1	3:171395539	G/T	Intron	4.21x10^{-6}	99/3/0	967/0/0
SP110	2:231067305	C/G	Synonymous	4.29x10^{-6}	100/2/0	984/0/0
GAS2L1	22:29708497	C/T	Utr3	4.35x10^{-6}	97/2/0	970/0/0
PKN3	9:131477723	C/T	Missense	4.38x10^{-6}	100/2/0	971/0/0
TNNI3	19:55668060	A/G	Utr5	4.50x10^{-6}	96/4/0	952/3/0
ALMS1	2:73675525	A/G	Missense	4.54x10^{-6}	92/9/1	950/27/0
USP43	17:9631849	A/G	Missense	4.71x10^{-6}	100/2/0	983/0/0
PLEKHO2	15:65147075	G/A	Intron	4.77x10^{-6}	100/2/0	951/0/0
VRK1	14:97327024	C/T	Synonymous	4.78x10^{-6}	100/2/0	983/0/0
IL4I1	19:50398456	C/T	Intron	4.87x10^{-6}	98/2/0	936/0/0
UNC80	2:210707144	A/C	Missense	4.87x10^{-6}	100/2/0	982/0/0
NUP85	17:73231116	C/T	Utr3	4.95x10^{-6}	98/2/0	977/0/0
WDR49	3:167320010	T/C	Missense	5.00x10^{-6}	100/2/0	982/0/0
DYNLL1	12:120934382	C/T	Missense	5.09x10^{-6}	100/2/0	929/0/0
NLRP8	19:56473458	C/T	Missense	5.16x10^{-6}	100/2/0	977/0/0
DCTPP1	16:30435652	G/A	Missense	5.27x10^{-6}	100/2/0	976/0/0
SEMG1	20:43837053	G/T	Missense	5.30x10^{-6}	100/2/0	999/0/0
DUOX1	15:45426437	C/T	Synonymous	5.45x10^{-6}	99/2/0	972/0/0
SEMG2	20:43851101	T/C	Synonymous	5.53x10^{-6}	100/2/0	986/0/0
TBL1XR1	3:176816233	T/C	Intron	5.54x10^{-6}	100/2/0	984/0/0
NLRP8	19:56473648	G/T	Intron	5.58x10^{-6}	100/2/0	945/0/0
RBPJL	20:43938284	G/A	Missense	5.61x10^{-6}	100/2/0	981/0/0
WHSC2	4:1991451	C/T	Synonymous	5.69x10^{-6}	100/2/0	959/0/0
SEMG2	20:43849998	C/T	Utr5	5.69x10^{-6}	100/2/0	974/0/0
NLRP8	19:56490695	C/T	Intron	5.84x10^{-6}	100/2/0	923/0/0
CUL4A	13:113918897	G/A	Utr3	5.96x10^{-6}	97/5/0	991/2/0
SPOCK1	5:136403477	G/A	Synonymous	6.15x10^{-6}	100/2/0	980/0/0
SLC12A7	5:1057572	G/A	Intron	6.23x10^{-6}	99/2/0	965/0/0
WRAP73	1:3552461	C/T	Intron	6.61x10^{-6}	99/3/0	962/0/0
MKL1	22:40813411	G/C	Synonymous	6.65x10^{-6}	91/3/0	960/0/0
RGS12	4:3419060	C/T	Intron	6.68x10^{-6}	99/3/0	855/0/0
VPS13D	1:12321134	A/G	Missense	6.81x10^{-6}	100/2/0	981/0/0
ERN2	16:23702233	G/A	Synonymous	6.86x10^{-6}	93/5/0	951/3/0

Continued on next page...

Gene	**Pos**	**Allele**	**Variant type**	P_{EMMAX}	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
EFHD2	1:15755067	C/T	Intron	6.92×10^{-6}	100/2/0	954/0/0
ATR	3:142269106	T/G	Synonymous	7.33×10^{-6}	99/3/0	981/0/0
CRB2	9:126133660	G/A	Missense	7.71×10^{-6}	51/2/0	920/0/0
USP33	1:78201654	C/T	Intron	7.83×10^{-6}	99/3/0	954/0/0
PLXNA3	X:153688602	G/A	Missense	7.90×10^{-6}	34/2/0	931/0/0
GSK3A	19:42736267	A/G	Synonymous	7.94×10^{-6}	0/2/100	0/0/987
ITGB1	10:33221404	C/T	Intron	8.11×10^{-6}	100/2/0	976/0/0
GFI1	1:92941539	G/A	Utr3	8.43×10^{-6}	100/2/0	966/0/0
EPS8	12:15793810	T/C	Intron	8.54×10^{-6}	96/6/0	956/7/0
EYS	6:64488060	A/G	Synonymous	8.80×10^{-6}	96/6/0	969/6/0
VILL	3:38039809	G/A	Missense	8.93×10^{-6}	100/2/0	979/0/0
CCT3	1:156280470	G/A	Missense	9.30×10^{-6}	100/2/0	983/0/0
AGBL1	15:86800041	G/C	Intron	9.41×10^{-6}	91/11/0	938/19/0
GTF2A1	14:81662596	A/G	Intron	9.51×10^{-6}	100/2/0	979/0/0
ERCC3	2:128046416	G/A	Missense	9.68×10^{-6}	98/4/0	968/3/0
TNIP3	4:122071313	C/T	Missense	9.76×10^{-6}	100/2/0	982/0/0
TEX15	8:30702739	G/C	Missense	1.00×10^{-5}	93/7/0	973/10/0
PLEKHF2	8:96168654	T/A	Utr3	1.04×10^{-5}	100/2/0	993/0/0
ERBB3	12:56493677	A/G	Missense	1.05×10^{-5}	100/2/0	983/0/0
HABP2	10:115348046	G/A	Missense	1.08×10^{-5}	80/21/1	895/76/2
OR9Q1	11:57947909	C/T	Utr3	1.09×10^{-5}	100/2/0	975/0/0
CDAN1	15:43020434	G/A	Missense	1.13×10^{-5}	99/3/0	979/0/0
PHF20	20:34515764	C/T	Synonymous	1.13×10^{-5}	100/2/0	982/0/0
LPHN3	4:62599073	T/C	Synonymous	1.14×10^{-5}	99/3/0	991/1/0
LLGL1	17:18146176	C/T	Intron	1.14×10^{-5}	95/2/0	880/0/0
KCNE4	2:223917983	T/G	Missense	1.15×10^{-5}	11/62/25	61/386/495
IGF1R	15:99507613	C/T	Utr3	1.17×10^{-5}	100/2/0	980/0/0
MMS19	10:99218913	G/A	Intron	1.18×10^{-5}	91/10/0	956/21/1
POLE	12:133218333	C/T	Missense	1.19×10^{-5}	100/2/0	979/0/0
CLCA3P	1:87115944	A/G	Non-coding transcript	1.21×10^{-5}	100/2/0	979/0/0
CCT2	12:69980614	A/G	Intron	1.21×10^{-5}	95/5/0	971/2/0
DDX60	4:169196430	T/G	Intron	1.22×10^{-5}	100/2/0	904/0/0
FBXO3	11:33763250	C/T	Utr3	1.26×10^{-5}	99/3/0	982/2/0
CTNNA2	2:80389729	C/T	Intron	1.26×10^{-5}	100/2/0	986/0/0
COPA	1:160265582	T/G	Missense	1.27×10^{-5}	100/2/0	987/0/0
NEB	2:152484315	C/T	Missense	1.32×10^{-5}	100/2/0	974/0/0
AHNAK2	14:105420116	C/G	Missense	1.39×10^{-5}	100/2/0	978/0/0
CDH1	16:68863590	G/A	Missense	1.43×10^{-5}	100/2/0	971/0/0
RTN1	14:60063472	A/G	Synonymous	1.47×10^{-5}	0/2/100	0/0/958

Continued on next page...

Gene	**Pos**	**Allele**	**Variant type**	P_{EMMAX}	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
FAM171A2	17:42437419	C/T	Missense	$1.51x10^{-5}$	98/3/0	971/0/0
POLR3A	10:79781739	G/A	Synonymous	$1.54x10^{-5}$	94/8/0	967/9/0
TBL3	16:2024208	C/T	Missense	$1.54x10^{-5}$	98/2/0	976/0/0
TAF12	1:28931901	T/A	Missense	$1.60x10^{-5}$	95/7/0	975/7/0
PIM1	6:37142034	C/A	Utr3	$1.67x10^{-5}$	97/3/0	975/2/0
CHRNA2	8:27324870	G/T	Intron	$1.75x10^{-5}$	100/2/0	961/0/0
FCGR2A	1:161481530	G/A	Intron	$1.80x10^{-5}$	79/23/0	908/71/0
ADRA1A	8:26627756	C/A	Synonymous	$1.82x10^{-5}$	100/2/0	980/0/0
WFDC12	20:43752953	G/A	Intron	$1.83x10^{-5}$	86/16/0	908/45/0
LAMA5	20:60889897	G/A	Synonymous	$1.91x10^{-5}$	78/4/0	862/3/0
BRD9	5:871547	C/T	Intron	$1.91x10^{-5}$	95/6/1	948/15/0
SLC17A6	11:22399187	C/T	Synonymous	$1.95x10^{-5}$	100/2/0	978/0/0
LYZL1	10:29580723	A/G	Intron	$1.96x10^{-5}$	97/0/0	758/149/20
ACP5	19:11685942	G/A	Missense	$1.98x10^{-5}$	99/3/0	971/0/0
DENND4A	15:65988636	A/G	Intron	$2.02x10^{-5}$	100/2/0	964/0/0
CA12	15:63618425	G/A	Utr3	$2.06x10^{-5}$	99/3/0	969/2/0
LRRC66	4:52862020	C/G	Missense	$2.07x10^{-5}$	99/2/0	978/0/0
MLL2	12:49444960	G/T	Missense	$2.10x10^{-5}$	94/2/0	956/0/0
NAA16	13:41894867	C/T	Synonymous	$2.14x10^{-5}$	97/5/0	978/2/0
ZNF469	16:88505635	A/G	Synonymous	$2.17x10^{-5}$	8/53/20	51/368/517
LRBA	4:151817649	A/G	Intron	$2.17x10^{-5}$	98/4/0	974/2/0
TTBK2	15:43038399	C/T	Missense	$2.17x10^{-5}$	100/2/0	979/0/0
ZNF692	1:249150022	G/A	Synonymous	$2.20x10^{-5}$	90/4/0	974/4/0
ZAN	7:100359748	G/C	Intron	$2.23x10^{-5}$	100/2/0	882/0/0
CHD1L	1:146758054	G/A	Missense	$2.33x10^{-5}$	97/5/0	979/6/0
SEZ6L2	16:29889591	G/A	Intron	$2.41x10^{-5}$	98/3/0	972/0/0
ARHGAP22	10:49687689	C/T	Synonymous	$2.44x10^{-5}$	100/2/0	883/0/0
TEX11	X:70093173	C/T	Intron	$2.45x10^{-5}$	100/2/0	983/0/0
TMCO4	1:20027423	T/G	Intron	$2.48x10^{-5}$	22/59/19	414/445/99
PKHD1	6:51824680	G/A	Synonymous	$2.49x10^{-5}$	100/2/0	966/1/0
AGR2	7:16832624	G/A	Intron	$2.51x10^{-5}$	98/4/0	944/2/0
OSTN	3:190930454	T/C	Intron	$2.54x10^{-5}$	98/4/0	973/2/0
PDCL3	2:101188227	C/G	Missense	$2.56x10^{-5}$	97/5/0	984/6/0
NCOA2	8:71039122	C/T	Missense	$2.64x10^{-5}$	100/2/0	985/0/0
CYP2C18	10:96495207	G/T	Utr3	$2.65x10^{-5}$	100/2/0	987/0/0
COL15A1	9:101824569	C/G	Missense	$2.66x10^{-5}$	100/2/0	985/1/0
GLI1	12:57863433	C/T	Missense	$2.68x10^{-5}$	97/5/0	932/5/0
ERMN	2:158176774	C/T	Utr3	$2.72x10^{-5}$	93/9/0	974/15/0
FAM47E	4:77184996	C/T	Missense	$2.81x10^{-5}$	98/4/0	973/4/0
IL31RA	5:55204221	G/A	Missense	$2.91x10^{-5}$	100/2/0	992/0/0

Continued on next page...

Gene	**Pos**	**Allele**	**Variant type**	P_{EMMAX}	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
ANPEP	15:90334146	G/T	Intron	2.91x10^{-5}	98/4/0	949/2/0
PCDHA4	5:140188401	A/G	Synonymous	2.97x10^{-5}	34/30/6	211/494/282
ABHD14A-ACY1	3:52023020	C/T	Missense	3.02x10^{-5}	97/2/0	976/0/0
SEMA3B	3:50308833	C/T	Non-coding transcript	3.05x10^{-5}	99/2/0	984/0/0
RPL5	1:93301736	T/G	Intron	3.06x10^{-5}	98/4/0	947/4/0
CPED1	7:120911388	A/C	Synonymous	3.17x10^{-5}	98/4/0	984/4/0
HRNR	1:152191918	G/A	Synonymous	3.19x10^{-5}	100/2/0	995/0/0
RARS2	6:88255413	A/G	Synonymous	3.21x10^{-5}	97/5/0	984/2/0
KRT78	12:53242641	C/T	Missense	3.39x10^{-5}	84/7/0	943/10/0
SMCP	1:152850906	G/A	Utr5	3.44x10^{-5}	99/3/0	983/1/0
NHSL2	X:71360679	T/C	Utr3	3.45x10^{-5}	99/0/3	849/6/0
SPATA21	1:16731609	G/C	Intron	3.54x10^{-5}	85/11/0	919/28/0
SEPSECS	4:25153606	T/C	Synonymous	3.54x10^{-5}	99/3/0	985/2/0
RIMS2	8:104943552	A/C	Missense	3.59x10^{-5}	96/6/0	988/5/0
HEATR8	1:55182448	G/A	Intron	3.64x10^{-5}	96/5/1	962/9/0
SLC24A6	12:113756897	G/A	Intron	3.74x10^{-5}	99/3/0	962/2/0
UBE2V2	8:48974121	G/A	Utr3	3.88x10^{-5}	100/2/0	986/0/0
CADPS	3:62751699	C/T	Intron	3.91x10^{-5}	88/14/0	934/31/0
CREB3L3	19:4157105	C/G	Synonymous	4.07x10^{-5}	95/7/0	846/6/0
EFCAB13	17:45447802	G/A	Splice_site	4.18x10^{-5}	90/11/0	971/16/0
HSPA12B	20:3726524	C/T	Intron	4.21x10^{-5}	99/3/0	967/2/0
XPR1	1:180853225	A/G	Utr3	4.34x10^{-5}	99/3/0	980/1/0
EPHA7	6:93967903	T/G	Missense	4.46x10^{-5}	99/3/0	982/1/0
ZNF135	19:58579679	C/A	Missense	4.66x10^{-5}	97/5/0	983/7/0
MED24	17:38178627	A/G	Intron	4.81x10^{-5}	47/25/0	350/470/139
TTC7B	14:91121268	A/T	Intron	4.84x10^{-5}	92/3/0	937/2/0
PNPLA3	22:44328870	C/T	Missense	4.93x10^{-5}	99/3/0	984/2/0
TTBK1	6:43220632	C/G	Splice_site	5.02x10^{-5}	95/2/0	915/0/0
KRT78	12:53240025	C/T	Missense	5.04x10^{-5}	95/7/0	977/10/0
SGSM2	17:2276388	T/C	Intron	5.06x10^{-5}	95/2/0	943/1/0
MMS19	10:99240838	A/T	Intron	5.15x10^{-5}	92/9/0	925/22/1
ADAT1	16:75642673	G/A	Intron	5.26x10^{-5}	90/11/0	878/30/0
KRT78	12:53240101	C/T	Intron	5.34x10^{-5}	95/7/0	971/10/0
DHX8	17:41568515	G/A	Intron	5.40x10^{-5}	98/4/0	972/5/0
HRH3	20:60791422	C/T	Synonymous	5.50x10^{-5}	19/32/1	698/232/19
B3GNT7	2:232263318	G/A	Synonymous	5.53x10^{-5}	33/1/0	967/0/0
ANGPT2	8:6372298	G/A	Missense	5.65x10^{-5}	97/5/0	975/5/0
SF3A2	19:2244646	C/T	Intron	5.65x10^{-5}	79/11/0	926/30/0

Continued on next page...

Gene	**Pos**	**Allele**	**Variant type**	P_{EMMAX}	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
OR2L3	1:248224489	A/G	Missense	5.85x10^{-5}	77/20/2	470/361/20
ACTN1	14:69351064	G/A	Intron	5.88x10^{-5}	93/8/1	929/18/0
KRT78	12:53238387	C/T	Missense	5.96x10^{-5}	95/7/0	951/10/0
KIF13A	6:17788001	A/C	Splice_site	5.99x10^{-5}	98/4/0	982/3/0
ZBED4	22:50278584	C/T	Missense	6.00x10^{-5}	98/3/0	957/1/0
XYLT2	17:48434513	C/T	Missense	6.03x10^{-5}	96/4/0	969/4/0
TMCO4	1:20027404	G/A	Intron	6.14x10^{-5}	30/55/14	495/407/66
CSHL1	17:61988585	C/G	Utr5	6.19x10^{-5}	99/3/0	990/2/0
T	6:166580188	G/A	Synonymous	6.23x10^{-5}	26/49/4	578/326/34
CBX1	17:46153319	A/G	Intron	6.26x10^{-5}	100/2/0	946/1/0
HEATR8	1:55117520	G/A	Intron	6.67x10^{-5}	70/24/2	769/91/9
KIAA1324L	7:86569438	A/G	Splice_site	6.67x10^{-5}	96/6/0	979/5/0
MPP5	14:67759390	G/A	Missense	6.71x10^{-5}	96/6/0	974/6/0
ZC3H3	8:144557739	G/A	Missense	6.73x10^{-5}	85/17/0	883/53/1
QKI	6:163876260	A/G	Intron	6.74x10^{-5}	96/5/0	959/8/0
ESPL1	12:53675412	T/C	Splice_site	6.79x10^{-5}	99/3/0	979/2/0
ADAM21P1	14:70714162	A/G	Non-coding transcript	6.81x10^{-5}	2/54/38	25/295/671
NDFIP1	5:141511942	T/G	Intron	6.91x10^{-5}	99/3/0	962/1/0
KCNK18	10:118969144	T/C	Synonymous	7.10x10^{-5}	99/3/0	978/3/0
TPCN2	11:68846198	G/A	Splice_site	7.11x10^{-5}	63/3/0	981/2/0
PRLR	5:35072685	C/T	Missense	7.14x10^{-5}	98/4/0	978/2/0
GCH1	14:55331990	T/C	Intron	7.31x10^{-5}	99/3/0	958/2/0
KCNA1	12:5026667	G/C	Utr3	7.51x10^{-5}	98/4/0	986/5/0
C6orf89	6:36867376	C/T	Synonymous	7.64x10^{-5}	97/5/0	976/4/0
OR10H1	19:15918647	G/T	Synonymous	7.75x10^{-5}	58/30/2	425/337/193
EPS8	12:15803904	A/G	Synonymous	7.75x10^{-5}	96/6/0	956/9/0
SPTA1	1:158639328	C/T	Missense	8.04x10^{-5}	98/4/0	986/4/0
CPS1	2:211518862	T/G	Intron	8.21x10^{-5}	98/3/0	973/3/0
KIAA1191	5:175777766	T/C	Intron	8.21x10^{-5}	99/3/0	968/2/0
RYR3	15:33922154	A/G	Missense	8.39x10^{-5}	99/3/0	976/2/0
EXOC4	7:133749118	C/T	Missense	8.41x10^{-5}	100/2/0	974/0/0
PMS2CL	7:6777043	C/T	Non-coding transcript	8.43x10^{-5}	98/4/0	991/3/0
DDX11	12:31249225	T/C	Intron	8.50x10^{-5}	51/34/5	353/425/172
SPAG9	17:49075832	G/T	Missense	8.53x10^{-5}	100/2/0	984/1/0
ELN	7:73475491	T/C	Intron	8.89x10^{-5}	98/4/0	975/3/0
CACNA2D4	12:1919556	G/A	Intron	8.90x10^{-5}	95/7/0	943/8/0
MOG	6:29627222	C/A	Missense	8.97x10^{-5}	97/4/1	977/7/0
PCNX	14:71577222	G/T	Utr3	9.04x10^{-5}	97/5/0	970/5/0

Continued on next page...

Gene	**Pos**	**Allele**	**Variant type**	P_{EMMAX}	**Case count** (ref/het/mut)	**Control count** (ref/het/mut)
PHLPP1	18:60563225	A/G	Missense	9.14×10^{-5}	94/8/0	945/15/0
IGSF21	1:18688779	G/T	Intron	9.14×10^{-5}	90/2/0	944/0/0
POLR1B	2:113304572	T/G	Intron	9.26×10^{-5}	99/3/0	971/2/0
WDR11	10:122640433	A/G	Intron	9.34×10^{-5}	98/4/0	971/5/0
C10orf68	10:32977956	G/A	Intron	9.36×10^{-5}	93/8/0	962/18/0
CCNE1	19:30303617	C/T	Synonymous	9.39×10^{-5}	98/2/0	975/1/0
CAPN3	15:42698177	C/G	Intron	9.61×10^{-5}	98/3/0	986/2/0
NRAP	10:115406674	G/T	Splice_site	9.76×10^{-5}	97/4/0	970/5/0
LGR6	1:202284047	T/G	Intron	9.81×10^{-5}	92/3/0	959/3/0

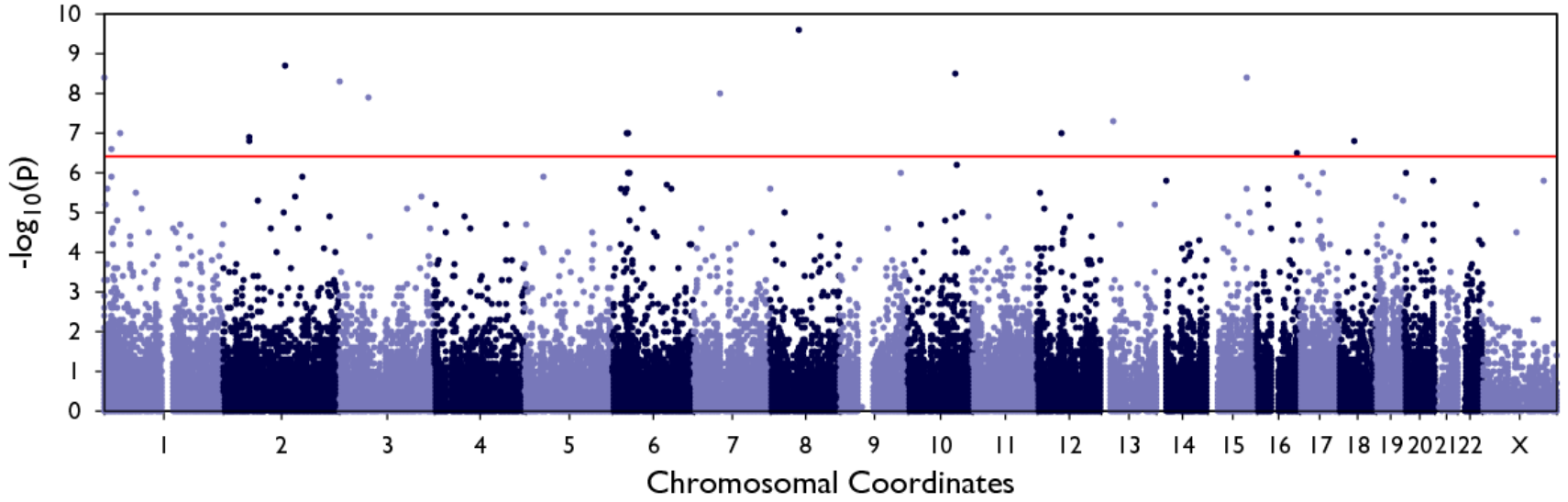
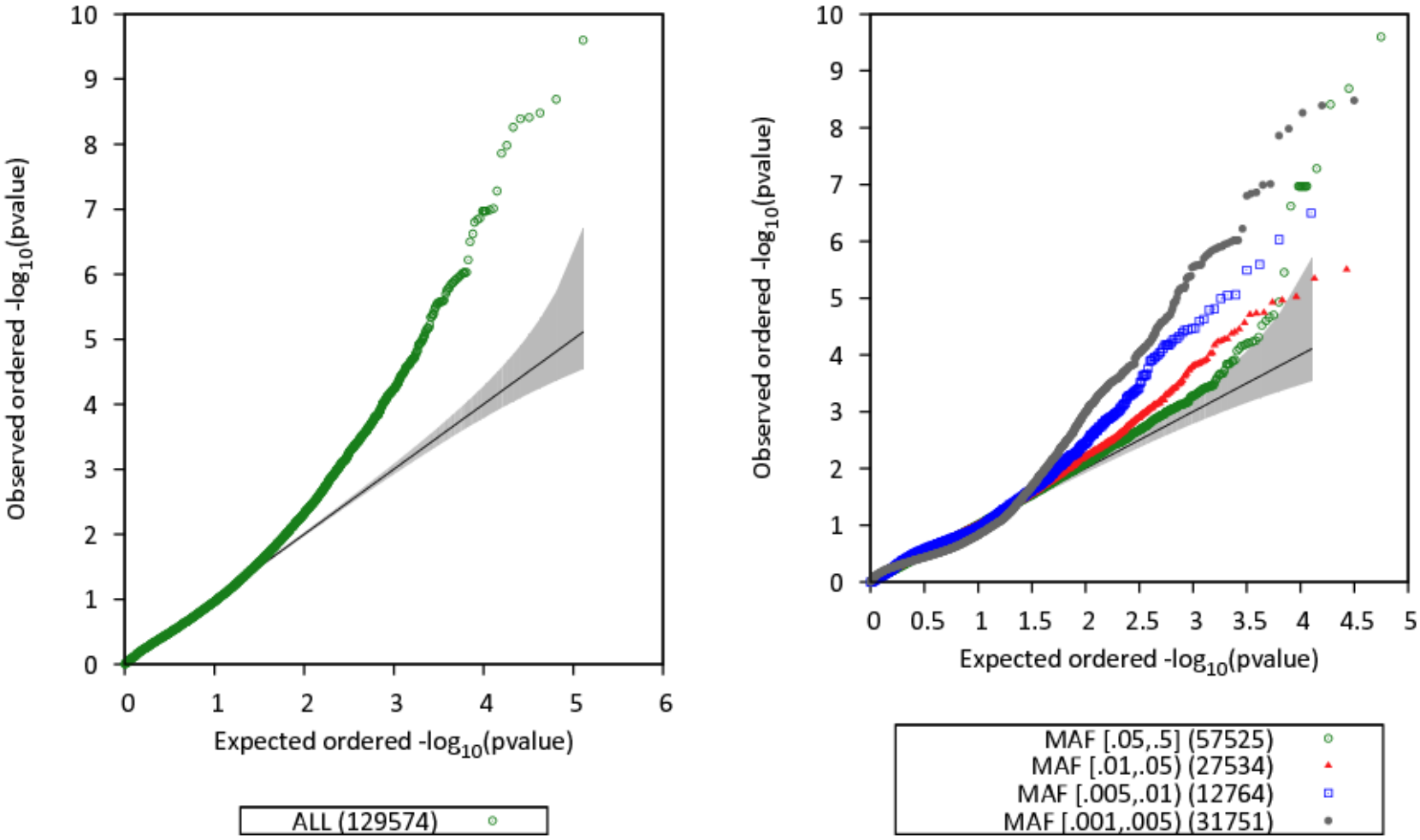


Figure 6.3: Manhattan plot for IPD single variant association analysis. P-values were derived from a mixed model with a kinship matrix calculated from common variants.



(a) IPD association analysis Q-Q plot

(b) IPD association analysis stratified Q-Q plot

Figure 6.4: Single variant analysis. Q-Q plots showing single variant association analysis for IPD. Genomic control revealed inflation of test statistic (λ =1.11) despite stringent QC, primarily due to singletons and doubleton (b).

6.2.4 Gene burden analysis

To assess for a cumulative effect of rare (<1% frequency) and low frequency (1-5%) variants on disease risk, variants were aggregated across each gene as described in Materials and Methods section 2.7.2. Frequency thresholds were set at 1% for rare variants and 5% to include both rare and low frequency variants in the analysis.

There were no genes that achieved statistical significance while collapsing just the rare variants or jointly aggregating rare and low frequency variants. Though not reaching statistical significance, the strongest association using the two collapsing strategies was in the *RNASE7* gene encoding a ribonuclease peptide with broad anti-microbial properties (Table 6.4). The mutation driving the association in the *RNASE7* gene was rare; this p.Val104Met missense mutation was predicted as deleterious by both PolyPhen-2 and SIFT and was identified in four IPD cases and absent in controls. In total, 42 genes had suggestive evidence of a burden of rare and low frequency protein altering mutations predisposing to IPD (Table 6.4). Pathway analysis of these set of genes using the program STRING did not reveal significant enrichment in any of the KEGG pathways after correction for multiple testing. The most significant uncorrected p-values were for 'complement and coagulase' (GO:0004610, P=0.0086) driven by *C9* and *KLKB1* genes and 'lysosome' (GO:0004141, P=0.025) due to *SLC11A2* and *CTSE* genes. Analysis of enriched GO biological processes did not reveal significant enrichment after correction for multiple testing. Analysis of protein-protein interactions in this set of genes revealed a statistically significant excess of interacting proteins ($P_{\text{CORRECTED}} = 4.60\text{x}10^{-3}$). Eight direct protein-protein interactions were observed in the 42 genes, a significant enrichment over three interactions expected from random networks of genes with the same binding degree sampled from the genome. One interaction (*RIMS2* and *UNC13C*) included two genes critical in exocytosis. Two interactions involved *EFCAB13*, *STK17B*, and *PSKH2*, genes with a role in phosphorylation of myosin light chains. Two further interactions involved key genes in the development and structure of ciliary and flagellar microtubules (*TEKT4*, *DRG2* and *KIF6*). A second burden analysis strategy involved aggregating

highly disruptive LoFs across a gene. This analysis did not reveal any genes with a statistically significant excess of LoFs associating with IPD. Eleven genes however, showed suggestive evidence of an association (Table 6.4). The strongest suggestive association was for low frequency variants in *EFCAB13* gene with $P = 6.51 \text{x} 10^{-6}$ where 15% of cases had disruptive mutations compared to 5% of controls.

Table 6.4: IPD gene-based analysis. Genes with suggestive evidence of association with IPD (P_{SKAT} <1.0x10^{-3}) in the protein altering gene-based analysis. Rare variants(<1% MAF[1]) were analysed separately as well as jointly with low frequency variants (1-5%)

	Rare (<1%)			Rare & low frequency (<5%)	
	Gene	**No. of variants**	**P_{SKAT}**	**No. of variants**	***P_{SKAT}***
Pra[2]	*RNASE7*	1	2.7510x^{-5}	1	1.5510x^{-5}
	KIAA1324L	2	4.0910x^{-5}	2	1.7310x^{-5}
	MPP5	3	4.7510x^{-5}	4	1.5110x^{-4}
	KRT78	6	5.3410x^{-5}	7	2.5510x^{-4}
	CTSE	4	1.1510x^{-4}	4	1.0410x^{-4}
	RIMS2	7	1.2610x^{-4}	7	1.3010x^{-4}
	GJB5	1	1.6610x^{-4}	1	1.2910x^{-4}
	ABI3	1	1.7010x^{-4}	1	1.8910x^{-4}
	TEX15	5	1.7910x^{-4}	6	1.5810x^{-4}
	SHROOM3	6	2.0910x^{-4}	7	3.0510x^{-3}
	STK17B	1	2.2110x^{-4}	2	1.7210x^{-4}
	COL15A1	9	2.6510x^{-4}	10	5.3510x^{-4}
	UNC13C	6	3.1210x^{-4}	6	3.4810x^{-4}
	SLC11A2	4	3.6110x^{-4}	5	5.7310x^{-4}
	MYL4	1	3.8910x^{-4}	1	1.0510x^{-4}
	DHDH	2	3.9110x^{-4}	2	1.8310x^{-4}
	ERCC3	4	4.0510x^{-4}	4	5.3110x^{-4}
	SLC30A9	1	4.1610x^{-4}	2	7.5510x^{-3}
	MOG	2	4.4110x^{-4}	3	2.1210x^{-3}
	CDAN1	6	4.9910x^{-4}	6	4.5110x^{-4}
	GLI1	5	5.0910x^{-4}	5	7.1910x^{-4}
	PCDHB1	2	5.6310x^{-4}	5	0.20691
	EPHA7	6	5.7410x^{-4}	7	2.8510x^{-2}
	CEACAM7	2	5.9310x^{-4}	2	6.2110x^{-4}
	UPB1	2	6.2310x^{-4}	2	6.0910x^{-4}
	DRG2	1	6.2410x^{-4}	1	8.6810x^{-4}
	KAT2B	3	6.7710x^{-4}	3	1.0710x^{-3}
	C9	4	8.2610x^{-4}	4	9.6010x^{-4}
	KLKB1	2	8.4510x^{-4}	2	8.1910x^{-4}
	ANGPT2	2	8.4810x^{-4}	2	5.9910x^{-4}
	OVGP1	2	8.8810x^{-4}	3	1.2810x^{-2}
	NAT8	1	9.0210x^{-4}	1	6.5910x^{-4}
	PSKH2	2	1.2910x^{-3}	3	7.3210x^{-4}
	KDM4E	1	6.1910x^{-3}	2	5.3310x^{-4}
	STK11IP	10	3.4310x^{-2}	11	8.7810x^{-4}
	SYNE1	28	3.6010x^{-2}	32	3.1410x^{-4}
	CCDC73	2	0.11514	3	1.2910x^{-4}
	KIF6	4	0.12954	5	2.1010x^{-4}

Continued on next page...

	Rare (less than 1%)			Rare & low frequency (<5%)	
	Gene	**No. of variants**	**P_{SKAT}**	**No. of variants**	***P*$_{SKAT}$**
	HABP2	8	0.28011	10	2.2210x^{-4}
	EFCAB13	1	0.50515	5	2.1310x^{-5}
	ZNF366	7	1	8	3.7110x^{-4}
	TEKT4	0	NA	1	6.7910x^{-4}
LoF[3]	*EFCAB13*	1	0.51	3	6.51x10^{-6}
	KIAA1324L	1	4.73x10^{-5}	1	6.01x10^{-5}
	MYL4	1	4.20x10^{-4}	1	6.89x10^{-5}
	KIF6	2	0.08	3	1.52x10^{-4}
	KIF13A	1	6.79x10^{-4}	1	5.76x10^{-4}
	ANGPT2	1	2.41x10^{-4}	1	6.21x10^{-4}
	UPB1	2	6.05x10^{-4}	2	6.33x10^{-4}
	OVGP1	2	8.63x10^{-4}	2	6.57x10^{-4}
	KLKB1	1	7.37x10^{-4}	1	7.11x10^{-4}
	NUP37	2	0.05	3	7.95x10^{-4}
	FAM47E-STBD1	1	1.47E-03	1	8.98x10^{-4}

[1]MAF:minor allele frequency

[2]Pra: Protein altering i.e missense mutation predicted as deleterious by either PolyPhen-2 or SIFT,stop-gain,stop-loss, splice site or frameshift mutations

[3]LoFs: Loss of Function mutations i.e stop-gain,stop-loss, splice site or frameshift mutations

6.2.5 Pathway enrichment analysis

A key characteristic of complex disease is genetic heterogeneity, whereby disruption in different genes with related function may lead to the same phenotype. To investigate this further in the context of IPD, genes were first grouped according to pathways or groups of genes annotated as having the some molecular function or participating in the same biological processes (GO terms) and analysis of deleterious mutations in defined groups was then performed.

In the first instance a burden of deleterious mutations was assessed for each of the 299 canonical pathways defined in the KEGG database, requiring a $P<1.67\text{x}10^{-4}$ for statistical significance [275]. There were no pathways with a statistically significant excess of rare deleterious mutations after correction for multiple testing for the 299 pathways. The top pathway was "Hippo signalling pathway" with $P=2.33\text{x}10^{-3}$. Of the 154 genes mapped to this pathway, 100 had rare deleterious mutations in the study samples. In the joint analysis of rare and low frequency variants, the top canonical pathway was "Tight junction" with $P=9.07\text{x}10^{-4}$. Of the 138 genes mapped to this pathway, 93 genes had protein altering mutation in the study samples. A complementary approach assessed for a burden of mutations in genes with the same ontology in the GO database associating with risk of IPD [274]. There were 14,698 gene ontology terms tested requiring a conservative $P<3.40\text{x}10^{-6}$ after Bonferroni correction. Though not reaching statistical significance, the strongest association for rare variants was in genes annotated with "response to manganese ion" (GO:0010042) term with the least p-value of $9.42\text{x}10^{-6}$. Seven of the nine genes with this annotation contained rare deleterious mutations. Joint analysis of rare and low frequency variants genes annotated with the GO term "membrane disruption in other organism" (GO:0051673) had the least p-value of $P=9.40\text{x}10^{-6}$. The *RNASE7* gene was the only gene with this annotation.

6.3 Discussion

This chapter presents insights from a population based study on the potential contribution of rare large effect genetic variants to risk of invasive pneumococcal disease.

Preliminary results from this study implicate deleterious variants in the *RNASE7* gene with increased risk of pneumococcal disease in adults. RNase 7 is a member of the RNase A family of antimicrobial peptides and is highly expressed on the skin and on mucosal surfaces with the highest expression levels detected on oesophageal mucosa and urinary tract [387, 388] [389]. Expression of RNase 7 on these surfaces is constitutive but significant upregulation occurs in response to IL-17, IFN-γ, TNF-α, IL-1β or by pathogen associated molecular patterns such as LPS on Gram negative bacteria and peptidoglycan on the cell wall of Gram positive bacteria [387, 390, 391]. A dose-dependent killing activity against *Staphylococcus aureus* as well as the respiratory pathogen *Pseudomonas aeruginosa* have been widely reported [391–394], however studies on the antimicrobial activity targeting pneumococcus are lacking. Moreover, orthologs of this gene in the mouse have not been identified, creating a challenge for *in vivo* studies. Antimicrobial activity against pneumococcus however has been demonstrated for RNase5 peptide, another member of the RNase A family [395]. A recent study demonstrated that basal cells on epithelium provide protection on mucosal surfaces primarily through production of RNase 7 [396]. RNase 7 activity is highly dependent on electrostatic interactions with the negatively charged pathogen cell surfaces [397, 398]. In the mutation identified in this study, a hydrophobic valine residue is substituted with a polar methionine residue. While hydrophobic to polar substitutions might not lead to large electrostatic differences, the sulphur molecule present in the methionine residue might alter RNase 7 interaction with microbial membranes. Nasopharyngeal colonisation by pneumococcus is a prerequisite for invasive disease and invasion is partly dependent on an inadequate host response. Modulated RNase 7 activity could result in impaired resistance to pneumococcal invasion. However replication in an independent cohort or functional validation would be necessary to establish a role for this gene in IPD with certainty.

To assess whether genes of similar function were enriched in rare variants associating with disease, the results highlighted two pathways: the "hippo signalling pathway" and the "tight junction pathway". The key gene common to both pathways and with supportive evidence in the gene burden analysis (Table 6.4) was the *PALS1* gene (also known as *MPP5*). The PALS1 protein is part of a three-protein complex that is essential for cell polarity [399] . This complex is crucial in the establishment and development of tight junctions in epithelial cells which are essential in determining permeability of the airways [400]. Indeed, PALS1 deficiency has been implicated in tight junction defects including reduced trans-epithelial resistance [401–403]. Moreover, there is evidence to suggest that SARS-coronavirus, which causes severe acute respiratory syndrome, binds PALS1 impairing formation of tight junctions and adherens junctions and consequently allowing the virus to breach the epithelial barrier and disseminate to systemic circulation [404]. A compromised integrity of the barrier function in epithelial cells of airways has been implicated as one of the pathogenic mechanisms in invasive pneumococcal disease [405, 406]. At the time of this analysis the *RNASE7* gene was not mapped to any of the 299 pathways in the KEGG database highlighting the incompleteness of this database and potential for other pathways to strongly associate with IPD.

Single variant analysis implicated a deleterious variant in *SLC11A2* an important transporter of iron in phagocytic cells. A similar protein SLC11A1- also known as NRAMP1- has been implicated in susceptibility to intracellular pathogens such as *Mycobacterium* and *Salmonella*; *SLC11A2* has not been implicated in susceptibility to any pathogens [407]. Due to its critical function in phagocytes however, an association with susceptibility to infection would not be improbable.

From both single variant and burden analysis, deleterious mutations in complement component 9 (*C9*) are implicated in IPD susceptibility. C9 is critical component of the membrane attack complex (MAC) of the complement system. Deficiencies in early complement components C1 to C4 are associated with increased risk of infection from encapsulated bacteria such as pneumococcal infections [408]. However, deficiencies in late complement components C5-C9 that form the membrane attack complex are

primarily associated with increased risk of infection from *Neisseria meningitidis* [408]. Increased risk of pneumococcal infections in C9 deficiency has not been reported.

This study provides preliminary evidence implicating novel genes and pathways of the innate immune system in susceptibility to invasive pneumococcal disease. Both antimicrobial peptides and the barrier function of epithelial cells are key first lines of defence against invading pathogens. Although the role of antimicrobial peptides in mucosal immunity is well established, genetic variation in these genes associating with disease susceptibility has not been previously described. Furthermore, evidence to suggest processes involving differentiation of epithelial cells and formation of tight junctions as possible determinants of pneumococcal invasion may provide further insight into IPD pathogenesis.

Chapter 7
Discussion

Despite the relatively widespread asymptomatic colonisation by pathogens described in this thesis, invasive bacterial disease occurs in just a minority of colonised individuals, a host genetic component could in part explain this variability. In the last few decades, studies on identification of susceptibility loci primarily focused on a relatively small number of pre-selected genes. Advances in next generation sequencing technologies have now made it possible to assess the human genome in an unbiased manner potentially enabling discovery of novel loci underlying disease susceptibility. In all studies described in this thesis, an exome-wide sequencing approach has been employed in order to identify rare variants underlying meningococcal disease, sepsis arising from community acquired pneumonia, pleural empyema in children and finally invasive pneumococcal disease in adults (described in Chapters 3, 4, 5 and 6 respectively).

Invasive meningococcal disease is characterised by a rapid onset with a high mortality rate and significant neurological sequelae in survivors. With the occurrence of frequent epidemics in the meningitis belt of sub-Saharan Africa as well as over 2,000 infections annually in the UK, a greater understanding of the role of genetic risk factors is necessary. In Chapter 3, I assessed the role of rare genetic variants in adult UK Caucasians admitted to ICU with invasive meningococcal disease. Using prioritization strategies modelled on early, successful exome sequencing studies of Mendelian diseases, key genes have been implicated in disease susceptibility. The role of the complement system in protection against *Neisseria meningitidis* was further highlighted with the discovery and subsequent functional validation of a novel missense

mutation in the *CFP* gene giving rise to type I properdin deficiency in one individual. Although the association between this mutation and disease is challenging to prove statistically in a single case, the prior evidence linking properdin deficiency to this phenotype together with the novel functional studies present compelling biological evidence that this novel variant is causative. Furthermore, although not functionally validated, highly disruptive loss of function mutations in *CD4* and *ZAP70* genes were implicated in meningococcal disease susceptibility. Although genetic mutations in these genes have not been previously reported in IMD susceptibility, both genes encode key proteins in the adaptive immune system and impaired function of these proteins has been shown to lead to increased susceptibility to infections. Firstly, ZAP70-related SCID is inherited in an autosomal recessive manner and typically presents early in life with reduced numbers of CD8 T cells, normal levels of dysfunctional CD4 T cells and increased susceptibility to bacterial, viral and fungal infections. However, atypical late-onset ZAP70 associated immunodeficiency has been described [409, 410]. In one individual with late onset ZAP70-associated immunodeficiency, a splice-region variant resulted in an alternative transcript and a rapidly degraded protein [409]. Individuals with partial ZAP70 defects presented with recurrent LRTIs and skin infections as well as low antibody titres after pneumococcal polysaccharide vaccine. In this study one individual had a heterozygous splice-region variant in this gene. Secondly, CD4 T cell depletion in HIV-positive individuals has been reported to correlate with a five-fold increased risk of meningococcal disease compared to HIV-positive individuals with higher CD4 T cell counts [411]. CD4 T cells are crucial in the development of memory to *Neisseria meningitidis* and indeed it has been known for well over four decades that antibody levels correlate with protection against invasive meningococcal disease [73]. Furthermore, it has been functionally demonstrated that poor proliferation of CD4 T cells correlated with failure of meningococcal serogroup C conjugate vaccine in adolescence [412]. Further investigation of these genes in meningococcal disease is warranted, we therefore intend to perform similar analysis in additional meningococcal cases collected as part of GenOSept from other countries in Europe.

Preliminary evidence presented in Chapter 4 implicates signaling via small GTPases in susceptibility to sepsis arising from community acquired pneumonia. Strikingly, four genes encoding Rho activator and downstream effector proteins had suggestive evidence of association. Rho GTPases are crucial in the normal functioning of the immune system with roles ranging from migration of phagocytes and apoptosis. Furthermore, Rho GTPases have been implicated in bacterial invasion processes, for example, the intestinal intracellular pathogen *Salmonella typhimurium* produces proteins that activate GTPases leading to cytoskeleton rearrangement and promoting pathogen internalization [346, 347]. These pathogenic processes that are characteristic of invasive serotypes of microbes have been associated with the activation of the host immune system [413]. Indeed the activation of Rho GTPases by *Salmonella* has been demonstrated to result in the activation of the NOD1 signaling pathway [349]. The functional validation of these genetic associations in the Rho GTPase pathway might provide further insight into the complex and poorly understood mechanisms leading to sepsis.

Novel candidate genes are reported in Chapters 5 and 6 as susceptibility loci for paediatric pleural empyema and adult IPD. In Chapter 5, the *CD109* gene encoding a negative regulator of TGF-β signaling presents an interesting candidate for further follow-up. In animal models, TGF-β levels have been shown to correlate with fibrosis during empyema and administration of anti-TGF-β antibodies markedly reduced progression to empyema in rabbits [362, 414]. Genetic studies of variation in TGF-β signaling genes and *CD109* in particular in relation to empyema have not been reported before. Interestingly, despite over half of empyema cases having *Streptococcus pneumoniae* as the identified causative pathogen, there lacked an obvious overlap of susceptibility loci when compared to adult IPD presented in Chapter 6. This is perhaps suggestive of differences in underlying pathophysiological processes resulting in empyema in children and IPD in adults despite the same causative pathogen. Also of note is the potential role of low frequency "Goldilocks' alleles" in disease susceptibility. These variants are usually excluded from GWAS as well as in the majority of studies on rare variants using a <1% frequency threshold. From these

results, the joint analysis of low frequency and rare variants in addition to analysing rare variants separately proved to be the best approach.

The innate immune system as a first line of defence against invading pathogens is highly emphasized in this project. Preliminary evidence from pathway analysis in Chapters 5 and 6 incriminates genes encoding proteins crucial in maintaining epithelial junction integrity in disease susceptibility. Coupled with complement deficiency reported in Chapter 3 as well as the antimicrobial peptide *RNASE7* in IPD susceptibility, the central role of this branch of immunity in preventing pathogen invasion is underscored.

Pathway analysis provides a biologically intuitive method of analysis in infectious diseases as mutations in genes with related biological function result in related phenotypes. Key examples are mutations in the Toll-like receptor (TLR)-Nuclear Factor-κB (NF-κB) signalling pathway [304, 322, 345] and risk of pyogenic infections, complement system deficiencies and increased risk of meningococcal infections and IL-12/Interferon-gamma (IFN-γ) pathway mutations and risk of infection with intracellular pathogens [226, 233–236]. Pathway analysis implicated epithelial barrier function integrity in empyema and IPD, highlighting the important role of this innate immune system component. However, these methods of analyses are still underdeveloped despite significant potential to identify hitherto unknown pathophysiological processes. Current databases such as KEGG, REACTOME and BIOCARTA are incomplete. Novel susceptibility genes from this analysis such as *ARHGAP18*, *CDC42BPG* and *CD109* are yet to be mapped to pathways in the KEGG database despite significant scientific literature. Indeed, of the approximately 20,000 genes in the genome, at the time of writing only 6,964 genes are mapped to the 299 pathways in the KEGG database.

The utility of using shared population controls whose sequence data is freely available for academic use is explored in Chapter 4. Although the cost of sequencing has dramatically reduced in recent years, it remains prohibitive for very large sample sizes. Shared population controls are well validated for use in GWAS with genotype data from individuals recruited into the 1958 birth cohort and National blood service

widely used as Caucasian population controls. Large scale sequencing projects such as 1000genomes and UK10K provide freely available sequence data for use in academic research, enabling research groups to focus available resources on disease cases. Chapter 4 highlights the significant challenges encountered during analysis of rare variants in these shared controls. A key finding from this initial stage of analysis was the importance of using the same bioinformatics pipeline for analysis of rare variants in order to reduce artefacts in resulting datasets. Furthermore, as rare variants are more likely to be of much recent origin and population specific, ethnicity matching of cases and controls should be strictly adhered to.

Next generation sequencing still has significant potential in uncovering rare variants underlying disease. As reported in Chapter 3, the small sample size in this study (13 individuals) precluded robust statistical analysis as is recommended in order to reduce false positive findings in genetic studies [153]. However, this study provides further evidence that the study of a small number of individuals with "extreme phenotypes" coupled to robust functional validation can uncover causal mutations. Indeed 21% of known human primary immunodeficiencies were originally identified from the study of single individuals [216]. In studies involving rare phenotypes where the accumulation of large samples sizes might prove difficult, the methods described in this chapter prove useful.

Of note however, next generation sequencing typically uncovers a significant number of plausible causal mutations whose biological significance is open to interpretation, a phenomenon described as the "narrative potential" of the human genome [415]. With the exception of the *CFP* gene described in Chapter 3 whose role in disease susceptibility had been previously documented, novel and biologically plausible candidates from the sepsis, empyema and IPD studies are reported in this thesis. These findings might be fortuitous primarily due to the small sample sizes, but nevertheless the biological plausibility of the novel genes and pathways implicated in these analyses merit further investigation through attempted replication in ongoing sample collections and/or study of their functional effects.

The inter-individual variation in susceptibility to infectious diseases is dependent on host, microbial and environmental factors. Partitioning the relative contribution of each has been a challenging task. Undoubtedly, in line with accumulating evidence and findings presented in this thesis, host genetics have a key role in infectious disease susceptibility. However, the complex interplay between the host and disease-causing pathogen in determining susceptibility to disease has largely been ignored. *Neisseria meningitidis* and pneumococcal strains have been shown to vary in virulence, both displaying evidence of hypervirulent clones [55, 416]. Furthermore, it has been long known that properdin deficient individuals present with meningococcal disease due to uncommon serogroups yet the precise mechanism for this is yet to be determined. For phenotypes such as sepsis and empyema, the causal organisms greatly vary and in many cases are undetermined. This heterogeneity confounds genetic analysis as pathogen-specific pathophysiological processes such as differences in elicited immune responses are not disentangled. With decreasing costs of sequencing as well as widespread use of MLST in characterising pathogens, future studies in genetics of infectious diseases should therefore consider the study of host and infecting pathogen genomes in parallel. This would greatly advance our understanding of the interactions that occur ultimately leading to disease.

In conclusion, results from these studies present preliminary evidence to indicate a role for rare variants in novel genes and pathways as determinants of infectious disease susceptibility. Further insight into the poorly understood mechanisms of bacterial invasion and disease pathophysiology might be gained through experimental assays or replication in independent studies.

Chapter 8
Exome sequencing

The three major steps in exome sequencing are library preparation, cluster generation and finally sequencing.

8.1 Library preparation

Library preparation involves the random fragmentation of DNA. Sonication is the most widely used method of DNA fragmentation; alternative protocols that make use of transposons are preferred in situations where the starting material is limited. The ends of the DNA are repaired and adenylated followed by ligation of 5' and 3' specific adapters to both ends of the DNA fragments. Fragments of at least 100 base pairs are then size selected. The fragments are denatured and one end of a single strand fragments hybridizes to a glass slide known as a flow cell. The flow cell surface will have oligonucleotides which are complementary to the adaptors on the fragments. A complementary strand is synthesised on the flow cell surface and the original strand is washed off.

8.2 Cluster generation

In cluster generation, the fragment folds over and the adaptor at the other end hybridises to an oligo on the flow cell surface, forming a bridge. A complimentary strand is then synthesised forming a double stranded bridge. The bridge is then denatured resulting in two strands hybridised to the surface. This process, known

as bridge amplification, is repeated resulting in thousands of identical fragment. It occurs simultaneously across all fragments generating millions of clusters. Finally, the reverse strands are cleaved and washed off.

8.3 Sequencing

Sequencing begins with annealing of a sequencing primer followed by polymerase extension to produce a read. Sequence is by synthesis, whereby for each step, attachment of a nucleotide releases a fluorescent signal characteristic to each nucleotide. This process occurs in parallel across thousands of fragments. In the case of paired end sequencing, a process similar to bridge amplification is repeated although in this step the forward strand is cleaved and washed off. Sequencing is repeated for the reverse strand.

Chapter 9

KEGG pathway analysis for genes with case only loss of function mutations in meningococcal cases

Table 9.1: KEGG pathways with case only mutations in 13 meningococcal disease cases. There were 94 pathways with genes that had LoF[1]only in cases, 24 of these were immune related.

GO-id[2]	Term	BriteHierachy	No.of Genes	$P_{uncorrected}$	$P_{corrected}$(FDR)
5340	Primary immunodeficiency	Immune diseases	2	$6.74x10^{-3}$	$6.27x10^{-1}$
3030	DNA replication	Replication and repair	2	$6.74x10^{-3}$	$6.27x10^{-1}$
5033	Nicotine addiction	Substance dependence	2	$8.32x10^{-3}$	$6.27x10^{-1}$
2010	ABC transporters	Membrane transport	2	$8.74x10^{-3}$	$6.27x10^{-1}$
3420	Nucleotide excision repair	Replication and repair	2	$1.14x10^{-2}$	$6.57x10^{-1}$
4020	Calcium signaling pathway	Signal transduction	3	$2.35x10^{-2}$	1.00
4350	TGF-beta signaling pathway	Signal transduction	2	$3.02x10^{-2}$	1.00
4727	GABAergic synapse	Nervous system	2	$3.46x10^{-2}$	1.00
4640	Hematopoietic cell lineage	Immune system	2	$3.62x10^{-2}$	1.00
5032	Morphine addiction	Substance dependence	2	$3.77x10^{-2}$	1.00
4723	Retrograde endocannabinoid signaling	Nervous system	2	$4.51x10^{-2}$	1.00
4660	T cell receptor signaling pathway	Immune system	2	$4.68x10^{-2}$	1.00
5162	Measles	Infectious diseases: Viral	2	$7.58x10^{-2}$	1.00
4514	Cell adhesion molecules (CAMs)	Signaling molecules and interaction	2	$8.30x10^{-2}$	1.00
4145	Phagosome	Transport and catabolism	2	$8.93x10^{-2}$	1.00

Continued on next page...

GO-id	Term	BriteHierachy	No.Of Genes	$P_{uncorrected}$	$P_{corrected}$(FDR)
3013	RNA transport	Translation	2	$9.91x10^{-2}$	1.00
1100	Metabolic pathways	Metabolic pathways	7	$1.13x10^{-1}$	1.00
5010	Alzheimer's disease	Neurodegenerative diseases	2	$1.14x10^{-1}$	1.00
5168	Herpes simplex infection	Infectious diseases: Viral	2	$1.25x10^{-1}$	1.00
5205	Proteoglycans in cancer	Cancers: Overview	2	$1.79x10^{-1}$	1.00
4080	Neuroactive ligand-receptor interaction	Signaling molecules and interaction	2	$2.44x10^{-1}$	1.00
4620	Toll-like receptor signaling pathway	Immune system	1	1.00	1.00
4060	Cytokine-cytokine receptor interaction	Signaling molecules and interaction	1	1.00	1.00
5161	Hepatitis B	Infectious diseases: Viral	1	1.00	1.00
5160	Hepatitis C	Infectious diseases: Viral	1	1.00	1.00
5202	Transcriptional misregulation in cancer	Cancers: Overview	1	1.00	1.00
52	Galactose metabolism	Carbohydrate metabolism	1	1.00	1.00
340	Histidine metabolism	Amino acid metabolism	1	1.00	1.00
5166	HTLV-I infection	Infectious diseases: Viral	1	1.00	1.00
4972	Pancreatic secretion	Digestive system	1	1.00	1.00
4110	Cell cycle	Cell growth and death	1	1.00	1.00
5169	Epstein-Barr virus infection	Infectious diseases: Viral	1	1.00	1.00
4974	Protein digestion and absorption	Digestive system	1	1.00	1.00
4976	Bile secretion	Digestive system	1	1.00	1.00

Continued on next page...

GO-id	Term	BriteHierachy	No.Of Genes	$P_{uncorrected}$	$P_{corrected}$(FDR)
4390	Hippo signaling pathway	Signal transduction	1	1.00	1.00
4064	NF-kappa B signaling pathway	Signal transduction	1	1.00	1.00
5164	Influenza A	Infectious diseases: Viral	1	1.00	1.00
4010	MAPK signaling pathway	Signal transduction	1	1.00	1.00
4012	ErbB signaling pathway	Signal transduction	1	1.00	1.00
4014	Ras signaling pathway	Signal transduction	1	1.00	1.00
380	Tryptophan metabolism	Amino acid metabolism	1	1.00	1.00
250	Alanine, aspartate and glutamate metabolism	Amino acid metabolism	1	1.00	1.00
5206	MicroRNAs in cancer	Cancers: Overview	1	1.00	1.00
511	Other glycan degradation	Glycan biosynthesis and metabolism	1	1.00	1.00
5213	Endometrial cancer	Cancers: Specific types	1	1.00	1.00
4612	Antigen processing and presentation	Immune system	1	1.00	1.00
4261	Adrenergic signaling in cardiomyocytes	Circulatory system	1	1.00	1.00
5134	Legionellosis	Infectious diseases: Bacterial	1	1.00	1.00
4915	Estrogen signaling pathway	Endocrine system	1	1.00	1.00
4260	Cardiac muscle contraction	Circulatory system	1	1.00	1.00

Continued on next page...

GO-id	Term	BriteHierachy	No.Of Genes	$P_{uncorrected}$	$P_{corrected}$(FDR)
5414	Dilated cardiomyopathy	Cardiovascular diseases	1	1.00	1.00
5132	Salmonella infection	Infectious diseases: Bacterial	1	1.00	1.00
5412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	Cardiovascular diseases	1	1.00	1.00
4919	Thyroid hormone signaling pathway	Endocrine system	1	1.00	1.00
4510	Focal adhesion	Cellular community	1	1.00	1.00
4913	Ovarian steroidogenesis	Endocrine system	1	1.00	1.00
561	Glycerolipid metabolism	Lipid metabolism	1	1.00	1.00
564	Glycerophospholipid metabolism	Lipid metabolism	1	1.00	1.00
4151	PI3K-Akt signaling pathway	Signal transduction	1	1.00	1.00
3410	Base excision repair	Replication and repair	1	1.00	1.00
4022	cGMP-PKG signaling pathway	Signal transduction	1	1.00	1.00
4650	Natural killer cell mediated cytotoxicity	Immune system	1	1.00	1.00
5410	Hypertrophic cardiomyopathy (HCM)	Cardiovascular diseases	1	1.00	1.00
3320	PPAR signaling pathway	Endocrine system	1	1.00	1.00

Continued on next page...

GO-id	Term	BriteHierachy	No.Of Genes	$P_{uncorrected}$	$P_{corrected}$(FDR)
5140	Leishmaniasis	Infectious diseases: Parasitic	1	1.00	1.00
5142	Chagas disease (American trypanosomiasis)	Infectious diseases: Parasitic	1	1.00	1.00
3015	mRNA surveillance pathway	Translation	1	1.00	1.00
5144	Malaria	Infectious diseases: Parasitic	1	1.00	1.00
5145	Toxoplasmosis	Infectious diseases: Parasitic	1	1.00	1.00
5146	Amoebiasis	Infectious diseases: Parasitic	1	1.00	1.00
5321	Inflammatory bowel disease (IBD)	Immune diseases	1	1.00	1.00
4750	Inflammatory mediator regulation of TRP channels	Sensory system	1	1.00	1.00
5323	Rheumatoid arthritis	Immune diseases	1	1.00	1.00
4340	Hedgehog signaling pathway	Signal transduction	1	1.00	1.00
534	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	Glycan biosynthesis and metabolism	1	1.00	1.00
230	Purine metabolism	Nucleotide metabolism	1	1.00	1.00
531	Glycosaminoglycan degradation	Glycan biosynthesis and metabolism	1	1.00	1.00
3430	Mismatch repair	Replication and repair	1	1.00	1.00
5152	Tuberculosis	Infectious diseases: Bacterial	1	1.00	1.00
3008	Ribosome biogenesis in eukaryotes	Translation	1	1.00	1.00
4142	Lysosome	Transport and catabolism	1	1.00	1.00

Continued on next page...

GO-id	Term	BriteHierachy	No.Of Genes	$P_{uncorrected}$	$P_{corrected}$(FDR)
4141	Protein processing in endoplasmic reticulum	Folding	1	1.00	1.00
4932	Non-alcoholic fatty liver disease (NAFLD)	Endocrine and metabolic diseases	1	1.00	1.00
4930	Type II diabetes mellitus	Endocrine and metabolic diseases	1	1.00	1.00
830	Retinol metabolism	Metabolism of cofactors and vitamins	1	1.00	1.00
4962	Vasopressin-regulated water reabsorption	Excretory system	1	1.00	1.00
4144	Endocytosis	Transport and catabolism	1	1.00	1.00
140	Steroid hormone biosynthesis	Lipid metabolism	1	1.00	1.00
604	Glycosphingolipid biosynthesis - ganglio series	Glycan biosynthesis and metabolism	1	1.00	1.00
3040	Spliceosome	Transcription	1	1.00	1.00
4724	Glutamatergic synapse	Nervous system	1	1.00	1.00
5100	Bacterial invasion of epithelial cells	Infectious diseases: Bacterial	1	1.00	1.00
600	Sphingolipid metabolism	Lipid metabolism	1	1.00	1.00
240	Pyrimidine metabolism	Nucleotide metabolism	1	1.00	1.00

Chapter 10
Inclusion criteria for the GenOSept study

The following criteria, adapted from the American College of Chest Physicians/Society of Critical Care Medicine consensus criteria for sepsis [5], were used in the selection of patients included in the GenOSept study. Meningococcal disease cases and community acquired pneumonia cases from this study were used for this project and results presented in Chapters 3 and 4 respectively.

10.1 Inclusion criteria

1. Clinical evidence of infection within the previous 72 hours (a, b or c)
   (a) Presence of polymorphonuclear cells in a normally sterile body fluid.
   (b) Culture or Gram stain of blood, sputum, urine or normally sterile body fluid positive for a pathogenic micro-organism.
   (c) Focus of infection identified by visual inspection (e.g. ruptured bowel with the presence of free air or bowel contents in the abdomen found at the time of surgery, wound with purulent drainage)
2. Evidence of a systemic response to infection as defined by the presence of three or more of the following signs within the previous 24 hours.
   (a) Fever (rectal temperature >38.0°C [>100.4°F]) or hypothermia (rectal temperature <36.0°C [96.8°F])
   (b) Tachycardia (heart rate of >90 beat/min)

(c) Tachypnea (respiratory rate > 20 breaths/min, $PaO_(2)$<32 mmHg) or patient requires mechanical ventilation

(d) Leukocytosis (>12,000 mm^3) or leukopenia (<4,000 mm^3) or >10% immature (band) forms.

3. Evidence of severe sepsis with hypotension, hypoperfusion or organ dysfunction which are not the results of underlying diseases but are attributable to sepsis within the previous 24 hours, including one of the following:

(a) Arterial hypoxemia (PaO_2<75 mmHg or $PaO_2/FI0_2$<250 in the absence of pneumonia)

(b) Hypotension (systolic blood pressure of <90 mmHg or reduction of >40mmHg from the baseline

(c) Metabolic acidosis (pH of <7.3, or a base deficit of >5.0 mmol/L, or an increased lactic acid concentration)

(d) Sustained oliguria (urine output < 0.5 ml/kg/hr for a minimum of 1 hour in the presence of adequate fluid resuscitation)

(e) Platelet count < 100,000/mm^3

(f) Coagulation abnormality (>20% increase in prothrombin time or a >20% increase in partial thromboplastin time)

(g) Altered mental status (Glasgow Coma Scale <14) in the absence of sedation or acute change from baseline in behaviour or cognitive function.

4. Diagnosis of one of the following diseases:

(a) Community acquired Pneumonia

(b) Faecal Peritonitis

(c) Pancreatitis

(d) Meningococcal disease

10.2 Exclusion criteria

1. Pregnancy
2. Age less than 18 years.
3. Underlying disease with a prognosis for survival of less than 2 months.
4. Cardiopulmonary resuscitation within 72 hours before study.
5. Immunosuppression, chemotherapy or radiation therapy within 4 weeks before the study.
6. Presence of an advanced directive to withhold or withdraw life sustaining treatment.
7. Lack of consent for GenOSept.

Chapter 11
IPD research protocol for ethics application

Study Title: Genetic Susceptibility to Invasive Pneumococcal Disease

Internal Reference Number / Short title: IPD STUDY

Ethics Ref:

Date and Version No: Version 1.0 Last revised: 27-May-14

Chief Investigator:	Dr Stephen J. Chapman , Wellcome Trust Centre for Human Genetics, Oxford, UK schapman@well.ox.ac.uk
Investigators:	Prof. Adrian Hill, Professor of Human Genetics, Wellcome Trust Centre For Human Genetics, Oxford, UK Prof. Derrick Crook, Professor of Microbiology, Microbiology Department, John Radcliffe Hospital, Oxford, UK Dr. Anna Rautanen, Senior Postdoctoral Research Scientist, Wellcome Trust Centre For Human Genetics, Oxford, UK Dr. Kate Elliott, Postdoctoral Research Scientist , Wellcome Trust Centre For Human Genetics, Oxford, UK Anne Ndungu, DPhil Student, Wellcome Trust Centre for Human Genetics, Oxford, UK
Sponsor:	University of Oxford
Funder:	Medical Research Council, UK

Stephen Chapman

Chief Investigator Signature: ----------

Conflict of Interest

There are no conflicts of interest to declare.

Confidentiality Statement

This document contains confidential information that must not be disclosed to anyone other than the Sponsor, the Investigator Team, host organisation, and members of the Research Ethics Committee, unless authorised to do so.

TABLE OF CONTENTS

1. SYNOPSIS 3
2. BACKGROUND AND RATIONALE 3
2.1. Epidemiology of Invasive Pneumococcal Disease (IPD) 3
2.2. Pathogenesis of IPD 3
2.3. The genetic basis of susceptibility to IPD 4
3. OBJECTIVES AND OUTCOME MEASURES/ENDPOINTS 5
4. STUDY DESIGN 5
5. PARTICIPANT IDENTIFICATION 6
5.1. Study Participants 6
5.2. Inclusion Criteria 6
5.3. Exclusion Criteria 6
5.4. Recruitment 6
5.5. Informed Consent 6
6. STATISTICS AND ANALYSIS 7
7. DATA MANAGEMENT 7
7.1. Access to Data 7
8. QUALITY ASSURANCE PROCEDURES 7
9. ETHICAL AND REGULATORY CONSIDERATIONS 7
9.1. Declaration of Helsinki 7
9.2. Approvals 8
9.3. Reporting 8
9.4. Participant Confidentiality 8
9.5. Other Ethical Considerations 8
10. FINANCE AND INSURANCE 8
10.1. Funding 8
11. PUBLICATION POLICY 9
12. APPENDIX A: STUDY FLOW CHART 11

1. SYNOPSIS

Study Title	Genetic Susceptibility to Invasive Pneumococcal Disease
Internal ref. no. / short title	IPD STUDY
Study Design	Case -Control study
Study Participants	Patients with invasive pneumococcal disease
Planned Sample Size	400 (samples already available)
Planned Study Period	Samples were collected between 1995 and 1999. The planned experimental work will take place between June 2014 and June 2019.
	Objectives
Primary	To identify underlying genetic correlates that associate with susceptibility to invasive pneumococcal disease
Secondary	To identify genetic polymorphisms and rare genetic variants that may affect an individual's outcome following the development of IPD To define functional immunological consequences of protein-altering variants contributing to susceptibility to invasive pneumococcal disease.

2. BACKGROUND AND RATIONALE

2.1. Epidemiology of Invasive Pneumococcal Disease (IPD)

Pneumonia is a leading cause of morbidity and mortality worldwide, it is attributable for more deaths than AIDS, malaria and tuberculosis combined [1]. *Streptococcus pneumoniae* (the pneumococcus) is the most common cause of pneumonia in Europe and the United States[2]. In addition to pneumonia, pneumococcal infections can manifest as invasive disease, defined by isolation of the pathogen from a normally sterile site, most commonly blood (bacteraemia) or cerebrospinal fluid (meningitis). Mortality from invasive pneumococcal disease (IPD) in developed countries is at least 20%, and this rate has remained unchanged for the last twenty years despite advances in medical treatments. The mortality associated with invasive bacterial disease in developing countries is even higher: bacteraemia is associated with a mortality rate of 28% among hospitalised Kenyan children, compared with 5.5% among other admissions[3].

2.2. Pathogenesis of IPD

Asymptomatic colonisation by *Streptococcus pneumoniae* is relatively widespread: estimated to peak at 50% for children 2-3 years in developed countries and 60-90% for children in developing countries. In a minority of individuals however, the pathogen is able to cross mucosal barrier and cause disease. As with other infectious diseases, the likelihood of developing invasive disease after exposure reflects an interaction between microbial virulence factors, acquired and genetic factors. Acquired host risk factors for IPD are well described and include extremes of age, smoking, malnutrition and co-morbidity such as HIV infection.

2.3. The genetic basis of susceptibility to IPD

A host genetic component to susceptibility was implied in a landmark study on Danish adoptees: the risk of death from infection in adoptees was increased fivefold if a biological parent died prematurely from infection particularly from invasive bacterial infection [4]. Additional data from the UK showed similar heritability of the likelihood of meningococcal disease[5]. In addition, there are an increasing number of rare single gene defects that confer susceptibility to a narrow range of pathogens. Despite the evidence for a role of host genetics in the development of invasive bacterial disease, the specific genes involved remain largely unknown. Identification of these genes may now be possible with the use of 'genome-wide' approaches which study markers across all human genes – such an approach has recently proven extremely successfully in the identification of genes responsible for common conditions such as asthma and diabetes.

The proposed research aims to identify previously unknown genes and genetic variants that may determine why some individuals are particularly susceptible to invasive pneumococcal disease. A crucial question relates to the relative contribution of both common and rare host genetic variation in determining susceptibility to IPD.

Methods in genome wide association studies (GWAS) to capture common variation associating with disease are now robust and well established, allowing for up to two million genetic markers to be genotyped in a single experiment. This approach has been successful in defining the role of common variation –defined as variation with greater that 5% minor allele frequency- in complex disease with almost two thousand published studies to date.

Although common human disease was previously considered to reflect common genetic variation, there is increasing evidence to support an important role for multiple, rare variants in susceptibility to common disease. Examples include the recent identification of rare genetic mutations underlying autism, schizophrenia, type I diabetes and other autoimmune conditions. In infectious disease, rare single gene mutations have been described that confer susceptibility to a narrow range of infections in otherwise healthy individuals. Examples include rare mutations in the IL-12/Interferon-gamma (IFN-γ) pathway that predispose to infection with intracellular pathogens such as non-tuberculous mycobacteria and *Salmonella* [11-13] . These examples provide proof of principle that rare mutations in single genes can underlie single cases of severe bacterial disease in otherwise healthy individuals.

With the great improvements in technology, it is now feasible to study the role of rare variants through whole genome and whole exome sequencing (all protein coding regions only). Protein coding regions constitute only 1% of the genome but contain 85% of known large-effect disease associated mutations. A recent large scale exome sequencing project concluded that the majority of protein-coding variation is rare and enriched for deleterious alleles[15]. In addition, exome sequencing has proven a robust tool in Mendelian disorders, for example a proof-of-concept study illustrated that it was possible to use exome sequencing to find the causative mutation for Freeman-Sheldon syndrome[16]. Several studies have used this approach to establish a genetic cause for other monogenic diseases[17,18]

The proposed research aims to apply this recently available genotyping and sequencing technology to an already established sample collection. This unique collection was established as part of an enhanced surveillance of invasive pneumococcal disease in three Oxfordshire hospitals between 1995 and 1999.

Knowledge gained from this study will be invaluable in advancing understanding of the genetic basis of susceptibility to severe infectious disease.

3. OBJECTIVES AND OUTCOME MEASURES/ENDPOINTS

Objectives	**Point of Completion**
Primary Objective To use current methods in genetics to identify both common and rare genetic variants contributing to susceptibility to invasive pneumococcal disease.	Completed analysis of genetic variants that associate with disease susceptibility and outcome reported in a peer-reviewed publication.
Secondary Objectives To identify genetic polymorphisms and rare genetic variants that may affect an individual's outcome following the development of IPD To define functional immunological consequences of protein-altering variants contributing to susceptibility to invasive pneumococcal disease.	Completed analysis of genetic variants that affect outcome of disease reported in a peer-reviewed publication Completed immunological follow-up through functional assays for genetic variants identified as significantly associated with IPD susceptibility and/or outcome reported in a peer-reviewed publication.

4. STUDY DESIGN

The study follows a case-control genetic association design involving anonymised, unrelated individuals. It will examine both common and rare genetic variants on a genome-wide level using both genotyping and sequencing technology.

- Common genetic variants: the frequency of single nucleotide polymorphisms with a frequency greater than 1% in the population will be compared between patients with invasive pneumococcal disease and healthy controls.
- Complete sequencing of protein coding regions (whole exome sequencing) and non-coding regions (whole genome sequencing) will be performed in order to detect rare protein-altering variants that may impact on disease susceptibility.
- Specific candidate genes of interest will also be selected for targeted deep sequencing. Selection of these candidate genes will be based on existing knowledge of their role in immune signalling or their association with disease from previous studies.

5. PARTICIPANT IDENTIFICATION

5.1. Study Participants

DNA samples are available from 400 UK individuals of European origin with IPD.

5.2. Inclusion Criteria

- Individuals with *S. pneumoniae* isolated from a sterile site (including blood, cerebrospinal fluid, pleural fluid, joint fluid).
- Male or Female, aged 18 years or above.

5.3. Exclusion Criteria

- Ethnicity: Only individuals of European origin were included in the study.

5.4. Recruitment

Samples were collected on diagnosis from all hospitalised adult patients with microbiologically-proven IPD as part of an enhanced active surveillance programme in three hospitals in Oxfordshire, coordinated by Professor Derrick Crook. Clinical details, including age, gender, diagnosis and co-morbidities, were recorded. Frequencies of clinical presentation were as follows: pneumonia 69%, isolated bacteraemia 15%, meningitis 11%, and other presentations 5%. Pneumococcal serotypes were identified; the distribution of serotypes was very similar to that of previous UK studies, with serotype 14 being the commonest.

5.5. Informed Consent

Samples in this study are existing holdings of DNA collected between 1995 and 1999 as per the regulations prior to the Human Tissue Act (2004).All samples were obtained from excess blood taken for routine diagnostics from living patients. All data was de-linked from individuals and samples were completely anonymised, identified only by a study number assigned on collection.

We wish to utilise this sample collection for comprehensive analysis into the genetic architecture of invasive pneumococcal disease. Current technology allows us to identify both common and rare genetic variants that may associate with disease. We are concerned however that informed consent for research was not requested for these archived samples. As there is no way for us to identify participants from this collection, it is not possible to retrospectively approach individuals to request informed consent. We believe that the proposed use of archival material without consent can be justified because this represents an extremely rare if not unique collection, with the potential to lead to significant advances in understanding of disease biology and ultimately improved future patient care. The relative rarity of invasive pneumococcal disease means that it would require a very prolonged period of time (and resource) to accumulate a similar collection. Furthermore the anonymisation procedure serves to

protect individual confidentiality. As a consequence, we do not believe that a reasonable person, if asked, would object to their tissue being used for this proposed research.

6. STATISTICS AND ANALYSIS

Genome-wide data set will be analysed using standard tools designed for genome-wide association studies (such as PLINK and SNPTEST2). Control genome-wide genotype results are freely available from 2000 healthy individuals from the United Kingdom, determined by the Sanger Institute in Cambridge and freely available to investigators on-line.

Poorly performing samples and variants, as well as any related individuals or population outliers, will be excluded from the analysis after stringent quality control measures. Differences in the genotype distributions between cases and controls will be compared using logistic regression with the first two principal components of population structure as covariates. Correction will be made for the multiple independent comparisons tested as part of a genome-wide study. A *P* value of $<5x10^{-8}$ is generally recognised as an appropriate threshold for significance in the context of genome-wide analysis.

In terms of analysis of rare variants, sequences will be aligned to a reference using algorithms such as Stampy and Burrows-Wheeler Aligner (BWA) [19]. Variants will then be called using either Platypus or Genome Analysis toolkit's UnifiedGenotyper [20,21]. Whole genome sequencing data is available from over 2000 healthy UK individuals from the United Kingdom as part of the UK10K project [22]. These data are available for us to use as controls for this project. Comparison of mutational loads with variants stratified on the basis of likely functional significance and frequency will be performed using developed methods such as Sequence Kernel Association Test (SKAT) within the framework of toolkits such as PLINK/SEQ and R [23,24].

7. DATA MANAGEMENT

7.1. Access to Data

All DNA samples were de-identified upon collection. There is no data that identifies the patients.

Genetic data generated will be stored securely in password protected encrypted databases whose access will be restricted to limited research staff directly involved in the study.

8. QUALITY ASSURANCE PROCEDURES

The study will be monitored, or audited in accordance with the approved protocol, relevant regulations and standard operating procedures.

9. ETHICAL AND REGULATORY CONSIDERATIONS

9.1. Declaration of Helsinki

The Investigator will ensure that this study is conducted in accordance with the principles of the Declaration of Helsinki. ICH Guidelines for Good Clinical Practice

The Investigator will ensure that this study is conducted in full conformity with relevant regulations and with the ICH Guidelines for Good Clinical Practice (CPMP/ICH/135/95) July 1996.

9.2. Approvals

The protocol and any proposed advertising material will be submitted to an appropriate Research Ethics Committee (REC), and host institution(s) for written approval.

The Investigator will submit and, where necessary, obtain approval from the above parties for all substantial amendments to the original approved documents.

9.3. Reporting

The Principal Investigator shall submit once a year throughout the study or on request, an Annual Progress report to the REC Committee, host organisation and Sponsor. In addition, an End of Study notification and final report will be submitted to the same parties.

9.4. Participant Confidentiality

All the data was de-identified to ensure that the patients' anonymity was maintained. The samples are identified only by a study number assigned during initial collection. All documents are stored securely and only accessible by study staff and authorised personnel. The study complies with the Data Protection Act, which requires data to be anonymised as soon as it is practical to do so.

9.5. Other Ethical Considerations

The main ethical issues involve sampling of human tissue and subsequent genetic analysis. Issues surrounding informed consent are discussed in section 5.5.

For all participants, blood samples were collected as a 'surplus' sample during routine venepuncture for clinical indications - that is, patients were not subjected to additional venepuncture for research purposes only.

With regard to the subsequent genetic studies, participants were de-identified upon collection of blood; all samples are only identifiable by study number. There is no way to link genetic data generated to individuals from whom the sample was collected. For this reason it will not be possible to feedback results from the study to them.

Although the study will not benefit the participants, it is likely to be successful in furthering the understanding of the development of IPD and may lead to medical benefits for future patients.

10. FINANCE AND INSURANCE

10.1. Funding

The initial enhanced surveillance of IPD was funded by Wellcome Trust, Oxford Vaccine Group and Zeneca Pharmaceutics. Genotyping and sequencing is funded by the Wellcome Trust and Medical Research Council

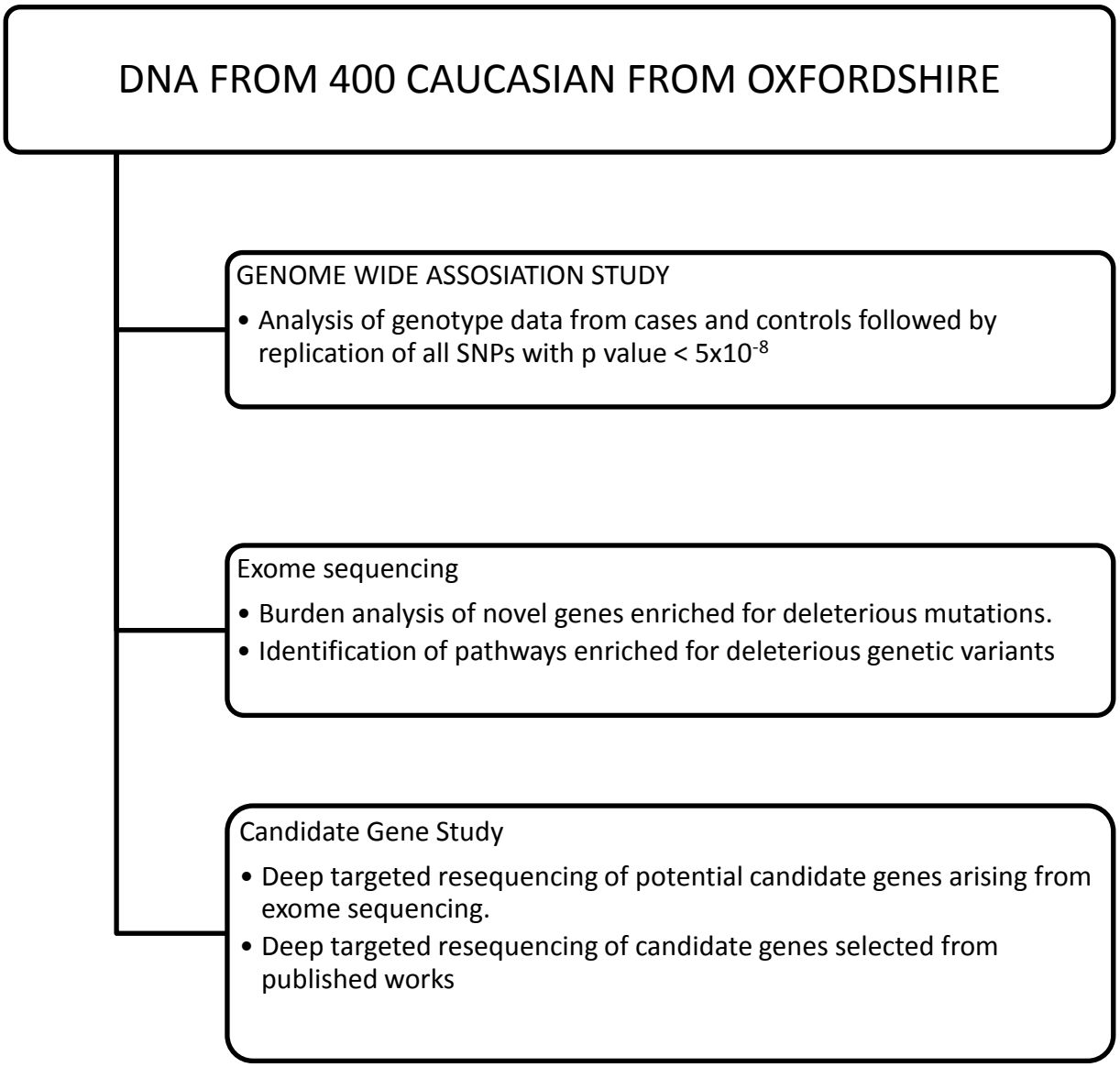
11. PUBLICATION POLICY

The Investigators will be involved in reviewing drafts of the manuscripts, abstracts, press releases and any other publications arising from the study. Authors will acknowledge that the study was funded by the Medical Research Council, Zeneca Pharmaceutics, Oxford Vaccine Group and the Wellcome Trust. Authorship will be determined in accordance with the ICMJE guidelines and other contributors will be acknowledged.

REFERENCES

1. http://www.who.int/mediacentre/factsheets/fs331/en/.
2. Lynch JP, 3rd, Zhanel GG. Streptococcus pneumoniae: epidemiology and risk factors, evolution of antimicrobial resistance, and impact of vaccines. Current opinion in pulmonary medicine 2010;16:217-25.
3. Berkley JA, Lowe BS, Mwangi I, et al. Bacteremia among children admitted to a rural hospital in Kenya. The New England journal of medicine 2005;352:39-47.
4. Sorensen TI, Nielsen GG, Andersen PK, Teasdale TW. Genetic and environmental influences on premature death in adult adoptees. The New England journal of medicine 1988;318:727-32.
5. Haralambous E, Weiss HA, Radalowicz A, Hibberd ML, Booy R, Levin M. Sibling familial risk ratio of meningococcal disease in UK Caucasians. Epidemiology and infection 2003;130:413-8.
6. Chapman SJ, Khor CC, Vannberg FO, et al. IkappaB genetic polymorphisms and invasive pneumococcal disease. American journal of respiratory and critical care medicine 2007;176:181-7.
7. Chapman SJ, Khor CC, Vannberg FO, et al. PTPN22 and invasive bacterial disease. Nature genetics 2006;38:499-500.
8. Khor CC, Chapman SJ, Vannberg FO, et al. A Mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. Nature genetics 2007;39:523-8.
9. Chapman SJ, Khor CC, Vannberg FO, et al. NFKBIZ polymorphisms and susceptibility to pneumococcal disease in European and African populations. Genes and immunity 2010;11:319-25.
10. Roy S, Knox K, Segal S, et al. MBL genotype and risk of invasive pneumococcal disease: a case-control study. Lancet 2002;359:1569-73.
11. Altare F, Durandy A, Lammas D, et al. Impairment of mycobacterial immunity in human interleukin-12 receptor deficiency. Science 1998;280:1432-5.
12. Altare F, Lammas D, Revy P, et al. Inherited interleukin 12 deficiency in a child with bacille Calmette-Guerin and Salmonella enteritidis disseminated infection. The Journal of clinical investigation 1998;102:2035-40.
13. de Jong R, Altare F, Haagen IA, et al. Severe mycobacterial and Salmonella infections in interleukin-12 receptor-deficient patients. Science 1998;280:1435-8.
14. Smirnova I, Mann N, Dols A, et al. Assay of locus-specific genetic load implicates rare Toll-like receptor 4 mutations in meningococcal susceptibility. Proceedings of the National Academy of Sciences of the United States of America 2003;100:6075-80.
15. Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 2012;337:64-9.
16. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature 2009;461:272-6.
17. Ng SB, Bigham AW, Buckingham KJ, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nature genetics 2010;42:790-3.
18. Ku CS, Naidoo N, Pawitan Y. Revisiting Mendelian disorders through exome sequencing. Human genetics 2011;129:351-70.
19. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome research 2011;21:936-9.
20. Rimmer A MI, Lunter G, McVean G. Platypus: An Integrated Variant Caller (www.well.ox.ac.uk/platypus). 2012.
21. http://www.broadinstitute.org/gatk/guide/best-practices.
22. http://www.uk10k.org/.
23. https://atgu.mgh.harvard.edu/plinkseq/.
24. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. American journal of human genetics 2011;89:82-93.

12. APPENDIX A: STUDY FLOW CHART



Chapter 12

Whole genome amplification in exome sequencing

An assessment on the suitability of whole genome-amplified DNA in exome sequencing was performed. As exome sequencing requires a high amount of high quality genomic DNA (>1µg), this method was explored for use in samples with an insufficient quantity of starting DNA. One sample was randomly selected from the meningococcal study presented in Chapter 3. This sample was whole genome amplified as described in Materials and Methods section 2.4.3. Due to the random primers used in amplification, the same sample was also whole genome amplified in duplicate and then pooled to assess for any bias of amplified regions due to random primers

12.1 Results

For the sample whereby two whole genome amplification reactions were pooled, library preparation (described in Appendix A.1) was unsuccessful. A comparison of the single whole genome amplified sample with the same sample non-amplified revealed a slight decrease in quality of sequence data in the whole genome amplified sample (appendix table E.1). The coverage for the genomic DNA was 86X while the whole genome amplified DNA had 75X coverage. The greatest discordance was observed in regions covered at greater than 30X, genomic DNA had 81% covered to a depth of at least 30X while whole genome amplified sample had 77% of regions covered at greater than 30X depth. This was still higher than the average coverage

of 73% at 30X in the 13 meningococcal sepsis exomes. The rate of duplicate reads in both samples was similar at 6%.

12.2 Discussion

While the sequence data from the whole genome amplified sample was of less coverage than the genomic DNA, the metrics observed are within the range observed from other studies as presented in Chapters 4, 5 and 6. However, the bias in genomic regions amplified by the random primers could not be determined from this experiment. Furthermore, while these results do not preclude whole genome amplified samples from exome sequencing studies, it is important to note that the starting material was high quality DNA with minimal degradation and of high concentration. Extrapolation of these results to studies with poor quality starting material would not be advised at this stage as further pilot analysis of such scenarios will be required.

Table 12.1: Exome sequencing of whole genome amplified DNA. Comparisons of exome sequencing genomic DNA and whole genome amplified DNA from the same sample was undertaken. The whole genome amplified samples had data of less quality but within the range observed from genomic DNA in other studies.

SampleID	RawReads	MappedReads	Duplicates(%)	Reads afterDedup	MeanCovofTarget	10xTarget	20xTarget	30xTarget
Genomic DNA	102.0M	101.4M	6.1	96.0M	86.2	92	87	81
Amplified DNA	89.0M	88.4M	6.5	83.3M	75.8	90	84	77

Bibliography

[1] UNICEF, "Levels and trends in child mortality report 2014: Estimates developed by the un inter-agency group for child mortality estimation," report, UNICEF, 2014.

[2] WHO, "http://www.who.int/mediacentre/factsheets/fs310/en/," 2014-05.

[3] "Classics in infectious diseases. childbed fever by Ignaz Philipp Semmelweis," *Rev Infect Dis*, vol. 3, no. 4, pp. 808–11, 1981.

[4] B. Jessney, "Joseph Lister (1827-1912): a pioneer of antiseptic surgery remembered a century after his death," *J Med Biogr*, vol. 20, no. 3, pp. 107–10, 2012.

[5] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. the ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine," *Chest*, vol. 101, no. 6, pp. 1644–55, 1992.

[6] M. M. Levy, M. P. Fink, J. C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S. M. Opal, J. L. Vincent, G. Ramsay, and Sccm/Esicm/Accp/Ats/Sis, "2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference," *Crit Care Med*, vol. 31, no. 4, pp. 1250–6, 2003.

[7] D. Rittirsch, M. A. Flierl, and P. A. Ward, "Harmful molecular mechanisms in sepsis," *Nat Rev Immunol*, vol. 8, no. 10, pp. 776–87, 2008.

[8] M. Singer, C. S. Deutschman, C. Seymour, and et al., "The third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, 2016.

[9] C. W. Seymour, V. X. Liu, T. J. Iwashyna, and et al., "Assessment of clinical criteria for sepsis: For the third international consensus definitions for sepsis and septic shock (sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 762–774, 2016.

[10] L. Thomas, "Germs," *N Engl J Med*, vol. 287, no. 11, pp. 553–5, 1972.

[11] L. Shapiro and J. A. Gelfand, "Cytokines and sepsis: pathophysiology and therapy," *New Horiz*, vol. 1, no. 1, pp. 13–22, 1993.

[12] D. F. Arnold, A. G. Roberts, A. Thomas, B. Ferry, B. P. Morgan, and H. Chapel, "A novel mutation in a patient with a deficiency of the eighth component of complement associated with recurrent meningococcal meningitis," *J Clin Immunol*, vol. 29, no. 5, pp. 691–5, 2009.

[13] J. Cohen, "Adjunctive therapy in sepsis: a critical analysis of the clinical trial programme," *Br Med Bull*, vol. 55, no. 1, pp. 212–25, 1999.

[14] R. E. Cunnion, "Clinical trials of immunotherapy for sepsis," *Crit Care Med*, vol. 20, no. 6, pp. 721–3, 1992.

[15] D. L. Dunn, "Antibody immunotherapy of gram-negative bacterial sepsis," *Pharmacotherapy*, vol. 7, no. 2, pp. S31–5, 1987.

[16] H. D. Spapen, M. Diltoer, and L. P. Huyghens, "Passive immunotherapy of gram-negative bacteremia, sepsis and septic shock," *Acta Clin Belg*, vol. 48, no. 1, pp. 20–9, 1993.

[17] J. S. Boomer, K. To, K. C. Chang, O. Takasu, D. F. Osborne, A. H. Walton, T. L. Bricker, n. Jarman, S. D., D. Kreisel, A. S. Krupnick, A. Srivastava, P. E. Swanson, J. M. Green, and R. S. Hotchkiss, "Immunosuppression in patients who die of sepsis and multiple organ failure," *JAMA*, vol. 306, no. 23, pp. 2594–605, 2011.

[18] R. S. Hotchkiss and I. E. Karl, "The pathophysiology and treatment of sepsis," *N Engl J Med*, vol. 348, no. 2, pp. 138–50, 2003.

[19] R. S. Hotchkiss, S. B. Osmon, K. C. Chang, T. H. Wagner, C. M. Coopersmith, and I. E. Karl, "Accelerated lymphocyte death in sepsis occurs by both the death receptor and mitochondrial pathways," *J Immunol*, vol. 174, no. 8, pp. 5110–8, 2005.

[20] R. S. Hotchkiss, K. W. Tinsley, P. E. Swanson, M. H. Grayson, D. F. Osborne, T. H. Wagner, J. P. Cobb, C. Coopersmith, and I. E. Karl, "Depletion of dendritic cells, but not macrophages, in patients with sepsis," *J Immunol*, vol. 168, no. 5, pp. 2493–500, 2002.

[21] D. C. Angus, W. T. Linde-Zwirble, J. Lidicker, G. Clermont, J. Carcillo, and M. R. Pinsky, "Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care," *Crit Care Med*, vol. 29, no. 7, pp. 1303–10, 2001.

[22] D. A. Harrison, C. A. Welch, and J. M. Eddleston, "The epidemiology of severe sepsis in England, Wales and Northern Ireland, 1996 to 2004: secondary analysis of a high quality clinical database, the ICnarc case mix programme database," *Crit Care*, vol. 10, no. 2, p. R42, 2006.

[23] G. S. Martin, D. M. Mannino, S. Eaton, and M. Moss, "The epidemiology of sepsis in the United States from 1979 through 2000," *N Engl J Med*, vol. 348, no. 16, pp. 1546–54, 2003.

[24] V. M. Ranieri, B. T. Thompson, P. S. Barie, J. F. Dhainaut, I. S. Douglas, S. Finfer, B. Gardlund, J. C. Marshall, A. Rhodes, A. Artigas, D. Payen, J. Tenhunen, H. R. Al-Khalidi, V. Thompson, J. Janes, W. L. Macias, B. Vangerow, M. D. Williams, and P.-S. S. Group, "Drotrecogin alfa (activated) in adults with septic shock," *N Engl J Med*, vol. 366, no. 22, pp. 2055–64, 2012.

[25] A. P. Walden, G. M. Clarke, S. McKechnie, P. Hutton, A. C. Gordon, J. Rello, J. D. Chiche, F. Stueber, C. S. Garrard, C. J. Hinds, and E. E. G. Investigators, "Patients with community acquired pneumonia admitted to European intensive care units: an epidemiological survey of the GenOSept cohort," *Crit Care*, vol. 18, no. 2, p. R58, 2014.

[26] M. S. Rangel-Frausto, D. Pittet, M. Costigan, T. Hwang, C. S. Davis, and R. P. Wenzel, "The natural history of the systemic inflammatory response syndrome (SIRS). A prospective study," *JAMA*, vol. 273, no. 2, pp. 117–23, 1995.

[27] A. A. Quartin, R. M. Schein, D. H. Kett, and P. N. Peduzzi, "Magnitude and duration of the effect of sepsis on survival. department of veterans affairs systemic sepsis cooperative studies group," *JAMA*, vol. 277, no. 13, pp. 1058–63, 1997.

[28] B. D. Winters, M. Eberlein, J. Leung, D. M. Needham, P. J. Pronovost, and J. E. Sevransky, "Long-term mortality and quality of life in sepsis: a systematic review," *Crit Care Med*, vol. 38, no. 5, pp. 1276–83, 2010.

[29] H. C. Prescott, K. M. Langa, and T. J. Iwashyna, "Readmission diagnoses after hospitalization for severe sepsis and other acute medical conditions," *JAMA*, vol. 313, no. 10, pp. 1055–7, 2015.

[30] T. Wang, A. Derhovanessian, S. De Cruz, J. A. Belperio, J. C. Deng, and G. S. Hoo, "Subsequent infections in survivors of sepsis: epidemiology and outcomes," *J Intensive Care Med*, vol. 29, no. 2, pp. 87–95, 2014.

[31] J. L. Vincent, J. Rello, J. Marshall, E. Silva, A. Anzueto, C. D. Martin, R. Moreno, J. Lipman, C. Gomersall, Y. Sakr, K. Reinhart, and E. I. G. o. Investigators, "International study of the prevalence and outcomes of infection in intensive care units," *JAMA*, vol. 302, no. 21, pp. 2323–9, 2009.

[32] J. A. Kellum, L. Kong, M. P. Fink, L. A. Weissfeld, D. M. Yealy, M. R. Pinsky, J. Fine, A. Krichevsky, R. L. Delude, D. C. Angus, and I. M. S. I. Gen, "Understanding the inflammatory cytokine response in pneumonia and sepsis: results of the genetic and inflammatory markers of sepsis (genims) study," *Arch Intern Med*, vol. 167, no. 15, pp. 1655–63, 2007.

[33] N. Johansson, M. Kalin, A. Tiveljung-Lindell, C. G. Giske, and J. Hedlund, "Etiology of community-acquired pneumonia: increased microbiological yield with new diagnostic methods," *Clin Infect Dis*, vol. 50, no. 2, pp. 202–9, 2010.

[34] F. B. Mayr, S. Yende, and D. C. Angus, "Epidemiology of severe sepsis," *Virulence*, vol. 5, no. 1, pp. 4–11, 2014.

[35] M. L. Mendu, S. Zager, F. K. Gibbons, and K. B. Christopher, "Relationship between neighborhood poverty rate and bloodstream infections in the critically ill," *Crit Care Med*, vol. 40, no. 5, pp. 1427–36, 2012.

[36] C. M. Kahler and D. S. Stephens, "Genetic basis for biosynthesis, structure, and function of meningococcal lipooligosaccharide (endotoxin)," *Critical Reviews in Microbiology*, vol. 24, no. 4, pp. 281–334, 1998.

[37] M. N. Swartz, "Bacterial meningitis–a view of the past 90 years," *N Engl J Med*, vol. 351, no. 18, pp. 1826–8, 2004.

[38] V. Manchanda, S. Gupta, and P. Bhalla, "Meningococcal disease: history, epidemiology, pathogenesis, clinical manifestations, diagnosis, antimicrobial susceptibility and prevention," *Indian J Med Microbiol*, vol. 24, no. 1, pp. 7–19, 2006.

[39] S. Flexner, "The results of the serum treatment in thirteen hundred cases of epidemic meningitis," *J Exp Med*, vol. 17, no. 5, pp. 553–76, 1913.

[40] J. Page, S. G. and M. J. Page, "Sulfadiazine in the treatment of meningococcic meningitis; a study of 60 cases-without a death," *Va Med Mon (1918)*, vol. 75, no. 8, pp. 397–402, 1948.

[41] D. M. Kuhns and H. A. Feldman, "Laboratory methods used in determining the value of sulfadiazine as a mass prophylactic against meningococcic infections," *Am J Public Health Nations Health*, vol. 33, no. 12, pp. 1461–5, 1943.

[42] J. W. Millar, E. E. Siess, H. A. Feldman, C. Silverman, and P. Frank, "In Vivo and in Vitro Resistance to Sulfadiazine in Strains of Neisseria Meningitidis," *JAMA*, vol. 186, pp. 139–41, 1963.

[43] H. Christensen, M. May, L. Bowen, M. Hickman, and C. L. Trotter, "Meningococcal carriage by age: a systematic review and meta-analysis," *Lancet Infect Dis*, vol. 10, no. 12, pp. 853–61, 2010.

[44] S. P. Yazdankhah and D. A. Caugant, "Neisseria meningitidis: an overview of the carriage state," *J Med Microbiol*, vol. 53, no. Pt 9, pp. 821–32, 2004.

[45] A. Sivonen, O. V. Renkonen, P. Weckstrom, K. Koskenvuo, V. Raunio, and P. H. Makela, "The effect of chemoprophylactic use of rifampin and minocycline on rates of carriage of neisseria meningitidis in army recruits in Finland," *J Infect Dis*, vol. 137, no. 3, pp. 238–44, 1978.

[46] J. MacLennan, G. Kafatos, K. Neal, N. Andrews, J. C. Cameron, R. Roberts, M. R. Evans, K. Cann, D. N. Baxter, M. C. Maiden, J. M. Stuart, and G. United Kingdom Meningococcal Carriage, "Social behavior and meningococcal carriage in British teenagers," *Emerg Infect Dis*, vol. 12, no. 6, pp. 950–7, 2006.

[47] R. E. Wolf and C. A. Birbara, "Meningococcal infections at an army training center," *Am J Med*, vol. 44, no. 2, pp. 243–55, 1968.

[48] R. Dwilow and S. Fanella, "Invasive Meningococcal Disease in the 21st Centuryâ€"An Update for the Clinician," *Current Neurology and Neuroscience Reports*, vol. 15, no. 3, pp. 1–9, 2015.

[49] D. Pace and A. J. Pollard, "Meningococcal disease: Clinical presentation and sequelae," *Vaccine*, vol. 30, Supplement 2, pp. B3–B9, 2012.

[50] N. E. Rosenstein, B. A. Perkins, D. S. Stephens, T. Popovic, and J. M. Hughes, "Meningococcal disease," *New England Journal of Medicine*, vol. 344, no. 18, pp. 1378–1388, 2001.

[51] L. H. Harrison, C. L. Trotter, and M. E. Ramsay, "Global epidemiology of meningococcal disease," *Vaccine*, vol. 27, Supplement 2, pp. B51–B63, 2009.

[52] K. A. Jolley, D. J. Wilson, P. Kriz, G. Mcvean, and M. C. J. Maiden, "The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in neisseria meningitidis," *Molecular Biology and Evolution*, vol. 22, no. 3, pp. 562–569, 2005.

[53] M. C. Maiden, "Multilocus Sequence Typing of Bacteria," *Annu Rev Microbiol*, vol. 60, 2006.

[54] M. C. J. Maiden, J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, and D. A. Caugant, "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms," *Proc Natl Acad Sci USA*, vol. 95, 1998.

[55] R. C. Read, "Neisseria meningitidis; clones, carriage, and disease," *Clin Microbiol Infect*, vol. 20, no. 5, pp. 391–5, 2014.

[56] K. A. Jolley and M. C. Maiden, "BIGSdb: Scalable analysis of bacterial genome variation at the population level," *BMC Bioinformatics*, vol. 11, no. 1, pp. 1–11, 2010.

[57] P. Boisier, P. Nicolas, S. Djibo, M. K. Taha, I. Jeanne, H. B. MaÃ¯nassara, B. Tenebray, K. K. Kairo, D. Giorgini, and S. Chanteau, "Meningococcal meningitis: Unprecedented incidence of serogroup X-related cases in 2006 in Niger," *Clinical Infectious Diseases*, vol. 44, no. 5, pp. 657–663, 2007.

[58] J. Chamot-Rooke, G. Mikaty, C. Malosse, M. Soyer, A. Dumont, J. Gault, A. F. Imhaus, P. Martin, M. Trellet, G. Clary, P. Chafey, L. Camoin, M. Nilges, X. Nassif, and G. Dumenil, "Posttranslational modification of pili upon cell contact triggers n. meningitidis dissemination," *Science*, vol. 331, no. 6018, pp. 778–82, 2011.

[59] F. P. De Vries, R. Cole, J. Dankert, M. Frosch, and J. P. M. Van Putten, "Neisseria meningitidis producing the opc adhesin binds epithelial cell proteoglycan receptors," *Molecular Microbiology*, vol. 27, no. 6, pp. 1203–1212, 1998.

[60] M. Virji, D. Evans, A. Hadfield, F. Grunert, A. M. Teixeira, and S. M. Watt, "Critical determinants of host receptor targeting by Neisseria meningitidis and Neisseria gonorrhoeaeâ : identification of Opa adhesiotopes on the n-domain of CD66 molecules," *Molecular Microbiology*, vol. 34, no. 3, pp. 538–551, 1999.

[61] R. C. Read, S. Zimmerli, C. Broaddus, D. A. Sanan, D. S. Stephens, and J. D. Ernst, "The (alpha2–>8)-linked polysialic acid capsule of group B Neisseria meningitidis modifies multiple steps during interaction with human macrophages," *Infection and Immunity*, vol. 64, no. 8, pp. 3210–3217, 1996.

[62] I. W. Devoe and J. E. Gilchrist, "RELEASE OF ENDOTOXIN IN THE FORM OF CELL WALL BLEBS DURING IN VITRO GROWTH OF NEISSERIA MENINGITIDIS," *The Journal of Experimental Medicine*, vol. 138, no. 5, pp. 1156–1167, 1973.

[63] M. van Deuren, J. van der Ven-Jongekrijg, A. K. M. Bartelink, R. van Dalen, R. W. Sauerwein, and J. W. M. van der Meer, "Correlation between Proinflammatory Cytokines and Antiinflammatory Mediators and the Severity of Disease in Meningococcal Infections," *Journal of Infectious Diseases*, vol. 172, no. 2, pp. 433–439, 1995.

[64] P. Brandtzaeg, P. Kierulf, P. Gaustad, A. Skulberg, J. N. Bruun, S. Halvorsen, and E. SÃ¸rensen, "Plasma Endotoxin as a Predictor of Multiple Organ Failure and Death in Systemic Meningococcal Disease," *Journal of Infectious Diseases*, vol. 159, no. 2, pp. 195–204, 1989.

[65] G. Madico, J. A. Welsch, L. A. Lewis, A. McNaughton, D. H. Perlman, C. E. Costello, J. Ngampasutadol, U. Vogel, D. M. Granoff, and S. Ram, "The meningococcal vaccine candidate GNA1870 binds the complement regulatory protein factor H and enhances serum resistance," *J Immunol*, vol. 177, no. 1, pp. 501–10, 2006.

[66] K. L. Seib, D. Serruto, F. Oriente, I. Delany, J. Adu-Bobie, D. Veggi, B. Arico, R. Rappuoli, and M. Pizza, "Factor H-binding protein is important for meningococcal survival in human whole blood and serum and in the presence of the antimicrobial peptide LL-37," *Infect Immun*, vol. 77, no. 1, pp. 292–9, 2009.

[67] K. L. Seib, M. Scarselli, M. Comanducci, D. Toneatto, and V. Masignani, "Neisseria meningitidis factor H-binding protein fHbp: a key virulence factor and vaccine antigen," *Expert Review of Vaccines*, vol. 14, no. 6, pp. 841–859, 2015.

[68] V. Masignani, M. Comanducci, M. M. Giuliani, S. Bambini, J. Adu-Bobie, B. AricÃ², B. Brunelli, A. Pieri, L. Santini, S. Savino, D. Serruto, D. Litt, S. Kroll, J. A. Welsch, D. M. Granoff, R. Rappuoli, and M. Pizza, "Vaccination against Neisseria meningitidis Using Three Variants of the Lipoprotein GNA1870," *The Journal of Experimental Medicine*, vol. 197, no. 6, pp. 789–799, 2003.

[69] L. D. Fletcher, L. Bernfield, V. Barniak, J. E. Farley, A. Howell, M. Knauf, P. Ooi, R. P. Smith, P. Weise, M. Wetherell, X. Xie, R. Zagursky, Y. Zhang, and G. W. Zlotnick, "Vaccine Potential of the Neisseria meningitidis 2086 Lipoprotein," *Infection and Immunity*, vol. 72, no. 4, pp. 2088–2100, 2004.

[70] L. K. McNeil, R. J. Zagursky, S. L. Lin, E. Murphy, G. W. Zlotnick, S. K. Hoiseth, K. U. Jansen, and A. S. Anderson, "Role of factor H binding protein in Neisseria meningitidis virulence and its potential as a vaccine candidate to broadly protect against meningococcal disease," *Microbiol Mol Biol Rev*, vol. 77, no. 2, pp. 234–52, 2013.

[71] A. R. Gorringe and R. PajÃ³n, "Bexsero," *Human Vaccines & Immunotherapeutics*, vol. 8, no. 2, pp. 174–183, 2012.

[72] J. Findlow, R. Borrow, M. D. Snape, T. Dawson, A. Holland, T. M. John, A. Evans, K. L. Telford, E. Ypma, D. Toneatto, P. Oster, E. Miller, and A. J. Pollard, "Multicenter, open-label, randomized phase II controlled trial of an investigational recombinant Meningococcal serogroup B vaccine with and without outer membrane vesicles, administered in infancy," *Clin Infect Dis*, vol. 51, no. 10, pp. 1127–37, 2010.

[73] I. Goldschneider, E. C. Gotschlich, and M. S. Artenstein, "Human immunity to the meningococcus. i. the role of humoral antibodies," *J Exp Med*, vol. 129, no. 6, pp. 1307–26, 1969.

[74] C. E. Frasch, R. Borrow, and J. Donnelly, "Bactericidal antibody is the immunologic surrogate of protection against meningococcal disease," *Vaccine*, vol. 27, Supplement 2, pp. B112–B116, 2009.

[75] C. Trotter, N. Gay, and W. Edmunds, "The natural history of meningococcal carriage and disease," *Epidemiology and Infection*, vol. 134, no. 3, pp. 556–566, 2006.

[76] P. Coen, K. Cartwright, and J. Stuart, "Mathematical modelling of infection and disease due to Neisseria meningitidis and Neisseria lactamica," *International Journal of Epidemiology*, vol. 29, no. 1, pp. 180–188, 2000.

[77] A. J. Pollard and C. Frasch, "Development of natural immunity to neisseria meningitidis," *Vaccine*, vol. 19, no. 11-12, pp. 1327–46, 2001.

[78] WHO, "Meningococcal vaccines: Who position paper, November 2011," *Wkly Epidemiol Rec*, vol. 86, no. 47, pp. 521–39, 2011.

[79] A. C. Cohn, J. R. MacNeil, T. A. Clark, I. R. Ortega-Sanchez, E. Z. Briere, H. C. Meissner, C. J. Baker, and N. E. Messonnier, "Prevention and control of meningococcal disease: recommendations of the Advisory Committee on

Immunization Practices (ACIP)," *MMWR Recomm Rep*, vol. 62, no. Rr-2, pp. 1–28, 2013.

[80] K. A. Geno, G. L. Gilbert, J. Y. Song, I. C. Skovsted, K. P. Klugman, C. Jones, H. B. Konradsen, and M. H. Nahm, "Pneumococcal capsules and their types: Past, present, and future," *Clin Microbiol Rev*, vol. 28, no. 3, pp. 871–99, 2015.

[81] M. Habib, B. D. Porter, and C. Satzke, "Capsular serotyping of Streptococcus pneumoniae using the Quellung reaction," *J Vis Exp*, no. 84, p. e51208, 2014.

[82] E. Jauneikaite, A. S. Tocheva, J. M. Jefferies, R. A. Gladstone, S. N. Faust, M. Christodoulides, M. L. Hibberd, and S. C. Clarke, "Current methods for capsular typing of Streptococcus pneumoniae," *J Microbiol Methods*, vol. 113, pp. 41–9, 2015.

[83] U. B. Sorensen, "Typing of pneumococci by using 12 pooled antisera," *J Clin Microbiol*, vol. 31, no. 8, pp. 2097–100, 1993.

[84] J. N. Weiser, R. Austrian, P. K. Sreenivasan, and H. R. Masure, "Phase variation in pneumococcal opacity: relationship between colonial morphology and nasopharyngeal colonization," *Infect Immun*, vol. 62, no. 6, pp. 2582–9, 1994.

[85] A. Kadioglu, J. N. Weiser, J. C. Paton, and P. W. Andrew, "The role of Streptococcus pneumoniae virulence factors in host respiratory colonization and disease," *Nat Rev Microbiol*, vol. 6, no. 4, pp. 288–301, 2008.

[86] I. M. Adetifa, M. Antonio, C. A. Okoromah, C. Ebruke, V. Inem, D. Nsekpong, A. Bojang, and R. A. Adegbola, "Pre-vaccination nasopharyngeal pneumococcal carriage in a Nigerian population: epidemiology and population biology," *PLoS One*, vol. 7, no. 1, p. e30548, 2012.

[87] J. N. Weiser, "The pneumococcus: why a commensal misbehaves," *J Mol Med (Berl)*, vol. 88, no. 2, pp. 97–102, 2010.

[88] S. J. Chapman, C. C. Khor, F. O. Vannberg, A. Frodsham, A. Walley, N. A. Maskell, C. W. Davies, S. Segal, C. E. Moore, S. H. Gillespie, P. Denny, N. P. Day, D. W. Crook, R. J. Davies, and A. V. Hill, "IkappaB genetic polymorphisms and invasive pneumococcal disease," *Am J Respir Crit Care Med*, vol. 176, no. 2, pp. 181–7, 2007.

[89] S. J. Chapman, C. C. Khor, F. O. Vannberg, N. A. Maskell, C. W. Davies, E. L. Hedley, S. Segal, C. E. Moore, K. Knox, N. P. Day, S. H. Gillespie, D. W. Crook, R. J. Davies, and A. V. Hill, "PTPN22 and invasive bacterial disease," *Nat Genet*, vol. 38, no. 5, pp. 499–500, 2006.

[90] S. Roy, K. Knox, S. Segal, D. Griffiths, C. E. Moore, K. I. Welsh, A. Smarason, N. P. Day, W. L. McPheat, D. W. Crook, A. V. Hill, and G. Oxford Pneumoccocal Surveillance, "MBL genotype and risk of invasive pneumococcal disease: a case-control study," *Lancet*, vol. 359, no. 9317, pp. 1569–73, 2002.

[91] W. P. Hausdorff, D. R. Feikin, and K. P. Klugman, "Epidemiological differences among pneumococcal serotypes," *Lancet Infect Dis*, vol. 5, no. 2, pp. 83–93, 2005.

[92] J. O. Kim, S. Romero-Steiner, U. B. Sorensen, J. Blom, M. Carvalho, S. Barnard, G. Carlone, and J. N. Weiser, "Relationship between cell surface carbohydrates and intrastrain variation on opsonophagocytosis of Streptococcus pneumoniae," *Infect Immun*, vol. 67, no. 5, pp. 2327–33, 1999.

[93] A. Kadioglu, W. Coward, M. J. Colston, C. R. A. Hewitt, and P. W. Andrew, "CD4-T-lymphocyte interactions with pneumolysin and pneumococci suggest a crucial protective role in the host response to pneumococcal infection," *Infection and Immunity*, vol. 72, no. 5, pp. 2689–2697, 2004.

[94] A. Wellmer, G. Zysk, J. Gerber, T. Kunst, M. von Mering, S. Bunkowski, H. Eiffert, and R. Nau, "Decreased virulence of a pneumolysin-deficient strain of streptococcus pneumoniae in murine meningitis," *Infection and Immunity*, vol. 70, no. 11, pp. 6504–6508, 2002.

[95] E. N. Janoff, J. B. Rubins, C. Fasching, D. Charboneau, J. T. Rahkola, A. G. Plaut, and J. N. Weiser, "Pneumococcal iga1 protease subverts specific protection by human iga1," *Mucosal Immunol*, vol. 7, no. 2, pp. 249–56, 2014.

[96] A. Di Sabatino, R. Carsetti, and G. R. Corazza, "Post-splenectomy and hyposplenic states," *The Lancet*, vol. 378, no. 9785, pp. 86–97, 2011.

[97] C. Theilacker, K. Ludewig, A. Serr, J. Schimpf, J. Held, M. BÃ¶gelein, V. Bahr, S. Rusch, A. Pohl, K. Kogelmann, S. Frieseke, R. Bogdanski, F. M. Brunkhorst, and W. V. Kern, "Overwhelming postsplenectomy infection: A prospective multicenter cohort study," *Clinical Infectious Diseases*, vol. 62, no. 7, pp. 871–878, 2016.

[98] CDC, "Use of 13-valent pneumococcal conjugate vaccine and 23-valent pneumococcal polysaccharide vaccine for adults with immunocompromising conditions: recommendations of the advisory committee on immunization practices (ACIP)," *MMWR Morb Mortal Wkly Rep*, vol. 61, no. 40, pp. 816–9, 2012.

[99] CDC, "Prevention of pneumococcal disease: recommendations of the Advisory Committee on Immunization Practices (ACIP)," *MMWR Recomm Rep*, vol. 46, no. Rr-8, pp. 1–24, 1997.

[100] K. B. Laupland, D. B. Gregson, D. A. Zygun, C. J. nonodoig, G. Mortis, and D. L. Church, "Severe bloodstream infections: a population-based assessment," *Crit Care Med*, vol. 32, no. 4, pp. 992–7, 2004.

[101] T. R. Talbot, T. V. Hartert, E. Mitchel, N. B. Halasa, P. G. Arbogast, K. A. Poehling, W. Schaffner, A. S. Craig, and M. R. Griffin, "Asthma as a risk factor for invasive pneumococcal disease," *N Engl J Med*, vol. 352, no. 20, pp. 2082–90, 2005.

[102] J. J. Drijkoningen and G. G. Rohde, "Pneumococcal infection in adults: burden of disease," *Clin Microbiol Infect*, vol. 20 Suppl 5, pp. 45–51, 2014.

[103] J. A. Berkley, B. S. Lowe, I. Mwangi, T. Williams, E. Bauni, S. Mwarumba, C. Ngetsa, M. P. Slack, S. Njenga, C. A. Hart, K. Maitland, M. English, K. Marsh, and J. A. Scott, "Bacteremia among children admitted to a rural hospital in Kenya," *N Engl J Med*, vol. 352, no. 1, pp. 39–47, 2005.

[104] M. Weisfelt, D. van de Beek, L. Spanjaard, J. B. Reitsma, and J. de Gans, "Clinical features, complications, and outcome in adults with pneumococcal meningitis: a prospective case series," *Lancet Neurol*, vol. 5, no. 2, pp. 123–9, 2006.

[105] P. B. McIntyre, K. L. O'Brien, B. Greenwood, and D. van de Beek, "Effect of vaccines on bacterial meningitis worldwide," *Lancet*, vol. 380, no. 9854, pp. 1703–11, 2012.

[106] M. Ledwith, "Pneumococcal conjugate vaccine," *Curr Opin Pediatr*, vol. 13, no. 1, pp. 70–4, 2001.

[107] T. Pilishvili, C. Lexau, M. M. Farley, J. Hadler, L. H. Harrison, N. M. Bennett, A. Reingold, A. Thomas, W. Schaffner, A. S. Craig, P. J. Smith, B. W. Beall, C. G. Whitney, M. R. Moore, and N. Active Bacterial Core Surveillance/Emerging Infections Program, "Sustained reductions in invasive pneumococcal disease in the era of conjugate vaccine," *J Infect Dis*, vol. 201, no. 1, pp. 32–41, 2010.

[108] M. R. Moore, R. Link-Gelles, W. Schaffner, R. Lynfield, C. Lexau, N. M. Bennett, S. Petit, S. M. Zansky, L. H. Harrison, A. Reingold, L. Miller, K. Scherzinger, A. Thomas, M. M. Farley, E. R. Zell, J. Taylor, T. H., T. Pondo, L. Rodgers, L. McGee, B. Beall, J. H. Jorgensen, and C. G. Whitney, "Effect of use of 13-valent pneumococcal conjugate vaccine in children on invasive pneumococcal disease in children and adults in the USA: analysis of multisite, population-based surveillance," *Lancet Infect Dis*, vol. 15, no. 3, pp. 301–9, 2015.

[109] R. E. Strachan, T. L. Snelling, and A. Jaffe, "Increased paediatric hospitalizations for empyema in Australia after introduction of the 7-valent pneumococcal conjugate vaccine," *Bull World Health Organ*, vol. 91, no. 3, pp. 167–73, 2013.

[110] R. W. Light, W. M. Girard, S. G. Jenkinson, and R. B. George, "Parapneumonic effusions," *The American Journal of Medicine*, vol. 69, no. 4, pp. 507–512, 1980.

[111] M. A. Elemraid, M. F. Thomas, A. P. Blain, S. P. Rushton, D. A. Spencer, A. R. Gennery, J. E. Clark, and U. K. North East of England Pediatric Respiratory Infection Study Group Newcastle upon Tyne, "Risk factors for the development of pleural empyema in children," *Pediatr Pulmonol*, vol. 50, no. 7, pp. 721–6, 2015.

[112] M. A. Elemraid, A. D. Sails, G. J. Eltringham, J. D. Perry, S. P. Rushton, D. A. Spencer, M. F. Thomas, K. M. Eastham, F. Hampton, A. R. Gennery, and J. E. Clark, "Aetiology of paediatric pneumonia after the introduction of pneumococcal conjugate vaccine," *European Respiratory Journal*, vol. 42, no. 6, pp. 1595–1603, 2013.

[113] C. G. Grijalva, J. P. Nuorti, Y. Zhu, and M. R. Griffin, "Increasing incidence of empyema complicating childhood community-acquired pneumonia in the United States," *Clin Infect Dis*, vol. 50, no. 6, pp. 805–13, 2010.

[114] D. A. Spencer, S. M. Iqbal, A. Hasan, and L. Hamilton, "Empyema thoracis is still increasing in uk children," *BMJ*, vol. 332, no. 7553, p. 1333, 2006.

[115] C. L. Byington, L. Y. Spencer, T. A. Johnson, A. T. Pavia, D. Allen, E. O. Mason, S. Kaplan, K. C. Carroll, J. A. Daly, J. C. Christenson, and M. H. Samore, "An epidemiological investigation of a sustained high rate of pediatric parapneumonic empyema: risk factors and microbiological associations," *Clin Infect Dis*, vol. 34, no. 4, pp. 434–40, 2002.

[116] N. Al-Kaabi, Z. Solh, S. Pacheco, L. Murray, I. Gaboury, and N. Le Saux, "A comparison of group a streptococcus versus Streptococcus pneumoniae pneumonia," *Pediatr Infect Dis J*, vol. 25, no. 11, pp. 1008–12, 2006.

[117] S. Saxena, C. Atchison, E. Cecil, M. Sharland, E. Koshy, and A. Bottle, "Additive impact of pneumococcal conjugate vaccines on pneumonia and empyema hospital admissions in England," *Journal of Infection*, vol. 71, no. 4, pp. 428–436, 2015.

[118] L. Simonsen, R. J. Taylor, C. Schuck-Paim, R. Lustig, M. Haber, and K. P. Klugman, "Effect of 13-valent pneumococcal conjugate vaccine on admissions to hospital 2 years after its introduction in the USA: a time series analysis," *The Lancet Respiratory Medicine*, vol. 2, no. 5, pp. 387–394, 2014.

[119] H. E. Davies, R. J. O. Davies, and C. W. H. Davies, "Management of pleural infection in adults: British Thoracic Society pleural disease guideline 2010," *Thorax*, vol. 65, no. Suppl 2, pp. ii41–ii53, 2010.

[120] L. McCauley and N. Dean, "Pneumonia and empyema: causal, casual or unknown," *Journal of Thoracic Disease*, vol. 7, no. 6, pp. 992–998, 2015.

[121] C. Kroegel and V. Antony, "Immunobiology of pleural inflammation: potential implications for pathogenesis, diagnosis and therapy," *European Respiratory Journal*, vol. 10, no. 10, pp. 2411–2418, 1997.

[122] D. E. Potts, D. A. Taryle, and S. A. Sahn, "The glucose-ph relationship in parapneumonic effusions," *Arch Intern Med*, vol. 138, no. 9, pp. 1378–80, 1978.

[123] S. Wilkosz, L. A. Edwards, S. Bielsa, C. Hyams, A. Taylor, R. J. Davies, G. J. Laurent, R. C. Chambers, J. S. Brown, and Y. C. Lee, "Characterization of a new mouse model of empyema and the mechanisms of pleural invasion by streptococcus pneumoniae," *Am J Respir Cell Mol Biol*, vol. 46, no. 2, pp. 180–7, 2012.

[124] F. Rosner, "Hemophilia in the talmud and rabbinic writings," *Ann Intern Med*, vol. 70, no. 4, pp. 833–7, 1969.

[125] F. Weiling, "Mendel, g.: Experiments in plant hybridisation. mendel's original paper in english translation with commentary and asessment by the late sir ronald a. fisher, edited by j. h. bennett. oliver & boyd, edinburgh and london 1965; ix, 95 S., preis 21 s," *Biometrische Zeitschrift*, vol. 9, no. 3, pp. 208–208, 1967.

[126] M. W. BATESON, "PROBLEMS OF HEREDITY AS A SUBJECT FOR HORTICULTURAL INVESTIGATION," *A Century of Mendelism in Human Genetics*, p. 153, 2004.

[127] A. E. Garrod, "The incidence of alkaptonuria: a study in chemical individuality. 1902 [classical article]," *The Yale Journal of Biology and Medicine*, vol. 75, no. 4, pp. 221–231, 2002.

[128] J. B. Herrick, "Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia. 1910," *Yale J Biol Med*, vol. 74, no. 3, pp. 179–84, 2001.

[129] T. L. Savitt and M. F. Goldberg, "Herrick's 1910 case report of sickle cell anemia. the rest of the story," *JAMA*, vol. 261, no. 2, pp. 266–71, 1989.

[130] L. Pauling, H. A. Itano, and et al., "Sickle cell anemia a molecular disease," *Science*, vol. 110, no. 2865, pp. 543–8, 1949.

[131] V. M. Ingram, "Gene mutations in human haemoglobin: the chemical difference between normal and sickle cell haemoglobin," *Nature*, vol. 180, no. 4581, pp. 326–8, 1957.

[132] D. Altshuler, M. J. Daly, and E. S. Lander, "Genetic Mapping in Human Disease," *Science*, vol. 322, no. 5903, pp. 881–888, 2008.

[133] A. V. Hill, "The genomics AND genetics of human infectious disease susceptibility," *Annual Review of Genomics and Human Genetics*, vol. 2, no. 1, pp. 373–400, 2001.

[134] N. Risch and K. Merikangas, "The future of genetic studies of complex human diseases," *Science*, vol. 273, no. 5281, pp. 1516–7, 1996.

[135] N. M. Laird and C. Lange, "Family-based designs in the age of large-scale gene-association studies," *Nat Rev Genet*, vol. 7, no. 5, pp. 385–394, 2006.

[136] J. N. Hirschhorn, K. Lohmueller, E. Byrne, and K. Hirschhorn, "A comprehensive review of genetic association studies," *Genet Med*, vol. 4, no. 2, pp. 45–61, 2002.

[137] A. V. Hill, "Genetics and genomics of infectious disease susceptibility," *British Medical Bulletin*, vol. 55, no. 2, pp. 401–413, 1999.

[138] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, Y. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, *et al.*, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[139] C. International HapMap, "The international hapmap project," *Nature*, vol. 426, no. 6968, pp. 789–96, 2003.

[140] C. International HapMap, "A haplotype map of the human genome," *Nature*, vol. 437, no. 7063, pp. 1299–320, 2005.

[141] C. Wellcome Trust Case Control, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, no. 7145, pp. 661–78, 2007.

[142] B. Howie, J. Marchini, and M. Stephens, "Genotype imputation with thousands of genomes," *G3 (Bethesda)*, vol. 1, no. 6, pp. 457–70, 2011.

[143] J. K. Pritchard and N. J. Cox, "The allelic architecture of human disease genes: common disease-common variant...or not?," *Hum Mol Genet*, vol. 11, no. 20, pp. 2417–23, 2002.

[144] D. E. Reich and E. S. Lander, "On the allelic spectrum of human disease," *Trends Genet*, vol. 17, no. 9, pp. 502–10, 2001.

[145] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. Ioannidis, and J. N. Hirschhorn, "Genome-wide association studies for complex traits: consensus, uncertainty and challenges," *Nat Rev Genet*, vol. 9, no. 5, pp. 356–69, 2008.

[146] D. Altshuler and M. Daly, "Guilt beyond a reasonable doubt," *Nat Genet*, vol. 39, no. 7, pp. 813–5, 2007.

[147] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. Mackay, S. A. McCarroll, and P. M. Visscher, "Finding the missing heritability of complex diseases," *Nature*, vol. 461, no. 7265, pp. 747–53, 2009.

[148] A. Franke, D. P. McGovern, J. C. Barrett, K. Wang, G. L. Radford-Smith, T. Ahmad, C. W. Lees, T. Balschun, J. Lee, R. Roberts, C. A. Anderson, J. C. Bis, S. Bumpstead, D. Ellinghaus, E. M. Festen, M. Georges, T. Green, T. Haritunians, L. Jostins, A. Latiano, C. G. Mathew, G. W. Montgomery, N. J. Prescott, S. Raychaudhuri, J. I. Rotter, P. Schumm, Y. Sharma, L. A. Simms, K. D. Taylor, D. Whiteman, C. Wijmenga, R. N. Baldassano, M. Barclay, T. M. Bayless, S. Brand, C. Buning, A. Cohen, J. F. Colombel, M. Cottone, L. Stronati, T. Denson, M. De Vos, R. D'Inca, M. Dubinsky, C. Edwards, T. Florin, D. Franchimont, R. Gearry, J. Glas, A. Van Gossum, S. L. Guthery, J. Halfvarson, H. W. Verspaget, J. P. Hugot, A. Karban, D. Laukens, I. Lawrance, M. Lemann, A. Levine, C. Libioulle, E. Louis, C. Mowat, W. Newman, J. Panes, A. Phillips, D. D. Proctor, M. Regueiro, R. Russell, P. Rutgeerts, J. Sanderson, M. Sans, F. Seibold, A. H. Steinhart, P. C. Stokkers, L. Torkvist, G. Kullak-Ublick, D. Wilson, T. Walters, S. R. Targan, S. R. Brant, J. D. Rioux, M. D'Amato, R. K. Weersma, S. Kugathasan, A. M. Griffiths, J. C. Mansfield, S. Vermeire, R. H. Duerr, M. S. Silverberg, J. Satsangi, S. Schreiber, J. H. Cho, V. Annese, H. Hakonarson, M. J. Daly, and M. Parkes, "Genome-wide meta-analysis increases to 71 the number of confirmed crohn's disease susceptibility loci," *Nat Genet*, vol. 42, no. 12, pp. 1118–25, 2010.

[149] P. M. Visscher, "Sizing up human height variation," *Nat Genet*, vol. 40, no. 5, pp. 489–90, 2008.

[150] E. E. Eichler, J. Flint, G. Gibson, A. Kong, S. M. Leal, J. H. Moore, and J. H. Nadeau, "Missing heritability and strategies for finding the underlying causes of complex disease," *Nat Rev Genet*, vol. 11, no. 6, pp. 446–50, 2010.

[151] O. Zuk, S. F. Schaffner, K. Samocha, R. Do, E. Hechter, S. Kathiresan, M. J. Daly, B. M. Neale, S. R. Sunyaev, and E. S. Lander, "Searching for missing heritability: designing rare variant association studies," *Proc Natl Acad Sci U S A*, vol. 111, no. 4, pp. E455–64, 2014.

[152] I. P. Gorlov, O. Y. Gorlova, S. R. Sunyaev, M. R. Spitz, and C. I. Amos, "Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms," *Am J Hum Genet*, vol. 82, no. 1, pp. 100–12, 2008.

[153] D. G. MacArthur, T. A. Manolio, D. P. Dimmock, H. L. Rehm, J. Shendure, G. R. Abecasis, D. R. Adams, R. B. Altman, S. E. Antonarakis, E. A. Ashley, J. C. Barrett, L. G. Biesecker, D. F. Conrad, G. M. Cooper, N. J. Cox, M. J. Daly, M. B. Gerstein, D. B. Goldstein, J. N. Hirschhorn, S. M. Leal, L. A. Pennacchio, J. A. Stamatoyannopoulos, S. R. Sunyaev, D. Valle, B. F. Voight, W. Winckler, and C. Gunter, "Guidelines for investigating causality of sequence variants in human disease," *Nature*, vol. 508, no. 7497, pp. 469–76, 2014.

[154] J. E. Huffman, P. S. de Vries, A. C. Morrison, M. Sabater-Lleal, T. Kacprowski, P. L. Auer, J. A. Brody, D. I. Chasman, M. H. Chen, X. Guo, L. A. Lin, R. E. Marioni, M. Muller-Nurasyid, L. R. Yanek, N. Pankratz, M. L. Grove, M. P. de Maat, M. Cushman, K. L. Wiggins, L. Qi, B. Sennblad, S. E. Harris, O. Polasek, H. Riess, F. Rivadeneira, L. M. Rose, A. Goel, K. D. Taylor, A. Teumer, A. G. Uitterlinden, D. Vaidya, J. Yao, W. Tang, D. Levy, M. Waldenberger, D. M. Becker, A. R. Folsom, F. Giulianini, A. Greinacher, A. Hofman, C. C. Huang, C. Kooperberg, A. Silveira, J. M. Starr, K. Strauch, R. J. Strawbridge, A. F. Wright, B. McKnight, O. H. Franco, N. Zakai, R. A. Mathias, B. M. Psaty, P. M. Ridker, G. H. Tofler, U. Volker, H. Watkins, M. Fornage, A. Hamsten, I. J. Deary, E. Boerwinkle, W. Koenig, J. I. Rotter, C. Hayward, A. Dehghan, A. P. Reiner, C. J. O'Donnell, and N. L. Smith, "Rare and low-frequency variants and their association with plasma levels of fibrinogen, fvii, fviii, and vwf," *Blood*, vol. 126, no. 11, pp. e19–29, 2015.

[155] A. C. Allison, "Protection afforded by sickle-cell trait against subtertian malareal infection," *Br Med J*, vol. 1, no. 4857, pp. 290–4, 1954.

[156] A. C. Allison, "The distribution of the sickle-cell trait in East Africa and elsewhere, and its apparent relationship to the incidence of subtertian malaria," *Trans R Soc Trop Med Hyg*, vol. 48, no. 4, pp. 312–8, 1954.

[157] G. V. Kryukov, L. A. Pennacchio, and S. R. Sunyaev, "Most rare missense alleles are deleterious in humans: implications for complex disease and association studies," *Am J Hum Genet*, vol. 80, no. 4, pp. 727–39, 2007.

[158] M. R. Nelson, D. Wegmann, M. G. Ehm, D. Kessner, P. St Jean, C. Verzilli, J. Shen, Z. Tang, S. A. Bacanu, D. Fraser, L. Warren, J. Aponte, M. Zawistowski, X. Liu, H. Zhang, Y. Zhang, J. Li, Y. Li, L. Li, P. Woollard, S. Topp, M. D. Hall, K. Nangle, J. Wang, G. Abecasis, L. R. Cardon, S. Zollner, J. C. Whittaker, S. L. Chissoe, J. Novembre, and V. Mooser, "An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people," *Science*, vol. 337, no. 6090, pp. 100–4, 2012.

[159] S. Nejentsev, N. Walker, D. Riches, M. Egholm, and J. A. Todd, "Rare variants of IFih1, a gene implicated in antiviral responses, protect against type 1 diabetes," *Science*, vol. 324, no. 5925, pp. 387–9, 2009.

[160] J. C. Cohen, A. Pertsemlidis, S. Fahmi, S. Esmail, G. L. Vega, S. M. Grundy, and H. H. Hobbs, "Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels," *Proc Natl Acad Sci U S A*, vol. 103, no. 6, pp. 1810–5, 2006.

[161] J. C. Cohen, E. Boerwinkle, J. Mosley, T. H., and H. H. Hobbs, "Sequence variations in PCSK9, low LDL, and protection against coronary heart disease," *N Engl J Med*, vol. 354, no. 12, pp. 1264–72, 2006.

[162] J. Cohen, A. Pertsemlidis, I. K. Kotowski, R. Graham, C. K. Garcia, and H. H. Hobbs, "Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9," *Nat Genet*, vol. 37, no. 2, pp. 161–5, 2005.

[163] N. Ahituv, N. Kavaslar, W. Schackwitz, A. Ustaszewska, J. Martin, S. Hebert, H. Doelle, B. Ersoy, G. Kryukov, S. Schmidt, N. Yosef, E. Ruppin, R. Sharan, C. Vaisse, S. Sunyaev, R. Dent, J. Cohen, R. McPherson, and L. A. Pennacchio, "Medical sequencing at the extremes of human body mass," *Am J Hum Genet*, vol. 80, no. 4, pp. 779–91, 2007.

[164] M. A. Rivas, M. Beaunonodoin, A. Gardet, C. Stevens, Y. Sharma, C. K. Zhang, G. Boucher, S. Ripke, D. Ellinghaus, N. Burtt, T. Fennell, A. Kirby, A. Latiano, P. Goyette, T. Green, J. Halfvarson, T. Haritunians, J. M. Korn, F. Kuruvilla, C. Lagace, B. Neale, K. S. Lo, P. Schumm, L. Torkvist, D. National Institute of, C. Digestive Kidney Diseases Inflammatory Bowel Disease Genetics, C. United Kingdom Inflammatory Bowel Disease Genetics, C. International Inflammatory Bowel Disease Genetics, M. C. Dubinsky, S. R. Brant, M. S. Silverberg, R. H. Duerr, D. Altshuler, S. Gabriel, G. Lettre, A. Franke, M. D'Amato, D. P. McGovern, J. H. Cho, J. D. Rioux, R. J. Xavier, and M. J. Daly, "Deep resequencing of gwas loci identifies independent rare variants associated with inflammatory bowel disease," *Nat Genet*, vol. 43, no. 11, pp. 1066–73, 2011.

[165] B. Li and S. M. Leal, "Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data," *Am J Hum Genet*, vol. 83, no. 3, pp. 311–21, 2008.

[166] J. K. Teer and J. C. Mullikin, "Exome sequencing: the sweet spot before whole genomes," *Hum Mol Genet*, vol. 19, no. R2, pp. R145–51, 2010.

[167] M. Pertea and S. L. Salzberg, "Between a chicken and a grape: estimating the number of human genes," *Genome Biol*, vol. 11, no. 5, p. 206, 2010.

[168] K. D. Pruitt, J. Harrow, R. A. Harte, C. Wallin, M. Diekhans, D. R. Maglott, S. Searle, C. M. Farrell, J. E. Loveland, B. J. Ruef, E. Hart, M. M. Suner, M. J. Landrum, B. Aken, S. Ayling, R. Baertsch, J. Fernandez-Banet, J. L. Cherry, V. Curwen, M. Dicuccio, M. Kellis, J. Lee, M. F. Lin, M. Schuster, A. Shkeda, C. Amid, G. Brown, O. Dukhanina, A. Frankish, J. Hart, B. L. Maidak, J. Mudge, M. R. Murphy, T. Murphy, J. Rajan, B. Rajput, L. D. Riddick, C. Snow, C. Steward, D. Webb, J. A. Weber, L. Wilming, W. Wu, E. Birney, D. Haussler, T. Hubbard, J. Ostell, R. Durbin, and D. Lipman, "The consensus coding sequence (CCDs) project: Identifying a common protein-coding gene set for the human and mouse genomes," *Genome Res*, vol. 19, no. 7, pp. 1316–23, 2009.

[169] D. Karolchik, A. S. Hinrichs, and W. J. Kent, "The ucsc genome browser," *Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]*, vol. CHAPTER, pp. Unit1.4–Unit1.4, 2009.

[170] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei, A. Astashyn, A. Badretdin, Y. Bao, O. Blinkova, V. Brover, V. Chetvernin, J. Choi, E. Cox,

O. Ermolaeva, C. M. Farrell, T. Goldfarb, T. Gupta, D. Haft, E. Hatcher, W. Hlavina, V. S. Joardar, V. K. Kodali, W. Li, D. Maglott, P. Masterson, K. M. McGarvey, M. R. Murphy, K. O'Neill, S. Pujar, S. H. Rangwala, D. Rausch, L. D. Riddick, C. Schoch, A. Shkeda, S. S. Storz, H. Sun, F. Thibaud-Nissen, I. Tolstoy, R. E. Tully, A. R. Vatsan, C. Wallin, D. Webb, W. Wu, M. J. Landrum, A. Kimchi, T. Tatusova, M. DiCuccio, P. Kitts, T. D. Murphy, and K. D. Pruitt, "Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation," *Nucleic Acids Res*, 2015.

[171] P. Flicek, B. L. Aken, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, G. Koscielny, E. Kulesha, D. Lawson, I. Longden, T. Massingham, W. McLaren, K. Megy, B. Overduin, B. Pritchard, D. Rios, M. Ruffier, M. Schuster, G. Slater, D. Smedley, G. Spudich, Y. A. Tang, S. Trevanion, A. Vilella, J. Vogel, S. White, S. P. Wilder, A. Zadissa, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. Hubbard, A. Parker, G. Proctor, J. Smith, and S. M. Searle, "Ensembl's 10th year," *Nucleic Acids Res*, vol. 38, no. Database issue, pp. D557–62, 2010.

[172] M. Choi, U. I. Scholl, W. Ji, T. Liu, I. R. Tikhonova, P. Zumbo, A. Nayir, A. Bakkaloglu, S. Ozen, S. Sanjad, C. Nelson-Williams, A. Farhi, S. Mane, and R. P. Lifton, "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing," *Proc Natl Acad Sci U S A*, vol. 106, no. 45, pp. 19096–101, 2009.

[173] S. B. Ng, E. H. Turner, P. D. Robertson, S. D. Flygare, A. W. Bigham, C. Lee, T. Shaffer, M. Wong, A. Bhattacharjee, E. E. Eichler, M. Bamshad, D. A. Nickerson, and J. Shendure, "Targeted capture and massively parallel sequencing of 12 human exomes," *Nature*, vol. 461, no. 7261, pp. 272–6, 2009.

[174] J. Becker, O. Semler, C. Gilissen, Y. Li, H. J. Bolz, C. Giunta, C. Bergmann, M. Rohrbach, F. Koerber, K. Zimmermann, P. de Vries, B. Wirth, E. Schoenau, B. Wollnik, J. A. Veltman, A. Hoischen, and C. Netzer, "Exome sequencing identifies truncating mutations in human SERPINF1 in autosomal-recessive osteogenesis imperfecta," *Am J Hum Genet*, vol. 88, no. 3, pp. 362–71, 2011.

[175] A. C. Need, J. P. McEvoy, M. Gennarelli, E. L. Heinzen, D. Ge, J. M. Maia, K. V. Shianna, M. He, E. T. Cirulli, C. E. Gumbs, Q. Zhao, C. R. Campbell, L. Hong, P. Rosenquist, A. Putkonen, T. Hallikainen, E. Repo-Tiihonen, J. Tiihonen, D. L. Levy, H. Y. Meltzer, and D. B. Goldstein, "Exome sequencing followed by large-scale genotyping suggests a limited role for moderately rare risk factors of strong effect in schizophrenia," *Am J Hum Genet*, vol. 91, no. 2, pp. 303–12, 2012.

[176] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, and M. J. Bamshad, "Exome sequencing identifies the cause of a mendelian disorder," *Nat Genet*, vol. 42, no. 1, pp. 30–5, 2010.

[177] L. C. D'Alessandro, S. Al Turki, A. K. Manickaraj, D. Manase, B. J. Mulder, L. Bergin, H. C. Rosenberg, T. Mondal, E. Gordon, J. Lougheed, J. Smythe, K. Devriendt, S. Bhattacharya, H. Watkins, J. Bentham, S. Bowdin, M. E. Hurles, and S. Mital, "Exome sequencing identifies rare variants in multiple genes in atrioventricular septal defect," *Genet Med*, 2015.

[178] D. A. Dyment, M. Tetreault, C. L. Beaulieu, T. Hartley, P. Ferreira, J. W. Chardon, J. Marcadier, S. L. Sawyer, S. J. Mosca, A. M. Innes, J. S. Parboosingh, D. E. Bulman, J. Schwartzentruber, J. Majewski, M. Tarnopolsky, K. M. Boycott, F. C. Consortium, and C. Care4Rare, "Whole-exome sequencing broadens the phenotypic spectrum of rare pediatric epilepsy: a retrospective study," *Clin Genet*, vol. 88, no. 1, pp. 34–40, 2015.

[179] J. Egawa, Y. Watanabe, C. Wang, E. Inoue, A. Sugimoto, T. Sugiyama, H. Igeta, A. Nunokawa, M. Shibuya, I. Kushima, N. Orime, T. Hayashi, T. Okada, Y. Uno, N. Ozaki, and T. Someya, "Novel rare missense variations and risk of autism spectrum disorder: whole-exome sequencing in two families with affected siblings and a two-stage follow-up study in a Japanese population," *PLoS One*, vol. 10, no. 3, p. e0119413, 2015.

[180] M. J. Emond, T. Louie, J. Emerson, W. Zhao, R. A. Mathias, M. R. Knowles, F. A. Wright, M. J. Rieder, H. K. Tabor, D. A. Nickerson, K. C. Barnes, L. National Heart, G. O. E. S. P. Blood Institute, G. O. Lung, R. L. Gibson, and M. J. Bamshad, "Exome sequencing of extreme phenotypes identifies DCTN4 as a modifier of chronic pseudomonas aeruginosa infection in cystic fibrosis," *Nat Genet*, vol. 44, no. 8, pp. 886–9, 2012.

[181] L. Gao, M. J. Emond, T. Louie, C. Cheadle, A. E. Berger, N. Rafaels, C. Vergara, Y. Kim, M. A. Taub, I. Ruczinski, S. C. Mathai, S. S. Rich, D. A. Nickerson, L. K. Hummers, M. J. Bamshad, P. M. Hassoun, R. A. Mathias, L. National Heart, G. O. E. S. P. Blood Institute, and K. C. Barnes, "Whole-exome sequencing identifies rare variants in ATP8B4 as a risk factor for systemic sclerosis," *Arthritis Rheumatol*, 2015.

[182] E. L. Heinzen, C. Depondt, G. L. Cavalleri, E. K. Ruzzo, N. M. Walley, A. C. Need, D. Ge, M. He, E. T. Cirulli, Q. Zhao, K. D. Cronin, C. E. Gumbs, C. R. Campbell, L. K. Hong, J. M. Maia, K. V. Shianna, M. McCormack, R. A. Radtke, G. D. O'Conner, M. A. Mikati, W. B. Gallentine, A. M. Husain, S. R. Sinha, K. Chinthapalli, R. S. Puranam, J. O. McNamara, R. Ottman, S. M. Sisodiya, N. Delanty, and D. B. Goldstein, "Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy," *Am J Hum Genet*, vol. 91, no. 2, pp. 293–302, 2012.

[183] O. L. Holmen, H. Zhang, W. Zhou, E. Schmidt, D. H. Hovelson, A. Langhammer, M. L. Lochen, S. K. Ganesh, E. B. Mathiesen, L. Vatten, C. Platou, T. Wilsgaard, J. Chen, F. Skorpen, H. Dalen, M. Boehnke, G. R. Abecasis, I. Njolstad, K. Hveem, and C. J. Willer, "No large-effect low-frequency coding variation found for myocardial infarction," *Hum Mol Genet*, vol. 23, no. 17, pp. 4721–8, 2014.

[184] E. Inoue, Y. Watanabe, J. Egawa, A. Sugimoto, A. Nunokawa, M. Shibuya, H. Igeta, and T. Someya, "Rare heterozygous truncating variations and risk of autism spectrum disorder: Whole-exome sequencing of a multiplex family and follow-up study in a Japanese population," *Psychiatry Clin Neurosci*, vol. 69, no. 8, pp. 472–6, 2015.

[185] A. Kiezun, K. Garimella, R. Do, N. O. Stitziel, B. M. Neale, P. J. McLaren, N. Gupta, P. Sklar, P. F. Sullivan, J. L. Moran, C. M. Hultman, P. Lichtenstein, P. Magnusson, T. Lehner, Y. Y. Shugart, A. L. Price, P. I. de Bakker, S. M. Purcell, and S. R. Sunyaev, "Exome sequencing and the genetic basis of complex traits," *Nat Genet*, vol. 44, no. 6, pp. 623–30, 2012.

[186] L. A. Lange, Y. Hu, H. Zhang, C. Xue, E. M. Schmidt, Z. Z. Tang, C. Bizon, E. M. Lange, J. D. Smith, E. H. Turner, G. Jun, H. M. Kang, G. Peloso, P. Auer, K. P. Li, J. Flannick, J. Zhang, C. Fuchsberger, K. Gaulton, C. Lindgren, A. Locke, A. Manning, X. Sim, M. A. Rivas, O. L. Holmen, O. Gottesman, Y. Lu, D. Ruderfer, E. A. Stahl, Q. Duan, Y. Li, P. Durda, S. Jiao, A. Isaacs, A. Hofman, J. C. Bis, A. Correa, M. E. Griswold, J. Jakobsdottir, A. V. Smith, P. J. Schreiner, M. F. Feitosa, Q. Zhang, J. E. Huffman, J. Crosby, C. L. Wassel, R. Do, N. Franceschini, L. W. Martin, J. G. Robinson, T. L. Assimes, D. R. Crosslin, E. A. Rosenthal, M. Tsai, M. J. Rieder, D. N. Farlow, A. R. Folsom, T. Lumley, E. R. Fox, C. S. Carlson, U. Peters, R. D. Jackson, C. M. van Duijn, A. G. Uitterlinden, D. Levy, J. I. Rotter, H. A. Taylor, J. Gudnason, V., D. S. Siscovick, M. Fornage, I. B. Borecki, C. Hayward, I. Rudan, Y. E. Chen, E. P. Bottinger, R. J. Loos, P. Saetrom, K. Hveem, M. Boehnke, L. Groop, M. McCarthy, T. Meitinger, C. M. Ballantyne, S. B. Gabriel, C. J. O'Donnell, W. S. Post, K. E. North, A. P. Reiner, E. Boerwinkle, B. M. Psaty, D. Altshuler, S. Kathiresan, D. Y. Lin, G. P. Jarvik, L. A. Cupples, C. Kooperberg, J. G. Wilson, D. A. Nickerson, G. R. Abecasis, S. S. Rich, *et al.*, "Whole-exome sequencing identifies rare and low-frequency coding variants associated with ldl cholesterol," *Am J Hum Genet*, vol. 94, no. 2, pp. 233–45, 2014.

[187] K. E. Lohmueller, T. Sparso, Q. Li, E. Andersson, T. Korneliussen, A. Albrechtsen, K. Banasik, N. Grarup, I. Hallgrimsdottir, K. Kiil, T. O. Kilpelainen, N. T. Krarup, T. H. Pers, G. Sanchez, Y. Hu, M. Degiorgio, T. Jorgensen, A. Sandbaek, T. Lauritzen, S. Brunak, K. Kristiansen, Y. Li, T. Hansen, J. Wang, R. Nielsen, and O. Pedersen, "Whole-exome sequencing of 2,000 danish individuals and the role of rare coding variants in type 2 diabetes," *Am J Hum Genet*, vol. 93, no. 6, pp. 1072–86, 2013.

[188] S. M. Purcell, J. L. Moran, M. Fromer, D. Ruderfer, N. Solovieff, P. Roussos, C. O'Dushlaine, K. Chambert, S. E. Bergen, A. Kahler, L. Duncan, E. Stahl, G. Genovese, E. Fernandez, M. O. Collins, N. H. Komiyama, J. S. Choudhary, P. K. Magnusson, E. Banks, K. Shakir, K. Garimella, T. Fennell, M. DePristo, S. G. Grant, S. J. Haggarty, S. Gabriel, E. M. Scolnick, E. S. Lander, C. M. Hultman, P. F. Sullivan, S. A. McCarroll, and P. Sklar, "A polygenic burden of rare disruptive mutations in schizophrenia," *Nature*, vol. 506, no. 7487, pp. 185–90, 2014.

[189] M. Ladouceur, Z. Dastani, Y. S. Aulchenko, C. M. Greenwood, and J. B. Richards, "The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals," *PLoS Genet*, vol. 8, no. 2, p. e1002496, 2012.

[190] D. J. Liu and S. M. Leal, "A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions," *PLoS Genet*, vol. 6, no. 10, p. e1001156, 2010.

[191] P. C. Sham and S. M. Purcell, "Statistical power and significance testing in large-scale genetic studies," *Nat Rev Genet*, vol. 15, no. 5, pp. 335–46, 2014.

[192] S. E. Antonarakis, A. Chakravarti, J. C. Cohen, and J. Hardy, "Mendelian disorders and multifactorial traits: the big divide or one for all?," *Nat Rev Genet*, vol. 11, no. 5, pp. 380–4, 2010.

[193] J. C. Cohen, "Emerging LDL therapies: Using human genetics to discover new therapeutic targets for plasma lipids," *J Clin Lipidol*, vol. 7, no. 3 Suppl, pp. S1–5, 2013.

[194] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, "New approaches to population stratification in genome-wide association studies," *Nat Rev Genet*, vol. 11, no. 7, pp. 459–63, 2010.

[195] J. Asimit and E. Zeggini, "Rare variant association analysis methods for complex traits," *Annu Rev Genet*, vol. 44, pp. 293–308, 2010.

[196] P. L. Auer and G. Lettre, "Rare variant association studies: considerations, challenges and opportunities," *Genome Med*, vol. 7, no. 1, p. 16, 2015.

[197] V. Bansal, O. Libiger, A. Torkamani, and N. J. Schork, "Statistical analysis strategies for association studies involving rare variants," *Nat Rev Genet*, vol. 11, no. 11, pp. 773–85, 2010.

[198] W. Bodmer and C. Bonilla, "Common and rare variants in multifactorial susceptibility to common diseases," *Nat Genet*, vol. 40, no. 6, pp. 695–701, 2008.

[199] R. Do, S. Kathiresan, and G. R. Abecasis, "Exome sequencing and complex disease: practical aspects of rare variant association studies," *Hum Mol Genet*, vol. 21, no. R1, pp. R1–9, 2012.

[200] S. Morgenthaler and W. G. Thilly, "A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CASt)," *Mutat Res*, vol. 615, no. 1-2, pp. 28–56, 2007.

[201] B. E. Madsen and S. R. Browning, "A groupwise association test for rare mutations using a weighted sum statistic," *PLoS Genet*, vol. 5, no. 2, p. e1000384, 2009.

[202] B. M. Neale, M. A. Rivas, B. F. Voight, D. Altshuler, B. Devlin, M. Orho-Melander, S. Kathiresan, S. M. Purcell, K. Roeder, and M. J. Daly, "Testing for an unusual distribution of rare variants," *PLoS Genet*, vol. 7, no. 3, p. e1001322, 2011.

[203] M. Abifadel, M. Varret, J. P. Rabes, D. Allard, K. Ouguerram, M. Devillers, C. Cruaud, S. Benjannet, L. Wickham, D. Erlich, A. Derre, L. Villeger, M. Farnier, I. Beucler, E. Bruckert, J. Chambaz, B. Chanu, J. M. Lecerf, G. Luc, P. Moulin, J. Weissenbach, A. Prat, M. Krempf, C. Junien, N. G. Seidah, and C. Boileau, "Mutations in PCSK9 cause autosomal dominant hypercholesterolemia," *Nat Genet*, vol. 34, no. 2, pp. 154–6, 2003.

[204] D. Allard, S. Amsellem, M. Abifadel, M. Trillard, M. Devillers, G. Luc, M. Krempf, Y. Reznik, J. P. Girardet, A. Fredenrich, C. Junien, M. Varret, C. Boileau, P. Benlian, and J. P. Rabes, "Novel mutations of the PCSK9 gene cause variable phenotype of autosomal dominant hypercholesterolemia," *Hum Mutat*, vol. 26, no. 5, p. 497, 2005.

[205] P. N. Hopkins, J. Defesche, S. W. Fouchier, E. Bruckert, G. Luc, B. Cariou, B. Sjouke, T. P. Leren, M. Harada-Shiba, H. Mabuchi, J. P. Rabes, A. Carrie, C. van Heyningen, V. Carreau, M. Farnier, Y. P. Teoh, M. Bourbon, M. A. Kawashiri, A. Nohara, H. Soran, A. D. Marais, H. Tada, M. Abifadel, C. Boileau, B. Chanu, S. Katsuda, I. Kishimoto, G. Lambert, H. Makino, Y. Miyamoto, M. Pichelin, K. Yagi, M. Yamagishi, Y. Zair, S. Mellis, G. D. Yancopoulos, N. Stahl, J. Mendoza, Y. Du, S. Hamon, M. Krempf, and G. D. Swergold, "Characterization of autosomal dominant hypercholesterolemia caused by PCsk9 gain of function mutations and its specific treatment with alirocumab, a PCsk9 monoclonal antibody," *Circ Cardiovasc Genet*, vol. 8, no. 6, pp. 823–31, 2015.

[206] T. P. Leren, "Mutations in the PCsk9 gene in norwegian subjects with autosomal dominant hypercholesterolemia," *Clin Genet*, vol. 65, no. 5, pp. 419–22, 2004.

[207] S. Basu and W. Pan, "Comparison of Statistical Tests for Disease Association with Rare Variants," *Genetic epidemiology*, vol. 35, no. 7, pp. 606–619, 2011.

[208] S. Lee, M. J. Emond, M. J. Bamshad, K. C. Barnes, M. J. Rieder, D. A. Nickerson, N. G. E. S. P.-E. L. P. Team, D. C. Christiani, M. M. Wurfel, and X. Lin, "Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies," *Am J Hum Genet*, vol. 91, no. 2, pp. 224–37, 2012.

[209] I. J. Barnett, S. Lee, and X. Lin, "Detecting rare variant effects using extreme phenotype sampling in sequencing association studies," *Genet Epidemiol*, vol. 37, no. 2, pp. 142–51, 2013.

[210] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-variant association testing for sequencing data with the sequence kernel association test," *Am J Hum Genet*, vol. 89, no. 1, pp. 82–93, 2011.

[211] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork, A. S. Kondrashov, and S. R. Sunyaev, "A method and server for predicting damaging missense mutations," *Nat Methods*, vol. 7, no. 4, pp. 248–9, 2010.

[212] P. Kumar, S. Henikoff, and P. C. Ng, "Predicting the effects of coding non-synonymous variants on protein function using the sift algorithm," *Nat Protoc*, vol. 4, no. 7, pp. 1073–81, 2009.

[213] E. V. Davydov, D. L. Goode, M. Sirota, G. M. Cooper, A. Sidow, and S. Batzoglou, "Identifying a high fraction of the human genome to be under selective constraint using GERP++," *PLoS Comput Biol*, vol. 6, no. 12, p. e1001025, 2010.

[214] J. C. Taylor, H. C. Martin, S. Lise, J. Broxholme, J.-B. Cazier, A. Rimmer, A. Kanapin, G. Lunter, S. Fiddy, C. Allan, A. R. Aricescu, M. Attar, C. Babbs, J. Becq, D. Beeson, C. Bento, P. Bignell, E. Blair, V. J. Buckle, K. Bull, O. Cais, H. Cario, H. Chapel, R. R. Copley, R. Cornall, J. Craft, K. Dahan, E. E. Davenport, C. Dendrou, O. Devuyst, A. L. Fenwick, J. Flint, L. Fugger, R. D. Gilbert, A. Goriely, A. Green, I. H. Greger, R. Grocock, A. V. Gruszczyk, R. Hastings, E. Hatton, D. Higgs, A. Hill, C. Holmes, M. Howard, L. Hughes, P. Humburg, D. Johnson, F. Karpe, Z. Kingsbury, U. Kini, J. C. Knight, J. Krohn, S. Lamble, C. Langman, L. Lonie, J. Luck, D. McCarthy, S. J. McGowan, M. F. McMullin, K. A. Miller, L. Murray, A. H. Nemeth, M. A. Nesbit, D. Nutt, E. Ormondroyd, A. B. Oturai, A. Pagnamenta, S. Y. Patel, M. Percy, N. Petousi, P. Piazza, S. E. Piret, G. Polanco-Echeverry, N. Popitsch, F. Powrie, C. Pugh, L. Quek, P. A. Robbins, K. Robson, A. Russo, N. Sahgal, P. A. van Schouwenburg, A. Schuh, E. Silverman, A. Simmons, P. S. Sorensen, E. Sweeney, J. Taylor, R. V. Thakker, I. Tomlinson, A. Trebes, S. R. F. Twigg, H. H. Uhlig, P. Vyas, T. Vyse, S. A. Wall, H. Watkins, M. P. Whyte, L. Witty, *et al.*, "Factors influencing success of clinical genome sequencing across a broad spectrum of disorders," *Nat Genet*, vol. 47, no. 7, pp. 717–726, 2015.

[215] L. A. Miosge, M. A. Field, Y. Sontani, V. Cho, S. Johnson, A. Palkova, B. Balakishnan, R. Liang, Y. Zhang, S. Lyon, B. Beutler, B. Whittle, E. M. Bertram, A. Enders, C. C. Goodnow, and T. D. Andrews, "Comparison of predicted and actual consequences of missense mutations," *Proceedings of the National Academy of Sciences*, vol. 112, no. 37, pp. E5189–E5198, 2015.

[216] J. L. Casanova, M. E. Conley, S. J. Seligman, L. Abel, and L. D. Notarangelo, "Guidelines for genetic studies in single patients: lessons from primary immunodeficiencies," *J Exp Med*, vol. 211, no. 11, pp. 2137–49, 2014.

[217] Y. Itan and J. L. Casanova, "Can the impact of human genetic variations be predicted?," *Proc Natl Acad Sci U S A*, vol. 112, no. 37, pp. 11426–7, 2015.

[218] J. B. S. Haldane, "The rate of mutation of human genes," *Hereditas*, vol. 35, no. S1, pp. 267–273, 1949.

[219] T. I. Sorensen, G. G. Nielsen, P. K. Andersen, and T. W. Teasdale, "Genetic and environmental influences on premature death in adult adoptees," *N Engl J Med*, vol. 318, no. 12, pp. 727–32, 1988.

[220] L. B. Barreiro and L. Quintana-Murci, "From evolutionary genetics to human immunology: how selection shapes host defence genes," *Nat Rev Genet*, vol. 11, no. 1, pp. 17–30, 2010.

[221] M. Fumagalli and M. Sironi, "Human genome variability, natural selection and infectious diseases," *Curr Opin Immunol*, vol. 30, pp. 9–16, 2014.

[222] E. K. Karlsson, D. P. Kwiatkowski, and P. C. Sabeti, "Natural selection and infectious disease in human populations," *Nat Rev Genet*, vol. 15, no. 6, pp. 379–93, 2014.

[223] O. C. Bruton, "Agammaglobulinemia," *Pediatrics*, vol. 9, no. 6, pp. 722–8, 1952.

[224] M. S. Horwitz, Z. Duan, B. Korkmaz, H. H. Lee, M. E. Mealiffe, and S. J. Salipante, "Neutrophil elastase in cyclic and severe congenital neutropenia," *Blood*, vol. 109, no. 5, pp. 1817–24, 2007.

[225] R. H. Buckley, "Molecular defects in human severe combined immunodeficiency and approaches to immune reconstitution," *Annu Rev Immunol*, vol. 22, pp. 625–55, 2004.

[226] J. L. Casanova, C. Fieschi, J. Bustamante, J. Reichenbach, N. Remus, H. von Bernuth, and C. Picard, "From idiopathic infectious diseases to novel primary immunodeficiencies," *J Allergy Clin Immunol*, vol. 116, no. 2, pp. 426–30, 2005.

[227] J. L. Casanova and L. Abel, "Primary immunodeficiencies: a field in its infancy," *Science*, vol. 317, no. 5838, pp. 617–9, 2007.

[228] C. L. Ku, C. Picard, M. Erdos, A. Jeurissen, J. Bustamante, A. Puel, H. von Bernuth, O. Filipe-Santos, H. H. Chang, T. Lawrence, M. Raes, L. Marodi, X. Bossuyt, and J. L. Casanova, "IRak4 and nemo mutations in otherwise healthy children with recurrent invasive pneumococcal disease," *J Med Genet*, vol. 44, no. 1, pp. 16–23, 2007.

[229] C. Picard, J. L. Casanova, and A. Puel, "Infectious diseases in patients with IRak-4, myd88, nemo, or ikappabalpha deficiency," *Clin Microbiol Rev*, vol. 24, no. 3, pp. 490–7, 2011.

[230] C. Picard, A. Puel, J. Bustamante, C. L. Ku, and J. L. Casanova, "Primary immunodeficiencies associated with pneumococcal disease," *Curr Opin Allergy Clin Immunol*, vol. 3, no. 6, pp. 451–9, 2003.

[231] A. Smahi, G. Courtois, S. H. Rabia, R. Doffinger, C. Bodemer, A. Munnich, J. L. Casanova, and A. Israel, "The NF-kappab signalling pathway in human diseases: from incontinentia pigmenti to ectodermal dysplasias and immune-deficiency syndromes," *Hum Mol Genet*, vol. 11, no. 20, pp. 2371–5, 2002.

[232] H. von Bernuth, C. Picard, Z. Jin, R. Pankla, H. Xiao, C. L. Ku, M. Chrabieh, I. B. Mustapha, P. Ghandil, Y. Camcioglu, J. Vasconcelos, N. Sirvent, M. Guedes, A. B. Vitor, M. J. Herrero-Mata, J. I. Arostegui, C. Rodrigo, L. Alsina, E. Ruiz-Ortiz, M. Juan, C. Fortuny, J. Yague, J. Anton, M. Pascal, H. H. Chang, L. Janniere, Y. Rose, B. Z. Garty, H. Chapel, A. Issekutz, L. Marodi, C. Rodriguez-Gallego, J. Banchereau, L. Abel, X. Li, D. Chaussabel, A. Puel, and J. L. Casanova, "Pyogenic bacterial infections in humans with MyD88 deficiency," *Science*, vol. 321, no. 5889, pp. 691–6, 2008.

[233] F. Altare, D. Lammas, P. Revy, E. Jouanguy, R. Doffinger, S. Lamhamedi, P. Drysdale, D. Scheel-Toellner, J. Girdlestone, P. Darbyshire, M. Wadhwa, H. Dockrell, M. Salmon, A. Fischer, A. Durandy, J. L. Casanova, and D. S. Kumararatne, "Inherited interleukin 12 deficiency in a child with bacille Calmette-Guerin and Salmonella enteritidis disseminated infection," *J Clin Invest*, vol. 102, no. 12, pp. 2035–40, 1998.

[234] R. de Jong, F. Altare, I. A. Haagen, D. G. Elferink, T. Boer, P. J. van Breda Vriesman, P. J. Kabel, J. M. Draaisma, J. T. van Dissel, F. P. Kroon, J. L. Casanova, and T. H. Ottenhoff, "Severe mycobacterial and salmonella infections in interleukin-12 receptor-deficient patients," *Science*, vol. 280, no. 5368, pp. 1435–8, 1998.

[235] S. E. Dorman and S. M. Holland, "Mutation in the signal-transducing chain of the interferon-gamma receptor and susceptibility to mycobacterial infection," *J Clin Invest*, vol. 101, no. 11, pp. 2364–9, 1998.

[236] M. J. Newport, C. M. Huxley, S. Huston, C. M. Hawrylowicz, B. A. Oostra, R. Williamson, and M. Levin, "A mutation in the interferon-gamma-receptor gene and susceptibility to mycobacterial infection," *N Engl J Med*, vol. 335, no. 26, pp. 1941–9, 1996.

[237] S. Barroso, M. Lopez-Trascasa, D. Merino, A. J. Alvarez, A. Nunez-Roldan, and B. Sanchez, "C7 deficiency and meningococcal infection susceptibility in two spanish families," *Scand J Immunol*, vol. 72, no. 1, pp. 38–43, 2010.

[238] A. P. Dale and R. C. Read, "Genetic susceptibility to meningococcal infection," *Expert Rev Anti Infect Ther*, vol. 11, no. 2, pp. 187–99, 2013.

[239] L. A. Lewis and S. Ram, "Meningococcal disease and the complement system," *Virulence*, vol. 5, no. 1, pp. 98–126, 2014.

[240] M. Schlesinger, Z. Nave, Y. Levy, P. E. Slater, and Z. Fishelson, "Prevalence of hereditary properdin, c7 and c8 deficiencies in patients with meningococcal infections," *Clin Exp Immunol*, vol. 81, no. 3, pp. 423–7, 1990.

[241] A. G. Sjoholm, J. H. Braconier, and C. Soderstrom, "Properdin deficiency in a family with fulminant meningococcal infections," *Clin Exp Immunol*, vol. 50, no. 2, pp. 291–7, 1982.

[242] C. Slade, J. Bosco, G. Unglik, K. Bleasel, M. Nagel, and I. Winship, "Deficiency in complement factor b," *N Engl J Med*, vol. 369, no. 17, pp. 1667–9, 2013.

[243] I. Smirnova, N. Mann, A. Dols, H. H. Derkx, M. L. Hibberd, M. Levin, and B. Beutler, "Assay of locus-specific genetic load implicates rare toll-like receptor 4 mutations in meningococcal susceptibility," *Proc Natl Acad Sci U S A*, vol. 100, no. 10, pp. 6075–80, 2003.

[244] A. Rautanen, T. C. Mills, A. C. Gordon, P. Hutton, M. Steffens, R. Nuamah, J. D. Chiche, T. Parks, S. J. Chapman, E. E. Davenport, K. S. Elliott, J. Bion, P. Lichtner, T. Meitinger, T. F. Wienker, M. J. Caulfield, C. Mein, F. Bloos, I. Bobek, P. Cotogni, V. Sramek, S. Sarapuu, M. Kobilay, V. M. Ranieri, J. Rello, G. Sirgo, Y. G. Weiss, S. Russwurm, E. M. Schneider, K. Reinhart, P. A. Holloway, J. C. Knight, C. S. Garrard, J. A. Russell, K. R. Walley, F. Stuber, A. V. Hill, C. J. Hinds, and E. E. G. Investigators, "Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study," *Lancet Respir Med*, vol. 3, no. 1, pp. 53–60, 2015.

[245] "https://www.hta.gov.uk/human-tissue-act-2004."

[246] M. Pembrey and A. S. Team, "The avon longitudinal study of parents and children (alspac): a resource for genetic epidemiology," *Eur J Endocrinol*, vol. 151 Suppl 3, pp. U125–9, 2004.

[247] S. E. Mumford, "Children of the 90s: ethical guidance for a longitudinal study," *Arch Dis Child Fetal Neonatal Ed*, vol. 81, no. 2, pp. F146–51, 1999.

[248] S. E. Mumford, "Children of the 90s ii: challenges for the ethics and law committee," *Arch Dis Child Fetal Neonatal Ed*, vol. 81, no. 3, pp. F228–31, 1999.

[249] B. H. Smith, A. Campbell, P. Linksted, B. Fitzpatrick, C. Jackson, S. M. Kerr, I. J. Deary, D. J. Macintyre, H. Campbell, M. McGilchrist, L. J. Hocking, L. Wisely, I. Ford, R. S. Lindsay, R. Morton, C. N. Palmer, A. F. Dominiczak, D. J. Porteous, and A. D. Morris, "Cohort profile: Generation Scotland: Scottish family health study (gs:sfhs). the study, its participants and their potential for genetic research on health and illness," *Int J Epidemiol*, vol. 42, no. 3, pp. 689–700, 2013.

[250] C. Power and J. Elliott, "Cohort profile: 1958 British birth cohort (national child development study)," *Int J Epidemiol*, vol. 35, no. 1, pp. 34–41, 2006.

[251] G. Lunter and M. Goodson, "Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads," *Genome Res*, vol. 21, no. 6, pp. 936–9, 2011.

[252] H. Li and R. Durbin, "Fast and accurate short read alignment with burrows-wheeler transform," *Bioinformatics*, vol. 25, no. 14, pp. 1754–60, 2009.

[253] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. Del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo, "From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline," *Curr Protoc Bioinformatics*, vol. 11, no. 1110, pp. 11 101–1110 33, 2013.

[254] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, "A framework for variation discovery and genotyping using next-generation DNA sequencing data," *Nat Genet*, vol. 43, no. 5, pp. 491–8, 2011.

[255] M. Pirooznia, M. Kramer, J. Parla, F. S. Goes, J. B. Potash, W. R. McCombie, and P. P. Zandi, "Validation and assessment of variant calling pipelines for next-generation sequencing," *Hum Genomics*, vol. 8, p. 14, 2014.

[256] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo, "The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data," *Genome Res*, vol. 20, no. 9, pp. 1297–303, 2010.

[257] D. H. Ghoneim, J. R. Myers, E. Tuttle, and A. R. Paciorkowski, "Comparison of insertion/deletion calling algorithms on human next-generation sequencing data," *BMC Res annotes*, vol. 7, p. 864, 2014.

[258] C. Genomes Project, G. R. Abecasis, A. Auton, L. D. Brooks, M. A. DePristo, R. M. Durbin, R. E. Handsaker, H. M. Kang, G. T. Marth, and G. A. McVean, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[259] GeraldineVdAuwera, "(howto) recalibrate variant quality scores = run VQsr," 2013-12.

[260] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and G. Genomes Project Analysis, "The variant call format and VCFtools," *Bioinformatics*, vol. 27, no. 15, pp. 2156–8, 2011.

[261] W. McLaren, B. Pritchard, D. Rios, Y. Chen, P. Flicek, and F. Cunningham, "Deriving the consequences of genomic variants with the ensembl API and snp effect predictor," *Bioinformatics*, vol. 26, no. 16, pp. 2069–70, 2010.

[262] D. MacArthur, "Exome aggregation consortium (exac)."

[263] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS Genet*, vol. 2, no. 12, p. e190, 2006.

[264] A. L. Price, G. V. Kryukov, P. I. de Bakker, S. M. Purcell, J. Staples, L. J. Wei, and S. R. Sunyaev, "Pooled association tests for rare variants in exon-resequencing studies," *Am J Hum Genet*, vol. 86, no. 6, pp. 832–8, 2010.

[265] B. Devlin and K. Roeder, "Genomic control for association studies," *Biometrics*, vol. 55, no. 4, pp. 997–1004, 1999.

[266] B. Devlin, K. Roeder, and L. Wasserman, "Genomic control, a new approach to genetic-based association studies," *Theor Popul Biol*, vol. 60, no. 3, pp. 155–66, 2001.

[267] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, "Plink: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, vol. 81, no. 3, pp. 559–75, 2007.

[268] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nat Genet*, vol. 38, no. 8, pp. 904–9, 2006.

[269] J. Yu, G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, S. Kresovich, and E. S. Buckler, "A unified mixed-model method for association mapping that accounts for multiple levels of relatedness," *Nat Genet*, vol. 38, no. 2, pp. 203–8, 2006.

[270] H. M. Kang, J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong, N. B. Freimer, C. Sabatti, and E. Eskin, "Variance component model to account for sample structure in genome-wide association studies," *Nat Genet*, vol. 42, no. 4, pp. 348–54, 2010.

[271] S. Purcell, "Plink/seq a library for the analysis of genetic variation data."

[272] L. Moutsianas, V. Agarwala, C. Fuchsberger, J. Flannick, M. A. Rivas, K. J. Gaulton, P. K. Albers, T. D. C. Go, G. McVean, M. Boehnke, D. Altshuler, and M. I. McCarthy, "The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease," *PLoS Genet*, vol. 11, no. 4, p. e1005165, 2015.

[273] A. Franceschini, D. Szklarczyk, S. Frankild, M. Kuhn, M. Simonovic, A. Roth, J. Lin, P. Minguez, P. Bork, C. von Mering, and L. J. Jensen, "String v9.1: protein-protein interaction networks, with increased coverage and integration," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D808–15, 2013.

[274] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nat Genet*, vol. 25, no. 1, pp. 25–9, 2000.

[275] J. Wixon and D. Kell, "The kyoto encyclopedia of genes and genomes–kegg," *Yeast*, vol. 17, no. 1, pp. 48–55, 2000.

[276] T. N. Petersen, S. Brunak, G. von Heijne, and H. Nielsen, "Signalp 4.0: discriminating signal peptides from transmembrane regions," *Nat Methods*, vol. 8, no. 10, pp. 785–6, 2011.

[277] F. Ponten, K. Jirstrom, and M. Uhlen, "The human protein atlas-a tool for pathology," *The Journal of Pathology*, vol. 216, no. 4, pp. 387–393, 2008.

[278] C. L. Vermont, R. de Groot, and J. A. Hazelzet, "Bench-to-bedside review: Genetic influences on meningococcal disease," *Critical Care*, vol. 6, no. 1, pp. 60–65, 2002.

[279] E. Haralambous, H. A. Weiss, A. Radalowicz, M. L. Hibberd, R. Booy, and M. Levin, "Sibling familial risk ratio of meningococcal disease in uk caucasians," *Epidemiol Infect*, vol. 130, no. 3, pp. 413–8, 2003.

[280] V. Wright, M. Hibberd, and M. Levin, "Genetic polymorphisms in host response to meningococcal infection: the role of susceptibility and severity genes," *Vaccine*, vol. 27 Suppl 2, pp. B90–102, 2009.

[281] S. Davila, V. J. Wright, C. C. Khor, K. S. Sim, A. Binder, W. B. Breunis, D. Inwald, S. Nadel, H. Betts, E. D. Carrol, R. de Groot, P. W. Hermans, J. Hazelzet, M. Emonts, C. C. Lim, T. W. Kuijpers, F. Martinon-Torres, A. Salas, W. Zenz, M. Levin, M. L. Hibberd, and C. International Meningococcal Genetics, "Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease," *Nat Genet*, vol. 42, no. 9, pp. 772–6, 2010.

[282] E. P. Owen, R. Wurzner, F. Leisegang, P. Rizkallah, A. Whitelaw, J. Simpson, A. D. Thomas, C. L. Harris, J. L. Giles, B. C. Hellerud, T. E. Mollnes, B. P. Morgan, P. C. Potter, and A. Orren, "A complement c5 gene mutation, c.754g>a:p.a252T, is common in the western cape, South Africa and found to be homozygous in seven percent of black African meningococcal disease cases," *Mol Immunol*, vol. 64, no. 1, pp. 170–6, 2015.

[283] D. Westra, R. A. Kurvers, L. P. van den Heuvel, R. Wurzner, E. P. Hoppenreijs, M. van der Flier, N. C. van de Kar, and A. Warris, "Compound heterozygous mutations in the c6 gene of a child with recurrent infections," *Mol Immunol*, vol. 58, no. 2, pp. 201–5, 2014.

[284] A. Piotrowski, J. Xie, Y. F. Liu, A. B. Poplawski, A. R. Gomes, P. Madanecki, C. Fu, M. R. Crowley, D. K. Crossman, L. Armstrong, D. Babovic-Vuksanovic, A. Bergner, J. O. Blakeley, A. L. Blumenthal, M. S. Daniels, H. Feit, K. Gardner, S. Hurst, C. Kobelka, C. Lee, R. Nagy, K. A. Rauen, J. M. Slopis, P. Suwannarat, J. A. Westman, A. Zanko, B. R. Korf, and L. M. Messiaen, "Germline loss-of-function mutations in lztr1 predispose to an inherited disorder of multiple schwannomas," *Nat Genet*, vol. 46, no. 2, pp. 182–7, 2014.

[285] J. T. Eppig, J. A. Blake, C. J. Bult, J. A. Kadin, and J. E. Richardson, "The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease," *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D726–36, 2015.

[286] S. M. Linton and B. P. Morgan, "Properdin deficiency and meningococcal diseaseâ€"identifying those most at risk," *Clinical and Experimental Immunology*, vol. 118, no. 2, pp. 189–191, 1999.

[287] M. Schlesinger, U. Mashal, J. Levy, and Z. Fishelson, "Hereditary properdin deficiency in three families of Tunisian jews," *Acta Paediatr*, vol. 82, no. 9, pp. 744–7, 1993.

[288] P. J. Spath, A. G. Sjoholm, G. N. Fredrikson, G. Misiano, R. Scherz, U. B. Schaad, B. Uhring-Lambert, G. Hauptmann, J. Westberg, M. Uhlen, C. Wadelius, and L. Truedsson, "Properdin deficiency in a large swiss family: identification of a stop codon in the properdin gene, and association of meningococcal disease with lack of the igg2 allotype marker g2m(n)," *Clin Exp Immunol*, vol. 118, no. 2, pp. 278–84, 1999.

[289] W. Fu, T. D. O'Connor, G. Jun, H. M. Kang, G. Abecasis, S. M. Leal, S. Gabriel, M. J. Rieder, D. Altshuler, J. Shendure, D. A. Nickerson, M. J. Bamshad, N. E. S. Project, and J. M. Akey, "Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants," *Nature*, vol. 493, no. 7431, pp. 216–20, 2013.

[290] M. Brouwer, *Bacterial meningitis in adults: clinical characteristics, risk factors and adjunctive treatment.* Thesis, 2010.

[291] D. T. Fearon and K. F. Austen, "Properdin: binding to c3b and stabilization of the c3b-dependent c3 convertase," *J Exp Med*, vol. 142, no. 4, pp. 856–63, 1975.

[292] W. Schwaeble, H. P. Huemer, J. Most, M. P. Dierich, M. Strobel, C. Claus, K. B. Reid, and H. W. Ziegler-Heitbrock, "Expression of properdin in human monocytes," *Eur J Biochem*, vol. 219, no. 3, pp. 759–64, 1994.

[293] C. A. Fijen, E. J. Kuijper, M. T. te Bulte, M. R. Daha, and J. Dankert, "Assessment of complement deficiency in patients with meningococcal disease in the Netherlands," *Clin Infect Dis*, vol. 28, no. 1, pp. 98–105, 1999.

[294] M.-n. Alcorlo, A.-n. Tortajada, S. RodrÃguez de CÃ³rdoba, and O. Llorca, "Structural basis for the stabilization of the complement alternative pathway C3 convertase by properdin," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 33, pp. 13504–13509, 2013.

[295] Z. Sun, K. B. M. Reid, and S. J. Perkins, "The dimeric and trimeric solution structures of the multidomain complement protein properdin by x-ray scattering, analytical ultracentrifugation and constrained modelling," *Journal of Molecular Biology*, vol. 343, no. 5, pp. 1327–1343, 2004.

[296] L. Ying, Y. Katz, M. Schlesinger, R. Carmi, H. Shalev, N. Haider, G. Beck, V. C. Sheffield, and D. Landau, "Complement factor h gene mutation associated with autosomal recessive atypical hemolytic uremic syndrome," *American Journal of Human Genetics*, vol. 65, no. 6, pp. 1538–1546, 1999.

[297] G. A. Silverman, S. C. Pak, and D. H. Perlmutter, "Disorders of protein misfolding: alpha-1-antitrypsin deficiency as prototype," *J Pediatr*, vol. 163, no. 2, pp. 320–6, 2013.

[298] G. Vogt, J. Bustamante, A. Chapgier, J. Feinberg, S. Boisson Dupuis, C. Picard, N. Mahlaoui, L. Gineau, A. Alcais, C. Lamaze, J. M. Puck, G. de Saint Basile, C. D. Khayat, R. Mikhael, and J. L. Casanova, "Complementation of a pathogenic IFngr2 misfolding mutation with modifiers of n-glycosylation," *J Exp Med*, vol. 205, no. 8, pp. 1729–37, 2008.

[299] D. G. MacArthur, S. Balasubramanian, A. Frankish, N. Huang, J. Morris, K. Walter, L. Jostins, L. Habegger, J. K. Pickrell, S. B. Montgomery, C. A. Albers, Z. D. Zhang, D. F. Conrad, G. Lunter, H. Zheng, Q. Ayub, M. A. DePristo, E. Banks, M. Hu, R. E. Handsaker, J. A. Rosenfeld, M. Fromer, M. Jin, X. J. Mu, E. Khurana, K. Ye, M. Kay, G. I. Saunders, M. M. Suner, T. Hunt, I. H. Barnes, C. Amid, D. R. Carvalho-Silva, A. H. Bignell, C. Snow, B. Yngvadottir, S. Bumpstead, D. N. Cooper, Y. Xue, I. G. Romero, C. Genomes Project, J. Wang, Y. Li, R. A. Gibbs, S. A. McCarroll, E. T. Dermitzakis, J. K. Pritchard, J. C. Barrett, J. Harrow, M. E. Hurles, M. B. Gerstein, and C. Tyler-Smith, "A systematic survey of loss-of-function variants in human protein-coding genes," *Science*, vol. 335, no. 6070, pp. 823–8, 2012.

[300] S. Petrovski, Q. Wang, E. L. Heinzen, A. S. Allen, and D. B. Goldstein, "Genic intolerance to functional variation and the interpretation of personal genomes," *PLoS Genet*, vol. 9, no. 8, p. e1003709, 2013.

[301] H. Wang, T. A. Kadlecek, B. B. Au-Yeung, H. E. S. Goodfellow, L.-Y. Hsu, T. S. Freedman, and A. Weiss, "Zap-70: An essential kinase in t-cell signaling," *Cold Spring Harbor Perspectives in Biology*, vol. 2, no. 5, p. a002279, 2010.

[302] A. A. Okoye and L. J. Picker, "CD4(+) t cell depletion in hiv infection: mechanisms of immunological failure," *Immunological reviews*, vol. 254, no. 1, pp. 54–64, 2013.

[303] W. Yu, M. Gwinn, M. Clyne, A. Yesupriya, and M. J. Khoury, "A navigator for human genome epidemiology," *Nat Genet*, vol. 40, no. 2, pp. 124–5, 2008.

[304] M. C. Brouwer, J. de Gans, S. G. B. Heckenberg, A. H. Zwinderman, T. van der Poll, and D. van de Beek, "Host genetic susceptibility to pneumococcal and meningococcal disease: a systematic review and meta-analysis," *The Lancet Infectious Diseases*, vol. 9, no. 1, pp. 31–44, 2009.

[305] N. C. Arbour, E. Lorenz, B. C. Schutte, J. Zabner, J. N. Kline, M. Jones, K. Frees, J. L. Watt, and D. A. Schwartz, "TLR4 mutations are associated with endotoxin hyporesponsiveness in humans," *Nat Genet*, vol. 25, no. 2, pp. 187–91, 2000.

[306] S. Gibot, A. Cariou, L. Drouet, M. Rossignol, and L. Ripoll, "Association between a genomic polymorphism within the CD14 locus and septic shock susceptibility and mortality rate," *Crit Care Med*, vol. 30, no. 5, pp. 969–73, 2002.

[307] A. C. Gordon, A. L. Lagan, E. Aganna, L. Cheung, C. J. Peters, M. F. McDermott, J. L. Millo, K. I. Welsh, P. Holloway, G. A. Hitman, R. D. Piper, C. S. Garrard, and C. J. Hinds, "Tnf and tnfr polymorphisms in severe sepsis and septic shock: a prospective multicentre study," *Genes Immun*, vol. 5, no. 8, pp. 631–40, 2004.

[308] M. Kompoti, A. Michopoulos, M. Michalia, P. M. Clouva-Molyvdas, A. E. Germenis, and M. Speletas, "Genetic polymorphisms of innate and adaptive immunity as predictors of outcome in critically ill patients," *Immunobiology*, vol. 220, no. 3, pp. 414–21, 2015.

[309] T. C. Mills, S. Chapman, P. Hutton, A. C. Gordon, J. Bion, J. D. Chiche, P. A. Holloway, F. Stuber, C. S. Garrard, C. J. Hinds, A. V. Hill, A. Rautanen, and E. E. G. Investigators, "Variants in the mannose-binding lectin gene mbl2 do not associate with sepsis susceptibility or survival in a large European cohort," *Clin Infect Dis*, 2015.

[310] T. A. Nakada, J. A. Russell, J. H. Boyd, and K. R. Walley, "Il17A genetic variation is associated with altered susceptibility to gram-positive infection and mortality of severe sepsis," *Crit Care*, vol. 15, no. 5, p. R254, 2011.

[311] B. Schluter, C. Raufhake, M. Erren, H. Schotte, F. Kipp, S. Rust, H. Van Aken, G. Assmann, and E. Berendes, "Effect of the interleukin-6 promoter polymorphism (-174 g/c) on the incidence and outcome of sepsis," *Crit Care Med*, vol. 30, no. 1, pp. 32–7, 2002.

[312] D. L. Zhang, H. M. Zheng, B. J. Yu, Z. W. Jiang, and J. S. Li, "Association of polymorphisms of il and CD14 genes with acute severe pancreatitis and septic shock," *World J Gastroenterol*, vol. 11, no. 28, pp. 4409–13, 2005.

[313] A. M. Sutherland and K. R. Walley, "Bench-to-bedside review: Association of genetic variation with sepsis," *Critical Care*, vol. 13, no. 2, pp. 210–210, 2009.

[314] M. Karakachoff, N. Duforet-Frebourg, F. Simonet, S. Le Scouarnec, N. Pellen, S. Lecointe, E. Charpentier, F. Gros, S. Cauchi, P. Froguel, N. Copin, T. Le Tourneau, V. Probst, H. Le Marec, S. Molinaro, B. Balkau, R. Redon, J. J. Schott, M. G. Blum, and C. Dina, "Fine-scale human genetic structure in western France," *Eur J Hum Genet*, vol. 23, no. 6, pp. 831–6, 2015.

[315] T. D. O'Connor, A. Kiezun, M. Bamshad, S. S. Rich, J. D. Smith, E. Turner, S. M. Leal, and J. M. Akey, "Fine-scale patterns of population stratification confound rare variant association tests," *PLoS One*, vol. 8, no. 7, p. e65834, 2013.

[316] C. Xing, G. A. Satten, and A. S. Allen, "A weighted accumulation test for associating rare genetic variation with quantitative phenotypes," *BMC Proc*, vol. 5 Suppl 9, p. S6, 2011.

[317] K. Q. Ye and C. D. Engelman, "Detecting multiple causal rare variants in exome sequence data," *Genet Epidemiol*, vol. 35 Suppl 1, pp. S18–21, 2011.

[318] M. C. Babron, M. de Tayrac, D. N. Rutledge, E. Zeggini, and E. Genin, "Rare and low frequency variant stratification in the UK population: description and impact on association tests," *PLoS One*, vol. 7, no. 10, p. e46519, 2012.

[319] I. Mathieson and G. McVean, "Differential confounding of rare and common variants in spatially structured populations," *Nat Genet*, vol. 44, no. 3, pp. 243–6, 2012.

[320] I. Ebermann, H. N. Scholl, P. Charbel Issa, E. Becirovic, J. Lamprecht, B. Jurklies, J. MillÃ¡n, E. Aller, D. Mitter, and H. Bolz, "A novel gene for usher syndrome type 2: mutations in the long isoform of whirlin are associated with retinitis pigmentosa and sensorineural hearing loss," *Human Genetics*, vol. 121, no. 2, pp. 203–211, 2007.

[321] P. Mburu, M. Mustapha, A. Varela, D. Weil, A. El-Amraoui, R. H. Holme, A. Rump, R. E. Hardisty, S. Blanchard, R. S. Coimbra, I. Perfettini, N. Parkinson, A. M. Mallon, P. Glenister, M. J. Rogers, A. J. Paige, L. Moir, J. Clay, A. Rosenthal, X. Z. Liu, G. Blanco, K. P. Steel, C. Petit, and S. D. Brown, "Defects in whirlin, a pdz domain molecule involved in stereocilia elongation, cause deafness in the whirler mouse and families with dfnb31," *Nat Genet*, vol. 34, no. 4, pp. 421–8, 2003.

[322] S. J. Chapman and A. V. Hill, "Human genetic susceptibility to infectious disease," *Nat Rev Genet*, vol. 13, no. 3, pp. 175–88, 2012.

[323] S. C. L. Gough and M. J. Simmonds, "The hla region and autoimmune disease: Associations and mechanisms of action," *Current Genomics*, vol. 8, no. 7, pp. 453–465, 2007.

[324] H. Gong, H. Zhou, G. W. McKenzie, Z. Yu, S. Clerens, J. M. Dyer, J. E. Plowman, M. W. Wright, R. Arora, C. S. Bawden, Y. Chen, J. Li, and J. G. H. Hickford, "An updated nomenclature for keratin-associated proteins (kaps)," *International Journal of Biological Sciences*, vol. 8, no. 2, pp. 258–264, 2012.

[325] M. A. Rogers and J. Schweizer, "Human kap genes, only the half of it? extensive size polymorphisms in hair keratin-associated protein genes," *J Investig Dermatol*, vol. 124, no. 6, pp. vii–ix, 2005.

[326] M. Maeda, H. Hasegawa, T. Hyodo, S. Ito, E. Asano, H. Yuang, K. Funasaka, K. Shimokata, Y. Hasegawa, M. Hamaguchi, and T. Senga, "Arhgap18, a gtpase-activating protein for rhoa, controls cell shape, spreading, and motility," *Molecular Biology of the Cell*, vol. 22, no. 20, pp. 3840–3852, 2011.

[327] G. Loirand, "Rho kinases in health and disease: From basic science to translational research," *Pharmacol Rev*, vol. 67, no. 4, pp. 1074–95, 2015.

[328] N. Heemskerk, J. van Rijssel, and J. D. van Buul, "Rho-gtpase signaling in leukocyte extravasation," *Cell Adhesion & Migration*, vol. 8, no. 2, pp. 67–75, 2014.

[329] G. van Nieuw Amerongen, C. Beckers, I. Achekar, S. Zeeman, R. Musters, and V. van Hinsbergh, "Involvement of rho kinase in endothelial barrier maintenance," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 27, no. 11, pp. 2332–2339, 2007.

[330] E. Oguz, B. Alasehirli, Y. Pehlivan, A. M. Onat, S. Oztuzcu, E. Ozkara, B. Kisacik, C. Camci, and A. T. DemiryÃ¼rek, "Association between rho-kinase (rock2) gene polymorphisms and behã§etâ€™s disease," *Translational Research*, vol. 160, no. 6, pp. 428–434, 2012.

[331] D. J. Angelini, S.-W. Hyun, D. N. Grigoryev, P. Garg, P. Gong, I. S. Singh, A. Passaniti, J. D. Hasday, and S. E. Goldblum, "TNf-î± increases tyrosine phosphorylation of vascular endothelial cadherin and opens the paracellular pathway through fyn activation in human lung endothelia," *American Journal of Physiology - Lung Cellular and Molecular Physiology*, vol. 291, no. 6, pp. L1232–L1245, 2006.

[332] B. Marcos-Ramiro, D. GarcÃa Weber, and J. MillÃ¡n, "Tnf-induced endothelial barrier disruption: beyond actin and rho," *Thrombosis and Haemostasis*, vol. 112, no. 12, pp. 1088–1102, 2014.

[333] S. Nadel, M. J. Newport, R. Booy, and M. Levin, "Variation in the tumor necrosis factor-alpha gene promoter region may be associated with death from meningococcal disease," *J Infect Dis*, vol. 174, no. 4, pp. 878–80, 1996.

[334] J. Piedra, S. Miravet, J. CastaÃ±o, H. G. PÃ¡lmer, N. Heisterkamp, A. GarcÃ-a de Herreros, and M. DuÃ±ach, "p120 Catenin-associated fer and fyn tyrosine kinases regulate î²-catenin tyr-142 phosphorylation and î²-catenin-î±-catenin interaction," *Molecular and Cellular Biology*, vol. 23, no. 7, pp. 2287–2297, 2003.

[335] E. H. Palacios and A. Weiss, "Function of the src-family kinases, lck and fyn, in t-cell development and activation," *Oncogene*, vol. 23, no. 48, pp. 7990–8000, 2004.

[336] M. Groger, W. Pasteiner, G. Ignatyev, U. Matt, S. Knapp, A. Atrasheuskaya, E. Bukin, P. Friedl, D. Zinkl, R. Hofer-Warbinek, K. Zacharowski, P. Petzelbauer, and S. Reingruber, "Peptide bbeta(15-42) preserves endothelial barrier function in shock," *PLoS One*, vol. 4, no. 4, p. e5391, 2009.

[337] X. Liang, N. A. Draghi, and M. D. Resh, "Signaling from integrins to fyn to rho family gtpases regulates morphologic differentiation of oligodendrocytes," *J Neurosci*, vol. 24, no. 32, pp. 7140–9, 2004.

[338] T. Palomero, L. Couronne, H. Khiabanian, M.-Y. Kim, A. Ambesi-Impiombato, A. Perez-Garcia, Z. Carpenter, F. Abate, M. Allegretta, J. E. Haydu, X. Jiang, I. S. Lossos, C. Nicolas, M. Balbin, C. Bastard, G. Bhagat, M. A. Piris, E. Campo, O. A. Bernard, R. Rabadan, and A. A. Ferrando, "Recurrent mutations in epigenetic regulators, rhoa and fyn kinase in peripheral t cell lymphomas," *Nat Genet*, vol. 46, pp. 166–170, Feb. 2014.

[339] Z. Poghosyan, S. M. Robbins, M. D. Houslay, A. Webster, G. Murphy, and D. R. Edwards, "Phosphorylation-dependent interactions between adam15 cytoplasmic domain and src family protein-tyrosine kinases," *J Biol Chem*, vol. 277, no. 7, pp. 4999–5007, 2002.

[340] W. R. Thompson, C. Guilluy, Z. Xie, B. Sen, K. E. Brobst, S. S. Yen, G. Uzer, M. Styner, N. Case, K. Burridge, and J. Rubin, "Mechanically activated fyn utilizes mtorc2 to regulate rhoa and adipogenesis in mesenchymal stem cells," *Stem Cells*, vol. 31, no. 11, pp. 2528–37, 2013.

[341] M. Unbekandt and M. F. Olson, "The actin-myosin regulatory MRCk kinases: regulation, biological functions and associations with human cancer," *Journal of Molecular Medicine (Berlin, Germany)*, vol. 92, no. 3, pp. 217–225, 2014.

[342] Z.-s. Zhao and E. Manser, "Pak and other rho-associated kinases â€“ effectors with surprisingly diverse mechanisms of regulation," *Biochemical Journal*, vol. 386, no. Pt 2, pp. 201–214, 2005.

[343] S. Etienne-Manneville and A. Hall, "Rho GTPases in cell biology," *Nature*, vol. 420, no. 6916, pp. 629–635, 2002.

[344] C. Nobes and M. Marsh, "Dendritic cells: New roles for cdc42 and rac in antigen uptake?," *Current Biology*, vol. 10, no. 20, pp. R739–R741, 2000.

[345] C. Picard, W. Al-Herz, A. Bousfiha, J.-L. Casanova, T. Chatila, M. Conley, C. Cunningham-Rundles, A. Etzioni, S. Holland, C. Klein, S. Nonoyama, H. Ochs, E. Oksenhendler, J. Puck, K. Sullivan, M. Tang, J. Franco, and H. B. Gaspar, "Primary immunodeficiency diseases: an update on the classification from the international union of immunological societies expert committee for primary immunodeficiency 2015," *Journal of Clinical Immunology*, pp. 1–31, 2015.

[346] R. J. Cain, R. D. Hayward, and V. Koronakis, "Deciphering Interplay between Salmonella Invasion Effectors," *PLoS Pathogens*, vol. 4, no. 4, p. e1000037, 2008.

[347] J. E. Galan and D. Zhou, "Striking a balance: modulation of the actin cytoskeleton by salmonella," *Proc Natl Acad Sci U S A*, vol. 97, no. 16, pp. 8754–61, 2000.

[348] M. Georg, L. Maudsdotter, R. Tavares, and A. B. Jonsson, "Meningococcal resistance to antimicrobial peptides is mediated by bacterial adhesion and host cell rhoa and cdc42 signalling," *Cell Microbiol*, vol. 15, no. 11, pp. 1938–54, 2013.

[349] A. M. Keestra, M. G. Winter, J. J. Auburger, S. P. Fraszle, M. N. Xavier, S. E. Winter, A. Kim, V. Poon, M. M. Ravesloot, J. F. T. Waldenmaier, R. M. Tsolis, R. A. Eigenheer, and A. J. Baumler, "Manipulation of small rho gtpases is a pathogen-induced process detected by nod1," *Nature*, vol. 496, no. 7444, pp. 233–237, 2013.

[350] M. Essler, J. M. Staddon, P. C. Weber, and M. Aepfelbacher, "Cyclic AMp blocks bacterial lipopolysaccharide-induced myosin light chain phosphorylation in endothelial cells through inhibition of rho/rho kinase signaling," *J Immunol*, vol. 164, no. 12, pp. 6543–9, 2000.

[351] F. Guo, J. Tang, Z. Zhou, Y. Dou, D. Van Lonkhuyzen, C. Gao, and J. Huan, "Gef-h1-rhoa signaling pathway mediates lps-induced NF-kappab transactivation and il-8 synthesis in endothelial cells," *Mol Immunol*, vol. 50, no. 1-2, pp. 98–107, 2012.

[352] M. Ruse and U. G. Knaus, "New players in tlr-mediated innate immunity: Pi3K and small rho gtpases," *Immunol Res*, vol. 34, no. 1, pp. 33–48, 2006.

[353] N. Sobreira, F. Schiettecatte, D. Valle, and A. Hamosh, "Genematcher: a matching tool for connecting investigators with an interest in the same gene.," *Human mutation*, vol. 36, pp. 928–30, Oct 2015.

[354] L. E. Salnikova, T. V. Smelaya, A. M. Golubev, A. V. Rubanovich, and V. V. Moroz, "Cyp1A1, gclc, agt, agtr1 gene-gene interactions in community-acquired pneumonia pulmonary complications," *Mol Biol Rep*, vol. 40, no. 11, pp. 6163–76, 2013.

[355] D. Wang, Y. Sun, P. Stang, J. A. Berlin, M. A. Wilcox, and Q. Li, "Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling," *BMC Proc*, vol. 3 Suppl 7, p. S109, 2009.

[356] L. Liu, A. Sabo, B. M. Neale, U. Nagaswamy, C. Stevens, E. Lim, C. A. Bodea, D. Muzny, J. G. Reid, E. Banks, H. Coon, M. Depristo, H. Dinh, T. Fennel, J. Flannick, S. Gabriel, K. Garimella, S. Gross, A. Hawes, L. Lewis, V. Makarov, J. Maguire, I. Newsham, R. Poplin, S. Ripke, K. Shakir, K. E. Samocha, Y. Wu, E. Boerwinkle, J. D. Buxbaum, J. Cook, E. H., B. Devlin, G. D. Schellenberg, J. S. Sutcliffe, M. J. Daly, R. A. Gibbs, and K. Roeder, "Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls," *PLoS Genet*, vol. 9, no. 4, p. e1003443, 2013.

[357] S. Novelli, C. Canals, N. Nogues, M. R. Julia, M. Gracia, I. Vinyets, and E. Muniz-Diaz, "Severe neonatal alloimmune thrombocytopaenia with anaemia," *Transfus Med*, vol. 20, no. 2, pp. 125–6, 2010.

[358] A. A. Bizet, K. Liu, N. Tran-Khanh, A. Saksena, J. Vorstenbosch, K. W. Finnson, M. D. Buschmann, and A. Philip, "The TGF-beta co-receptor, CD109, promotes internalization and degradation of TGF-beta receptors," *Biochim Biophys Acta*, vol. 1813, no. 5, pp. 742–53, 2011.

[359] M. Lin, D. R. Sutherland, W. Horsfall, N. Totty, E. Yeo, R. Nayar, X. F. Wu, and A. C. Schuh, "Cell surface antigen CD109 is a novel member of the alpha(2) macroglobulin/c3, c4, c5 family of thioester-containing proteins," *Blood*, vol. 99, no. 5, pp. 1683–91, 2002.

[360] N. Ferrara, "Molecular and biological properties of vascular endothelial growth factor," *J Mol Med (Berl)*, vol. 77, no. 7, pp. 527–43, 1999.

[361] Y. C. Lee and K. B. Lane, "The many faces of transforming growth factor-beta in pleural diseases," *Curr Opin Pulm Med*, vol. 7, no. 4, pp. 173–9, 2001.

[362] S. A. Sasse, M. R. Jadus, and G. D. Kukes, "Pleural fluid transforming growth factor-beta1 correlates with pleural fibrosis in experimental empyema," *Am J Respir Crit Care Med*, vol. 168, no. 6, pp. 700–5, 2003.

[363] P. Bui, S. Imaizumi, S. R. Beedanagari, S. T. Reddy, and O. Hankinson, "Human CYP2S1 metabolizes cyclooxygenase- and lipoxygenase-derived eicosanoids," *Drug Metab Dispos*, vol. 39, no. 2, pp. 180–90, 2011.

[364] L. E. Salnikova, T. V. Smelaya, V. V. Moroz, A. M. Golubev, and A. V. Rubanovich, "Functional polymorphisms in the cyp1A1, ACe, and il-6 genes contribute to susceptibility to community-acquired and nosocomial pneumonia," *Int J Infect Dis*, vol. 17, no. 6, pp. e433–42, 2013.

[365] M. Ikegami, A. Falcone, and J. A. Whitsett, "Stat-3 regulates surfactant phospholipid homeostasis in normal lung and during endotoxin-mediated lung injury," *J Appl Physiol (1985)*, vol. 104, no. 6, pp. 1753–60, 2008.

[366] S. M. Holland, F. R. DeLeo, H. Z. Elloumi, A. P. Hsu, G. Uzel, N. Brodsky, A. F. Freeman, A. Demidowich, J. Davis, M. L. Turner, V. L. Anderson, D. N. Darnell, P. A. Welch, D. B. Kuhns, D. M. Frucht, H. L. Malech, J. I. Gallin, S. D. Kobayashi, A. R. Whitney, J. M. Voyich, J. M. Musser, C. Woellner, A. A. Schaffer, J. M. Puck, and B. Grimbacher, "Stat3 mutations in the hyper-ige syndrome," *N Engl J Med*, vol. 357, no. 16, pp. 1608–19, 2007.

[367] M. Severgnini, S. Takahashi, L. M. Rozo, R. J. Homer, C. Kuhn, J. W. Jhung, G. Perides, M. Steer, P. M. Hassoun, B. L. Fanburg, B. H. Cochran, and A. R. Simon, "Activation of the stat pathway in acute lung injury," *Am J Physiol Lung Cell Mol Physiol*, vol. 286, no. 6, pp. L1282–92, 2004.

[368] G. A. Bezerra, E. Dobrovetsky, A. Seitova, S. Fedosyuk, S. Dhe-Paganon, and K. Gruber, "Structure of human dipeptidyl peptidase 10 (dppy): a modulator of neuronal kv4 channels," *Sci Rep*, vol. 5, p. 8769, 2015.

[369] S. H. Kim, H. Choi, M. G. Yoon, Y. M. Ye, and H. S. Park, "Dipeptidyl-peptidase 10 as a genetic biomarker for the aspirin-exacerbated respiratory disease phenotype," *Ann Allergy Asthma Immunol*, vol. 114, no. 3, pp. 208–13, 2015.

[370] A. A. Bizet, N. Tran-Khanh, A. Saksena, K. Liu, M. D. Buschmann, and A. Philip, "CD109-mediated degradation of TGF-beta receptors and inhibition of TGF-beta responses involve regulation of SMAD7 and Smurf2 localization and function," *J Cell Biochem*, vol. 113, no. 1, pp. 238–46, 2012.

[371] F. Delolme, C. Anastasi, L. B. Alcaraz, V. Mendoza, S. Vadon-Le Goff, M. Talantikite, R. Capomaccio, J. Mevaere, L. Fortin, D. Mazzocut, O. Damour, I. Zanella-Cleon, D. J. Hulmes, C. M. Overall, U. Valcourt, F. Lopez-Casillas, and C. Moali, "Proteolytic control of TGF-beta co-receptor activity by BMP-1/tolloid-like proteases revealed by quantitative iTRAQ proteomics," *Cell Mol Life Sci*, vol. 72, no. 5, pp. 1009–27, 2015.

[372] K. W. Finnson, B. Y. Tam, K. Liu, A. Marcoux, P. Lepage, S. Roy, A. A. Bizet, and A. Philip, "Identification of CD109 as part of the tgf-beta receptor system in human keratinocytes," *FASEB J*, vol. 20, no. 9, pp. 1525–7, 2006.

[373] S. Hagiwara, Y. Murakumo, S. Mii, T. Shigetomi, N. Yamamoto, H. Furue, M. Ueda, and M. Takahashi, "Processing of CD109 by furin and its role in the regulation of tgf-beta signaling," *Oncogene*, vol. 29, no. 15, pp. 2181–91, 2010.

[374] M. Hashimoto, M. Ichihara, T. Watanabe, K. Kawai, K. Koshikawa, N. Yuasa, T. Takahashi, Y. Yatabe, Y. Murakumo, J. M. Zhang, Y. Nimura, and M. Takahashi, "Expression of CD109 in human cancer," *Oncogene*, vol. 23, no. 20, pp. 3716–20, 2004.

[375] R. S. Haun, C. Y. Fan, S. G. Mackintosh, H. Zhao, and A. J. Tackett, "CD109 overexpression in pancreatic cancer identified by cell-surface glycoprotein capture," *J Proteomics Bioinform*, vol. Suppl 10, p. S10003, 2014.

[376] X. Y. Man, K. W. Finnson, M. Baron, and A. Philip, "CD109, a tgf-beta co-receptor, attenuates extracellular matrix production in scleroderma skin fibroblasts," *Arthritis Res Ther*, vol. 14, no. 3, p. R144, 2012.

[377] P. Mancuso, A. Calleri, G. Gregato, V. Labanca, J. Quarna, P. Antoniotti, L. Cuppini, G. Finocchiaro, M. Eoli, V. Rosti, and F. Bertolini, "A subpopulation of circulating endothelial cells express CD109 and is enriched in the blood of cancer patients," *PLoS One*, vol. 9, no. 12, p. e114713, 2014.

[378] J. Mokrosinski and W. M. Krajewska, "[tgf beta signalling accessory receptors]," *Postepy Biochem*, vol. 54, no. 3, pp. 264–73, 2008.

[379] J. Vorstenbosch, C. Gallant-Behm, A. Trzeciak, S. Roy, T. Mustoe, and A. Philip, "Transgenic mice overexpressing CD109 in the epidermis display decreased inflammation and granulation tissue and improved collagen architecture during wound healing," *Wound Repair Regen*, vol. 21, no. 2, pp. 235–46, 2013.

[380] S. Winocour, J. Vorstenbosch, A. Trzeciak, L. Lessard, and A. Philip, "CD109, a novel tgf-beta antagonist, decreases fibrotic responses in a hypoxic wound model," *Exp Dermatol*, vol. 23, no. 7, pp. 475–9, 2014.

[381] J. M. Zhang, Y. Murakumo, S. Hagiwara, P. Jiang, S. Mii, E. Kalyoncu, S. Saito, C. Suzuki, Y. Sakurai, Y. Numata, T. Yamamoto, and M. Takahashi, "CD109 attenuates tgf-beta1 signaling and enhances egf signaling in sk-mg-1 human glioblastoma cells," *Biochem Biophys Res Commun*, vol. 459, no. 2, pp. 252–8, 2015.

[382] P. Francois, A. Desrumaux, C. Cans, I. Pin, P. Pavese, and J. Labarere, "Prevalence and risk factors of suppurative complications in children with pneumonia," *Acta Paediatr*, vol. 99, no. 6, pp. 861–6, 2010.

[383] M. Le Bourgeois, A. Ferroni, M. Leruez-Ville, E. Varon, C. Thumerelle, F. Bremont, M. J. Fayon, C. Delacourt, C. Ligier, L. Watier, and D. Guillemot, "Nonsteroidal anti-inflammatory drug without antibiotics for acute viral infection increases the empyema risk in children: A matched case-control study," *J Pediatr*, 2016.

[384] S. J. Chapman, C. C. Khor, F. O. Vannberg, A. Rautanen, S. Segal, C. E. Moore, R. J. Davies, N. P. Day, N. Peshu, D. W. Crook, J. A. Berkley, T. N. Williams, J. A. Scott, and A. V. Hill, "NFKBIZ polymorphisms and susceptibility to pneumococcal disease in European and African populations," *Genes Immun*, vol. 11, no. 4, pp. 319–25, 2010.

[385] M. Hubeau, F. Ngadjeua, A. Puel, L. Israel, J. Feinberg, M. Chrabieh, K. Belani, C. Bodemer, I. Fabre, A. Plebani, S. Boisson-Dupuis, C. Picard, A. Fischer, A. Israel, L. Abel, M. Veron, J. L. Casanova, F. Agou, and J. Bustamante, "New mechanism of x-linked anhidrotic ectodermal dysplasia with immunodeficiency: impairment of ubiquitin binding despite normal folding of nemo protein," *Blood*, vol. 118, no. 4, pp. 926–35, 2011.

[386] M. K. Ellis, K. S. Elliott, A. Rautanen, D. W. Crook, A. V. Hill, and S. J. Chapman, "Rare variants in MYD88, IRAK4 and IKBKG and susceptibility to invasive pneumococcal disease: a population-based case-control study," *PLoS One*, vol. 10, no. 4, p. e0123532, 2015.

[387] J. Harder and J. M. Schroder, "Rnase 7, a novel innate immune defense antimicrobial protein of healthy human skin," *J Biol Chem*, vol. 277, no. 48, pp. 46779–84, 2002.

[388] M. Laudien, S. Dressel, J. Harder, and R. Glaser, "Differential expression pattern of antimicrobial peptides in nasal mucosa and secretion," *Rhinology*, vol. 49, no. 1, pp. 107–11, 2011.

[389] J. Lonsdale, J. Thomas, M. Salvatore, R. Phillips, E. Lo, S. Shad, R. Hasz, G. Walters, F. Garcia, N. Young, B. Foster, M. Moser, E. Karasik, B. Gillard, K. Ramsey, S. Sullivan, J. Bridge, H. Magazine, J. Syron, J. Fleming, L. Siminoff, H. Traino, M. Mosavel, L. Barker, S. Jewell, D. Rohrer, D. Maxim, D. Filkins, P. Harbach, E. Cortadillo, B. Berghuis, L. Turner, E. Hudson, K. Feenstra, L. Sobin, J. Robb, P. Branton, G. Korzeniewski, C. Shive, D. Tabor, L. Qi, K. Groch, S. Nampally, S. Buia, A. Zimmerman, A. Smith, R. Burges, K. Robinson, K. Valentino, D. Bradbury, M. Cosentino, N. Diaz-Mayoral, M. Kennedy, T. Engel, P. Williams, K. Erickson, K. Ardlie, W. Winckler, G. Getz, D. DeLuca, D. MacArthur, M. Kellis, A. Thomson, T. Young, E. Gelfand, M. Donovan, Y. Meng, G. Grant, D. Mash, Y. Marcus, M. Basile, J. Liu, J. Zhu, Z. Tu, N. J. Cox, D. L. Nicolae, E. R. Gamazon, H. K. Im, A. Konkashbaev, J. Pritchard, M. Stevens, T. Flutre, X. Wen, E. T. Dermitzakis, T. Lappalainen, R. Guigo, J. Monlong, M. Sammeth, D. Koller, A. Battle, S. Mostafavi, M. McCarthy, M. Rivas, J. Maller, I. Rusyn, A. Nobel, F. Wright, A. Shabalin, M. Feolo, N. Sharopova, *et al.*, "The genotype-tissue expression (gtex) project," *Nat Genet*, vol. 45, no. 6, pp. 580–585, 2013.

[390] M. Simanski, F. Rademacher, L. Schroder, H. M. Schumacher, R. Glaser, and J. Harder, "Il-17A and IFn-gamma synergistically induce rnase 7 expression via stat3 in primary keratinocytes," *PLoS One*, vol. 8, no. 3, p. e59531, 2013.

[391] M. Torrent, M. Badia, M. Moussaoui, D. Sanchez, M. V. Nogues, and E. Boix, "Comparison of human rnase 3 and rnase 7 bactericidal action at the gram-negative and gram-positive bacterial cell wall," *FEBS J*, vol. 277, no. 7, pp. 1713–25, 2010.

[392] J. S. Cho, C. Xuan, and L. S. Miller, "Lucky number seven: RNase 7 can prevent Staphylococcus aureus skin colonization," *J Invest Dermatol*, vol. 130, no. 12, pp. 2703–6, 2010.

[393] M. Simanski, S. Dressel, R. Glaser, and J. Harder, "Rnase 7 protects healthy skin from staphylococcus aureus colonization," *J Invest Dermatol*, vol. 130, no. 12, pp. 2836–8, 2010.

[394] P. Zanger, J. Holzer, R. Schleucher, H. Steffen, B. Schittek, and S. Gabrysch, "Constitutive expression of the antimicrobial peptide rnase 7 is associated with staphylococcus aureus infection of the skin," *J Infect Dis*, vol. 200, no. 12, pp. 1907–15, 2009.

[395] L. V. Hooper, T. S. Stappenbeck, C. V. Hong, and J. I. Gordon, "Angiogenins: a new class of microbicidal proteins involved in innate immunity," *Nat Immunol*, vol. 4, no. 3, pp. 269–73, 2003.

[396] G. D. Amatngalim, Y. van Wijck, Y. de Mooij-Eijk, R. M. Verhoosel, J. Harder, A. N. Lekkerkerker, R. A. Janssen, and P. S. Hiemstra, "Basal cells contribute to innate immunity of the airway epithelium through production of the antimicrobial protein RNase 7," *J Immunol*, vol. 194, no. 7, pp. 3340–50, 2015.

[397] Y. C. Huang, Y. M. Lin, T. W. Chang, S. J. Wu, Y. S. Lee, M. D. Chang, C. Chen, S. H. Wu, and Y. D. Liao, "The flexible and clustered lysine residues of human ribonuclease 7 are critical for membrane permeability and antimicrobial activity," *J Biol Chem*, vol. 282, no. 7, pp. 4626–33, 2007.

[398] F. Rademacher, M. Simanski, and J. Harder, "Rnase 7 in cutaneous defense," *Int J Mol Sci*, vol. 17, no. 4, 2016.

[399] S. Iden and J. G. Collard, "Crosstalk between small gtpases and polarity proteins in cell polarization," *Nat Rev Mol Cell Biol*, vol. 9, no. 11, pp. 846–59, 2008.

[400] S. Ganesan, A. T. Comstock, and U. S. Sajjan, "Barrier function of airway tract epithelium," *Tissue Barriers*, vol. 1, no. 4, p. e24997, 2013.

[401] T. W. Hurd, L. Gao, M. H. Roh, I. G. Macara, and B. Margolis, "Direct interaction of two polarity complexes implicated in epithelial tight junction assembly," *Nat Cell Biol*, vol. 5, no. 2, pp. 137–42, 2003.

[402] K. Shin, V. C. Fogg, and B. Margolis, "Tight junctions and cell polarity," *Annual Review of Cell and Developmental Biology*, vol. 22, no. 1, pp. 207–235, 2006.

[403] S. W. Straight, K. Shin, V. C. Fogg, S. Fan, C. J. Liu, M. Roh, and B. Margolis, "Loss of PALs1 expression leads to tight junction and polarity defects," *Mol Biol Cell*, vol. 15, no. 4, pp. 1981–90, 2004.

[404] K.-T. Teoh, Y.-L. Siu, W.-L. Chan, M. A. SchlÃ¹/4ter, C.-J. Liu, J. S. M. Peiris, R. Bruzzone, B. Margolis, and B. Nal, "The SARs coronavirus e protein interacts with PALs1 and alters tight junction formation and epithelial morphogenesis," *Molecular Biology of the Cell*, vol. 21, no. 22, pp. 3838–3852, 2010.

[405] C. Beisswenger, C. B. Coyne, M. Shchepetov, and J. N. Weiser, "Role of p38 MAp kinase and transforming growth factor-beta signaling in transepithelial migration of invasive bacterial pathogens," *J Biol Chem*, vol. 282, no. 39, pp. 28700–8, 2007.

[406] T. B. Clarke, N. Francella, A. Huegel, and J. N. Weiser, "Invasive bacterial pathogens exploit TLR-mediated downregulation of tight junction components to facilitate translocation across the epithelium," *Cell Host Microbe*, vol. 9, no. 5, pp. 404–14, 2011.

[407] R. Bellamy, "The natural resistance-associated macrophage protein and susceptibility to intracellular pathogens," *Microbes Infect*, vol. 1, no. 1, pp. 23–7, 1999.

[408] S. Ram, L. A. Lewis, and P. A. Rice, "Infections of people with complement deficiencies and patients who have undergone splenectomy," *Clinical Microbiology Reviews*, vol. 23, no. 4, pp. 740–780, 2010.

[409] C. Picard, S. Dogniaux, K. Chemin, Z. Maciorowski, A. Lim, F. Mazerolles, F. Rieux-Laucat, M.-C. Stolzenberg, M. Debre, J.-P. Magny, F. Le Deist, A. Fischer, and C. Hivroz, "Hypomorphic mutation of zap70 in human results in a late onset immunodeficiency and no autoimmunity," *European Journal of Immunology*, vol. 39, no. 7, pp. 1966–1976, 2009.

[410] T. Turul, I. Tezcan, H. Artac, S. de Bruin-Versteeg, B. H. Barendregt, I. Reisli, O. Sanal, J. J. van Dongen, and M. van der Burg, "Clinical heterogeneity can hamper the diagnosis of patients with zap70 deficiency," *Eur J Pediatr*, vol. 168, no. 1, pp. 87–93, 2009.

[411] L. Miller, L. Arakaki, A. Ramautar, S. Bodach, S. L. Braunstein, J. Kennedy, L. Steiner-Sichel, S. Ngai, C. Shepard, and D. Weiss, "Elevated risk for invasive meningococcal disease among persons with hiv," *Ann Intern Med*, vol. 160, no. 1, pp. 30–7, 2014.

[412] R. A. Foster, J. Carlring, A. Lees, R. Borrow, M. Ramsay, E. Kacsmarski, E. Miller, M. W. McKendrick, A. W. Heath, and R. C. Read, "Functional t-cell deficiency in adolescents who experience serogroup c meningococcal disease despite receiving the meningococcal serogroup c conjugate vaccine," *Clinical and Vaccine Immunology : CVI*, vol. 17, no. 7, pp. 1104–1110, 2010.

[413] R. E. Vance, R. R. Isberg, and D. A. Portnoy, "Patterns of pathogenesis: discrimination of pathogenic and non-pathogenic microbes by the innate immune system," *Cell host & microbe*, vol. 6, no. 1, pp. 10–21, 2009.

[414] C. R. Kunz, M. R. Jadus, G. D. Kukes, F. Kramer, V. N. Nguyen, and S. A. Sasse, "Intrapleural injection of transforming growth factor-beta antibody inhibits pleural fibrosis in empyema," *Chest*, vol. 126, no. 5, pp. 1636–44, 2004.

[415] D. B. Goldstein, A. Allen, J. Keebler, E. H. Margulies, S. Petrou, S. Petrovski, and S. Sunyaev, "Sequencing studies in human genetics: design and interpretation," *Nat Rev Genet*, vol. 14, no. 7, pp. 460–470, 2013.

[416] M. Antonio, I. Hakeem, T. Awine, O. Secka, K. Sankareh, D. Nsekpong, G. Lahai, A. Akisanya, U. Egere, G. Enwere, S. M. Zaman, P. C. Hill, T. Corrah, F. Cutts, B. M. Greenwood, and R. A. Adegbola, "Seasonality and outbreak of a predominant Streptococcus pneumoniae serotype 1 clone from the Gambia: expansion of ST217 hypervirulent clonal complex in west Africa," *BMC Microbiol*, vol. 8, p. 198, 2008.