Human Genetic Susceptibility to Common Infectious Diseases in Europe

Tara Carolyn Mills
St. Catherine's College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Michaelmas Term 2012

Human Genetic Susceptibility to Common Infectious Diseases in Europe

Tara Carolyn Mills
St. Catherine's College, University of Oxford, D.Phil, Michaelmas 2012

Abstract

Lower respiratory tract infections (LRTIs) are one of the most common infectious diseases in Europe and worldwide. Very little is known about the genetic factors associated with susceptibility to LRTI and so far studies have only analysed candidate gene loci.

This work aimed to find genetic loci associated with susceptibility to severe and mild LRTI. To analyse severe LRTI, a European cohort of community-acquired pneumonia (CAP) sepsis patients was used. To analyse mild LRTI, patients attending primary care with a cough were recruited across Europe.

Two important candidate genes for susceptibility to severe and mild LRTI were studied. None of the four functional polymorphisms analysed in *MBL2* were associated individually or when combined with susceptibility to LRTI or survival from severe LRTI. Rare homozygotes for rs12252 in the *IFITM3* gene were found to associate with susceptibility to mild and severe viral LRTI, particularly influenza infection.

Using a combination of genome-wide and candidate gene methods, 93 single nucleotide polymorphisms (SNPs) were chosen to genotype and analyse for susceptibility to mild LRTI. A gene region containing the *MPHOSPH8*, *TPTE2*, *PSPC1* and *ZMYM5* genes was found to associate with susceptibility to rhinovirus infection. A functional promoter SNP in the *IL6* gene and a gene region containing the *SAP18*, *MRP63* and *SKA3* genes were associated with susceptibility to mild LRTI in smokers.

A GWAS was performed for susceptibility to CAP sepsis in Europe. Genotypes were imputed for eight GWAS data sets comprising over 1000 CAP sepsis cases and over 4000 controls. Preliminary analysis results suggest an association in a chromosome 1 region containing the genes *VASH2*, *ANGEL2* and *RPS6KC1*.

This thesis presents results of the largest known genetic association study for susceptibility to CAP sepsis, and the only known genetic association study for mild LRTI. These novel findings will facilitate further research which may lead to better treatment and outcome for LRTI.

Acknowledgements

Firstly I would like to say an enormous thanks to my supervisors Prof. Adrian Hill and Dr. Anna Rautanen, for their constant support and encouragement. I couldn't have done this without them, and I think I have been lucky to have the best combination of supervisors I could hope for.

I would also like to thank all members of the Hill genetics group, especially Sunny Wong for his enthusiasm and brilliance in genetics and life in general; Magda Ellis for her support, insight, and a trip of a lifetime; Fredrik Vannberg for persuading me to start this D.Phil. in the first place, and for always seeing things in an alternative way; Stephen Chapman for his amazing knowledge bank and for always putting things into perspective and Kate Elliott for all the help with script writing, and for a fantastic house in London.

I could never have managed this without the companionship of the girls in the centre I have come to call my friends: Rachel Craik for her endless DNA knowledge, incredible lab skills and for the most boring and yet hilarious weekends spent in the lab; Gemma Buck for all the film nights and her incredible photography; Katharine Plant for her endless support and kindness, and for the best reaction ever and Emma Davenport for help in every aspect of everything. And not forgetting Hayley Davis for incomparable guidance with the formatting of this thesis.

These acknowledgements would not be complete without recognising the undeniable influence of my parents in helping me to reach this point in my life. For always helping me believe in myself, for guiding and teaching me in every aspect of my life,

for teaching me how to learn and teach, and for far too many things to list here. And of course I would never have come this far without Jonny (the best brother in the world) to compete against!

Finally I want to thank my Oliver. These have been some of the hardest years of my life, but we always had a place in the world where we could be together, and I will never stop thanking you for making that place here. I don't even love genetics as much as I love you.

Attributions

This thesis is a collection of research work carried out by the author in the Hill group laboratory, Wellcome Trust Centre for Human Genetics, Department of Clinical Medicine, University of Oxford, between October 2009 and December 2012, under the supervision of Prof. Adrian Hill and Dr. Anna Rautanen.

All research presented in this thesis is original, although any parts of this work which were not conducted by the author are stated in the text. No part of this thesis has been accepted or is currently being submitted for any degree, diploma or certificate or other qualification at the University of Oxford or elsewhere.

This research would not have been possible without the work of those people involved in the collection of the sample cohorts. I would like to thank all those individuals involved in the GenOSept, GAinS and GRACE projects for their hard work setting up successful sample collections, and for the collection of samples for the extraction of high quality DNA.

The core genomics group at the Wellcome Trust Centre for Human Genetics have been involved in the genotyping of samples using the iPLEX technology from Sequenom. I would like to thank Lyn-Louise Johnson, Joseph Trakalo, Helen Butler and Christine Blancher for this service and for their excellent work.

I would also like to thank Angel Carracedo and Inés Quintela for their help with the Spanish control GWAS data. I am also very grateful to Sara Lupoli and Daniele Cusi for their help with the Italian control GWAS data.

Associated publications

Some of the work described in this thesis is under preparation in the following manuscripts:

Mills, T.C., Garrard, C.S., Hinds, C., Chapman, S., Hill, A.V., Rautanen, A. (2012) Mannose binding lectin genotypes are not associated with sepsis survival or susceptibility. *Manuscript in preparation.*

Mills, T.C., Rautanen, A., Elliott, K.S., Parks, T., Naranbhai, V., Ieven, M.I., Butler C.C., Little, P., Verheij, T., Garrard, C.S., Hinds, C., Goossens, H., Chapman, S., Hill, A.V. (2012) *IFITM3* and susceptibility to respiratory viral infections. *Manuscript in preparation.*

Mills, T.C., Ieven, M.I., Lammens, C., Butler C.C., Little, P., Verheij T., Henriques Normack, B., Claas, E., Crook, D., Goossens, H., Chapman, S., Hill, A.V., Rautanen, A. (2012) Genetic susceptibility to mild lower respiratory tract infections. *Manuscript in preparation.*

Additionally, I have contributed to the following publications that are not covered in this thesis:

George, J., Motshwene, P.G., Wang, H., Kubarenko, A.V., Rautanen, A., **Mills, T.C.**, Hill, A.V., Gay, N.J., Weber, A.N. (2011) Two human *MYD88* variants, S34Y and R98C, interfere with MyD88-IRAK4-myddosome assembly. *J Biol Chem.* **286**, 1341-53.

Kubarenko, A.V., Ranjan, S., Rautanen, A., **Mills, T.C.**, Wong, S., Vannberg, F., Neumaier, M., Bekeredjian-Ding, I., Hill, A.V., Ahmad-Nejad, P., Weber, A.N. (2010) A naturally occurring variant in human *TLR9*, P99L, is associated with loss of CpG oligonucleotide responsiveness. *J Biol Chem.* **285**, 36486-94.

Wong, S.H., Gochhait, S., Malhotra, D., Pettersson, F.H., Teo, Y.Y., Khor, C.C., Rautanen, A., Chapman, S.J., **Mills, T.C.**, Srivastava, A., Rudko, A., Freidin, M.B., Puzyrev, V.P., Ali, S., Aggarwal, S., Chopra, R., Reddy, B.S., Garg, V.K., Roy, S., Meisner, S., Hazra, S.K., Saha, B., Floyd, S., Keating, B.J., Kim, C., Fairfax, B.P., Knight, J.C., Hill, P.C., Adegbola, R.A., Hakonarson, H., Fine, P.E., Pitchappan, R.M., Bamezai, R.N., Hill, A.V., Vannberg, F.O. (2010) Leprosy and the adaptation of human toll-like receptor 1. *PLoS Pathog.* **6**:e1000979.

Khor, C.C., Vannberg, F.O., Chapman, S.J., Guo, H., Wong, S.H., Walley, A.J., Vukcevic, D., Rautanen, A., **Mills, T.C.**, Chang, K.C., Kam, K.M., Crampin, A.C., Ngwira, B., Leung, C.C., Tam, C.M., Chan, C.Y., Sung, J.J., Yew, W.W., Toh, K.Y., Tay, S.K., Kwiatkowski, D., Lienhardt, C., Hien, T.T., Day, N.P., Peshu, N., Marsh, K., Maitland, K., Scott, J.A., Williams, T.N., Berkley, J.A., Floyd, S., Tang, N.L., Fine, P.E., Goh, D.L., Hill, A.V. (2010) *CISH* and susceptibility to infectious diseases. *N Engl J Med.* **362**, 2092-101.

George, J., Kubarenko, A.V., Rautanen, A., **Mills, T.C.**, Colak, E., Kempf, T., Hill, A.V., Nieters, A., Weber, A.N. (2010) MyD88 adaptor-like D96N is a naturally occurring loss-of-function variant of *TIRAP*. *J Immunol.* **184**, 3025-32.

Contents

List of Figures

List of Tables

List of Abbreviations

aeCOPD	Acute exacerbations of chronic obstructive pulmonary disease
AIDS	Acquired immune deficiency syndrome
ARDS	Acute respiratory distress syndrome
CAP	Community-acquired pneumonia
CD-CV	Common disease-common variant
CEPH	Centre d'Etude du Polymorphisme Humain
Chr	Chromosome
CI	Confidence interval
COPD	Chronic obstructive pulmonary disease
DALYs	Disability-adjusted life years
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
FISH	Fluorescence in situ hybridisation
FP	Faecal peritonitis
GAinS	Genomic Advances in Sepsis
GenOSept	Genetics Of Sepsis and Septic Shock
GP	General practitioner
GRACE	Genomics to combat Resistance against Antibiotics in Community-acquired lower respiratory tract infections (LRTI) in Europe
GWAS	Genome-wide association study/studies

HA	Hemagglutinin
HIV	Human immunodeficiency virus
HRMA	High resonance melting analysis
HRV	Human rhinovirus
HuGE	Human genome epidemiology
HWE	Hardy-Weinberg equilibrium
IBD	Identity by descent
IBS	Identity by state
ICU	Intensive care unit
IFN	Interferon
IL	Interleukin
INDEL	Insertion/deletion
IPD	Invasive pneumococcal disease
KEMRI	Kenya Medical Research Institute
LD	Linkage disequilibrium
LEV	Low elution volume
LRTI	Lower respiratory tract infection
MAF	Minor allele frequency
MBL	Mannose-binding lectin
MDS	Multi-dimensional scaling
MHC	Major histocompatibility complex
MS	Multiple sclerosis
NA	Neuraminidase
NK	Natural killer
OR	Odds ratio
PAH	Phenylalanine hydroxylase
PCA	Principal components analysis

PCR	Polymerase chain reaction
PKU	Phenylketonuria
QC	Quality control
QQ	Quantile-quantile
RFLP	Restriction fragment length polymorphism
RNA	Ribonucleic acid
RSV	Respiratory syncytial virus
SEV	Standard elution volume
SHAPEIT	Segmented HAPlotype Estimation and Imputation Tool
SIRS	Systemic inflammatory response syndrome
SNP	Single nucleotide polymorphism
TLR	Toll-like receptor
UK	United Kingdom
USA	United States of America
WGA	Whole genome amplified
WHO	World Health Organization
WTCCC	Wellcome Trust case control consortium

Chapter 1

Introduction

1.1 Common infectious diseases in Europe

The two infectious diseases with the highest incidence worldwide are diarrhoeal diseases and lower respiratory infections. These are also responsible for the greatest incidence in every region worldwide, except for Africa where malaria is more common than lower respiratory infections (World Health Organisation, 2008). Lower respiratory infections are also associated with the greatest number of deaths worldwide of any infectious disease. They also cause more deaths than other infectious diseases in all regions except for Africa and South-East Asia where HIV/AIDS and diarrhoeal disease cause slightly more deaths respectively (World Health Organisation, 2008). The distribution of other infectious diseases varies considerably worldwide. For example, the incidence of measles is very high in South-East Asia but effectively non-existent in the Americas, whereas Chagas disease is present only in the Americas and absent everywhere else in the world (World Health Organisation, 2008).

"Lung infection", including influenza, pneumonia and other acute lower respiratory infections accounts for a greater global burden of disease than HIV/AIDS, malaria, ischemic heart disease and tuberculosis. This category of disease does not

include HIV disease resulting in infectious disease, so it could be argued that the global burden of lung infection is even higher than this estimate (Mizgerd, 2006).

Studying all infectious disease in Europe would be too much for the breadth of this thesis. Therefore, in this thesis I have studied lower respiratory tract infections (LRTIs), which account for 19 million cases of infectious disease in Europe per year (World Health Organisation, 2008). Lower respiratory infections account for by far the greatest number of deaths of any infectious disease in Europe, accounting for over 50% of all infectious disease deaths in 2008 estimates (World Health Organisation, 2008). Lower respiratory infections are also the leading cause of death in low-income countries, and still account for 0.35 million deaths (the 5th leading cause of death) in high income countries, despite improved health care (World Health Organisation, 2011).

This thesis analyses the genetics of susceptibility to lower respiratory tract infection (LRTI) using two disease groups. One is a cohort of severe community-acquired pneumonia (CAP) cases with sepsis. The other is a cohort of mild LRTI cases collected in primary care. In this way, I am studying both ends of the spectrum of diseases which are classified as LRTI, with very mild and very severe disease.

1.1.1 Sepsis

Sepsis is defined as the systemic response to infection and is often confused with overlapping terms such as infection, bacteraemia, septicaemia and septic shock. An American College of Chest Physicians/Society of Critical Care Medicine Consensus Conference was held in 1991, and their agreed definitions for sepsis and related terms were published in 1992 (Bone *et al.*, 1992). Sepsis is defined as a clinical inflammatory response arising from infection. However, even in the absence of infection, a similar or identical response can arise. Therefore, systemic inflammatory response syndrome (SIRS) is defined as the inflammatory response, with or without infection. SIRS is

diagnosed, according to Bone *et al.* (1992) when a patient presents with more than one of the following clinical manifestations:

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 32mm Hg
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 percent immature neutrophils

Figure 1.1 shows the interrelationship between these related terms. This diagram clearly shows that sepsis can only be defined when a patient has SIRS and indication of infection. The related term bacteraemia, which is defined as a bacterial invasion of the bloodstream, will therefore only occur due to infection, and may or may not result in sepsis (Bone *et al.*, 1992).

The mortality rate for sepsis varies widely between studies and the time point analysed (Winters *et al.*, 2010). Mortality has been estimated between 5% and 90% in different studies, time points and continents (Casalino *et al.*, 1998; Shapiro *et al.*, 2007; Winters *et al.*, 2010). A large study observing sepsis at more than 200 sites in Europe and the USA was published recently. This study included 25,375 patients, and showed that raw hospital mortality was higher in Europe than in the USA (41% vs. 28%), but that adjusted mortality (adjusted for descriptive measures of severity of illness, admission source and survival status) was not significantly higher (32% vs. 31%). Differences in patient and hospital management, intensive care unit (ICU) admission and treatment protocols may be the cause of this difference, but further large studies are required to accurately determine this (Levy *et al.*, 2012).

Mortality after sepsis increases dramatically from discharge to 1 year after admission (Korosec Jagodic *et al.*, 2006; Sasse *et al.*, 1995). Long-term mortality from

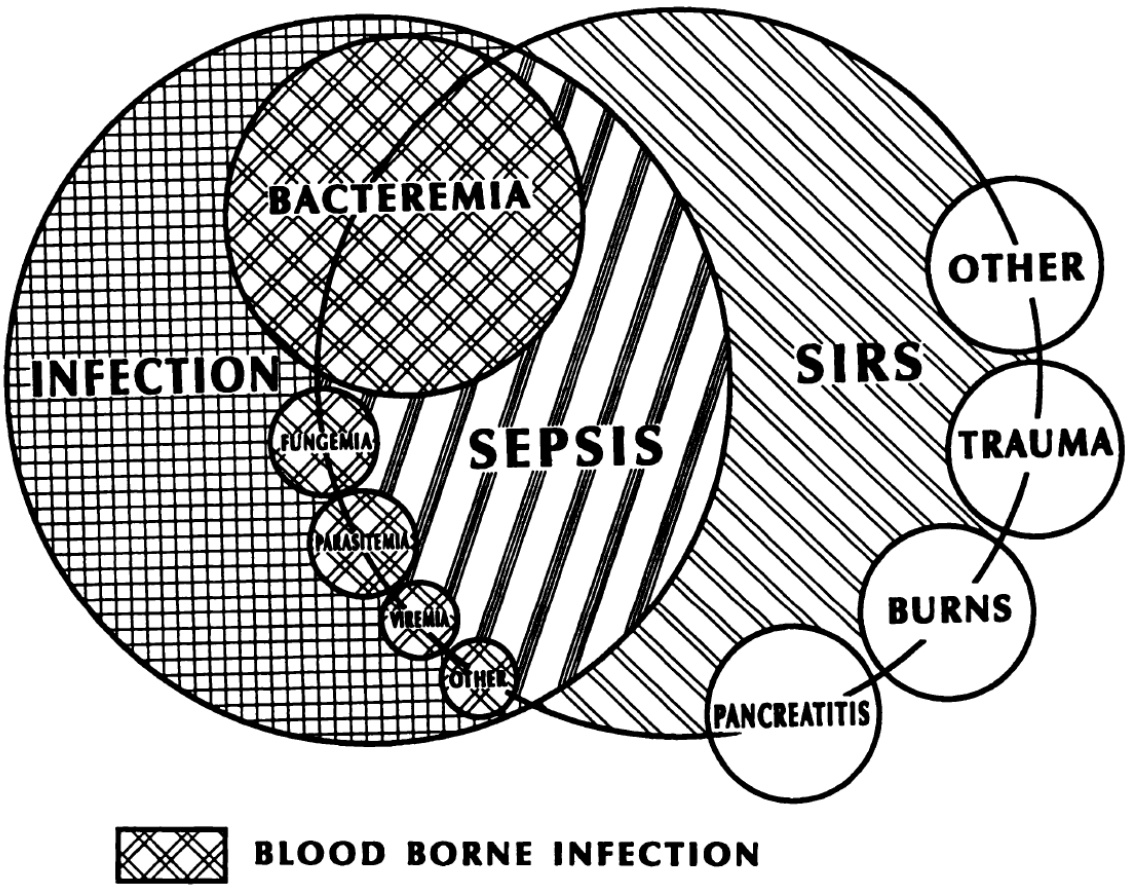


Figure 1.1: The overlap between infection, sepsis and systemic inflammatory response syndrome (SIRS). The corresponding overlap with bacteraemia, fungemia, parasitemia and viremia is also shown, where the presence of infection may or may not lead to SIRS and therefore sepsis. In addition, conditions which lead to SIRS without infection are shown: trauma, burns and pancreatitis. Reproduced with permission from the American College of Chest Physicians (Bone *et al.*, 1992)

sepsis is almost invariably high (Korosec Jagodic *et al.*, 2006; Perl *et al.*, 1995; Sasse *et al.*, 1995; Weycker *et al.*, 2003; Winters *et al.*, 2010), and was estimated at 74% in one very large study (much higher than the 21% hospital mortality recorded) (Weycker *et al.*, 2003). Sepsis is responsible for an estimated 36,800 deaths annually in the UK alone; more than three times as many as breast cancer (Daniels, 2011).

Sepsis mortality for six countries in Europe (UK, France, Germany, Italy, Spain and The Netherlands) was estimated between 99,000 and 128,000 per year, which equates to between 118,000 and 153,000 deaths from sepsis in Europe annually (Davies *et al.*, 2001). Estimates have also shown that 10-14% of European ICU patients that develop sepsis account for approximately 40% of the total costs in the ICU, a total burden estimate of between €6.2 billion and €8.3 billion (Davies *et al.*, 2001).

The sepsis occurrence in acutely ill patients (SOAP) study enrolled 1,177 sepsis patients (37.4% of intensive care unit patients) from 24 European countries to identify the frequency of sepsis in European ICUs and identify etiological, diagnostic, therapeutic and prognostic factors associated with sepsis in the European population. Considerable variation in rates of sepsis was observed between countries in Europe (from 18% in Switzerland to 73% in Portugal). This study showed that there was no difference in ICU mortality whether or not patients had suspected clinical infection and whether or not pathogens were identified. ICU and hospital mortality also varied between countries, with Switzerland exhibiting the lowest, and Portugal the highest mortality (Vincent *et al.*, 2006).

The frequency of hospitalisations with severe sepsis has increased relentlessly: more than doubling each year between 2000 and 2007, with sepsis complicated by organ failure accounting for one in 40 hospitalisations in 2007 (Kumar *et al.*, 2011). Pneumonia or lower respiratory infection is the predominant cause of sepsis (Angus *et al.*, 2001; Levy *et al.*, 2012; Shapiro *et al.*, 2007; Vincent *et al.*, 2006; Weycker *et al.*, 2003).

1.1.2 LRTI in primary care

Clinical definitions of LRTI in primary care vary. LRTIs can be classified into different subcategories, each of which is distinct, but still falls under the bracket of LRTI. The four most common LRTIs presenting in general practice are acute bronchitis, acute exacerbations of chronic obstructive pulmonary disease (aeCOPD), community-acquired pneumonia (CAP) and acute infective exacerbations of asthma (Greene *et al.*, 2011).

LRTIs are the most common reason for consulting a general practitioner (GP) in the UK, but the management and characterisation of these infections is still poorly developed (Greene *et al.*, 2011; Macfarlane *et al.*, 2001). This has an impact on antibiotic prescribing behaviour for LRTI. A recent study showed that antibiotic prescribing (whether or not antibiotics are prescribed, as well as antibiotic choice) is not associated with clinically important differences in recovery, when clinical presentation was taken into account. Variation in prescribing behaviour therefore wastes resources, leads to selection of resistant organisms, puts patients at risk of side effects and undermines self care procedures with no improved outcome for patients (Butler *et al.*, 2009). In addition, non-adherence to antibiotics is common, with an average of only 60% of individuals across Europe taking the antibiotics they are prescribed (Francis *et al.*, 2012b). There are no effective treatments for viral LRTI (Goetghebuer *et al.*, 2004), which is the greatest defined cause of mild LRTI.

Although hospitalisation following a primary consultation for LRTI is low (3%), the median duration of "moderately bad" symptoms is six days, and 6% of individuals experience moderately bad symptoms for a prolonged period of 21-28 days (Francis *et al.*, 2012a). Based on an overall incidence of 44 cases per 1000 in the adult population per year, mild LRTI affects more than 2 million individuals in Europe per year (Macfarlane *et al.*, 2001; World Health Organisation, 2008). Therefore, despite not being associated with very severe disease or high mortality rates, LRTI in primary care still has a large impact on the well-being of a large number of people across

Europe.

1.2 Immunity to infectious disease

Immunity is the process by which the body defends itself against infectious pathogens. Immunity can be divided into two main areas: innate and adaptive immunity.

1.2.1 Innate immunity

Innate immunity is the first line of defence against infection, and consists of mechanisms that are in place before infection, and respond in essentially the same way to every pathogen and repeated infection. Mechanisms of innate immunity include physical barriers (such as epithelial surfaces), soluble and cellular factors.

One of the soluble factors of innate immunity is the complement system. This uses antibodies and mannose binding lectin, as well as direct recognition, to recognise microbes and result in activation. Complement proteins direct the phagocytosis of microbes and lead to inflammation. Complement can be activated by the alternative, classical and lectin pathways.

The cellular components of innate immunity include macrophages and neutrophils, which phagocytose invading microbes and natural killer (NK) cells, which kill infected host cells. Different receptors, including toll-like receptors (TLRs) can recognise molecular patterns of microbes (such as double stranded RNA and lipopolysaccharide) and activate cells to result in killing or secretion of cytokines. Cytokines are secreted by macrophages and NK cells to activate phagocytes and stimulate inflammation. Cytokines are classified into various families, including interleukins (IL) and interferons (IFN), and these proteins are named with this prefix and a descriptive suffix (IL-6 and IFN-α for instance). NK cells recognise major histocompatibility complex (MHC) molecules on the surface of normal host cells, to prevent killing of

uninfected host cells. When cells are infected, the expression of MHC is inhibited, allowing infected cells to be destroyed (Abbas and Lichtman, 2005).

1.2.2 Adaptive immunity

Conversely, adaptive immunity is stimulated in reaction to infectious agents, and due to the creation of immunological memory, responses gain magnitude and defensive capabilities with repeated exposure. It is this principle which underlies the mechanism of vaccination. Adaptive immunity mechanisms can specifically recognise distinct molecules, and "remember" a microbe, responding more quickly to an infection by the same organism many years later. The main cells involved in the adaptive immune response are B and T lymphocytes. B cells produce antibodies, which have regions of high variation that can recognise different antigens and thereby initiate specific B cell responses and effector mechanisms to eliminate the recognised antigens.

There are two major classes of T cells: CD4+ and CD8+ T cells. CD4+ cells have CD4 molecules but no CD8 molecules expressed on their surface, and CD8+ cells have the opposite. Both types of T cells recognise peptide fragments derived from protein antigens expressed on the surface of host cells. This recognition can then lead to macrophage activation in the case of CD4+ T cells, or the killing of virus-infected cells in the case of CD8+ T cells. The MHC is involved with displaying antigens for recognition by T cells. The recognition of antigen leads to an expansion of the types of antibodies and cells that can recognise that antigen, thus increasing the magnitude of the specific immune response (Abbas and Lichtman, 2005).

1.3 Pathogens in LRTI

Streptococcus pneumoniae was the most common pathogen identified in the sepsis cohort used in this thesis. Other pathogens identified include *Haemophilus influenzae*,

Staphylococcus aureus, *Pseudomonas aeruginosa* and influenza virus. *S. pneumoniae* (commonly referred to as "pneumococcus") is a capsulated gram-positive bacteria. The pneumococcus capsule protects the bacteria from the host immune response of phagocytosis, whereby invading microbes are engulfed and destroyed by phagocytes. The polysaccharide composition of the capsule determines the serotype of the pneumococcus, of which there are at least 93. Modern vaccines only confer resistance to certain serotypes (Catterall, 1999; Song *et al.*, 2012).

Rhinovirus is the most common pathogen identified in the mild LRTI cohort used in this thesis. Other pathogens identified include influenza virus, coronavirus, parainfluenza virus, human metapneumovirus, respiratory syncytial virus, *S. pneumoniae* and *H. influenzae*. Rhinovirus is part of the picornavirus family and is therefore a small RNA-containing virus with an icosahedral shell surrounding its RNA core (Rossmann *et al.*, 1985). There are over 160 types of human rhinovirus (HRV) which belong to three species: HRV-A, HRV-B and HRV-C. HRV-A and HRV-C have been shown to cause more severe illness in infants (Lee *et al.*, 2012).

Influenza virus is a pathogen of major public interest, in part due to recent pandemic strains of the virus. Susceptibility to influenza virus has been studied in this thesis. The influenza viruses A, B and C all infect humans, although types A and B are far more common and were identified in the cohorts used in this thesis. Influenza A and B viruses have segmented, negative-strand RNA genomes, and microscopically look extremely similar, but have a slightly different composition of envelope proteins. Both viruses have glycoprotein spikes which project from the virus envelope. Subtypes of influenza virus are named after the subtype of their surface glycoproteins hemagglutinin (HA) and neuraminidase (NA), such that the H1N1 subtype has HA subtype 1 and NA subtype 1 (Bouvier and Palese, 2008; Lamb and Choppin, 1983).

1.4 LRTI and immunity

The host mechanisms involved in innate immunity to particular pathogens are not fully understood. However, *S. pneumoniae* is known to activate innate immunity through complement activation and recognition by TLRs. Studies in mice suggest complement activation is mainly through the classical pathway (Paterson and Mitchell, 2006). Influenza viruses are known to activate host immune responses through the production of interferons via the signalling pathway of TLR7, but also upregulate several other cytokines (Fukuyama and Kawaoka, 2011). Very little is known about the exact innate immune mechanisms associated with rhinovirus infection, although interferons are probably involved (as in viral infections in general) (Bartlett *et al.*, 2012).

However, immunity mechanisms can cause tissue injury and disease in some situations. For instance in the case of sepsis, abnormalities in inflammation and immunity mechanisms lead to the sepsis phenotype (Nee, 2006; Tang *et al.*, 2010). It has been suggested that both a lack of sufficient regulation of the adaptive immune response and increased apoptosis of important immune cells could be responsible for the immune response responsible for sepsis phenotypes. However, further understanding of how the innate and adaptive immune responses contribute to sepsis is necessary (Siner, 2009).

1.5 Genetics of disease

Records as early as the 2nd century A.D. describe circumcision regulations which most likely refer to haemophilia, and demonstrate that its maternal transmission and limitation to males were known even this long ago (Harper, 2004; Ingram, 1976). This heredity was graphically described and the inheritance pattern of the disease was clearly appreciated in a paper by Hay in 1813, which was still very early to be

accurately describing genetic inheritance (Harper, 2004; Hay, 1813).

Genetics associated with disease was therefore appreciated long before heredity was explained, long before genes and DNA were discovered, and long before modern genetics studies. There are many distinct ways that genetics can define both Mendelian disease (disease caused by one genetic locus) and complex disease (disease where susceptibility is influenced by multiple genetic loci and other factors). Genetics has been helpful in defining and explaining disease through a range of evolving methods and techniques. A number of examples of different diseases, the genetic polymorphisms which are associated with them and the techniques used to define them are presented here.

1.5.1 Genetics of disease: Cytogenetics

One of the most dramatic genetic differences that has been identified is altered chromosome number. Amazingly, the structure of DNA was solved before the number of human chromosomes was correctly determined (Tijo and Levan, 1956; Watson and Crick, 1953). Following this discovery, however, several syndromes were found to be associated with altered chromosome number. Lejeune *et al.* (1959) identified an extra small telocentric chromosome in children with Down's syndrome (who were at the time referred to as "Mongoloid" children). Lejeune *et al.* hypothesise (correctly) that this extra chromosome could be the result of non-disjunction during meiosis. They also (correctly) suggest that this could account for the increase in the frequency of Down's syndrome when the mother is older, in concordance with previous studies in *Drosophila melanogaster* (Boyer, 1963; Harper, 2004; Lejeune *et al.*, 1959).

In the same year, sex chromosome abnormalities were also described. An individual with Klinefelter's syndrome was described with an extra sex chromosome (XXY) (Jacobs and Strong, 1959). Later that year, an individual with Turner's syndrome was shown to be missing a sex chromosome (XO) (Ford *et al.*, 1959). In a later study

of chromosomal number in spontaneous abortions, it was shown that 11 out of the 200 spontaneous abortion samples studied also had a missing sex chromosome, implying (by comparison with the incidence of this trait) that only 1 in 40 individuals with this karyotype reach full term. In addition, 33 other samples were shown to have abnormal chromosome number. These were in the form of trisomy (3 copies of one chromosome, as in Down's syndrome; 22 samples), triploidy (3 copies of all chromosomes; 9 samples) and tetraploidy (4 copies of all chromosomes; 2 samples). This study showed that chromosome number is an important factor in spontaneous abortions, and that abnormal chromosome number is surprisingly common (Carr, 1965; Harper, 2004).

Cytogenetics (the microscopic study of chromosome structure) has advanced more recently, and is now used not only to determine chromosome number, but also to identify microdeletions and microduplications, more recently in conjunction with a method called fluorescence in situ hybridization (FISH) (Dave and Sanger, 2007; Lawrence, 2000). Before FISH, however, chromosomes could still be stained to assess translocations, deletions and duplications. This led to the discovery of a deletion in chromosome 15, which can either result in Angelman syndrome or Prader-Willi syndrome, depending whether the deletion was inherited maternally (Angelman syndrome) or paternally (Prader-Willi syndrome) (Kaplan *et al.*, 1987; Knoll *et al.*, 1989; Ledbetter *et al.*, 1981). Experiments defined the inheritance of the deletion using a restriction fragment length polymorphism (RFLP) assay to identify the allele present in affected children and their parents, and showed that the maternal allele was not inherited in individuals with Angelman syndrome (and therefore this defines the deletion), and vice versa in Prader-Willi syndrome (Knoll *et al.*, 1989). This demonstrates that sometimes even when polymorphisms have been characterised with disease, DNA sequence is not the only important factor influencing the genetic impact on disease and epigenetic effects such as this can also influence inheritance.

1.5.2 Genetics of disease: Linkage for Huntington's

The genetic features of Huntington's disease were first described in 1872 alongside the first description of the disease by George Huntington (Harper, 2004; Huntington, 1872). However, it wasn't until over 100 years later that a linkage study identified the genetic loci responsible for the disease. In 1980, a proposal for the first type of genome-wide linkage study was published, which would use linkage to look for genetic loci leading to disease. Just three years after this, a genetic marker for Huntington's was discovered, using RFLPs as DNA markers to analyse polymorphisms throughout the genome in family pedigrees, to determine polymorphisms which segregate with disease (Botstein *et al.*, 1980; Gusella *et al.*, 1983; Harper, 2004). RFLP is still used in genetic association studies, and its invention was wholly dependent on the discovery of restriction enzymes, which specifically cut DNA at a defined sequence site (Roberts, 2005). The exact genetic polymorphism responsible for Huntington's disease wasn't discovered until ten years later, but this genetic linkage mapping was utilised in its discovery. It was shown that Huntington's disease is defined by the number of CAG repeats in a gene which was named Huntingtin following this discovery (Huntington's Disease Collaborative Research Group, 1993).

1.5.3 Genetics of disease: Phenylketonuria

Conversely, phenylketonuria (PKU) can be caused by any one of a great many mutations in the phenylalanine hydroxylase (*PAH*) gene; 548 separate mutations in the gene were recorded as of 2007. Roughly 5% of these do not affect the function of the *PAH* gene, but the rest lead to a spectrum of the PKU condition (Blau *et al.*, 2010).

PKU was first described in 1934, when F∅lling identified that "feeble-minded" children (which we now know had phenylketonuria) had phenylpyruvic acid excreted in their urine, and hypothesised (correctly) that this was due to an inability to properly metabolise phenylalanine (Boyer, 1963; Folling, 1934; Harper, 2004). It was later

shown that raised phenylalanine concentrations in the brain can lead to impaired neuropsychological function, and thereby the mental retardation seen in patients suffering from PKU (Blau *et al.*, 2010).

The inheritance of this condition was identified as autosomal recessive by 1946, and the differing incidence of the condition in different geographical populations was also noted. Ways of altering the metabolism of affected individuals as a form of treatment were already being tested, and the idea of changing the diet of those with the condition was raised in this year (Penrose, 1946). This idea was tested in 1953 with great success, despite the exact cause of the condition still being unknown. A two-year old child with phenylketonuria was given a low-phenylalanine diet for 10 months and showed improvements in mental state. However, when phenylalanine was put back into her diet, the child "lost in a few days all the ground gained in the previous ten months", and within 24 hours of beginning the new diet "could no longer stand" (Bickel *et al.*, 1953). Screening programmes for PKU were underway in the 1960s, and these rapidly led to programmes of population screening which still take place worldwide (Blau *et al.*, 2010; Guthrie and Susi, 1963; Harper, 2004; Levy *et al.*, 1970; National Health Service, 2012).

This story from the discovery of PKU to worldwide screening is proof that detection and treatment of genetic conditions is possible even without the identification of the exact genetic mutations involved. However, for complex disease, the advancement of sophisticated techniques has been necessary to identify genetic polymorphisms associated with disease.

1.5.4 PCR

The Polymerase Chain Reaction (PCR) was first fully described in 1987 (Mullis and Faloona, 1987). This remarkably simple technology can amplify specific fragments of DNA and paved the way for rapid advances in genetics in recent times. Without the

development of PCR, recent genetic discoveries would not have been possible. All current genotyping technologies utilise PCR.

PCR involves the addition of deoxyribonucleotide triphosphates (dNTPs), oligo primers, buffer and Taq polymerase to DNA samples. Taq polymerase is a thermostable polymerase discovered within the organism *Thermus aquaticus*. It remains relatively unaffected by the high temperatures in the PCR procedure. The DNA and PCR reagents mixture is cycled through different temperature steps. The first step heats the mixture to separate DNA into its two strands. A cooling stage then allows the hybridisation of designed oligonucleotides that act as primers, to complementary sequences in the DNA strands. DNA polymerase then performs DNA synthesis starting from the two primers, using the dNTPs in the mixture. The procedure is then performed in repetitive cycles, through which the sequence of DNA bracketed by the primers is amplified, producing double the number of specific double-stranded DNA molecules in each cycle. Therefore, as the cycles are completed multiple times, the amount of DNA of a particular sequence can be increased dramatically (Saiki *et al.*, 1988).

1.5.5 SNPs and linkage disequilibrium

Single nucleotide polymorphisms (SNPs) are single base positions which are found to have a different nucleotide between individuals. SNPs found within the coding regions of a gene can affect the function of the resulting protein. For instance, if a base pair position is usually a T, and is the second base in a codon for leucine (CTC for instance), if this base pair position can also be found as a G, then this will replace the leucine at that position with arginine (CGC). This is called a non-synonymous, missense SNP. SNPs can also change codons to stop codons (non-synonymous, nonsense SNPs), resulting in a truncated protein, and can be silent changes which do not affect the amino acid composition of the resulting protein

(synonymous SNPs).

Linkage disequilibrium (LD) describes the phenomenon whereby polymorphisms under selection or close to each other are inherited together. Therefore, by mapping one locus in a chromosomal region, information about the surrounding polymorphisms can be inferred. SNP genotypes are not randomly distributed between individuals, they follow the rules of LD and are found on specific haplotypes. LD describes a correlation between pairs of variants. High LD between two SNPs means that the observation of a particular allele at one SNP location would mean a high likelihood of observing a particular allele at the other SNP location. When a new variant arises through mutation, it is initially only found on the unique chromosome on which the mutation occurred, meaning that it is only found in the presence of surrounding variation. As recombination and mutation occur, this association between variation is slowly reduced. Recombination is less likely to occur between SNPs that are very close together than between SNPs that are further apart. Therefore, the association between a SNP and its neighbours usually drops the further the neighbouring variation from the SNP of interest, and will drop after a recombination hotspot (The International HapMap Consortium, 2005).

1.5.6 Candidate gene studies

Linkage studies were very successful in determining genetic regions associated with Mendelian diseases, but for complex diseases these studies have been less informative. Focus therefore turned towards other methods to determine genetic associations with disease, and candidate gene studies were the main approach.

Candidate gene studies focus on genes which are implicated as important in the disease of interest, usually based on knowledge of gene function. Rather than scanning the whole genome in a hypothesis-free manner such as is done with linkage studies, hypotheses are made prior to genotyping. Although vast numbers of candidate genes

have been implicated in a whole range of diseases, these associations rarely replicate, unless they use large sample sets and are very statistically significant. There are a number of reasons why this could occur, including heterogeneity between patient populations, a lack of population stratification analysis meaning that related individuals and individuals of differing ethnicity are not removed and therefore act to skew the data, or a lack of linkage disequilibrium (LD) between the causal polymorphism and the genotyped polymorphism (Tabor *et al.*, 2002).

In addition, the less that is known about a disease, the fewer options there are for candidate genes to genotype. Due to publication bias, studies showing positive results are more likely to be published than studies showing negative results, and studies have shown that the first study to report an association often implies a stronger risk than that observed in later publications (Tabor *et al.*, 2002). There is also a lack of correction for multiple testing in candidate gene studies. Even though previous candidate gene studies will have been conducted in the same cohort, individual associations are published with no correction. Following results from the Human Genome Project (Collins *et al.*, 2003) and The International HapMap Project (The International HapMap Consortium, 2005), genome-wide association studies (GWAS) became possible, which have largely made candidate gene studies obsolete.

1.5.7 The Human Genome Project

The Human Genome Project produced the first draft of the full human genome sequence in 2001 (Lander *et al.*, 2001), with the final sequence completed in 2003 (Collins *et al.*, 2003), just 50 years after the structure of DNA was described (Watson and Crick, 1953). This extraordinary feat was the largest genome sequenced at the time, and was eight times larger than the sum of all previously sequenced genomes. The sequence of just one human genome gleaned an enormous amount of information about the broad genomic landscape, including information about genes, transposable

elements, GC content, CpG islands, segmental duplications, simple sequence repeats and more. Comparisons between the human genome sequence and already completed genome sequences could also be performed, allowing discoveries related to the origin of genes and the number and complexity of genes compared to other species (Lander *et al.*, 2001).

The human genome sequence was generated from genomic DNA of multiple individuals, which meant that differences in DNA between people could also be observed. Thereby the number of SNPs identified in the human genome increased one-thousand fold after analysis of the human genome sequence, to describe 1.42 million SNPs. It was also possible to analyse the distribution of these SNPs, with observations that the density of SNPs is higher in regions containing genes, lower on sex chromosomes, and strikingly higher or lower in some regions than average (for example there is a high level of variation observed in the HLA region on chromosome 6) (Chakravarti, 2001; Lander *et al.*, 2001; Sachidanandam *et al.*, 2001).

1.5.8 The International HapMap Project

The International HapMap Project aimed to further this characterisation of variation, by identifying human genetic variation between four populations with European, African and Asian ancestry. These populations were from the Yoruba people of Ibadan, Nigeria, individuals from Tokyo, Japan, Han Chinese individuals from Beijing, China and United States residents with Northern and Western European ancestry collected by the Centre d'Etude du Polymorphisme Humain (CEPH). Phase I of the project aimed to genotype at least one SNP with a minor allele frequency (MAF) ≥ 0.05 every 5kb in every sample from these four populations (a total of 269 individuals).

By mapping common variation in the human genome between populations, The International HapMap Project characterised LD between SNPs and investigated the

evolutionary forces that have shaped variation in natural populations. They showed evidence of selective sweeps in particular genomic regions (for instance the *LCT* gene, which influences the ability to digest dairy products), confirmed that LD is low near telomeres and elevated near centromeres and showed that genes involved in immune responses are more often located in regions of low LD, to name a few examples.

The primary aim of the project, however, was to create a public, genome-wide database of common human sequence variation which could be used to better guide the design and analysis of medical genetic studies (The International HapMap Consortium, 2005). This project was a fundamental step towards the generation of chip arrays for GWAS.

1.5.9 Genome-wide association studies

GWAS became possible due to the International HapMap Project described above, the design and manufacture of dense genotyping chips to genotype hundreds of thousands of SNPs, and the assembly of large, well-characterised sample collections for many common diseases (de Bakker *et al.*, 2005; Hirschhorn and Daly, 2005; The Wellcome Trust Case Control Consortium, 2007). GWAS are based on the common disease-common variant (CD-CV) hypothesis, which states that common diseases are associated with common variants (Gibson, 2012; Lander, 1996). Modern GWAS involve the genotyping of 300,000 - 2,500,000 SNPs in many samples. In a case-control study, each SNP is analysed in cases and controls, with the aim of detecting differences in genotype or allele frequencies between healthy individuals and cases with the disease or trait of interest. GWAS make use of LD in the genome, which means that not all SNPs in the genome need to be typed in order to detect an association (de Bakker *et al.*, 2005). Figure 1.2 explains this further, and shows examples where SNPs are either directly or indirectly (as in a GWAS) tested for an association.

Fine-mapping around a SNP that associates with disease can often identify the

causal SNP for an association, which will have been tagged by the SNPs genotyped in the GWAS. A landmark paper from the Wellcome Trust Case Control Consortium (WTCCC) presented some of the earliest and largest case-control GWAS, and demonstrated the power of these studies to detect genetic associations with common disease (The Wellcome Trust Case Control Consortium, 2007). This paper also highlighted important considerations with these studies in relation to their design and analysis.

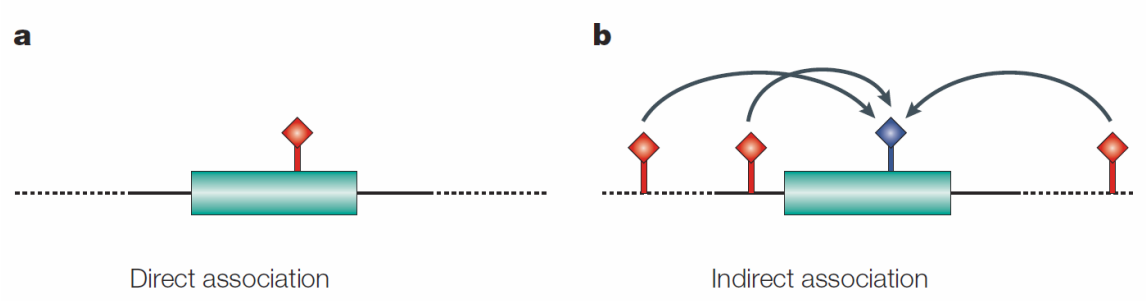


Figure 1.2: Testing SNPs directly or indirectly for association. **(a)** A case in which the functional SNP (in red) responsible for an association is directly typed and tested for an association. This scenario would occur, for instance, in a candidate gene study, where a SNP is chosen for genotyping based on prior knowledge about its possible function. For example it could be a missense SNP which is likely to affect the function of the candidate gene (green rectangle) of interest. **(b)** This describes a case such as that likely to be found in a genome-wide association study (GWAS). Genotyped SNPs (in red) are chosen to be genotyped based on LD in the region, such that each genotyped SNP is in high LD with as many ungenotyped SNPs as possible. In this example, the red genotyped SNPs provide information about the blue SNP, such that the blue causal SNP is tested for association indirectly. If the red SNPs associate with a disease/trait in a GWAS, the blue SNP most likely will too. In this way, fewer SNPs need to be directly genotyped in the first steps of a GWAS. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics (Hirschhorn and Daly, 2005), copyright 2005.

1.5.9.1 False positives and multiple testing

Due to the large number of tests performed in any GWAS (one for each SNP, including additional analyses for different phenotypes within one study), conservative significance levels have been established to distinguish true from false-positive results. A p-value of 0.05 by its very definition means that the association observed would occur by chance alone 5% of the time. Therefore, if this p-value threshold was used

in a GWAS, it would be expected that 5% of results (25,000 SNPs in a GWAS using a 0.5 million SNP chip) would reach this p-value threshold just by chance and would therefore be false positive results. Thus, to correct for this multiple testing problem, a much lower p-value threshold of $5x10^{-8}$ has been set, above which GWAS results are not considered to be genome-wide statistically significant (Kraft *et al.*, 2009).

1.5.9.2 Advantages and disadvantages of GWAS

GWAS has been enormously successful in determining genetic loci which associate with a wide range of diseases and traits. As of 15th November 2012, the GWAS catalogue contains 9277 published associations with p-values of $1x10^{-5}$ and lower, including associations with aging, alcohol consumption, Alzheimer's disease, eye colour, heart failure, height, multiple sclerosis, Parkinson's disease, psoriasis, rheumatoid arthritis, type 1 and 2 diabetes and ulcerative colitis, to name just some (Hindorff *et al.*, 2012).

By design, GWAS requires no prior information about gene function, and is essentially a hypothesis-free method for determining genetic associations across the whole genome. This means that when dense genotyping chips are used with appropriate sample sizes, GWAS have the potential to find any common genetic variation associated with disease in the genome. GWAS are susceptible to the effects of stratification, but in terms of population stratification, this can usually be corrected for using the data available from a GWAS (Anderson *et al.*, 2010; Hirschhorn and Daly, 2005). However, GWAS are unable to assess the affect of rare variants on genetic risk and could therefore be missing a large and important aspect of genetic susceptibility to disease. In addition, GWAS are often underpowered to detect associations with very small effect sizes (odds ratio between 1 and 1.2) without extremely large sample sizes (Spencer *et al.*, 2009).

1.5.9.3 Imputation and phasing

The design of GWAS chips has meant that imputation can be performed using genotypes spanning the entire genome. Imputation is an important process in the analysis of any GWAS dataset. This two-step process allows the inference of millions of ungenotyped SNPs at high accuracy, meaning that GWAS involving for instance one million genotyped SNPs can be analysed for many millions of SNPs. This increases the power of GWAS, can be used to aid fine-mapping, and means that meta-analysis can be performed for sample sets genotyped on different SNP sets (Howie *et al.*, 2012, 2011; Li *et al.*, 2009; Marchini and Howie, 2010). It was recently shown that imputation can be performed for data from a wide range of human populations, despite a lack of specific reference panels for these populations (Howie *et al.*, 2011). Figure 1.3 shows the main principles of genotype imputation.

In order for imputation to be accurately performed from GWAS data, data must first be phased to construct haplotypes which can then be used to impute genotypes. IMPUTE2 performs phasing as the first step of the imputation process (Howie *et al.*, 2009). However, phasing can now be performed before imputation (pre-phasing) which can reduce the computational cost of imputation while maintaining high levels of accuracy. Performing this step prior to imputation also means that, as reference panels for imputation evolve, imputation can more easily be repeated with more sophisticated reference panels without implicitly repeating phasing for no reason (Howie *et al.*, 2012).

Phasing works by looking at the genotypes of an individual in chunks and comparing them to the haplotypes of other individuals in the sample set and/or reference panel (depending on the exact phasing method). This establishes best-guess phased haplotype data for each individual. This means that genotype data is converted to haplotype data for every individual, where the genotypes of neighbouring SNPs on the same chromosome of a chromosome pair are inferred (Delaneau *et al.*, 2012; Howie

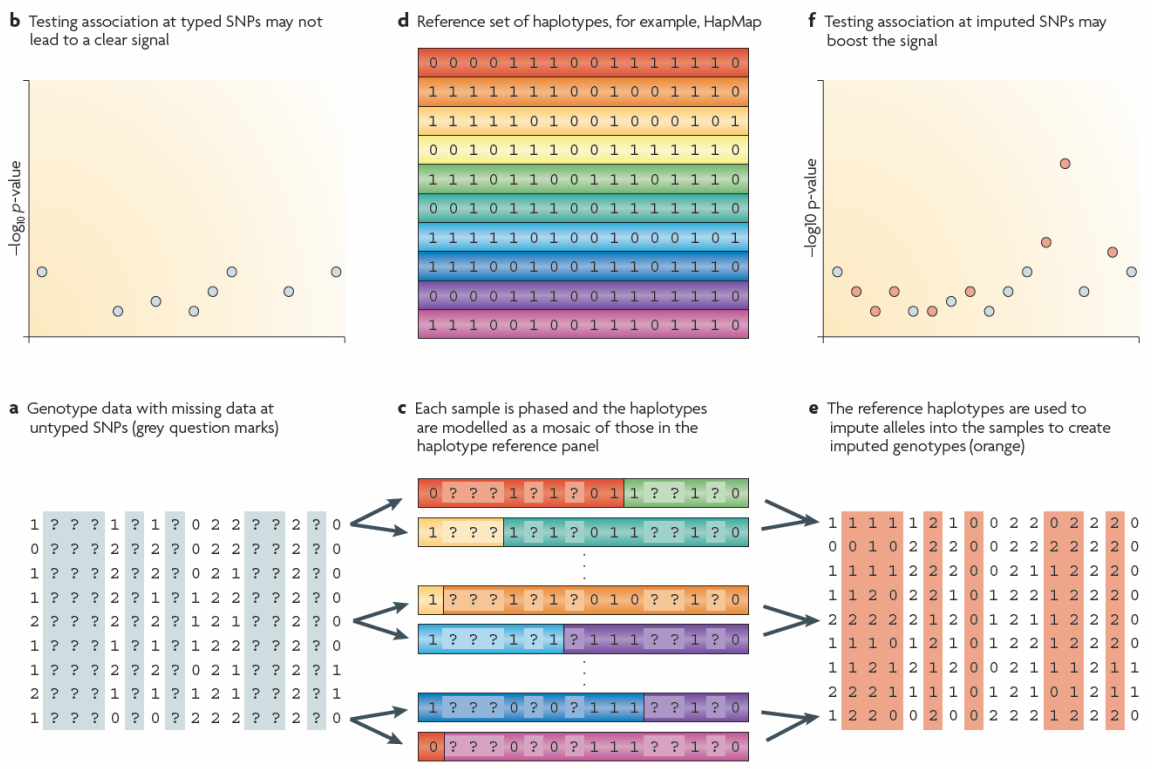


Figure 1.3: This figure illustrates the steps involved in imputation from GWAS data. Data in **(a)** has been genotyped with a genome-wide chip, and therefore there are SNPs which have not been typed in the data. This data may show an association plot such as that shown in **(b)**. Sample data in **(a)** is phased to generate haplotypes as in **(c)** (three samples shown). A reference set of haplotypes from HapMap or thousand genomes, as in **(d)**, is then used to impute the data and infer genotypes with high frequency. The phased data in **(c)** have been coloured to match the haplotype which they correspond to in **(d)**. Imputation therefore generates the genotype data in **(e)**. An association plot for this combined genotyped and imputed data may resemble **(f)**, following association analysis which takes the uncertainty associated with each imputed SNP into account. Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics (Marchini and Howie, 2010), copyright 2010.

et al., 2012, 2009).

Different phasing methods perform phasing in slightly different ways, and therefore require different computational loads and result in differing accuracy of phased data (Delaneau *et al.*, 2012; Howie *et al.*, 2012). Segmented HAPlotype Estimation and Imputation Tool (SHAPEIT) (Delaneau *et al.*, 2012) is probably the fastest and most accurate pre-phasing method available at the time of writing (Delaneau *et al.*, 2012; Howie *et al.*, 2012). Therefore, the process by which this method works to phase data is described next.

SHAPEIT (Delaneau *et al.*, 2012) improves on previously available phasing methods (IMPUTE2 and MaCH) to create a method which scales linearly as sample size and SNP number increase. This method uses the data in a sample set in a way which is simplified but does not ignore any portion of the data. Haplotypes of individuals are collapsed into a graph structure, such that the same amount of data can be represented in a smaller space, thereby simplifying how it can be used. Haplotypes of all samples are split into chunks, so each chunk has a number of possible haplotypes observed in the sample set. Paths between chunks are then defined and weighted depending how often haplotypes in neighbouring chunks are found together. In a similar way, the genotypes of individuals are split into chunks along the genome, where each chunk contains a defined number of heterozygotes. These chunks can then be divided into the possible haplotypes that would represent each chunk of genotypes, and paths between chunks can be defined and their probabilities calculated. These haplotype defining processes occur as a loop, improving with each iteration, and using an average of the iterations to define the final haplotypes for each individual (Delaneau *et al.*, 2012, 2011).

Following phasing, imputation can be performed using a reference panel of haplotypes such as those available from The HapMap Project, or more recently from the 1000 Genomes Project (described below). This reference panel is used to accurately

infer the genotypes of SNPs which are found in the reference panel but not in the dataset to be imputed. Using the haplotypes available for the samples from phasing, and comparing these to the haplotypes in the reference for the same region, imputation methods can then define the probability of seeing each genotype at positions ungenotyped in the sample set (Howie *et al.*, 2009; Marchini and Howie, 2010). This means that output genotype data from imputation is formatted as probabilities rather than absolute genotypes. The uncertainty level for each genotype can be taken into consideration when analysing data, and a reliability score for each association result is calculated from these probabilities. For genotypes which were genotyped in the sample set rather than imputed, or for ungenotyped SNPs which are found in perfect LD with genotyped SNPs, the probability of the genotype will be 1 (Marchini *et al.*, 2010; Pei *et al.*, 2010).

1.5.10 The 1000 Genomes Project

The 1000 Genomes Project is an ongoing venture which aims to "discover, genotype and provide accurate haplotype information on all forms of human DNA polymorphism in multiple human populations" (1000 Genomes Project Consortium *et al.*, 2010). Phase 2 of the HapMap Project obtained genotypes for 210 unrelated individuals in 4 populations at approximately four million SNPs. As a comparison of scale, the phase 1 variant call set from the 1000 Genomes Project (released in March 2012) includes 1,092 individuals from 14 populations typed at >38 million SNP sites (1000 Genomes Project Consortium, 2012b; Howie *et al.*, 2012). This vast data set has been used to characterise genetic variation within and between populations, and more specifically rare variants, and their locations in the genome (1000 Genomes Project Consortium, 2012b).

The latest release of 1000 Genomes Project data has meant that imputation can be performed more accurately for a wider range of human populations (Howie *et al.*,

2012). The continuing evolution of data from the 1000 Genomes Project is a prime reason for performing phasing before imputation as mentioned earlier; data can more easily be re-imputed as new data freezes become available. Data from this project is also utilised when screening data from medical genetics next-generation and exome sequencing projects. Comparisons between 1000 Genomes data and sequencing data from cases can illuminate rare variants which are over-represented in the case population, or gene regions with more rare variants than expected (1000 Genomes Project Consortium, 2012b).

1.6 Genetics of infectious disease

It has long been apparent that host genetic factors can significantly influence mortality from infectious disease; one landmark study showed that adoptees who had a biological parent who died prematurely from infectious disease had a relative risk of death from infectious disease of 5.81, while the relative risk of death from all causes was just 1.71 (Sorensen *et al.*, 1988). More recent similar studies have also shown a strong genetic aspect to infectious disease (Petersen *et al.*, 2010, 2002).

The earliest example of a specific genetic variant conferring susceptibility to infectious disease is that of the sickle-cell gene and susceptibility to malaria (Allison, 1954). In 1978, the restriction endonuclease HpaI was shown to be useful in the diagnosis of sickle-cell trait, as this enzyme would cut β-globin gene DNA in a different pattern in sickle-cell carriers, sufferers and unaffected individuals (Kan and Dozy, 1978). This is an example of balancing selection, as sickle-cell trait carriers are protected from malaria, which maintains the polymorphism and therefore sickle-cell trait in the population. A large West African GWAS demonstrated a strong association ($p=4.5x10^{-14}$) between rs334, the sickle haemoglobin (HbS) causal polymorphism, and risk of disease caused by the malaria parasite *Plasmodium falciparum* (Jallow

et al., 2009).

Another similar example was discovered and published more recently. Two independent variants in the *APOL1* gene were found to confer risk of kidney disease. These variants seem to be only present in individuals of African descent, African Americans have higher rates of kidney disease than European Americans, and the *APOL1* gene region on chromosome 22 harbours signals of strong positive selection. It has been shown that only *APOL1* gene products with these variants are able to lyse trypanosomes and these variants probably confer resistance to trypanosomal infection and protect from African sleeping sickness (Freedman *et al.*, 2010; Genovese *et al.*, 2010). Both of these findings are probably examples of balancing selection, a phenomena also proposed between inflammatory bowel disease and mycobacterial infection (Jostins *et al.*, 2012).

Another example of an infectious disease association is a 32 base pair deletion in the *CCR5* gene which was shown to confer resistance to HIV infection, even in individuals who have been multiply-exposed (Liu *et al.*, 1996; Samson *et al.*, 1996). This deletion generates a truncated dysfunctional protein which is not expressed on the cell surface, and can therefore not be used for HIV entry into cells (Lederman *et al.*, 2006). This phenomenon was utilised in a stem-cell transplant to treat a HIV-1 positive patient with acute myeloid leukaemia. Stem cells from a donor who was homozygous for this 32 base pair deletion were used for the transplant, and the patient was still HIV-1 virus free 20 months after transplantation and the discontinuation of antiretroviral therapy (Hutter *et al.*, 2009).

Recently, GWAS has proved extremely useful in defining genetic variants associated with infectious disease (Chapman and Hill, 2012; Hill, 2012). Studies in populations across the world have found associations with tuberculosis, leprosy, dengue, hepatitis B and more (Chapman and Hill, 2012; Kamatani *et al.*, 2009; Khor *et al.*, 2011; Thye *et al.*, 2010; Wong *et al.*, 2010a,b; Zhang *et al.*, 2009). The first GWAS of

an infectious disease was published in 2007, and studied host control of HIV-1 (Chapman and Hill, 2012; Fellay *et al.*, 2007). Many strong associations were identified in the first GWAS of leprosy susceptibility in a Chinese population, and a further GWAS of susceptibility to leprosy in an Indian population identified a further association in *TLR1* (Wong *et al.*, 2010b; Zhang *et al.*, 2009).

GWAS of infectious disease has been limited to a selection of diseases and populations, however. Of 36 infectious disease GWAS in the GWAS catalogue (updated 11/4/12), 12 (33%) of these are studies of HIV or AIDS (Hindorff *et al.*, 2012). However, according to the World Health Organization (WHO), HIV accounts for only 15% of deaths (estimates for 2008), 15% of Disability-adjusted life years (DALYs) (2004 estimates) and just 0.05% of incidence (2004 estimates) of infectious disease worldwide (World Health Organisation, 2008). In addition, ten of these GWAS have been performed solely in individuals with European ancestry (Hindorff *et al.*, 2012). That accounts for 28% of the published infectious disease studies in the GWAS catalogue, despite only 4% of infectious disease cases (according to 2004 estimates) and 4% of infectious disease deaths (according to 2008 estimates) occurring in Europe (World Health Organisation, 2008).

1.7 Genetics of common infectious diseases in Europe

GWAS has been used to identify genetic variants associated with infectious disease in Europe (Chapman and Hill, 2012). The first GWAS of an infectious disease was conducted in eight European and one Australian cohorts, for host control of HIV-1. This GWAS identified three hit regions associated with control of infection; *HCP5* (which tags *HLA-B*), *HLA-C* and *ZNRD1* (Chapman and Hill, 2012; Fellay *et al.*, 2007). Further GWAS in Europeans have also studied HIV, and have analysed viral

load, viral control and disease progression (Chapman and Hill, 2012). Apart from those genes already mentioned, the most significant observed association in these studies, and the finding with the greatest odds ratio, is for the *MICA* gene and HIV control ($p=1.4x10^{-34}$, OR=4.4) (Pereyra *et al.*, 2010).

The largest European infectious disease GWAS published so far, in terms of the total number of samples in the study, is for meningococcal disease, with 475 cases and 4703 controls in the discovery sample set, and 968 cases and 1376 controls in the replication cohort. This GWAS found a significant association between variants in the *CFH* gene region and susceptibility to meningococcal disease (Davila *et al.*, 2010).

Several GWAS in Europeans and the rest of the world have identified an association between genetic variation in *IL28B* and both treatment-induced hepatitis C viral clearance and spontaneous clearance of the hepatitis C virus, showing that these two characteristics overlap in terms of their genetics. Many patients with chronic hepatitis C will not be cured by treatment, and those patients with African ancestry are less likely to be cured than Europeans. In addition, approximately 30% of individuals spontaneously clear hepatitis C infection (Ge *et al.*, 2009; Rauch *et al.*, 2010; Thomas *et al.*, 2009). *IL28B* encodes IFN-λ3. IFN-λ proteins are up-regulated in the presence of viruses, and can activate a host of genes with immunomodulatory function, via the JAK/STAT pathway. *IL28B* genotyping is now in wide clinical use to predict treatment responses in hepatitis C (Urban *et al.*, 2012).

Despite these successes in GWAS of common infectious diseases in Europe, there are infectious diseases which are more neglected in terms of genetics studies, in comparison to others and considering the impact of disease. Figure 1.4 shows genetics publications in the human genome epidemiology (HuGE) navigator database compared to the disability-adjusted life years lost (DALYs) worldwide for a selection of infectious and chronic diseases. Of the three European GWAS discussed here,

HIV/AIDS and hepatitis C are among the six most frequently studied infectious diseases. Lower respiratory infection is the disease with the highest morbidity worldwide in this plot, and according to this data, is the most neglected disease group in terms of genetics studies (Rowell *et al.*, 2012).

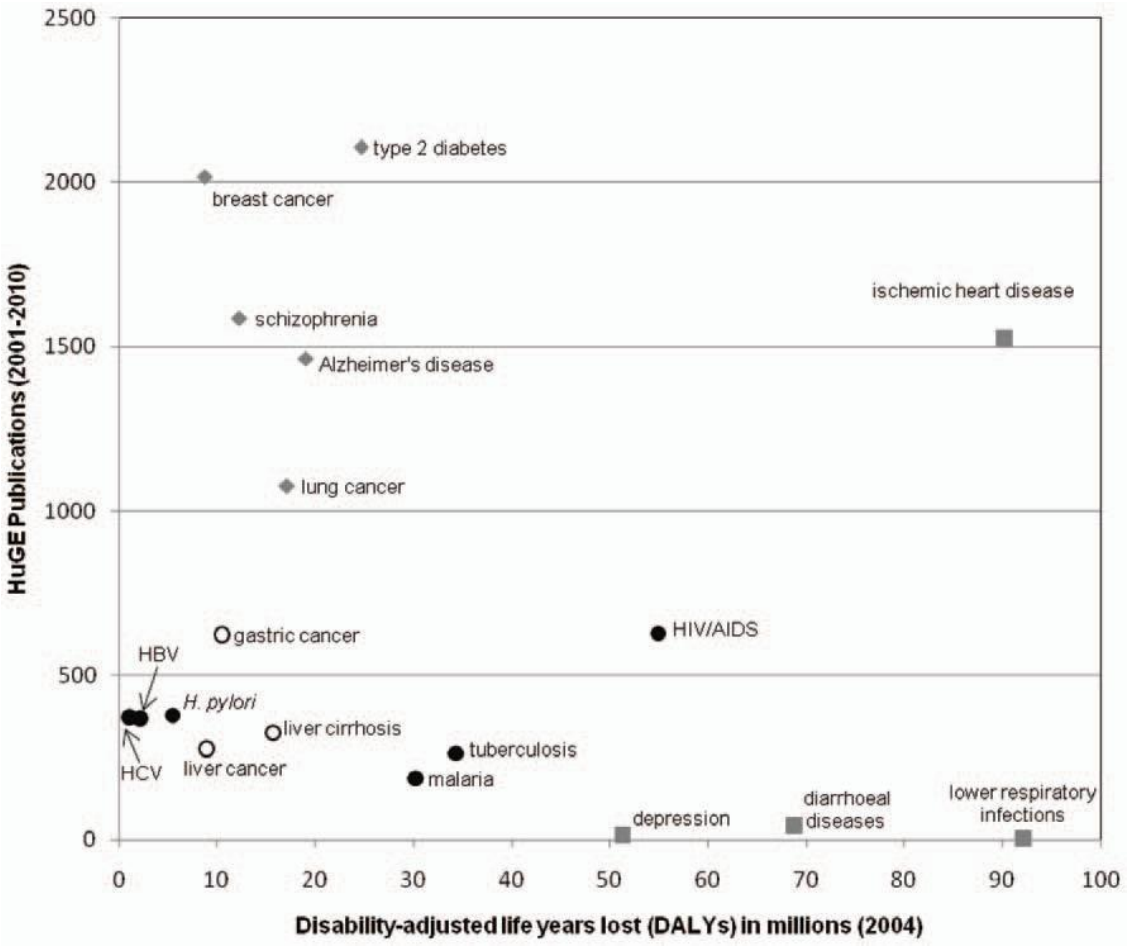


Figure 1.4: Infectious and chronic diseases are represented according to worldwide morbidity (disability-adjusted life-years, DALYs) and publication frequency in human genome epidemiology (HuGE), 2001-2010. Solid black circles represent the six most studied infectious diseases. Open black circles represent three chronic diseases that are often part of the natural course of infection with hepatitis B virus, hepatitis C virus and *Helicobacter pylori* infection. Grey diamonds denote the most frequently studied diseases overall, and grey squares denote the diseases with the greatest morbidity worldwide (Rowell *et al.*, 2012).

Although it is known that there is a genetic component for susceptibility to infectious disease, and although some variants have been defined in other diseases, there

are no known genetic variants which convincingly confer susceptibility to sepsis or LRTI.

1.7.1 Genetics of susceptibility to CAP and sepsis

Susceptibility to sepsis is a neglected phenotype within modern genetics and has not to my knowledge been studied using any high-throughput technique. A number of candidate gene studies exist, but these are usually underpowered and collectively have only studied a limited number of SNPs within the genome. In addition, due to the heterogeneity in sepsis, it is rare to find studies which precisely overlap in terms of phenotype.

There have been many studies of candidate polymorphisms and sepsis, with polymorphisms in 38 genes implicated as potential contributors to susceptibility to pulmonary infection and sepsis (Chung and Waterer, 2011). Most studies only report analysis for one polymorphism at a time, however, and usually with small sample sizes.

A recent paper investigated host genetic risk factors for susceptibility to CAP in Russian patients requiring hospitalisation. This study is one of very few which studies many polymorphisms for susceptibility to CAP at once. The authors studied 334 patients with CAP, and compared these to two different control groups. The first group were a group of healthy individuals that were medical workers repeatedly exposed to infectious agents through contact with patients, who had not contracted CAP during or before the study (141 individuals). The second group were a group of healthy population controls (314 individuals). The patients and controls were genotyped for 13 variants (including SNPs and INDELs) in 11 genes. A recessive association with one SNP in the *CYP1A1* gene remained statistically significant after Bonferroni correction (post-correction p=0.01 with the first control group, p=0.0036 with the second control group). *CYP1A1* was included in the study because it is a

xenobiotics detoxification gene. These genes encode enzymes which are involved in the detoxification and excretion of a range of compounds, and thereby participate in resistance to harmful metabolites. The associating SNP is functional and determines gene expression of *CYP1A1* (Salnikova *et al.*, 2012).

A Spanish group have published findings with numerous candidate genes and CAP phenotypes in one of the largest CAP sample sets utilised for genetics studies. They have published studies for 22 variants in 11 genes, and have studied susceptibility, survival and severity in CAP phenotypes (Garcia-Laorden *et al.*, 2011, 2008; Martin-Loeches *et al.*, 2012; Sole-Violan *et al.*, 2010, 2011). Their susceptibility analyses include large numbers of patients and controls (1262 CAP patients and 1224 healthy controls in the largest study (Sole-Violan *et al.*, 2011)). However, they only report statistically significant associations with susceptibility for surfactant protein haplotypes (lowest p-value=$7\text{x}10^{-5}$) (Garcia-Laorden *et al.*, 2011). With Bonferroni correction for the 22 published variants, which does not account for any other variants that have been studied but not published, a p-value threshold of 0.002 should be achieved before a result is considered statistically significant. Apart from the susceptibility result for surfactant protein haplotypes mentioned above, there are two published associations from the group which meet this criterion. One is an association with bacteraemia in pneumococcal CAP for a functional SNP in the *FCGR2A* gene (p=0.00016), and the other is an association with acute respiratory distress syndrome (ARDS) for a promoter SNP in the *IL6* gene (p=0.002). However, both these analyses include very small numbers of samples (85 bacteraemic vs. 234 non-bacteraemic pneumococcal CAP cases; 20 ARDS patients vs. 286 non-ARDS patients) (Martin-Loeches *et al.*, 2012; Sole-Violan *et al.*, 2011), and so genotype frequencies (in cases in particular) are likely to be inaccurate.

Studies of genetics of survival from sepsis are slightly more advanced; one GWAS paper has been published for prognosis and response to therapy in sepsis (Man *et al.*,

2012). Primarily this GWAS was used to look for genetic markers that predict drug response to the drug Drotrecogin alfa (DAA), a recombinant form of activated protein C. One SNP (in the 5' untranslated region of LOC222052) was shown to associate with treatment response (p-value $< 5\text{x}10^{-7}$). This study also analysed genetic factors associated with mortality at 28 days. This survival analysis "yielded fewer interesting genetic candidates", however. This analysis included quite a small number of samples (total n=1446), though, which could explain the lack of findings.

Despite extensive research very few robust associations between genetic polymorphisms and sepsis phenotypes have been described to date (Sutherland and Walley, 2009). Possible explanations for these inconsistent findings include inadequate sample sizes (and hence a lack of study power), unidentified population sub-structure, failure to correct for multiple testing during statistical analysis, variations in phenotype definition and unrecognised host-pathogen interactions.

1.7.2 Genetics of susceptibility to LRTI in primary care

Genetics studies published for LRTIs focus on severe disease leading to hospitalisation, are usually conducted in children, and usually involve the analysis of one pathogen causing disease. Respiratory syncytial virus (RSV) is probably the most extensively studied subgroup of LRTI, and is by far the most common cause of viral LRTI in infants and children (Miyairi and DeVincenzo, 2008; van Woensel *et al.*, 2003). One UK-based study utilised data from 1,308 children, including 436 children under the age of five who were admitted to hospital with an LRTI. This study examined familial and environmental risk factors for LRTI, and showed that there is some familial susceptibility to severe LRTI early in life (Goetghebuer *et al.*, 2004). In 2007, a study looked at 384 SNPs in 220 candidate genes for genetic susceptibility to RSV bronchiolitis in 470 hospitalised children compared to 1008 population controls. The four SNPs with the lowest p-values for significance were in the genes *VDR* (p=0.0017),

JUN (p=0.0093), *IFNA5* (p=0.0093) and *NOS2* (p=0.0031) (Janssen *et al.*, 2007). However, almost all other studies for RSV are far too small to accurately define genetic effects (Miyairi and DeVincenzo, 2008).

Studies have also been performed with severe acute respiratory syndrome (SARS) coronavirus, another severe form of LRTI. These studies have mostly used larger sample collections, but have only found associations with polymorphisms in *CLEC4M* (p=0.005) and *FCGR2A* (p=0.03), despite studying over 34 polymorphisms (Chan *et al.*, 2007; Ip *et al.*, 2005; Khoo *et al.*, 2008; Li *et al.*, 2008; Wang *et al.*, 2008; Yuan *et al.*, 2005; Zhang *et al.*, 2005).

Chronic obstructive pulmonary disease (COPD) is a risk factor for LRTI, and as mentioned earlier, acute exacerbations of COPD is one of the four commonest types of LRTI. GWAS have been performed, and have identified three genetic loci for susceptibility to COPD (CHRNA3/5, HHIP and FAM13A) (Pillai *et al.*, 2009). However, a further study then showed that the polymorphisms in these regions are associated with subtypes and risk factors for COPD. For instance, the CHRNA3/5 region is associated with increased smoking intensity, airflow obstruction and emphysema in individuals with COPD. It is therefore possible that these SNPs are actually associated with an increase in behaviour or clinical symptoms which lead to COPD (Pillai *et al.*, 2010).

Rhinovirus is one of the leading causes of LRTI in primary care. A recent paper showed that host susceptibility to illness caused by rhinovirus varies significantly between individuals (Lee *et al.*, 2012). However, it is unknown whether this inter-individual difference in susceptibility has a genetic component and any specific genes involved remain unknown.

1.8 Outline of this thesis

This thesis explores human genetic susceptibility to common severe and mild lower respiratory infections in Europe, in a variety of ways. Using a combination of genome-wide and candidate gene approaches, this work aims to identify genetic factors associated with susceptibility to severe sepsis and mild LRTI. Challenges associated with modern studies for genetic susceptibility to these diseases are also addressed. I have used both old and new technologies and methods to extract DNA, genotype SNPs and analyse associations.

Chapter 3 describes experiments to find the optimal methods to extract DNA for three different sample types, for different types of genetic association study. As the basis of all association studies, DNA extraction is a very important step, and this report of methods for this important procedure could be of use to anyone starting a new sample collection or deciding the best way to process their samples.

Chapter 4 investigates polymorphisms in two candidate genes for susceptibility to severe sepsis and mild LRTI. I analyse subgroups within the large sepsis and mild LRTI cohorts, and highlight the difficulties associated with candidate gene studies and the findings published from them. I have genotyped four SNPs in the *MBL2* gene, and shown a lack of association. In addition, I have found a recessive association for a SNP in the *IFITM3* gene with severe H1N1 influenza sepsis and mild viral LRTI.

Chapter 5 combines modern GWAS approaches and classic candidate gene approaches to present what is probably the first reported study of its kind. This study analyses genetic susceptibility to mild LRTI in primary care using SNPs chosen from candidate genes and GWAS findings of severe disease, and shows that some genetic loci do affect susceptibility to these mild infections. I replicate a finding from an unpublished invasive pneumococcal disease GWAS dataset using mild bacterial LRTI cases, demonstrate two separate associations for smokers with LRTI and identify a new locus which associates with genetic susceptibility to rhinovirus.

Chapter 6 presents an approach to study susceptibility to severe sepsis genome-wide. This study combines eight GWAS datasets from across Europe, analysing susceptibility to sepsis across Europe in the largest study of its kind in infectious disease. Having processed the data and imputed genotypes for all sample sets, I have preliminary results from this study, which describe novel, promising associations with susceptibility to sepsis.

In the final chapter, I summarise these findings and consider the future of the field. It is hoped that, through an increased understanding of susceptibility, protection and disease mechanisms, these findings could eventually lead to new therapies and vaccines, and thereby lower incidence of disease.

Chapter 2

Materials and methods

2.1 Sample collections

The following sample collections are used in this thesis.

2.1.1 GRACE

(Genomics to combat Resistance against Antibiotics in Community-acquired LRTIs in Europe)

This collection includes approximately 3000 LRTI cases from 16 networks across 11 countries in Europe, with approximately 3000 controls matched by age, location and time of collection. Patients were being assessed in a primary care setting for an illness where an acute or worsened cough was the main symptom, had a clinical presentation that suggested a LRTI with a duration of up to and including 28 days, were consulting for the first time within this illness episode, were seen within normal consulting hours, and were considered immunocompetent. Controls were individuals attending the same GP surgery for reasons other than infectious causes. Controls were matched for age, sex, location and time of collection. All individuals included in analysis were of self-reported Caucasian ancestry. Individuals were between 18 and 88

years of age, and the median age was 53 years; 40% were male. GRACE was funded by the 6th Framework Programme of the European Commission under the reference LSHM-CT-2005-518226.

2.1.2 GenOSept

(Genetics Of sepsis and Septic shock)

The GenOSept sample collection consists of 1770 sepsis cases from across Europe. GenOSept aims to assess genetic factors associated with outcome from sepsis. Sepsis was diagnosed based on the International Consensus Criteria, 2003 (Bone *et al.*, 1992). Patients admitted to ICU with sepsis caused by community acquired pneumonia (CAP) or faecal peritonitis (FP) were recruited to the GenOSept study between 13/6/2006 and 31/12/2008 in 118 different hospitals in 17 different countries in Europe. Patients all had clinical evidence of infection, evidence of a systemic response to infection and evidence of severe sepsis, as defined in appendix chapter B. Individuals were between 18 and 98 years of age, and the median age was 67. 58% of the cases were male. GenOsept was funded by the European Union 6th Framework-Programme of RTD Funding, The Sainsbury Family Trust, The Intensive Care Society and the Wellcome Trust.

2.1.3 GAinS

(Genomic Advances in Sepsis)

GAinS is a UK collaborative group and part of the European GenOSept study. The GAinS group has overseen the collection of DNA from over 1800 sepsis patients from around the UK. GAinS began recruiting on 30-9-2005 and the study is still actively recruiting. Many of the patients collected in GAinS were included in the GenOSept GWAS. However, many samples have also been collected since the GenOSept study ended, and these have been used in further GWAS and candidate

gene genotyping, and will be used for the replication of GWAS results.

2.1.4 Pneumococcal Empyema

More than 600 cases of thoracic empyema have been collected from children across the UK. A subset of these cases is caused by *S. pneumoniae*. This collection is ongoing and further phenotypic information is not available at this time.

2.1.5 Kenyan Bacteraemia

Approximately 2000 bacteraemia cases (as defined by a blood culture positive bacterial infection) from children <13 years, and approximately 3000 cord blood samples to use as controls have been collected in Kilifi District Hospital, Kenya. Ethical approval for the collection of the Kenyan samples was given by the Kenya Medical Research Institute (KEMRI) National Scientific Steering and Research Committees. The pneumococcal samples used in this thesis all had confirmed infection by *S. pneumoniae* (Scott *et al.*, 2011).

2.1.6 IPD

(Invasive Pneumococcal Disease)

Cases were collected from patients with IPD (294 samples) from three hospitals in Buckinghamshire and Oxfordshire, UK (John Radcliffe, Horton General and Wycombe General hospitals). IPD was defined as the isolation of *Streptococcus pneumoniae* from a normally sterile site. This collection was part of an enhanced active surveillance programme coordinated by Dr Derrick Crook, John Radcliffe Hospital Department of Microbiology. Consecutive IPD cases were enrolled between June 1995 and May 2001. The mean age of the patients was 58 years, ranging from 0 to 94 years, and 50% were male. 69% of the cases were pneumonia, 15% bacteraemia,

11% meningitis and 5% other (Roy *et al.*, 2002).

2.2 Cohorts with available genome-wide genotyping data

2.2.1 WTCCC2 UK controls

(Wellcome Trust Case Control Consortium 2)

Samples from the 1958 birth cohort were genotyped genome-wide as part of the WTCCC2 project. These samples were obtained from the National Child Development Study which was based on all individuals born during one week in 1958 in England, Wales and Scotland. DNA was obtained from EBV-transformed cell lines. Genotyping was performed using the Affymetrix 6.0 chip. Half of the samples from this cohort were used as controls in the IPD GWAS performed by Dr. Anna Rautanen at the Wellcome Trust Centre for Human Genetics, University of Oxford. The other half were used the GAinS UK CAP sepsis susceptibility GWAS presented in chapter 5, and also in the GenOSept European CAP sepsis GWAS presented in chapter 6. The entire cohort contains approximately 3000 samples.

2.2.2 HYPERGENES Italian controls

Individuals were recruited from across continental Italy (excluding Sardinia) as part of a study of hypertension. All individuals self-reported being of Italian origin for more than two generations and were more than 55 years of age. Samples were quantified and normalised to 50ng/μl prior to genotyping and were genotyped using the Illumina Infinium II 1M duo BeadChips (Illumina, San Diego, CA, USA), in the Genomic Laboratory of the Genomic and Bioinformatic platform at the University of Milan. Data from 1280 individuals from this cohort were studied in this thesis.

2.2.3 EPICOLON Spanish controls

Colorectal cancer cases and controls matched by age and gender were collected in 25 centres across Spain. Controls were collected from healthy unrelated individuals. Appropriate informed consent was collected from all participants, and approval for the study was gained from the corresponding ethical committees. DNA was extracted from most samples in the Spanish National Genotyping Center, and all genotyping was performed here using the Affymetrix 6.0 chip. Data from 572 control individuals from this cohort were studied in this thesis.

2.2.4 GABRIEL German controls

(Gabriel Advanced Survey)

GABRIEL (GABRIEL, 2012; Moffatt *et al.*, 2010) was funded by the European commission and Wellcome Trust Award. German control individuals collected as part of the GABRIEL ADVANCED SURVEY were used. This is a cross-sectional population-based survey conducted in rural areas of Austria, Germany and Switzerland. Children aged 6-12 years (135,359 individuals in total) were recruited through schools. DNA was extracted from blood samples using Puregene (QIAGEN, Hilden, Germany) on an Autopure LS instrument (QIAGEN, Hilden, Germany). These samples were genotyped with the Illumina Human610 quad array, at the Commissariat à L'Energie Atomique, Institut de Génomique, Centre National de Génotypage, Evry, France. Data from 858 control individuals from this cohort were studied in this thesis.

2.2.5 EGEA French controls

(Epidemiological study of Genetics and Environment of Asthma)

EGEA is part of GABRIEL (GABRIEL, 2012; Moffatt *et al.*, 2010). All individuals recruited in EGEA were of European ancestry and were born in France. Samples

were genotyped using the Illumina Human610 quad array at the Commissariat à L'Energie Atomique, Institut de Génomique, Centre National de Génotypage, Evry, France. Data from 357 control individuals from this cohort were studied in this thesis.

2.2.6 SIALS Irish controls

(Studies in Amyotrophic Lateral Sclerosis)

Control data from the SIALS study (Cronin *et al.*, 2008) were accessed through dbGaP (accession number phs000127.v1.p1) at www.ncbi.nlm.nih.gov/sites/entrez?db=gap. Control samples were from individuals with no personal or family history of neurological disease that were accompanying patients to neurology clinics in Dublin, Ireland. All controls had self-reported Irish ancestry for at least three generations. Genotyping was provided by NINDS (National Institute of Neurological Disorders and Stroke) and was performed on the Illumina Infinium II 500K chip. Funding was provided by the Muscular Dystrophy Association, Irish Institute of Clinical Neurosciences and Intramural funding at NIH. Data from 211 control individuals from this cohort were studied in this thesis.

2.3 Laboratory methods

2.3.1 DNA extraction

DNA extraction was performed in a slightly different way for the sample sets processed in this thesis. Details for each cohort are shown below.

2.3.1.1 GRACE

Blood was collected in EDTA tubes for all GRACE individuals and stored at $-80\,^{\circ}\mathrm{C}$. Methods were tested to find the most appropriate technique to extract DNA from

these samples (see section 3.3.1). Maxwell ®16 (Promega UK Ltd) Blood DNA Purification kits were used with the Maxwell ®16 instrument (Promega UK Ltd) to extract DNA. Rachel Craik extracted approximately 1000 of these samples, while I extracted the rest (approximately 5000).

2.3.1.2 GAinS

Blood was collected in EDTA tubes, transported at room temperature and finally stored at $-80\,^{\circ}\mathrm{C}$. All samples which were not genotyped in the GenOSept GWAS were extracted using the Maxwell ®16 (Promega UK Ltd) Blood DNA Purification kits with the Maxwell ®16 instrument (Promega UK Ltd), or using the QIAamp DNA Blood Midi kit (Qiagen).

2.3.1.3 Pneumococcal Empyema

Samples were collected in Oragene DNA OG-250 kits (DNA Genotek Inc) and transported and stored at room temperature before extraction. DNA extraction was performed using the standard extraction method from Oragene (DNA Genotek Inc). More details on the method used for extraction can be found in section 3.3.2.

2.3.2 DNA quantification

All DNA collections were quantified using PicoGreen ®(Invitrogen). The QuantIT ™PicoGreen ®reagent will fluoresce in the presence of double stranded DNA, while fluorescence in the presence of RNA or single stranded DNA is minimal (Invitrogen, 2008). Therefore, by measuring the fluorescence of an aliquot of a sample following the addition of PicoGreen ®, the concentration of DNA in the sample can be quantified by comparison to a standard curve.

2.3.3 Whole genome amplification

Samples from the IPD and Pneumococcal Empyema collections were whole genome amplified using GenomiPhi (GE Healthcare). GenomiPhi uses a DNA polymerase, random hexamer primers and dNTPs with input DNA, to polymerise DNA from many random locations in the genome, based on where the primers bind non-specifically. As polymerisation continues over the course of the procedure, DNA in the sample is amplified and thus concentration of DNA is increased (GE Healthcare, 2006). The standard protocol was used, but for samples with starting concentrations lower than 10ng/μl, 3μl starting DNA was used, and 7μl sample buffer added before continuing with the standard protocol.

2.3.4 Agarose gel electrophoresis

Agarose gel electrophoresis was used to check the presence of DNA extracted from serum as in chapter 3 and for visualisation of restriction fragment length polymorphism (RFLP) results in chapter 4. Agarose gels (2%) were made using agarose powder (Sigma) and 1x Tris acetate ethylenediaminetetraacetic acid (TAE) buffer (VWR). After dissolving agarose in 600ml TAE buffer and cooling, 25μl ethidium bromide was added to the solution, mixed, and poured into a large tray. Samples were loaded with loading dye from New England Biolabs, and electrophoresis was performed at 200 volts for 30 minutes. Images were visualised under ultraviolet light.

2.3.5 Genotyping methods

2.3.5.1 *MBL2* HRMA assay methods

High resonance melting analysis (HRMA) uses fluorescent dyes which incorporate into the DNA fragments amplified by PCR. Following PCR amplification of a region of interest, the DNA is heated in a machine which can measure fluorescence as the

temperature increases. As the temperature increases and double stranded DNA separates, fluorescence is released and measured. DNA with a different genotype will separate and release fluorescence at slightly different temperatures, and therefore the pattern of fluorescence release can be used to define genotype. Probes can also be designed which bind to the amplified fragments and can enhance the difference in temperature which different genotypes of DNA separate at (Reed *et al.*, 2007; Vossen *et al.*, 2009).

Two HRMA assays were used to genotype four SNPs in the *MBL2* gene, as described by Vossen *et al.* (2010). To genotype three of these SNPs, primers amplify a 103 bp fragment spanning the three SNPs (B, C, D) in exon 1 of the *MBL2* gene. Sequences of the PCR primers for the exon were CTGCAGTGATTGCCTGTAGC (F) and GCCCAACACGTACCTGGTTC (R). The unlabelled probe was designed according to the wildtype (AA) genotype and had the following sequence: GGCAAAGATGGGCGTGATGGCACCAAGGGA. The probe had a 3'-amino-C6 modification to prevent DNA polymerase extension during PCR. Primer sequences for amplifying the Y/X polymorphism in the *MBL2* promoter region were TTCCCTAAGCTAACAGGCATAA (F) and GCACTATGATGAGCAGTGGG (R). The probe was designed to match the Y-variant and had the following sequence: CCACGGAAAGCATGTTTATAGTCTT. The probe was 3'-modified with a phosphate group to prevent extension during PCR (Vossen *et al.*, 2010). Control samples of known genotype (confirmed by sequencing) were included on every plate for exon genotyping. Genotypes were assigned by 96-well plate. All sepsis samples (survivors and non-survivors) were mixed on the same plates when genotypes were assigned. PCR primers and unlabelled probes were manufactured by Metabion.

Amplification was performed in a 10μl reaction volume containing 1μl 10x PCR buffer without $MgCl_2$ (Roche Diagnostics), 0.8μl 25mM $MgCl_2$ (Roche Diagnostics), 1μl dNTP mix (BioLine; 2mM of each dNTP), 0.05μl forward primer (10μM), 0.5μl

reverse primer (10μM), 0.5μl unlabelled probe (10μM), 1μl 10x LC Green Plus (Idaho Technology), 0.1μl Fast Start Taq DNA Polymerase (Roche Diagnostics), 3.05μl water and 2μl 10ng/μl DNA template.

PCR was performed in a 96 well plate (FrameStar 96, skirted PCR plate, white wells, black frame; 4titude, UK) and reactions were overlaid with mineral oil (Sigma). Amplification was performed with an initial denaturation of 95 °C for 10 minutes followed by 55 cycles of 20 seconds at 95 °C, 30 seconds at 60 °C and 40 seconds at 72 °C. A final extension step of 5 minutes at 72 °C was performed followed by 1 minute at 95 °C. Finally, the samples were cooled to room temperature.

High resolution melting was performed using a LightScanner (HR-96, Idaho Technology). Samples were melted from 55 °C to 98 °C at 0.1 °C/second. Melting curves were analysed with Lightscanner Software using Call-IT 1.5. Probe melting data were displayed as derivative melting plots in unlabelled probe scanning mode of the Lightscanner software and automatic grouping of the different genotypes was performed. Amplicon melting analysis was performed to complement the probe melting data and the data were displayed as difference plots of temperature-overlaid normalized melting curves (Wittwer *et al.*, 2003). The baseline (standard) in amplicon scanning mode was set at variant AA, the genotype with the highest T_m.

2.3.5.2 Sequencing of *MBL2* exon and promoter polymorphisms

Primers for exon polymorphism sequencing were GCACCCAGATTGTAGGACAGAG (F) and CAGGCAGTTTCCTCTGGAAGG (R) (Chapman *et al.*, 2010a). PCR conditions were: 94 °C for 14 minutes, 38 cycles of 94 °C for 30 seconds, 58 °C for 30 seconds and 72 °C for 30 seconds, then 72 °C for 5 minutes. Primers for promoter sequencing were AGCTGGGGATTTGGAAAAGT (F) and GGATCACCAGCTTTCAGCTC (R). PCR conditions were: 95 °C for 15 minutes, 43 cycles of 94 °C for 30 seconds, 63 °C for 30 seconds and 72 °C for 1 minute, then 72 °C for 10 minutes.

All other conditions in the sequencing protocol were the same for both promoter and exon sequencing:

PCR was performed in a 10μl reaction with final concentrations of 1.2x PCR buffer (Qiagen), 1.2mM $MgCl_2$ (Qiagen), 240μM each dNTP (BioLine), 240μM of each primer (Metabion), 0.3 units HotStar Taq DNA Polymerase (Qiagen) and 20ng DNA template. 6μl PCR product was run on a 2% (w/v) ethidium bromide agarose gel for 1 hour to check that PCR was successful.

ExoSAP was used to clean PCR products, which were then subject to sequencing reactions. 4μl PCR product was treated with 1 unit of Shrimp Alkaline phosphatase (Promega) and 2 units of Exonuclease I (New England Biolabs) in a final volume of 8μl with 0.5x Shrimp alkaline phosphatase buffer (Promega),

Purified PCR product (2μl) was used in a 10μl reaction containing 0.25μl Big Dye Terminator Sequencing Mix v3.1 (ABI), 0.94x Big Dye Dilution Buffer, 0.5μM primer (forward or reverse depending on which direction the reaction is for). Primers were the same as those used for PCR. Sequencing conditions were as follows: 25 cycles of 96 °C for 10 seconds, 50 °C for 5 seconds and 60 °C for 4 minutes.

Ethanol precipitation was used to clean products, which were then run on an ABI 3730xl capillary sequencer, in the Zoology department at the University of Oxford. Trace and sequence analysis was performed using Sequencher 4.2 software. Traces were checked for any unreliable calls, and forward and reverse strands were aligned. The genotype of the SNPs was manually checked on chromatograms.

2.3.5.3 *IFITM3* RFLP

Genotyping of rs12252 was performed by restriction fragment length polymorphism (RFLP) using primers CAGGAAAAGGAAACTGTTGAGAACC (F) and CTCCTG-GAGCCTCCTCCTCA (R). Primers had 3' penultimate base mismatches to *IFITM3* to ensure specific amplication of *IFITM3* rather than *IFITM2*. PCR was performed

in a 6μl reaction containing 0.6μl 10x PCR buffer (Qiagen), 0.6μl dNTP mix (Bio-Line; 2mM of each dNTP), 0.12μl of each primer (10μM; Metabion), 0.03μl HotStar Taq DNA Polymerase (Qiagen), 3.73μl water and 0.8μl 25ng/μl DNA template. PCR conditions were: 95 °C for 15 minutes, 45 cycles of 94 °C for 30 seconds, 62 °C for 30 seconds and 72 °C for 1 minute, then a final step of 72 °C for 10 minutes. PCR products were incubated with MScI at 37 °C for 4 hours, which cuts the PCR product in the presence of the wild type T allele. Fragments of length 572, 426 and 146bp were visualised on agarose gels, according to the genotype of the sample. All minor homozygotes and heterozygotes were Sanger sequenced in the reverse direction to confirm their genotype, using the same IFITM3 RFLP primers and the same methods used for *MBL2* sequencing (see section 2.3.5.2). Analysis was performed using PLINK v1.07 (Purcell, 2012; Purcell *et al.*, 2007) and SPSS v18 (PASW Statistics 18).

2.3.5.4 Sequenom iPLEX genotyping

The Sequenom iPLEX system makes use of matrix-assisted laser desorption/ionisation, time-of-flight (MALDI-TOF) mass spectrometry to accurately genotype up to 40 SNPs at a time in hundreds to thousands of samples. Carefully designed primers amplify approximately 100bp around all SNPs in an assay in a multiplex PCR reaction. These PCR products are treated to remove unused nucleotides, and then an extension reaction is performed. Extension primers are designed to end at the base before the SNP site. Extension is performed with mass-modified terminator nucleotides so that only one base is added to the extension primer in the reaction. Products are cleaned again, dispensed onto a chip, and then analysed using a MassARRAY Analyzer instrument. This measures time-of-flight, which depends on the mass of the extended primer, and can therefore demonstrate differences in mass between primers extended with different mass-modified terminator nucleotides, which are visualised

with spectra for each sample (Oeth *et al.*, 2007, 2009; Sequenom, Inc., 2010a).

Results were visualised using the MassARRAY Typer 4.0. Results were clustered automatically using the Typer software, after input of expected mass peaks (Sequenom, Inc., 2010b). All cluster plots for all plates were visually inspected and sample genotypes were changed appropriately, in a conservative manner, to ensure that all called genotypes were reliable. Data were exported for quality control analysis using PLINK (Purcell, 2012; Purcell *et al.*, 2007).

2.3.5.5 *MPHOSPH8* region fine-mapping in GRACE

Quality control was performed using PLINK (Purcell, 2012; Purcell *et al.*, 2007), for the Sequenom genotyping data from efforts to fine-map the *MPHOSPH8* gene region. Two SNPs were removed for having a call rate less than 80%. An additional three SNPs were removed because the clustering from genotyping was very unclear. Genotyping was performed in two separate assays- half the SNPs were genotyped in one assay, and half in the other. Call rate thresholds for sample exclusion were slightly different for the two assays. The first assay used a call rate threshold of 80%, which removed 16 samples from the analysis. The second assay used a call rate threshold of 70%, which removed 37 samples. This gave similar genotyping rates in the remaining samples for both assays; 0.974 and 0.978 in assays one and two respectively. Two further SNPs were removed with HWE p-values $<1\text{x}10^{-5}$ in controls. Both of these SNPs showed p-values of 0.01.

All samples passing QC were used for imputation in the region, using IMPUTE2.2 (Howie *et al.*, 2011, 2009). The 1,000 Genomes Phase I interim release in NCBI build 37 (hg19) coordinates was used as a reference panel.

2.3.5.6 GWAS genotyping

Many of the sample cohorts described above have been genotyped in GWAS.

The genome-wide chip used for each cohort is shown below:

GenOSept: Affymetrix 5.0

IPD: Affymetrix 6.0

Kenyan bacteraemia: Affymetrix 6.0

GAinS 2011: Illumina Omni Express

WTCCC2: Affymetrix 6.0

EPICOLON: Affymetrix 6.0

GABRIEL: Illumina 610 Quad

SIALS: Illumina Infinium II 550K

HYPERGENES: Illumina 1M

2.4 Analytical methods

2.4.1 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium (HWE) was almost simultaneously first described in 1908 by both G.H. Hardy and W. Weinberg. This equilibrium describes the constant proportions of genotype frequencies in successive generations of a population. The equilibrium stands as long as the population is large, mates at random, and is under no selection or migration pressure (Crow, 1988; Hardy, 1908; Stern, 1943; Wigginton *et al.*, 2005).

Equation 2.1 shows the equilibrium, where p is the allele frequency of the major allele, and q is the allele frequency of the minor allele. Therefore the expected genotype frequencies of major homozygotes are defined by p^2, heterozygotes by 2pq and

minor homozygotes by q^2. If this equilibrium deviates from 1, the population is not in equilibrium.

$$p^2 + 2pq + q^2 = 1 \tag{2.1}$$

Deviations from equilibrium can be due to small sampling numbers, non-random mating, selection or migration pressure in the sample, and can be an indication of poor genotyping quality (Hosking *et al.*, 2004). An exact test should be used to calculate HWE, due to the inflated type I error rates associated with using a simple χ^2 goodness-of-fit test (Wigginton *et al.*, 2005). An exact test is used to calculate HWE in the PLINK and SNPTEST softwares. An HWE threshold of $<1x10^{-5}$ in controls was used in this thesis, except for in GWAS analysis, where a threshold of $<1x10^{-10}$ was used.

2.4.2 Linkage disequilibrium measures

There are two widely used measures of LD; D' and r^2. LD between any pair of SNPs can be defined with either of these measures as a value between 0 and 1, where 1 is perfect LD. r^2 is a much stricter measure and will only be observed as 1 if the allele frequencies of both SNPs are exactly the same. It is a two-way measure which looks at how often a particular allele in SNP one is found with a particular allele in SNP two, and how often the same allele is found in SNP two when the same allele is found in SNP one. Conversely, D' will reach 1 for instance when one rare allele is only found with a particular allele in another SNP, even though the opposite will not be true.

2.4.3 Power calculations

The power to detect genetic association was calculated using a non-centrality parameter in a non-central χ^2 distribution. The non-centrality parameter (λ) is a function

of the sample size, the allele frequency in controls, the allele frequency in cases (if known) and the LD between the marker of interest and the disease-causing variant, and is given by equation 2.2.

$$\lambda = \frac{2r^2 N_{case} N_{ctrl} (F_{case} - F_{ctrl})^2}{NF(1-F)} \tag{2.2}$$

where N is the number of cases and controls, F if the frequency of the risk allele, and r^2 is the LD between the marker of interest and the disease-causing variant. Where the frequency of the risk allele in cases is unknown, this is given by equation 2.3, where the estimated effect size (odds ratio) is used.

$$F_{case} = \frac{F_{ctrl} OR}{1 - F_{ctrl} + F_{ctrl} OR} \tag{2.3}$$

All power calculations in this thesis were performed using an R script written by Damjan Vukcevic (Clinical Epidemiology and Biostatistics Unit, The Royal Children's Hospital Melbourne, Australia) which follows these formulae.

2.4.4 Statistical tests for association

2.4.4.1 Pearson's χ^2 test

Pearson's χ^2 test compares the distribution of observed values with that of expected values under the χ^2 distribution (Pearson, 1900). In a case control analysis, genotype data are applied to a contingency table, where case data are found in one row, and control data in the other row. Based on the totals of rows and columns in this table, the expected frequency in each cell of the table can be calculated, and the deviation from this expectation can be tested for significance. Pearson's χ^2 test was used for matched analysis in GRACE in chapters 4 and 5, and for the analysis of the GAinS UK CAP sepsis GWAS in chapter 5.

2.4.4.2 Fisher's exact test

Fisher's exact test was devised by Fisher to improve upon Pearson's χ^2 test. Pearson's test is inaccurate when cells of a contingency table contain small numbers of observations (it is used for all instances where the observation in at least one cell of the table is less than 5). Fisher's exact test exactly calculates the distribution of the data and can therefore more accurately determine the likelihood of observing this distribution by chance (Fisher, 1922). Fisher's test is much more computationally heavy than Pearson's χ^2 test, and is therefore not practical to compute for larger numbers of observations. The difference in accuracy between the two tests is greatly reduced for larger numbers of observations. Fisher's exact test was used for all *IFITM3* H1N1 sepsis analyses in chapter 4.

2.4.4.3 Logistic regression

Logistic regression works to fit data to a regression line, and then tests how well the data fit to this line (for the significance of the association), and the slope of the line (for the odds ratio). Logistic regression can be binomial (for 2 categories to analyse-as in the allelic test) or multinomial (with 3 or more categories to analyse-as in the genotypic test, see section 2.4.4.5). Covariates can also be used in logistic regression, which will be corrected for when forming the regression line through the data. Logistic regression was used for all analyses where covariates were included: all GRACE analyses except the matched analysis (for *MBL2* and *IFITM3* in chapter 4 and the SNPs genotyped by iPLEX in chapter 5) and all MBL sepsis analyses in chapter 4.

2.4.4.4 SNPTEST methods

SNPTEST is a software which can be used to analyse imputed genotype data, in order to take into account the uncertainty associated with each genotype, for in-

formative analysis results. Statistical theory for missing data looks at the observed data and missing data (and thereby the uncertainty) for each genotype. Tests are then performed to calculate the score data and an information matrix for each SNP. This score data and information matrix are required to perform the score test on the data, to determine the statistical significance of an association. Information measures ("info") are also calculated for each SNP. These are measures of the reliability of the information for each SNP, and values are between 0 and 1, where 1 is the most reliable measure and will be assigned to directly genotyped SNPs. SNPTEST was used to perform all SNP association analyses in chapter 6 (Marchini *et al.*, 2010), and the *MPHOSPH8* imputed analyses in chapter 5.

2.4.4.5 Models of association

Genotype data can be analysed for association using a number of different models. Each model analyses the differences between genotypes in cases and controls in a different way. Where AA denotes the major homozygote, Aa denotes the heterozygote and aa denotes the minor homozygote, the models are defined as follows:

Allelic test: Compares all A alleles to all a alleles

Genotypic test: Compares all 3 genotypes; AA to Aa to aa

Dominant test: Compares Aa and aa to AA

Recessive test: Compares aa to Aa and AA

Heterozygous test: Compares Aa to AA and aa

In chapter 4, MBL analysis was performed for each genotype in every subgroup for all of the tests above and IFITM3 analysis was performed for all tests except the heterozygous test (any difference in the heterozygous test will be reflected in the genotypic test). In GRACE chapter 5, the allelic, genotypic, dominant and recessive

tests were performed for all major subgroups. In chapter 6, all tests were performed, but only the allelic results have currently been visualised.

2.4.5 Identity by state and descent

Unknown relatedness in a sample set can create bias in genetic association studies. However, using genome-wide data, it is possible to use analytical methods to calculate relatedness between samples and detect closely related samples. Methods in PLINK can be used to estimate the underlying identity by descent (IBD) state between pairs of individuals given the observed identity by state (IBS) sharing and genome-wide level of relatedness between them (Purcell, 2012).

Equation 2.4 shows the calculation for $\hat{\pi}$ (pi hat). This denotes the proportion of alleles shared IBD, and therefore denotes relatedness between pairs of samples. The expected probability of IBD states given IBS states is specified in terms of allele frequency and averaged over all SNPs, and can then be used to calculate P(Z=1) and P(Z=2). These are the probabilities of two individuals sharing one or two alleles respectively (Purcell *et al.*, 2007).

$$\hat{\pi} = \frac{P(Z = 1)}{2} + P(Z = 2) \tag{2.4}$$

A $\hat{\pi}$ score of 1 denotes that all alleles are shared IBD, and therefore represents duplicate samples. A score of 0.5 denotes that half of the alleles are shared IBD, and therefore denotes siblings (Anderson *et al.*, 2010). A $\hat{\pi}$ score threshold of 0.2 was used in this thesis to remove related individuals in GWAS analysis.

2.4.6 Multi-dimensional scaling

Following calculations of relatedness between samples, these data can be used for further analysis to determine population stratification within a sample-set. This can

be performed by two main methods: principal component analysis (PCA) and multi-dimensional scaling (MDS). Both of these methods give similar results (Nelis *et al.*, 2009). Using PLINK, MDS was performed for the datasets in this thesis (Purcell, 2012; Purcell *et al.*, 2007).

Population stratification is an important confounder which must be accounted for in association studies. Without correction for this confounder, false positive results can be generated because of differences between cases and controls due to differences in ancestry, rather than differences in disease risk (Campbell *et al.*, 2005; Cardon and Palmer, 2003). Studies have been performed to analyse population structure within Europe, and fine-scale population stratification within individual populations in Europe. This analysis can be used to create genetic maps, infer population history and ancestry and aid the design of future genome-wide studies (Di Gaetano *et al.*, 2012; Nelis *et al.*, 2009; Salmela *et al.*, 2011).

MDS is used to visualise population substructure and to provide quantitative measures of population stratification which can be used as covariates in downstream analyses, as well as to identify individuals with different population backgrounds to the majority of the sample-set (outliers) that need to be excluded from analysis. MDS is a statistical method to quantify similarities and differences in data. Indices which represent these similarities and differences are defined in multiple dimensions, such that more similar pairs of samples have more similar values and less similar pairs of samples have less similar values (Purcell *et al.*, 2007; Young, 1985). PLINK uses MDS to produce a dimensional representation (the number of dimensions can be specified) of substructure in a dataset based on the pairwise comparisons of the IBS analysis (Purcell *et al.*, 2007). All MDS plots in this thesis were made using SPSS v18 (PASW Statistics 18).

2.4.7 GWAS QC

The GenOSept, IPD and Kenyan bacteraemia GWAS datasets underwent quality control and were analysed by Dr. Anna Rautanen, Wellcome Trust Centre for Human Genetics, University of Oxford. The quality control, processing and analysis of all data sets used in the CAP sepsis susceptibility GWAS described in chapter 6 are presented below, and in tables 2.1 and 2.2. Quality control for the GenOSept dataset for use in the *MBL2* association analysis was also performed as described below.

All sample sets were processed to put them in the same format before QC. All steps did not need to be performed for all datasets, but all datasets were processed so that:

1. All SNP locations matched to NCBI build 37 (hg19) coordinates.
2. All SNPs were named with rs numbers (rather than chip codes).
3. All alleles were coded as A,C,G,T and 0 (missing genotype).
4. All phenotypes were coded as 1 for controls, 2 for cases and -9 for missing.
5. Gender was coded as 1 for male, 2 for female and -9 for missing.
6. All SNPs were flipped to the positive strand after QC.

Quality control (QC) for SNPs and samples was performed in the same way for all sample sets and in the following order:

1. SNPs with clustersep value <0.3 were removed where information was available for cohorts genotyped with Illumina chips.
2. SNPs which could not be matched to NCBI build 37 (hg19) coordinates using LiftOver (University of California, Santa Cruz, 2012a) were removed.
3. Samples with call rate <98% were removed.

4. Samples were removed when computed gender did not match reported gender.

5. SNPs with call rate <98% were removed.

6. SNPs with HWE $<1x10^{-10}$ were removed.

7. SNPs with minor allele frequency (MAF) <1% were removed.

8. SNPs with MAF between 1% and 5% and with call rate <99% were removed.

9. Related samples (with $\hat{\pi}$ score >0.2) were removed.

10. In a window of 50 SNPs, one SNP of each pair was removed if LD between the SNPs was greater than 0.2, to create a list of SNPs not in LD with each other. A new window was analysed every five SNPs along the genome.

11. Using only SNPs which were not in LD with each other, MDS components were calculated.

12. The following MDS plots were made: component (C) 1 vs. C2, C1 vs. C3, C2 vs. C3 and C3 vs. C4. Any outliers were removed.

13. Heterozygosity was plotted against missingness and any samples more than 3 standard deviations away from the mean for heterozygosity in either direction were removed.

2.4.8 GWAS processing and analysis procedures

The LiftOver tool available from UCSC (University of California, Santa Cruz, 2012a) was used to convert SNP locations from NCBI build 36 coordinates to NCBI build 37.

The SHAPEIT program was used (Delaneau *et al.*, 2012) to phase all GWAS data before imputation. An effective population size of 11418 (which is equivalent to the

Table 2.1: The number of samples removed from different sample sets for different QC reasons. Although the GenOSept sample set used the same parameters for QC, all parameters were assessed individually rather than sequentially as in the other sample sets. This means that although it is known how many samples were removed for each reason, many of these samples overlap. This is shown by brackets in the table. Het: samples removed for heterozygosity. More information about QC thresholds is given in section 2.4.7.

Name of dataset	**Number of samples**	**Sample QC removal reason**					**Whole set MDS sample removal**	**Duplicates removed**	**Samples remaining**
		Gender	Call rate	Relateds	MDS	Het			
GenOSept	1689	(49)	(52)	(17)	(49)	(58)	66	0	1517
GAinS 2011	288	9	7	1	1	0	1	29	240
WTCCC2 UK	1390	0	0	2	0	92	5	1	1290
EPICOLON Spain	572	37	7	0	0	9	31	0	488
GABRIEL France	357	1	0	0	1	2	2	0	351
GABRIEL Germany	858	5	31	9	0	3	0	0	810
SIALS Ireland	211	0	0	0	0	2	0	0	209
HYPERGENES Italy	1280	0	51	0	0	18	13	0	1198

Table 2.2: The number of SNPs removed from different sample sets for different QC reasons. 86,311 SNPs were removed in GenOSept for call rate, HWE, and MAF. *SNPs with a MAF >5% were filtered for a call rate threshold of 98%, while SNPs with a MAF between 1% and 5% were filtered for a call rate threshold of 99%. HWE: Hardy-Weinberg equilibrium. MAF: minor allele frequency. HWE refers to SNPs removed for a HWE p-value < $1x10^{-10}$. MAF refers to SNPs removed for call rate <1%.

Name of dataset	**Number of SNPs before QC**	**SNP QC removal reason**							**Number of SNPs after QC**
		liftOver	Cluster sep	Call rate <98%*	HWE	MAF	Call rate < 99%*	No strand information	
GenOSept	440,794	8,005	-			86,311		-	346,478
GAinS 2011	730,525	-	1,643	3,754	380	79,281	692	-	644,775
WTCCC2 UK	900,120	1,545	-	3,372	29,846	132,470	73	-	732,814
EPICOLON Spain	905,525	-	-	76,412	241	125,496	4,354	-	699,022
GABRIEL France	582,892	15,438	-	5,514	0	30,534	441	-	530,964
GABRIEL Germany	582,892	15,438	-	10,931	76	29,639	700	-	526,107
SIALS Ireland	561,466	14,134	16,037	2,016	5	23,754	83	2,768	502,669
HYPERGENES Italy	1,111,215	240	-	33,217	377	163,768	1,769	8,059	903,783

effective size of the HapMap phase II population) was used, and the genetic map file was from the 1,000 Genomes Phase I integrated variant set release (v1) in NCBI build 37 (hg19) coordinates. All other settings were performed as default (states-phase 100, burn 10, prune 10, main 50).

IMPUTE2.2 (Howie *et al.*, 2011, 2009) was used to impute polymorphisms for all datasets. 1,000 Genomes Phase I integrated variant set release (v1) in NCBI build 37 (hg19) coordinates was used as a reference panel. Options of buffer 250 and Ne 20000 were used. Accurate strand information could not be obtained for the Italian HYPERGENES dataset. Therefore, the align_by_maf_g option was specified during imputation, to flip the strand of all SNPs which IMPUTE2 deemed to be on the negative strand. This compares the alleles in the genotyped data to those in the reference data, and flips the strand of the SNP if the alleles do not match. In addition, it uses the allele frequency of the genotyped and reference data to determine the strand of A/T and C/G SNPs.

SNPTEST version 2.4.0 was used for the analysis of all imputed data. The score method was used, and frequentist tests 1-5.

Lambda was calculated with the R package GenABEL.

Chapter 3

Modern DNA extraction techniques

3.1 Introduction

Poor quality DNA is a recognised problem leading to poor genotyping reliability (Lahermo *et al.*, 2006), and genotyping errors are likely to lead to spurious associations. Even very slight differences between samples can lead to differences in genotyping (Clayton *et al.*, 2005; The Wellcome Trust Case Control Consortium, 2007). Therefore, samples of poor quality must be removed from genetic analysis. These can be identified in ways such as lower genotyping call rates and deviating heterozygosity rates in genome wide association studies (Anderson *et al.*, 2010). Contamination of DNA samples can lead to an increased number of heterozygous genotypes, and can even be a cause of spurious associations.

In order to provide enough statistical power to detect associations with small effects, large sample collections are needed, particularly for GWAS (The Wellcome Trust Case Control Consortium, 2007; Zondervan and Cardon, 2004). This necessity is becoming increasingly apparent for whole genome and whole exome sequencing

too. Large sample sizes are also required for candidate gene studies in order to more reliably settle controversial susceptibility associations and accurately measure allele frequencies within different populations or disease groups. The collection, management and storage of thousands of high quality DNA samples is extremely laborious, requiring patience and skill to carefully avoid contamination or degradation of DNA that can lead to poor results in genotyping.

One of the most important steps in sample processing is DNA extraction, since a good quality of resulting DNA is necessary for all genetic analyses. Fairly high quantities of DNA are also required for new genetic analyses including whole genome sequencing. Comparisons of extraction techniques to find the most cost and labour effective techniques to extract good quality DNA from blood, saliva and serum are presented here.

3.2 Aim

The aim of the work in this chapter was to find the modern DNA extraction techniques which lead to the highest yields and concentrations of high quality DNA for genotyping from three different sample types (blood, saliva and serum).

3.3 Results

3.3.1 DNA extraction from blood

DNA extractions were performed from two different collections of blood samples: the GRACE European LRTI collection and the GAinS UK sepsis collection. These samples are described in more detail in Materials and methods section 2.1. Table 3.1 describes factors associated with the different methods used for both sample collections. Manual and automated methods were compared.

Table 3.1: Comparison of different methods for extracting DNA from blood in two sample collections. SEV: standard elution volume. Approximately 1000 of the GRACE samples were extracted by Rachel Craik, Wellcome Trust Centre for Human Genetics, University of Oxford. The rest were extracted by the author of this thesis.

Starting material	**DNA extraction method**	**Cost per sample (£)**	**Cost of robot (£)**	**Time required per sample (mins)**	**Volume starting material required**	**Volume resulting DNA (μl)**	**n**	**Mean concentration DNA (ng/μl)**	**Total average yield (μg)**
GRACE Whole Blood	QIAamp mini kit	2	None	12	200μl	200	175	36	7
GRACE Whole Blood	QIAcube robot	2	9,200	10	200μl	100	180	40	4
GRACE Whole Blood	Maxwell 16 robot Blood SEV kit	2	9,500	4	400μl	260	5602	46	12
GAinS Whole Blood	Maxwell 16 robot Blood SEV kit	2	9,500	4	400μl	260	184	34	8.8
GAinS Whole Blood	QIAamp midi kit	3.77	None	10	2ml	300	408	264	79

DNA requirements for the two sample collections were different. For the GRACE collection, extracted DNA needed to be of a minimum concentration of 25ng/μl (the required concentration for genotyping using the Sequenom iPLEX system). The yield needed to be as high as possible, but the technique needed to be quick to perform due to the large sample collection and time constraints for the project. For the GAinS collection, DNA needed to be of a minimum concentration of 50ng/μl (the required concentration for whole exome sequencing), with as high a yield as possible.

Therefore, different methods were compared for the different sample collections. Two Qiagen DNA extraction kits, the QIAamp mini and midi, were tested. In addition, the QIAcube was tested. This is the Qiagen automation method which uses the QIAamp mini kit reagents and components in a robot which performs the extractions for 12 samples at a time. Finally, the Maxwell 16 robot from Promega was tested, using the Blood SEV (standard elution volume) kit.

The Qiagen and Promega kits use different methods to extract DNA. The Qiagen kits use a column which will bind DNA alone at the pH of the lysis and wash buffers, and release DNA at the pH of the elution buffer or water. This allows the DNA to be caught on the column, washed with buffer and then eluted at a lower pH. The Promega kit uses specially designed MagneSil® Paramagnetic Particles (PMPs) which will bind negatively charged DNA and can then bind a magnet to be transferred between wash steps. This allows the DNA to be separated from all other components of the sample, and effectively cleaned through repeated wash steps and separation from the remainder of the sample (Dan Kephart and Shenoi, 2006).

The Maxwell method obtained lower average yields from the GAinS whole blood than from the GRACE whole blood (8.8μg compared to 12μg). The QIAamp midi kit obtained by far the highest concentrations and yields from blood (264ng/μl mean concentration and 79μg mean yield). The QIAcube and QIAamp mini kit obtained very similar DNA concentrations to each other and to that obtained using the Maxwell,

but yields were lower, and the time required to process each sample was higher.

3.3.2 DNA extraction from saliva

The Hill group has collected saliva samples from children with empyema across the UK, and this collection is ongoing. Most of these children will have empyema caused by *Streptococcus pneumoniae*. This collection will be used for further replication of findings presented in this thesis. Results from tests to determine the best methods for extracting the highest possible concentrations and yields of DNA from these saliva samples are presented here.

The standard manual method for DNA extraction from Oragene saliva collection kits (DNA Genotek) was compared to automated methods using the Maxwell 16 robot, with the SEV blood and tissue kits. Figure 3.1 shows a comparison of DNA yields obtained when comparing these methods. Similar yields are obtained for all three methods when the amount of DNA in the saliva sample is low. However, when the amount of DNA in the saliva is higher, the manual method can extract higher yields. The Maxwell LEV (low elution volume) kit was also tested using the Maxwell 16 robot for extractions from saliva. However, although both the cell and blood kits were tested, high quality DNA could not be extracted from saliva using this LEV Maxwell method.

When performing manual DNA extractions from saliva, there are further variations in the method which have been tested in order to try and improve the yields and concentrations obtained from samples. The standard extraction procedure was performed with varying elution volumes and starting volumes of saliva.

Table 3.2 shows the difference in yields and concentrations obtained when DNA is extracted from saliva aliquots of different volumes. A starting volume of 500μl leads to by far the highest average yield of DNA per μl of saliva, and has a similar percentage of samples with the concentration of resulting DNA <5ng/μl to the method using the

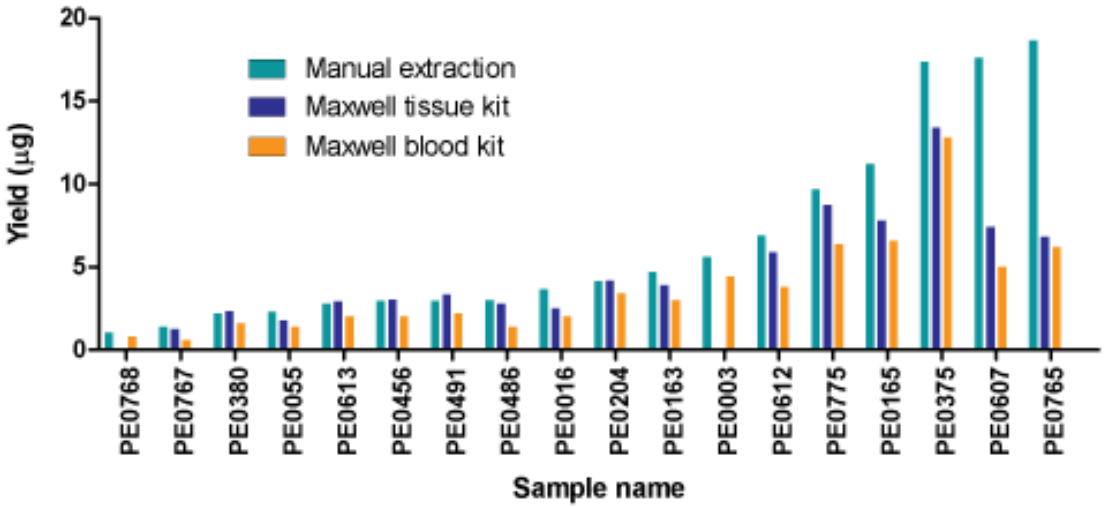


Figure 3.1: DNA yield comparison of manual and Maxwell 16 automated methods for DNA extraction from saliva

whole sample and eluting in 500μl.

Further tests were then performed to determine the best elution volume to use when extracting from 500μl saliva. Fifty-eight samples were processed from 500μl saliva in duplicate, with differing elution volumes. One version of the sample was eluted in 50μl and the other in 100μl. Figures 3.2 and 3.3 show the differing concentrations and yields obtained from the samples with different elution volumes. DNA concentrations in the 50μl elution samples were at least twice as high as in the 100μl elution samples in 65% of cases. A further 23% of samples had concentrations at least 1.5 times higher in the 50μl elution than the 100μl elution.

3.3.3 DNA extraction from serum

A number of collaborators with the Hill group have large serum collections available which were collected as part of other studies but never used. With the correct ethics regulations in place, it would be possible to use these samples for genetics studies,

Table 3.2: Comparison of methods for extracting DNA from saliva with different starting volumes and elution volumes. The whole sample average volume was 1407μl.

Volume starting saliva	**Volume eluted in**	**Fraction of samples with concentration of resulting DNA <5ng/μl**	**Average yield (μg)**	**Average yield per μl saliva extracted from (ng)**
500μl	100μl	135/381 (35%)	0.53	5.32
Whole sample	200μl	20/40 (50%)	0.53	2.63
Whole sample	500μl	12/40 (30%)	1.94	3.88

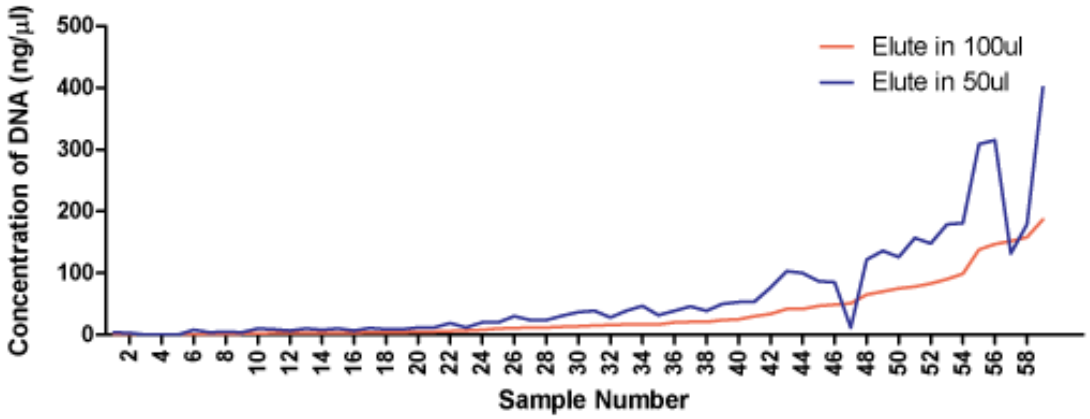


Figure 3.2: Comparison of elution volumes for DNA extraction from saliva in terms of resulting concentration

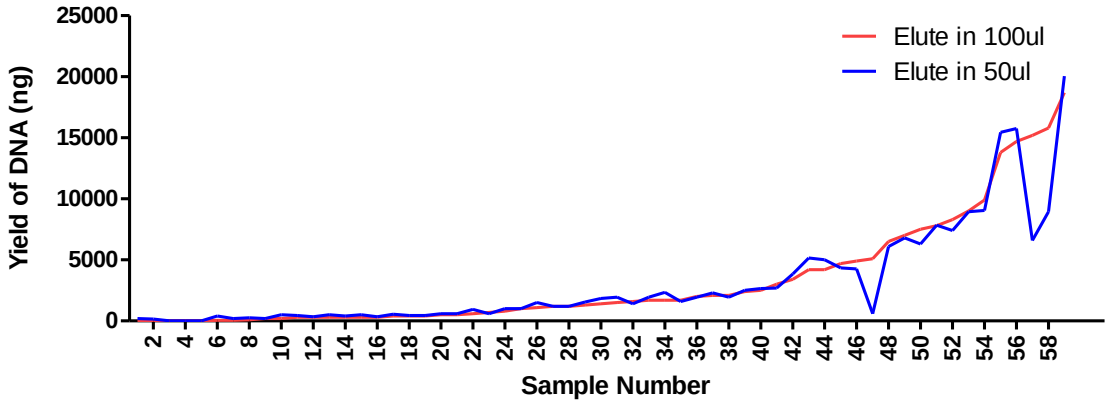


Figure 3.3: Comparison of elution volumes for DNA extraction from saliva in terms of resulting yield

as long as it is possible to extract DNA from the samples. Therefore, a number of methods were tested to determine if and thereby how DNA can be extracted from serum for modern genetics studies.

Table 3.3 shows a comparison of four different techniques for extracting DNA from serum samples. Two of these techniques (using the Qiagen circulating nucleic acid kit) involved using the same kit but with serum that had been freshly separated from blood compared to serum that had been stored at $-80\,^{\circ}\mathrm{C}$ for more than two years. These techniques were able to extract 20-30ng of DNA per ml sera. Of the other two techniques, one used a different commercially available kit designed specifically for isolating DNA from serum (Invitrogen kit), and the other was a modified technique from a paper comparing DNA extraction techniques from serum (Fong *et al.*, 2009). These techniques were tested using old sera too. The results from an RFLP assay for the genomic DNA and whole genome amplified (using GenomiPhi) DNA obtained from these techniques are shown in figure 3.4.

The Maxwell LEV (low elution volume) kit was also tested in the Maxwell 16 robot for extractions from serum. Despite attempts to optimise this procedure for the extraction of DNA from serum (including using different pre-extraction steps,

Table 3.3: Comparison of different methods for extracting DNA from serum. Conc: concentration. RFLP: restriction fragment length polymorphism. WGA: whole genome amplified.

Extraction method	**Number of samples processed**	**Average volume sera used**	**Elution volume (μl)**	**Average conc (ng/μl)**	**Genomic RFLP result**	**Average Genomiphi conc (ng/μl)**	**WGA RFLP result**
Qiagen circulating nucleic acid-new sera	4	1.7ml	30	1	strong band in all samples	752	strong band in 3 samples
Qiagen circulating nucleic acid-old sera	4	393μl	30	0.43	strong band in all samples	143	strong band in 2 samples
Invitrogen kit	4	513μl	50	0	very faint band in 3 samples	76	strong band in 1 sample, faint band in 2
NaI method	4	355μl	50	1	no band in any sample	0	no band in any sample

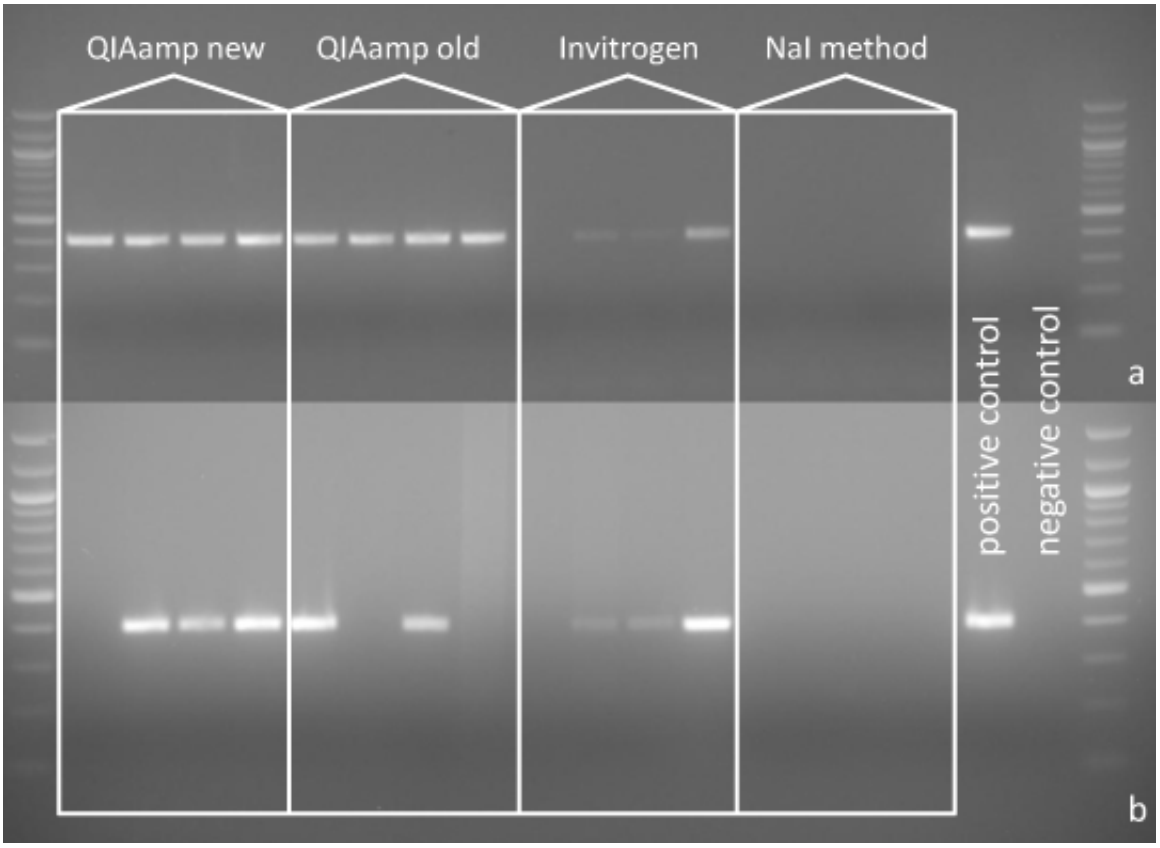


Figure 3.4: RFLP results run on an agarose gel to asses DNA quality before (a) and after (b) whole genome amplification. The four samples were tested from each technique. Positive and negative controls were run at the same time.

different cartridges in the Maxwell and different programs in the robot), no DNA could be obtained from serum samples using this method.

3.4 Discussion

Blood is the traditional material used for collection of samples for DNA analysis, and with good reason. The results presented here show that this material is the one which most consistently leads to good amounts of DNA after extraction; concentrations obtained from DNA extracted from saliva are often less than 5ng/μl and yields of DNA from serum are consistently low. However, there are safety issues surrounding the collection and handling of blood, and it is often a lot more difficult to obtain consent for a collection of blood than saliva or buccal swabs, especially when the collection is from young children. In addition, blood needs to be properly stored upon collection to ensure the extraction of good yields of DNA, as demonstrated here.

3.4.1 DNA extraction from blood

For the GRACE project DNA extractions, it is clear that the Maxwell 16 robot with the SEV blood kit was the best extraction method to use. This method gave a higher mean concentration and yield for GRACE blood samples than the other methods tested, and the mean concentration was well above the 25ng/μl concentration required. Although the QIAamp mini kit and QIAcube robot required smaller volumes of blood for each extraction, performing two extractions would not have improved the concentration obtained, and would only lead to a marginally improved yield compared to the Maxwell method. In addition, it took far less time to extract DNA using the Maxwell 16 robot than the QIAamp mini kit and QIAcube robot.

For the GAinS project, two methods were tested. The QIAamp midi kit was

almost twice as expensive as the Maxwell 16 robot with the SEV blood kit, used a lot more starting material, and took more time per extraction. However, the mean concentration and yield obtained with the QIAamp midi kit was much higher. Since the minimum concentration required for this sample set was a lot higher than the mean obtained using the Maxwell 16 robot with the SEV blood kit for this collection, the QIAamp midi kit is clearly a better choice for this purpose.

In the GRACE project all blood samples were immediately frozen following collection, and carefully transported in a frozen state too. However, blood samples in the GAinS collection were posted to a collection point in the standard post. It is likely that DNA in the sample degraded during this period of transportation at room temperature. This could account for the difference in mean concentration and average yield obtained from samples extracted using the Maxwell 16 robot from the GRACE and GAinS collections. This difference is statistically significant ($p=8.3x10^{-17}$ for concentration, $p=1.3x10^{-16}$ for yield).

3.4.2 DNA extraction from saliva

The non invasive nature of saliva collection for genetics studies has higher compliance and fewer ethical implications particularly in the recruitment of children. The main problem in processing saliva samples turned out to be low DNA yields, but this varied highly from sample to sample. It was shown that the tissue and blood kits for use with the Maxwell 16 instrument are very comparable, and at lower yields these are also very comparable with the manual saliva DNA extraction method from Oragene. However, at higher concentrations, manual extraction generates substantially higher yields. The manual extraction method is also cheaper, as it does not require the purchase of Maxwell kits. Manual extraction is performed with the purifier solution that is provided free of charge when Oragene kits are purchased. Manual extraction was therefore chosen as the method of choice for DNA extraction from saliva.

Further tests were performed in order to improve the DNA yields and concentrations obtained from the manual extraction method. These tests showed that the most reliable method for obtaining DNA at a higher concentration and yield was to extract from 500μl saliva and elute into 50μl elution buffer. It seems counter intuitive that extracting from a smaller volume of saliva could give greater yields per μl. It is likely that this is because when extracting from 500μl saliva, extractions can be performed in 1.5ml eppendorf tubes rather than 15ml falcon tubes, and so the DNA extracted is less likely to be missed during the elution process. This was also a concern when reducing the elution volume. DNA must be lost in those cases where there is not an equal yield of DNA in a 50μl and 100μl elution from the same sample. However, using a 50μl elution does often lead to higher yields, and since it is very important in this collection that concentrations be kept as high as possible, the 50μl elution volume is certainly the best.

Unsurprisingly, there is a significant correlation between the amount of saliva collected in each sample and the yield obtained from extraction. Therefore it is extremely important in a saliva collection to ensure that enough saliva is provided by each individual in the study. As every participant in this collection was a young child, this was particularly difficult to achieve in this case. This does demonstrate though, that even though saliva should be the ideal material for collection of DNA from children (given its non-invasive nature), it does pose problems after collection when the resulting quantity of DNA can be low.

Since the concentrations of DNA obtained from the saliva samples tested is on average very low, it is necessary to whole genome amplify the extracted DNA. This has been performed for all samples and most samples performed extremely well, giving much higher amounts of DNA from each sample which can be used in downstream applications including GWAS and high-throughput sequencing.

3.4.3 DNA extraction from serum

One of the more difficult aspects of working in infectious disease genetics is the requirement to obtain cohorts large enough for modern genetic studies. Collections are largely made in remote locations where ethical considerations may be more complex and where people with disease are in medical settings for short lengths of time, thereby making it harder to accumulate very large collections.

When a sample collection has already been made, it is therefore very important that it can be used for genetics studies in addition to the collection's primary purpose. With ethical approval, samples already collected for other purposes could also be used for genetics studies. Serum, a component of blood which separates after coagulation, is often collected for other purposes and can remain unused after a study is complete. Although by its very nature, serum should be largely cell free, it is still possible to extract small amounts of DNA from these samples. Although none of the work in this thesis uses DNA from serum samples, it is important to study the best way to extract this as a means of obtaining sample collections for the future.

As previously mentioned, the main difficulty in extracting DNA from serum is that there is very little DNA in the sample in the first place. Therefore it is important to use a technique designed to capture all DNA in a sample. Of the three techniques tested for DNA extraction from serum, one stood out as being far better than the others. The Qiagen circulating nucleic acid kit more consistently obtained higher yields with better quality DNA (as judged by its performance in PCR before and after whole genome amplification) than the other techniques tested. This was even the case when extracting from serum that had been stored for long periods of time.

Before implementing a technique for the extraction of DNA from a large sample collection, more tests should be performed to find the best method for performing these extractions. However, these tests work well as a proof of principle to show that serum collections probably could be used for large genetic studies.

3.4.4 Other collection materials

The Hill group has also been involved with the processing of samples collected as buccal swabs, blood spots on cards similar to Guthrie cards, and the extraction of DNA from blood using the traditional salting out method. Although these methods have not been tested as thoroughly as the methods here, it is generally considered that the salting out method for DNA extraction from blood is cheap and often leads to high yields, but requires high volumes of blood and is very labour intensive. Buccal swabs are a good non-invasive way of collecting DNA samples, but this has largely been superseded by the new Oragene saliva collection method which gives higher DNA yields with lower bacterial content and is easier to perform collections for.

Blood spots on Guthrie cards are collected routinely after birth worldwide, and were originally designed to use to check for the genetic disorder phenylketonuria (Blau *et al.*, 2010; Guthrie and Susi, 1963). Just as saliva may be a preferable alternative to blood collection due to consent issues, participation in genetic studies may be increased with this less invasive collection procedure (Bhatti *et al.*, 2009). In addition, some collections of blood spots on cards are kept and are available as an important potential resource for genetics studies (with the correct ethics procedures in place), in a similar way to serum collections. However, it is often difficult to extract DNA from these cards, due to the presence of PCR inhibitors and the low quantity of DNA in a blood spot (Makowski *et al.*, 1997). Methods to whole genome amplify DNA directly from the card could improve the yield of DNA and have been trialled in the Hill group laboratory, though, with some success.

3.5 Conclusion

I have demonstrated here the value of thoroughly testing DNA extraction techniques, both to discover which techniques work best, and also to figure out the best method for

any individual collection. I have shown an analysis to determine the best techniques for DNA extraction from blood for two different types of study, as well as extractions from saliva and serum. I have also shown the importance of storage conditions for blood samples for DNA extraction. The results presented in this chapter have been key to obtaining good quantities of extracted DNA for collections used in this thesis. The findings are also of value for future collections.

Chapter 4

Candidate genes for susceptibility to sepsis and mild LRTI

4.1 Introduction

The introduction of genome-wide association studies (GWAS) and high-throughput sequencing in human genetics studies has meant that candidate gene studies are sometimes considered to be outdated and obsolete. Candidate gene studies have been described as "woefully inadequate" and "a shot in the dark" (Altshuler *et al.*, 2008). Since these studies only look at a limited number of polymorphisms for any given trait, they are certainly not adequate to determine the main genetic loci associated with any given disease. In addition, the limited number of SNPs genotyped in a candidate gene study mean that it is a lot harder to detect evidence of confounding within a dataset. DNA quality, population ancestry and familial relationships can therefore not be accurately calculated, which could bias the study's results (Anderson *et al.*, 2010). However, there are still many SNPs which cannot be accurately genotyped in GWAS, examples of which are presented here.

This chapter presents studies for SNPs in two candidate genes for infectious dis-

ease. Four SNPs were genotyped in the *MBL2* gene. One of these SNPs can be accurately genotyped using a GWAS chip, but the other three reside very close together in exon 1 of the gene, making it impossible to design primers to specifically genotype these SNPs individually using a GWAS. One SNP was genotyped in the *IFITM3* gene (rs12252). The DNA sequence for this gene is extremely similar (95% identical) to that of the *IFITM2* gene (National Center for Biotechnology Information, 2012). Therefore, the challenge associated with genotyping this SNP is in designing primers which specifically amplify the fragment of interest in *IFITM3* but not *IFITM2*.

The *MBL2* SNPs have previously been associated with bacterial disease (de Benedetti *et al.*, 2007; Eisen *et al.*, 2006; Frakking *et al.*, 2007; Garcia-Laorden *et al.*, 2008; Garred *et al.*, 2003; Roy *et al.*, 2002) while the *IFITM3* SNP was recently shown to associate with susceptibility to viral infection (Everitt *et al.*, 2012). These SNPs have all been studied and are presented here with respect to susceptibility to mild lower respiratory tract infection and severe sepsis, in both bacterial and viral subgroups. In addition, the *MBL2* SNPs are analysed with respect to survival from severe sepsis.

Interestingly, the SNPs in both of these genes show very different allele frequencies in different regions of the world. The allele frequency of the *MBL2* variants can range from 0 to 80% between South Africa and Peru (Garred, 2008), while the allele frequency of the *IFITM3* SNP rs12252 ranges from 4% in Europe to 59% in Japan (1000 Genomes Project Consortium *et al.*, 2010).

4.2 Aim

The aim of this work was to assess associations between SNPs in two candidate genes (*MBL2* and *IFITM3*) and susceptibility to sepsis and mild LRTI. Associations between *MBL2* and survival from sepsis were also assessed. The results for these two

candidate genes are presented separately below.

4.3 *MBL2* association study

4.3.1 Introduction

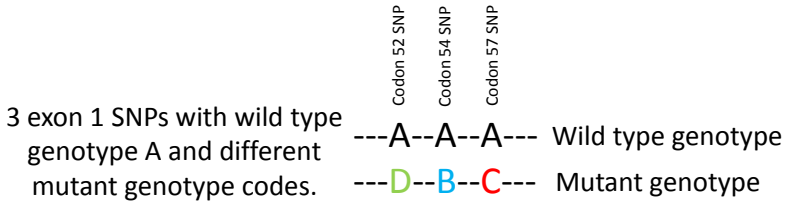
MBL2, which encodes mannose-binding lectin (MBL), is one of the most extensively studied but still controversial candidate genes reported to be associated with sepsis and infectious disease phenotypes. MBL binds carbohydrates on the surface of a wide variety of viruses, bacteria, yeasts, fungi and protozoa and subsequently triggers host innate immune responses through activation of the complement system, in addition to promoting opsonophagocytosis by a complement-independent pathway (Turner, 2003).

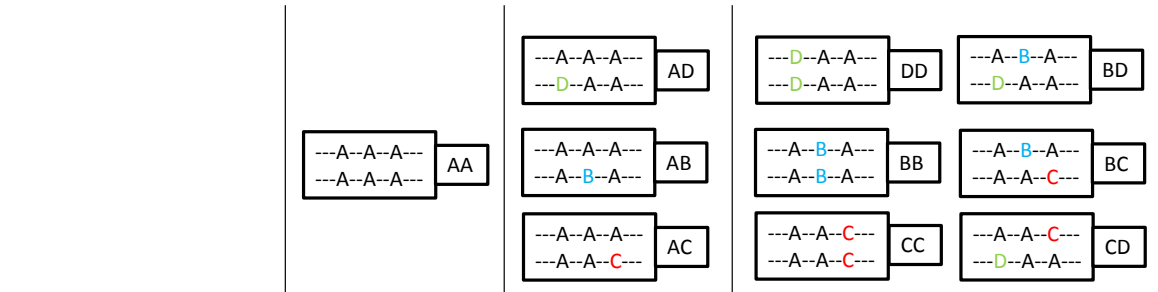
The function of MBL is known to be influenced by common SNPs found both in the promoter region and in exon 1 of the gene, which affect the level of gene expression and structure of the MBL protein (Chapman *et al.*, 2010a; Garred, 2008; Garred *et al.*, 2006; Larsen *et al.*, 2004; Lipscombe *et al.*, 1992; Madsen *et al.*, 1994, 1995, 1998; Matsushita *et al.*, 1995; Mullighan *et al.*, 2008; Naito *et al.*, 1999; Sumiya *et al.*, 1991).

Three functional SNPs (denoted D-rs5030737-codon 52; B-rs1800450-codon 54 and C-rs1800451-codon 57) in exon 1 of *MBL2* affect the stability, ligand binding capacity, complement activating ability and half-life of the encoded protein (Garred, 2008; Garred *et al.*, 2006; Larsen *et al.*, 2004; Lipscombe *et al.*, 1992; Madsen *et al.*, 1994; Matsushita *et al.*, 1995; Naito *et al.*, 1999; Sumiya *et al.*, 1991). Because their effect on serum MBL levels is very similar, wild type alleles for these SNPs are denoted A, while the B,C and D variants are pooled and given the designation O. If an individual is heterozygous (A/O) for any one of these three variants, MBL levels are 10-20% that of wild type individuals, whereas in variant homozygotes or compound heterozygotes

(O/O), serum MBL concentrations are virtually undetectable (Chapman *et al.*, 2010a; Mullighan *et al.*, 2008).

MBL2 promoter polymorphisms have also been associated with MBL levels independent of these exonic SNPs, with the SNP denoted X/Y at position -221 (rs7096206) having the strongest down-regulating effect (Garred, 2008; Madsen *et al.*, 1995, 1998). Figure 4.1 shows the names given to combined genotypes for the three exon SNPs and this promoter SNP.





	A/A	A/O	O/O
With Y/Y promoter SNP	YA/YA	YA/YO	YO/YO
With X/Y promoter SNP	XA/YA	XA/YO	XO/YO
With X/X promoter SNP	XA/XA	This combined genotype has not been observed	This combined genotype has not been observed

Figure 4.1: A diagram to show the different observed combinations of *MBL2* genotypes. Two combined genotypes have not been observed in any published work that I know of. The XO/YO combined genotype has not been observed in any published work that I know of, but is observed in this study.

Previous studies have reported both the presence and the absence of an association between these SNPs and susceptibility to, and outcome from, infectious disease, including invasive pneumococcal disease (IPD), community-acquired pneumonia (CAP) and sepsis (Eisen *et al.*, 2008; Garcia-Laorden *et al.*, 2008; Gordon *et al.*, 2006; Roy

et al., 2002). Upwards of 30 studies have reported the presence or absence of associations with *MBL2* polymorphisms and/or MBL levels in sepsis phenotypes, with sepsis sample sizes ranging from 44 to 848 (de Benedetti *et al.*, 2007; Eisen *et al.*, 2008, 2006; Frakking *et al.*, 2007; Garcia-Laorden *et al.*, 2008; Garred *et al.*, 2003; Gordon *et al.*, 2006; Ozdemir *et al.*, 2010). In an attempt to confirm or refute a role for common *MBL2* genetic variation in relation to sepsis susceptibility and outcome, an association study was performed in the large cohort of well-phenotyped patients with sepsis recruited from intensive care units (ICUs) across Europe as part of the GenOSept consortium. This is the largest known study investigating the influence of *MBL2* polymorphisms on susceptibility to, and survival from sepsis. In addition, in order to investigate the role of *MBL2* polymorphisms in susceptibility to milder infectious disease, these polymorphisms were studied in the GRACE mild lower respiratory tract infection (LRTI) cohort. No studies of *MBL2* variation have been published with mild LRTI in a primary care setting to my knowledge.

4.3.2 Methods

4.3.2.1 Genotyping

The three exon SNPs were genotyped in 1846 community-acquired pneumonia (CAP) and faecal peritonitis (FP) sepsis samples from the GenOSept and GAinS cohorts using high resonance melting analysis (HRMA). Seven of these failed genotyping. Genotypes of eleven samples were confirmed by sequencing. All of these samples were also genotyped for the -221 promoter SNP; 369 by HRMA and 1477 by GWAS on the Affymetrix chip 5.0. Promoter genotypes were confirmed in two samples by sequencing. GWAS data were cleaned using the parameters described in Materials and methods section 2.4.7. Twenty-eight day status was not available for 107 sepsis patients and so these were not included in the survival analysis performed. Every participant except three in the GRACE study (n=5869) was successfully genotyped

for all four MBL SNPs described here using HRMA. Sequencing was used to confirm some genotypes. See Material and methods section 2.3.5 for details of all genotyping procedures.

4.3.2.2 Subgroups and statistical analysis

Both survival from and susceptibility to sepsis were analysed. The main end point for survival was chosen as 28 days because most studies of sepsis patients have used 28-day mortality as a clinical end-point (Dhainaut *et al.*, 2003; Huh *et al.*, 2009; Man *et al.*, 2012; Nakada *et al.*, 2011; Winters *et al.*, 2010). Specific age groups and UK individuals were analysed separately, to remove any age or population bias in the data. In addition to sepsis caused by CAP and FP, other phenotypes were studied, in order to establish if there was an effect of *MBL2* polymorphisms on specific phenotypes within sepsis. These phenotypes were: sepsis with CAP alone, sepsis with FP alone and sepsis caused by *Streptococcus pneumoniae* (the largest subgroup defined by organism) alone. Survival at 6 months was also analysed because a previous study (Garcia-Laorden *et al.*, 2008) found that *MBL2* genotypes were associated with death after 28 days.

In the susceptibility analysis, all UK individuals with sepsis were analysed against UK healthy controls from the GRACE study. CAP sepsis susceptibility was also studied separately, but valid controls to analyse susceptibility to FP sepsis alone were not available. CAP sepsis caused by *S. pneumoniae* only was also analysed. All subgroups within GRACE were analysed for susceptibility, as described in section 4.3.3.3.

All analyses were performed by logistic regression using SPSS v18 (PASW Statistics 18). Analysis defining deficient and sufficient genotypes was performed in three different ways (see results). Analysis looking at each exon and promoter combined genotype (as in figure 4.1) against the YA/YA combined genotype, and analysis of the

distribution of all seven combined genotypes (with XO/YO and YO/YO genotypes grouped together) were performed in every subgroup. Allelic, genotypic, recessive, dominant and heterozygous tests were carried out for each SNP individually, and on the A/A, A/O and O/O combined exon genotypes.

Age associated significantly with sepsis 28 day survival ($p=1.51x10^{-19}$), whereas other confounding factors such as gender, CAP/FP diagnosis and presence of respiratory disease did not. Therefore age was included as a covariate in all the sepsis analyses. Since many of the sepsis samples analysed had undergone genome-wide genotyping, population outliers could be removed based on multi-dimensional scaling (MDS) data. In GRACE, as no GWAS analysis has been performed, population stratification could not be corrected for in this way. Therefore, all GRACE analyses (except the matched analysis) used country as a covariate. LRTI overall and rhinovirus LRTI subgroups were also analysed with smoking as an additional covariate, as smoking status differed significantly between cases and controls.

4.3.3 Results

The observed genotype frequencies in the sepsis cases, GRACE mild LRTI cases and GRACE healthy controls were comparable to those observed in previously-published, large population-based European studies (Dahl *et al.*, 2004; Garcia-Laorden *et al.*, 2008).

4.3.3.1 Sepsis survival analysis

The number of deaths and survivors within the subgroups analysed for sepsis survival are shown in table 4.1.

MBL2 genotype distributions for the main survival analyses are presented in table 4.2. MBL functional status (where the XO/YO, YO/YO and XA/YO combined genotypes were defined as MBL deficient and all other combined genotypes as MBL

Table 4.1: Survival analyses performed for the sepsis data set, and the numbers of deaths and survivors included in each subgroup. The 6 month survival group is a subset of samples from the whole set, for which 6 month survival information was available.

	Analysis	**Number of deaths**	**Number of survivors**
28 day survival	Whole set	357	1482
	Aged 70 and under	134	940
	Aged 60 and under	67	578
	CAP sepsis	178	805
	FP sepsis	179	677
	Pneumococcal sepsis	45	200
	UK sepsis samples	158	673
6 month survival	Whole set	265	704

sufficient) did not associate with 28 day sepsis mortality (p=0.169), 6 month sepsis mortality (p=0.229) or pneumococcal sepsis 28 day mortality (p=0.643). Significant associations were not seen in any of the other subgroups either (p>0.05).

The specific combinations of *MBL2* genotypes that define functional sufficiency or deficiency remain somewhat controversial. Therefore, sufficient/deficient analysis was also performed by defining genotypes XO/YO, YO/YO, XA/YO and XA/XA (Chapman *et al.*, 2010a) or XO/YO, YO/YO, XA/YO and YA/YO (Frakking *et al.*, 2007) as MBL deficient. There was no association in the whole study set or any of the subgroups (p>0.05) in these variations of the analysis.

Each genotype was also examined individually, combined genotypes were compared between cases and controls, and the exon genotypes combined (A/A, A/O, O/O) were analysed. The lowest p-value observed in all tests was 0.017 for the promoter polymorphism heterozygous test for FP sepsis survival. Genotype frequencies are listed for survivors and deaths for the main survival subgroups analysed in this cohort, in table 4.3.

Table 4.2: MBL functional status and the combined genotype distribution (for the -221 promoter SNP and 3 functional exonic SNPs) among survivors and deaths in 28 day and 6 month survival analyses, and in 28 day pneumococcal survival analysis. The XO/YO genotypes both included one polymorphism in the -221 promoter, and variant B and D exonic SNPs, and were confirmed by sequencing. P-values given are for the main sufficient/deficient analysis. Association tests for deficient/sufficient status were non-significant for all subgroups listed in table 4.1.

MBL functional status	**Genotype**	**CAP and FP sepsis 28 day deaths (%)**	**CAP and FP sepsis 28 day survivors (%)**	**Sepsis 6 month deaths (%)**	**Sepsis 6 month survivors (%)**	**Pneumococcal sepsis 28 day deaths (%)**	**Pneumococcal sepsis 28 day survivors (%)**
MBL Deficient	XO/YO	1 (0.3)	1 (0.1)	0	1 (0.1)	0	0
	YO/YO	19 (5.3)	81 (5.5)	12 (4.5)	35 (5.0)	2 (4.4)	11 (5.5)
	XA/YO	30 (8.4)	162 (10.9)	28 (10.6)	73 (10.4)	5 (11.1)	24 (12)
	Total	**50 (14)**	**244 (16.5)**	**40 (15.1)**	**109 (15.5)**	**7 (15.5)**	**35 (17.5)**
MBL Sufficient	XA/XA	17 (4.8)	75 (5.0)	10 (3.8)	34 (4.8)	1 (2.2)	15 (7.5)
	XA/YA	74 (20.7)	355 (24.0)	67 (25.3)	166 (23.6)	12 (26.7)	40 (20)
	YA/YO	98 (27.4)	381 (25.7)	68 (25.7)	189 (26.8)	9 (20)	57 (28.5)
	YA/YA	118 (33.1)	427 (28.8)	80 (30.2)	206 (29.3)	16 (35.6)	53 (26.5)
	Total	**307 (86)**	**1238 (83.5)**	**225 (85)**	**595 (84.5)**	**38 (84.5)**	**165 (82.5)**
	Overall Total	**357**	**1482**	**265**	**704**	**45**	**200**
p-value		**0.169**		**0.229**		**0.643**	

Table 4.3: Genotype counts for *MBL2* SNPs are shown for each of the three main survival analysis subgroups, with percentages shown in brackets.

***MBL2* Genotype**	**Sepsis 28 day deaths (%)**	**Sepsis 28 day survivors (%)**	**Sepsis 6 month deaths (%)**	**Sepsis 6 month survivors (%)**	**Pneumococcal sepsis 28 day deaths (%)**	**Pneumococcal sepsis 28 day survivors (%)**
A/A	**209 (58.5)**	**857 (57.8)**	**157 (59.2)**	**406 (57.7)**	**29 (64.4)**	**108 (54)**
A/B	81 (22.7)	336 (22.7)	58 (21.9)	164 (23.3)	8 (17.8)	46 (23)
A/C	10 (2.8)	43 (2.9)	6 (2.3)	20 (2.8)	0	8 (4)
A/D	37 (10.4)	164 (11.1)	32 (12.1)	78 (11.1)	6 (13.3)	27 (13.5)
Sum A/O	**128 (35.9)**	**543 (36.7)**	**96 (36.2)**	**262 (37.2)**	**14 (31.1)**	**81 (40.5)**
B/B	3 (0.8)	39 (2.6)	5 (1.9)	17 (2.4)	0	9 (4.5)
B/C	2 (0.6)	5 (0.3)	0	0	0	0
B/D	9 (2.5)	30 (2)	4 (1.5)	16 (2.3)	2 (4.4)	2 (1)
C/C	0	1 (0.1)	0	1 (0.1)	0	0
C/D	3 (0.8)	4 (0.3)	2 (0.8)	2 (0.3)	0	0
D/D	3 (0.8)	3 (0.2)	1 (0.4)	0	0	0
Sum O/O	**20 (5.6)**	**82 (5.5)**	**12 (4.5)**	**36 (5.1)**	**2 (4.4)**	**11 (5.5)**
Total	**357**	**1482**	**265**	**704**	**45**	**200**

4.3.3.2 Sepsis susceptibility analysis

The number of cases and controls within the subgroups analysed for sepsis susceptibility are shown in table 4.4.

Table 4.4: Number of cases and controls within the subgroups analysed for susceptibility to sepsis. UK controls were healthy controls from the GRACE study.

Analysis	Number of cases	Number of UK controls
UK sepsis	831	477
UK CAP sepsis	496	477
UK pneumococcal sepsis	95	477

MBL2 genotype distributions for the cases and controls in the susceptibility analyses are presented in table 4.5. MBL functional status did not associate with UK sepsis overall (p=0.51), UK CAP sepsis susceptibility (p=0.343) or with UK pneumococcal sepsis susceptibility (p=0.587).

As with survival analysis, functionally sufficient/deficient genotypes were analysed according to their alternative definitions. There was no significant association in any of these alternative analyses (p>0.05). Each genotype was also examined individually, combined genotypes were compared between cases and controls, and the exon genotypes combined (A/A, A/O, O/O) were analysed. The lowest p-value observed in all tests was 0.029 for the B genotype recessive test in the UK pneumococcal susceptibility subgroup. Genotype frequencies are listed for cases and controls in table 4.6.

4.3.3.3 Mild LRTI susceptibility analysis

The same analyses were performed for all GRACE subgroups (as listed in table 4.7). The main association analyses are presented in table 4.8. MBL functional status did not associate with LRTI overall (p=0.653), bacterial LRTI (p=0.181) or viral LRTI (p=0.844), or any of the other subgroups. Functionally sufficient/deficient genotypes

Table 4.5: MBL functional status and the combined genotype distribution (for the -221 promoter SNP and 3 functional exonic SNPs) among cases and controls in UK sepsis, UK CAP sepsis and UK pneumococcal sepsis susceptibility analyses. The XO/YO genotypes both included one polymorphism in the -221 promoter, and variant B and D exonic SNPs, and were confirmed by sequencing. Association tests for deficient/sufficient status for each set of cases compared to the controls were non-significant.

MBL functional status	**Genotypes**	**UK healthy controls (%)**	**UK sepsis cases (%)**	**UK CAP sepsis cases (%)**	**UK pneumococcal sepsis cases (%)**
MBL Deficient	XO/YO	0	2 (0.2)	1 (0.2)	0
	YO/YO	19 (4)	51 (6.1)	27 (5.4)	6 (6.3)
	XA/YO	57 (11.9)	74 (8.9)	45 (9.1)	12 (12.6)
	Total	**76 (15.9)**	**127 (15.2)**	**73 (14.7)**	**18 (18.9)**
MBL Sufficient	XA/XA	15 (3.1)	42 (5.1)	22 (4.4)	6 (6.3)
	XA/YA	107 (22.4)	202 (24.3)	128 (25.8)	21 (22.1)
	YA/YO	139 (29.1)	210 (25.3)	128 (25.8)	22 (23.2)
	YA/YA	140 (29.4)	250 (30.1)	145 (29.2)	28 (29.5)
	Total	**401 (84)**	**704 (84.8)**	**423 (85.2)**	**77 (81.1)**
	Overall Total	**477**	**831**	**496**	**95**
p-value for deficient/sufficient status		NA	0.51	0.343	0.587

Table 4.6: *MBL2* genotype counts are shown for each susceptibility subgroup, with percentages shown in brackets.

***MBL2* Genotype**	**UK healthy controls (%)**	**UK sepsis cases (%)**	**UK CAP sepsis cases (%)**	**UK pneumococcal cases (%)**
A/A	**262 (54.9)**	**494 (59.4)**	**295 (59.5)**	**55 (57.9)**
A/B	122 (25.6)	182 (21.9)	114 (23)	23 (24.2)
A/C	20 (4.2)	17 (2)	10 (2)	2 (2.1)
A/D	54 (11.3)	85 (10.2)	49 (9.9)	9 (9.5)
Sum A/O	**196 (41.1)**	**284 (34.1)**	**173 (34.9)**	**34 (35.8)**
B/B	6 (1.3)	22 (2.6)	13 (2.6)	5 (5.3)
B/C	2 (0.4)	2 (0.2)	0	0
B/D	8 (1.7)	19 (2.3)	10 (2)	1 (1.1)
C/C	0	0	0	0
C/D	1 (0.2)	6 (0.7)	4 (0.8)	0
D/D	2 (0.4)	4 (0.5)	1 (0.2)	0
Sum O/O	**19 (4.0)**	**53 (6.3)**	**28 (5.6)**	**6 (6.4)**
Total	**477**	**831**	**496**	**95**

were also analysed according to their alternative definitions. One subgroup had a p-value <0.05 for these main analyses (*Haemophilus influenzae* subgroup, O/O and XA/O as deficient, p=0.048). Each genotype was also examined individually, combined genotypes were compared between cases and controls, and the exon genotypes combined (A/A, A/O, O/O) were analysed. The lowest p-value observed in all tests was 0.012 for the B genotype heterozygous test in the coronavirus subgroup. Genotype frequencies are listed for controls and the main subgroups of cases in table 4.9.

4.3.4 *MBL2* discussion

Overall, across all subgroups and with all tests included within the sepsis analysis alone, 374 tests were performed. With Bonferroni correction, a p-value of 1.34×10^{-4} would be necessary to be able to consider a result significant, but this is a very stringent correction, since the tests are not all independent. However, the lowest p-

Table 4.7: GRACE subgroups analysed for *MBL2* SNPs with the covariates used in each analysis, and the number of cases and controls included.

Type of analysis	**Covariates**	**Number of cases**	**Number of controls**
Basic case control	country	2864	2686
Matched	none	2449	2449
All bacterial	country	396	2686
Viral	country	1283	2686
Non-viral	country	1581	2686
Consolidation	country	205	2686
Recurrent LRTI	country	850	2686
Haemophilus influenzae	country	177	2686
Streptococcus pnuemoniae	country	156	2686
H. influenzae and *S. pneumoniae*	country	307	2686
Rhinovirus	country	541	2686
Coronavirus	country	200	2686
Influenza A and B	country	273	2686
Rhinovirus	country and smoking	541	2679
Rhinovirus smokers	country	226	599
Rhinovirus non-smokers	country	315	2080
Case control	country and smoking	2863	2679
Case control smokers	country	805	599
Case control non-smokers	country	2058	2080

Table 4.8: MBL functional status and the combined genotype distribution (for the -221 promoter SNP and 3 functional exonic SNPs) among cases and controls in GRACE overall, bacterial and viral subgroup susceptibility analyses. Association tests for this deficient/sufficient status for each set of cases compared to the controls were non-significant (p-values are shown).

MBL functional status	**Genotypes**	**GRACE healthy controls (%)**	**GRACE cases (%)**	**GRACE bacterial cases (%)**	**GRACE viral cases (%)**
MBL Deficient	YO/YO	119 (4.4)	160 (5.6)	19 (4.8)	67 (5.2)
	XA/YO	285 (10.6)	262 (9.1)	31 (7.8)	122 (9.5)
	Total	**404 (15)**	**422 (14.7)**	**50 (12.6)**	**189 (14.7)**
MBL Sufficient	XA/XA	127 (4.7)	137 (4.8)	19 (4.8)	54 (4.2)
	XA/YA	669 (24.9)	709 (24.8)	102 (25.8)	325 (25.3)
	YA/YO	675 (25.1)	718 (25.1)	100 (25.3)	317 (24.7)
	YA/YA	811 (30.2)	878 (30.7)	125 (31.6)	398 (31)
	Total	**2282 (84.9)**	**2442 (85.4)**	**346 (87.5)**	**1094 (85.2)**
	Overall Total	**2686**	**2864**	**396**	**1283**
p-value for deficient/sufficient status		NA	0.653	0.181	0.844

Table 4.9: *MBL2* genotype counts are shown for each susceptibility subgroup, with percentages shown in brackets.

***MBL2* Genotype**	**GRACE healthy controls (%)**	**GRACE cases (%)**	**GRACE bacterial cases (%)**	**GRACE viral cases (%)**
A/A	**1607 (59.8)**	**1724 (60.2)**	**246 (62.1)**	**777 (60.6)**
A/B	590 (22)	617 (21.5)	81 (20.5)	279 (21.7)
A/C	85 (3.2)	81 (2.8)	12 (3)	29 (2.3)
A/D	285 (10.6)	282 (9.8)	38 (9.6)	131 (10.2)
Sum A/O	**960 (35.8)**	**980 (34.1)**	**131 (33.1)**	**439 (34.2)**
B/B	52 (1.9)	57 (2)	5 (1.3)	22 (1.7)
B/C	20 (0.7)	19 (0.7)	3 (0.8)	8 (0.6)
B/D	38 (1.4)	68 (2.4)	10 (2.5)	28 (2.2)
C/C	1 (0)	0	0	0
C/D	2 (0.1)	6 (0.2)	0	4 (0.3)
D/D	6 (0.2)	10 (0.3)	1 (0.3)	5 (0.4)
Sum O/O	**119 (4.3)**	**160 (5.6)**	**19 (4.9)**	**67 (5.2)**
Total	**2686**	**2864**	**396**	**1283**

value of 0.017 (for the FP sepsis survival promoter polymorphism heterozygous test) would still not be considered a significant result, even with less stringent correction.

In GRACE, a total of 580 tests have been performed, giving a Bonferroni corrected p-value threshold of 8.62×10^{-5}. Again, these tests are not strictly independent, but the most significant p-value (0.012 for the B genotype heterozygous test in the coronavirus subgroup) would still not be considered significant, even with less stringent correction.

Despite being the largest known cohort of severe reaction to infection that has been analysed according to *MBL2* polymorphisms, this study of 28 day survival from severe sepsis across Europe did not find a significant association between *MBL2* deficiency genotypes and either survival from severe sepsis (at 28 days and six months), susceptibility to severe sepsis or susceptibility to mild LRTI. The mild LRTI study is the only reported research studying *MBL2* polymorphisms and mild infectious disease.

These large sample collections enabled the examination of *MBL2* genotype associations in a number of predefined subgroups. Even when examining these smaller subgroups, these analyses still have sufficient power to detect any strong associations. Odds ratios of 1.45, 1.57 and 1.52 or more can be detected with 80% power in the largest subgroups (28 day survival, six month survival and CAP sepsis susceptibility respectively). Even in the most underpowered subgroup, (pneumococcal sepsis 28 day survival) there is 80% power to detect odds ratios of 2.5 or more. Although age was corrected for in the primary sepsis analyses, association analysis was also performed in two groups of younger patients (less than 60 and less than 70 years of age) in whom it might be expected that MBL deficiency would play a greater role in mortality. Twenty-eight day survival was also analysed in the predefined groups of CAP sepsis, FP sepsis, and pneumococcal sepsis (caused by *S. pneumoniae*). Similarly, susceptibility associations were examined in the subgroups of patients with CAP and pneumococcal sepsis. No associations were found in any of these analyses. 28 day survival in UK cases alone was also analysed, with the aim of avoiding any population stratification which might have obscured associations, but again *MBL2* genotype did not influence outcome. Moreover, population outliers were removed from these sepsis analyses based on multidimensional-scaling using genome-wide genotypes. Therefore it is unlikely that population stratification was a major confounder in this study.

In the main subgroups of LRTI overall, bacterial LRTI and viral LRTI there is 80% power to detect odds ratios of 1.20, 1.39 and 1.25 respectively. In the smallest GRACE subgroup (smokers with rhinovirus) there is 80% power to detect an odds ratio of 1.62 or more. The GRACE sample set is therefore powered to see associations of smaller effects than in the sepsis cohort. Smaller effects are probably what would be expected in a cohort of milder phenotype.

One of the limitations of the sepsis collection is that microbiological information is not available for all cases (approximately 60% of sepsis patients had a microbe iden-

tified), and therefore the causal organism of infection is not known for all individuals. Without this information, thorough analysis of subgroups within the data cannot be performed (with the exception of the pneumococcal analysis). However, even with this information, pathogen based subgroups (other than *S. pneumoniae*) are likely to be too small to have sufficient power to detect an association.

Due to the study design, the patients with the most severe outcome from sepsis who died very quickly on admission to ICU were not included in this study (for instance if there wasn't time to collect consent from next of kin). Although this will have diluted the sample cohort, there is still adequate power, and so it is unlikely that this will have affected the results.

The effect of *MBL2* functional genetic variants has been studied widely in a number of sepsis related phenotypes. Garcia-Laorden *et al.* (2008) found, in one of the largest studies, that *MBL2* genotypes do not associate with susceptibility to CAP in adults, which is consistent with these findings. They also concluded, as this study did, that there was no association between 28 day mortality and *MBL2* deficient genotypes. They did, however, find an association with 90 day survival. As the 6 month survival analysis shows, however, these data do not support such an association with longer term survival, even though there is 100% power to detect an association with the odds ratio they observe (OR=2.34).

Gordon *et al.* (2006) have reported the study with the most similar sepsis phenotype to that of GenOSept, but with many fewer patients. They found an association between A/O and O/O exonic polymorphisms and susceptibility to sepsis ($p=0.001$), and also showed an even more significant association between the distribution of combined genotypes and susceptibility to sepsis ($p<0.0001$). These findings were not replicated in this study, despite the phenotype being very comparable. This suggests that Gordon *et al.*'s findings were false positives, since studies with fewer patients are more likely to be unreliable.

Eisen *et al.* (2008) performed a meta analysis of six studies of *MBL2* polymorphisms in patients with severe bacterial infection. They found an association between low circulating levels of MBL and mortality, and an increased risk of death from *S. pneumoniae* infection (OR=5.62). In the study presented here, an association was not observed between *MBL2* deficient genotypes and death from sepsis due to *S. pneumoniae* infection, despite having 100% power to find an association of this strength. However, Eisen *et al.* (2008) also showed that *MBL2* genotype poorly determines MBL levels, and in fact, the proportion of MBL deficient individuals is underestimated when defined by genotype. It is therefore possible that all MBL deficient individuals were not captured in this study, and that MBL levels, rather than genotype, might correlate with survival in the GenOSept cohort. Nevertheless, the findings presented here strongly suggest that an important effect of MBL status on sepsis phenotypes is highly unlikely.

Since there are no published studies of *MBL2* and susceptibility to mild lower respiratory tract infection, it is impossible to compare the data presented here with previous similar work. A study of MBL genotypes and levels in a very small number of children showed no association with mild respiratory syncytial virus (RSV), which is equivalent to a very small proportion of my data. This small study is far from conclusive, but it does support my findings (Kristensen *et al.*, 2004). Since the GRACE study is well-powered and many subgroups have been analysed, it is unlikely that further work would find an association between *MBL2* genotypes and susceptibility to mild lower respiratory tract infection.

Variant *MBL2* is very common in certain populations among healthy individuals (Garred *et al.*, 2006) and it is known that alternative immune mechanisms can replace the function of MBL (Roos *et al.*, 2004) thereby rendering deficient individuals as immune competent. This could explain the lack of a significant association between *MBL2* functional polymorphisms and the phenotypes analysed in the present study.

Two individuals were identified with variant alleles X, B and D, which were confirmed by sequencing. Previous reports have suggested that due to linkage disequilibrium between the exon and promoter SNP, only the Y allele in the promoter is described with exonic polymorphisms (Garcia-Laorden *et al.*, 2008; Madsen *et al.*, 1995). This study shows that other combinations are also possible, although they are very rare.

Although connections between MBL levels or *MBL2* genotypes with sepsis have been extensively studied, most of the studies are statistically underpowered and therefore it is not surprising that findings have been contradictory. In this study, it has been shown that there is no significant association between *MBL2* genotypes and susceptibility to or survival from severe sepsis in a sample set that has adequate statistical power to show associations with effect sizes previously reported. In addition, the GRACE LRTI sample set has shown that there is no association between *MBL2* genotypes and susceptibility to LRTI overall, bacterial LRTI, viral LRTI or many further subgroups of the GRACE sample set.

4.4 *IFITM3* association study

4.4.1 Introduction

Interferon-inducible transmembrane proteins 1,2 and 3 (IFITM1,2 and 3) are viral restriction factors that mediate cellular resistance to a collection of viruses, including influenza, West Nile virus and dengue virus (Brass *et al.*, 2009). IFITM3 is necessary and sufficient for interferon (IFN) regulation of viral replication through interferon effector genes. IFN prevents the transferral of influenza A viral components to the cytosol, thereby inhibiting viral infection (Feeley *et al.*, 2011). There are currently 15 non-synonymous, 15 synonymous, one stop gaining and one possible splice-site altering SNPs known in the IFITM3 gene. The possible splice-site altering SNP

(rs12252) is found close to a splice site for one transcript of the IFITM3 gene (1000 Genomes Project Consortium, 2012a,b).

Everitt *et al.* (2012) report functional data and a genetic association with variation in the gene *IFITM3*. They found a significant additive association for susceptibility to severe H1N1 influenza infection with rs12252. However, their genetic analysis only included a small number of H1N1 influenza cases (n=53) and the association result used imputed data for the control population. Small sample sizes often imprecisely estimate genotype frequencies and imputed data may be inaccurate for rare SNPs such as this (Marchini and Howie, 2010). In an attempt to reassess this association, additional H1N1 influenza cases from the GAinS study, and over 5000 European cases of lower respiratory tract infection and matched controls from the GRACE study were genotyped for rs12252 (for details of genotyping methods, see Materials and methods section 2.3.5.3) .

4.4.2 Subgroups and statistical analysis

Severe H1N1 influenza cases with sepsis collected as part of the UK GAinS sepsis study, mild lower respiratory tract infection (LRTI) cases and healthy controls collected as part of the GRACE study were genotyped. Viral and bacterial subgroups within GRACE were studied (viral overall, rhinovirus, coronavirus, influenza, rhinovirus smokers, rhinovirus non-smokers, bacterial overall, *H. influenzae*, *S. pnuemoniae*, and *H. influenzae* and *S. pneumoniae* combined). The GAinS H1N1 cases were compared to GRACE UK and GRACE European controls. The H1N1 cases from Everitt *et al.* (2012) were studied against the European GRACE controls, and in combination with the GAinS H1N1 cases.

All severe H1N1 analyses were performed using a two-tailed Fisher's exact test. All analyses in GRACE were performed using logistic regression with country as a covariate. The allelic, dominant, recessive and genotypic model were tested in every

subgroup.

4.4.3 Results

2623 healthy European controls were genotyped for rs12252, providing the largest available collection of genotyped Caucasians for this SNP. Genotype counts and association results for all subgroups which showed p-values <0.05 are shown in table 4.10. No other subgroups showed any associations.

A slightly smaller number (n=34) of hospitalised sepsis cases with H1N1 infection were genotyped from GAinS, but different allele and genotype frequencies were observed to those shown by Everitt *et al.* (2012) (table 4.10). Analysis was performed for Everitt *et al.*'s cases, and a combinations of Everitt *et al.*'s cases and the GAinS H1N1 cases, compared to the healthy European controls in this study. These analyses clearly showed that the model of association is recessive for this SNP.

2730 European mild LRTI cases were also genotyped, and analysis found that this SNP does not associate with bacterial infections overall (379 cases), nor with specific infections with *H. influenzae* (152 cases) or *S. pneumoniae* (131 cases). However, there was evidence of a recessive association for this SNP with general mild viral infections of the respiratory tract (caused by rhinovirus (498), influenza (240), coronavirus (169), respiratory syncytial virus (116), human metapneumovirus (110) and parainfluenza virus (61), or a combination of two of the above viruses (54)). A recessive association is seen with viral LRTI as a whole (OR 3.59, 95% CI 1.01-12.79). Association results for the largest subgroups in this group were: rhinovirus (OR 3.93, 95% CI 0.87-17.77), influenza virus (OR 7.12, 95% CI 1.28-39.58) and coronavirus (OR 4.12, 95% CI 0.45-37.51).

Table 4.10: All analyses included 2623 Caucasian European controls from GRACE. Subjects of non-European origin were excluded from the analysis. H1N1 analysis was performed with the two-sided Fisher's exact test. Analyses with GRACE cases from across Europe were performed using logistic regression in PLINK (Purcell, 2012; Purcell *et al.*, 2007) with country as a covariate. The GRACE rhinovirus non-smokers were compared to 2029 control non-smokers. Association analysis p-values less than 0.05 are indicated with a *. OR=odds ratio.

Cases	**GRACE controls**	**Everitt *et al.* (2012) H1N1**	**GAinS H1N1**	**GAinS and Everitt *et al.* (2012) H1N1**	**GRACE viral cases**	**GRACE influenza cases**	**GRACE rhinovirus cases**	**GRACE rhinovirus non smokers**
n	2623	53	34	87	1248	259	532	306
HWE	1	0.00005	1	0.004	0.007	0.12	0.015	1
TT (%)	2417 (92.1)	46 (86.8)	31 (91.2)	77 (88.5)	1160 (92.9)	235 (90.7)	500 (94)	293 (95.8)
CT (%)	202 (7.7)	4 (7.4)	3 (8.8)	7 (8)	82 (6.6)	22 (8.5)	29 (5.5)	13 (4.2)
CC (%)	4 (0.15)	3 (5.6)	0	3 (3.4)	6 (0.48)	2 (0.77)	3 (0.5)	0
allelic P		*0.011	0.753	0.032	0.764	0.154	0.287	*0.038
allelic OR		2.498	1.107	1.936	0.96	1.358	0.821	0.543
recessive P		$*7.4x10^{-4}$	1	*0.001	*0.049	*0.025	0.076	0.999
recessive OR		39.29	NA	23.38	3.59	7.126	3.93	NA
dominant P		0.154	0.747	0.224	0.461	0.28	0.148	*0.041
dominant OR		0.56	0.881	0.6856	0.907	1.279	0.752	0.545
genotypic P		$*3x10^{-4}$	0.756	*0.002	0.077	0.062	*0.043	0.138

4.4.4 *IFITM3* discussion

Genotype data from this large collection of controls, compared to the combined data from the GAinS H1N1 cases and those from Everitt *et al.* (2012), indicate clearly that the model of association for this SNP is recessive (table 4.10), rather than the additive model described by Everitt *et al.* (2012). Rare homozygotes for this SNP are driving this association and show an OR of 23.38, with no association for heterozygotes compared to major homozygotes (OR 1.05 95% CI 0.48-2.30).

Without directly genotyped controls and with small sample sizes, it is difficult to define genetic associations. By extending the case series reported by Everitt *et al.*, a different genetic model for severe H1N1 influenza susceptibility has been found. This recessive susceptibility effect is important in characterising possible inter-population differences in susceptibility given that only 1 in 600 Europeans has the susceptibility genotype compared to as many as 1 in 4 East Asians and 1 in 12 Africans (1000 Genomes Project Consortium *et al.*, 2010). This association relies on three homozygotes genotyped by Everitt *et al.*, however, and not all of these individuals have undergone population stratification analysis to exclude consequent bias. This is particularly important given the wide variation in genotype frequency between populations. No minor homozygotes were observed in the Caucasian GAinS H1N1 cases, which is perhaps expected given the low frequency of the genotype in this population.

Moreover, by analysing large numbers of cases of community-acquired mild LRTIs and matched controls, this study has shown the same recessive susceptibility association for mild viral infections to that identified for severe H1N1 influenza. This indicates that the *IFITM3* locus does not provide a specific susceptibility locus for H1N1 influenza but, instead, this interferon pathway gene likely associates with altered susceptibility to several common viral infections. There was no evidence that this SNP is associated with mild bacterial infection, which is unsurprising since IFITM3 has not been implicated in the control of bacterial infection. IFITM3 has implicated in

infection by multiple viruses, including influenza A, HIV-1, yellow fever virus and west nile virus (Brass *et al.*, 2009; Feeley *et al.*, 2011; Huang *et al.*, 2011; Jiang *et al.*, 2010; Schoggins *et al.*, 2011). Therefore future studies with other large viral sample sets are necessary to further define this genetic association.

4.5 Discussion

A total of five SNPs in two candidate genes were genotyped in over 5000 samples. SNPs in *MBL2* have been extensively studied in infectious disease phenotypes previously, and have shown differing associations in bacterial disease subgroups. A SNP in *IFITM3* has only recently been implicated in susceptibility to viral disease, and the study presented here represents an extension of that study. All SNPs (with the exception of the *MBL2* promoter SNP) are difficult to genotype and show differing allele frequencies in populations across the world. It is therefore harder to form conclusions for associations found with these polymorphisms, since any association found can easily be confounded by genotyping error or population structure in sample sets. It is unlikely that these issues have confounded the results presented here, since control allele frequencies in the *MBL2* and *IFIMT3* studies match those previously published for Caucasian populations.

Despite extensive analysis of many different subgroups in mild and severe disease, none of the SNPs in *MBL2* associated with any subgroup with a low enough p-value to be considered statistically significant. In the main sufficient/deficient analysis, no p-value was lower than 0.05 across the 32 subgroups analysed. In contrast, the SNP in *IFITM3* shows a statistically significant recessive association with severe H1N1 influenza infection, mild viral infection and mild influenza infection.

This work highlights the problems associated with publishing genetic association results using small sample sizes. *MBL2* polymorphisms have shown differing as-

sociation results in different studies, largely due to the small sample sizes used in these studies. The *IFITM3* H1N1 association presented here is less significant with a greater number of samples, indicating that study of larger numbers of severe cases would be useful.

These association studies should aid the future study of these candidate genes. This work demonstrates that *MBL2* is very unlikely to show a true association in future sepsis and mild LRTI phenotype studies of a similar or smaller size. This *IFITM3* work should pave the way for future studies (which should have larger sample numbers) of association between the rs12252 SNP and susceptibility to a range of viral infections.

Chapter 5

GRACE: Genetic susceptibility to mild LRTI

5.1 Introduction

GRACE (Genomics to combat Resistance against Antibiotics in Community-acquired lower respiratory tract infections (LRTI) in Europe) is a multi-centre, Europe-wide network of excellence. GRACE studied individuals in 11 countries across Europe that presented at primary care with their primary symptom being an acute cough, and collected extensive clinical data to better understand their illness, and the way they were treated. GRACE is probably the largest study of LRTI ever undertaken in primary care, and has presented many findings from the network in peer-reviewed literature (Brookes-Howell *et al.*, 2012a,b; Finch *et al.*, 2012; Francis *et al.*, 2012b). One of the main papers showed that individuals presenting with a cough did not have a statistically significant difference in their chance of recovery whether they were given antibiotics or not, and no matter which antibiotic was prescribed (Butler *et al.*, 2009). GRACE has contributed to a change in clinical practice and to the development of new guidelines for the prescription of antibiotic drugs when individuals present at

their general practitioner (GP) with a cough.

The human 'genomics' part of the GRACE study is what will be presented here. LRTIs are one of the leading reasons for seeking primary medical care in Europe. They also account for far greater worldwide morbidity than a large collection of infectious and chronic diseases including breast cancer, type 2 diabetes, hepatitis B and C, malaria, tuberculosis and HIV/AIDS (Rowell *et al.*, 2012). However, mild LRTI has been largely neglected as a phenotype for human genetic association studies. Very little work has been done previously in this area and funding for this study was limited. Therefore, an approach was used which is more similar to a candidate gene approach than a modern genome-wide association study or high-throughput sequencing analysis. This can be considered a first step in discovering loci conferring protection from or susceptibility to mild LRTI. SNPs to genotype were chosen based on GWAS results for studies of susceptibility to more severe LRTIs and from the literature for previous associations with more severe bacterial and viral diseases. A selection of these SNPs were then genotyped in samples from the GRACE study; approximately 3000 LRTI cases and 3000 healthy controls matched by age, gender, date and location of collection.

There are a number of examples of the same SNP associating with different but related diseases (Hindorff *et al.*, 2012). For example, gain-of-function SNP rs2476601 in the *PTPN22* gene has been associated with type I diabetes, rheumatoid arthritis, systemic lupus erythematosus, Crohn's disease, vitiligo, as well as susceptibility to invasive pneumococcal disease (IPD) and leprosy (Barrett *et al.*, 2009, 2008; Begovich *et al.*, 2004; Bottini *et al.*, 2004; Chapman *et al.*, 2006; Gregersen *et al.*, 2009; Hakonarson *et al.*, 2007; Jin *et al.*, 2010; Kyogoku *et al.*, 2004; Plenge *et al.*, 2007; Rani *et al.*, 2009; Smyth *et al.*, 2004; Stahl *et al.*, 2010; The Wellcome Trust Case Control Consortium, 2007; Todd *et al.*, 2007). In addition, polymorphisms in the *TIRAP* and *CISH* genes have been associated with different infectious diseases

(including IPD, bacteraemia, malaria and tuberculosis) across multiple populations (Khor *et al.*, 2007, 2010).

5.2 Aim

The aim of this work was to use a two step approach to find genetic variants associated with susceptibility to LRTIs or subgroups of LRTI. The first step was to use a range of GWAS and candidate methods to choose SNPs to genotype. The second step was to genotype (using Sequenom) and analyse these SNPs in LRTI cases and controls.

5.3 Choosing SNPs for genotyping in GRACE

Seven methods were used to choose SNPs to genotype in GRACE, five of which utilised GWAS data. Two GWAS were performed: one in IPD and one for community-acquired pneumonia (CAP) sepsis. In addition, two meta analyses of GWAS were performed, one for these IPD and CAP sepsis GWAS, and another for these studies and an additional Kenyan bacteraemia GWAS. An additional GWAS comparison was performed between the top associating genes in a further IPD GWAS, and the IPD and CAP sepsis GWAS presented here. The two other methods used were candidate gene methods, to choose viral and bacterial candidate SNPs to genotype. These methods are all detailed below.

5.3.1 GWAS for susceptibility to IPD

Dr. Anna Rautanen performed a GWAS to study genetic susceptibility to IPD. IPD cases had *Streptococcus pneumoniae* bacteria cultured from a normally sterile site, and were collected in hospitals in Buckinghamshire and Oxfordshire, UK. Further information about this collection can be found in Materials and methods section 2.1.6.

Control GWAS data were obtained from the Wellcome Trust Case Control Consortium 2 (WTCCC2), and half of the individuals in these data were used as controls for this study. There were 271 cases, 1264 controls and 544,438 SNPs analysed in this GWAS. The Manhattan plot, QQ plot and a list of the top associating SNPs are presented in appendix figures A.1 and A.2 and table A.1. Top associating SNPs were prioritised based on p-value and a list of SNPs was created for Sequenom assay design.

5.3.2 GWAS for susceptibility to CAP sepsis

A GWAS analysis was performed for susceptibility to community-acquired pneumonia (CAP) sepsis, using GWAS data from UK individuals in the GenOSept (Genetics of Sepsis and Septic Shock) GWAS (GAinS study individuals). GAinS (Genomic Advances in Sepsis) is a collection of DNA from sepsis patients in the UK, and is a part of GenOSept, which is a Europe-wide study of genetic factors affecting outcome from sepsis and septic shock. Further information about these collections can be found in Materials and methods section 2.1. The other half (to that used in the IPD GWAS) of the individuals from the WTCCC2 project were used as controls for this analysis.

Figures 5.1 and 5.2 show the QC procedures used for these data before analysis. Figure 5.3 shows the first two multi-dimensional scaling (MDS) components for cases and controls. Figures 5.4 and 5.5 show the QQ plot and Manhattan plot from the analysis of this data. Table 5.1 shows the top associating SNPs from this analysis, and figure 5.6 shows regional association plots for the five best associating regions. Individual associating SNPs (with no associating SNPs nearby) were not considered reliable enough for further consideration in GRACE, as they are more likely to be spurious associations than those associations found in an association peak. Top associating SNPs were prioritised based on p-value, odds ratio and whether there are

other associating SNPs in the region, and added to the list of SNPs for Sequenom assay design.

5.3.3 Meta analysis of IPD and GAinS CAP sepsis GWAS

In order to identify SNPs associated with susceptibility to severe bacterial infection regardless of the organism involved, a meta analysis was performed with the results from the IPD and GAinS CAP sepsis GWAS. This meta analysis was performed using PLINK (Purcell, 2012; Purcell *et al.*, 2007) and 280,423 SNPs overlapped between the two analyses. Top associating SNPs were prioritised based on p-value and added to the list of SNPs for Sequenom assay design.

5.3.4 Meta analysis of IPD GWAS, GAinS CAP sepsis GWAS and Kenyan pneumococcal bacteraemia GWAS

In addition to the meta analysis performed above, a meta analysis was also performed with the results from the IPD GWAS, GAinS CAP sepsis GWAS and an analysis of susceptibility to bacteraemia caused by *S. pneumoniae* in Kenya. The Kenyan association analysis results were part of a GWAS analysis performed by Dr. Anna Rautanen. This study is part of the Wellcome Trust Case Control Consortium 2, and further information about the sample collection can be found in Materials and methods section 2.1. Top associating SNPs were prioritised based on p-value and added to the list of SNPs for Sequenom assay design.

5.3.5 Collaborative GWAS comparison

Thomas Benfield's group (Department of Infectious Diseases and Clinical Research Centre, Hvidovre University Hospital, Denmark) have also performed a GWAS for susceptibility to IPD in Danish children with pneumococcal meningitis (unpublished

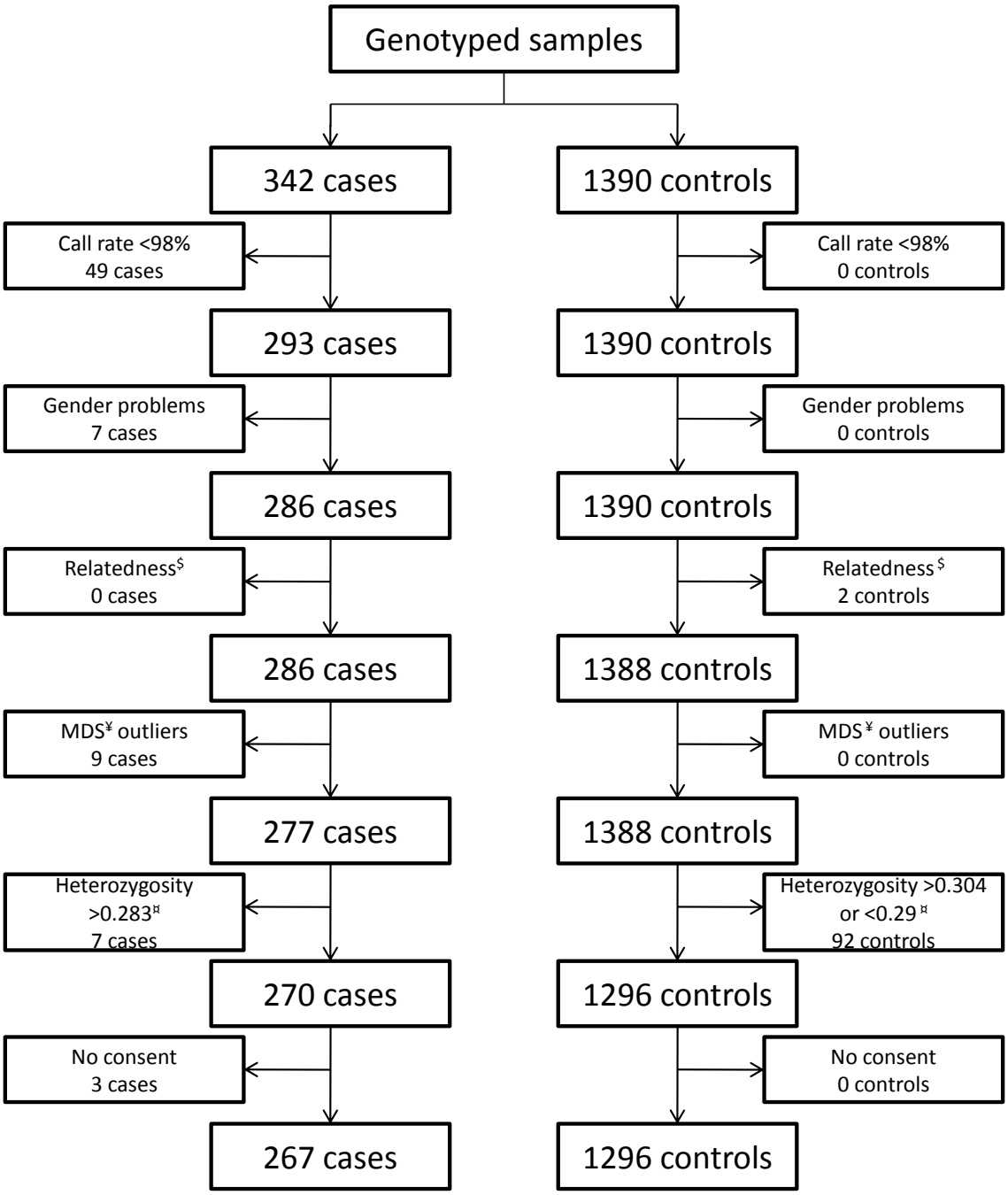


Figure 5.1: GAinS CAP sepsis GWAS sample QC procedures for cases and controls. $ When the proportion identical by state ≥ 0.2 between samples, one of each pair of samples was removed for high levels of relatedness. ¥ MDS: multi-dimensional scaling. MDS components were compared (component 1 vs. component 2, component 1 vs. component 3, component 1 vs. component 4 and component 3 vs. component 4) and any outliers in any comparison were removed from the analysis. ¤ Samples with high and low heterozygosity usually have low call rate. High heterozygosity implies contamination of the sample, and low heterozygosity can be a sign of uneven amplification of alleles following whole-genome amplification. Heterozygosity thresholds were chosen based on plots of heterozygosity against missingness.

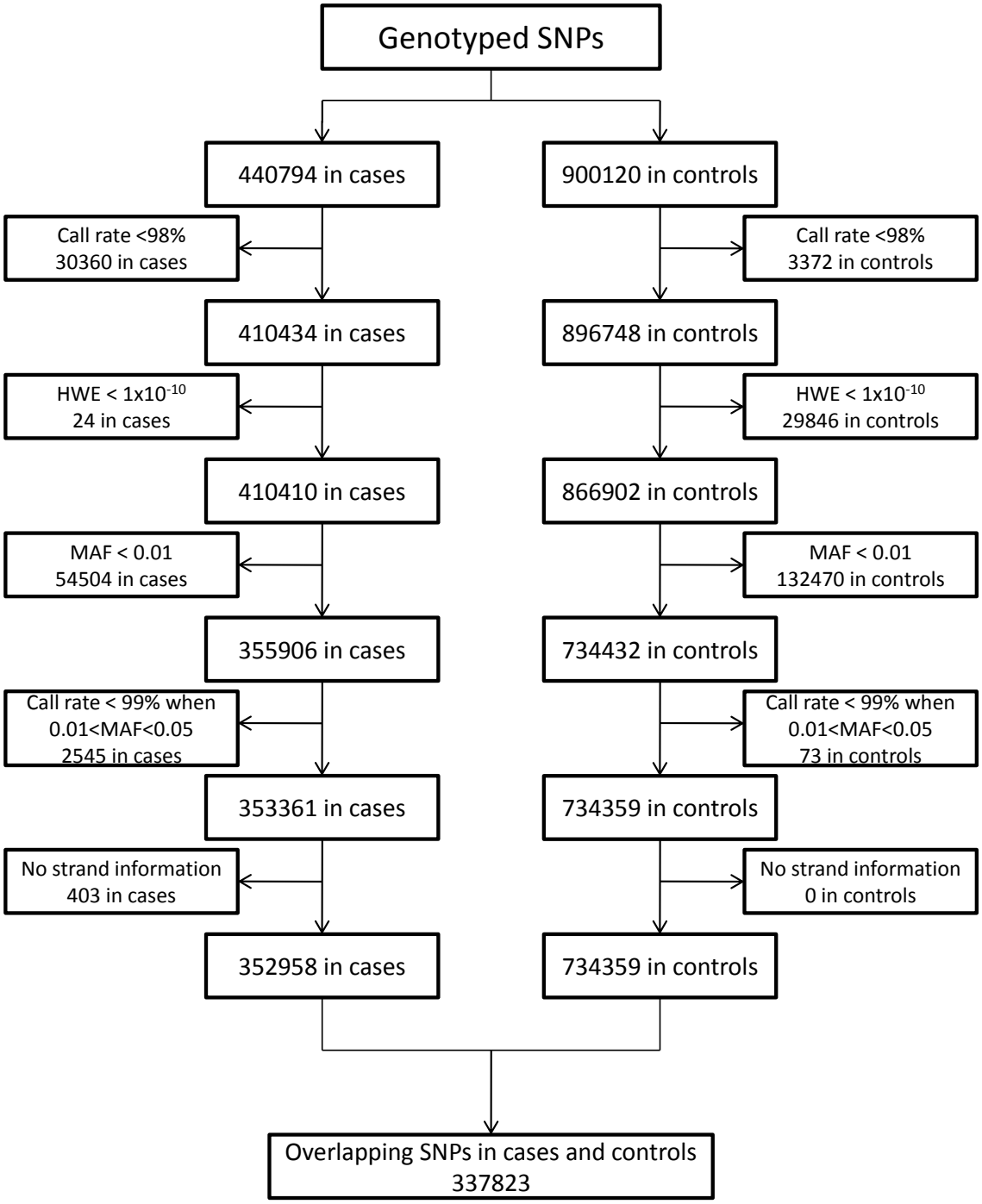


Figure 5.2: GAinS CAP sepsis GWAS SNP QC procedures for cases and controls. An additional 24 SNPs among the top hits were removed after inspection of the genotyping cluster plots. HWE=Hardy Weinberg Equilibrium. MAF=Minor Allele Frequency.

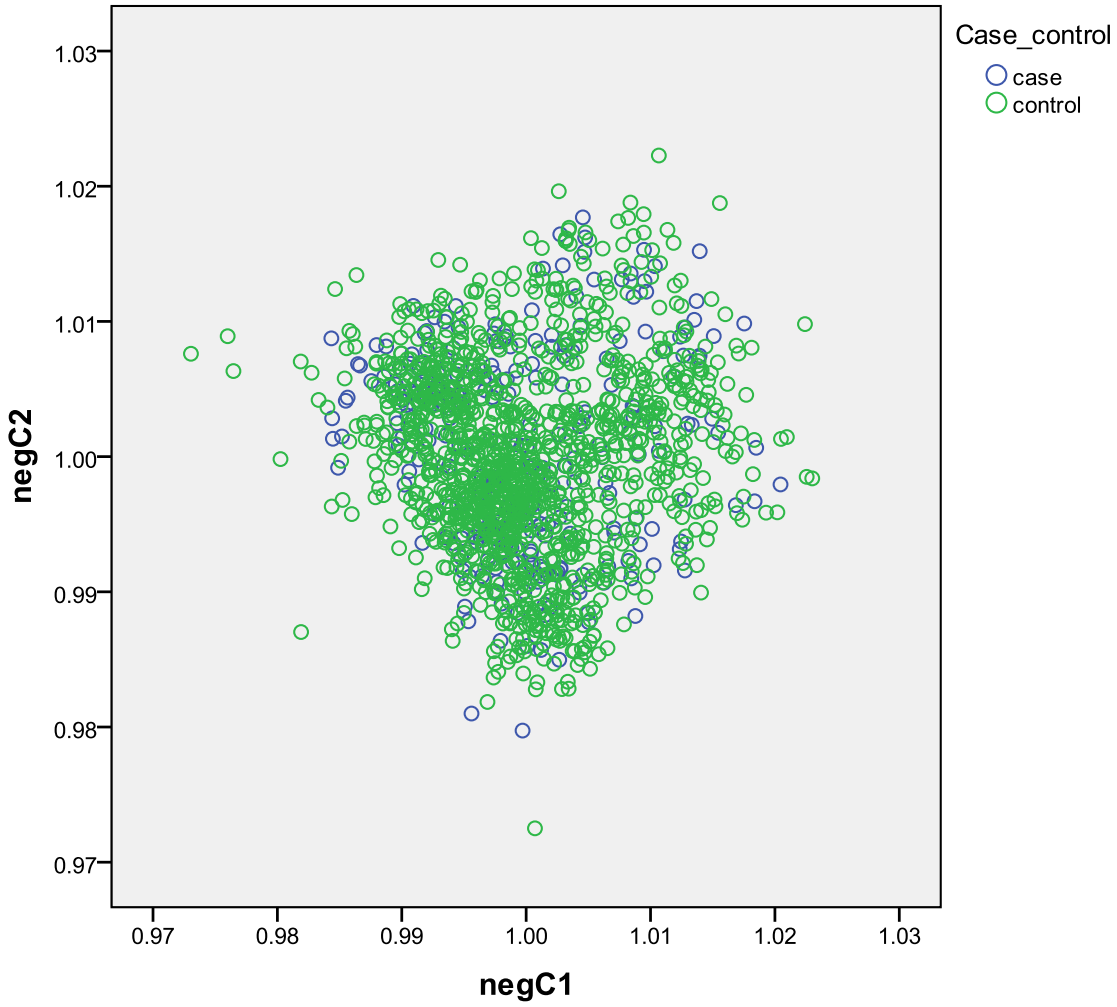


Figure 5.3: GAinS CAP sepsis GWAS multi-dimensional scaling (MDS) plot for the first two components, showing cases and controls. NegC1 is the inverse of the first component, and negC2 is the inverse of the second component.

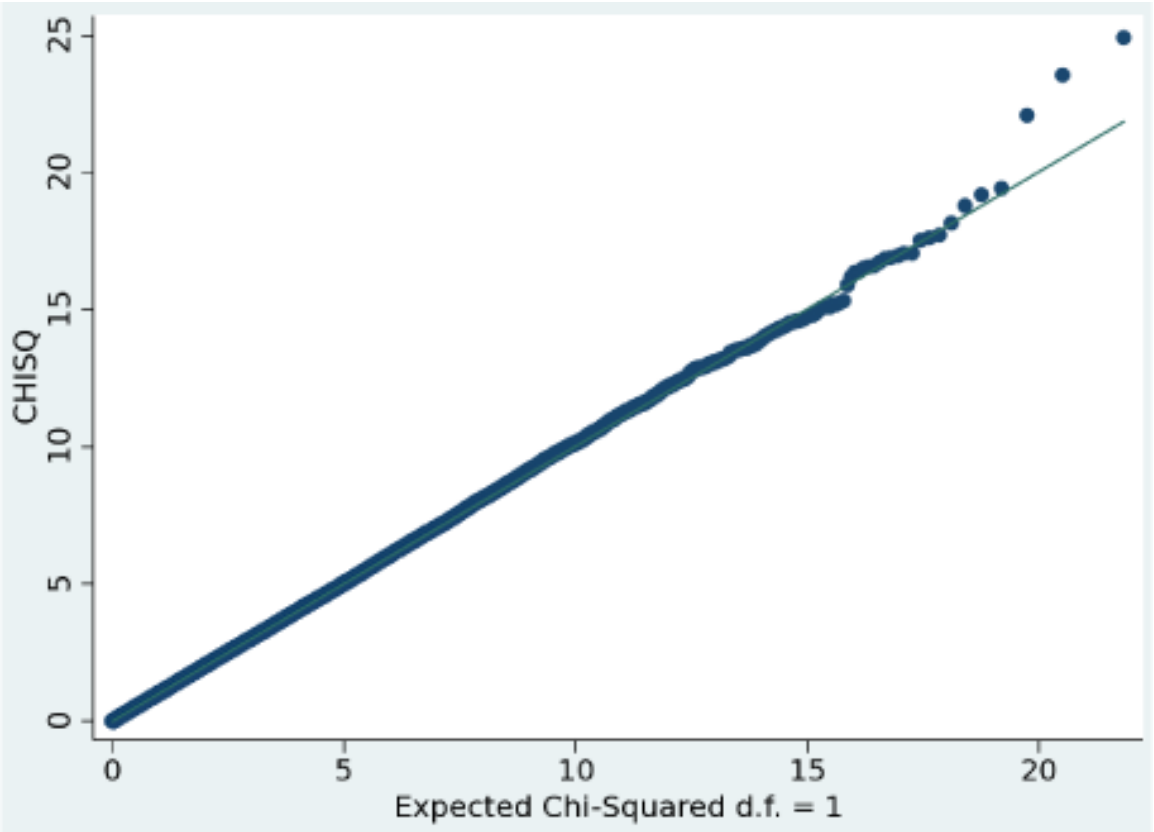


Figure 5.4: GAinS GWAS QQ plot

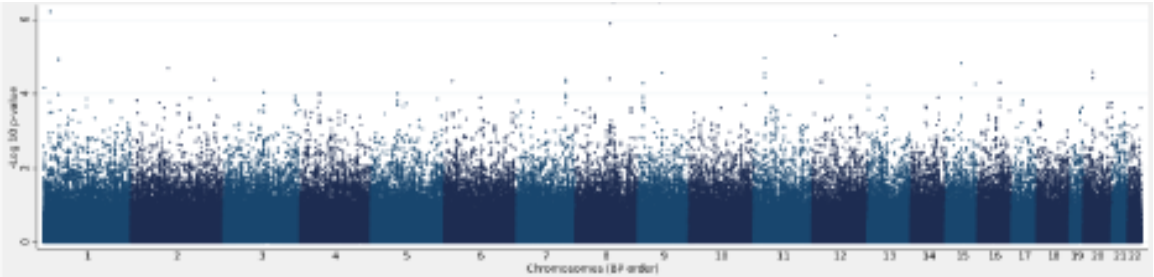


Figure 5.5: GAinS GWAS Manhattan plot

Table 5.1: The best regions of association found with the GAinS susceptibility to CAP sepsis GWAS. Genes in the regions were identified using HapMap Genome Browser release 28 at http://hapmap.ncbi.nlm.nih.gov. Gene descriptions found using www.genecards.org. The best regions of association (as determined by 2 or more SNPs within 100kb with p-values of 9.9×10^{-5} or less, or 5 or more SNPs within 100kb with p-values of 9.9×10^{-4} or less) are listed. [$] The number in brackets indicates the number of SNPs in the region with p-values of 9.9×10^{-5} or less. The most promising looking peaks are indicated with a * and are shown in figure 5.6.

	Chr	Number of SNPs [$]	Top SNP	Top p-value	Gene	Gene description
	1	2	rs2484714	1.2×10^{-5}	*WDR65*, *TMEM125*, *EBNA1BP2*	WD repeat-containing protein 65; transmembrane protein 125; EBNA1 binding protein 2
	1	6	rs12128857	4.1×10^{-4}	*SGIP1*	SH3-containing GRB2-like protein 3-interacting protein 1
	1	5	rs1982530	6.2×10^{-4}	*FMN2*	Formin 2-like
	2	4 (1)	rs10932875	4.0×10^{-5}	none	
	3	5	rs7651090	1.1×10^{-4}	*IGF2BP2*	IGF2 mRNA-binding protein 2
*	4	8 (1)	rs9995416	9.7×10^{-5}	*GABRA4*, *COX7B2*	Gamma-aminobutyric acid (GABA-A) receptor, subunit alpha 4; cytochrome c oxidase polypeptide VIIb2
	5	3 (1)	rs2707767	9.7×10^{-5}	none	
	6	6	rs293489	1.2×10^{-4}	*DOPEY1*, C6orf157	
*	7	8 (2)	rs17167808	4.0×10^{-5}	*SLC35B4*, *LRGUK*	UDP-Xylose/N-Acetylglucosamine transporter; leucine-rich repeat and guanylate kinase domain-containing protein
	9	6 (1)	rs7021164	5.2×10^{-5}	C9orf123	
	10	8	rs3950310	6.3×10^{-4}	*MAPK8*	MAP kinase 8
*	11	3 (3)	rs12278965	1.1×10^{-5}	*NELL1*	Protein kinase C-binding protein
	11	3 (1)	rs7925363	9.3×10^{-5}	*LUZP2*	Leucine zipper protein 2
	13	3 (1)	rs7992703	5.7×10^{-5}	*SAP18*, *MRP63*	Sin3-associated polypeptide p18; mitochondrial ribosomal protein bMRP63
*	15	4 (2)	rs17238047	1.5×10^{-5}	*VPS13C*	Vacuolar protein sorting-associated protein 13C
	16	6	rs1423982	4.3×10^{-4}	none	
	16	5 (1)	rs3826248	4.8×10^{-5}	*CALB2*, *FLJ11171*	Calbindin 2
	18	10	rs6506348	3.0×10^{-4}	*TTMA*, *L3MBTL4*	Transmembrane protein, lethal(3)malignant brain tumour-like protein 4
	19	6	rs381983	2.8×10^{-4}	*ZNF155*, *ZNF230*, *ZNF222*, *ZNF223*, *ZNF45*	Zinc finger proteins
*	20	3 (2)	rs2760549	2.5×10^{-5}	*MACROD2*, *KIF16B*	MACRO domain-containing protein 2, kinesin-like protein

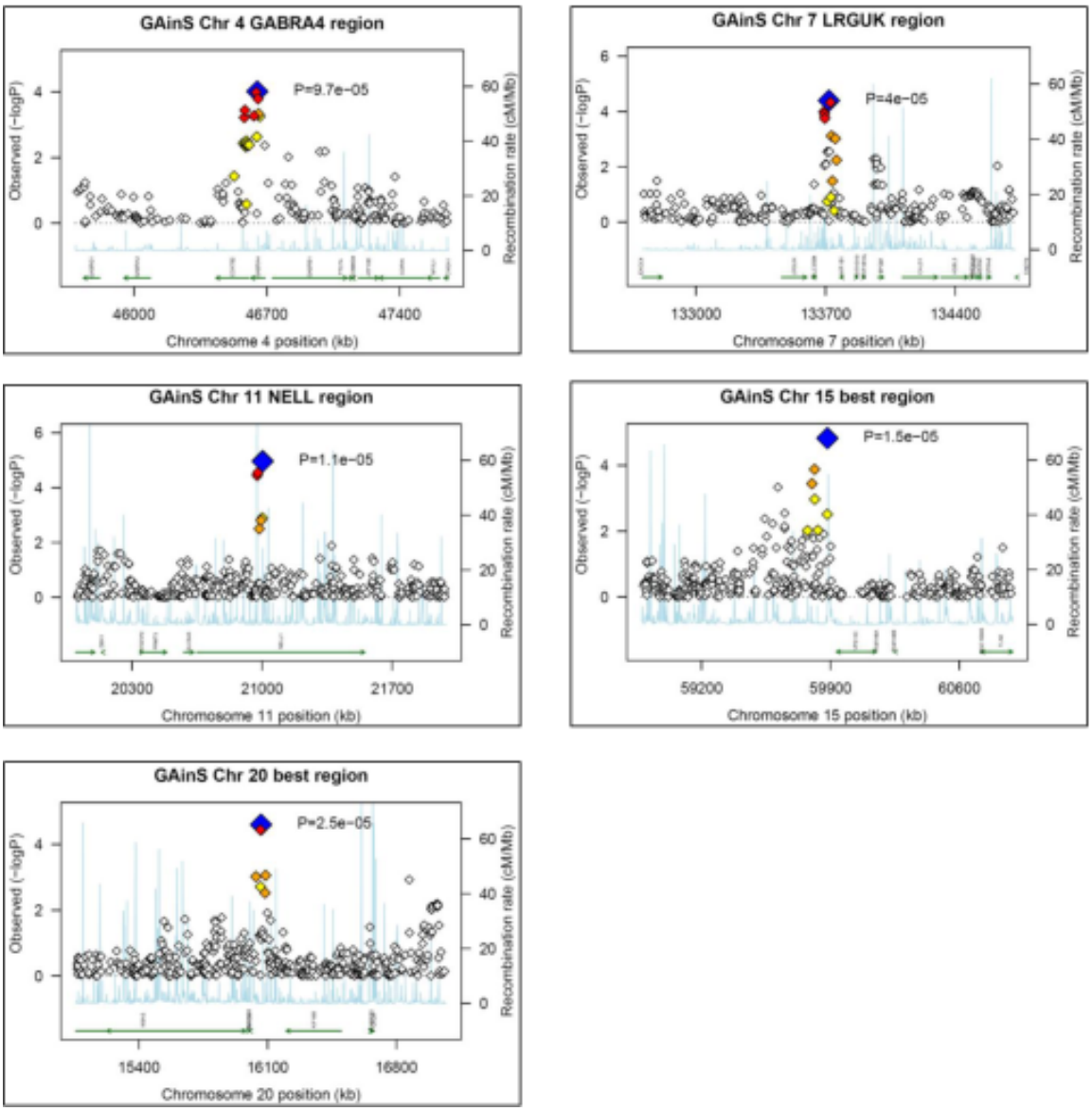


Figure 5.6: GAinS CAP sepsis GWAS regional association plots for the five best looking peaks in loci identified from the GAinS CAP sepsis GWAS. P-values (as -log10 values) are plotted as a function of genomic position (with NCBI Build 36). The SNP with the most significant p-value in each region is shown as a blue diamond, and its p-value is shown. Other SNPs are coloured based on their r^2 with the top SNP, as calculated with PLINK from GWAS data (red: $r^2 \geq 0.8$; orange: $0.5 \leq r^2 < 0.8$; yellow: $0.2 \leq r^2 < 0.5$; white: $r^2 < 0.2$). Estimated recombination rates are plotted in pale blue (as taken from HapMap). Genes within the region of the hit SNPs are labelled and the direction of transcription for each gene is indicated by the corresponding green arrow.

data). The top 500 hits from this GWAS were compared with all hits in the meta analysis of IPD and GAinS CAP sepsis that had p-values less than 0.05 (approximately 10,000 SNPs). The three GWAS had been performed with different chips, and there were no overlapping SNPs between the two lists. Therefore, the genes associated with SNPs in both lists were compared, and any overlapping genes from the two GWAS were considered further. There were 195 genes associated with Thomas Benfield's SNPs and 70 of these are found in both lists. All genes where at least two SNPs (one from each screen) were found with D' linkage disequilibrium (LD) >0.8 and where both SNPs had p-values <0.01 were considered further. The best associating SNPs in these regions from either screen, and non-synonymous SNPs in high LD with these SNPs were added to the list of SNPs for Sequenom assay design.

5.3.6 Candidate gene SNPs

Candidate gene SNPs which have shown associations previously in the Hill group (Wellcome Trust Centre for Human Genetics, University of Oxford, UK) in severe bacterial infection sample sets were selected and prioritised based on the strength and significance of the association previously seen. In addition, a selection of other SNPs which had shown associations in bacterial infection or autoimmune sample sets in the literature were also considered. This list was added to the list of SNPs for Sequenom assay design.

5.3.7 Literature review for viral candidate SNPs

Microbiological characterisation of each GRACE sample showed that a much higher than anticipated fraction of the individuals had a LRTI caused by a virus (45%). Therefore, a literature review was undertaken to identify viral candidate SNPs of interest to genotype in the GRACE samples. Genetic association studies for viral infection were searched for using PubMed, and a list of SNPs which associate with

viral infection was compiled. HIV infection associations were not included, as the biology of HIV infection is very different to that of viral LRTI. These SNPs were prioritised based on the number of samples in the study, the number of different studies which had shown an association with the SNP in the same direction, the significance and strength of the association seen, and whether there were other studies which showed a lack of association with the SNP. This list was added to the list of SNPs for Sequenom assay design.

5.4 Assay design for genotyping candidate SNPs in GRACE

A list of SNPs to try to genotype was created for Sequenom iPLEX assay design based on the seven procedures detailed above (section 5.3). In theory, a maximum of 40 SNPs can be multiplexed in a Sequenom iPLEX assay. A preliminary assay of 27 SNPs was designed using candidate gene SNPs (section 5.3.6), and top hits from the IPD GWAS (section 5.3.1). Subsequently, all other SNPs on the list were considered for the design of two further assays. Top priority SNPs (the SNPs with the lowest p-values from each procedure described above) were used to generate two assays containing as many SNPs as possible. Further SNPs were then added to these assays, in order of priority (again based on p-values from each procedure above), until two assays contained as many SNPs as possible, with as many top priority SNPs included as possible. Table 5.2 lists all SNPs included in the assays, the assay they were included in, the gene associated with the SNP and why the SNP was chosen to be genotyped in GRACE. There were 93 SNPs genotyped in 67 genes, of which 12 failed genotyping QC and could not be analysed. Sequenom genotyping was performed by Genomics Services in the Wellcome Trust Centre for Human Genetics, Oxford.

Table 5.2: All SNPs genotyped by Sequenom in the GRACE cohort, including associated genes, and why each SNP was chosen for genotyping. SNPs shown in brackets are those which could not be analysed due to bad clustering in Sequenom genotyping or failure to pass QC. SNPs labelled as being shortlisted from 'IPD GAinS KBpneumo meta analysis' are those from the meta analysis of the IPD GWAS, GAinS CAP sepsis GWAS and Kenyan pneumococcal bacteraemia GWAS. Candidate gene SNPs are listed with a reference for the publication which has shown an association for the same SNP, or list which infectious disease has shown a preliminary association in the Hill group. SNPs chosen based on the viral literature review list the reference associated with the previous study showing the best association. Where no gene is present in the region, the chromosome and base pair position are given in the format chromosome:base pair position.

SNP	Associated gene(s)	Assay	SNP shortlisted from
rs10991437	*ABCA1*	2	IPD GWAS
rs17079516	*CSMD1*	2	IPD GWAS
rs17390811	*CSMD1*	2	IPD GWAS
rs12545740	*DUSP26*, *RNF122*	1	IPD GWAS
rs2898575	*EMID2*	1	IPD GWAS
rs6963365	*EMID2*	1	IPD GWAS
rs6947185	*EMID2*	3	IPD GWAS
rs4693264	*FAM190A*	1	IPD GWAS
rs12040948	*IL19*	1	IPD GWAS
(rs11119602)	*IL19*	1	IPD GWAS
rs2317280	*KHDRBS3*	2	IPD GWAS
(rs6578241)	*KHDRBS3*	1	IPD GWAS
rs9508811	*MPHOSPH8*	1	IPD GWAS
(rs2112626)	*ODZ2*	2	IPD GWAS
(rs4802515)	*RYR1*	2	IPD GWAS
rs10142146	*SLC24A4*	3	IPD GWAS
rs9296536	*TNFRSF21*	1	IPD GWAS
rs485861	*TRERF1*	2	IPD GWAS

Continued on next page

SNP	Associated gene(s)	Assay	SNP shortlisted from
rs4714591	*TRERF1*	3	IPD GWAS
rs9966505	*ZNF521*	1	IPD GWAS
(rs6953653)	7:89637017	3	IPD GWAS
rs6041512	20:12630784	2	IPD GWAS
(rs9995416)	*GABRA4*, *COX7B2*	3	GAinS CAP sepsis GWAS
rs2760549	*MACROD2*, *KIF16B*	3	GAinS CAP sepsis GWAS
rs12278965	*NELL1*	2	GAinS CAP sepsis GWAS
rs7992703	*SAP18*, *MRP63*, *SKA3*	2	GAinS CAP sepsis GWAS
rs9580025	*SAP18*, *MRP63*, *SKA3*	2	GAinS CAP sepsis GWAS
rs17167808	*SLC35B4*, *LRGUK*	3	GAinS CAP sepsis GWAS
rs293489	*UBE3D*, *DOPEY1*	3	GAinS CAP sepsis GWAS
rs2484714	*WDR65*, *TMEM125*, EBNA1BP2	2	GAinS CAP sepsis GWAS
rs4607596	8:78902651	3	GAinS CAP sepsis GWAS
rs16906665	8:80186148	2	GAinS CAP sepsis GWAS
rs10124357	9:79429800	3	GAinS CAP sepsis GWAS
rs10508857	10:37304120	2	GAinS CAP sepsis GWAS
rs1395513	12:58386125	3	GAinS CAP sepsis GWAS
rs704582	*CMSS1*, *FILIP1L*	2	IPD GAinS meta analysis
rs1416079	1:83097138	3	IPD GAinS meta analysis
rs10171575	2:52844226	3	IPD GAinS meta analysis
rs2062848	3:6132954	2	IPD GAinS meta analysis
rs11762595	7:8366512	3	IPD GAinS meta analysis

Continued on next page

SNP	Associated gene(s)	Assay	SNP shortlisted from
rs10238794	7:31215377	3	IPD GAinS meta analysis
(rs10508347)	10:8232422	3	IPD GAinS meta analysis
rs1938859	11:101547723	2	IPD GAinS meta analysis
(rs1431310)	17:54111270	2	IPD GAinS meta analysis
rs3778348	*HSF2*	3	IPD GAinS KBpneumo meta analysis
rs1003845	*SPOCK1*	2	IPD GAinS KBpneumo meta analysis
rs1434642	*SPOCK1*	2	IPD GAinS KBpneumo meta analysis
rs11726723	4:26456267	3	IPD GAinS KBpneumo meta analysis
rs13107432	4:143444744	2	IPD GAinS KBpneumo meta analysis
rs554531	10:8201728	3	IPD GAinS KBpneumo meta analysis
rs9989461	17:10216686	3	IPD GAinS KBpneumo meta analysis
rs497116	*CASP12*	1	Candidate: preliminary evidence of association in bacteraemia
rs414171	*CISH*	1	Khor *et al.* (2010)
rs622502	*CISH*	1	Khor *et al.* (2010)
rs2239751	*CISH*	1	Khor *et al.* (2010)
rs6768330	*CISH*	1	Khor *et al.* (2010)
rs148685070	*CISH*	1	Khor *et al.* (2010)
rs35337543	*IFIH1*	3	Nejentsev *et al.* (2009)
rs35667974	*IFIH1*	3	Nejentsev *et al.* (2009)
rs12006030	*LCN2*	3	Candidate: preliminary evidence of association in IPD
rs3138053	*NFKBIA*	1	Chapman *et al.* (2007)
rs529948	*NFKBIE*	1	Chapman *et al.* (2007)
rs600718	*NFKBIZ*	1	Chapman *et al.* (2009)

Continued on next page

SNP	Associated gene(s)	Assay	SNP shortlisted from
rs616597	*NFKBIZ*	1	Chapman *et al.* (2009)
rs2476601	*PTPN22*	1	Chapman *et al.* (2006)
rs8177374	*TIRAP*	1	Khor *et al.* (2007)
rs5743551	*TLR1*	1	Candidate: preliminary evidence of association in IPD
rs5743595	*TLR1*	1	Dixon *et al.* (2007)
rs4986791	*TLR4*	1	Yuan *et al.* (2008)
rs137992076	*TLR6*	1	Candidate: preliminary evidence of association in IPD
(rs760477)	*TONSL*	1	Chapman *et al.* (2010b)
rs8034191	*AGPHD1*	3	Pillai *et al.* (2009)
rs7671167	*FAM13A*	2	Cho *et al.* (2010)
rs601338	*FUT2*	2	Thorven *et al.* (2005)
rs1828591	*HHIP*	3	Pillai *et al.* (2009)
(rs10757212)	*IFNA5*	2	Janssen *et al.* (2007)
(rs643070)	*IFNA13*	2	Janssen *et al.* (2007)
rs2430561	*IFNG*	3	Gentile *et al.* (2003)
rs1800872	*IL10*	2	Janssen *et al.* (2007)
(rs12979860)	*IL28B*	2	Thomas *et al.* (2009)
rs1800795	*IL6*	2	Doyle *et al.* (2010)
rs2071430	*MX1*	2	He *et al.* (2006)
rs2660	*OAS1*	3	Knapp *et al.* (2003)
rs731236	*VDR*	2	Janssen *et al.* (2007)
rs10735810	*VDR*	2	Janssen *et al.* (2007)
rs4263503	*ADAMTS12*	3	Collaborative GWAS comparison

Continued on next page

SNP	Associated gene(s)	Assay	SNP shortlisted from
rs4504387	*ADAMTS12*	2	Collaborative GWAS comparison
rs6481955	*C10orf112*	3	Collaborative GWAS comparison
rs6722763	*DGKD*	2	Collaborative GWAS comparison
rs17139769	*HDAC9*	3	Collaborative GWAS comparison
rs734784	*KCNS1*	3	Collaborative GWAS comparison
rs10964832	*IFNB1*	3	T. Benfield most replicating in IPD

5.5 Quality control for Sequenom SNP genotyping data

Cluster plots were checked for every SNP and for every plate of samples genotyped. Any badly clustering SNPs were removed from the analysis. Figure 5.7 shows some examples of the best and worst clustering SNPs. Table 5.3 shows the number of SNPs and samples removed from the analysis during each step of the QC procedure, for each assay. Unlike in a GWAS analysis, with relatively few SNPs genotyped in this study, it was impossible to analyse population stratification in the dataset. Individuals could only be removed based on self-assessed ethnicity. All QC and data analysis was performed using PLINK (Purcell, 2012; Purcell *et al.*, 2007).

5.6 Subgroup analyses

Genotyping was completed on the approximately 3000 GRACE LRTI cases and 3000 matched controls. Many subgroups were analysed within the GRACE cohort. Table 5.4 lists the subgroups which were analysed, and the number of cases and controls included in each subgroup analysis for each assay (since QC was performed separately

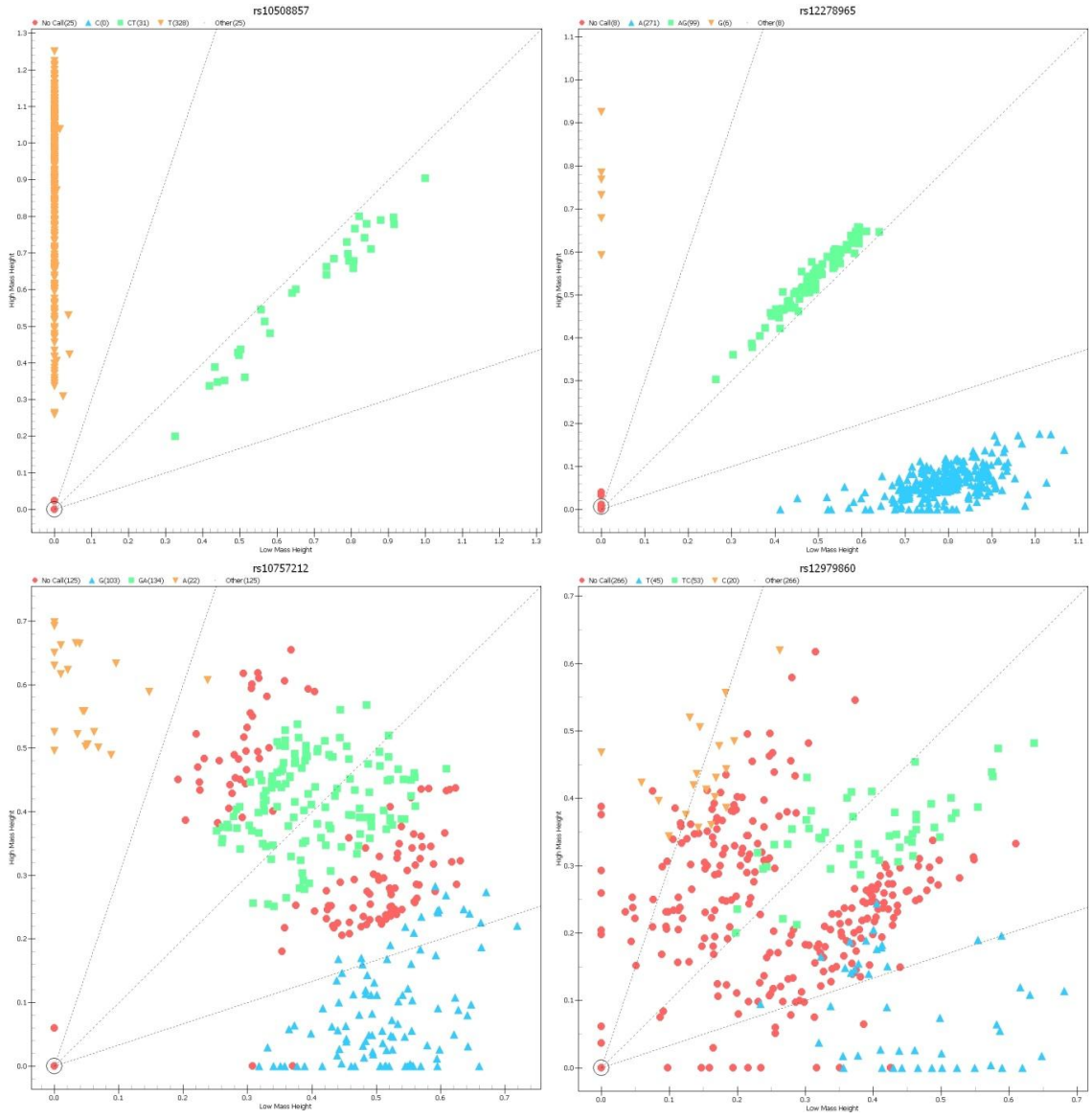


Figure 5.7: Examples of very good cluster plots: one rarer SNP and one common (top panels), and very bad cluster plots (bottom panels) which demonstrate bad genotyping and cannot be used as reliable genotype calls

Table 5.3: A summary of the number of samples and SNPs removed before analysis of GRACE genotyping data, and for what reason.

	Sample removal		
	Assay 1	**Assay 2**	**Assay 3**
Number of samples genotyped	5869	5869	5869
Ethnicity	254	254	254
Duplicate	16	16	16
Bad sample	9	9	9
Genotyping rate $< 80\%$	56	140	179
Number of samples remaining	5534	5450	5411
	SNP removal		
	Assay 1	**Assay 2**	**Assay 3**
Number of SNPs genotyped	27	34	32
Genotyping rate <90%	3	4	3
HWE $<9\text{x}10^{-4}$	2	0	0
Number of SNPs analysed	22	30	29

for each assay, each analysis contained slightly different individuals).

Table 5.4: The subgroups analysed in the GRACE study, with the number of cases and controls in each subgroup for each assay shown. All analyses except the matched analysis include country as a covariate.

	Assay 1		**Assay 2**		**Assay 3**	
Analysis	Number of cases	Number of controls	Number of cases	Number of controls	Number of cases	Number of controls
Basic case control analysis	2839	2662	2801	2619	2771	2608
Matched analysis	2407	2407	2346	2346	2306	2306
Consolidation	205	2662	200	2620	197	2609
Recurrent LRTI	847	2662	833	2620	827	2609
Viral	1271	2662	1255	2620	1240	2609
Non-viral	1568	2662	1546	2620	1531	2609
Rhinovirus	533	2662	534	2620	518	2609
Influenza A B	271	2662	269	2620	266	2609
Coronavirus	199	2662	195	2620	194	2609
Rhinovirus, smoking as covariate	533	2655	534	2613	518	2603
Rhinovirus, smokers	223	596	224	585	212	582
Rhinovirus, non-smokers	310	2059	310	2028	306	2021
Case control, smoking as covariate	2838	2655	2800	2612	2770	2602
Case control, smokers	799	596	788	585	770	582
Case control, non-smokers	2039	2059	2012	2027	2000	2020
Haemophilus influenzae	176	2662	175	2620	170	2609
Streptococcus pnuemoniae	155	2662	148	2620	145	2609
H. influenzae and *S. pneumoniae*	306	2662	298	2620	291	2609
Samples positive for bacteria in sputum or nasopharangeal swab	393	2662	384	2620	378	2609

5.6.1 Viral analysis

Table 5.5 lists the numbers of viral positive cases and controls for all viruses identified in individuals in the GRACE project. Those viruses with more than 200 cases were analysed separately as a subgroup on their own. Only the viruses that were overrep-

resented among cases compared to controls were combined for the viral analysis.

Table 5.5: Viruses identified in GRACE cases and controls. Viruses indicated with a * are those included in the viral analysis.

	Virus	Number of cases virus positive	Number of controls virus positive
*	Rhinovirus	586	72
*	Parainfluenza virus (PIV)	73	8
*	Human metapneumovirus (hMPV)	126	3
*	Respiratory syncytial virus (RSV)	143	11
*	Influenza A and B	299	6
*	Coronavirus	208	27
	Adenovirus	38	24
	Bocavirus	13	16
	WU (Washington University)	45	39
	KI (Karolinska Institute)	27	17

5.6.2 Bacterial analysis

Four bacterial subgroups were analysed in GRACE. The only individual bacterial subgroups considered large enough for analysis on their own were *H. influenzae* and *S. pneumoniae*. Cases with infections caused by these organisms were also combined for an analysis of susceptibility to *H. influenzae* and *S. pneumoniae*. In addition, any samples positive for any bacteria in a sputum sample or nasopharangeal swab were combined for the biggest bacterial subgroup (the "all bacterial" group). The bacteria found in GRACE were *H. influenzae*, *S. pneumoniae*, *Mycoplasma pneumoniae*, *Bordetella pertussis*, *Legionella pneumophila* and *Chlamydophila pneumoniae*.

5.6.3 Taking smokers into account

When analysing possible confounding factors for this study, a significant difference in the proportion of smokers between cases and controls was noticed, in particular in the rhinovirus subgroup. The main difference between cases and controls was in the proportion of current smokers. Therefore, all non-smokers and all individuals that had smoked in the past were grouped together as non-current smokers. Table 5.6 shows p-values and odds ratios for the association between smoking status and infection in the different subgroups within GRACE. Smoking associates with a number of these groups. The rhinovirus group was large enough that it was possible to reliably perform analysis of susceptibility to rhinovirus with just current smokers or non-current smokers in this subgroup. This was also performed for the basic case control analysis.

Table 5.6: P-values and odds ratios for the association between smoking status (current vs. non-current smokers) and infection in different subgroups within GRACE

Group	**p-value**	**OR**
Basic case control	5×10^{-7}	1.358
Matched analysis	8.5×10^{-7}	1.376
Consolidation	0.005	1.537
Recurrent	6.3×10^{-7}	1.543
Viral	9×10^{-8}	1.50
Non-viral	0.002	1.248
Rhinovirus	2×10^{-19}	2.485
Influenza A and B	0.355	1.068
Coronavirus	0.078	1.287
All bacterial	2×10^{-4}	1.547
H. influenzae	0.02	1.453
S. pneumoniae	0.001	1.749
H. influenzae and *S. pneumoniae*	0.001	1.560

5.7 Results

As 81 SNPs were analysed in 65 different genes, a Bonferroni corrected p-value of 7.7×10^{-4} (0.05/65) was considered to be statistically significant. All such results are presented in table 5.7 and below.

5.7.1 *MPHOSPH8* gene region association

The SNP which associates across the most GRACE subgroups is rs9508811 in the gene *MPHOSPH8* (table 5.8). This SNP was one of the top hits from the IPD GWAS ($p=4.3 \times 10^{-5}$). In GRACE, this SNP associates with viral and rhinovirus infections, and has the lowest p-value (2.3×10^{-4}) in the rhinovirus analysis with smoking as a covariate. Although the association is seen in all individuals with rhinovirus infection, it is clear that this SNP is more important in non-current smokers, as the highest odds ratio was seen in this subgroup, while the association was not significant in the current smokers group.

Figure 5.8 shows the linkage disequilibrium (LD) between rs9508811 and the surrounding SNPs, as calculated from the 1000 Genomes project. This figure implicates three possible causal genes for this association, since the high LD around the top SNP is so extensive. Fine-mapping was performed with the aim of narrowing down the association region (see section 5.8).

5.7.2 *IL6* gene association

Table 5.9 shows the association of the *IL6* -174 promoter polymorphism SNP (rs1800795) with different subgroups in GRACE, the most significant of which is in the analysis of smokers with LRTI. This *IL6* SNP was chosen for genotyping based on two studies showing an association between genotype and illness severity in respiratory syncytial virus (RSV) (Gentile *et al.*, 2003) and rhinovirus (Doyle *et al.*,

Table 5.7: GRACE SNPs and subgroups passing the Bonferroni corrected threshold of $p<7.7x10^{-4}$. All analyses use country as a covariate. NS stands for non-significant.

Gene/s in SNP region	**SNP**	**SNP chosen from**		**Viral analysis**	**Rhinovirus analysis**	**Rhinovirus smoking as covariate**	**Rhinovirus smokers only**	**Rhinovirus non-smokers only**	**All smokers**
MPHOSPH8	rs9508811	IPD GWAS	**p-value**	0.00068	0.00029	0.00023	NS	0.00055	NS
			OR	1.22	1.332	1.346	NS	1.411	NS
IL6	rs1800795	viral susceptibility literature review	**p-value**	NS	NS	NS	0.00093	NS	0.00027
			OR	NS	NS	NS	1.468	NS	1.344
SAP18, *MRP63*, *C13orf3*	rs7992703	GAinS susceptibility GWAS	**p-value**	NS	NS	NS	NS	NS	0.00072
			OR	NS	NS	NS	NS	NS	1.579

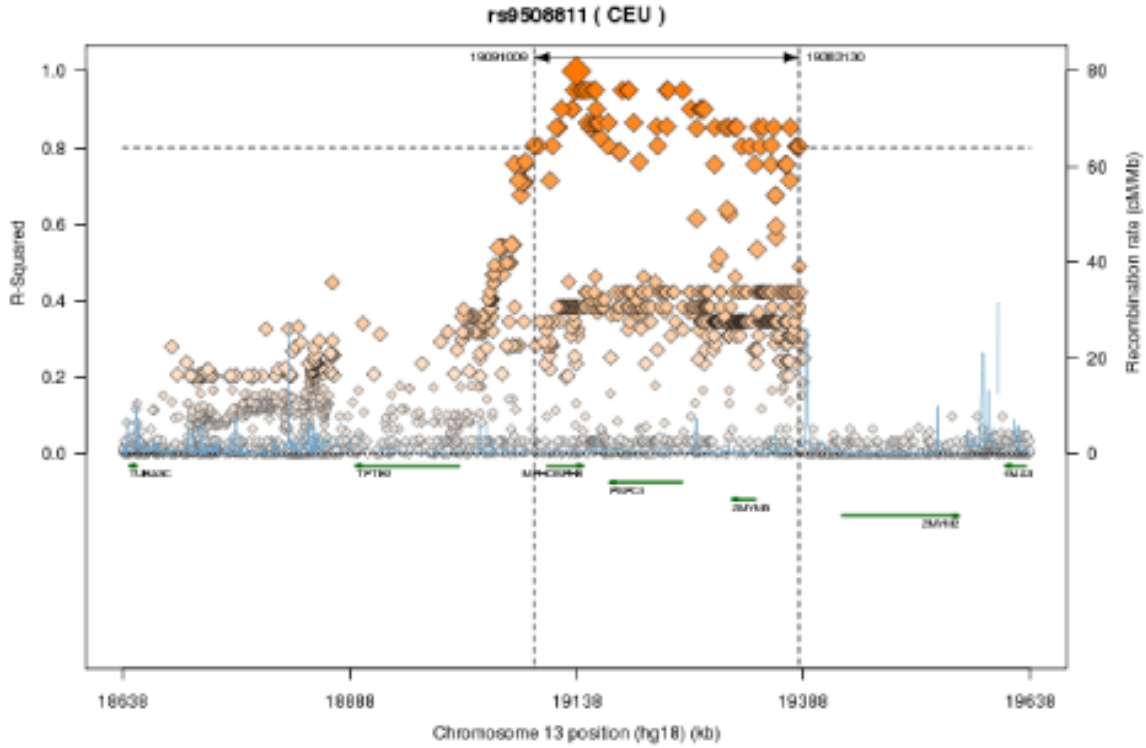


Figure 5.8: A linkage disequilibrium (LD) plot for SNPs with GRACE top SNP rs9508811 using data from the 1000 Genomes project. Each orange diamond represents a SNP, with the SNP of interest as the largest orange diamond. The y-axis represents the r-squared measure of LD with the top SNP. Each diamond's horizontal location is defined by its position on chromosome 13 (x-axis). Pale blue lines on the graph denote the recombination rate as shown by the right hand y-axis. Genes are shown with green arrows to denote the direction of transcription. All data are from the 1000 Genomes CEU (individuals from Utah, USA of European descent) data pilot 1, and the plot was made using SNAP (Johnson *et al.*, 2008)

Table 5.8: rs9508811 SNP associations for various subgroups in the GRACE samples. Odds ratios (OR) are for the minor allele A.

Subgroup	p-value	OR
Case control analysis	0.00464	1.142
Matched analysis	0.00700	1.143
Viral analysis	0.000681	1.220
Non-viral analysis	0.1606	1.082
Rhinovirus analysis	0.000292	1.332
Rhinovirus, smoking as covariate	0.000228	1.346
Rhinovirus, smokers only	0.0879	1.270
Rhinovirus, non-smokers only	0.000546	1.411
Case control, smoking as covariate	0.00335	1.149
Case control, smokers only	0.631	1.047
Case control, non-smokers only	0.00165	1.187

2010) infection. The original studies showed an association in the recessive test for this SNP. In GRACE, the effect size was greater in the recessive test than the allelic test (OR=1.74 in the rhinovirus smokers subgroup), although the association was less significant (p=0.0056 in the rhinovirus smokers subgroup).

Table 5.9: SNP associations for rs1800795 in *IL6* with different subgroups within GRACE. OR is the odds ratio for the minor allele C.

	All Smokers	All non-smokers	Rhinovirus smokers	Rhinovirus non-smokers
p-value	0.000273	0.7882	0.000932	0.02956
OR	1.344	0.9877	1.468	0.8193

5.7.3 *SAP18* gene region association

Two SNPs in the *SAP18* gene region were genotyped in the GRACE samples because these SNPs had two of the lowest p-values seen in the GAinS CAP sepsis susceptibility GWAS (regional association plot shown in figure 5.9). The causal gene for this association is unclear, as there are three genes in the region, any of which could be responsible for the signal given the high LD in the region, as shown in figure 5.10.

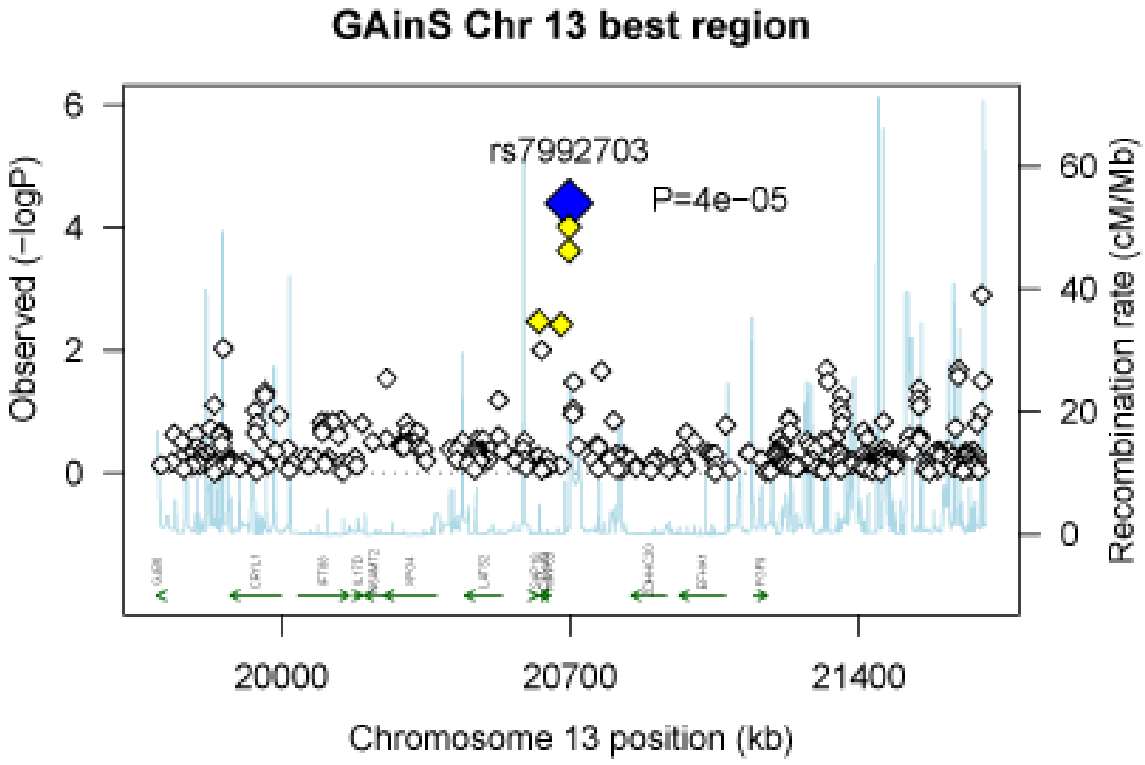


Figure 5.9: A regional association plot for the association seen in the GAinS CAP sepsis susceptibility GWAS for SNPs in the *SAP18* gene region. P-values (as -log10 values) are plotted as a function of genomic position (with NCBI Build 36). The SNP with the most significant p-value is shown as a blue diamond, and its p-value is shown. Other SNPs are coloured based on their r^2 with the top SNP, as calculated with PLINK from GWAS data (yellow: $0.2 \leq r^2 < 0.5$; white: $r^2 < 0.2$). Estimated recombination rates are plotted in pale blue (as taken from HapMap). Genes within the region of the hit SNP are labelled and the direction of transcription for each gene is indicated by the corresponding green arrow. The three genes in the region are *SAP18*, *C13orf3* and *MRP63*.

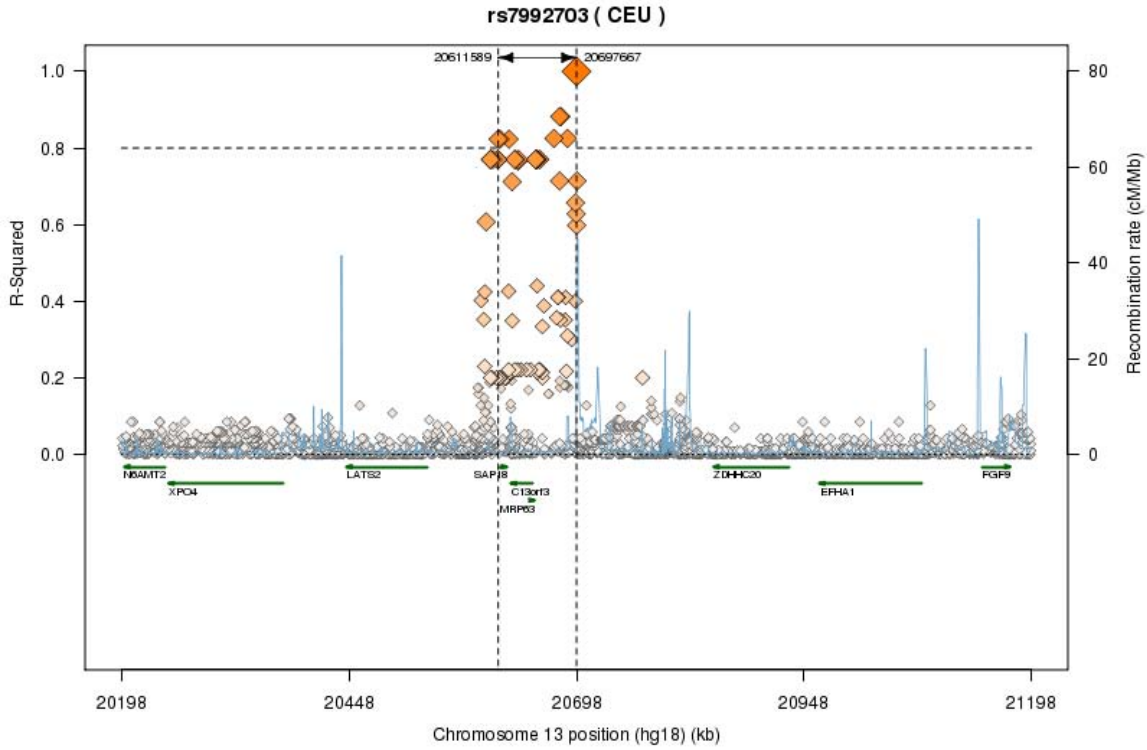


Figure 5.10: A linkage disequilibrium (LD) plot for SNPs with GRACE top SNP rs7992703 using data from the 1000 Genomes project. Each orange diamond represents a SNP, with the SNP of interest as the largest orange diamond. The y-axis represents the r-squared measure of LD with the top SNP. Each diamond's horizontal location is defined by its position on chromosome 13 (x-axis). Pale blue lines on the graph denote the recombination rate as shown by the right hand y-axis. Genes are shown with green arrows to denote the direction of transcription. All data are from the 1000 Genomes CEU (individuals from Utah, USA of European descent) data pilot 1, and the plot was made using SNAP (Johnson *et al.*, 2008)

Table 5.10 shows p-values and odds ratios for different subgroups for the two SNPs in this gene region that were genotyped in GRACE samples. The top SNP shows a significant association with susceptibility to LRTI but, like the *IL6* SNP, only among smokers. Like in *IL6*, this association is more significant in the analysis of all infection in smokers, but the odds ratio is stronger in the samples with rhinovirus infection that smoke.

Table 5.10: SNP associations for two SNPs in the *SAP18* gene region with different subgroups within GRACE. OR is the odds ratio for the minor allele.

SNP		**All smokers**	**All non-smokers**	**Rhinovirus smokers**	**Rhinovirus non-smokers**
rs7992703	p-value	0.000717	0.8901	0.003007	0.1097
	OR	1.579	1.01	1.779	0.7745
rs9580025	p-value	0.04404	0.3403	0.04819	0.1914
	OR	1.22	0.9487	1.32	0.8658

Since this association pattern is so similar to that seen in *IL6*, a pairwise epistasis analysis was performed using PLINK (Purcell, 2012; Purcell *et al.*, 2007) to investigate whether the effects of the SNP in *IL6* and those in the *SAP18* gene region complement each other. There was no significant epistatic association when just smokers were included in the analysis. When all individuals were included there was a statistically significant epistatic association result ($p=3.2x10^{-5}$) for the *IL6* SNP and the rs9580025 SNP in the *SAP18* gene region, but not with rs7992703.

The genes in this region are *SAP18* (sin3A associated protein, 18kDa), *MRP63* (mitochondrial ribosomal protein 63) and *C13orf3* (also known as *SKA3*; spindle and kinetochore associated complex subunit 3).

5.7.4 *EMID2* gene region association

Three SNPs in the *EMID2* gene region were genotyped in GRACE as they were top hits in the IPD GWAS. These SNPs did not reach the Bonferroni corrected p-value threshold of $7.7\text{x}10^{-4}$ (considered significant after correction for 65 genes in GRACE) in any subgroup. However, two of the SNPs (rs2898575 and rs6963365) did associate with p-values at the 10^{-3} level in bacterial subgroups (see table 5.11). Therefore, a meta analysis was performed for the two bacterial subgroups most similar to the phenotype in the IPD GWAS. The results of this analysis are also shown in table 5.11.

Table 5.11: Association analysis and meta analysis results for IPD GWAS and GRACE bacterial subgroups for three SNPs genotyped in the *EMID2* gene region. OR is the odds ratio for the reference allele.

Analysis	SNP	Reference allele	p-value	OR
IPD	rs2898575	G	$1.9\text{x}10^{-5}$	1.506
	rs6963365	A	$6.08\text{x}10^{-6}$	1.543
	rs6947185	C	$3.37\text{x}10^{-6}$	1.571
GRACE *S. pneumoniae*	rs2898575	G	0.08433	1.226
	rs6963365	A	0.05386	1.255
	rs6947185	C	0.3811	1.116
GRACE *H. influenzae* and *S. pneumoniae*	rs2898575	G	0.009094	1.253
	rs6963365	A	0.007946	1.258
	rs6947185	C	0.3202	1.095
Meta analysis of IPD and GRACE *S. pneumoniae*	rs2898575	G	$1.16\text{x}10^{-5}$	1.388
	rs6963365	A	$2.46\text{x}10^{-6}$	1.421
	rs6947185	C	$2.93\text{x}10^{-5}$	1.381
Meta analysis of IPD and GRACE *H. influenzae* and *S. pneumoniae*	rs2898575	G	$1.76\text{x}10^{-6}$	1.361
	rs6963365	A	$6.15\text{x}10^{-7}$	1.378
	rs6947185	C	0.000106	1.295

5.7.5 Association between associating SNPs and smoking

Since the main findings of this study have been shown to associate with infection in only smokers or ony non-smokers, it is important to assess whether these SNPs are in fact associated with smoking itself. Therefore, I analysed these SNPs for current smokers compared to non-current smokers. The results of this analysis are shown in table 5.12. These analyses were non-significant for all three SNPs.

Table 5.12: Analysis was performed for the associating SNPs and susceptibility to smoking, with current smokers labelled as cases and non-current smokers labelled as controls. Country was used as a covariate for this analysis.

SNP	Gene/s in SNP region	P value	OR
rs9508811	*MPHOSPH8*	0.2863	0.9422
rs1800795	*IL6*	0.34	1.045
rs7992703	*SAP18*, *MRP63*, *C13orf3*	0.1511	1.112

5.8 *MPHOSPH8* fine-mapping

5.8.1 Choosing SNPs for fine-mapping

Fine-mapping was performed with the aim of narrowing down the *MPHOSPH8* association region. In order to economically fine-map such a large region, SNPs to genotype were chosen very carefully. To be sure of including every possible SNP in the region which could be the cause of the association, a list was made of all SNPs in the region from 18380-19440kb on chromosome 13 (3812 SNPs in 1000 Genomes). SNPs were chosen for genotyping which were in medium LD (r^2>0.2 when D'>0.8 or r^2>0.5) with the top SNP rs9508811 and SNP rs9508887 (which was almost as significant in the IPD GWAS). The r^2 LD between these SNPs is 0.95. This list was then filtered by the following criteria:

1. Keep all SNPs in r^2>0.2 when D'>0.8 or r^2>0.5 with rs9508811 and rs9508887
2. Remove SNPs genotyped in the IPD GWAS which were less significant than the two hit SNPs
3. Remove SNPs which are in higher LD with non-associating IPD GWAS SNPs than with rs9508811 or rs9508887
4. Remove SNPs which are in complete (r^2=1) LD with each other
5. Remove rare SNPs from the list (MAF<0.05)

This list of 121 SNPs was then prioritised by their LD with the hit SNP. Then two assays were designed to include as many of these SNPs as possible. When a SNP could not fit into the iPLEX, one of the SNPs in complete LD with it was tried, until all possibilities were used up. Finally, all remaining SNPs left after the first removal criteria in the list above were tried, to have as many SNPs as possible in the designed iPLEXes. The final design included 44 SNPs which fit into two iPLEXes for genotyping.

5.8.2 Sequenom fine-mapping genotyping analysis

Rhinovirus cases (540) and a similar number (576) of controls (most were matched to the cases) were genotyped for the chosen SNPs using Sequenom. Following cluster plot, sample and SNP QC for the genotyping data (see section 2.3.5.5), analysis was performed using PLINK (Purcell, 2012; Purcell *et al.*, 2007). Analysis was also performed again for the genotypes of rs9508811 using only the samples used in this fine-mapping study. As fewer controls were genotyped in fine-mapping than in the original study, the p-value for rs9508811 is less significant than originally seen (p=0.011 rather than p=0.000292). The effect size for this SNP (OR=1.298) is, however, very sim-

ilar to what was seen with the full control set (OR=1.332). Figure 5.11 shows the association results for this and the other genotyped SNPs in the fine-mapping.

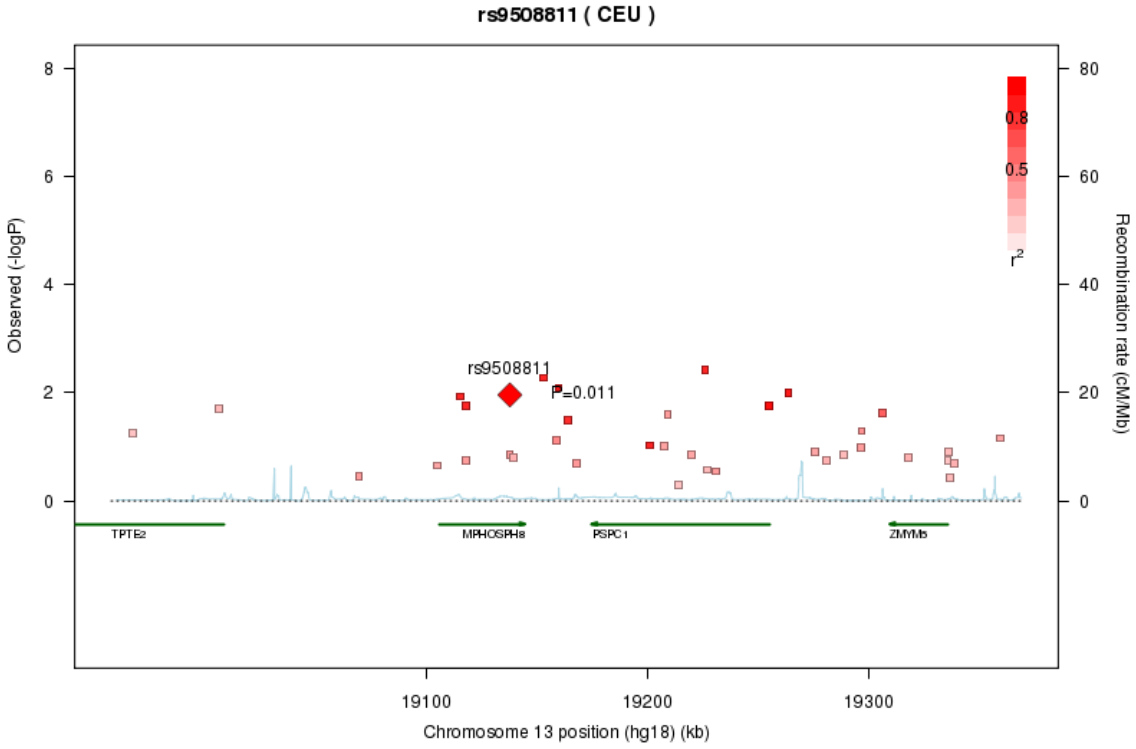


Figure 5.11: Regional association plot for fine-mapping of SNPs around rs9508811 in the *MPHOSPH8* gene. Association results are for susceptibility to rhinovirus infection in mild LRTI in Europe. SNP rs9508811 is indicated by a large red diamond.

There are four SNPs in this fine-mapping analysis with a lower p-value than rs9508811. All these SNPs also had stronger odds ratios than rs9508811, and not all of them were found in the *MPHOSPH8* gene. The highest OR was seen with rs117170334 (OR=1.88), which is located in the *TPTE2* (transmembrane phosphoinositide 3-phosphatase and tensin homolog 2) gene, therefore implicating another possible functional gene association. In order to attempt to pinpoint the functional variant for this association, imputation was performed.

5.8.3 Imputation of more SNPs

Imputation was performed for a 1Mb region around the fine-mapped SNPs, using IMPUTE v2.2 (Howie *et al.*, 2009). Further details about this method can be found in Materials and methods section 2.3.5.5. Association analysis was then performed for 13183 SNPs using SNPTEST. SNPs were filtered for minor allele frequency (8984 were removed for MAF<0.01), the control Hardy Weinberg equilibrium p-value (none were removed for $p<1x10^{-10}$) and for the proper info value (3718 SNPs were removed for info<0.8). Figure 5.12 shows the association results from imputation using this cleaned set of SNPs. None of the imputed SNPs are more significant that any of the genotyped SNPs.

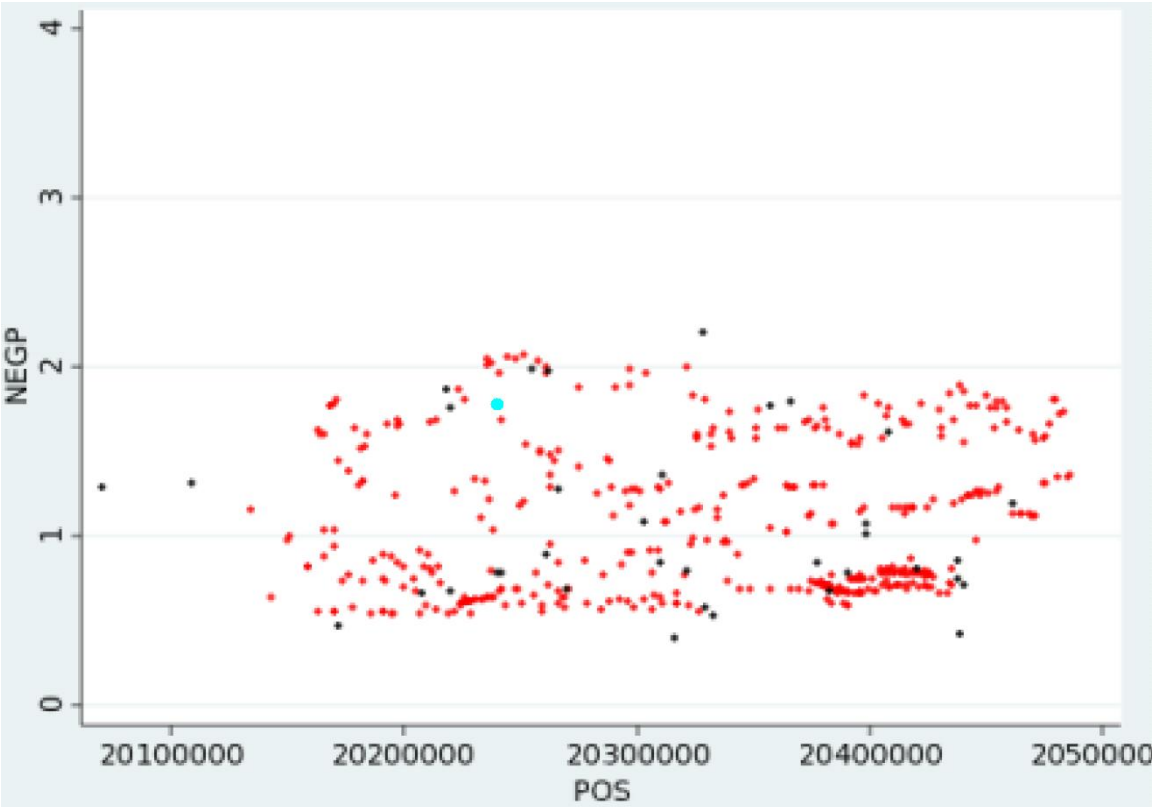


Figure 5.12: Association plot for genotyped (black) and imputed (red) SNPs in the *MPHOSPH8* region. Genetic position (NCBI build 37) on chromosome 13 is shown on the x axis, and -log10 p-value (NEGP) on the y-axis. The SNP rs9508811 is indicated by a pale blue dot.

5.9 Discussion

The aim of this study was to identify SNPs which associate with lower respiratory tract infections (LRTIs) in Europe. The results of a large Europe-wide study analysing associations between 81 SNPs and LRTI are presented here. Statistically significant (after correction for the 65 genes studied) associations were found in three different gene regions, and one additional association can be considered as replication of the results of a GWAS for susceptibility to IPD.

The lowest observed p-value ($2.3\text{x}10^{-4}$) in this study was for the SNP rs9508811 in the *MPHOSPH8* gene region (table 5.8), in an analysis of susceptibility to rhinovirus infection with smoking as a covariate. This SNP showed the highest odds ratio (1.411) in the rhinovirus subgroup with only non-smokers included. In order to narrow down the region of association, and determine the causal gene, fine-mapping followed by imputation was performed. However, this was unsuccessful in determining a causal gene, and if anything, the associating region was shown to be even larger than initially suspected (figure 5.11).

There are four genes in the *MPHOSPH8* region; *MPHOSPH8* itself (M-phase phosphoprotein 8), *TPTE2* (transmembrane phosphoinositide 3-phosphatase and tensin homolog 2), *PSPC1* (paraspeckle component 1) and *ZMYM5* (zinc finger, MYM-type 5). At first glance, none of these genes seem to be of particular interest in terms of susceptibility to rhinovirus infection in mild LRTI. However, *MPHOSPH8* is expressed more highly in CD4+ and CD8+ T-cells than in any other cell type, which implicates its importance in the immune response. In addition, MPHOSPH8 has been shown to mediate E-cadherin gene silencing (Kokura *et al.*, 2010).

E-cadherins are important adhesion molecules that contribute to the structural and immunological function of the airway epithelium (Nawijn *et al.*, 2011). The airway epithelium is an important defence against invading microbial pathogens. Rhinovirus infection has been shown to significantly reduce the mRNA levels, protein

levels and expression levels of E-cadherin in primary cultured nasal epithelial cells (Yeo and Jang, 2010), which would certainly affect the immunological function of the airway epithelium. Interestingly, cigarette smoke has also been shown to decrease E-cadherin expression (Liu *et al.*, 2010). With this potential joint alteration in the function of E-cadherins by rhinovirus and by the *MPHOSPH8* gene (and an effect of cigarette smoke), it seems likely that the *MPHOSPH8* gene is most likely the causal gene for this association result. A recent GWAS for susceptibility to malaria identified an associating SNP close to the gene for another adhesion molecule; the tight-junction protein MARVELD3 (Timmann *et al.*, 2012). It is therefore possible that adhesion molecules in general may be involved in susceptibility to a range of infectious diseases.

The strongest odds ratio observed for any SNP associating to a statistically significant level after correction for the 65 genes studied was 1.579 for a SNP in the *SAP18* gene region (table 5.10), for the case control analysis just including smokers ($p=7.2x10^{-4}$). A SNP in *IL6* also associated strongly with this subgroup ($p=2.7x10^{-4}$, OR=1.344) (table 5.9). Interestingly, both these SNPs showed associations with stronger odds ratios (though the association in the *SAP18* SNP was not significant after correction) in the rhinovirus subgroup which only included smokers.

In addition, an epistasis analysis of the two SNPs showed no association when just smokers were included, nor when all samples were included. However, an epistasis analysis of all samples for the *IL6* SNP and another SNP in the *SAP18* gene region did show a highly significant association ($p=3.2x10^{-5}$). Although it is difficult to dissect this result, it certainly shows that there is something of interest in these two genes in relation to susceptibility to lower respiratory tract infections.

The genes in the *SAP18* region are *SAP18* (sin3A associated protein, 18kDa), *MRP63* (mitochondrial ribosomal protein 63) and *SKA3* (spindle and kinetochore associated complex subunit 3). The LD in this gene region is very high (figure 5.10),

and so in anticipation of the same result as that from the *MPHOSPH8* fine-mapping, fine-mapping was not performed for this association result. *SAP18* seems to be the most likely candidate gene in this gene region. Whole genome expression profiling analyses in *Drosophila melanogaster* embryos showed differentially expressed genes in *dsap18* mutants compared to embryos with a *dsap18* transgene in the mutant background. Many of these downregulated genes are involved in the immune response. It was also shown that *dsap18* mutant flies were more susceptible to fungal infection by *Aspergillus fumigatus* but not to *Escherichia coli* or *Micrococcus luteus* bacteria (Costa *et al.*, 2011). In addition, another study showed that SAP18 directly binds to INI1 in yeast, *in vitro* and *in vivo*. INI1 is a component of the SWI/SNF complex, which interacts with HIV-1 IN, a virally encoded protein required for integration of viral cDNA into host chromosomes. SAP18 also binds to IN itself *in vitro* and *in vivo*. SAP18 is a component of the Sin3a-HDAC1 complex, and incorporation of an inactive mutant of HDAC1 was shown to decrease infectivity of HIV-1 (Sorin *et al.*, 2009). Although these results are not directly applicable to the findings presented here, they do show that it is plausible that SAP18 is involved in susceptibility to mild LRTI. It is not clear why the original associations observed with *SAP18* SNPs should only occur in smokers within GRACE, however.

Since the *IL6* SNP was a viral candidate chosen based on previous studies for infectious disease, there is no need to hypothesise its role in infection. IL6 is a proinflammatory cytokine involved in innate and adaptive immunity (Abbas and Lichtman, 2005; Foster *et al.*, 2000). The genotyped SNP is found in the promoter region of the *IL6* gene, and has been linked to the plasma levels of IL6 (Bennermo *et al.*, 2011; Fishman *et al.*, 1998; Gentile *et al.*, 2003). Smoking has also been shown to determine plasma levels of IL6 (Bennermo *et al.*, 2011). Since smoking is a proinflammatory process which will affect levels of many cytokines (not just IL6), and increases susceptibility to rhinovirus, it is certainly likely that further modulation of IL6 levels

is likely to interfere with the balance of cytokines in the body, leading to increased susceptibility to rhinovirus.

Considering that the *IL6* SNP was chosen for genotyping based on studies for severity of illness, it would be very interesting to be able to study this phenotype in GRACE. Clinicians in GRACE recorded the presence or absence of 14 symptoms, and patients filled in a symptom diary (Francis *et al.*, 2012b). However, empirical data was not routinely collected, so severity of illness could not be accurately analysed in this dataset.

Two SNPs (rs2898575 and rs6963365) in the *EMID2* (EMI domain containing 2) gene region associated with susceptibility to bacterial LRTI in two subgroups (see table 5.11). This association was not considered significant after correction for the 65 genes analysed in this study. However, using bacterial subgroups with similar phenotypes to the discovery sample set for these SNPs (the IPD samples), meta analysis was performed. The GRACE sample set can therefore act as a replication set for the IPD study, although the association does not reach genome-wide significance ($p<5x10^{-8}$). In this meta analysis, the lowest observed p-value was $6.2x10^{-7}$ for rs6963365 with the GRACE *H. influenzae* and *S. pneumoniae* subgroup and the IPD GWAS data.

It is difficult to say which gene in the region is responsible for this signal, because of the extensive LD in the region, and since very little is known about the two genes (*EMID2* and *RABL5*). Two different SNPs in the *EMID2* gene region (rs4729705 and rs4729707, which associate at the 10^{-5} level following imputation in the IPD GWAS) are associated with the expression of *CNPY4* (canopy 4 homolog) in B-cells (Fairfax *et al.*, 2012). CNPY4 regulates the cell surface expression of toll-like receptor (TLR)4 (Konno *et al.*, 2006) and TLR1 (Hart and Tapping, 2012). Toll-like receptors are a family of proteins which act as pattern recognition receptors for microbe-derived molecules and stimulate innate immune responses to these microbes

(Abbas and Lichtman, 2005). TLR4 recognises ligands from mainly gram-negative bacteria (Abbas and Lichtman, 2005), and TLR1 recognises cell wall components from gram-positive bacteria (Hart and Tapping, 2012). *H. influenzae* is a gram-negative bacteria and *S. pneumoniae* is gram-positive. Interestingly, a non-synonymous SNP in the *TLR1* gene (I602S) is a protective allele for tuberculosis and leprosy (Johnson *et al.*, 2007; Ma *et al.*, 2007; Wong *et al.*, 2010a), and this SNP prevents the cell surface expression of TLR1. However, TLR1 cell surface expression can be rescued by knock-down of *CNPY4* (Hart and Tapping, 2012). Therefore, *CNPY4* plays an integral role in the innate immune response and could explain the association observed in the *EMID2* gene region.

This is the only genetic study for susceptibility to mild LRTI that I know of. It is therefore difficult to say whether these findings are likely to be applicable to individuals outside of Europe. Since I have been unable to fully correct for population stratification in the sample set, it is possible that unknown outlying ethnicities within the dataset are skewing the results, although this is less likely due to the close matching of controls to cases in this cohort. It is also impossible to say that these are the only gene regions which associate with susceptibility to LRTI, since I have only studied a very small number of SNPs. Many of these concerns could be addressed with further study using this important sample set, studying more SNPs using for instance a GWAS approach. Unfortunately some of the subgroups within GRACE (for instance for particular viral infections) are too small to reliably analyse and it would be very difficult to collect more cases for these subgroups.

This study has identified three novel genetic regions which associate with subgroups of LRTI. I have identified two genes/regions (*IL6* and the *SAP18* gene region) with SNPs that associate with susceptibility to infection in smokers only, and a gene region (*MPHOSPH8*) with an unknown causal gene which associates with susceptibility to rhinovirus infection in LRTI overall and in non-smokers, but not current

smokers. I have performed analysis for the three SNPs in these regions and shown that these SNPs do not associate with smoking.

I have also replicated a finding from a GWAS for susceptibility to IPD, in a much milder, related sample set (LRTI caused by *S. pneumoniae*). It is interesting to note that each of the associations found was genotyped in GRACE for a different reason, which shows that choosing many sources of SNPs to genotype in the study design was effective. It is curious that the *MPHOSPH8* association was found with rhinovirus despite being based on a SNP discovered from a bacterial infection GWAS. It is possible that this reflects that these mild viral infections predispose to severe bacterial infection, although this hypothesis is only speculation at this point.

Cold and flu remedies are the second largest over-the-counter category after pain relief, and reached sales of £428 million in the year 2010-2011 in the UK alone. Cough remedies account for the largest sales for a single symptom in this category, and have shown the largest growth of any symptom within the category (nearly twice as high as the next symptom) (Nielsen, 2011). This, combined with the observation that LRTIs account for far greater worldwide morbidity than a large collection of infectious and chronic diseases including breast cancer, type 2 diabetes, hepatitis B and C, malaria, TB and HIV/AIDS (Rowell *et al.*, 2012), suggest that LRTI is an important disease to study.

Unfortunately it has not been possible to identify causal genes for some of the associations found, but functional work could elucidate the causal mechanisms affecting susceptibility to mild LRTI. In addition, the knowledge that different mechanisms could make smokers susceptible to mild LRTI could lead to further study to find the best remedies for smokers and non-smokers separately, thereby improving the morbidity associated with mild LRTI.

Functional work certainly needs to be undertaken to identify the causal genes for some of the associations presented here. Further genotyping using the GRACE

cohort, even including a GWAS scale study would be important for identifying further genes of interest for susceptibility to mild LRTI, which could then be targeted for therapy. A GWAS for susceptibility to rhinovirus could be performed first, as this is the largest homogenous subgroup. This study has acted as proof of principle that genetic associations with mild disease do exist, and could pave the way for future work with genetics and mild disease.

Chapter 6

CAP sepsis susceptibility GWAS

6.1 Introduction

It is likely that genes affecting susceptibility are different to genes affecting outcome from infectious disease (Petersen *et al.*, 2010), and therefore it is important to study both survival from and susceptibility to the same diseases. A GWAS has not yet been performed for susceptibility to sepsis that I know of. Therefore, a project was undertaken to utilise the genome-wide genotypes from the GenOSept sepsis survival study, to analyse susceptibility to CAP sepsis.

The process to analyse data in this study of susceptibility to CAP sepsis was similar to that of a recent study of the genetics of multiple sclerosis (MS). This MS study utilised GWAS data from MS cases from around the world, and collected GWAS data for healthy controls from matching countries from different studies. These data were then combined to analyse susceptibility to MS in 9,772 MS cases and 17,376 controls, at 465,434 SNPs across the genome which were genotyped in all the analysed cohorts (International Multiple Sclerosis Genetics Consortium *et al.*, 2011).

All community-acquired pneumonia (CAP) cases in the GenOSept study were used as cases in this study. The GenOSept samples were collected from patients all

over Europe, and therefore control data were required from the same countries to study susceptibility. Six healthy control datasets from the control arms of studies across Europe were used as controls.

6.2 Aim

The aim of this work was to develop a reliable pipeline to combine GWAS datasets and analyse these for susceptibility to community-acquired pneumonia sepsis.

6.3 Methods

6.3.1 Cohorts for analysis

Table 6.1 lists the cohorts used in this analysis, the chips used to genotype them, and the number of samples and SNPs in the sample set before and after QC. Details of QC procedures are given in Materials and methods section 2.4.7. Details of how the data were processed are given in section 6.3.2 below. Further details for all cohorts are presented in Materials and methods sections 2.1 and 2.2. Cases were genotyped on two separate chips. The GenOSept study used the Affymetrix 5.0 chip. This included individuals from the GAinS study which were separately analysed for susceptibility to CAP sepsis in the UK in the GRACE study in chapter 5. Additional GAinS study individuals (the GAinS 2011 study set) were genotyped in 2011 using the Illumina Omni Express chip, and these are also used as cases in this study.

6.3.2 Pipeline for QC and analysis

All eight cohorts in this study underwent the same QC procedures, to ensure that all genotyped data were of a high quality. Since the datasets were genotyped on a range of different GWAS chips, imputation was performed to increase the number of

Table 6.1: Table showing the project each study set came from, the country which the samples were from, the chip used, and the number of samples and SNPs before and after QC. Further information about all sample cohorts can be found in Materials and methods sections 2.1 and 2.2

Name of dataset	**Country**	**Number of samples before QC**	**Number of samples after QC**	**Chip used**	**Number of SNPs before QC**	**Number of SNPs after QC**
GenOSept cases	Europe-wide	1770 (924 CAP)	1583 (765 CAP)	Affy 5.0	440,794	346,478
GAinS 2011 cases	UK	288	240	Illumina Omni Express	730,525	644,775
WTCCC2 controls	UK	1390	1290	Affy 6.0	900,120	732,814
EPICOLON controls	Spain	572	488	Affy 6.0	905,525	699,022
EGEA controls	France	357	351	Illumina 610 Quad	582,892	530,964
GABRIEL controls	Germany	858	810	Illumina 610 Quad	582,892	526,107
SIALS controls	Ireland	211	209	Illumina Infinium II 550K	561,466	502,669
HYPERGENES controls	Italy	1280	1198	Illumina 1M	1,111,215	903,783

SNPs which overlapped between datasets and could be analysed. Figure 6.1 shows the pipeline used for QC and processing of all cohorts before final QC and analysis.

Further details of procedures for formatting and QC are given in Materials and methods section 2.4.7. SHAPEIT was used to phase genotyped data, and IMPUTE2.2 was used to impute genotypes for millions of SNPs. Further details about these processes are also in Materials and methods section 2.4.7.

Insertion/deletions (INDELs) and triallelic SNPs were removed from the sample sets before analysis. All the data in this study were imputed using a release of the 1000 Genomes data which was discovered to have data quality issues for INDELs and structural variants. In addition, when using SNPTEST to analyse data with multiple sample sets such as this, an option must be applied which does not allow the analysis of variants with the same chromosomal position. In impute output, triallelic SNPs are presented as two separate SNPs with the same chromosomal location, and INDELs often have the same chromosomal location as other SNPs.

Details of MDS cleaning and plots are found in section 6.3.3 below.

6.3.3 Multi-dimensional scaling

Due to the way these data have been gathered, it was particularly important to make sure that there would be no bias between cases and controls, especially in terms of population stratification. Using the 28,452 SNPs which overlap between all datasets and are not in LD with each other, multi-dimensional scaling (MDS) was performed to look for stratification within all eight cohorts. MDS was chosen to assess this stratification as this can be easily performed using PLINK, and gives similar results to principal component analysis (PCA) (Nelis *et al.*, 2009).

Plots for the first and second MDS components are shown in figures 6.2, 6.3 and 6.4 below. There was very little population specific stratification for the third and fourth components. Figure 6.2 shows the samples labelled according to whether

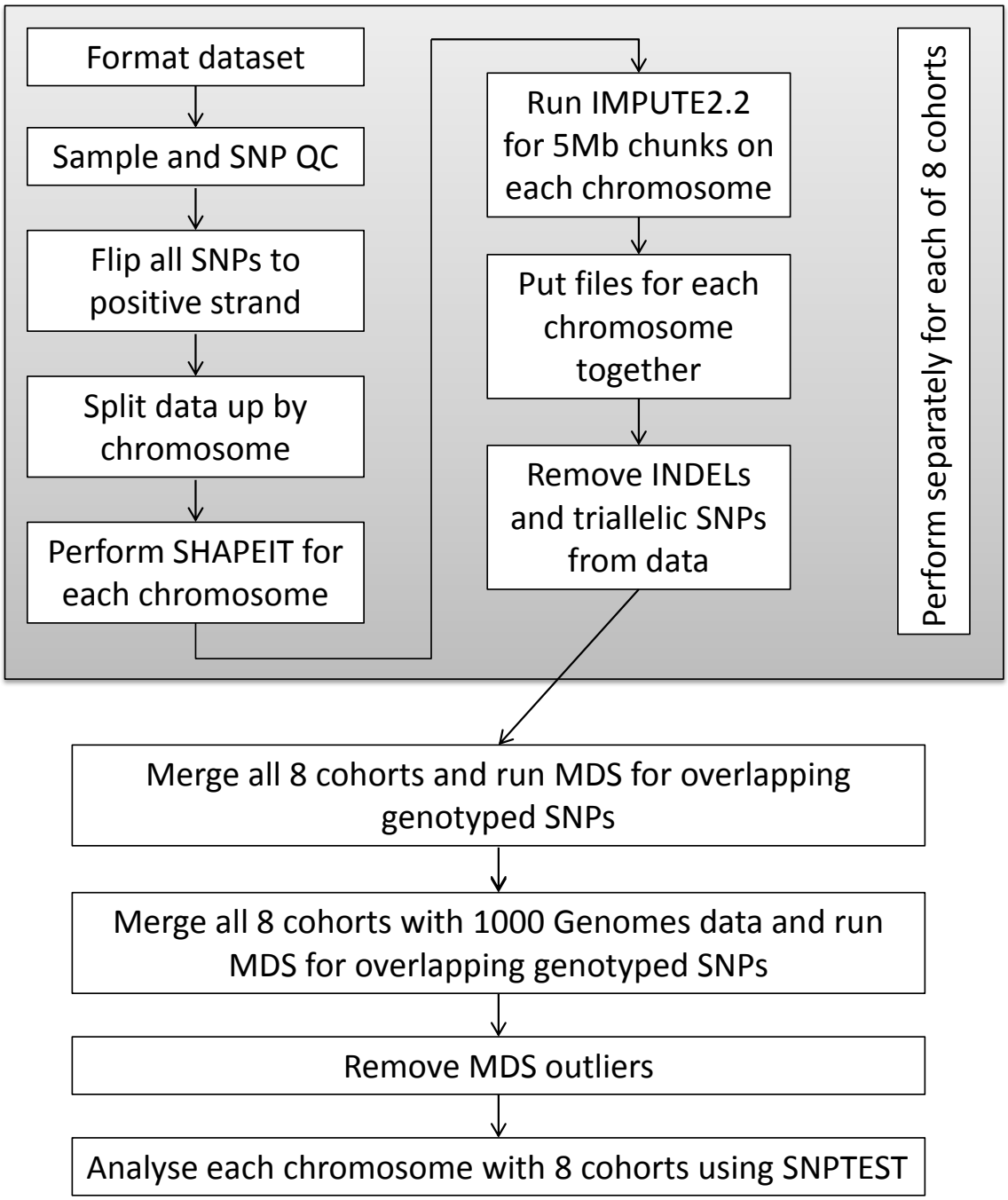


Figure 6.1: Pipeline to describe how each cohort was processed before final analysis. Further details for each step are given in Materials and methods section 2.4.7 and section 6.3.3 below. MDS: multi-dimensional scaling. INDELs: insertion/deletions. PLINK (Purcell, 2012; Purcell *et al.*, 2007), SHAPEIT (Delaneau *et al.*, 2012), IMPUTE2.2 (Howie *et al.*, 2011, 2009), SPSS v18 (PASW Statistics 18) and SNPTEST were used in these processes.

they are cases or controls. Samples found towards the edges of these clusters were removed from further analysis, as there was not good overlap of cases and controls. These are shown in the plot. Figure 6.3 shows the results of an analysis of all eight cohorts combined with all 1000 Genomes data (using data from 24,277 SNPs not in LD). African, Asian and European populations segregate clearly in this plot, and samples from this study are most similar to the 1000 Genomes samples of European descent. Figure 6.4 shows an MDS plot for just the samples in this study, with samples labelled by the country in which they were collected. Samples in this plot cluster well by country, which suggests that this analysis is successfully stratifying samples by population. The number of cases and controls from each country that are included in the final analysis after all QC are shown in table 6.2.

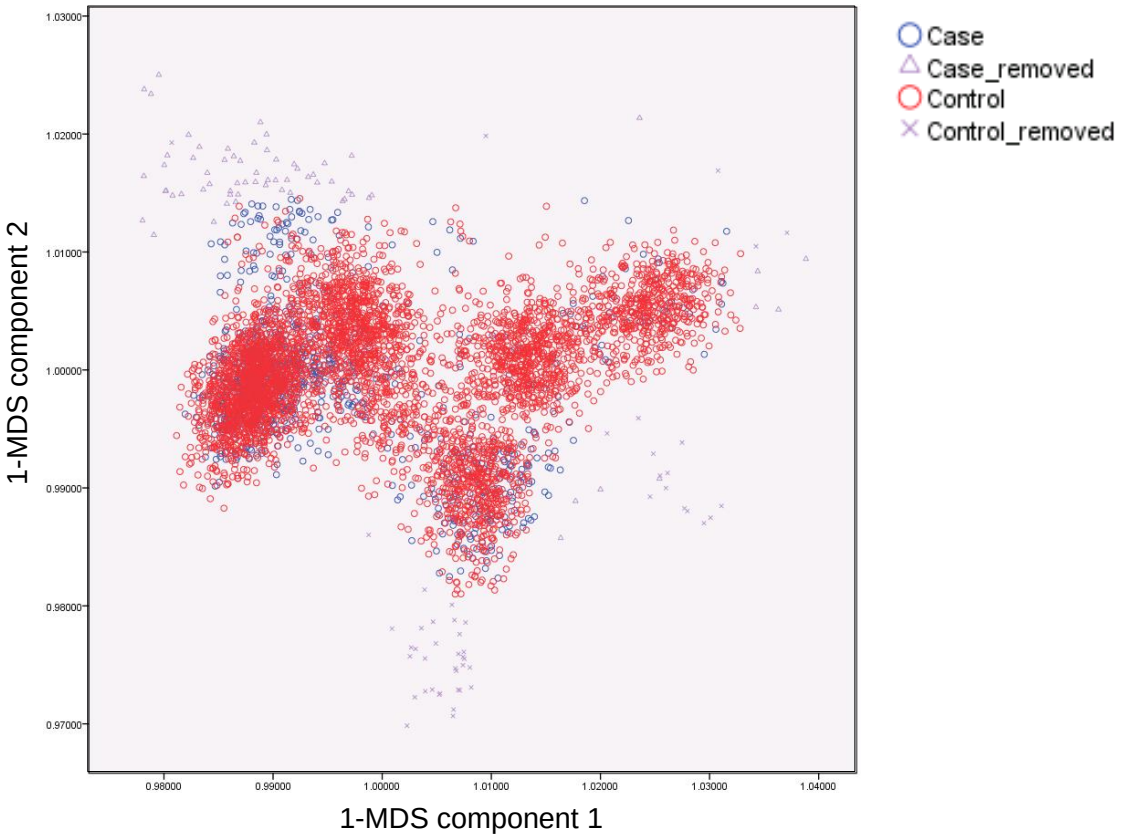


Figure 6.2: Multi-dimensional scaling (MDS) plot of the first two MDS components for all samples passing initial QC from all eight cohorts. Samples are labelled by case control status, and whether they were removed following this whole set MDS QC step.

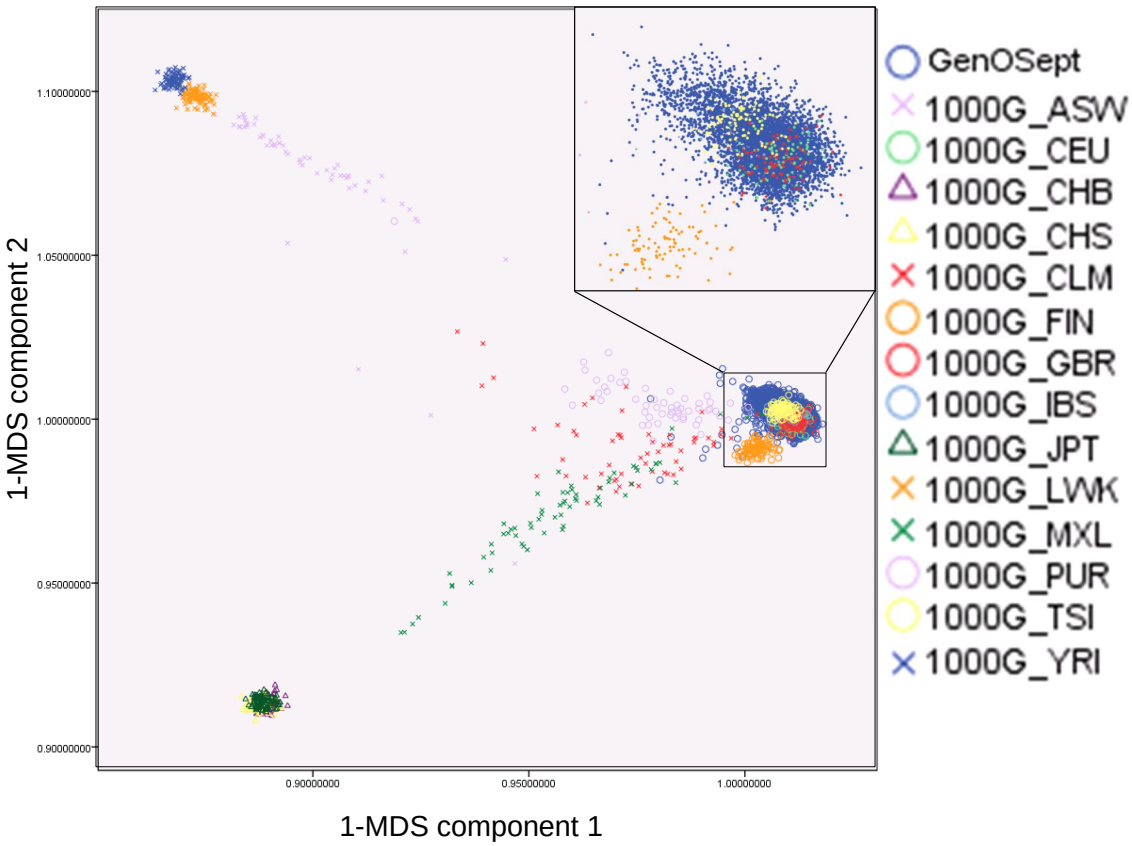


Figure 6.3: Multi-dimensional scaling (MDS) plot of the first two MDS components for different populations from 1000 Genomes and all samples in this study passing final QC. ASW: HapMap African Ancestry individuals from SW USA, CEU: Utah residents (CEPH) with Northern and Western European ancestry; CHB: Han Chinese in Beijing; CHS: Han Chinese South; CLM: Colombians in Medellin, Colombia; FIN: HapMap Finnish individuals from Finland; GBR: British individuals from England and Scotland; IBS: Iberian populations in Spain; JPT: Japanese individuals; LWK: Luhya individuals in Webuye, Kenya; MXL: Mexican ancestry in Los Angeles, USA; PUR: Puerto Ricans in Puerto Rico; TSI: Toscan individuals in Italy; YRI: Yoruban individuals in Ibadan, Nigeria.

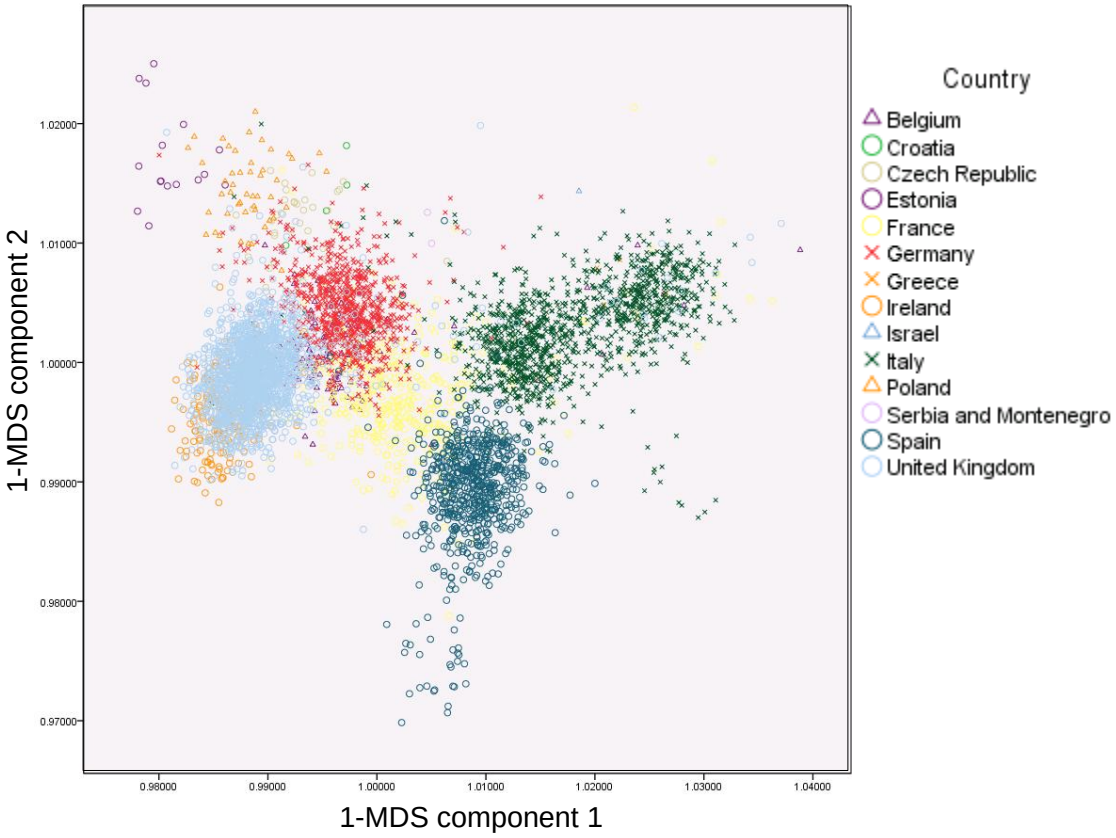


Figure 6.4: Multi-dimensional scaling (MDS) plot of the first two MDS components for all samples passing initial QC from all eight cohorts. Samples are labelled by the country which they were collected in.

Table 6.2: All samples included in GWAS after QC, as separated by case control status, and by the country they were collected in

Country	**Number of CAP sepsis cases after QC**	**Number of healthy controls after QC**
UK	530	1290
Spain	163	488
Belgium	119	
France	61	351
Italy	38	1198
Czech Republic	28	
Poland	21	
Germany	20	810
Ireland	16	209
Greece	4	
Croatia	2	
Serbia and Montenegro	2	
Israel	1	
Total	**1005**	**4346**

6.4 Results

There were 37,005 overlapping SNPs which were genotyped in all eight cohorts. These SNPs were analysed using logistic regression in PLINK (Purcell, 2012; Purcell *et al.*, 2007), with the first four MDS components as covariates, in order to assess the quality and inflation of the genotyped data. The QQ-plot and Manhattan plot for this analysis is shown in figure 6.5. There are no associating peaks in the Manhattan plot.

Following imputation, genotypes were available for 36,707,462 SNPs in all cohorts. These were analysed using SNPTEST, and the data were filtered as below, to only keep association data for reliable SNPs. Many of the imputed SNPs in the data are very rare, and therefore not reliable. Furthermore, reliability scores for imputed data can be used to remove additional SNPs which have not imputed well. This score is called an info score, and can be output from SNPTEST as an average score for all datasets.

Table 6.3 details the different thresholds used to clean the data, the number of SNPs left using each criteria, and the lambda value for the remaining SNP associations in each way of cleaning. In addition to cleaning for overall MAF and overall info score, data were also cleaned for the MAF in separate cohorts. In some cases, a SNP was very rare in some cohorts but slightly less rare in others, meaning that these SNPs would not be removed using an overall MAF threshold. However, the data would still be unreliable for the cohorts in which the SNP is rare, so these needed to be removed. In all cleaning methods, SNPs were also cleaned for control HWE, using a p-value threshold of $1\text{x}10^{-10}$.

QQ-plots and Manhattan plots were made for each way of cleaning the data. Some of these are shown here. The thresholds used in the "cleaned 1" dataset are quite stringent, and are what has been used for the final analysis of other imputed datasets in the Hill group. However, as can be seen by the inflated lambda value shown in table 6.3, and by the QQ and Manhattan plots shown in figure 6.6, these thresholds

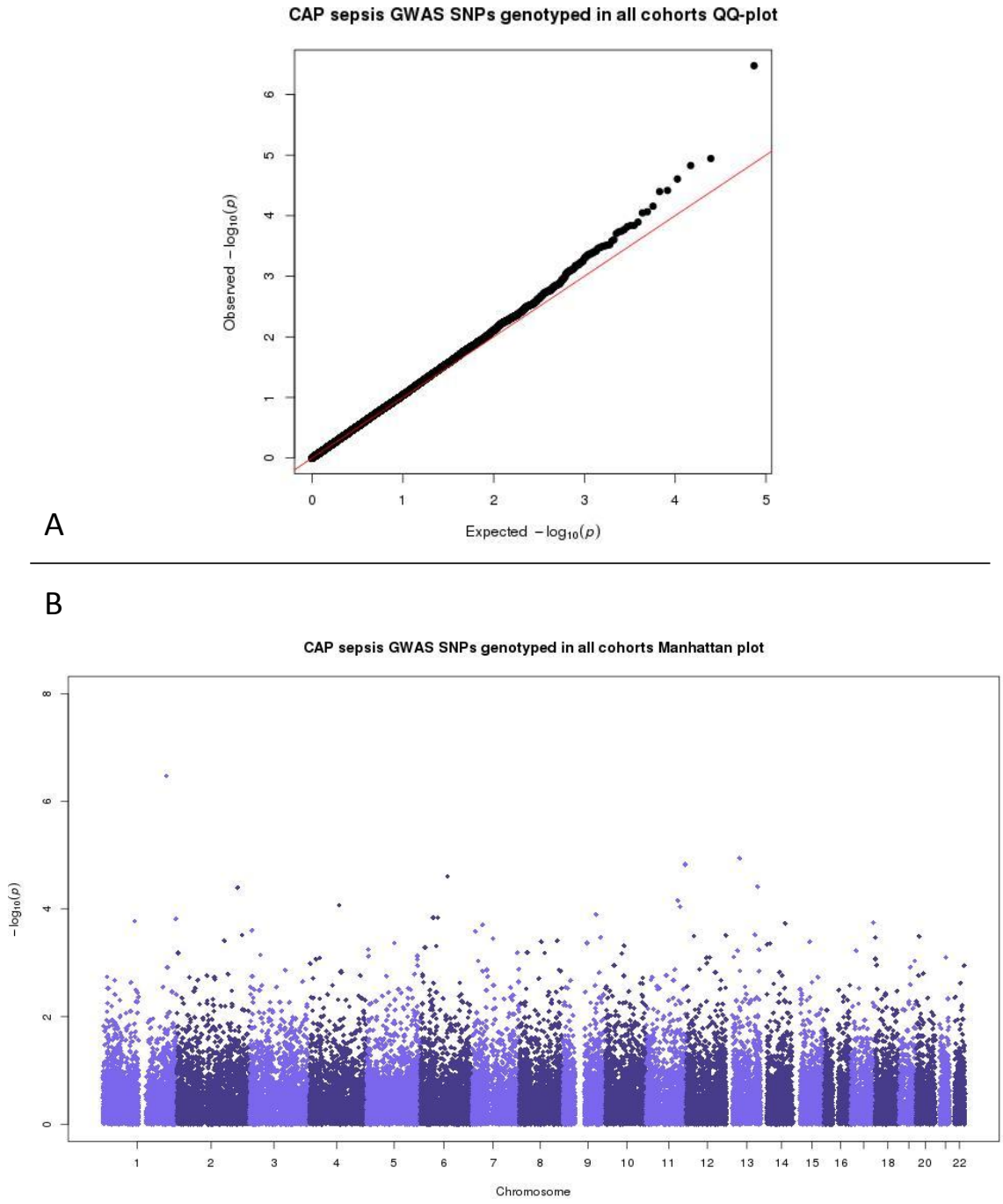


Figure 6.5: QQ-plot **(A)** and Manhattan plot **(B)** for SNPs genotyped in all eight cohorts (37,005 SNPs). SNPs were analysed using logistic regression in PLINK, using the first 4 MDS components as covariates. Lambda for these data is 1.067. Plots were made using an R script available from Getting Genetics Done (Turner, 2012).

are not stringent enough for these data.

Following further and more stringent cleaning, the lambda value for the data is much improved. Figure 6.7 shows the association results for the "cleaned 7" data. Even after this very stringent level of cleaning, which left nearly 3 million SNPs in the analysis, there are still many spurious significant associations on this plot. All SNPs with p-values $< 1\text{x}10^{-4}$ in this "cleaned 7" analysis were inspected to assess the allele frequency in all cohorts. During this inspection, SNPs were removed if the allele frequency across case or control cohorts was not consistent.

Table 6.4 details some of the SNPs removed from the "cleaned 7" data, after individual assessment. With the exception of the two SNPs at the bottom of the table, all SNPs in this table were removed from the "cleaned 7" data. The allele frequencies across cohorts for these SNPs show that the genotyping or imputation of these SNPs has performed poorly, generating an association due to unreliable data. The two SNPs at the bottom of the table are the top hits for the chromosome 1 and chromosome 16 hit regions in the data. The chromosome 1 hit SNP in particular has very consistent allele frequencies between case cohorts, and across control cohorts.

Following the removal of 218 SNPs after assessment of allele frequency as in table 6.4, QQ-plots and Manhattan plots for the "cleaned 8" data were generated, as in figure 6.8. Regional association plots for the two best peaks in the plot are shown in figures 6.9 and 6.10.

The top SNP in the chromosome 1 hit region is rs2936011 ($p=2.02\text{x}10^{-7}$). This SNP was genotyped in the GenOSept, UK controls and Spanish controls datasets, but imputed in all others. The p-value of this SNP in the GAinS UK CAP susceptibility GWAS presented in chapter 5 was 0.0025. The MAF of this SNP ranges from 10% in Africa to 40% in Asia and is 26% in Europe, according to 1000 Genomes data (1000 Genomes Project Consortium, 2012b). Therefore, the MAF of the control data in this study is very similar to that seen in the 1000 Genomes European data, which

Table 6.3: The number of SNPs and lambda values for the imputed GWAS data following cleaning using different thresholds. In addition to the criteria shown, SNPs were also cleaned for control HWE, using a p-value threshold of $> 1x10^{-10}$, in every analysis. *218 SNPs removed with p-values $< 1x10^{-4}$, after assessment of allele frequency across cohorts.

	SNPs genotyped in all cohorts	**Number of snps after imputation**	**Cleaned 1**	**Cleaned 2**	**Cleaned 3**	**Cleaned 4**	**Cleaned 5**	**Cleaned 6**	**Cleaned 7**	**Cleaned 8**
SNPTEST Info threshold			> 0.8	> 0.8	> 0.9	> 0.999	> 0.95	> 0.98	> 0.99	> 0.99
Overall MAF threshold			> 0.01	> 0.02	> 0.02	> 0.02	> 0.02	> 0.02	> 0.02	> 0.02
Cleaning cohorts separately for MAF							> 0.02	> 0.02	> 0.02	> 0.02
Additional cleaning										Yes*
Number of SNPs remaining	37,005	36,707,462	7,170,630	6,513,533	5,851,395	642,491	4,796,728	3,663,099	2,745,963	2,745,745
Lambda	1.067000	NA	1.624747	1.601475	1.384956	1.081530	1.203316	1.105043	1.074317	1.061839

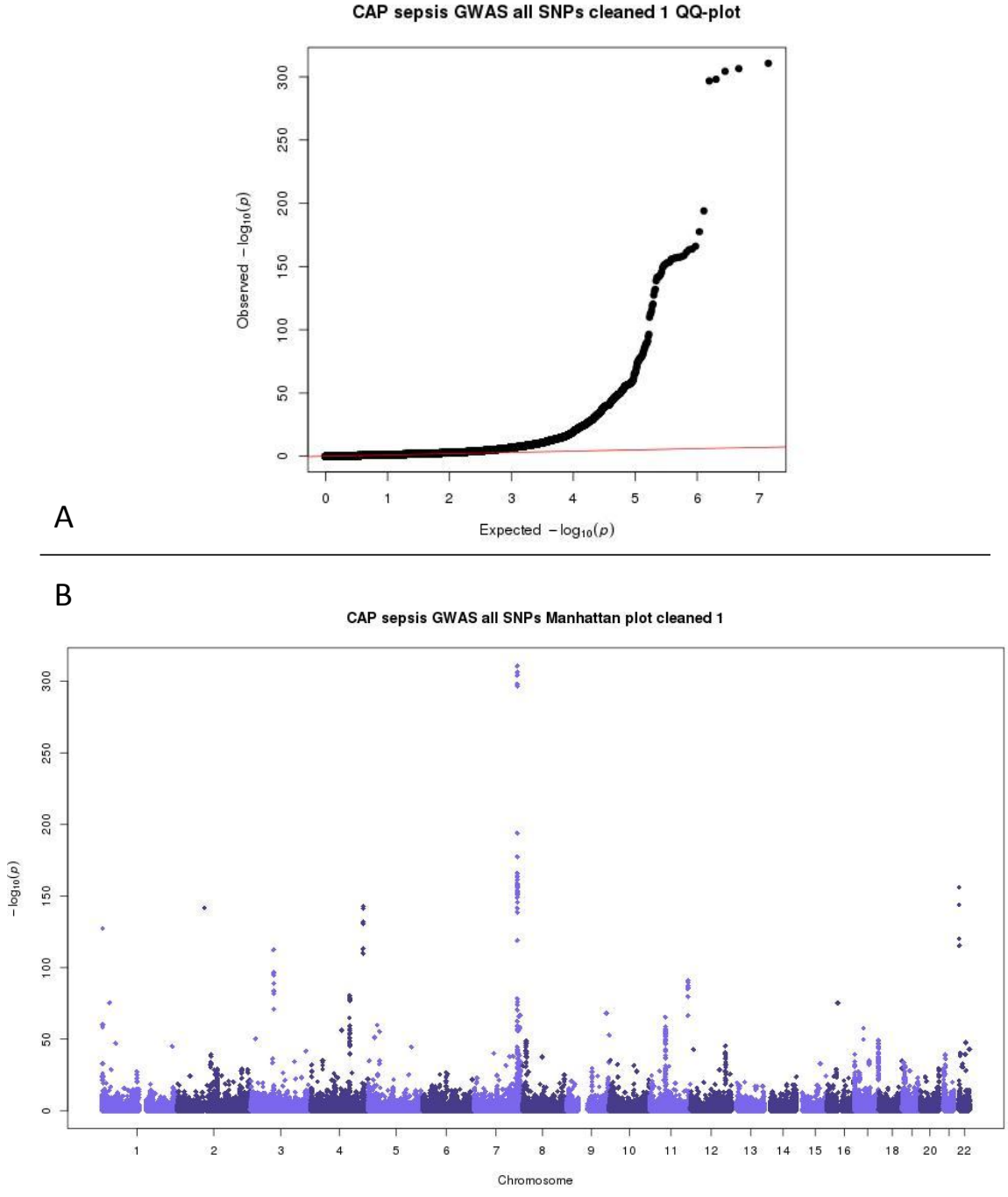


Figure 6.6: QQ-plot **(A)** and Manhattan plot **(B)** for all SNPs (genotyped and imputed in all eight cohorts) which passed the cleaning thresholds for "cleaned 1", as specified in table 6.3. Thresholds used were an overall info score > 0.8, overall MAF > 0.01 and HWE p-value in controls $> 1\text{x}10^{-10}$. 7,170,630 SNPs passed these thresholds. SNPs were analysed using the score test in SNPTEST, using the first four MDS components as covariates. Lambda for these data is 1.6247. Plots were made using an R script available from Getting Genetics Done (Turner, 2012).

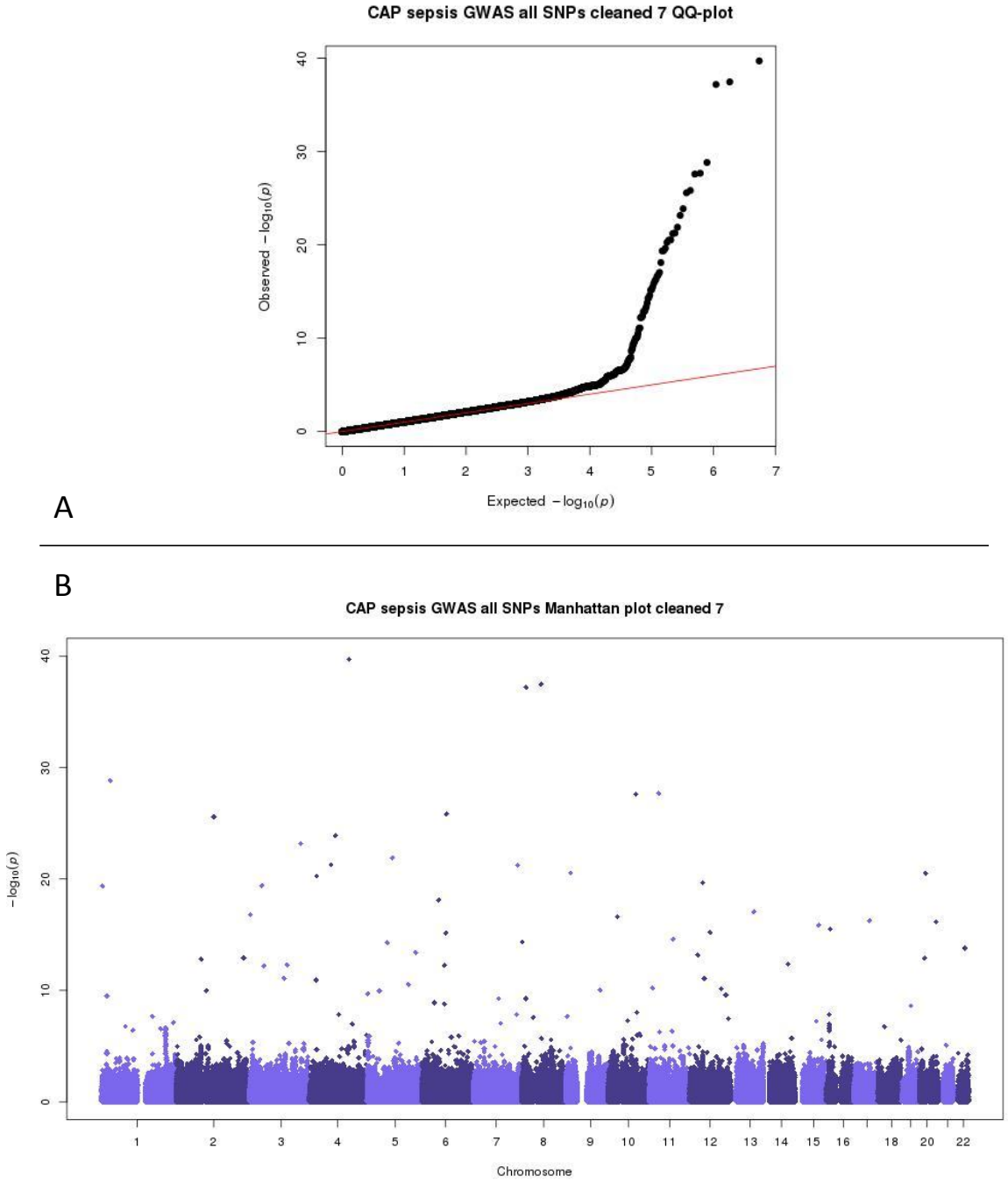


Figure 6.7: QQ-plot **(A)** and Manhattan plot **(B)** for all SNPs (genotyped and imputed in all eight cohorts) which passed the cleaning thresholds for "cleaned 7", as specified in table 6.3. Thresholds used were an overall info score > 0.99, overall MAF > 0.02, MAF in individual cohorts > 0.02 and HWE p-value in controls > $1x10^{-10}$. 2,745,963 SNPs passed these thresholds. SNPs were analysed using the score test in SNPTEST, using the first four MDS components as covariates. Lambda for these data is 1.0743. Plots were made using an R script available from Getting Genetics Done (Turner, 2012).

Table 6.4: The allele frequencies observed in each of the eight cohorts included in this analysis, for SNPs removed from the data, and for the top associating SNPs. HWE: Hardy-Weinberg equilibrium p-value. AAF: allele frequency of the A allele. Chr:pos lists the chromosome number and genome build 37 position of each SNP. *The two top associating SNPs. All other SNPs in this table were removed after inspection of the allele frequencies across cohorts.

							cases		controls					
	Chr:pos	**RS ID**	**Allele A/B**	**Info**	**HWE controls**	**P-value**	**GenOSept AAF**	**GAinS 2011 AAF**	**French AAF**	**German AAF**	**Irish AAF**	**UK AAF**	**Spanish AAF**	**Italian AAF**
	1:4506845	rs10799233	C/T	0.998	$6.88\text{x}10^{-6}$	$4.37\text{x}10^{-20}$	0.817	0.815	0.809	0.811	0.847	0.565	0.835	0.827
	1:29475394	rs2230677	C/G	0.997	0.96474	$1.47\text{x}10^{-29}$	0.202	0.796	0.214	0.237	0.170	0.195	0.221	0.245
	1:196942621	rs13376702	C/T	0.991	0.003195	$2.81\text{x}10^{-7}$	0.717	0.763	0.715	0.750	0.729	0.860	0.675	0.710
	3:113233188	rs16861103	C/A	0.998	$3.53\text{x}10^{-9}$	$8.00\text{x}10^{-12}$	0.975	0.783	0.771	0.763	0.799	0.966	0.964	0.746
	5:1826075	rs4975732	A/T	0.997	$1.55\text{x}10^{-5}$	$1.93\text{x}10^{-10}$	0.324	0.364	0.345	0.362	0.383	0.392	0.341	0.663
	7:91113955	rs4626524	A/G	0.993	0.066008	$8.90\text{x}10^{-8}$	0.211	0.340	0.282	0.323	0.284	0.316	0.293	0.281
	8:65449168	rs6472146	G/C	0.998	$4.84\text{x}10^{-8}$	$3.40\text{x}10^{-38}$	0.797	0.798	0.808	0.762	0.819	0.459	0.791	0.731
	10:31298325	rs1690624	G/A	0.998	0.39756	$2.43\text{x}10^{-17}$	0.331	0.815	0.316	0.338	0.356	0.324	0.353	0.354
	11:11540861	rs4520594	G/A	1.000	0.25325	$6.22\text{x}10^{-11}$	0.777	0.388	0.805	0.746	0.720	0.745	0.822	0.820
	11:79063721	rs498820	C/T	0.995	$3.69\text{x}10^{-6}$	$9.71\text{x}10^{-6}$	0.233	0.209	0.205	0.219	0.253	0.347	0.240	0.229
	12:128404530	rs11059416	C/A	0.990	0.19403	$3.43\text{x}10^{-8}$	0.909	0.938	0.813	0.833	0.816	0.910	0.921	0.808
	14:82471133	rs930129	G/T	0.997	0.005458	$4.13\text{x}10^{-13}$	0.138	0.165	0.159	0.150	0.117	0.029	0.061	0.177
	15:76435640	rs2955726	G/A	0.997	0.000926	$1.46\text{x}10^{-16}$	0.881	0.883	0.894	0.892	0.911	0.967	0.980	0.920
	17:52595418	rs1835221	A/G	0.996	$6.44\text{x}10^{-9}$	$5.30\text{x}10^{-17}$	0.501	0.477	0.496	0.509	0.496	0.236	0.502	0.510
*	1:213211519	rs2936011	C/T	0.996	0.87926	$2.02\text{x}10^{-7}$	0.669	0.645	0.731	0.738	0.723	0.714	0.720	0.725
*	16:9802763	rs10852234	G/A	0.996	1	$1.09\text{x}10^{-7}$	0.444	0.432	0.388	0.389	0.426	0.367	0.353	0.373

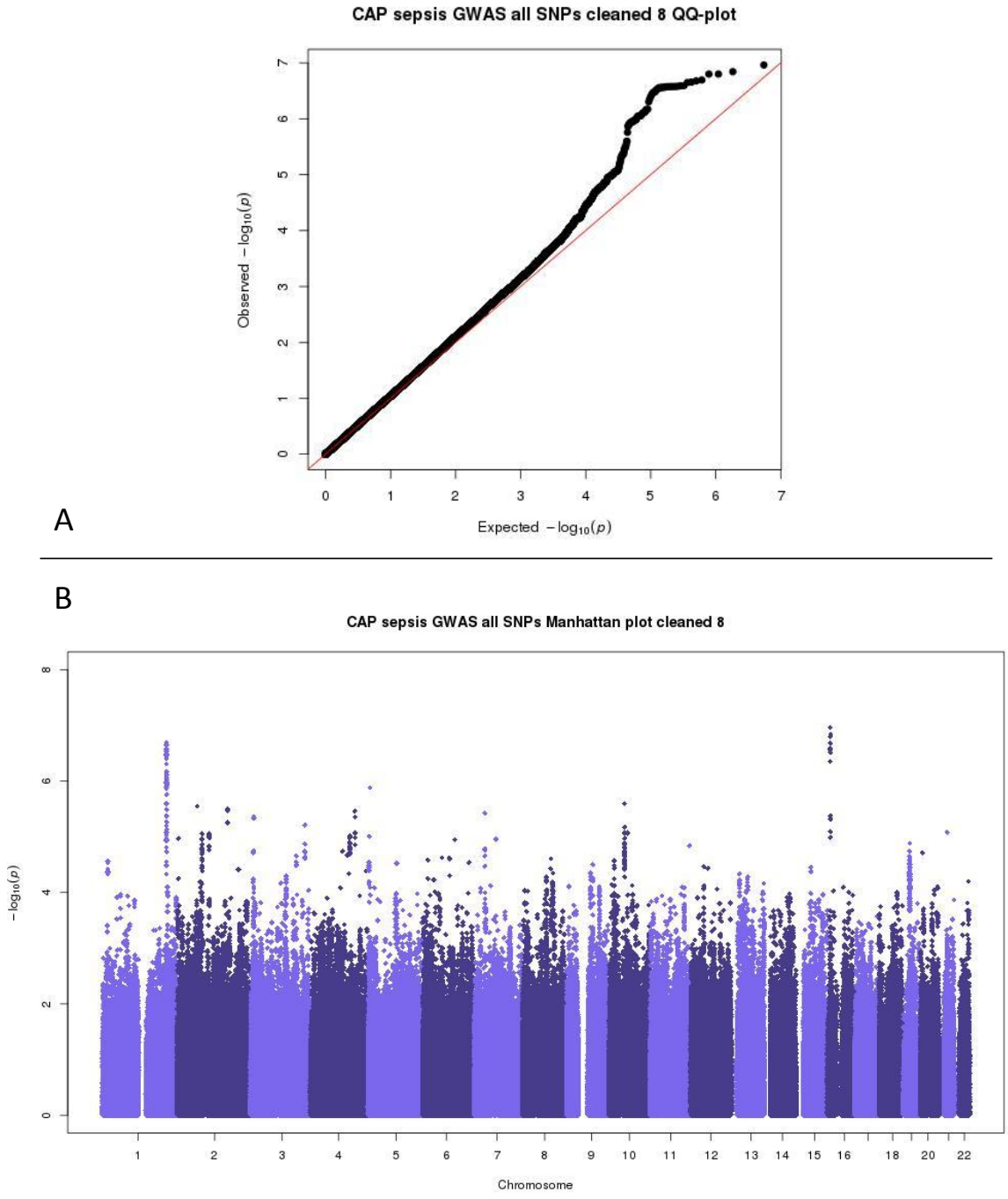


Figure 6.8: QQ-plot **(A)** and Manhattan plot **(B)** for all SNPs (genotyped and imputed in all eight cohorts) which passed the cleaning thresholds for "cleaned 8", as specified in table 6.3. Thresholds used were an overall info score > 0.99, overall MAF > 0.02, MAF in individual cohorts > 0.02 and HWE p-value in controls $> 1\text{x}10^{-10}$. In addition, 218 SNPs were removed after inspection of the allele frequencies in all cohorts. 2,745,745 SNPs passed these thresholds. SNPs were analysed using the score test in SNPTEST, using the first four MDS components as covariates. Lambda for these data is 1.0618. Plots were made using an R script available from Getting Genetics Done (Turner, 2012).

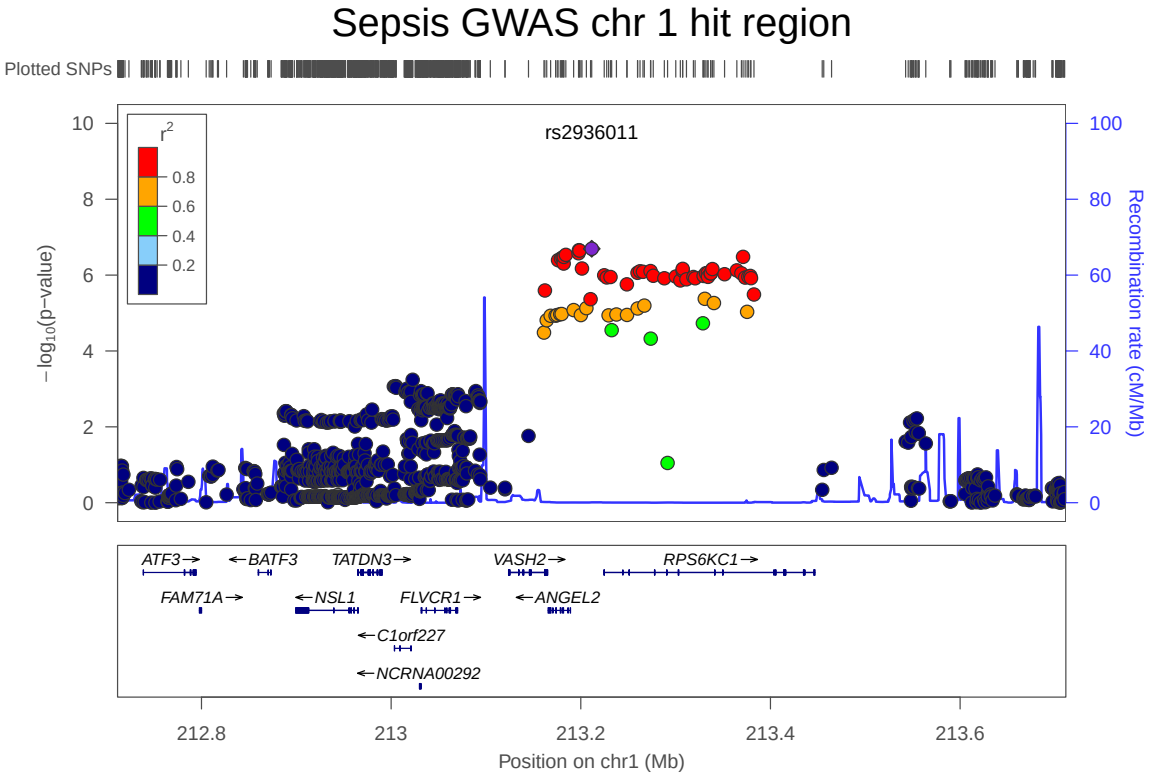


Figure 6.9: Regional association plot for the hit region on chromosome 1. Position on chromosome 1 is indicated by the x-axis. Genes in the region are shown at the bottom of the plot, with exons indicated by vertical bars, and the direction of transcription shown by an arrow. The -log^{10}(p-value) for all SNPs is indicated by the y-axis, as is the recombination rate, represented by the blue line on the plot. SNPs are coloured according to their LD with the top SNP (rs2936011). LD measures are from 1000 Genomes data from Europeans released in November 2010. The top SNP is indicated by a purple diamond. Regional association plots were made using LocusZoom (Pruim *et al.*, 2010).

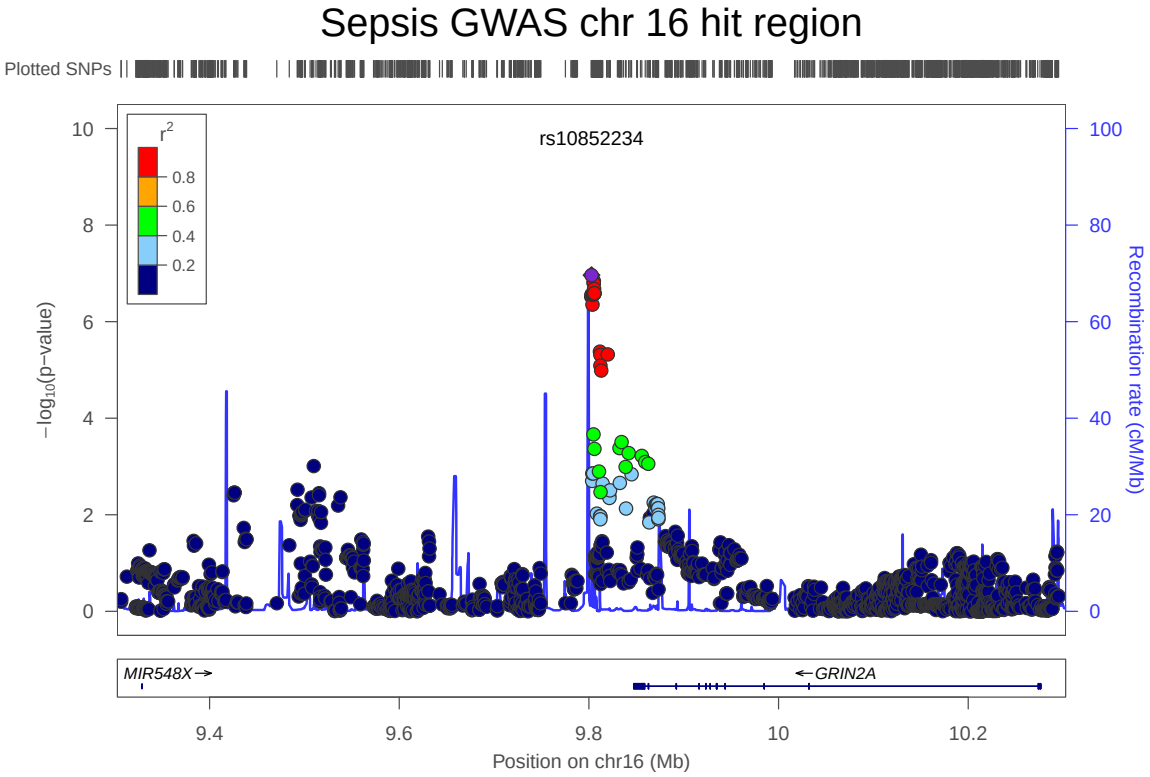


Figure 6.10: Regional association plot for the hit region on chromosome 16. Position on chromosome 16 is indicated by the x-axis. Genes in the region are shown at the bottom of the plot, with exons indicated by vertical bars, and the direction of transcription shown by an arrow. The -log[10](p-value) for all SNPs is indicated by the y-axis, as is the recombination rate, represented by the blue line on the plot. SNPs are coloured according to their LD with the top SNP (rs10852234). LD measures are from 1000 Genomes data from Europeans released in November 2010. The top SNP is indicated by a purple diamond. Regional association plots were made using LocusZoom (Pruim *et al.*, 2010).

means that it is likely to be reliable.

The top SNP in the chromosome 1 region which was genotyped in all eight sample sets is rs3124668 ($p=2.92x10^{-7}$). The p-value of this SNP in the GAinS UK CAP sepsis susceptibility GWAS presented in chapter 5 was 0.0034. Cluster plots for this SNP were assessed in the GenOSept, GAinS 2011 and Spanish control data (as I do not have access to the other cluster plots), and all cluster plots look very clear. The cluster plot for half of the Spanish data is shown in figure 6.11. There are three genes in this region; *VASH2* (vasohibin 2: vascular endothelial growth factor (VEGF)-inducible endothelium-derived angiogenesis inhibitor 2), *ANGEL2* (angel homolog 2 (drosophila)) and *RPS6KC1* (ribosomal protein S6 kinase, 52kDa, polypeptide 1).

The top SNP in the chromosome 16 hit region is rs10852234 ($p=1.09x10^{-7}$). This SNP was imputed in all eight cohorts. The MAF of this SNP in the 1000 Genomes data from Europe is 42%, which is very similar to that seen in the GenOSept, GAinS 2011 and Irish control datasets (1000 Genomes Project Consortium, 2012b). There are no associating SNPs in this region that have been genotyped in all cohorts. The closest gene to this hit region is *GRIN2A* (glutamate receptor, ionotropic, N-methyl D-aspartate 2A).

Using the "cleaned 8" data, the regions of the top associations from the GAinS UK CAP sepsis susceptibility analysis presented in chapter 5 were assessed. None of the top SNPs from the GAinS UK CAP sepsis susceptibility study were more significant in this analysis (see table 6.5). Furthermore, in the 50kb window around each GAinS UK CAP sepsis susceptibility association top SNP, there are no SNPs which are more significant than the top SNPs in the GAinS UK CAP sepsis susceptibility GWAS.

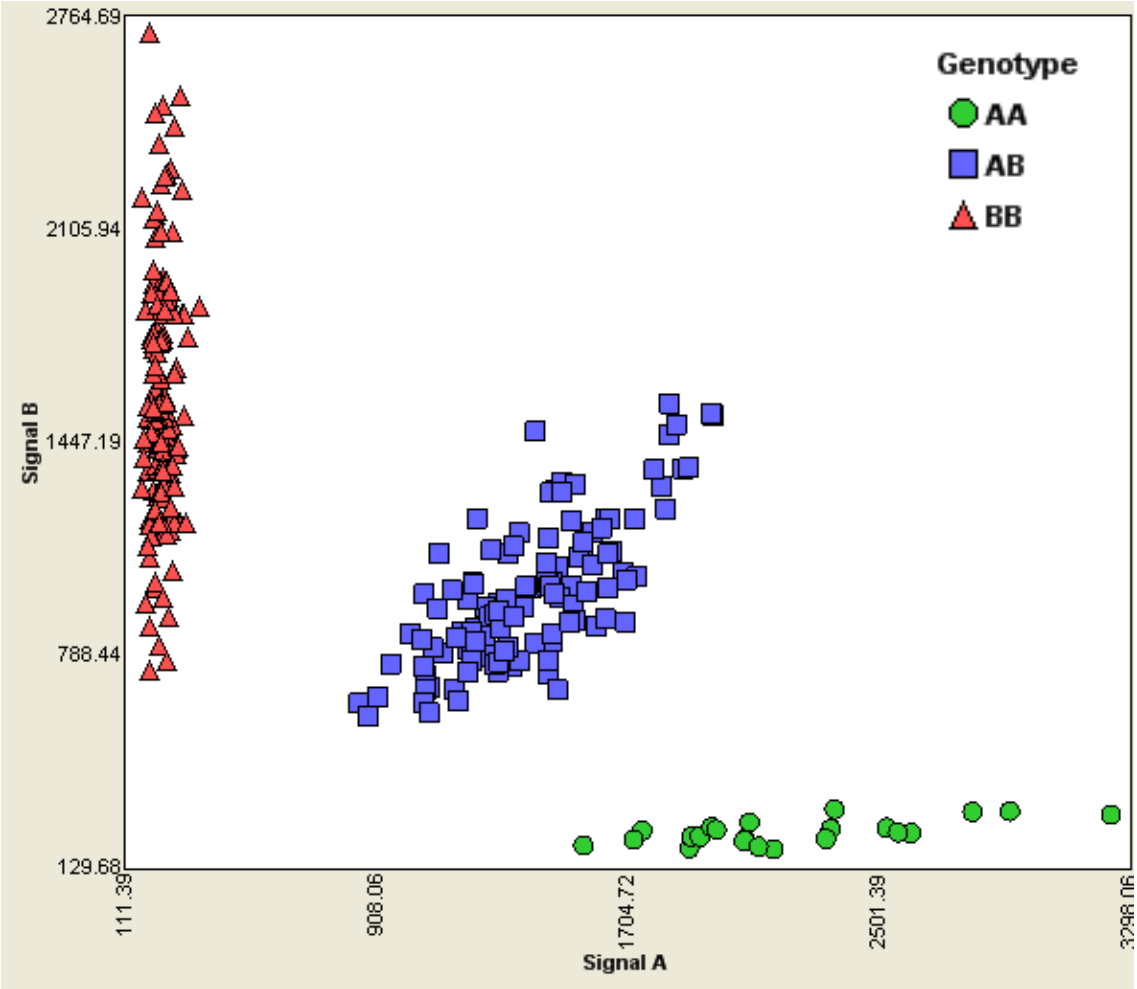


Figure 6.11: Cluster plot for half of the Spanish data for the top genotyped SNP in the chromosome 1 hit. In contrast to the data in table 6.4, in this plot allele A is the rare allele. The data for this SNP will have been flipped to the positive strand, therefore changing which allele is defined as allele A in the final analysis. This plot was created using the Affymetrix genotyping console, which can only handle a small number of samples at once in processing, which is why this plot only shows data for half of the Spanish samples.

Table 6.5: The best regions of association found with the GAinS CAP sepsis susceptibility GWAS, with p-values for the SNPs from this European CAP sepsis susceptibility GWAS. Genes in each region were found using HapMap Genome Browser release 28 at http://hapmap.ncbi.nlm.nih.gov.

Top SNP in GAinS GWAS	**Top p-value in GAinS GWAS**	**P-value in GenOSept CAP sepsis GWAS**	**Gene**
rs2484714	$1.2\text{x}10^{-5}$	$6.3\text{x}10^{-3}$	*WDR65*, *TMEM125*, *EBNA1BP2*
rs12128857	$4.1\text{x}10^{-4}$	0.019	*SGIP1*
rs1982530	$6.2\text{x}10^{-4}$	0.049	*FMN2*
rs10932875	$4.0\text{x}10^{-5}$	0.088	none
rs7651090	$1.1\text{x}10^{-4}$	$8.0\text{x}10^{-3}$	*IGF2BP2*
rs9995416	$9.7\text{x}10^{-5}$	0.401	*GABRA4*, *COX7B2*
rs2707767	$9.7\text{x}10^{-5}$	0.0027	none
rs293489	$1.2\text{x}10^{-4}$	0.0543	*DOPEY1*, C6orf157
rs17167808	$4.0\text{x}10^{-5}$	0.0494	*SLC35B4*, *LRGUK*
rs7021164	$5.2\text{x}10^{-5}$	$2.8\text{x}10^{-3}$	C9orf123
rs3950310	$6.3\text{x}10^{-4}$	0.0146	*MAPK8*
rs12278965	$1.1\text{x}10^{-5}$	NA	*NELL1*
rs7925363	$9.3\text{x}10^{-5}$	$5.5\text{x}10^{-3}$	*LUZP2*
rs7992703	$5.7\text{x}10^{-5}$	NA	*SAP18*, *MRP63*
rs17238047	$1.5\text{x}10^{-5}$	0.0191	*VPS13C*
rs1423982	$4.3\text{x}10^{-4}$	0.0270	none
rs3826248	$4.8\text{x}10^{-5}$	$5.3\text{x}10^{-3}$	*CALB2*, *FLJ11171*
rs6506348	$3.0\text{x}10^{-4}$	0.0156	*TTMA*, *L3MBTL4*
rs381983	$2.8\text{x}10^{-4}$	0.141	*ZNF155*, *ZNF230*, *ZNF222*, *ZNF223*, *ZNF45*
rs2760549	$2.5\text{x}10^{-5}$	0.114	*MACROD2*, *KIF16B*

6.5 Discussion

This chapter presents preliminary analysis for a large GWAS for susceptibility to CAP sepsis in Europe, and two novel genetic regions have shown statistically significant associations in this analysis.

Due to the nature of this analysis, quality control and sample set processing must be performed with great care. When combining the data of so many cohorts that have been genotyped on different platforms and imputed separately, the chance of false positive associations is greatly increased. Although the current cleaning process has been extremely stringent and promising results have been found, further cleaning of the data should still be performed to ensure the reliability of results, as detailed below. In addition, by cleaning the data in a different way, it may be possible to include more reliable SNPs in the final analysis and thereby identify further associating loci in this dataset. There is no defined way to handle data such as this, and so further exploratory analyses need to be performed to make sure that the data is analysed in a way which provides reliable results, as detailed below. In addition, replication must be performed once reliable data cleaning and analysis methods have been established. This is also discussed below.

6.5.1 Cleaning genotyped data

It was very important that the original genotyped data for every cohort was cleaned with strict QC, since the inclusion of unreliable genotypes in any one cohort could lead to a false positive association. Stringent QC thresholds were used in all datasets, but further, more stringent cleaning could be performed for the data in order to clean the data even further. For instance, a HWE p-value threshold in cases could also be included, and the threshold in controls could be increased. A further criteria for QC which can be included is to compare the call rate between cases and controls, and to

discard SNPs where the call rate is quite different. This will reduce the likelihood of observing false positive associations. However, as relatively very few SNPs overlap between all sample sets in this study, this could limit the effectiveness of this QC step. The lambda value for the analysis of the SNPs which had been genotyped in all cohorts was inflated quite considerably (λ=1.067, see table 6.3). Lambda values of less than 1.05 are generally considered acceptable. Although four MDS components were included as covariates in this analysis, components 3 and 4 did not have a significant effect on the data. Therefore, it is unlikely that including further components as covariates would correct this inflation, although this could be tested.

Furthermore, although all of the MDS outliers were removed from the analysis based on the MDS plot of these data, there were some additional outliers when MDS was performed together with the 1000 Genomes data. Roughly ten additional samples should perhaps be removed based on this plot, although it is strange that these samples are not outliers in the plot for just cases and controls. This is not what would be expected, so further exploratory analysis for these samples should be performed, and the analysis should be repeated without these additional outliers, to test whether this can reduce the inflation in the data.

6.5.2 Cleaning imputed data

Imputed data should also be analysed with care. As shown in table 6.3, different methods for cleaning these data result in very different levels of inflation as represented by lambda. The criteria chosen for the final preliminary analysis results in a lambda value slightly lower than that seen in the genotyped SNPs analysis (λ=1.0618). This could mean that with improved cleaning of the imputed data, the lambda value could be reduced still further, meaning that the data is a lot more reliable.

In the current analysis, imputed data have been cleaned based on the overall info value for reliability of imputed data. This value is from SNPTEST and is a value

to represent all the datasets. However, this does mean that if one sample set has performed rather badly, but the rest have behaved very well, this SNP could still be included in the analysis after cleaning, and could produce a false positive association result. Therefore, it would be important to also clean the data based on the info value for each cohort individually. Info values of this kind are output from imputation for each chunk of data imputed.

It is important when cleaning the data that it is not cleaned too much, as otherwise true positive associations can be removed from the data. For instance, in the "cleaned 4" criteria, an info threshold of > 0.999 was used. This meant that there were only 642,491 SNPs left in the analysis; fewer than most modern GWAS. This meant that most significant association results were removed from the data, which may be due to over cleaning. Further exploratory analyses for the processing of these data must be performed before the final results can be determined.

6.5.3 Performing analysis

The SNPTEST analysis presented in this thesis was performed using the score test. The score test makes a number of assumptions about the data, and increased genotype uncertainty from imputation can lead to spuriously low p-values when performing the analysis with this test. Therefore, the analysis should also be performed using different tests, to assess the effect this has on the association results. The threshold test would be the first test to perform. This test will only analyse genotypes which pass a certain threshold of certainty, and will therefore only analyse SNPs which are reliably imputed, if the correct threshold is chosen. This means that defined genotypes, rather than probabilities are used in the association analysis (Marchini *et al.*, 2010).

During the course of this analysis, improved reference datasets have become available, and methods used for data processing have been improved. For instance, the

latest 1000 Genomes release has been updated to include accurate data for INDELs, which could potentially be analysed for this dataset if I used this updated reference panel for imputation. Additionally, SHAPEIT2 has been developed, which vastly improves on SHAPEIT1 and could be used to more reliably phase and therefore impute genotypes in my datasets.

6.5.4 Preliminary association results

Two genomic regions have been shown to associate with susceptibility in this preliminary analysis, in chromosomes 1 and 16.

There are three possible candidate genes in the region of the chromosome 1 hit. These are *VASH2* (vasohibin 2: vascular endothelial growth factor (VEGF)-inducible endothelium-derived angiogenesis inhibitor 2), *ANGEL2* (angel homolog 2 (drosophila)) and *RPS6KC1* (ribosomal protein S6 kinase, 52kDa, polypeptide 1).

There is very little known about any of these genes. *VASH2* has been shown to be involved in angiogenesis (the development of new blood-vessels from pre-existing ones), and homozygous null mice for *VASH2* are deficient for this process (Kimura *et al.*, 2009; Lawrence, 2000). The related gene *VASH1* has been shown to also protect endothelial cells from premature senescence and cell death by strengthening resistance against oxidative stress (Miyashita *et al.*, 2012). Lung tissue is the most susceptible tissue to oxidative injury and oxidative stress is thought to enhance the inflammatory response in sepsis and lead to greater cell death (Miyashita *et al.*, 2012; Taylor and Piantadosi, 1995). If *VASH2* is also involved in resistance against oxidative stress, this could explain the observed association. *ANGEL2* is a protein coding gene with unknown function. It seems to be expressed most strongly in B cells, T cells and NK cells, however. Protein domain information suggests that the resulting protein is part of an endonuclease/exonuclease/phosphatase family (University of California, Santa Cruz, 2012b). *RPS6KC1* is involved in nucleotide, ATP and protein binding, and cell

communication. It is weakly expressed in the lung (University of California, Santa Cruz, 2012b). None of these genes have been shown to associate in genetics studies of disease before (Hindorff *et al.*, 2012).

The difference in p-value for these top SNPs between the GAinS UK CAP sepsis susceptibility GWAS and this GWAS demonstrates the added power of this study. Whereas these SNPs associated at the 10^{-3} level in the previous GWAS, they now show p-values at the 10^{-7} level. With additional cases and controls, it is possible to identify this association which was hidden amongst thousands of other associations of the same strength in the previous study.

Two SNPs (rs1045868; p=1.2x10^{-5} and rs10864033; p=1.6x10^{-5}) in the hit region were shown to be cis eQTL SNPs for *ANGEL2* in monocytes in a study of German individuals (Zeller *et al.*, 2010). Since there is so little information about the function of any of these genes, it is impossible to hypothesise which may be affecting susceptibility to CAP sepsis. Functional work would need to be performed to assess this further.

There are a number of factors which make it more likely that this chromosome 1 association is not a false positive. There is an associating SNP in this region which was directly genotyped in all eight cohorts, which associates with a p-value almost as low as the top SNP in the region. The MAF of this SNP is consistent across cases and separately across controls, and is consistent with 1000 Genomes European data in controls. The cluster plots for this SNP in both case cohorts and one control cohort look very clear, meaning that genotyping error is less likely to be the cause of this association. The data will need further cleaning as described above, however, which may affect the observed association in this region.

The top SNP in the chromosome 16 hit region is rs10852234 (p=1.09x10^{-7}). The MAF of this SNP in 1000 Genomes is 42%, which is very similar to that seen in the GenOSept, GAinS 2011 and Irish datasets (1000 Genomes Project Consortium,

2012b). There are no associating SNPs in this region that have been directly genotyped in all cohorts. The closest gene to this hit region is *GRIN2A* (glutamate receptor, ionotropic, N-methyl D-aspartate subunit 2A).

GRIN2A encodes a receptor which is an ionotropic glutamate-gated ion channel (University of California, Santa Cruz, 2012b). The gene is frequently mutated in tumour DNA from melanoma patients, as described by a recent exome sequencing study (Wei *et al.*, 2011). In addition, *GRIN2A* has been shown to be disrupted by chromosome translocation breakpoints in an individual with leprosy, and point mutations have been discovered in the gene in individuals with seizures (Endele *et al.*, 2010).

This hit region is difficult to explain. It seems strange that one set of the controls would show the same allele frequency as the two sets of cases, and the function of the gene does not easily fit with a role in sepsis. It will be interesting to see whether this association still exists following further processing of these data.

The GenOSept survival analysis results were examined for associations in the region of these association peaks. There are no SNPs with p-values less than $1x10^{-4}$ in the regions of these hits (or in fact anywhere on chromosomes 1 or 16) in the CAP survival analysis using the GenOSept and GAinS 2011 datasets. In addition, there were no significant associations in these regions in an extended survival analysis of lung infection survival using two additional datasets.

6.5.5 Comparison with GAinS UK CAP sepsis susceptibility GWAS results

None of the hit regions from the GAinS UK CAP sepsis susceptibility GWAS showed reduced p-values in this study. Since this study includes the same samples as the GAinS UK CAP sepsis susceptibility GWAS, with many additional samples, this is surprising. It is possible that the data has been over cleaned in the final preliminary

analysis presented here, and therefore these hit regions do not seem to associate. However, since most of the top SNPs from the GAinS UK study are still included in this analysis after cleaning, this is unlikely. Since these regions do not significantly associate in a very similar but larger analysis, it is likely that the initial findings were false positives.

6.5.6 Further genotyping and replication

Usually, when imputation is performed in a study, it is good practice to go back to the analysed samples and genotype the top imputed SNP, to ensure that the association is not due to poor imputation. This will be impossible to do in all the datasets in this study. However, top SNPs can be genotyped in both case cohorts, and in the UK cohort, to confirm some of the results from this analysis. In the case of the chromosome 1 hit region, this process is less important, since one of the top SNPs was genotyped in all eight cohorts.

Replication is extremely important in any GWAS, but perhaps more so in this study, since there are so many factors which can lead to false positive results. There are more than 700 UK CAP sepsis samples available for replication from the GAinS sepsis study which is still collecting in the UK. It is very important that replication is performed in as similar way as possible to the original GWAS (Kraft *et al.*, 2009), so it is particularly useful that the cases that will be used for replication are from an extension of the same study as cases in the GWAS presented here are. These cases are from around the UK and are collected in the same centres and in the same way as the UK sepsis cases in this cohort. In addition, the UK controls for replication are from the same cohort that the UK controls in this analysis are from. Replication will be performed using the Sequenom genotyping system, once reliable associations from this study have been established.

6.5.7 Conclusion

When complete, this study will be larger than any European infectious disease GWAS published to date. Currently, the largest European GWAS published is the meningococcal GWAS. This CAP sepsis study has twice as many cases but almost as many controls as the meningococcal GWAS, and will have almost as many samples in replication (Davila *et al.*, 2010; Hindorff *et al.*, 2012).

This work presents an important and ambitious analysis of susceptibility to sepsis, with large sample sizes and large numbers of SNPs to analyse. I have presented two novel regions of association from this analysis, but further work needs to be performed to reliably clean and analyse this large quantity of data, and reliably filter the association results.

Chapter 7

Discussion

This thesis aimed to study human genetic susceptibility to common infectious diseases in Europe, and undertook this study with LRTI; one of the most common infectious diseases in Europe. Mild and severe LRTI cohorts were studied to look at both ends of the spectrum for disease. Genetics studies for susceptibility to LRTI have so far been limited, and have largely used candidate gene techniques.

The rapid advancement of genetics techniques in recent years has provided more exhaustive and sophisticated methods with which to study genetic susceptibility to disease. Instead of choosing genes and loci to study in candidate gene studies, hypothesis-free studies can be conducted by studying loci genome-wide. Genome-wide association studies have been a major landmark in this respect, and have provided enormous amounts of information about genetic loci which associate with a whole range of traits and diseases.

The importance of DNA extraction is often overlooked in the success of these genome-wide studies, despite the universal acknowledgement that GWAS require large sample sizes to have the power to detect associations. Without high-throughput techniques for high quality DNA extraction in large sample cohorts, recent progress could have been rather slower.

I have tested modern methods to extract DNA from blood, saliva and serum. These tests demonstrated the importance of testing DNA extraction techniques for every cohort, and suggest that the handling of blood samples prior to DNA extraction affects the quantity of DNA that can be extracted from blood.

Once DNA has been reliably extracted from a large cohort, an association study can be conducted. I have utilised genome-wide and candidate gene techniques to study genetic susceptibility to mild and severe LRTI.

The main findings of this thesis are the following: *MBL2* genotypes are not associated with mild or severe LRTI with the high effect sizes previously observed for *MBL2* associations; a SNP in the *IFITM3* gene is associated with mild and severe viral infection in the recessive model; the *MPHOSPH8* gene region associates with susceptibility to mild rhinovirus infection; a functional SNP in the *IL6* gene and the *SAP18* gene region associate with susceptibility to mild LRTI but only in smokers, and there may be an epistatic relationship between these genes with regards to susceptibility to LRTI; and preliminary findings from a large GWAS suggest that a gene region on chromosome 1 is associated with susceptibility to CAP sepsis.

The main findings in this thesis are all novel. Previous studies have shown positive and negative associations with SNPs in the *MBL2* gene, but this is the largest and thereby most reliable study to show a lack of association with severe LRTI, and the only study that I know of which analyses susceptibility to mild LRTI. To my knowledge, genetic studies have not been performed before for susceptibility to mild LRTI. Therefore, the mild LRTI study presented here is the first of its kind and the only account of genetic variants which associate with LRTI in primary care. As far as I know, there are no published GWAS of susceptibility to sepsis. The CAP sepsis susceptibility GWAS presented here is also the largest (in terms of sample numbers and populations included) study for susceptibility to sepsis that I know. The most promising finding from this study so far is a region of chromosome 1 which has not

been implicated for susceptibility to disease before.

There was some overlap between genetic susceptibility loci for mild and severe LRTI. Polymorphisms in the candidate genes *MBL2* and *IFITM3* were studied for mild and severe LRTI, and demonstrated very similar results for both. One of the associations from the UK CAP sepsis susceptibility GWAS was replicated in the mild LRTI cohort in the GRACE study. However, the hit SNPs from the UK CAP sepsis susceptibility GWAS show higher p-values in the larger, better-powered European CAP sepsis susceptibility GWAS. It is possible that the hit SNPs in the original study were false positives, but this doesn't explain why an association exists with mild LRTI. It will be very interesting to genotype any SNPs which replicate from the European CAP sepsis GWAS in the GRACE mild LRTI cohort.

It is intriguing that the three top associating SNPs from the GRACE study all show a different association between smokers and non-smokers. Genetic studies such as those for COPD and lung cancer have analysed data from smokers or individuals that have never smoked, or with respect to smoking (Cho *et al.*, 2012; Dong *et al.*, 2012; Lan *et al.*, 2012). The mild LRTI study presented in this thesis shows the importance of also taking smoking into account when analysing genetic susceptibility to LRTI. Unfortunately, although smoking status was collected in the sepsis cases used in this thesis, smoking status is not available for the GWAS controls, and therefore it is not possible to take smoking into account in the large CAP sepsis susceptibility GWAS presented in chapter 6.

The inclusion of covariates in genetic association studies is important to avoid confounding. However, the inclusion of non-confounding covariates can reduce the power of a genetic association study, and therefore the choice of covariates to include is important (Pirinen *et al.*, 2012). Age is included as a covariate in the sepsis survival analyses in the *MBL2* study presented in chapter 4. Age is known to be significantly associated with survival in sepsis. Country was used as a covariate in

the GRACE mild LRTI analyses, as well as smoking in the rhinovirus subgroup where smoking associated significantly with susceptibility. In the European CAP sepsis susceptibility GWAS, the first four MDS components were used as covariates. Besides sex and further MDS components it is impossible to use additional covariates in this analysis at this stage, as no further information is known about the controls in the study. However, other covariates could be used in the GRACE study. Although smoking appears to be the main confounding factor within subgroups in this study, other factors such as the presence of asthma, COPD and other underlying conditions may also confound results to a lesser extent. The difference in prevalence of these conditions between cases and controls is smaller than the difference in frequency of smokers, but these factors could still be important confounders. Further analysis should be performed to assess which of these covariates should be included for more in-depth analysis of this mild LRTI cohort.

The importance of directly genotyped rather than imputed genotypes has been highlighted in this thesis. The study of a SNP in the *IFITM3* gene presented in chapter 4 demonstrated the importance of directly genotyped controls in defining the model of a genetic association. Additionally, in the European CAP sepsis susceptibility GWAS presented in chapter 6, the difficulties associated with false associations due to low imputation quality are highlighted. Imputation is an extremely important tool for the discovery of genetic variants associated with disease, but will never be as reliable as direct genotyping.

The main findings from the mild LRTI study and the CAP sepsis GWAS are in regions of high LD, such that the causal gene for the association probably cannot be determined by genetics alone. This highlights the importance of further study for these genetic associations. Functional work could be performed for the genes of interest to determine which of the genes is responsible for the observed association and affects susceptibility to the disease of interest. In addition, new sequencing

technologies such as next-generation and exome sequencing could be utilised to aid fine-mapping efforts. These technologies would also be helpful to define new regions of association for susceptibility to disease.

Next-generation and exome sequencing have become increasingly popular in recent years, partly due to dropping costs associated with these technologies. These new strategies involve distinct challenges to those associated with GWAS strategies, particularly in the processing and analysis of data. Exome sequencing has been used to look for rare causal variants for as yet uncharacterised Mendelian disorders. The same experimental and analytical principles are now being applied to study complex diseases to find rare variants (as opposed to the common variants studied with GWAS) which contribute to susceptibility (Bamshad *et al.*, 2011).

Despite the success of GWAS for a whole range of complex diseases, much of the heritability associated with disease is yet to be explained with these findings. For example, the 18 loci first associated with type 2 diabetes from GWAS only explained 6% of the heritability associated with the disease. For many traits and diseases, attention has thereby turned towards rare variants that are undetectable using GWAS (Gibson, 2012; Manolio *et al.*, 2009).

The sepsis GWAS presented here should be finalised first, but it is very likely that it will not explain all of the heritability associated with CAP sepsis susceptibility. Therefore, sequencing technologies could be utilised to further investigate rare genetic variants that are associated with this disease. In fact, exome sequencing analysis is already underway to study the most severe sepsis cases in this cohort.

Since studies of the genetics of mild LRTI are in their infancy, it could be argued that it would be more cost-efficient to use GWAS as the next step in identifying genetic associations associated with mild disease, instead of sequencing. Next-generation sequencing can be more informative than GWAS, however. Should the cost of sequencing continue to fall dramatically, this would be a more useful method to study

susceptibility to mild LRTI. The current study of the genetics of susceptibility to LRTI is far from exhaustive, and further work is certainly required in this area.

The genetics of infectious disease susceptibility is rather more complicated than the genetics of many other complex diseases. Studies are complicated by the fact that not all controls may have been exposed to the infectious organism of interest, and thereby the genetic differences between cases and controls may be diluted. However, a similar problem exists in studies which utilise birth cohorts as controls for a study. In these studies, some of the control individuals will go on to develop the disease of interest later in life, and the power of the study is therefore reduced. Logistical problems involved with the collection of the perfect controls prevent this from being a viable option for most studies.

Infectious disease genetics studies have an added level of complexity due to genetic variation of the infectious organism. Susceptibility to infectious disease can be altered by the genetics of the infectious organism and the genetics of the host. Future studies should be developed to utilise information about the genetics of the host and infectious organism in parallel, to better understand the relationship between host and invading organism.

These added complications in genetics studies of infectious disease can reduce the power of these studies to detect real genetic effects, especially those with smaller effect sizes and those associated with rarer variants. Without taking these complications into account, the sepsis susceptibility GWAS has 80% power to detect an association with effect size greater than 1.39 at a p-value threshold of $5x10^{-8}$ when the variant is common (minor allele frequency of 0.3). When the variant is rare (minor allele frequency of 0.05), there is 80% power to detect an association with an odds ratio of 1.79 or higher at a p-value threshold of $5x10^{-8}$. The main case control analysis in the GRACE study both has 80% power to detect an association with effect size greater than 1.19 at a p-value threshold of $7.7x10^{-4}$ when the variant is common (minor allele

frequency of 0.3). When the variant is rare (minor allele frequency of 0.05), there is 80% power to detect an association with an odds ratio of 1.41 or higher at a p-value threshold of $7.7\text{x}10^{-4}$. Therefore, these studies do not have the power to detect associations with very low effect sizes, or with moderate effect sizes for rarer alleles. Increased sample numbers would increase the power of these studies. In the subgroups within the GRACE analysis, power is therefore even lower. In the rhinovirus analysis in the GRACE study there is 80% power to detect an odds ratio of 1.71 or higher with a rare variant at a p-value threshold of $7.7\text{x}10^{-4}$. The lowest minor allele frequency of the observed associating SNPs is 0.21. There is 80% power to detect an association with an odds ratio of 1.39 or higher with SNPs at this frequency in this study.

However, these power calculations are not accurate for those instances where the causal variant is in less than perfect LD with the variant being analysed, when power is reduced. In addition, the power of these studies is reduced by the fact that the healthy controls may go on to develop the disease under study. This is a problem especially in the study of mild LRTI, since the prevalence of mild LRTI in the general population is so high. Hence it was important to use carefully matched controls that are slightly more likely to have been exposed to the same pathogen as cases, although this is far from guaranteed.

This issue of reduced power in the GRACE study is particularly problematic in the overall analysis since it might be expected that genetic associations for mild, common, heterogeneous disease which is difficult to define would have low effect sizes. Therefore, the power to detect true associations in the GRACE study overall is likely a lot lower than the estimations above. However, when analysing LRTI caused by a specific pathogen (for instance in the case of mild LRTI caused by rhinovirus), the phenotype is more homogeneous than with many other complex diseases, and therefore higher odd ratios could be expected. The GRACE study associations identified and discussed in this thesis all have moderately high odds ratios (1.22-1.58), which is

unsurprising since there is only power to detect associations with large effects.

The p-value threshold chosen for the GRACE study only takes into account the different loci analysed in the GRACE study. It does not correct for the multiple subgroups that were analysed, nor the multiple genetic models analysed for these subgroups. The subgroups and genetic models analysed in this study are related to each other and are therefore not independent tests. Since these tests should not be corrected for as such, it is difficult to determine the correct p-value threshold to use in this study, Therefore, it is possible that some of these associations could have arisen by chance, but this could only be determined by further replication studies.

I cannot rule out possible undetected associations for the variants and phenotypes studied in this thesis, as there is not 100% power to detect an association with a small effect size. It is safe to conclude that the studied SNPs in the *MBL2* gene do not associate with the phenotypes studied at the effect size previously identified in a similar large study (2.34) (Garcia-Laorden *et al.*, 2008). However, it is possible that some/all of the positive associations that I have found and discussed are not in fact real, and these could be shown to be false positives following study with larger sample sizes, with further population stratification analysis, or with a more stringent p-value threshold to correct for the additional multiple testing performed in the GRACE study.

As the mild LRTI study was based on a candidate gene approach and followed up hits from GWAS of more severe bacterial phenotypes rather than a GWAS of mild LRTI itself, there are still potentially many more associations to be identified. Once the susceptibility to sepsis study has been completed, any replicating associations found will probably be the associations with common variants with the largest effect sizes but this study will likely miss associations with rare variants.

At this stage, these results are still suggestive, and more confident conclusions can only be made following further study. However, the associations I have found

are novel and interesting, and result from (to my knowledge) the most well-powered studies performed for these phenotypes at this time.

This thesis presents novel findings for human genetic susceptibility to common infectious disease in Europe. These findings cannot be directly translated into improved treatment for infectious disease, but it is hoped that they can generate further research which will eventually lead to improved outcome from severe LRTI and perhaps better therapy for mild LRTI.

Appendix A

IPD GWAS

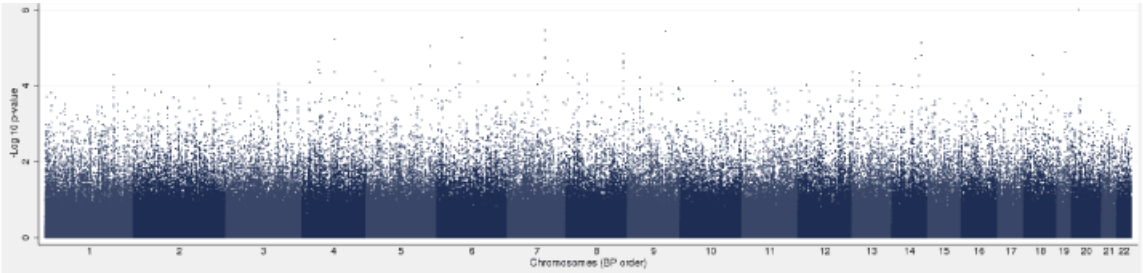


Figure A.1: Manhattan plot for susceptibility to IPD GWAS. Analysis was performed and this plot was made by Dr. Anna Rautanen.

Table A.1: IPD GWAS top 50 SNPs and their association genes. Analysis was performed by Dr. Anna Rautanen. Chr: Chromosome. bp position: base pair co-ordinate as genome build 36. OR is the odds ratio for the minor allele.

Chr	SNP	bp position	p-value	OR	Gene
1	rs1928447	57219033	$5.57\text{x}10^{-7}$	0.4886	
20	rs6041512	12578784	$9.94\text{x}10^{-7}$	0.5914	
7	rs6947185	100828478	$3.37\text{x}10^{-6}$	1.571	*EMID2*
9	rs10991437	106775741	$3.57\text{x}10^{-6}$	1.841	
14	rs10483438	32905671	$4.26\text{x}10^{-6}$	0.5993	*NPAS3*
14	rs8014138	23405979	$5.17\text{x}10^{-6}$	1.708	

Continued on next page

Chr	SNP	bp position	p-value	OR	Gene
6	rs9296536	47238947	$5.20\text{x}10^{-6}$	1.66	
4	rs17018021	92213444	$5.87\text{x}10^{-6}$	1.926	
7	rs6963365	100796435	$6.08\text{x}10^{-6}$	1.543	*EMID2*
14	rs10142146	91958094	$7.33\text{x}10^{-6}$	1.529	*SLC24A4*
5	rs2112626	166898288	$8.89\text{x}10^{-6}$	3.356	
19	rs4802515	43685916	$1.28\text{x}10^{-5}$	0.5913	*RYR1*
8	rs6578241	136511031	$1.44\text{x}10^{-5}$	0.5863	
18	rs9966505	21240053	$1.54\text{x}10^{-5}$	1.747	
14	rs10142321	91958260	$1.55\text{x}10^{-5}$	1.508	*SLC24A4*
7	rs6465799	100794277	$1.71\text{x}10^{-5}$	1.51	*EMID2*
14	rs2284230	77085650	$1.93\text{x}10^{-5}$	1.578	*SPTLC2*
7	rs2898575	100804090	$1.94\text{x}10^{-5}$	1.506	*EMID2*
8	rs2317280	136639684	$2.15\text{x}10^{-5}$	0.5975	*KHDRBS3*
8	rs17390811	2975473	$2.17\text{x}10^{-5}$	3.156	*CSMD1*
8	rs4398958	136689673	$2.26\text{x}10^{-5}$	0.5942	*KHDRBS3*
4	rs6832151	39998408	$2.26\text{x}10^{-5}$	0.618	
10	rs17691239	76938749	$2.36\text{x}10^{-5}$	0.5948	
7	rs1468596	77857412	$2.37\text{x}10^{-5}$	0.5147	*MAGI2*
8	rs6986340	136564507	$2.41\text{x}10^{-5}$	0.5956	*KHDRBS3*
12	rs838332	113479638	$2.45\text{x}10^{-5}$	1.817	
6	rs485861	42389225	$2.45\text{x}10^{-5}$	0.6347	*TRERF1*
8	rs13261331	136615955	$2.61\text{x}10^{-5}$	0.5955	*KHDRBS3*
5	rs6889293	166842883	$2.94\text{x}10^{-5}$	3.19	
8	rs6996533	136558961	$3.27\text{x}10^{-5}$	0.6019	*KHDRBS3*
4	rs4078470	40000520	$3.74\text{x}10^{-5}$	0.6349	

Continued on next page

Chr	SNP	bp position	p-value	OR	Gene
1	rs10798415	173990701	$4.00\text{x}10^{-5}$	1.71	
5	rs403186	16580431	$4.09\text{x}10^{-5}$	0.5064	*FAM134B*
4	rs4693264	92252540	$4.20\text{x}10^{-5}$	1.877	
13	rs9508811	19137913	$4.31\text{x}10^{-5}$	1.549	*MPHOSPH8*
7	rs7799285	100755083	$4.32\text{x}10^{-5}$	1.477	*EMID2*
4	rs6855691	42140836	$4.53\text{x}10^{-5}$	0.6516	*ATP8A1*
8	rs17079516	3011547	$4.63\text{x}10^{-5}$	1.931	*CSMD1*
13	rs9591390	32350358	$4.66\text{x}10^{-5}$	3.086	
8	rs12545740	33574842	$4.73\text{x}10^{-5}$	1.472	*DUSP26*
18	rs17663145	46572280	$4.88\text{x}10^{-5}$	1.58	*MRO*
7	rs6953653	89474953	$4.97\text{x}10^{-5}$	1.472	
1	rs12042283	205041402	$4.98\text{x}10^{-5}$	1.476	*IL19*
14	rs4904481	88485810	$5.14\text{x}10^{-5}$	0.6284	
7	rs2461123	45701021	$5.21\text{x}10^{-5}$	1.466	*ADCY1*
7	rs2520345	18786880	$5.25\text{x}10^{-5}$	0.5351	*HDAC9*
8	rs872137	8157397	$5.90\text{x}10^{-5}$	1.516	
9	rs560843	18734367	$5.91\text{x}10^{-5}$	1.574	
3	rs9834227	55340333	$6.49\text{x}10^{-5}$	1.492	
7	rs10263109	89483440	$6.92\text{x}10^{-5}$	1.466	

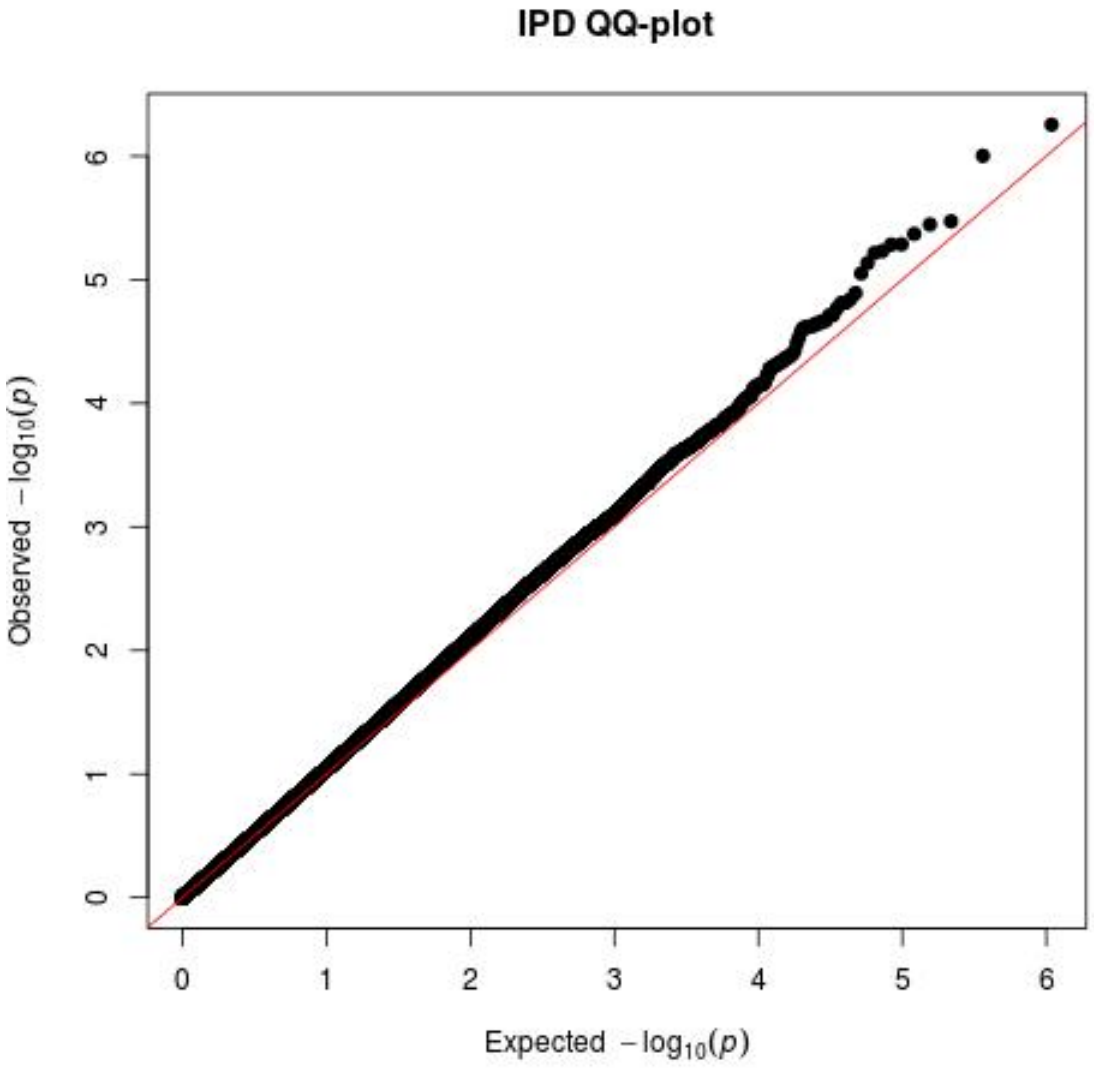


Figure A.2: QQ plot for susceptibility to IPD GWAS. Analysis performed and plot made by Dr. Anna Rautanen.

Appendix B

GenOSept study sepsis inclusion criteria

The following inclusion criteria were used for the GenOSept sepsis study collection. Only those individuals with community-acquired pneumonia or faecal peritonitis were used in this thesis. This criteria is adapted from that specified by Bone *et al.* (1992), and was compiled by the GenOSept consortium.

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 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) A body temperature greater than 38 °C or less than 36 °C

(b) A heart rate greater than 90 beats per minute

(c) A respiratory rate greater than 20 breaths per minute or a $PaCO_2$ of less than 32mm Hg

(d) A white blood cell count greater than 12,000/mm^3, less than 4,000/mm^3 or the presence of more than 10 percent immature neutrophils

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 mm Hg or PaO_2/FIO_2 <250 in the absence of pneumonia)

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

(c) Metabolic acidosis (pH of $\leq$ 7.3, or a base deficit of $\geq$ 5.0 mmol/L, or an increased lactic acid concentration)

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

(e) Platelet count $\leq$ 100,000/mm^3

(f) Coagulation abnormality $\geq$ 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

References

1000 Genomes Project Consortium (2012a). 1000 genomes: A deep catalog of human genetic variation. Available from: `http://browser.1000genomes.org/index.html`.

1000 Genomes Project Consortium (2012b). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65.

1000 Genomes Project Consortium, Durbin, R. M., Abecasis, G. R., Altshuler, D. L., Auton, A., Brooks, L. D., Gibbs, R. A., Hurles, M. E., and McVean, G. A. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73.

Abbas, A. K. and Lichtman, A. H. (2005). *Cellular and Molecular Immunology*. Elsevier Saunders, fifth edition.

Allison, A. C. (1954). Protection afforded by sickle-cell trait against subtertian malareal infection. *Br Med J*, 1(4857):290–4.

Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science*, 322(5903):881–8.

Anderson, C. A., Pettersson, F. H., Clarke, G. M., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2010). Data quality control in genetic case-control association studies. *Nat Protoc*, 5(9):1564–73.

Angus, D. C., Linde-Zwirble, W. T., Lidicker, J., Clermont, G., Carcillo, J., and Pinsky, M. R. (2001). Epidemiology of severe sepsis in the united states: analysis of incidence, outcome, and associated costs of care. *Crit Care Med*, 29(7):1303–10.

Bamshad, M. J., Ng, S. B., Bigham, A. W., Tabor, H. K., Emond, M. J., Nickerson, D. A., and Shendure, J. (2011). Exome sequencing as a tool for mendelian disease gene discovery. *Nat Rev Genet*, 12(11):745–55.

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., Julier, C., Morahan, G., Nerup, J., Nierras, C., Plagnol, V., Pociot, F., Schuilenburg, H., Smyth, D. J., Stevens, H., Todd, J. A., Walker, N. M., and Rich, S. S. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet*, 41(6):703–7.

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., Brant, S. R., Silverberg, M. S., Taylor, K. D., Barmada, M. M., Bitton, A., Dassopoulos, T., Datta, L. W., Green, T., Griffiths, A. M., Kistner, E. O., Murtha, M. T., Regueiro, M. D., Rotter, J. I., Schumm, L. P., Steinhart, A. H., Targan, S. R., Xavier, R. J., Libioulle, C., Sandor, C., Lathrop, M., Belaiche, J., Dewit, O., Gut, I., Heath, S., Laukens, D., Mni, M., Rutgeerts, P., Van Gossum, A., Zelenika, D., Franchimont, D., Hugot, J. P., de Vos, M., Vermeire, S., Louis, E., Cardon, L. R., Anderson, C. A., Drummond, H., Nimmo, E., Ahmad, T., Prescott, N. J., Onnie, C. M., Fisher, S. A., Marchini, J., Ghori, J., Bumpstead, S., Gwilliam, R., Tremelling, M., Deloukas, P., Mansfield, J., Jewell, D., Satsangi, J., Mathew, C. G., Parkes, M., Georges, M., and Daly, M. J. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat Genet*, 40(8):955–62.

Bartlett, N. W., Slater, L., Glanville, N., Haas, J. J., Caramori, G., Casolari, P., Clarke, D. L., Message, S. D., Aniscenko, J., Kebadze, T., Zhu, J., Mallia, P., Mizgerd, J. P., Belvisi, M., Papi, A., Kotenko, S. V., Johnston, S. L., and Edwards, M. R. (2012). Defining critical roles for NF-kappaB p65 and type I interferon in innate immunity to rhinovirus. *EMBO Mol Med*, 4(12):1244–60.

Begovich, A. B., Carlton, V. E., Honigberg, L. A., Schrodi, S. J., Chokkalingam, A. P., Alexander, H. C., Ardlie, K. G., Huang, Q., Smith, A. M., Spoerke, J. M., Conn, M. T.,

Chang, M., Chang, S. Y., Saiki, R. K., Catanese, J. J., Leong, D. U., Garcia, V. E., McAllister, L. B., Jeffery, D. A., Lee, A. T., Batliwalla, F., Remmers, E., Criswell, L. A., Seldin, M. F., Kastner, D. L., Amos, C. I., Sninsky, J. J., and Gregersen, P. K. (2004). A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet*, 75(2):330–7.

Bennermo, M., Nordin, M., Lundman, P., Boqvist, S., Held, C., Samnegard, A., Ericsson, C. G., Silveira, A., Hamsten, A., Nastase, M. M., and Tornvall, P. (2011). Genetic and environmental influences on the plasma interleukin-6 concentration in patients with a recent myocardial infarction: a case-control study. *J Interferon Cytokine Res*, 31(2):259–64.

Bhatti, P., Kampa, D., Alexander, B. H., McClure, C., Ringer, D., Doody, M. M., and Sigurdson, A. J. (2009). Blood spots as an alternative to whole blood collection and the effect of a small monetary incentive to increase participation in genetic association studies. *BMC Med Res Methodol*, 9:76.

Bickel, H., Gerrard, J., and Hickmans, E. M. (1953). Influence of phenylalanine intake on phenylketonuria. *Lancet*, 265(6790):812–3.

Blau, N., van Spronsen, F. J., and Levy, H. L. (2010). Phenylketonuria. *Lancet*, 376(9750):1417–27.

Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R. M., and Sibbald, W. J. (1992). Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. *Chest*, 101(6):1644–55.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*, 32(3):314–31.

Bottini, N., Musumeci, L., Alonso, A., Rahmouni, S., Nika, K., Rostamkhani, M., MacMurray, J., Meloni, G. F., Lucarelli, P., Pellecchia, M., Eisenbarth, G. S., Comings, D., and Mustelin, T. (2004). A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes. *Nat Genet*, 36(4):337–8.

Bouvier, N. M. and Palese, P. (2008). The biology of influenza viruses. *Vaccine*, 26 Suppl 4:D49–53.

Boyer, S., editor (1963). *Papers on Human Genetics*. Prentice Hall.

Brass, A. L., Huang, I. C., Benita, Y., John, S. P., Krishnan, M. N., Feeley, E. M., Ryan, B. J., Weyer, J. L., van der Weyden, L., Fikrig, E., Adams, D. J., Xavier, R. J., Farzan, M., and Elledge, S. J. (2009). The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. *Cell*, 139(7):1243–54.

Brookes-Howell, L., Elwyn, G., Hood, K., Wood, F., Cooper, L., Goossens, H., Ieven, M., and Butler, C. C. (2012a). 'The body gets used to them': patients' interpretations of antibiotic resistance and the implications for containment strategies. *J Gen Intern Med*, 27(7):766–72.

Brookes-Howell, L., Hood, K., Cooper, L., Coenen, S., Little, P., Verheij, T., Godycki-Cwirko, M., Melbye, H., Krawczyk, J., Borras-Santos, A., Jakobsen, K., Worby, P., Goossens, H., and Butler, C. C. (2012b). Clinical influences on antibiotic prescribing decisions for lower respiratory tract infection: a nine country qualitative study of variation in care. *BMJ Open*, 2(3):e000795.

Butler, C. C., Hood, K., Verheij, T., Little, P., Melbye, H., Nuttall, J., Kelly, M. J., Molstad, S., Godycki-Cwirko, M., Almirall, J., Torres, A., Gillespie, D., Rautakorpi, U., Coenen, S., and Goossens, H. (2009). Variation in antibiotic prescribing and its impact on recovery in patients with acute cough in primary care: prospective study in 13 countries. *BMJ*, 338:b2242.

Campbell, C. D., Ogburn, E. L., Lunetta, K. L., Lyon, H. N., Freedman, M. L., Groop, L. C.,

Altshuler, D., Ardlie, K. G., and Hirschhorn, J. N. (2005). Demonstrating stratification in a European American population. *Nat Genet*, 37(8):868–72.

Cardon, L. R. and Palmer, L. J. (2003). Population stratification and spurious allelic association. *Lancet*, 361(9357):598–604.

Carr, D. H. (1965). Chromosome studies in spontaneous abortions. *Obstet Gynecol*, 26:308–26.

Casalino, E., Mendoza-Sassi, G., Wolff, M., Bedos, J. P., Gaudebout, C., Regnier, B., and Vachon, F. (1998). Predictors of short- and long-term survival in HIV-infected patients admitted to the ICU. *Chest*, 113(2):421–9.

Catterall, J. R. (1999). *Streptococcus pneumoniae*. *Thorax*, 54(10):929–37.

Chakravarti, A. (2001). To a future of genetic medicine. *Nature*, 409(6822):822–3.

Chan, K. Y., Ching, J. C., Xu, M. S., Cheung, A. N., Yip, S. P., Yam, L. Y., Lai, S. T., Chu, C. M., Wong, A. T., Song, Y. Q., Huang, F. P., Liu, W., Chung, P. H., Leung, G. M., Chow, E. Y., Chan, E. Y., Chan, J. C., Ngan, H. Y., Tam, P., Chan, L. C., Sham, P., Chan, V. S., Peiris, M., Lin, S. C., and Khoo, U. S. (2007). Association of *ICAM3* genetic variant with severe acute respiratory syndrome. *J Infect Dis*, 196(2):271–80.

Chapman, S., Vannberg, F., Khor, C., Rautanen, A., Maskell, N., Davies, C., Moore, C., Day, N., Crook, D., Davies, R., and Hill, A. (2010a). Mannose-binding lectin genotypes: lack of association with susceptibility to thoracic empyema. *BMC Medical Genetics*, 11(1):5.

Chapman, S. J. and Hill, A. V. (2012). Human genetic susceptibility to infectious disease. *Nat Rev Genet*, 13(3):175–88.

Chapman, S. J., Khor, C. C., Vannberg, F. O., Frodsham, A., Walley, A., Maskell, N. A., Davies, C. W. H., Segal, S., Moore, C. E., Gillespie, S. H., Denny, P., Day, N. P., Crook, D. W., Davies, R. J. O., and Hill, A. V. S. (2007). IkappaB Genetic Polymorphisms and Invasive Pneumococcal Disease. *Am. J. Respir. Crit. Care Med.*, 176(2):181–187.

Chapman, S. J., Khor, C. C., Vannberg, F. O., Maskell, N. A., Davies, C. W. H., Hedley, E. L., Segal, S., Moore, C. E., Knox, K., Day, N. P., Gillespie, S. H., Crook, D. W., Davies, R. J. O., and Hill, A. V. S. (2006). *PTPN22* and invasive bacterial disease. *Nat Genet*, 38(5):499–500.

Chapman, S. J., Khor, C. C., Vannberg, F. O., Rautanen, A., Segal, S., Moore, C. E., Davies, R. J., Day, N. P., Peshu, N., Crook, D. W., Berkley, J. A., Williams, T. N., Scott, J. A., and Hill, A. V. (2009). *NFKBIZ* polymorphisms and susceptibility to pneumococcal disease in European and African populations. *Genes Immun*, 11(4):319–25.

Chapman, S. J., Khor, C. C., Vannberg, F. O., Rautanen, A., Walley, A., Segal, S., Moore, C. E., Davies, R. J., Day, N. P., Peshu, N., Crook, D. W., Berkley, J. A., Williams, T. N., Scott, J. A., and Hill, A. V. (2010b). Common *NFKBIL2* polymorphisms and susceptibility to pneumococcal disease: a genetic association study. *Crit Care*, 14(6):R227.

Cho, M. H., Boutaoui, N., Klanderman, B. J., Sylvia, J. S., Ziniti, J. P., Hersh, C. P., DeMeo, D. L., Hunninghake, G. M., Litonjua, A. A., Sparrow, D., Lange, C., Won, S., Murphy, J. R., Beaty, T. H., Regan, E. A., Make, B. J., Hokanson, J. E., Crapo, J. D., Kong, X., Anderson, W. H., Tal-Singer, R., Lomas, D. A., Bakke, P., Gulsvik, A., Pillai, S. G., and Silverman, E. K. (2010). Variants in *FAM13A* are associated with chronic obstructive pulmonary disease. *Nat Genet*, 42(3):200–2.

Cho, M. H., Castaldi, P. J., Wan, E. S., Siedlinski, M., Hersh, C. P., Demeo, D. L., Himes, B. E., Sylvia, J. S., Klanderman, B. J., Ziniti, J. P., Lange, C., Litonjua, A. A., Sparrow, D., Regan, E. A., Make, B. J., Hokanson, J. E., Murray, T., Hetmanski, J. B., Pillai, S. G., Kong, X., Anderson, W. H., Tal-Singer, R., Lomas, D. A., Coxson, H. O., Edwards, L. D., MacNee, W., Vestbo, J., Yates, J. C., Agusti, A., Calverley, P. M., Celli, B., Crim, C., Rennard, S., Wouters, E., Bakke, P., Gulsvik, A., Crapo, J. D., Beaty, T. H., and Silverman, E. K. (2012). A genome-wide association study of COPD identifies a susceptibility locus on chromosome 19q13. *Hum Mol Genet*, 21(4):947–57.

Chung, L. P. and Waterer, G. W. (2011). Genetic predisposition to respiratory infection and sepsis. *Crit Rev Clin Lab Sci*, 48(5-6):250–68.

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., Smink, L. J., Lam, A. C., Ovington, N. R., Stevens, H. E., Nutland, S., Howson, J. M. M., Faham, M., Moorhead, M., Jones, H. B., Falkowski, M., Hardenbol, P., Willis, T. D., and Todd, J. A. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet*, 37(11):1243–1246.

Collins, F. S., Green, E. D., Guttmacher, A. E., and Guyer, M. S. (2003). A vision for the future of genomics research. *Nature*, 422(6934):835–47.

Costa, E., Beltran, S., and Espinas, M. L. (2011). Drosophila melanogaster SAP18 protein is required for environmental stress responses. *FEBS Lett*, 585(2):275–80.

Cronin, S., Berger, S., Ding, J., Schymick, J. C., Washecka, N., Hernandez, D. G., Greenway, M. J., Bradley, D. G., Traynor, B. J., and Hardiman, O. (2008). A genome-wide association study of sporadic ALS in a homogenous Irish population. *Hum Mol Genet*, 17(5):768–74.

Crow, J. F. (1988). Eighty years ago: the beginnings of population genetics. *Genetics*, 119(3):473–6.

Dahl, M., Tybjaerg-Hansen, A., Schnohr, P., and Nordestgaard, B. G. (2004). A population-based study of morbidity and mortality in mannose-binding lectin deficiency. *J Exp Med*, 199(10):1391–9.

Dan Kephart, Steve Krueger, T. G. and Shenoi, H. (2006). Introducing the Maxwell® 16 Instrument: A Simple, Robust and Flexible Tool for DNA Purification. *Promega Notes*, 92:20–23.

Daniels, R. (2011). Surviving the first hours in sepsis: getting the basics right (an intensivist's perspective). *J Antimicrob Chemother*, 66 Suppl 2:ii11–23.

Dave, B. J. and Sanger, W. G. (2007). Role of cytogenetics and molecular cytogenetics in the diagnosis of genetic imbalances. *Semin Pediatr Neurol*, 14(1):2–6.

Davies, A., Green, C., Hutton, J., Chinn, C., European Office, Medtap International, London, UK, and Health Outcomes Res, Eli Lilly and Company Ltd, Surrey, UK (2001). Severe Sepsis: A European Estimate of the Burden of Disease in ICU. *Intensive Care Med*, 27(13):S284.

Davila, S., Wright, V. J., Khor, C. C., Sim, K. S., Binder, A., Breunis, W. B., Inwald, D., Nadel, S., Betts, H., Carrol, E. D., de Groot, R., Hermans, P. W., Hazelzet, J., Emonts, M., Lim, C. C., Kuijpers, T. W., Martinon-Torres, F., Salas, A., Zenz, W., Levin, M., and Hibberd, M. L. (2010). Genome-wide association study identifies variants in the *CFH* region associated with host susceptibility to meningococcal disease. *Nat Genet*, 42(9):772–6.

de Bakker, P. I., Yelensky, R., Pe'er, I., Gabriel, S. B., Daly, M. J., and Altshuler, D. (2005). Efficiency and power in genetic association studies. *Nat Genet*, 37(11):1217–23.

de Benedetti, F., Auriti, C., D'Urbano, L. E., Ronchetti, M. P., Rava, L., Tozzi, A., Ugazio, A. G., and Orzalesi, M. M. (2007). Low serum levels of mannose binding lectin are a risk factor for neonatal sepsis. *Pediatr Res*, 61(3):325–8.

Delaneau, O., Marchini, J., and Zagury, J. F. (2012). A linear complexity phasing method for thousands of genomes. *Nat Methods*, 9(2):179–81.

Delaneau, O., Zagury, J. F., and Marchini, J. (2011). Segmented haplotype estimation and imputation tool: Algorithm overview. Available from: `http://www.shapeit.fr/`.

Dhainaut, J. F., Laterre, P. F., Janes, J. M., Bernard, G. R., Artigas, A., Bakker, J., Riess, H., Basson, B. R., Charpentier, J., Utterback, B. G., and Vincent, J. L. (2003). Drotrecogin alfa (activated) in the treatment of severe sepsis patients with multiple-organ dysfunction: data from the PROWESS trial. *Intensive Care Med*, 29(6):894–903.

Di Gaetano, C., Voglino, F., Guarrera, S., Fiorito, G., Rosa, F., Di Blasio, A. M., Manzini, P., Dianzani, I., Betti, M., Cusi, D., Frau, F., Barlassina, C., Mirabelli, D., Magnani, C., Glorioso, N., Bonassi, S., Piazza, A., and Matullo, G. (2012). An Overview of the Genetic Structure within the Italian Population from Genome-Wide Data. *PLoS ONE*, 7(9):e43759.

Dixon, A. L., Liang, L., Moffatt, M. F., Chen, W., Heath, S., Wong, K. C., Taylor, J., Burnett, E., Gut, I., Farrall, M., Lathrop, G. M., Abecasis, G. R., and Cookson, W. O. (2007). A genome-wide association study of global gene expression. *Nat Genet*, 39(10):1202–7.

Dong, J., Hu, Z., Wu, C., Guo, H., Zhou, B., Lv, J., Lu, D., Chen, K., Shi, Y., Chu, M., Wang, C., Zhang, R., Dai, J., Jiang, Y., Cao, S., Qin, Z., Yu, D., Ma, H., Jin, G., Gong, J., Sun, C., Zhao, X., Yin, Z., Yang, L., Li, Z., Deng, Q., Wang, J., Wu, W., Zheng, H., Zhou, G., Chen, H., Guan, P., Peng, Z., Chen, Y., Shu, Y., Xu, L., Liu, X., Liu, L., Xu, P., Han, B., Bai, C., Zhao, Y., Zhang, H., Yan, Y., Amos, C. I., Chen, F., Tan, W., Jin, L., Wu, T., Lin, D., and Shen, H. (2012). Association analyses identify multiple new lung cancer susceptibility loci and their interactions with smoking in the Chinese population. *Nat Genet*, 44(8):895–9.

Doyle, W. J., Casselbrant, M. L., Li-Korotky, H. S., Doyle, A. P., Lo, C. Y., Turner, R., and Cohen, S. (2010). The interleukin 6 -174 C/C genotype predicts greater rhinovirus illness. *J Infect Dis*, 201(2):199–206.

Eisen, D. P., Dean, M. M., Boermeester, M. A., Fidler, K. J., Gordon, A. C., Kronborg, G., Kun, J. F., Lau, Y. L., Payeras, A., Valdimarsson, H., Brett, S. J., Ip, W. K., Mila, J., Peters, M. J., Saevarsdottir, S., van Till, J. W., Hinds, C. J., and McBryde, E. S. (2008). Low serum mannose-binding lectin level increases the risk of death due to pneumococcal infection. *Clin Infect Dis*, 47(4):510–6.

Eisen, D. P., Dean, M. M., Thomas, P., Marshall, P., Gerns, N., Heatley, S., Quinn, J., Minchinton, R. M., and Lipman, J. (2006). Low mannose-binding lectin function is associated with sepsis in adult patients. *FEMS Immunol Med Microbiol*, 48(2):274–82.

Endele, S., Rosenberger, G., Geider, K., Popp, B., Tamer, C., Stefanova, I., Milh, M., Kortum, F., Fritsch, A., Pientka, F. K., Hellenbroich, Y., Kalscheuer, V. M., Kohlhase, J., Moog, U., Rappold, G., Rauch, A., Ropers, H. H., von Spiczak, S., Tonnies, H., Villeneuve, N., Villard, L., Zabel, B., Zenker, M., Laube, B., Reis, A., Wieczorek, D., Van Maldergem, L., and Kutsche, K. (2010). Mutations in *GRIN2A* and *GRIN2B* encoding regulatory subunits of NMDA receptors cause variable neurodevelopmental phenotypes. *Nat Genet*, 42(11):1021–6.

Everitt, A. R., Clare, S., Pertel, T., John, S. P., Wash, R. S., Smith, S. E., Chin, C. R., Feeley, E. M., Sims, J. S., Adams, D. J., Wise, H. M., Kane, L., Goulding, D., Digard, P., Anttila, V., Baillie, J. K., Walsh, T. S., Hume, D. A., Palotie, A., Xue, Y., Colonna, V., Tyler-Smith, C., Dunning, J., Gordon, S. B., Smyth, R. L., Openshaw, P. J., Dougan, G., Brass, A. L., and Kellam, P. (2012). IFITM3 restricts the morbidity and mortality associated with influenza. *Nature*, 484(7395):519–23.

Fairfax, B. P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., Ellis, P., Langford, C., Vannberg, F. O., and Knight, J. C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet*, 44(5):502–10.

Feeley, E. M., Sims, J. S., John, S. P., Chin, C. R., Pertel, T., Chen, L. M., Gaiha, G. D., Ryan, B. J., Donis, R. O., Elledge, S. J., and Brass, A. L. (2011). IFITM3 inhibits influenza A virus infection by preventing cytosolic entry. *PLoS Pathog*, 7(10):e1002337.

Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., Cozzi-Lepri, A., De Luca, A., Easterbrook, P., Francioli, P., Mallal, S., Martinez-Picado, J., Miro, J. M., Obel, N., Smith, J. P., Wyniger, J., Descombes, P., Antonarakis, S. E., Letvin, N. L., McMichael, A. J., Haynes, B. F., Telenti, A., and Goldstein, D. B. (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science*, 317(5840):944–7.

Finch, R. G., Blasi, F. B., Verheij, T. J., Goossens, H., Coenen, S., Loens, K., Rohde, G.,

Saenz, H., and Akova, M. (2012). GRACE and the development of an education and training curriculum. *Clin Microbiol Infect*, 18(9):E308–13.

Fisher, R. (1922). On the interpretation of Chi^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94.

Fishman, D., Faulds, G., Jeffery, R., Mohamed-Ali, V., Yudkin, J. S., Humphries, S., and Woo, P. (1998). The effect of novel polymorphisms in the interleukin-6 (*IL-6*) gene on IL-6 transcription and plasma IL-6 levels, and an association with systemic-onset juvenile chronic arthritis. *J Clin Invest*, 102(7):1369–76.

Folling, A. (1934). Uber ausscheidung von phenylbrenztraubensaure in dem harn als stoffwechselanomalie in verbindung mit imbezillitat. *Zeitschrift fur Physiologische Chemie*, 227:169–176.

Fong, S. L., Zhang, J. T., Lim, C. K., Eu, K. W., and Liu, Y. (2009). Comparison of 7 methods for extracting cell-free DNA from serum samples of colorectal cancer patients. *Clin Chem*, 55(3):587–9.

Ford, C. E., Jones, K. W., Polani, P. E., De Almeida, J. C., and Briggs, J. H. (1959). A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet*, 1(7075):711–3.

Foster, C. B., Lehrnbecher, T., Samuels, S., Stein, S., Mol, F., Metcalf, J. A., Wyvill, K., Steinberg, S. M., Kovacs, J., Blauvelt, A., Yarchoan, R., and Chanock, S. J. (2000). An *IL6* promoter polymorphism is associated with a lifetime risk of development of Kaposi sarcoma in men infected with human immunodeficiency virus. *Blood*, 96(7):2562–7.

Frakking, F. N., Brouwer, N., van Eijkelenburg, N. K., Merkus, M. P., Kuijpers, T. W., Offringa, M., and Dolman, K. M. (2007). Low mannose-binding lectin (MBL) levels in neonates with pneumonia and sepsis. *Clin Exp Immunol*, 150(2):255–62.

Francis, N. A., Cals, J. W., Butler, C. C., Hood, K., Verheij, T., Little, P., Goossens, H., and

Coenen, S. (2012a). Severity assessment for lower respiratory tract infections: potential use and validity of the CRB-65 in primary care. *Prim Care Respir J*, 21(1):65–70.

Francis, N. A., Gillespie, D., Nuttall, J., Hood, K., Little, P., Verheij, T., Coenen, S., Cals, J. W., Goossens, H., and Butler, C. C. (2012b). Antibiotics for acute cough: an international observational study of patient adherence in primary care. *Br J Gen Pract*, 62(599):429–37.

Freedman, B. I., Kopp, J. B., Langefeld, C. D., Genovese, G., Friedman, D. J., Nelson, G. W., Winkler, C. A., Bowden, D. W., and Pollak, M. R. (2010). The apolipoprotein L1 (*APOL1*) gene and nondiabetic nephropathy in African Americans. *J Am Soc Nephrol*, 21(9):1422–6.

Fukuyama, S. and Kawaoka, Y. (2011). The pathogenesis of influenza virus infections: the contributions of virus and host factors. *Curr Opin Immunol*, 23(4):481–6.

GABRIEL (2012). Available from: `http://www.cng.fr/gabriel/index.html`.

Garcia-Laorden, M. I., Rodriguez de Castro, F., Sole-Violan, J., Rajas, O., Blanquer, J., Borderias, L., Aspa, J., Briones, M. L., Saavedra, P., Marcos-Ramos, J. A., Gonzalez-Quevedo, N., Sologuren, I., Herrera-Ramos, E., Ferrer, J. M., Rello, J., and Rodriguez-Gallego, C. (2011). Influence of genetic variability at the surfactant proteins A and D in community-acquired pneumonia: a prospective, observational, genetic study. *Crit Care*, 15(1):R57.

Garcia-Laorden, M. I., Sole-Violan, J., Rodriguez de Castro, F., Aspa, J., Briones, M. L., Garcia-Saavedra, A., Rajas, O., Blanquer, J., Caballero-Hidalgo, A., Marcos-Ramos, J. A., Hernandez-Lopez, J., and Rodriguez-Gallego, C. (2008). Mannose-binding lectin and mannose-binding lectin-associated serine protease 2 in susceptibility, severity, and outcome of pneumonia in adults. *J Allergy Clin Immunol*, 122(2):368–74, 374 e1–2.

Garred, P. (2008). Mannose-binding lectin genetics: from A to Z. *Biochem Soc Trans*, 36(Pt 6):1461–6.

Garred, P., J, J. S., Quist, L., Taaning, E., and Madsen, H. O. (2003). Association of mannose-binding lectin polymorphisms with sepsis and fatal outcome, in patients with systemic inflammatory response syndrome. *J Infect Dis*, 188(9):1394–403.

Garred, P., Larsen, F., Seyfarth, J., Fujita, R., and Madsen, H. O. (2006). Mannose-binding lectin and its genetic variants. *Genes Immun*, 7(2):85–94.

Ge, D., Fellay, J., Thompson, A. J., Simon, J. S., Shianna, K. V., Urban, T. J., Heinzen, E. L., Qiu, P., Bertelsen, A. H., Muir, A. J., Sulkowski, M., McHutchison, J. G., and Goldstein, D. B. (2009). Genetic variation in *IL28B* predicts hepatitis C treatment-induced viral clearance. *Nature*, 461(7262):399–401.

GE Healthcare (2006). illustra GenomiPhi V2 DNA Amplification Kit. Technical report.

Genovese, G., Friedman, D. J., Ross, M. D., Lecordier, L., Uzureau, P., Freedman, B. I., Bowden, D. W., Langefeld, C. D., Oleksyk, T. K., Uscinski Knob, A. L., Bernhardy, A. J., Hicks, P. J., Nelson, G. W., Vanhollebeke, B., Winkler, C. A., Kopp, J. B., Pays, E., and Pollak, M. R. (2010). Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science*, 329(5993):841–5.

Gentile, D. A., Doyle, W. J., Zeevi, A., Howe-Adams, J., Kapadia, S., Trecki, J., and Skoner, D. P. (2003). Cytokine gene polymorphisms moderate illness severity in infants with respiratory syncytial virus infection. *Hum Immunol*, 64(3):338–44.

Gibson, G. (2012). Rare and common variants: twenty arguments. *Nat Rev Genet*, 13(2):135–45.

Goetghebuer, T., Kwiatkowski, D., Thomson, A., and Hull, J. (2004). Familial susceptibility to severe respiratory infection in early life. *Pediatr Pulmonol*, 38(4):321–8.

Gordon, A. C., Waheed, U., Hansen, T. K., Hitman, G. A., Garrard, C. S., Turner, M. W., Klein, N. J., Brett, S. J., and Hinds, C. J. (2006). Mannose-binding lectin polymorphisms in severe sepsis: relationship to levels, incidence, and outcome. *Shock*, 25(1):88–93.

Greene, G., Hood, K., Little, P., Verheij, T., Goossens, H., Coenen, S., and Butler, C. C. (2011). Towards clinical definitions of lower respiratory tract infection (LRTI) for research and primary care practice in Europe: an international consensus study. *Prim Care Respir J*, 20(3):299–306.

Gregersen, P. K., Amos, C. I., Lee, A. T., Lu, Y., Remmers, E. F., Kastner, D. L., Seldin, M. F., Criswell, L. A., Plenge, R. M., Holers, V. M., Mikuls, T. R., Sokka, T., Moreland, L. W., Bridges, S. L., J., Xie, G., Begovich, A. B., and Siminovitch, K. A. (2009). REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nat Genet*, 41(7):820–3.

Gusella, J. F., Wexler, N. S., Conneally, P. M., Naylor, S. L., Anderson, M. A., Tanzi, R. E., Watkins, P. C., Ottina, K., Wallace, M. R., Sakaguchi, A. Y., and et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, 306(5940):234–8.

Guthrie, R. and Susi, A. (1963). A simple phenylalanine method for detecting phenylketonuria in large populations of newborn infants. *Pediatrics*, 32:338–43.

Hakonarson, H., Grant, S. F., Bradfield, J. P., Marchand, L., Kim, C. E., Glessner, J. T., Grabs, R., Casalunovo, T., Taback, S. P., Frackelton, E. C., Lawson, M. L., Robinson, L. J., Skraban, R., Lu, Y., Chiavacci, R. M., Stanley, C. A., Kirsch, S. E., Rappaport, E. F., Orange, J. S., Monos, D. S., Devoto, M., Qu, H. Q., and Polychronakos, C. (2007). A genome-wide association study identifies *KIAA0350* as a type 1 diabetes gene. *Nature*, 448(7153):591–4.

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science*, 28(706):49–50.

Harper, P. (2004). *Landmarks in Medical Genetics*. Oxford University Press.

Hart, B. E. and Tapping, R. I. (2012). Cell surface trafficking of TLR1 is differentially regulated by the chaperones PRAT4A and PRAT4B. *J Biol Chem*, 287(20):16550–62.

Hay, J. (1813). Account of a remarkable haemorrhagic disposition, existing in many individuals of the same family. *New England Journal of Medicine*, 2:221–225.

He, J., Feng, D., de Vlas, S. J., Wang, H., Fontanet, A., Zhang, P., Plancoulaine, S., Tang, F., Zhan, L., Yang, H., Wang, T., Richardus, J. H., Habbema, J. D., and Cao, W. (2006). Association of SARS susceptibility with single nucleic acid polymorphisms of *OAS1* and *MxA* genes: a case-control study. *BMC Infect Dis*, 6:106.

Hill, A. V. (2012). Evolution, revolution and heresy in the genetics of infectious disease susceptibility. *Philos Trans R Soc Lond B Biol Sci*, 367(1590):840–9.

Hindorff, L., MacArthur, J., Morales, J., Junkins, H., Hall, P., Klemm, A., and Manolio, T. (2012). A catalog of published genome-wide association studies. Available from: `www.genome.gov/gwastudies`.

Hirschhorn, J. N. and Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet*, 6(2):95–108.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Purvis, I., and Xu, C. F. (2004). Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *Eur J Hum Genet*, 12(5):395–9.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*, 44(8):955–9.

Howie, B., Marchini, J., and Stephens, M. (2011). Genotype imputation with thousands of genomes. *G3 (Bethesda)*, 1(6):457–70.

Howie, B. N., Donnelly, P., and Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, 5(6):e1000529.

Huang, I. C., Bailey, C. C., Weyer, J. L., Radoshitzky, S. R., Becker, M. M., Chiang, J. J., Brass, A. L., Ahmed, A. A., Chi, X., Dong, L., Longobardi, L. E., Boltz, D., Kuhn, J. H.,

Elledge, S. J., Bavari, S., Denison, M. R., Choe, H., and Farzan, M. (2011). Distinct patterns of IFITM-mediated restriction of filoviruses, SARS coronavirus, and influenza A virus. *PLoS Pathog*, 7(1):e1001258.

Huh, J. W., Song, K., Yum, J. S., Hong, S. B., Lim, C. M., and Koh, Y. (2009). Association of mannose-binding lectin-2 genotype and serum levels with prognosis of sepsis. *Crit Care*, 13(6):R176.

Huntington, G. (1872). On chorea. *Medical and Surgical Reporter*, 26:320–321.

Huntington's Disease Collaborative Research Group (1993). A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. *Cell*, 72(6):971–83.

Hutter, G., Nowak, D., Mossner, M., Ganepola, S., Mussig, A., Allers, K., Schneider, T., Hofmann, J., Kucherer, C., Blau, O., Blau, I. W., Hofmann, W. K., and Thiel, E. (2009). Long-Term Control of HIV by CCR5 Delta32/Delta32 Stem-Cell Transplantation. *N Engl J Med*, 360(7):692–698.

Ingram, G. I. (1976). The history of haemophilia. *J Clin Pathol*, 29(6):469–79.

International Multiple Sclerosis Genetics Consortium, Wellcome Trust Case Controls Consortium 2, Sawcer, S., Hellenthal, G., Pirinen, M., Spencer, C. C., Patsopoulos, N. A., Moutsianas, L., Dilthey, A., Su, Z., Freeman, C., Hunt, S. E., Edkins, S., Gray, E., Booth, D. R., Potter, S. C., Goris, A., Band, G., Oturai, A. B., Strange, A., Saarela, J., Bellenguez, C., Fontaine, B., Gillman, M., Hemmer, B., Gwilliam, R., Zipp, F., Jayakumar, A., Martin, R., Leslie, S., Hawkins, S., Giannoulatou, E., D'Alfonso, S., Blackburn, H., Martinelli Boneschi, F., Liddle, J., Harbo, H. F., Perez, M. L., Spurkland, A., Waller, M. J., Mycko, M. P., Ricketts, M., Comabella, M., Hammond, N., Kockum, I., McCann, O. T., Ban, M., Whittaker, P., Kemppinen, A., Weston, P., Hawkins, C., Widaa, S., Zajicek, J., Dronov, S., Robertson, N., Bumpstead, S. J., Barcellos, L. F., Ravindrarajah, R., Abraham, R., Alfredsson, L., Ardlie, K., Aubin, C., Baker, A., Baker, K.,

Baranzini, S. E., Bergamaschi, L., Bergamaschi, R., Bernstein, A., Berthele, A., Boggild, M., Bradfield, J. P., Brassat, D., Broadley, S. A., Buck, D., Butzkueven, H., Capra, R., Carroll, W. M., Cavalla, P., Celius, E. G., Cepok, S., Chiavacci, R., Clerget-Darpoux, F., Clysters, K., Comi, G., Cossburn, M., Cournu-Rebeix, I., Cox, M. B., Cozen, W., Cree, B. A., Cross, A. H., Cusi, D., Daly, M. J., Davis, E., de Bakker, P. I., Debouverie, M., D'Hooghe M, B., Dixon, K., Dobosi, R., Dubois, B., Ellinghaus, D., *et al.* (2011). Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, 476(7359):214–9.

Invitrogen (2008). Quant-iT PicoGreen dsDNA Reagent and Kits. Technical report.

Ip, W. K., Chan, K. H., Law, H. K., Tso, G. H., Kong, E. K., Wong, W. H., To, Y. F., Yung, R. W., Chow, E. Y., Au, K. L., Chan, E. Y., Lim, W., Jensenius, J. C., Turner, M. W., Peiris, J. S., and Lau, Y. L. (2005). Mannose-binding lectin in severe acute respiratory syndrome coronavirus infection. *J Infect Dis*, 191(10):1697–704.

Jacobs, P. A. and Strong, J. A. (1959). A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature*, 183(4657):302–3.

Jallow, M., Teo, Y. Y., Small, K. S., Rockett, K. A., Deloukas, P., Clark, T. G., Kivinen, K., Bojang, K. A., Conway, D. J., Pinder, M., Sirugo, G., Sisay-Joof, F., Usen, S., Auburn, S., Bumpstead, S. J., Campino, S., Coffey, A., Dunham, A., Fry, A. E., Green, A., Gwilliam, R., Hunt, S. E., Inouye, M., Jeffreys, A. E., Mendy, A., Palotie, A., Potter, S., Ragoussis, J., Rogers, J., Rowlands, K., Somaskantharajah, E., Whittaker, P., Widden, C., Donnelly, P., Howie, B., Marchini, J., Morris, A., SanJoaquin, M., Achidi, E. A., Agbenyega, T., Allen, A., Amodu, O., Corran, P., Djimde, A., Dolo, A., Doumbo, O. K., Drakeley, C., Dunstan, S., Evans, J., Farrar, J., Fernando, D., Hien, T. T., Horstmann, R. D., Ibrahim, M., Karunaweera, N., Kokwaro, G., Koram, K. A., Lemnge, M., Makani, J., Marsh, K., Michon, P., Modiano, D., Molyneux, M. E., Mueller, I., Parker, M., Peshu, N., Plowe, C. V., Puijalon, O., Reeder, J., Reyburn, H., Riley, E. M., Sakuntabhai, A., Singhasivanon, P., Sirima, S., Tall, A., Taylor, T. E., Thera, M., Troye-Blomberg, M.,

Williams, T. N., Wilson, M., and Kwiatkowski, D. P. (2009). Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat Genet*, 41(6):657–65.

Janssen, R., Bont, L., Siezen, C. L., Hodemaekers, H. M., Ermers, M. J., Doornbos, G., van 't Slot, R., Wijmenga, C., Goeman, J. J., Kimpen, J. L., van Houwelingen, H. C., Kimman, T. G., and Hoebee, B. (2007). Genetic susceptibility to respiratory syncytial virus bronchiolitis is predominantly associated with innate immune genes. *J Infect Dis*, 196(6):826–34.

Jiang, D., Weidner, J. M., Qing, M., Pan, X. B., Guo, H., Xu, C., Zhang, X., Birk, A., Chang, J., Shi, P. Y., Block, T. M., and Guo, J. T. (2010). Identification of five interferon-induced cellular proteins that inhibit west nile virus and dengue virus infections. *J Virol*, 84(16):8332–41.

Jin, Y., Birlea, S. A., Fain, P. R., Gowan, K., Riccardi, S. L., Holland, P. J., Mailloux, C. M., Sufit, A. J., Hutton, S. M., Amadi-Myers, A., Bennett, D. C., Wallace, M. R., McCormack, W. T., Kemp, E. H., Gawkrodger, D. J., Weetman, A. P., Picardo, M., Leone, G., Taieb, A., Jouary, T., Ezzedine, K., van Geel, N., Lambert, J., Overbeck, A., and Spritz, R. A. (2010). Variant of *TYR* and autoimmunity susceptibility loci in generalized vitiligo. *N Engl J Med*, 362(18):1686–97.

Johnson, A. D., Handsaker, R. E., Pulit, S. L., Nizzari, M. M., O'Donnell, C. J., and de Bakker, P. I. (2008). SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics*, 24(24):2938–9.

Johnson, C. M., Lyle, E. A., Omueti, K. O., Stepensky, V. A., Yegin, O., Alpsoy, E., Hamann, L., Schumann, R. R., and Tapping, R. I. (2007). Cutting edge: A common polymorphism impairs cell surface trafficking and functional responses of TLR1 but protects against leprosy. *J Immunol*, 178(12):7520–4.

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., Lee, J. C., Philip Schumm, L., Sharma, Y., Anderson, C. A., Essers, J., Mitrovic, M., Ning, K.,

Cleynen, I., Theatre, E., Spain, S. L., Raychaudhuri, S., Goyette, P., Wei, Z., Abraham, C., Achkar, J.-P., Ahmad, T., Amininejad, L., Ananthakrishnan, A. N., Andersen, V., Andrews, J. M., Baidoo, L., Balschun, T., Bampton, P. A., Bitton, A., Boucher, G., Brand, S., Buning, C., Cohain, A., Cichon, S., D'Amato, M., De Jong, D., Devaney, K. L., Dubinsky, M., Edwards, C., Ellinghaus, D., Ferguson, L. R., Franchimont, D., Fransen, K., Gearry, R., Georges, M., Gieger, C., Glas, J., Haritunians, T., Hart, A., Hawkey, C., Hedl, M., Hu, X., Karlsen, T. H., Kupcinskas, L., Kugathasan, S., Latiano, A., Laukens, D., Lawrance, I. C., Lees, C. W., Louis, E., Mahy, G., Mansfield, J., Morgan, A. R., Mowat, C., Newman, W., Palmieri, O., Ponsioen, C. Y., Potocnik, U., Prescott, N. J., Regueiro, M., Rotter, J. I., Russell, R. K., Sanderson, J. D., Sans, M., Satsangi, J., Schreiber, S., Simms, L. A., Sventoraityte, J., Targan, S. R., Taylor, K. D., Tremelling, M., Verspaget, H. W., De Vos, M., Wijmenga, C., Wilson, D. C., Winkelmann, J., Xavier, R. J., Zeissig, S., Zhang, B., Zhang, C. K., Zhao, H., Silverberg, M. S., Annese, V., Hakonarson, H., Brant, S. R., Radford-Smith, G., Mathew, C. G., Rioux, J. D., Schadt, E. E., *et al.* (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature*, 491(7422):119–124.

Kamatani, Y., Wattanapokayakit, S., Ochi, H., Kawaguchi, T., Takahashi, A., Hosono, N., Kubo, M., Tsunoda, T., Kamatani, N., Kumada, H., Puseenam, A., Sura, T., Daigo, Y., Chayama, K., Chantratita, W., Nakamura, Y., and Matsuda, K. (2009). A genome-wide association study identifies variants in the *HLA-DP* locus associated with chronic hepatitis B in Asians. *Nat Genet*, 41(5):591–5.

Kan, Y. W. and Dozy, A. M. (1978). Antenatal diagnosis of sickle-cell anaemia by D.N.A. analysis of amniotic-fluid cells. *Lancet*, 2(8096):910–2.

Kaplan, L. C., Wharton, R., Elias, E., Mandell, F., Donlon, T., and Latt, S. A. (1987). Clinical heterogeneity associated with deletions in the long arm of chromosome 15: report of 3 new cases and their possible genetic significance. *Am J Med Genet*, 28(1):45–53.

Khoo, U. S., Chan, K. Y., Chan, V. S., Ching, J. C., Yam, L., Chu, C. M., Lai, S. T.,

Wong, T. Y., Tam, P., Yip, S. P., Leung, G. M., Lin, C. L., and Peiris, J. S. (2008). Role of polymorphisms of the inflammatory response genes and *DC-SIGNR* in genetic susceptibility to SARS and other infections. *Hong Kong Med J*, 14 Suppl 4:31–5.

Khor, C. C., Chapman, S. J., Vannberg, F. O., Dunne, A., Murphy, C., Ling, E. Y., Frodsham, A. J., Walley, A. J., Kyrieleis, O., Khan, A., Aucan, C., Segal, S., Moore, C. E., Knox, K., Campbell, S. J., Lienhardt, C., Scott, A., Aaby, P., Sow, O. Y., Grignani, R. T., Sillah, J., Sirugo, G., Peshu, N., Williams, T. N., Maitland, K., Davies, R. J. O., Kwiatkowski, D. P., Day, N. P., Yala, D., Crook, D. W., Marsh, K., Berkley, J. A., O'Neill, L. A. J., and Hill, A. V. S. (2007). A mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. *Nat Genet*, 39(4):523–528.

Khor, C. C., Chau, T. N., Pang, J., Davila, S., Long, H. T., Ong, R. T., Dunstan, S. J., Wills, B., Farrar, J., Van Tram, T., Gan, T. T., Binh, N. T., Tri le, T., Lien le, B., Tuan, N. M., Tham, N. T., Lanh, M. N., Nguyet, N. M., Hieu, N. T., Van, N. V. C. N., Thuy, T. T., Tan, D. E., Sakuntabhai, A., Teo, Y. Y., Hibberd, M. L., and Simmons, C. P. (2011). Genome-wide association study identifies susceptibility loci for dengue shock syndrome at *MICB* and *PLCE1*. *Nat Genet*, 43(11):1139–41.

Khor, C. C., Vannberg, F. O., Chapman, S. J., Guo, H., Wong, S. H., Walley, A. J., Vukcevic, D., Rautanen, A., Mills, T. C., Chang, K.-C., Kam, K.-M., Crampin, A. C., Ngwira, B., Leung, C.-C., Tam, C.-M., Chan, C.-Y., Sung, J. J., Yew, W.-W., Toh, K.-Y., Tay, S. K., Kwiatkowski, D., Lienhardt, C., Hien, T.-T., Day, N. P., Peshu, N., Marsh, K., Maitland, K., Scott, J. A., Williams, T. N., Berkley, J. A., Floyd, S., Tang, N. L., Fine, P. E., Goh, D. L., and Hill, A. V. (2010). *CISH* and Susceptibility to Infectious Diseases. *New England Journal of Medicine*, 362(22):2092–2101.

Kimura, H., Miyashita, H., Suzuki, Y., Kobayashi, M., Watanabe, K., Sonoda, H., Ohta, H., Fujiwara, T., Shimosegawa, T., and Sato, Y. (2009). Distinctive localization and

opposed roles of vasohibin-1 and vasohibin-2 in the regulation of angiogenesis. *Blood*, 113(19):4810–8.

Knapp, S., Yee, L. J., Frodsham, A. J., Hennig, B. J., Hellier, S., Zhang, L., Wright, M., Chiaramonte, M., Graves, M., Thomas, H. C., Hill, A. V., and Thursz, M. R. (2003). Polymorphisms in interferon-induced genes and the outcome of hepatitis C virus infection: roles of *MxA*, *OAS-1* and *PKR*. *Genes Immun*, 4(6):411–9.

Knoll, J. H., Nicholls, R. D., Magenis, R. E., Graham, J. M., J., Lalande, M., and Latt, S. A. (1989). Angelman and Prader-Willi syndromes share a common chromosome 15 deletion but differ in parental origin of the deletion. *Am J Med Genet*, 32(2):285–90.

Kokura, K., Sun, L., Bedford, M. T., and Fang, J. (2010). Methyl-H3K9-binding protein MPP8 mediates E-cadherin gene silencing and promotes tumour cell motility and invasion. *EMBO J*, 29(21):3673–87.

Konno, K., Wakabayashi, Y., Akashi-Takamura, S., Ishii, T., Kobayashi, M., Takahashi, K., Kusumoto, Y., Saitoh, S., Yoshizawa, Y., and Miyake, K. (2006). A molecule that is associated with toll-like receptor 4 and regulates its cell surface expression. *Biochem Biophys Res Commun*, 339(4):1076–82.

Korosec Jagodic, H., Jagodic, K., and Podbregar, M. (2006). Long-term outcome and quality of life of patients treated in surgical intensive care: a comparison between sepsis and trauma. *Crit Care*, 10(5):R134.

Kraft, P., Zeggini, E., and Ioannidis, J. P. (2009). Replication in genome-wide association studies. *Stat Sci*, 24(4):561–573.

Kristensen, I. A., Thiel, S., Steffensen, R., Madhi, S., Sorour, G., and Olsen, J. (2004). Mannan-binding lectin and RSV lower respiratory tract infection leading to hospitalization in children: a case-control study from Soweto, South Africa. *Scand J Immunol*, 60(1-2):184–8.

Kumar, G., Kumar, N., Taneja, A., Kaleekal, T., Tarima, S., McGinley, E., Jimenez, E., Mohan, A., Khan, R. A., Whittle, J., Jacobs, E., and Nanchal, R. (2011). Nationwide trends of severe sepsis in the 21st century (2000-2007). *Chest*, 140(5):1223–31.

Kyogoku, C., Langefeld, C. D., Ortmann, W. A., Lee, A., Selby, S., Carlton, V. E., Chang, M., Ramos, P., Baechler, E. C., Batliwalla, F. M., Novitzke, J., Williams, A. H., Gillett, C., Rodine, P., Graham, R. R., Ardlie, K. G., Gaffney, P. M., Moser, K. L., Petri, M., Begovich, A. B., Gregersen, P. K., and Behrens, T. W. (2004). Genetic association of the R620W polymorphism of protein tyrosine phosphatase *PTPN22* with human SLE. *Am J Hum Genet*, 75(3):504–7.

Lahermo, P., Liljedahl, U., Alnaes, G., Axelsson, T., Brookes, A. J., Ellonen, P., Groop, P. H., Hallden, C., Holmberg, D., Holmberg, K., Keinanen, M., Kepp, K., Kere, J., Kiviluoma, P., Kristensen, V., Lindgren, C., Odeberg, J., Osterman, P., Parkkonen, M., Saarela, J., Sterner, M., Stromqvist, L., Talas, U., Wessman, M., Palotie, A., and Syvanen, A. C. (2006). A quality assessment survey of SNP genotyping laboratories. *Hum Mutat*, 27(7):711–4.

Lamb, R. A. and Choppin, P. W. (1983). The gene structure and replication of influenza virus. *Annu Rev Biochem*, 52:467–506.

Lan, Q., Hsiung, C. A., Matsuo, K., Hong, Y. C., Seow, A., Wang, Z., Hosgood, H. D., r., Chen, K., Wang, J. C., Chatterjee, N., Hu, W., Wong, M. P., Zheng, W., Caporaso, N., Park, J. Y., Chen, C. J., Kim, Y. H., Kim, Y. T., Landi, M. T., Shen, H., Lawrence, C., Burdett, L., Yeager, M., Yuenger, J., Jacobs, K. B., Chang, I. S., Mitsudomi, T., Kim, H. N., Chang, G. C., Bassig, B. A., Tucker, M., Wei, F., Yin, Z., Wu, C., An, S. J., Qian, B., Lee, V. H., Lu, D., Liu, J., Jeon, H. S., Hsiao, C. F., Sung, J. S., Kim, J. H., Gao, Y. T., Tsai, Y. H., Jung, Y. J., Guo, H., Hu, Z., Hutchinson, A., Wang, W. C., Klein, R., Chung, C. C., Oh, I. J., Chen, K. Y., Berndt, S. I., He, X., Wu, W., Chang, J., Zhang, X. C., Huang, M. S., Zheng, H., Wang, J., Zhao, X., Li, Y., Choi, J. E., Su, W. C., Park, K. H., Sung, S. W., Shu, X. O., Chen, Y. M., Liu, L., Kang, C. H., Hu, L., Chen, C. H.,

Pao, W., Kim, Y. C., Yang, T. Y., Xu, J., Guan, P., Tan, W., Su, J., Wang, C. L., Li, H., Sihoe, A. D., Zhao, Z., Chen, Y., Choi, Y. Y., Hung, J. Y., Kim, J. S., Yoon, H. I., Cai, Q., Lin, C. C., Park, I. K., Xu, P., Dong, J., Kim, C., He, Q., Perng, R. P., Kohno, T., Kweon, S. S., *et al.* (2012). Genome-wide association analysis identifies new lung cancer susceptibility loci in never-smoking women in Asia. *Nat Genet*, 44(12):1330–5.

Lander, E. S. (1996). The new genomics: global views of biology. *Science*, 274(5287):536–9.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J. P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J. C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R. H., Wilson, R. K., Hillier, L. W., McPherson, J. D., Marra, M. A., Mardis, E. R., Fulton, L. A., Chinwalla, A. T., Pepin, K. H., Gish, W. R., Chissoe, S. L., Wendl, M. C., Delehaunty, K. D., Miner, T. L., Delehaunty, A., Kramer, J. B., Cook, L. L., Fulton, R. S., Johnson, D. L., Minx, P. J., Clifton, S. W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J. F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921.

Larsen, F., Madsen, H. O., Sim, R. B., Koch, C., and Garred, P. (2004). Disease-associated mutations in human mannose-binding lectin compromise oligomerization and activity of the final protein. *J Biol Chem*, 279(20):21302–11.

Lawrence, E., editor (2000). *Henderson's Dictionary of Biological Terms*. Pearson Education Limited, 12th edition.

Ledbetter, D. H., Riccardi, V. M., Airhart, S. D., Strobel, R. J., Keenan, B. S., and Crawford, J. D. (1981). Deletions of chromosome 15 as a cause of the Prader-Willi syndrome. *N Engl J Med*, 304(6):325–9.

Lederman, M. M., Penn-Nicholson, A., Cho, M., and Mosier, D. (2006). Biology of CCR5 and its role in HIV infection and treatment. *JAMA*, 296(7):815–26.

Lee, W. M., Lemanske, R. F., J., Evans, M. D., Vang, F., Pappas, T., Gangnon, R., Jackson, D. J., and Gern, J. E. (2012). Human rhinovirus species and season of infection determine illness severity. *Am J Respir Crit Care Med*, 186(9):886–91.

Lejeune, J., Gautier, M., and Turpin, R. (1959). Etude des chromosomes somatiques de neuf enfants mongoliens. *Comptes rendus des seances de l'Academie des Sciences*, 248:1721–1722.

Levy, H. L., Shih, V. E., Karolkewicz, V., and MacCready, R. A. (1970). Screening for phenylketonuria. *Lancet*, 2(7671):522–3.

Levy, M. M., Artigas, A., Phillips, G. S., Rhodes, A., Beale, R., Osborn, T., Vincent, J. L., Townsend, S., Lemeshow, S., and Dellinger, R. P. (2012). Outcomes of the Surviving Sepsis Campaign in intensive care units in the USA and Europe: a prospective cohort study. *Lancet Infect Dis*, 12(12):919–24.

Li, H., Tang, N. L., Chan, P. K., Wang, C. Y., Hui, D. S., Luk, C., Kwok, R., Huang, W., Sung, J. J., Kong, Q. P., and Zhang, Y. P. (2008). Polymorphisms in the C-type lectin genes cluster in chromosome 19 and predisposition to severe acute respiratory syndrome coronavirus (SARS-CoV) infection. *J Med Genet*, 45(11):752–8.

Li, Y., Willer, C., Sanna, S., and Abecasis, G. (2009). Genotype imputation. *Annu Rev Genomics Hum Genet*, 10:387–406.

Lipscombe, R. J., Sumiya, M., Hill, A. V., Lau, Y. L., Levinsky, R. J., Summerfield, J. A., and Turner, M. W. (1992). High frequencies in African and non-African populations of independent mutations in the mannose binding protein gene. *Hum Mol Genet*, 1(9):709–15.

Liu, R., Paxton, W. A., Choe, S., Ceradini, D., Martin, S. R., Horuk, R., MacDonald, M. E., Stuhlmann, H., Koup, R. A., and Landau, N. R. (1996). Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell*, 86(3):367–77.

Liu, Y., Gao, W., and Zhang, D. (2010). Effects of cigarette smoke extract on A549 cells and human lung fibroblasts treated with transforming growth factor-beta1 in a coculture system. *Clin Exp Med*, 10(3):159–67.

Ma, X., Liu, Y., Gowen, B. B., Graviss, E. A., Clark, A. G., and Musser, J. M. (2007). Full-exon resequencing reveals toll-like receptor variants contribute to human susceptibility to tuberculosis disease. *PLoS ONE*, 2(12):e1318.

Macfarlane, J., Holmes, W., Gard, P., Macfarlane, R., Rose, D., Weston, V., Leinonen, M., Saikku, P., and Myint, S. (2001). Prospective study of the incidence, aetiology and outcome of adult lower respiratory tract illness in the community. *Thorax*, 56(2):109–14.

Madsen, H. O., Garred, P., Kurtzhals, J. A., Lamm, L. U., Ryder, L. P., Thiel, S., and Svejgaard, A. (1994). A new frequent allele is the missing link in the structural polymorphism of the human mannan-binding protein. *Immunogenetics*, 40(1):37–44.

Madsen, H. O., Garred, P., Thiel, S., Kurtzhals, J. A., Lamm, L. U., Ryder, L. P., and Svejgaard, A. (1995). Interplay between promoter and structural gene variants control basal serum level of mannan-binding protein. *J Immunol*, 155(6):3013–20.

Madsen, H. O., Satz, M. L., Hogh, B., Svejgaard, A., and Garred, P. (1998). Different Molecular Events Result in Low Protein Levels of Mannan-Binding Lectin in Populations from Southeast Africa and South America. *J Immunol*, 161(6):3169–3175.

Makowski, G. S., Davis, E. L., and Hopfer, S. M. (1997). Amplification of Guthrie card DNA: effect of guanidine thiocyanate on binding of natural whole blood PCR inhibitors. *J Clin Lab Anal*, 11(2):87–93.

Man, M., Close, S. L., Shaw, A. D., Bernard, G. R., Douglas, I. S., Kaner, R. J., Payen, D., Vincent, J. L., Fossceco, S., Janes, J. M., Leishman, A. G., O'Brien, L., Williams, M. D., and Garcia, J. G. (2012). Beyond single-marker analyses: mining whole genome scans for insights into treatment responses in severe sepsis. *Pharmacogenomics J*, Epub ahead of print.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet*, 11(7):499–511.

Marchini, J., Howie, B., Myers, S., McVean, G. A., and Donnelly, P. (2010). SNPTEST v2. Available from: `https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html`.

Martin-Loeches, I., Sole-Violan, J., Rodriguez de Castro, F., Garcia-Laorden, M. I., Borderias, L., Blanquer, J., Rajas, O., Briones, M. L., Aspa, J., Herrera-Ramos, E., Marcos-Ramos, J. A., Sologuren, I., Gonzalez-Quevedo, N., Ferrer-Aguero, J. M., Noda, J., and Rodriguez-Gallego, C. (2012). Variants at the promoter of the interleukin-6 gene are associated with severity and outcome of pneumococcal community-acquired pneumonia. *Intensive Care Med*, 38(2):256–62.

Matsushita, M., Ezekowitz, R. A., and Fujita, T. (1995). The Gly-54–>Asp allelic form

of human mannose-binding protein (MBP) fails to bind MBP-associated serine protease. *Biochem J*, 311 (Pt 3):1021–3.

Miyairi, I. and DeVincenzo, J. P. (2008). Human genetic factors and respiratory syncytial virus disease severity. *Clin Microbiol Rev*, 21(4):686–703.

Miyashita, H., Watanabe, T., Hayashi, H., Suzuki, Y., Nakamura, T., Ito, S., Ono, M., Hoshikawa, Y., Okada, Y., Kondo, T., and Sato, Y. (2012). Angiogenesis Inhibitor Vasohibin-1 Enhances Stress Resistance of Endothelial Cells via Induction of SOD2 and SIRT1. *PLoS ONE*, 7(10):e46459.

Mizgerd, J. P. (2006). Lung infection–a public health priority. *PLoS Med*, 3(2):e76.

Moffatt, M. F., Gut, I. G., Demenais, F., Strachan, D. P., Bouzigon, E., Heath, S., von Mutius, E., Farrall, M., Lathrop, M., and Cookson, W. O. (2010). A large-scale, consortium-based genomewide association study of asthma. *N Engl J Med*, 363(13):1211–21.

Mullighan, C. G., Heatley, S. L., Danner, S., Dean, M. M., Doherty, K., Hahn, U., Bradstock, K. F., Minchinton, R., Schwarer, A. P., Szer, J., and Bardy, P. G. (2008). Mannose-binding lectin status is associated with risk of major infection following myeloablative sibling allogeneic hematopoietic stem cell transplantation. *Blood*, 112(5):2120–8.

Mullis, K. B. and Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol*, 155:335–50.

Naito, H., Ma, Y., Uemura, K., Arano, Y., and Kawasaki, T. (1999). Metabolic properties of normal and mutant mannan-binding proteins in mouse plasma. *Biochem Biophys Res Commun*, 256(1):231–4.

Nakada, T. A., Russell, J. A., Boyd, J. H., and Walley, K. R. (2011). *IL17A* genetic variation is associated with altered susceptibility to Gram-positive infection and mortality of severe sepsis. *Crit Care*, 15(5):R254.

National Center for Biotechnology Information, U. N. L. o. M. (2012). Basic local alignment search tool.

National Health Service (2012). Phenylketonuria. Available from: `www.nhs.uk/conditions/Phenylketonuria/Pages/Introduction.aspx`.

Nawijn, M. C., Hackett, T. L., Postma, D. S., van Oosterhout, A. J., and Heijink, I. H. (2011). E-cadherin: gatekeeper of airway mucosa and allergic sensitization. *Trends Immunol*, 32(6):248–55.

Nee, P. A. (2006). Critical care in the emergency department: severe sepsis and septic shock. *Emerg Med J*, 23(9):713–7.

Nejentsev, S., Walker, N., Riches, D., Egholm, M., and Todd, J. A. (2009). Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science*, 324(5925):387–9.

Nelis, M., Esko, T., Magi, R., Zimprich, F., Zimprich, A., Toncheva, D., Karachanak, S., Piskackova, T., Balascak, I., Peltonen, L., Jakkula, E., Rehnstrom, K., Lathrop, M., Heath, S., Galan, P., Schreiber, S., Meitinger, T., Pfeufer, A., Wichmann, H. E., Melegh, B., Polgar, N., Toniolo, D., Gasparini, P., D'Adamo, P., Klovins, J., Nikitina-Zake, L., Kucinskas, V., Kasnauskiene, J., Lubinski, J., Debniak, T., Limborska, S., Khrunin, A., Estivill, X., Rabionet, R., Marsal, S., Julia, A., Antonarakis, S. E., Deutsch, S., Borel, C., Attar, H., Gagnebin, M., Macek, M., Krawczak, M., Remm, M., and Metspalu, A. (2009). Genetic structure of Europeans: a view from the North-East. *PLoS ONE*, 4(5):e5472.

Nielsen (2011). Electronic point of sale data recorded and published by AC Nielsen GB. 52 week moving annual total to w/e 16th April 2011 based on scanning data with supplementary audits (covering 60,000 retail outlets in Great Britain). Technical report.

Oeth, P., Beaulieu, M., Park, C., Kosman, D., Del Mistro, G., Van den Boom, D., and Jurinke, C. (2007). iPLEX™ Assay: Increased Plexing Efficiency and Flexibility for MassARRAY® System Through Single Base Primer Extension with Mass-Modified Terminators.

Oeth, P., del Mistro, G., Marnellos, G., Shi, T., and van den Boom, D. (2009). Qualitative and quantitative genotyping using single base primer extension coupled with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MassARRAY). *Methods Mol Biol*, 578:307–43.

Ozdemir, O., Dinleyici, E. C., Tekin, N., Colak, O., and Aksit, M. A. (2010). Low-mannose-binding lectin levels in susceptibility to neonatal sepsis in preterm neonates with fetal inflammatory response syndrome. *J Matern Fetal Neonatal Med*, 23(9):1009–13.

Paterson, G. K. and Mitchell, T. J. (2006). Innate immunity and the pneumococcus. *Microbiology*, 152(Pt 2):285–93.

Pearson, K. (1900). On the criterion that a given suste, of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50(302):157–175.

Pei, Y. F., Zhang, L., Li, J., and Deng, H. W. (2010). Analyses and comparison of imputation-based association methods. *PLoS ONE*, 5(5):e10827.

Penrose, L. S. (1946). Phenylketonuria; a problem in eugenics. *Lancet*, 1(6409):949–53.

Pereyra, F., Jia, X., McLaren, P. J., Telenti, A., de Bakker, P. I., Walker, B. D., Ripke, S., Brumme, C. J., Pulit, S. L., Carrington, M., Kadie, C. M., Carlson, J. M., Heckerman, D., Graham, R. R., Plenge, R. M., Deeks, S. G., Gianniny, L., Crawford, G., Sullivan, J., Gonzalez, E., Davies, L., Camargo, A., Moore, J. M., Beattie, N., Gupta, S., Crenshaw, A., Burtt, N. P., Guiducci, C., Gupta, N., Gao, X., Qi, Y., Yuki, Y., Piechocka-Trocha, A., Cutrell, E., Rosenberg, R., Moss, K. L., Lemay, P., O'Leary, J., Schaefer, T., Verma, P., Toth, I., Block, B., Baker, B., Rothchild, A., Lian, J., Proudfoot, J., Alvino, D. M., Vine, S., Addo, M. M., Allen, T. M., Altfeld, M., Henn, M. R., Le Gall, S., Streeck, H., Haas, D. W., Kuritzkes, D. R., Robbins, G. K., Shafer, R. W., Gulick, R. M., Shikuma, C. M., Haubrich, R., Riddler, S., Sax, P. E., Daar, E. S., Ribaudo, H. J., Agan, B., Agarwal, S., Ahern, R. L., Allen, B. L., Altidor, S., Altschuler, E. L., Ambardar, S.,

Anastos, K., Anderson, B., Anderson, V., Andrady, U., Antoniskis, D., Bangsberg, D., Barbaro, D., Barrie, W., Bartczak, J., Barton, S., Basden, P., Basgoz, N., Bazner, S., Bellos, N. C., Benson, A. M., Berger, J., Bernard, N. F., Bernard, A. M., Birch, C., Bodner, S. J., Bolan, R. K., Boudreaux, E. T., Bradley, M., Braun, J. F., Brndjar, J. E., Brown, S. J., Brown, K., Brown, S. T., *et al.* (2010). The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science*, 330(6010):1551–7.

Perl, T. M., Dvorak, L., Hwang, T., and Wenzel, R. P. (1995). Long-term survival and function after suspected gram-negative sepsis. *JAMA*, 274(4):338–45.

Petersen, L., Andersen, P. K., and Sorensen, T. I. (2010). Genetic influences on incidence and case-fatality of infectious disease. *PLoS ONE*, 5(5):e10603.

Petersen, L., Nielsen, G. G., Andersen, P. K., and Sorensen, T. I. (2002). Case-control study of genetic and environmental influences on premature death of adult adoptees. *Genet Epidemiol*, 23(2):123–32.

Pillai, S. G., Ge, D., Zhu, G., Kong, X., Shianna, K. V., Need, A. C., Feng, S., Hersh, C. P., Bakke, P., Gulsvik, A., Ruppert, A., Lodrup Carlsen, K. C., Roses, A., Anderson, W., Rennard, S. I., Lomas, D. A., Silverman, E. K., and Goldstein, D. B. (2009). A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet*, 5(3):e1000421.

Pillai, S. G., Kong, X., Edwards, L. D., Cho, M. H., Anderson, W. H., Coxson, H. O., Lomas, D. A., and Silverman, E. K. (2010). Loci identified by genome-wide association studies influence different disease-related phenotypes in chronic obstructive pulmonary disease. *Am J Respir Crit Care Med*, 182(12):1498–505.

Pirinen, M., Donnelly, P., and Spencer, C. C. (2012). Including known covariates can reduce power to detect genetic effects in case-control studies. *Nat Genet*, 44(8):848–51.

Plenge, R. M., Seielstad, M., Padyukov, L., Lee, A. T., Remmers, E. F., Ding, B., Liew, A., Khalili, H., Chandrasekaran, A., Davies, L. R., Li, W., Tan, A. K., Bonnard, C., Ong,

R. T., Thalamuthu, A., Pettersson, S., Liu, C., Tian, C., Chen, W. V., Carulli, J. P., Beckman, E. M., Altshuler, D., Alfredsson, L., Criswell, L. A., Amos, C. I., Seldin, M. F., Kastner, D. L., Klareskog, L., and Gregersen, P. K. (2007). TRAF1-C5 as a risk locus for rheumatoid arthritis–a genomewide study. *N Engl J Med*, 357(12):1199–209.

Pruim, R. J., Welch, R. P., Sanna, S., Teslovich, T. M., Chines, P. S., Gliedt, T. P., Boehnke, M., Abecasis, G. R., and Willer, C. J. (2010). LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*, 26(18):2336–7.

Purcell, S. (2012). PLINK v1.07. Available from: `http://pngu.mgh.harvard.edu/purcell/plink/`.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., de Bakker, P. I., Daly, M. J., and Sham, P. C. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–75.

Rani, R., Singh, A., Israni, N., Sharma, P., and Kar, H. K. (2009). The role of polymorphic protein tyrosine phosphatase non-receptor type 22 in leprosy. *J Invest Dermatol*, 129(11):2726–8.

Rauch, A., Kutalik, Z., Descombes, P., Cai, T., Di Iulio, J., Mueller, T., Bochud, M., Battegay, M., Bernasconi, E., Borovicka, J., Colombo, S., Cerny, A., Dufour, J. F., Furrer, H., Gunthard, H. F., Heim, M., Hirschel, B., Malinverni, R., Moradpour, D., Mullhaupt, B., Witteck, A., Beckmann, J. S., Berg, T., Bergmann, S., Negro, F., Telenti, A., and Bochud, P. Y. (2010). Genetic variation in *IL28B* is associated with chronic hepatitis C and treatment failure: a genome-wide association study. *Gastroenterology*, 138(4):1338–45, 1345 e1–7.

Reed, G. H., Kent, J. O., and Wittwer, C. T. (2007). High-resolution DNA melting analysis for simple and efficient molecular diagnostics. *Pharmacogenomics*, 8(6):597–608.

Roberts, R. J. (2005). How restriction enzymes became the workhorses of molecular biology. *Proc Natl Acad Sci U S A*, 102(17):5905–8.

Roos, A., Garred, P., Wildenberg, M. E., Lynch, N. J., Munoz, J. R., Zuiverloon, T. C., Bouwman, L. H., Schlagwein, N., Fallaux van den Houten, F. C., Faber-Krol, M. C., Madsen, H. O., Schwaeble, W. J., Matsushita, M., Fujita, T., and Daha, M. R. (2004). Antibody-mediated activation of the classical pathway of complement may compensate for mannose-binding lectin deficiency. *Eur J Immunol*, 34(9):2589–98.

Rossmann, M. G., Arnold, E., Erickson, J. W., Frankenberger, E. A., Griffith, J. P., Hecht, H. J., Johnson, J. E., Kamer, G., Luo, M., Mosser, A. G., R.R., R., B., S., and G., V. (1985). Structure of a human common cold virus and functional relationship to other picornaviruses. *Nature*, 317(6033):145–53.

Rowell, J. L., Dowling, N. F., Yu, W., Yesupriya, A., Zhang, L., and Gwinn, M. (2012). Trends in population-based studies of human genetics in infectious diseases. *PLoS ONE*, 7(2):e25431.

Roy, S., Knox, K., Segal, S., Griffiths, D., Moore, C. E., Welsh, K. I., Smarason, A., Day, N. P., McPheat, W. L., Crook, D. W., and Hill, A. V. S. (2002). Mbl genotype and risk of invasive pneumococcal disease: a case-control study. *The Lancet*, 359(9317):1569–1573.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Coggill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P. Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomann, N., Zody, M. C., Linton, L., Lander, E. S., and Altshuler, D. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822):928–33.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, 239(4839):487–91.

Salmela, E., Lappalainen, T., Liu, J., Sistonen, P., Andersen, P. M., Schreiber, S., Savontaus, M. L., Czene, K., Lahermo, P., Hall, P., and Kere, J. (2011). Swedish population substructure revealed by genome-wide single nucleotide polymorphism data. *PLoS ONE*, 6(2):e16747.

Salnikova, L. E., Smelaya, T. V., Moroz, V. V., Golubev, A. M., and Rubanovich, A. V. (2012). Host genetic risk factors for community-acquired pneumonia. *Gene*, Epub ahead of print.

Samson, M., Libert, F., Doranz, B. J., Rucker, J., Liesnard, C., Farber, C. M., Saragosti, S., Lapoumeroulie, C., Cognaux, J., Forceille, C., Muyldermans, G., Verhofstede, C., Burtonboy, G., Georges, M., Imai, T., Rana, S., Yi, Y., Smyth, R. J., Collman, R. G., Doms, R. W., Vassart, G., and Parmentier, M. (1996). Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the *CCR-5* chemokine receptor gene. *Nature*, 382(6593):722–5.

Sasse, K. C., Nauenberg, E., Long, A., Anton, B., Tucker, H. J., and Hu, T. W. (1995). Long-term survival after intensive care unit admission with sepsis. *Crit Care Med*, 23(6):1040–7.

Schoggins, J. W., Wilson, S. J., Panis, M., Murphy, M. Y., Jones, C. T., Bieniasz, P., and Rice, C. M. (2011). A diverse range of gene products are effectors of the type I interferon antiviral response. *Nature*, 472(7344):481–5.

Scott, J. A., Berkley, J. A., Mwangi, I., Ochola, L., Uyoga, S., Macharia, A., Ndila, C., Lowe, B. S., Mwarumba, S., Bauni, E., Marsh, K., and Williams, T. N. (2011). Relation between falciparum malaria and bacteraemia in Kenyan children: a population-based, case-control study and a longitudinal study. *Lancet*, 378(9799):1316–23.

Sequenom, Inc. (2010a). iPLEX Pro Application Guide. Technical report, Sequenom, Inc.

Sequenom, Inc. (2010b). Typer 4.0.20. Technical report, Sequenom, Inc.

Shapiro, N. I., Howell, M. D., Talmor, D., Donnino, M., Ngo, L., and Bates, D. W. (2007). Mortality in Emergency Department Sepsis (MEDS) score predicts 1-year mortality. *Crit Care Med*, 35(1):192–8.

Siner, J. (2009). Sepsis: Definitions, epidemiology, etiology and pathogenesis. Available from: `http://69.36.35.38/accp/pccsu/sepsis-definitions-epidemiology-etiology-and-pathogenesis?page=0,3`.

Smyth, D., Cooper, J. D., Collins, J. E., Heward, J. M., Franklyn, J. A., Howson, J. M., Vella, A., Nutland, S., Rance, H. E., Maier, L., Barratt, B. J., Guja, C., Ionescu-Tirgoviste, C., Savage, D. A., Dunger, D. B., Widmer, B., Strachan, D. P., Ring, S. M., Walker, N., Clayton, D. G., Twells, R. C., Gough, S. C., and Todd, J. A. (2004). Replication of an association between the lymphoid tyrosine phosphatase locus (*LYP/PTPN22*) with type 1 diabetes, and evidence for its role as a general autoimmunity locus. *Diabetes*, 53(11):3020–3.

Sole-Violan, J., de Castro, F., Garcia-Laorden, M. I., Blanquer, J., Aspa, J., Borderias, L., Briones, M. L., Rajas, O., Carrondo, I. M., Marcos-Ramos, J. A., Ferrer Aguero, J. M., Garcia-Saavedra, A., Fiuza, M. D., Caballero-Hidalgo, A., and Rodriguez-Gallego, C. (2010). Genetic variability in the severity and outcome of community-acquired pneumonia. *Respir Med*, 104(3):440–7.

Sole-Violan, J., Garcia-Laorden, M. I., Marcos-Ramos, J. A., de Castro, F. R., Rajas, O., Borderias, L., Briones, M. L., Herrera-Ramos, E., Blanquer, J., Aspa, J., Florido, Y., Garcia-Bello, M. A., Ferrer-Aguero, J. M., Sologuren, I., and Rodriguez-Gallego, C. (2011). The Fcgamma receptor IIA-H/H131 genotype is associated with bacteremia in pneumococcal community-acquired pneumonia. *Crit Care Med*, 39(6):1388–93.

Song, J. H., Dagan, R., Klugman, K. P., and Fritzell, B. (2012). The relationship between pneumococcal serotypes and antibiotic resistance. *Vaccine*, 30(17):2728–37.

Sorensen, T. I., Nielsen, G. G., Andersen, P. K., and Teasdale, T. W. (1988). Genetic and environmental influences on premature death in adult adoptees. *N Engl J Med*, 318(12):727–32.

Sorin, M., Cano, J., Das, S., Mathew, S., Wu, X., Davies, K. P., Shi, X., Cheng, S. W., Ott, D., and Kalpana, G. V. (2009). Recruitment of a SAP18-HDAC1 complex into HIV-1 virions and its requirement for viral replication. *PLoS Pathog*, 5(6):e1000463.

Spencer, C. C., Su, Z., Donnelly, P., and Marchini, J. (2009). Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*, 5(5):e1000477.

Stahl, E. A., Raychaudhuri, S., Remmers, E. F., Xie, G., Eyre, S., Thomson, B. P., Li, Y., Kurreeman, F. A., Zhernakova, A., Hinks, A., Guiducci, C., Chen, R., Alfredsson, L., Amos, C. I., Ardlie, K. G., Barton, A., Bowes, J., Brouwer, E., Burtt, N. P., Catanese, J. J., Coblyn, J., Coenen, M. J., Costenbader, K. H., Criswell, L. A., Crusius, J. B., Cui, J., de Bakker, P. I., De Jager, P. L., Ding, B., Emery, P., Flynn, E., Harrison, P., Hocking, L. J., Huizinga, T. W., Kastner, D. L., Ke, X., Lee, A. T., Liu, X., Martin, P., Morgan, A. W., Padyukov, L., Posthumus, M. D., Radstake, T. R., Reid, D. M., Seielstad, M., Seldin, M. F., Shadick, N. A., Steer, S., Tak, P. P., Thomson, W., van der Helm-van Mil, A. H., van der Horst-Bruinsma, I. E., van der Schoot, C. E., van Riel, P. L., Weinblatt, M. E., Wilson, A. G., Wolbink, G. J., Wordsworth, B. P., Wijmenga, C., Karlson, E. W., Toes, R. E., de Vries, N., Begovich, A. B., Worthington, J., Siminovitch, K. A., Gregersen, P. K., Klareskog, L., and Plenge, R. M. (2010). Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet*, 42(6):508–14.

Stern, C. (1943). The Hardy-Weinberg Law. *Science*, 97(2510):137–8.

Sumiya, M., Super, M., Tabona, P., Levinsky, R. J., Arai, T., Turner, M. W., and Summerfield, J. A. (1991). Molecular basis of opsonic defect in immunodeficient children. *Lancet*, 337(8757):1569–70.

Sutherland, A. M. and Walley, K. R. (2009). Bench-to-bedside review: Association of genetic variation with sepsis. *Crit Care*, 13(2):210.

Tabor, H. K., Risch, N. J., and Myers, R. M. (2002). Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet*, 3(5):391–7.

Tang, B. M., Huang, S. J., and McLean, A. S. (2010). Genome-wide transcription profiling of human sepsis: a systematic review. *Crit Care*, 14(6):R237.

Taylor, D. E. and Piantadosi, C. A. (1995). Oxidative metabolism in sepsis and sepsis syndrome. *J Crit Care*, 10(3):122–35.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–320.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678.

Thomas, D. L., Thio, C. L., Martin, M. P., Qi, Y., Ge, D., O'Huigin, C., Kidd, J., Kidd, K., Khakoo, S. I., Alexander, G., Goedert, J. J., Kirk, G. D., Donfield, S. M., Rosen, H. R., Tobler, L. H., Busch, M. P., McHutchison, J. G., Goldstein, D. B., and Carrington, M. (2009). Genetic variation in *IL28B* and spontaneous clearance of hepatitis C virus. *Nature*, 461(7265):798–801.

Thorven, M., Grahn, A., Hedlund, K. O., Johansson, H., Wahlfrid, C., Larson, G., and Svensson, L. (2005). A homozygous nonsense mutation (428G–>A) in the human secretor (*FUT2*) gene provides resistance to symptomatic norovirus (GGII) infections. *J Virol*, 79(24):15351–5.

Thye, T., Vannberg, F. O., Wong, S. H., Owusu-Dabo, E., Osei, I., Gyapong, J., Sirugo, G., Sisay-Joof, F., Enimil, A., Chinbuah, M. A., Floyd, S., Warndorff, D. K., Sichali, L., Malema, S., Crampin, A. C., Ngwira, B., Teo, Y. Y., Small, K., Rockett, K., Kwiatkowski, D., Fine, P. E., Hill, P. C., Newport, M., Lienhardt, C., Adegbola, R. A., Corrah, T.,

Ziegler, A., Morris, A. P., Meyer, C. G., Horstmann, R. D., and Hill, A. V. (2010). Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat Genet*, 42(9):739–41.

Tijo, J. and Levan, A. (1956). The chromosome number of man. *Hereditas*, 42:1–6.

Timmann, C., Thye, T., Vens, M., Evans, J., May, J., Ehmen, C., Sievertsen, J., Muntau, B., Ruge, G., Loag, W., Ansong, D., Antwi, S., Asafo-Adjei, E., Nguah, S. B., Kwakye, K. O., Akoto, A. O., Sylverken, J., Brendel, M., Schuldt, K., Loley, C., Franke, A., Meyer, C. G., Agbenyega, T., Ziegler, A., and Horstmann, R. D. (2012). Genome-wide association study indicates two novel resistance loci for severe malaria. *Nature*, 489(7416):443–6.

Todd, J. A., Walker, N. M., Cooper, J. D., Smyth, D. J., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S. F., Payne, F., Lowe, C. E., Szeszko, J. S., Hafler, J. P., Zeitels, L., Yang, J. H., Vella, A., Nutland, S., Stevens, H. E., Schuilenburg, H., Coleman, G., Maisuria, M., Meadows, W., Smink, L. J., Healy, B., Burren, O. S., Lam, A. A., Ovington, N. R., Allen, J., Adlem, E., Leung, H. T., Wallace, C., Howson, J. M., Guja, C., Ionescu-Tirgoviste, C., Simmonds, M. J., Heward, J. M., Gough, S. C., Dunger, D. B., Wicker, L. S., and Clayton, D. G. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet*, 39(7):857–64.

Turner, M. W. (2003). The role of mannose-binding lectin in health and disease. *Mol Immunol*, 40(7):423–9.

Turner, S. (2012). Getting genetics done. Available from: `http://gettinggeneticsdone.blogspot.co.uk/`.

University of California, Santa Cruz (2012a). UCSC Genome Bioinformatics. Available from: `http://hgdownload.cse.ucsc.edu/downloads.html`.

University of California, Santa Cruz (2012b). UCSC Genome Bioinformatics. Available from: `http://genome.cse.ucsc.edu/index.html`.

Urban, T., Charlton, M. R., and Goldstein, D. B. (2012). Introduction to the genetics and biology of interleukin-28B. *Hepatology*, 56(1):361–6.

van Woensel, J. B., van Aalderen, W. M., and Kimpen, J. L. (2003). Viral lower respiratory tract infection in infants and young children. *BMJ*, 327(7405):36–40.

Vincent, J. L., Sakr, Y., Sprung, C. L., Ranieri, V. M., Reinhart, K., Gerlach, H., Moreno, R., Carlet, J., Le Gall, J. R., and Payen, D. (2006). Sepsis in European intensive care units: results of the SOAP study. *Crit Care Med*, 34(2):344–53.

Vossen, R., Aten, E., Roos, A., and den Dunnen, J. (2009). High-Resolution Melting Analysis (HRMA) - More than just sequence variant screening. *Human Mutation*, 30(6):860–866.

Vossen, R. H., Duijn, M. v., Daha, M. R., Dunnen, J. T. d., and Roos, A. (2010). High-throughput genotyping of mannose-binding lectin variants using high-resolution DNA-melting analysis. *Human Mutation*, 31(4):E1286–93.

Wang, S., Wei, M., Han, Y., Zhang, K., He, L., Yang, Z., Su, B., Zhang, Z., Hu, Y., and Hui, W. (2008). Roles of TNF-alpha gene polymorphisms in the occurrence and progress of SARS-Cov infection: a case-control study. *BMC Infect Dis*, 8:27.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8.

Wei, X., Walia, V., Lin, J. C., Teer, J. K., Prickett, T. D., Gartner, J., Davis, S., Stemke-Hale, K., Davies, M. A., Gershenwald, J. E., Robinson, W., Robinson, S., Rosenberg, S. A., and Samuels, Y. (2011). Exome sequencing identifies *GRIN2A* as frequently mutated in melanoma. *Nat Genet*, 43(5):442–6.

Weycker, D., Akhras, K. S., Edelsberg, J., Angus, D. C., and Oster, G. (2003). Long-term mortality and medical care charges in patients with severe sepsis. *Crit Care Med*, 31(9):2316–23.

Wigginton, J. E., Cutler, D. J., and Abecasis, G. R. (2005). A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet*, 76(5):887–93.

Winters, B. D., Eberlein, M., Leung, J., Needham, D. M., Pronovost, P. J., and Sevransky, J. E. (2010). Long-term mortality and quality of life in sepsis: a systematic review. *Crit Care Med*, 38(5):1276–83.

Wittwer, C. T., Reed, G. H., Gundry, C. N., Vandersteen, J. G., and Pryor, R. J. (2003). High-resolution genotyping by amplicon melting analysis using LCGreen. *Clin Chem*, 49(6 Pt 1):853–60.

Wong, S. H., Gochhait, S., Malhotra, D., Pettersson, F. H., Teo, Y. Y., Khor, C. C., Rautanen, A., Chapman, S. J., Mills, T. C., Srivastava, A., Rudko, A., Freidin, M. B., Puzyrev, V. P., Ali, S., Aggarwal, S., Chopra, R., Reddy, B. S., Garg, V. K., Roy, S., Meisner, S., Hazra, S. K., Saha, B., Floyd, S., Keating, B. J., Kim, C., Fairfax, B. P., Knight, J. C., Hill, P. C., Adegbola, R. A., Hakonarson, H., Fine, P. E., Pitchappan, R. M., Bamezai, R. N., Hill, A. V., and Vannberg, F. O. (2010a). Leprosy and the adaptation of human toll-like receptor 1. *PLoS Pathog*, 6:e1000979.

Wong, S. H., Hill, A. V., and Vannberg, F. O. (2010b). Genomewide association study of leprosy. *N Engl J Med*, 362(15):1446–7; author reply 1447–8.

World Health Organisation (2008). World health organisation health statistics and heath information system. Available from: `http://www.who.int/healthinfo/global_burden_disease/estimates_regional/en/index.html`.

World Health Organisation (2011). The top 10 causes of death, fact sheet no 310. Available from: `http://www.who.int/mediacentre/factsheets/fs310/en/index.html`.

Yeo, N. K. and Jang, Y. J. (2010). Rhinovirus infection-induced alteration of tight junction and adherens junction components in human nasal epithelial cells. *Laryngoscope*, 120(2):346–52.

Young, F. (1985). Multidimensional scaling. In Kotz, S. and Johnson, N., editors, *Encyclopedia of Statistical Science*, volume 5. John Wiley and Sons, Inc.

Yuan, F. F., Marks, K., Wong, M., Watson, S., de Leon, E., McIntyre, P. B., and Sullivan, J. S. (2008). Clinical relevance of *TLR2*, *TLR4*, *CD14* and *FcgammaRIIA* gene polymorphisms in *Streptococcus pneumoniae* infection. *Immunol Cell Biol*, 86(3):268–270.

Yuan, F. F., Tanner, J., Chan, P. K., Biffin, S., Dyer, W. B., Geczy, A. F., Tang, J. W., Hui, D. S., Sung, J. J., and Sullivan, J. S. (2005). Influence of *FcgammaRIIA* and *MBL* polymorphisms on severe acute respiratory syndrome. *Tissue Antigens*, 66(4):291–6.

Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K., Rossmann, H., Eleftheriadis, M., Sinning, C. R., Schnabel, R. B., Lubos, E., Mennerich, D., Rust, W., Perret, C., Proust, C., Nicaud, V., Loscalzo, J., Hubner, N., Tregouet, D., Munzel, T., Ziegler, A., Tiret, L., Blankenberg, S., and Cambien, F. (2010). Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PLoS ONE*, 5(5):e10693.

Zhang, F. R., Huang, W., Chen, S. M., Sun, L. D., Liu, H., Li, Y., Cui, Y., Yan, X. X., Yang, H. T., Yang, R. D., Chu, T. S., Zhang, C., Zhang, L., Han, J. W., Yu, G. Q., Quan, C., Yu, Y. X., Zhang, Z., Shi, B. Q., Zhang, L. H., Cheng, H., Wang, C. Y., Lin, Y., Zheng, H. F., Fu, X. A., Zuo, X. B., Wang, Q., Long, H., Sun, Y. P., Cheng, Y. L., Tian, H. Q., Zhou, F. S., Liu, H. X., Lu, W. S., He, S. M., Du, W. L., Shen, M., Jin, Q. Y., Wang, Y., Low, H. Q., Erwin, T., Yang, N. H., Li, J. Y., Zhao, X., Jiao, Y. L., Mao, L. G., Yin, G., Jiang, Z. X., Wang, X. D., Yu, J. P., Hu, Z. H., Gong, C. H., Liu, Y. Q., Liu, R. Y., Wang, D. M., Wei, D., Liu, J. X., Cao, W. K., Cao, H. Z., Li, Y. P., Yan, W. G., Wei, S. Y., Wang, K. J., Hibberd, M. L., Yang, S., Zhang, X. J., and Liu, J. J. (2009). Genomewide association study of leprosy. *N Engl J Med*, 361(27):2609–18.

Zhang, H., Zhou, G., Zhi, L., Yang, H., Zhai, Y., Dong, X., Zhang, X., Gao, X., Zhu, Y., and He, F. (2005). Association between mannose-binding lectin gene polymorphisms and

susceptibility to severe acute respiratory syndrome coronavirus infection. *J Infect Dis*, 192(8):1355–61.

Zondervan, K. T. and Cardon, L. R. (2004). The complex interplay among factors that influence allelic association. *Nat Rev Genet*, 5(2):89–100.