Genetic susceptibility to common mycobacterial diseases

Hei Sunny Wong

Green Templeton College

Nuffield Department of Clinical Medicine

Thesis submitted for the degree of Doctor of Philosophy

Trinity Term, 2010

Abstract

Common mycobacterial diseases, including tuberculosis and leprosy, contribute to major mortality and morbidity worldwide. Despite evidence of an important role of host genetic factors in susceptibility to these infections, few compelling genetic associations have been identified with previous candidate gene and linkage approaches.

This thesis investigates the genetic factors of human immunity to these mycobacterial diseases using a high-throughput approach of association testing. To assess genetic susceptibility to tuberculosis, I have conducted a genome-wide association study in the Gambian population as part of the Wellcome Trust Case Control Consortium (WTCCC). The study reveals the region flanking *CADM1* as a potential susceptibility locus. Combining this study with a Ghanaian cohort further implicates two genetic loci at chromosome 18q11.2 ($P = 9.2 \times 10^{-9}$) and *PARD3B* ($P = 1.4 \times 10^{-6}$). For leprosy, I have performed a gene-centric association study in the New Delhi Indian population. Evidence of significant association was observed in the *HLA-DRB1/DQA1* ($P = 4.9 \times 10^{-14}$) and *TLR1* ($P = 1.7 \times 10^{-9}$) loci. These studies identify important genomic regions that may be involved in immunity to tuberculosis and leprosy. Further analysis revealed a significant immunogenetic overlap between tuberculosis and leprosy. This provides proof-of-principle for the subsequent aggregate analysis for mycobacterial susceptibility, which suggests that the steroid biosynthesis pathway may be important in anti-mycobacterial immunity.

This thesis represents one of the largest studies to identify the genetic factors for human immunity against mycobacteria. These novel findings will further enhance vaccine and pharmaceutical efforts into prevention and treatment of these mycobacterial diseases.

Acknowledgements

I would like to thank my supervisor Professor Adrian Hill for his enormous support throughout my study. It has been a wonderful experience working in the group.

I am indebted to my co-supervisor and great friend Dr Fredrik Vannberg for his excellent mentorship. He has made research fun by introducing the Saturday experiments. I also thank my colleagues Anna Rautanen, Sophie Roetynck, David Wyllie and Elma Tchilian for their teaching, as well as Tara Mills, Magda Ellis, Chiea-Chuen Khor, Yik-Ting Teo, Andrew Morris, Damjan Vukcevic, Benjamin Fairfax, Julian Knight, Louise van der Weyden, Nelson Tang and Chiu-Yeung Chan for their insights. It would have been a much harder time without their support.

I would like to thank my undergraduate supervisors, Professors Dennis Lo and Rossa Chiu, for their enlightening teaching and education. Without them I would not have started my doctoral study. I am thankful to Professors Joseph Sung and Ho-Keung Ng for their teaching and guidance.

I am very grateful to Dr Deanna Lee for her generous support and great education. She has taught me the meaning of benevolence.

I thank Jasmine Lim for her great friendship. It would have been a far more stressful time without her support. I also thank Queenie Wong, Angeline Lo, Clarice Kwok, May Chan, Lynn Yu, David Cooper, Yves Schymura, David Chan, Ye Dee Tay, Larry

Lai, Vincent Wong, Chris Wong and James Kim for their incessant encouragements.

I thank my girlfriend, Angie Au, for her constant love. She has given me the motivation to carry on my research. Lastly, I would also like to thank my mother for cultivating me the empathy and compassion for others, and my father showing me the meaning of integrity and wisdom. I have learnt from them that love can indeed make a better world for others.

Attributions

The cohort collections used for the genetic studies in this thesis were collected and managed over many years. I would like to thank the following individuals for their work on the study samples: Richard Adegbola, Fatou Sisay-Joof, Tumani Corrah, Christian Lienhardt, Melanie Newport, Philip Hill, Giorgio Sirugo, Kirk Rockett and Dominic Kwiatkowski for the samples from The Gambia; Baboucar Bah and Oumou Bah-Sow for the samples from Conakry; Per Gustafson and Peter Aaby for the samples from Bissau; Sian Floyd, David Warndorff, Lifted Sichali, Simon Malema, Amelia Crampin, Bagrey Ngwira and Paul Fine for the samples from Malawi; Thorsten Thye, Ellis Owusu-Dabo, Ivy Osei, John Gyapong, Anthony Enimil, Margaret Chinbuah, Christian Meyer and Rolf Horstmann for the samples from Ghana; Sailesh Gochlait, Dheeraj Malhotra, Shafat Ali, Shweta Aggarwal, Belum Reddy, Vijay Garg, Rupali Chopra and Rameshwar Bamezei for the samples from New Delhi, India; Sunil Hazra, Bibhuti Saha and Suchismita Roy for the samples from Kolkata, India; and Ramasamy Pitchappan for the samples from Madurai, India.

The core genomics team in the Wellcome Trust Centre for Human Genetics helped in the genotyping and sequencing of the samples in the studies. Credit must be given to David Buck, Ioannis Ragoussis, Lyn-Louise Johnson, Lorna Gregory, Christine Blancher, Lorne Lornie, Chris Allan, Piazza Paolo for their excellent work.

The Wellcome Trust Case Control Consortium tuberculosis study was conducted as

part of the main study. I would like to thank Yik-Ying Teo and Chris Spencer for their genotyping calling, Jeff Barrett and John Broxholme for formatting the genotype data, Andrew Morris, Emily Lyons and Fredrik Vannberg for performing the ethnicity correction and analysis for the genome-wide data, Damjan Vukcevic for the programming script for power calculation, and Sophie Roetynk for designing the replication experiments. I would also like to express gratitude to Thorsten Thye for preparing the genotype file for the combined data, as well as genotyping the Ghanaian samples. Moreover, thanks to Sailesh Gochhait, Dheeraj Malhotra, Cecilia Kim, Brendan Keating and Hakon Hakonarson who contributed to genotyping the New Delhi samples.

Associated Publications

Wong S.H., Hill A.V., Vannberg F.O., India-Africa-United Kingdom Leprosy Genetics Consortium. Comment on"Genomewide association study of leprosy". *N Engl J Med* 2010;362(15):1446-7.

Wong S.H., Vannberg F.O., Spencer A.J., van der Weyden L., Hill A.V., et al. Comment on "CRTAM confers late-stage activation of $CD8^+$ T-cells to regulate retention within lymph node". *J Immunol* 2010;184(8):4052-3.

Wong S.H., Gochhait S., Malhotra D., Pettersson F.P., Teo Y.Y., et al. Leprosy and adaption of human Toll-like receptor 1. *PLoS Pathog, accepted.*

Khor C.C., Vannberg F.O., Chapman S.J., Guo H., **Wong S.H.**, et al. CISH and susceptibility to infectious diseases. *N Engl J Med* 2010;362(22):2092-3101.

The African tuberculosis Consortium and the Wellcome Trust Case Control Consortium. Combined analysis of tuberculosis genome-wide association studies reveals a novel susceptibility locus on chromosome 18q11.2. *Nat Genet, submitted.*

The Wellcome Trust Case Control Consortium and the African tuberculosis Genetics Consortium. Genome-wide association study of tuberculosis susceptibility in West Africa. *Manuscript in preparation.*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Common mycobacterial diseases

Human mycobacterial diseases consist of a wide range of diseases caused by the genus of bacteria *Mycobacterium*. Tuberculosis and leprosy, caused by *Mycobacterium tuberculosis* (Koch, 1882) and *Mycobacterium leprae* (Hansen, 1874) respectively contribute to a major global health burden. The tubercle bacilli infect more than 2 billion people, equivalent to one-third of the world's population (Dye et al., 1999), about 10% of which will develop the active disease in their lifetime (Selwyn et al., 1989). A survey by the World Health Organization (WHO) showed that tuberculosis had an incidence of 9.4 million cases, and caused 1.8 million deaths in 2008 in which many are in developing countries (The World Health Organization, 2009). Tuberculosis is also the leading killer in individuals infected with the human immunodeficiency virus (HIV), who are 20-40 times more likely to develop active tuberculosis than their counterparts (Selwyn et al., 1989). As for leprosy, although it does not kill as many people as tuberculosis, it is endemic in many parts of the world, and remains a major global health concern with an incidence of more than 200,000 in 2007 (The World

Health Organization, 2007).

Atypical mycobacterial diseases consist of a wide variety of infections affecting the skin, lungs, joints and bones. Although these infections are well documented especially with *M. ulcerans*, *M. avium-intracellulare*, *M. marinum* and *M. kansaiii*, they are uncommon in immunocompetent individuals and therefore do not contribute to a significant health burden in the population at large. For this reason they are beyond the scope of this thesis. It is important to remember that, because of the similar characteristics among certain mycobacteria, research on atypical mycobacteria may provide insight into the pathophysiology of tuberculosis and leprosy.

1.1.1 Tuberculosis

Tuberculosis is one of the oldest and deadliest diseases of mankind. The earliest known human remains showing signs of *M. tuberculosis* are 9,000 years old (Hershkovitz et al., 2008). Studies of human skeletons have shown features characteristic of Pott's disease, showing collapse and anterior fusion of adjacent mid-thoracic vertebra as a sequela of disseminated tuberculosis (Sager et al., 1972). The tubercle bacilli have been identified in human skeletons from Germany and Denmark dating back to 5,000 B.C. (Sager et al., 1972; Zimmerman, 1979). Tuberculosis has allegedly killed more people than any other infectious disease (Bloom and Murray, 1992). Although its exact prevalence before the modern times cannot be ascertained, the description of tuberculosis (previously known as *phthisis* or consumption) as the *white plague* indicates that the disease may have had a high incidence and mortality in medieval Europe. This led to the widespread establishment of sanitaria, where patients were admitted for an indefinite period under strict discipline. A vast number of therapies were tried but with little success, including iron salts supplement, oxygen and sunlight therapies and the later surgical collapse therapies. The late nineteenth century

saw the discovery of *M. tuberculosis* by Robert Koch. The Bacille Calmette-Guérin (BCG) vaccine was first used in humans in 1921 (Figure 1.1), and was developed by subculturing *M. bovis* to give an attenuated strain. Despite this being very effective against tuberculous meningitis and military tuberculosis, its efficacy against pulmonary tuberculosis appears variable (Pönnighaus et al., 1992; Rodrigues et al., 1993; Colditz et al., 1994). It was not until the discovery of streptomycin in 1944 (Jones et al., 1944) that it was possible to envisage a cure for pulmonary tuberculosis (Figure 1.1). Combinational chemotherapy remains the mainstay of treatment (Crofton, 1959).

Tuberculosis commonly infects the lung, causing pulmonary tuberculosis, but it can also disseminate to other organs including lymph nodes, bone and joints, intestines and meninges. The primary transmission route is air-borne. Upon infection, small mycobacterial bacilli (1-2μm) are inhaled and a local inflammatory reaction develops in the lung. A Ghon focus is established when the bacilli reach the mediastinal lymph nodes, recruiting further inflammatory cells and enlarging as a primary complex. The patient may be asymptomatic in the primary infection stage. Over a course of 3-8 weeks, the adaptive immune system begins to recognise the pathogen, causing a positive tuberculin reactivity which can be demonstrated using the Mantoux test. After this primary infection, the tubercle bacilli remain dormant in the lung parenchyma. The infection is reactivated when immunity is weak, presenting as pulmonary tuberculosis or less commonly disseminated tuberculosis. The risk of reactivation is highest in the first year after infection (1.5-2%) (Ferebee et al., 1963) and is about 5-10% in the first five years. The remaining life long risk is about 0.1% per annum, and some individuals remain asymptomatic for life. The current recommended treatment for pulmonary tuberculosis involves a Directly Observed Therapy, Short-course (DOTS) with isoniazid, pyrazinamide, rifampicin, and ethambutol or streptomycin for several months.

1.1.2 Leprosy

Like tuberculosis, the history of leprosy is long and intertwined with the history of human civilization. The pathological changes caused by leprosy have been observed in a human skeleton from 2,000 B.C. in India (Robbins et al., 2009), and an early description of the disease has been found on Egyptian papyrus dating back to 1,550 B.C. (Hulse, 1972). Clinical descriptions consistent with leprosy have been described in ancient Sanskrit, Chinese, Greek and Latin languages (Haak et al., 2005; Robbins et al., 2009). The arrival of leprosy in Europe was evident in the fourth century B.C. (Monot et al., 2005), and appeared to be a serious public health problem in Europe during the Middle Ages (Robbins et al., 2009). Numerous leprosaria, or leprosy hospitals, were built throughout Europe to treat the patients as well as to segregate them from the public.

Throughout history, leprosy has been associated with intense social stigma. The term leprosy originates from the Latin word *lepros*, meaning defilement. The disease was conceived of as a divine punishment for sin, and the patients were regarded as unclean or cursed by God. This is illustrated by the sentence "*he is a leprous man, he is unclean*" in the Book of Leviticus. Leprosy patients were ostracised from society and condemned to living in seclusion in Medieval Europe. Once identified as a leper, a Mass of Separation was performed by the local priest which would declare the sufferer legally dead and forbid him or her from sharing church, home or market with the public. Considerable social stigma was also common in ancient society in India, China and Japan (Jopling, 1991), and it remains commonplace in some parts of the world today (Senior, 2009).

Leprosy, also known as Hansen's disease, is a chronic infection affecting mainly the skin and peripheral nerves. It commonly manifests with skin nodules and depigmentation, nerve thickening, loss of sensation and muscle weakness, and can result

in complications like deformities and blindness. The clinical disease can be classified into the tuberculoid and lepromatous types. The tuberculoid form is a subtype where the infection is relatively well-contained and manifests with a limited number of skin nodules. The lepromatous form represents the other end of the clinical spectrum with widespread skin lesions and profound neurological deficits. Distinct types of histoid leprosy and diffuse lepromatous leprosy were also described as variants of the lepromatous type. Most patients, however, present with aspects of tuberculoid and lepromatous leprosy. The Ridley-Jopling classification system classifies the disease into several types (Ridley and Jopling, 1962, 1966): tuberculoid, borderline-tuberculoid, borderline-borderline, boderline-lepromatous, lepromatous and indeterminate, according to clinical and immunopathological characteristics. The WHO simplified the classification into two types: paucibacillary leprosy with a negative skin smear and multibacillary leprosy with a skin smear positive for acid-fast bacilli at any lesion site. Multidrug therapy (MDT) of dapsone with rifampicin alone, or in conjunction with clofazimine, is used as a standard regimen for the treatment of leprosy.

1.2 The *Mycobacteriaceae* family

Understanding the pathogens is the first step in studying the diseases they cause. In this section, I briefly describe the *Mycobacteriaceae* family with emphasis on the microbiology, bacterial genomes and genetic diversity of *M. tuberculosis* and *M. leprae*. These characteristics provide insight into their pathogenicity, virulence and interaction with the human host.

Mycobacterium is a genus of rod-shaped, aerobic, Gram positive bacteria with its own family, the *Mycobacteriaceae*. More than 100 species have been identified, and many of them are ubiquitous in water, soil and food sources. Some mycobacteria, including

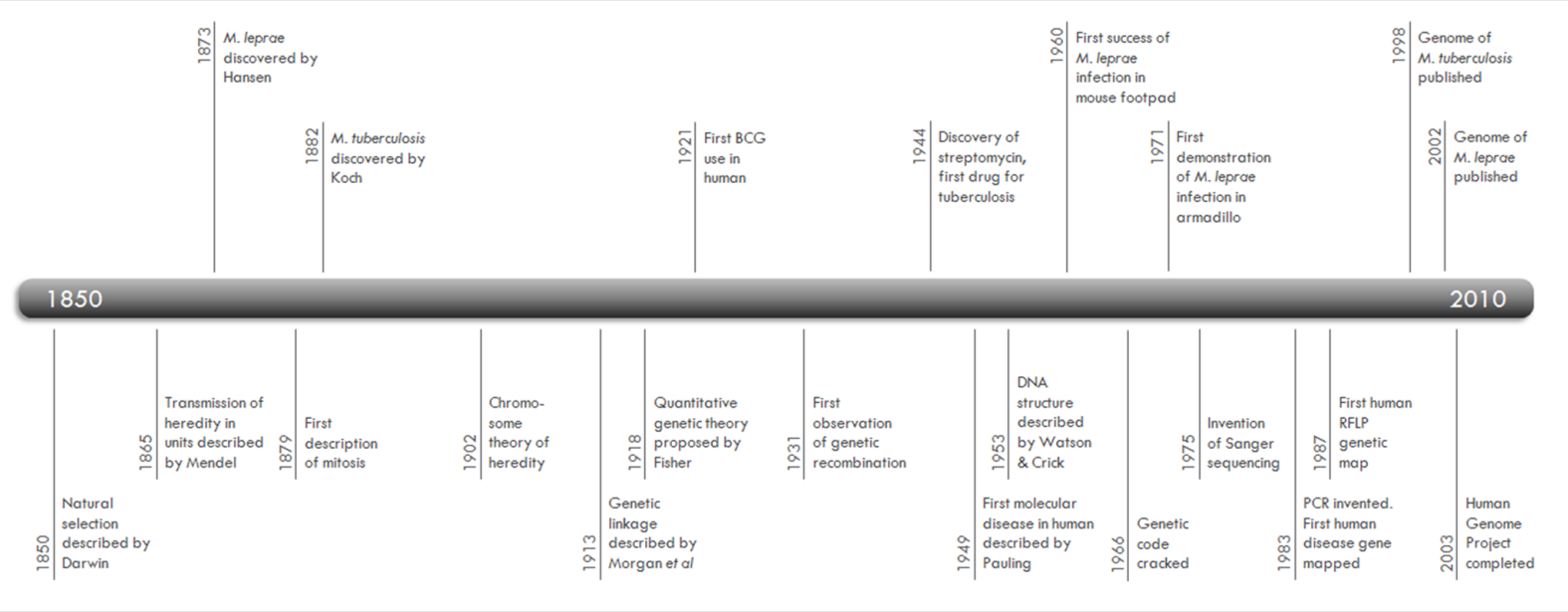


Figure 1.1: A timeline of major milestones in mycobacteria and genetic research in modern human history

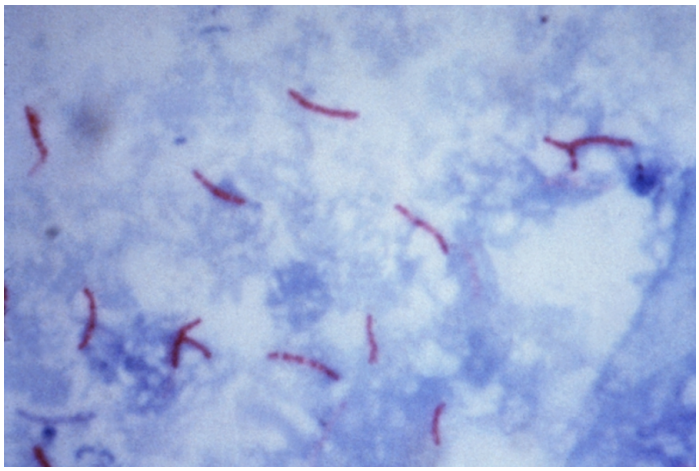


Figure 1.2: Staining for *M. tuberculosis* using the Ziehl-Neelsen stain

M. tuberculosis and *M. leprae* however, are obligate intracellular parasites and are not found free-living outside living cells. This genus is unique in that all *Mycobacterium* species share a characteristic cell wall thicker than many other bacteria. The cell wall consists of an inner layer of peptidoglycan covalently bound to arabinogalactan, which is in turn linked to the outer layer of glycolipids and mycolic acids. The mycolic acids are strong hydrophobic molecules that form a waxy layer, reducing permeability to chemicals. This renders the bacterium resistant to desiccation as well as lysozymal digestion, hydrogen ions and oxygen radicals in phagocytic granules (Russell, 2001). The cell wall is also impermeable to usual stains including the Gram's stain. For this reason mycobacteria are visualised by the special Ziehl-Neelsen stain (Figure 1.2), which consists of carbolfuchsin that binds to mycolic acid to give a contrast to the background.

1.2.1 *Mycobacterium tuberculosis*

Microbiology of *M. tuberculosis*

The discovery of *M. tuberculosis* by Robert Koch in 1882 could be considered as the beginning of the microbiology of mycobacteria (Figure 1.1). In his experiments,

Robert Koch was able to fulfil the postulates of Henle (1840) to identify *M. tuberculosis* as the causative agent of tuberculosis (Koch, 1882). The postulates was later ascribed to him as Koch's postulates because of the successful application in infectious disease research.

M. tuberculosis is a slow growing species with the characteristic thick cell wall of the *Mycobacterium* genus. Apart from the cell wall components of peptidoglycan, arabinoglactan, glycolipids and mycolic acids, *M. tuberculosis* also processes immunogenic proteins like lipoarabinomannan, lipomannan and phosphatidyl-myoinositol mannosides (Quesniaux et al., 2004). *M. tuberculosis* is also known to secrete a variety of different proteins, including the antigen 85 complex (85A, 85B, 85C), early secretory antigenic target 6 (ESAT-6) and culture filtrate protein 10 (CFP-10) (Lewis et al., 2003). Many of these proteins are immunogenic, and are likely to be important as part of its virulence mechanisms and in resistance to phagocytic killing.

M. tuberculosis can be stained with the Ziehl-Neelsen stain to look for pink staining against a blue background. The definitive diagnosis, however, is the isolation of *M. tuberculosis* in culture. The most commonly used medium is the Lowenstein-Jensen medium. The presence of *M. tuberculosis* is indicated by rough and brown-coloured colonies within 8 weeks at 37°C. Apart from microscopy and culture, molecular techniques including nucleic acid amplification have been used to detect and trace *M. tuberculosis* in both clinical and research settings (Piersimoni and Scarparo, 2003; Greco et al., 2006; Laraque et al., 2009). This technique has been widely used especially for tracing the multidrug resistant (MDR) and extremely drug resistant (XDR) strains of *M. tuberculosis*.

The *M. tuberculosis* genome

The complete genome sequence of a virulent strain of *M. tuberculosis*, H37Rv, was published in 1998 (Cole et al., 1998) (Figure 1.1). The bacterial genome has a size of 4.4 Mb and about 4,000 genes, and is rich in guanine and cytosine (65.6%). *M. tuberculosis* differs from other bacteria in that an unusually large portion of its coding capacity is devoted to enzymes involved in lipogenesis and lipolysis. A total of about 250 distinct enzymes involved in fatty acid metabolism were found, compared to only 50 in *Escherichia coli* (Cole et al., 1998). This accounts for the diverse array of lipophilic molecules synthesised by *M. tuberculosis*. The sequencing effort also identified two unrelated families of acidic, glycine-rich proteins carrying the Pro-Glu (PE) and Pro-Pro-Glu (PPE) motifs, which are encoded by 10% of the coding capacity of the genome. These proteins represent a major source of antigenic variation (Choudhary et al., 2003; Chakhaiyar et al., 2004), and may interfere with immune responses by inhibiting antigen processing (Talarico et al., 2005). These variations may lead to the predominant differences in pathogenesis between the *Mycobacterium* species.

Genetic diversity of *M. tuberculosis*

Recent work with comparative whole-genome hybridization as well as deoxyribonucleic acid (DNA) sequencing has revealed the population genetics of *M. tuberculosis*. This species is highly geographically structured and can be divided into six main phylogeographical lineages (Gagneux et al., 2006) despite a low level of diversity compared to other pathogens (Sreevatsan et al., 1997; Comas et al., 2010). Importantly, these lineages include two West African lineages that are also known as *M. africanum* and do not occur elsewhere. This shows that the *M. tuberculosis* pathogen has likely adapted to specific human populations. This is also of particular interest

Characteristic	Tuberculosis	Leprosy
Pathogen		
Main species	*M. tuberculosis*	*M. leprae*
Doubling time	16-24 hours	14 days
Growth *in vitro*	4-6 weeks in medium	Cannot grow *in vitro*
Experimental hosts	Mouse, rabbit, guinea pig	Armadillo
Pathogen genome		
Genome size	4.41 Mb	3.27 Mb
GC content	65.6%	57.8%
Protein coding proportion	90.8%	49.5%
Encoding genes	3,959	1,604
Pseudogenes	6	1,116
Clinical disease		
Transmission	Respiratory	Respiratory or contact
Primary affected organs	Lungs	Skin, peripheral nerves
Other affected organs	Lymph nodes, meninges	Eyes, testicles
Classification	Pulmonary / Extra-pulmonary	Paucibacillary / Multibacillary

Table 1.1: A brief comparsion between tuberculosis and leprosy

to the studies in this thesis, since some of the case-control samples come from the West Africa. This has important implications in genetic studies where genetic effects may be population specific due to host-pathogen compatibility.

1.2.2 *Mycobacterium leprae*

Microbiology of *M. leprae*

M. leprae was identified by Armauer Hansen in 1873 as the first pathogen to cause a major human disease (Figure 1.1) (Hansen, 1874). Despite this early identification, microbiological research for the bacillus was slow due to the inability to culture or infect animals with *M. leprae*. It was not until 1960, more than 80 years after its identification, that limited infection of *M. leprae* was first demonstrated in mouse footpads. The nine-banded armadillo was found to be the only immunocompetent animal apart from humans that can be reliably infected (Kirchheimer and Storrs, 1971;

Kirchheimer et al., 1972). Today, *M. leprae* remains the only disease-causing bacterium that cannot be cultured in artificial media. Similar to other mycobacteria, *M. leprae* demonstrates strong acid-fast staining with carolfuchsin. Under microscopy, the bacilli are often observed grouping together to form clumps known as globi, and in smaller clumps they characteristically appear like "*bundles of cigars*". *M. leprae* has a uniquely slow rate of multiplication even compared to other the slow-growing mycobacteria, with a generation time of 11-13 days compared to 20 hours for *M. tuberculosis*.

The *M. leprae* genome

The genome sequence of the leprosy bacillus was completed in 2001 (Figure 1.1). The sequence from an armadillo-derived Indian isolate of the bacillus revealed an extreme case of reductive evolution (Cole et al., 2001). At a size of 3.3 Mb, the genome of *M. leprae* is significantly smaller than that of *M. tuberculosis*, and just 49.5% of this is functional and containing coding genes. Another 27% of the genome is recognised as pseudogenes, and the remaining 23.5% of the genome does not appear to be coding and may correspond to regulatory sequences or mutated gene remnants. It is estimated that the bacillus may have lost more than 2,000 genes since diverging from the last common mycobacterial ancestor with *M. tuberculosis*.

This sequencing effort identified *M. leprae* as an organism with the most degraded genome known to date. It also provided explanations for the unique properties of the bacterium. Due to the loss of genes coding for proteins in the metabolic pathway, the ability of *M. leprae* to generate energy is very limited. This genetic reductionism is thought to account for its slow rate of growth. Such reductive evolution has been documented in obligate intracellular parasites such as *Rickettsia* and *Chlamydia* species, due to the inactivation of genes whose functions are no longer required. As

for *M. leprae*, having committed itself to a highly specialised niche in humans and armadillos, the bacterium is relatively free from the evolutionary pressure experienced by other microbes (Young and Robertson, 2001). As mutations are rarely reversed in organisms that asexually reproduce without exchange of genetic material, deleterious mutations accumulate and result in a highly degraded genome (a phenomenon known as Muller's ratchet (Felsenstein, 1974)). This has likely contributed to the degradative genome of *M. leprae*.

Genetic diversity of *M. leprae*

Recent phylogeographic analysis has suggested that the population of *M. leprae* arises from a single clone, due to its asexual reproduction by binary fission (Monot et al., 2005, 2009). A total of 16 subtypes belonging to four main lineages have been identified, which show strong geographical association that reflects early human migration patterns. This study suggests that the progenitor strain may have originated in East Africa and spread to Europe and Asia. Notably, *M. leprae* shows a lack of diversity with only 215 polymorphic sites throughout its 3.3 Mb genome. This may have important implications in host genetic studies, and points to the postulation that the phenotypic variation may be more attributable to host factors.

1.2.3 Other important mycobacterial species

The *Mycobacterium tuberculosis* complex

Apart from *M. tuberculosis*, tuberculosis also be caused by few other closely related mycobacteria: *M. africanum*, *M. bovis*, *M. canetti* and *M. microti*. *M. africanum* is most closely related to *M. tuberculosis* and is often regarded as a lineage of *M. tuberculosis*. It is common in West Africa and causes symptoms resembling those of

infections by *M. tuberculosis*, except that the infection is less likely to progress to clinical disease in immunocompetent individuals. *M. bovis* was once a common cause of tuberculosis but has been largely eliminated by milk pasteurisation. *M. canetti* is rare and *M. microti* is mostly seen in immunodeficient individuals. Together with *M. tuberculosis*, these mycobacteria are collectively known as the *Mycobacterium tuberculosis* complex. Previous work has shown that this complex lacks interstrain genetic diversity with a rarity of nucleotide change (Sreevatsan et al., 1997).

Mycobacterium lepromatosis

Recently, a new *Mycobacterium* species *M. lepromatosis* was identified to cause a distinct type of diffuse lepromatous leprosy, also known as diffuse leprosy of Lucio and Latapi (Han et al., 2008). This type of leprosy is characterised by multiple, well-defined purpuric lesions evolving into massive ulcerations, and is restricted to patients from Mexico and the Caribbean. Phylogenetic analysis revealed significant divergence from *M. leprae* with an overall level of nucleotide diversity of 90.9% (Han et al., 2009), below the 95% level cut-off used to distinguish between species (Goris et al., 2007).

1.3 Human immunity to mycobacteria

The immune system is the direct interface of the interaction between mycobacteria and human. A proper understanding of immmunology is important for mycobacterial research. In the context of genetics, understanding immunity against mycobacteria has two advantages: to guide candidate gene studies and help interpret results of genome-wide studies. In this section I briefly describe human immunity against mycobacteria.

The mammalian immune system uses two separate types of immunity to protect the host against pathogens: innate and adaptive. The differences between the two systems are attributable to their *specificity*. The innate immune system is encoded by the germ-line to provide a non-specific but immediate response, whereas the adaptive immune system is specific and is generated *de novo* through somatic rearrangements during B- and T-cell development (Medzhitov and Janeway Jr., 2000; Medzhitov, 2001). The subsequent B- and T-cell expansion is subject to the recognition of antigens presented on major histocompatibility complex (MHC) on antigen-presenting cells (APCs).

1.3.1 Immunology of tuberculosis

M. tuberculosis is considered one of the most successful pathogens, and has managed to latently infect one third of the human population. Despite the highly sophisticated mammalian immune system, *M. tuberculosis* has the ability to manipulate the immune response to escape killing by host cells. Understanding the intricacies of this interaction would help to identify genes affecting tuberculosis susceptibility and may lead to vaccine and therapeutic development for disease control.

Innate immunity

Recognition of *M. tuberculosis* Immunity against *M. tuberculosis* starts with the recognition of pathogen associated molecular patterns (PAMPs) on the mycobacterial cell wall by pattern recognition receptors on the host cells. Studies have found that the human macrophage uses the mannose receptor and complement receptor to recognise lipoarabinomannans on *M. tuberculosis* and phagocytose the bacterium into the cells (Schlesinger et al., 1990; Schlesinger, 1993), whereas the dendritic cell uses the receptor dendritic-cell specific intercellular molecule-3-grabbing non-integrin

(DC-SIGN) to achieve the same result (Tailleux et al., 2003). These interactions have been shown to inhibit the macrophage response to *M. tuberculosis* (Nigou et al., 2001; Kang et al., 2005; Doz et al., 2007), suggesting that the bacterium may have utilised these interactions to subvert the host defences as part of their virulence mechanisms. Consistent with this, genetic associations between variants in *DC-SIGN* and susceptibility to tuberculosis have been reported (Barreiro et al., 2006; Vannberg et al., 2008).

Toll-like receptors (TLRs) are important receptors in the detection of pathogens and activation of the immune response (Medzhitov and Janeway Jr., 2000; Medzhitov, 2001). Once stimulated by respective ligands, the receptor binds to an adaptor molecule, myeloid differentiation factor 88 (MyD88) or MyD88-adaptor-like (MAL) proteins, to activate a downstream signal cascade that stimulates pro-inflammatory nuclear factor kappa B (NF-κB) mediated transcription (Horng et al., 2002). *In vivo* studies have suggested that MyD88 mediates the immune response to *M. tuberculosis*, with a greater than hundred-fold increased number of bacteria in the lung among knockout animals (Fremond et al., 2004; Scanga et al., 2004). Several *in vitro* studies also showed a central role of TLR2 in recognising *M. tuberculosis* (Brightbill et al., 1999; Means et al., 1999; Underhill et al., 1999) by cooperating with TLR1 and TLR6 (Bulut et al., 2001; Takeuchi et al., 2002). This is supported by several studies suggesting an association between variants in *TLR2* and susceptibility to tuberculosis (Ben-Ali et al., 2004; Ogus et al., 2004; Thuong et al., 2007). These show that the TLR pathway is important in the immunity to *M. tuberculosis*.

Intracellular events The innate immune system has developed several mechanisms active in intracellular defence against *M. tuberculosis*. The human macrophage is able to generate reactive oxygen intermediates (ROIs) and reactive nitrogen intermediates (RNIs), which are unstable radicals that can kill mycobacteria by damag-

ing their DNA. The importance of this pathway is demonstrated by children with chronic granulomatous disease (CGD), who are hyper-susceptible to tuberculosis due to a defective oxidative burst mechanism (Bustamante et al., 2007; Lee et al., 2008). Phagosomes containing bacteria are normally diverted for lysomsomal fusion upon phagocytosis, killing the bacteria by acidification and lysosomal digestion. However, being an obligate intracellular pathogen, *M. tuberculosis* has developed several mechanisms to ensure its survival inside host cells. This includes its ability to exclude radical generating enzymes (Davis et al., 2007) and to synthesise proteins to protect its DNA (Martinez and Kolter, 1997; Colangeli et al., 2009). Certain virulent strains of *M. tuberculosis* have been shown to arrest phagosomal maturation (Russell, 2001; Vergne et al., 2005) and with acidification inside the phagosomes (Vandal et al., 2008).

Nevertheless, autophagy represents another major host defence mechanism to kill *M. tuberculosis*. It is an evolutionarily conserved process in which cell components are degraded by the lysosomal machinery (Deretic et al., 2009; Deretic, 2010), and has been recently shown to be an effective defence mechanism against intracellular pathogens including mycobacteria (Gutierrez et al., 2004; Levine and Deretic, 2007; Biswas et al., 2008). Autophagy can be induced by the immunity-related guanine triphosphatase family M (IRGM) (Singh et al., 2006) and TLRs (Sanjuan et al., 2007; Delgado et al., 2008), and is regulated by cytokines as an integral part of the immune response (Gutierrez et al., 2004; Singh et al., 2006; Djavaheri-Mergny et al., 2006; Harris et al., 2007).

Adaptive immunity

Effector cells and cytokines The adaptive immune system is closely intertwined with the innate immune system and is characterised by a specific response against

pathogens by the B-cells or T-cells. In humans, the activation of adaptive immunity against *M. tuberculosis* requires antigen presentation by MHC molecules on antigen presenting cells (APCs) to T-cells, which are essential in the protective immunity against tuberculosis. In particular, it is clear that $CD4^+$ T-cells are central to immunity against *M. tuberculosis* (Orme et al., 1992; Havlir and Barnes, 1999; Chackerian et al., 2001; Khader et al., 2007). Several studies suggested that $CD8^+$ T-cells may also play a role (Flynn et al., 1992; Stenger et al., 1997; Liebana et al., 1999; Cho et al., 2000). As for the cytokines, interferon-γ (IFN-γ) is an important cytokine which can activate macrophages to kill mycobacteria, through regulating autophagy (Gutierrez et al., 2004; Singh et al., 2006) and overcoming phagolysosomal block (MacMicking et al., 2003). The release of IFN-γ is enhanced by interleukin-12 (IL-12) which differentiates $CD4^+$ T-cells into T_h1-cells. Consistent with this, mutations in genes encoding IFN-γ receptors (*IFNGR1* and *IFNGR2*) and an IL-12 subunit (*IL12B*) have been identified to cause unusual Mendelian susceptibility to environmental mycobacteria (Al-Muhsen and Casanova, 2008), highlighting the importance of type 1 cytokines in immunity against *M. tuberculosis*. The cytokine tumour necrosis factor α (TNF-α) is important for the formation of granulomas (Kindler et al., 1989; Roach et al., 2002; Algood et al., 2005). Individuals receiving anti-TNF therapy for chronic inflammatory disease are at increased risk of developing tuberculosis (Gardam et al., 2003).

1.3.2 Immunology of leprosy

Research in leprosy immunology has been hampered by the inability to culture the bacterium *in vitro* and to infect laboratory animals with the exception of armadillos. Nevertheless, certain overlap is observed between the immunological response to *M. leprae* and *M. tuberculosis*. They are both intracellular pathogens recognised by the

innate immune system, with a central importance of an adaptive T-cell response in both diseases.

Innate immunity

Recognition of *M. leprae* Like *M. tuberculosis*, the human TLRs are crucial for the recognition of *M. leprae* by macrophages and dendritic cells during innate immunity. *In vitro* studies have shown that lipoproteins from *M. leprae* could activate monocytes and monocyte-derived dendritic cells through TLR1 and TLR2 (Krutzik et al., 2003), initiating a downstream signalling cascade for transcription of pro-inflammatory cytokines (Tapping and Tobias, 2003; Quesniaux et al., 2004). Genetic variants in *TLR2* have been found to associate with susceptibility to lepromatous leprosy (Kang and Chae, 2001; Kang et al., 2005). Other receptors likely to be involved in recognizing *M. leprae* include the mannose receptor (Astarie-Dequeker et al., 1999), complement receptors 1 and 3 (Schlesinger and Horwitz, 1991) and DC-SIGN (Maeda et al., 2003; Koppel et al., 2004).

Adaptive immunity

Effector cells and cytokines The essential role of T-cells immunity against *M. leprae* is supported by evidence of accelerated bacterial growth in neonatally thymectomised (Rees et al., 1967) and congenitally athymic (Colston and Hilson, 1976) mice. Furthermore, the wide range of clinical spectrum observed in leprosy is dictated by the T-cell immune response. Tuberculoid leprosy has a predominant cell-mediated immunity component with a Th1 immune response, whereas lepromatous leprosy has a dominant humoral Th2 immune response (Modlin, 1994). Such a difference may be determined by the cytokine milieu (Spellberg and Edwards Jr., 2001). A preponderance of $CD4^+$ T-cells with IFN-γ, IL-2 and lymphotoxin-α are observed in

lesions from tuberculoid leprosy patients (Modlin et al., 1983b). Under the influence of cytokines, particularly TNF-α, macrophages fuse to form the Langhans giant cells in granuloma (Algood et al., 2005). In contrast, the $CD8^+$ T-cells seem to predominate in lepromatous leprosy with poor granuloma formation (Scollard et al., 2006). The cytokines which predominate are IL-4, IL-5 and IL-10, showing a Th2 immmune response profile.

Immunology of leprosy reactions

Type 1 reaction The type 1 reactions, also known as reversal reactions (RR), are caused by enhancement of cellular immunity and delayed hypersensitivity to *M. leprae* antigens. They usually affects about 30% individuals with borderline to tuberculoid leprosy, and present with acute inflammation of the skin lesions or nerves. They can arise spontaneously or follow start of therapy. Functional studies showed increased number of $CD4^+$ T-cells (Modlin et al., 1983a, 1988) with elevation in a wide spectrum of cytokines including IL-2, IL-6, IL-12 and IFN-γ in reacting skin lesions (Yamamura et al., 1992; Moraes et al., 1999).

Type 2 reaction The type 2 reactions are also known as the erythema nodosum leprosum (ENL) reaction. They occur in about 10% of borderline lepromatous and about 50% of lepromatous leprosy patients, and presents with systemic inflammation affecting multiple organs. They may result in profound neurological deficit, skin ulceration, fever, lymphadenitis, scleritis, uveitis and polyarthritis. Elevated levels of cytokines, particularly TNF-α, are observed (Sarno et al., 1991; Barnes et al., 1992). There is evidence that the type 2 reaction results from an inability to develop cell-mediated immunity upon IFN-γ stimulation, the effect of which is channelled into pathways that lead to enhanced humoral immune activity (Sampaio et al., 1992).

1.4 The science of heredity

Genetics is the study of heredity. Genetics defines the relationship between genotype and phenotype in a biological system, and seeks to understand the basis of how information in DNA is transcribed to ribonucleic acid (RNA) and then translated to protein (molecular genetics). Genetics in a border sense, however, embraces many different branches of studies including population genetics, evolutionary genetics, developmental genetics and others. Here I describe the general background of the field and several key concepts pertinent in disease mapping.

1.4.1 General background of genetics

The field of genetics began in 1865 when Gregor Mendel described his experiments with peas showing the inheritance of phenotypic traits. Working in the garden of his monastery, he observed that the colours of the second generation peas consistently showed a 3:1 ratio when he crossed pure breeding lines with each other. His observation led him to come up with the theory that there were two alternative forms of a factor controlling the pea colour, with one form being dominant and the other being recessive. These alternative forms were later known as alleles. During the formation of gametes, alleles are separated such that each gamete receives one copy of the allele (law of segregation). The diploid state in somatic cells is restored when two gametes, one from each parent, combine during reproduction. As one allele is dominant and masks its recessive counterpart in the heterozygous state, the offspring will show a phenotypic ratio of 3:1 when parents of the two homozygous states are mixed. This important work showed that hereditary information is transmitted in discrete units, and illustrated the concept of allele dominance affecting a phenotypic trait.

In 1949, Linus Pauling described the first molecular disease in humans, when he

and his colleagues attributed sickle cell anaemia to a change in the structure of haemoglobin (Pauling et al., 1949). Using electrophoresis, they demonstrated that individuals with sickle cell disease had a modified form of haemoglobin (HbS) in their red blood cells. Individuals with the sickle cell trait showed the heterozygous state with both the normal (HbA) and modified (HbS) forms. The cause of this structural change was later identified to be due to a mutation resulting in the substitution of glutamic acid with valine in the sixth amino acid position in the beta globin chain (Ingram, 1956; Saiki et al., 1985). This was an important finding because it was the first time that the molecular mechanism of a disease was understood. Moreover, this mutation was also shown to be under natural selection, with an increased frequency in regions where malaria was endemic due to the increased resistance to the parasite of the heterozygous individuals (Allison, 1954).

The double helix structure of DNA was elucidated by James Watson and Francis Crick in 1953 (Watson and Crick, 1953). Prior to their discovery, DNA was known to be made of nucleotides with a phosphate residue, a deoxyribose sugar and one of the four nitrogenous bases of adenine (A), cytosine (C), guanine and thymine (T). It was also known that the DNA is highly symmetrical with an equal amount of adenine and thymine, and cytosine and guanine (Wilkins et al., 1953). The model described by Watson and Crick correctly described the double helix structure of DNA, with sugar and phosphate forming two anti-parallel strands and the bases pointing into the helix. It also described the hydrogen bonds linking the bases specifically, with an adenine always pairing with a thymine, and a cytosine pairing with a guanine. This discovery of the structure of DNA was essential to understanding how DNA is duplicated, and how information is transcribed from DNA to RNA and translated to protein which is commonly described as the "*central dogma of molecular biology*" (Crick, 1958).

The chain-termination method of DNA sequencing was first described by Frederick Sanger in 1975 (Sanger and Coulson, 1975). By eliminating the 3-hydroxyl group from nucleotides, he produced dideoxynucleotides incapable of forming phosphodiester bonds for extension. These dideoxynucleotides (ddNTPs) are randomly incorporated into the DNA strands, resulting in fragments of varying length separable by gel electrophoresis. The DNA bands can then be visualised by autoradiography to determine the sequence. Another major breakthrough in molecular technology was the invention and the development of polymerase chain reaction (PCR) by Kary Mullis (Mullis and Faloona, 1987). The PCR allows *in vitro* amplification of DNA by a cyclic three-step process of denaturation, annealing and extension, producing billions of copies of a specific DNA segment in a few hours (Saiki et al., 1988).

The Human Genome Project was launched in 1990 (Watson, 1990). The aim of the project was to map and sequence the human genome, which was estimated to contain about three billion nucleotides. The project was completed in 2003, two years ahead of schedule (Collins et al., 2003). The data provided a reference sequence for the human genome, and covered 99% of the human genome with an accuracy of 99.99%. At the time near the completion of the Human Genome Project, the International HapMap Project (HapMap) was commenced with a goal to catalogue common genetic variants, using a reference panel of 270 individuals from the USA, China, Japan and Nigeria (The International HapMap Consortium, 2003, 2005). Over three million variations were documented at the completion of phase II of the project (Frazer et al., 2007). This effort provided the foundation for the subsequent genome-wide association (GWA) studies which have unravelled many common variants affecting common disease traits.

1.4.2 Important concepts in disease mapping

One of the main goals of geneticists is to find genes that cause diseases and to understand their underlying molecular mechanisms. This process involves quantifying a genetic effect and determining the nature of the disease phenotype. This information is then taken forward to the design of the experiment. In this section I discuss several important concepts relevant to disease mapping.

Measuring genetic effects

The first step in mapping a disease is to determine whether the disease is genetic in nature, that is, how much do genetic factors contribute to the disease. Heritability (H^2 or h^2) refers to the proportion of phenotypic variation in a trait that is attributable to genetic variation (Visscher et al., 2008). The heritability of a disease is classically determined by twin studies. If a disease is heritable, the disease concordance will be higher in monozygotic twins which have an identical genetic make-up, than dizygotic twins who share only half of their genomes. Heritability can therefore be determined by comparing disease concordance between the two types of twins, which allows dissection of the genetic effects from the environmental effects. The sibling relative risk (λ_S), which is the ratio of an individual's risk given an affected sibling compared to the general population, measures the combined effects of the genes and environment shared by the siblings.

Defining the nature of phenotypes

The original work described by Mendel described inheritance of a dichotomous trait determined by one genetic locus. However, organisms have many quantitative traits such as height and weight that do not follow the simple Mendelian pattern of inher-

itance. This became a source of debate between geneticists and biometricians, until 1918 when Fisher applied his statistical model to reconcile these concepts (Fisher, 1918). He proposed that a quantitative trait can be explained in Mendelian terms as if it is determined by multiple genes. The theory predicts that a quantitative trait is contributed by the additive effect of multiple loci each having a small effect. Such a trait will approximate a normal Gaussian distribution, and the approximation improves with more loci involved. Variability of the trait can be described by the mean and variance of the distribution. This polygenic theory was later applied to dichotomous traits like coronary artery disease and diabetes, that occur when the additive sum of genetic effects surpasses a susceptibility threshold (Falconer, 1960). These disease traits caused by multiple factors are called complex diseases and likely include tuberculosis and leprosy. This will be discussed in more detail in Section 1.5.3.

Genetic variants and phenotypes

Humans are unique and are shaped by the environment as well as genetic variants. These variants account for substantial phenotypic variability between humans, including distinctive morphological characteristics (such as skin pigmentation and eye colour) (Relethford, 2002; Liu et al., 2009) as well as susceptibility to diseases. Genetic variations in humans occurs on many different scales, and can take the form of single nucleotide polymorphisms (SNPs), copy number variations (CNV) or epigenetic changes. Single nucleotide polymorphisms (SNPs) are sequence variations in the genome and account for a substantial proportion of genetic variability between humans (Collins et al., 1998; Brookes, 1999). These variants are the main source of variation studied in this thesis.

The last few decades saw a boom in our understanding of genetic variants in suscep-

Characteristic	Complex disease	Mendelian disease
Disease		
Relative incidence	Common	Rare
Disease onset	Often late	Variable
Public health impact	Large	Small
Example	Hypertension	Cystic fibrosis
Genetic variant		
Genetic model	Polygenic	Monogenic
Allele frequency	Common	Rare
Effect size	Small	Large

Table 1.2: A generalised comparison of complex and Mendelian diseases

tibility to Mendelian diseases, with over 2,000 traits mapped to the human genome to date. These diseases are relatively rare in the population and are typically caused by one or few rare mutations of large effect sizes (Bodmer and Bonilla, 2008; Hindorff et al., 2009). This is due to the fact that deleterious mutations are often lost to purifying selection during evolution, and are therefore rare in nature (Altshuler et al., 2008). Due to this force of nature, common diseases, on the other hand, are usually milder in severity with an onset beyond the reproductive age. This has led to the *common disease-common variant* hypothesis, which proposes that common variants would confer susceptibility to common diseases (Lander, 1996; Risch and Merikangas, 1996). This hypothesis is supported by candidate gene and subsequent GWA studies identifying putative causal variants to common diseases.

It is noteworthy that despite identification of numerous common variants that contribute to common diseases, these variants only contribute a part of the total genetic effects. Indeed, work from recent GWA studies suggests that less than 20% of total genetic effects are caused by common variants (McCarthy and Hirschhorn, 2008; Manolio et al., 2009). The remaining effects could be due to rarer variants of greater effect sizes, structural variants, epigenetic imprinting and other mechanisms.

Mapping by linkage and association

Once the phenotype is defined, genetic mapping can be performed in either of the two approaches of *linkage* and *association*, each having its own merits and weaknesses. The former approach is built upon the concept of *linkage*, which describes the tendency of two or more genetic loci to be inherited together. This is because genes on the same chromosome are aligned in a linear fashion (Morgan et al., 1920) and do not assort independently of each other during meiosis. Linkage can be disrupted by homologous recombination between the sister chromatids. Disease mapping by linkage analysis involves genotyping markers that co-segregate with the disease status. A marker is linked with a disease trait if it is inherited by diseased individuals more often than expected by chance. The result can be reflected by the logarithm of odds (LOD) score which describes the likelihood of linkage.

This linkage approach has been successful in identifying genetic variants of more than 2,000 Mendelian traits. This has led scientists to envision genetic mapping to common diseases which also show familial clustering (Altshuler et al., 2008). However, although Mendelian subtypes of common diseases like breast cancer (Welcsh and King, 2001), hypertension (Lifton, 1996; Ji et al., 2008) and diabetes (Bell and Polonsky, 2001) have been elucidated, these genes explain a tiny proportion of the cases in the population. The use of linkage studies is also limited by their insufficient power to detect genetic variants of modest effect size (Dawn Teare and Barrett, 2005; Laird and Lange, 2006), which affect susceptibility to common disease traits according to the *common disease-common variant hypothesis* (Lander, 1996; Risch and Merikangas, 1996; Lohmueller et al., 2003). Therefore, the association approach was advocated as a more powerful way to detect variants contributing to common disease traits (Risch and Merikangas, 1996). This approach involves the collection of unrelated samples in a population, followed by marker genotyping and comparison

of the frequencies between cases and controls. It is built upon the concept of *linkage disequilibrium*, which refers to the strong correlation of nearby genetic markers in a population. The association signal of the disease variant can then be captured by nearby tagging variants. This approach elicits a powerful way to survey the genome by a subset of tagging markers in cases and controls. However, it has the disadvantage of being prone to spurious results, if the cases and controls are not properly matched with respect to ethnicity or other factors that influence the genetic composition (Devlin and Roeder, 1999).

The candidate gene and genome-wide approaches

Apart from linkage and association, genetic studies can also be categorised into the candidate gene or the genome-wide approach. The fundamental difference between the approaches lies with the *a priori* hypothesis of the experiment. The first step in conducting candidate gene studies is the choice of a suitable candidate gene that may plausibly play a relevant role in the disease process. Flanking markers in the gene are then tested for evidence of linkage or association. This requires prior knowledge on the function of the gene before the experiment. On the other hand, genome-wide studies seek to survey the whole genome and therefore do not require *a prior* hypothesis about any candidate gene. This has the advantage of being an unbiased search for genetic variants, and can potentially identify novel genes that contribute to disease susceptibility.

The advent of microarray technology has allowed many markers to be genotyped in one experiment, and has taken association studies from a candidate gene to a genome-wide approach. This was also fuelled by the International HapMap Project which has identified numerous SNPs to allow alleles tagging (The International HapMap Consortium, 2003, 2005). As such, the Wellcome Trust Case Control Consortium

(WTCCC) was established in 2005 to explore the utility, design and analysis of GWA studies as a proof-of-concept experiment, and has identified many novel variants affecting susceptibility to different diseases (The Wellcome Trust Case Control Consortium, 2007). Chapters 3 and 4 in this thesis describe the studies to determine tuberculosis susceptibility as part of the WTCCC experiment.

1.4.3 Genetics and human health

Genetics has helped us to understand the heredity of life and the basis of variations between organisms. The use of genetic fingerprinting has been used in forensic investigations, and recently sequencing technology has been used to expand our knowledge in the field of palaeoarchaeology (Rasmussen et al., 2010). In the medical context, progress in genetics has enabled molecular diagnosis of numerous diseases ever since 1983 when the first gene was mapped in Huntington Disease. Human genetics has expanded our knowledge in more than 2,000 Mendelian traits and provided a method to diagnose diseases and predict risks.

Until recently our knowledge in disease genetics was largely on Mendelian traits. The plethora of GWA studies in recent years has unveiled the genetic basis of many common diseases, and provided insights into novel disease mechanisms (Donnelly, 2008). For example, variants associated with type 2 diabetes were found in genes that regulate pancreatic β-cell functioning, suggesting the importance of insulin secretion rather than resistance in the disease process (Florez, 2008). Findings from GWA studies of Crohn's disease have pointed to autophagy and innate immune mechanisms as being important in disease aetiology (Barrett et al., 2008). As for infectious diseases, a recent GWA study has identified overlapping variants between Crohn's disease and susceptibility to leprosy (Zhang et al., 2009). Genetic variants determining the response to hepatitis C treatment have also been identified (Ge et al., 2009;

Thomas et al., 2009). The finding of the near-complete resistance of *CCR5* $\triangle$32 deletion to HIV-1 in a candidate gene study (Dean et al., 1996; Liu et al., 1996) has provided insight for the development of novel therapeutic agents against the infection (Fatkenheuer et al., 2008; Gulick et al., 2008; Hutter et al., 2009). These important findings will increase our understanding of common diseases and ultimately improve human health.

1.4.4 Genetics in infectious diseases

The above-mentioned concepts and techniques in genetics have been applied to studies of infectious diseases, which carry a major health burden for mankind. In this section I discuss the background and progress in the field of genetics of infectious diseases.

The role of host genetics

The role of host genetics in infectious diseases is supported by the epidemiological observation that individuals vary largely in their response upon exposure to pathogens (Cooke and Hill, 2001). The accidental inoculation of *M. tuberculosis* to 249 infants in 1930 in Lubeck, Germany, left some individuals unaffected while killing many others. Haemophiliac individuals who were erroneously transfused with HIV in the 1980s have also shown significant differences in rates of progression to acquired immunodeficiency syndrome (AIDS). Apart from these specific incidents, the variability in disease manifestation is also observed in routine clinical settings. Only 10% of individuals infected with *M. tuberculosis* develop clinical disease (Bloom and Small 1998) with a highly variable presentation. Some patients rapidly progress to disseminated tuberculosis while some develop chronic localised disease. Such variability applies to leprosy which has a wide clinical spectrum between the tuberculoid and lepromatous

forms, where the former is associated with few demarcated lesions and the latter with disseminated infiltrates.

Well-conducted epidemiological studies have provided more direct evidence for the role of host genetics in infectious diseases. Several studies in Burkina Faso in Africa have shown consistently different *P. falciparum* infection rates between ethnic groups (Modiano et al., 1996, 1999). A Danish study on the causes of death among adopted children suggested that early death of a biological but not the adoptive parent from infection increased the risk of death of the child from an infectious disease nearly sixfold (Sorensen et al., 1988). This magnitude is greater than the respective risks for cancers, cardiovascular and cerebrovascular diseases.

Major genes in infectious diseases

There are notable examples of common genetic variants affecting susceptibility to infectious diseases. The observation that haemoglobinopathies occur more commonly in places where *P. falciparum* infection is endemic has led John B. S. Haldane to hypothesise that red blood cell disorders may protect against severe malaria. This hypothesis was later confirmed by Anthony C. Allison, showing that sickle cell heterozygosity (HbAS) protects against falciparum malaria infection (Allison, 1954), with a risk of only about 10% that of normal (HbAA) individuals. Furthermore, absolute protection against *P. vivax*, another pathogen that causes malaria, is conferred by the Duffy blood group negativity caused by homozygosity for a promoter variant of the chemokine receptor gene *DARC*. Other well-known examples include the above-mentioned near-complete resistance of homozygotes for a 32-bp deletion in the *CCR5* chemokine receptor gene to HIV-1 infection (Liu et al., 1996), and the absolute resistance to norovirus infection in individuals with the non-secretor phenotype conferred by inactivating *FUT2* mutations (Lindesmith et al., 2003). As these

mutations are at reasonably common frequencies, they account for a significant proportion of the phenotypic variation in the population and therefore are justified as major genes.

1.5 Human genetics and mycobacterial diseases

Like many other diseases, susceptibility to mycobacterial diseases has a genetic component. Numerous studies of different scales have been conducted to determine the genetic factors for tuberculosis and leprosy susceptibility. Despite years of work, progress in the field has been slow with very few examples of consistent and reproducible genetic loci. In this section I describe the current progress in the field of human genetics in susceptibility to mycobacterial diseases.

1.5.1 Genetic role in mycobacterial diseases

Early studies consistently suggested an important role for host genetics in susceptibility to mycobacterial diseases. This is supported by a study showing significant differences in infection rates among different ethnic groups after a tuberculosis outbreak in a nursing home (Stead et al., 1990). Individuals coming from areas of low tuberculosis prevalence appeared to be more susceptible to tuberculosis (Stead, 1992). Moreover, twin studies have shown consistently higher disease concordance rates for monozygotic compared to dizygotic pairs in tuberculosis (Comstock, 1978) and leprosy (Chakravartti and Vogel, 1973).

1.5.2 Mendelian susceptibility to mycobacterial diseases

Studies of families with unusual mycobacterial infection have shown examples of rare variants affecting susceptibility to certain mycobacteria. These patients usually

present with susceptibility to mildly virulent species of mycobacteria like Bacille Calmette-Guérin (BCG) or environmental mycobacteria, with a Mendelian pattern of transmission in the family. Six genes have been identified to date and are intriguingly all related to the type 1 cytokine pathway (Al-Muhsen and Casanova 2008). These are the two chains of the IFN-γ receptor (*IFNGR1* and *IFNGR2*) (Newport et al., 1996; Jouanguy et al., 1999; Döffinger et al., 2000), a subunit of IL-12 (*IL12B*) (Elloumi-Zghal et al., 2002; Picard et al., 2002) and its receptor (*IL12RB1*) (de Jong et al., 1998), the signal transducer and activator of transcription (*STAT1*) (Dupuis et al., 2001) and the NF-κB Essential Modulator (*NEMO*) (Filipe-Santos et al., 2006). Although these mutations are rare and unlikely to account for common infections, they support the genetic basis of mycobacterial diseases and reflect the importance of the interferon-γ pathway in immunity against mycobacteria.

1.5.3 Common mycobacterial diseases as a complex phenotype

Although Mendalian susceptibility to mycobacterial diseases has been well described, it only accounts for a small number of cases due to the rarity of the mutations. Previous family linkage studies have suggested no major gene in tuberculosis (Bellamy et al., 2000; Miller et al., 2004; Cooke et al., 2008). Thus, the host genetic component is likely to be shared by multiple loci each exerting small effects (Kramnik et al., 1998). This has an important implication for disease mapping, as a large sample size is required to detect genetic variants of small effect sizes. Extensive follow-up and combined analysis are therefore of paramount importance in finding true disease loci (Donnelly, 2008).

1.5.4 Genetic susceptibility to tuberculosis

Candidate genes from animal studies

Laboratory strains of mice show differential susceptibility to infections and may provide insight into the immunology of human tuberculosis (Blackwell, 1998). Two genetic loci affecting tuberculosis susceptibility have been identified through animal work. The first locus corresponds to the human gene *SLC11A1* (solute carrier family 11; also known as *NRAMP*, natural resistance-associated macrophage protein), which encodes a divalent cation transporter protein that appears to be important in macrophage activation (Nevo and Nelson, 2004). Significant associations between variants in this gene and susceptibility to tuberculosis have been reported (Bellamy et al., 1998; Greenwood et al., 2000; Li et al., 2006). The second locus corresponds to the human gene *SP110*. Despite significant associations being reported for several variants in the gene in West Africa (Tosh et al., 2006a), consistent replication was not observed in subsequent studies (Thye et al., 2006; Babb et al., 2007; Szeszko et al., 2007). Further surveys are necessary to determine the role of these genes in susceptibility to tuberculosis.

Candidate genes studies in humans

The highly polymorphic human *HLA* genes encode proteins that present antigenic peptides to T-cells. Given the importance of T-cell immunity in tuberculosis, alleles of these *HLA* genes have been studied with tuberculosis susceptibility. Significant associations have been observed at the *HLA-DR2* locus in an Indian population (Rajalingam et al., 1996) and the *HLA-DQB1* loci in a Cambodian population (Goldfeld et al., 1998). Apart from *HLA*, a number of non-*HLA* genes have also been studied regarding susceptibility to tuberculosis. Significant associations have been reported

in variants flanking *VDR* (Bellamy et al., 1999; Wilkinson et al., 2000; Selvaraj et al., 2000; Bornman et al., 2004; Liu et al., 2004; Lewis et al., 2005), *CD209* (Barreiro et al., 2006; Vannberg et al., 2008), *MBL* (Mombo et al., 2003; Soborg et al., 2003; El Sahly et al., 2004), *TLR2* (Ben-Ali et al., 2004; Ogus et al., 2004), *IFN-*γ (Tso et al., 2005; Etokebe et al., 2006), *IL12B* (Morahan et al., 2007; Sahiratmadja et al., 2007), *IL12RB1* (Akahoshi et al., 2003; Remus et al., 2004; Kusuhara et al., 2007) and others (Cooke and Hill, 2001; Hill, 2006). However, none of these associations are consistently replicated and none clearly reaches genome-wide levels of statistical significance.

Genome-wide linkage studies

The genome-wide linkage approach has been used to study susceptibility to tuberculosis using microsatellite markers. A scan on patients with pulmonary tuberculosis from The Gambia and South Africa has showed suggestive linkage with chromosomal regions 15q11-13 and Xq (Bellamy et al., 2000). Evidence of linkage was observed for chromosomal regions 6p21-6q23 and 20q13 (Cooke et al., 2008). Another linkage study has showed signals at regions 10q26.13, 11q12.3 and 20p21.1 (Miller et al., 2004). These inconsistent results are probably due to insufficient sample sizes and low sensitivity of detection with linkage studies (Lin and Zou, 2004). It is of note that these weaknesses in linkage studies are potentially ameliorated by the association approach.

1.5.5 Genetic susceptibility to leprosy

Candidate genes studies in humans

Due to the difficulty in growing *M. leprae* in animal models, candidate genes in leprosy susceptibility were mainly identified through human studies. An early study in an Indian population has shown significant associations between certain alleles at the *HLA-DR/DQ* locus with leprosy susceptibility (de Vries et al., 1980; van Eden et al., 1982; Xu et al., 1985), particularly with those corresponding to the HLA-DR2 serotype group (van Eden et al., 1980; Schauf et al., 1985; Rani et al., 1993). These observations are consistent with the results of the gene-centric study described in this thesis in Chapter 5. Various class I antigen associations were also suggested (Shankarkumar et al., 2003), but no clear or consistent associations have been identified. Apart from the *HLA* region, significant associations have been reported in *VDR* (Roy et al., 1999), *TNF-α* (Roy et al., 1997; Levee et al., 1997; Santos et al., 2002; Vejbaesya et al., 2007; Franceschi et al., 2009), *LTA* (Alcais et al., 2007), *PARK2/PACRG* (Mira et al., 2004) and others (Cooke and Hill, 2001; Hill, 2006), of which several will be assessed and discussed in Chapter 5 in this thesis.

Genome-wide linkage studies

Genome-wide linkage studies have been performed to detect linkage in leprosy using microsatellite markers. A family study on South Indians has found suggestive linkage on chromosomal regions 10p13 and 20p12.3 (Siddiqui et al., 2001; Tosh et al., 2002b). A consistent signal at chromosome 10p13 was observed in a separate Vietnamese study, which also identified chromosome 6q25 as a susceptibility locus (Mira et al., 2003). A subsequent study in Brazil identified three other chromosomal regions at 6p21.32, 17q22 and 20p13 with a non-significant linkage to leprosy (Miller et al.,

2004). Overall, the strongest evidence of linkage was observed at the chromosome 10p13 locus.

Genome-wide association study

A GWA study on susceptibility to leprosy was conducted on a Chinese population and recently published (Zhang et al., 2009). The primary study was conducted on 706 cases and 1225 controls, genotyped for more than 600,000 markers throughout the genome. The study identified significant associations in the *HLA-DR* locus as well as in five other genes, namely *NOD2*, *RIPK2*, *TNFSF15* and *CCDC122/C13orf31*. Interestingly, the genes *NOD2*, *RIPK2* and *TNFSF15* encode proteins that are involved in the pro-inflammatory NF-κB pathway that activates IFN-γ, which is important in immunity against mycobacteria. Genetic variants of *NOD2*, *TNFSF15* and *CCDC122/C13orf31* have been reported to associate with Crohn's disease (Barrett et al., 2008), which has been speculated to have an infectious cause related to *Mycobacterium avium paratuberculosis* infection (Pierce, 2009).

1.5.6 Genetic susceptibility to common mycobacterial diseases

Despite the identification of many genetic loci, the field of genetic susceptibility to mycobacterial diseases is far from mature. Consistent and reproducible associations across populations are scarce, and at best explain only a modest proportion of the phenotypic variation. However, due to the relatively low level of diversity among *M. tuberculosis* and *M. leprae* (particularly the latter), host genetics is likely to play a major role in susceptibility to these mycobacterial diseases. This thesis presents continuing work in the search for genetic variants affecting common mycobacterial diseases in humans.

1.6 Interactions between humans and mycobacteria

Understanding the host-pathogen interaction is important for disease mapping, as pathogen infectivity may depend on both host and pathogen strains and their compatibility. In this section, I describe aspects of mycobacterial and human genetics relevant to their interactions, and discuss evidence for the long-standing co-evolutionary relationship between the two species.

1.6.1 Mycobacterial genetics and adaptation

Co-evolution is the process of reciprocal adaptive genetic change in two or more species. Although the co-evolutionary relationship between humans and mycobacteria is difficult to prove, several observations suggest that such a relationship is possible. Firstly, mycobacterial infections have a long history in humans with the earliest evidence dating to several thousand years ago. This long term interaction has allowed selection to take place over time. Secondly, tuberculosis and leprosy are both chronic diseases, in which the manifestations are heavily dependent on the intimate interactions between host cells and pathogens. Thirdly, tuberculosis and leprosy are major infectious diseases with significant mortality and morbidity (The World Health Organization, 2007, 2009), with a considerable proportion of patients at their reproductive age. Fourthly, susceptibility to mycobacteria is a heritable trait and significant genetic variation exists between individuals (Stead et al., 1990). Collectively, these factors have provided a favourable condition for a long term co-evolutionary relationship to build up between humans and mycobacteria. This relationship is consistent with the hypothesis by John B. S. Haldane that pathogens can impose evolutionary effects on their hosts. In fact, the evolutionary pressure of *M. leprae* on humans may be underestimated, given the physical disfigurement and intense social segregation associated with leprosy.

The interaction between host and pathogen is a powerful determinant of the genetic susceptibility to the infection (Woolhouse et al., 2002). The Red Queen hypothesis suggests that ecological relationships between hosts and pathogens have strong evolutionary effects on one another (Haldane, 1949; Lively and Dybdahl, 2000), which would reciprocally select adaptive genetic changes on both species (Hamilton et al., 1990; Howard and Lively, 1994; Peters and Lively, 1999). Examples of such reciprocal traits in co-evolution include pathogen infectivity and host resistance, and functionally, the ability of the host to clear an infection versus the ability of the pathogen to evade or suppress host defences (Woolhouse et al., 2002). The relationship between humans and mycobacteria shows features compatible with host-pathogen co-evolution. Both *M. tuberculosis* and *M. leprae* are obligate intracellular parasites and their virulence is associated with decreased fitness of the human host. Such struggle for survival has likely imposed evolutionary pressure on both species resulting in readily observable reciprocal traits. Humans have developed a sophisticated, multi-layered immune system to combat infectious agents including mycobacteria, whereas certain mycobacteria have developed mechanisms to counteract host defence through attenuation of host immunity (Doz et al., 2007; Singh et al., 2009), intracellular phagolysosomal block (Russell, 2001; Vergne et al., 2005; Rohde et al., 2007), inhibition of radical generation (Davis et al., 2007) and resistance to acidification (Vandal et al., 2008).

Apart from reciprocal traits, recent genetic studies of mycobacteria also support this co-evolutionary relationship between humans and mycobacteria. The genome sequence of *M. tuberculosis* shows abundant genes encoding the PE/PPE families of proteins (Cole et al., 1998) which may modulate inflammation upon infection of the host (Nair et al., 2009). Moreover, substantial host-pathogen compatibility has been demonstrated between different *M. tuberculosis* lineages and human populations (Hirsh et al., 2004; Gagneux et al., 2006). As for *M. leprae*, molecular studies have

shown its highly degradative genome compatible with its specialised parasitic niche (Cole et al., 2001). These studies provide support for a co-evolutionary relationship between humans and mycobacteria in nature.

1.7 Outline of this thesis

The aim of this thesis is to identify genetic loci contributing to susceptibility to the common mycobacterial diseases of tuberculosis and leprosy. This thesis depicts the efforts to dissect human immunity to mycobacteria, and marks the era of transition from the candidate gene to the hypothesis-free genome-wide approach in mapping infectious disease genes. I describe my work on several aspects of these projects.

Chapter 3 describes the WTCCC tuberculosis study, which is a GWA study based on samples from the The Gambia. I describe the finding of a potential susceptibility locus, as well as some difficulties encountered during the course of the study which are informative on studying genetics in African populations.

Chapter 4 combines the results of the WTCCC Gambian tuberculosis study with a Ghanaian tuberculosis study. This results in a combined study with a greater power to detect genetic variants. Together with the replication cohorts, this study involves over 10,000 samples and identifies two novel susceptibility loci. The results of this study entail the most comprehensive survey of host genetic factors to tuberculosis to-date.

Chapter 5 investigates genetic susceptibility to leprosy using a gene-centric microarray. I present the results of this study with successful identification of two susceptibility loci. In addition, I describe work on replicating several genetic loci from a recent Chinese GWA study in the Indian populations.

Chapter 6 describes an analysis on evidence of overlapping genetic loci between tuberculosis and leprosy. This lays the foundation to study host genetic factors for both mycobacterial diseases. I further present work on a functional gene annotation and pathway analysis. This identifies potentially important pathways for human immunity against mycobacteria.

Finally, I summarise these results and discuss their potential impacts on improving our understanding of anti-mycobacterial immunity. I also discuss the current development and future direction for the field of medical genetics. It is hoped that these scientific advances will bring us one step closer to the invention of new vaccines and therapeutics for effective control of mycobacterial diseases.

Chapter 2

Materials and Methods

2.1 Introduction

Population based genetic studies rely heavily on the collection of samples with well-defined phenotypes, followed by accurate sample genotyping and statistical analysis. This chapter describes the various tuberculosis and leprosy cohorts, laboratory methods and statistical analyses used in the studies described in this thesis.

2.2 Disease cohorts

2.2.1 Tuberculosis

The Gambia case-control cohort

The Gambia tuberculosis cohort consists of 1,498 pulmonary tuberculosis cases and 1,496 controls recruited from The Gambia. The diagnosis of tuberculosis was based on compatible clinical features confirmed by culture or smear positivity. All cases that were smear positive but culture negative had a radiographic confirmation, and were

collected at major urban health centres (Lienhardt et al., 2002; Bennett et al., 2002). All cases were above 15 years of age. The controls were newborns recruited from routine births at local clinics (Jallow et al., 2009). The majority (>95%) of the cases was screened for HIV-1, with positive cases excluded from the study because HIV infection increases the risk of tuberculosis. All samples were obtained with informed consent from the individuals who has donated blood, or their parents or guardians. Ethical approval for the study was granted by the Medical Research Council (MRC) and the Gambian government joint ethical committee.

Guinea case-control cohorts

Two tuberculosis cohorts were recruited from the West African countries of Guinea-Bissau (Bissau) and the Republic of Guinea (Conakry) (Lienhardt et al., 2002; Bennett et al., 2002). The tuberculosis cases were recruited at major urban health centres and included patients over 15 years of age with two consecutive positive sputum smears or positive culture. Community members with no history of tuberculosis were recruited as controls. The total number of HIV-1 negative samples in the Bissau-Guinean set was 254 with 138 cases and 116 controls, and the Conakry-Guinean set had a total of 345 samples with 178 cases and 167 controls. All samples were obtained with informed consent for this tuberculosis study. Ethical approval was provided by the Ministry of Public Health, Guinea-Bissau and the National Ethics Committee, Ministry of Health, Republic of Guinea.

Malawi case-control cohort

Tuberculosis case and control samples were recruited as part of the Karonga Prevention Study (KPS) based in the Karonga district in northern Malawi (Fitness et al., 2004). Tuberculosis cases were recruited in health centres with the diagnosis con-

firmed by sputum smear, bacteriological culture or lymph node aspirate histology. Extra-pulmonary tuberculosis cases were excluded from this genetic study. Samples with a single smear of less than 10 bacilli were excluded. Controls included individuals with no history of tuberculosis or leprosy and were matched for age, sex and area of residence. HIV-1 samples were excluded from the analysis. The total number of samples for analysis was 1,512 with 320 cases and 1,192 controls. All samples were obtained with informed consent for this tuberculosis study. Ethical approval was obtained from the National Health Sciences Research Committee of Malawi and the London School of Hygiene and Tropical Medicine Ethics Committee.

Ghana case-control and family cohorts

Tuberculosis patients were enrolled at seventeen different hospitals in Accra and Kumasi, Ghana (Herb et al., 2008; Thye et al., 2009). Tuberculosis cases were diagnosed with by clinical pictures with radiological confirmation and consisted of 2,147 individuals. All patients were screened for HIV status, and only HIV negative cases were included in this study. The control samples consisted of 5,565 samples collected in two batches - 3,551 healthy adults with no history of tuberculosis, and 2,014 children with a median age of 20 months. In addition, 332 parent child trios or duos were also collected for family studies. The study protocol was approved by the Committee on Human Research, Publications and Ethics of the Kwame Nkrumah University of Science and Technology, and the Ethics Committee of the Ghana Health Service, Accra. Venous blood samples were taken with informed consent of the patients or their guardians.

2.2.2 Leprosy

New Delhi case-control cohort

The New Delhi leprosy cohort consists of 258 leprosy patients and 300 controls recruited from the Hansens out-patient unit of Lok Nayak Jai Prakash Hospital in New Delhi, India (Malhotra et al., 2005, 2006). Diagnosis of leprosy was made by at least two independent leprologists with standard histopathological examination of affected skin lesions. Classification of leprosy was based on clinical (Norman et al., 2004) and histological criteria (Ridley and Jopling, 1962, 1966). Patients were classified either as multibacillary (MB, with a bacterial index >0), a group that included patients with lepromatous (LL), borderline lepromatous (BL) and borderline (BB) leprosy; or as paucibacillary (PB, with a bacterial index of 0) which included patients with borderline tuberculoid (BT) and tuberculoid (TT) leprosy. Multibacillary leprosy was characterised by the presence of six or more skin lesions that were smaller in size, tending to be bilaterally symmetrical with about 10-40% loss of sensation. Paucibacillary leprosy was characterised by the presence of large well-defined skin lesions which were less than five in number, dry and with almost 90-100% loss of sensation. The controls consisted of unrelated individuals with no history of leprosy or tuberculosis, and were matched to the same geographical area. An informed written consent was obtained from all the individuals. The study was approved by the Jawaharlal Nehru University ethics committee.

Kolkata case-control cohort

This cohort consists of 220 cases and 162 controls recruited at the School of Tropical Medicine, Kolkata, India (Roy et al., 1997, 1999). Diagnosis was based on the clinical assessment of skin lesions and peripheral neuropathy, plus the presence of acid-fast

bacilli in slit-skin smear examination. Diagnosis of lepromatous leprosy required the presence of diffuse skin lesions with numerous acid-fast bacilli in slit-skin smears taken from at least three parts of the body, including one from an ear lobe. Tuberculoid leprosy was defined by the presence of asymmetrical well-defined lesions with a dry surface and the absence of detectable acid-fast bacilli. Borderline leprosy patients were not recruited for this study. The control samples were from blood donors visiting the Swasti Blood Clinic, Kolkata. Controls included for this cohort were unrelated and seronegative for antibodies to the HIV virus, *Treponema pallidum*, and hepatitis B surface antigen. Both patients and control groups included Hindus, Muslims, and Christians; and the Hindus included four castes: Brahmins, Vaidyas, Kaisthas, and Shudras. The cases and controls were matched for the ethnic groups. Informed consent was obtained from all the individuals in this study.

Madurai family cohort

The Madurai leprosy cohort consists of families affected by leprosy recruited from Sakthi Nagar and Kumbakonam in Tamil Nadu, India (Siddiqui et al., 2001; Tosh et al., 2002a, 2006b). All patients were diagnosed based on the World Health Organisation (WHO) guidelines with clinical signs including hypopigmented skin lesions and peripheral neuropathy, and were classified according to the Ridley and Jopling classification (Ridley and Jopling, 1962, 1966). A total of 754 individuals from 189 nuclear families, including 17 families with one affected child, 161 families with two affected children and 11 families with more than two affected children were included in this study.

2.3 Laboratory Methods

2.3.1 DNA extraction and quantification

DNA extractions for the tuberculosis samples were performed from blood using the Nucleon II kit (Scotlab Bioscience, Buckingham, UK). The extractions for the leprosy samples was performed using phenol chloroform (Malhotra et al., 2005, 2006) or sodium chloride salt (Roy et al., 1997, 1999) extraction. DNA concentrations were determined using PicoGreen (Invitrogen, Carlsbad, USA).

2.3.2 Polymerase chain reaction

Polymerase chain reactions (PCR) were performed prior to genotyping or sequencing. The PCRs for the genotyping assays were performed with the Qiagen HotStarTaq DNA Polymerase with a mixture of 2μL of 10ng/μL DNA, 0.63μL of 10X Qiagen PCR buffer (contains 15mM $MgCl_2$), 0.33μL of 25mM $MgCl_2$, 0.31μL of 8mM dNTP mix, 0.5μL of primer mix at 1μM each, 0.5 units of HotStarTaq DNA polymerase and deionised water to a 5μL final volume. The reactions were performed on Tetrad thermocyclers (MJ Research, USA) with conditions as follows: 95°C for 15 minutes, followed by 35 cycles of 94°C for 15 seconds, 55-65°C for 30 seconds, 72°C for 30 seconds, and a final extension of 72°C for 5 minutes.

The long-range PCRs for sequencing were performed with the Invitrogen SequalPrep Long PCR Kit with a mixture of 1μg of 3.3ng/μL DNA, 2μL of 10X SequalPrep reaction buffer (contains 15mM $MgCl_2$), 0.4μL of dimethyl sulfoxide (DMSO), 0-2μL of 10X SequalPrep enhancer A, 0-2μL of 10X SequalPrep enhancer B, 1μL of primer mix at 10μM each, 1.8 units of SequalPrep DNA polymerase and deionised water to a 20μL final volume. The reactions were performed on Tetrad thermocyclers (MJ Research, USA) with conditions as follows: 94°C for 2 minutes, followed by 10 cycles

of 94°C for 10 seconds, 55-65°C for 30 seconds, 68°C for 1 minute / Kb, followed by 25 cycles of 94°C for 10 seconds, 55-65°C for 30 seconds, 68°C for 1.33 minute / Kb, and a final extension of 72°C for 5 minutes.

2.3.3 Whole genome DNA amplification

Samples with low DNA quantity were amplified with the GenomiPhi Amplification Kit (GE Healthcare, UK). This uses the multiple displacement amplification technology with ϕ29 DNA polymerase and random hexamers that non-specifically amplify the input DNA. Successful amplification was indicated by a positive smear on agarose gel electrophoresis of the undiluted product, and confirmed with Picogreen quantification. The amplified products were diluted to the desired concentrations before further downstream applications.

2.3.4 Agarose gel electrophoresis

Agarose gels were prepared at concentrations of 2-3% by mixing agarose powder (Sigma-Aldrich, UK) to 1X Tris acetate (TAE) buffer (Sigma-Aldrich, UK). The mixture was heated in a microwave oven until complete dissolution. After cooling 5μL of ethidium bromide per 100μL of solution was added and swirled, and the gel was poured to a suitable gel tray to set. Samples were loaded with xylenol orange dye and electrophoresis was performed at 100 volts for 30-60 minutes before visualization by ultraviolet transillumination.

2.3.5 Genotyping assays

Restriction fragment length polymorphism

Restriction fragment length polymorphism (RFLP) assay was used for genotyping by incubating PCR products with a suitable restriction enzyme, for 4 hours to overnight at a temperature specified by the manufacturer. The digested products were separated by agarose gel electrophoresis and visualised with ultraviolet transillumination.

Sequenom primer extension assay

The Sequenom MassArray primer extension assay uses a multi-staged approach to genotype SNPs based on mass spectrometry (Gabriel and Ziaugra, 2004). The first stage involves a multiplexed PCR resulting in products of approximately 100bp. This is followed by extension of an oligonucleotide probe over a polymorphic site with a mixture of deoxynucleotides and dideoxynucleotides, producing extended products of different sizes detectable by matrix-assisted laser desorption / ionisation time-of-flight (MALDI-TOF) mass spectrometry.

LightTyper melting curve assay

The LightTyper melting curve assay works by detecting a difference in the melting patterns between the post-amplified DNA strands of different genotypes (Bennett et al., 2003), using fluorescence resonance energy transfer in a LightTyper device (Roche Diagnostics, Germany).

DNA microarray genotyping

Recent advances in DNA microarray technology have allowed a large number of polymorphisms to be genotyped at one time. These chips often consist of oligonucleotide probes arrayed at high density, which are hybridised to the DNA templates to give fluorescence or chemoluminescence corresponding to the genotype. This technology has greatly increased throughput and enabled genome wide studies to be conducted.

Two different microarray chips have been used in the studies described in this thesis. The Gambian samples in the tuberculosis study in Chapter 3 were genotyped with the GeneChip Human Mapping 500K Array Set (Affymetrix, USA), which was made up of two arrays (Nsp and Sty) collectively capable of genotyping about 500,000 random SNPs throughout the genome. The New Delhi Indian samples in the leprosy study in Chapter 5 were genotyped with the HumanCVD Genotyping BeadChip (Illumina, USA) which had about 50,000 markers covering approximately 2,100 genes in the genome.

2.3.6 Solexa DNA sequencing

DNA sequencing was performed using the Genome Analyzer IIx (Illumina, USA) in the WTCCC tuberculosis study in Chapter 3. The first stage of the experiment involved preparing libraries of genomic DNA. The target regions for resequencing were amplified as described in Section 2.3.2. The purified PCR products were sheared to less than 800bp fragments using an adaptive focused acoustic nebuliser (Covaris, USA), and the overhangs were repaired using T4 DNA polymerase and Klenow DNA polymerase. These blunt ended phosphorylated DNA fragments were then extended and ligated to adapters. After further purification and amplification, the DNA templates were loaded onto a flow cell for analysis.

The second stage of the experiment involved the sequencing-by-synthesis reaction on the flow cell. The reaction was performed with four fluorescent labelled nucleotides (A, C, T, G), which were incorporated into the DNA fragments according to the sequence on the template. The image of the emitted fluorescence from each cluster was captured after laser excitation. Because of the high densities of molecules on the flow cell, tens of millions of DNA templates can be sequenced at the same time.

2.4 Statistical and analytical methods

2.4.1 Power calculation

The power to detect genetic association was calculated using a summarised non-centrality parameter (λ) in a non-central χ^2 distribution (Patnaik, 1949). The non-centrality parameter λ is a function of the sample size, allele frequencies in the cases and controls, and the correlation between the marker and the disease-causing variant which is given by

$$\lambda = \frac{2r^2 N_{case} N_{ctrl} (F_{case} - F_{ctrl})^2}{N(F)(1-F)} \tag{2.1}$$

where N is the number of cases and controls, F is the frequency of the risk allele, and r^2 is the correlation between the marker and the disease-causing variant. The probability of a type II error (β) is given by $\chi_1'^2(\lambda)$, which is the cumulative distribution function of the non-central χ^2 distribution with 1 degree of freedom with a non-central parameter λ. The power of the test can be calculated by by 1-β.

2.4.2 Genotype calling

Digested PCR products in the RFLP assays were transilluminated with ultraviolet light. The sizes of the bands were determined by comparing side-by-side to the DNA ladder, and genotypes were called by visual inspection. Genotypes for the Sequenom

assay were called with the MassARRAY Typer software (Sequenom, USA). After automated calling, visual inspection of the cluster plots were carried out for every SNP. Genetic variants with non-segregating genotype clusters were discarded from analysis. Individual genotypes with ambiguous calls were marked as no call. Samples with genotype calls of less than 70% in an assay were discarded from the analysis.

Genotypes for the GeneChip Human Mapping 500K Array Set were called using the CHIAMO algorithm, which uses a Bayesian hierarchical four-class mixture model to call genotypes (The Wellcome Trust Case Control Consortium, 2007). All genotypes were based on *a posteriori* call probability threshold of 90%, which resulted in a discordance rate of 0.48% in the tuberculosis study, based on about 50 million duplicate genotypes. Genotypes for the HumanCVD Genotyping BeadChip were called using the BeadStudio Genotyping Module (Illumina, USA) with a calling threshold for 0.15 on the GenCall score (Browning and Yu, 2009).

2.4.3 Hardy-Weinberg equilibrium

The Hardy-Weinberg equilibrium describes the relationship between allele frequencies and genotype frequencies (Hardy, 1908; Weinberg, 1908). In a large non-random mating population under no selection, migration or non-random mating, these frequencies are constant at an equilibrium from generation to generation. Let p and q represent the frequencies of two alleles M and m of a polymorphism such that $p + q = 1$, the genotype frequencies under the Hardy-Weinberg equilibrium are given as follow

$$p^2 + 2pq + q^2 = 1 \tag{2.2}$$

Where p^2 and q^2 represent frequencies of the MM and mm homozygotes, and $2pq$ represents the frequency of the Mm heterozygote. In the context of association studies, the genotype frequencies of the SNPs can be compared to their expected

frequencies under the Hardy-Weinberg equilibrium with the Pearson's χ^2 test at 1 degree of freedom. Significant deviation may indicate genotype errors and is therefore used as a quality control criterion (Gomes et al., 1999; Hosking et al., 2004). The threshold of $p < 0.001$ was used to indicate significant deviation in this thesis unless otherwise specified.

2.4.4 Genome-wide analytical methods

Identity-by-state analysis

The crux of association studies lies in comparing gene frequencies between unrelated cases and controls. The presence of duplicated samples or closely related individuals may skew the frequency estimates through linkage and therefore give rise to spurious associations. To identify these samples in genome-wide studies, the identity-by-state (IBS) relationships between each pair of individuals were analysed. The genome-wide identity-by-descent (IBD) can be calculated from the IBS information. Each pair of samples showing evidence of duplication or cryptic relatedness was investigated, and the sample of the lower genotype rate in each pair was discarded for subsequent analysis. The thresholds used in the studies in this thesis will be discussed in the relevant result chapters.

Multidimensional scaling analysis

Population stratification can give rise to false positive associations (Devlin and Roeder, 1999; McCarthy et al., 2008). To understand the population structure between samples, multidimensional scaling (MDS) was used to analyse the identity-by-state relationships of each pair of individuals in genome-wide studies. It is a multivariate statistical method to explore similarities and dissimilarities between each pair of data,

and approximate them into low-dimensional spaces through preservation of these relationships. The location of these items, representing individuals in the studies, can then be visualised on two-dimensional plots. It has been shown to have comparable type 1 error, power and bias with principle component analysis (PCA) (Zhu and Yu, 2009) which is an equivalent method used in exploratory data analysis. The MDS analysis was performed for the study samples either with or without the reference samples from the International HapMap Project, which includes 270 individuals from four human populations of European Amerian (CEU), Han Chinese (CHB), Japanese (JPT) and Yoruban Nigerian (YRI) (The International HapMap Consortium, 2003, 2005).

Quality controls and significance thresholds

Apart from the identity-by-state and multidimensional scaling analyses, genome-wide studies have other quality control measures including genotype and sample call rates, overall heterozygosity and cluster plot inspection. Different thresholds should be applied to the significance tests of genetic association and Hardy-Weinberg equilibrium, due to multiple testing in GWA studies. Details of these measures and thresholds will be discussed in the relevant result chapters.

2.4.5 Single locus tests of significance

Several tests were used for testing single locus association. These include the Pearson's χ^2 test, the Cochran-Armitage test for trend and the logistic regression model. The null hypothesis of no association was rejected at a conventional threshold of $p < 0.05$, except in high throughput genotyping where multiple testing warranted otherwise specified thresholds.

Pearson's χ^2 test

Genotype or allele counts can be tested for association using the Pearson's χ^2 test. The χ^2 statistic tests for departure of the observed values from the expected values across cells in a contingency table. It is the sum of the squared differences between the observed values (O_i) and the expected values (E_i) as given by

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{2.3}$$

where n is the number of possible outcomes, and X^2 is the test statistic that asymptotically approaches a χ^2 distribution. In Pearson's χ^2 test, a 2 x 3 contingency table with 2 degrees of freedom can be used to determine a genotypic association, whereas a 2 x 2 contingency table with 1 degree of freedom is used for allelic, dominant and recessive models of associations. Using the notations from Table 2.1 for the allele counts in a case-control study, the test statistic for the allelic test can be calculated by

$$X^2_{allelic} = \frac{r(p_1 q_2 - p_2 q_1)^2}{pqr_1 r_2} \tag{2.4}$$

which follows a χ^2_1 distribution under the null hypothesis. However, it should be noted that the allelic test is incorrect if the Hardy-Weinberg equilibrium does not hold in the combined case-control population. Furthermore, the more conservative Fisher's exact (2 x 2 table) and the Fisher-Freeman-Halton exact (2 x 3 table) tests should be used for a table with any cell count less than five.

Cochran-Armitage test for trend

Apart from the Pearson's χ^2 test, inference about the null hypothesis can be based on the Cochran-Armitage test for trend (Cochran, 1954; Armitage, 1955). It tests for a linear trend of disease risk across the genotypes, and has the advantage of being robust against departure from the Hardy-Weinberg equilibrium (Sasieni, 1997). Taking the

Genotype count	mm	Mm	MM	Total
Case	a_0	a_1	a_2	a
Control	b_0	b_1	b_2	b
Total	c_0	c_1	c_2	c

Allele count	m	M	Total
Case	p_1	p_2	p
Control	q_1	q_2	q
Total	r_1	r_2	r

Table 2.1: Notation for genotype and allele counts in a case-control study

notations for the genotype counts in Table 2.1, the general form of the test is

$$X^2 = \frac{c(c\sum a_i x_i - a\sum c_i x_i)^2}{a(c-a)[c\sum c_i x_i^2 - (\sum c_i x_i)^2]} \tag{2.5}$$

where x_i are weights assigned to each genotype. To assess the additive effect of an allele on disease risk, substitute $x_i = i, i = 0, 1, 2$ into the equation to get the trend statistic

$$X^2_{trend} = \frac{c[c(a_1 + 2a_2) - a(c_1 + 2c_2)]^2}{a(c-a)[c(c_1 + 4c_2) - (c_1 + 2c_2)^2]} \tag{2.6}$$

which follows a χ^2_1 distribution under the null hypothesis. The trend test is most powerful if and only if the allele effect is exactly co-dominant, i.e., the homozygous odds ratio is the square of the heterozygous one. As the trend test is robust to departure from the Hardy-Weinberg equilibrium, it was used as a primary test of association for the study in Chapter 5.

Generalised linear model

The generalised linear model uses ordinary least square regression to relate an outcome to independent variables with a link function. In disease association, the outcome is a binary phenotype which is related to the genotype and other variables through a *logit* link in a logistic regression model. The variables can be either

categorical or continuous. The logistic regression model offers a way to study the relationships between genotypes and phenotypes after correcting for possible confounders, such as age, gender, ethnicity and HIV status in the context of infectious disease genetic studies. This is important in association studies with populations of diverse ethnic background, which may necessitate correction for ethnicity to study the independent effects of SNPs. It can also be used to study the effect of the SNPs under additive, dominant and recessive models. The logistic regression was used in the studies in Chapters 3 and 4.

The logits and natural logarithm of odds of the binomial probabilities are modelled as a linear function of X_i as follows:

$$logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \cdots + \beta_k X_{k,i} \tag{2.7}$$

The unknown parameters β_j can be determined by the maximum likelihood estimate.

Transmission disequilibrium test

The genotype data in the family studies were analysed with the transmission disequilibrium test (TDT) (Spielman et al., 1993). It is a family based association test which measures the over-transmission of an allele from heterozygous parents to affected offspring. This is robust to population stratification and detects linkage only in the presence of genetic association (Laird and Lange, 2006). However, it is worth noting that this test has a low power compared to case-control analyses, making it specific but not sensitive in detecting association (Carey and Williamson, 1991; Laird and Lange, 2006).

The test is built upon Mendelian inheritance that, each parent only passes one of their two alleles to their offspring. Let M and m be the two alleles at a SNP. For a set of n heterozygous parents of an affected offspring with alleles M and m at a SNP,

let b be the number of parents transmitting M but not m, and c be the number of parents transmitting m but not M to the offspring, the TDT statistic is given by the McNemar statistic as:

$$X^2_{tdt} = \frac{(b-c)^2}{b+c} \tag{2.8}$$

which follows a χ^2_1 distribution under the null hypothesis. This test was used to analyse the family data in the study in Chapter 5.

2.4.6 Combining different studies

Mantel-Haenszel test

Genotype or allele counts from different study populations were combined and analysed using the Mantel-Haenszel test (Mantel and Haenszel, 1959), which evaluates the common effect size and evidence of association across K populations. For 2 x 2 contingency tables in K populations, taking notations for the allele counts in Table 2.1, the Mantel-Haenszel estimator of the effect size is

$$\hat{\psi}_{mh} = \frac{\sum(p_{1k}q_{2k}/r_k)}{\sum(p_{2k}q_{1k}/r_k)} \tag{2.9}$$

where p_{jk}, q_{jk} and r_k are the observed frequencies in the k^{th} population. The Mantel-Haenszel test statistic is given by

$$X^2_{mh} = \frac{[\sum(p_{1k}q_{2k} - p_{2k}q_{1k})/r_k]^2}{\sum\{(r_{1k}r_{2k}p_kq_k)/[(r_k-1)r_k^2]\}} \tag{2.10}$$

which follows a χ^2_1 distribution under the null hypothesis. It should be noted that as the Mantel-Haenszel test assumes a fixed effect for each of the included studies, the test is only valid when there is no significant heterogeneity between them. The homogeneity of odds ratio can be tested with the Breslow-Day test (Breslow and Day, 1980). The Mantel-Haenszel test was used for combining studies from different case-control populations in Chapter 5.

Fisher's combined test

The Fisher's combined test can be used to combine statistics from several studies (Fisher, 1922), and is given by

$$X^2_{fisher} = -2\sum log_e P_k \tag{2.11}$$

χ^2_1 where P_k is the P-value for the k^{th} hypothesis test. The test statistic follows a χ^2_{2K} distribution under the null hypothesis. The meta-analysis null hypothesis is that all of the separate null hypotheses are true, and the meta-analysis alternative hypothesis is that at least one of the separate alternative hypotheses is true. Although this method gives a combined statistics, it does not take effect size and direction. This test was therefore only used to assess the aggregate gene-based statistics without inference about the direction of effect in Chapter 6.

Inverse variance method

The inverse variance method represents another approach to combine results from different studies (Fleiss, 1993; Cantor et al., 2010). This method relies on assigning a weight to each study to determine the combined test statistic and effect size. Assuming there is a fixed effect shared by all the K included studies and let T_k denote the effect size of an allele for the k^{th} study, the weight is equal to the reciprocal of the study variance $w_k = 1/\sigma^2$. The meta-analysis effect size is given by

$$T_{ma} = \frac{\sum w_k T_k}{\sum w_k} \tag{2.12}$$

with the corresponding variance given by $\sigma^2_{ma} = (\sum w_k)^{-1}$. The inverse variance test statistic is given by

$$X^2_{ma} = \frac{(T_{ma})^2}{\sigma^2_{ma}} \tag{2.13}$$

which follows a χ^2_1 distribution under the null hypothesis. However, as with the Mantel-Haenszel test, the fixed effect assumption is violated when there is significant

heterogeneity between studies. In such cases, the statistic can be modified to a random effects model (DerSimonian and Laird, 1986) which allows the effect in each study to vary, by incorporating the between-studies variance τ^2 to the weights given by $w_k = 1/(\sigma^2 + \tau^2)$. The combined statistic in the random effects model estimates the mean of the distribution of effect sizes among the included studies and accounts for the between-studies variance. The inverse variance method was used to combine studies of different populations in Chapters 3 and 4.

Heterogeneity assessment

The assessment of heterogeneity is important because it can affect the choice of statistical model for combining data (Ioannidis, 2007; Sagoo et al., 2009; Cantor et al., 2010). The Q statistic has been proposed as a measure to test for heterogeneity (Cochran, 1954) and is given by $Q = \sum w_k(T_k - T_M)^2$ taking the notations from the inverse variance method. Under null hypothesis, the Q statistic follows a χ^2_{K-1} distribution. Furthermore, the I^2 index has been used to quantify the extent of heterogeneity (Higgins and Thompson, 2002) and is given by

$$I^2 = \begin{cases} \frac{Q-(K-1)}{c} & \text{for } Q > (K-1) \\ 0 & \text{for } Q \leq (K-1) \end{cases} \tag{2.14}$$

2.4.7 Haplotypic analysis

Linkage disequilibrium

Linkage disequilibrium (LD) is the non-random association of alleles at two or more genetic loci. Under the conditions of linkage equilibrium, alleles are passed to offspring independently of each other and without any association. However because of incomplete recombination, alleles in close proximity to each other on the same chro-

mosome are more likely passed to be passed on together. This results in a non-random association of the two markers at a population level described as LD.

Although LD can be described by two commonly used estimates D' and r^2 (Hill and Robertson, 1966; Lewontin, 1964), D' values are known to fluctuate with small numbers of samples or rare alleles. The coefficient r^2 is most relevant in the context of association mapping as there is a simple inverse proportion relationship between r^2 and the power to detect association (Mueller, 2004). Therefore the coefficient r^2 has been used in the studies described in this thesis. Let M and m be alleles on the first locus and N and n be alleles on the second locus so that the coefficient r^2 is given by:

$$r^2 = \frac{(P_{MN} - P_M P_N)^2}{P_M P_m P_N P_n} \tag{2.15}$$

where P_M, P_m, P_N and P_n represent the frequencies of the corresponding alleles. The assessment of LD is important in fine-mapping disease associations, as well as haplotypic analysis.

Haplotype reconstruction and test of significance

A haplotype is a combination of closely linked markers on the same chromosome, which tend to be inherited together and not interrupted by recombination. Haplotypic analysis is important in genetic mapping, because the association signal of a casual variant may appear inconspicuous among its tagging markers until haplotypes of different populations are compared. In addition, multiple casual variants may interact to form a haplotype as a functional unit. In the studies described in this thesis, blocks of haplotypes were reconstructed using the confidence interval definition (Gabriel et al., 2002). The Pearson's χ^2 test was performed to compare the haplotype frequencies between the cases and controls.

2.4.8 Sequence analysis

Base calling quality

As described in Section 2.3.6, the DNA samples were prepared and resequenced using the Genome Analyzer IIx. The resequencing was performed in single-end with a read-length of 52 base pairs, with a quality score for each base. This score can be converted to the standard Phred quality score Q (Ewing and Green, 1998; Ewing et al., 1998) which is related to the probability of incorrect base calling by

$$Q_{phred} = -10log_{10}P_{error} \tag{2.16}$$

For example, a Phred quality score of 20 will be translated as having 1 in 100 probability of being incorrect, with an accuracy of 99%.

Sequence alignment and variant calling

Sequence alignment was performed with the Novoalign software (www.novocraft.com). Quality control filters were applied at several stages. The first 25 bases of each read were used for analysis, since the base quality decreases with more errors further along the sequence reads. Sequences with at least 13 good bases ($Q_{phred} \geq 20$) were analysed. The alignment was performed with the NeedlemanWunsch algorithm (Needleman and Wunsch, 1970) which assigns a positive score to a correct match and a penalty to a gap or mismatch according to a similarity matrix. Satisfactory alignments required reads have at least 95% identity with the reference sequence, uniquely aligned with an alignment score of greater than 50. Furthermore, an alignment quality score was computed based on the probability of the alignment given the read and the reference sequence. Reads with an alignment quality score of greater than 20 were accepted.

Finally, variant calling was performed by the VarScan software (Koboldt et al., 2009) and required the region to have at least ten-fold coverage by the sequence reads. Genetic variants with a frequency of more than 3% were called. The minor allele frequency of a SNP is given by

$$MAF_{snp} = \frac{R_2}{R_1 + R_2} \tag{2.17}$$

where R_1 is the number of reads carrying the most frequent base, and R_2 is the number of reads carrying the second most frequent base at a particular position.

2.4.9 Gene-based and pathway analysis

Gene-based statistic

The gene-based analysis involves integrating the SNP statistics into a single statistical value for each gene, allowing the effects of the genes to be compared and studied. To compute the gene-based statistics, all SNPs located within 50 kb of a gene were considered, and gene regions with a SNP density of more than 50 SNPs/Mb were analysed to ensure a satisfactory coverage. The statistical value for a gene was calculated by Sidak's combination method (Peng et al., 2010) given by

$$P_{gene} = 1 - (1 - P_{snp})^n \tag{2.18}$$

where P_{snp} is the P-value of the best SNP and n is the number of SNPs in the gene. This gene-based approach using the best SNP statistics have been described (Wang et al., 2007; Torkamani et al., 2008; Baranzini et al., 2009). This analysis was performed to show the associations at previous candidate genes in Chapters 4 and 5, as well as the gene-based analysis in Chapter 6.

Identifying shared susceptibility

The evidence for shared susceptibility was tested by looking for over-representation of genes significantly associated with both diseases. Chapter 6 describes the efforts to identify potential shared genetic components between tuberculosis and leprosy, as well as Crohn's disease which is an inflammatory bowel disease with a possible mycobacterial cause. This analysis was performed by the permutation and parametric methods on the gene-based statistics.

The permutation analysis was performed by first calculating the proportion of genes associated in both diseases in the study, at two thresholds of $P < 0.05$ and $P < 0.01$. This gave the observed value for the proportion of genes associated with both diseases (double positive genes). This was followed randomizing the gene labels 100,000 times, each time calculating the proportion of double positive genes to obtain an empirical distribution. The observed value was then compared to the empirical distribution to give the permutation P-value

$$P_{permute} = P(D > D_{obs}) \tag{2.19}$$

where D represents the proportion values in the empirical distribution and D_{obs} represents the observed value.

The parametric analysis was performed by comparing the counts of double positive genes against a hypergeometric probability distribution. The probability of observing k double positive genes is given by

$$P_{hypergeo} = 1 - \sum_{i=0}^{k} \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}} \tag{2.20}$$

where N is the total number of genes in comparison, m is the number of associated genes in one disease and n the number of associated genes in the other disease respectively.

Gene annotation analysis

The genes analysed in Chapter 6 were annotated in reference to the Gene Ontology project, a major bioinformatic initiative which aims to standardise the description of gene products by providing a controlled set of vocabularies (Ashburner et al., 2000). The annotation was performed by the Gene Ontology Enrichment Analysis and Visualization Tool (GOrilla) (Eden et al., 2009) for the three main categories of terms: Biological Processes (BP), Molecular Functions (MF) and Cellular Components (CC). It determines whether a set of genes of a biological theme may be over-represented in the dataset, by comparing the significantly associated genes at $P < 0.001$ compared to the background list containing the genes in the study.

Pathway analysis

The pathway analysis in Chapter 6 was performed with the Ingenuity Pathway Analysis software (Ingenuity System Incoporation, USA) using a ranked list of genes with their P-values. It tests for an over-representation of significantly associated genes in a pathway. The canonical network was constructed by connecting the 25 most associated genes at $P < 0.001$ in the aggregate mycobacterial susceptibility analysis, with the unsupervised addition of several key molecules identified by the software.

2.4.10 Population differentiation

The fixation index F_{ST} is a measure of population differentiation based on frequencies of genetic variants (Wright, 1921, 1952). Let p_k denote the allele frequency of the k^{th} population and p denote the mean allele frequency of all the n included populations, the F_{ST} statistic is given by the ratio of the variances as

$$F_{ST} = \frac{\sum_{i=1}^{k}(p_k - p)^2}{np(1-p)} \tag{2.21}$$

Population differentiation can arise due to local adaptation in geographically separated populations, and is used as an indicator for recent natural selection. However, it should be noted that it can also occur upon genetic drift and demographic history, especially population bottleneck and migration.

2.4.11 Analytical softwares

The genotypes for the GeneChip Human Mapping 500K Array Set were called using the CHIAMO algorithm (The Wellcome Trust Case Control Consortium, 2007), and genotypes for the HumanCVD Genotyping BeadChip were called using the BeadStudio Genotyping Module (Illumina, USA). The enumeration of genotype and allele counts, quality control measures and association tests of significance were conducted with PLINK v1.06 (Purcell et al., 2007), SNPTEST (Marchini et al., 2007) and the R Project for Statistical Computing (http://www.r-project.org/). Haplotype reconstruction and analysis were performed with Haploview v4.1 (Barrett et al., 2005). The formatting and alignment of sequencing reads were performed with the Perl Programming Language v5.10.0 (www.perl.org) and Novoalign 2.05 (www.novocraft.com), and variant calling was performed with VarScan v1.02 (Koboldt et al., 2009).

Chapter 3

WTCCC tuberculosis study

3.1 Introduction

Genome-wide association (GWA) studies offer a powerful approach to identify genetic variants that contribute to common traits. This methodology relies on the collection of unrelated samples from a population, followed by high density genotyping of polymorphic markers and testing for association with a given phenotype. This genome-wide approach is built upon the *common disease-common variant* hypothesis, which suggests that the genetic component of a common disease can be accounted by common genetic variants (Lander, 1996; Risch and Merikangas, 1996). The presence of LD allows a genome-wide survey of variants by genotyping a selected set of tagging markers (Gabriel et al., 2002). The WTCCC was established in 2005 to utilise this approach to identify genetic variants in common diseases, with an ultimate goal to improve our understanding of common disease aetiology. It has successfully identified novel susceptibility loci to seven common diseases (The Wellcome Trust Case Control Consortium, 2007), and this landmark study was followed by a plethora of various GWA studies on other common diseases (Goldstein, 2009; Hindorff et al., 2009). This

chapter describes the effort to investigate the genetic susceptibility to tuberculosis as part of the WTCCC main project, using similar experimental design and statistical methods.

3.2 Cohorts and study design

This study was conducted with sample cohorts from The Gambia, Bissau, Conakry and Malawi. The sample cohorts have been described in Chapter 2. Primary genotyping and analysis was performed with the Gambian cohort, which consists of 1,498 pulmonary cases and 1,496 controls with a male proportion of 58.6%. The largest ethnic group is Mandinka (31.7%), followed by Jola (16.8%), Fula (16.8%) and Wollof (12.8%). HIV-I positive individuals were excluded from the analysis. Replication studies were performed with samples from Bissau (138 cases and 116 controls), Conakry (178 cases and 167 controls) and Malawi (320 cases and 1,192 controls).

Whole genome amplified samples were genotyped in the primary study with the GeneChip 500K Mapping Array Set (Affymetrix, USA) comprising 500,568 SNPs randomly distributed across the genome. These SNPs were estimated to capture approximately 41% of all common variations in the African Yoruban population in the International HapMap Project (The International HapMap Consortium, 2003; Barrett and Cardon, 2006). Genetic variants showing evidence of association with $P < 1.0 \times 10^{-4}$ in the primary study were carried forward for replication, and were genotyped by the Sequenom primer extension assay (Gabriel and Ziaugra, 2004) in the Bissau-Guinean, Conakry-Guinean and Malawian cohorts. Subsequent resequencing was performed on 55 Gambian samples with the Genome Analyzer IIx (Illumina, USA), using a reversible terminator based sequencing technology. The resequencing reads

were aligned to the reference sequence with Novoalign (http://www.novocraft.com/), and the variants were called by Varscan (Koboldt et al., 2009).

3.3 Quality control

Genotypes were called using the CHIAMO algorithm (The Wellcome Trust Case Control Consortium, 2007) with *a posteriori* call probability threshold of 90%, which resulted in a discordance rate of 0.48% based on approximately 50 million duplicate genotypes. Quality control filters were applied to minimise the chance of spurious associations.

3.3.1 Sample filters

Samples with a low call rate ($< 92.5\%$) or an extreme overall heterozygosity (< 0.24 or > 0.34) were excluded (Figure 3.1). As this study aimed to test association in unrelated individuals, samples showing evidence of duplication (IBS $> 98\%$) or cryptic relatedness (IBS $> 85\%$) were excluded. Finally, MDS analysis was performed on these samples with the reference samples from the HapMap populations (The International HapMap Consortium, 2003, 2005). Samples showing outlying ancestry were excluded as in Figure 3.2. This left a total of 2,704 individuals (1,320 cases and 1,384 controls) for analysis (Table 3.1).

3.3.2 SNP filters

Single nucleotide polymorphisms with a low call rate ($< 95\%$) or an extreme deviation from the Hardy-Weinberg equilibrium ($P < 5.7 \times 10^{-7}$) were excluded. In addition, SNPs with a low minor allele frequency ($< 1\%$) were excluded since their calling is less accurate compared to polymorphic SNPs. A total of 405,266 SNPs were available

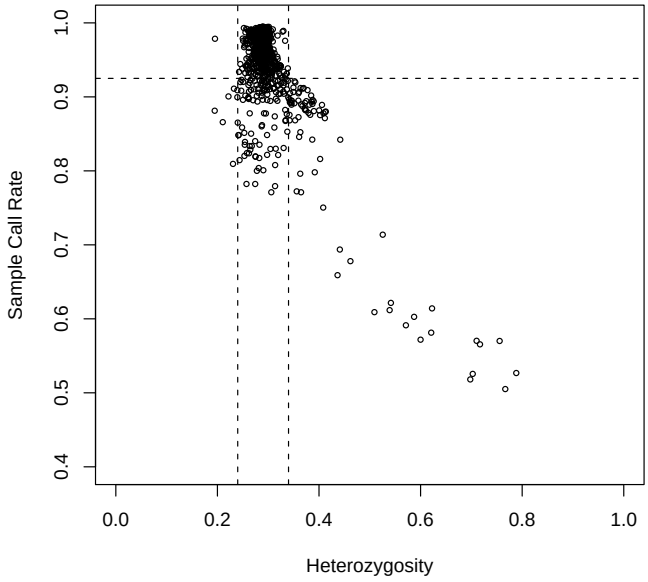


Figure 3.1: Call rate and heterozygosity per sample. Dotted lines delimit the thresholds for samples exclusion

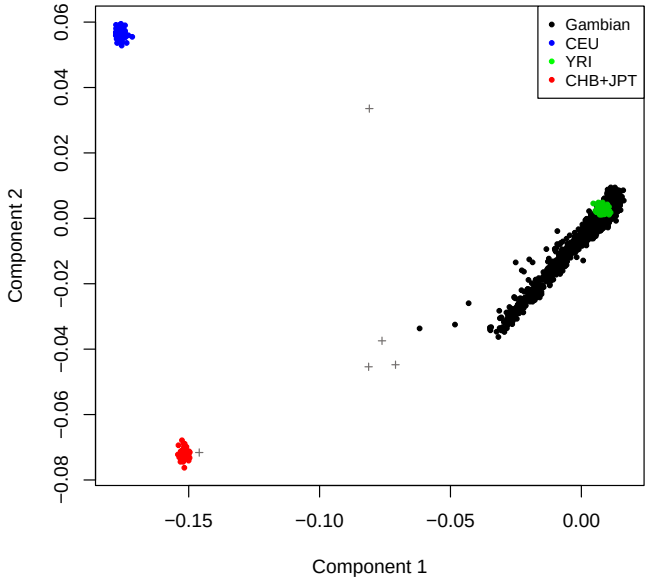


Figure 3.2: MDS plot of the Gambian and the HapMap populations. Crosses denote samples excluded due to outlying ancestry

Exclusion criteria	No of SNPs	No of samples
Sample filters		
Sample call rate ($< 92.5\%$)	—	139
Heterozygosity (< 0.24 or > 0.34)	—	5
Duplication (IBS $> 98\%$)	—	98
Relatedness (IBS $> 85\%$)	—	43
Outlying ancestry	—	5
Retained for analysis	—	*2,704*
SNPs filters		
SNP call rate ($< 95\%$)	66,753	—
Hardy-Weinberg equilibrium ($P < 5.7 \times 10^{-7}$)	4,412	—
Minor allele frequency ($< 1\%$)	24,177	—
Retained for analysis	*405,226*	—

Table 3.1: Samples and SNPs exclusion criteria

for analysis (Table 3.1). Cluster plots of all SNPs showing evidence of association with $P < 1.0 \times 10^{-4}$ were visually inspected for genotype calling quality, and SNPs with poorly separating clusters were excluded from further analysis.

3.4 Population structure

Multidimensional scaling (MDS) was performed to study the genetic structure of the populations. As illustrated in Figure 3.2, the genetic structure of the Gambian population is distinct from the European and Asian populations, and is more similar to the African Yoruba population. However, significant differences can be observed among the Gambian individuals, as indicated by the non-discrete clustering of individuals in the MDS plot. Additional MDS analysis of the different ethnic groups also showed significant population structure (Figure 3.3), with the first component separating the Fula and the second component separating the Jola from the other ethnic groups.

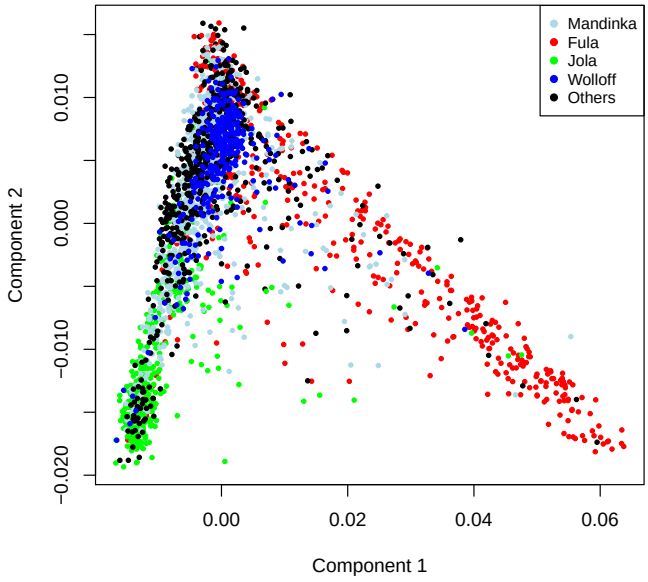


Figure 3.3: MDS plot of the Gambian ethnic groups

3.5 Primary association results

Two statistical tests were used for the primary analysis. These included the trend test which tests for the additive effect of an allele on disease risk, and the genotypic test which compares the genotype frequencies between cases and controls. Both tests were adjusted for three axes of MDS components in a logistic regression model. This reduced the over-dispersion (λ) from 1.13 to 1.05 (Figure 3.4).

A total of 56 SNPs in 38 discrete regions were associated with $P < 1.0 \times 10^{-4}$ in either the trend (P_{trend}) or genotypic (P_{geno}) test in the primary analysis (Figure 3.5, Table 3.2). The strongest evidence of association was observed with SNP rs2446890 in the region flanking the gene *CADM1*, with $P_{geno} = 2.0 \times 10^{-6}$. The trend test statistics was $P = 7.9 \times 10^{-6}$ with an odds ratio of 0.78 (95% CI=0.70-0.87) for the minor G allele. This association was supported by four other nearby SNPs showing evidence of association (Table 3.2). Apart from the *CADM1* locus, evidence of association was

also observed in markers flanking *CADM2* and *CADM3*, two other members of the cell adhesion molecule (CADM) family, as well as *MC3R* and *CD40LG* which have been reported to associate with tuberculosis susceptibility (Campbell et al., 2003; Cooke et al., 2008).

3.6 Replication results

To replicate these findings, SNPs with $P < 1.0 \times 10^{-4}$ in the primary analysis were carried forward for replication in samples from the West African countries of Bissau and Conakry, and the Southeast African country of Malawi (Table 3.3). Significant associations were observed in three SNPs flanking the *CADM1* region in the Conakry-Guinean population, with rs1563899 being most significant with $P_{geno} = 0.002$ (Table 3.3). There was also a consistent association of rs2446890, the most significant SNP in the primary study, in the Conakry-Guinean cohort with $P_{geno} = 0.007$ (Table 3.3). Marginal associations were observed for SNPs at *CADM3* (rs12057331, $P_{geno} = 0.032$ in Malawi), *DEFB112* ($P_{trend} = 0.020$ in Malawi) and *TXNDC10* (rs507533, $P_{geno} = 0.044$), although none of these SNPs showed convincing evidence of replication across populations. No evidence of consistent association was observed in SNPs flanking *CADM2*, *MC3R* or *CD40LG* in the replication cohorts.

3.7 The *CADM1* locus

3.7.1 Results of primary and replication studies

As above-mentioned, the SNP rs2446890 near the *CADM1* gene showed evidence of association in the primary Gambian study ($P_{geno} = 2.0 \times 10^{-6}$). The trend test

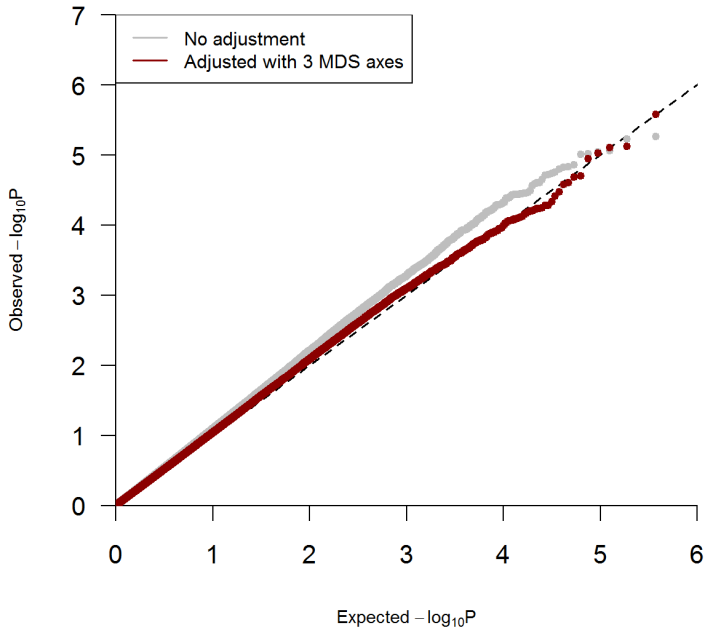


Figure 3.4: Quantile-quantile plot of the trend test statistics in the WTCCC tuberculosis study

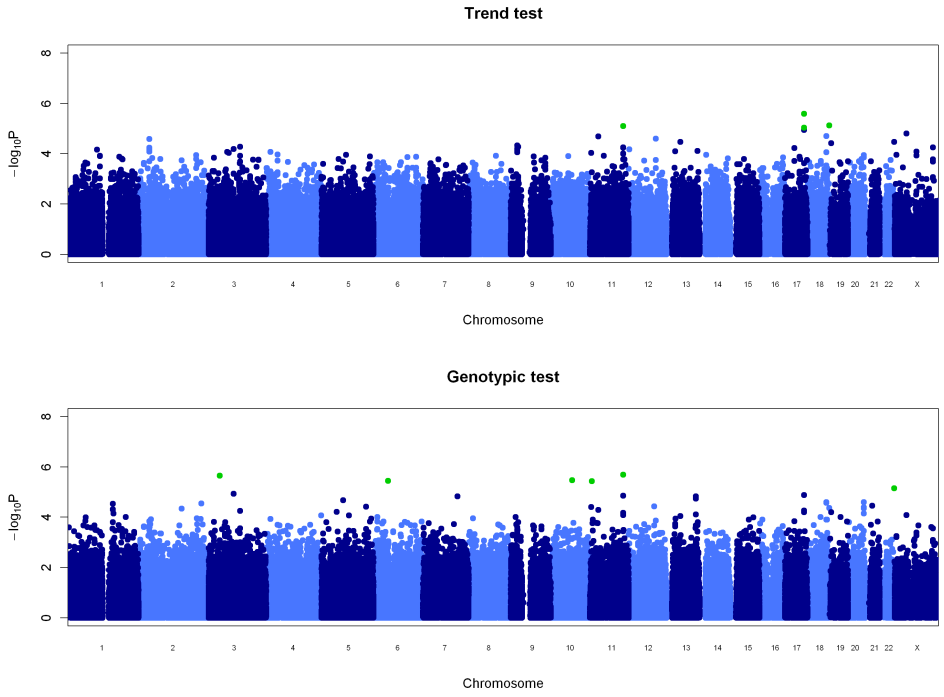


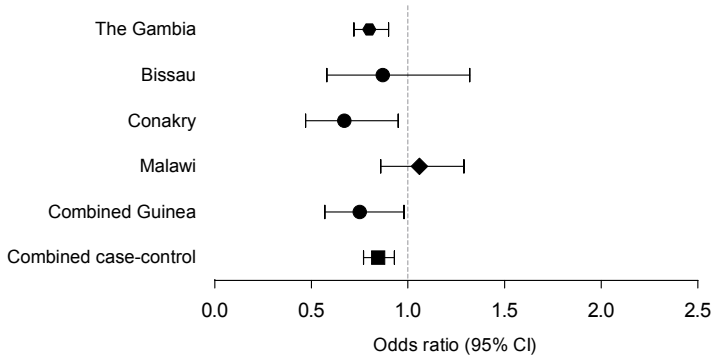
Figure 3.5: Manhattan plot of the WTCCC tuberculosis study. SNPs with $P < 1.0 \times 10^{-5}$ are highlighted in green

SNP	Chr	Position	Gene	Minor	MAF (case)	MAF (ctrl)	OR	95% CI	P_{trend}	P_{geno}
rs6676334	1	101372143	∼*EDG1*	C	0.467	0.525	0.80	0.72-0.89	7.0×10^{-5}	3.3×10^{-4}
rs12068892	1	155895443	∼*CADM3*	T	0.267	0.291	0.89	0.79-1.00	0.057	4.8×10^{-5}
rs12057331	1	155903630	∼*CADM3*	C	0.265	0.290	0.88	0.78-0.99	0.035	2.9×10^{-5}
rs16841729	1	155904171	∼*CADM3*	A	0.263	0.289	0.88	0.78-0.99	0.034	4.9×10^{-5}
rs10753758	1	158820989	*NOS1AP*	A	0.147	0.183	0.78	0.67-0.90	5.6×10^{-4}	7.1×10^{-5}
rs2160394	2	37411057	*PRKD3*	T	0.099	0.137	0.71	0.60-0.84	6.2×10^{-5}	2.5×10^{-4}
rs11124576	2	37430367	*PRKD3*	G	0.097	0.135	0.71	0.60-0.84	5.8×10^{-5}	2.2×10^{-4}
rs2268987	2	37444826	*PRKD3*	C	0.169	0.217	0.74	0.65-0.86	2.6×10^{-5}	1.4×10^{-4}
rs1979038	2	149044630	∼*MDB5*	G	0.179	0.147	1.32	1.14-1.52	1.8×10^{-4}	4.6×10^{-5}
rs6780349	3	38530182	*ENDOGL1*	C	0.431	0.399	1.14	1.02-1.27	0.018	2.2×10^{-6}
rs9822541	3	64317269	∼*ADAMTS9*	A	0.177	0.140	1.34	1.16-1.56	8.5×10^{-5}	2.6×10^{-4}
rs2324979	3	86202099	∼*CADM2*	C	0.211	0.257	0.77	0.68-0.88	6.5×10^{-5}	1.1×10^{-5}
rs1437237	3	109135362	∼*CD47*	T	0.260	0.213	1.30	1.15-1.48	5.2×10^{-5}	2.7×10^{-4}
rs6437758	3	109146525	∼*CD47*	A	0.137	0.106	1.39	1.17-1.64	1.2×10^{-4}	5.5×10^{-5}
rs4074082	4	14910135	∼*C1QTNF7*	C	0.186	0.230	0.77	0.67-0.88	8.5×10^{-5}	1.1×10^{-4}
rs6553232	4	190085264	-	C	0.129	0.141	0.91	0.78-1.07	0.250	8.5×10^{-5}
rs6880850	5	52801726	∼*FST*	T	0.254	0.291	0.84	0.75-0.95	0.005	6.1×10^{-5}
rs1917753	5	75160395	∼*LOC441087*	T	0.142	0.116	1.27	1.08-1.50	0.003	2.1×10^{-5}
rs261230	5	95932477	∼*CAST*	C	0.450	0.497	0.83	0.74-0.92	5.2×10^{-4}	8.4×10^{-5}
rs6880985	5	154761413	-	T	0.123	0.137	0.87	0.74-1.02	0.087	3.8×10^{-5}
rs6902413	6	50189261	∼*DEFB112*	C	0.383	0.416	0.85	0.76-0.95	0.004	3.5×10^{-6}
rs7790895	7	119896238	*KCND2*	G	0.160	0.135	1.20	1.04-1.40	0.015	1.4×10^{-5}
rs2123338	9	21684039	∼*MTAP*	G	0.177	0.138	1.34	1.16-1.56	7.5×10^{-5}	3.8×10^{-4}
rs1889680	9	21685460	∼*MTAP*	T	0.178	0.138	1.35	1.17-1.57	4.6×10^{-5}	2.5×10^{-4}
rs12380505	9	21685893	∼*MTAP*	A	0.170	0.131	1.35	1.16-1.57	8.7×10^{-6}	4.2×10^{-4}
rs10811584	9	21691005	∼*MTAP*	A	0.234	0.188	1.30	1.14-1.48	7.6×10^{-5}	3.9×10^{-4}
rs6475552	9	21691674	∼*MTAP*	C	0.232	0.186	1.31	1.15-1.49	6.3×10^{-5}	3.3×10^{-4}
rs7038176	9	24966322	-	T	0.498	0.450	1.25	1.12-1.39	5.2×10^{-5}	1.5×10^{-4}
rs1245574	10	73478473	∼*SPOCK2*	A	0.280	0.310	0.85	0.76-0.96	0.010	3.4×10^{-6}
rs7105156	11	4672522	*OR51E2*	A	0.319	0.267	1.27	1.13-1.43	9.3×10^{-5}	3.9×10^{-5}
rs2939233	11	28916116	-	T	0.292	0.241	1.31	1.16-1.48	2.0×10^{-5}	5.1×10^{-5}
rs1460911	11	114937935	∼*CADM1*	T	0.433	0.484	0.81	0.72-0.90	9.9×10^{-5}	6.6×10^{-5}
rs1563900	11	114943631	∼*CADM1*	C	0.432	0.483	0.80	0.72-0.90	1.0×10^{-4}	1.4×10^{-5}
rs1563899	11	114943656	∼*CADM1*	T	0.432	0.484	0.80	0.71-0.89	5.7×10^{-5}	1.3×10^{-5}
rs2446890	11	114952656	∼*CADM1*	G	0.427	0.487	0.78	0.70-0.87	7.9×10^{-6}	2.0×10^{-6}
rs2515327	11	114955759	∼*CADM1*	C	0.426	0.481	0.80	0.72-0.89	5.7×10^{-5}	8.3×10^{-5}
rs216009	12	2592794	*CACNA1C*	C	0.146	0.193	0.75	0.65-0.86	6.8×10^{-5}	3.5×10^{-4}
rs2769431	12	94306426	∼*METAP2*	C	0.315	0.264	1.29	1.15-1.45	2.5×10^{-5}	1.3×10^{-4}
rs4270069	13	47093766	-	A	0.037	0.063	0.59	0.46-0.76	3.3×10^{-5}	9.1×10^{-5}
rs7985451	13	105631206	-	G	0.431	0.379	1.25	1.12-1.39	7.7×10^{-5}	2.0×10^{-4}
rs853217	17	32929668	*DUSP14*	T	0.523	0.469	1.25	1.12-1.39	6.0×10^{-5}	1.4×10^{-4}
rs17176195	17	66215798	-	C	0.429	0.368	1.28	1.15-1.43	9.3×10^{-6}	5.4×10^{-5}
rs7213806	17	66235577	-	A	0.285	0.341	0.76	0.67-0.85	2.6×10^{-6}	1.3×10^{-5}
rs1029754	17	66236699	-	C	0.279	0.332	0.77	0.69-0.87	1.1×10^{-5}	6.3×10^{-5}
rs507533	18	64472145	∼*TXNDC10*	G	0.141	0.161	0.84	0.72-0.98	0.023	2.6×10^{-5}
rs17080156	18	64868442	*C18orf14*	A	0.200	0.240	0.77	0.67-0.88	8.6×10^{-5}	1.2×10^{-4}
rs17080166	18	64868782	*C18orf14*	C	0.187	0.229	0.74	0.64-0.85	2.0×10^{-5}	2.5×10^{-5}
rs4799232	18	74706295	∼*SALL3*	C	0.420	0.367	1.29	1.15-1.43	7.4×10^{-6}	4.3×10^{-5}
rs17363814	19	4795633	*M6PRBP1*	A	0.150	0.111	1.39	1.19-1.63	3.8×10^{-5}	6.4×10^{-5}
rs6092276	20	54255495	∼*MC3R*	G	0.452	0.503	0.83	0.75-0.93	7.1×10^{-4}	4.1×10^{-5}
rs3746619	20	54257212	∼*MC3R*	C	0.460	0.517	0.81	0.73-0.91	1.4×10^{-4}	4.4×10^{-5}
rs3827103	20	54257436	∼*MC3R*	T	0.477	0.424	1.24	1.11-1.38	1.1×10^{-4}	2.5×10^{-5}
rs8184921	21	21294594	*NCAM2*	T	0.346	0.385	0.82	0.74-0.92	8.2×10^{-4}	3.5×10^{-5}
rs7357969	X	641188	∼*SHOX*	T	0.044	0.076	0.62	0.49-0.78	3.4×10^{-5}	7.0×10^{-6}
rs2064070	X	43364936	*MAOA/B*	T	0.381	0.316	1.22	1.11-1.33	1.6×10^{-5}	8.3×10^{-5}
rs3092947	X	135453672	∼*CD40LG*	C	0.297	0.365	0.83	0.76-0.91	5.6×10^{-5}	2.8×10^{-4}

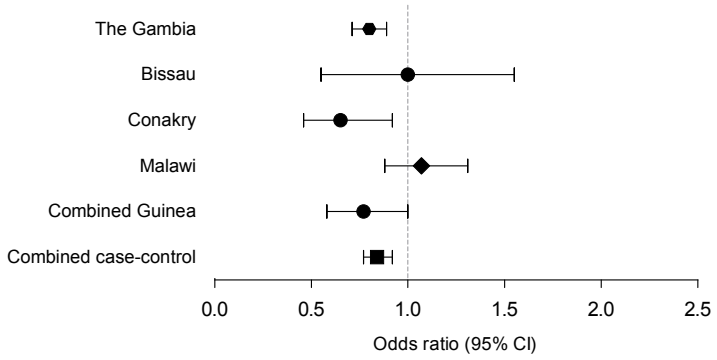
Table 3.2: Primary association statistics of SNPs with $P < 1.0 \times 10^{-4}$

SNP	Gene	Gambia	Bissau			Conakry			Malawi		
		OR	OR (95% CI)	P_{trend}	P_{geno}	OR (95% CI)	P_{trend}	P_{geno}	OR (95% CI)	P_{trend}	P_{geno}
rs6676334	∼*EDG1*	0.80	1.09 (0.73-1.64)	0.660	0.042	0.98 (0.71-1.36)	0.918	0.772	0.87 (0.70-1.08)	0.212	0.426
rs12068892	∼*CADM3*	0.89	1.11 (0.68-1.82)	0.665	0.898	1.03 (0.66-1.59)	0.911	0.554	0.97 (0.72-1.31)	0.847	0.079
rs12057331	∼*CADM3*	0.88	1.28 (0.78-2.08)	0.323	0.343	1.13 (0.77-1.64)	0.541	0.092	0.85 (0.67-1.08)	0.189	0.032
rs16841729	∼*CADM3*	0.88	1.33 (0.82-2.16)	0.247	0.386	1.15 (0.74-1.79)	0.545	0.174	0.87 (0.68-1.12)	0.269	0.097
rs10753758	*NOS1AP*	0.78	1.08 (0.63-1.85)	0.772	0.669	1.05 (0.69-1.58)	0.823	0.974	1.00 (0.78-1.29)	0.981	0.192
rs2160394	*PRKD3*	0.71	2.81 (1.10-7.15)	0.023	0.023	1.27 (0.76-2.13)	0.363	0.381	1.23 (0.81-1.85)	0.334	0.432
rs11124576	*PRKD3*	0.71	1.55 (0.75-3.21)	0.230	0.230	1.38 (0.86-2.22)	0.178	0.122	0.96 (0.67-1.39)	0.835	0.827
rs2268987	*PRKD3*	0.74	0.84 (0.42-1.65)	0.607	0.607	1.19 (0.68-2.07)	0.537	0.551	1.02 (0.76-1.37)	0.886	0.966
rs1979038	∼*MDB5*	1.32	0.99 (0.58-1.70)	0.977	0.637	1.14 (0.71-1.83)	0.580	0.840	0.97 (0.76-1.22)	0.776	0.950
rs6780349	*ENDOGL1*	1.14	0.84 (0.54-1.30)	0.434	0.353	0.84 (0.59-1.20)	0.344	0.639	0.92 (0.76-1.13)	0.446	0.731
rs9822541	∼*ADAMTS9*	1.34	1.48 (0.90-2.45)	0.121	0.279	0.82 (0.52-1.28)	0.383	0.555	1.04 (0.76-1.42)	0.792	0.731
rs2324979	∼*CADM2*	0.77	1.23 (0.79-1.92)	0.365	0.643	0.93 (0.65-1.33)	0.699	0.728	Assay failed		
rs1437237	∼*CD47*	1.30	1.32 (0.83-2.07)	0.235	0.431	1.07 (0.73-1.57)	0.723	0.517	0.96 (0.76-1.21)	0.747	0.862
rs6437758	∼*CD47*	1.39	0.91 (0.46-1.81)	0.788	0.662	0.95 (0.58-1.55)	0.824	0.589	0.88 (0.65-1.18)	0.387	0.417
rs4074082	∼*C1QTNF7*	0.77	Not genotyped			Not genotyped			Not genotyped		
rs6553232	-	0.91	1.88 (1.07-3.31)	0.025	0.050	1.09 (0.7-1.71)	0.703	0.731	1.00 (0.76-1.33)	0.983	0.212
rs6880850	∼*FST*	0.84	0.84 (0.54-1.29)	0.426	0.729	0.85 (0.60-1.22)	0.375	0.212	1.28 (1.03-1.59)	0.030	0.094
rs1917753	∼*LOC441087*	1.27	0.83 (0.45-1.55)	0.559	0.961	1.02 (0.58-1.80)	0.947	0.994	0.98 (0.75-1.28)	0.899	0.587
rs261230	∼*CAST*	0.83	0.72 (0.47-1.10)	0.131	0.287	0.90 (0.62-1.3)	0.575	0.851	1.06 (0.87-1.29)	0.574	0.501
rs6880985	-	0.87	1.57 (0.84-2.91)	0.149	0.347	0.75 (0.47-1.21)	0.237	0.278	1.03 (0.73-1.44)	0.882	0.421
rs6902413	∼*DEFB112*	0.85	1.51 (0.99-2.32)	0.054	0.145	0.87 (0.61-1.24)	0.452	0.697	0.78 (0.62-0.96)	0.020	0.061
rs7790895	*KCND2*	1.20	1.05 (0.56-1.98)	0.882	0.537	0.52 (0.28-0.95)	0.031	0.074	0.92 (0.70-1.22)	0.572	0.327
rs2123338	∼*MTAP*	1.34	0.84 (0.47-1.52)	0.574	0.785	0.91 (0.48-1.75)	0.788	0.964	0.84 (0.62-1.13)	0.239	0.428
rs1889680	∼*MTAP*	1.35	0.70 (0.40-1.22)	0.211	0.446	0.99 (0.55-1.77)	0.971	0.998	0.85 (0.63-1.15)	0.297	0.433
rs12380505	∼*MTAP*	1.35	0.77 (0.45-1.29)	0.316	0.535	1.18 (0.65-2.14)	0.580	0.832	0.92 (0.69-1.23)	0.557	0.310
rs10811584	∼*MTAP*	1.30	0.60 (0.32-1.09)	0.960	0.997	0.83 (0.46-1.50)	0.538	0.767	1.04 (0.76-1.43)	0.797	0.216
rs6475552	∼*MTAP*	1.31	1.01 (0.59-1.74)	0.960	0.997	0.83 (0.47-1.47)	0.531	0.815	0.94 (0.73-1.22)	0.667	0.863
rs7038176	-	1.25	0.80 (0.51-1.24)	0.311	0.598	0.50 (0.32-0.78)	0.001	0.003	1.07 (0.84-1.37)	0.578	0.399
rs1245574	∼*SPOCK2*	0.85	1.21 (0.78-1.88)	0.400	0.654	1.02 (0.68-1.53)	0.927	0.335	1.00 (0.82-1.23)	0.965	0.329
rs7105156	*OR51E2*	1.27	1.02 (0.62-1.68)	0.935	0.353	1.05 (0.56-1.95)	0.884	0.966	0.84 (0.63-1.11)	0.217	0.326
rs2939233	-	1.31	1.25 (0.80-1.95)	0.324	0.585	0.95 (0.65-1.38)	0.777	0.688	1.24 (0.99-1.57)	0.067	0.186
rs1460911	∼*CADM1*	0.81	1.07 (0.74-1.54)	0.716	0.290	0.74 (0.50-1.10)	0.135	0.083	1.21 (0.97-1.51)	0.095	0.189
rs1563900	∼*CADM1*	0.80	0.87 (0.58-1.32)	0.519	0.752	0.67 (0.47-0.95)	0.024	0.003	1.06 (0.86-1.29)	0.606	0.256
rs1563899	∼*CADM1*	0.80	1.00 (0.65-1.55)	1.000	0.653	0.65 (0.46-0.92)	0.014	0.002	1.07 (0.88-1.31)	0.490	0.390
rs2446890	∼*CADM1*	0.78	0.94 (0.62-1.41)	0.753	0.826	0.70 (0.50-0.99)	0.043	0.007	1.06 (0.87-1.30)	0.539	0.356
rs2515327	∼*CADM1*	0.80	1.25 (0.85-1.83)	0.262	0.532	0.81 (0.58-1.12)	0.204	0.164	1.05 (0.85-1.30)	0.645	0.822
rs216009	*CACNA1C*	0.75	1.00 (0.61-1.62)	0.987	0.570	0.94 (0.63-1.41)	0.783	0.089	1.14 (0.88-1.46)	0.326	0.506
rs2769431	∼*METAP2*	1.29	1.34 (0.75-2.37)	0.313	0.329	0.81 (0.43-1.52)	0.509	0.691	1.19 (0.96-1.48)	0.117	0.282
rs4270069	-	0.59	1.75 (0.67-4.56)	0.246	0.246	1.52 (0.60-3.86)	0.369	0.369	1.46 (0.94-2.27)	0.100	0.152
rs7985451	-	1.25	1.21 (0.82-1.77)	0.334	0.481	1.03 (0.73-1.46)	0.863	0.773	1.02 (0.83-1.26)	0.821	0.957
rs853217	*DUSP14*	1.25	1.02 (0.68-1.52)	0.934	0.933	0.85 (0.61-1.18)	0.321	0.604	1.01 (0.81-1.26)	0.937	0.906
rs17176195	-	1.28	0.91 (0.60-1.39)	0.670	0.590	1.25 (0.86-1.81)	0.238	0.454	0.84 (0.67-1.04)	0.113	0.271
rs7213806	-	0.76	1.63 (0.91-2.93)	0.098	0.253	1.03 (0.71-1.49)	0.886	0.525	1.11 (0.90-1.36)	0.340	0.476
rs1029754	-	0.77	0.90 (0.55-1.48)	0.672	0.207	1.01 (0.69-1.47)	0.956	0.502	0.83 (0.67-1.03)	0.083	0.167
rs507533	∼*TXNDC10*	0.84	0.70 (0.38-1.26)	0.229	0.384	0.84 (0.5-1.43)	0.532	0.044	0.79 (0.6-1.04)	0.094	0.086
rs17080156	*C18orf14*	0.77	0.63 (0.38-1.05)	0.074	0.180	1.33 (0.86-2.05)	0.201	0.048	1.04 (0.81-1.33)	0.771	0.958
rs17080166	*C18orf14*	0.74	0.75 (0.37-1.50)	0.416	0.704	0.85 (0.51-1.42)	0.535	0.342	1.07 (0.80-1.44)	0.651	0.724
rs4799232	∼*SALL3*	1.29	1.29 (0.84-1.97)	0.246	0.248	1.01 (0.72-1.40)	0.966	0.925	0.96 (0.77-1.20)	0.736	0.342
rs17363814	*M6PRBP1*	1.39	0.99 (0.51-1.93)	0.984	0.165	0.51 (0.29-0.90)	0.019	0.035	0.91 (0.60-1.36)	0.638	0.888
rs6092276	∼*MC3R*	0.83	0.89 (0.56-1.42)	0.637	0.856	1.02 (0.71-1.47)	0.915	0.981	1.25 (1.03-1.52)	0.026	0.079
rs3746619	∼*MC3R*	0.81	Not genotyped			Not genotyped			Not genotyped		
rs3827103	∼*MC3R*	1.24	1.08 (0.69-1.68)	0.736	0.731	1.00 (0.70-1.43)	0.991	0.920	0.92 (0.75-1.11)	0.382	0.547
rs8184921	*NCAM2*	0.82	1.36 (0.89-2.07)	0.150	0.028	1.42 (0.99-2.03)	0.054	0.135	1.17 (0.94-1.45)	0.163	0.055
rs7357969	∼*SHOX*	0.62	Not genotyped			Not genotyped			Not genotyped		
rs2064070	*MAOA/B*	1.22	0.83 (0.59-1.17)	0.296	0.258	1.27 (0.93-1.73)	0.131	0.317	0.96 (0.80-1.15)	0.630	0.727
rs3092947	∼*CD40LG*	0.83	1.09 (0.71-1.66)	0.687	0.842	1.13 (0.8-1.59)	0.489	0.353	0.98 (0.77-1.25)	0.876	0.848

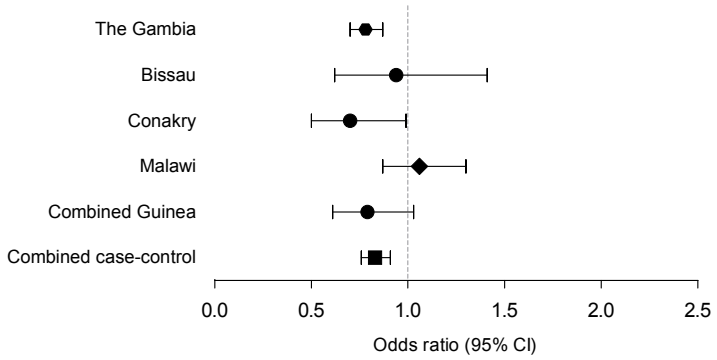
Table 3.3: Association statistics in the replication studies



(a) rs1563900



(b) rs1563899



(c) rs2446890

Figure 3.6: Forest plots showing associations of rs1563900, rs1563899 and rs2446890 in different populations

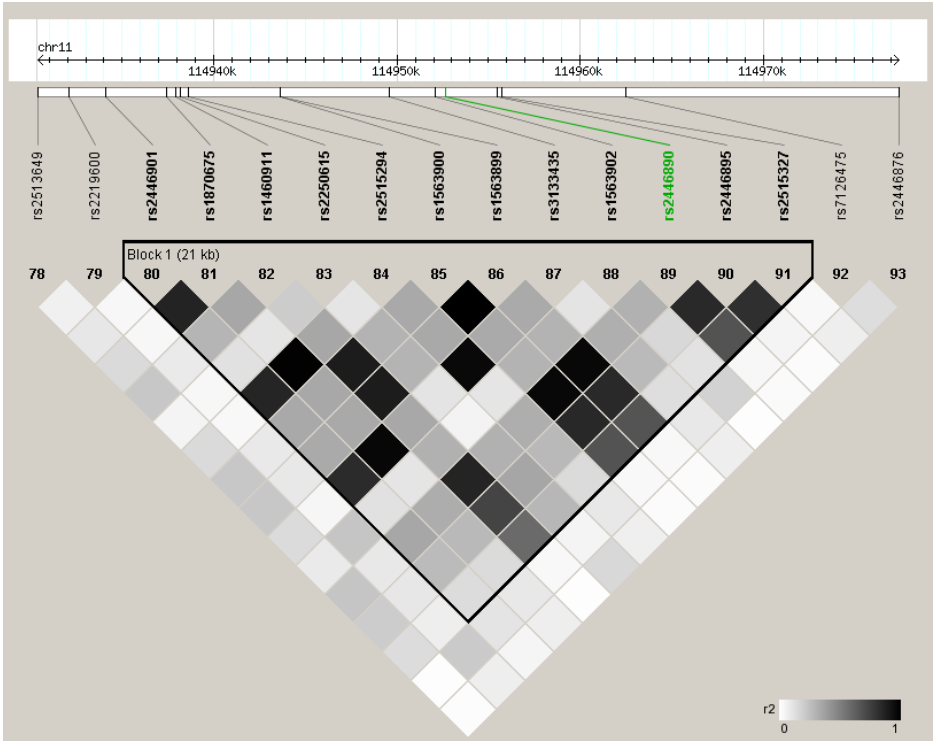


Figure 3.7: LD plot of the *CADM1* locus. The block indicates the minimal region to include all SNPs with $r^2 > 0.2$ with rs2446890

statistic in the Gambian population was $P = 8.0 \times 10^{-6}$ with an odds ratio of 0.78 (95% CI=0.70-0.87) for the minor allele (Table 3.2).

Significant association was also observed in the Conakry-Guinean population in the replication study with $P_{geno} = 0.007$. The trend test statistic was $P = 0.043$ with an odds ratio of 0.70 (95% CI=0.05-0.99) (Table 3.3, Figure 3.6). Similar to the Gambian population, the minor allele was associated with protection against tuberculosis in the Conakry-Guinean population. Although no significant association was observed in the Bissau-Guinean population, the genotypic test statistic for rs2446890 for the combined Guinea populations was significant with $P = 0.036$. The combined association statistic for the West African populations of the Gambian, Bissau-Guinean and Conakry-Guinean is $P = 1.7 \times 10^{-6}$ (OR=0.78, 95% CI=0.71-0.87). However, no evidence of association was observed in the Malawian cohort. The tests of heterogeneity showed a significant difference in the effect size of the Malawian cohort

(P_Q = 0.045, I^2 = 62.7%, 95% CI=0-87.5%) upon addition of this population to the West African populations. Similar association statistics were also observed at the two other SNPs in this locus in the Conakry-Guinean population (rs1563899, $P_{geno} = 0.002$ and rs1563900, $P_{geno} = 0.007$, Table 3.3).

3.7.2 Pattern of linkage disequilibrium

The local LD pattern was studied to dissect the association signal. The peak of association was restricted within a 21 kb LD block (Figure 3.7), which includes the minimal region to include all the SNPs with $r^2 > 0.2$ with rs2446890, the most associated SNP in the experiment. The block is located 72.3 kb away from the 5' end of the *CADM1* gene (Figure 3.10). As the size of the block is much smaller than the distance from the gene, it is more likely that the signal has arisen from variants outside rather than inside *CADM1*. This is supported by the association data, which showed that the main signal was clustered within nearby recombination hotspots, with no convincing association within the gene itself. This suggests that the associated SNPs, if they are real, may exert their effects by modulating long-range *cis*-regulatory elements of *CADM1*.

3.7.3 Resequencing and fine-mapping

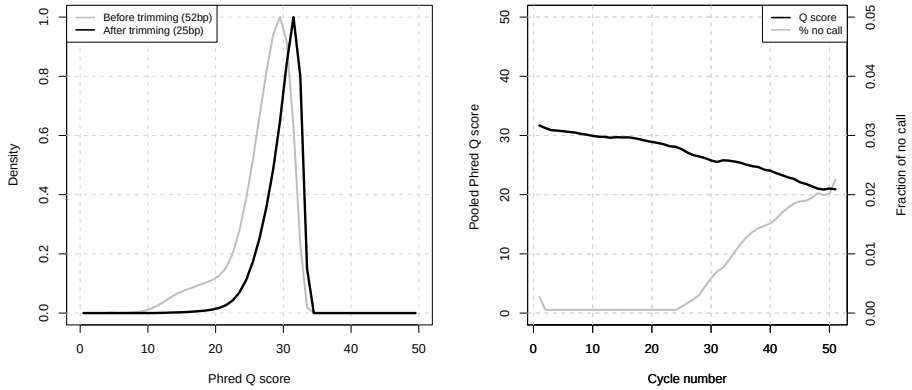
The associations of the SNPs in *CADM1* in the primary and replication studies showed heterogeneous results. Statistically significant associations were observed in the Gambian and Conakry-Guinean populations, but not in the Bissau-Guinean and Malawian populations. Although this can represent a chance finding, the possibility of real associations cannot be excluded. Furthermore, the different LD patterns between the populations can lead to different alleles tagging, resulting in the heterogeneous findings even if a functional variant does exist in this locus. This is because

of the substantial genetic diversities among the African populations (Campbell and Tishkoff, 2008), especially between the Gambian and Malawian populations which are geographically far apart. The work presented in this section was carried out in attempt to resolve this heterogeneity, by finding the causal variant which could be more consistently associated across populations. Genotyping the causal variant can boost the association signal by several orders of magnitude, especially in African populations with low LD or poorly tagging SNPs for the causal variant (Jallow et al., 2009). The aim of this section is to identify the causal variant with a stronger association in the Gambian population which may then be genotyped in the replication cohorts.

Resequencing experiment

This experiment is built on the hypothesis that there is a causal SNP that accounts for the association in this locus. To resequence the region, the 28kb block at the associated region was divided into eight smaller blocks (sized 1.68 to 4.87kb). The region was selected to cover all common variants with $r^2 > 0.2$ with the most associated SNP (rs2446890) in the primary study (Figure 3.7). The template consisted of equal amounts of DNA from 55 Gambian individuals. This gives $> 80\%$ power to detect at least two copies of a variant at a frequency of $> 3\%$. The methods for this experiment were described in Chapter 2.

This resequencing exercise yielded a total of 16.76 million reads of 52 base pairs in length. The first 25 bases of each read were trimmed for analysis (Figure 3.8), since the base quality decreases with more errors further along the Illumina reads (Druley et al., 2009). After quality control filters on reads and alignment qualities (Chapter 2), there were a total of 16.45 million reads (98.2%) uniquely mapped to the reference sequence. The average Phred quality score (Q) of all reads passing quality control



(a) Probabilty density of quality scores (b) Average scores on each position

Figure 3.8: Quality statistics for the sequencing reads

criteria was 29.73 corresponding to a base calling accuracy of 99.89%. The average coverage per haploid chromosome was 133-fold with higher coverage at overlapping region between the PCR blocks (Figure 3.9).

After alignment, a consensus sequence was assembled and variants were called. A total of 128 SNPs with frequency $> 3\%$ were identified, averaging to 4.72 SNPs per kb over the resequenced region. These included all 39 SNPs with a minor allele frequency of $> 5\%$ in the HapMap Yoruba population (The International HapMap Consortium, 2003, 2005). Thirteen insertion-deletion (InDels) polymorphisms were detected.

Fine-mapping experiment

A subset of 29 SNPs was selected for fine-mapping. As it is likely that the causal variant is at least partially correlated with the associated SNPs in the primary study, priority was given to SNPs in LD ($r^2 > 0.5$) with rs2446890 in the HapMap Yoruba population (The International HapMap Consortium, 2003, 2005) and novel SNPs with an estimated frequency close to that of rs2446890 ($20 - 80\%$). Together with

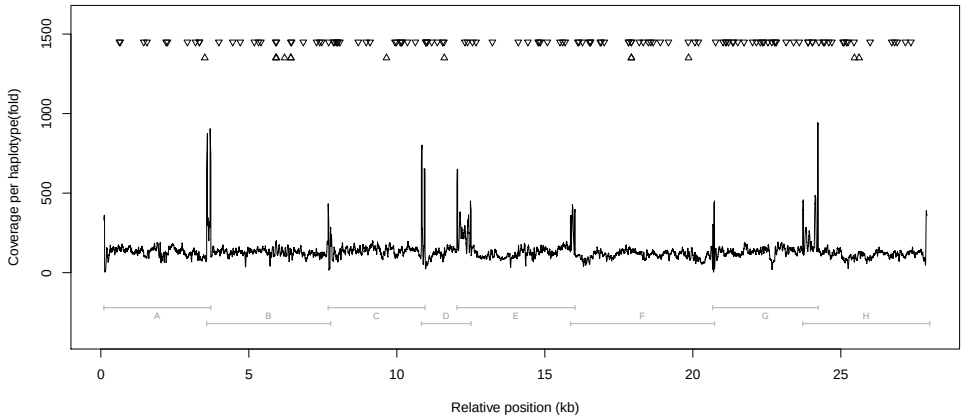


Figure 3.9: Sequence coverage per haplotype over the 28 kb region. Arrows indicate positions of the long-range PCR blocks and markers indicate polymorphic SNPs (∇) and InDels ($\triangle$) called

SNPs genotyped in the microarray, these included 19 out of 21 SNPs with $r^2 >$ 0.2 with rs2446890 in the HapMap Yoruba population (The International HapMap Consortium, 2003, 2005). The analysis showed four SNPs with stronger evidence of association in the trend test (rs2515324, rs979219, rs2515325 and rs2515322, Table 3.4), with the most significant $P = 4.2 \times 10^{-6}$ with the SNP rs2515324 identified in the resequencing experiment. The G allele was minor and associated with protection from tuberculosis (OR=0.79, 95% CI=0.71-0.88). However, the SNP rs2446890 in the original assay remained most significant in the genotypic test (Figure 3.10 and Table 3.4).

3.7.4 Conditional analysis

Conditional analysis was performed to further dissect the signal of association. In addition to the MDS components, the tests of association were performed with correction to best SNP rs2446890 in the Gambian cohort, thus looking at independent effects of the SNPs. However, this analysis did not identify any significant SNP in

SNP	Position	Type	MAF	r^2-rs2446890	OR (95% CI)	P_{trend}	P_{geno}
rs2446901	114934105	Original	0.311	0.32	1.14 (1.01-1.28)	0.007	0.027
rs2446900	114934130	Fine-map	0.478	0.68	0.83 (0.75-0.93)	2.0×10^{-4}	4.0×10^{-4}
rs1870675	114937465	Original	0.279	0.30	1.11 (0.99-1.25)	0.034	0.082
rs1460911	114937935	Original	0.459	0.88	0.81 (0.73-0.91)	9.9×10^{-5}	6.6×10^{-5}
rs2250615	114938189	Original	0.197	0.29	0.82 (0.71-0.93)	0.007	0.016
rs2515294	114938663	Original	0.275	0.29	1.15 (1.02-1.30)	0.009	0.027
rs2515295	114938885	Fine-map	0.198	0.29	0.80 (0.70-0.92)	0.003	0.007
cadm1_7596	114941175	Fine-map	0.222	0.33	0.79 (0.69-0.90)	7.0×10^{-4}	0.001
rs2513641	114942175	Fine-map	0.230	0.30	0.80 (0.70-0.91)	0.002	0.002
rs12279683	114942574	Fine-map	0.140	0.13	0.97 (0.82-1.15)	0.714	0.745
rs1563901	114943595	Fine-map	0.454	0.95	0.81 (0.72-0.90)	7.8×10^{-5}	2.7×10^{-5}
rs1563900	114943631	Original	0.458	0.97	0.80 (0.72-0.90)	1.0×10^{-4}	1.4×10^{-5}
rs1563899	114943656	Original	0.459	0.97	0.80 (0.71-0.89)	5.8×10^{-5}	1.4×10^{-5}
rs979219	114943835	Fine-map	0.435	0.88	0.79 (0.71-0.88)	4.9×10^{-6}	5.2×10^{-6}
rs2446882	114944508	Fine-map	0.463	0.93	0.80 (0.72-0.89)	3.1×10^{-5}	1.9×10^{-5}
rs2515310	114948564	Fine-map	0.280	0.29	1.10 (0.98-1.25)	0.049	0.126
rs2513631	114948991	Fine-map	0.454	0.94	0.80 (0.72-0.89)	4.0×10^{-5}	1.2×10^{-5}
rs2513630	114949165	Fine-map	0.198	0.29	0.80 (0.69-0.91)	0.002	0.005
rs3133434	114949596	Fine-map	0.431	0.87	0.80 (0.72-0.89)	2.9×10^{-5}	1.4×10^{-5}
rs3133435	114949627	Original	0.281	0.31	1.11 (0.99-1.25)	0.043	0.114
rs2515322	114950379	Fine-map	0.435	0.89	0.79 (0.71-0.88)	5.8×10^{-6}	3.3×10^{-6}
rs2515324	114951306	Fine-map	0.461	0.96	0.78 (0.70-0.87)	4.2×10^{-6}	3.1×10^{-6}
rs2515325	114951402	Fine-map	0.458	0.97	0.78 (0.70-0.87)	5.0×10^{-6}	2.1×10^{-6}
rs2515326	114951670	Fine-map	0.456	0.97	0.79 (0.70-0.88)	8.3×10^{-6}	2.6×10^{-6}
rs1008952	114951976	Fine-map	0.233	0.35	0.89 (0.79-1.02)	0.026	0.060
rs1563902	114952141	Original	0.202	0.29	0.93 (0.81-1.05)	0.081	0.061
rs2446890	114952656	Original	0.458	NA	0.78 (0.70-0.87)	8.0×10^{-6}	2.0×10^{-6}
rs922205	114953325	Fine-map	0.474	0.93	0.79 (0.71-0.88)	1.6×10^{-5}	4.1×10^{-6}
rs34243180	114953666	Fine-map	0.201	0.28	0.93 (0.82-1.06)	0.119	0.152
rs2446893	114954248	Fine-map	0.457	0.64	0.79 (0.71-0.88)	3.1×10^{-5}	3.3×10^{-5}
rs11215603	114954500	Fine-map	0.146	0.15	1.05 (0.89-1.24)	0.570	0.668
rs1008950	114954827	Fine-map	0.465	0.62	0.84 (0.76-0.94)	0.002	0.007
rs2446895	114955498	Original	0.407	0.80	0.82 (0.73-0.91)	1.7×10^{-4}	3.6×10^{-4}
rs2513633	114955626	Fine-map	0.200	0.29	0.80 (0.70-0.92)	0.003	0.002
rs2515327	114955759	Original	0.454	0.66	0.80 (0.72-0.89)	5.7×10^{-5}	8.3×10^{-5}
cadm1_22256	114955835	Fine-map	0.033	0.04	0.84 (0.62-1.14)	0.229	0.382
rs2446896	114956016	Fine-map	0.408	0.78	0.81 (0.73-0.90)	8.1×10^{-5}	2.4×10^{-4}
rs2446894	114956632	Fine-map	0.200	0.30	0.78 (0.68-0.89)	9.6×10^{-4}	0.001
rs2446899	114957373	Fine-map	0.265	0.21	0.82 (0.73-0.93)	0.007	0.016

Table 3.4: Association statistics of the SNPs in the fine-mapped region

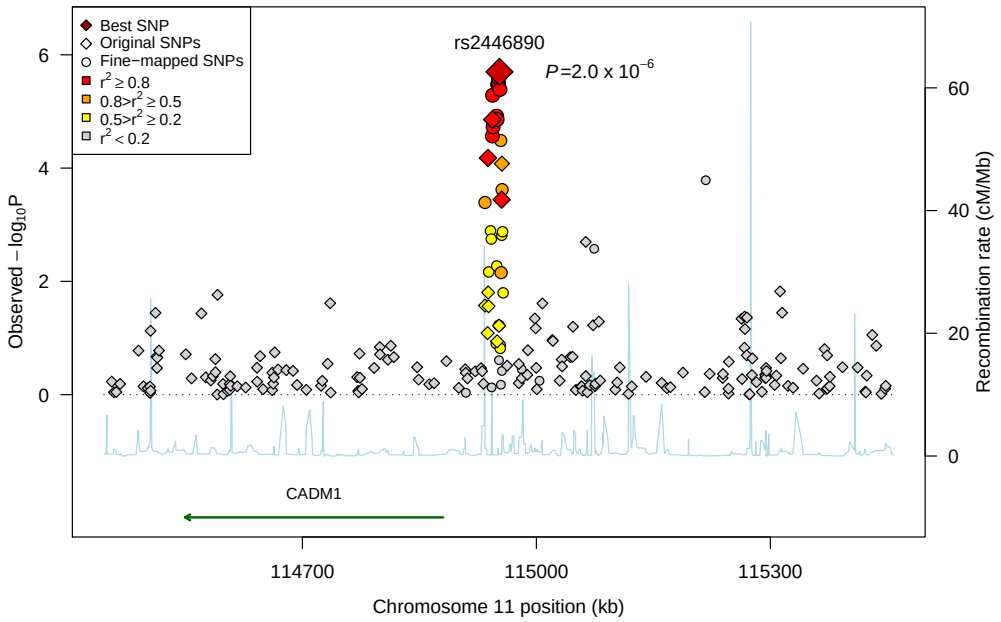


Figure 3.10: Association plot of the *CADM1* locus

the fine-mapped region, suggesting that the association signal at this locus may come from a single source well captured by rs2446890.

3.7.5 Haplotypic analysis

The high density of SNPs genotyped in the fine-mapping experiment allowed a detailed investigation of this locus. Haplotypes in this region were reconstructed using the confidence interval algorithm (Gabriel et al., 2002). The haplotype block containing the top SNP rs2446890 consisted of 29 SNPs within a 19 kb region (Figure 3.11). Analysis revealed five common haplotypes with a frequency greater than 5%, two of which were significantly associated with tuberculosis susceptibility (Table 3.5).

Haplotype	Freq	OR	95% CI	P-value
ACTCGAGCACGTCGGACGTTTGGTCGATG	0.250	1.21	1.05-1.40	0.009
TTCTGCGCACCCTAAGTGCCCAAATAGAA	0.178	0.86	0.73-1.02	0.079
TTCTACATGCCCTAAGTTCCCAAACGGAG	0.168	0.73	0.61-0.86	2.8×10^{-4}
TCCCGCGCAAGTCGGGCGTCTGGTCGATG	0.094	1.08	0.87-1.34	0.484
TCCCGCGCACGTCGGGCGTCTGGTCGATG	0.085	1.21	0.96-1.52	0.104

Table 3.5: Haplotypic association analysis of the fine-mapped region

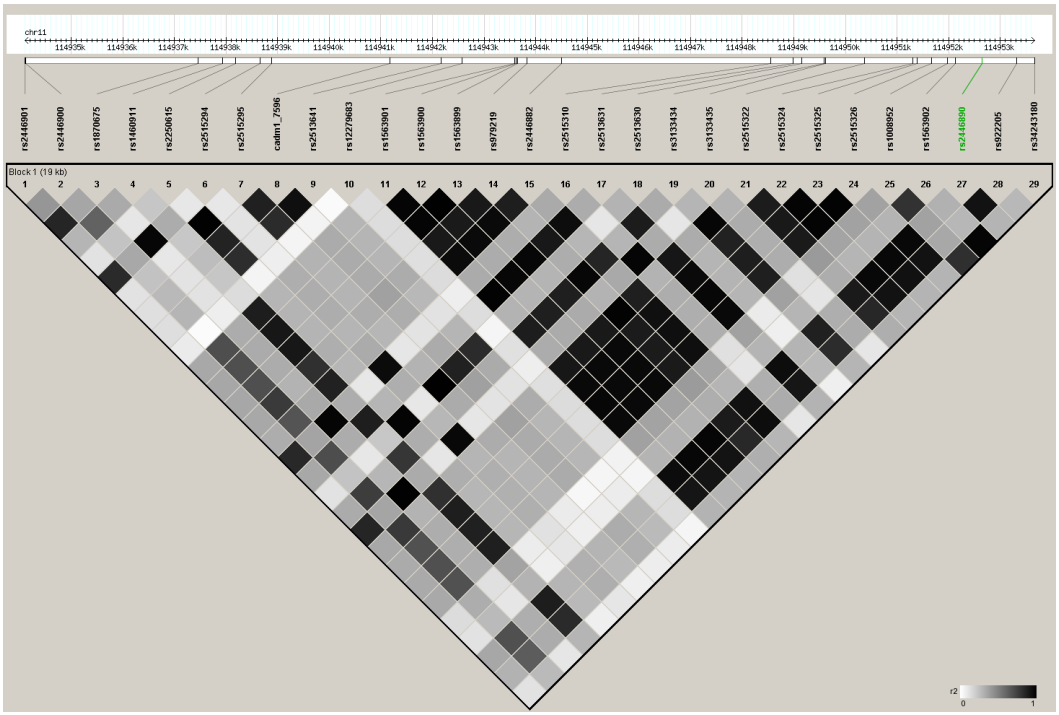


Figure 3.11: LD pattern of the fine-mapped *CADM1* locus. The block indicates the haplotype reconstructed containing rs2446890

3.8 Discussion

Genetic studies of tropical diseases are often limited by several major difficulties, some of which were encountered during the course of this study. Firstly, tropical diseases like tuberculosis are often endemic in developing countries where important infrastructure is limited. This poses difficulties in the collection of large disease cohorts with well characterised phenotypes. Given the modest effect sizes of most common variants (Hindorff et al., 2009; Manolio et al., 2009), the sample size is very important in providing a sufficient power to detect associations. It was estimated that sample sizes close to 5,000 maybe required for detecting moderate-to-low effect sizes of common genetic variants with reasonable statistical power (Bhangale et al., 2008).

Secondly, African populations are characterised by extensive genetic diversity (Campbell and Tishkoff, 2008) and substantially shorter LD in the populations (Gabriel et al., 2002; Conrad et al., 2006). As GWA studies rely on tagging markers to achieve a high genome coverage, due to the lower LD and less efficient alleles tagging, the coverage in the Gambian population is reduced. It was estimated that the coverage of the microarray used in this study was only 41% in the African Yoruba population (Barrett and Cardon, 2006), and another study reported even lower estimates (Bhangale et al., 2008). Moreoever, the extensive diversity of African populations precluded the possibility of imputation using the HapMap Yoruba population, as an attempted imputation for the Gambian data showed a disconcordant rate of 10.7% between imputed and genotyped data. This further limits the amount of informative data for analysis.

Nevertheless, this study identified *CADM1* as a possible candidate for tuberculosis susceptibility. CADM1, also known as IgSF4, TSLC1, Necl2 or SynCAM, is an immunoglobulin containing transmembraneous protein that defines a subset of dendritic

cells (Galibert et al., 2005). Through binding to the class I restricted T-cell associated molecule (CRTAM), it mediates release of cytokines, including IFN-γ which is crucially important for resistance to mycobacteria (Dalton et al., 1993; MacMicking et al., 2003). Recent studies suggested that this CRTAM protein is essential for the development and differentiation of T-cells (Yeh et al., 2008; Medina-Contreras et al., 2010), and *crtam*$^{-/-}$ mice have been observed to have more than ten-fold increase in bacterial load upon challenge with *Citrobacter rodentium*. Given that CADM1 is expressed on dendritic cells and is the only ligand of CRTAM (Wong et al., 2010), it is plausible that CADM1 may have an important immunological role in the defence against tuberculosis. In addition to its role in the immune system, CAMD1 also acts as a tumour suppressor protein in various cancers (Kuramochi et al., 2001; Ando et al., 2008; Dewan et al., 2008; Michels et al., 2008; Overmeer et al., 2008; Fu et al., 2009) and has important roles in spermatogenesis (van Der Weyden et al., 2006), neuronal synapse formation (Biederer et al., 2002) and mast cell interaction (Furuno et al., 2005; Hollins et al., 2008).

The associations of SNPs in *CADM1* in the primary and replication studies showed different results. Significant associations were observed in the Gambian and Conakry-Guinean populations, but not in the Bissau-Guinean and Malawian populations. Such observations reflect a tension between possibilities of true versus chance associations. On one side, given the small sample sizes of the replication cohorts, they have a low power of detecting real associations. Assuming the effect in the primary analysis is real, the power to detect the signal at rs2446890 at $P < 0.01$ was 11.8% and 17.2% for the Bissau-Guinean and Conakry-Guinean cohort, and was at best 58.2% for the Malawian cohort. Therefore the lack of association in these cohorts does not negate the possibility of a real association. Secondly, apart from *CADM1*, evidence of associations were also observed in the regions flanking two of the three other cell adhesion (CADM) family members, including one SNP in *CADM2* ($P_{geno} = 1.2 \times 10^{-5}$)

and three SNPs in *CADM3* (minimal $P_{geno} = 2.9 \times 10^{-5}$). An association was also seen with a SNP in *NCAM2* ($P_{geno} = 3.5 \times 10^{-5}$) which shares 65% homology with the Ig domains of CADM1. To evaluate the possibility of such an observation, a gene family based analysis by permutation was carried out by assigning the associated SNPs ($P < 1.0 \times 10^{-4}$) to the nearest gene within 100 kb. The analysis revealed that the probability of observing random associated regions clustering within a gene family is $P = 1.4 \times 10^{-4}$. These add to the genetic evidence and support *CADM1* as a possible candidate gene in tuberculosis.

On the other side, it remains possible that these associations represent a chance finding. The simultaneous survey of numerous SNPs in a microarray experiment necessitates an adjusted threshold for statistical significance, due to the multiple testing involved. Although the SNP rs2446890 near *CADM1* showed a significant association at $P_{geno} = 2.0 \times 10^{-6}$, this magnitude is below the commonly used threshold for genome-wide level of significance at $P < 5.0 \times 10^{-7}$, or even $P < 5.0 \times 10^{-8}$ (Hoggart et al., 2008). Despite it having the most significant association among all $> 400,000$ SNPs tested, it is still reasonably possible that the association is a chance finding. This, however, is unlikely to represent spurious association secondary to the population structure given the correction for three MDS components resulting in over-dispersion (λ) of 1.05. Moreover, significant association was not observed in the Ghanaian GWA study (rs2446890, $P_{geno} = 0.368$ and $P_{trend} = 0.248$) which forms parts of the combined analysis to be described in the next chapter; although a consistent direction of effect was detected (OR=0.92, 95% CI=0.78-1.07).

Finally, it is possible that the *CADM1* association represent a different genetic effect between populations. This population specific effect can arise due to differences in LD, or true biological heterogeneity. The former phenomenon results when there is a differential LD of the genetic marker with the true causal variant in different

populations. Assuming there is a real causal variant at the association region, such differences in LD would result in differential tagging and manifest as heterogeneous results in the genotyped markers. This is a possible scenario especially among African populations which harbour considerable genetic diversity (Campbell and Tishkoff, 2008). However, the resequencing and fine-mapping experiment failed to identify any possible casual variant, although it could have been missed as only a subset of all 128 SNPs were genotyped. As to the later phenomenon of true biological heterogeneity, it is possible that the genetic effect is specific to the West African populations, especially with those from Ghana and Malawi which are located at considerable geographical distances away from The Gambia. Given the diversity of *M. tuberculosis* (Gagneux et al., 2006), such distances would signify substantially different strains of the bacteria infecting individuals of different genetic compositions, resulting in a different host-pathogen compatibility as discussed in Chapter 1.

In conclusion, this study has identified *CADM1* as a potential candidate for susceptibility to tuberculosis. The work represents one of the largest efforts to study the host genetics of tuberculosis, and emphasises the genetic complexities of African populations. Further genetic studies from more West African and other populations are needed to determine the specificity and generalisability of this association. Finally, this work has also emphasised the importance of sample size in the detection of common variants of modest effect sizes. The next chapter describes a combined study, using data from this Gambian GWA study and a collaborative Ghana GWA study, to maximise the power to identify genetic variants in tuberculosis susceptibility.

Chapter 4

Combined analysis of two tuberculosis studies

4.1 Introduction

Genome-wide association (GWA) studies can identify common genetic variants associated with complex diseases and provide insight into the biology of these traits (The Wellcome Trust Case Control Consortium, 2007). However, because of the modest effect size of most common variants, individual studies are limited by insufficient power to detect true disease associations (Barrett and Cardon, 2006). In addition, power is reduced when the allele frequency is low. The most efficient way to increase sample size in order to detect loci of modest effect is to combine different studies to achieve greater power (Altshuler and Daly, 2007). This approach increases the sensitivity to detect genetic variants of smaller effect sizes. Presented here is the work of the combined analysis of two West African GWA studies, one being the WTCCC tuberculosis study described in the previous chapter and the other being a Ghanaian GWA tuberculosis study.

4.2 Cohorts and study designs

The combined GWA analysis consisted of samples recruited from The Gambia and Ghana. Details of the sample cohorts have been discussed in Chapter 2. Briefly, after quality control, the Gambian cohort consisted of 1,316 cases and 1,382 controls and the Ghanaian cohort consisted of 921 cases and 1,740 controls. These combined to give a total of 2,217 cases and 3,113 controls. The samples were genotyped on the Affymetrix 500K array (Gambians) and the Affymetrix 6.0 SNP array (Ghanaians) respectively, and the combined dataset consisted of 333,754 autosomal SNPs. These SNPs were tested for evidence of association using logistic regression, correcting for the first six MDS components. Seventeen SNPs with $P < 1.0 \times 10^{-4}$ in the combined analysis and $P < 1.0 \times 10^{-2}$ in both GWA studies were carried forward for replications in separate cohorts from Malawi, Bissau, Conakry and Ghana. The study design algorithm of the combined analysis is illustrated in Figure 4.1.

4.3 Power calculation

The primary goal of the combined analysis is to maximise the power to detect genetic variants. As illustrated in Figure 4.2, the combined dataset has more power than any of the individual cohorts, with at least 95% power to detect associations of $P < 1.0 \times 10^{-4}$ at an odds ratio of 1.3. The increase in power is most significant for genetic variants of odds ratio between 1.2-1.3. For a genetic variant with an odds ratio of 1.25, the power of detection increases from 50% in the Gambian cohort to 88% in the combined dataset. This significantly increases the ability to detect common variants with modest effect sizes.

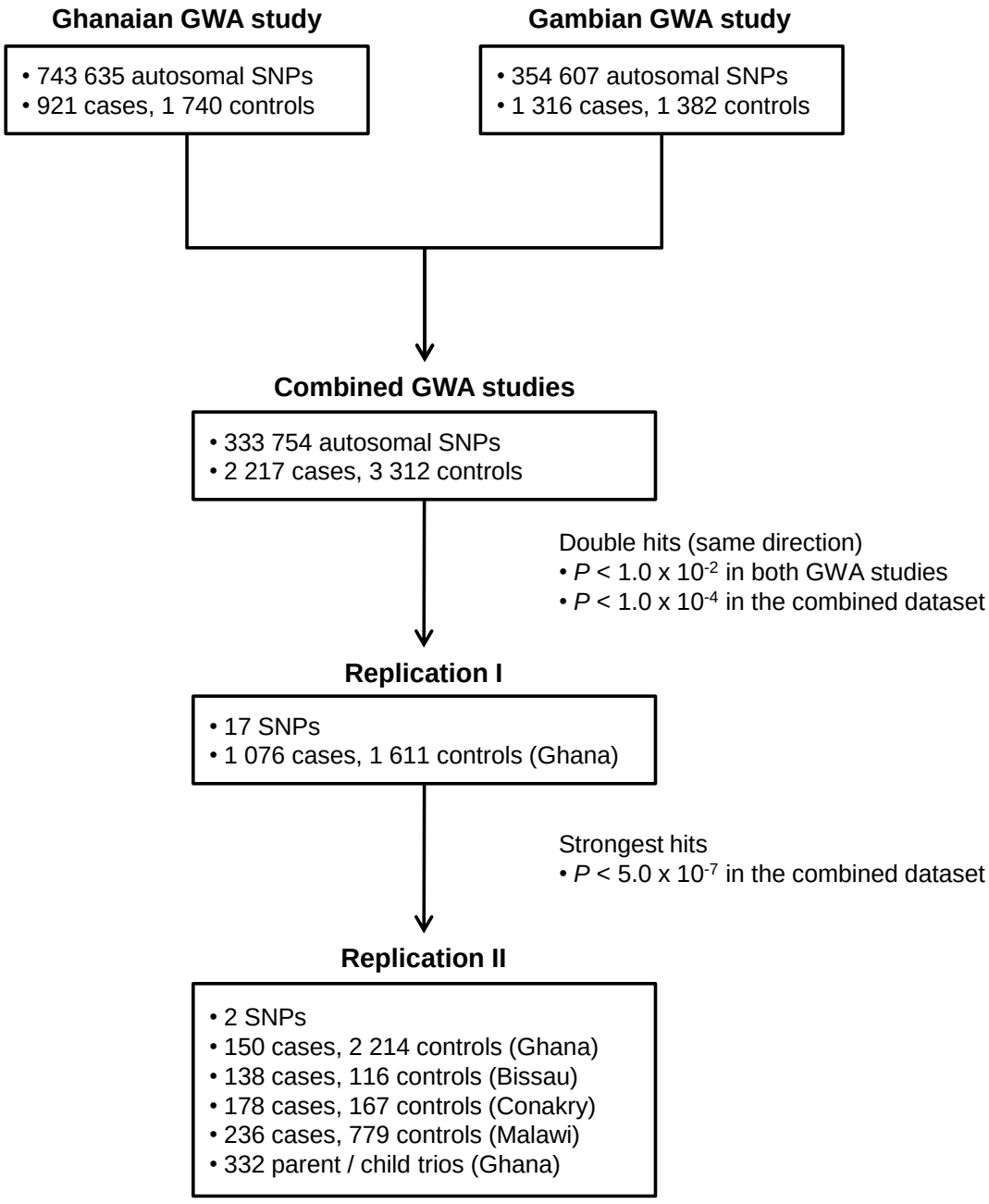


Figure 4.1: Flow chart showing the strategy of combining the Gambian and Ghanaian tuberculosis studies.

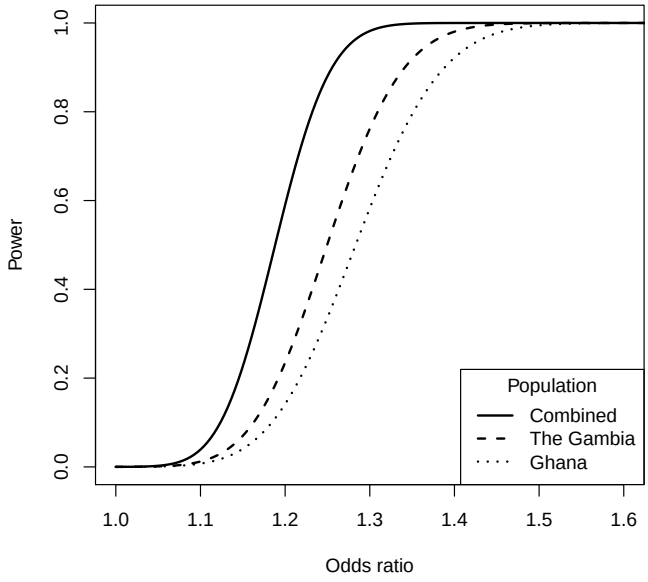


Figure 4.2: Power to detect association at $P < 1.0 \times 10^{-4}$ at an allele frequency of 25% in the combined and individual studies.

4.4 Quality control

4.4.1 Sample filters

The quality control criteria for the Gambian GWA study have been described in Chapter 3. As for the Ghanaian GWA study, genotypes were called using the Birdseed algorithm (Carvalho et al., 2010), which uses a contrast quality control to measure the separation of allele intensities into three genotype clusters. Fifteen samples yielded a contrast quality control below the recommended threshold of 0.4 and were excluded from further analyses. Comparing annotated with computed sex identified 56 individuals with discordant results and these samples were removed from the dataset. Eleven samples with an autosomal heterozygosity rate $> 32\%$ or a call rate of $< 95\%$ were excluded from further analysis. Genetic outliers were eliminated by the nearest neighbour allele sharing algorithm (Purcell et al., 2007). A total of 49

individuals with $Z < -3$ and five subjects with $Z > 1$ in the normalised distribution were excluded. Finally there were 260 related pairs of individuals, and samples with a higher SNP missing frequency were excluded. In all, 376 individuals were excluded from the data set resulting in 921 cases and 1,740 controls.

4.4.2 SNP filters

As for the SNP quality control, SNPs with a minor allele frequency of $< 1\%$, Hardy-Weinberg equilibrium of $P < 1.0 \times 10^{-4}$ and genotype rate of $< 98\%$ were removed from the analysis. These criteria were also applied to the Gambian data to standardise the two studies. For all the SNPs that met these criteria and yielded association signals of $P < 1.0 \times 10^{-5}$ in the combined analysis, signal intensity plots were visually inspected to exclude SNPs with ambiguous calling.

4.5 Population structure

Multidimensional scaling (MDS) analysis was carried out to investigate the population structure. Significant population structure was observed in both the Gambian and Ghanaian populations, with the samples clustering according to their self-reported ethnic groups (Figure 4.3). To look at these findings more globally we subjected the Ghanaian and Gambian results to MDS analysis with the reference samples from the HapMap populations (The International HapMap Consortium, 2003, 2005). As expected, genetic differences between the Asian and European populations were large, but the three populations originating from West Africa could also be distinguished clearly (Figure 4.4). These population-specific differences may preclude imputation of the Ghanaian or Gambian data using the HapMap Yoruba population genotypes.

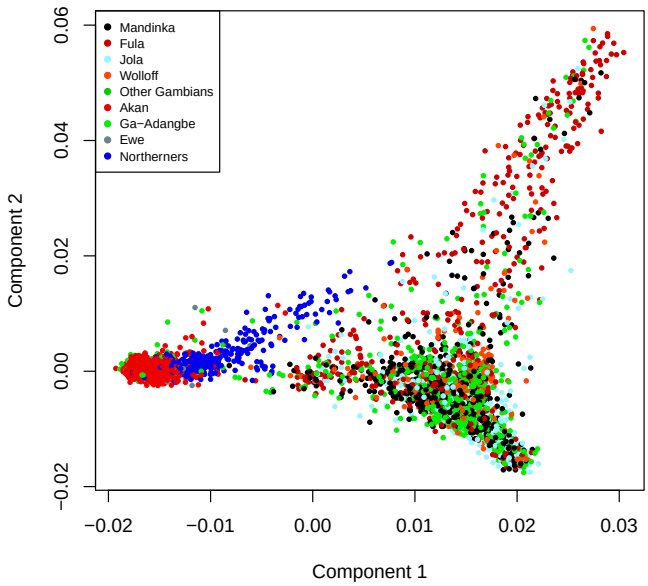


Figure 4.3: MDS plot of the Gambian and Ghanaian ethnic groups

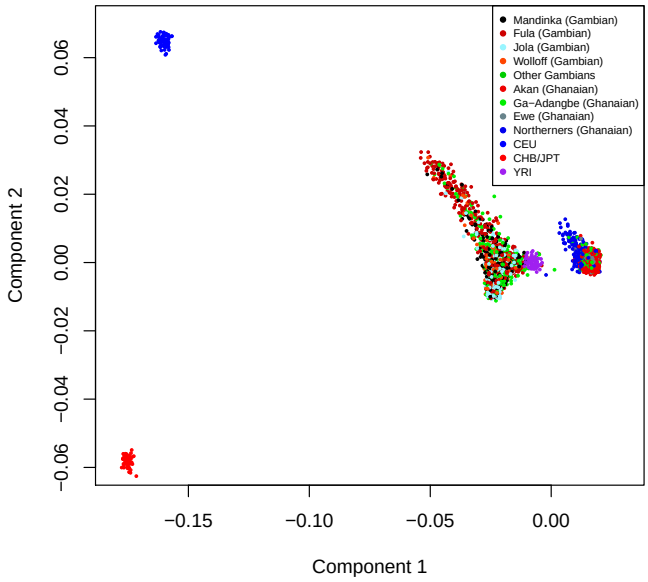


Figure 4.4: MDS plot of the Gambian, Ghanaian and the HapMap populations

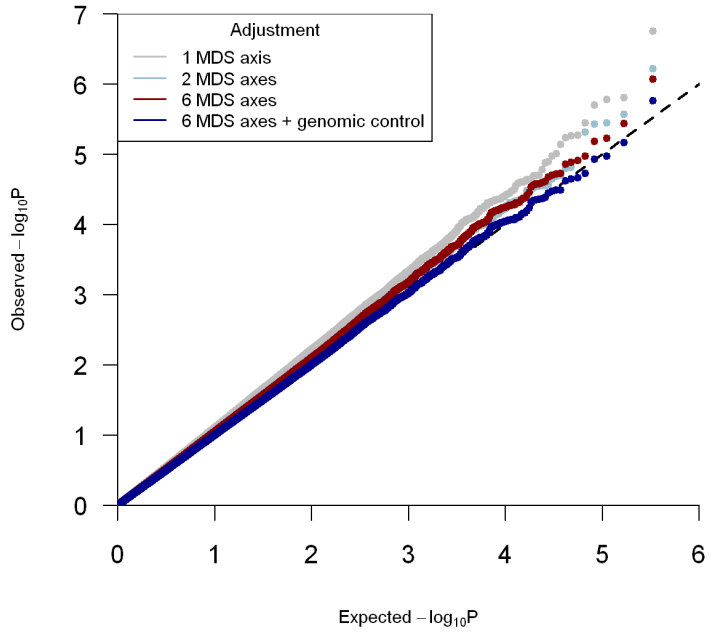


Figure 4.5: Quantile-quantile plot of the combined tuberculosis study. The final analysis utilised correction with six MDS axes

These data are compatible with the African populations being genetically diverse (Campbell and Tishkoff, 2008), and indicate that significant population structure exists in the dataset. To correct for the population stratification, the first six MDS components derived from the combined dataset were included in the logistic regression analysis of the combined dataset. The subsequent quantile-quantile plot showed no significant inflation of the association statistics after correction using the MDS components (λ=1.06, Figure 4.5).

4.6 Primary association and replication results

In the primary analysis of the combined dataset, a total of 17 SNPs showed evidence of association with $P < 1.0 \times 10^{-4}$ in the combined dataset and $P < 1.0 \times 10^{-2}$ in both groups with consistent directions of effects (Figure 4.6 and Table 4.1). These

SNPs were genotyped in an additional 1,076 cases and 1,611 controls from Ghana in the first phase of replication (Replication I). These replication results revealed two SNPs (rs4331426 and rs2335704) with $P < 5.0 \times 10^{-7}$ (Tables 4.1 and 4.2). These two SNPs were further genotyped in additional cohorts from Ghana, Bissau and Conakry and Malawi in the second phase of replication (Replication II), as well as the family samples from Ghana.

The SNP rs4331426, which maps to chromosome 18q11.2, showed association signal in the combined GWA analysis with $P = 2.2 \times 10^{-5}$ (Table 4.1). Consistent association was observed in the Ghanaian cohort in Replication I, with the minor allele associated with an increased risk of tuberculosis in the combined statistic ($P = 1.1 \times 10^{-7}$, OR=1.19, 95% 1.12-1.27, Tables 4.1). Further genotyping in Replication II showed that a consistent direction of effect (Table 4.2). Combining the GWA and all the replication data, the overall association statistic was strongly significant with $P = 9.2 \times 10^{-9}$ (OR = 1.18, 95% CI=1.11-1.24, Figure 4.7). The analysis of Ghanaian nuclear families also supported the association ($P = 0.016$, OR=1.33, 95% CI 1.05-1.68).

The SNP rs2335704 in the gene *PARD3B* showed evidence of association with $P = 4.6 \times 10^{-6}$ in the combined GWA studies. However, although this SNP passed the predefined threshold of significance after Replication I ($P = 3.0 \times 10^{-7}$, OR=1.23, 95% CI=1.14-1.33), it showed an opposite effect in the Ghanaian population in Replication II (Table 4.2). The overall association statistic was $P = 1.4 \times 10^{-6}$ (Figure 4.7).

4.7 The chromosome 18q11.2 locus

As above-mentioned, the SNP rs4331426 showed consistent associations in the different studies with an overall $P = 9.2 \times 10^{-9}$ for the case-control studies (Table 4.2).

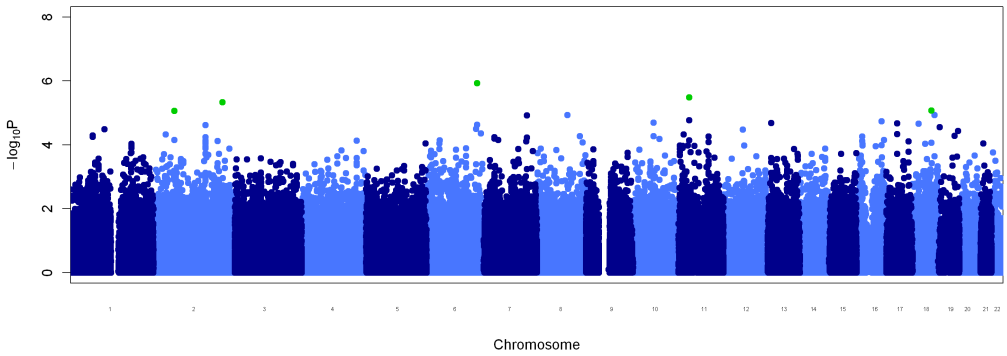
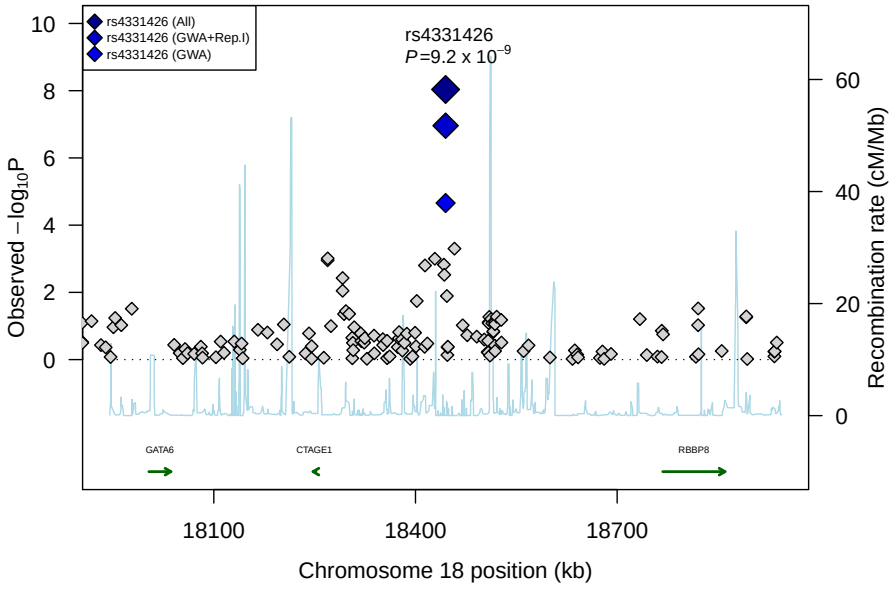


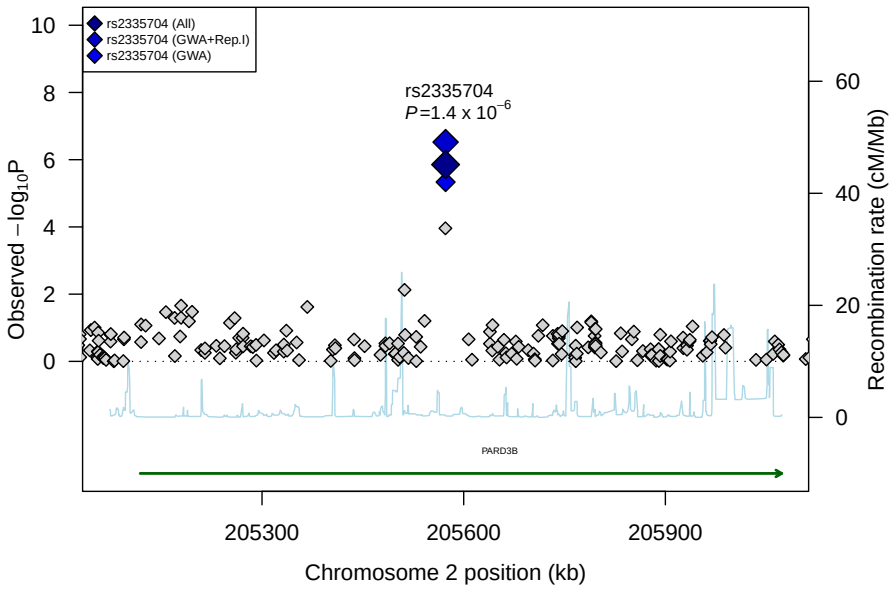
Figure 4.6: Manhattan plot of the combined tuberculosis study. SNPs with $P < 1.0 \times 10^{-5}$ are highlighted in green

SNP	Chr	Position	Gene	Gambia		Ghana		Combined GWA			GWA & Replication I		
				OR	*P*	OR	*P*	OR	95% CI	*P*-value	OR	95% CI	*P*-value
rs10493412	1	66855347	*SGIP1*	1.24	0.005	1.36	0.003	1.28	1.14-1.44	5.6×10^{-5}	1.19	1.07-1.33	0.002
rs17436311	1	66856380	*SGIP1*	1.25	0.004	1.35	0.003	1.28	1.14-1.44	5.2×10^{-5}	1.19	1.07-1.32	0.001
rs10490534	2	154646176	*GALNT13*	1.18	0.004	1.18	0.006	1.18	1.09-1.28	9.1×10^{-5}	1.09	1.02-1.16	0.010
rs2335704	2	205573077	*PARD3B*	1.23	0.003	1.28	4.3×10^{-4}	1.26	1.14-1.39	4.6×10^{-6}	1.23	1.14-1.33	3.0×10^{-7}
rs7769812	6	38846812	*DNAH8*	0.82	0.005	0.82	0.005	0.82	0.74-0.91	9.1×10^{-5}	0.87	0.80-0.95	0.001
rs4896905	6	147713687	*STXBP5*	0.76	0.002	0.77	0.006	0.76	0.67-0.86	3.2×10^{-5}	0.83	0.75-0.92	5.1×10^{-4}
rs1875148	10	61011997	-	1.28	0.003	1.28	0.003	1.29	1.15-1.45	2.0×10^{-5}	1.16	1.05-1.28	0.003
rs1503442	11	16382358	*SOX6*	0.81	0.004	0.82	0.004	0.82	0.75-0.90	4.8×10^{-5}	0.88	0.82-0.95	0.001
rs11173067	12	58203089	-	0.84	0.006	0.82	0.003	0.83	0.76-0.91	3.3×10^{-5}	0.92	0.85-0.99	0.033
rs7190310	16	17179363	*XYLT1*	1.19	0.010	1.24	0.003	1.21	1.10-1.33	7.6×10^{-5}	1.12	1.03-1.22	0.009
rs1353690	16	74576356	-	1.50	0.002	1.49	0.005	1.50	1.25-1.81	1.8×10^{-5}	1.40	1.19-1.65	4.5×10^{-5}
rs4331426	18	18444793	-	1.18	0.003	1.18	0.004	1.18	1.09-1.27	2.2×10^{-5}	1.19	1.12-1.27	1.1×10^{-7}
rs1943238	18	56266216	∼*SLC16A9*	1.43	0.006	1.72	0.002	1.52	1.23-1.87	8.7×10^{-5}	1.41	1.18-1.68	1.2×10^{-4}
rs1943240	18	56267422	∼*SLC16A9*	1.50	0.001	1.79	0.001	1.58	1.29-1.93	8.4×10^{-6}	1.47	1.24-1.74	8.7×10^{-6}
rs3937015	18	65750496	*CD226*	0.83	0.001	0.84	0.004	0.83	0.76-0.90	1.2×10^{-5}	0.90	0.84-0.96	0.002
rs35387445	19	59141468	∼*CACNG7*	1.42	0.001	1.46	0.003	1.41	1.20-1.66	3.7×10^{-5}	1.32	1.15-1.51	7.8×10^{-5}
rs9981165	21	22262330	-	0.73	0.001	0.48	0.002	0.66	0.54-0.81	9.1×10^{-5}	0.78	0.65-0.94	0.009

Table 4.1: Primary association and Replication I statistics of the combined tuberculosis study



(a) Chromosome 18q11.2



(b) *PARD3B*

Figure 4.7: Association plots of the chromosome 18q11.2 and *PARD3B* loci

Population	MAF (case, ctrl)	OR (95% CI)	$P_{logistic}$
rs4331426			
The Gambia (GWA)	0.521, 0.476	1.18 (1.06-1.31)	2.9×10^{-3}
Ghana (GWA)	0.491, 0.448	1.18 (1.05-1.32)	4.3×10^{-3}
Ghana (Replication I)	0.477, 0.429	1.19 (1.06-1.31)	2.8×10^{-3}
Ghana (Replication II)	0.476, 0.442	1.18 (0.92-1.51)	0.190
Bissau (Replication II)	0.513, 0.492	1.09 (0.76-1.56)	0.644
Conakry (Replication II)	0.440, 0.435	1.02 (0.73-1.43)	0.906
Malawi (Replication II)	0.563, 0.525	1.15 (0.91-1.45)	0.237
Ghana Families	-	1.33 (1.05-1.68)	0.016
Combined GWA	-	1.18 (1.09-1.27)	2.2×10^{-5}
Combined GWA and Replication I	-	1.19 (1.11-1.27)	1.1×10^{-7}
Combined all results	-	1.18 (1.11-1.24)	9.2×10^{-9}
rs2335704			
The Gambia (GWA)	0.206, 0.172	1.23 (1.07-1.42)	3.2×10^{-3}
Ghana (GWA)	0.232, 0.192	1.28 (1.12-1.47)	4.3×10^{-4}
Ghana (Replication I)	0.230, 0.204	1.19 (1.04-1.37)	0.014
Ghana (Replication II)	0.188, 0.202	0.92 (0.68-1.26)	0.630
Bissau (Replication II)	0.230, 0.193	1.25 (0.81-1.92)	0.317
Conakry (Replication II)	0.227, 0.202	1.15 (0.86-1.56)	0.333
Malawi (Replication II)	0.166, 0.163	1.02 (0.75-1.38)	0.910
Ghana Families	-	0.78 (0.60-1.03)	0.082
Combined GWA	-	1.26 (1.14-1.39)	4.6×10^{-6}
Combined GWA and Replication I	-	1.23 (1.14-1.33)	3.0×10^{-7}
Combined all results	-	1.20 (1.12-1.29)	1.4×10^{-6}

Table 4.2: Association statistics of the SNPs at chromosome 18q11.2 and *PARD3B*

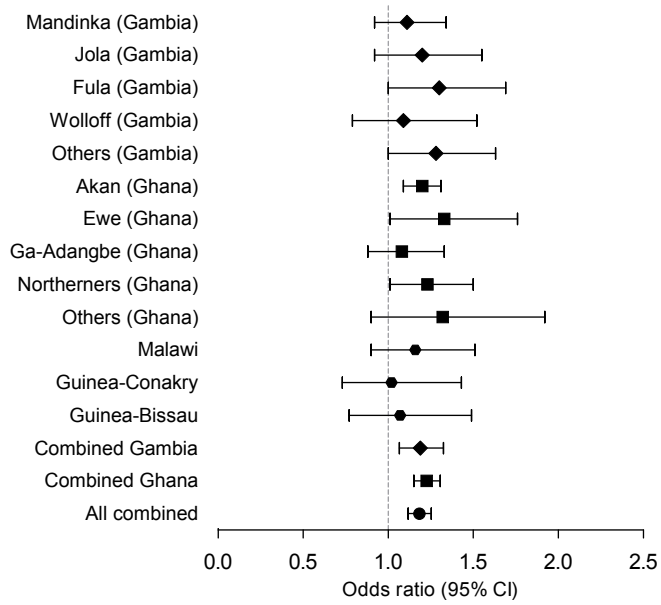
SNP	Chr	Position	Type	Gambia		Ghana		Combined GWA			GWA & Replication I		
				OR	*P*	OR	*P*	OR	95% CI	*P*-value	OR	95% CI	*P*-value
rs4331426	18	18444793	Original	1.18	0.003	1.18	0.004	1.18	1.09-1.27	2.2×10^{-5}	1.19	1.12-1.27	1.1×10^{-7}
rs11874936	18	18457951	Additional	1.11	0.140	1.21	0.001	1.16	1.07-1.28	5.0×10^{-4}	1.18	1.04-1.27	1.9×10^{-6}
rs11877287	18	18428795	Additional	1.12	0.073	1.20	0.002	1.15	1.05-1.25	0.001	1.16	1.09-1.25	7.9×10^{-6}
rs7231506	18	18442878	Additional	0.90	0.110	0.84	0.002	0.88	0.81-0.96	0.003	0.86	0.81-0.92	1.8×10^{-5}
rs2335704	2	205573077	Original	1.23	0.003	1.28	4.0×10^{-4}	1.26	1.14-1.39	4.6×10^{-6}	1.23	1.14-1.33	3.0×10^{-7}
rs2335705	2	205572777	Additional	1.19	0.045	1.31	2.0×10^{-4}	1.25	1.13-1.40	1.1×10^{-4}	1.22	1.11-1.33	7.9×10^{-6}

Table 4.3: Association statistics of the additional SNPs at chromosome 18q11.2 and *PARD3B*

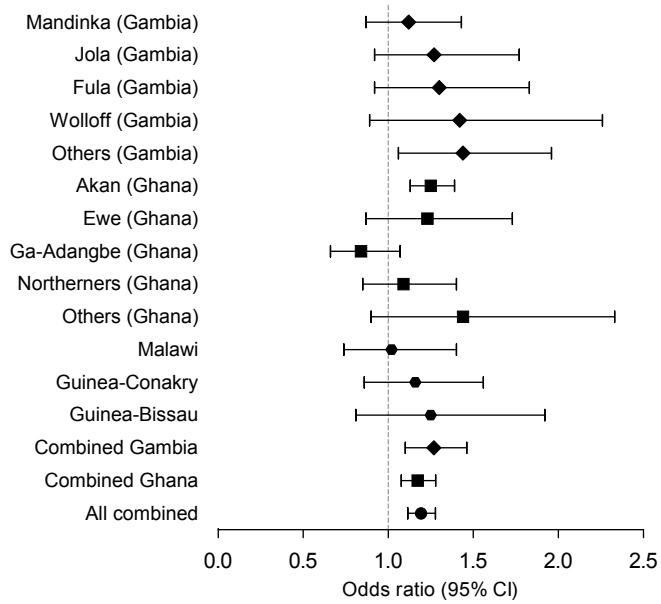
The minor G allele was associated with an increased risk of tuberculosis (OR=1.18, 95% CI=1.11-1.24). The associations were significant in the Ghanaian cohort in Replication I ($P = 0.003$), as well as the Ghanaian family study ($P = 0.016$) which should be robust to population stratification (Laird and Lange, 2006). Although the associations did not reach statistical significance in Replication II probably due to the smaller sample sizes, a consistent direction of effect was observed in each of these populations. In fact, the odds ratio for all of the studies show a consistent direction of effect with negligible heterogeneity between them ($P_Q = 0.961$, $I^2 = 0$, Table 4.2). Additional genotyping of three neighbouring SNPs did not identify any SNP with a stronger association (Table 4.3). To study the effect of this variant in different populations, the genotype data were stratified according to ethnic groups. A consistent effect was observed in all of 13 ethnic groups, with the minor G allele conferring an increased risk to tuberculosis (Figure 4.8).

4.8 The *PARD3B* locus

The SNP rs2335704 at *PARD3B* showed evidence of association in the combined GWA and Replication I ($P = 3.0 \times 10^{-7}$), with the minor C allele found to associate with an increased risk to tuberculosis (OR=1.23, 95% CI=1.14-1.33). However, heterogeneous results were observed in Replication II, with an opposite effect in the Ghanaian case-control cohort ($P = 0.630$, OR=0.92, 95% CI=0.68-1.26) and in the Ghanaian family cohort ($P = 0.082$, OR=0.78, 95% CI=0.60-1.03, Table 4.2). This inconsistency was reflected in the statistical tests of heterogeneity with $P_Q = 0.062$ ($I^2 = 47.9$, 95% CI=0-78.6%), indicating a moderate level of heterogeneity (Huedo-Medina et al., 2006). Combining all the case-control populations, the overall evidence was weakened with $P = 1.4 \times 10^{-6}$ (OR=1.20, 95% CI=1.12-1.29, Table 4.2). Further stratified analysis suggested that the effect of this SNP is distinctly opposite in the



(a) Chr 18q11.2, rs4331426



(b) *PARD3B*, rs2335704

Figure 4.8: Forest plots showing associations of rs4331426 and rs2335704 in different populations

Chr	Gene	Size (kb)	No of SNPs	Best SNP	OR (95% CI)	P_{SNP}	P_{Sidak}
1	*IL10*	4.9	21	rs3024505	1.23 (1.00-1.51)	0.052	0.676
2	*SLC11A1*	14.9	5	rs11904786	1.18 (0.96-1.44)	0.119	0.470
4	*TLR2*	21.8	19	rs10517578	1.19 (0.98-1.44)	0.082	0.801
5	*IL12B*	15.7	21	rs6874870	1.23 (0.97-1.56)	0.094	0.874
6	*HLA-DQA1*	6.2	9	rs9469220	1.15 (1.05-1.25)	0.002	0.015
6	*HLA-DRB1*	11	5	rs9272346	1.10 (1.01-1.19)	0.022	0.106
6	*IFNGR1*	21.9	17	rs9373180	1.57 (1.16-2.13)	0.004	0.064
6	*TNF*	2.8	5	rs1052248	0.96 (0.87-1.06)	0.401	0.923
10	*MBL2*	6.3	32	rs12573117	1.18 (1.04-1.35)	0.013	0.337
12	*IFNG*	5	29	rs2193047	1.09 (1.00-1.18)	0.038	0.676
12	*VDR*	63.5	23	rs11168312	0.86 (0.72-1.02)	0.082	0.861
15	*UBE3A*	101.7	27	rs7169070	1.12 (0.96-1.30)	0.158	0.990
16	*NOD2*	35.9	16	rs2066849	1.07 (0.99-1.16)	0.077	0.722
17	*CCL18*	7.2	20	rs1614133	1.11 (1.02-1.19)	0.012	0.219
17	*CCL2*	1.9	19	rs3917891	0.90 (0.81-1.00)	0.047	0.598
17	*CCL4*	1.8	11	rs1614133	1.11 (1.02-1.19)	0.012	0.127
17	*NOS2A*	43.8	12	rs3729508	0.90 (0.81-1.00)	0.042	0.403
17	*STAT5B*	77.2	7	rs9912576	1.05 (0.96-1.14)	0.283	0.903
19	*CD209*	7.5	10	rs11260030	0.93 (0.84-1.03)	0.164	0.832
19	*IL12RB1*	27.3	4	rs7255589	0.92 (0.85-1.01)	0.071	0.256
20	*MC3R*	1.1	35	rs6014657	0.88 (0.79-0.97)	0.012	0.355

Table 4.4: Association statistics of reported tuberculosis susceptibility candidates

Ga-Adangbe group in Ghana compared with other ethnic groups (OR=0.84, Figure 4.8), where the minor C allele was associated with protection instead of risk as in other populations. Finally, an additional nearby SNP was genotyped but did not identify a stronger association (Table 4.3).

4.9 Previous candidates for tuberculosis

The large dataset in this study allows previous candidate genes to be investigated. The SNP with the best association statistic within 50 kb of the gene was identified. The association statistic was corrected with the number of SNPs in the gene, using Sidak's correction method as described in Chapter 2. The analysis showed a weak

evidence of association at *HLA-DQA1* with $P_{Sidak} = 0.015$. The strongest association was observed at the SNP rs9469220 with $P = 0.002$ (OR=1.15, 95% CI=1.05-1.25). However, significant association was not observed in any of the other genes in Table 4.4.

4.10 Discussion

As illustrated in Figure 4.2, the power of genetic association studies follows a sigmoid curve and changes dramatically over a particular range of effect size. Given the modest effect sizes of most common genetic variants (Hindorff et al., 2009; Manolio et al., 2009), the sample size is critical in determining the power of detecting genetic associations. A large sample size is important in identifying novel associations of modest effects, especially for genetic variants with low allele frequencies with which the power is further reduced. The most efficient way to increase sample size in order to detect loci of modest effect is to combine different studies to achieve greater power (Altshuler and Daly, 2007). This approach has been utilised in several studies to successfully identify novel genetic variants in common diseases (Zeggini et al., 2007; Barrett et al., 2008, 2009).

This study combines two GWA studies to achieve greater power to detect genetic variants in susceptibility to tuberculosis. The power to detect a genetic variant with an odds ratio of 1.25 increases from 50% in the Gambian cohort to 88% in the combined dataset. The benefit of such an increase in power is best illustrated by the results of the experiment. Both the chromosome 18q11.2 and *PARD3B* variants have modest effect sizes (OR=1.18 and 1.20 respectively). Despite evidence of association in the combined analysis, the individual GWA studies have failed to identify these variants beyond the genome-wide threshold of significance (Table 4.1). These genetic

variants would not have been detected without collaborative efforts to combine the two studies.

This study uses a two-stage approach in the experimental design. The initial stage includes combining the data from the two GWA studies. Although the initial threshold of $P < 1.0 \times 10^{-4}$ is not very stringent, it reduces the chance of missing true genetic variants at the cost of potentially including some false positive variants. To reduce the possibility of false positive findings, a more stringent threshold of $P < 5.0 \times 10^{-7}$ was used in Replication II, after combining with the results with the Ghanaian cohort in Replication I. As this cohort has a bigger size with 1,076 cases and 1,611 controls, it has a greater power to detect real associations and therefore minimises the chance of having false positive results beyond this second stage. The two SNPs passing this stage were carried forward for further replications in the smaller cohorts in Replication II.

The data from this study suggest that the chromosome 18q11.2 variant is a genuine susceptibility locus for tuberculosis. This is supported by the strongly significant association of $P = 9.2 \times 10^{-9}$ in the final analysis, backed up by the striking consistency in its genetic effect among all the 8 population cohorts and 13 ethnic groups under study (Table 4.2, Figure 4.8), including the southeast African population from Malawi. The overall odds ratio is 1.18 and the minor allele is associated with an increased risk to tuberculosis. The effect size of this variant is small but is comparable to most findings from other GWA studies (Hindorff et al., 2009), possibly due to the fact that variants of larger effect are selected against and therefore disappear from the population (Manolio et al., 2009). Even though population stratification can result in spurious associations, the absence of significant inflation in the quantile-quantile plot and the consistent association in the family study argue against such a possibility. This combined study identifies the genetic variant at chromosome 18q11.2 as a novel

susceptibility locus for tuberculosis. The chromosome 18q11.2 variant is located in a gene desert. The nearest genes are *CTAGE1* and *RBBP8*, located 192.9 kb and 322.5 kb away from the SNP rs4331426 respectively.

In contrast to the chromosome 18q11.2 variant, heterogeneous results were observed for the *PARD3B* association at rs2335704. Despite the evidence of association in the combined GWA studies and Replication I, the results were inconsistent in the studies in Replication II. An opposite direction of effect was observed in the Ghanaian case-control cohort as well as the Ghanaian family cohort (Table 4.2, Figure 4.8). Although this can be due to a population specific effect in the Ga-Adangbe population or pathogen diversity with *M. tuberculosis* as discussed in Chapter 3, this also suggests that the association at *PARD3B* may be a chance finding. The effect of this locus should therefore be interpreted with caution, and further replication work is needed to establish the genetic role of this locus in immunity to tuberculosis.

In conclusion, this study aggregates the power of two GWA studies to identify chromosome 18q11.2, and with less evidence on *PARD3B*, as susceptibility loci to tuberculosis, despite the difficulties in African studies due to the extensive genetic diversity and low LD (Campbell and Tishkoff, 2008; Jallow et al., 2009). Such population heterogeneity necessitates careful dissection of the genetic structure, and has resulted in a lower coverage for common variants in the genome. Nevertheless, this study represents the largest human genetic study of tuberculosis known to date, and suggests that functional studies should be pursued on the chromosome 18q11.2 locus. This study also emphasises the need for collaborative studies to increase the sample size, which is critically important in genetic studies to identify novel genetic loci of smaller to moderate effect sizes. Such efforts will be invaluable in providing novel biological insight to help humans fight against this ancient disease of tuberculosis.

Chapter 5

Leprosy gene-centric study

5.1 Introduction

Leprosy is the second commonest mycobacterial infection after tuberculosis. It is a chronic granulomatous disease affecting the skin and peripheral nerves. Despite its causative agent *Mycobacterial leprae* being the first identified human pathogen (Hansen, 1874), natural immunity to the bacterium remains poorly understood. The disease remains endemic in central Africa, south and southeast Asia and south America with more than 200,000 new cases per year (The World Health Organization, 2007). Twin studies (Chakravartti and Vogel, 1973), familial clustering (Shields et al., 1987) and segregation analyses (Abel et al., 1995) suggest that host genetics play an important role in susceptibility to this infectious disease. The heritability for leprosy was estimated as up to 57% with a sibling relative risk (λ_S) of 6.4. (Bakker et al., 2005).

Given the global health importance and heritability of leprosy, an association study was conducted on three different cohorts recruited from India where the disease is prevalent. Primary analysis involved genotyping samples recruited from New Delhi,

using a microarray with approximately 50,000 SNPs in 2,092 genes with a potential role in inflammatory and metabolic processes. This was followed by replication in two other cohorts from Kolkata and Madurai. The aim of this study was to identify genetic loci in leprosy, which may provide insight into the disease susceptibility and resistance. Furthermore, as *M. leprae* and *M. tuberculosis* are both mycobacteria, results of this study may potentially be extrapolated to improve our understanding of immunity to tuberculosis. This chapter details the results of this study.

5.2 Cohorts and study designs

The study was conducted using sample cohorts from New Delhi, Kolkata and Madurai of India as described in Chapter 2. The New Delhi samples were genotyped using the HumanCVD Genotyping BeadChip microarray (Keating et al., 2008). This gene-centric microarray was designed to target potentially relevant loci across a wide range of inflammatory and metabolic syndromes, covering 2,092 genes of which 435 have a high likelihood of functional significance. Although the total number of SNPs is less than most high-throughput genotyping platforms intended for genome-wide studies, this array has a greater depth of coverage with an average density of 36.5 SNPs in priority regions. Given the pronounced tissue inflammation seen in leprosy patients (Modlin et al., 1988; Krutzik et al., 2003), it is conceivable that genes in the inflammatory and metabolic pathways may be involved in susceptibility to leprosy. Genetic variants associated with $P < 1.0 \times 10^{-4}$ in the primary analysis were carried forward for replication in the Kolkata and Madurai cohorts, and were genotyped using Sequenom primer extension assays (Gabriel and Ziaugra, 2004).

In addition to these Indian samples, DNA samples of 1,463 individuals from the United Kingdom, The Gambia, Malawi and Russia were genotyped for population

differentiation analysis. The genotyping was performed using Sequenom primer extension assays (Gabriel and Ziaugra, 2004).

5.3 Quality control

Genotypes from the microarray were called using the BeadStudio Genotyping Module (Dunning et al., 2008). A first-pass criterion was performed to remove all grossly failing SNPs with a call rate of $< 90\%$. The data were then subjected to further quality control as described below.

5.3.1 Sample filters

A total of 558 individuals were genotyped. Samples with a low call rate (<90%) or extreme overall heterozygosity (<0.22 or >0.27) were excluded (Figure 5.1). Furthermore, samples showing evidence of duplication, contamination or cryptic relatedness (IBD proportion >0.2), as well as samples with an outlying ancestry (Figure 5.2) were removed. A total of 448 individuals were retained for association analysis.

5.3.2 SNP filters

Genetic variants with a low call rate (<95%) or extreme deviation from the Hardy-Weinberg equilibrium ($P < 1 \times 10^{-6}$) were excluded, as well as those with low minor allele frequency (<1%). The resulting set contained 29,020 SNPs for association analysis. Cluster plots of all SNPs with association $P < 1.0 \times 10^{-4}$ in the primary analysis were visually inspected, and SNPs with poorly separating clusters were excluded from further analysis (Table 5.1).

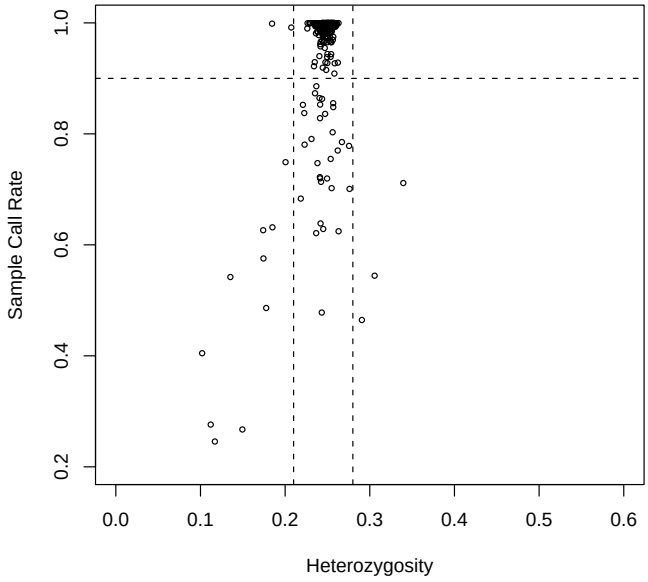


Figure 5.1: Sample call rate and heterozygosity. Dotted lines delimit the thresholds for sample exclusion.

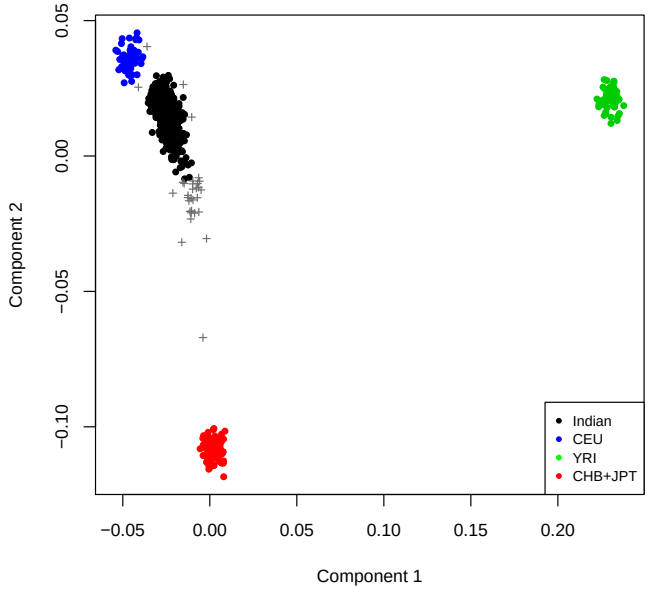


Figure 5.2: MDS plot of the New Delhi Indians and HapMap populations. Crosses denote samples excluded due to outlying ancestry

Exclusion criteria	No of SNPs	No of samples
Sample filters		
Sample call rate (<90%)	—	44
Heterozygosity (<0.22 or >0.27)	—	2
Duplication or relatedness (IBD proportion >0.2)	—	33
Outlying ancestry	—	31
Retained for analysis	—	*448*
SNPs filters		
SNP call rate (<95%)	752	—
Hardy Weinberg equilibrium ($P < 1 \times 10^{-6}$)	551	—
Minor allele frequency (<1%)	10,837	—
Retained for analysis	*29,020*	—

Table 5.1: Samples and SNPs exclusion criteria

5.4 Primary association results

The primary association analysis was performed with the Pearson's χ^2 allelic test (P_{alle}) and the Cochran-Armitage test for trend (P_{trend}). Both tests assumed a additive model for disease inheritance. A total of 49 SNPs from 30 different loci showed evidence of association at $P < 1.0 \times 10^{-4}$ (Figure 5.3, Table 5.2). The strongest evidence of association was observed at the human leukocyte antigen (*HLA*) class II locus at chromosome 6p21, with two SNPs showing a genome-wide level of association (rs9270650, $P_{alle} = 6.4 \times 10^{-10}$, $P_{trend} = 1.2 \times 10^{-9}$ and rs1071630, $P_{alle} = 8.5 \times 10^{-10}$, $P_{trend} = 4.8 \times 10^{-9}$). The genotype distributions of these two SNPs were in Hardy-Weinberg equilibrium in both the cases and controls (rs9270650, $P_{case} = 0.99$, $P_{ctrl} = 0.705$ and rs1071630, $P_{case} = 0.267$, $P_{ctrl} = 0.527$). Several SNPs close to but outside of the *HLA* region in *AGER*, *RNF5* and *BAK1* were also significantly associated. Significant associations were also observed for SNPs in *SLC12A1*, *MGLL*, *TLR1* and *RBP1*.

SNP	Chr	Position	Gene	Minor	MAF (case)	MAF (ctrl)	OR	95% CI	P_{alle}
rs3753242	1	2059541	*PRKCZ*	A	0.25	0.14	2.10	1.49-2.97	1.7×10^{-5}
rs3790562	1	67567011	*IL12RB2*	G	0.12	0.04	2.82	1.66-4.80	7.2×10^{-5}
rs6673833	1	85648210	*DDAH1*	G	0.15	0.27	0.49	0.35-0.68	2.4×10^{-5}
rs2244510	1	204661692	*SRGAP2*	C	0.27	0.40	0.55	0.42-0.73	3.5×10^{-5}
rs9429893	1	204667615	*SRGAP2*	G	0.27	0.40	0.56	0.42-0.75	6.4×10^{-5}
rs11118087	1	204723200	*IKBKE*	A	0.39	0.26	1.82	1.37-2.42	2.9×10^{-5}
rs6707461	2	164817572	-	A	0.27	0.39	0.57	0.43-0.76	9.5×10^{-5}
rs11465897	3	10229797	*IRAK2*	A	0.12	0.22	0.46	0.32-0.67	3.2×10^{-5}
rs664910	3	128956720	*MGLL*	G	0.46	0.33	1.73	1.32-2.28	6.7×10^{-5}
rs16830415	3	128956957	*MGLL*	C	0.21	0.10	2.35	1.61-3.42	6.8×10^{-6}
rs547801	3	128964929	*MGLL*	A	0.24	0.12	2.39	1.67-3.42	1.2×10^{-6}
rs295470	3	140696593	*RBP1*	G	0.58	0.42	1.90	1.46-2.49	2.0×10^{-6}
rs2071388	3	140719373	*RBP1*	G	0.58	0.42	1.90	1.45-2.47	2.2×10^{-6}
rs10008492	4	38442115	*TLR10*	C	0.03	0.11	0.23	0.12-0.44	1.6×10^{-6}
rs5743618	4	38475043	*TLR1*	C	0.04	0.13	0.27	0.15-0.47	1.3×10^{-6}
rs10033900	4	110878516	*PLA2G12A*	C	0.26	0.38	0.57	0.42-0.75	9.9×10^{-5}
rs9358391	6	21215088	*CDKAL1*	T	0.40	0.27	1.78	1.35-2.36	5.3×10^{-5}
rs3134947	6	32253183	*RNF5*	A	0.09	0.19	0.41	0.27-0.61	1.1×10^{-5}
rs3134940	6	32257794	*AGER*	G	0.08	0.18	0.41	0.27-0.63	2.8×10^{-5}
rs9270650	6	32673832	*HLA-DRB1*	C	0.44	0.24	2.43	1.83-3.23	6.4×10^{-10}
rs477515	6	32677669	*HLA-DQA1*	A	0.13	0.27	0.41	0.29-0.57	2.2×10^{-7}
rs2516049	6	32678378	*HLA-DRB1/DQA1*	G	0.14	0.27	0.42	0.29-0.59	5.5×10^{-7}
rs9270986	6	32682038	*HLA-DRB1/DQA1*	A	0.41	0.22	2.45	1.83-3.29	1.4×10^{-9}
rs482044	6	32684042	*HLA-DRB1/DQA1*	G	0.35	0.48	0.58	0.44-0.76	9.3×10^{-5}
rs3104369	6	32710460	*HLA-DQA1*	A	0.48	0.31	2.01	1.53-2.64	4.8×10^{-7}
rs1071630	6	32717104	*HLA-DQA1*	T	0.31	0.52	0.42	0.32-0.56	8.5×10^{-10}
rs9273363	6	32734250	*HLA-DQA1*	A	0.15	0.25	0.50	0.36-0.71	7.1×10^{-5}
rs3891175	6	32742445	*HLA-DQA1*	A	0.11	0.23	0.43	0.29-0.63	9.6×10^{-6}
rs210137	6	33650456	*BAK1*	T	0.43	0.30	1.76	1.33-2.31	5.9×10^{-5}
rs210138	6	33650516	*BAK1*	G	0.44	0.30	1.79	1.36-2.37	3.2×10^{-5}
rs7009163	8	11714373	*FDFT1*	C	0.07	0.17	0.34	0.22-0.55	3.4×10^{-6}
rs16902359	8	128812033	*MYC*	T	0.39	0.26	1.86	1.40-2.47	1.9×10^{-5}
rs1799998	8	143996602	*CYP11B2*	C	0.32	0.46	0.55	0.42-0.73	2.5×10^{-5}
rs4879816	9	34660128	*CCL27*	A	0.09	0.19	0.43	0.28-0.64	2.3×10^{-5}
rs3740237	10	32597598	*EPC1*	C	0.33	0.21	1.87	1.38-2.53	5.2×10^{-5}
rs17121799	10	108697641	*SORCS1*	A	0.15	0.07	2.39	1.53-3.73	8.7×10^{-5}
rs666004	11	120903788	*SORL1*	G	0.43	0.56	0.58	0.45-0.76	6.7×10^{-5}
rs525017	12	121781687	*GPR81*	A	0.25	0.14	2.06	1.47-2.89	2.6×10^{-5}
rs8006042	14	66899325	*EIF2S1*	G	0.32	0.20	1.89	1.39-2.56	4.1×10^{-5}
rs12588458	14	66900609	*EIF2S1*	G	0.37	0.24	1.87	1.40-2.50	2.4×10^{-5}
rs861537	14	103236828	*XRCC3*	A	0.40	0.54	0.57	0.44-0.75	3.6×10^{-5}
rs2675345	15	46187491	*SLC12A1*	G	0.29	0.16	2.09	1.51-2.88	5.7×10^{-6}
rs1320052	15	46282800	*SLC12A1*	C	0.22	0.11	2.19	1.52-3.15	2.2×10^{-5}
rs16960661	15	46293975	*SLC12A1*	C	0.20	0.09	2.51	1.69-3.73	3.1×10^{-6}
rs9920281	15	46301601	*SLC12A1*	G	0.25	0.12	2.53	1.77-3.61	1.9×10^{-7}
rs8032941	15	46302144	*SLC12A1*	C	0.12	0.05	2.83	1.68-4.76	4.8×10^{-5}
rs1013316	16	24042613	*PRKCB1*	A	0.41	0.55	0.58	0.45-0.76	6.0×10^{-5}
rs7501702	17	19234320	*MFAP4*	A	0.28	0.42	0.53	0.4-0.71	1.1×10^{-5}
rs11083841	19	51811873	*PTGIR*	G	0.45	0.32	1.73	1.32-2.27	7.5×10^{-5}

Table 5.2: Primary association statistics of SNPs with $Palle < 1.0 \times 10^{-4}$

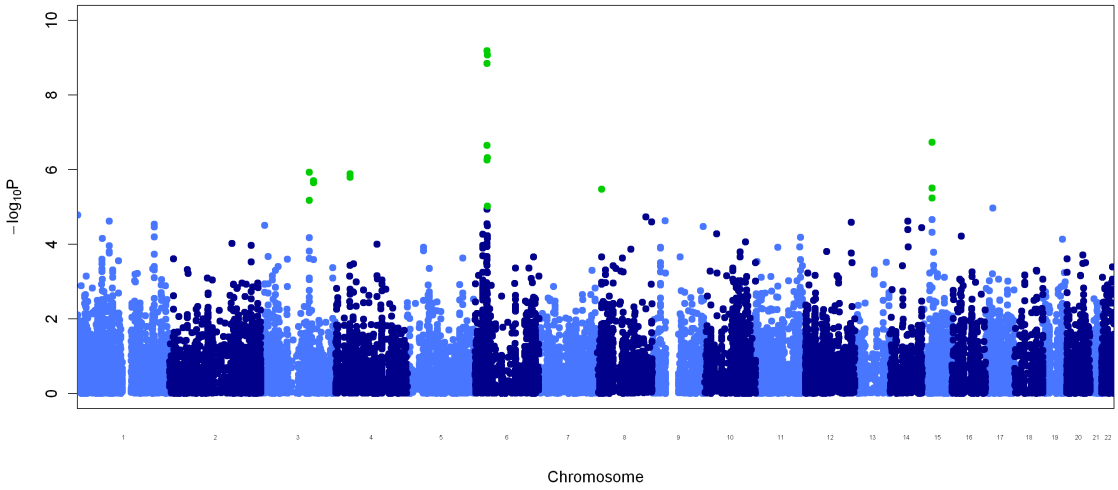


Figure 5.3: Manhattan plot of the leprosy gene-centric study. SNPs with $P < 1.0 \times 10^{-5}$ are highlighted in green

5.5 Replication results

A total of 40 SNPs with $P_{alle} < 1.0 \times 10^{-4}$ were successfully genotyped in the replication cohorts. The allelic test was used for the association test in the case-control cohorts, and the transmission-disequilibrium-test (TDT) was used for the association test in the family-based cohort.

Several SNPs showed evidence of association in the replication cohorts in the same direction as in the primary study (Table 5.3). The most convincing evidence of association was observed in chromosome 6, where 3 SNPs at the *HLA* class II locus (rs9270650, rs1071630, rs9270986) and 1 SNP at *RNF5* (rs3134947) showed consistent evidence of association. Combining results from the two case-control populations, the strongest signal was observed in rs1071630 at *HLA-DQA1* (*HLA-DQA1* F41S) with a overall $P = 4.9 \times 10^{-14}$ in the Mantel-Haenszel statistic (Figures 5.4 & 5.5). The Breslow-Day test showed no significant heterogeneity between these studies ($P =$

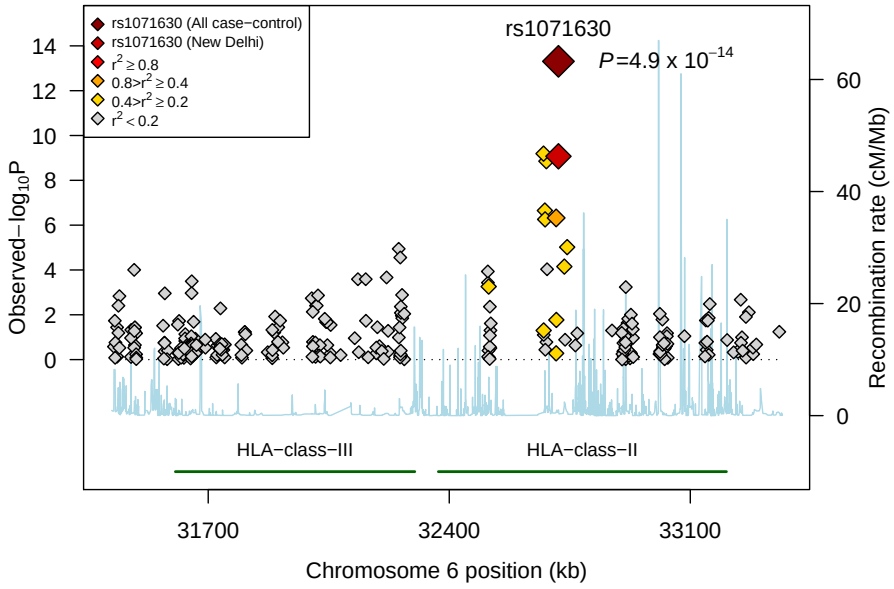
0.686). The minor allele significantly decreases the odds of infection (OR=0.43, 95% CI=0.35-0.54). The second strongest association was observed at rs9270650 in *HLA-DRB1* ($P_{mh} = 3.1 \times 10^{-11}$, OR=2.24, 95% CI=1.76-2.86, Figure 5.5).

Consistent association was also observed at the Toll-like receptor (*TLR*) gene cluster at chromosome 4p14. The association peaked at a non-synonymous variant rs5743618 (*TLR1* I602S) where the minor allele was significantly associated protection against leprosy. The association statistic was $P_{alle} = 1.3 \times 10^{-6}$ and $P_{trend} = 3.9 \times 10^{-6}$ for the New Delhi cohort, with the genotypes distributed in accordance to the Hardy-Weinberg equilibrium ($P_{case} = 0.565$, $P_{ctrl} = 0.088$). The association statistic was $P_{mh} = 5.7 \times 10^{-8}$ (OR=0.31, 95% CI=0.20-0.48) upon addition of data from the Kolkata case-control population, and a consistent effect was observed in the Madurai family cohort ($P_{tdt} = 0.090$, OR=0.51, 95% 0.35-1.09). Analysed together with a previously reported Turkish set of 57 cases and 90 controls (Johnson et al., 2007), the best estimate of risk for the minor serine variant was OR=0.37 (95% CI=0.26-0.51) with $P_{mh} = 1.7 \times 10^{-9}$ for all the case-control studies (Figure 5.4 & 5.5). The Breslow-Day test showed no significant heterogeneity between these studies ($P = 0.243$). This association was strongest among all 75 SNPs genotyped in the 80kb flanking region.

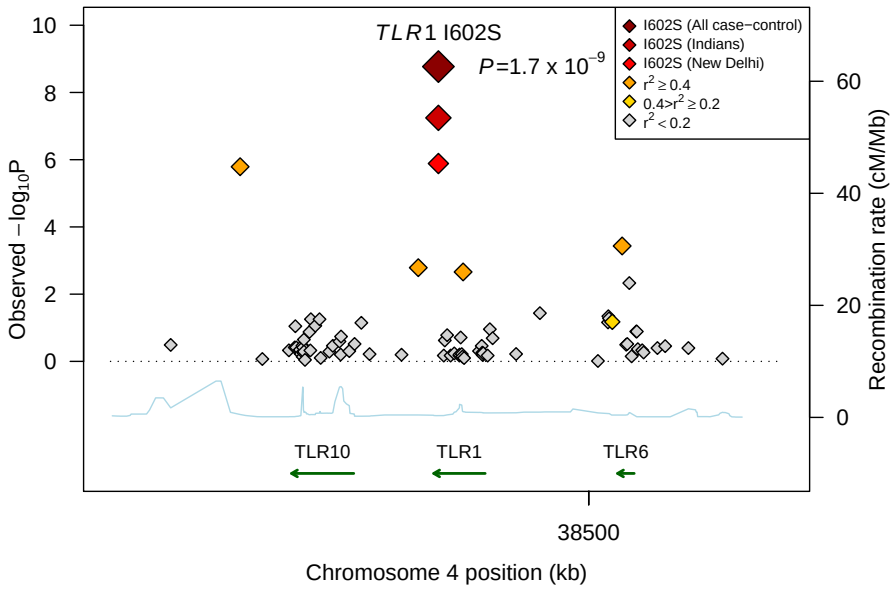
Some evidence of association was observed for SNPs at *CCL27* (rs4879816) and *PRKCB1* (rs1013316) with $P < 0.05$ in one of the replication cohorts. The gene *CCL27* encodes a cutaneous chemokine that attracts T-cells. However, none of these associations was consistent among all the three populations.

SNP	Chr	Position	Gene	New Delhi	Kolkata			Madurai		
				OR	OR	95% CI	P_{alle}	OR	95% CI	P_{tdt}
rs3753242	1	2059541	*PRKCZ*	2.10	Not genotyped			Not genotyped		
rs3790562	1	67567011	*IL12RB2*	2.82	0.62	0.35-1.10	0.102	1.19	0.84-1.70	0.325
rs6673833	1	85648210	*DDAH1*	0.49	0.96	0.65-1.43	0.846	0.79	0.61-1.02	0.067
rs2244510	1	204661692	*SRGAP2*	0.55	0.99	0.69-1.41	0.944	0.98	0.74-1.29	0.888
rs9429893	1	204667615	*SRGAP2*	0.56	0.98	0.69-1.41	0.928	1.05	0.81-1.36	0.694
rs11118087	1	204723200	*IKBKE*	1.82	1.12	0.79-1.57	0.528	1.08	0.87-1.34	0.506
rs6707461	2	164817572	-	0.57	Not genotyped			Not genotyped		
rs11465897	3	10229797	*IRAK2*	0.46	0.88	0.51-1.52	0.653	0.91	0.62-1.34	0.622
rs664910	3	128956720	*MGLL*	1.73	0.98	0.71-1.37	0.917	0.96	0.74-1.26	0.788
rs16830415	3	128956957	*MGLL*	2.35	1.05	0.67-1.65	0.834	0.92	0.70-1.22	0.572
rs547801	3	128964929	*MGLL*	2.39	Not genotyped			Not genotyped		
rs295470	3	140696593	*RBP1*	1.90	1.11	0.77-1.59	0.577	1.10	0.86-1.42	0.444
rs2071388	3	140719373	*RBP1*	1.90	0.87	0.61-1.24	0.441	1.03	0.79-1.33	0.842
rs10008492	4	38442115	*TLR10*	0.23	0.50	0.25-1.00	0.047	1.00	0.61-1.63	1.000
rs5743618	4	38475043	*TLR1*	0.27	0.40	0.20-0.83	0.011	0.61	0.35-1.09	0.090
rs10033900	4	110878516	*PLA2G12A*	0.57	0.86	0.60-1.24	0.424	1.10	0.86-1.39	0.460
rs9358391	6	21215088	*CDKAL1*	1.78	1.01	0.72-1.41	0.968	1.02	0.80-1.31	0.851
rs3134947	6	32253183	*RNF5*	0.41	0.54	0.35-0.85	0.007	0.68	0.48-0.98	0.038
rs3134940	6	32257794	*AGER*	0.41	0.71	0.38-1.33	0.279	0.69	0.45-1.08	0.100
rs9270650	6	32673832	*HLA-DRB1*	2.43	1.83	1.28-2.59	0.001	1.30	1.03-1.65	0.030
rs477515	6	32677669	*HLA-DRB1/DQA1*	0.41	Not genotyped			Not genotyped		
rs2516049	6	32678378	*HLA-DRB1/DQA1*	0.42	0.63	0.33-1.17	0.142	0.78	0.56-1.08	0.131
rs9270986	6	32682038	*HLA-DRB1/DQA1*	2.45	1.66	1.18-2.34	0.003	1.28	1.00-1.64	0.046
rs482044	6	32684042	*HLA-DRB1/DQA1*	0.58	Not genotyped			Not genotyped		
rs3104369	6	32710460	*HLA-DQA1*	2.01	Not genotyped			Not genotyped		
rs1071630	6	32717104	*HLA-DQA1*	0.42	0.46	0.32-0.65	1.1×10^{-5}	0.81	0.65-1.00	0.050
rs9273363	6	32734250	*HLA-DQA1*	0.50	Not genotyped			Not genotyped		
rs3891175	6	32742445	*HLA-DQA1*	0.43	Not genotyped			Not genotyped		
rs210137	6	33650456	*BAK1*	1.76	1.10	0.77-1.56	0.594	1.05	0.83-1.34	0.671
rs210138	6	33650516	*BAK1*	1.79	0.98	0.71-1.37	0.923	1.01	0.81-1.24	0.957
rs7009163	8	11714373	*FDFT1*	0.34	Not genotyped			Not genotyped		
rs16902359	8	128812033	*MYC*	1.86	1.02	0.71-1.48	0.910	1.04	0.79-1.36	0.785
rs1799998	8	143996602	*CYP11B2*	0.55	1.04	0.72-1.49	0.836	1.06	0.84-1.33	0.641
rs4879816	9	34660128	*CCL27*	0.43	1.42	0.82-2.46	0.211	0.58	0.42-0.80	0.001
rs3740237	10	32597598	*EPC1*	1.87	1.19	0.82-1.73	0.359	0.94	0.73-1.22	0.647
rs17121799	10	108697641	*SORCS1*	2.39	0.73	0.45-1.18	0.195	0.96	0.68-1.34	0.795
rs666004	11	120903788	*SORL1*	0.58	1.42	1.03-1.98	0.035	1.02	0.82-1.27	0.865
rs525017	12	121781687	*GPR81*	2.06	0.61	0.32-1.15	0.122	1.24	0.73-2.10	0.423
rs8006042	14	66899325	*EIF2S1*	1.89	1.25	0.77-2.03	0.362	0.60	0.38-0.94	0.025
rs12588458	14	66900609	*EIF2S1*	1.87	1.37	0.94-2.00	0.106	0.83	0.64-1.09	0.178
rs861537	14	103236828	*XRCC3*	0.57	0.91	0.61-1.36	0.648	0.92	0.71-1.20	0.541
rs2675345	15	46187491	*SLC12A1*	2.09	1.26	0.89-1.79	0.192	0.85	0.68-1.05	0.127
rs1320052	15	46282800	*SLC12A1*	2.19	1.29	0.88-1.91	0.193	0.82	0.66-1.01	0.064
rs16960661	15	46293975	*SLC12A1*	2.51	1.19	0.81-1.77	0.377	0.79	0.63-1.00	0.050
rs9920281	15	46301601	*SLC12A1*	2.53	1.10	0.76-1.58	0.616	0.82	0.65-1.03	0.089
rs8032941	15	46302144	*SLC12A1*	2.83	1.47	0.73-2.93	0.278	0.74	0.52-1.04	0.082
rs1013316	16	24042613	*PRKCB1*	0.58	0.70	0.49-0.98	0.039	1.25	1.00-1.55	0.046
rs7501702	17	19234320	*MFAP4*	0.53	0.99	0.67-1.46	0.950	1.05	0.82-1.33	0.715
rs11083841	19	51811873	*PTGIR*	1.73	1.04	0.73-1.48	0.850	1.23	0.99-1.53	0.066

Table 5.3: Association statistics in the replication studies

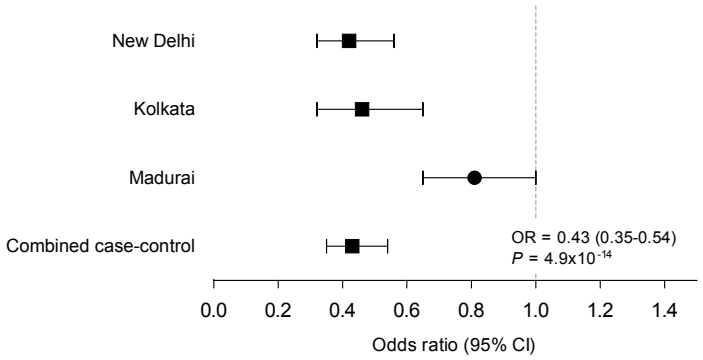


(a) *HLA-DRB1/DQA1*

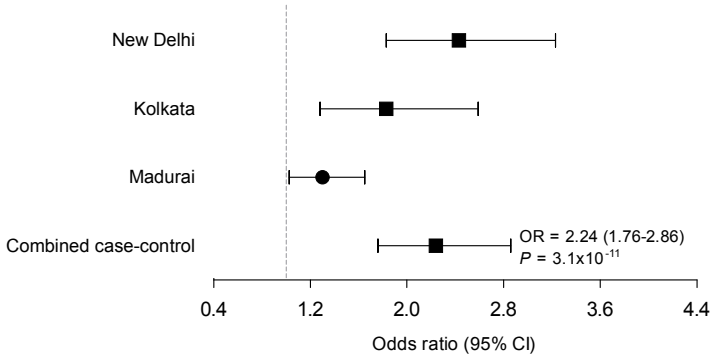


(b) *TLR 10/1/6*

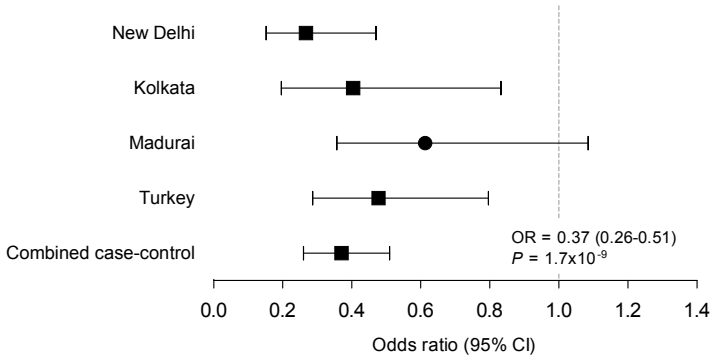
Figure 5.4: Association plots of the *HLA* class II and *TLR 10/1/6* loci



(a) rs1071630



(b) rs9270650



(c) rs5743618

Figure 5.5: Forest plots showing associations of rs1071630, rs9270650 and rs5743618 in different populations

5.6 The *HLA* class II locus

5.6.1 Conditional analysis

Significant associations showing evidence of replication were observed in four SNPs around the *HLA* locus, three at *HLA-DRB1/DQA1* and one at *RNF5*. Although the *RNF5* SNP is located outside the *HLA* region, its close proximity raises the possibility of its signal being related to the *HLA* associations. Linkage disequilibrium analysis showed that this *RNF5* SNP (rs3134947) was in partial LD with SNPs in the *HLA* class II locus (rs3134947-rs2516049, $r^2 = 0.50$ at 425kb away and rs3134947-rs1071630, $r^2 = 0.15$ at 464 kb away), and logistic regression showed that the association at rs3134947 became non-significant after conditioning upon rs2516049 and rs1071630 ($P = 0.380$ and $P = 0.077$). This suggests that the association at *RNF5* might be due to signal arising from the *HLA* class II locus, consistent with the existence of long haplotypes extending up to one million base pairs in *HLA* (de Bakker et al., 2006).

Further logistic regression analysis was carried out to study the signal at the three replicated SNPs in the *HLA* class II region. Analysis showed that the associations at rs1071630 and rs9270650 did not account for each other ($P = 0.001$ and $P = 0.004$). However, the signal at rs9270986 might be statistically accounted by the rs9270650 association ($P = 0.711$). In other words, the data suggests that the associations at this *HLA* locus might be due to two independent signals best captured by rs1071630 and rs9270650.

5.6.2 Haplotypic analysis

Two haplotype blocks were reconstructed in the *HLA* locus II region using the confidence intervals definition (Table 5.4) (Gabriel et al., 2002). The first block contained

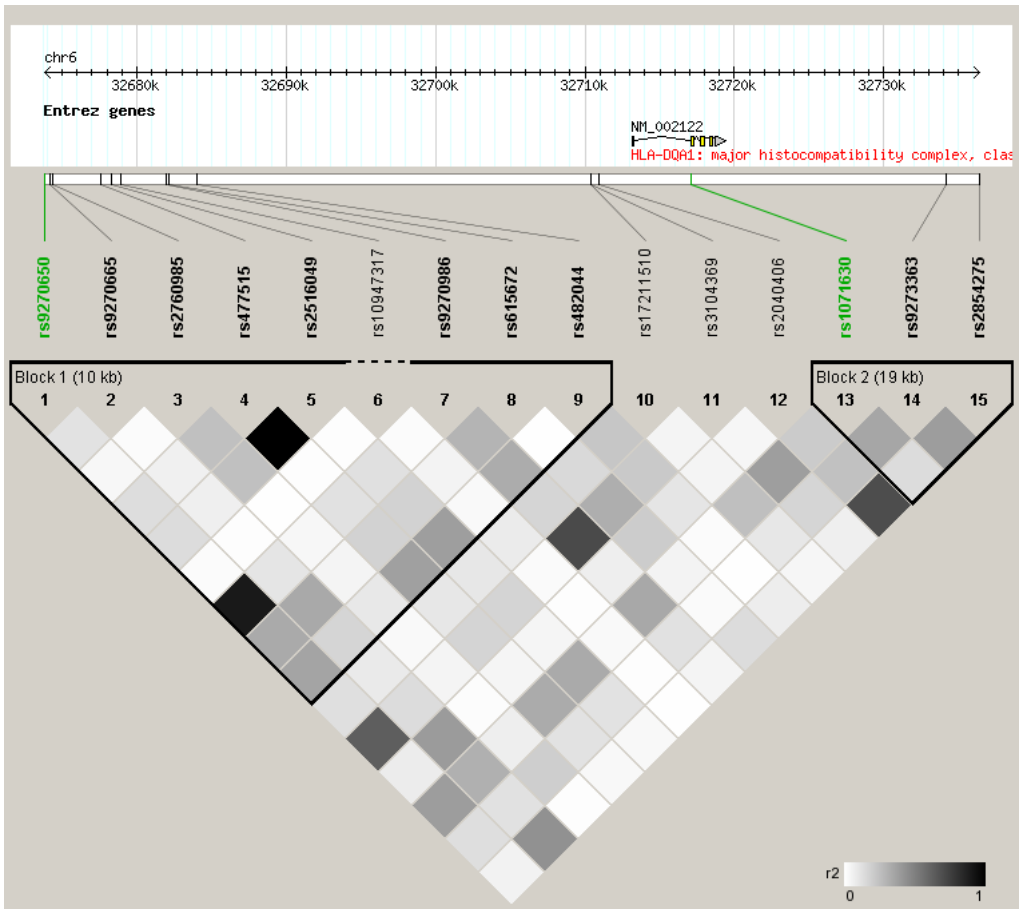


Figure 5.6: LD plot of the *HLA* class II locus. The blocks indicate the reconstructed haplotypes containing the top SNPs rs1071630 and rs9270650

eight common haplotypes with a frequency greater than 1%, three of which were significantly associated with susceptibility to leprosy at $P < 0.05$. Increased risk to leprosy was observed in the most common haplotype carrying the minor C allele for rs9270650 (41.1% in cases and 36.5% in controls), whereas significant protection was observed in two haplotypes containing its alternative allele (Table 5.4). The second block contained four common haplotypes, three of which were associated with leprosy susceptibility. These results were highly consistent with the single locus associations, and support the evidence that variations in this locus contribute to the disease phenotype.

Haplotype	Frequency	Odds ratio	95% CI	P-value
Block 1				
CGGCAAGC	0.311	2.43	1.82-3.24	1.5×10^{-9}
TTGCACCC	0.182	0.69	0.49-0.98	0.037
TGGCACCG	0.151	1.27	0.88-1.83	0.197
TGGTGCGG	0.147	0.38	0.26-0.56	1.4×10^{-6}
TGGCACCC	0.067	0.78	0.46-1.33	0.364
TGGCACGG	0.059	0.71	0.41-1.24	0.231
TGATGCGG	0.043	0.58	0.32-1.05	0.071
CGGCACGC	0.023	1.43	0.59-3.49	0.432
Block 2				
TCG	0.574	2.26	1.73-2.96	3.2×10^{-9}
CCG	0.222	0.59	0.42-0.82	0.002
CAG	0.116	0.45	0.29-0.69	3.0×10^{-4}
CAT	0.088	0.68	0.42-1.09	0.108

Table 5.4: Haplotypic association analysis at the *HLA* class II locus. The haplotype blocks were defined as in Figure 5.6

5.6.3 Subtype analysis

The clinical manifestion of leprosy can be broadly categorised into multibacillary and paucibacillary subtypes. The association of rs1071630 and rs9270650 remained robust in each of the two subtypes with similar effect size. The Breslow-Day test of heterogeneity did not reject the null hypothesis of homogeneity between the multibacillary and paucibacillary groups ($P > 0.05$, Table 5.5).

5.6.4 Allelic dosage analysis

Allelic dosage analysis was performed to look at the relationship between number of risk alleles and disease phenotype. The rationale of this analysis relied on the hypothesis that if the two SNPs interact in an additive manner, the odds of leprosy infection may increase with the number of risk alleles in a linear fashion, which could be demonstrated by the Cochran-Armitage test of trend. The number of risk alleles in these two SNPs carried by each individual was counted. Analysis showed that the

Leprosy subtype	OR (95% CI)	Allelic P-value
rs1071630		
Multibacillary leprosy	0.37 (0.26-0.53)	2.0×10^{-8}
Paucibacillary leprosy	0.48 (0.33-0.68)	3.0×10^{-5}
Combined	0.42 (0.32-0.56)	8.5×10^{-10}
$P_{Breslow-Day} = 0.33$		
rs9270650		
Multibacillary leprosy	2.27 (1.61-3.20)	1.6×10^{-6}
Paucibacillary leprosy	2.61 (1.84-3.69)	3.8×10^{-8}
Combined	2.43 (1.83-3.23)	6.4×10^{-10}
$P_{Breslow-Day} = 0.58$		
rs5743618		
Multibacillary leprosy	0.22 (0.10-0.50)	1.3×10^{-6}
Paucibacillary leprosy	0.31 (0.15-0.64)	9.1×10^{-4}
Combined	0.27 (0.15-0.47)	1.3×10^{-6}
$P_{Breslow-Day} = 0.55$		

Table 5.5: Leprosy subtype analysis for rs1071630, rs9270650 and rs5743618

number of risk alleles (ranged 0-4) was positively correlated with susceptibility to leprosy ($P_{trend} = 3.2 \times 10^{-11}$, $P_{logistic} = 1.1 \times 10^{-11}$). Individuals with homozygous risk alleles for both variants were highly susceptible to leprosy infection, with an odds ratio of 8.7 compared to their homozygous protective counterparts (Figure 5.7).

5.6.5 Epistasis analysis

Epistasis analysis was performed to look at the interaction (SNP1 x SNP2) between rs9270650 and rs1071630. The analysis statistic was not significant ($P = 0.44$), suggesting that rs9270650 and rs1071630 might not have a synergistic effect with each other on top of their summed independent effects.

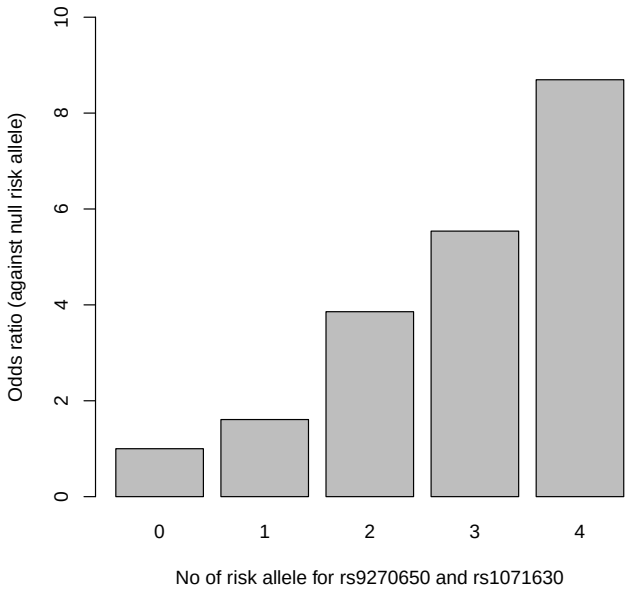


Figure 5.7: Odds of leprosy in individuals carrying different numbers of risk allels for rs9270650 and rs1071630

5.7 The *TLR 10/1/6* locus

5.7.1 Conditional analysis

Eight SNPs showed evidence of association at the *TLR 10/1/6* locus in chromosome 4p14 in the New Delhi cohort. To dissect the association signal, conditional logistic regresion was performed with correction for the SNP rs5743618 which associated with leprosy susceptibility most strongly. Analysis showed that the signals of all other associated SNPs in this region were lost after correcting for rs5743618, suggesting that these SNPs did not contribute to the association signal independently.

5.7.2 Haplotypic analysis

Haplotypes at the associated region at *TLR 10/1/6* were reconstructed using the confidence interval definition (Figure 5.8) (Gabriel et al., 2002). This gave six com-

Haplotype	Frequency	Odds ratio	95% CI	*P*-value
CTTTAT	0.429	1.23	0.94-1.60	0.125
CCATAC	0.172	0.89	0.64-1.24	0.495
CTAGAC	0.130	1.22	0.82-1.82	0.331
CTTTAC	0.088	1.13	0.71-1.79	0.604
CTTTCT	0.087	0.26	0.15-0.49	1.3×10^{-6}
GTAGAC	0.085	1.41	0.89-2.24	0.145

Table 5.6: Haplotypic association analysis at the *TLR 10/1/6* locus. The haplotype block was defined as in Figure 5.8

mon haplotypes containing rs5743618, with one haplotype carrying the protective C allele conferring significant protection against leprosy ($P = 1.3 \times 10^{-6}$, OR=0.26, 95% CI=0.15-0.47). This was consistent with the result from the single locus association.

5.7.3 Subtype analysis

The rs5743618 association was significant for both multibacillary and paucibacillary subtypes (Table 5.5), with a slightly stronger effect in multibacillary leprosy. However the Breslow-Day test of heterogeneity did not reveal any significant difference between the two subtypes ($P = 0.55$).

5.7.4 Population differentiation analysis

Infectious disease has been a major selective force during human evolution (Haldane, 1949; Lively and Dybdahl, 2000; Sabeti et al., 2006) and has contributed to the diversity of the mammalian *MHC* genes (Hughes and Yeager, 1998). Recently, several studies have shown that the *TLR 10/1/6* gene cluster is highly differentiated between populations (Smyth et al., 2006; The Wellcome Trust Case Control Consortium, 2007; Pickrell et al., 2009). To further investigate the population differentiation at this locus, the SNP rs5743618 was genotyped in the HapMap populations. Analysis showed extreme population differentiation of rs5743618 among over 3 million

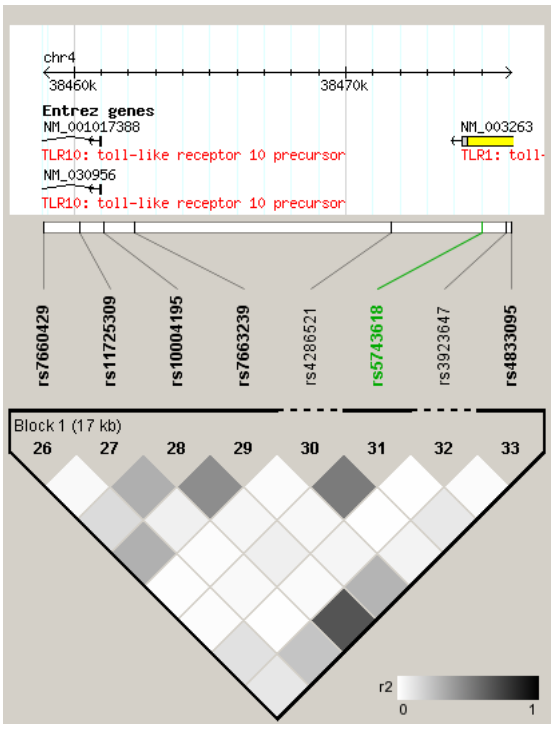


Figure 5.8: LD plot of the *TLR1* locus. The block indicates the reconstructed haplotype containing the top SNP rs5743618

polymorphic SNPs in the genome ($> 99^{th}$ percentile, Figure 5.9). Further literature review and genotyping of samples from India, Malawi, The Gambia, Russia and the United Kingdom showed a maximal pairwise F_{ST} of 0.57 between European American and Yoruban (Figure 5.10). The derived C allele was rare in African populations but was dominant in individuals of European descent, with more than 40% being homozygous for this allele (Figures 5.11).

5.8 Previous candidates for leprosy susceptibility

This microarray allowed genic SNPs to be interrogated at a high density. As such, SNPs located in genes previously reported to associate with susceptibility to leprosy were re-investigated in this New Delhi cohort. Evidence of association was observed at genes on chromosome 6 flanking *MICA*, *TNF*, *MICB* and *IL10*, which remained

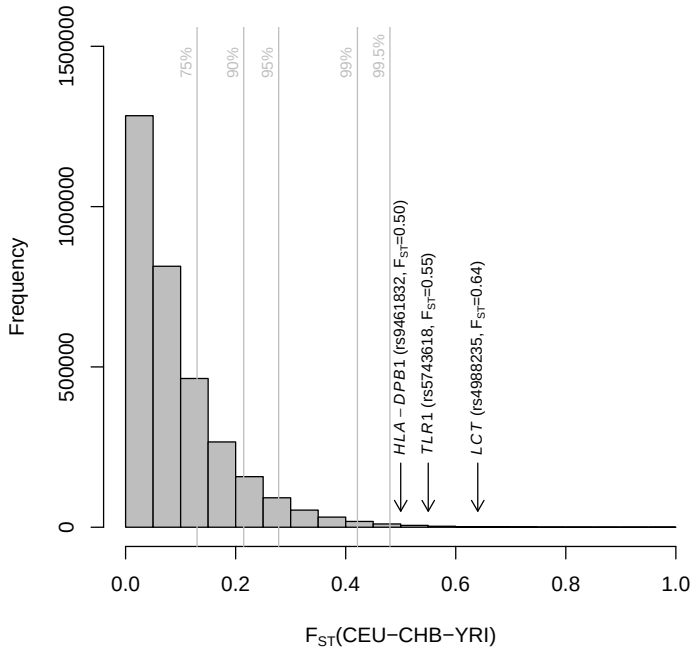


Figure 5.9: F_{ST} value of rs5743618 against more than three million autosomal SNPs in the HapMap populations

significant after Sidak's correction for multiple testing (Table 5.7) and conditioning for the *HLA* variants rs9270650 and rs1071630. However, no convincing association was observed in SNPs flanking *LTA* or *TLR4* despite the high density of coverage across these genes.

5.8.1 Genes from a recent genome-wide study

During the course of this study, a GWA study on leprosy in a Chinese population of 706 cases and 1,225 controls was published (Zhang et al., 2009). The data identified a consistent signal of association at the *HLA* class II locus, and also implicated the genes *NOD2*, *TNFSF15*, *RIPK2* and *C13orf31/CCDC122* in leprosy susceptibility. To attempt to replicate these findings, we genotyped the 15 SNPs in the non-*HLA* regions reported in the study (Zhang et al., 2009). Consistent associations for SNPs

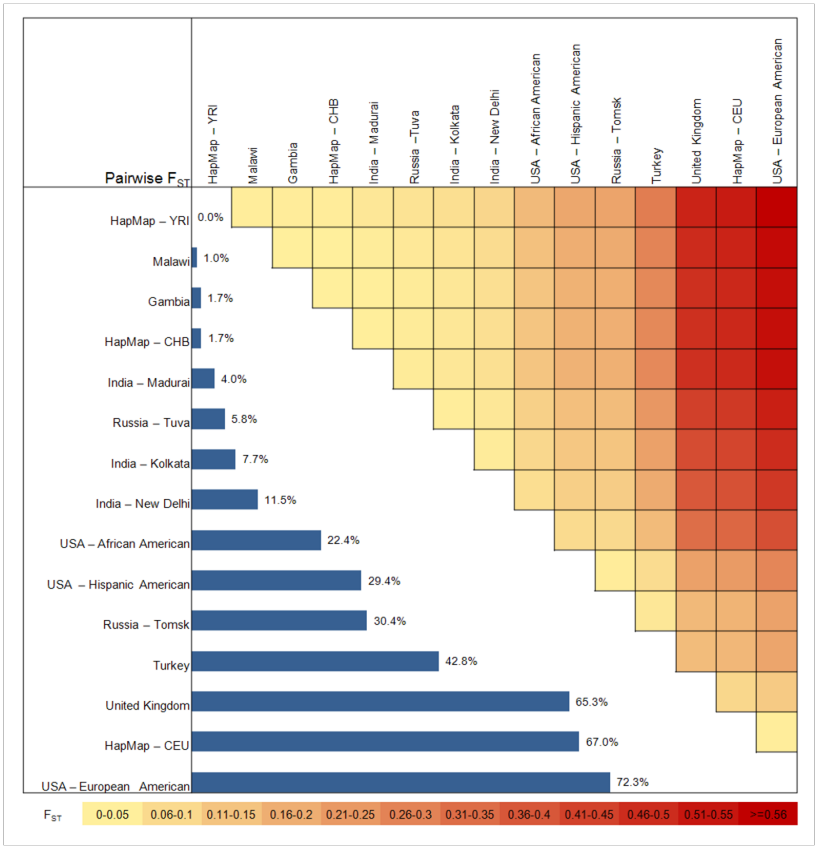


Figure 5.10: Pairwise F_{ST} of rs5743618 between the 15 populations as indicated

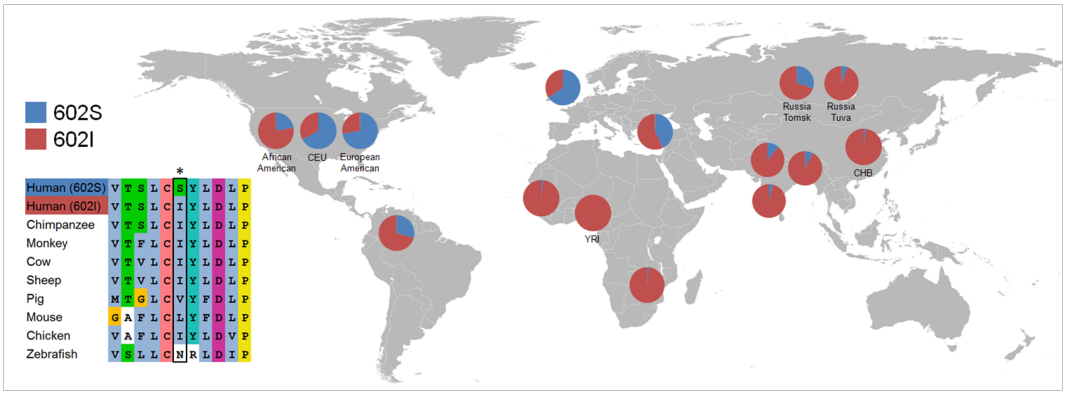


Figure 5.11: Global amino acid frequency corresponding to rs5743618. Side panel shows the amino acid sequence alignment around the SNP

Chr	Gene	Size (kb)	SNPs	Density	Best SNP	OR (95% CI)	P_{alle}	P_{Sidak}
1	*IL10*	4.89	13	2.66	rs1554286	1.52 (1.15-1.99)	0.003	0.039
2	*SLC11A1*	14.87	9	0.61	rs13062	1.39 (1.03-1.88)	0.032	0.254
4	*TLR2*	21.80	8	0.37	rs1339	0.72 (0.47-1.09)	0.116	0.627
5	*IL12B*	15.69	15	0.96	rs919766	2.82 (1.22-6.52)	0.011	0.152
6	*MICA*	11.72	13	1.11	rs12660741	2.15 (1.45-3.18)	9.9×10^{-5}	0.001
6	*MICB*	13.05	8	0.61	rs3828917	0.32 (0.16-0.66)	0.001	0.008
6	*LTA*	2.01	15	7.48	rs1121800	1.20 (0.92-1.56)	0.190	0.958
6	*TNF*	2.76	15	5.43	rs3093662	0.33 (0.17-0.62)	3.2×10^{-4}	0.005
9	*TLR4*	13.31	28	2.10	rs7044464	1.57 (1.05-2.35)	0.029	0.561
12	*VDR*	63.50	62	0.98	rs2254210	0.6 (0.41-0.88)	0.008	0.392
16	*NOD2*	35.94	25	0.70	rs2067085	0.65 (0.46-0.91)	0.011	0.222
19	*IL12RB1*	27.33	9	0.33	rs436857	0.77 (0.51-1.16)	0.207	0.876

Table 5.7: Association statistics of reported leprosy susceptibility candidates

at the *C13orf31/CCDC122* at chromosome 13q14.11 were observed, with the SNP rs3764147 showing a consistent effect in the New Delhi ($P_{alle} = 0.003$, OR=1.54, 95% CI=1.16-2.04) and Kolkata ($P_{alle} = 0.064$, OR=1.36, 95% CI=0.98-1.87) cohorts, although the family cohort did not detect any significant association ($P_{tdt} = 0.168$) probably due to its low power. The distribution of genotypes in the case-control populations followed the Hardy-Weinberg equilibrium ($P_{case} = 0.495$, $P_{ctrl} = 0.956$ for New Delhi, $P_{case} = 0.243$ and $P_{ctrl} = 0.069$ for Kolkata). The best effect size estimate for the combined case-control cohorts was OR=1.46 ($P_{mh} = 5.2 \times 10^{-4}$, 95% CI=1.18-1.80), compared with OR=1.68 in the original study. However, significant association was not detected in the reported SNPs in other genes (*NOD2*, *RIPK2*, *TNFSF15* and *LRRK2*).

5.9 Discussion

Despite the long history of leprosy and medical advances, leprosy remains a scourge in many parts of the world. Our understanding of the pathogenesis of *M. leprae* and its interaction with humans is limited, in part due to the inability to culture the

SNP	Gene	Original	New Delhi		Kolkata		Madurai	
		OR	OR (95% CI)	P_{alle}	OR (95% CI)	P_{alle}	OR (95% CI)	P_{alle}
rs42490	*RIPK2*	0.76	0.78 (0.60-1.03)	0.080	0.86 (0.62-1.19)	0.351	1.20 (0.84-1.73)	0.313
rs40457	*RIPK2*	0.77	0.83 (0.63-1.11)	0.209	1.04 (0.74-1.46)	0.829	1.04 (0.71-1.53)	0.845
rs10982385	*TNFSF15*	1.19	0.90 (0.69-1.18)	0.433	1.23 (0.89-1.69)	0.207	1.36 (0.91-2.02)	0.132
rs4574921	*TNFSF15*	1.31	1.04 (0.76-1.41)	0.818	1.21 (0.83-1.76)	0.319	0.83 (0.56-1.24)	0.361
rs10114470	*TNFSF15*	1.28	0.95 (0.71-1.26)	0.721	0.83 (0.59-1.17)	0.290	1.13 (0.78-1.63)	0.514
rs6478108	*TNFSF15*	1.37	1.04 (0.78-1.39)	0.782	0.82 (0.57-1.16)	0.258	1.29 (0.87-1.90)	0.200
rs1873613	*LRRK2*	0.86	0.75 (0.57-0.98)	0.038	1.02 (0.75-1.40)	0.881	1.19 (0.83-1.71)	0.353
rs9533634	*CCDC122*	0.76	0.65 (0.50-0.85)	0.001	0.86 (0.63-1.18)	0.355	1.16 (0.83-1.64)	0.384
rs3088362	*CCDC122*	1.52	1.35 (1.00-1.82)	0.051	1.25 (0.89-1.75)	0.205	1.04 (0.7-1.56)	0.837
rs3764147	*C13orf31*	1.68	1.54 (1.16-2.04)	0.003	1.36 (0.98-1.87)	0.064	1.34 (0.88-2.04)	0.168
rs10507522	*C13orf31*	0.68	1.04 (0.75-1.44)	0.813	0.85 (0.59-1.22)	0.367	1.21 (0.79-1.86)	0.383
rs9302752	*NOD2*	1.59	0.83 (0.63-1.09)	0.187	0.95 (0.68-1.32)	0.744	1.30 (0.89-1.90)	0.178
rs7194886	*NOD2*	1.63	Not genotyped		Not genotyped		Not genotyped	
rs8057341	*NOD2*	1.17	0.85 (0.64-1.11)	0.228	0.87 (0.62-1.23)	0.436	1.29 (0.87-1.90)	0.200
rs3135499	*NOD2*	1.16	0.83 (0.60-1.15)	0.264	0.58 (0.39-0.86)	0.007	1.16 (0.75-1.80)	0.502

Table 5.8: Association statistics of the SNPs reported in the recent genome-wide study (Zhang et al., 2009) in the Indian populations

bacterium *in vitro*. For this reason, there is greater reliance on genetics to understand the disease and the causative pathogen. Some progress has been made to identify genetic loci which associate with susceptibility to leprosy, however the number of candidates showing consistent replication across populations are few. This chapter describes a population based case-control association study surveying genetic variants in more than 2,000 genes throughout the genome.

This gene-centric study identified *HLA-DRB1/DQA1* and *TLR1* as major susceptibility loci for leprosy. The human *HLA* class II genes are expressed by antigen-presenting cells to present peptide antigens to activate T-cell immune response. The locus showed the most convincing evidence of association among all genetic loci tested in our microrray, confirming previous association results (Chiewsilp et al., 1979; Miyanaga et al., 1981; van Eden et al., 1982, 1985; Visentainer et al., 1997). This is also consistent with results from the recently published GWA study on leprosy (Zhang et al., 2009). With the large effect sizes indicated by the high odds ratio of the variants, it is justified that this *HLA* locus is a major determinant of disease

status. Importantly, the association analysis identified two distinct SNPs neither of which could explain the signal at this locus alone. Given the extreme polymorphism of the *HLA* and the results of the conditional logistic analysis, the observed association is likely to be contributed by multiple functional interacting variants tagged by rs1071630 and rs9270650.

Apart from the *HLA* locus, we also observed convincing evidence of association with *TLR1* in susceptibility to leprosy. The gene *TLR1* encodes for the TLR1 receptor, part of the Toll-like receptor (TLR) family which senses pathogen associated patterns present on the surface of pathogens (Akira et al., 2001). Upon activation, TLRs activates a downstream signaling cascade leading to activation of nuclear factor kappa-B (NF-κB), which promotes transcription of pro-inflammatory genes. Previous studies have suggested that TLR1 interact with TLR2 to recognise bacterial lipopeptides (Takeuchi et al., 2002; Jin et al., 2007) and mediate the immune response to *M. leprae* (Krutzik et al., 2003).

The results presented here further advance our understanding of the immunity against leprosy. The high density of SNPs in the microarray at the *TLR 10/1/6* locus allow detailed investigation at this locus. Although the genetic association of *TLR1* has been reported in a Turkish population of 57 cases and 90 controls (Johnson et al., 2007), this study is the first to show a consistent and reproducible association between *TLR1* and leprosy at a genome-wide level of statistical significance. The association is unlikely to be spurious, due to the robustness of family studies to population stratification. This association is most consistent after the *HLA* class II locus among the 2,092 genes surveyed. Although this does not exclude the possibility of other genes involved in leprosy, it supports the critical role of the TLR1 pathway in the immunity against *M. leprae* in humans. Moreover, the conditional logistic regression analysis did not detect any other contributing variant, suggesting rs5743618 may

account for the majority if not all of the association signal at this locus. These results provide insight into the genetic architecture of this locus, and are in line with previous studies showing the functional role of this SNP with down-regulation of NF-κB signalling and IL-6 production in individuals with the serine substitution (*TLR1* 602S) (Hawn et al., 2007; Johnson et al., 2007; Misch et al., 2008).

Together with these functional studies, the genetic association results give clues to the pathogenesis of *M. leprae* in humans. Despite the hypofunctional phenotype (Hawn et al., 2007; Johnson et al., 2007; Misch et al., 2008), the allele coding for the serine amino acid (rs5743618G, or *TLR1* I602S) is associated with significant protection to leprosy (OR=0.37). Conversely, the functional isoleucine form is associated with a higher risk. This suggests that *M. leprae* may have utilised TLR1 as part of its pathogenic mechanism, as certain bacteria are known to produce a homologue of the TLR1 signalling domain to subvert the innate immune response (Cirl et al., 2008). Although further work is required to confirm this speculative hypothesis, this suggests that modulation of TLR1 may be a useful and safe treatment in mycobacterial diseases, as more than 40% of individuals of European descent are homozygous for this hypofunctional variant but yet normal in phenotype. This also provides an example of hypofunctional human TLR1 protecting against a common pathogen.

The population differentiation analysis provides insight into the host-pathogen relationship between humans and *M. leprae*. Consistent with previous genome-wide studies (Todd et al., 2007; The Wellcome Trust Case Control Consortium, 2007; Pickrell et al., 2009), the results here show that rs5743618 is extremely differentiated when compared to other SNPs in the genome. Population differentiation can result from local adaptation when geographically separate populations are subject to distinct environmental pressures which change the allele frequency of one population but not another (Sabeti et al., 2006), and may therefore reflect natural selection (Barreiro

et al., 2008). Although population demographics can also give rise to allele frequency differences between populations, given the extrene degree of differentiation and the importance of Toll-like receptors in immunity (Iwasaki and Medzhitov, 2004; Akira and Takeda, 2004), the out-of-African expansion of the hypofunctional serine variant is likely to have occurred by natural selection. This is further supported by recent studies showing clear signatures of selection at this locus (Barreiro et al., 2009; Grossman et al., 2010). Although it is difficult to ascertain the historical selective agents behind this event, the consistent association with a large effect size suggests that *M. leprae* may be one such selection agent. Several studies have associated leprosy with a reduced reproductive fitness (Smith and Guinto, 1978; Saporta and Yuksel, 1994), and the all-cause mortality of lepromatous leprosy was reported to be at least 3.5 times of the general population (Guinto et al., 1954; Noordeen, 1972). These factors, together with the long period of human interaction with *M. leprae* (Monot et al., 2005, 2009) and intense social stigma associated with the disease (Boldsen, 2005), suggest that leprosy may have contributed to the present global distribution of human *TLR1*. The extreme population differentiation suggests that the selection event may have occurred after human dispersion, although the possibility of earlier adaptation cannot be excluded. Other infectious diseases, especially tuberculosis, could have also contributed, given the overlapping antigenic repertoire and immunogenicity of mycobacterial species, despite a lack of association at rs5743618 in a Gambian tuberculosis case-control population due to low allele frequency (MAF=1.72%, $P = 0.52$). Nevertheless, this study provides a plausible cause which drives the selection at this TLR1 locus, and supports the hypothesis that common variants with large effect sizes are likely to be under selection (Goldstein, 2009).

In the last section of this chapter, I have described the effort to replicate findings from a recently published GWA study (Zhang et al., 2009). Both consistent and discrepant results were observed compared to the original study. Reassuringly, both

studies confirmed *HLA* class II as a major leprosy susceptibility locus, a finding consistent with early studies reporting association of *HLA* with leprosy (Chiewsilp et al., 1979; Miyanaga et al., 1981; van Eden et al., 1982, 1985; Visentainer et al., 1997). Consistent findings were also observed for genetic variants at *C13orf31/CCDC122* at chromsome 13q14.11. However, consistent associations were not observed in the genes related to the NOD2 pathway, including *NOD2*, *TNFSF15* or *RIPK2*; nor did the Zhang et al report any notable association at the *TLR1* locus. The disparity at the *TLR1* locus is likely due to the low allele frequency of rs5743618 in the Chinese population (MAF=1.7%). For the other associated genes there could be a number of reasons for the non-replication. The reported effect sizes of variants in these genes are moderate compared to the *HLA* and *C13orf31/CCDC122*. Assuming these effect estimates are real, the present study would have less power to detect these associations. Another possibility is a population specific effect whereby variants in these genes confer susceptibility in the Chinese but not in the Indian populations. Given the significant differences in the genetic composition between the Indian and Chinese (Reich et al., 2009), this may have contributed to the heterogeneous signals between the studies. Furthermore, despite the relative homogenous population of M. leprae, the 16 mycobacterial subtypes highly correlated with the geography can be readily identified (Monot et al., 2009). Given the highly specialised nature of the immune system, these mycobacterial strains may activate the host immune response in different ways, accounting for overlapping yet some discordant results between the studies. Nevertheless, these results provide an independent replication for the association at *C13orf31/CCDC122*. Interestingly this locus, together with *NOD2* and *TNFSF15*, have been implicated in Crohn's disease (Barrett et al., 2008), an inflammatory bowel disease hypothesised to have a mycobacterial cause (Greenstein, 2003; Chamberlin et al., 2007; Pierce, 2009), raising the possibility of shared genetic susceptibility between leprosy and Crohn's disease. Although significant as-

sociations were not observed in *NOD2* and *TNFSF15*, the consistent associations at *C13orf31/CCDC122* supports the hypothesis of a common genetic fingerprint between leprosy and Crohn's disease (Schurr and Gros, 2009). This will be analysed and discussed in more details in the next chapter in Section 6.4.

In conclusion, this chapter shows that *HLA-DRB1/DQA1* and *TLR1* are major genetic determinants of susceptibility to leprosy. It shows a protective effect of a hypofunctional TLR1 variant which has possibly evolved out of adaptation. This highlights the key role of the Toll-like receptors in mediating response to microbial pathogens (Medzhitov and Janeway Jr., 2000; Medzhitov, 2001; Khor et al., 2007) and suggests that modulation of the TLR1 pathway may be valuable in the future treatment of mycobacterial diseases. In addition, I have extended the association of the chromosome 13q14.11 locus identified in China (Zhang et al., 2009) to the Indian populations.

Chapter 6

Integrative analysis in mycobacterial susceptibility

6.1 Introduction

The initial wave of high throughput association studies has identified numerous genetic variants that associate with susceptibility to various diseases (The Wellcome Trust Case Control Consortium, 2007). Many of these studies have increased the number of genetic susceptibility loci for diseases such as diabetes (Sladek et al., 2007; Steinthorsdottir et al., 2007; Dupuis et al., 2010), coronary heart disease (McPherson et al., 2007; Samani et al., 2007; Helgadottir et al., 2008) and Crohn's disease (Parkes et al., 2007; Barrett et al., 2008), and provided novel insight into the molecular mechanisms underlying these diseases (Donnelly, 2008). Nevertheless, these genetic studies only represent the first step in understanding the biology of disease, and functional experiments are required to understand the molecular mechanisms underlying these phenotypes (McCarthy et al., 2008). Such progress was demonstrated in recently published functional work following the initial genetic discovery in coronary artery

disease (McPherson et al., 2007) and Crohn's disease (Cooney et al., 2010).

Despite this progress, functional studies following GWA studies in many cases still represent a daunting task. Many association signals map to genes of unknown functions or unannotated regions in the genome (Eeles et al., 2008; Cho et al., 2009; Zhang et al., 2009), as shown in the chromosome 18q11.2 locus identified in the combined tuberculosis study described in Chapter 4. Prioritising genomic loci, therefore, is important for guiding follow-up studies to unravel the aetiology of complex diseases. With the rapid expansion of public depositories like the International HapMap Project (The International HapMap Consortium, 2003, 2005), the Ensembl Project (Hubbard, 2002), the National Institutes of Health (NIH) database of genotypes and phenotypes (dbGAP) and the European Genome-phenome Archive (EGA), additional data exploration and *in silico* analysis represent rapid and cost-effective ways to provide clues in disease biology and to guide functional studies. The success of GWA studies has largely been fuelled by the International HapMap Project through allele tagging (Gabriel et al., 2002; Barrett and Cardon, 2006) and imputation (Marchini et al., 2007), and the ongoing 1000 Genome Project will likely help in fine mapping disease-causing variants. Such integrated approaches have also been applied to pathway and network analyses, using data from microarray studies and from public depositories to identify important biological pathways affecting human traits (Kuijl et al., 2007; Yang et al., 2009; König et al., 2010).

This chapter describes efforts to further pursue the role of host genetics in mycobacterial diseases, by exploring and integrating datasets with a more functional perspective. It also illustrates how genetic data from genetic association experiments can be analysed by incorporating data from existing knowledge on biological pathways. This chapter follows the results of the tuberculosis and leprosy studies, to look for evidence of shared susceptibility and to identify common genetic loci affecting susceptibility to

mycobacteria in general. This is followed by *in silico* functional work, which includes gene ontology, pathway and network analyses using publicly available bioinformatic tools. Finally, using the WTCCC Crohn's disease data, I assess the relationships between these mycobacterial diseases and Crohn's disease, based on existing evidence for their shared susceptibility. Collectively, this work helps to pinpoint important genomic regions that may guide future functional studies to elucidate the mechanisms underlying human immunity to mycobacteria.

6.2 Identifying genes in mycobacterial diseases

6.2.1 Introduction

Despite tuberculosis and leprosy being the two most common mycobacterial diseases in humans, there are few genetic studies on mycobacteria targeting both diseases. Although this is partly due to their different geographical distributions and distinct clinical phenotypes, such combined studies could allow dissection of anti-mycobacterial immunity and may open new fields of preventive and therapeutic measures to combat mycobacterial diseases (Remus et al., 2003). In this section I present evidence for their shared susceptibility and identify genetic loci associated with mycobacterial susceptibility as a single phenotype.

There are several reasons to suggest that there may be an immunological overlap between tuberculosis and leprosy. Despite the distinct clinical manifestations, the majority of cases present with a chronic, indolent disease phase, due to the thick mycobacterial cell wall conferring the slow growing property of both *M. tuberculosis* and *M. leprae*. The cell wall is rich in mycolic acid and expresses common antigens that show potent immunological reactivity (Melancon-Kaplan et al., 1988) including lipoarabinomannan and muramyl dipeptide (Means et al., 1999; Inohara et al., 2003).

Moreover, mycobacteria are able to secrete proteins such as the antigen 85 complex proteins and the ESAT6 proteins, which are highly conserved (Hermans et al., 1995; Ronning et al., 2000, 2004) and are likely to be immunodominant in activating T-cell immunity (Launois et al., 1993). Therefore, it is possible that *M. tuberculosis* and *M. leprae* may stimulate host immunity with homologous components (Husson and Young, 1987; Young et al., 1992). Furthermore, antibodies reacting to both pathogens have been identified (Kolk et al., 1984; Shinnick et al., 1987), and the BCG vaccine has been shown to be also effective in protecting against leprosy (The Karonga Prevention Trial Group, 1996; Setia et al., 2006). These are consistent with a hypothesis that overlapping genetic loci exist which associate with both tuberculosis and leprosy, and that aggregate analysis of these two diseases may help to dissect human anti-mycobacterial immunity.

6.2.2 Methods

The study was performed using the combined Gambian and Ghanaian dataset for tuberculosis susceptibility (Chapter 4) and the Indian dataset for leprosy susceptibility (Chapter 5). As these studies were carried out on different populations and were genotyped on different microarrays, direct comparison of single SNP association statistics was not feasible. This is primarily because of the distinct LD patterns observed in different populations, resulting in different allele tagging and therefore incomparable signals of association. In addition, as the SNPs were genotyped using different platforms, a SNP based comparison would necessitate discarding a significant proportion of the genotype data, resulting in loss of power. Gene-based analysis has been used to interpret results from GWA studies (Neale and Sham, 2004; Wang et al., 2007; Peng et al., 2010; Zhong et al., 2010). In view of these reasons, a gene-based analysis was performed to investigate the genetic susceptibility to both

mycobacterial diseases, using the 2,092 genes assayed in the smaller leprosy study.

The methods for the gene-based and shared susceptibility analysis have been described in Chapter 2. In brief, the statistic of the best SNP was used to compute a statistic for each gene. Such an approach using the best SNP statistics has been described in several studies (Wang et al., 2007; Torkamani et al., 2008). In this study, the statistic of the best SNP within 50 kb of a gene was adjusted with the number of SNPs, using the Sidak's method (Peng et al., 2010) which corrects for the multiple testing of SNPs in each gene. This reduces the bias for large genes which are likely to have more SNPs genotyped in the microarray. However, it should be noted that the same association signal may contribute to the statistics of two or more nearby genes due to the 50 kb window. The results of all the genes attributable to a significant association signal were presented in this case. Evidence for shared susceptibility between tuberculosis and leprosy was tested with the permutation and parametric approaches, and genes showing evidence of association in both diseases in the aggregate statistics were identified.

6.2.3 Results

A total of 1,407 and 1,352 genes out of 2,092 on the smaller microarray had satisfactory coverage in the tuberculosis and leprosy studies respectively, among which 961 were shared between the studies. The data showed evidence of shared susceptibility between the two diseases, with $P_{permute} = 1.0 \times 10^{-4}$ and $P_{hypergeo} = 1.0 \times 10^{-5}$ at a threshold of $P < 0.01$ for a positive gene association (Figure 6.1). The statistics were also significant with $P_{permute} = 0.007$ and $P_{hypergeo} = 0.015$, at a threshold of $P < 0.05$. These analyses support the hypothesis that there are overlapping genetic loci affecting tuberculosis and leprosy, and provide proof-of-concept for the following work to look at common loci affecting mycobacterial diseases.

Aggregate analysis combining the tuberculosis and leprosy data revealed a total of 14 genes in 9 discrete regions showing evidence of association at $P < 0.001$ (Table 6.1). These include the *HLA* class II region which showed the an evidence of association with the best $P = 7.0 \times 10^{-9}$. Although the signal is driven by the association in leprosy, there is also a weak signal in tuberculosis with $P = 0.015$ (Table 6.1). To look for genetic loci involved in both diseases, the analysis was focused on regions with $P < 0.01$ in both diseases (Table 6.2). This revealed six genes in four discrete regions, including *CYP11B1/2*, *CTSB/FDFT1* and *CYP26A1* and *AGER* (Figure 6.2). All of these genes have an aggregate association statistic of $P < 1.0 \times 10^{-4}$ in the combined mycobacteria data, with the most significant gene *CYP11B2* having $P = 0.001$, $P = 2.5 \times 10^{-4}$ and $P = 5.1 \times 10^{-6}$ in tuberculosis, leprosy and both diseases combined respectively. Several other regions showed weaker but possible evidence of associations at a threshold of $P < 0.05$ (Table 6.2). These represent a novel set of genes that may be important in susceptibility to mycobacteria unrevealed by the previous analysis.

6.2.4 Discussion

The results presented here demonstrate a possible way to combine genetic studies using gene-based analysis. This gene-based analysis offers several advantages over SNP-based analysis. Firstly, the different populations in the studies have different LD patterns. Therefore the SNPs are likely to capture signals at different regions of the genome, making direct comparisons difficult. Secondly, as the individual studies were genotyped on different microarray platforms, only a small set of SNPs were genotyped on all platforms. Analysing only the common SNPs would mean discarding most of the genotype data, resulting in a significant loss of power. Thirdly, as genes are basic units that encode proteins to carry out physiological and biochemical functions,

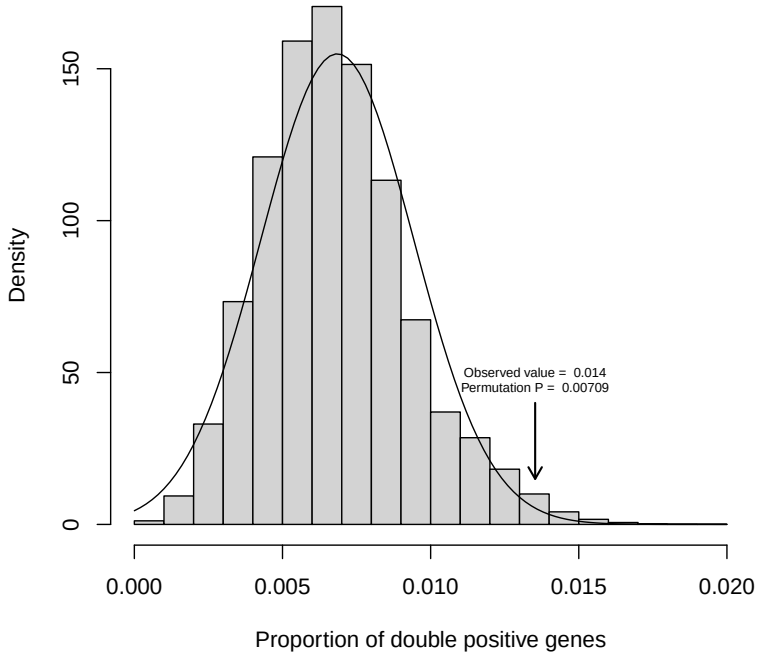


Figure 6.1: Histogram of proportion of double positive genes in 100,000 times of randomisation in the permutation analysis

Gene	Chr	Position	$P_{aggregate}$	Tuberculosis study		Leprosy study	
				SNPs/Mb	P_{gene}	SNPs/Mb	P_{gene}
HLA-DQA1	6	32,713,161-32,719,407	7.0×10^{-9}	85	0.015	160	2.0×10^{-8}
HLA-DQB1	6	32,735,642-32,742,419	2.3×10^{-8}	131	0.024	84	4.3×10^{-8}
CYP11B2	8	143,988,977-143,996,261	5.1×10^{-6}	56	0.001	84	2.5×10^{-4}
CYP11B1	8	143,950,775-143,958,238	9.8×10^{-6}	111	0.003	84	2.5×10^{-4}
RBP1	3	140,718,970-140,741,180	7.2×10^{-6}	164	0.023	90	2.0×10^{-5}
CTSB	8	11,737,442-11,763,055	9.9×10^{-6}	175	0.003	310	2.2×10^{-4}
FDFT1	8	11,697,599-11,734,227	1.7×10^{-5}	205	0.004	403	3.1×10^{-4}
CYP26A1	10	94,823,222-94,827,631	4.2×10^{-5}	67	0.004	57	6.7×10^{-4}
AGER	6	32,256,724-32,260,001	9.6×10^{-5}	78	0.008	194	9.8×10^{-4}
SLC12A1	15	46,287,190-46,382,417	1.5×10^{-4}	77	0.987	113	1.2×10^{-5}
MYC	8	128,817,498-128,822,856	4.9×10^{-4}	152	0.477	66	9.3×10^{-5}
TLR1	4	38,474,271-38,482,807	5.1×10^{-4}	166	0.166	654	2.8×10^{-4}
TLR10	4	38,450,629-38,460,984	5.5×10^{-4}	190	0.190	607	2.6×10^{-4}
TLR6	4	38,504,803-38,507,555	8.3×10^{-4}	107	0.389	506	2.0×10^{-4}

Table 6.1: Association statistics of genes with $P < 0.001$ in the aggregate analysis

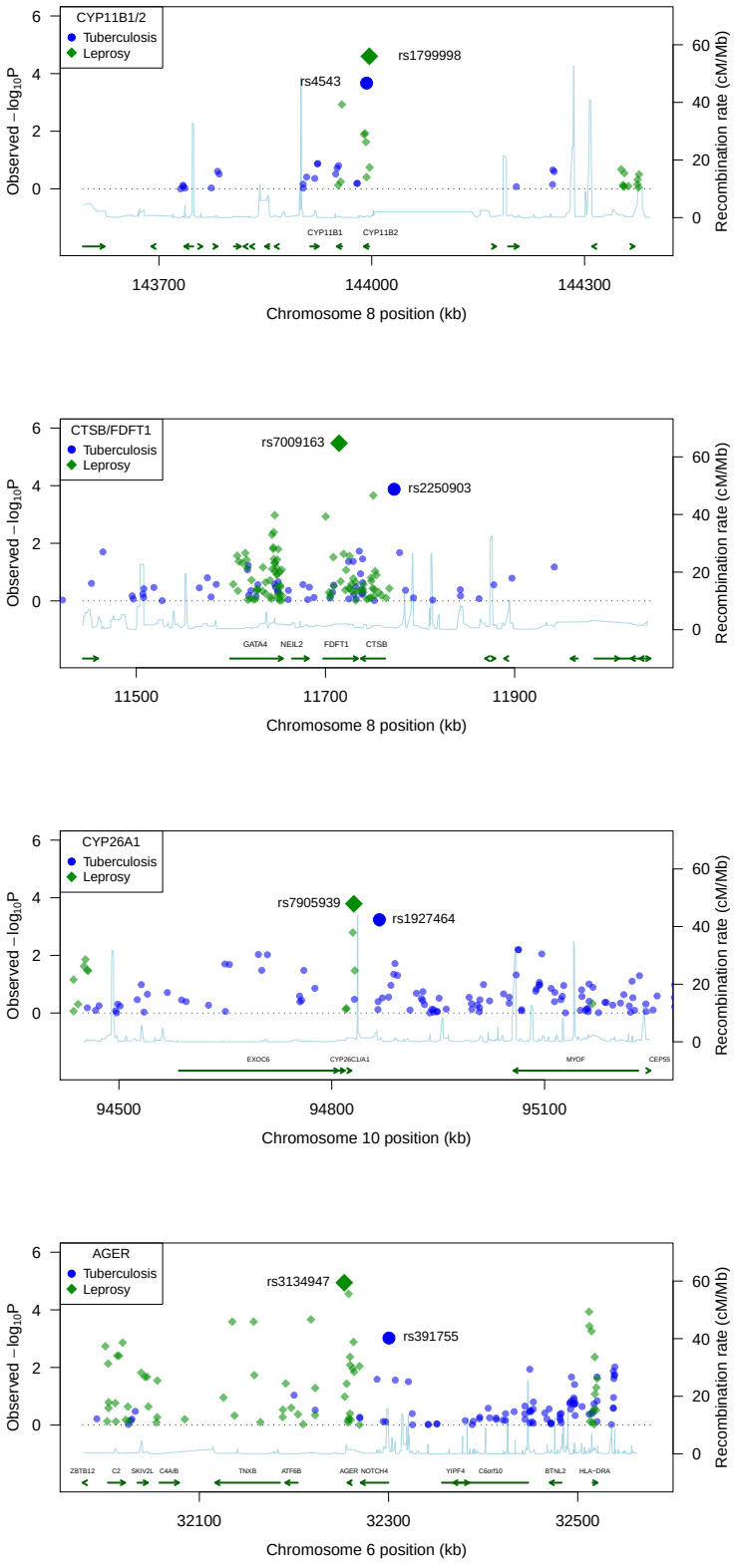


Figure 6.2: Association plots at the *CYP11B1/2*, *CTSB/FDFT1*, *CYP26A1* and *AGER* loci

Gene	Chr	Position	$P_{aggregate}$	Tuberculosis study		Leprosy study	
				SNPs/Mb	P_{gene}	SNPs/Mb	P_{gene}
Threshold $P < 0.01$							
CYP11B2	8	143,988,977-143,996,261	5.1×10^{-6}	56	0.001	84	2.5×10^{-4}
CYP11B1	8	143,950,775-143,958,238	9.8×10^{-6}	111	0.003	84	2.5×10^{-4}
CTSB	8	11,737,442-11,763,055	9.9×10^{-6}	175	0.003	310	2.2×10^{-4}
FDFT1	8	11,697,599-11,734,227	1.7×10^{-5}	205	0.004	403	3.1×10^{-4}
CYP26A1	10	94,823,222-94,827,631	4.2×10^{-5}	67	0.004	57	6.7×10^{-4}
AGER	6	32,256,724-32,260,001	9.6×10^{-5}	78	0.008	194	9.8×10^{-4}
Threshold $P < 0.05$							
HLA-DQA1	6	32,713,161-32,719,407	7.0×10^{-9}	85	0.015	160	2.0×10^{-8}
HLA-DQB1	6	32,735,642-32,742,419	2.3×10^{-8}	131	0.024	84	4.3×10^{-8}
RBP1	3	140,718,970-140,741,180	7.2×10^{-6}	164	0.023	90	2.0×10^{-5}
MYH8	17	10,234,367-10,265,747	0.001	99	0.014	53	0.011
PLAU	10	75,340,896-75,347,261	0.002	85	0.046	85	0.005
FOXC1	6	1,555,680-1,557,341	0.006	59	0.040	108	0.017
TTN	2	179,098,962-179,380,394	0.012	94	0.042	28	0.039

Table 6.2: Association statistics of genes with $P < 0.01$ or $P < 0.05$ in both tuberculosis and leprosy studies

the gene-based approach is more likely to help us to unravel the complex genetic structure of common diseases. This also applies to the pathway analysis which will be discussed in the next section. Although the Sidak's method of correction is based on the assumption that the SNP statistics are independent and uniformly distributed under their null hypotheses, which may be violated due to the presence of LD, such violation is unlikely to lead to significant errors (Peng et al., 2010). Indeed, even if such errors may occur, they tend to be over-conservative and therefore would have a minimal chance of introducing spurious results. This gene-based analysis approach using the best SNP statistics has been widely used and published (Wang et al., 2007; Baranzini et al., 2009; Torkamani et al., 2008).

The first part of this study was a proof-of-concept experiment to see whether there is overlapping susceptibility between tuberculosis and leprosy, in other words, whether genes associated with one disease are more likely to associate with another. Here

I have used a permutation approach to generate an empirical distribution using a randomisation procedure. This approach is robust to any statistical inflation that may have occurred in the individual studies, as it randomised the same set of observed data to generate the empirical distribution to which the count of double positive genes was compared. Reassuringly, the results of the permutation analysis are comparable to the tests against the hypergeometric distribution, and support the hypothesis of shared susceptibility between the two mycobacterial diseases. This hypothesis is consistent with literature showing common susceptibility genes in Asia including the *HLA* class II genes (Mehra et al., 1995; Remus et al., 2003).

The results presented here implicate four discrete regions showing evidence of association with both tuberculosis and leprosy, with $P < 1.0 \times 10^{-4}$ in the aggregate analysis and $P < 0.01$ in both of the individual sets. The strongest association was observed at *CYP11B1* and *CYP11B2* which encode the enzymes 11β-hydroxylase and aldosterone synthase respectively, both of which are members of the cytochrome P450 protein family that catalyse the synthesis of steroid hormones. Individuals who are immunosuppressed with steroid hormones, especially glucocorticoids, are prone to various infections including tuberculosis. Interestingly, the enzymes farnesyldiphosphate farnesyltransferase and retinoic acid 4-hydroxylase, encoded by two other significantly associated gene regions *FDFT1* and *CYP26A1*, are also involved in catalysing the metabolism of lipids. Given recent studies reporting the important role of lipids in the pathogenesis of mycobacteria (Steinberg and Grinstein, 2008), it is plausible that the metabolism of lipids may play a role in immunity against mycobacteria. This will be discussed in more detail in the next section on pathway analysis. In addition, cathepsin B, the gene product of *CTSB*, is a downstream mediator of the TNF-α triggered apoptosis cascade (Foghsgaard et al., 2001) and may be necessary for the response to *M. tuberculosis* infection (Kakegawa et al., 2004; O'Sullivan et al., 2007; Li et al., 2009). Lastly, the gene *AGER* encodes a receptor protein be-

longing to the immunoglobulin superfamily (Neeper et al., 1992), which is expressed on lymphocytes and phagocytes. Ligation with advanced glycosylation end products induces multiple signal transduction pathways, including the pro-inflammatory NF-κB pathway (Lander et al., 1997). Despite several studies reporting genetic associations of this gene with diabetic complications (Lindholm et al., 2006, 2008) and multiple sclerosis (Caillier et al., 2008); its definitive role in these diseases has not been established, and its role in the immune system remains unclear.

This work demonstrates a feasible approach to combine genetic studies using a gene-based analysis. However, while these results identified several novel genes as potential mycobacterial susceptibility candidates, further replication and fine mapping work is essential to establish the definitive role of these genes. It is premature to comment on the possible role of these genes without attempting to replicate these findings in multiple populations. Furthermore, this combined analysis is restricted by the smaller microarray in the leprosy study, which focuses on genes important for inflammatory and metabolic processes. It is therefore essential that these genes are not over-interpreted due to this focused genotyping approach. The following section describes gene annotation and pathway analysis, using a single ranked list to compare the significantly associated genes against the background list.

6.3 Gene annotation and pathway analysis

6.3.1 Introduction

As mentioned before, the gene-based approach offers several advantages in analysing genetic studies. Genes are the basic unit of life that code for proteins. These proteins interact in extensive networks to carry out different biological functions (Rual et al., 2005; Stelzl et al., 2005; Stumpf et al., 2008), and complex diseases are thought to

be contributed by the interaction of different genes in a functional module (Lage et al., 2007; Köhler et al., 2008). The current SNP based approach largely based on the additive model is unlikely to be able to capture such complex relationships in an interactome, and may therefore miss important non-additive gene-gene or gene-environment interactions. These interactions may account for the missing heritability of complex diseases that remains unexplained by current GWA findings (McCarthy and Hirschhorn, 2008; Manolio et al., 2009). Furthermore, the gene-based approach has been found to be less susceptible to erroneous findings (Neale and Sham, 2004). For these reasons, a gene-based or a pathway-based approach may allow us to gain better insight into the intricate biological network and to unravel the polygenic basis of common diseases. This section carries the association results forward for gene annotation and pathway analysis.

6.3.2 Methods

The genes analysed in Section 6.2 of this chapter were annotated with the Gene Ontology Enrichment Analysis and Visualization Tool (GOrilla) (Eden et al., 2009). It determines whether genes of a particular biological theme may be over-represented in the dataset. The pathway analysis was performed with the Ingenuity Pathway Analysis (Ingenuity System Incoporation, USA), and the canonical network was constructed by connecting the 14 most associated genes at $P < 0.001$ in the aggregate mycobacterial susceptibility analysis (Table 6.1). Details of the methodology have been described in Chapter 2.

6.3.3 Results

Significant enrichments were detected with several Gene Ontology terms (Figure 6.3). The most significant association was observed with the term *antigen processing*

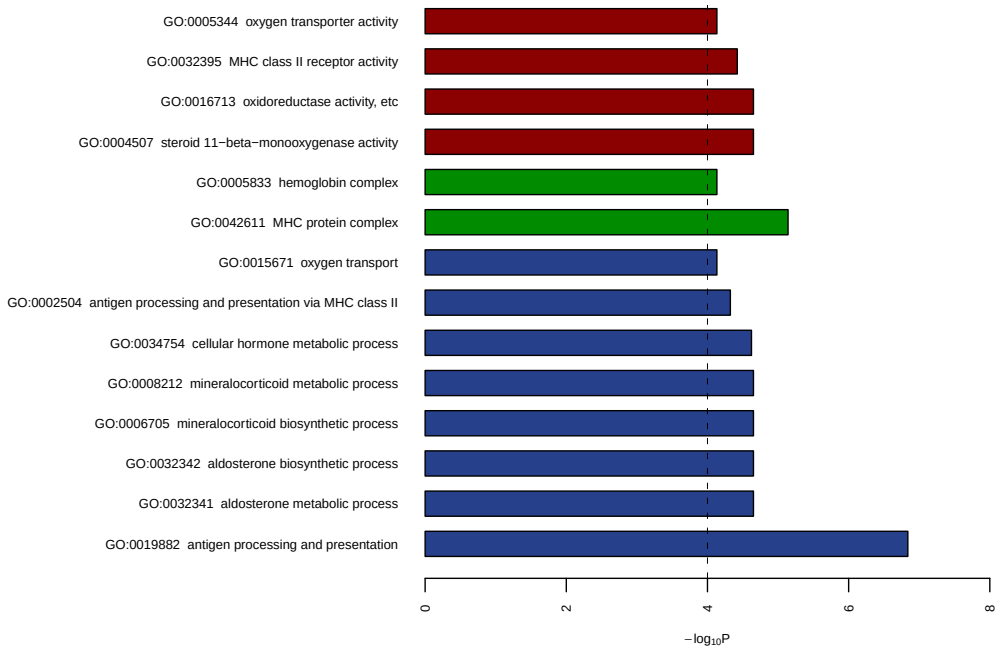


Figure 6.3: Gene Ontology analysis showing enriched terms at $P < 1.0 \times 10^{-4}$. Colours indicate term categories (red: biological processes, green: molecular functions, blue: cellular components)

and presentation (GO:0019882) with $P = 1.5 \times 10^{-7}$, indicating a significant over-representation of genes in this pathway. Notably, enrichment was also suggested for Gene Ontology terms relating to steroid hormone metabolism ($P_{best} = 2.2 \times 10^{-5}$). These results are consistent with the strongest association signal at the *HLA* class II locus, and suggest the possible involvement of the steroid hormone metabolism pathway.

To visualise the genes important in mycobacterial susceptibility, a canonical network was built from the protein products of the 14 associated genes with $P < 0.001$ in the aggregate analysis, with the addition of several key molecules such as IFN-γ, TNF-α and IL-6 (Figure 6.4). The analysis showed that the innate immune receptors, represented by TLR1, TLR6 and TLR10, were connected with the antigen presentation pathway through effector molecules like IFN-γ and IL-6. These molecules were in turn linked to the lipid metabolism pathway, represented by CYP11B1/2, CYP26A1

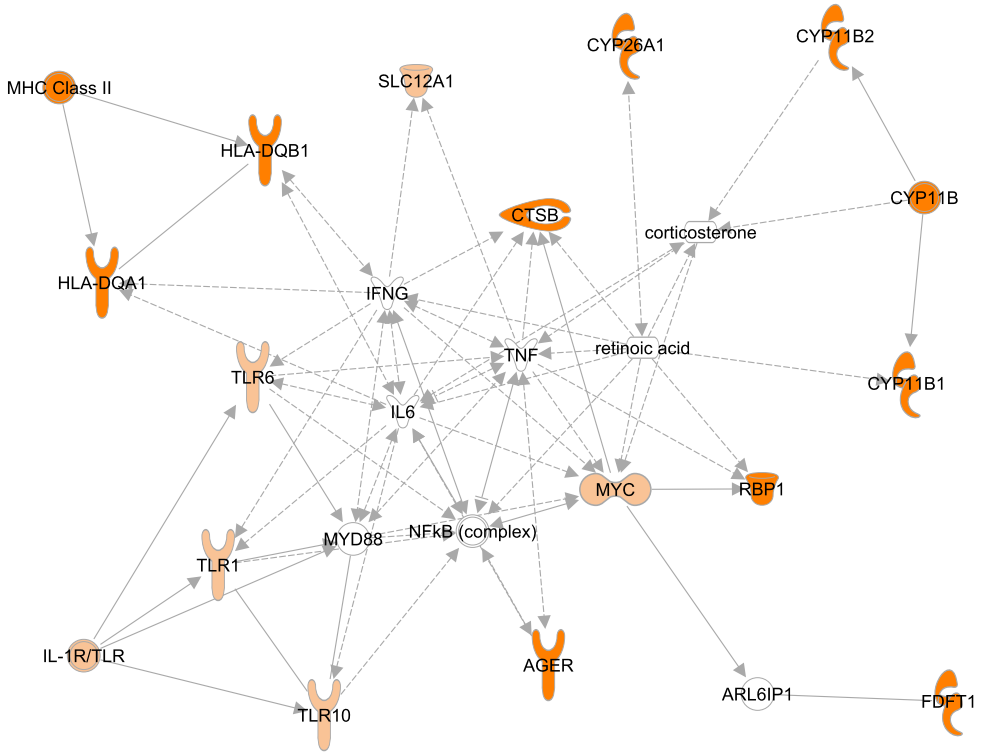


Figure 6.4: Canonical network built from the 14 products of genes with $P < 0.001$ (shaded in orange) and other molecules in the aggregate mycobacterial susceptibility analysis. The darker orange shade indicates genes with $P < 1.0 \times 10^{-4}$

and RBP1 through retinoic acid and corticosterone. Although the central molecules IFN-γ, IL-6 and some others did now show significant association in the aggregate analysis, *TNF* was associated with susceptibility to leprosy ($P = 0.015$). Moreoever, *IFNGR1*, which encodes the ligand-binding chain of the IFN-γ receptor, showed a weak association in the aggregate analysis ($P = 0.046$). Figures 6.5–6.7 illustrate the three important components of the network (Toll-like receptor, antigen presentation and steroid biosynthesis pathways) in more detail.

6.3.4 Discussion

The analysis presented in this section identified key areas which may be important in human immunity against mycobacteria. These include the Toll-like receptor sig-

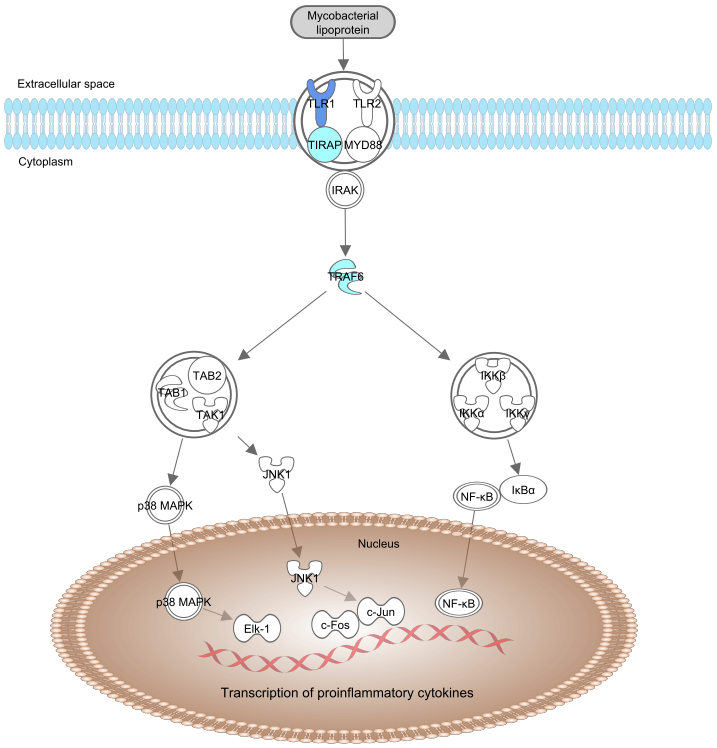


Figure 6.5: Diagram showing the Toll-like receptor pathway. Protein products of associated genes are highlighted (dark blue: $P < 0.001$ in the aggregate analysis, light blue: $P < 0.05$ in either tuberculosis, leprosy or the aggregate analysis)

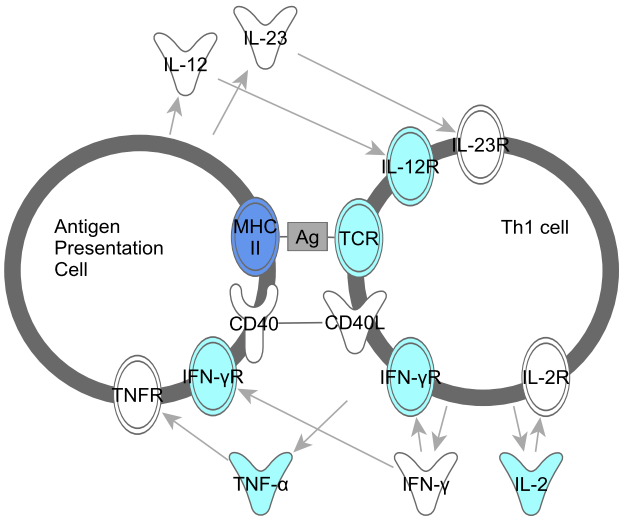


Figure 6.6: Diagram showing the antigen presentation pathway. Protein products of associated genes are highlighted (dark blue: $P < 0.001$ in the aggregate analysis, light blue: $P < 0.05$ in either tuberculosis, leprosy or the aggregate analysis)

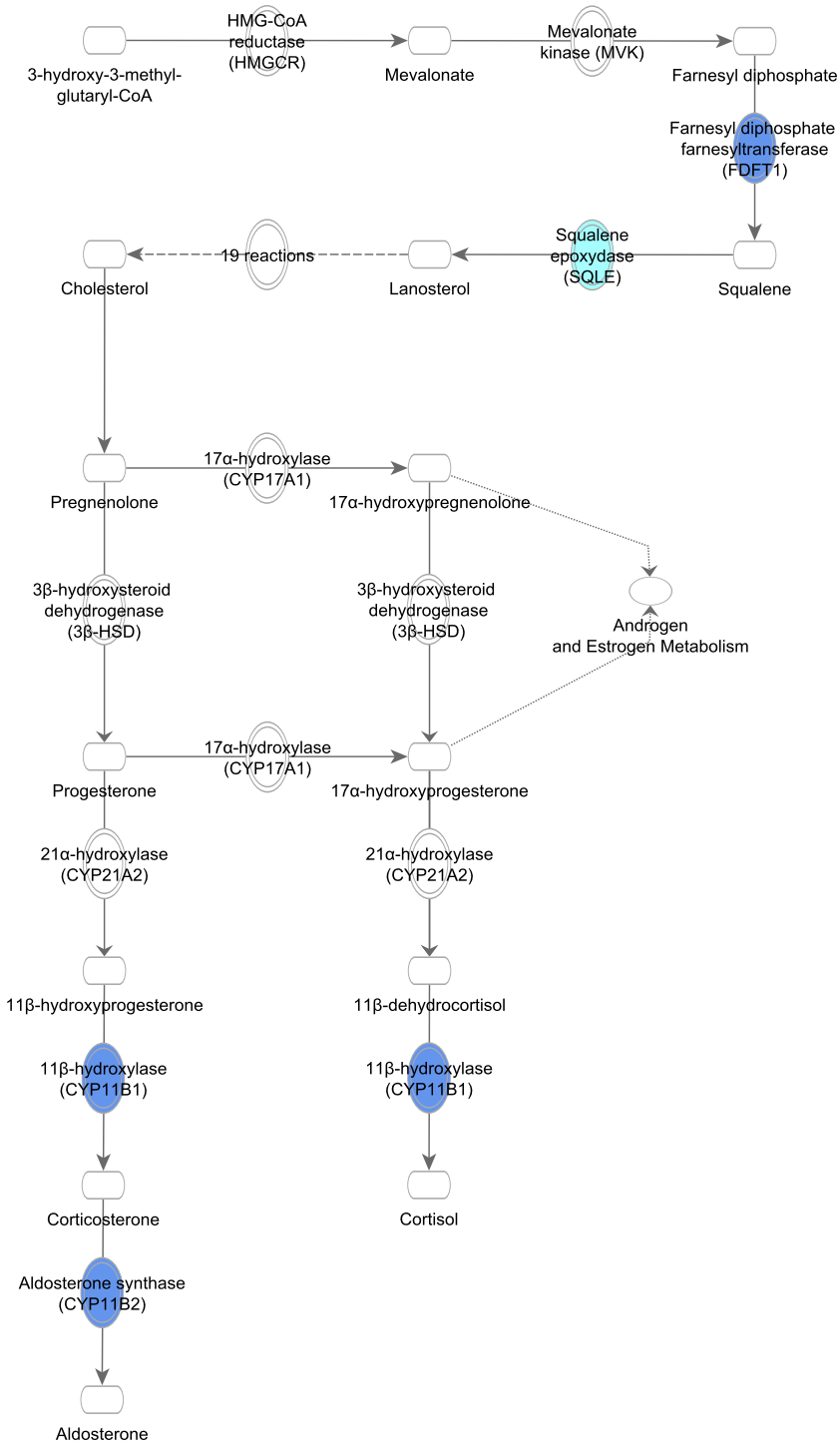


Figure 6.7: Diagram showing the steroid biosynthesis pathway. Protein products of associated genes are highlighted (dark blue: $P < 0.001$ in the aggregate analysis, light blue: $P < 0.05$ in either tuberculosis, leprosy or the aggregate analysis)

nalling, antigen presentation and steroid biosynthesis pathways which make up the key components in the canonical network.

These results highlight the importance of both the innate and adaptive immune systems in anti-mycobacterial immunity. Despite a weak association at the *HLA* locus in the tuberculosis study, the aggregate gene-based and pathway analyses do support a possible role of the *HLA* class II presentation complex in susceptibility to mycobacteria. This is consistent with early genetic studies suggesting an association between the *HLA* and mycobacterial diseases including both tuberculosis (Singh et al., 1983, 1984; Chandanayingyong et al., 1988; Ravikumar et al., 1999) and leprosy (Chiewsilp et al., 1979; de Vries et al., 1980; van Eden et al., 1980; Mehra et al., 1995).

Another important finding is the implication of the steroid biosynthesis pathway. This involves several key enzymes in the glucocorticoid and the mineralocorticoid pathways, including 11-beta hydroxylase (CYP11A1), aldosterone synthesis (CYP11B2), as well as farnesyldiphosphate farnesyltransferase (FDFT1). There was also weak evidence of association with *SQLE*, the gene encoding squalene epoxydase, in the tuberculosis study ($P = 0.046$). The possible involvement of these pathways is biologically plausible for a few reasons. Firstly, cholesterol is a key molecule in the synthesis of steroid hormones (Figure 6.7) which governs diverse metabolic functions in humans including the immune response. Individuals who are immunosuppressed with steroid hormones, especially glucocorticoids, are prone to various infections including tuberculosis; although paradoxically steroids may be used as an adjunct therapy for tuberculous pericarditis and meningitis (Fowler, 1991; Marras, 2005; Chan, 2008). Perturbations in the steroid hormone biosynthesis pathway have been suggested to account for the sexual inequality in the incidence of tuberculosis (Neyrolles and Quintana-Murci, 2009), with men being clearly more predisposed to the disease than women (Grossman, 1985; Yamamoto et al., 1990; Borgdorff et al.,

2000). Secondly, cholesterol is an important molecule for mycobacteria as well as for humans. It has been shown that cholesterol is required for mycobacteria to enter the host cell (Gatfield and Pieters, 2000; Kaul et al., 2004; Munoz et al., 2009) and to survive within the phagosome (de Chastellier and Thilo, 2006; van Der Geize et al., 2007; Pandey and Sassetti, 2008). As mycobacteria are obligate intracellular parasites, it is foreseeable that cholesterol metabolism is of vital importance to the pathogens. This is also reflected in the complete genome sequences of *M. tuberculosis* and *M. leprae*, showing significant proportions of their coding capacities related to lipid metabolism with at least 20 cytochrome P450 enzymes in *M. tuberculosis* (Cole et al., 1998, 2001). Although this study suggests a possible involvement in the host lipid metabolism pathway, certain bacteria have been shown to produce a homologue of host substrates to subvert the immune response (Cirl et al., 2008; Nesic et al., 2010), which could be possible with mycobacteria given their extensive coding machinery in lipid metabolism. Lastly, the canonical network analysis suggests that the lipid metabolism pathways may be related to the immune system through retinoic acid. Retinoic acid belongs to a class of chemical compounds called retinoids which are related to vitamin A. Through the retinoic acid receptors (RARs) and retinoid x receptor (RXRs) (Renaud et al., 1995; Bourguet et al., 1995), both members of the steroid hormone superfamily, retinoids act as transcription factors to promote transcription of target genes including cytokines IL-6 (Balmer and Blomhoff, 2002) and IFN-γ (Blalock and Gifford, 1977) in the canonical network. Retinoids have been found to alter $CD4^+$ T-cell differentiation (Stephensen et al., 2002; Iwata et al., 2003; Ivanov et al., 2006; Yang et al., 2008) and thus have a direct and major effect on the adaptive immune system.

Nevertheless, while these results support the genetic roles of these pathways in susceptibility to mycobacteria, caution must be taken not to over-interpret these results. It is possible that several genes out of a large pathway are associated purely

by chance, especially in this analysis which starts with predominantly inflammatory and metabolic pathways. Moreover, although the potential selection bias should have been adjusted by the comparison of significant genes to the background list, the statistical significance may still be overstated. This is because the association at one locus may contribute to the statistics of several adjacent genes, which may be homologous due to gene duplication and reside within the same functional pathway. Further work is needed to determine the roles of these pathways in mycobacterial diseases.

6.4 Identifying shared susceptibility with Crohn's disease

6.4.1 Introduction

In Section 6.2 of this chapter, I provided evidence of overlapping susceptibility between tuberculosis and leprosy using permutation and parametric methods. In this section I investigate such a possibility between mycobacterial diseases and Crohn's disease, an inflammatory bowel disease with a strong genetic component which has been thoroughly investigated in recent GWA studies (Parkes et al., 2007; Raelson et al., 2007; Barrett et al., 2008; Wang et al., 2009), using similar permutation and parametric methods.

The hypothesis of shared susceptibility between mycobacterial diseases and Crohn's disease is based on several arguments. Firstly, early studies suggested that Crohn's diseases has a mycobacterial cause. *Mycobacterium avium paratuberculosis* causes a chronic granulomatous inflammation in the intestines in cows called Johne's disease, and the bacterium has been detected in the intestines and peripheral blood of individuals with Crohn's disease (Naser et al., 2004; Sechi et al., 2005; Feller

et al., 2007). This has led to the argument that the pathogen may be an aetiological agent for Crohn's disease (Greenstein, 2003; Chamberlin et al., 2007; Pierce, 2009). Secondly, genetic studies have provided indirect evidence for the role of *M. avium paratuberculosis* in Crohn's disease. Recent GWA studies of Crohn's disease have identified numerous susceptibility genes relating to immunity against intracellular bacteria (Parkes et al., 2007; Barrett et al., 2008), including *NOD2*, *IRGM* and *IL23R*. The nucleotide-binding oligomerization domain 2 (NOD2) protein is a non-redundant pattern recognition receptor for mycobacteria (Ferwerda et al., 2005, 2007; Gandotra et al., 2007) and is involved in autophagy induction (Cooney et al., 2010), an important host defence mechanism against mycobacteria (MacMicking et al., 2003; Gutierrez et al., 2004; Alonso et al., 2007; Levine and Deretic, 2007; Kumar et al., 2010) which requires the protein immunity-related GTPase family, M (IRGM) (Feng et al., 2004; Singh et al., 2006). The gene *IL23R* encodes the receptor of IL-23, which is important in generating anti-mycobacterial $CD4^+$ T-cell responses (Khader et al., 2007). Consistent with this hypothesis, a recent GWA study of leprosy has identified several overlapping susceptibility genes with Crohn's disease (Barrett et al., 2008; Zhang et al., 2009). Based on these findings, I have conducted a formal analysis to look for shared susceptibility between mycobacterial diseases and Crohn's disease.

6.4.2 Methods

The analysis was conducted in a similar fashion to the methods described in Section 6.2 of this chapter. An association statistic was calculated for each of the 2,092 genes described, using SNPs within 50 kb of a gene. The number of associated genes in each disease was counted, as well as the number of double positive genes between leprosy and Crohn's disease, or tuberculosis and Crohn's disease. Evidence of shared susceptibility was tested with permutation and parametric tests.

6.4.3 Results and Discussion

Results of the analysis support an overlapping susceptibility between leprosy and Crohn's disease, with $P_{permute} = 0.031$ and $P_{hypergeo} = 0.027$ at a threshold of $P < 0.01$ for a positive gene association. The evidence was more significant at a threshold of $P < 0.05$, with $P_{permute} < 1.0 \times 10^{-6}$ and $P_{hypergeo} = 2.4 \times 10^{-7}$. This is in line with the hypothesis that there are overlapping genetic loci affecting leprosy and Crohn's disease, and the differences between the two thresholds suggest that there may be an expanded set of genetic loci shared between leprosy and Crohn's disease. However, analysis of the tuberculosis and Crohn's disease data does not support an overlap between these two diseases, as shown by non-significant P-values at either threshold of $P < 0.01$ ($P_{permute} = 0.447$ and $P_{hypergeo} = 0.394$) or threshold of $P < 0.05$ ($P_{permute} = 0.571$ and $P_{hypergeo} = 0.180$).

The data presented here support the hypothesis of shared susceptibility between leprosy and Crohn's disease, and between leprosy and tuberculosis, but not between tuberculosis and Crohn's disease. These results are in line with previous genetic findings identifying common susceptibility loci between leprosy and Crohn's disease (Parkes et al., 2007; Barrett et al., 2008; Zhang et al., 2009), yet less so between tuberculosis and Crohn's disease. One possible explanation for these findings is that there truly exist shared susceptibility between leprosy and each of the other diseases, yet with different genes between the pairs. Such a hypothesis is biologically plausible given the highly specialised nature of the immune system, with different components to generate immune responses to different pathogens.

6.5 Discussion

This chapter describes continuing efforts to pursue the role of host genetics in mycobacterial diseases, by integrating datasets with a more functional perspective. In the first section of this chapter, I have presented evidence for shared susceptibility between the two mycobacterial diseases, tuberculosis and leprosy. This provides proof-of-concept to study host genetic susceptibility to the two mycobacterial diseases as a single phenotype. In addition to the findings described in earlier chapters, these analyses identified several genes in the steroid biosynthesis pathways as potential susceptibility loci. These novel findings highlight the important roles of the Toll-like receptor pathway and T-cell immunity in defence against mycobacteria, and implicate the steroid biosynthesis pathway in mycobacterial susceptibility. These analyses show the power of data conglomeration in identifying potential biological pathways in human diseases. It is important to note, however, that the definitive roles of these genes cannot be established without further replication work. While these findings may provide insight into novel pathways, they should be thoroughly tested in different populations to show robust associations.

I have also investigated the overlap of susceptibility loci between mycobacterial diseases and Crohn's disease in this chapter. The results suggest a possible genetic overlap between leprosy and Crohn's disease. This is consistent with literature which suggests a possible mycobacterial cause of Crohn's disease (Naser et al., 2004; Feller et al., 2007). In this analysis of leprosy and Crohn's disease, more significant statistics were observed when a less stringent threshold of $P < 0.05$ was used to define a positive gene association, with $P_{permute} < 1.0 \times 10^{-6}$ and $P_{hypergeo} = 2.4 \times 10^{-7}$. In contrast, when the threshold of $P < 0.01$ was used, there was much weaker evidence with $P_{permute} = 0.031$ and $P_{hypergeo} = 0.027$. The reason for this difference may be related to the underlying genetic models as well as the overlap between the

two diseases as explained below. Despite most GWA findings explaining only a fraction of disease heritability (McCarthy and Hirschhorn, 2008; Manolio et al., 2009), these studies support the hypothesis that many common variants with very low effect sizes may affect susceptibility to complex diseases, some with substantial polygenic components (Weedon et al., 2008; Barrett et al., 2009; Purcell et al., 2009). Given recent GWA studies of Crohn's disease identifying numerous distinct loci (Alonso et al., 2007; Barrett et al., 2008), it is possible that its genetic architecture may consist of numerous variants each exerting a small effect on disease susceptibility. Assuming a polygenic model for both diseases with a substantial genetic overlap, the over-representation of double positive genes will preferentially occur with a less stringent threshold, indicated by the more significant statistics in the permutation or parametric analysis. Conversely, if the diseases have an oligogenic model or only have a limited set of overlapping loci, random genes may be included by chance at a less stringent threshold, rendering the tests insensitive in detecting genetic overlap. Although it is premature to comment on the genetic architecture of these mycobacterial diseases, results of these analyses suggest that genetic overlap between them may be limited to genes that are more significantly associated; while the list between leprosy and Crohn's disease may be more extended. It may therefore be useful to study the genes implicated in Crohn's disease in the context of leprosy.

In conclusion, this chapter demonstrates how data integration can provide novel insight into the biology of diseases, using genetic data from the mycobacterial studies, Crohn's disease data from public depositories, and gene annotation and pathway data from bioinformatic tools. I have also shown how genetic relationships between two diseases can be investigated. These analyses, however, represent only a few examples of how data integration and exploration can be performed. Large amounts of data from public depositories including gene expression, genome diversity and protein structure data can be further explored to infer the functional impact of genetic

variations. The ongoing 1000 Genome Project will provide a deep catalogue of human genetic variation, with a vast amount of sequence data in different populations which will facilitate the identification of functional variants and help to understand disease mechanisms.

Chapter 7

Discussion

7.1 Overview of this thesis

This thesis presents work investigating the role of host genetics in common mycobacterial diseases. The genetic component of mycobacterial diseases has long been recognised. The clustering of tuberculosis and leprosy in families initially led physicians and scientists to regard both as primarily inherited diseases (Scollard et al., 2006; Wilson, 2006). However, it was not until the late 1970s that scientists began to discover the host genetic variants responsible for this inheritance, with studies reporting associations of *HLA* alleles with tuberculosis and leprosy (van Eden et al., 1980; de Vries et al., 1980; van Eden et al., 1982; Singh et al., 1983, 1984; Schauf et al., 1985). The association results for leprosy were largely consistent likely due to the large effect sizes of the *HLA* alleles; whereas the results for tuberculosis were less consistent (Hawkins et al., 1988; Sanjeevi et al., 1992). The following decades saw further development in the field, aided by improvements in genotyping techniques, with studies reporting the identification of a number of candidate genes in the 1990s (Cooke and Hill, 2001; Hill, 2006). Some of these candidates came from studies of

animal models, while others came from human genetic studies based on hypotheses derived from gene function. Although the genome wide approach was used in linkage studies with micro-satellite markers, it had low power and resolution to detect genetic variants.

The availability of high-density microarray genotyping technology, together with the initial catalogue of human genetic variations created by the International HapMap Project, has fuelled exponential growth in the number of GWA studies conducted over the past few years. These studies have identified numerous novel genetic loci in common diseases. This genome-wide approach has also been used to study the host genetics of infectious diseases like malaria (Jallow et al., 2009), hepatitic C (Thomas et al., 2009; Ge et al., 2009) and HIV infection (Fellay et al., 2007). Due to the substantial heritable component in tuberculosis and leprosy, this approach represents a feasible way to dissect the genetic bases of these mycobacterial diseases. I have therefore conducted several studies on the host genetics of these diseases as described in this thesis. These findings can potentially provide an insight into the human immunity against these mycobacteria, and to help to devise novel therapeutics to alleviate the massive global health burden imposed by these diseases.

7.2 Genetic susceptibility to tuberculosis

In Chapters 3 and 4, I have assessed genetic susceptibility to tuberculosis as part of the WTCCC study. The study was conducted by genotyping 1,498 tuberculosis cases and 1,496 controls recruited from The Gambia, with the Affymetrix 500K array comprising 500,568 SNPs randomly distributed across the genome. The study identified the region flanking *CADM1* as a potential susceptibility locus. This gene codes for the protein CADM1 which mediates T-cell differentiation and cytokine release (Galibert et al., 2005). This association was supported by the consistent associations in

the Gambian and Conakry-Guinean cohorts. However, robust associations were not observed in the Malawian or Bissau-Guinean populations. Further fine mapping of the associated region did not identify significantly more associated variants. While it remains possible that *CADM1* is a true locus affecting susceptibility to tuberculosis, further studies from different populations are needed to establish its genetic role in tuberculosis.

Several major difficulties were encountered during the course of this study. Firstly, due to the short-range LD in African populations (Gabriel et al., 2002), the number of SNPs tagging the possible functional variants and therefore the coverage was significantly reduced (Barrett and Cardon, 2006). The extensive diversity of African populations discouraged the possibility of imputation using the HapMap Yoruba population, as an attempted imputation showed a disconcordant rate of 10.7%. This is consistent with a study showing the need for population specific data for imputation in African populations (Jallow et al., 2009). This has further limited the amount of informative data points; and despite the resequencing effort to overcome these shortcomings, this did not identify any significantly more associated variant. Secondly, tropical diseases including tuberculosis are often endemic in developing countries where infrastructure is lacking. This poses difficulties in collecting a large number of well characterised samples. Given the modest effect sizes of most common variants (Hindorff et al., 2009; Manolio et al., 2009), sample size is very important in yielding sufficient power to detect associations. These problems may have underlain the difficulty experienced with finding robust associations in this WTCCC study.

Given the importance of the sample size in associations studies, I conducted a joint analysis by combining the WTCCC Gambian data with a Ghanaian GWA tuberculosis study as described in Chapter 4. Previous studies have found success when using this strategy to identify novel genetic loci showing robust association (Zeggini

et al., 2008; Barrett et al., 2009). The joint analysis increased the sample size to a total of 2,217 cases and 3,113 controls, which increased the power of detecting a variant with an odds ratio of 1.25 from 50% to nearly 90%. This study identified the locus at chromosome 18q11.2 with a SNP showing strong evidence of association in the combined study, as well as a case control and a family population from Ghana in the replication studies. A consistent direction of effect was also observed in the Malawian, Bissau-Guinean and Conakry-Guinean populations, giving a highly significant overall statistic ($P = 9.2 \times 10^{-9}$). In fact, the association showed a consistent effect in all the 13 sub-populations in the study. These results present a novel locus that affect susceptibility to tuberculosis, and suggest that functional studies should be considered. Furthermore, a potential signal of association was detected at a SNP in *PARD3B*. This study represents the largest human genetic study of tuberculosis known-to date, and shows the power of combining genetic studies in identifying novel genetic loci not identifiable with individual studies.

7.3 Genetic susceptibility to leprosy

Chapter 5 describes the efforts to assess genetic susceptibility to leprosy. The importance of this study is twofold. Firstly, leprosy is the only other common mycobacterial disease affecting humans. Studying leprosy can potentially provide new insight into immunity against mycobacteria, including *M. tuberculosis*. Given the present difficulty in culturing *M. leprae in vitro*, genetic studies represent one of the few feasible ways to study the disease. Secondly, despite the falling incidence, leprosy is still endemic in many parts of the world with a global incidence of over 200,000. This study may help to develop novel therapeutics to reduce the health burden of the disease.

This study involved the genotyping of 258 leprosy cases and 300 controls recruited from New Delhi, India, using the HumanCVD Genotyping BeadChip microarray

which covers 2,092 genes in the human genome. This experiment identified *HLA-DRB1/DQA1* and *TLR1* as major determinants of susceptibility to leprosy. The results of this study identified two association signals in *HLA* class II genes that do not account for each other, suggesting the possibility of more than one functional allele that affects disease susceptibility (rs1071630, $P = 4.9 \times 10^{-14}$ and rs9270650, $P = 3.1 \times 10^{-11}$). This study also identified a *TLR1* variant protecting against leprosy, at a genome-wide threshold of significance (rs5743618, $P = 1.7 \times 10^{-9}$). This supports an important role of TLR1 in immunity against *M. leprae*, and suggests that its modulation may be helpful in the treatment of mycobacterial infections. Lastly, using these Indian population cohorts, I have replicated the locus at *C13orf31* that was reported in a recent GWA study of leprosy (Zhang et al., 2009). This provides validation for this locus in populations independent from the original study in China, and supports the need for further functional studies to characterise the function of this gene.

7.4 Integrative analysis of mycobacterial susceptibility

As mentioned before, one reason to study host genetics in leprosy is to provide new insight to resistance against tuberculosis. Although some evidence exists for overlapping susceptibility between the two diseases (Kolk et al., 1984; Shinnick et al., 1987; Hermans et al., 1995), formal analysis on the overlap is lacking. The availability of genome-wide genotype data in tuberculosis, along with the selected gene centric set in the smaller leprosy study, has enabled me to assess the genetic overlap between the two diseases as described in Chapter 6. The results of this analysis are consistent with the hypothesis of shared susceptibility between these two diseases, and provide a proof-of-concept for experiments to look for common genes important for both diseases.

The aggregate analysis was performed using the gene-based statistics of the 2,092 genes in both studies. Despite a weaker signal in the tuberculosis study, the *HLA* class II locus showed the most significant association statistics in the aggregate analysis. The analysis also implicated several genes including *CYP11B1/2*, *FDFT1*, *CYP26A1* and *AGER* in susceptibility to mycobacteria. This work demostrates a feasible approach to combine genetic studies using a gene-based analysis. Nevertheless, further genetic work is needed to investigate the roles of these genes in tuberculosis and leprosy in different populations.

These results show the power of integrating genetic and functional information, including microarray data, gene annotations, protein interaction networks and other association data, to help to identify genomic regions or biological pathways that are important for the phenotype under study. However, this work only provides a few examples of integrative analyses. With the rapid expansion of public bioinformatic databases, it is possible to further explore the human genome; for example, using expression databases, protein structure information and comparative genomic and human genome diversity data. These analyses can help to identify important regions of the genome, and provide a way to prioritise the order and nature of functional experiments that follow genetic studies. The 1000 Genome Project, an ongoing international collaboration to provide a deep catalogue of genetic variation, will generate a vast amount of data which could help to localise the functional variants in different disease genes.

7.5 Concluding remarks

The approach of testing common variants for association in this thesis is largely based on the *common disease-common variant* hypothesis (Lander, 1996; Risch and Merikangas, 1996), which suggests that common diseases are largely determined by

genetic variants common in the population. This hypothesis is supported by findings from association studies which have identified numerous common genetic variants affecting common disease susceptibility (Donnelly, 2008). However, due to the moderate effect sizes of most common variants, they often explain only a small fraction of the disease heritability (Manolio et al., 2009; McCarthy, 2009). Identifying this missing heritability is one of the biggest challenges forced by contemporary geneticists.

Despite agreement on the importance of identifying this unexplained heritability, there are different views regarding where to find it. The contrasting view to the common variant hypothesis suggests that common traits are affected by rare or low frequency variants. These variants may have large effect sizes and therefore account for a greater proportion of the unexplained heritability despite their lower frequencies. For example, it has been estimated that only twenty variants of an odds ratio of 3 and a minor allele frequency of 1% would be required to account for most of the unexplained heritability in type 2 diabetes (McCarthy, 2009). This *common disease rare variant* hypothesis has been supported by a recent study identifying low frequency variants in type 1 diabetes (Nejentsev et al., 2009). These views are not mutually exclusive, however, and it is likely that the genetic architecture of common disease actually embraces causal variants with a wide range of frequencies and effect sizes (McCarthy, 2009). In addition, other genetic factors that contribute to missing heritability may include copy number variation (CNVs), epigenetic change and gene-gene and gene-environment interactions apart from SNPs.

Understanding this missing heritability is of paramount importance in guiding the future of common disease genetics. If common genetic variants do account for the bulk of the missing heritability, efforts may be spent in expanding the current GWA studies with even larger sample sizes. Such a strategy, however, is highly ineffective in

identifying low frequency or rare variants. Therefore, if the large fraction of missing heritability lies with rare variants, the attention should be focused on tracking down the these variants through different resequencing strategies. Nevertheless, given the higher sequencing capacities now available at lower costs, high-throughput sequencing is likely to be more informative in defining the functional variants of human traits; as it is already making inroads into pinpointing the causal variant for some Mendelian diseases with either whole-genome (Lupski et al., 2010; Roach et al., 2010) or exome sequencing (Ng et al., 2009, 2010) strategies. These second and third generation sequencing technologies will likely remain centrally important in the genetic field. Moreover, several studies have taken the technique further, to assess molecular functions at the gene transcription (Ozsolak et al., 2009; Pickrell et al., 2010; Montgomery et al., 2010), protein translation (Ingolia et al., 2009) and DNA-protein interaction (Schmidt et al., 2009) levels. As these techniques continue to inform us about the unknown territories of the human genome, the latest example being the unexpected discovery of extensive untranslated regions (Pickrell et al., 2010), it is possible that this high-throughput sequencing technique may help to reveal the remaining missing heritability.

This thesis presents major efforts to understand genetic susceptibility to common mycobacterial diseases. The collaborative tuberculosis study represents the largest genetic study of tuberculosis currently known to date, involving a total of more than 10,000 samples. This identifies the chromosome 18q11.2 and, less strongly, the *PARD3B* loci as susceptibility loci, and demonstrated that combining genetic studies to achieve a larger sample size can help to identify novel genetic loci. The leprosy study reaffirms the association at *HLA* class II locus and represents the first study to show robust association at *TLR1* beyond a genome-wide threshold of statistical significance. Although these genetic findings may not be immediately translated into benefits of clinical relevance, they can help to improve our understanding of the

molecular mechanism by which diseases arise. Understanding these disease mechanisms can lead to the development of new therapeutics and vaccines; for example molecules in the Toll-like receptor pathway have been used as DNA vaccine adjuvants to enhance immunogenicity in the Grand Challenges in Global Health Project (Larsen et al., 2009). This progress can set examples for other genetic findings described in this thesis, so as to prioritise functional studies to help to understand their biological roles in mycobacterial diseases. These findings are likely to provide new footholds on the long and difficult path to better treatment and prevention for mycobacterial diseases.

Bibliography

Abel, L., Vu, D. L., Oberti, J., Nguyen, V. T., Van, V. C., Guilloud-Bataille, M., Schurr, E., and Lagrange, P. H. (1995). Complex segregation analysis of leprosy in southern Vietnam. *Genet Epidemiol*, 12(1):63–82.

Akahoshi, M., Nakashima, H., Miyake, K., Inoue, Y., Shimizu, S., Tanaka, Y., Okada, K., Otsuka, T., and Harada, M. (2003). Influence of interleukin-12 receptor beta1 polymorphisms on tuberculosis. *Hum Genet*, 112(3):237–43.

Akira, S. and Takeda, K. (2004). Toll-like receptor signalling. *Nat Rev Immunol*, 4(7):499–511.

Akira, S., Takeda, K., and Kaisho, T. (2001). Toll-like receptors: critical proteins linking innate and acquired immunity. *Nat Immunol*, 2(8):675–680.

Al-Muhsen, S. and Casanova, J. L. (2008). The genetic heterogeneity of mendelian susceptibility to mycobacterial diseases. *J Allergy Clin Immunol*, 122(6):1043.

Alcais, A., Alter, A., Antoni, G., Orlova, M., Nguyen, V. T., Singh, M., Vanderborght, P. R., Katoch, K., Mira, M. T., Vu, H. T., Ngyuen, T. H., Nguyen, N. B., Moraes, M., Mehra, N., Schurr, E., and Abel, L. (2007). Stepwise replication identifies a low-producing lymphotoxin-alpha allele as a major risk factor for early-onset leprosy. *Nat Genet*, 39(4):517–522.

Algood, H. M., Lin, P. L., and Flynn, J. L. (2005). Tumor necrosis factor and

chemokine interactions in the formation and maintenance of granulomas in tuberculosis. *Clin Infect Dis*, 41 Suppl 3:S189–93.

Allison, A. C. (1954). Protection afforded by sickle-cell trait against subtertian malareal infection. *Br Med J*, 1(4857):290–294.

Alonso, S., Pethe, K., Russell, D. G., and Purdy, G. E. (2007). Lysosomal killing of Mycobacterium mediated by ubiquitin-derived peptides is enhanced by autophagy. *Proc Natl Acad Sci U S A*, 104(14):6031–6.

Altshuler, D. and Daly, M. (2007). Guilt beyond a reasonable doubt. *Nat Genet*, 39(7):813–5.

Altshuler, D., Daly, M. J., and Lander, E. S. (2008). Genetic mapping in human disease. *Science*, 322(5903):881–888.

Ando, K., Ohira, M., Ozaki, T., Nakagawa, A., Akazawa, K., Suenaga, Y., Nakamura, Y., Koda, T., Kamijo, T., Murakami, Y., and Nakagawara, A. (2008). Expression of TSLC1, a candidate tumor suppressor gene mapped to chromosome 11q23, is downregulated in unfavorable neuroblastoma without promoter hypermethylation. *Int J Cancer*, 123(9):2087–94.

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11(3):375–386.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–9.

Astarie-Dequeker, C., N'Diaye, E. N., Le Cabec, V., Rittig, M. G., Prandi, J., and Maridonneau-Parini, I. (1999). The mannose receptor mediates uptake of

pathogenic and nonpathogenic mycobacteria and bypasses bactericidal responses in human macrophages. *Infect Immun*, 67(2):469–477.

Babb, C., Keet, E. H., van Helden, P. D., and Hoal, E. G. (2007). SP110 polymorphisms are not associated with pulmonary tuberculosis in a South African population. *Hum Genet*, 121(3-4):521–522.

Bakker, M. I., May, L., Hatta, M., Kwenang, A., Klatser, P. R., Oskam, L., and Houwing-Duistermaat, J. J. (2005). Genetic, household and spatial clustering of leprosy on an island in Indonesia: a population-based study. *BMC Med Genet*, 6:40.

Balmer, J. E. and Blomhoff, R. (2002). Gene expression regulation by retinoic acid. *J Lipid Res*, 43(11):1773–808.

Baranzini, S. E., Galwey, N. W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., Wu, W., Uitdehaag, B. M. J., Kappos, L., Polman, C. H., Matthews, P. M., Hauser, S. L., Gibson, R. A., Oksenberg, J. R., and Barnes, M. R. (2009). Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum Mol Genet*, 18(11):2078–90.

Barnes, P. F., Chatterjee, D., Brennan, P. J., Rea, T. H., and Modlin, R. L. (1992). Tumor necrosis factor production in patients with leprosy. *Infect Immun*, 60(4):1441–1446.

Barreiro, L. B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J. K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B., Kidd, J. R., Kidd, K. K., Alcais, A., Ragimbeau, J., Pellegrini, S., Abel, L., Casanova, J. L., and Quintana-Murci, L. (2009). Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet*, 5(7):e1000562.

Barreiro, L. B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. (2008).

Natural selection has driven population differentiation in modern humans. *Nat Genet*, 40(3):340–345.

Barreiro, L. B., Neyrolles, O., Babb, C. L., Tailleux, L., Quach, H., McElreavey, K., Helden, P. D., Hoal, E. G., Gicquel, B., and Quintana-Murci, L. (2006). Promoter variation in the DC-SIGN-encoding gene CD209 is associated with tuberculosis. *PLoS Med*, 3(2):e20.

Barrett, J. C. and Cardon, L. R. (2006). Evaluating coverage of genome-wide association studies. *Nat Genet*, 38(6):659–662.

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*.

Barrett, J. C., Fry, B., Maller, J., and Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, 21(2):263–265.

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):955–962.

Bell, G. I. and Polonsky, K. S. (2001). Diabetes mellitus and genetically programmed defects in beta-cell function. *Nature*, 414(6865):788–791.

Bellamy, R., Beyers, N., McAdam, K. P., Ruwende, C., Gie, R., Samaai, P., Bester, D., Meyer, M., Corrah, T., Collin, M., Camidge, D. R., Wilkinson, D., Hoal-Van Helden, E., Whittle, H. C., Amos, W., van Helden, P., and Hill, A. V. (2000). Genetic susceptibility to tuberculosis in Africans: a genome-wide scan. *Proc Natl Acad Sci U S A*, 97(14):8005–8009.

Bellamy, R., Ruwende, C., Corrah, T., McAdam, K. P., Thursz, M., Whittle, H. C., and Hill, A. V. (1999). Tuberculosis and chronic hepatitis B virus infection in Africans and variation in the vitamin D receptor gene. *J Infect Dis*, 179(3):721–724.

Bellamy, R., Ruwende, C., Corrah, T., McAdam, K. P., Whittle, H. C., and Hill, A. V. (1998). Variations in the NRAMP1 gene and susceptibility to tuberculosis in West Africans. *N Engl J Med*, 338(10):640–644.

Ben-Ali, M., Barbouche, M. R., Bousnina, S., Chabbou, A., and Dellagi, K. (2004). Toll-like receptor 2 Arg677Trp polymorphism is associated with susceptibility to tuberculosis in Tunisian patients. *Clin Diagn Lab Immunol*, 11(3):625–626.

Bennett, C. D., Campbell, M. N., Cook, C. J., Eyre, D. J., Nay, L. M., Nielsen, D. R., Rasmussen, R. P., and Bernard, P. S. (2003). The LightTyper: high-throughput genotyping using fluorescent melting curve analysis. *BioTechniques*, 34(6):1288–92, 1294–5.

Bennett, S., Lienhardt, C., Bah-Sow, O., Gustafson, P., Manneh, K., Del Prete, G., Gomes, V., Newport, M., McAdam, K., and Hill, A. (2002). Investigation of envi-

ronmental and host-related risk factors for tuberculosis in Africa. II. Investigation of host genetic factors. *Am J Epidemiol*, 155(11):1074–1079.

Bhangale, T. R., Rieder, M. J., and Nickerson, D. A. (2008). Estimating coverage and power for genetic association studies using near-complete variation data. *Nat Genet*, 40(7):841–3.

Biederer, T., Sara, Y., Mozhayeva, M., Atasoy, D., Liu, X., Kavalali, E. T., and Sudhof, T. C. (2002). SynCAM, a synaptic adhesion molecule that drives synapse assembly. *Science*, 297(5586):1525–1531.

Biswas, D., Qureshi, O. S., Lee, W. Y., Croudace, J. E., Mura, M., and Lammas, D. A. (2008). ATP-induced autophagy is associated with rapid killing of intracellular mycobacteria within human monocytes/macrophages. *BMC Immunol*, 9:35.

Blackwell, J. M. (1998). Genetics of host resistance and susceptibility to intramacrophage pathogens: a study of multicase families of tuberculosis, leprosy and leishmaniasis in north-eastern Brazil. *Int J Parasitol*, 28(1):21–8.

Blalock, J. E. and Gifford, G. E. (1977). Retinoic acid (vitamin A acid) induced transcriptional control of interferon production. *Proc Natl Acad Sci U S A*, 74(12):5382–6.

Bloom, B. R. and Murray, C. J. (1992). Tuberculosis: commentary on a reemergent killer. *Science*, 257(5073):1055–64.

Bodmer, W. and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet*, 40(6):695–701.

Boldsen, J. L. (2005). Leprosy and mortality in the Medieval Danish village of Tirup. *Am J Phys Anthropol*, 126(2):159–168.

Borgdorff, M. W., Nagelkerke, N. J., Dye, C., and Nunn, P. (2000). Gender and

tuberculosis: a comparison of prevalence surveys with notification data to explore sex differences in case detection. *Int J Tuberc Lung Dis*, 4(2):123–32.

Bornman, L., Campbell, S. J., Fielding, K., Bah, B., Sillah, J., Gustafson, P., Manneh, K., Lisse, I., Allen, A., Sirugo, G., Sylla, A., Aaby, P., McAdam, K. P., Bah-Sow, O., Bennett, S., Lienhardt, C., and Hill, A. V. (2004). Vitamin D receptor polymorphisms and susceptibility to tuberculosis in West Africa: a case-control and family study. *J Infect Dis*, 190(9):1631–1641.

Bourguet, W., Ruff, M., Chambon, P., Gronemeyer, H., and Moras, D. (1995). Crystal structure of the ligand-binding domain of the human nuclear receptor RXR-alpha. *Nature*, 375(6530):377–82.

Breslow, N. E. and Day, N. E. (1980). Statistical methods in cancer research. Volume I - The analysis of case-control studies. *IARC Sci Publ*, (32):5–338.

Brightbill, H. D., Libraty, D. H., Krutzik, S. R., Yang, R. B., Belisle, J. T., Bleharski, J. R., Maitland, M., Norgard, M. V., Plevy, S. E., Smale, S. T., Brennan, P. J., Bloom, B. R., Godowski, P. J., and Modlin, R. L. (1999). Host defense mechanisms triggered by microbial lipoproteins through toll-like receptors. *Science*, 285(5428):732–736.

Brookes, A. J. (1999). The essence of SNPs. *Gene*, 234(2):177–86.

Browning, B. L. and Yu, Z. (2009). Simultaneous genotype calling and haplotype phasing improves genotype accuracy and reduces false-positive associations for genome-wide association studies. *Am J Hum Genet*, 85(6):847–61.

Bulut, Y., Faure, E., Thomas, L., Equils, O., and Arditi, M. (2001). Cooperation of Toll-like receptor 2 and 6 for cellular activation by soluble tuberculosis factor and Borrelia burgdorferi outer surface protein A lipoprotein: role of Toll-interacting

protein and IL-1 receptor signaling molecules in Toll-like receptor 2 s. *J Immunol*, 167(2):987–994.

Bustamante, J., Aksu, G., Vogt, G., de Beaucoudrey, L., Genel, F., Chapgier, A., Filipe-Santos, O., Feinberg, J., Emile, J. F., Kutukculer, N., and Casanova, J. L. (2007). BCG-osis and tuberculosis in a child with chronic granulomatous disease. *J Allergy Clin Immunol*, 120(1):32–38.

Caillier, S. J., Briggs, F., Cree, B. A. C., Baranzini, S. E., Fernandez-Viña, M., Ramsay, P. P., Khan, O., Royal, W., Hauser, S. L., Barcellos, L. F., and Oksenberg, J. R. (2008). Uncoupling the roles of HLA-DRB1 and HLA-DRB5 genes in multiple sclerosis. *J Immunol*, 181(8):5473–80.

Campbell, M. C. and Tishkoff, S. A. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet*, 9:403–33.

Campbell, S. J., Sabeti, P., Fielding, K., Sillah, J., Bah, B., Gustafson, P., Manneh, K., Lisse, I., Sirugo, G., Bellamy, R., Bennett, S., Aaby, P., McAdam, K. P., Bah-Sow, O., Lienhardt, C., and Hill, A. V. (2003). Variants of the CD40 ligand gene are not associated with increased susceptibility to tuberculosis in West Africa. *Immunogenetics*, 55(7):502–507.

Cantor, R. M., Lange, K., and Sinsheimer, J. S. (2010). Prioritizing GWAS results: A review of statistical methods and recommendations for their application. *Am J Hum Genet*, 86(1):6–22.

Carey, G. and Williamson, J. (1991). Linkage analysis of quantitative traits: increased power by using selected samples. *Am J Hum Genet*, 49(4):786–96.

Carvalho, B. S., Louis, T. A., and Irizarry, R. A. (2010). Quantifying uncertainty in genotype calls. *Bioinformatics*, 26(2):242–9.

Chackerian, A. A., Perera, T. V., and Behar, S. M. (2001). Gamma interferon-producing CD4+ T lymphocytes in the lung correlate with resistance to infection with Mycobacterium tuberculosis. *Infect Immun*, 69(4):2666–2674.

Chakhaiyar, P., Nagalakshmi, Y., Aruna, B., Murthy, K. J., Katoch, V. M., and Hasnain, S. E. (2004). Regions of high antigenicity within the hypothetical PPE major polymorphic tandem repeat open-reading frame, Rv2608, show a differential humoral response and a low T cell response in various categories of patients with tuberculosis. *J Infect Dis*, 190(7):1237–1244.

Chakravartti, M. R. and Vogel, F. (1973). *A twin study on leprosy*, volume 1. Georg Thieme, Stuttgart.

Chamberlin, W., Borody, T., and Naser, S. (2007). MAP-associated Crohn's disease MAP, Koch's postulates, causality and Crohn's disease. *Dig Liver Dis*, 39(8):792–4.

Chan, E. D. (2008). Corticosteroids for bacterial meningitis. *N Engl J Med*, 358(13):1399; author reply 1400–1.

Chandanayingyong, D., Maranetra, N., and Bovornkitti, S. (1988). HLA antigen profiles in Thai tuberculosis patients. *Asian Pac J Allergy Immunol*, 6(2):77–80.

Chiewsilp, P., Athkambhira, S., Chirachariyavej, T., Bhamarapravati, N., and Entwistle, C. (1979). The HLA antigens and leprosy in Thailand. *Tissue Antigens*, 13(3):186–188.

Cho, S., Mehra, V., Thoma-Uszynski, S., Stenger, S., Serbina, N., Mazzaccaro, R. J., Flynn, J. L., Barnes, P. F., Southwood, S., Celis, E., Bloom, B. R., Modlin, R. L., and Sette, A. (2000). Antimicrobial activity of MHC class I-restricted CD8+ T cells in human tuberculosis. *Proc Natl Acad Sci U S A*, 97(22):12210–12215.

Cho, Y. S., Go, M. J., Kim, Y. J., Heo, J. Y., Oh, J. H., Ban, H. J., Yoon, D., Lee, M. H., Kim, D. J., Park, M., Cha, S. H., Kim, J. W., Han, B. G., Min, H., Ahn, Y.,

Park, M. S., Han, H. R., Jang, H. Y., Cho, E. Y., Lee, J. E., Cho, N. H., Shin, C., Park, T., Park, J. W., Lee, J., Cardon, L., Clarke, G., McCarthy, M. I., Lee, J. Y., Lee, J. K., Oh, B., and Kim, H. L. (2009). A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet*, 41(5):527–34.

Choudhary, R. K., Mukhopadhyay, S., Chakhaiyar, P., Sharma, N., Murthy, K. J. R., Katoch, V. M., and Hasnain, S. E. (2003). PPE antigen Rv2430c of Mycobacterium tuberculosis induces a strong B-cell response. *Infect Immun*, 71(11):6338–43.

Cirl, C., Wieser, A., Yadav, M., Duerr, S., Schubert, S., Fischer, H., Stappert, D., Wantia, N., Rodriguez, N., Wagner, H., Svanborg, C., and Miethke, T. (2008). Subversion of Toll-like receptor signaling by a unique family of bacterial Toll/interleukin-1 receptor domain-containing proteins. *Nat Med*, 14(4):399–406.

Cochran, W. G. (1954). Some methods for strengthening the common chi-square tests. *Biometrics*, 10:417–451.

Colangeli, R., Haq, A., Arcus, V. L., Summers, E., Magliozzo, R. S., McBride, A., Mitra, A. K., Radjainia, M., Khajo, A., Jacobs Jr., W. R., Salgame, P., and Alland, D. (2009). The multifunctional histone-like protein Lsr2 protects mycobacteria against reactive oxygen intermediates. *Proc Natl Acad Sci U S A*, 106(11):4414–4418.

Colditz, G. A., Brewer, T. F., Berkey, C. S., Wilson, M. E., Burdick, E., Fineberg, H. V., and Mosteller, F. (1994). Efficacy of BCG vaccine in the prevention of tuberculosis. Meta-analysis of the published literature. *JAMA*, 271(9):698–702.

Cole, S. T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S. V., Eiglmeier, K., Gas, S., Barry 3rd, C. E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T.,

Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M. A., Rajandream, M. A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J. E., Taylor, K., Whitehead, S., and Barrell, B. G. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685):537–544.

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Wheeler, P. R., Honoré, N., Garnier, T., Churcher, C., Harris, D., Mungall, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R. M., Devlin, K., Duthoy, S., Feltwell, T., Fraser, A., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Lacroix, C., Maclean, J., Moule, S., Murphy, L., Oliver, K., Quail, M. A., Rajandream, M. A., Rutherford, K. M., Rutter, S., Seeger, K., Simon, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Taylor, K., Whitehead, S., Woodward, J. R., and Barrell, B. G. (2001). Massive gene decay in the leprosy bacillus. *Nature*, 409(6823):1007–11.

Collins, F. S., Brooks, L. D., and Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res*, 8(12):1229–31.

Collins, F. S., Morgan, M., and Patrinos, A. (2003). The Human Genome Project: lessons from large-scale biology. *Science*, 300(5617):286–90.

Colston, M. J. and Hilson, G. R. (1976). Growth of Mycobacterium leprae and M. marinum in congenitally athymic (nude) mice. *Nature*, 262(5567):399–401.

Comas, I. n., Chakravartti, J., Small, P. M., Galagan, J., Niemann, S., Kremer, K., Ernst, J. D., and Gagneux, S. (2010). Human T cell epitopes of Mycobacterium tuberculosis are evolutionarily hyperconserved. *Nat Genet*.

Comstock, G. W. (1978). Tuberculosis in twins: a re-analysis of the Prophit survey. *Am Rev Respir Dis*, 117(4):621–4.

Conrad, D. F., Jakobsson, M., Coop, G., Wen, X., Wall, J. D., Rosenberg, N. A., and Pritchard, J. K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet*, 38(11):1251–60.

Cooke, G. S., Campbell, S. J., Bennett, S., Lienhardt, C., McAdam, K. P., Sirugo, G., Sow, O., Gustafson, P., Mwangulu, F., van Helden, P., Fine, P., Hoal, E. G., and Hill, A. V. (2008). Mapping of a novel susceptibility locus suggests a role for MC3R and CTSZ in human tuberculosis. *Am J Respir Crit Care Med*, 178(2):203–207.

Cooke, G. S. and Hill, A. V. (2001). Genetics of susceptibility to human infectious disease. *Nat Rev Genet*, 2(12):967–977.

Cooney, R., Baker, J., Brain, O., Danis, B., Pichulik, T., Allan, P., Ferguson, D. J. P., Campbell, B. J., Jewell, D., and Simmons, A. (2010). NOD2 stimulation induces autophagy in dendritic cells influencing bacterial handling and antigen presentation. *Nat Med*, 16(1):90–7.

Crick, F. (1958). On protein synthesis. *Symp Soc Exp Biol*, 12:138–63.

Crofton, J. (1959). Chemotherapy of pulmonary tuberculosis. *Br Med J*, 1(5138):1610–1614.

Dalton, D. K., Pitts-Meek, S., Keshav, S., Figari, I. S., Bradley, A., and Stewart, T. A. (1993). Multiple defects of immune cell function in mice with disrupted interferon-gamma genes. *Science*, 259(5102):1739–42.

Davis, A. S., Vergne, I., Master, S. S., Kyei, G. B., Chua, J., and Deretic, V. (2007). Mechanism of inducible nitric oxide synthase exclusion from mycobacterial phagosomes. *PLoS Pathog*, 3(12):e186.

Dawn Teare, M. and Barrett, J. H. (2005). Genetic linkage studies. *Lancet*, 366(9490):1036–1044.

de Bakker, P. I., McVean, G., Sabeti, P. C., Miretti, M. M., Green, T., Marchini, J., Ke, X., Monsuur, A. J., Whittaker, P., Delgado, M., Morrison, J., Richardson, A., Walsh, E. C., Gao, X., Galver, L., Hart, J., Hafler, D. A., Pericak-Vance, M., Todd, J. A., Daly, M. J., Trowsdale, J., Wijmenga, C., Vyse, T. J., Beck, S., Murray, S. S., Carrington, M., Gregory, S., Deloukas, P., and Rioux, J. D. (2006). A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC. *Nat Genet*, 38(10):1166–1172.

de Chastellier, C. and Thilo, L. (2006). Cholesterol depletion in Mycobacterium avium-infected macrophages overcomes the block in phagosome maturation and leads to the reversible sequestration of viable mycobacteria in phagolysosome-derived autophagic vacuoles. *Cell Microbiol*, 8(2):242–56.

de Jong, R., Altare, F., Haagen, I. A., Elferink, D. G., Boer, T., van Breda Vriesman, P. J., Kabel, P. J., Draaisma, J. M., van Dissel, J. T., Kroon, F. P., Casanova, J. L., and Ottenhoff, T. H. (1998). Severe mycobacterial and Salmonella infections in interleukin-12 receptor-deficient patients. *Science*, 280(5368):1435–1438.

de Vries, R. R., Mehra, N. K., Vaidya, M. C., Gupte, M. D., Meera Khan, P., and Van Rood, J. J. (1980). HLA-linked control of susceptibility to tuberculoid leprosy and association with HLA-DR types. *Tissue Antigens*, 16(4):294–304.

Dean, M., Carrington, M., Winkler, C., Huttley, G. A., Smith, M. W., Allikmets, R., Goedert, J. J., Buchbinder, S. P., Vittinghoff, E., Gomperts, E., Donfield, S., Vlahov, D., Kaslow, R., Saah, A., Rinaldo, C., Detels, R., and O'Brien, S. J. (1996). Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multi-

center AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE . *Science*, 273(5283):1856–1862.

Delgado, M. A., Elmaoued, R. A., Davis, A. S., Kyei, G., and Deretic, V. (2008). Toll-like receptors control autophagy. *EMBO J*, 27(7):1110–21.

Deretic, V. (2010). Autophagy in infection. *Curr Opin Cell Biol*.

Deretic, V., Delgado, M., Vergne, I., Master, S., De Haro, S., Ponpuak, M., and Singh, S. (2009). Autophagy in immunity against mycobacterium tuberculosis: a model system to dissect immunological roles of autophagy. *Curr Top Microbiol Immunol*, 335:169–188.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Control Clin Trials*, 7(3):177–88.

Devlin, B. and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, 55(4):997–1004.

Dewan, M. Z., Takamatsu, N., Hidaka, T., Hatakeyama, K., Nakahata, S., Fujisawa, J.-i., Katano, H., Yamamoto, N., and Morishita, K. (2008). Critical role for TSLC1 expression in the growth and organ infiltration of adult T-cell leukemia cells in vivo. *J Virol*, 82(23):11958–63.

Djavaheri-Mergny, M., Amelotti, M., Mathieu, J., Besancon, F., Bauvy, C., Souquere, S., Pierron, G., and Codogno, P. (2006). NF-kappaB activation represses tumor necrosis factor-alpha-induced autophagy. *J Biol Chem*, 281(41):30373–30382.

Döffinger, R., Jouanguy, E., Dupuis, S., Fondanèche, M. C., Stephan, J. L., Emile, J. F., Lamhamedi-Cherradi, S., Altare, F., Pallier, A., Barcenas-Morales, G., Meinl, E., Krause, C., Pestka, S., Schreiber, R. D., Novelli, F., and Casanova, J. L. (2000). Partial interferon-gamma receptor signaling chain deficiency in a

patient with bacille Calmette-Guérin and Mycobacterium abscessus infection. *J Infect Dis*, 181(1):379–84.

Donnelly, P. (2008). Progress and challenges in genome-wide association studies in humans. *Nature*, 456(7223):728–31.

Doz, E., Rose, S., Nigou, J., Gilleron, M., Puzo, G., Erard, F., Ryffel, B., and Quesniaux, V. F. (2007). Acylation determines the toll-like receptor (TLR)-dependent positive versus TLR2-, mannose receptor-, and SIGNR1-independent negative regulation of pro-inflammatory cytokines by mycobacterial lipomannan. *J Biol Chem*, 282(36):26014–26025.

Druley, T. E., Vallania, F. L. M., Wegner, D. J., Varley, K. E., Knowles, O. L., Bonds, J. A., Robison, S. W., Doniger, S. W., Hamvas, A., Cole, F. S., Fay, J. C., and Mitra, R. D. (2009). Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods*, 6(4):263–5.

Dunning, M. J., Barbosa-Morais, N. L., Lynch, A. G., Tavaré, S., and Ritchie, M. E. (2008). Statistical issues in the analysis of Illumina data. *BMC Bioinformatics*, 9:85.

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., Wheeler, E., Glazer, N. L., Bouatia-Naji, N., Gloyn, A. L., Lindgren, C. M., Mägi, R., Morris, A. P., Randall, J., Johnson, T., Elliott, P., Rybin, D., Thorleifsson, G., Steinthorsdottir, V., Henneman, P., Grallert, H., Dehghan, A., Hottenga, J. J., Franklin, C. S., Navarro, P., Song, K., Goel, A., Perry, J. R. B., Egan, J. M., Lajunen, T., Grarup, N., Sparsø, T., Doney, A., Voight, B. F., Stringham, H. M., Li, M., Kanoni, S., Shrader, P., Cavalcanti-Proença, C., Kumari, M., Qi, L., Timpson, N. J., Gieger, C., Zabena, C., Rocheleau, G., Ingelsson, E., An, P., O'Connell, J., Luan, J., Elliott, A., McCarroll, S. A., Payne, F., Roccasecca, R. M., Pattou, F., Sethupathy, P., Ardlie, K., Ariyurek, Y., Balkau, B., Barter, P., Beilby, J. P.,

Ben-Shlomo, Y., Benediktsson, R., Bennett, A. J., Bergmann, S., Bochud, M., Boerwinkle, E., Bonnefond, A., Bonnycastle, L. L., Borch-Johnsen, K., Böttcher, Y., Brunner, E., Bumpstead, S. J., Charpentier, G., Chen, Y.-D. I., Chines, P., Clarke, R., Coin, L. J. M., Cooper, M. N., Cornelis, M., Crawford, G., Crisponi, L., Day, I. N. M., de Geus, E. J. C., Delplanque, J., Dina, C., Erdos, M. R., Fedson, A. C., Fischer-Rosinsky, A., Forouhi, N. G., Fox, C. S., Frants, R., Franzosi, M. G., Galan, P., Goodarzi, M. O., Graessler, J., Groves, C. J., Grundy, S., Gwilliam, R., Gyllensten, U., Hadjadj, S., Hallmans, G., Hammond, N., Han, X., Hartikainen, A.-L., Hassanali, N., Hayward, C., Heath, S. C., Hercberg, S., Herder, C., Hicks, A. A., Hillman, D. R., Hingorani, A. D., Hofman, A., Hui, J., Hung, J., Isomaa, B., Johnson, P. R. V., Jø rgensen, T., Jula, A., Kaakinen, M., Kaprio, J., Kesaniemi, Y. A., Kivimaki, M., Knight, B., Koskinen, S., Kovacs, P., Kyvik, K. O., Lathrop, G. M., Lawlor, D. A., Le Bacquer, O., Lecoeur, C., Li, Y., Lyssenko, V., Mahley, R., Mangino, M., Manning, A. K., Martínez-Larrad, M. T., McAteer, J. B., McCulloch, L. J., McPherson, R., Meisinger, C., Melzer, D., Meyre, D., Mitchell, B. D., Morken, M. A., Mukherjee, S., Naitza, S., Narisu, N., Neville, M. J., Oostra, B. A., Orrù, M., Pakyz, R., Palmer, C. N. A., Paolisso, G., Pattaro, C., Pearson, D., Peden, J. F., Pedersen, N. L., Perola, M., Pfeiffer, A. F. H., Pichler, I., Polasek, O., Posthuma, D., Potter, S. C., Pouta, A., Province, M. A., Psaty, B. M., Rathmann, W., Rayner, N. W., Rice, K., Ripatti, S., Rivadeneira, F., Roden, M., Rolandsson, O., Sandbaek, A., Sandhu, M., Sanna, S., Sayer, A. A., Scheet, P., Scott, L. J., Seedorf, U., Sharp, S. J., Shields, B., Sigurethsson, G., Sijbrands, E. J. G., Silveira, A., Simpson, L., Singleton, A., Smith, N. L., Sovio, U., Swift, A., Syddall, H., Syvänen, A.-C., Tanaka, T., Thorand, B., Tichet, J., Tönjes, A., Tuomi, T., Uitterlinden, A. G., van Dijk, K. W., van Hoek, M., Varma, D., Visvikis-Siest, S., Vitart, V., Vogelzangs, N., Waeber, G., Wagner, P. J., Walley, A., Walters, G. B., Ward, K. L., Watkins, H., Weedon, M. N., Wild, S. H., Willemsen, G., Witteman,

J. C. M., Yarnell, J. W. G., Zeggini, E., Zelenika, D., Zethelius, B., Zhai, G., Zhao, J. H., Zillikens, M. C., Borecki, I. B., Loos, R. J. F., Meneton, P., Magnusson, P. K. E., Nathan, D. M., Williams, G. H., Hattersley, A. T., Silander, K., Salomaa, V., Smith, G. D., Bornstein, S. R., Schwarz, P., Spranger, J., Karpe, F., Shuldiner, A. R., Cooper, C., Dedoussis, G. V., Serrano-Ríos, M., Morris, A. D., Lind, L., Palmer, L. J., Hu, F. B., Franks, P. W., Ebrahim, S., Marmot, M., Kao, W. H. L., Pankow, J. S., Sampson, M. J., Kuusisto, J., Laakso, M., Hansen, T., Pedersen, O., Pramstaller, P. P., Wichmann, H. E., Illig, T., Rudan, I., Wright, A. F., Stumvoll, M., Campbell, H., Wilson, J. F., Bergman, R. N., Buchanan, T. A., Collins, F. S., Mohlke, K. L., Tuomilehto, J., Valle, T. T., Altshuler, D., Rotter, J. I., Siscovick, D. S., Penninx, B. W. J. H., Boomsma, D. I., Deloukas, P., Spector, T. D., Frayling, T. M., Ferrucci, L., Kong, A., Thorsteinsdottir, U., Stefansson, K., van Duijn, C. M., Aulchenko, Y. S., Cao, A., Scuteri, A., Schlessinger, D., Uda, M., Ruokonen, A., Jarvelin, M.-R., Waterworth, D. M., Vollenweider, P., Peltonen, L., Mooser, V., Abecasis, G. R., Wareham, N. J., Sladek, R., Froguel, P., Watanabe, R. M., Meigs, J. B., Groop, L., Boehnke, M., McCarthy, M. I., Florez, J. C., and Barroso, I. (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet*, 42(2):105–16.

Dupuis, S., Dargemont, C., Fieschi, C., Thomassin, N., Rosenzweig, S., Harris, J., Holland, S. M., Schreiber, R. D., and Casanova, J. L. (2001). Impairment of mycobacterial but not viral immunity by a germline human STAT1 mutation. *Science*, 293(5528):300–303.

Dye, C., Scheele, S., Dolin, P., Pathania, V., Raviglione, M. C., and for the WHO Global Surveillance and Monitoring Project (1999). Global Burden of Tuberculosis: Estimated Incidence, Prevalence, and Mortality by Country. *JAMA*, 282(7):677–686.

Eden, E., Navon, R., Steinfeld, I., Lipson, D., and Yakhini, Z. (2009). GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10:48.

Eeles, R. A., Kote-Jarai, Z., Giles, G. G., Olama, A. A. A., Guy, M., Jugurnauth, S. K., Mulholland, S., Leongamornlert, D. A., Edwards, S. M., Morrison, J., Field, H. I., Southey, M. C., Severi, G., Donovan, J. L., Hamdy, F. C., Dearnaley, D. P., Muir, K. R., Smith, C., Bagnato, M., Ardern-Jones, A. T., Hall, A. L., O'Brien, L. T., Gehr-Swain, B. N., Wilkinson, R. A., Cox, A., Lewis, S., Brown, P. M., Jhavar, S. G., Tymrakiewicz, M., Lophatananon, A., Bryant, S. L., Horwich, A., Huddart, R. A., Khoo, V. S., Parker, C. C., Woodhouse, C. J., Thompson, A., Christmas, T., Ogden, C., Fisher, C., Jamieson, C., Cooper, C. S., English, D. R., Hopper, J. L., Neal, D. E., and Easton, D. F. (2008). Multiple newly identified loci associated with prostate cancer susceptibility. *Nat Genet*, 40(3):316–21.

El Sahly, H. M., Reich, R. A., Dou, S. J., Musser, J. M., and Graviss, E. A. (2004). The effect of mannose binding lectin gene polymorphisms on susceptibility to tuberculosis in different ethnic groups. *Scand J Infect Dis*, 36(2):106–108.

Elloumi-Zghal, H., Barbouche, M. R., Chemli, J., Bejaoui, M., Harbi, A., Snoussi, N., Abdelhak, S., and Dellagi, K. (2002). Clinical and genetic heterogeneity of inherited autosomal recessive susceptibility to disseminated Mycobacterium bovis bacille calmette-guerin infection. *J Infect Dis*, 185(10):1468–1475.

Etokebe, G. E., Bulat-Kardum, L., Johansen, M. S., Knezevic, J., Balen, S., Matakovic-Mileusnic, N., Matanic, D., Flego, V., Pavelic, J., Beg-Zec, Z., and Dembic, Z. (2006). Interferon-gamma gene (T874A and G2109A) polymorphisms are associated with microscopy-positive tuberculosis. *Scand J Immunol*, 63(2):136–141.

Ewing, B. and Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, 8(3):186–94.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, 8(3):175–85.

Falconer, D. S. (1960). *Introduction to Quantitative Genetics*. Oliver & Boyd, Edinburgh/London.

Fatkenheuer, G., Nelson, M., Lazzarin, A., Konourina, I., Hoepelman, A. I., Lampiris, H., Hirschel, B., Tebas, P., Raffi, F., Trottier, B., Bellos, N., Saag, M., Cooper, D. A., Westby, M., Tawadrous, M., Sullivan, J. F., Ridgway, C., Dunne, M. W., Felstead, S., Mayer, H., and van der Ryst, E. (2008). Subgroup analyses of maraviroc in previously treated R5 HIV-1 infection. *N Engl J Med*, 359(14):1442–1455.

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):944–947.

Feller, M., Huwiler, K., Stephan, R., Altpeter, E., Shang, A., Furrer, H., Pfyffer, G. E., Jemmi, T., Baumgartner, A., and Egger, M. (2007). Mycobacterium avium subspecies paratuberculosis and Crohn's disease: a systematic review and meta-analysis. *Lancet Infect Dis*, 7(9):607–613.

Felsenstein, J. (1974). The evolutionary advantage of recombination. *Genetics*, 78(2):737–56.

Feng, C. G., Collazo-Custodio, C. M., Eckhaus, M., Hieny, S., Belkaid, Y., Elkins, K., Jankovic, D., Taylor, G. A., and Sher, A. (2004). Mice deficient in LRG-47 display increased susceptibility to mycobacterial infection associated with the induction of lymphopenia. *J Immunol*, 172(2):1163–8.

Ferebee, S. H., Mount, F. W., Murray, F. J., and Livesay, V. T. (1963). A controlled trial of isoniazid prophylaxis in mental institutions. *Am Rev Respir Dis*, 88:161–75.

Ferwerda, G., Girardin, S. E., Kullberg, B.-J., Le Bourhis, L., de Jong, D. J., Langenberg, D. M. L., van Crevel, R., Adema, G. J., Ottenhoff, T. H. M., Van Der Meer, J. W. M., and Netea, M. G. (2005). NOD2 and toll-like receptors are nonredundant recognition systems of Mycobacterium tuberculosis. *PLoS Pathog*, 1(3):279–85.

Ferwerda, G., Kullberg, B. J., de Jong, D. J., Girardin, S. E., Langenberg, D. M. L., van Crevel, R., Ottenhoff, T. H. M., Van Der Meer, J. W. M., and Netea, M. G. (2007). Mycobacterium paratuberculosis is recognized by Toll-like receptors and NOD2. *J Leukoc Biol*, 82(4):1011–8.

Filipe-Santos, O., Bustamante, J., Haverkamp, M. H., Vinolo, E., Ku, C. L., Puel, A., Frucht, D. M., Christel, K., von Bernuth, H., Jouanguy, E., Feinberg, J., Durandy, A., Senechal, B., Chapgier, A., Vogt, G., de Beaucoudrey, L., Fieschi, C., Picard, C., Garfa, M., Chemli, J., Bejaoui, M., Tsolia, M. N., Kutukculer, N., Plebani, A., Notarangelo, L., Bodemer, C., Geissmann, F., Israel, A., Veron, M., Knackstedt, M., Barbouche, R., Abel, L., Magdorf, K., Gendrel, D., Agou, F., Holland, S. M., and Casanova, J. L. (2006). X-linked susceptibility to mycobacteria is caused by mutations in NEMO impairing CD40-dependent IL-12 production. *J Exp Med*, 203(7):1745–1759.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans Roy Soc Edinb*, 52:399–433.

Fisher, R. A. (1922). On the interpretation of chi-square from contingency tables, and the calculation of P. *J R Stat Soc*, 85(1):87–94.

Fitness, J., Floyd, S., Warndorff, D. K., Sichali, L., Malema, S., Crampin, A. C., Fine, P. E., and Hill, A. V. (2004). Large-scale candidate gene study of tuberculosis susceptibility in the Karonga district of northern Malawi. *Am J Trop Med Hyg*, 71(3):341–349.

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Stat Methods Med Res*, 2(2):121–45.

Florez, J. C. (2008). Newly identified loci highlight beta cell dysfunction as a key cause of type 2 diabetes: where are the insulin resistance genes? *Diabetologia*, 51(7):1100–1110.

Flynn, J. L., Goldstein, M. M., Triebold, K. J., Koller, B., and Bloom, B. R. (1992). Major histocompatibility complex class I-restricted T cells are required for resistance to Mycobacterium tuberculosis infection. *Proc Natl Acad Sci U S A*, 89(24):12013–12017.

Foghsgaard, L., Wissing, D., Mauch, D., Lademann, U., Bastholm, L., Boes, M., Elling, F., Leist, M., and Jäättelä, M. (2001). Cathepsin B acts as a dominant execution protease in tumor cell apoptosis induced by tumor necrosis factor. *J Biol Chem*, 153(5):999–1010.

Fowler, N. O. (1991). Tuberculous pericarditis. *JAMA*, 266(1):99–103.

Franceschi, D. S., Mazini, P. S., Rudnick, C. C., Sell, A. M., Tsuneto, L. T., Ribas, M. L., Peixoto, P. R., and Visentainer, J. E. (2009). Influence of TNF and IL10 gene polymorphisms in the immunopathogenesis of leprosy in the south of Brazil. *Int J Infect Dis*, 13(4):493–498.

Frazer, K. A., Ballinger, D. G., Cox, D. R., Hinds, D. A., Stuve, L. L., Gibbs, R. A., Belmont, J. W., Boudreau, A., Hardenbol, P., Leal, S. M., Pasternak, S., Wheeler, D. A., Willis, T. D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhao, H., Zhou, J., Gabriel, S. B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R. C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Shen, Y., Sun, W., Wang, H., Wang, Y., Wang, Y., Xiong, X., Xu, L., Waye, M. M. Y., Tsui, S. K. W., Xue, H., Wong, J. T.-F., Galver, L. M., Fan, J.-B., Gunderson, K., Murray, S. S., Oliphant, A. R., Chee, M. S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.-F., Phillips, M. S., Roumy, S., Sallée, C., Verner, A., Hudson, T. J., Kwok, P.-Y., Cai, D., Koboldt, D. C., Miller, R. D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.-C., Mak, W., Song, Y. Q., Tam, P. K. H., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., Tsunoda, T., Deloukas, P., Bird, C. P., Delgado, M., Dermitzakis, E. T., Gwilliam, R., Hunt, S., Morrison, J., Powell, D., Stranger, B. E., Whittaker, P., Bentley, D. R., Daly, M. J., de Bakker, P. I. W., Barrett, J., Chretien, Y. R., Maller, J., McCarroll, S., Patterson, N., Pe'er, I., Price, A., Purcell, S., Richter, D. J., Sabeti, P., Saxena, R., Schaffner, S. F., Sham, P. C., Varilly, P., Stein, L. D., Krishnan, L., Smith, A. V., Tello-Ruiz, M. K., Thorisson, G. A., Chakravarti, A., Chen, P. E., Cutler, D. J., Kashuk, C. S., Lin, S., Abecasis, G. R., Guan, W., Li, Y., Munro, H. M., Qin, Z. S., Thomas, D. J., McVean, G., Auton, A., Bottolo, L., Cardin, N., Eyheramendy, S., Freeman, C., Marchini, J., Myers, S., Spencer, C., Stephens, M., Donnelly, P., Cardon, L. R., Clarke, G., Evans, D. M., Morris, A. P., Weir, B. S., Mullikin, J. C., Sherry, S. T., Feolo, M., Skol, A., Zhang, H., Matsuda, I., Fukushima, Y., Macer, D. R., Suda,

E., Rotimi, C. N., Adebamowo, C. A., Ajayi, I., Aniagwu, T., Marshall, P. A., Nkwodimmah, C., Royal, C. D. M., Leppert, M. F., Dixon, M., Peiffer, A., Qiu, R., Kent, A., Kato, K., Niikawa, N., Adewole, I. F., Knoppers, B. M., Foster, M. W., Clayton, E. W., Watkin, J., Muzny, D., Nazareth, L., Sodergren, E., Weinstock, G. M., Yakub, I., Birren, B. W., Wilson, R. K., Fulton, L. L., Rogers, J., Burton, J., Carter, N. P., Clee, C. M., Griffiths, M., Jones, M. C., McLay, K., Plumb, R. W., Ross, M. T., Sims, S. K., Willey, D. L., Chen, Z., Han, H., Kang, L., Godbout, M., Wallenburg, J. C., L'Archevêque, P., Bellemare, G., Saeki, K., Wang, H., An, D., Fu, H., Li, Q., Wang, Z., Wang, R., Holden, A. L., Brooks, L. D., McEwen, J. E., Guyer, M. S., Wang, V. O., Peterson, J. L., Shi, M., Spiegel, J., Sung, L. M., Zacharia, L. F., Collins, F. S., Kennedy, K., Jamieson, R., and Stewart, J. (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–61.

Fremond, C. M., Yeremeev, V., Nicolle, D. M., Jacobs, M., Quesniaux, V. F., and Ryffel, B. (2004). Fatal Mycobacterium tuberculosis infection despite adaptive immune response in the absence of MyD88. *J Clin Invest*, 114(12):1790–1799.

Fu, L., Gao, Z., Zhang, X., Tsang, Y. H., Goh, H. K., Geng, H., Shimizu, N., Tsuchiyama, J., Srivastava, G., and Tao, Q. (2009). Frequent concomitant epigenetic silencing of the stress-responsive tumor suppressor gene CADM1, and its interacting partner DAL-1 in nasal NK/T-cell lymphoma. *Int J Cancer*, 124(7):1572–8.

Furuno, T., Ito, A., Koma, Y., Watabe, K., Yokozaki, H., Bienenstock, J., Nakanishi, M., and Kitamura, Y. (2005). The spermatogenic Ig superfamily/synaptic cell adhesion molecule mast-cell adhesion molecule promotes interaction with nerves. *J Immunol*, 174(11):6934–6942.

Gabriel, S. and Ziaugra, L. (2004). SNP genotyping using Sequenom MassARRAY

7K platform. *Curr Protoc Hum Genet*, Chapter 2:Unit 2.12.

Gabriel, S. B., Schaffner, S. F., Nguyen, H., Moore, J. M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., Liu-Cordero, S. N., Rotimi, C., Adeyemo, A., Cooper, R., Ward, R., Lander, E. S., Daly, M. J., and Altshuler, D. (2002). The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229.

Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B. C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M. C., Hilty, M., Hopewell, P. C., and Small, P. M. (2006). Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A*, 103(8):2869–2873.

Galibert, L., Diemer, G. S., Liu, Z., Johnson, R. S., Smith, J. L., Walzer, T., Comeau, M. R., Rauch, C. T., Wolfson, M. F., Sorensen, R. A., Van der Vuurst de Vries, A. R., Branstetter, D. G., Koelling, R. M., Scholler, J., Fanslow, W. C., Baum, P. R., Derry, J. M., and Yan, W. (2005). Nectin-like protein 2 defines a subset of T-cell zone dendritic cells and is a ligand for class-I-restricted T-cell-associated molecule. *J Biol Chem*, 280(23):21955–21964.

Gandotra, S., Jang, S., Murray, P. J., Salgame, P., and Ehrt, S. (2007). Nucleotide-binding oligomerization domain protein 2-deficient mice control infection with Mycobacterium tuberculosis. *Infect Immun*, 75(11):5127–34.

Gardam, M. A., Keystone, E. C., Menzies, R., Manners, S., Skamene, E., Long, R., and Vinh, D. C. (2003). Anti-tumour necrosis factor agents and tuberculosis risk: mechanisms of action and clinical management. *Lancet Infect Dis*, 3(3):148–55.

Gatfield, J. and Pieters, J. (2000). Essential role for cholesterol in entry of mycobacteria into macrophages. *Science*, 288(5471):1647–50.

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.

Goldfeld, A. E., Delgado, J. C., Thim, S., Bozon, M. V., Uglialoro, A. M., Turbay, D., Cohen, C., and Yunis, E. J. (1998). Association of an HLA-DQ allele with clinical tuberculosis. *JAMA*, 279(3):226–228.

Goldstein, D. B. (2009). Common genetic variation and human traits. *N Engl J Med*, 360(17):1696–1698.

Gomes, I., Collins, A., Lonjou, C., Thomas, N. S., Wilkinson, J., Watson, M., and Morton, N. (1999). Hardy-Weinberg quality control. *Ann Hum Genet*, 63(Pt 6):535–8.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P., and Tiedje, J. M. (2007). DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol*, 57(Pt 1):81–91.

Greco, S., Girardi, E., Navarra, A., and Saltini, C. (2006). Current evidence on diagnostic accuracy of commercially based nucleic acid amplification tests for the diagnosis of pulmonary tuberculosis. *Thorax*, 61(9):783–790.

Greenstein, R. J. (2003). Is Crohn's disease caused by a mycobacterium? Comparisons with leprosy, tuberculosis, and Johne's disease. *Lancet Infect Dis*, 3(8):507–14.

Greenwood, C. M., Fujiwara, T. M., Boothroyd, L. J., Miller, M. A., Frappier, D., Fanning, E. A., Schurr, E., and Morgan, K. (2000). Linkage of tuberculosis to chromosome 2q35 loci, including NRAMP1, in a large aboriginal Canadian family. *Am J Hum Genet*, 67(2):405–416.

Grossman, C. J. (1985). Interactions between the gonadal steroids and the immune system. *Science*, 227(4684):257–61.

Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., Lander, E. S., Schaffner, S. F., and Sabeti, P. C. (2010). A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science*.

Guinto, R. S., Doull, J. A., and De Guia, L. (1954). Mortality of persons with leprosy prior to sulfone therapy, Cordova and Talisay, Cebu, Philippines. *Int J Lepr*, 22(3):273–284.

Gulick, R. M., Lalezari, J., Goodrich, J., Clumeck, N., DeJesus, E., Horban, A., Nadler, J., Clotet, B., Karlsson, A., Wohlfeiler, M., Montana, J. B., McHale, M., Sullivan, J., Ridgway, C., Felstead, S., Dunne, M. W., van der Ryst, E., and Mayer, H. (2008). Maraviroc for previously treated patients with R5 HIV-1 infection. *N Engl J Med*, 359(14):1429–1441.

Gutierrez, M. G., Master, S. S., Singh, S. B., Taylor, G. A., Colombo, M. I., and Deretic, V. (2004). Autophagy is a defense mechanism inhibiting BCG and Mycobacterium tuberculosis survival in infected macrophages. *Cell*, 119(6):753–766.

Haak, W., Forster, P., Bramanti, B., Matsumura, S., Brandt, G., Tänzer, M., Villems, R., Renfrew, C., Gronenborn, D., Alt, K. W., and Burger, J. (2005). Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science*, 310(5750):1016–8.

Haldane, J. B. S. (1949). Disease and Evolution. *Ric Sci Suppl*, 19:68–76.

Hamilton, W. D., Axelrod, R., and Tanese, R. (1990). Sexual reproduction as an adaptation to resist parasites (a review). *Proc Natl Acad Sci U S A*, 87(9):3566–3573.

Han, J. W., Zheng, H. F., Cui, Y., Sun, L. D., Ye, D. Q., Hu, Z., Xu, J. H., Cai, Z. M., Huang, W., Zhao, G. P., Xie, H. F., Fang, H., Lu, Q. J., Li, X. P., Pan, Y. F., Deng, D. Q., Zeng, F. Q., Ye, Z. Z., Zhang, X. Y., Wang, Q. W., Hao, F., Ma, L., Zuo, X. B., Zhou, F. S., Du, W. H., Cheng, Y. L., Yang, J. Q., Shen, S. K., Li, J., Sheng, Y. J., Zuo, X. X., Zhu, W. F., Gao, F., Zhang, P. L., Guo, Q., Li, B., Gao, M., Xiao, F. L., Quan, C., Zhang, C., Zhang, Z., Zhu, K. J., Li, Y., Hu, D. Y., Lu, W. S., Huang, J. L., Liu, S. X., Li, H., Ren, Y. Q., Wang, Z. X., Yang, C. J., Wang, P. G., Zhou, W. M., Lv, Y. M., Zhang, A. P., Zhang, S. Q., Lin, D., Low, H. Q., Shen, M., Zhai, Z. F., Wang, Y., Zhang, F. Y., Yang, S., Liu, J. J., and Zhang, X. J. (2009). Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat Genet*, 41(11):1234–1237.

Han, X. Y., Seo, Y. H., Sizer, K. C., Schoberle, T., May, G. S., Spencer, J. S., Li, W., and Nair, R. G. (2008). A new Mycobacterium species causing diffuse lepromatous leprosy. *Am J Clin Pathol*, 130(6):856–864.

Hansen, G. H. A. (1874). Undersøgelser Angående Spedalskhedens Årsager. *Norsk Mag Laegervidenskaben*, (4):1–88.

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

Harris, J., De Haro, S. A., Master, S. S., Keane, J., Roberts, E. A., Delgado, M., and Deretic, V. (2007). T helper 2 cytokines inhibit autophagic control of intracellular Mycobacterium tuberculosis. *Immunity*, 27(3):505–517.

Havlir, D. V. and Barnes, P. F. (1999). Tuberculosis in patients with human immunodeficiency virus infection. *N Engl J Med*, 340(5):367–373.

Hawkins, B. R., Higgins, D. A., Chan, S. L., Lowrie, D. B., Mitchison, D. A., and

Girling, D. J. (1988). HLA typing in the Hong Kong Chest Service/British Medical Research Council study of factors associated with the breakdown to active tuberculosis of inactive pulmonary lesions. *Am Rev Respir Dis*, 138(6):1616–1621.

Hawn, T. R., Misch, E. A., Dunstan, S. J., Thwaites, G. E., Lan, N. T., Quy, H. T., Chau, T. T., Rodrigues, S., Nachman, A., Janer, M., Hien, T. T., Farrar, J. J., and Aderem, A. (2007). A common human TLR1 polymorphism regulates the innate immune response to lipopeptides. *Eur J Immunol*, 37(8):2280–2289.

Helgadottir, A., Thorleifsson, G., Magnusson, K. P., Grétarsdottir, S., Steinthorsdottir, V., Manolescu, A., Jones, G. T., Rinkel, G. J. E., Blankensteijn, J. D., Ronkainen, A., Jääskeläinen, J. E., Kyo, Y., Lenk, G. M., Sakalihasan, N., Kostulas, K., Gottsäter, A., Flex, A., Stefansson, H., Hansen, T., Andersen, G., Weinsheimer, S., Borch-Johnsen, K., Jorgensen, T., Shah, S. H., Quyyumi, A. A., Granger, C. B., Reilly, M. P., Austin, H., Levey, A. I., Vaccarino, V., Palsdottir, E., Walters, G. B., Jonsdottir, T., Snorradottir, S., Magnusdottir, D., Gudmundsson, G., Ferrell, R. E., Sveinbjornsdottir, S., Hernesniemi, J., Niemelä, M., Limet, R., Andersen, K., Sigurdsson, G., Benediktsson, R., Verhoeven, E. L. G., Teijink, J. A. W., Grobbee, D. E., Rader, D. J., Collier, D. A., Pedersen, O., Pola, R., Hillert, J., Lindblad, B., Valdimarsson, E. M., Magnadottir, H. B., Wijmenga, C., Tromp, G., Baas, A. F., Ruigrok, Y. M., van Rij, A. M., Kuivaniemi, H., Powell, J. T., Matthiasson, S. E., Gulcher, J. R., Thorgeirsson, G., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (2008). The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nat Genet*, 40(2):217–24.

Herb, F., Thye, T., Niemann, S., Browne, E. N. L., Chinbuah, M. A., Gyapong, J., Osei, I., Owusu-Dabo, E., Werz, O., Rüsch-Gerdes, S., Horstmann, R. D., and Meyer, C. G. (2008). ALOX5 variants associated with susceptibility to human

pulmonary tuberculosis. *Hum Mol Genet*, 17(7):1052–60.

Hermans, P. W., Abebe, F., Kuteyi, V. I., Kolk, A. H., Thole, J. E., and Harboe, M. (1995). Molecular and immunological characterization of the highly conserved antigen 84 from Mycobacterium tuberculosis and Mycobacterium leprae. *Infect Immun*, 63(3):954–60.

Hershkovitz, I., Donoghue, H. D., Minnikin, D. E., Besra, G. S., Lee, O. Y., Gernaey, A. M., Galili, E., Eshed, V., Greenblatt, C. L., Lemma, E., Bar-Gal, G. K., and Spigelman, M. (2008). Detection and molecular characterization of 9,000-year-old Mycobacterium tuberculosis from a Neolithic settlement in the Eastern Mediterranean. *PLoS ONE*, 3(10):e3426.

Higgins, J. P. T. and Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Stat Med*, 21(11):1539–58.

Hill, A. V. S. (2006). Aspects of genetic susceptibility to human infectious diseases. *Annu Rev Genet*, 40:469–86.

Hill, W. G. and Robertson, A. (1966). The effect of linkage on limits to artificial selection. *Genet Res*, 8(3):269–94.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106(23):9362–9367.

Hirsh, A. E., Tsolaki, A. G., DeRiemer, K., Feldman, M. W., and Small, P. M. (2004). Stable association between strains of Mycobacterium tuberculosis and their human host populations. *Proc Natl Acad Sci U S A*, 101(14):4871–6.

Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C., and Balding, D. J.

(2008). Genome-wide significance for dense SNP and resequencing data. *Genet Epidemiol*, 32(2):179–85.

Hollins, F., Kaur, D., Yang, W., Cruse, G., Saunders, R., Sutcliffe, A., Berger, P., Ito, A., Brightling, C. E., and Bradding, P. (2008). Human airway smooth muscle promotes human lung mast cell survival, proliferation, and constitutive activation: cooperative roles for CADM1, stem cell factor, and IL-6. *J Immunol*, 181(4):2772–2780.

Horng, T., Barton, G. M., Flavell, R. A., and Medzhitov, R. (2002). The adaptor molecule TIRAP provides signalling specificity for Toll-like receptors. *Nature*, 420(6913):329–333.

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.

Howard, R. S. and Lively, C. M. (1994). Parasitism, mutation accumulation and the maintenance of sex. *Nature*, 367(6463):554–557.

Hubbard, T. (2002). The Ensembl genome database project. *Nucleic Acids Res*, 30(1):38–41.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., and Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychol Methods*, 11(2):193–206.

Hughes, A. L. and Yeager, M. (1998). Natural selection at major histocompatibility complex loci of vertebrates. *Annu Rev Genet*, 32:415–435.

Hulse, E. V. (1972). Leprosy and ancient Egypt. *Lancet*, 2(7778):659.

Husson, R. N. and Young, R. A. (1987). Genes for the major protein antigens of

Mycobacterium tuberculosis: the etiologic agents of tuberculosis and leprosy share an immunodominant antigen. *Proc Natl Acad Sci U S A*, 84(6):1679–83.

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.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science*, 324(5924):218–23.

Ingram, V. M. (1956). A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin. *Nature*, 178(4537):792–794.

Inohara, N., Ogura, Y., Fontalba, A., Gutierrez, O., Pons, F., Crespo, J., Fukase, K., Inamura, S., Kusumoto, S., Hashimoto, M., Foster, S. J., Moran, A. P., Fernandez-Luna, J. L., and Nuñez, G. (2003). Host recognition of bacterial muramyl dipeptide mediated through NOD2. Implications for Crohn's disease. *J Biol Chem*, 278(8):5509–12.

Ioannidis, J. P. A. (2007). Non-replication and inconsistency in the genome-wide association setting. *Hum Hered*, 64(4):203–13.

Ivanov, I. I., McKenzie, B. S., Zhou, L., Tadokoro, C. E., Lepelley, A., Lafaille, J. J., Cua, D. J., and Littman, D. R. (2006). The orphan nuclear receptor RORgammat directs the differentiation program of proinflammatory IL-17+ T helper cells. *Cell*, 126(6):1121–33.

Iwasaki, A. and Medzhitov, R. (2004). Toll-like receptor control of the adaptive immune responses. *Nat Immunol*, 5(10):987–995.

Iwata, M., Eshima, Y., and Kagechika, H. (2003). Retinoic acids exert direct effects on T cells to suppress Th1 development and enhance Th2 development via retinoic acid receptors. *Int Immunol*, 15(8):1017–25.

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*.

Ji, W., Foo, J. N., O'Roak, B. J., Zhao, H., Larson, M. G., Simon, D. B., Newton-Cheh, C., State, M. W., Levy, D., and Lifton, R. P. (2008). Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet*, 40(5):592–9.

Jin, M. S., Kim, S. E., Heo, J. Y., Lee, M. E., Kim, H. M., Paik, S.-G., Lee, H., and Lee, J.-O. (2007). Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. *Cell*, 130(6):1071–82.

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):7520–7524.

Jones, D., Metzger, H. J., Schatz, A., and Waksman, S. A. (1944). Control of Gram-negative bacteria in experimental animals by streptomycin. *Science*, 100(2588):103–105.

Jopling, W. H. (1991). Leprosy stigma. *Lepr Rev*, 62(1):1–12.

Jouanguy, E., Lamhamedi-Cherradi, S., Lammas, D., Dorman, S. E., Fondanèche, M. C., Dupuis, S., Döffinger, R., Altare, F., Girdlestone, J., Emile, J. F., Ducoulombier, H., Edgar, D., Clarke, J., Oxelius, V. A., Brai, M., Novelli, V., Heyne, K., Fischer, A., Holland, S. M., Kumararatne, D. S., Schreiber, R. D., and Casanova, J. L. (1999). A human IFNGR1 small deletion hotspot associated with dominant susceptibility to mycobacterial infection. *Nat Genet*, 21(4):370–8.

Kakegawa, H., Matano, Y., Inubushi, T., and Katunuma, N. (2004). Significant accumulations of cathepsin B and prolylendopeptidase in inflammatory focus of delayed-type hypersensitivity induced by Mycobacterium tuberculosis in mice. *Biochem Biophys Res Commun*, 316(1):78–84.

Kang, P. B., Azad, A. K., Torrelles, J. B., Kaufman, T. M., Beharka, A., Tibesar, E., DesJardin, L. E., and Schlesinger, L. S. (2005). The human macrophage mannose receptor directs Mycobacterium tuberculosis lipoarabinomannan-mediated phagosome biogenesis. *J Exp Med*, 202(7):987–999.

Kang, T. J. and Chae, G. T. (2001). Detection of Toll-like receptor 2 (TLR2) mutation in the lepromatous leprosy patients. *J Exp Med*, 31(1):53–58.

Kaul, D., Anand, P. K., and Verma, I. (2004). Cholesterol-sensor initiates M. tuberculosis entry into human macrophages. *Mol Cell Biol*, 258(1-2):219–22.

Keating, B. J., Tischfield, S., Murray, S. S., Bhangale, T., Price, T. S., Glessner, J. T., Galver, L., Barrett, J. C., Grant, S. F., Farlow, D. N., Chandrupatla, H. R., Hansen, M., Ajmal, S., Papanicolaou, G. J., Guo, Y., Li, M., Derohannessian, S., de Bakker, P. I., Bailey, S. D., Montpetit, A., Edmondson, A. C., Taylor, K., Gai, X., Wang, S. S., Fornage, M., Shaikh, T., Groop, L., Boehnke, M., Hall, A. S., Hattersley, A. T., Frackelton, E., Patterson, N., Chiang, C. W., Kim, C. E., Fabsitz, R. R., Ouwehand, W., Price, A. L., Munroe, P., Caulfield, M., Drake, T., Boerwinkle, E., Reich, D., Whitehead, A. S., Cappola, T. P., Samani, N. J., Lusis, A. J., Schadt, E., Wilson, J. G., Koenig, W., McCarthy, M. I., Kathiresan, S., Gabriel, S. B., Hakonarson, H., Anand, S. S., Reilly, M., Engert, J. C., Nickerson, D. A., Rader, D. J., Hirschhorn, J. N., and Fitzgerald, G. A. (2008). Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PLoS ONE*, 3(10):e3583.

Khader, S. A., Bell, G. K., Pearl, J. E., Fountain, J. J., Rangel-Moreno, J., Cilley, G. E., Shen, F., Eaton, S. M., Gaffen, S. L., Swain, S. L., Locksley, R. M., Haynes, L., Randall, T. D., and Cooper, A. M. (2007). IL-23 and IL-17 in the establishment of protective pulmonary CD4+ T cell responses after vaccination and during Mycobacterium tuberculosis challenge. *Nat Immunol*, 8(4):369–377.

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., Kwiatkowski, D. P., Day, N. P., Yala, D., Crook, D. W., Marsh, K., Berkley, J. A., O'Neill, L. A., and Hill, A. V. (2007). A Mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. *Nat Genet*, 39(4):523–528.

Kindler, V., Sappino, A. P., Grau, G. E., Piguet, P. F., and Vassalli, P. (1989). The inducing role of tumor necrosis factor in the development of bactericidal granulomas during BCG infection. *Cell*, 56(5):731–740.

Kirchheimer, W. F. and Storrs, E. E. (1971). Attempts to establish the armadillo (Dasypus novemcinctus Linn.) as a model for the study of leprosy. I. Report of lepromatoid leprosy in an experimentally infected armadillo. *Int J Lepr Other Mycobact Dis*, 39(3):693–702.

Kirchheimer, W. F., Storrs, E. E., and Binford, C. H. (1972). Attempts to establish the Armadillo (Dasypus novemcinctus linn.) as a model for the study of leprosy. II. Histopathologic and bacteriologic post-mortem findings in lepromatoid leprosy in the Armadillo. *Int J Lepr Other Mycobact Dis*, 40(3):229–42.

Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K., and Ding, L. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*, 25(17):2283–2285.

Koch, R. (1882). Die Atiologie der Tuberkulose. *Berl Klin Wochenschr*, (15):221–230.

Köhler, S., Bauer, S., Horn, D., and Robinson, P. N. (2008). Walking the interactome for prioritization of candidate disease genes. *Am J Hum Genet*, 82(4):949–58.

Kolk, A. H., Ho, M. L., Klatser, P. R., Eggelte, T. A., Kuijper, S., de Jonge, S., and van Leeuwen, J. (1984). Production and characterization of monoclonal antibodies to Mycobacterium tuberculosis, M. bovis (BCG) and M. leprae. *Clin Exp Immunol*, 58(3):511–21.

König, R., Stertz, S., Zhou, Y., Inoue, A., Hoffmann, H.-H., Bhattacharyya, S., Alamares, J. G., Tscherne, D. M., Ortigoza, M. B., Liang, Y., Gao, Q., Andrews, S. E., Bandyopadhyay, S., De Jesus, P., Tu, B. P., Pache, L., Shih, C., Orth, A.,

Bonamy, G., Miraglia, L., Ideker, T., García-Sastre, A., Young, J. A. T., Palese, P., Shaw, M. L., and Chanda, S. K. (2010). Human host factors required for influenza virus replication. *Nature*, 463(7282):813–7.

Koppel, E. A., Ludwig, I. S., Hernandez, M. S., Lowary, T. L., Gadikota, R. R., Tuzikov, A. B., Vandenbroucke-Grauls, C. M., van Kooyk, Y., Appelmelk, B. J., and Geijtenbeek, T. B. (2004). Identification of the mycobacterial carbohydrate structure that binds the C-type lectins DC-SIGN, L-SIGN and SIGNR1. *Immunobiology*, 209(1-2):117–127.

Kramnik, I., Demant, P., and Bloom, B. B. (1998). Susceptibility to tuberculosis as a complex genetic trait: analysis using recombinant congenic strains of mice. *Novartis Found Symp*, 217:120–127.

Krutzik, S. R., Ochoa, M. T., Sieling, P. A., Uematsu, S., Ng, Y. W., Legaspi, A., Liu, P. T., Cole, S. T., Godowski, P. J., Maeda, Y., Sarno, E. N., Norgard, M. V., Brennan, P. J., Akira, S., Rea, T. H., and Modlin, R. L. (2003). Activation and regulation of Toll-like receptors 2 and 1 in human leprosy. *Nat Med*, 9(5):525–532.

Kuijl, C., Savage, N. D. L., Marsman, M., Tuin, A. W., Janssen, L., Egan, D. A., Ketema, M., van Den Nieuwendijk, R., van Den Eeden, S. J. F., Geluk, A., Poot, A., van Der Marel, G., Beijersbergen, R. L., Overkleeft, H., Ottenhoff, T. H. M., and Neefjes, J. (2007). Intracellular bacterial growth is controlled by a kinase network around PKB/AKT1. *Nature*, 450(7170):725–30.

Kumar, D., Nath, L., Kamal, M. A., Varshney, A., Jain, A., Singh, S., and Rao, K. V. S. (2010). Genome-wide analysis of the host intracellular network that regulates survival of Mycobacterium tuberculosis. *Cell*, 140(5):731–43.

Kuramochi, M., Fukuhara, H., Nobukuni, T., Kanbe, T., Maruyama, T., Ghosh, H. P., Pletcher, M., Isomura, M., Onizuka, M., Kitamura, T., Sekiya, T., Reeves,

R. H., and Murakami, Y. (2001). TSLC1 is a tumor-suppressor gene in human non-small-cell lung cancer. *Nat Genet*, 27(4):427–430.

Kusuhara, K., Yamamoto, K., Okada, K., Mizuno, Y., and Hara, T. (2007). Association of IL12RB1 polymorphisms with susceptibility to and severity of tuberculosis in Japanese: a gene-based association analysis of 21 candidate genes. *Int J Immunogenet*, 34(1):35–44.

Lage, K., Karlberg, E. O., Stø rling, Z. M., Olason, P. I., Pedersen, A. G., Rigina, O., Hinsby, A. M., Tümer, Z., Pociot, F., Tommerup, N., Moreau, Y., and Brunak, S. r. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol*, 25(3):309–16.

Laird, N. M. and Lange, C. (2006). Family-based designs in the age of large-scale gene-association studies. *Nat Rev Genet*, 7(5):385–94.

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

Lander, H. M., Tauras, J. M., Ogiste, J. S., Hori, O., Moss, R. A., and Schmidt, A. M. (1997). Activation of the receptor for advanced glycation end products triggers a p21(ras)-dependent mitogen-activated protein kinase pathway regulated by oxidant stress. *J Biol Chem*, 272(28):17810–4.

Laraque, F., Griggs, A., Slopen, M., and Munsiff, S. S. (2009). Performance of nucleic acid amplification tests for diagnosis of tuberculosis in a large urban setting. *Clin Infect Dis*, 49(1):46–54.

Larsen, K. C., Spencer, A. J., Goodman, A. L., Gilchrist, A., Furze, J., Rollier, C. S., Kiss-Toth, E., Gilbert, S. C., Bregu, M., Soilleux, E. J., Hill, A. V., and Wyllie, D. H. (2009). Expression of tak1 and tram induces synergistic pro-inflammatory signalling and adjuvants DNA vaccines. *Vaccine*, 27(41):5589–5598.

Launois, P., Niang, M. N., Sarthou, J. L., Rivier, F., Drowart, A., Van Vooren, J. P., Millan, J., and Huygen, K. (1993). T-cell stimulation with purified mycobacterial antigens in patients and healthy subjects infected with Mycobacterium leprae: secreted antigen 85 is another immunodominant antigen. *Scand J Immunol*, 38(2):167–76.

Lee, P. P. W., Chan, K.-W., Jiang, L., Chen, T., Li, C., Lee, T.-L., Mak, P. H. S., Fok, S. F. S., Yang, X., and Lau, Y.-L. (2008). Susceptibility to mycobacterial infections in children with X-linked chronic granulomatous disease: a review of 17 patients living in a region endemic for tuberculosis. *Pediatr Infect Dis J*, 27(3):224–30.

Levee, G., Schurr, E., and Pandey, J. P. (1997). Tumor necrosis factor-alpha, interleukin-1-beta and immunoglobulin (GM and KM) polymorphisms in leprosy. A linkage study. *Exp Clin Immunogenet*, 14(2):160–165.

Levine, B. and Deretic, V. (2007). Unveiling the roles of autophagy in innate and adaptive immunity. *Nat Rev Immunol*, 7(10):767–777.

Lewis, K. N., Liao, R., Guinn, K. M., Hickey, M. J., Smith, S., Behr, M. A., and Sherman, D. R. (2003). Deletion of RD1 from Mycobacterium tuberculosis mimics bacille Calmette-Guérin attenuation. *J Infect Dis*, 187(1):117–23.

Lewis, S. J., Baker, I., and Davey Smith, G. (2005). Meta-analysis of vitamin D receptor polymorphisms and pulmonary tuberculosis risk. *Int J Tuberc Lung Dis*, 9(10):1174–1177.

Lewontin, R. C. (1964). The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. *Genetics*, 49(1):49–67.

Li, H. T., Zhang, T. T., Zhou, Y. Q., Huang, Q. H., and Huang, J. (2006). SLC11A1 (formerly NRAMP1) gene polymorphisms and tuberculosis susceptibility: a meta-analysis. *Int J Tuberc Lung Dis*, 10(1):3–12.

Li, J. H., D'Alessio, A., and Pober, J. S. (2009). Lipopolysaccharide can trigger a cathepsin B-dependent programmed death response in human endothelial cells. *Am J Pathol*, 175(3):1124–35.

Liebana, E., Girvin, R. M., Welsh, M., Neill, S. D., and Pollock, J. M. (1999). Generation of CD8(+) T-cell responses to Mycobacterium bovis and mycobacterial antigen in experimental bovine tuberculosis. *Infect Immun*, 67(3):1034–1044.

Lienhardt, C., Bennett, S., Del Prete, G., Bah-Sow, O., Newport, M., Gustafson, P., Manneh, K., Gomes, V., Hill, A., and McAdam, K. (2002). Investigation of environmental and host-related risk factors for tuberculosis in Africa. I. Methodological aspects of a combined design. *Am J Epidemiol*, 155(11):1066–1073.

Lifton, R. P. (1996). Molecular genetics of human blood pressure variation. *Science*, 272(5262):676–80.

Lin, D. Y. and Zou, F. (2004). Assessing genomewide statistical significance in linkage studies. *Genet Epidemiol*, 27(3):202–214.

Lindesmith, L., Moe, C., Marionneau, S., Ruvoen, N., Jiang, X., Lindblad, L., Stewart, P., LePendu, J., and Baric, R. (2003). Human susceptibility and resistance to Norwalk virus infection. *Nat Med*, 9(5):548–553.

Lindholm, E., Bakhtadze, E., Cilio, C., Agardh, E., Groop, L., and Agardh, C.-D. (2008). Association between LTA, TNF and AGER polymorphisms and late diabetic complications. *PLoS ONE*, 3(6):e2546.

Lindholm, E., Bakhtadze, E., Sjögren, M., Cilio, C. M., Agardh, E., Groop, L., and Agardh, C.-D. (2006). The -374 T/A polymorphism in the gene encoding RAGE is associated with diabetic nephropathy and retinopathy in type 1 diabetic patients. *Diabetologia*, 49(11):2745–55.

Liu, F., van Duijn, K., Vingerling, J. R., Hofman, A., Uitterlinden, A. G., Janssens, A. C. J. W., and Kayser, M. (2009). Eye color and the prediction of complex phenotypes from genotypes. *Curr Biol*, 19(5):R192–3.

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):367–377.

Liu, W., Cao, W. C., Zhang, C. Y., Tian, L., Wu, X. M., Habbema, J. D., Zhao, Q. M., Zhang, P. H., Xin, Z. T., Li, C. Z., and Yang, H. (2004). VDR and NRAMP1 gene polymorphisms in susceptibility to pulmonary tuberculosis among the Chinese Han population: a case-control study. *Int J Tuberc Lung Dis*, 8(4):428–434.

Lively, C. M. and Dybdahl, M. F. (2000). Parasite adaptation to locally common host genotypes. *Nature*, 405(6787):679–681.

Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S., and Hirschhorn, J. N. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet*, 33(2):177–82.

Lupski, J. R., Reid, J. G., Gonzaga-Jauregui, C., Rio Deiros, D., Chen, D. C. Y., Nazareth, L., Bainbridge, M., Dinh, H., Jing, C., Wheeler, D. A., McGuire, A. L., Zhang, F., Stankiewicz, P., Halperin, J. J., Yang, C., Gehman, C., Guo, D., Irikat, R. K., Tom, W., Fantin, N. J., Muzny, D. M., and Gibbs, R. A. (2010). Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *N Engl J Med*, 362(13):1181–91.

MacMicking, J. D., Taylor, G. A., and McKinney, J. D. (2003). Immune control of tuberculosis by IFN-gamma-inducible LRG-47. *Science*, 302(5645):654–9.

Maeda, Y., Gidoh, M., Ishii, N., Mukai, C., and Makino, M. (2003). Assessment

of cell mediated immunogenicity of Mycobacterium leprae-derived antigens. *Cell Immunol*, 222(1):69–77.

Malhotra, D., Darvishi, K., Lohra, M., Kumar, H., Grover, C., Sood, S., Reddy, B. S., and Bamezai, R. N. (2006). Association study of major risk single nucleotide polymorphisms in the common regulatory region of PARK2 and PACRG genes with leprosy in an Indian population. *Eur J Hum Genet*, 14(4):438–442.

Malhotra, D., Darvishi, K., Sood, S., Sharma, S., Grover, C., Relhan, V., Reddy, B. S., and Bamezai, R. N. (2005). IL-10 promoter single nucleotide polymorphisms are significantly associated with resistance to leprosy. *Hum Genet*, 118(2):295–300.

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., Chakravarit, 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. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461:747–753.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst*, 22(4):719–48.

Marchini, J., Howie, B., Myers, S., McVean, G., and Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–913.

Marras, T. K. (2005). Dexamethasone for tuberculous meningitis. *N Engl J Med*, 352(6):628–30; author reply 628–30.

Martinez, A. and Kolter, R. (1997). Protection of DNA during oxidative stress by the nonspecific DNA-binding protein Dps. *J Bacteriol*, 179(16):5188–5194.

McCarthy, M. I. (2009). Exploring the unknown: assumptions about allelic architecture and strategies for susceptibility variant discovery. *Genome Med*, 1(7):66.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., and Hirschhorn, J. N. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet*, 9(5):356–69.

McCarthy, M. I. and Hirschhorn, J. N. (2008). Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet*, 17(R2):R156–65.

McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R., Hinds, D. A., Pennacchio, L. A., Tybjaerg-Hansen, A., Folsom, A. R., Boerwinkle, E., Hobbs, H. H., and Cohen, J. C. (2007). A common allele on chromosome 9 associated with coronary heart disease. *Science*, 316(5830):1488–91.

Means, T. K., Wang, S., Lien, E., Yoshimura, A., Golenbock, D. T., and Fenton, M. J. (1999). Human toll-like receptors mediate cellular activation by Mycobacterium tuberculosis. *J Biol Chem*, 163(7):3920–7.

Medina-Contreras, O., Soldevila, G., Patiño Lopez, G., Canche-Pool, E., Valle-Rios, R., and Ortiz-Navarrete, V. (2010). Role of CRTAM during mouse early T lymphocytes development. *Dev Comp Immunol*, 34(2):196–202.

Medzhitov, R. (2001). Toll-like receptors and innate immunity. *Nat Rev Immunol*, 1(2):135–145.

Medzhitov, R. and Janeway Jr., C. (2000). Innate immunity. *N Engl J Med*, 343(5):338–344.

Mehra, N. K., Rajalingam, R., Mitra, D. K., Taneja, V., and Giphart, M. J. (1995). Variants of HLA-DR2/DR51 group haplotypes and susceptibility to tuberculoid leprosy and pulmonary tuberculosis in Asian Indians. *Int J Lepr Other Mycobact Dis*, 63(2):241–248.

Melancon-Kaplan, J., Hunter, S. W., McNeil, M., Stewart, C., Modlin, R. L., Rea, T. H., Convit, J., Salgame, P., Mehra, V., Bloom, B. R., and al., E. (1988). Immunological significance of Mycobacterium leprae cell walls. *Proc Natl Acad Sci U S A*, 85(6):1917–1921.

Michels, E., Hoebeeck, J., De Preter, K., Schramm, A., Brichard, B., De Paepe, A., Eggert, A., Laureys, G., Vandesompele, J., and Speleman, F. (2008). CADM1 is a strong neuroblastoma candidate gene that maps within a 3.72 Mb critical region of loss on 11q23. *BMC Cancer*, 8:173.

Miller, E. N., Jamieson, S. E., Joberty, C., Fakiola, M., Hudson, D., Peacock, C. S., Cordell, H. J., Shaw, M. A., Lins-Lainson, Z., Shaw, J. J., Ramos, F., Silveira, F., and Blackwell, J. M. (2004). Genome-wide scans for leprosy and tuberculosis susceptibility genes in Brazilians. *Genes Immun*, 5(1):63–67.

Mira, M. T., Alcais, A., Nguyen, V. T., Moraes, M. O., Di Flumeri, C., Vu, H. T., Mai, C. P., Nguyen, T. H., Nguyen, N. B., Pham, X. K., Sarno, E. N., Alter, A., Montpetit, A., Moraes, M. E., Moraes, J. R., Dore, C., Gallant, C. J., Lepage, P., Verner, A., Van De Vosse, E., Hudson, T. J., Abel, L., and Schurr, E. (2004). Susceptibility to leprosy is associated with PARK2 and PACRG. *Nature*, 427(6975):636–640.

Mira, M. T., Alcaïs, A., Van Thuc, N., Thai, V. H., Huong, N. T., Ba, N. N., Verner, A., Hudson, T. J., Abel, L., and Schurr, E. (2003). Chromosome 6q25 is linked to susceptibility to leprosy in a Vietnamese population. *Nat Genet*, 33(3):412–5.

Misch, E. A., Macdonald, M., Ranjit, C., Sapkota, B. R., Wells, R. D., Siddiqui, M. R., Kaplan, G., and Hawn, T. R. (2008). Human TLR1 Deficiency Is Associated with Impaired Mycobacterial Signaling and Protection from Leprosy Reversal Reaction. *PLoS Negl Trop Dis*, 2(5):e231.

Miyanaga, K., Juji, T., Maeda, H., Nakajima, S., and Kobayashi, S. (1981). Tuberculoid leprosy and HLA in Japanese. *Tissue Antigens*, 18(5):331–334.

Modiano, D., Chiucchiuini, A., Petrarca, V., Sirima, B. S., Luoni, G., Roggero, M. A., Corradin, G., Coluzzi, M., and Esposito, F. (1999). Interethnic differences in the humoral response to non-repetitive regions of the Plasmodium falciparum circumsporozoite protein. *Am J Trop Med Hyg*, 61(4):663–667.

Modiano, D., Petrarca, V., Sirima, B. S., Nebie, I., Diallo, D., Esposito, F., and Coluzzi, M. (1996). Different response to Plasmodium falciparum malaria in west African sympatric ethnic groups. *Proc Natl Acad Sci U S A*, 93(23):13206–13211.

Modlin, R. L. (1994). Th1-Th2 paradigm: insights from leprosy. *J Invest Dermatol*, 102(6):828–832.

Modlin, R. L., Gebhard, J. F., Taylor, C. R., and Rea, T. H. (1983a). In situ characterization of T lymphocyte subsets in the reactional states of leprosy. *Clin Exp Immunol*, 53(1):17–24.

Modlin, R. L., Hofman, F. M., Taylor, C. R., and Rea, T. H. (1983b). T lymphocyte subsets in the skin lesions of patients with leprosy. *J Am Acad Dermatol*, 8(2):182–189.

Modlin, R. L., Melancon-Kaplan, J., Young, S. M., Pirmez, C., Kino, H., Convit, J., Rea, T. H., and Bloom, B. R. (1988). Learning from lesions: patterns of tissue inflammation in leprosy. *Proc Natl Acad Sci U S A*, 85(4):1213–1217.

Mombo, L. E., Lu, C. Y., Ossari, S., Bedjabaga, I., Sica, L., Krishnamoorthy, R., and Lapoumeroulie, C. (2003). Mannose-binding lectin alleles in sub-Saharan Africans and relation with susceptibility to infections. *Genes Immun*, 4(5):362–367.

Monot, M., Honore, N., Garnier, T., Araoz, R., Coppee, J. Y., Lacroix, C., Sow, S., Spencer, J. S., Truman, R. W., Williams, D. L., Gelber, R., Virmond, M., Flageul,

B., Cho, S. N., Ji, B., Paniz-Mondolfi, A., Convit, J., Young, S., Fine, P. E., Rasolofo, V., Brennan, P. J., and Cole, S. T. (2005). On the origin of leprosy. *Science*, 308(5724):1040–1042.

Monot, M., Honore, N., Garnier, T., Zidane, N., Sherafi, D., Paniz-Mondolfi, A., Matsuoka, M., Taylor, G. M., Donoghue, H. D., Bouwman, A., Mays, S., Watson, C., Lockwood, D., Khamispour, A., Dowlati, Y., Jianping, S., Rea, T. H., Vera-Cabrera, L., Stefani, M. M., Banu, S., Macdonald, M., Sapkota, B. R., Spencer, J. S., Thomas, J., Harshman, K., Singh, P., Busso, P., Gattiker, A., Rougemont, J., Brennan, P. J., and Cole, S. T. (2009). Comparative genomic and phylogeographic analysis of Mycobacterium leprae. *Nat Genet*, 41(12):1282–1289.

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., Guigo, R., and Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, 464(7289):773–7.

Moraes, M. O., Sarno, E. N., Almeida, A. S., Saraiva, B. C., Nery, J. A., Martins, R. C., and Sampaio, E. P. (1999). Cytokine mRNA expression in leprosy: a possible role for interferon-gamma and interleukin-12 in reactions (RR and ENL). *Scand J Immunol*, 50(5):541–549.

Morahan, G., Kaur, G., Singh, M., Rapthap, C. C., Kumar, N., Katoch, K., Mehra, N. K., and Huang, D. (2007). Association of variants in the IL12B gene with leprosy and tuberculosis. *Tissue Antigens*, 69 Suppl 1:234–236.

Morgan, T. H., Sturtevant, A. H., and Bridge, C. B. (1920). The Evidence for the Linear Order of the Genes. *Proc Natl Acad Sci U S A*, 6(4):162–4.

Mueller, J. C. (2004). Linkage disequilibrium for different scales and applications. *Brief Bioinform*, 5(4):355–64.

Mullis, K. B. and Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol*, 155:335–350.

Munoz, S., Rivas-Santiago, B., and Enciso, J. A. (2009). Mycobacterium tuberculosis entry into mast cells through cholesterol-rich membrane microdomains. *Scand J Immunol*, 70(3):256–63.

Nair, S., Ramaswamy, P. A., Ghosh, S., Joshi, D. C., Pathak, N., Siddiqui, I., Sharma, P., Hasnain, S. E., Mande, S. C., and Mukhopadhyay, S. (2009). The PPE18 of Mycobacterium tuberculosis interacts with TLR2 and activates IL-10 induction in macrophage. *J Immunol*, 183(10):6269–6281.

Naser, S., Ghobrial, G., Romero, C., and Valentine, J. (2004). Culture of subspecies from the blood of patients with Crohn's disease. *Lancet*, 364(9439):1039–1044.

Neale, B. M. and Sham, P. C. (2004). The future of association studies: gene-based analysis and replication. *Am J Hum Genet*, 75(3):353–62.

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48(3):443–53.

Neeper, M., Schmidt, A. M., Brett, J., Yan, S. D., Wang, F., Pan, Y. C., Elliston, K., Stern, D., and Shaw, A. (1992). Cloning and expression of a cell surface receptor for advanced glycosylation end products of proteins. *J Biol Chem*, 267(21):14998–5004.

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.

Nesic, D., Miller, M. C., Quinkert, Z. T., Stein, M., Chait, B. T., and Stebbins,

C. E. (2010). Helicobacter pylori CagA inhibits PAR1-MARK family kinases by mimicking host substrates. *Nat Struct Mol Biol*, 17(1):130–2.

Nevo, Y. and Nelson, N. (2004). The mutation F227I increases the coupling of metal ion transport in DCT1. *J Biol Chem*, 279(51):53056–53061.

Newport, M. J., Huxley, C. M., Huston, S., Hawrylowicz, C. M., Oostra, B. A., Williamson, R., and Levin, M. (1996). A mutation in the interferon-gamma-receptor gene and susceptibility to mycobacterial infection. *N Engl J Med*, 335(26):1941–1949.

Neyrolles, O. and Quintana-Murci, L. (2009). Sexual inequality in tuberculosis. *PLoS Med*, 6(12):e1000199.

Ng, S. B., Buckingham, K. J., Lee, C., Bigham, A. W., Tabor, H. K., Dent, K. M., Huff, C. D., Shannon, P. T., Jabs, E. W., Nickerson, D. A., Shendure, J., and Bamshad, M. J. (2010). Exome sequencing identifies the cause of a mendelian disorder. *Nat Genet*, 42(1):30–5.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., Lee, C., Shaffer, T., Wong, M., Bhattacharjee, A., Eichler, E. E., Bamshad, M., Nickerson, D. A., and Shendure, J. (2009). Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*, 461(7261):272–6.

Nigou, J., Zelle-Rieser, C., Gilleron, M., Thurnher, M., and Puzo, G. (2001). Mannosylated lipoarabinomannans inhibit IL-12 production by human dendritic cells: evidence for a negative signal delivered through the mannose receptor. *J Immunol*, 166(12):7477–7485.

Noordeen, S. K. (1972). Mortality in leprosy. *Indian J Med Res*, 60(3):439–445.

Norman, G., Joseph, G., and Richard, J. (2004). Validity of the WHO Operational

Classification and Value of Other Clinical Signs in the Classification of Leprosy. *Int J Lepr Other Mycobact Dis*, 72(3):278.

Ogus, A. C., Yoldas, B., Ozdemir, T., Uguz, A., Olcen, S., Keser, I., Coskun, M., Cilli, A., and Yegin, O. (2004). The Arg753GLn polymorphism of the human toll-like receptor 2 gene in tuberculosis disease. *Eur Respir J*, 23(2):219–223.

Orme, I. M., Miller, E. S., Roberts, A. D., Furney, S. K., Griffin, J. P., Dobos, K. M., Chi, D., Rivoire, B., and Brennan, P. J. (1992). T lymphocytes mediating protection and cellular cytolysis during the course of Mycobacterium tuberculosis infection. Evidence for different kinetics and recognition of a wide spectrum of protein antigens. *J Immunol*, 148(1):189–196.

O'Sullivan, M. P., O'Leary, S., Kelly, D. M., and Keane, J. (2007). A caspase-independent pathway mediates macrophage cell death in response to Mycobacterium tuberculosis infection. *Infect Immun*, 75(4):1984–93.

Overmeer, R. M., Henken, F. E., Snijders, P. J. F., Claassen-Kramer, D., Berkhof, J., Helmerhorst, T. J. M., Heideman, D. A. M., Wilting, S. M., Murakami, Y., Ito, A., Meijer, C. J. L. M., and Steenbergen, R. D. M. (2008). Association between dense CADM1 promoter methylation and reduced protein expression in high-grade CIN and cervical SCC. *J Pathol*, 215(4):388–97.

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., and Milos, P. M. (2009). Direct RNA sequencing. *Nature*, 461(7265):814–8.

Pandey, A. K. and Sassetti, C. M. (2008). Mycobacterial persistence requires the utilization of host cholesterol. *Proc Natl Acad Sci U S A*, 105(11):4376–80.

Parkes, M., Barrett, J. C., Prescott, N. J., Tremelling, M., Anderson, C. A., Fisher, S. A., Roberts, R. G., Nimmo, E. R., Cummings, F. R., Soars, D., Drummond,

H., Lees, C. W., Khawaja, S. A., Bagnall, R., Burke, D. A., Todhunter, C. E., Ahmad, T., Onnie, C. M., McArdle, W., Strachan, D., Bethel, G., Bryan, C., Lewis, C. M., Deloukas, P., Forbes, A., Sanderson, J., Jewell, D. P., Satsangi, J., Mansfield, J. C., Cardon, L., and Mathew, C. G. (2007). Sequence variants in the autophagy gene IRGM and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nat Genet*, 39(7):830–832.

Patnaik, P. B. (1949). The non-central chi2- and F-distributions and their applications. *Biometrika*, 36(1-2):202–32.

Pauling, L., Itano, H. A., and al., E. (1949). Sickle cell anemia a molecular disease. *Science*, 110(2865):543–548.

Peng, G., Luo, L., Siu, H., Zhu, Y., Hu, P., Hong, S., Zhao, J., Zhou, X., Reveille, J. D., Jin, L., Amos, C. I., and Xiong, M. (2010). Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet*, 18(1):111–7.

Peters, A. D. and Lively, C. M. (1999). The Red Queen and Fluctuating Epistasis: A Population Genetic Analysis of Antagonistic Coevolution. *Am Nat*, 154(4):393–405.

Picard, C., Fieschi, C., Altare, F., Al-Jumaah, S., Al-Hajjar, S., Feinberg, J., Dupuis, S., Soudais, C., Al-Mohsen, I. Z., Genin, E., Lammas, D., Kumararatne, D. S., Leclerc, T., Rafii, A., Frayha, H., Murugasu, B., Wah, L. B., Sinniah, R., Loubser, M., Okamoto, E., Al-Ghonaium, A., Tufenkeji, H., Abel, L., and Casanova, J. L. (2002). Inherited interleukin-12 deficiency: IL12B genotype and clinical phenotype of 13 patients from six kindreds. *Am J Hum Genet*, 70(2):336–348.

Pickrell, J. K., Coop, G., Novembre, J., Kudaravalli, S., Li, J. Z., Absher, D., Srinivasan, B. S., Barsh, G. S., Myers, R. M., Feldman, M. W., and Pritchard, J. K.

(2009). Signals of recent positive selection in a worldwide sample of human populations. *Genome Res*, 19(5):826–37.

Pickrell, J. K., Marioni, J. C., Pai, A. A., Degner, J. F., Engelhardt, B. E., Nkadori, E., Veyrieras, J.-B., Stephens, M., Gilad, Y., and Pritchard, J. K. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464(7289):768–72.

Pierce, E. S. (2009). Where are all the Mycobacterium avium subspecies paratuberculosis in patients with Crohn's disease? *PLoS Pathog*, 5(3):e1000234.

Piersimoni, C. and Scarparo, C. (2003). Relevance of commercial amplification methods for direct detection of Mycobacterium tuberculosis complex in clinical samples. *J Clin Microbiol*, 41(12):5355–5365.

Pönnighaus, J. M., Fine, P. E., Sterne, J. A., Wilson, R. J., Msosa, E., Gruer, P. J., Jenkins, P. A., Lucas, S. B., Liomba, N. G., and Bliss, L. (1992). Efficacy of BCG vaccine against leprosy and tuberculosis in northern Malawi. *Lancet*, 339(8794):636–9.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., 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.

Purcell, S. M., Wray, N. R., Stone, J. L., Visscher, P. M., O'Donovan, M. C., Sullivan, P. F., and Sklar, P. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–752.

Quesniaux, V. J., Nicolle, D. M., Torres, D., Kremer, L., Guerardel, Y., Nigou, J., Puzo, G., Erard, F., and Ryffel, B. (2004). Toll-like receptor 2 (TLR2)-dependent-

positive and TLR2-independent-negative regulation of proinflammatory cytokines by mycobacterial lipomannans. *J Immunol*, 172(7):4425–4434.

Raelson, J. V., Little, R. D., Ruether, A., Fournier, H., Paquin, B., Van Eerdewegh, P., Bradley, W. E. C., Croteau, P., Nguyen-Huu, Q., Segal, J., Debrus, S., Allard, R., Rosenstiel, P., Franke, A., Jacobs, G., Nikolaus, S., Vidal, J.-M., Szego, P., Laplante, N., Clark, H. F., Paulussen, R. J., Hooper, J. W., Keith, T. P., Belouchi, A., and Schreiber, S. (2007). Genome-wide association study for Crohn's disease in the Quebec Founder Population identifies multiple validated disease loci. *Proc Natl Acad Sci U S A*, 104(37):14747–52.

Rajalingam, R., Mehra, N. K., Jain, R. C., Myneedu, V. P., and Pande, J. N. (1996). Polymerase chain reaction–based sequence-specific oligonucleotide hybridization analysis of HLA class II antigens in pulmonary tuberculosis: relevance to chemotherapy and disease severity. *J Infect Dis*, 173(3):669–676.

Rani, R., Fernandez-Vina, M. A., Zaheer, S. A., Beena, K. R., and Stastny, P. (1993). Study of HLA class II alleles by PCR oligotyping in leprosy patients from north India. *Tissue Antigens*, 42(3):133–137.

Rasmussen, M., Li, Y., Lindgreen, S., Pedersen, J. S., Albrechtsen, A., Moltke, I., Metspalu, M., Metspalu, E., Kivisild, T., Gupta, R., Bertalan, M., Nielsen, K., Gilbert, M. T. P., Wang, Y., Raghavan, M., Campos, P. F., Kamp, H. M., Wilson, A. S., Gledhill, A., Tridico, S., Bunce, M., Lorenzen, E. D., Binladen, J., Guo, X., Zhao, J., Zhang, X., Zhang, H., Li, Z., Chen, M., Orlando, L., Kristiansen, K., Bak, M., Tommerup, N., Bendixen, C., Pierre, T. L., Grø nnow, B., Meldgaard, M., Andreasen, C., Fedorova, S. A., Osipova, L. P., Higham, T. F. G., Ramsey, C. B., Hansen, T. V. O., Nielsen, F. C., Crawford, M. H., Brunak, S. r., Sicheritz-Pontén, T., Villems, R., Nielsen, R., Krogh, A., Wang, J., and Willerslev, E. (2010). Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature*,

463(7282):757–62.

Ravikumar, M., Dheenadhayalan, V., Rajaram, K., Lakshmi, S. S., Kumaran, P. P., Paramasivan, C. N., Balakrishnan, K., and Pitchappan, R. M. (1999). Associations of HLA-DRB1, DQB1 and DPB1 alleles with pulmonary tuberculosis in south India. *Tuber Lung Dis*, 79(5):309–317.

Rees, R. J., Waters, M. F., Weddell, A. G., and Palmer, E. (1967). Experimental lepromatous leprosy. *Nature*, 215(5101):599–602.

Reich, D., Thangaraj, K., Patterson, N., Price, A. L., and Singh, L. (2009). Reconstructing Indian population history. *Nature*, 461(7263):489–94.

Relethford, J. H. (2002). Apportionment of global human genetic diversity based on craniometrics and skin color. *Am J Phys Anthropol*, 118(4):393–8.

Remus, N., Alcais, A., and Abel, L. (2003). Human Genetics of Common Mycobacterial Infections. *Immunol Res*, 28(2):109–130.

Remus, N., El Baghdadi, J., Fieschi, C., Feinberg, J., Quintin, T., Chentoufi, M., Schurr, E., Benslimane, A., Casanova, J.-L., and Abel, L. (2004). Association of IL12RB1 polymorphisms with pulmonary tuberculosis in adults in Morocco. *J Infect Dis*, 190(3):580–7.

Renaud, J. P., Rochel, N., Ruff, M., Vivat, V., Chambon, P., Gronemeyer, H., and Moras, D. (1995). Crystal structure of the RAR-gamma ligand-binding domain bound to all-trans retinoic acid. *Nature*, 378(6558):681–9.

Ridley, D. S. and Jopling, W. H. (1962). A classification of leprosy for research purposes. *Lepr Rev*, 33:119–28.

Ridley, D. S. and Jopling, W. H. (1966). Classification of leprosy according to immunity. A five-group system. *Int J Lepr Other Mycobact Dis*, 34(3):255–273.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–7.

Roach, D. R., Bean, A. G., Demangel, C., France, M. P., Briscoe, H., and Britton, W. J. (2002). TNF regulates chemokine induction essential for cell recruitment, granuloma formation, and clearance of mycobacterial infection. *J Immunol*, 168(9):4620–4627.

Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Hubley, R., Shannon, P. T., Rowen, L., Pant, K. P., Goodman, N., Bamshad, M., Shendure, J., Drmanac, R., Jorde, L. B., Hood, L., and Galas, D. J. (2010). Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing. *Science*.

Robbins, G., Tripathy, V. M., Misra, V. N., Mohanty, R. K., Shinde, V. S., Gray, K. M., and Schug, M. D. (2009). Ancient skeletal evidence for leprosy in India (2000 B.C.). *PLoS ONE*, 4(5):e5669.

Rodrigues, L. C., Diwan, V. K., and Wheeler, J. G. (1993). Protective effect of BCG against tuberculous meningitis and miliary tuberculosis: a meta-analysis. *Int J Epidemiol*, 22(6):1154–1158.

Rohde, K., Yates, R. M., Purdy, G. E., and Russell, D. G. (2007). Mycobacterium tuberculosis and the environment within the phagosome. *Immunol Rev*, 219:37–54.

Ronning, D. R., Klabunde, T., Besra, G. S., Vissa, V. D., Belisle, J. T., and Sacchettini, J. C. (2000). Crystal structure of the secreted form of antigen 85C reveals potential targets for mycobacterial drugs and vaccines. *Nat Struct Biol*, 7(2):141–6.

Ronning, D. R., Vissa, V., Besra, G. S., Belisle, J. T., and Sacchettini, J. C. (2004). Mycobacterium tuberculosis antigen 85A and 85C structures confirm binding orientation and conserved substrate specificity. *J Biol Chem*, 279(35):36771–7.

Roy, S., Frodsham, A., Saha, B., Hazra, S. K., Mascie-Taylor, C. G., and Hill, A. V. (1999). Association of vitamin D receptor genotype with leprosy type. *J Infect Dis*, 179(1):187–191.

Roy, S., McGuire, W., Mascie-Taylor, C. G., Saha, B., Hazra, S. K., Hill, A. V., and Kwiatkowski, D. (1997). Tumor necrosis factor promoter polymorphism and susceptibility to lepromatous leprosy. *J Infect Dis*, 176(2):530–532.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–8.

Russell, D. G. (2001). Mycobacterium tuberculosis: here today, and here tomorrow. *Nat Rev Mol Cell Biol*, 2(8):569–577.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., and Lander, E. S. (2006). Positive natural selection in the human lineage. *Science*, 312(5780):1614–1620.

Sager, P., Schalimtzer, M., and Moller-Christensen, V. (1972). A case of spondylitis tuberculosa in the Danish Neolithic Age. *Dan Med Bull*, 19(5):176–180.

Sagoo, G. S., Little, J., and Higgins, J. P. T. (2009). Systematic reviews of genetic association studies. Human Genome Epidemiology Network. *PLoS Med*, 6(3):e28.

Sahiratmadja, E., Baak-Pablo, R., de Visser, A. W., Alisjahbana, B., Adnan, I., van Crevel, R., Marzuki, S., van Dissel, J. T., Ottenhoff, T. H. M., and van De Vosse,

E. (2007). Association of polymorphisms in IL-12/IFN-gamma pathway genes with susceptibility to pulmonary tuberculosis in Indonesia. *Tuberculosis*, 87(4):303–11.

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):487–491.

Saiki, R. K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G. T., Erlich, H. A., and Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732):1350–4.

Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B., Dixon, R. J., Meitinger, T., Braund, P., Wichmann, H. E., Barrett, J. H., Konig, I. R., Stevens, S. E., Szymczak, S., Tregouet, D. A., Iles, M. M., Pahlke, F., Pollard, H., Lieb, W., Cambien, F., Fischer, M., Ouwehand, W., Blankenberg, S., Balmforth, A. J., Baessler, A., Ball, S. G., Strom, T. M., Braenne, I., Gieger, C., Deloukas, P., Tobin, M. D., Ziegler, A., Thompson, J. R., and Schunkert, H. (2007). Genomewide association analysis of coronary artery disease. *N Engl J Med*, 357(5):443–453.

Sampaio, E. P., Moreira, A. L., Sarno, E. N., Malta, A. M., and Kaplan, G. (1992). Prolonged treatment with recombinant interferon gamma induces erythema nodosum leprosum in lepromatous leprosy patients. *J Exp Med*, 175(6):1729–1737.

Sanger, F. and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol*, 94(3):441–8.

Sanjeevi, C. B., Narayanan, P. R., Prabakar, R., Charles, N., Thomas, B. E., Balasubramaniam, R., and Olerup, O. (1992). No association or linkage with HLA-DR or -DQ genes in south Indians with pulmonary tuberculosis. *Tuber Lung Dis*,

73(5):280–284.

Sanjuan, M. A., Dillon, C. P., Tait, S. W., Moshiach, S., Dorsey, F., Connell, S., Komatsu, M., Tanaka, K., Cleveland, J. L., Withoff, S., and Green, D. R. (2007). Toll-like receptor signalling in macrophages links the autophagy pathway to phagocytosis. *Nature*, 450(7173):1253–1257.

Santos, A. R., Suffys, P. N., Vanderborght, P. R., Moraes, M. O., Vieira, L. M., Cabello, P. H., Bakker, A. M., Matos, H. J., Huizinga, T. W., Ottenhoff, T. H., Sampaio, E. P., and Sarno, E. N. (2002). Role of tumor necrosis factor-alpha and interleukin-10 promoter gene polymorphisms in leprosy. *J Infect Dis*, 186(11):1687–1691.

Saporta, L. and Yuksel, A. (1994). Androgenic status in patients with lepromatous leprosy. *Br J Urol*, 74(2):221–224.

Sarno, E. N., Grau, G. E., Vieira, L. M., and Nery, J. A. (1991). Serum levels of tumour necrosis factor-alpha and interleukin-1 beta during leprosy reactional states. *Clin Exp Immunol*, 84(1):103–108.

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics*, 53(4):1253–61.

Scanga, C. A., Bafica, A., Feng, C. G., Cheever, A. W., Hieny, S., and Sher, A. (2004). MyD88-deficient mice display a profound loss in resistance to Mycobacterium tuberculosis associated with partially impaired Th1 cytokine and nitric oxide synthase 2 expression. *Infect Immun*, 72(4):2400–2404.

Schauf, V., Ryan, S., Scollard, D., Jonasson, O., Brown, A., Nelson, K., Smith, T., and Vithayasai, V. (1985). Leprosy associated with HLA-DR2 and DQw1 in the population of northern Thailand. *Tissue Antigens*, 26(4):243–247.

Schlesinger, L. S. (1993). Macrophage phagocytosis of virulent but not attenuated strains of Mycobacterium tuberculosis is mediated by mannose receptors in addition to complement receptors. *J Immunol*, 150(7):2920–2930.

Schlesinger, L. S., Bellinger-Kawahara, C. G., Payne, N. R., and Horwitz, M. A. (1990). Phagocytosis of Mycobacterium tuberculosis is mediated by human monocyte complement receptors and complement component C3. *J Immunol*, 144(7):2771–2780.

Schlesinger, L. S. and Horwitz, M. A. (1991). Phenolic glycolipid-1 of Mycobacterium leprae binds complement component C3 in serum and mediates phagocytosis by human monocytes. *J Exp Med*, 174(5):1031–1038.

Schmidt, D., Wilson, M. D., Spyrou, C., Brown, G. D., Hadfield, J., and Odom, D. T. (2009). ChIP-seq: using high-throughput sequencing to discover protein-DNA interactions. *Nat Methods*, 48(3):240–8.

Schurr, E. and Gros, P. (2009). A common genetic fingerprint in leprosy and Crohn's disease? *N Engl J Med*, 361(27):2666–8.

Scollard, D. M., Adams, L. B., Gillis, T. P., Krahenbuhl, J. L., Truman, R. W., and Williams, D. L. (2006). The continuing challenges of leprosy. *Clin Microbiol Rev*, 19(2):338–81.

Sechi, L. A., Scanu, A. M., Molicotti, P., Cannas, S., Mura, M., Dettori, G., Fadda, G., and Zanetti, S. (2005). Detection and Isolation of Mycobacterium avium subspecies paratuberculosis from intestinal mucosal biopsies of patients with and without Crohn's disease in Sardinia. *Am J Gastroenterol*, 100(7):1529–36.

Selvaraj, P., Narayanan, P. R., and Reetha, A. M. (2000). Association of vitamin D receptor genotypes with the susceptibility to pulmonary tuberculosis in female patients & resistance in female contacts. *Indian J Med Res*, 111:172–179.

Selwyn, P. A., Hartel, D., Lewis, V. A., Schoenbaum, E. E., Vermund, S. H., Klein, R. S., Walker, A. T., and Friedland, G. H. (1989). A prospective study of the risk of tuberculosis among intravenous drug users with human immunodeficiency virus infection. *N Engl J Med*, 320(9):545–50.

Senior, K. (2009). Stigma, chemoprophylaxis, and leprosy control. *Lancet Infect Dis*, 9(1):10.

Setia, M., Steinmaus, C., Ho, C., and Rutherford, G. (2006). The role of BCG in prevention of leprosy: a meta-analysis. *Lancet Infect Dis*, 6(3):162–170.

Shankarkumar, U., Ghosh, K., Badakere, S., and Mohanty, D. (2003). Novel HLA Class I Alleles Associated with Indian Leprosy Patients. *J Biomed Biotechnol*, 2003(3):208–211.

Shields, E. D., Russell, D. A., and Pericak-Vance, M. A. (1987). Genetic epidemiology of the susceptibility to leprosy. *J Clin Invest*, 79(4):1139–1143.

Shinnick, T. M., Sweetser, D., Thole, J., van Embden, J., and Young, R. A. (1987). The etiologic agents of leprosy and tuberculosis share an immunoreactive protein antigen with the vaccine strain Mycobacterium bovis BCG. *Infect Immun*, 55(8):1932–5.

Siddiqui, M. R., Meisner, S., Tosh, K., Balakrishnan, K., Ghei, S., Fisher, S. E., Golding, M., Shanker Narayan, N. P., Sitaraman, T., Sengupta, U., Pitchappan, R., and Hill, A. V. (2001). A major susceptibility locus for leprosy in India maps to chromosome 10p13. *Nat Genet*, 27(4):439–41.

Singh, A., Crossman, D. K., Mai, D., Guidry, L., Voskuil, M. I., Renfrow, M. B., and Steyn, A. J. C. (2009). Mycobacterium tuberculosis WhiB3 maintains redox homeostasis by regulating virulence lipid anabolism to modulate macrophage response. *PLoS Pathog*, 5(8):e1000545.

Singh, S. B., Davis, A. S., Taylor, G. A., and Deretic, V. (2006). Human IRGM induces autophagy to eliminate intracellular mycobacteria. *Science*, 313(5792):1438–1441.

Singh, S. P., Mehra, N. K., Dingley, H. B., Pande, J. N., and Vaidya, M. C. (1983). Human leukocyte antigen (HLA)-linked control of susceptibility to pulmonary tuberculosis and association with HLA-DR types. *J Infect Dis*, 148(4):676–681.

Singh, S. P., Mehra, N. K., Dingley, H. B., Pande, J. N., and Vaidya, M. C. (1984). HLA haplotype segregation study in multiple case families of pulmonary tuberculosis. *Tissue Antigens*, 23(1):84–86.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T. J., Montpetit, A., Pshezhetsky, A. V., Prentki, M., Posner, B. I., Balding, D. J., Meyre, D., Polychronakos, C., and Froguel, P. (2007). A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–5.

Smith, D. G. and Guinto, R. S. (1978). Leprosy and fertility. *Hum Biol*, 50(4):451–460.

Smyth, D. J., Cooper, J. D., Bailey, R., Field, S., Burren, O., Smink, L. J., Guja, C., Ionescu-Tirgoviste, C., Widmer, B., Dunger, D. B., Savage, D. A., Walker, N. M., Clayton, D. G., and Todd, J. A. (2006). A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat Genet*, 38(6):617–619.

Soborg, C., Madsen, H. O., Andersen, A. B., Lillebaek, T., Kok-Jensen, A., and Garred, P. (2003). Mannose-binding lectin polymorphisms in clinical tuberculosis. *J Infect Dis*, 188(5):777–782.

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.

Spellberg, B. and Edwards Jr., J. E. (2001). Type 1/Type 2 immunity in infectious diseases. *Clin Infect Dis*, 32(1):76–102.

Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*, 52(3):506–16.

Sreevatsan, S., Pan, X., Stockbauer, K. E., Connell, N. D., Kreiswirth, B. N., Whittam, T. S., and Musser, J. M. (1997). Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination. *Proc Natl Acad Sci U S A*, 94(18):9869–9874.

Stead, W. W. (1992). Genetics and resistance to tuberculosis. Could resistance be enhanced by genetic engineering? *Ann Intern Med*, 116(11):937–941.

Stead, W. W., Senner, J. W., Reddick, W. T., and Lofgren, J. P. (1990). Racial differences in susceptibility to infection by Mycobacterium tuberculosis. *N Engl J Med*, 322(7):422–427.

Steinberg, B. E. and Grinstein, S. (2008). Pathogen destruction versus intracellular survival: the role of lipids as phagosomal fate determinants. *J Clin Invest*, 118(6):2002–11.

Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G. B., Styrkarsdottir, U., Gretarsdottir, S., Emilsson, V., Ghosh, S., Baker, A., Snorradottir, S., Bjarnason, H., Ng, M. C. Y., Hansen, T., Bagger, Y., Wilensky, R. L., Reilly, M. P., Adeyemo, A., Chen, Y., Zhou, J., Gudnason, V., Chen, G., Huang, H., Lashley, K., Doumatey, A., So, W.-Y., Ma, R. C. Y., Andersen, G., Borch-Johnsen, K., Jorgensen, T., van Vliet-Ostaptchouk, J. V.,

Hofker, M. H., Wijmenga, C., Christiansen, C., Rader, D. J., Rotimi, C., Gurney, M., Chan, J. C. N., Pedersen, O., Sigurdsson, G., Gulcher, J. R., Thorsteinsdottir, U., Kong, A., and Stefansson, K. (2007). A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat Genet*, 39(6):770–5.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H., and Wanker, E. E. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–68.

Stenger, S., Mazzaccaro, R. J., Uyemura, K., Cho, S., Barnes, P. F., Rosat, J. P., Sette, A., Brenner, M. B., Porcelli, S. A., Bloom, B. R., and Modlin, R. L. (1997). Differential effects of cytolytic T cell subsets on intracellular infection. *Science*, 276(5319):1684–1687.

Stephensen, C. B., Rasooly, R., Jiang, X., Ceddia, M. A., Weaver, C. T., Chandraratna, R. A. S., and Bucy, R. P. (2002). Vitamin A enhances in vitro Th2 development via retinoid X receptor pathway. *J Immunol*, 168(9):4495–503.

Stumpf, M. P. H., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. *Proc Natl Acad Sci U S A*, 105(19):6959–64.

Szeszko, J. S., Healy, B., Stevens, H., Balabanova, Y., Drobniewski, F., Todd, J. A., and Nejentsev, S. (2007). Resequencing and association analysis of the SP110 gene in adult pulmonary tuberculosis. *Hum Genet*, 121(2):155–60.

Tailleux, L., Schwartz, O., Herrmann, J. L., Pivert, E., Jackson, M., Amara, A., Legres, L., Dreher, D., Nicod, L. P., Gluckman, J. C., Lagrange, P. H., Gicquel,

B., and Neyrolles, O. (2003). DC-SIGN is the major Mycobacterium tuberculosis receptor on human dendritic cells. *J Exp Med*, 197(1):121–127.

Takeuchi, O., Sato, S., Horiuchi, T., Hoshino, K., Takeda, K., Dong, Z., Modlin, R. L., and Akira, S. (2002). Cutting edge: role of Toll-like receptor 1 in mediating immune response to microbial lipoproteins. *J Immunol*, 169(1):10–14.

Talarico, S., Cave, M. D., Marrs, C. F., Foxman, B., Zhang, L., and Yang, Z. (2005). Variation of the Mycobacterium tuberculosis PE_PGRS 33 gene among clinical isolates. *J Clin Microbiol*, 43(10):4954–60.

Tapping, R. I. and Tobias, P. S. (2003). Mycobacterial lipoarabinomannan mediates physical interactions between TLR1 and TLR2 to induce signaling. *J Endotoxin Res*, 9(4):264–268.

The International HapMap Consortium (2003). The International HapMap Project. *Nature*, 426(6968):789–96.

The International HapMap Consortium (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

The Karonga Prevention Trial Group (1996). Randomised controlled trial of single BCG, repeated BCG, or combined BCG and killed Mycobacterium leprae vaccine for prevention of leprosy and tuberculosis in Malawi. Karonga Prevention Trial Group. *Lancet*, 348(9019):17–24.

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.

The World Health Organization (2007). Global leprosy situation, 2007. *Wkly Epidemiol Rec*, 82(25):225–232.

The World Health Organization (2009). Crunch time for tuberculosis control. *Lancet*, 373(9670):1145.

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.

Thuong, N. T., Hawn, T. R., Thwaites, G. E., Chau, T. T., Lan, N. T., Quy, H. T., Hieu, N. T., Aderem, A., Hien, T. T., Farrar, J. J., and Dunstan, S. J. (2007). A polymorphism in human TLR2 is associated with increased susceptibility to tuberculous meningitis. *Genes Immun*, 8(5):422–428.

Thye, T., Browne, E. N., Chinbuah, M. A., Gyapong, J., Osei, I., Owusu-Dabo, E., Niemann, S., Rusch-Gerdes, S., Horstmann, R. D., and Meyer, C. G. (2006). No associations of human pulmonary tuberculosis with Sp110 variants. *J Med Genet*, 43(7):e32.

Thye, T., Nejentsev, S., Intemann, C. D., Browne, E. N., Chinbuah, M. A., Gyapong, J., Osei, I., Owusu-Dabo, E., Zeitels, L. R., Herb, F., Horstmann, R. D., and Meyer, C. G. (2009). MCP-1 promoter variant -362C associated with protection from pulmonary tuberculosis in Ghana, West Africa. *Hum Mol Genet*, 18(2):381–8.

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):857–864.

Torkamani, A., Topol, E. J., and Schork, N. J. (2008). Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, 92(5):265–72.

Tosh, K., Campbell, S. J., Fielding, K., Sillah, J., Bah, B., Gustafson, P., Manneh, K., Lisse, I., Sirugo, G., Bennett, S., Aaby, P., McAdam, K. P., Bah-Sow, O., Lienhardt, C., Kramnik, I., and Hill, A. V. (2006a). Variants in the SP110 gene are associated with genetic susceptibility to tuberculosis in West Africa. *Proc Natl Acad Sci U S A*, 103(27):10364–10368.

Tosh, K., Meisner, S., Siddiqui, M. R., Balakrishnan, K., Ghei, S., Golding, M., Sengupta, U., Pitchappan, R. M., and Hill, A. V. (2002a). A region of chromosome 20 is linked to leprosy susceptibility in a South Indian population. *J Infect Dis*, 186(8):1190–1193.

Tosh, K., Meisner, S., Siddiqui, M. R., Balakrishnan, K., Ghei, S., Golding, M., Sengupta, U., Pitchappan, R. M., and Hill, A. V. S. (2002b). A region of chromosome 20 is linked to leprosy susceptibility in a South Indian population. *J Infect Dis*, 186(8):1190–3.

Tosh, K., Ravikumar, M., Bell, J. T., Meisner, S., Hill, A. V., and Pitchappan, R. (2006b). Variation in MICA and MICB genes and enhanced susceptibility to paucibacillary leprosy in South India. *Hum Mol Genet*, 15(19):2880–2887.

Tso, H. W., Ip, W. K., Chong, W. P., Tam, C. M., Chiang, A. K., and Lau, Y. L. (2005). Association of interferon gamma and interleukin 10 genes with tuberculosis in Hong Kong Chinese. *Genes Immun*, 6(4):358–363.

Underhill, D. M., Ozinsky, A., Smith, K. D., and Aderem, A. (1999). Toll-like receptor-2 mediates mycobacteria-induced proinflammatory signaling in

macrophages. *Proc Natl Acad Sci U S A*, 96(25):14459–14463.

van Der Geize, R., Yam, K., Heuser, T., Wilbrink, M. H., Hara, H., Anderton, M. C., Sim, E., Dijkhuizen, L., Davies, J. E., Mohn, W. W., and Eltis, L. D. (2007). A gene cluster encoding cholesterol catabolism in a soil actinomycete provides insight into Mycobacterium tuberculosis survival in macrophages. *Proc Natl Acad Sci U S A*, 104(6):1947–52.

van Der Weyden, L., Arends, M. J., Chausiaux, O. E., Ellis, P. J., Lange, U. C., Surani, M. A., Affara, N., Murakami, Y., Adams, D. J., and Bradley, A. (2006). Loss of TSLC1 causes male infertility due to a defect at the spermatid stage of spermatogenesis. *Mol Cell Biol*, 26(9):3595–3609.

van Eden, W., de Vries, R. R., D'Amaro, J., Schreuder, I., Leiker, D. L., and van Rood, J. J. (1982). HLA-DR-associated genetic control of the type of leprosy in a population from surinam. *Hum Immunol*, 4(4):343–350.

van Eden, W., de Vries, R. R., Mehra, N. K., Vaidya, M. C., D'Amaro, J., and van Rood, J. J. (1980). HLA segregation of tuberculoid leprosy: confirmation of the DR2 marker. *J Infect Dis*, 141(6):693–701.

van Eden, W., Gonzalez, N. M., de Vries, R. R., Convit, J., and van Rood, J. J. (1985). HLA-linked control of predisposition to lepromatous leprosy. *J Infect Dis*, 151(1):9–14.

Vandal, O. H., Pierini, L. M., Schnappinger, D., Nathan, C. F., and Ehrt, S. (2008). A membrane protein preserves intrabacterial pH in intraphagosomal Mycobacterium tuberculosis. *Nat Med*, 14(8):849–854.

Vannberg, F. O., Chapman, S. J., Khor, C. C., Tosh, K., Floyd, S., Jackson-Sillah, D., Crampin, A., Sichali, L., Bah, B., Gustafson, P., Aaby, P., McAdam, K. P.,

Bah-Sow, O., Lienhardt, C., Sirugo, G., Fine, P., and Hill, A. V. (2008). CD209 genetic polymorphism and tuberculosis disease. *PLoS ONE*, 3(1):e1388.

Vejbaesya, S., Mahaisavariya, P., Luangtrakool, P., and Sermduangprateep, C. (2007). TNF alpha and NRAMP1 polymorphisms in leprosy. *J Med Assoc Thai*, 90(6):1188–1192.

Vergne, I., Chua, J., Lee, H. H., Lucas, M., Belisle, J., and Deretic, V. (2005). Mechanism of phagolysosome biogenesis block by viable Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A*, 102(11):4033–4038.

Visentainer, J. E., Tsuneto, L. T., Serra, M. F., Peixoto, P. R., and Petzl-Erler, M. L. (1997). Association of leprosy with HLA-DR2 in a Southern Brazilian population. *Braz J Med Biol Res*, 30(1):51–59.

Visscher, P. M., Hill, W. G., and Wray, N. R. (2008). Heritability in the genomics era–concepts and misconceptions. *Nat Rev Genet*, 9(4):255–66.

Wang, K., Li, M., and Bucan, M. (2007). Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am J Hum Genet*, 81(6).

Wang, K., Zhang, H., Kugathasan, S., Annese, V., Bradfield, J. P., Russell, R. K., Sleiman, P. M. A., Imielinski, M., Glessner, J., Hou, C., Wilson, D. C., Walters, T., Kim, C., Frackelton, E. C., Lionetti, P., Barabino, A., Van Limbergen, J., Guthery, S., Denson, L., Piccoli, D., Li, M., Dubinsky, M., Silverberg, M., Griffiths, A., Grant, S. F. A., Satsangi, J., Baldassano, R., and Hakonarson, H. (2009). Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn Disease. *Am J Hum Genet*, 84(3):399–405.

Watson, J. D. (1990). The human genome project: past, present, and future. *Science*, 248(4951):44–9.

Watson, J. D. and Crick, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., Freathy, R. M., Perry, J. R. B., Stevens, S., Hall, A. S., Samani, N. J., Shields, B., Prokopenko, I., Farrall, M., Dominiczak, A., Johnson, T., Bergmann, S., Beckmann, J. S., Vollenweider, P., Waterworth, D. M., Mooser, V., Palmer, C. N. A., Morris, A. D., Ouwehand, W. H., Zhao, J. H., Li, S., Loos, R. J. F., Barroso, I., Deloukas, P., Sandhu, M. S., Wheeler, E., Soranzo, N., Inouye, M., Wareham, N. J., Caulfield, M., Munroe, P. B., Hattersley, A. T., McCarthy, M. I., and Frayling, T. M. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet*, 40(5):575–83.

Weinberg, W. (1908). Ueber den Nachweis der Vererbung beim Menschen. *Jh Ver vaterl Naturk Württemb*, 64:369–82.

Welcsh, P. L. and King, M. C. (2001). BRCA1 and BRCA2 and the genetics of breast and ovarian cancer. *Hum Mol Genet*, 10(7):705–13.

Wilkins, M. H. F., Stokes, A. R., and Wilson, H. R. (1953). Molecular structure of deoxypentose nucleic acids. *Nature*, 171(4356):738–40.

Wilkinson, R. J., Llewelyn, M., Toossi, Z., Patel, P., Pasvol, G., Lalvani, A., Wright, D., Latif, M., and Davidson, R. N. (2000). Influence of vitamin D deficiency and vitamin D receptor polymorphisms on tuberculosis among Gujarati Asians in west London: a case-control study. *Lancet*, 355(9204):618–621.

Wilson, P. K. (2006). Confronting "hereditary" disease: eugenic attempts to eliminate tuberculosis in progressive era America. *J Med Humanit*, 27(1):19–37.

Wong, S. H., Vannberg, F. O., Spencer, A. J., van Der Weyden, L., Hill, A. V. S., and Wyllie, D. H. (2010). Comment on "CRTAM Confers Late-Stage Activa-

tion of CD8+ T Cells to Regulate Retention within Lymph Node". *J Immunol*, 184(8):4052–4053.

Woolhouse, M. E., Webster, J. P., Domingo, E., Charlesworth, B., and Levin, B. R. (2002). Biological and biomedical implications of the co-evolution of pathogens and their hosts. *Nat Genet*, 32(4):569–577.

Wright, S. (1921). Systems of Mating. V. General Considerations. *Genetics*, 6(2):167–78.

Wright, S. (1952). The Theoretical Variance within and among Subdivisions of a Population That Is in a Steady State. *Genetics*, 37(3):312–21.

Xu, K. Y., de Vries, R. R., Fei, H. M., van Leeuwen, A., Chen, R. B., and Ye, G. Y. (1985). HLA-linked control of predisposition to lepromatous leprosy. *Int J Lepr Other Mycobact Dis*, 53(1):56–63.

Yamamoto, Y., Tomioka, H., Sato, K., Saito, H., Yamada, Y., and Setogawa, T. (1990). Sex differences in the susceptibility of mice to infection induced by Mycobacterium intracellulare. *Am Rev Respir Dis*, 142(2):430–3.

Yamamura, M., Wang, X. H., Ohmen, J. D., Uyemura, K., Rea, T. H., Bloom, B. R., and Modlin, R. L. (1992). Cytokine patterns of immunologically mediated tissue damage. *J Immunol*, 149(4):1470–1475.

Yang, X., Deignan, J. L., Qi, H., Zhu, J., Qian, S., Zhong, J., Torosyan, G., Majid, S., Falkard, B., Kleinhanz, R. R., Karlsson, J., Castellani, L. W., Mumick, S., Wang, K., Xie, T., Coon, M., Zhang, C., Estrada-Smith, D., Farber, C. R., Wang, S. S., van Nas, A., Ghazalpour, A., Zhang, B., Macneil, D. J., Lamb, J. R., Dipple, K. M., Reitman, M. L., Mehrabian, M., Lum, P. Y., Schadt, E. E., Lusis, A. J., and Drake, T. A. (2009). Validation of candidate causal genes for obesity that affect shared metabolic pathways and networks. *Nat Genet*, 41(4):415–23.

Yang, X. O., Pappu, B. P., Nurieva, R., Akimzhanov, A., Kang, H. S., Chung, Y., Ma, L., Shah, B., Panopoulos, A. D., Schluns, K. S., Watowich, S. S., Tian, Q., Jetten, A. M., and Dong, C. (2008). T helper 17 lineage differentiation is programmed by orphan nuclear receptors ROR alpha and ROR gamma. *Immunity*, 28(1):29–39.

Yeh, J. H., Sidhu, S. S., and Chan, A. C. (2008). Regulation of a late phase of T cell polarity and effector functions by Crtam. *Cell*, 132(5):846–859.

Young, D. and Robertson, B. (2001). Genomics: leprosy - a degenerative disease of the genome. *Curr Biol*, 11(10):R381–3.

Young, D. B., Kaufmann, S. H., Hermans, P. W., and Thole, J. E. (1992). Mycobacterial protein antigens: a compilation. *Mol Microbiol*, 6(2):133–45.

Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T., de Bakker, P. I., Abecasis, G. R., Almgren, P., Andersen, G., Ardlie, K., Bostrom, K. B., Bergman, R. N., Bonnycastle, L. L., Borch-Johnsen, K., Burtt, N. P., Chen, H., Chines, P. S., Daly, M. J., Deodhar, P., Ding, C. J., Doney, A. S., Duren, W. L., Elliott, K. S., Erdos, M. R., Frayling, T. M., Freathy, R. M., Gianniny, L., Grallert, H., Grarup, N., Groves, C. J., Guiducci, C., Hansen, T., Herder, C., Hitman, G. A., Hughes, T. E., Isomaa, B., Jackson, A. U., Jorgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F. G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C. M., Lyssenko, V., Marvelle, A. F., Meisinger, C., Midthjell, K., Mohlke, K. L., Morken, M. A., Morris, A. D., Narisu, N., Nilsson, P., Owen, K. R., Palmer, C. N., Payne, F., Perry, J. R., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N. W., Rees, M., Roix, J. J., Sandbaek, A., Shields, B., Sjogren, M., Steinthorsdottir, V., Stringham, H. M., Swift, A. J., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N. J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R. M., Weedon, M. N., Willer, C. J., Illig, T., Hveem, K., Hu, F. B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N. J., Barroso, I., Hattersley,

A. T., Collins, F. S., Groop, L., McCarthy, M. I., Boehnke, M., and Altshuler, D. (2008). Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*, 40(5):638–645.

Zeggini, E., Weedon, M. N., Lindgren, C. M., Frayling, T. M., Elliott, K. S., Lango, H., Timpson, N. J., Perry, J. R., Rayner, N. W., Freathy, R. M., Barrett, J. C., Shields, B., Morris, A. P., Ellard, S., Groves, C. J., Harries, L. W., Marchini, J. L., Owen, K. R., Knight, B., Cardon, L. R., Walker, M., Hitman, G. A., Morris, A. D., Doney, A. S., McCarthy, M. I., and Hattersley, A. T. (2007). Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, 316(5829):1336–1341.

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):2609–2618.

Zhong, H., Yang, X., Kaplan, L. M., Molony, C., and Schadt, E. E. (2010). Integrating pathway analysis and genetics of gene expression for genome-wide association studies. *Am J Hum Genet*, 86(4):581–91.

Zhu, C. and Yu, J. (2009). Nonmetric multidimensional scaling corrects for pop-

ulation structure in association mapping with different sample types. *Genetics*, 182(3):875–88.

Zimmerman, M. R. (1979). Pulmonary and osseous tuberculosis in an Egyptian mummy. *Bull N Y Acad Med*, 55(6):604–608.