GENETICS OF TUBERCULOSIS RESISTANCE

By

RAFAL S. SOBOTA

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

May, 2015

Nashville, Tennessee

Approved:

Professor Scott M. Williams

Professor Dana C. Crawford

Professor Jonathan L. Haines

Professor Luc Van Kaer

Professor Bingshan Li

Professor Carl H. Johnson

Copyright © 2015 by Rafal Sebastian Sobota

All Rights Reserved

DEDICATION

To my mom.

To the patients in Dar es Salaam and Kampala who made this study possible.

ACKNOWLEDGEMENTS

I would like to thank the members of my thesis committee, Drs. Dana Crawford, Jonathan Haines, Bingshan Li, Luc Van Kaer, and the Chair, Dr. Carl Johnson. Their critiques, comments, and suggestions have been invaluable throughout the course of this project. I would also like to recognize Dr. Terry Dermody who is a mentor, an inspiration, and a role model.

I would like to acknowledge Nuri Kodaman for his intellectual contributions to this and my other projects, and Tulsi Roy for her unwavering support throughout the many challenges I encountered along the way. Thank you for making each departure from Nashville increasingly more difficult. I also would like to extend a special thank you to my parents, family, and friends, both in the US and in Poland, who supported me and my commitment to this project despite considerable logistical difficulties and too many months away from home.

And finally, I would like to recognize Dr. Scott Williams, who from the very day I started treated me as a colleague. His professionalism, moral rectitude, and commitment to the patients we encountered along the way has left an indelible mark in my growth as a person and a physician scientist. He has taught me much more than genetics.

My research was supported by Public Health Service award T32 GM07347 from the National Institute of General Medical Studies for the Vanderbilt Medical-Scientist Training Program, and by the NIH grant P20 GM103534. This work was also supported by the National Institutes of Health Office of the Director, Fogarty International Center, Office of AIDS Research, National Cancer Center, National Eye Institute, National Heart, Blood, and Lung Institute, National Institute of Dental and Craniofacial Research, National Institute On Drug Abuse, National Institute of Mental Health, National Institute of Allergy and Infectious Diseases, and National Institutes of Health Office of Women's Health and Research through the Fogarty

International Clinical Research Scholars and Fellows Program at Vanderbilt University (R24 TW007988) and the American Relief and Recovery Act.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Ab	Antibody
AFB	Acid fast bacilli
Ag	Antigen
Ag85	MTB antigen 85
AIDS	Acquired immunodeficiency syndrome
ASW	African ancestry in Southwest United States HapMap population
BCG	Bacillus Calmette-Guerin
CCL2	C-C motif ligand 2
CD4	Cluster of differentiation 4 co-receptor protein
CEU	Utah residents with northern and western European ancestry HapMap population
CFP	Culture filtered protein
CFU	Colony forming units
CHB	Han Chinese in Beijing, China HapMap population
CHD	Chinese in Metropolitan Denver, Colorado HapMap population
CHR	Chromosome
CNS	Central nervous system
CR1	Complement receptor 1
CXFT	MTB culture filtrate
ELISA	Enzyme linked immunosorbent assay
EMB	Ethambutol
ESAT-6	Early secretory antigenic target 6
GAS2	Growth arrest-specific protein 2

GIH	Gujarati Indians in Houston, Texas HapMap population
GLI2	GLI family zinc finger 2
GWAS	Genome-wide association study
HHC	Household Contacts Study
HIV	Human immunodeficiency virus
IFN-γ	Interferon gamma
IGRA	Interferon gamma release assay
iHS	Integrated haplotype score
IL9	Interleukin 9
IL12	Interleukin 12
IL12B	Interleukin 12 beta subunit
IL23	Interleukin 23
INH	Isoniazid
JPT	Japanese in Tokyo, Japan HapMap population
LAM	Lipoarabinomannan
LD	Linkage disequilibrium
LPA	Lymphocyte proliferation assay
LWK	Luhya in Webuye, Kenya HapMap population
MAF	Minor allele frequency
MDR	Multifactor dimensionality reduction
MDR-TB	Mutlidrug resistant tuberculosis
MHC	Major histocompatibility complex
MKK	Maasai in Kinyawa, Kenya HapMap population

MTB	*Mycobacterium tuberculosis*
MXL	Mexican ancestry in Los Angeles, California HapMap population
NRAMP	Natural resistance-associated macrophage protein
PAMP	Pathogen-associated molecular pattern
PBMC	Peripheral blood mononuclear cells
PHA	Phytohaemagglutinin
PLK2	Polo-like kinase 2
PPD	Purified protein derivative
PTX3	Pentraxin 3
PZA	Pyrazinamide
RAB6C	Ras-related protein Rab-6C
RIF	Rifampin
SLC6A3	Solute carrier family 6, member 3
SLC25A48	Solute carrier family 25, member 48
SNP	Single nucleotide polymorphism
TB	Tuberculosis disease
T_H1	T helper cell type 1
TIRAP	Toll-interleukin 1 receptor domain containing protein
TLR2	Toll-like receptor 2
TLR4	Toll-like receptor 4
TNF-α	Tumor necrosis factor alpha
TSI	Toscani in Italia HapMap population
TST	Tuberculin skin test

UBLCP1	Ubiquitin-like domain containing CTD phosphatase
VDR	Vitamin D receptor
WCL	Whole cell lysate
XDR-TB	Extensively drug-resistant tuberculosis
YRI	Yoruba in Ibadan, Nigeria HapMap population

CHAPTER I

OVERVIEW

Mycobacterium tuberculosis (MTB) infection and subsequent tuberculosis (TB) is the second-leading cause of mortality from a single infectious agent worldwide, after the human immunodeficiency virus (HIV)[1,2]. In 2013, 9 million new cases of clinical tuberculosis were diagnosed and 1.5 million deaths were attributed to the disease[2]. An estimated 1.1 million new cases and 360,000 of the deaths occurred in people co-infected with HIV[2]. The immunosuppression resulting from HIV infection increases the risk of progression to active disease following new exposure to MTB, or reactivation of latent MTB in patients with prior infection[3,4]. Sub-Saharan Africa is the location where most HIV/TB co-infection occurs, with 75% of all cases reporting coinfection[1,2]. The influence of host genetics on tuberculosis disease has been extensively studied, mostly in HIV-negative patients, and revealed that variation in pathways pertinent to macrophage and Type 1 helper T cell (T_H1) signaling, among others, modulate disease risk[5-11]. Generally, HIV seropositive status has been viewed as a confounder in such studies and it is either adjusted for or used as an exclusion criterion.

In the current project we present a novel hypothesis for studying resistance to either TB disease or MTB infection, using the immunosuppression of HIV-positive patients to identify an extreme phenotype. Namely, we posit that HIV-positive patients living in areas endemic for MTB who do not develop TB are resistant, and that this protection has a genetic component with an effect size large enough to permit using a smaller sample size than those seen in prior genome-wide association studies (GWAS) on TB[10,11]. The goal of the following chapters is to evaluate this hypothesis as it pertains to resistance to TB disease and MTB infection, in single variant and epistatic models.

Background information pertinent to this project is described in part A of Chapter II, including clinical presentation, diagnostics, treatments, a summary of the worldwide burden of tuberculosis as well as a description of prior genetic variants associating with TB. Previous study designs are discussed, as well as the approach we are taking to analyze this complex phenotype. Explicit hypotheses and specific aims are presented in part B of Chapter II.

A genome-wide association study of common genetic variants and with TB resistance in HIV-positive patients is described in Chapter III. Of note, I recruited patients on-site and isolated DNA in one of the cohorts, the DarDar vaccine trial extended follow up, in Dar es Salaam, Tanzania. I also isolated DNA from samples collected in two other studies, the DarDar Women's Nutrition Study from Tanzania and the Household Contacts Study from Kampala, Uganda. The samples for the Nutrition and Household Contacts Studies were made available to us by fellow investigators. In a study combining the cohorts, we found a common variant associating at the genome-wide significance threshold in the *IL12B* region. Linkage disequilibrium (LD) patterns in the region suggested that the region is conserved and integrated haplotype score analyses using sub-Saharan populations demonstrated that the LD block containing rs4921437 has undergone selection. The single nucleotide polymorphism (SNP) of interest is located in an area enriched for a histone acetylation mark often found in active regulatory elements, suggesting possible functionality and a genetic-epigenetic interaction at the site. Further studies of this interaction are warranted.

Chapter IV describes two approaches used to evaluate the genetics of MTB infection. In Part A, we used a genome-wide approach to identify variants associating with MTB infection. We discovered a novel association to MTB infection, variant rs877356 near *IL9*, a gene with a substantial role in airway inflammation and bronchial asthma. Observational studies have

reported a pattern of inverse incidence of asthma and tuberculosis, and we believe that *IL9*, a gene whose over expression plays a significant role in the pathogenesis of asthma, also prevents MTB infection by the same mechanism. In Part B, we validated and fine-mapped regions previously associated through genome-wide linkage analyses, *SLC6A3* +/- 0.5 mb[12], 2q14[13], 2q21-2q24[13], 5p13-5q22[13] and chromosome 11 p14.1[12]. We discovered an infection susceptibility locus rs10834029 on chromosome 11 near *GAS2*, a gene previously associated with the apoptotic pathway.

The immune response to MTB infection and resultant disease is complex[8,14-17]. The molecular signaling profile and extent tissue involvement change over the course of disease. Therefore, it is likely that no single factor alone can adequately explain risk of TB disease or MTB infection risk. In Chapter V we present a study in which we examined multi-locus relationships between previously associated candidate genes and TB disease in Part A, and MTB infection in Part B, using Multifactor Dimensionality Reduction (MDR) software[18,19]. We found an association of TB disease and a three-locus interaction of variants near *IL12B, TNF-α,* and *CR1*. The products of *IL12B* and *TNF-α* have some redundant and some concordant effects in the inflammatory response to MTB, while *CR1* plays a role in phagocytosis of MTB by macrophages, also a key step in immune response to MTB. In Part B, we found an association between MTB infection and a three-locus interaction between variants near *IL9, SLC6A3,* and *RAB6C*. Importantly, each three-locus interaction had consistent effects in the Ugandan and Tanzanian cohorts.

Since the majority of our Tanzanian participants were previously enrolled in a Phase III whole cell SRL 172 vaccine trial, we wanted to evaluate whether host genetic components affect protective outcomes. The vaccine confers protection against definite TB in HIV positive adults

who had a childhood BCG vaccination. Response to the vaccine was variable with SRL 172 showing a 39% efficacy; therefore it is of interest whether its effectiveness is modulated by polymorphisms in our either our candidate gene/loci list, or any other available variants[17]. However, as described in Chapter VI, we were unable to recruit enough definite or probable TB cases from the extended follow up cohort of this trial to have adequate power for this aim.

Chapter VII summarizes the conclusions of this project and proposes future directions.

CHAPTER II

INTRODUCTION

A. BACKGROUND

Clinical presentation of tuberculosis

Pathogenesis

Tuberculosis is an infectious disease caused by the inhalation of 1-3 droplet nuclei containing *Mycobacterium tuberculosis*, emitted by individuals with active respiratory disease via coughing, sneezing or spitting[2,20]. Differentiating TB infection from TB disease is essential, in that infection simply means the presence of *M. tuberculosis* organisms in a host, whether it leads to disease or not. In fact, 90-95% of people infected with *M. tuberculosis* will contain the initial infection without symptoms and develop a latent infection, while the remaining 5-10% will develop active TB disease soon after exposure (primary tuberculosis), with symptoms such as cough, fever, weight loss and night sweats[21]. The pattern of lung involvement in primary tuberculosis results in consolidation of the lower and middle lung lobes, along with hilar adenopathy, and very rarely cavitation. Latent infections can be activated (secondary tuberculosis) by immunosuppression caused by age, malnutrition, or HIV infection, among other factors. Cavitation occurs readily in secondary tuberculosis, and it is generally localized in upper lobe apices of the lungs[21]. TB disease is lethal in two thirds of the patients who do not receive proper treatment[2]. The causal organism, *M. tuberculosis*, is an aerobic acid fast bacterium that divides very slowly relative to other bacteria, once every 15-20 hours, which makes it very difficult to grow in culture and consequently diagnose. Like Gram positive bacteria, *M. tuberculosis* has an outer membrane lipid bilayer containing peptidoglycan; however the bilayer

is filled with other lipids (liparabinomannan) and the peptidoglycan is linked to arabinogalactan bound to mycolic acids that makes it very difficult to Gram stain[15]. The acid fast Ziehl-Neelsen stain is usually used instead of Gram staining when TB is a part of the differential diagnosis. The small shape and unique cell wall lipid profile make *M. tuberculosis* very resilient, even when exposed to desiccation or disinfectants[15].

Once the droplet nuclei containing *M. tuberculosis* move into the lungs of a naïve individual, the bacteria enter macrophages and grow in intracellular membrane bound vesicles (Figure 2-1). This allows *M. tuberculosis* to avoid degradation by the immune system because these vesicles do not bind to lysosomes. Since the vesicles are intracellular, the *M. tuberculosis* molecules are not exposed to cytosolic proteasomes and less efficient intravesicular proteases must degrade the bacteria instead. Therefore, in the first 3 weeks after infection in a non-sensitized individual, the bacteremia is largely unchecked by innate immunity and the mycobacteria can seed multiple sites and airspaces[21].

Roughly three weeks following the infection, the host starts to mount an adaptive response by inducing the differentiation of T cells. There are two main types of CD4 expressing T helper cells, T_H1 and T_H2[22]. T_H1 cells control infections by intracellular pathogens while T_H2 cells are responsible for extracellular pathogens and they induce B cells to produce antibodies[22]. Upon *M. tuberculosis* digestion and peptide presentation with major histocompatibility complex (MHC) class II molecules, Toll-like receptor 2 (TLR2) binding to dendritic cells stimulates an IL-12 response which in turn activates the differentiation of T_H1 CD4 T cells[21]. The mature T_H1 cells mount an interferon gamma (IFN-γ) response and activate phagolysosomes in *M. tuberculosis* containing macrophages, and induce the production of nitric oxide and reactive oxygen species[21]. These activated macrophages also use tumor necrosis factor (TNF) signaling in

a localized inflammatory response that leads to the formation of granuloma. The granuloma is comprised of activated macrophages that differentiate into epithelioid histiocytes, and may fuse forming giant cells. In the case of *M. tuberculosis* TB, the granuloma consists of a central core of macrophages, organized into progressively necrotic multinucleated giant cells, and surrounded by CD4 T cells[21]. The *M. tuberculosis* bacteria survive within the granulomas even in the presence of widespread necrosis[22]. If the granuloma is macroscopic in size, the central portion is yellowish/white upon dissection, and the term caseous necrosis is used[21]. Complete clearance of the bacteria is rarely achieved, and the lesions become fibrocalcific nodules known as Ghon foci (Ghon complexes if lymph node granulomas are involved) forming at the site of the infection. Viable *M. tuberculosis* organisms can survive there for many years until reactivation, thereby establishing a latent infection[21]. In the 5-10% of patients who progress to active TB disease, the immune system is unable to wall off the infection via the granulomas and the bacteria multiply, with significant caseation, cavitation and tissue destruction[21]. 80% of such disease is restricted to the lungs, with symptoms such as prolonged cough, hemoptysis and chest pain[23]. In rare cases, mostly in children and immunosuppressed patients, *M. tuberculosis* can enter systemic circulation by draining into systemic veins via lymph nodes or directly entering the pulmonary vein and cause disseminated disease usually localizing to pleura, bones, and joints (miliary tuberculosis) or the CNS (tuberculosis meningitis)[21].

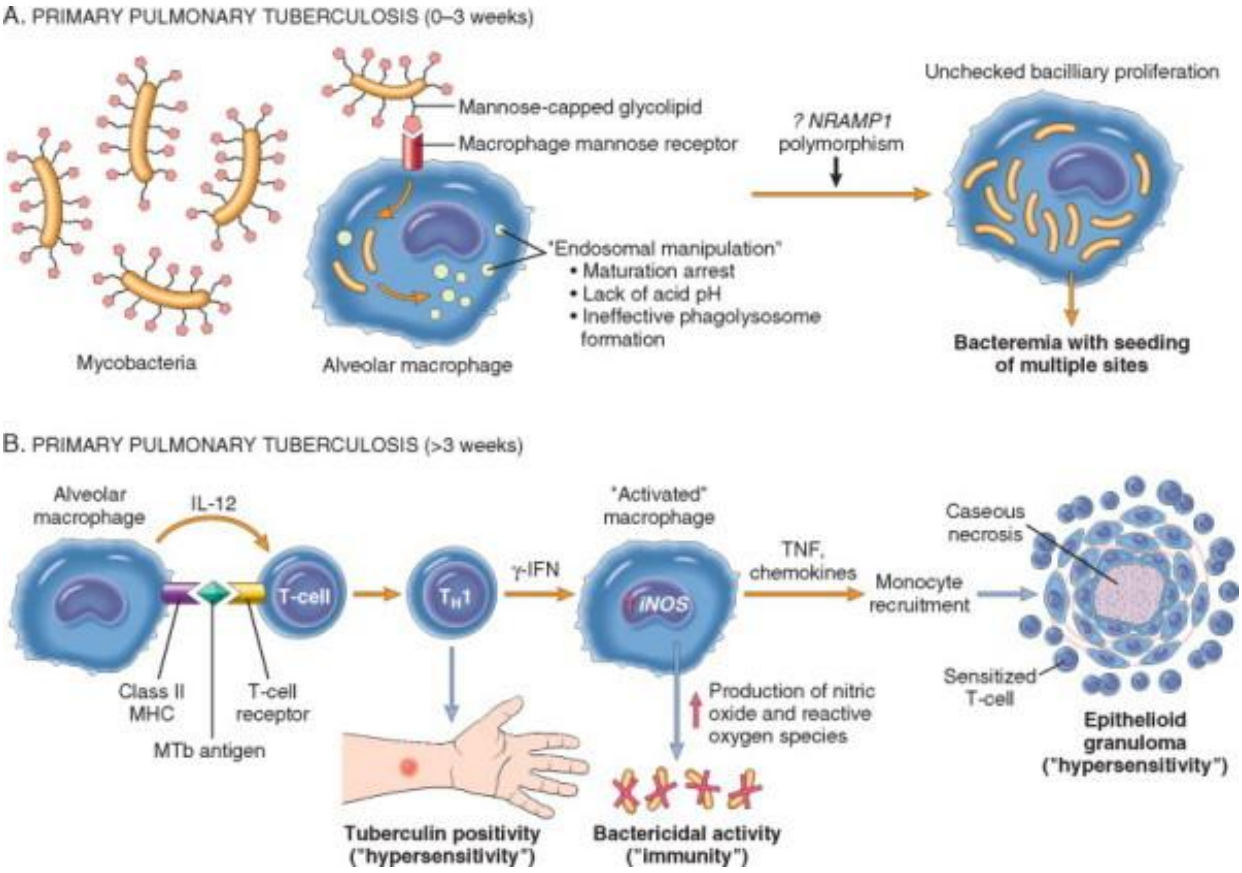


Figure 2-1. Progression of primary pulmonary tuberculosis upon inhalation of *Mycobacterium tuberculosis* containing droplet nuclei, prior (A) and following (B) adaptive immune response[21]; image from: http://dr.klinikbtp.com/tuberculosis-pathogenesis-and-immunity/

Diagnostics and treatment

The broad spectrum of possible symptoms for both pulmonary and extra-pulmonary tuberculosis makes it difficult to diagnose based solely on clinical presentation, as the disease can mimic many conditions such as other bacterial infections or a variety of malignancies[15]. The above mentioned problems with staining and culturing of *M. tuberculosis* compound this problem. Due to the predominant involvement of lower lobes, diagnosis of primary tuberculosis is difficult because it resembles primary progressive pneumonia[21]. On the other hand, secondary tuberculosis usually features the consolidation and cavitation of tissues found in apices of the upper lobes of the lungs, which can be detected with a chest X ray[21].

While most cases of primary TB are restricted to the lungs, in HIV infected individuals the disease causes extra-pulmonary disease more often, with such localizations being found in 53-63% of such patients[24,25]. The presentation is usually contingent on the $CD4^+$ count; patients with >300 cells/mm^3 will resemble secondary tuberculosis symptoms, while those <200 cells/mm^3 resemble progressive primary tuberculosis[21]. Primary prophylaxis of daily or twice weekly isoniazid for 9 months has been recommended for HIV-positive individuals with latent TB or recent TB exposure. This has been proven effective in preventing subsequent disease and it does not cause any adverse interactions with highly active anti-retroviral therapy[26].

The Mantoux tuberculin skin test (TST) is the most common technique used to ascertain *M. tuberculosis* exposure and subsequent response. The involvement of T-cell mediated immunity in combating *M. tuberculosis* infection allows for the use of delayed hypersensitivity to its antigens as a diagnostic[21]. A TST consists of an intradermal injection of a standard dose of *M. tuberculosis* purified protein derivatives (PPD) into the forearm of a patient; in those previously exposed to the bacteria, this will induce a palpable and visible induration that attains

its maximal size within 48-72 hours post injection[21]. TST results are, therefore, classified in the 48-72 hour period after the injection, and are interpreted in light of recent travel to endemic countries, HIV status and contact with TB patients, among others, and in the highest risk group an induration >5mm indicates a positive reaction.

The TST does not distinguish between *M. tuberculosis* infection and TB disease, and a childhood Bacillus Calmette-Guerin (BCG) vaccination or exposure to other mycobacteria can lead to false positive results[21]. False negative results can be a result of anergy caused by severe immunosuppression, as caused by HIV or malnutrition. In recent years alternatives to the long used TST have been developed. Two Interferon Gamma Release Assay (IGRA) tests, the QuantiFERON®-TB Gold In-Tube test (QFT-GIT) and T-SPOT®.*TB* test (T-Spot), are now commercially available and approved by the U.S. Food and Drug Administration. The Centers for Disease Control and Prevention recommend the use of IGRAs in certain populations, especially in settings where people do not return for TST readings or in those with previous BCG vaccination[27]. The QFT-GIT uses ELISA to detect the concentration of IFN-γ in response to the use of a mixture of TB-specific antigens, culture filtered protein (CFP)-10, TB7.7 and early secretory antigenic target (ESAT)-6 on whole blood[27,28]. Whole blood samples must be processed within 16 hours of the phlebotomy[27]. The specificity of the test is >99% regardless of BCG vaccination status. The QFT-GIT antigens are encoded by a region of the *M. tuberculosis* genome not present in the BCG strains and most other mycobacteria[27]. However, like the TST, QFT-GIT does not differentiate between latent TB infection and TB disease. The T-spot utilizes the enzyme-linked immunospot (ELISpot) assay, and requires peripheral blood mononuclear cells (PBMCs) that need to be processed 8-30 hours after the phlebotomy depending on the technique used. The test uses separate mixtures of synthetic peptides CFP-10 and ESAT-6, and

measures the proportion of IFN-γ producing T-cells to the total number of PBMCs on the plate[29]. The T-Spot has a sensitivity of ~95%; however, as with other described tests, it cannot discriminate between latent TB and TB disease[29,30].

Treatment modalities for TB disease vary depending on the antibiotic resistance of the infecting MTB strain and the presence of other comorbidities[31]. It is important to note that many regimens have been approved, and only a few are listed below. There are four first line drugs approved by the Food and Drug Administration for treatment of TB in otherwise healthy individuals in the USA: isoniazid (INH), rifampin (RIF), ethambutol (EMB) and pyrazinamide (PZA)[31]. The standard regimen calls for daily use of each of the first line medications for 8 weeks, followed by daily INH and RIF for 18 months[31]. An important caveat to treating patients coinfected with TB and HIV is the interaction between RIF and nonnucleoside reverse transcriptase inhibitors or certain protease inhibitors[31]. In those cases, rifabutin replaces RIF[31]. For latent TB infection, a daily nine-month regimen of INH is indicated. In cases of drug resistant TB, EMB is indicated along with a combination of fluoroquinilones (levofloxacin, moxyfloxacin) and aminoglycosides (amikacin, streptomycin, kanamycin), despite the substantial ototoxicity/nephrotoxicity associated with prolonged use of the latter class[31,32].

Epidemiology of tuberculosis

Global statistics

Mycobacterium tuberculosis infection and subsequent tuberculosis is the second-leading single infectious cause of mortality worldwide, after the human immunodeficiency virus (HIV)[1,2]. *M. tuberculosis* is the most common cause of human *Mycobacterial* disease and it alone is responsible for most of the 1.5 million deaths annually attributed to clinical

tuberculosis[2,15]. In 2013, 9 million new cases of clinical tuberculosis were diagnosed[2]. Over half of the new cases (56%) developed in South East Asia and Oceania, with 24% occurring in India alone[2]. Another quarter of the global TB incidence that year was reported in Africa[2].

The spread of drug resistant tuberculosis is an important aspect of TB epidemiology. While rates of TB have decreased since 1990, multi-drug resistant tuberculosis (MDR-TB) is rising. MDR-TB is defined as tuberculosis resistant to rifampicin and isoniazid, and it requires a longer treatment course using drugs with worse toxicity profiles[31]. 3.7% of new TB cases are MDR, as are 20% of previously treated cases[2]. This translates to 480,000 new cases of MDR-TB worldwide in 2013. Furthermore, 9% of MDR cases are resistant to other TB drugs, and are considered extensively drug resistant (XDR) TB cases. Most XDR cases are seen in Russia, China, and India[2].

An estimated 60% of incident cases and mortality from tuberculosis is observed in males[2]. In 2013, 5.7 million new cases of TB were diagnosed in males. The cause of this gender bias remains unexplained. Underreporting of cases in females due to social barriers in access to healthcare in the developing world has been suggested as a possible cause; however large prevalence studies suggest that biological differences between the sexes most likely account for this disparity[33].

TB/HIV coinfection

An estimated 360,000 of the 1.5 million TB deaths in 2013 occurred in people co-infected with HIV[2]. Of the 9 million new cases, 1.1 million developed in patients who were HIV-positive[2]. The immunosuppression resulting from HIV infection increases the risk of progression to active disease following new exposure to MTB, or reactivation of latent TB in patients with prior infection[3,4]. Sub-Saharan Africa is the location where most HIV/TB co-

infections occur, with 75% of all reported cases[1,2]. Tuberculosis in turn is the most common cause of death in people infected with HIV worldwide[34]. Of note, 50% of the HIV/TB positive patient deaths occurred in females; therefore, the gender bias in mortality is not observed in this patient population[2,33]. The gender distribution of the 1.1 million cases of tuberculosis in HIV positive patients was not reported.

SRL 172 vaccine

Considering the extent of morbidity and mortality caused by tuberculosis in HIV-positive patients, finding clinical solutions specific to that population is a public health priority[34]. Recently, a new means of preventing the development of tuberculosis in HIV positive patients has been developed. An inactivated whole cell SRL 172 vaccine was shown to be effective in preventing definite tuberculosis when administered as a booster for patients who received the Bacillus Calmette-Guérin vaccine as children in the DarDar vaccine trial cohort in Tanzania[17]. All diagnoses were obtained via a thorough evaluation comprised of a physical examination, a chest radiograph, three sputum samples for staining and culture, as well as an automated mycobacterial blood culture. Definite tuberculosis was defined as one or more of the following: at least one *M. tuberculosis* positive blood culture; sputum culture (with ≥10 colony-forming units, CFU); or at least 2 sputum cultures with CFU 1-9[17]. Other definite tuberculosis criteria involved at least 2 sputum smears with ≥2 acid fast bacilli (AFB) per 100 oil immersion fields; ≥1 positive culture or AFB smear and caseous necrosis from a sterile site other than blood[17]. The criteria for probable tuberculosis were: caseous necrosis on biopsy; a positive chest radiograph with either one AFB positive sputum film or one sputum culture with CFU 1-9; or clinical symptoms with either one AFB positive sputum film or one sputum culture with CFU 1-9. Other criteria included a positive chest X ray or other imaging, clinical signs and symptoms along with

response to TB therapy; as well as a positive AFB smear from a sterile site with clinical signs and symptoms[17]. The trial was randomized, double blind, and placebo-controlled. The five dose regimen was shown to provide 39% protection from definite tuberculosis in patients with a CD4 cell count greater than 200 cells/μl[17].

The original end point for the trial was the prevention of disseminated tuberculosis; however the number of cases in both the vaccine and placebo arm was insufficient to definitively prove effectiveness, even though the trend supported the original hypothesis that inactivated whole cell SRL 172 does indeed protect from developing disseminated symptoms[17]. The host genetics underlying the efficacy of the SRL 172 booster in this population were not part of the original study, and such an analysis would also be helpful in identifying the target population for its use as well as helping explain the mechanism of its action.

Genetic studies of tuberculosis

Genome-wide association studies

Two genome-wide association studies for tuberculosis susceptibility have been published to date[10,11]. A 2010 study discovered an association in a gene desert on chromosome 18q11.2, with a combined p value of 6.8×10^{-9} and an odds ratio of 1.19 of the G allele of rs4331426 in a combined Ghanaian, Gambian and Malawian cohort[10]. This cross sectional case-control study analyzed a total of 7,501 controls and 3,632 cases. Availability of 1000 Genomes Project data released in August of 2010 allowed the authors to impute SNPs into the Ghanaian cohort of the 2010 study and a genome-wide significant association for the A allele of variant rs2057178 was discovered, with a p value of 2.63×10^{-9} and an odds ratio of 0.77, that was replicated in the Gambia, Indonesia and Russia[11]. The variant is 46 kb downstream of *WT1,* which encodes a transcription factor known to activate the vitamin D receptor and suppress IL10 levels[35,36]. The

initial cohort in which genome-wide significance was attained was 5,636 controls and 2,127 cases, and the cohort including the replication analyses totaled 13,859 controls and 8,821 cases.

Overview of established candidate genes

Human genetic variants can impact susceptibility/resistance to *M. tuberculosis* infection/disease and if strong effect sizes are present, they could guide public health policy. Previous murine and human studies have described genetic variants that affect tuberculosis pathogenesis. The oldest studies of human genetic association with tuberculosis are in the major histocompatibility complex region, in particular the *HLA-DR2* region, the importance of which was first recognized in Northern India in 1983[15,37].

In 1993, polymorphisms in *NRAMP1* (also known as *SLC11A1*) were found to impact resistance/susceptibility to a spectrum of *Mycobacterial* infections in a murine model[5,38]. NRAMP1 is a transmembrane protein in the lysosomal and endosomal membranes that functions as a divalent ion pump removing them into the cytosol[15,21]. It was postulated that in doing so, it deprives bacteria such as *M. tuberculosis* of essential elements such as iron, thereby inhibiting growth[21]. Polymorphisms that impair proper function of *NRAMP1* inhibit the ability of an infected macrophage to clear the infection, and individuals carrying such mutations are more likely to develop TB disease. This murine finding has been recapitulated in human population studies from the Gambia[39], Japan[40], Cambodia[41], South Korea[42] and the United States[43].

The *SP110* gene product has also been shown to associate with tuberculosis susceptibility in mice, as it potentiates macrophage apoptosis *in vivo* and *in vitro.*[44] This finding was later confirmed in its human homologue in a Gambian population[45]. To date, genes from various other pathways have been associated with tuberculosis. Polymorphisms in cytokines/chemokines utilized in macrophage and T_H1 cell signaling such as *IFN-γ*[15], *CCL2*[46], *IL-8*[9], *IL-1B*[47], *IL-10*[48],

IL-4[49], and the *IL-4delta2*[50] splice variant, have been shown to affect *M. tuberculosis* infection outcomes in an array of populations. Variants in immunomodulatory receptors such as *PTX3*[51], *CR1*[52], *VDR*[51,53], *CD209*[51,54], *P2RX7*[55], also show modest, population specific effects. *CR1* is of particular interest for this study because it was shown to modulate *M. tuberculosis* infection in an HIV dependent manner. Homozygosity for a particular mutation associated with TB susceptibility in an HIV-negative population from Northern Malawi, but not in an HIV-positive population[52]. Pattern recognition molecules are another important category associated with response to *M. tuberculosis* infections. Complement activator *MBL2*[56] or *TLR2*[57,58] and *TLR4*[59], as well as their downstream effector *TIRAP*[60], are all essential in activating immune cell responses, and have variants that have been associated with tuberculosis outcomes.

Importantly, variants in these candidate genes did not associate with active tuberculosis below a genome-wide significant threshold in the GWAS studies by Thye et al[10,11]. The possible reasons for this observation are manifold. Some of the candidate genes have been discovered with linkage methods in individuals carrying rare mutations that cause a loss of function of a gene, as in the case of *IL12*[7]. Rare variants were removed in the quality control steps of the Thye et al studies; therefore, their effects could not be observed. Another possibility for why variants in candidate genes were not statistically significant at a multiple testing corrected threshold in the GWAS studies is that their effects might be population specific and, thereby, not relevant in modulating disease risk in Ghana or Gambia. Also of note, since the study from 2010 was not able to identify the significant imputed region reported in 2012, using a very similar patient population, it is possible that the array used simply was not dense enough to cover all relevant variants in or around the candidate genes.

The genetics of Tuberculin Skin Test reactivity have also been studied. Using a genome-wide linkage approach, Cobat et al. showed that *TST1* exclusively governs TST positivity in a binary fashion while *SLC6A3* controls the extent of reactivity to TST as a quantitative trait[12]. In a similar study of household contacts in Kampala, Uganda, Stein et al discovered associations with persistently negative TST scores in the 2q14, 2q21-2q24 and 5p13-5q22 regions[13]. In their review of ~650 thousand TST results of Navy recruits Rose et al. found that a negative TST rules out *M. tuberculosis* disease; therefore, it has been postulated that a gene variant that guarantees a TST response of zero is a marker of protection[12,61]. Unfortunately polymorphisms in genes that differentiate disease manifestation on a population scale have shown to have only modest effects, proving that the genetics of tuberculosis susceptibility and resistance are complex. To date, candidate gene approaches have been most effective in finding these variants[15], and even though recent GWAS have identified new loci, they carried very small effect sizes. While the overarching goal of finding a polygenic variant of predisposition/resistance remains elusive, such studies aid in understanding the pathogenesis of TB disease and the consequent immunological response.

B. HYPOTHESES AND SPECIFIC AIMS

To date, many studies of tuberculosis susceptibility have been performed, including GWAS, but the majority have identified variants of weak genetic effects[10,11,15]. **We hypothesize that HIV-positive patients who live in areas endemic for MTB but continue to remain disease free, and separately infection free, have genetic protection from doing so, and, moreover, they present an extreme phenotype that will allow the discovery of variants with stronger effect sizes.**

HIV positive patients can remain TB negative if they do not get exposed to MTB, if their immune system upon exposure is strong enough to prevent infection, if their immune system remains strong enough to prevent disease, or if they have *strong protective host genetic variants to prevent* either infection or disease. Since the patients in our studies all live in highly endemic regions, it is reasonable to assume that virtually all have been exposed to *M. tuberculosis.* Furthermore, the design of one of our cohorts, the Household Contacts Study, guarantees it. Studying HIV positive individuals who do not develop TB in this setting allows us to study those truly resistant to TB disease, and studying those who do not get infected allows insight into resistance to MTB infection. We propose a two-tiered approach. First, we will carry out single SNP associations in a genome wide analysis. Second, we will leverage prior knowledge that TB disease and infection are complex biological processes involving many pathways, and look to epistatic models in a candidate gene approach to evaluate the effect of higher order interactions on TB disease and MTB infection.

Since one of the cohorts used in this study was a follow up to a Phase 3 vaccine trial, we want to elaborate on the whole cell SRL 172 trial results by studying the host genetic components that govern protective outcomes. The vaccine confers protection against definite TB in HIV positive adults with a childhood BCG vaccine, and it is of interest to determine if its effectiveness is modulated by polymorphisms in our candidate gene/loci list to further understanding the variability of vaccine efficacy[17].

Specific Aim 1: Determine whether HIV positive subjects who do not show signs of TB disease are protected from active TB based on a genetic resistance.

1.1 Recruit HIV-positive subjects from the extended follow up of the DarDar vaccine trial. Obtain additional samples from HIV-positive patients the DarDar Nutrition Study, and the Uganda Household Contacts Study.

1.2 Genotype available samples using the Human Exome Beadchip platform.

1.3 Using a genome-wide approach, compare TB disease risk in cases versus controls, adjusting for relevant covariates and principal components, if necessary.

Separate logistic regression analyses will be performed for the all study participants from Uganda, for both studies from Tanzania, and for the combined cohort. Available covariates will be associated with the phenotype in univariate analyses, and those with significant association at the 0.05 level will be adjusted for in the logistic regression analyses. Since available samples are from major urban centers, Dar es Salaam, Tanzania, and Kampala, Uganda, principal components will be separately calculated for samples from each country and adjusted for in the logistic regression analyses. Haplotype analyses will be performed using UNPHASED[62]. Functional annotation will be carried out using data from the ENCODE Project.

Specific Aim 2: Determine whether HIV positive subjects who did not develop latent infection with *Mycobacterium tuberculosis* despite living in endemic areas are protected from MTB infection based on a genetic resistance.

2.1 Test the association of TB skin test results, both continuous and binary, <5mm versus ≥5mm, using a genome-wide approach.

2.2 Fine map associations from genome-wide linkage studies of TB skin test results, continuous and binary corresponding to the outcome in prior studies. Using a candidate region approach of *SLC6A3* encoding dopamine transporter 1, the *TST1* region on

chromosome 11p14, 2q14, 2q21-2q24 and 5p13-5q22 regions, evaluate association with TST results.

Carry out separate logistic regression analyses using the binary outcome, and linear regression analyses for the continuous skin test induration measurement for study participants from Uganda, for both studies from Tanzania, and for the combined cohort to correspond with the study that reported the association. Available covariates will be associated with the phenotype in univariate analyses, and those with significant association at the 0.05 level will be adjusted for in the logistic regression analyses. Since available samples are from major urban centers, Dar es Salaam, Tanzania, and Kampala, Uganda, principal components will be calculated and adjusted for in the logistic regression analyses. Use TST results at time of enrollment to limit possible immune anergy caused by immunosuppression in patients with low CD4 counts. Assess IFN-γ response to verify whether immune anergy occurred, and if present, use anergy as an exclusion criterion.

Specific Aim 3: Evaluate epistatic effects on TB disease and MTB infection using a candidate gene approach

3.1 Use multifactor dimensionality reduction software to evaluate epistasis between genes previously associated with TB disease and associated in our current single SNP study (Chapter III)

3.2 Use multifactor dimensionality reduction software to evaluate epistasis between genes previously associated with MTB infection and associated in our current single SNP study (Chapter IV)

The candidate gene list for TB disease is comprised of the following: *HLA-DR2*[15,37], *SLC11A1*[5,38], *SP110*[44,45], *IFN-γ*[15], *IFNGR1*[63], *IL12B*[7,8], *IL12BR1*[64], *UBE3A*[65], *TNF-α*[66], *CCL2*[46],

IL-8[9], *IL-1B*[47], *IL-10*[48,66], *IL-4*[49], *IL-4delta2*[50], *PTX3*[51], *CR1*[52], *VDR*[51,53], *CD209*[51,54], *P2RX7*[55], *CR1*[52], *MBL2*[56], *TLR2*[57,58], *TLR4*[59], and *TIRAP*[60]. The candidate gene list for MTB infection includes *SLC6A3*[12,61], the *TST1*[12,61] region on chromosome 11p14, and the 2q14, 2q21-2q24 and 5p13-5q22 regions[13]. Analyses will be performed for the all study participants from Uganda, for both studies from Tanzania, and for the combined cohort.

Specific Aim 4: Determine the genetic factors that govern response to a multiple-dose series of an inactivated SRL 172 whole cell vaccine in its prevention of HIV-associated tuberculosis in patients with a prior BCG vaccination.

4.1 Perform logistic regression analyses using subsequent TB disease status in the vaccine arm of the trial cohort as the outcome using a candidate gene approach

Available covariates from the vaccine arm of the study will be associated with TB disease in univariate analyses, and those with significant association at the 0.05 level will be adjusted for in the logistic regression analyses.

Identification of genes that affect risk of tuberculosis disease and MTB infection has many potential clinical implications. Candidate genes with large effect sizes provide targets for the development of therapeutics and diagnostics, or if relevant drugs already exist, assessing their efficacy for treatment of tuberculosis. SNPs with large effect sizes can also be used as genetic markers for developing disease. Moreover, discovery of novel genes associating with TB disease or MTB infection can further our understanding of the pathophysiology and resultant immune response, as well as provide novel targets for functional follow up studies in cell lines and model organisms.

CHAPTER III

GENOME-WIDE ASSOCIATION STUDY IDENTIFIES TUBERCULOSIS RESISTANCE LOCUS NEAR *IL12B* IN IMMUNOSUPPRESSED PATIENTS FROM TANZANIA AND UGANDA

Introduction

Mycobacterium tuberculosis (MTB) infection and subsequent tuberculosis (TB) is the second-leading infectious cause of mortality worldwide, after the human immunodeficiency virus (HIV)[1,2]. In 2013, 9 million new cases of clinical tuberculosis were diagnosed and 1.5 million deaths were attributed to the disease[2]. An estimated 360,000 deaths occurred in people co-infected with HIV[2]. The immunosuppression resulting from HIV infection increases the risk of progression to active disease following exposure to MTB, or reactivation of latent TB in patients with prior infection[3,4]. Sub-Saharan Africa harbors the majority of HIV/TB co-infection, with 75% of all reported cases[1,2]. The influence of host genetics on tuberculosis disease has been extensively studied, mostly in HIV-negative patients, and revealed that variation in pathways pertinent to macrophage and T_H1 signaling, among others, modulate disease risk[5-11]. In these, HIV seropositive was viewed as a confounder and was either adjusted for or used as an exclusion criterion.

In the current study, we hypothesize that HIV-positive patients living in areas endemic for MTB who do not develop TB can be defined as highly resistant. We posit that these individuals carry significant genetic protection and will represent a more phenotypically homogenous group than those studied in prior genome-wide association studies (GWAS) on TB[10,11]. If true, this will increase the measured effect sizes, permitting a much smaller sample size. To test this hypothesis we genotyped 639 HIV-positive individuals from East Africa using

the Human Exome Beadchip and had data available from the Illumina HumanOmni5-Quad BeadChip for 65 others. After quality control, data for 175,906 variants were available for 267 cases of TB and 314 controls followed throughout recently completed prospective cohorts of TB, the DarDar Vaccine Trial and Nutrition Study both carried out in Dar es Salaam, Tanzania, and the Household Contacts study in Kampala, Uganda[17,67,68].

Methods

Study populations

Tanzania

We recruited patients from the extended follow-up cohort of the DarDar vaccine trial and used previously collected samples from the DarDar Nutrition Study in Dar es Salaam, Tanzania. Both cohorts have been previously described[17,67]. Briefly, the DarDar trial was a randomized double blind phase III trial of an inactivated whole cell mycobacterial vaccine (SRL 172). Subject enrollment occurred between 2001 and 2005, and the study was concluded in 2008. All enrolled patients were HIV-positive adults (>18 years old) with a CD4 count >200 cells/μl, had to have a Bacille Calmette-Guérin childhood vaccination scar, and were TB-negative at the time of enrollment. The final study population consisted of 1007 patients in the placebo arm, and 1006 receiving the vaccine. Participants were routinely followed up every 3 months and evaluated for active TB through a physical examination, a chest radiograph, sputum samples for culture and AFB stain, and phlebotomy for an automated mycobacterial blood culture. At the conclusion of the study, a cohort of 800 participants in both the placebo and vaccine arm was selected for extended follow-up. These patients have been evaluated for active TB once a year since the conclusion of the study. We recruited 304 of the extended follow-up participants

between September and December, 2013, during their routine visits. 36 of the patients had been diagnosed with definite or probable tuberculosis since the onset of the trial. Diagnostic criteria described in von Reyn et al.[17] Patients who did not develop tuberculosis during the trial, but stated that they had previous active TB were excluded from the study, as those diagnoses could not be confirmed. The DarDar Nutrition Study was a randomized, controlled trial assessing the effectiveness of a protein-calorie supplement (PCS) to standard tuberculosis and HIV treatment in women. All enrolled women were HIV-positive adults (>18 years old) and had newly diagnosed active TB. 150 participants were randomized to either a PCS arm or a multivitamin control. WHO recommended treatment protocols were followed through the Tanzanian Ministry of Health National Tuberculosis and Leprosy Program. Subjects were followed monthly until the completion of the study in 2014. Samples from 85 participants of the DarDar Nutrition Study were used in this analysis.

Uganda

We used samples from participants of the Household Contact Study (HHC), conducted in Kampala, Uganda from 2002 to 2014. This cohort has been previously described[14,68]. Briefly, the HHC combines a case-control and a family based design to analyze the genetic epidemiology of tuberculosis. Patients diagnosed with new active tuberculosis were referred to the study through the Uganda National Tuberculosis and Leprosy Programme, and those who consented were enrolled as index cases. Relatives and unrelated individuals living within the same household were subsequently enrolled and evaluated for active TB, latent TB, and HIV. Both index and incident cases of active TB were given the recommended therapy[69]. Importantly, selecting cases and controls from the HHC guarantees exposure of the controls to the index case during the follow up[68]. We analyzed samples from 263 HIV-positive adult patients (>18 years old). Patients

who did not develop tuberculosis during the course of the study, but stated that they had previous active TB were excluded from the study, as those diagnoses could not be confirmed.

DNA isolation and genotyping

For participants from the DarDar vaccine trial, 5ml of whole blood was drawn into EDTA coated tubes (BD Biosciences) and immediately stored at 4°C in Dar es Salaam. DNA was extracted the day of the phlebotomy using the Gentra Puregene Blood kit (QIAGEN) in accordance with the manufacturer's recommendations. For participants of the DarDar Nutrition and the Household Contacts Study, buffy coats were isolated on site and shipped to Dartmouth College for DNA extraction. DNA was isolated from buffy coats using the QIAamp DNA Blood Mini Kit (QIAGEN). All DNA samples were stored at -80°C prior to genotyping.

Samples from the DarDar vaccine trial (n=304), the DarDar Nutrition Study (n=85) and the Household Contact Study (n=263) were submitted for genotyping. DNA quality was evaluated with the 260/280 ratio using a NanoDrop 2000 spectrophotometer (Thermo Scientific) and an Electrophoresis Quality Score. Following quality control, a total of 639 samples were genotyped using the Human Core Exome Beadchip (542,585 SNPs) at the Hussman Institute for Human Genetics, Miami, Florida. SNPs with a genotyping call rate > 0.95 and a Hardy-Weinberg equilibrium p value > 1*10E-4 were retained. All remaining participants had a per individual genotyping call rate > 0.95. A sex check was performed, and cryptic relatedness was evaluated in PLINK(v1.07)[70,71]. One individual was randomly removed in pairs of related study participants (pi-hat > 0.20). The final study population genotyped on the Exome Beadchip included 278 participants from the DarDar vaccine trial extended follow up, 65 participants from the DarDar Nutrition study, and 213 participants from the Household Contact Study.

Samples from 64 additional HIV-positive Ugandan HHC participants were previously genotyped using an Illumina HumanOmni5-Quad BeadChip (4.8 million SNPs) and made available for this study. DNA was extracted from buffy coats using the QIAamp mini DNA kit (Qiagen), and quantified using Nanodrop and Qubit. Genotyping and DNA quality checking were done at the Genomics Core at Case Western Reserve University. Further quality control steps for samples included checking for sex mismatch errors, relationship errors (within the larger dataset of 483 samples), consistency of blind duplicates, call rate (> 95%), 10th percentile GenCall score (> 0.42), visual inspection of BAF plot of samples with < 98% call rate, and unusually high autosomal heterozygosity. Individuals <18 years old and those related to the participants genotyped with the Exome Beadchip (pi-hat > 0.20) were excluded. 25 of the additional 64 individuals remained after applying these exclusion criteria.

Statistical Analyses

Logistic regression was used to test the association between single SNPs in an additive model and active tuberculosis case/control status using PLINK(v1.07)[70,71]. Results from dominant and recessive models are presented in the Appendix (Appendix Tables 3-1, 3-2, 3-3). *A priori* power analyses were carried out using Quanto[72] and revealed that to discover an association with an odds ratio of 2.0 below two-sided p value threshold of 0.05, we needed to set the minor allele frequency (MAF) at > 0.2 for the single SNP association analyses (Appendix Table 3-4). Summary statistics for available covariates were calculated in STATA(v11.2)[73]. Covariates significant in univariate logistic regression analyses with case/control status were included in the final models. The Tanzanian cohorts were recruited in Dar es Salaam, and the Ugandan cohort was recruited in Kampala, both large urban centers, making admixture a likely confounder in this study. Importantly, vaccine status did not associate with active tuberculosis in

the DarDar extended follow-up cohort (p 0.107); therefore it was not adjusted for. To adjust for population structure within each cohort, principal components were calculated using SNPs with $r^2 < 0.1$, MAF > 0.2, and a genotype call rate > 0.95 using the SNPRelate package in R[74,75]. Self-reported tribal identity data was available for the patients from Uganda. The first, sixth and seventh principal components were significant in predicting membership from the predominant tribe, Muganda (63% of the participants), versus all others; therefore all analyses were adjusted for a standard of 10 principal components (Appendix Tables 3-5, 3-6). Manhattan and qq plots were generated using the qqman package in R[76]. Locus zoom was used to plot the region of the SNP with the strongest association[77]. SNPs in the region of interest (rs4921437 position +/- 1 megabase) were imputed with IMPUTE2 (v2.3.1), using one phased reference panel from all populations of the 1000 Genomes project[78-80].

Unimputed SNPs with an MAF > 0.05 were included in the haplotype analyses of the region of interest. Haplotype plots were generated using Haploview[81]. Haplotype association analyses were performed using UNPHASED(v3.1.7)[62], adjusting for the same covariates as in the single SNP association analyses above. We studied all pairwise haplotypes for 11 available SNPs spanning the 2 genes in the region of interest in each cohort, and in the combined cohort. We then performed a 3-way haplotype analysis for the 2 haplotypes with the most significant associations in pairwise analyses. For each cohort, haplotypes from cases only, controls only, and the entire dataset were compared to corresponding haplotypes in all available Phase 3 samples of the HapMap project[82,83], using a chi squared test in STATA(v11.2)[73].

Immune Assays

All patients from the Nutrition study were cases of tuberculosis, and the design of the Household Contacts Study in Uganda guaranteed exposure to MTB. To address confounding of

our results by possible lack of exposure in the extended follow up of the DarDar vaccine trial, we leveraged available interferon-gamma release assay (IGRA) data. Immune response to *Mycobacteria* was assessed with an interferon gamma (IFN-γ) enzyme linked immunosorbent assay (ELISA), a tritiated thymidine lymphocyte proliferation assay (LPA) and ELISA for antibodies to the MTB glycolipid lipoarabinomannan (LAM). The assays used in this study have been previously described[84]. Briefly, IFN-γ and LPA assays used four different antigens: *Mycobacterium vaccae* sonicate (2 mcg/ml), MTB Antigen 85 (Ag85; 1 mcg/ml), MTB early secretory antigenic target 6 (ESAT-6; 2 mcg/ml), and MTB whole cell lysate (WCL; 1 mcg/ml)[84]. Media alone served as a negative control and phytohemagglutinin (PHA, 2.5 mcg/mL; Sigma, St. Louis, MO) was used as a positive control[84].

Selection Analysis

Phase was inferred using the Beagle software package[85]. Unrelated individuals from two HapMap3 release 2 populations[83], the Yoruba from Ibadan, Nigeria (YRI) and Luhya from Webuye, Kenya (LWK), 25 males and 25 females from each, were used to generate a fine-scale recombination map with LDhat(v2.1)[86]. The ancestral allele for each SNP included in this analysis was established using genome-wide sequences of non-human primates, the chimpanzee, orangutan, and rhesus macaque, downloaded from the UCSC Genome Browser website[87]. Approximately 5% of available SNPs could not be assigned an unambiguous ancestral state, and were removed prior to selection analysis, as were SNPs with MAF <0.05[88]. Selection was assessed using the integrated haplotype score test (iHS)[88]. iHS scores were standardized to a mean of zero and a unit variance with respect to SNPs with similar derived allele frequencies[88]. iHS values in the upper 0.1% of the distribution of absolute values were considered top candidates for having undergone recent selection.

In this analysis, we included individuals from a previous selection study of six populations in Cameroon (n = 125)[89]. The Baka, Bakola, and Bedzan are Niger-Kordofanian Bantu-speaking hunting and gathering populations designated here as Western Pygmy, and the Ngumba, Southern Tikar, and Lemande are neighboring Niger-Kordofanian Bantu-speaking agricultural populations, designated as Niger-Congo West. Data was also available for an additional 12 Yoruba individuals, living in Nigeria, and for 18 Datog living in Tanzania. Following phlebotomy, we isolated white blood cells using a modified salting out procedure[90], and DNA was extracted with a Purgene DNA extraction kit (Gentra Systems Inc., Minneapolis, MN)[89].

The samples were genotyped using the Illumina Human 1M-Duo BeadChip[89]. We evaluated cryptic relatedness, randomly removing individuals with pi-hat>0.25, and SNP with call rates less than 95% were removed. All quality control measures were done in PLINK(v1.07)[70,71]. Admixture was evaluated using Principal Component Analysis in R[74], and STRUCTURE, as previously described[89].

Functional Annotation

The ENCODE Project[91] was accessed via the UCSC Genome Browser[87] and Histone modification data provided by the Bernstein lab at the Broad Institute was used for functional annotation.

Ethics

Informed consent was obtained directly from all patients in the DarDar follow-up cohort and the Women's Nutrition Study at the Dar es Salaam clinic. This study was approved by the research ethics committee at the Muhimbili University of Health and Allied Sciences and the

Committee for the Protection of Human Subjects at Dartmouth College and the Dartmouth-Hitchcock Medical Center. Informed consent was obtained directly from the index case and household members in the HHC study. This study was approved by institutional review boards at the Uganda Council for Science and Technology and the University Hospitals of Cleveland.

Appropriate IRB approval for the selection cohorts used in this project was obtained from both the University of Maryland and the University of Pennsylvania. Prior to sample collection, informed consent was obtained from all research participants, and permits were received from the Ministry of Health and National Committee of Ethics in Cameroon, Nigeria and Tanzania.

Results and Discussion

Gender was significantly associated with active tuberculosis in the Uganda cohort (OR for males 2.06, 95% CI = 1.15-3.69, p value = 0.015; Table 3-1C), and when the two cohorts were combined (OR 1.78, 95% CI = 1.23-2.57, p value = 0.002; Table 3-1D). Gender was not significant in the DarDar vaccine trial (p value = 0.858; Table 3-1A); however since we had to add samples from the DarDar Women's Nutrition Study (Table 3-1B) to increase the number of cases and consequently statistical power, we decided to adjust the combined Tanzania cohort for gender as well. Age was not a significant covariate in either of the cohorts individually, or combined (Table 3-1A-D); therefore it was not adjusted for. The studies were conducted in major urban centers; therefore population stratification was a concern. Following adjustment for principal components calculated separately for samples originating in each country, the genomic inflation factor (λ) was 1.023 for samples from Tanzania, 1.056 for Uganda, and 1.029 for analysis of the cohorts combined.

In logistic regression analyses, we observed an association between common variant rs4921437 and active TB in Uganda (OR = 0.28, 95% CI = 0.17-0.47, $p = 1.18 \times 10^{-6}$; Appendix Table 3-7A, Appendix Figure 3-5) and Tanzania (OR = 0.48, 95% CI = 0.30-0.79, $p = 3.84 \times 10^{-3}$; Appendix Table 3-7B, Appendix Figure 3-6). When we analyzed the cohorts combined, rs4921437 associated with TB below the genome-wide significance threshold (OR = 0.37, 95% CI = 0.27-0.53, $p = 2.11 \times 10^{-8}$; Table 3-2, Appendix Figures 3-3 through 3-6).

Table 3-1. Summary statistics of patients with active tuberculosis, cases, and study participants who did not develop tuberculosis during the duration of follow up, controls, from A) the DarDar vaccine trial in Tanzania, B) the DarDar Nutrition Study in Tanzania, C) the Household Contacts Study in Uganda, and D) the combined cohorts

A)

		Cases	Controls	Odds Ratio	95% Confidence Interval	p value
n^		31 (11.3)	243 (88.7)			
Age* (years)		35.76 ± 7.97	34.23 ± 8.02	1.02	(0.98, 1.07)	0.298
Gender^	Male	6 (11.5)	46 (88.5)			
	Female	25 (11.3)	197 (88.7)	0.92	(0.36, 2.35)	0.858

B)

		Cases
n		69
Age* (years)		39.01 ± 8.51
Gender	Male	0
	Female	69

C)

		Cases	Controls	Odds Ratio	95% Confidence Interval	p value
n^		167 (70.2)	71 (29.8)			
Age* (years)		32.38 ± 7.92	31.23 ± 7.58	1.02	(0.98, 1.06)	0.296
Gender^	Male	83 (78.3)	23 (21.7)			
	Female	84 (63.6)	48 (36.4)	2.06	(1.15, 3.69)	0.015

D)

		Cases	Controls	Odds Ratio	95% Confidence Interval	p value
n^		267 (46.0)	314 (54.0)			
Age* (years)		34.46 ± 8.53	33.55 ± 8.01	1.01	(0.99, 1.03)	0.186
Gender^	Male	89 (56.3)	69 (43.7)			
	Female	178 (42.1)	245 (57.9)	1.78	(1.23, 2.57)	0.002

*mean ± standard deviation; ^ n (% of row)

Table 3-2. Single nucleotide polymorphisms associating with active tuberculosis below a $1x10^{-5}$ p value threshold in the combined cohort adjusted for principal components, sex, and cohort of origin

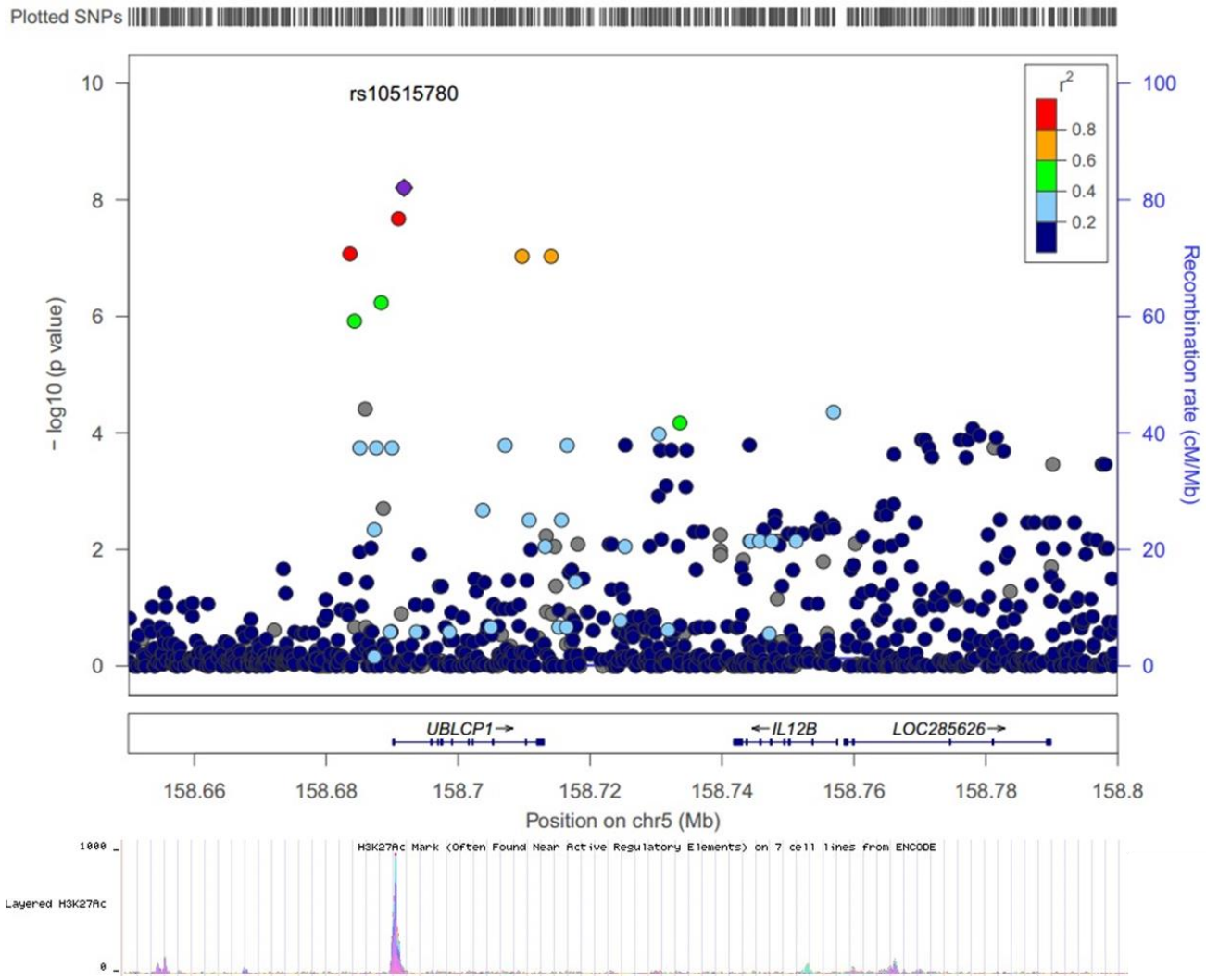
SNP	CHR	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value	Nearest gene
rs4921437	5	T	0.21	0.37	(0.27, 0.53)	2.11E-08	*UBLCP1/IL12B**
rs8028149	15	C	0.42	0.51	(0.39, 0.67)	1.96E-06	*VPS13C*
rs1616723	6	G	0.20	0.44	(0.31, 0.62)	4.13E-06	*GLO1*
rs955263	4	T	0.34	1.85	(1.39, 2.45)	2.26E-05	*SORBS2*
rs12636260	3	T	0.26	1.98	(1.44, 2.71)	2.26E-05	*ZPLD1*
rs4768760	12	C	0.41	0.56	(0.43, 0.73)	2.57E-05	*SLC38A4*
rs844669	7	G	0.30	0.52	(0.38, 0.71)	2.68E-05	*CALN1*
rs4236914	8	T	0.24	1.91	(1.41, 2.59)	2.84E-05	*SFRP1**
rs4860106	4	G	0.46	0.58	(0.45, 0.75)	3.03E-05	*LPHN3*
rs1482868	6	T	0.28	0.54	(0.41, 0.73)	4.29E-05	*F13A1*
rs2346943	16	G	0.33	1.79	(1.35, 2.37)	4.48E-05	*RBFOX1*

*previously associated with tuberculosis disease

In order to evaluate SNPs in the region not included on the genotyping array, we used IMPUTE2 (v2.3.1)[79,80] software to impute SNPs within 1 megabase of rs4921437. rs10515780, in high linkage disequilibrium (LD) with rs4921437 (D' = 1 in both cohorts, r^2 = 0.84 in Tanzania, 0.89 in Uganda) and 842 bases away, is the variant with the most significant association to TB in the region (combined cohort OR = 0.34, 95% CI = 0.24-0.49, p = 6.21×10^{-9}; Figure 3-1, Appendix Table 3-8, 3-9). Both rs4921437 and rs10515780 map to an intron of ubiquitin-like domain containing C-terminal domain phosphatase 1, *UBLCP1*. Variants in this gene have been previously associated with carotid stenosis[92], psoriasis[93], and platelet aggregation[94]. Importantly, rs4921437 is 51,240 base pairs away from the 3' UTR of *IL12B*, a gene extensively associated with TB[8,16,64,95,96] and leprosy tuberculosis[97]. *IL12B* variants also have been associated with carotid stenosis[98] and psoriasis[99,100], suggesting possible linkage disequilibrium between SNPs in the two genes. *IL12B* encodes for p40, one of two subunits of IL12[101] and IL23[102].

IL12 is a cytokine secreted by phagocytes and dendritic cells that causes differentiation of naïve T cells into T_H1 cells, and stimulates interferon-gamma production from T cells and natural killer cells[103,104]. In experimental models, IL12 restricted MTB proliferation throughout the course of infection as a necessary factor in granuloma formation and antigen-specific delayed type hypersensitivity[105-107]. Previous studies demonstrated that IL12 deficiency in humans was associated with predisposition to infections with *Mycobacteria*[7,108,109]. IL23 contributes to the differentiation of T helper 17 cells (T_H17) responsible for the production of IL17. IL17 signaling is essential in control of MTB infection, as it is necessary for targeting of neutrophils to the site of infection and mice deficient for this interleukin are unable to control highly virulent MTB strains[110-112].

Figure 3-1. Locus zoom of imputed data using the combined cohort with a 50 kb region around rs4921437, along with the corresponding UCSC genome browser window displaying the layered H3K27Ac histone mark; imputed SNP with the most significant association in a purple



The Encyclopedia of DNA Elements (ENCODE) was used for functional annotation of rs4921437, and it revealed that the SNP is located in a region of substantial regulatory potential, an H3K27Ac mark (Figure 3-1, bottom panel)[87,91]. This type of histone modification has been shown to act in combination with H3K4Me1 and differentiate active enhancer elements from poised/inactive ones[113,114]. Such chromatin modifications are associated with higher regulatory responses at enhancer transcription factor binding sites[113], possibly potentiating the effect of DNA polymorphisms in the region. Enhancers have highly cell-type-specific histone modification patterns, and cause cell-type-specific gene expression profiles[113]. It is therefore noteworthy that two of the cell lines in which the H3K27Ac mark was found have implications in immunity, the GM12878 and K562, from B-lymphocytes and chronic myelogenous leukemia, respectively[91].

We studied the LD pattern in the *UBCLP1*/*IL12B* region in the unimputed dataset and found blocks of LD between rs4921437 and variants in *IL12B* in the Ugandan and Tanzanian cohorts, as well as all African origin data from Phase 3 of the HapMap Project[83] (Figure 3-2, Appendix Figure 3-7). Association analyses between haplotypes in the region and TB were performed using UNPHASED(v3.1.7)[62]. A three variant haplotype of rs4921437, rs4921468 (intergenic), and rs3213094 (intronic in *IL12B*) had a more significant association than rs4921437 alone in the Ugandan and Tanzanian cohorts ($p = 1.98 \times 10^{-11}$ and 4.56×10^{-5}, respectively), and the combined cohort (p value = 4.56×10^{-15}) (Table 3-3, Appendix Table 3-10). When we conditioned the analysis on the main effect variant, rs4921437, the combined cohort haplotype was still significant at the multiple testing corrected level (p value = 1.19×10^{-7}). A chi-squared analysis testing the difference between the frequencies of the three variant haplotype in the Ugandan and Tanzanian cohorts was not significant (Appendix Table 3-11). No

significant differences were discovered when comparing the three variant haplotypes between the study cohorts and founders of the Luhya in Webuye, Kenya and African ancestry in Southwest USA populations from Phase 3 of HapMap[78] (Appendix Table 3-11). Comparisons of the frequencies of this three variant haplotype observed in Uganda and Tanzania was significantly different from all other Phase 3 HapMap populations (Appendix Tables 3-11, 3-12). The conservation of the haplotype frequencies in several African populations was unexpected given the diverse history of these populations.

We also noted that the LD patterns of the entire *IL12B* region, not only the three variant haplotype, were similar between the Ugandan and Tanzanian cohorts. Considering previously noted patterns of shorter LD ranges in African populations[115], this suggested possible selection in the region (Figure 3-2, Appendix Figures 3-7 through 3-9). T, the ancestral, minor allele of rs4912437 associated with the TB resistance effect in this analysis, has a minor allele frequency of ~0.21 in our study cohorts and an MAF between 0.11 and 0.36 in the African, European and South Asian origin populations from Phase 3 of the HapMap Project[83] and the 1000 Genomes Project[78] (Appendix Figure 3-10, Appendix Tables 3-13, 3-14). The minor allele frequency reaches as high as 0.454 in the Biaka Pygmy and and 0.461 Mbuti Pygmy populations from the Human Genome Diversity Project[116,117]. Of note, the variant has nearly disappeared in East Asian populations (MAF < 0.025 in all available populations), suggesting either a loss due to an out of Africa bottleneck event or its maintenance due to balancing selection in Europe, South Asia and Africa.

Table 3-3. Association of the rs4921437, rs4921468 and rs3213094 haplotype in the IL12B region using an additive model in A) the combined cohort, B) Tanzanian and C) Ugandan cohort

A)

Haplotype	Case	Control	Case frequency	Control frequency
C-A-T	5	6	0.0098	0.0098
C-A-C	96	102	0.18	0.16
C-G-T	242	223	0.45	0.36
C-G-C	115	130	0.21	0.21
T-G-T	5	7	0.0071	0.012
T-A-C	6	12	0.0094	0.02
T-G-C	65	147	0.13	0.23
Likelihood ratio chisq = 79.49 df = 6 p-value = 4.56E-15*				

B)

Haplotype	Case	Control	Case frequency	Control frequency
C-A-T	5	6	0.024	0.013
C-A-C	36	81	0.18	0.17
C-G-T	92	176	0.46	0.36
C-G-C	40	107	0.20	0.22
T-G-T	1	4	0.005	0.0087
T-A-C	1	8	0.005	0.017
T-G-C	25	103	0.12	0.21
Likelihood ratio chisq = 31.75 df = 6 p-value = 1.82E-05^				

C)

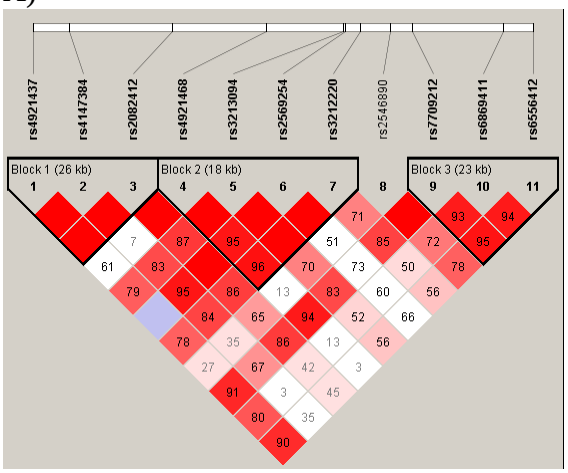
Haplotype	Case	Control	Case frequency	Control frequency
C-A-C	60	21	0.18	0.15
C-G-T	150	47	0.45	0.33
C-G-C	75	23	0.22	0.16
T-G-T	4	3	0.012	0.022
T-A-C	5	4	0.015	0.028
T-G-C	40	44	0.12	0.31
Likelihood ratio chisq = 58.98 df = 5 p-value = 1.98E-11^				

* adjusted for principal components, sex, and cohort of origin

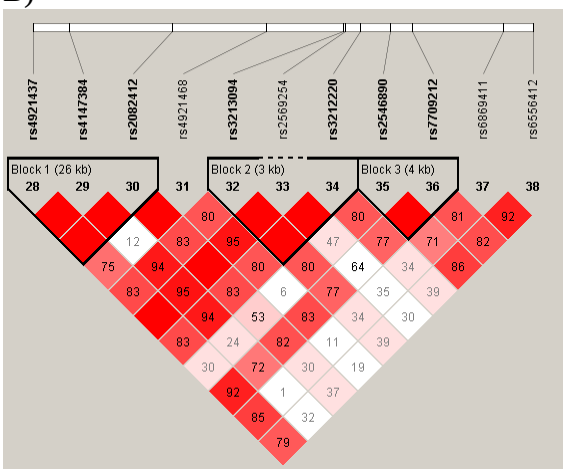
^ adjusted for principal components and sex

Figure 3-2. Haploview plots of D' in the IL12B region (96 kb range) of the study populations from A) Uganda and B) Tanzania, and in African HapMap populations C) MKK, D) LWK, E) ASW, and F) YRI

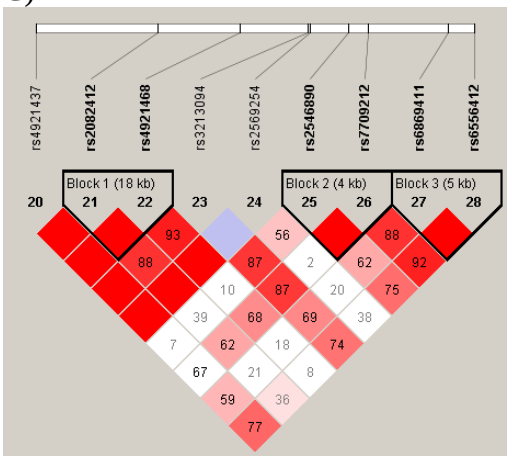
A)



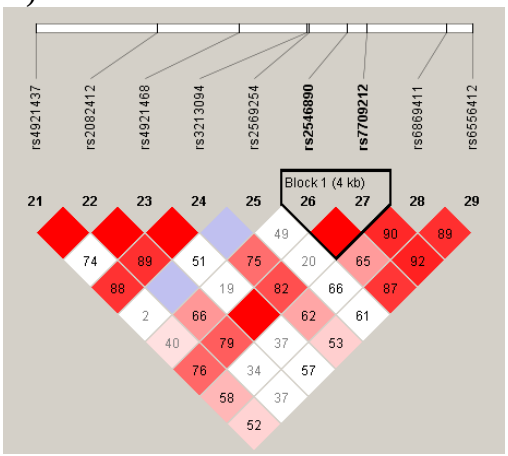
B)



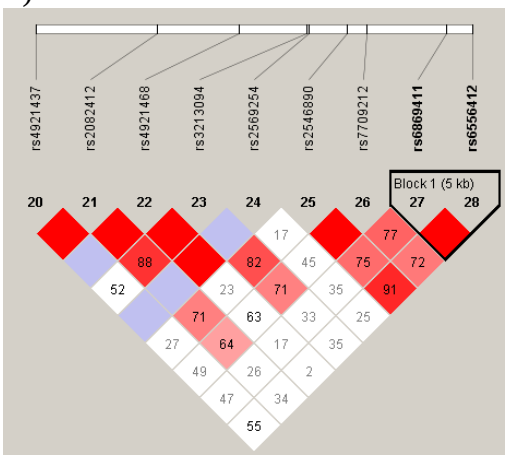
C)



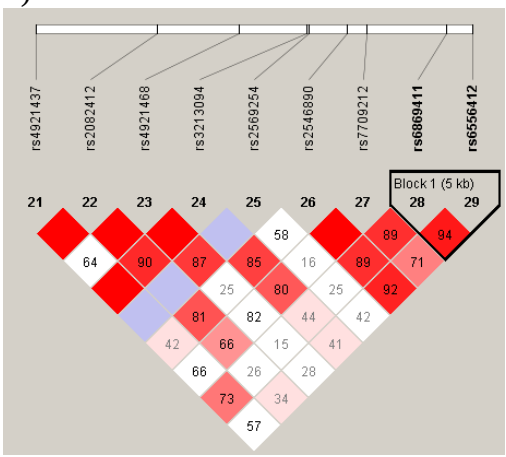
D)



E)



F)



We evaluated signatures of selection using the absolute value of iHS[88] in less admixed equatorial African populations for which we had available data with greater coverage than provided by the Human Exome Beadchip[118]. The iHS test is most sensitive to recent adaptive events (<25 thousand years ago), which is appropriate for TB as its selective pressure became more prominent following recent increases in population density associated with agriculture[88,118]. We found two intronic SNPs in *IL12B* in the upper 0.1% of the distribution of absolute values of iHS scores of 1 million genotyped SNPs in two of these populations. The region surrounding rs3213093 has a signature of selection in the upper 0.1% of the distribution in the Datog population sample and nearly meets the criterion in the Niger-Kordofanian-west population samples, while the region centered on rs2421047 has a signature of selection in the upper 0.1% of the distribution in the Niger Kordofanian West population samples and nearly meets the criterion in the Datog (Appendix Table 3-15). The |iHS| scores were not as extreme in the Western Pygmy populations, though they were both >2.5. The two SNPs are 4,672 bases apart. In our study cohorts, rs3213093 and rs2421047 are in high LD with each other (D' = 0.98 - 1.0), and, importantly, with rs4921437 (D' = 0.79 - 0.92). The iHS scores of rs4921437 were not as extreme, potentially because the SNP is near the end of an *UBCLP1*/*IL12B* haplotype block (Appendix Table 3-15)[88]. The Datog live in Tanzania, practice agro-pastoral subsistence, and speak a language belonging to the Nilo-Saharan language family that includes the majority of languages spoken in Uganda and Tanzania. The Niger Kordofanian West population sample used in this study includes Yoruba (living in Nigeria), Southern Tikar, Ngumba, Lemande (living in Cameroon), all speaking languages belonging to the Niger-Kordofanian language family and all practicing agriculture. *In toto*, this demonstrates strong selection in the region of interest in

populations related to our study cohorts, although the selection signature appears to be strongest within the *IL12B* gene and not in the regulatory region in which rs4921437 is located.

The Household Contacts Study design guarantees exposure of the TB-negative patients to MTB. All patients from the DarDar Nutrition Study were cases of tuberculosis. Direct contact with index cases of active TB was not assessed in the DarDar vaccine trial. Importantly, 8 years elapsed between the completion of enrollment for the vaccine trial and the recruitment into this study; therefore these participants were HIV-positive and living in an MTB hyperendemic area for at least that long. To further address possible confounding by lack of MTB exposure, we assessed response to interferon gamma release assays and lymphocyte proliferation assays by case control status, and no statistically significant differences were observed (Appendix Table 3-16).

We validated previously reported TB-association variants present on the Human Exome Beadchip. Variants rs2057178 and rs4331426, previously associated with TB in genome-wide analyses had similar odds ratios to those previously reported (rs2057178, OR = 0.84 in our combined cohort versus 0.77 in Thye et al, 2012, and rs4331426, OR = 1.16 in our combined cohort versus 1.18 in Thye et al, 2010). However, our sample size is smaller than the reported studies and the SNPs were not significant at the replication threshold of 0.05 in our analyses (Table 3-4). We also observed that a SNP in *PTX3*, rs3816527, associated with TB at a p value of 4.69×10^{-4}. This SNP is 715 bases away from rs1840680, a previously associated variant, and it is in strong LD with it in the LWK 1000 Genomes population ($r^2 = 0.86$). We imputed rs1840680, and the SNP associated with a p value of 0.0013, and an odds ratio of 1.63 (Table 3-4).

Table 3-4. Test of replication of polymorphisms from other TB studies adjusted for principal components, sex, and cohort of origin

SNP	Study	Chr.	Minor Allele	n	Odds Ratio	95% CI	p value	Gene
rs2057178	Thye T. et al 2012	11	A	581	0.84	(0.64, 1.10)	0.200	*WT1*
rs4331426	Thye T. et al 2010	18	G	556	1.16	(0.68, 1.51)	0.259	*GATA6*
rs3212227	Morris G.A.J. et al	5	G	581	1.37	(1.05, 1.79)	0.0209	*IL12B*
rs1840680*	Olesen R. et al 2007	3	A	581	1.63	(1.21, 2.21)	0.0013	*PTX3*

*imputed SNP

There are several weaknesses in the described analyses. The choice of an extreme phenotype limited the recruitment into our study. As a consequence, a small sample size was used in this study relative to the other published GWAS of tuberculosis. Furthermore, we were only able to recruit very few cases of active tuberculosis in the DarDar vaccine trial extended follow up. To increase the statistical power in analyses of the Tanzanian study population, we added samples from the Nutrition Study, which enrolled only female participants. This recruitment bias makes it likely that the gender balance between cases and controls in our study cohort is not representative of the general population in Tanzania. However, in the combined cohort analyses, the odds of active tuberculosis associated with male sex are roughly the same as that observed in the general population. Another weakness stems from the fact that the low density of the Exome Beadchip prevented us from using our own data in selection analyses. Instead we used other available data from equatorial Africa, and while we were able to demonstrate that populations in that region have recently undergone positive selection around *IL12B*, we cannot definitively state that the same was observed in our own data. Finally, we did not have CD4 data available for most participants from the Household Contacts Study. Consequently, we did not adjust for the variable, as it would drop our sample size by 160 participants. In the DarDar vaccine trial extended follow up, the mean CD4 was higher in cases than it was in controls, but since we did not have the variable for all participants, it is possible that it would confound results.

In summary, we present a novel approach to the study of TB genetics, examining resistance as opposed to susceptibility. Specifically, we hypothesized that immunocompromised individuals who resist TB represent an extreme phenotype that can provide new insights into pathogenesis. Within our HIV-positive cohorts we found a variant in the previously associated

IL12B region that had a much larger effect size than all previously associated common variants, an effect remarkably detected with a sample size far below generally-accepted GWAS power criteria. Selection of extreme phenotypes can be a powerful strategy for unravelling genetic susceptibility to complex infectious disease, as it reduces noise associated with both phenotypic and genetic heterogeneity and facilitates detection of signals from major genes even in apparently "underpowered" samples. In our study, we identified a locus encompassing a classic gene, *IL-12B*, already shown to underlie monogenic immunodeficiency and high susceptibility to mycobacterial infections[7,108,109]. Immunocompromised by HIV *in lieu* of congenital immune deficiency, our patients are positioned at the opposite end of the genetic- phenotypic spectrum, carrying variants for resistance rather than susceptibility to infection. Unlike immunodeficiency-causing mutations identified *because* they determine disease, such variants can only be found when looking for protection *against* disease, i.e. under strong selective pressure provided by HIV, from which they emerge conferring noticeable resistance in subjects otherwise expected to be diseased. Furthermore, our results in the *IL12B* region indicated unexpectedly high concordance of LD patterns across several African-descent populations, and analyses using other sub-Saharan populations demonstrated that the *IL12B* region has undergone positive selection. Our SNP of interest is located in an area enriched for a histone acetylation mark often found in active regulatory elements, suggesting possible functionality and a genetic-epigenetic interaction at the site. Further studies of this interaction are warranted.

CHAPTER IV

GENOME-WIDE ASSOCIATION STUDY IDENTIFIES A RESISTANCE LOCUS TO *MYCOBACTERIUM TUBERCULOSIS* INFECTION IN THE *SLC25A48/IL9* REGION PREVIOUSLY ASSOCIATED WITH BRONCHIAL HYPERRESPONSIVENESS

A. Genome-wide association study of tuberculin skin test response

Introduction

One third of the world's population has been infected with *Mycobacterium tuberculosis* (MTB)[2,119]. Subsequent tuberculosis disease (TB) occurs during the lifespan of about 10% of those infected[2,119,120]. Tuberculosis is a major cause of morbidity and mortality worldwide, with 1.5 million deaths and 9 million new cases of active disease reported in 2013[2]. Tuberculosis is the primary cause of death in people co-infected with the human immunodeficiency virus (HIV), and 360,000 of the TB deaths occurred in this patient population[1,2]. The immunosuppression from HIV facilitates progression to active disease directly following infection, or by the reactivation of a latent MTB infection[3,4]. While the clinical trajectory of a given MTB infection has many determinants and possible outcomes, infection itself is a necessary prerequisite. Of note, about 20% of people living in areas hyperendemic for MTB, virtually guaranteeing repeated exposure, appear to be resistant to infection[12,61,121,122].

Historically, MTB infection has been evaluated with a tuberculin skin test (TST) that measures the induration caused by a delayed type hypersensitivity reaction to an intradermal injection of an MTB purified protein derivative (PPD)[21,123]. In endemic areas, induration ≥ 5mm measured between 48 and 72 hours after the injection is indicative of an MTB infection. A sibling study of TST reactivity demonstrated high heritability, suggesting a possible genetic component to the MTB infection resistance phenotype[124]. Several studies have capitalized on this

finding and identified loci relevant to the MTB infection phenotype. A family-based linkage analysis of TST response identified *SLC6A3* and a region on chromosome 11 (p14) as loci that associate with MTB infection[12]. A full genome microsatellite scan comparing persistent MTB negative patients to those with latent infections identified an association with the *SLC11A1* gene, and candidate regions on chromosomes 2 (q14, q21-q24) and 5 (p13-q22)[13].

Recently, novel methods for evaluating MTB infection status have been developed. Interferon-gamma release assays (IGRA) detect the concentration of IFN-γ in response a mixture of MTB-specific antigens[27,28]. The purified protein derivative used in TST has some antigenic overlap with the Bacille Calmette-Guérin (BCG) vaccination; therefore it is impossible to distinguish between a TST reaction to an MTB infection or to childhood BCG in individuals who received the vaccine. IGRA antigens have no overlap with the BCG vaccine, and maintain excellent specificity in individuals who had childhood BCG vaccinations[27,28]. Additionally, in people with compromised immune systems, anergy may prevent detection of a positive TST and/or IGRA in individuals previously exposed to the MTB. Lack of response to an IGRA positive control identifies individuals with possible immune anergy. Using these methods, we can identify a more refined phenotype that minimizes both false positives and false negatives.

We used a genome-wide approach to evaluate common variants that associate with TST response in a patient population that hypothetically allows us to identify extreme effects. Namely, we hypothesized that HIV-positive individuals who live in areas endemic for MTB but do not get infected are genetically resistant. Using two recently concluded prospective cohorts of tuberculosis disease from Tanzania and Uganda, with available TST and interferon gamma release assay results, we identified a variant in the *SLC25A48/IL9* region that confers resistance to MTB infection.

Methods

Study populations

Tanzania

Patients from the extended follow-up cohort of the DarDar vaccine trial in Dar es Salaam, Tanzania were recruited for this study. The full cohort has been described elsewhere[17]. Briefly, the DarDar trial was a phase III randomized trial of SRL 172, an inactivated whole cell mycobacterial vaccine booster to a childhood Bacille Calmette-Guérin (BCG) vaccination. Subjects were enrolled between 2001 and 2005. Follow-up continued until the study was concluded in 2008. Recruited patients had to be HIV-positive adults (≥18 years old) with a BCG scar, and had to have a CD4 count >200 cells/μl and be TB-negative at the time of enrollment. TST reactivity was measured at enrollment, preventing any confounding by the effects of the vaccine. A saline placebo was administered to 1007 patients, while 1006 patients received 5 doses of the vaccine. A routine follow up for active TB (physical examination, chest radiograph, sputum samples for culture and AFB stain, and phlebotomy for an automated mycobacterial blood culture) was performed every 3 months for the duration of the study. Upon conclusion of the trial, an extended follow up cohort of 800 participants from both the placebo and vaccine arm was selected for annual evaluation for active TB. Between September and December of 2013, 304 patients from the extended follow-up were recruited during their routine visits.

Uganda

We obtained 263 samples from HIV-positive participants from the Household Contact Study (HHC), conducted in Kampala, Uganda. This cohort has been previously described in detail[14,68]. Briefly, the Uganda National Tuberculosis and Leprosy Programme referred patients diagnosed with new active tuberculosis to the study, and patients who consented were enrolled as

index cases. Household contacts were defined as individuals living in the same household as the index case for at least 7 consecutive days in the 3 month period leading up to the diagnosis of the index case[122]. Household contacts were subsequently enrolled and evaluated for active TB, latent TB, and HIV. Recommended therapy was administered to all cases of active TB[69]. Of note, the HHC study design guarantees exposure of the controls to MTB during the follow up[68]. We only analyzed adult participants (≥18 years old) of the HCC.

DNA isolation and genotyping

For participants from the extended follow-up of the DarDar vaccine trial, 5ml of whole blood was drawn upon enrollment, and DNA was extracted the day of the phlebotomy using the Gentra Puregene Blood kit (QIAGEN) in accordance with the manufacturer's recommendations. For participants of the Household Contacts Study, buffy coats were isolated on site and shipped to Dartmouth College for DNA extraction. The QIAamp DNA Blood Mini Kit (QIAGEN) was used to isolate DNA from the buffy coats. DNA samples were stored at -80°C before genotyping.

Samples from the DarDar vaccine trial (n=304) and the Household Contact Study (n=263) were submitted for genotyping at the Hussman Institute for Human Genetics, Miami, Florida. DNA quality was evaluated with the 260/280 ratio using a NanoDrop 2000 spectrophotometer at Dartmouth College (Thermo Scientific) and an Electrophoresis Quality Score at the University of Miami. A total of 567 samples passed quality control measures and were genotyped using the Human Core Exome Beadchip (542,585 SNPs). SNPs with a genotyping call rate < 0.95 and a Hardy-Weinberg equilibrium p value < $1*10x10^{-4}$ were excluded. Participants with a per individual genotyping call rate < 0.95 were excluded. Concordance of reported and genotypic gender was verified. In case of relatedness among study participants (pi-hat > 0.20), one individual was randomly removed. All quality controls were

performed in PLINK(v1.07)[70,71]. The final study population included 278 participants from the extended follow up of the DarDar vaccine trial and 213 participants from the Household Contact Study.

Immune Assays

Tanzania

Intradermal injections of purified protein derivative (0.1 ml, RT-23, State Serum Institute, Copenhagen) on the forearm were administered to all enrolled patients prior to vaccination, and resultant skin induration size was measured by trained personnel after 48–72 hours. Preventative isoniazid treatment (300 mg daily for 6 months) was offered to subjects with a positive reaction using the criterion for HIV-positive patients (≥5 mm as positive).

Immune response to *Mycobacteria* was assessed with an interferon gamma (IFN-γ) enzyme linked immunosorbent assay (ELISA), a tritiated thymidine lymphocyte proliferation assay (LPA) and an ELISA for antibodies to the glycolipid lipoarabinomannan of MTB (LAM). The assays used in this study have been described in detail elsewhere[84]. Briefly, phlebotomy was performed prior to vaccination and at the conclusion of the study, and peripheral blood mononuclear cells (PBMCs) were isolated by ficoll density gradient centrifugation for IFN-γ and LPA assays, performed on site. Centrifuged, frozen serum was sent to Dartmouth College for LAM assays.

IFN-γ and LPA assays used four different antigens: *Mycobacterium Vaccae* sonicate (2 mcg/ml), MTB Antigen 85 (Ag85; 1 mcg/ml), MTB early secretory antigenic target 6 (ESAT-6; 2 mcg/ml), and MTB whole cell lysate (WCL; 1 mcg/ml)[84]. Importantly, ESAT-6 is not present in the childhood Bacille Calmette-Guérin (BCG) vaccine that is commonplace in East Africa; therefore confounding by BCG status can be controlled for with these additional data.

Media alone was used as a negative control and phytohemagglutinin (PHA, 2.5 mcg/mL; Sigma, St. Louis, MO) was used as a positive control[84].

Uganda

Intradermal injections of purified protein derivative (5 tuberculin units) on the forearm were administered to study participants at enrollment, after 3, 6, 12, and 24 months if the tests were negative at earlier time points[122]. The size of skin induration was measured by trained personnel 48–72 hours after each injection. For patients measured at multiple time points the largest TST reaction was used in analysis. In Phase II of the HHC study, daily isoniazid treatment was offered to all participants with positive TST results (≥5 mm)[13,122].

The immune assays used in this study have been detailed elsewhere[13,122,125]. Briefly, phlebotomy was performed at enrollment. Whole blood was stimulated with MTB antigens: MTB culture filtrate CXFT, ESAT-6, and CFP10[122,125], and the IFN-γ response was measured by ELISA (Thermo Scientific, Rockford, IL). Whole blood cultured without antigen stimulation served as a negative control. Phytohemagglutinin (PHA; Sigma, St. Louis, MO) was used as a positive control, while the IFN-γ response to media was subtracted from antigen-stimulated readings[122,125]. Negative differences were considered a 0.

Statistical Analyses

TST data were evaluated using additive, dominant, and recessive genetic models both as a continuous variable using linear regression, and as a binary variable (< versus ≥ 5mm) with logistic regression in PLINK(v1.07)[70,71]. *A priori* power analyses in Quanto[72] revealed that to discover an association with an odds ratio of 2.0 for the binary TST outcome below two-sided p value threshold of 0.05, we needed to set the minor allele frequency (MAF) at > 0.2 for the

single SNP association analyses (Appendix Table 4-1). A total of 162,228 SNPs passed the inclusion criteria at an MAF > 0.20. If we consider the SNPs passing quality control as independent tests, the corresponding Bonferroni corrected multiple testing threshold is 3.08×10^{-7}. Summary statistics and univariate logistic regression models of TST case/control status with available covariates were calculated in STATA(v11.2)[73]. Covariates associating with case/control status at the 0.05 level were included in the final models.

Participants from the extended follow up of the DarDar vaccine trial were recruited in Dar es Salaam, Tanzania, and the patients in the Household Contact study were recruited in Kampala, Uganda, both large urban centers. To adjust for possible admixture within each cohort, principal components were calculated using SNPs with $r^2 < 0.1$ and MAF > 0.2 using the SNPRelate package in R[74,75]. All analyses were adjusted for 10 principal components, and analyses of the cohorts combined were adjusted for a cohort variable. The qqman package in R was used to generate Manhattan and qq plots[76]. Locus zoom was used to plot the region of the SNP with the strongest association[77]. SNPs in the region of interest (rs877356 position +/- 0.5 megabase) were imputed with IMPUTE2 (v2.3.1), using one phased reference panel from the 1000 Genomes project[78-80]. All SNPs with an MAF > 0.05 in each cohort were included in the haplotype analyses of the region of interest. Haplotype plots were generated using Haploview[81]. We used UNPHASED(v3.1.7)[62] to perform two and three SNP haplotype association analyses, adjusting for the same covariates as in the single SNP association analyses above. We studied all pairwise haplotypes for 30 available SNPs with a minor allele frequency >0.05, within 250kb of the SNP with the most significant association in the combined cohort. We then performed a three-SNP haplotype analysis using the two most significant haplotypes in pairwise analyses, as one of the SNPs was common to both.

There are three major ways to remain TST negative given an exposure to *M. tuberculosis*: 1) MTB can be inhaled but mechanically prevented from seeding the lungs, 2) it can seed the lungs but be cleared before immune memory is invoked, or 3) it can establish a latent infection but host immunosuppression and an inability to mount a delayed type hypersensitivity response can prevent a positive TST test, i.e. anergy. To adjust for possible immune anergy in Tanzania, we removed all patients who had negative responses to all interferon-gamma release assay (IGRA) antigens, and negative responses to the positive control antigen, PHA, in IGRA (defined as a PHA < 300pg/mL) and lymphocyte proliferation assays (LPA) (defined as a proliferative index < 3). LPA data was not available for the Ugandan cohort; therefore we removed all patients who were IGRA negative to all available antigens and to PHA (PHA < 300pg/mL). Patients who were PHA positive but IGRA negative for all antigens remained in the study. IGRA assays were not performed on 102 patients from the HHC cohort and 33 patients from the DarDar vaccine trial extended follow up. Of the patients with missing IGRA assays, 71 had TST measurements ≥ 5mm, and 31 < 5mm in Uganda, and 10 had TST ≥ 5mm, and 23 < 5mm in Tanzania. Logistic regression models of TST case/control status adjusting for missing IGRA data were performed to prevent confounding by missing data. Patients who stated that they had previous active TB, but had a TST of 0mm were excluded from the main analyses presented below. Logistic regression models of TST case/control status including this patient population are presented in the supplement.

To evaluate the effect of possible false negative TST responses on our association results, additional logistic regression analyses were performed removing individuals with a 0mm TST induration who had a substantial INF-γ response (> mean in cases) to any of the tested antigens at the time of the TST induration measurement. The effect of possible false positive TST results

due to BCG vaccination was evaluated by performing logistic regression analyses removing individuals with a positive TST scores but low IGRA response (< mean of controls) to any of the tested antigens.

Functional Annotation

The ENCODE Project[91] was accessed via the UCSC Genome Browser[87] and used for functional annotation.

Ethics

Informed consent was obtained from all patients in the extended DarDar follow-up cohort, at the Dar es Salaam clinic. The research ethics committee at the Muhimbili University of Health and Allied Sciences and the Committee for the Protection of Human Subjects at Dartmouth College and the Dartmouth-Hitchcock Medical Center approved this study. Informed consent was obtained from all subjects in the Household Contacts study in Kampala, Uganda. The institutional review boards at the Uganda Council for Science and Technology and the University Hospitals of Cleveland approved this study.

Results

Gender was not significantly associated with TST case/control status in Uganda (p = 0.76; Table 4-1A) nor Tanzania (p = 0.349; Table 4-1B), but it did associate in the combined cohort (Odds ratio for males 1.91, 95% CI = 1.27-2.86, p value = 0.002; Table 4-1C). Age was not significantly associated with TST case/control status in Uganda (p = 0.384; Table 4-1A), Tanzania (p = 0.153; Table 4-1B), nor in the combined cohort (p = 0.108; Table 4-1C). Therefore, all analyses below were adjusted for gender, along with 10 principal components, and cohort of origin, where appropriate.

In logistic regression analysis, we observed a genome-wide significant association between a dominant genetic model of common variation in rs877356 and binary TST status in the combined cohort (Odds ratio = 0.27, 95% CI = 0.17-0.42, p = 1.22×10^{-8}; Table 4-2A, Appendix Figures 4-1 through 4-3). The variant had consistent effects in Uganda (OR = 0.17, 95% CI = 0.078-0.37, p = 9.18×10^{-6}; Table 4-2B, Appendix Figure 4-4) and Tanzania (OR = 0.3295, 95% CI 0.1843-0.5892, p = 1.81×10^{-4}; Table 4-2C, Appendix Figure 4-5). Linear regression analyses of continuous TST scores in a dominant genetic model produced similar results (combined cohort beta = -4.14, 95% CI = -5.55 to -2.74, p = 1.45×10^{-8}; Appendix Tables 4-2). Variant rs877356 was below the multiple testing adjusted threshold for this study (3.08×10^{-7}) and nearly genome-wide significant in an additive model using a binary TST designation (OR = 0.33, 95% CI = 0.22-0.49, p = 5.45×10^{-8}; Appendix Table 4-3), and continuous TST scores (combined cohort beta = -3.34, 95% CI = -4.53 to -2.14, p = 6.95×10^{-8}; Appendix Table 4-4). No other SNPs were significant at the multiple testing corrected threshold in any of the genetic models tested (Table 4-2, Appendix Tables 4-2 through 4-6).

Table 4-1. Summary statistics of study participants from by TST case/control status (≥ versus < 5mm, respectively) in A) the extended follow up of the DarDar vaccine trial in Tanzania, B) the Household Contacts Study in Uganda, and C) the combined cohorts

A)

		Cases	Controls	Odds Ratio	95% Confidence Interval	p value
n^		150 (75.4)	49 (24.6)			
Age* (years)		32.72 ± 7.78	31.57 ± 8.76	1.02	(0.98, 1.06)	0.384
Gender^	Male	68 (76.4)	21 (23.6)			
	Female	82 (74.5)	28 (25.5)	1.11	(0.58, 2.12)	0.762

B)

		Cases	Controls	Odds Ratio	95% Confidence Interval	p value
n^		94 (34.8)	176 (65.2)			
Age* (years)		33.50 ± 8.56	34.95 ± 7.61	0.98	(0.95, 1.01)	0.153
Gender^	Male	21 (40.4)	31 (59.6)			
	Female	73 (33.5)	145 (66.5)	1.35	(0.72, 2.50)	0.349

C)

		Cases	Controls	Odds Ratio	95% Confidence Interval	p value
n^		244 (50.9)	235 (49.1)			
Age* (years)		32.43 ± 8.02	34.49 ± 7.96	0.98	(0.96, 1.01)	0.108
Gender^	Male	89 (63.1)	52 (36.9)			
	Female	155 (47.3)	173 (52.7)	1.91	(1.27, 2.86)	0.002

*mean ± standard deviation; ^ n (% of row)

Table 4-2. Single nucleotide polymorphisms associating with case/control tuberculin skin test induration status (< versus ≥ 5mm) below a $5x10^{-5}$ p value using a dominant genetic model in A) the combined cohort*, B) the Ugandan cohort^, and C) the Tanzanian cohort^

A)

SNP	Chr.	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Nearby gene
rs877356	5	T	0.23	469	0.27	(0.17, 0.42)	1.22E-08	*SLC25A48/IL9*
rs7808481	7	A	0.22	469	2.52	(1.63, 3.91)	3.33E-05	*Loc340268*
rs1880386	10	A	0.21	469	2.46	(1.59, 3.80)	4.85E-05	*GRID1*

B)

SNP	Chr.	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Nearby gene
rs877356	5	T	0.23	199	0.17	(0.078, 0.37)	9.18E-06	*SLC25A48/IL9*
rs654718	11	G	0.21	199	0.19	(0.089, 0.41)	1.81E-05	*MRE11A*
rs7944514	11	C	0.41	199	5.28	(2.46, 11.36)	2.03E-05	*POLD3*
rs7837658	8	T	0.45	199	4.84	(2.32, 10.11)	2.67E-05	*RNF19A*

C)

SNP	Chr.	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Nearby gene
rs17062122	6	C	0.33	270	0.28	(0.16, 0.49)	6.20E-06	*Loc285735*
rs8142256	22	C	0.35	270	0.31	(0.18, 0.54)	2.87E-05	*FAM19A5*
rs10998959	10	T	0.25	270	0.31	(0.17, 0.54)	4.33E-05	*Loc100129281*
rs11736841	4	T	0.26	270	3.09	(1.79, 5.33)	4.96E-05	*ODZ3*
…	…	…	…	…	…	…	…	...
rs877356	5	T	0.23	270	0.33	(0.18, 0.59)	1.81E-04	*SLC25A48/IL9*

* adjusted for 10 principal components, sex, and cohort of origin

^ adjusted for 10 principal components and sex

To evaluate SNPs in the region not included on our genotyping array, we used IMPUTE2 (v2.3.1)[79,80] software to impute SNPs within 0.5 megabases of rs877356. rs17169187, in high linkage disequilibrium (LD) with rs877356 (D' = 1 in both cohorts, r^2 = 0.99 in Tanzania, 0.98 in Uganda) and 2,340 bases away, is the variant with the most significant association to TST case/control status using a dominant model (combined cohort OR = 0.25, 95% CI = 0.16-0.40, p = 4.57×10^{-9}; Figure 4-1, Appendix Table 4-7A). The results were consistent when we used linear regression on TST induration as a continuous variable (combined cohort beta = -4.28, 95% CI = -5.69 to -2.88, p = 4.58×10^{-9}; Appendix Table 4-8A). The variant is also genome-wide significant in additive modeling of both a dichotomous TST designation (combined cohort OR = 0.3196, 95% CI = 0.22-0.48, p = 2.56×10^{-8}; Figure 4-1, Appendix Table 4-7B) and continuous TST induration (combined cohort beta = -3.43, 95% CI = -4.62 to -2.24, p = 2.84×10^{-8}; Appendix Table 4-8B).

In the Tanzanian cohort, IGRA responses did not differ between TST cases/controls in positive (PHA) and negative controls (MEDIUM), but were significantly higher in TST cases for all test antigens (Appendix Table 4-9B). In Uganda, we observed the same trends; however, due to smaller sample sizes, the comparisons were not statistically significant (Appendix Table 4-9A). When we removed 16 patients with TST results of 0 mm but at least one IGRA > than the mean observed in patients with TST > 5 mm, the results of the association remained consistent and a dominant model of rs8773556 was significant at the genome-wide significance threshold with both the binary and continuous TST outcomes (Appendix Table 4-10). When we removed 20 patients from the analysis because of possible positive TST reaction to a childhood BCG vaccination, a dominant model of rs877356 was significant adjusted for multiple testing for both outcomes and nearly genome-wide significant for binary TST (p value = 5.58×10^{-8}; Appendix

Table 4-11). The association was also robust to removing all 36 patients with either suspected false positive or false negative TST results, as the associations remained significant at a multiple testing adjusted threshold (Appendix Table 4-12). The dominant model of rs877356 remained genome-wide significant for either outcome when the analysis was adjusted for a missing IGRA variable (Appendix Table 4-13). The variant was also genome-wide significant when we included patients with prior tuberculosis into the study cohort (Appendix Table 4-14).

We found the strongest single variant association using a dominant model of rs877356; therefore, we used dominant coding of the SNP in 2-variant haplotype in the *SLC25A48* region while using additive models of all other SNPs. An rs877356-rs2069885 haplotype had the strongest association in this analysis (p value 1.59×10^{-12} in the combined cohort; Table 4-3). The haplotype had similar effects in the Ugandan (p value = 2.51×10^{-8}; Table 4-3B) and Tanzanian cohorts (p value = 1.37×10^{-11}; Table 4-3C), with the protective T-G haplotype case/control frequencies of 0.32/0.60 and 0.20/0.45, respectively. The risk haplotype, C-G, also had a consistent distribution between the cohorts, with a case/control frequency of 0.58/0.33 in Uganda and 0.68/0.48 in Tanzania (Table 4-3B and C). The results were consistent in additive modeling of both SNPs (p value = 2.59×10^{-9} in the combined cohort; Table 4-3). The haplotype had similar effects in the Ugandan (p value = 1.03×10^{-5}; Table 4-3B) and Tanzanian cohorts (p value = 6.35×10^{-5}; Table 4-3C).

Table 4-3. Association of the 2-variant haplotype using a dominant model of rs877356 with an additive model of rs2069885 with TST induration case/control status (< versus ≥ 5mm) in the *SLC25A48/IL9* region in A) the combined cohort, B) the Ugandan cohort and C) the Tanzanian cohort

A)

Haplotype	Case	Control	Ca-Freq	Co-Freq
C-A	33	12	0.068	0.027
C-G	301	202	0.62	0.45
T-A	17	17	0.035	0.038
T-G	135	219	0.28	0.49
Likelihood ratio chisq = 57.97 df = 3 p-value = 1.59E-12*				

B)

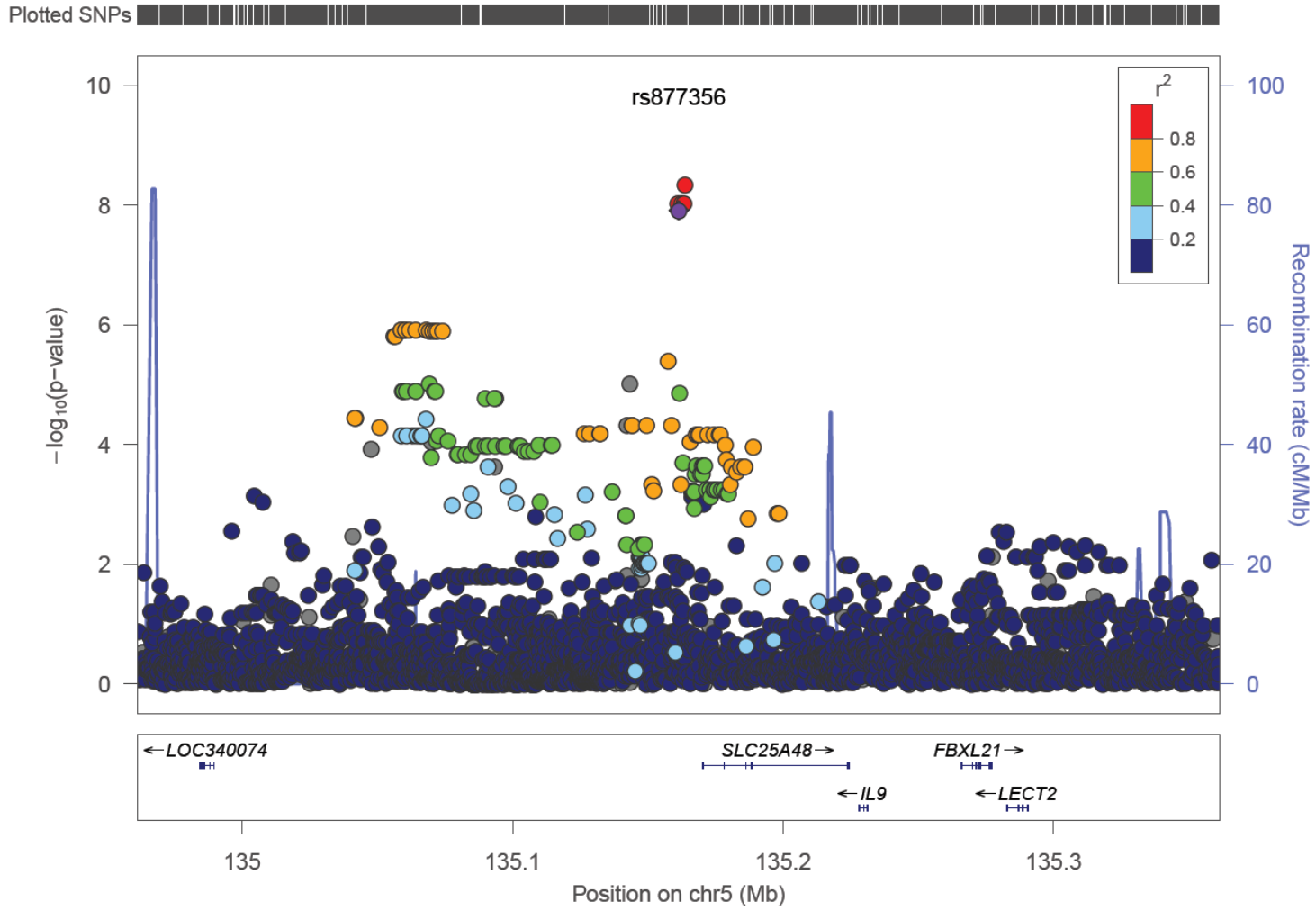
Haplotype	Case	Control	Ca-Freq	Co-Freq
C-A	19	2	0.063	0.020
C-G	175	32	0.58	0.33
T-A	9	5	0.030	0.051
T-G	97	59	0.32	0.60
Likelihood ratio chisq = 38.25 df = 3 p-value = 2.51E-08^				

C)

Haplotype	Case	Control	Ca-Freq	Co-Freq
C-A	14	10	0.075	0.028
C-G	126	170	0.68	0.48
T-A	8	12	0.043	0.034
T-G	38	160	0.20	0.45
Likelihood ratio chisq = 53.59 df = 3 p-value = 1.37E-11^				

* adjusted for principal components, sex, and cohort of origin
^ adjusted for principal components and sex

Figure 4-1. Locus zoom plot of results from a logistic regression association of case/control tuberculin skin test induration status (< versus ≥ 5mm) with a dominant genetic model of imputed SNPs in the *SLC25A48/IL9* region in the in the combined cohort, adjusted for 10 principal components, sex, and cohort of origin; SNP with the most significant association in the Exome Beadchip analysis in purple



Discussion

In this study we examined the association of common genetic variants with *Mycobacterium tuberculosis* infection in HIV-positive patients from the extended follow up of the DarDar vaccine trial in Tanzania and the Household Contacts study in Uganda. We hypothesized that immunosuppressed patients who live in MTB endemic areas but do not get infected have genetic resistance. We identified a novel association between resistance to MTB infection and rs877356, a variant 9,119 bases upstream of the coding region of *SLC25A48*[87], a *Homo sapiens* solute carrier family 25, member 48. *SLC25A48* is a mitochondrial carrier of amino acids[126,127]. The variant of interest is also 57,662 bases downstream from *IL9.*

The involvement of *IL9* as the potentially causal gene in our association study was further supported by our haplotype analyses. The rs877356-rs2069885 haplotype had the most significant association in the *SLC25A48/IL9* region. rs2069885, 66kb away from rs877356, is a missense variant in *IL9* (Threonine (A**C**G) ->Methionine (A**T**G))[87]. While rs2069885 was not significant in univariate analyses (p value 0.0907 in the combined cohort, binary TST, additive model), the association of the haplotype was more significant than that of rs877356 alone.

IL9 is a particularly interesting candidate for MTB infection because it is a cytokine previously associated with bronchial hyperresponsiveness, a heightened bronchoconstriction in response to external stimuli[128]. Bronchial hyperresponsiveness is hereditary, and is a risk factor for subsequent development of asthma[128-130]. Asthma is a chronic condition with symptoms of wheezing, shortness of breath, chest tightness and cough[131]. Of note, the prevalence of asthma in East Africa is very high, especially in urban settings[132]. Childhood MTB infection protects patients from developing asthma, and an inverse relationship between incidence of active TB and asthma has been reported[133,134]. We posit that our variant associates with heightened

inflammation and bronchoconstriction, and that this hyperresponsiveness plays a key role in preventing MTB infection.

A recent review of *Mycobacterium tuberculosis* host evasion suggests a possible mechanism for how airway inflammation prevents MTB infection[20]. In an animal model, large aerosolized particles containing >10,000 bacteria do not readily initiate infection because they get trapped in the upper airway[135]. Microbicidal macrophages in the upper airway are recruited by toll-like receptor signaling upon immune system recognition of bacterial pathogen-associated molecular patterns (PAMP). This mechanism is used to keep mucosal commensal pathogens from migrating into the lower respiratory tract. MTB readily infect macrophages without activating interferon-gamma signaling because they mask their PAMP using a surface lipid phthiocerol dimycocerosate[136]. However, the presence of other TLR-stimulating commensals in the upper airway activates the microbicidal macrophages and MTB is killed as collateral[20]. In order to circumvent this response, MTB travels in small droplet nuclei containing as few as 1-3 bacteria, delivering them directly to the lower lung. There, with less immune stimulation by other pathogens, MTB initiates infection in unactivated macrophages[137]. A persistent state of increased upper airway inflammation would make passage into the preferred alveolar spaces more difficult, and MTB droplet nuclei would be sequestered and consequently killed before reaching the lower lung.

The IL9 cytokine was originally described as a T cell and mast cell growth factor, but has since been found to have further pleiotropic effects on the immune system[138-140]. It is produced by T_H2 cells, naïve CD4+ cells, T_h17 cells, Mast cells, T_H9 cells and possibly T_{reg} cells[138,141-143]. T_H9 cells are primed for IL9 production and are stimulated to differentiate by pleural mesothelial cells in the presence of MTB infection[144,145]. In B cells, IL9 promotes IL4-mediated production

of IgE and IgG antibodies[146,147]. Of note, bronchial hyperresponsiveness is associated with elevated serum IgE levels[128,148]. IL9 also promotes proliferation of hematopoietic progenitor cells[149,150], and it has specific effects on lungs. In airway smooth muscle cells, IL9 induces the expression of chemokine CCL11, thereby inducing eosinophil chemotaxis and allergic reactions, and in airway epithelial cells, IL9 directly induces mucous production and stimulates IL13, which leads to further airway inflammation[138,139,151,152].

The TST phenotype can be studied both as a binary variable, < versus ≥ 5mm induration, or as a continuous outcome. Our single SNP association results were consistent using both outcome types. Variant rs877356 was genome-wide significant in both logistic and linear regression models in the combined cohort using a dominant genetic model. The variant was also significant at a multiple testing corrected level in additive modeling. The most significant imputed variant in the region, rs17169187, was genome-wide significant for both outcomes in additive and dominant modeling.

Immune anergy is an important confounder in studies of TST reactivity in an HIV-positive context. TST responses can be < 5mm because a patient has not been infected with MTB, or in case of infection, they are unable to mount a hypersensitivity reaction to purified protein derivatives on account of their immunosuppresion. We leveraged existing interferon gamma release assay data in both cohorts to evaluate confounding by immunosuppression. We removed all patients suspected of immune anergy prior to analysis, and further adjustment for a missing IGRA variable did not affect the association of our variant, demonstrating the robustness of our findings.

This study demonstrates that the choice of an extreme phenotype, HIV-positive patients who live in MTB endemic areas but do not get infected, allows for the use of relatively small

sample size even in a genome-wide association study. Although the small sample size is the biggest weakness in this study, the large effect size observed in this unique study population allowed us to find significant associations in an apparently relevant region of the genome. The variant with the most significant association is near *IL9*, a gene with a substantial role in airway inflammation, bronchial asthma, and other respiratory infections[153,154]. This, along with some observational studies of the inverse incidence of asthma and tuberculosis, leads to the conclusion that the same gene whose over expression plays a significant role in the pathogenesis of asthma, also prevents MTB infection by the same mechanism.

B. Fine mapping of regions previously associated through genome-wide linkage studies

Introduction

Availability of tuberculin skin test measurements and dense variant coverage of our genotyping array allowed us to follow up on prior studies of MTB infection. A study by Cobat et al. used a genome-wide linkage approach to identify a signal in the chromosome 11p14.1 region for tuberculin skin test reactivity as a binary variable, and the *SLC6A3* region on chromosome 5 when using a TST as a continuous variable[12]. The *SLC6A3* region was fine mapped in a follow up analysis, and SNP rs250682 was the most significant variant. The region on chromosome 11 spanned from 22.35 to 28.82mb and was termed *TST1*. No fine mapping follow up analysis was performed on *TST1*; therefore, no single SNP could be identified as associating with the phenotype. The study evaluated TST reactivity in 128 families, comprised of 186 parents and 350 offspring from the Western Cape region of South Africa. Importantly, HIV positive participants were excluded from analyses[12].

A genome-wide linkage analysis by Stein et al. compared persistently TST negative patients to individuals with latent infection[13]. None of the analyses were significant when adjusted for multiple testing; however three regions approached significance, 2q14, 2q21-2q24 and 5p13-5q22 (p value < 0.005 for all). The study population was comprised of 193 families and a total of 803 participants, 130 of whom were HIV-positive. HIV-positive patients were included in the analysis, and the final model was adjusted for HIV status[13].

We used the genome-wide data from the Human Exome Beadchip to evaluate if common variants in these previously identified regions associate with TST response in HIV-positive individuals who live in areas endemic for MTB but do not get infected. Using two recently

concluded prospective cohorts of tuberculosis disease from Tanzania and Uganda, we fine mapped each area, and identified candidate SNPs in each region.

Methods

The study populations, genotyping, exclusion criteria, and covariates are described in the Methods section of Chapter IV Part A. SNPs in each respective region of interest (*SLC6A3* +/- 0.5 mb, chr11:22.35-28.82mb, 2q14, 2q21-2q24 and 5p13-5q22) with a minor allele frequency >0.10 were tested for association using the Exome Beadchip data in the combined cohort. Analyses were performed using the same outcomes (binary versus continuous TST) as described in the initial analyses; however, only covariates associating with the phenotype in our study cohorts were used in the final models (described in Part A). After identifying the most significant SNP within each region using our data from the Exome Beadchip, we imputed all ungenotyped SNPs +/- 0.5mb of the index SNP with IMPUTE2 (v2.3.1), using one phased reference panel from the 1000 Genomes project[78-80]. Association analyses were then performed on the imputed datasets.

A prior study by Sobota and Shriner et al. demonstrated that linkage disequilibrium patterns make a strict Bonferroni adjustment (0.05/number of SNPs) of a p value threshold too conservative in genetic studies[115]. Since polymorphisms in the same region are not independent of each other, adjusting the p value by their total number is statistically punitive. Even in African populations, known for shorter LD blocks in comparison to other global populations, the Bonferroni adjusted p value threshold was observed to be about half an order of magnitude too conservative. Consequently, for fine mapped polymorphisms with a p value within a half order of magnitude for the multiple testing adjusted threshold, we used a linkage disequilibrium

pruning-based method to adjust the conservative Bonferroni significance threshold to a more natural one.

Results

Linear regression in the SLC6A3 region

We were unable to validate any of the quantitative trait loci fine mapped in the *SLC6A3* region by Cobat et al using a standard statistical significance criterion. The most significant association was reported rs250682[12]; however, this variant did not associate with a continuous TST phenotype in our combined cohort analysis (p value = 0.72; Table 4-4). Of the most significant variants reported by Cobat et al, rs4975579 was closest to replicating in our study (p value = 0.0805; Table 4-4). Unfortunately, the authors did not disclose the risk allele or the effect size; therefore, we cannot even assess whether we are reporting an effect trending in the same direction.

The most significant association for the Human Exome Beadchip variants within 0.5mb of rs250682 was rs10056116 (Beta = 1.89, 95% CI = 0.63-3.15, p value = 0.00351; Table 4-5A). There were 117 variants passing quality control in that region. Therefore, the top SNP did not pass a multiple testing criterion (threshold 4.27×10^{-4}). After imputation, rs13186183 is the variant with the most significant association in the region (Beta = -2.85, 95% CI = -4.56 to -1.23, p value = 7.10×10^{-4}; Table 4-5A).

Table 4-4. Association of variants in the *SLC6A3* region significant in Cobat et al. with continuous TST scores in an additive genetic model in the combined Ugandan-Tanzanian cohort; adjusted for 10 principal components, sex, and cohort of origin

SNP	Minor Allele	MAF	Beta	95% Confidence Interval	p value
rs4975579	G	0.43	-0.86	(-1.81, 0.10)	0.080
rs6554677	A	0.21	-0.57	(-1.81, 0.66)	0.361
rs1801075	C	0.16	-0.27	(-1.62, 1.08)	0.697
rs250682	C	0.48	-0.19	(-1.19, 0.81)	0.715
rs10475030	T	0.49	0.013	(-0.95, 0.97)	0.978
rs1018120	T	0.088	-1.16	(-2.99, 0.67)	0.214
rs2232376	T	0.24	0.31	(-0.85, 1.48)	0.600

Table 4-5. Fine mapping of available and imputed variants using an additive genetic model and continuous TST scores in A) the *SLC6A3* region and binary TST scores in regions, B) chromosome 11p14, C) chromosome 2q14, D) chromosome 2q21-q24, and E) chromosome 5p13-q22; adjusted for 10 principal components, sex, and cohort of origin

A)

SNP	Source	Position	Minor Allele	MAF	Beta	95% Confidence Interval	p value
rs10056116	Exome Beadchip	1676638	A	0.18	1.89	(0.63, 3.15)	0.00351
rs13186183	Imputed	1791180	T	0.096	-2.89	(-4.56, -1.23)	7.10E-04
rs7715002	Imputed	1499057	T	0.067	3.36	(1.37, 5.35)	0.00102
rs58371035	Imputed	1505771	A	0.062	3.39	(1.32, 5.46)	0.00144
rs186894420	Imputed	1734126	C	0.015	5.97	(2.24, 9.70)	0.0018

B)

SNP	Source	Position	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value
rs17234274	Exome Beadchip	23214366	C	0.37	0.58	(0.43, 0.78)	2.85E-04
rs10834029	Imputed	23306545	G	0.49	1.80	(1.34, 2.43)	1.07E-04
rs76936560	Imputed	23242837	A	0.39	0.55	(0.40, 0.75)	1.57E-04
rs1384479	Imputed	23243454	T	0.37	0.55	(0.41, 0.76)	1.88E-04
rs2449427	Imputed	23215101	C	0.39	0.57	(0.42, 0.77)	2.71E-04
rs17234274	Imputed	23214366	C	0.37	0.57	(0.42, 0.77)	2.93E-04

C)

SNP	Source	Position	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value
rs4848637	Exome Beadchip	121579245	A	0.16	0.52	(0.35, 0.77)	0.0013
rs4848638	Imputed	121579604	A	0.16	0.51	(0.34, 0.77)	0.00122

rs12612236	Imputed	121459243	A	0.32	0.60	(0.44, 0.83)	0.00168
rs199672409	Imputed	121522007	CT	0.18	1.88	(1.26, 2.80)	0.00181
rs12996197	Imputed	121464285	A	0.33	0.60	(0.43, 0.83)	0.00198

D)

SNP	Source	Position	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value
rs2521933	Exome Beadchip	130457465	T	0.20	2.02	(1.41, 2.88)	1.13E-04
rs2521927	Imputed	130456688	A	0.20	2.03	(1.42, 2.91)	1.12E-04
rs2704538	Imputed	130403009	T	0.20	0.50	(0.34, 0.74)	5.03E-04
rs2521927	Imputed	130404123	T	0.20	0.51	(0.34, 0.75)	6.56E-04
rs2704538	Imputed	130401545	T	0.20	0.51	(0.35, 0.75)	7.38E-04

E)

SNP	Source	Position	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value
rs13156567	Exome Beadchip	57122376	A	0.11	2.37	(1.50, 3.74)	2.12E-04
rs72751331	Imputed	57125178	T	0.11	2.44	(1.53, 3.91)	1.95E-04
rs10487616	Imputed	57105894	G	0.10	2.48	(1.53, 4.01)	2.21E-04
rs1968422	Imputed	57138990	C	0.31	1.80	(1.31, 2.46)	2.44E-04
rs13169816	Imputed	57145564	T	0.11	2.46	(1.50, 4.02)	3.46E-04

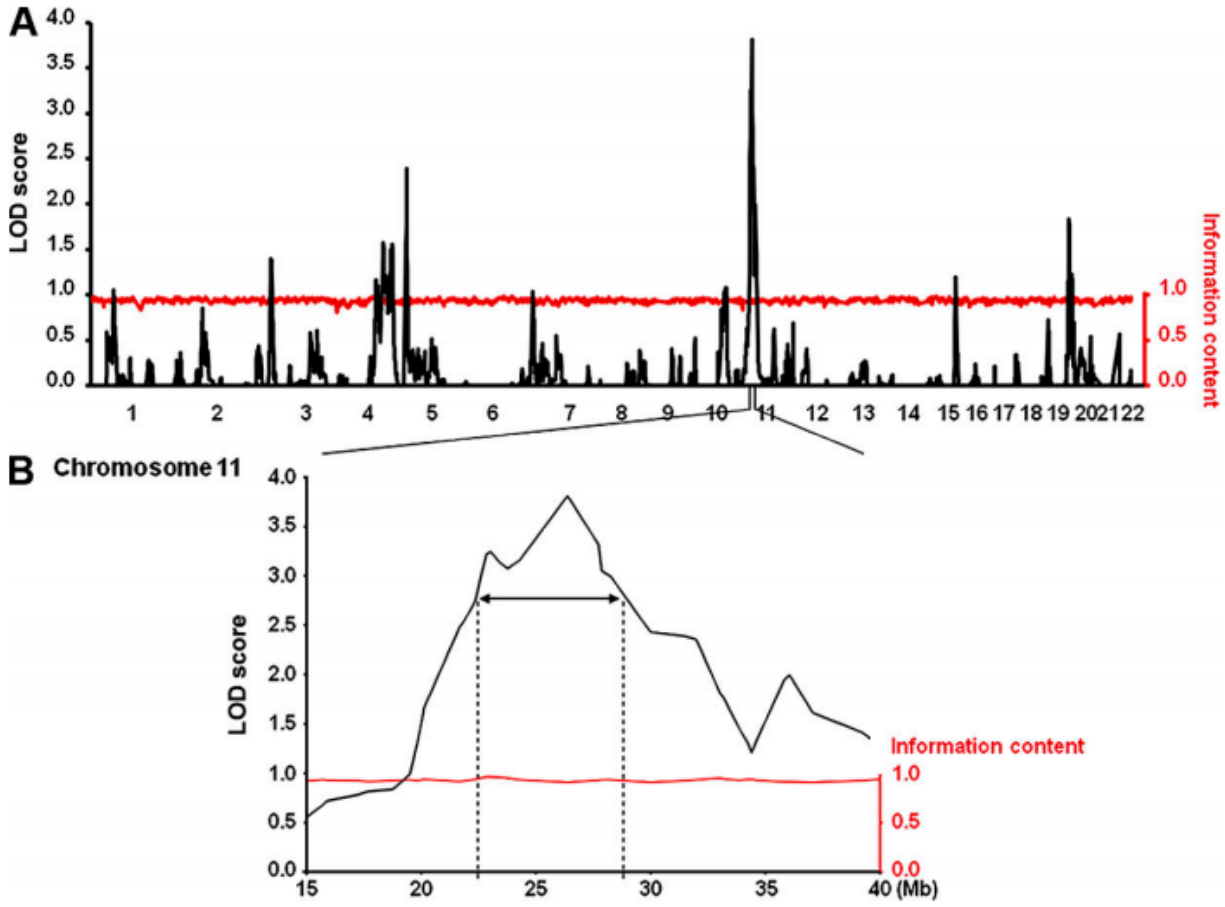
Logistic regression in the chromosome 11p14.1 region

A total of 714 variants from the Human Exome Beadchip passed quality control in the chromosome 11p14 region (22.35 to 28.82mb). SNP rs17234274 had the strongest association with a binary TST outcome in logistic regression analyses (OR = 0.58, 95% = CI 0.43-0.78, p value = 2.85×10^{-4}; Table 4-5B). This variant did not pass a Bonferroni-adjusted multiple testing criterion (threshold 7.00×10^{-5}). When the LD pruning method for adjusting the multiple testing threshold was applied, there were 402 independent SNPs in the region, and the new criterion became 1.24×10^{-4}. After imputation, rs10834029 was the variant with the most significant association in the region (OR = 1.80, 95% = CI 1.34-2.43, p value = 1.07×10^{-4}; Table 4-5B). The variant is significant when an LD-pruning based multiple testing criterion is used (LD pruning of imputed data also yielded 402 independent SNPs in the region). Importantly rs10834029 has the genomic position of 23306545, which roughly corresponds to one of the apices of LOD scores observed by Cobat et al. (Figure 4-2)[12].

Logistic regression in the chromosome 2q14 region

The most significant association for the Human Exome Beadchip variants in the chromosome 2q14 region (114.40 to 129.90mb) was rs4848637 (OR = 0.52, 95% = CI 0.35-0.77, p value = 0.0013; Table 4-5C). There were 1103 variants passing quality control in that region, and the most significant SNP did not pass a multiple testing adjusted criterion (threshold 4.53×10^{-5}). After imputation, rs4848637 was the variant with the most significant association in the region (OR = 0.51, 95% = CI 0.34-0.77, p value = 0.00122; Table 4-5C). The two SNPs are in perfect LD with each other, and the slight discrepancies in the reported odds ratio and p value are due to a missing observation for rs4848637 in our study cohorts.

Figure 4-2. Linkage map from Cobat et al using a binary TST outcome of A) the entire genome and B) the chromosome 11 region with the highest LOD score



Logistic regression in the chromosome 2q21-2q24 region

The most significant association for the available Human Exome Beadchip variants in the chromosome 2q21-q24 region (129.90 – 169.70mb) was rs2521933 (OR = 2.02, 95% = CI 1.41-2.88, p value = $1.13x10^{-4}$; Table 4-5D). There were 2731 variants passing quality control in the chromosome 2q21-q24 region, and the most significant SNP did not pass a multiple testing adjusted criterion (threshold $1.83x10^{-5}$). After imputation, rs2521927 was the variant with the most significant association in the region (OR = 2.03, 95% CI = 1.42-2.91, p value = $1.12x10^{-4}$; Table 4-5D). The most significant SNP from the Exome Beadchip and the most significant imputed SNP are in very high LD.

Logistic regression in the chromosome 5p13-5q22 region

The chromosome 5p13-q22 region does not contain rs877356, the variant associated with TST in Part A of this Chapter. The most significant association for the Human Exome Beadchip variants in this region (28.90 to 115.20mb) was rs13156567 (OR = 2.37, 95% = CI 1.50-3.74, p value = $2.12x10^{-4}$; Table 4-5E). There were 5683 variants passing quality control on chromosome 5p13-q22, and the most significant SNP did not pass a multiple testing adjusted criterion (threshold $8.80x10^{-6}$). After imputation, rs72751331 is the variant with the most significant association in the region (OR = 2.44, 95% CI = 1.53-3.91, p value = $1.95x10^{-4}$; Table 4-5E).

Discussion

We performed a fine mapping follow-up using an association approach in all genomic regions previously discovered in genome-wide linkage studies of TST. The SNP with the strongest association in the chromosome 11p14 region, rs10834029, overlapped with the second

highest LOD score observed in the original analysis by Cobat et al. The imputed variant was not significant when using a strict Bonferroni multiple testing correction, but became significant when a more natural threshold was used[115]. The variant is located in a gene desert, ~500mb away from the closest genes *GAS2* and *CCDC179*. *GAS2* encodes growth arrest-specific 2 protein that was originally identified in murine fibroblasts when cell cycle progression was stopped[155]. It is a caspase-3 substrate that regulates cell shape changes in apoptosis, and this association with cell death makes it a very good candidate for modulating MTB infection. Tumor necrosis factor alpha mediated apoptosis of MTB infected macrophages deprives the bacteria of the intracellular environment it needs for replication[156]. Several SNPs in the *GAS2* coding region associate with the phenotype at a 0.05 threshold (Appendix Table 4-16), but not at any of the multiple testing adjusted thresholds.

Variants with the most significant associations with TST outcomes in the other fine mapped regions are not in or near genes associated with immune function or tuberculosis. Variant rs4848637 in the 2q14 region is in an intron of *GLI2*, a zinc finger gene previously associated with cleft lip and cleft palate[157]. Variant rs2521933 in the 2q21-24 region is located in a gene desert. The only annotated gene within 0.5mb of the SNP is *Loc151121*, previously associated with alcohol abuse[158]. Variant rs13156567 in the 5p13-q22 region is closest to *PLK2*, a gene encoding a serine/threonine kinase that has been previously associated with multiple myeloma and breast cancer[159,160].

The size of each region identified in genome-wide linkage analyses was a major impediment in identifying the SNPs truly associating with the phenotype in fine mapping. The regions spanned up to 85 megabases and, consequently, the multiple testing adjusted thresholds were difficult to surpass. The large number of SNPs in those regions permitted us to find variants

significant at the standard validation criterion, p value of 0.05, but we were otherwise unable to discern type I errors from true findings. Our approach was tailored for finding resistance loci; therefore, the fact that neither study reported the direction of the observed effect may mean in some instances we were attempting to replicate susceptibility loci in a study population not suited to do so. HIV status is also an important consideration in attempting to validate findings from other studies in our Ugandan and Tanzanian cohorts. While Cobat et al. excluded these patients, and Stein et al. adjusted for HIV as a variable, our patients population was comprised entirely of HIV-seropositive adults[12,13,68]. This permitted us to find a resistance locus using a very small sample size for a genome-wide analysis, but it might not be ideal for replicating findings from the general patient population, where odds ratios are usually not as extreme.

CHAPTER V

EXAMINING MULTILOCUS INTERACTIONS IN CANDIDATE GENES FOR TUBERCULOSIS RESISTANCE

A. Multifactor Dimensionality Reduction Analysis of Active Tuberculosis Disease

Introduction

Tuberculosis (TB) has a significant heritable component[161]. The identification of genetic variants affecting disease risk has substantial implications in designing diagnostics, therapeutics, and vaccines. To date, most genetic studies of TB have used approaches associating individual variants to disease, assuming strong effect sizes of single genes[10,11,14,58]. Despite the use of dense genomic coverage and large sample sizes, most TB heritability remains unexplained. Recently, analyses of gene-gene interactions, which assume a complex genetic architecture to TB susceptibility, identified several important effects and demonstrated the potential of epistasis in explaining the missing heritability[162,163].

The immunological response to *Mycobacterium tuberculosis* (MTB) infection and resultant tuberculosis is complex. Numerous cells types, proteins, and signaling molecules are responsible for limiting symptoms, and the biological networks involved change during the course of disease[20]. To infect a naïve host, MTB must move through the upper respiratory system without being sequestered. If the MTB is trapped by macrophages in the upper respiratory system, the presence of commensal bacteria stimulates the macrophages through toll like receptors (TLR) to digest all phagocytosed bacteria including MTB[20]. However, once a lower respiratory tract infection is established, MTB enter macrophages in the alveolar spaces and use C-C motif ligand 2 (CCL2) signaling to recruit other macrophages to the region. MTB prevents

the binding of the phagosome in which it resides and the lysosome containing hydrolytic enzymes designed to degrade it; therefore, in the absence of commensal bacteria that stimulate macrophages, MTB survives[21]. The bacterium is thereby largely unrestricted by innate immunity early in the course of infection[21].

Subsequently, the host organism starts to mount an adaptive response. MTB are slowly digested by endosomal proteases and MTB peptides are loaded onto MHC class II molecules. TLR2 bind to antigen presenting cells and stimulate an IL-12 response which in turn activates the differentiation of T_H1 CD4 T cells[21]. Mature T_H1 cells mount an interferon gamma (IFN-γ) response and activate the binding of lysosomes with MTB containing endosomes in macrophages[21]. The recognition of MTB peptides also stimulates tumor necrosis factor (TNF) signaling, generating a localized inflammatory response that leads to the formation of granuloma consisting of infected macrophages walled off by surrounding T_H1 CD4 T cells. MTB survive within the granulomas even in the presence of widespread necrosis[22]; thereby maintaining a latent infection[21]. In 5-10% of patients, the immune system is unable to contain MTB with granulomas, the bacteria multiply, and patients progress to active TB disease shortly after infection, with significant caseation, cavitation and tissue destruction in the lungs[21]. The pattern of lung involvement in primary tuberculosis results in consolidation of the lower and middle lung lobes, along with hilar adenopathy, and very rarely cavitation.

Latent infections can be activated by immunosuppression caused by age, malnutrition, or HIV infection, among other factors. The depletion of CD4 cells seen in HIV patients prevents the containment of MTB in granulomas. Cavitation occurs readily in secondary tuberculosis, and it is generally localized in upper lobe apices of the lungs[21]. In immunosuppressed patients, MTB

more readily enters the systemic circulation, causing disseminated disease usually localizing to pleura, bones, and joints (miliary tuberculosis) or the CNS (tuberculosis meningitis)[21].

The complex signaling involved in the pathogenesis of tuberculosis along with the array of tissues that are exposed to the bacteria, especially in HIV-positive patients, make single variants unlikely to account for most of the heritability observed for this disease[161]. Here, we proposed a two-tiered approach for studying gene-gene interactions in TB. We identified candidate genes with a literature search, and separated available variants in these genes by chromosome to identify *cis* acting interactions. We then applied a filter to lower the number of included variants and studied interactions between all candidate genes.

Methods

Study Populations

The participants of the Household Contact Study, DarDar Women's Nutrition Study and the extended follow up of the DarDar vaccine trial used in the genome-wide association study of tuberculosis resistance in Chapter III were used to carry out this aim.

Candidate Gene Selection

We used a candidate gene approach selecting variants with a minor allele frequency > 0.05 in and near genes (+/- 100 kb of coding region) previously associated with tuberculosis disease. The candidate gene list for TB disease is comprised of the following: the major histocompatibility complex region *HLA-DR2*[15,37]; a proton-coupled divalent metal ion transporter *SLC11A1*[5,38]; a solute carrier associated with MTB infection *SLC6A3*[12,14]; a ligase in the ubiquitin proteasome pathway: *UBE3A*[65]; cytokines and chemokines used in macrophage-T_H1 cell signaling: *IFN-γ*[15], *IFNGR1*[63], *IL12B*[7,8], *IL12BR1*[64], *CCL2*[46], *IL-8*[9], *IL-1B*[47], *IL-10*[48,66], *IL-*

4[49], and *TNF-α*[66]; a mediator of pleiotropic effects of *IFN-γ*: *SP110*[44,45]; variants in immunomodulatory receptors: *PTX3*[51], *CR1*[52], *VDR*[51,53], *CD209*[51,54], *P2RX7*[55]; complement activating pattern recognition molecules *MBL2*[56], *TLR2*[57,58], *TLR4*[59], and their downstream effector *TIRAP*[60]. A total of 808 variants met the inclusion criteria.

Multifactor Dimensionality Reduction

We used Multifactor Dimensionality Reduction (MDR) software to study the association of gene-gene interactions with active tuberculosis[18,19]. Missing genotypes were imputed using the MDR data tool. All analyses were performed using the same random seed and up to 3 locus models of susceptibility, which allows the study of single variant effects along with 2 and 3 SNP interactions. The default setting assigning tied cells as cases and ten-fold cross-validation was used throughout.

We evaluated epistasis in the combined cohort, as well as the Ugandan and Tanzanian cohorts individually. Analyses of the combined cohort were adjusted for variable designating Tanzanian or Ugandan origin. Top testing accuracy and cross validation consistency results in each model are reported. The combined cohort model with the highest testing accuracy with a cross-validation consistency of at least 8/10 was forced into the Ugandan and Tanzanian cohorts to assess replication. Using MDRpt, we performed 1000-fold permutation tests on all 2 or 3 variant models from the combined cohort that replicated in both Ugandan and Tanzanian cohorts.

We first restricted the analyses to SNPs on the same chromosome. We then used the ReliefF filter to select the top 40 SNPs for MDR analyses of interactions between variants from all candidate genes. ReliefF uses a computational algorithm to filter data and generate a subset of most informative SNPs. Statistically significant 2 or 3 locus interactions in the combined cohort

were followed up with logistic regression analyses in STATA(v11.2)[73]. BioGRID(v3.3) was used to assess biological interactions between significant genes[164].

All analyses were carried out on two sets of SNPs. We first performed MDR on all available variants, including rs4921437, which associated with TB disease at a genome wide significant threshold, as shown in Chapter III. We then evaluated epistasis without significant main effect SNPs, i.e. removing rs4921437 from analyses.

Results

Single chromosome analyses

The two-locus interaction between rs4921437 and rs7737692 was the only higher order (>1 variant) model that met our criteria for follow up in the combined cohort analysis of variants from the same chromosome (Table 5-1A). The testing accuracy of this model was 0.63, and the cross-validation consistency was 10 out of 10. When we used cohort of origin as a covariate, instead of adjusting for it, the testing accuracy was 0.68, and the cross-validation consistency was 10 out of 10 (Appendix Table 5-1A). The rs4921437-rs7737692 interaction had a testing accuracy of 0.60 and a cross validation consistency of 10 out of 10 in the Ugandan cohort. In Tanzania, the testing accuracy was lower at 0.53; however, the cross-validation consistency was still 10 out of 10. The permutation test p value for the combined cohort analysis of the rs4921437 and rs7737692 interaction was <0.001.

Chromosome 5 analyses omitting rs4921437 did not produce results matching the MDR metric criteria for follow up permutation test analyses. None of the models had a testing accuracy >0.5, or cross-validation criteria > 4 out of 10, indicating the rs4921437 is the major factor for an effect either marginal or via an interaction (Appendix Table 5-2).

Table 5-1. MDR metrics of top loci and interactions associating with TB disease for variants on the same chromosome in A) the combined*, B) the Ugandan, and C) the Tanzanian cohort

A)

Variant Combinations	Chr.	Attribute Count	Balanced Accuracy Training	Balanced Accuracy Testing	Cross Validation Consistency
rs12622683	2	1	0.58	0.58	10/10
rs4680367	3	1	0.58	0.54	8/10
rs4921437	5	1	0.62	0.62	10/10
rs7737692, rs4921437	5	2	0.64	0.63	10/10
rs7095115	10	1	0.57	0.52	8/10
rs7496458	15	1	0.55	0.54	10/10
rs8069624	17	1	0.56	0.56	10/10

*adjusted for cohort of origin

B)

Variant Combinations	Chr.	Attribute Count	Balanced Accuracy Training	Balanced Accuracy Testing	Cross Validation Consistency
rs1032980	1	1	0.61	0.61	10/10
rs4921437	5	1	0.66	0.66	10/10
rs7737692, rs4921437	5	2	0.66	0.60	10/10
rs7852394	9	1	0.62	0.62	10/10
rs7095115	10	1	0.60	0.54	8/10
rs7917895, rs7474987	10	2	0.66	0.61	9/10
rs570984	12	1	0.61	0.57	8/10
rs8075846	17	1	0.61	0.61	10/10

C)

Variant Combinations	Chr.	Attribute Count	Balanced Accuracy Training	Balanced Accuracy Testing	Cross Validation Consistency
rs7540516	1	1	0.59	0.54	8/10
rs4921437	5	1	0.59	0.59	10/10
rs4921437, rs4262088	5	2	0.64	0.60	9/10
rs7737692, rs4921437	5	2	0.60	0.53	10/10
rs6916394	6	1	0.59	0.59	10/10
rs7034845	9	1	0.57	0.57	10/10
exm997700, rs4760673	12	2	0.67	0.67	10/10
rs12441992, rs2340634, rs7496458	15	3	0.66	0.56	8/10
rs8069624	17	1	0.58	0.53	8/10

Analyses of variants in all candidate genes

A three-locus interaction between rs4921437, rs2242656, and rs2940252 had the best MDR metrics from a dataset filtered to 40 variants by ReliefF. The cross-validation consistency was 10 out of 10 in the combined cohort and in Uganda and Tanzania separately. The testing accuracy was 0.65 in the combined cohort, 0.67 in Uganda, and 0.58 in Tanzania (Table 5-2A, Figure 5-1). When we used cohort of origin as a covariate, instead of adjusting for it, the testing accuracy was 0.66, and the cross-validation consistency was 10 out of 10 (Appendix Table 5-1B). Importantly, the testing accuracy for the 3-locus interaction performed better than the single locus rs4921437 model in the combined cohort and in Uganda (Table 5-2B). The permutation test p value for the top three-locus model in the combined cohort was estimated to be <0.001.

Analyses omitting rs4921437 did not produce results matching the MDR metric criteria for follow up permutation test analyses. None of the Tanzanian nor combined cohort models had a testing accuracy >0.56, or cross-validation criteria > 4 out of 10 (Appendix Table 5-3). A two-locus model, rs3129943 and rs12788021, had a testing accuracy of 0.70 and cross-validation consistency of 10 out of 10 in the Ugandan cohort. A three-locus model, rs6793453, rs1265761, and rs12788021, had a testing accuracy of 0.68 but a cross-validation consistency of 4 out of 10; therefore it was not generalizable.

Table 5-2. MDR metrics of the top model associating with TB disease using A) variants in or near *IL12B* (rs4921437), *TNF-α* (rs2242656), and *CR1* (rs2940252), and B) the corresponding metrics for the top single SNP model of rs4921437 alone

A)

Cohort	Testing balanced accuracy	Cross-validation consistency
Combined*	0.65	10/10
Uganda	0.67	10/10
Tanzania	0.58	10/10

*****adjusted for cohort of origin

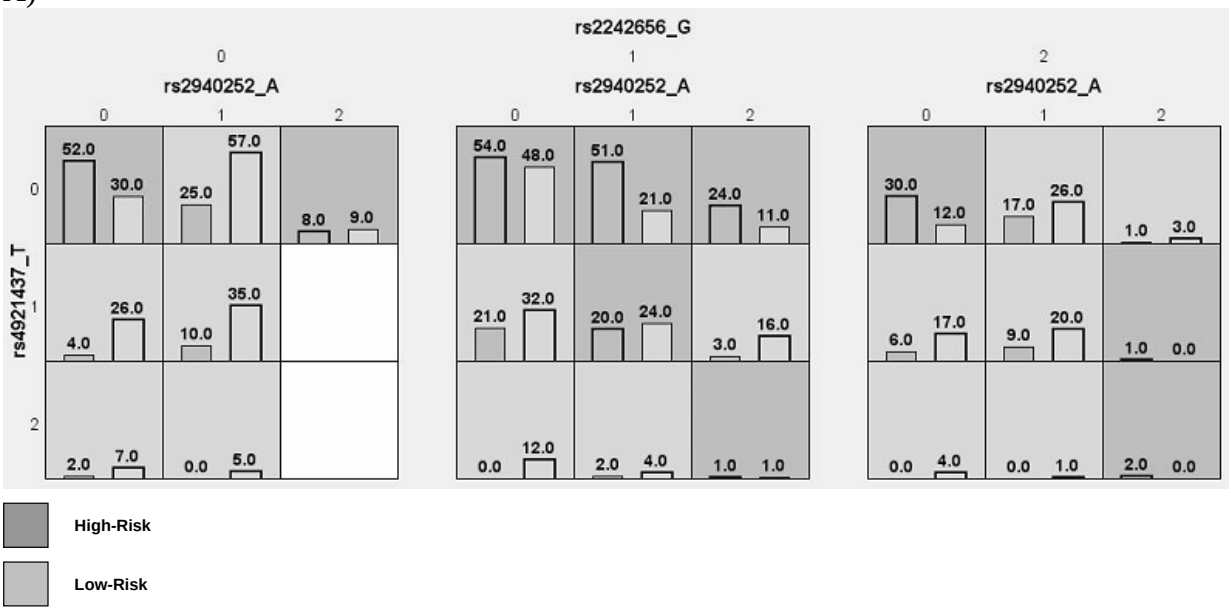
B)

Cohort	Testing balanced accuracy	Cross-validation consistency
Combined*	0.62	10/10
Uganda	0.66	10/10
Tanzania	0.59	10/10

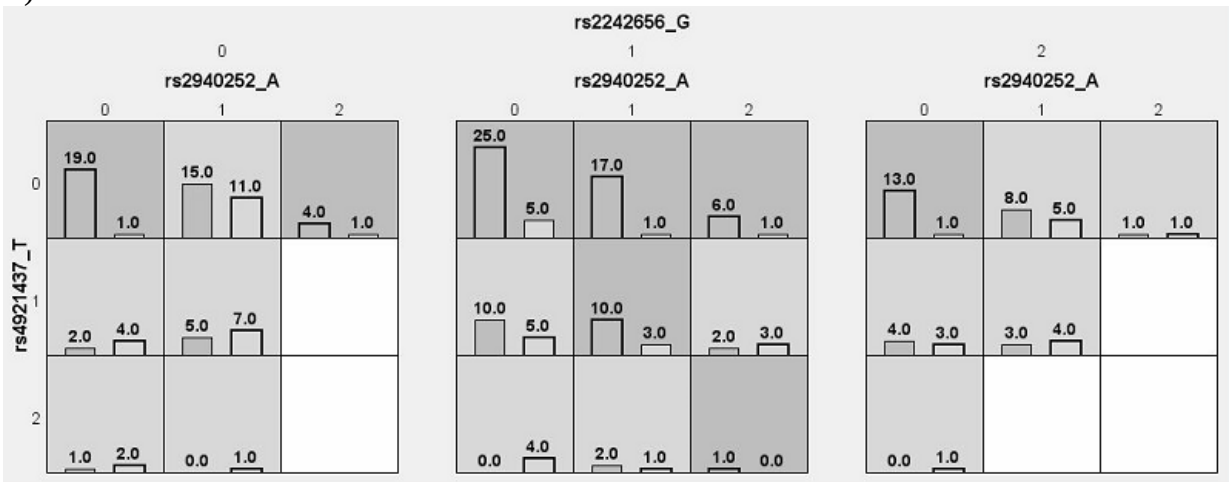
*****adjusted for cohort of origin

Figure 5-1. Three locus MDR model of variants in or near *IL12B* (rs4921437), *TNF-α* (rs2242656), and *CR1* (rs2940252) in A) the combined cohort, B) the Ugandan cohort, and C) the Tanzanian cohort; each cell is labeled with counts of the listed allele and represents a three variant haplotype that is designated as "high risk" or "low risk" based on a hypothetical distribution of cases *(left bar in cell)* and controls *(right bar in cell)*

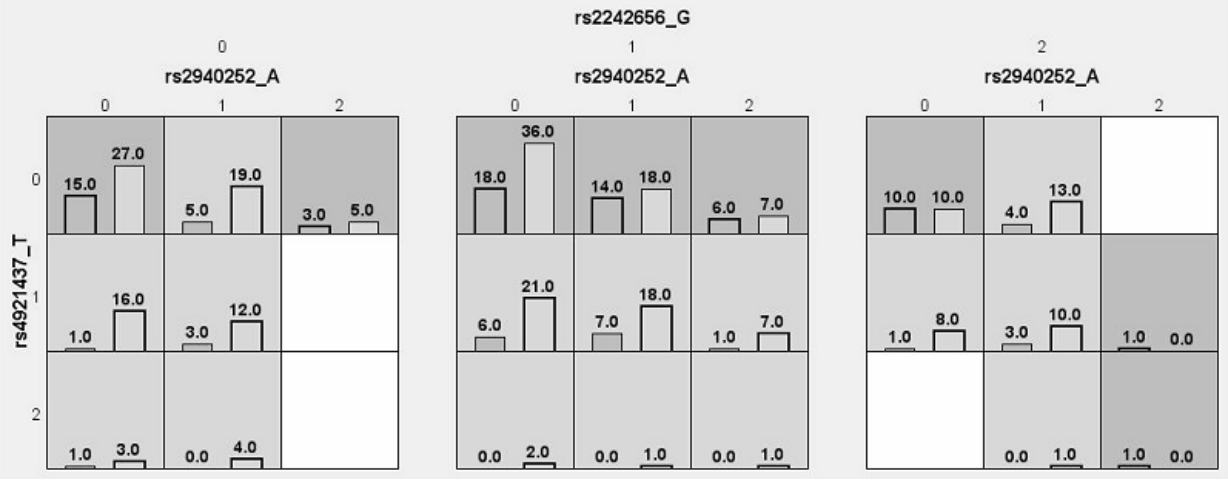
A)



B)



C)


rs2242656_G
0
rs2940252_A
0 1 2
rs4921437_T
0
15.0 27.0
5.0 19.0
3.0 5.0
1
1.0 16.0
3.0 12.0
2
1.0 3.0
0.0 4.0
1
rs2940252_A
0 1 2
0
18.0 36.0
14.0 18.0
6.0 7.0
1
6.0 21.0
7.0 18.0
1.0 7.0
2
0.0 2.0
0.0 1.0
0.0 1.0
2
rs2940252_A
0 1 2
0
10.0 10.0
4.0 13.0
1
1.0 8.0
3.0 10.0
1.0 0.0
2
0.0 1.0
1.0 0.0

Logistic Regression

In the combined cohort, logistic regression of TB disease status was modeled with cohort of origin, rs4921437, rs7737692, and an interaction between the SNPs. Cohort of origin and variant rs4921437 were significant in this analysis (p values = $1.91x10^{-21}$ and $3.32x10^{-5}$, respectively), while rs7737692 and the interaction term were not (p values = 0.27 and 0.24, respectively).

To follow up on the top MDR three-locus interaction model in the combined cohort, logistic regression of TB disease status was modeled with cohort of origin, rs4921437, rs2242656, rs2940252, all possible two locus interactions and an interaction between the three SNPs. Cohort of origin and variant rs4921437 were significant in this analysis (p values $7.51x10^{-20}$ and 0.009, respectively), while variants rs2242656, rs2940252, and two-locus interactions rs2940252*rs4921437, rs2940252*rs2242656, rs4921437*rs2242656 were not (p values = 0.23, 0.79, 0.69, 0.48, and 0.37, respectively). The three-locus interaction term was nearly significant (Odds ratio = 2.04, 95% = CI 0.90-4.63, p value = 0.087). The minor, protective allele of rs4921437 described in Chapter III had a similar resistance effect in this model (Odds ratio = 0.37, 95% CI = 0.18-0.78).

Discussion

The most significant three locus model was found for variants from all candidate genes, namely rs4921437 near *IL12B* on chromosome 5, rs2242656 near *TNF-α* on chromosome 6, and rs2940252 near *CR1* on chromosome 1. The functions of IL12 and TNF-α are complementary and in some instances redundant. The product of *IL12B* is a subunit of cytokines IL12 and IL23, secreted by macrophages and dendritic cells. IL12 causes differentiation of naïve T cells into T_H1 cells and stimulates interferon-gamma production from T cells and natural killer cells[103,104]. In the initial response to MTB infection, TNF-α potentiates the inflammatory response by recruiting macrophages and CD4+ T cells to the site of the infection[165]. TNF-α is also essential in the long-term containment of MTB, as it is necessary in the formation and maintenance of the granuloma[165]. TNF-α signaling is used in the granuloma to induce apoptosis, which prohibitive to bacterial growth[166]. Since there is some redundancy in TNF-α and IL12 function, it is possible that diminished production of one could be compensated by the overexpression of the other leading to the observed interaction. Complement receptor 1, the product of *CR1*, is important in the phagocytosis of MTB by macrophages[167]. It is, therefore, essential in both the establishment of an MTB infection and in the initiation of an innate immune response. Antibodies against CR1 substantially decrease MTB macrophage adherence[167].

The *IL12B, TNF-α, CR1* interaction was identified in the analyses of the combined cohort; separate analyses of the Ugandan and Tanzanian participants demonstrate that the cross validation and testing accuracy replicated, which is rare in studies of epistasis. The three way interaction term was significant with a p value <0.001 in the MDR permutation testing and was nearly significant in logistic regression analyses. There are several possible reasons for this disparity. MDR is a model free, non-parametric approach to studying epistasis[18]. It uses

combinatorial-partitioning to reduce the genetic predictor space to a one dimensional variable, which allows for its use in small study populations without loss of power due to increasing the degrees of freedom[18]. Since a logistic regression model for a three-way interaction requires the inclusion of each possible two-way interaction, the number of degrees of freedom increases. Furthermore, empty cells, which are frequent in the study of higher-order interactions, inflate the standard errors during this estimation, and statistical significance becomes difficult to attain, especially in smaller studies[168].

The most significant two-locus interaction was found for variants on chromosome 5 only, rs4921437 near *IL12B* and rs7737692 near *SLC6A3*. The genes are ~157 megabases apart and the proteins they encode for appear functionally disparate, as no interactions have been reported in BioGRID. *SLC6A3* encodes a transporter for the neurotransmitter dopamine. While the gene has been previously associated with MTB infection, no mechanism of action with respect to TB disease has been proposed[12,14]. Since the *SLC6A3* product is a transmembrane protein involved in the removal of dopamine from the synaptic cleft, any direct interaction with IL12 is unlikely given current knowledge. This dopamine transporter has also been associated with addictive behavior, including tobacco use, and neoplasms in the lungs[169,170]. The testing accuracy of the identified *IL12-SLC6A3* interaction was very good in the combined cohort, and the Ugandan cohort; however it was slightly higher than what is expected by chance in the Tanzanian cohort; therefore this might not be a true replication. The interaction term between rs4921437 and rs7737692 was not statistically significant in follow up logistic regression analyses.

When we removed the SNP with the strongest main effect, rs4921437, none of the models in the combined cohort produced testing accuracy >0.60 or cross validation consistency >4 out of 10. Of note, the two-locus model rs3129943 and rs12788021 had good MDR metrics in

the Ugandan cohort, but did not replicate in Tanzania. Since epistasis studies often do not replicate, we evaluated the possible biological consequences of this interaction. Variant rs3129943 is in the *HLA-DRA* region on chromosome 6 which encodes for the DR alpha chain of the HLA class II major histocompatibility complex cell surface receptor. MHC class II receptors are present on macrophages and other antigen presenting cells, and they present foreign peptides to CD4+ helper T-cells which eventually leads to the production of antibodies against the peptide protein it is derived from[22]. HLA-DR is a heterodimer comprised of an alpha and beta subunit. The beta subunit has multiple variants, but the alpha protein is invariable in humans[64]. Polymorphisms in this receptor play a role in dictating the specificity, extent and type of T cell response to MTB[171]. The other variant in this interaction, rs12788021, is near *TIRAP* on chromosome 11. TIRAP is an adapter molecule downstream of Toll-like receptors, used in the MTB context to activate a cytokine response leading to the fusion of an endosome and a lysosome[20]. TLRs are important in eliminating MTB in the upper respiratory system because they are activated by the presence of commensal bacteria[20]. Once MTB is in the alveolar spaces, TLRs contain chronic MTB infection[172,173]. The interacting variants in this analysis are both involved in the chronic containment of tuberculosis. This aspect of immune response to MTB is particularly important in HIV-positive patients.

In summary we found significant two and three-variant interactions in HIV-positive tuberculosis cases and controls from Uganda and Tanzania. Importantly, the three-locus interaction of variants near *IL12B*, near *TNF-α,* and near *CR1* was consistent in both cohorts, which is rare in studies of epistasis. Functional follow-up analyses are necessary to evaluate the effects we observed with MDR.

B. Multifactor Dimensionality Reduction Analysis of Host Resistance to *Mycobacterium tuberculosis* Infection

Introduction

One in three people have been infected with *Mycobacterium tuberculosis* (MTB)[2,119]. Many individuals who remain uninfected live in highly endemic areas, virtually guaranteeing exposure to the bacteria. The tuberculin skin test (TST) is the standard for assessing exposure to MTB. TST measures delayed type hypersensitivity to a purified protein derivative; thereby assessing whether a person has immune memory from an MTB exposure[21]. There are three major ways to remain TST negative given an exposure to *M. tuberculosis*: MTB can be inhaled but mechanically prevented from seeding the lungs, it can seed the lungs but be cleared before immune memory is invoked, or it can establish a latent infection but host immunosuppression can lead to a false negative result[21]. We hypothesize that the clearance of MTB either before or after it reaches the alveolar spaces of lungs has a genetic component with a complex architecture. To date, studies of MTB infection genetics have predominantly used single SNP association approaches or genome-wide linkage, assuming detectable single locus effects[12,13]. Identifying the genetic variants associated with MTB infection predisposition/resistance is of great global health importance.

The immune response to an MTB infection is complex; therefore, it is likely that single variants cannot account for most of the heritability observed in the MTB infection phenotype. We used two recently concluded prospective cohorts of tuberculosis from Uganda and Tanzania to evaluate epistatic effects on MTB infection. Study design and available immunological data were leveraged to assure that all patients in this analysis have been exposed to MTB, and that immunosuppression did not confound the results with false negative data. We used a two-tiered approach leveraging candidate genes for studying epistasis. First, we studied effects for all

variants on the same chromosome. We then applied a filter to lower the number of included variants and studied interactions between all candidate genes within and among chromosomes. The first step served to reduce the number of potentially interacting variants along a chromosome, effectively identifying key SNPs on each, and the second to reduce the total number of total comparisons made.

Methods

Study Populations

The participants of the Household Contact Study and the extended follow up of the DarDar vaccine trial used in the genome-wide association study of resistance to *Mycobacterium tuberculosis* infection in Chapter IV were used to carry out this aim. Individuals who reported prior tuberculosis disease but had a negative TST result were removed from this analysis.

Candidate Gene Selection

We used a candidate gene approach selecting variants with a minor allele frequency > 0.05 in genes (+/- 100 kb of coding region) and regions previously associated with MTB infection. The candidate gene/region list is comprised of the dopamine neurotrasporter *SLC6A3*, the proton-coupled divalent cation pump *SLC11A1*, and regions on chromosomes 2q14, 2q21-q24, 5p13-q22, and 11p14. The *SLC25A48/IL9* region associated with MTB infection in Chapter IV Part A was included in this analysis.

Multifactor Dimensionality Reduction

Binary tuberculin skin test status (induration size < versus ≥ 5mm) was used as the outcome in this study. The MDR modeling and variant selection criteria were the same as those described in Chapter V Part A. Follow up logistic regression analyses were also carried out in

STATA(v11.2)[73], and BioGRID(v3.3)[164] was used to assess prior evidence of interactions, as described in Part A.

Results

Single chromosome analyses

The two-locus interaction between rs931709 and rs877356 was the only multilocus model (>1 variant) that met our criteria for follow up in the combined cohort analysis of variants from the same chromosome (Table 5-3A, Figure 5-2). The testing accuracy of this model was 0.65, and the cross-validation consistency was 10 out of 10. When we used cohort of origin as a covariate, instead of adjusting for it, the testing accuracy was 0.66, and the cross-validation consistency was 10 out of 10 (Appendix Table 5-4A). In the Ugandan cohort the rs931709-rs877356 model had a testing accuracy of 0.65 and in Tanzania it had a testing accuracy of 0.64. The cross-validation consistency was at 10 out of 10 in both cohorts. Examining the high and low risk two-locus genotypes, it is apparent that there is substantial consistency across these locations. The permutation test p value for the top two-locus model in the combined cohort was estimated to be 0.003-0.004.

Chromosome 5 analyses omitting rs877356 did not produce two-locus results with a cross validation consistency greater or equal to 8 out of 10 (Appendix Table 5-5A). An interaction between rs12515850, rs981883, and rs17169129 had a testing accuracy of 0.66 and a cross-validation consistency of 9 out of 10 in the combined cohort (Appendix Table 5-5B). However, the model did not meet the replication criteria in the Ugandan and Tanzanian cohorts when analyzed separately.

Table 5-3. MDR metrics of top loci and interactions associating with MTB infection for variants on the same chromosome in A) the combined*, B) the Ugandan, and C) the Tanzanian cohort

A)

Variant Combinations	Chr.	Attribute Count	Balanced Accuracy Training	Balanced Accuracy Testing	Cross Validation Consistency
rs2521933	2	1	0.59	0.59	10/10
rs877356	5	1	0.64	0.64	10/10
rs931709,rs877356	5	2	0.65	0.65	10/10
rs10833965	11	1	0.59	0.59	10/10

*****adjusted for cohort of origin

B)

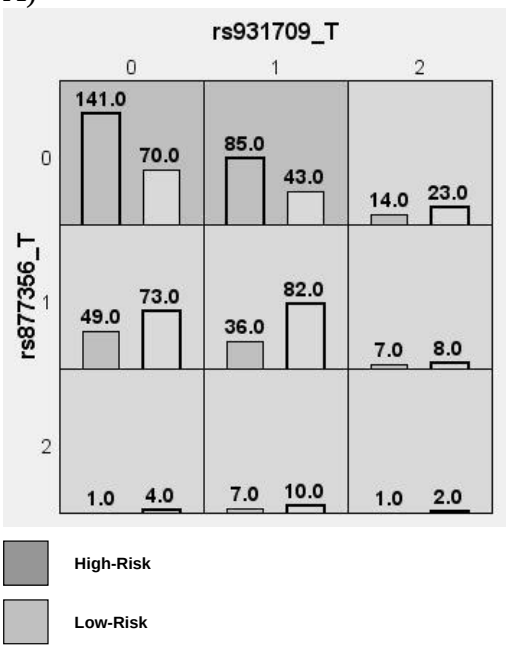
Variant Combinations	Chr.	Attribute Count	Balanced Accuracy Training	Balanced Accuracy Testing	Cross Validation Consistency
rs2521933	2	1	0.61	0.58	9/10
rs877356	5	1	0.65	0.65	10/10
rs931709,rs877356	5	2	0.67	0.65	10/10
rs1835874,rs877356	5	2	0.67	0.63	9/10
rs10833965,rs922858	11	2	0.70	0.61	8/10
rs10833965	11	1	0.65	0.62	9/10

C)

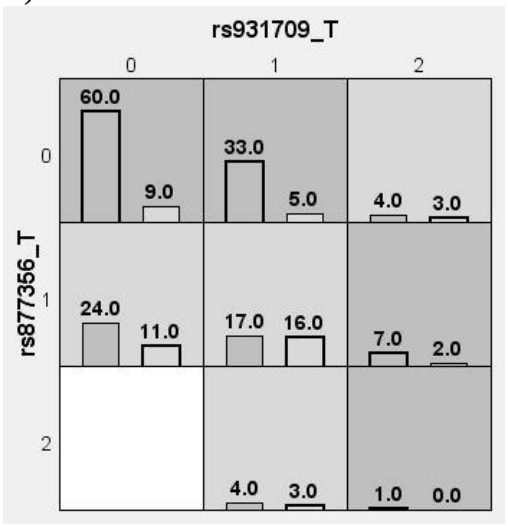
Variant Combinations	Chr.	Attribute Count	Balanced Accuracy Training	Balanced Accuracy Testing	Cross Validation Consistency
rs4848628	2	1	0.61	0.59	9/10
rs877356	5	1	0.62	0.62	10/10
rs1835874,rs877356	5	2	0.67	0.63	9/10
rs931709,rs877356	5	2	0.64	0.64	10/10
rs17306503	11	1	0.56	0.51	8/10

Figure 5-2. Two locus MDR model of variants associating with MTB infection in or near *IL9* (rs877356), and *SLC6A3* (rs931709) in A) the combined cohort, B) the Ugandan cohort, and C) the Tanzanian cohort; each cell is labeled with counts of the listed allele and represents a two variant haplotype that is designated as "high risk" or "low risk" based on a hypothetical distribution of cases *(left bar in cell)* and controls *(right bar in cell)*

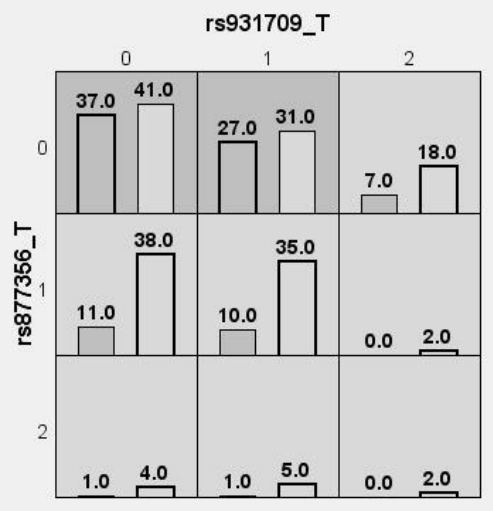
A)



B)



C)


rs931709_T
0
1
2
rs877356_T
0
1
2
37.0
41.0
27.0
31.0
7.0
18.0
11.0
38.0
10.0
35.0
0.0
2.0
1.0
4.0
1.0
5.0
0.0
2.0

Analyses of variants in all candidate genes

A three-variant interaction between SNPs rs877356, rs2521933, and rs17597967 had the best MDR metrics in analyses of all candidate genes (Table 5-4, Figure 5-3). The testing accuracy was 0.68 in the combined cohort, 0.73 in Uganda, and 0.60 in Tanzania. Cross-validation consistency was 10 out of 10 in each cohort individually and in the combined analysis. The permutation test p value for this three-locus model in the combined cohort was estimated to be <0.001. When we used cohort of origin as a covariate in the combined cohort analysis, instead of adjusting for it, the testing accuracy was 0.67, and the cross-validation consistency was 10 out of 10 (Appendix Table 5-4B).

Analyses omitting rs877356 in the combined cohort did not result in any two or three-locus model meeting the prerequisite MDR metric criteria for permutation test follow up. A two-locus model rs6865443 and rs10833965 had a testing accuracy of 0.65 and a cross-validation of 8 out of 10 in the Tanzanian cohort, but similar effects were not observed in the Ugandan cohort (Appendix Table 5-6).

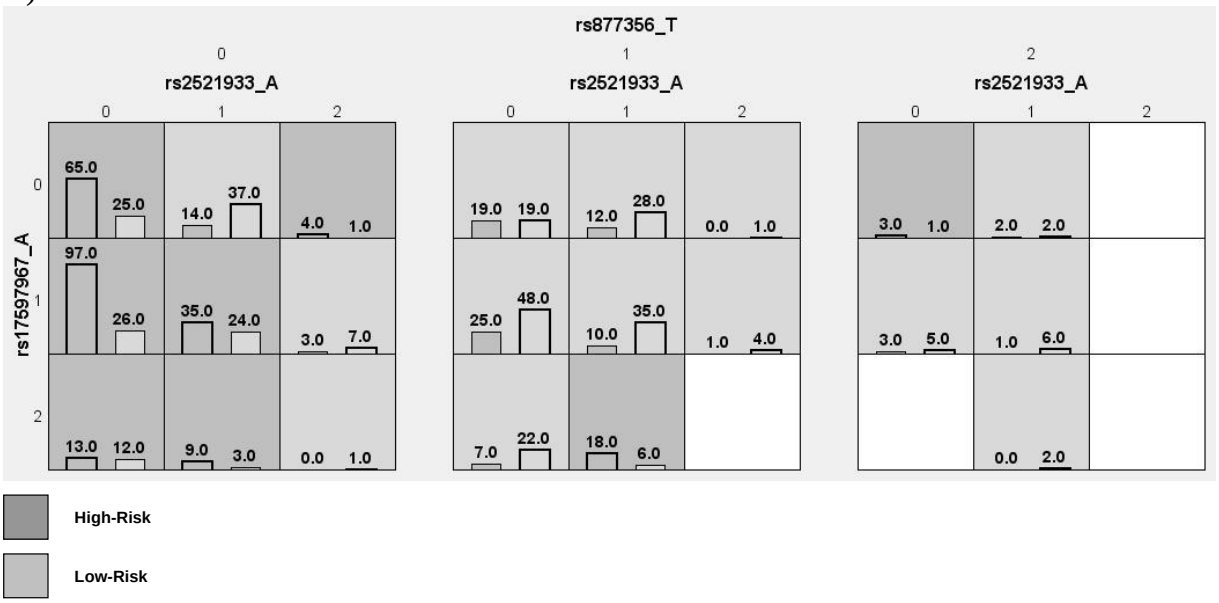
Table 5-4. MDR metrics of the top model using all candidate gene variants associating with MTB infection in or near *IL9* (rs877356), *RAB6C* (rs2521933), and *SLC6A3* (rs17597967)

Cohort	Training balanced accuracy	Testing balanced accuracy	Cross-validation consistency
Combined*	0.70	0.68	10/10
Uganda	0.74	0.73	10/10
Tanzania	0.68	0.60	10/10

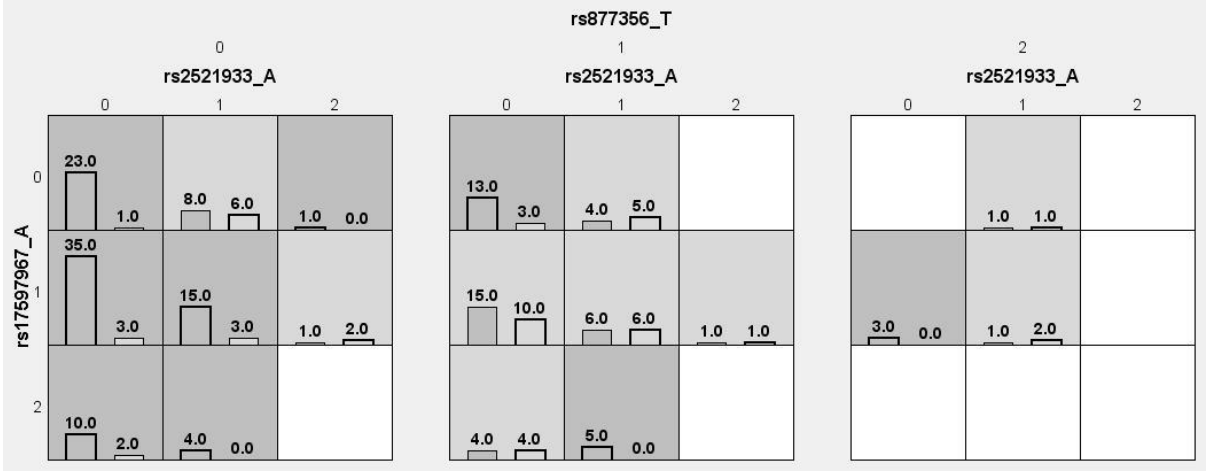
*****adjusted for cohort of origin

Figure 5-3. Three locus MDR model of variants associating with MTB infection in or near *IL9* (rs877356), *RAB6C* (rs2521933), and *SLC6A3* (rs17597967) in A) the combined cohort, B) the Ugandan cohort, and C) the Tanzanian cohort; each cell is labeled with counts of the listed allele and represents a three variant haplotype that is designated as "high risk" or "low risk" based on a hypothetical distribution of cases *(left bar in cell)* and controls *(right bar in cell)*

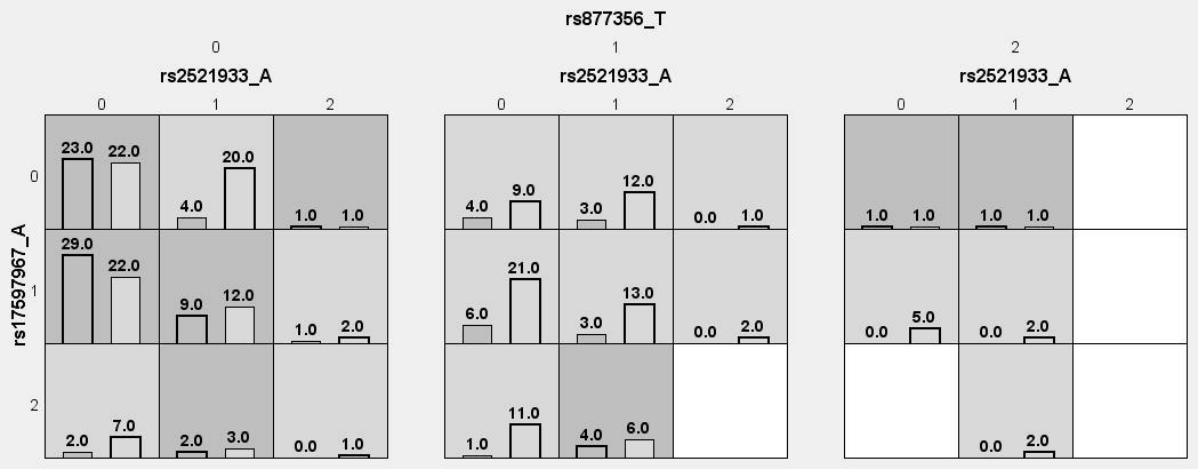
A)



B)



C)


rs877356_T
0
rs2521933_A
0 1 2
rs17597967_A
0
23.0 22.0
4.0 20.0
1.0 1.0
1
29.0 22.0
9.0 12.0
1.0 2.0
2
2.0 7.0
2.0 3.0
0.0 1.0
1
rs2521933_A
0 1 2
0
4.0 9.0
3.0 12.0
0.0 1.0
1
6.0 21.0
3.0 13.0
0.0 2.0
2
1.0 11.0
4.0 6.0
2
rs2521933_A
0 1 2
0
1.0 1.0
1.0 1.0
1
0.0 5.0
0.0 2.0
2
0.0 2.0

Logistic Regression

In the combined cohort, logistic regression of MTB infection status was modeled with cohort of origin, rs877356, rs931709, and an interaction between the SNPs. Variants rs877356, rs931709 and cohort of origin were significant in this analysis (p values = $2.55x10^{-5}$, 0.042 and $1.91x10^{-21}$, respectively), while the interaction term was not (p value = 0.24).

To follow up on the top MDR three-locus interaction model in the combined cohort, logistic regression of MTB infection status was modeled with cohort of origin, variants rs877356, rs2521933, and rs17597967, all possible two-variant interaction terms and an interaction between the three SNPs. Cohort of origin, variant rs2521933, and the two-SNP interaction rs2521933*rs17597967 were significant in this analysis (p values = $1.64x10^{-17}$, $2.43x10^{-4}$, and 0.044 respectively), while variants rs17597967, rs877356, two-locus interactions rs2521933*rs877356, rs877356*rs17597967, and the three-way interaction term were not (p values = 0.54, 0.15, 0.98, 0.12, and 0.44, respectively).

Discussion

In analyses of variants on the same chromosome, the two-locus interaction between rs931709 and rs877356 had the best MDR metrics and a permutation p value <0.05. Variant rs931709 is near *SLC6A3*, a gene encoding a dopamine transporter previously associated with the phenotype[12,14]. No mechanism was proposed for how *SLC6A3* affects MTB infection, although its deletion in mice has been associated with diminished delayed type hypersensitivity[174]. Variant rs877356 is in the newly discovered *IL9* region, described in detail in Chapter IV Part A. Briefly, we posit that this polymorphism is associated with bronchial hyperresponsiveness. The over-reactive airway allows for MTB to be sequestered in the upper airway where microbicidal macrophages kill the bacteria before it can seed the alveolar spaces and multiply[20]. The lack of mechanistic information on the role of *SLC6A3* prevents us from making an informed speculation about the nature of this interaction.

In analyses of all available variants, a three-locus interaction between rs877356 near *IL9* on chromosome 5, rs2521933 in the 2q21 region, and rs17597967 near *SLC6A3* had the most consistent results. The roles of *IL9* and *SLC6A3* have been discussed above. The closest gene to rs2521933 is *RAB6C* (~0.5mb away), which encodes for RAS-related protein Rab6C. This protein has been shown to interact with p53 and generate a pro-apoptotic phenotype in human cells[175]. While necrosis of macrophages furthers the spread of MTB, the apoptotic pathway limits the spread of the bacteria making *RAB6C* a possible candidate for preventing infection.

Both the two and three-locus interactions identified here were consistent between the Ugandan and Tanzanian cohorts. Logistic regression analyses failed to validate these findings; however, this is expected considering the small sample size used in this study. A major obstacle in evaluating the potential effects of these interactions is the lack of well-defined gene targets in

the prior studies. The known loci associating with MTB infection are not as well defined as those for TB disease; therefore biological interpretation of significant interactions is more speculative. Even though biological mechanisms cannot be completely understood from these results, they suggest that preventing MTB infection occurs both on a macro level, with airway inflammation preventing the bacteria from reaching the lower lungs, and a micro level, with stimulation of genes which induce apoptosis of the infected macrophages.

CHAPTER VI

GENETICS OF SRL 172 VACCINE RESPONSE

Introduction

The SRL 172 vaccine is the only placebo-controlled vaccine to show efficacy in preventing adult onset tuberculosis in HIV-positive patients[17]. In patients with a childhood BCG vaccination, the SRL 172 vaccine was shown to provide a 39% protection from definite tuberculosis. Details of the vaccine trial, patient population and diagnostic criteria were reported in von Reyn et al[17]. Briefly, the vaccine trial recruited HIV-positive adults from Dar es Salaam, Tanzania, with CD4 cell counts >200 cells/mm^3. Patients were randomly assigned into a placebo (n = 1007) and vaccine (n = 1006) arm, and were followed up every 3 months. Diagnoses of active tuberculosis were assessed via a thorough evaluation comprised of a physical examination, a chest radiograph, three sputum samples for staining and culture, as well as an automated mycobacterial blood culture. The biological determinants of the effectiveness of this 5 dose booster vaccination are still unclear.

We hypothesized that human genetic variants influence the efficacy of the inactivated SRL 172 whole cell vaccine in preventing definite tuberculosis in an HIV-positive population living in an area endemic for *Mycobacterium tuberculosis.*

Methods

Study population

The original DarDar vaccine trail was concluded in 2008. At the time of our project design, surveillance of a 1535 patient extended follow up group continued. The number of patients in the DarDar extended follow up cohort who received the full regimen of SRL 172 vaccine was 761, 17 of whom developed definite tuberculosis, and 50 who developed definite or probable tuberculosis. *A priori* power analyses were carried out in Quanto[72].

Statistical analyses

The number of cases was prohibitive of a genome-wide association approach. We proposed a candidate gene approach, using the variants described in Chapter V, to assess whether genetic variants in the *M. tuberculosis* associated genes influence the efficacy of the vaccine. Using subjects who received the full regimen of 5 vaccine doses, we proposed logistic regression analyses to test for genetic differences between those who did and did not develop definite tuberculosis during the trial. Age, gender, and principal components calculated from available genetic data were proposed as relevant covariates. Linear regression analyses using the same independent variables, but enrollment level corrected IFN-γ levels as the dependent variable were proposed as a secondary outcome.

Results

A summary of the *a priori* power analyses, assuming recruitment of all available cases of tuberculosis in the vaccine arm, is presented in Table 6-1A. Prior to recruitment, we were underpowered to pursue the aim of evaluating the genetic determinants of the effectiveness of the SRL 172 booster vaccine on the primary outcome of the DarDar vaccine trial, definite tuberculosis, as it required an odds ratio >5. We were, however, adequately powered to study the secondary DarDar vaccine trial outcome, a combined probable and definite tuberculosis outcome variable. We deemed an odds ratio of 2.5 reasonable enough to pursue this aim, especially since data were available and it did not require any additional clinical tests. This was the first study of the genetics of TB vaccine response; therefore we did not have an expected estimate of effect size in this study.

The recruitment of DarDar extended follow up participants into this study occurred between August and December of 2013, five years after the conclusion of the vaccine trial. We recruited 151 patients from the vaccine arm of the DarDar trial, and 153 patients from the placebo arm. Importantly, only 4 of the recruited patients who finished the 5 dose regimen of the SRL 172 vaccine had definite tuberculosis during the trial. The case number increases to 14 if the phenotype is adjusted to include patients who developed either definite or probable tuberculosis. Following recruitment, we carried out a power analysis to determine the effect size needed to identify a single SNP association (Table 6-1 B). We found that an odds ratio >5 would have to be present for either definite tuberculosis, or definite/probable tuberculosis. A multiple testing adjustment would further lower the power of this approach.

Table 6-1. Odds ratio estimates for SRL 172 vaccinogenetics assuming an additive genetic model, an alpha level of 0.05, a minor allele frequency of 0.2, and 80% power in an A) *a priori* and B) post recruitment power analyses

A)

Case definition	Controls	Cases	Power	Alpha	Effect Size
Definite TB	761	17	0.8	0.05	>5
Definite or Probable TB	761	50	0.8	0.05	2.5

B)

Case definition	Controls	Cases	Power	Alpha	Effect Size
Definite TB	761	4	0.8	0.05	>5
Definite or Probable TB	761	14	0.8	0.05	>5

Discussion

We were unable to recruit enough cases in the vaccine arm of the DarDar trial to carry out this aim. The possible reasons for loss to follow up are manifold. Patients could have moved away from Dar es Salaam and a return to clinic was no longer feasible. Available contact information for patients or family members might have become obsolete. Given their HIV status, it is also possible that participants succumbed to tuberculosis or other AIDS diseases.

We decided not to perform any association analyses for this aim due to the limitations of statistical comparisons with the available sample. Furthermore, even in the unlikely event that we found a SNP with an odds ratio large enough to be significant at a multiple-testing corrected level, we did not believe that a strong enough case about the importance of such an association could be made given either only 4 or 14 cases.

CHAPTER VII

CONCLUSIONS AND FUTURE DIRECTIONS

A. Summary and Significance

Mycobacterium tuberculosis is the second leading infectious cause of death from a single agent[1,2]. While the annual incidence of tuberculosis has remained constant in recent years, the burden of disease is very high and increases in the number of multidrug resistant tuberculosis necessitate novel pharmacological solutions[1,2]. Studies of genetic determinants of both tuberculosis disease and MTB infection provide promising drug and vaccine targets. This is especially important as evolution of MTB renders current treatment modalities obsolete.

We proposed to study the genetics of an extreme TB resistance phenotype, to both disease and infection, as opposed to the more standard approaches tailored towards finding loci associated with susceptibility. We presented a novel approach to carrying out genome-wide association studies, also applicable to whole genome sequencing studies, where the use of an extreme resistance phenotype allowed us to identify large effect sizes and attain genome-wide significance in a relatively small sample size. The use of HIV status as a central feature of our hypothesis as opposed to a confounder or exclusion criterion is also unique to these studies.

In Chapter III, we hypothesized that immunocompromised individuals living in MTB hyperendemic areas who do not develop TB disease represent an extreme resistance phenotype. Within our HIV-positive cohorts we found a genome-wide significant association of a common variant, rs4921437, with TB resistance in the *IL12B* region previously associated with susceptibility to disease. The effect size estimated for rs4921437 in the combined cohort (Odds ratio 0.3745) is larger than all previously associated common variants, a result detected with a sample size far below generally-accepted GWAS power criteria[10,11]. Furthermore, when we

imputed variants in the region, we discovered that the strongest association signal was indeed in close proximity and in linkage disequilibrium with rs4921437, and annotation of the region revealed a histone H3K27Ac mark encompassing both SNPs. We speculate that our resistance genotypes modulate this active enhancer element and lead to increased expression of *IL12B.* We also performed haplotype association analyses, and the most significant three-variant haplotype in the region was comprised of rs4921437, an intergenic SNP, and a variant in an intron of *IL12B.* We noticed substantial concordance of linkage disequilibrium patterns in patients from the Ugandan and Tanzanian cohorts in this genomic region, which is unusual for African, and especially East African populations. Using genetic data with denser coverage from nearby populations, we also demonstrated that the *IL12B* region has undergone selection. It is pertinent to point out, however, that due to the pleiotropic effects of the IL12 cytokine, we cannot definitively state that the observed signature of selection is a result of pressure from tuberculosis alone, as IL12 plays a role in the response to other infections. In conclusion, results from this chapter are noteworthy because they reveal an important novel association of resistance to TB with common variation in a putative regulatory region of *IL12B,* and also because they serve as a proof of principle for the utility of extreme phenotypes in genetic association studies.

Rates of tuberculosis in Africa are high because of very high prevalence of HIV. Globally, most cases of tuberculosis are observed in Asia. We posit that if rs4921437 modulates *IL12B* and the variant is, therefore, truly protective, its absence of the protective allele in East Asia is at could be partially responsible for the high incidence of disease there despite low HIV rates. Furthermore, the IL12 cytokine is available as a pharmacological agent, experimentally used as an injectable therapy in cancer treatment. Therefore, our discovery of a protective role of IL12 suggests an alternative indication for the cytokine as a possible treatment modality for

active TB, which hypothetically would be most useful in Asia. Further studies of IL12 safety in cells and animal models as well as its effectiveness as an inhalant are merited.

In Chapter IV, we hypothesized that HIV-positive individuals living in MTB hyperendemic areas who do not get infected with MTB also represent an extreme resistance phenotype. The MTB infection phenotype is nuanced in the HIV-positive patient population because immunosuppression can generate false negative results of the tuberculin skin test diagnostic. We were able to leverage additional immune data available in our patient populations and exclude those with suspected immune anergy to minimize confounding by immunosuppression. Part A of this chapter describes a genome-wide association study of both binary and continuous TST outcomes. We found a genome-wide significant association of a resistance variant, rs877356, with both the continuous and binary infection related outcomes. Variant rs877356 is near *IL9*, a gene with a substantial role in airway inflammation, bronchial asthma, and other respiratory infections. This is an exciting discovery, as observational studies of the inverse incidence of asthma and tuberculosis have been reported[133]. We speculate that over expression of *IL9* prevents MTB infection by the same mechanism that leads to constitutive inflammation of the upper airway in the pathogenesis of asthma. MTB sequestered in the upper airway is killed by microbicidal macrophages, and is unable to establish a latent infection[20].

Of note, the MTB infection resistance variant, rs877356, was not significant at even the 0.05 level in conferring resistance from active TB disease in Chapter III, with a p value of 0.537 and an odds ratio of 0.907. While the odds ratio is still protective, this weak effect size for TB disease is not surprising. Over two billion people have been infected with MTB worldwide, but only about 10% will develop active TB over the course of their lifetime. Therefore, it stands to reason that the variants affecting infection differ from those modulating disease.

In Part B of Chapter IV, we used regression to fine map genes and regions previously associated with MTB infection through genome-wide linkage studies. We found variant rs10834029, in the chromosome 11p14 region, to be significantly associated with MTB infection using a multiple testing corrected adjustment for the number of independent SNPs in the region[115]. The position of the variant also overlaps with the second highest LOD score observed in the original study[12]. *GAS2* is the closest gene to the SNP, and it plays a role in the induction of apoptosis. Programmed cell death deprives MTB of the intracellular nidus for infection; therefore this is a promising candidate for follow up. Overall, the fine mapping follow up was hindered by the broad target ranges defined in the original studies. The conventional 0.05 statistical threshold for replication could not be used on account of type I error inflation, while a conservative Bonferroni adjusted threshold for SNPs in each region could not be attained.

The immune response to both MTB infection and resultant TB disease is multifaceted, and the cells types, cytokines, and even affected tissues change during the course of the disease. It is, therefore, likely that single locus associations are not adequate to explain these phenotypes. In Chapter V, we used multifactorial dimensionality reduction software to explore higher order interactions between variants in candidate genes as they pertain to TB disease and MTB infection. In Part A of the chapter, looking at the genetic architecture of TB disease, we found a three way interaction between variants near *IL12B*, *TNF-α,* and *CR1*. The products of *IL12B* and *TNF-α* have some redundant and some concordant effects in the inflammatory response to MTB; therefore, their interaction is biologically plausible. *CR1* plays a role in phagocytosis of MTB by macrophages, which is also a key step in immune response to MTB. Importantly, this interaction was consistent in the Ugandan and Tanzanian cohorts, which is rare in studies of epistasis. In Part B of the chapter, we evaluated higher order genetic architecture of MTB infection, and

found an interaction between variants near *IL9, SLC6A3,* and *RAB6C.* Of note, *RAB6C* interacts with p53 in a pro-apoptotic pathway; therefore it is a very good candidate for resistance to MTB infection. Unfortunately the mechanism of action of *SLC6A3* as it pertains to MTB infection is still unclear.

Statistical power considerations prevented us from studying genetic determinants of the SRL 172 vaccine, as briefly detailed in Chapter VI.

B. Future Directions

The findings of this project provide numerous avenues for follow up analyses. Based on the current and previous results, the role of the *IL12B* gene in tuberculosis pathogenesis appears to be well established; therefore further replication of this gene may not be necessary to substantiate its association with TB. However, the role of variants in or near the H3K27Ac mark has not been explored pertaining to TB disease. Since we posited that our genome-wide significant variants rs4921437 and rs10515780 might indeed be causal, a functional analysis relating variants in the H3K27Ac region to *IL12B* expression is a logical follow up. Since the histone mark was discovered in immunity related cell lines GM12878 and K562, from B-lymphocytes and chronic myelogenous leukemia, respectively, they would be good candidates for initial functional experiments. Replication of the exact resistance loci we identified in different cohorts would also be suggestive of causality for variants in the region, especially in patient populations from different regions of the world. Deep resequencing of the *IL12B* region in our sample would allow us to evaluate any association of rare variants with TB. While we imputed the region to attain greater coverage, the procedure is not reliable for SNPs with very low minor allele frequencies, especially in non-European populations. Therefore, if the variants

we identified are in fact not causal, we would be able to discover other relevant SNPs by resequencing. Finally, it is also important to demonstrate that the rs4921437/H3K27Ac finding is consistent beyond the HIV-positive patient niche; therefore, replication in an HIV-negative population is worth pursuing.

The association of variants in the *IL9* region with resistance to MTB infection is a novel discovery; hence, follow up analyses examining this region are merited. Resequencing of the region is a logical follow up, as is a functional analysis of *IL9* and *SLC25A48* expression as it pertains to MTB infection, since either or both genes could possibly play a role in resistance. Again, replicating this association in cohorts from different populations and in an HIV-negative cohort would be optimal. A study linking variation in rs877356 with expression of *IL9* or the other gene in the region, *SLC25A48*, would provide additional evidence that the variant is indeed causal and complement the suggested functional analyses.

The epistatic effects discovered for both TB disease and MTB infection need to be corroborated in other cohorts. A functional analysis linking *SLC6A3* to a mechanism of MTB infection resistance would facilitate interpretation of MDR results for these phenotypes. Candidate gene analyses of epistasis revealed that the replicating models in MDR include the best single hits from the TB disease and MTB infection studies, it is, therefore, a logical follow up assess the interaction of the respective large main effect SNPs with all other variants available on the chip.

Finally, although we likely missed the window for recruiting the SRL 172 vaccine arm patients from the DarDar vaccine trial into our study, a new formulation of the vaccine is being tested. We plan to obtain blood samples and consent patients for genetic analyses on enrollment

into this study, which would guarantee a much larger sample size, and permit us to carry out the vaccinogenetics aim in the new cohort.

APPENDIX

Appendix Table 3-1. Single nucleotide polymorphisms associating with active tuberculosis below a $5x10^{-5}$ p value in the combined cohort using a dominant model

SNP	CHR	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value	Gene
rs4921437	5	T	0.21	0.34	(0.23, 0.50)	1.14E-07	*UBLCP1/IL12B*
rs11652378	17	A	0.34	0.40	(0.27, 0.58)	2.21E-06	*TUSC5*
rs10826834	10	C	0.32	2.48	(1.69, 3.62)	2.74E-06	*LYZL2*
rs8028149	15	C	0.42	0.40	(0.27, 0.59)	4.93E-06	*VPS13C*
rs12636260	3	T	0.26	2.42	(1.65, 3.54)	5.95E-06	*ZPLD1**
rs1537434	13	G	0.27	0.43	(0.29, 0.63)	1.73E-05	*RPL21*
rs57568	2	G	0.41	0.42	(0.28, 0.62)	1.80E-05	*LTBP1*
rs1437747	2	T	0.32	0.44	(0.30, 0.64)	1.81E-05	*CAMKMT*
rs1616723	6	G	0.20	0.42	(0.28, 0.63)	2.03E-05	*GLO1*
rs1525738	7	A	0.40	0.43	(0.29, 0.64)	2.31E-05	*AGR3*
rs10938264	4	G	0.33	2.25	(1.54, 3.28)	2.79E-05	*GRXCR1*
rs12693365	2	A	0.32	0.45	(0.31, 0.66)	3.77E-05	*NUP35**
rs844669	7	G	0.30	0.46	(0.32, 0.67)	4.89E-05	*CALN1*

*>500kb away

Appendix Table 3-2. Single nucleotide polymorphisms associating with active tuberculosis below a $5x10^{-5}$ p value using a dominant model in the A) Ugandan and B) Tanzanian (rs4921437 added) cohorts

A)

SNP	CHR	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value	Gene
rs4921437	5	T	0.21	0.23	(0.12, 0.42)	3.68E-06	*UBLCP1/IL12B*
rs859063	1	G	0.35	4.57	(2.40, 8.72)	4.01E-06	*SLC44A3*
rs11800062	1	C	0.40	0.18	(0.084, 0.39)	1.58E-05	*CDC42BPA*
rs11811176	1	G	0.47	0.13	(0.051, 0.33)	2.14E-05	*CDC42BPA*
rs12429777	13	C	0.21	0.27	(0.15, 0.50)	2.40E-05	*KLHL1*
rs293392	15	G	0.32	3.84	(2.04, 7.23)	3.17E-05	*ABHD2*
rs2006724	6	G	0.47	3.93	(2.05, 7.55)	3.92E-05	*BC035400*
rs2395300	6	A	0.25	4.13	(2.09, 8.13)	4.20E-05	*HLA-DOA*
rs4724086	7	G	0.28	3.83	(2.01, 7.28)	4.23E-05	*GLI3*
rs345397	7	A	0.22	0.28	(0.15, 0.51)	4.95E-05	*SEPT7*

B)

SNP	CHR	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value	Gene
rs9566115	13	T	0.48	0.26	(0.14, 0.48)	1.07E-05	*SLITRK6*
rs8111608	19	A	0.33	3.39	(1.95, 5.89)	1.58E-05	*SLC1A6*
rs11733040	4	G	0.26	0.28	(0.16, 0.50)	1.60E-05	*PCDH18**
rs557438	5	T	0.37	0.31	(0.18, 0.53)	1.63E-05	*MCC*
rs2807348	1	C	0.26	0.30	(0.17, 0.52)	2.18E-05	*WNT4*
rs1158059	6	T	0.20	3.09	(1.80, 5.29)	3.99E-05	*SNAP91*
…	…	…	…	…	…	…	...
rs4921437	5	T	0.21	0.45	(0.26, 0.78)	4.66E-03	*UBLCP1/IL12B*

*>500kb away

Appendix Table 3-3. Single nucleotide polymorphisms associating with active tuberculosis below a $5x10^{-5}$ p value using a recessive model in the A) combined, B) Tanzanian, and C) Ugandan cohort

A)

SNP	CHR	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value	Gene
rs1573219	9	T	0.43	0.34	(0.20, 0.56)	2.29E-05	*NTRK2*
rs3763787	10	A	0.33	0.25	(0.13, 0.48)	2.77E-05	*ACBD7*
rs955263	4	T	0.34	3.41	(1.91, 6.06)	3.10E-05	*SORBS2*

B)

SNP	CHR	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value	Gene
rs10521221	16	T	0.30	5.74	(2.55, 12.94)	2.49E-05	*SALL1*
rs7797545	7	T	0.43	3.90	(2.07, 7.34)	2.57E-05	*PKD1L1*
rs10203511	2	A	0.45	3.44	(1.92, 6.17)	3.41E-05	*NRXN1*

C)

SNP	CHR	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value	Gene
rs1573219	9	T	0.43	0.19	(0.096, 0.40)	6.22E-06	*NTRK2*
rs3763787	10	A	0.36	0.16	(0.064, 0.38)	3.51E-05	*ACBD7*

Appendix Table 3-4. Power calculation using a log additive model, a genetic effect of 2.0, a baseline risk of 0.1, an alpha of 0.05, and power of 0.8 in A) Uganda with a 2 to 1 case to control ratio and B) Tanzania with a 1 to 2 case to control ratio

A)

Minor allele frequency	N cases needed
0.05	203
0.10	112
0.15	82
0.20	68
0.25	60

B)

Minor allele frequency	N cases needed
0.05	432
0.10	236
0.15	172
0.20	141
0.25	124

Appendix Table 3-5. Summary of self-reported tribal membership in the Ugandan cohort

Tribe	Count
Muganda	151
Mutooro	10
Mufumbira	8
Mukiga	7
Ankole	4
Musoga	3
Munyarwanda	5
Tanzanian	3
Muhororo	1
Lugbara	2
Soga	1
Alur	2
Lugwara	1
Mwamba	1
Munyoro	7
Nyankole	7
Rwandese	2
Nkole	4
Chapadhola	3
Munyankole	7
Langi	4
Itesof	1
Toro	1
Acholi	1
Kakwa	1
Luo	1
total n	238

Appendix Table 3-6. Association of principal components with self-reported tribal membership in the Ugandan cohort using logistic regression predicting Muganda versus other

Variable	OR	95% CI	p value
PC1	1.22E-13	(5.42E-17, 2.73E-10)	4.79E-13
PC2	0.004	(4.50E-05, 0.34)	0.015
PC3	0.002	(6.67E-06, 0.58)	0.032
PC4	0.17	(0.0014, 22.08)	0.478
PC5	1.07	(0.017, 65.17)	0.975
PC6	7659	(80, 742677)	1.28E-4
PC7	2.65E-04	(9.26-07, 0.076)	0.004
PC8	4.97	(0.075, 330.09)	0.454
PC9	1.40	(0.021, 92.46)	0.876
PC10	0.12	(0.0016, 8.20)	0.322

Appendix Table 3-7. Single nucleotide polymorphisms associating with active tuberculosis below a $5x10^{-5}$ p value in the A) Ugandan and B) Tanzanian (rs4921437 also listed) cohorts

A)

SNP	CHR	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value	Gene
rs4921437	5	T	0.21	0.28	(0.17, 0.47)	1.18E-06	*UBLCP1/IL12B*
rs7297313	12	C	0.34	0.35	(0.22, 0.55)	8.6E-06	*GRIN2B*
rs859063	1	G	0.35	3.06	(1.82, 5.13)	2.40E-05	*SLC44A3*

B)

SNP	CHR	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	p value	Gene
rs2681052	7	T	0.39	2.65	(1.72, 4.09)	1.07E-05	*THSD7A*
rs17048476	3	A	0.42	2.43	(1.62, 3.66)	1.91E-05	*AK124857*
rs2587469	10	C	0.46	0.42	(0.28, 0.63)	2.26E-05	*ADAMTS14*
rs9893385	17	G	0.47	0.41	(0.28, 0.62)	2.26E-05	*ABCA8*
rs3860173	9	T	0.36	2.28	(1.56, 3.33)	2.27E-05	*RGS3*
rs1524713	1	T	0.28	2.43	(1.59, 3.69)	3.47E-05	*DAB1*
rs2807348	1	C	0.26	0.35	(0.22, 0.58)	3.48E-05	*WNT4*
rs4700255	5	G	0.32	2.34	(1.56, 3.50)	3.87E-05	*ACTBL2*
rs16870583	5	C	0.25	2.32	(1.55, 3.47)	4.35E-05	*IRX2*
rs930205	5	G	0.29	2.39	(1.57, 3.64)	4.64E-05	*SH3TC2*
rs557438	5	T	0.37	0.41	(0.27, 0.63)	4.70E-05	*MCC*
…	…	…	…	…	…	…	...
rs4921437	5	T	0.21	0.48	(0.30, 0.79)	3.84E-03	*UBLCP1/IL12B*

Appendix Table 3-8. Association of imputed SNPs in the rs4921437 +/-1 mb window on chromosome 5 with active tuberculosis in the combined cohort

SNP	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	Imputation Certainty Uganda	Imputation Certainty Tanzania	p value	Gene
rs10515780	C	0.19	0.34	(0.24, 0.49)	0.99	0.99	6.21E-09	intron *UBLCP1*
rs4921437	T	0.21	0.37	(0.27, 0.53)	Not Imputed	Not Imputed	2.11E-08	intron *UBLCP1*
rs56296726	A	0.18	0.37	(0.25, 0.53)	0.97	0.97	8.46E-08	~6kb upstr. *UBLCP1*
rs7736656	A	0.15	0.33	(0.22, 0.50)	0.97	0.97	9.32E-08	intron *UBLCP1*
rs6895626	T	0.15	0.33	(0.22, 0.50)	0.97	0.97	9.32E-08	~1kb dwnstr. *UBLCP1*
rs73307745	G	0.14	0.35	(0.23, 0.53)	0.96	0.97	5.82E-07	~2kb upstr. *UBLCP1*
rs73307742	T	0.13	0.35	(0.23, 0.54)	0.96	0.97	1.21E-06	~6kb upstr. *UBLCP1*
rs200213893	INS	0.32	0.55	(0.42, 0.73)	0.97	0.97	3.88E-05	~4kb upstr. *UBLCP1*
rs10040411	C	0.46	0.57	(0.43, 0.74)	0.97	0.98	3.93E-05	~54kb upstr. *IL12B*
rs115740168	G	0.10	0.38	(0.24, 0.60)	0.95	0.96	4.39E-05	intron *IL12B*

Appendix Table 3-9. Association of imputed SNPs in the rs4921437 +/-1 mb window on chromosome 5 with active tuberculosis in the A) Ugandan and B) Tanzanian cohort

A)

SNP	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	Imputation Certainty Uganda	p value	Gene
rs10515780	C	0.19	0.27	(0.16, 0.55)	0.99	9.92E-07	intron *UBLCP1*
rs4921437	T	0.21	0.28	(0.17, 0.47)	Not Imputed	1.18E-06	intron *UBLCP1*
rs56296726	A	0.19	0.28	(0.17, 0.48)	0.97	3.49E-06	~6kb upstr. *UBLCP1*
rs6895626	T	0.15	0.27	(0.16, 0.48)	0.97	8.36E-06	~1kb downstr. *UBLCP1*
rs7736656	A	0.15	0.27	(0.16, 0.48)	0.97	8.36E-06	intron *UBLCP1*
rs73307742	T	0.14	0.27	(0.15, 0.48)	0.96	1.25E-05	~6kb upstr. *UBLCP1*
rs115740168	G	0.11	0.22	(0.11, 0.44)	0.95	1.46E-05	intron *IL12B*
rs73307745	G	0.15	0.29	(0.16, 0.51)	0.96	1.99E-05	~2kb upstr. *UBLCP1*
rs10077752	T	0.47	0.37	(0.24, 0.60)	0.99	3.97E-05	~65kb upstr. *UBLCP1*

B)

SNP	Minor Allele	MAF	Odds Ratio	95% Confidence Interval	Imputation Certainty Uganda	p value	Gene
rs58574384	A	0.25	0.47	(0.29, 0.76)	0.86	0.0021	~112kb downstr. *UBLCP1*
rs10515780	C	0.18	0.43	(0.25, 0.74)	0.99	0.0021	intron *UBLCP1*
rs56318149	A	0.18	2.00	(1.27, 3.15)	0.93	0.0027	~200kb downstr. *UBLCP1*
rs7736656	A	0.15	0.40	(0.22, 0.73)	0.97	0.0029	intron *UBLCP1*
rs6895626	T	0.15	0.40	(0.22, 0.73)	0.97	0.0029	~1kb downstr. *UBLCP1*
rs6556400	T	0.16	2.07	(1.27, 3.37)	0.91	0.0036	~111kb upstr. *UBLCP1*
rs6883501	C	0.17	0.44	(0.25, 0.77)	0.95	0.0037	~100kb downstr. *UBLCP1*
rs4921437	T	0.21	0.48	(0.30, 0.79)	Not Imputed	0.0038	intron *UBLCP1*

Appendix Table 3-10. Association of the rs4921437 and rs4921468 haplotype in A) the combined cohort, B) Tanzanian and C) Ugandan cohorts, and the rs4921437 and rs3213094 haplotype in D) the combined cohort, E) Tanzanian and F) Ugandan cohorts

A)

Haplotype	Case	Control	Ca-Freq	Co-Freq
C-A	102	108	0.19	0.17
C-G	356	353	0.67	0.56
T-A	4	7	0.01	0.01
T-G	72	160	0.13	0.25
Likelihood ratio chisq = 34.2397 df = 3 p-value = 1.76328E-007*				

B)

Haplotype	Case	Control	Ca-Freq	Co-Freq
C-A	42	87	0.21	0.18
C-G	131	283	0.66	0.58
T-A	0	4	0.00	0.01
T-G	27	112	0.14	0.23
Likelihood ratio chisq = 12.3668 df = 3 p-value = 0.00622658^				

C)

Haplotype	Case	Control	Ca-Freq	Co-Freq
C-A	60	21	0.18	0.15
C-G	225	70	0.67	0.50
T-A	4	3	0.01	0.02
T-G	45	48	0.13	0.34
Likelihood ratio chisq = 39.023 df = 3 p-value = 1.71634E-008^				

D)

Haplotype	Case	Control	Ca-Freq	Co-Freq
C-T	247	229	0.46	0.36
C-C	211	232	0.40	0.37
T-T	6	13	0.02	0.02
T-C	70	154	0.13	0.24
Likelihood ratio chisq = 37.3328 df = 3 p-value = 3.91261E-008*				

E)

Haplotype	Case	Control	Ca-Freq	Co-Freq
C-T	97	182	0.49	0.37
C-C	76	188	0.38	0.39
T-T	1	9	0.01	0.02
T-C	26	107	0.13	0.22
Likelihood ratio chisq = 14.8324 df = 3 p-value = 0.0019656^				

F)

Haplotype	Case	Control	Ca-Freq	Co-Freq
C-T	150	47	0.45	0.33
C-C	135	44	0.40	0.31
T-T	5	4	0.01	0.03
T-C	44	47	0.13	0.33
Likelihood ratio chisq = 45.0414 df = 3 p-value = 9.0671E-010^				

* adjusted for principal components, sex, and cohort of origin

^ adjusted for principal components and sex

Appendix Table 3-11. Chi squared comparison of haplotypes between the study samples from Uganda and Tanzania with African origin samples from Phase 3 HapMap using A) rs4921437 and rs3213094, B) rs4921437 and rs4921468, and C) rs4921437, rs3213094, and rs4921468

A)

Population Comparisons	p value using full Ugandan, Tanzanian cohorts	p value using Tanzanian, Ugandan cases only	p value using for Tanzanian, Ugandan controls only
Uganda v Tanzania	0.93	0.65	0.04
Uganda v YRI	0.001	1.41E-06	0.066
Tanzania v YRI	0.001	4.06E-05	0.016
Uganda v LWK	0.91	0.25	0.028
Tanzania v LWK	0.98	0.23	0.73
Uganda v MKK	1.08E-05	2.12E-05	1.34E-06
Tanzania v MKK	1.57E-05	3.96E-05	7.71E-05
Uganda v ASW	0.077	0.016	0.039
Tanzania v ASW	0.067	0.003	0.20

B)

Population Comparisons	p value using full Ugandan, Tanzanian cohorts	p value using Tanzanian, Ugandan cases only	p value using for Tanzanian, Ugandan controls only
Uganda v Tanzania	0.15	0.20	0.033
Uganda v YRI	0.026	3.27E-05	0.81
Tanzania v YRI	0.011	0.00019	0.20
Uganda v LWK	0.65	0.13	0.024
Tanzania v LWK	0.86	0.17	0.76
Uganda v MKK	7.71E-08	1.23E-07	1.35E-06
Tanzania v MKK	3.74E-06	0.001	3.79E-06
Uganda v ASW	0.14	0.011	0.01
Tanzania v ASW	0.50	0.065	0.48

C)

Population Comparisons	p value using full Ugandan, Tanzanian cohorts	p value using Tanzanian, Ugandan cases only	p value using for Tanzanian, Ugandan controls only
Uganda v Tanzania	0.12	0.085	0.088
Uganda v YRI	0.002	3.98E-06	0.20
Tanzania v YRI	0.002	6.14E-05	0.025
Uganda v LWK	0.87	0.41	0.067
Tanzania v LWK	0.30	0.12	0.17
Uganda v MKK	2.34E-08	2.84E-08	7.14E-07
Tanzania v MKK	2.52E-08	3.07E-06	3.1E-07
Uganda v ASW	0.19	0.035	0.12
Tanzania v ASW	0.24	0.024	0.46

Appendix Table 3-12. Chi squared comparison of haplotypes between the study samples from Uganda and Tanzania with non-African origin samples from Phase 3 HapMap using A) rs4921437 and rs3213094, B) rs4921437 and rs4921468, and C) rs4921437, rs3213094, and rs4921468

A)

Population Comparisons	p value using full Ugandan, Tanzanian cohorts	p value using Tanzanian, Ugandan cases only	p value using for Tanzanian, Ugandan controls only
Uganda v CEU	5.93E-10	7.42E-10	7.64E-08
Tanzania v CEU	5.45E-10	1.71E-09	2.18E-08
Uganda v CHB	1.57E-08	1.86E-05	7.37E-14
Tanzania v CHB	1.57E-08	3.83E-05	2.16E-09
Uganda v GIH	0.001	0.008	1.23E-06
Tanzania v GIH	0.001	0.007	0.001
Uganda v MEX	NA	NA	NA
Tanzania v MEX	NA	NA	NA
Uganda v TSI	2.95E-06	1.48E-05	9.93E-07
Tanzania v TSI	3.58E-06	1.46E-05	1.61E-05
Uganda v JPT	1.96E-09	5.32E-08	9.90E-16
Tanzania v JPT	1.19E-09	1.03E-05	6.19E-11

B)

Population Comparison	p value using full Ugandan, Tanzanian cohorts	p value using Tanzanian, Ugandan cases only	p value using for Tanzanian, Ugandan controls only
Uganda v CEU	0.01	0.002	0.001
Tanzania v CEU	0.14	0.11	0.067
Uganda v CHB	4.62E-08	3.94E-05	1.45E-13
Tanzania v CHB	1.02E-07	0.000665	4.6E-09
Uganda v GIH	0.035	0.29	7.24E-06
Tanzania v GIH	0.023	0.31	0.004
Uganda v MEX	6.71E-05	0.002	7.6E-07
Tanzania v MEX	3.62E-04	0.025	8.17E-05
Uganda v TSI	8.79E-07	5.01E-06	2.19E-06
Tanzania v TSI	1.77E-05	0.002	9.82E-06
Uganda v JPT	<1.1E-16	1.07E-11	<1.1E-16
Tanzania v JPT	<1.1E-16	1.01E-10	<1.1E-16

C)

Population Comparisons	p value using full Ugandan, Tanzanian cohorts	p value using Tanzanian, Ugandan cases only	p value using for Tanzanian, Ugandan controls only
Uganda v CEU	1.99E-09	1.54E-09	1.7E-07
Tanzania v CEU	1.49E-08	2.76E-08	1.32E-09
Uganda v CHB	3.07E-09	9.26E-06	1.46E-12
Tanzania v CHB	5.63E-07	0.001	8.32E-08
Uganda v GIH	2.02E-05	0.001	4.65E-07
Tanzania v GIH	3.23E-05	0.001	7.93E-05
Uganda v MEX	NA	NA	NA
Tanzania v MEX	NA	NA	NA
Uganda v TSI	1.98E-07	9.2E-07	1.27E-06
Tanzania v TSI	3.7E-07	1.88E-05	1.85E-06
Uganda v JPT	1.05E-09	3.56E-06	2.12E-14
Tanzania v JPT	4.31E-08	0.000249	1.52E-09

Appendix Table 3-13. Minor allele frequencies of rs4921437 in the HapMap Phase 3 populations

Population	MAF (T allele)	Proportion CC	Proportion CT	Proportion TT
ASW	0.24	0.63	0.27	0.10
CEU	0.20	0.63	0.33	0.04
CHB	0.01	0.98	0.03	0
GIH	0.13	0.77	0.21	0.02
HCB	0.02	0.95	0.05	0
JPT	0.01	0.99	0.01	0
LWK	0.21	0.64	0.30	0.06
MEX	0.07	0.86	0.14	0
MKK	0.17	0.68	0.29	0.03
TSI	0.16	0.70	0.29	0.01
YRI	0.31	0.49	0.41	0.10

Appendix Table 3-14. Minor allele frequencies of rs4921437 in the 1000 Genomes Project populations

Population	Alleles C	Alleles T	Genotypes C|C	Genotypes C|T	Genotypes T|T
All Populations	0.86	0.14	0.75	0.22	0.03
ACB	0.79	0.21	0.62	0.34	0.04
All African Populations	0.74	0.26	0.57	0.35	0.08
All American Populations	0.90	0.11	0.81	0.16	0.02
All East Asian Populations	1.00	0	1.00	0	
ASW	0.74	0.26	0.56	0.36	0.08
BEB	0.84	0.16	0.67	0.33	
CDX	1.00		1.00		
CEU	0.80	0.20	0.63	0.34	0.03
CHB	1.00	0.01	0.99	0.01	
CHS	1.00		1.00		
CLM	0.82	0.18	0.69	0.26	0.05
ESN	0.73	0.27	0.53	0.40	0.07
All European Populations	0.83	0.17	0.68	0.30	0.02
FIN	0.87	0.13	0.76	0.23	0.01
GBR	0.82	0.18	0.68	0.29	0.03
GIH	0.89	0.11	0.81	0.17	0.03
GWD	0.76	0.24	0.61	0.30	0.09
IBS	0.80	0.20	0.63	0.35	0.03
ITU	0.89	0.11	0.80	0.18	0.02
JPT	1.00		1.00		
KHV	1.00		1.00		
LWK	0.83	0.17	0.70	0.26	0.04
MSL	0.64	0.37	0.44	0.40	0.17
MXL	0.93	0.07	0.86	0.14	
PEL	0.94	0.06	0.89	0.09	0.01
PJL	0.87	0.14	0.75	0.23	0.02
PUR	0.90	0.10	0.83	0.15	0.02
All South Asian Populations	0.87	0.13	0.76	0.23	0.02
STU	0.86	0.14	0.75	0.22	0.03
TSI	0.79	0.21	0.62	0.34	0.04
YRI	0.74	0.26	0.57	0.35	0.08

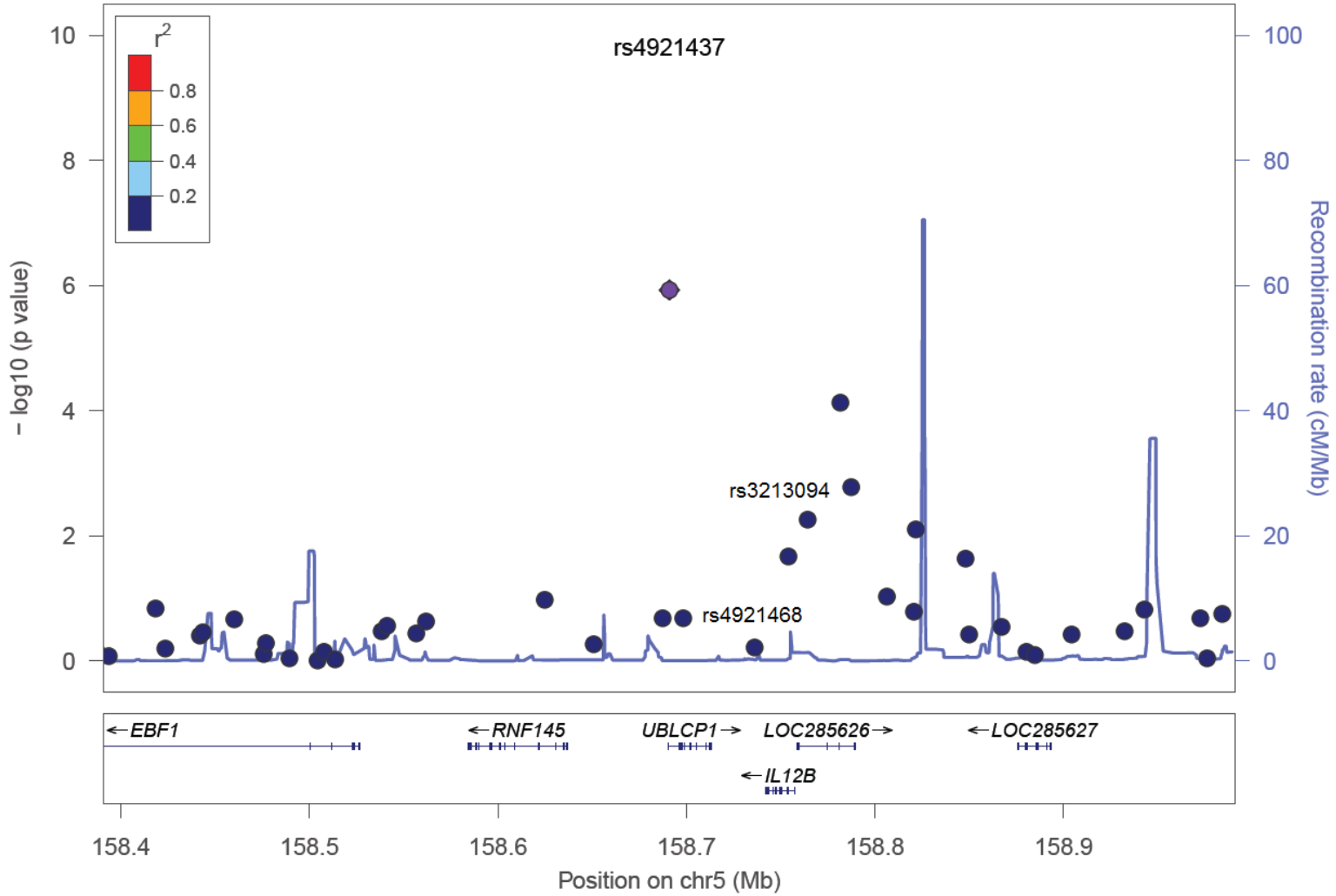
Appendix Table 3-15. 2 SNPs in the 0.1 percentile of the distribution of absolute values of normalized iHS scores in 2 African populations, along with corresponding values for rs4921437

SNP	|iHS| in the Datog	|iHS| in the Niger Kordofanian West	|iHS| in the Western Pygmy
rs3213093	3.915	3.591	2.599
rs2421047	3.088	3.616	2.636
rs4921437	0.391	0.742	0.997

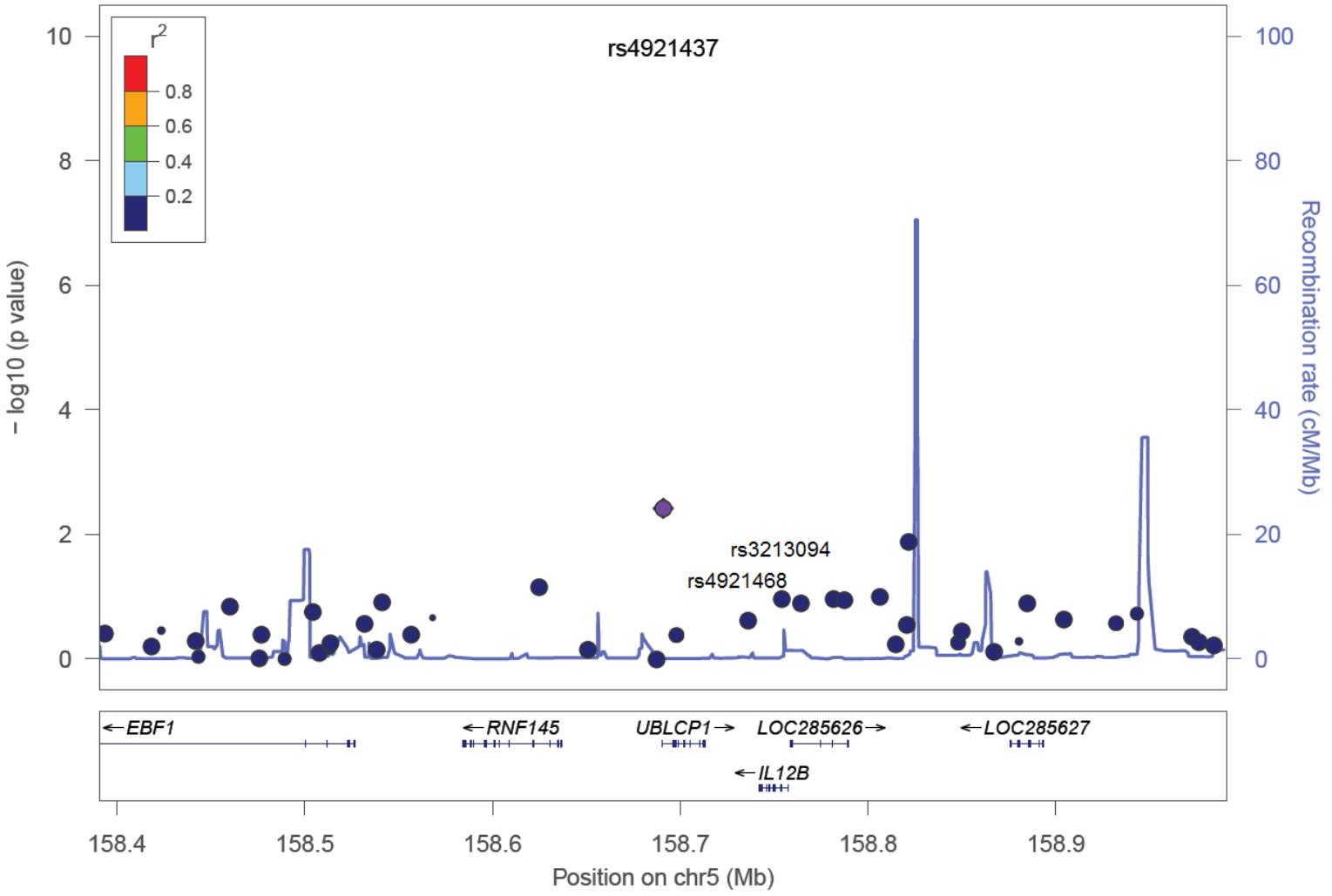
Appendix Table 3-16. DarDar vaccine trial CD4 count, Interferon Gamma Release Assay and Lymphocyte Proliferation Assay data in cases versus controls; B – baseline values, F – final values upon completion of the trial, pi – proliferation index

Variable	n	n cases	TB cases mean (st. dev)	n controls	TB controls mean (st. dev.)	p value
CD4 count	288	36	580.78 (224.90)	252	540.68 (235.30)	0.34
B IFN med	243	29	148.64 (211.20)	214	204.73 (471.08)	0.53
B IFN pha	243	29	10778.93 (13367.56)	214	21028.32 (31904.27)	0.09
B IFN mvs	243	29	201.36 (345.97)	214	334.20 (976.42)	0.47
B IFN ag85	243	29	341.78 (482.99)	214	735.60 (2062.01)	0.31
B IFN esat	240	29	1254.02 (3482.90)	211	1637.08 (6131.51)	0.74
B IFN wcl	204	22	2748.05 (4200.48)	182	2145.87 (3855.69)	0.50
B lam	304	36	0.4976 (0.32)	268	0.5040 (0.45)	0.93
B lpa med	281	36	2154.06 (2513.33)	245	3774.52 (4787.78)	0.05
B lpa pha	281	36	11163.05 (11912.21)	245	15313.85 (13652.40)	0.08
B lpa mvs pi	281	36	1.85 (1.12)	245	1.71 (1.49)	0.58
B lpa ag85 pi	281	36	3.91 (8.07)	245	2.68 (5.43)	0.24
B lpa esat pi	279	36	2.58 (6.24)	243	2.47 (5.17)	0.91
B lpa wcl pi	281	36	8.14 (15.20)	245	5.67 (10.76)	0.23
F IFN mvs	204	22	203.67 (263.75)	182	445.67 (2008.17)	0.57
F IFN ag85	204	22	387.38 (812.68)	182	536.79 (1319.86)	0.60
F IFN esat	204	22	474.15 (1008.68)	182	952.11 (3359.77)	0.51

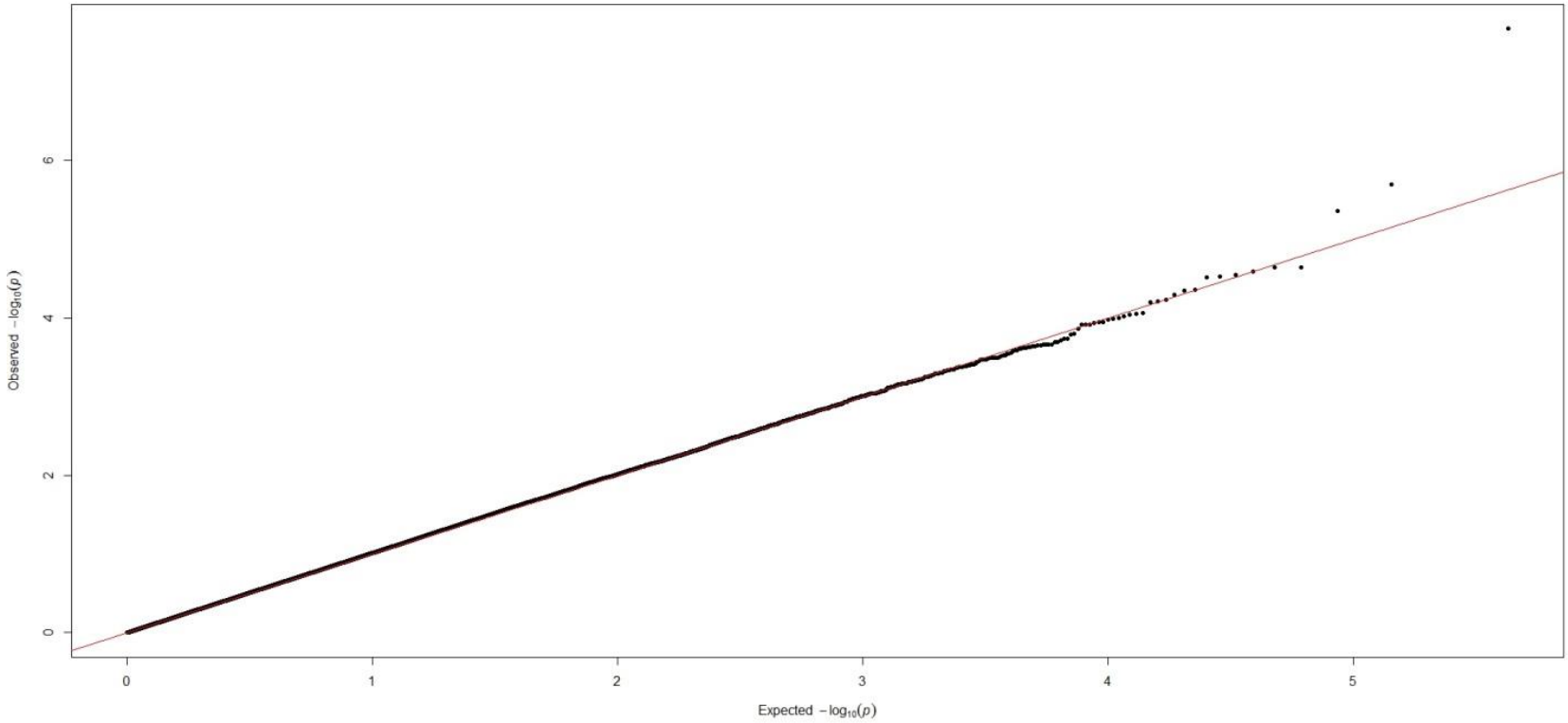
Appendix Figure 3-1. Locus zoom plot of IL12B region in the Ugandan cohort



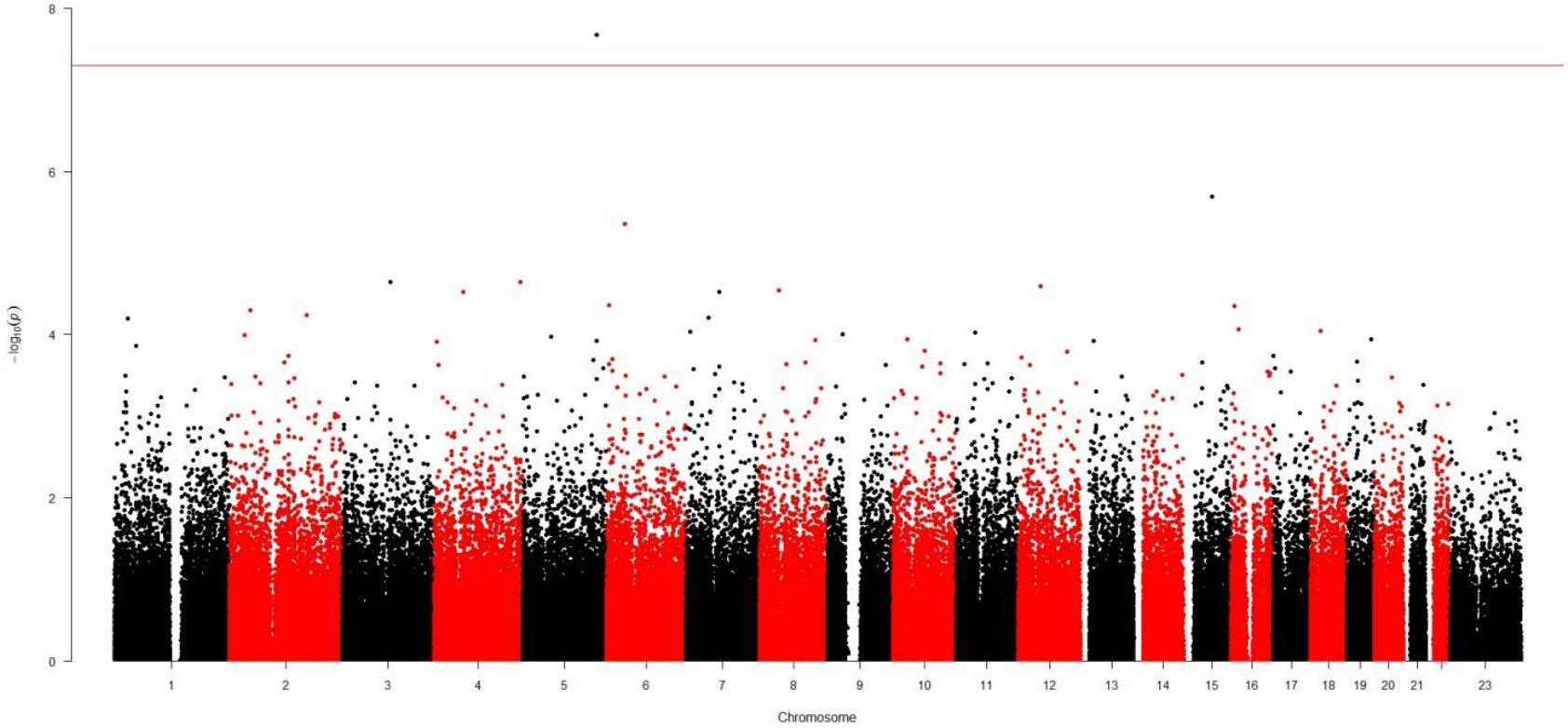
Appendix Figure 3-2. Locus zoom plot of IL12B region in the Tanzanian cohort



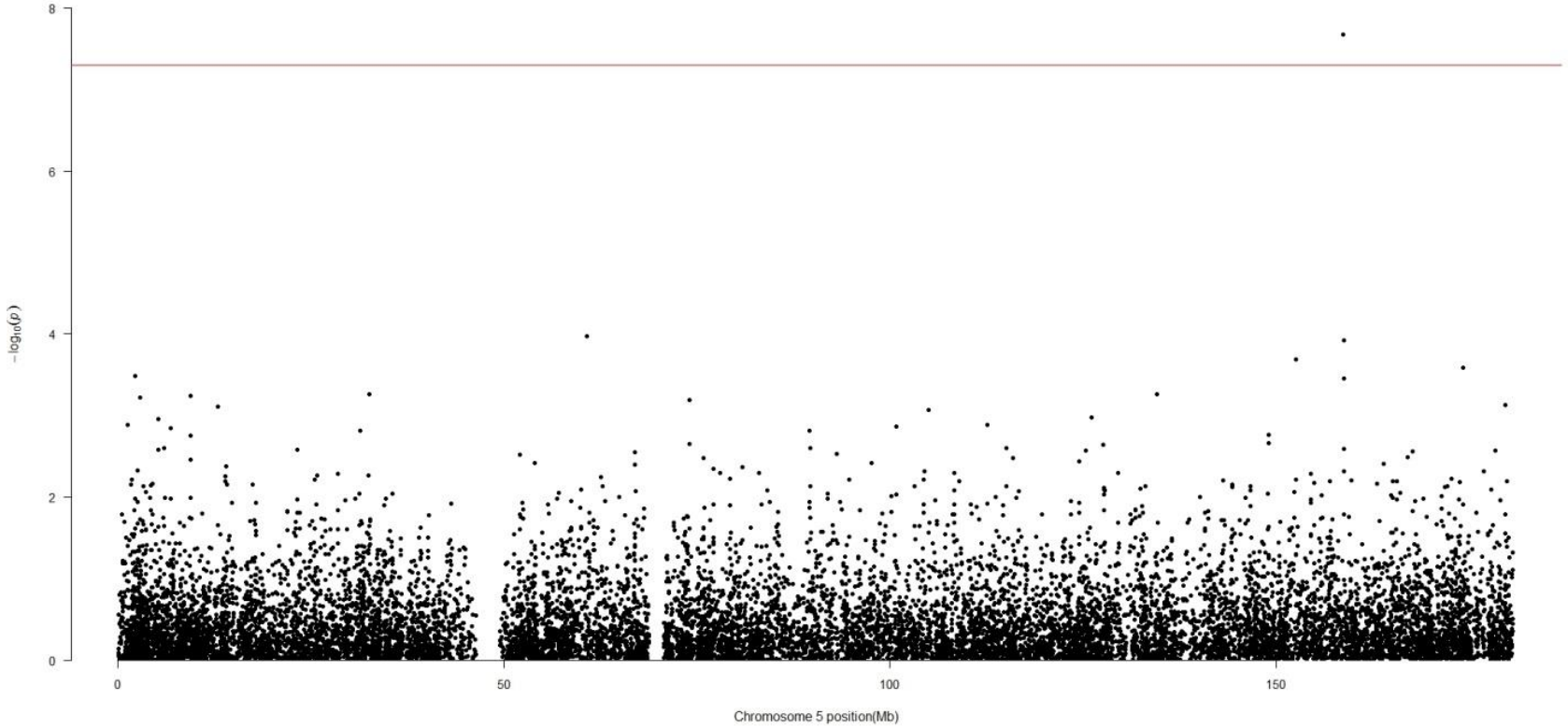
Appendix Figure 3-3. QQ plot of logistic regression results using an additive model for the combined Ugandan and Tanzanian datasets



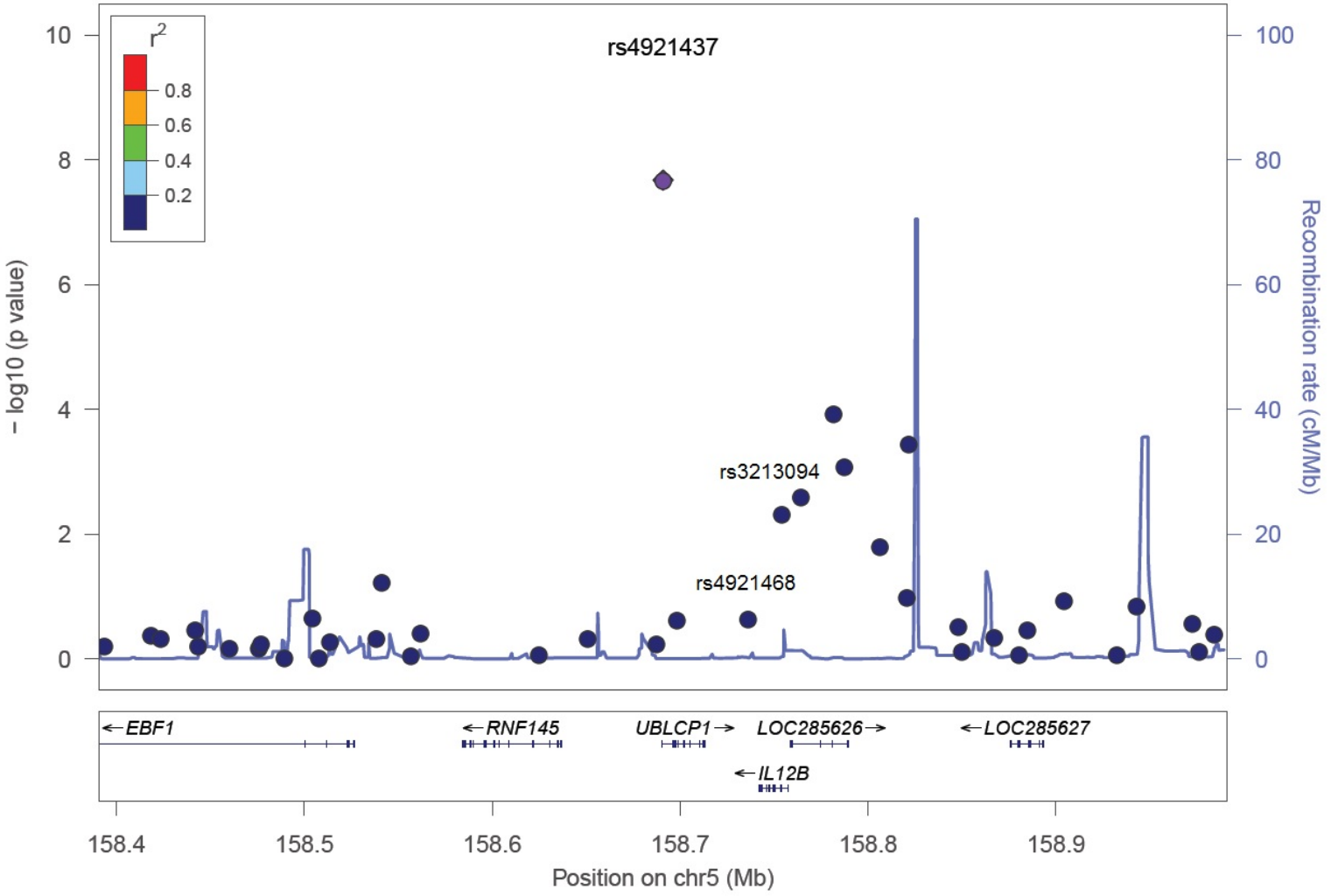
Appendix Figure 3-4. Manhattan plot of logistic regression using an additive model for the combined Ugandan and Tanzanian datasets. Red – genome-wide significance threshold.



Appendix Figure 3-5. Manhattan plot chromosome 5 only using an additive model for the combined Ugandan and Tanzanian datasets. Red – genome-wide significance threshold.

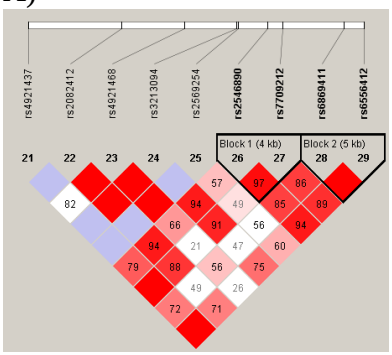


Appendix Figure 3-6. Locus zoom plot of IL12B region in the combined cohort

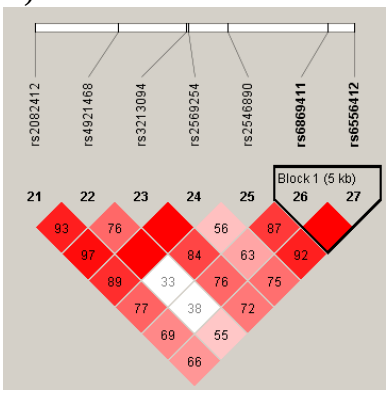


Appendix Figure 3-7. Haploview plots of D' the IL12B region (94 kb range) in non-African Phase 3 HapMap populations A) CEU, B) CHD, C) TSI, D) GIH, E) JPT, F) MEX, and G) CHB

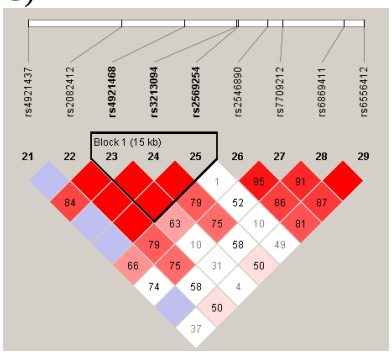
A)



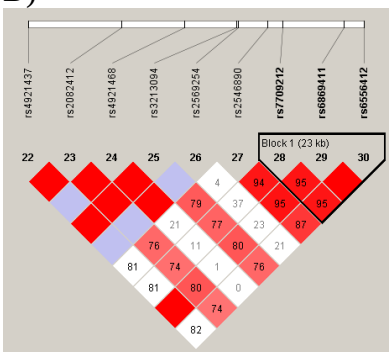
B)



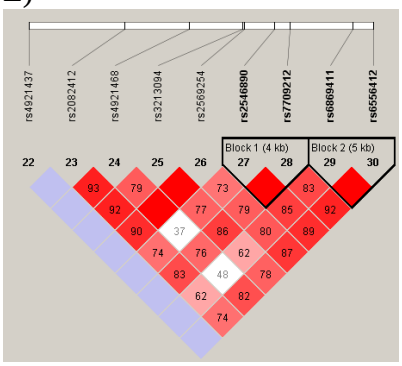
C)



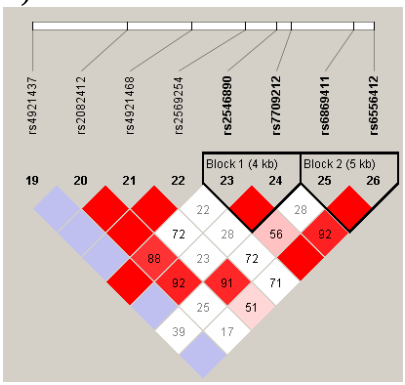
D)



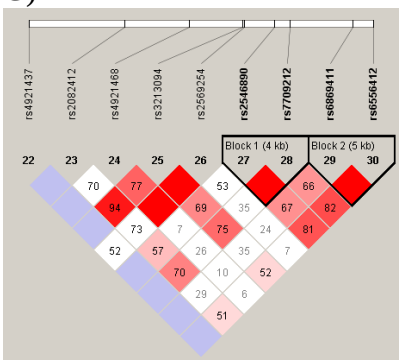
E)



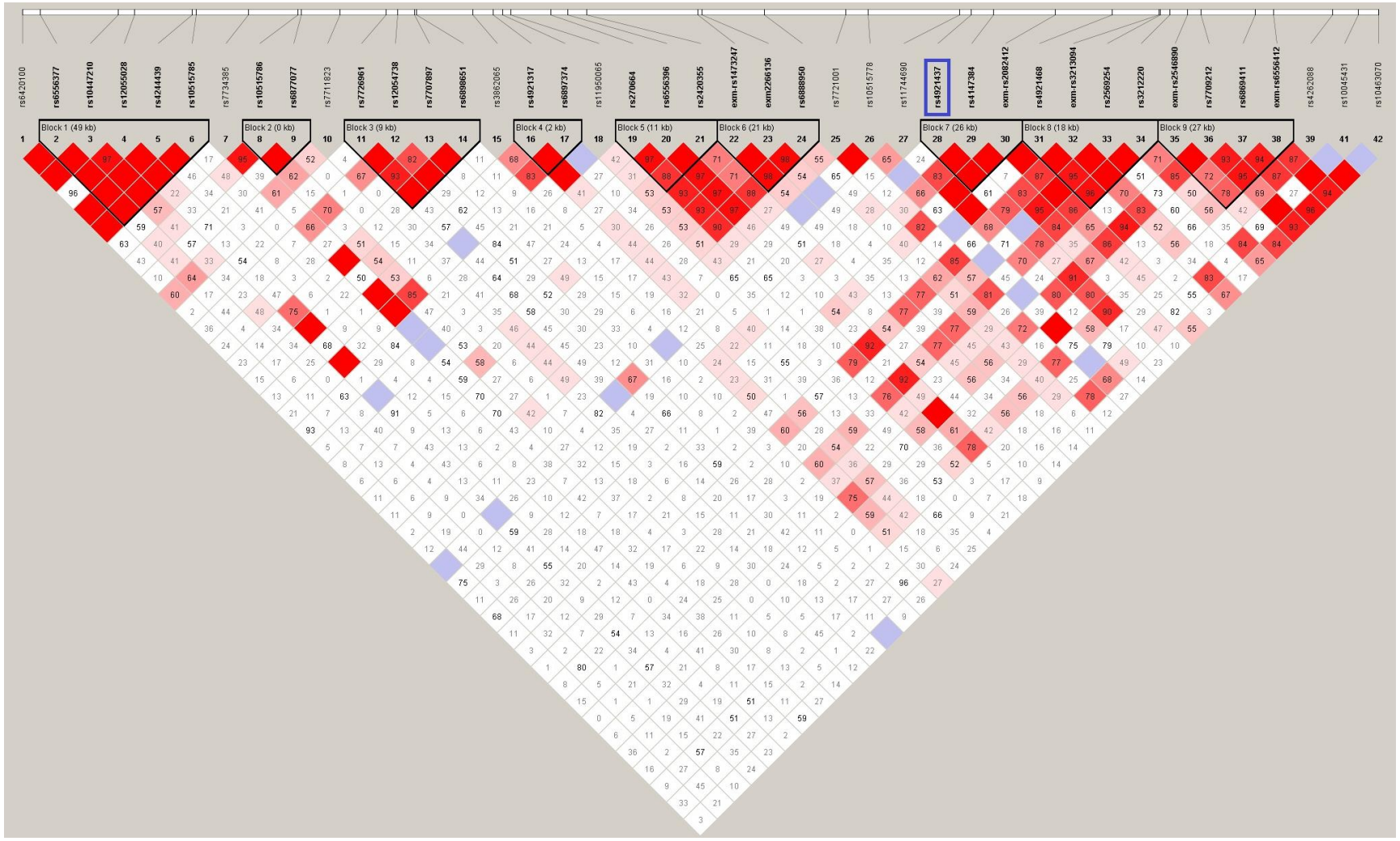
F)



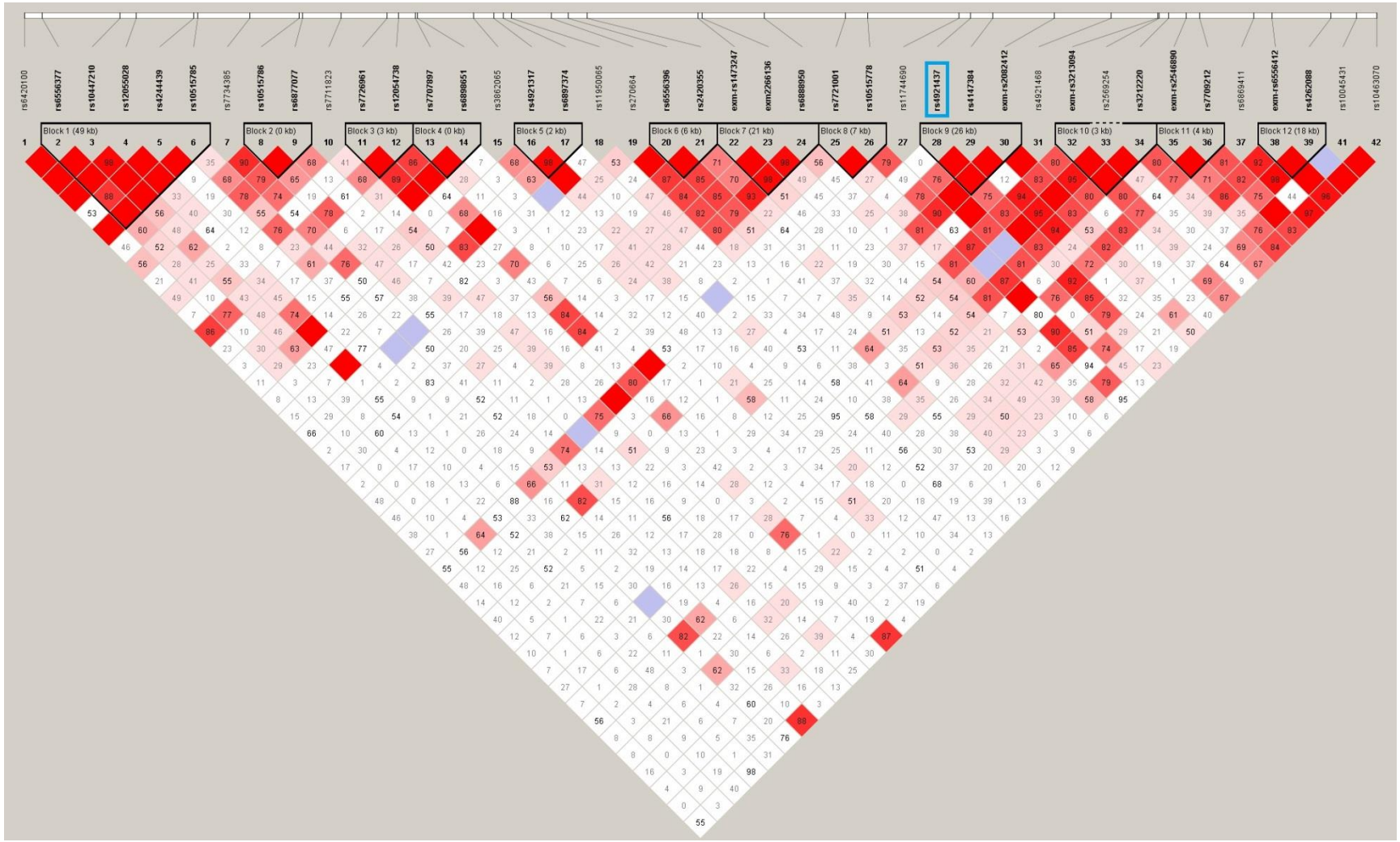
G)



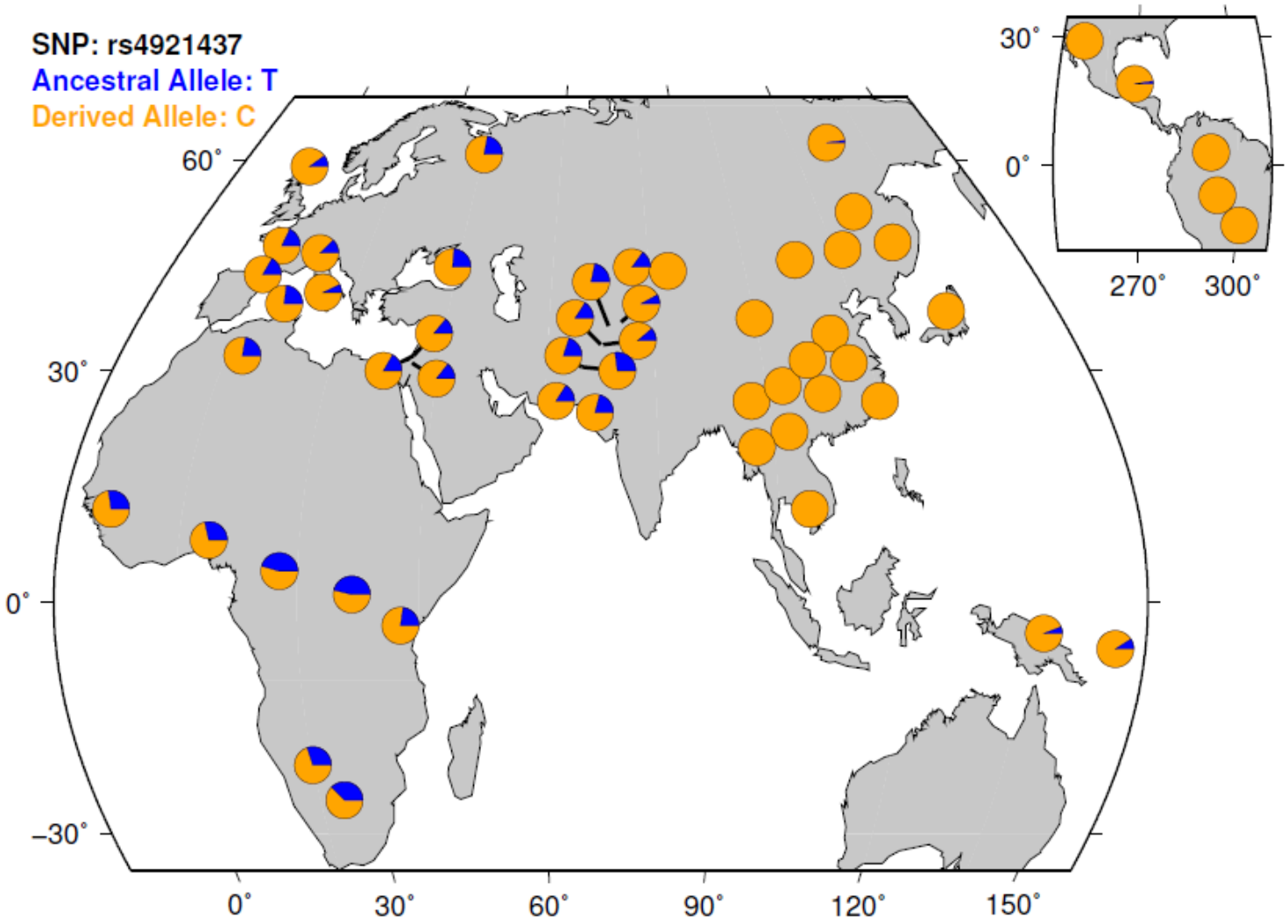
Appendix Figure 3-8. Extended Haploview plot of D’ in the IL12B region (433 kb range) in participants from the Uganda cohort



Appendix Figure 3-9. Extended Haploview plot of D' in the IL12B region (433 kb range) in participants from the Tanzanian cohort



Appendix Figure 3-10. Minor allele frequencies of rs4921437, from the HGDP genome browser[116,117]



Appendix Table 4-1. Power calculation using a log additive model, a genetic effect of 2.0, a baseline risk of 0.33, an alpha of 0.05, and power of 0.8 in A) Uganda with a 3 to 1 case to control ratio and B) Tanzania with a 1 to 2 case to control ratio

A)

Minor allele frequency	N cases needed
0.05	699
0.10	374
0.15	267
0.20	215
0.25	185

B)

Minor allele frequency	N cases needed
0.05	240
0.10	130
0.15	94
0.20	76
0.25	66

Appendix Table 4-2. Single nucleotide polymorphisms associating with continuous tuberculin skin test induration below a $5x10^{-5}$ p value in a dominant genetic model in A) the combined cohort*, B) the Ugandan cohort^, and C) the Tanzanian cohort^

A)

SNP	Chr.	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs877356	5	T	0.23	469	-4.14	(-5.55, -2.74)	1.45E-08	*SLC25A48/IL9*
rs7239554	18	A	0.28	469	-3.01	(-4.44, -1.59)	3.97E-05	*C18orf10*
rs6974557	7	T	0.26	469	-2.98	(-4.39, -1.57)	4.05E-05	*Loc646614*
rs6733728	2	C	0.38	469	-3.04	(-4.49, -1.60)	4.13E-05	*Loc402093*
rs2389096	13	T	0.24	468	3.03	(1.58, 4.47)	4.69E-05	*GPC6*
rs267280	3	G	0.47	469	-3.29	(-4.86, -1.72)	4.70E-05	*LARS2*

B)

SNP	Chr.	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs7326145	13	A	0.23	199	4.95	(2.81, 7.09)	1.01E-05	*COL4A2*
rs7837658	8	T	0.45	199	5.12	(2.86, 7.39)	1.62E-05	*RNF19A*
rs877356	5	T	0.23	199	-4.72	(-6.82, -2.61)	1.84E-05	*SLC25A48/IL9*
rs7944514	11	C	0.41	199	4.83	(2.65, 7.00)	2.19E-05	*POLD3*
rs2839520	21	A	0.27	199	4.45	(2.38, 6.56)	4.01E-05	*UBASH3A*
rs16872344	5	A	0.22	199	4.51	(2.40, 6.63)	4.51E-05	*IRX1*
rs13174381	5	A	0.23	199	4.47	(2.37, 6.58)	4.86E-05	*IRX1*
rs10085086	5	C	0.20	197	-4.76	(-7.01, -2.52)	4.90E-05	*Loc391738*

C)

SNP	Chr.	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs17062122	6	C	0.33	270	-4.69	(-6.56, -2.82)	1.54E-06	*Loc285735*
rs903281	10	G	0.29	270	-4.36	(-6.21, -2.51)	6.28E-06	*RAB18*
rs331086	5	C	0.46	270	-4.58	(-6.67, -2.49)	2.52E-05	*FBN2*

rs7137335	12	T	0.33	269	-4.02	(-5.89, -2.15)	3.53E-05	*SLC16A7*
rs7074813	10	G	0.33	270	-4.02	(-5.89, -2.14)	3.84E-05	*RAB18*
…	…	…	…	…	…	…	…	...
rs877356	5	T	0.23	270	-3.59	(-5.52, -1.67)	3.09E-04	*SLC25A48/IL9*

* adjusted for 10 principal components, sex, and cohort of origin

^ adjusted for 10 principal components and sex

Appendix Table 4-3. Single nucleotide polymorphisms associating with case/control tuberculin skin test induration status (< versus ≥ 5mm) below a $5x10^{-5}$ p value in an additive genetic model in A) the combined cohort*, B) the Ugandan cohort^, and C) the Tanzanian cohort^

A)

SNP	Chr.	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Gene
rs877356	5	T	0.23	469	0.33	(0.22, 0.49)	5.45E-08	*SLC25A48/IL9*
rs7808481	7	A	0.22	469	2.31	(1.60, 3.34)	8.67E-06	*Loc340268*
rs10804666	3	G	0.44	469	2.04	(1.47, 2.84)	2.28E-05	*NMNAT3*
rs8179938	3	A	0.37	469	2.01	(1.45, 2.79)	2.77E-05	*Loc643634*
rs4705073	5	C	0.45	469	1.91	(1.41, 2.59)	2.79E-05	*MIRN145*
rs964739	4	T	0.38	469	1.93	(1.41, 2.63)	3.67E-05	*KLHL8*

B)

SNP	Chr.	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Gene
rs877356	5	T	0.23	199	0.23	(0.12, 0.44)	1.41E-05	*SLC25A48/IL9*
rs9989936	21	G	0.40	199	0.25	(0.13, 0.48)	3.07E-05	*SAMSN1*

C)

SNP	Chr.	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Gene
rs642774	1	G	0.44	270	0.37	(0.24, 0.57)	6.26E-06	*UOX*
rs6589880	11	C	0.40	270	2.23	(1.52, 3.28)	4.18E-05	*Loc283155*
rs7808481	7	A	0.23	270	2.53	(1.62, 3.95)	4.74E-05	*Loc340268*
…	…	…	…	…	…	…	…	...
rs877356	5	T	0.23	270	0.38	(0.23, 0.65)	3.14E-04	*SLC25A48/IL9*

* adjusted for 10 principal components, sex, and cohort of origin

^ adjusted for 10 principal components and sex

Appendix Table 4-4. Single nucleotide polymorphisms associating with continuous tuberculin skin test induration below a $5x10^{-5}$ p value in an additive genetic model in A) the combined cohort*, B) the Ugandan cohort^, and C) the Tanzanian cohort^

A)

SNP	Chr.	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs877356	5	T	0.23	469	-3.34	(-4.53, -2.14)	6.95E-08	*SLC25A48/IL9*
rs7082209	10	G	0.40	469	-2.56	(-3.66, -1.46)	6.45E-06	*CXCL12*
rs7163504	15	G	0.43	467	-2.27	(-3.26, -1.28)	9.51E-06	*NEO1*
rs8179938	3	A	0.37	469	2.37	(1.30, 3.43)	1.56E-05	*Loc643636*
rs10881240	1	T	0.37	469	-2.23	(-3.25, -1.21)	2.25E-05	*Loc642337*
rs290185	11	A	0.46	469	2.18	(1.17, 3.19)	2.66E-05	*CCDC89*
rs7239554	18	A	0.28	469	-2.45	(-3.59, -1.32)	2.80E-05	*C18orf10*
rs11120119	1	G	0.45	466	-2.15	(-3.14, -1.15)	2.89E-05	*RPS6KC1*
rs7808481	7	A	0.22	469	2.57	(1.37, 3.77)	3.05E-05	*Loc340268*
rs6780136	3	T	0.33	468	-2.21	(-3.24, -1.18)	3.12E-05	*CPNE4*
rs12634351	3	C	0.33	469	-2.21	(-3.25, -1.18)	3.50E-05	*CPNE4*
rs10804666	3	G	0.44	469	2.25	(1.18, 3.31)	4.06E-05	*NMNAT3*
rs7326145	13	A	0.25	469	2.48	(1.30, 3.66)	4.29E-05	*COL4A2*

B)

SNP	Chr.	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs7326145	13	A	0.23	199	4.30	(2.44, 6.15)	1.00E-05	*COL4A2*
rs10085086	5	C	0.20	197	-4.21	(-6.13, -2.29)	2.80E-05	*Loc391738*
rs6545560	2	G	0.40	199	-3.20	(-4.68, -1.73)	3.15E-05	*CCDC85A*
…	…	…	…	…	…	…	…	...
rs877356	5	T	0.23	199	0.23	(0.12, 0.44)	1.25E-04	*SLC25A48/IL9*

C)

SNP	Chr.	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs642774	1	G	0.44	270	-3.10	(-4.46, -1.74)	1.11E-05	*UOX*
rs2727529	7	C	0.38	270	2.93	(1.60, 4.25)	2.07E-05	*PRKAG2*
rs312305	5	G	0.43	269	2.90	(1.57, 4.22)	2.55E-05	*GABRG2*
rs2141372	2	G	0.40	270	-3.10	(-4.52, -1.68)	2.62E-05	*ZNF512*
rs7808481	7	A	0.23	270	3.35	(1.80, 4.90)	3.06E-05	*Loc340268*
rs8103597	19	A	0.26	270	3.26	(1.74, 4.77)	3.48E-05	*SIPA1L3*
rs17062122	6	C	0.33	270	-3.13	(-4.60, -1.67)	3.85E-05	*Loc285735*
rs12469734	2	A	0.45	270	2.99	(1.58, 4.41)	4.58E-05	*ARL4C*
rs7906180	10	T	0.36	270	2.98	(1.57, 4.39)	4.81E-05	*Loc100128641*
…	…	…	…	…	…	…	…	...
rs877356	5	T	0.23	270	-3.05	(-4.65, -1.45)	2.22E-04	*SLC25A48/IL9*

* adjusted for 10 principal components, sex, and cohort of origin

^ adjusted for 10 principal components and sex

Appendix Table 4-5. Single nucleotide polymorphisms associating with case/control tuberculin skin test induration status (< versus ≥ 5mm) below a $5x10^{-5}$ p value in a recessive genetic model in A) the combined cohort*, B) the Ugandan cohort^, and C) the Tanzanian cohort^

A)

SNP	Chr.	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Gene
rs1293940	6	A	0.32	469	0.17	(0.080, 0.37)	7.21E-06	*ESR1*
rs2285513	19	A	0.40	469	0.26	(0.14, 0.49)	2.58E-05	*SBSN*
rs2434785	5	G	0.46	469	0.32	(0.18, 0.55)	3.73E-05	*Loc266786*
rs10804666	3	G	0.44	469	3.57	(1.95, 6.54)	3.94E-05	*NMNAT3*
rs4705073	5	C	0.45	469	2.99	(1.77, 5.05)	4.14E-05	*MIRN145*
rs2489772	1	C	0.43	469	3.43	(1.90, 6.20)	4.53E-05	*KAZN*

B)

SNP	Chr.	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Gene
rs4989483	16	G	0.49	199	0.16	(0.069, 0.38)	2.94E-05	*FLJ32252*
rs12412686	10	A	0.33	199	0.090	(0.028, 0.29)	4.81E-05	*CLRN3*

C)

SNP	Chr.	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Gene
rs7239336	18	C	0.48	270	3.94	(1.91, 6.37)	4.52E-05	*MIRN924*

* adjusted for 10 principal components, sex, and cohort of origin

^ adjusted for 10 principal components and sex

Appendix Table 4-6. Single nucleotide polymorphisms associating with continuous tuberculin skin test induration below a $5x10^{-5}$ p value in a recessive genetic model in A) the combined cohort*, B) the Ugandan cohort^, and C) the Tanzanian cohort^

A)

SNP	Chr.	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs2520696	13	A	0.47	469	-4.00	(-5.64, -2.36)	2.51E-06	*FRY*
rs2333021	14	A	0.33	469	5.16	(2.95, 7.36)	5.77E-06	*ZFYVE1*
rs10804666	3	G	0.44	469	4.30	(2.41, 6.19)	1.07E-05	*NMNAT3*
rs798957	13	A	0.37	469	-4.49	(-6.48, -2.51)	1.15E-05	*FRY*
rs2434785	5	G	0.46	469	-3.88	(-5.62, -2.15)	1.46E-05	*Loc266786*
rs2489772	1	C	0.43	469	4.22	(2.33, 6.11)	1.50E-05	*KAZN*
rs10957982	8	T	0.26	468	-6.27	(-9.18, -3.35)	3.05E-05	*ZBTB10*
rs753927	20	C	0.35	469	4.64	(2.48, 6.80)	3.07E-05	*FERMT1*
rs2285513	19	A	0.40	469	-4.22	(-6.19, -2.25)	3.33E-05	*SBSN*
rs1293940	6	A	0.32	469	-4.74	(-6.96, -2.52)	3.50E-05	*ESR1*
rs10776801	1	C	0.31	469	-5.09	(-7.48, -2.70)	3.53E-05	*Loc100130948*
rs10881240	1	T	0.37	469	-4.27	(-6.27, -2.26)	3.73E-05	*Loc642337*
rs7623698	3	A	0.24	469	-6.58	(-9.70, -3.46)	4.27E-05	*Loc285303*

B)

SNP	Chr.	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs10800363	1	T	0.44	199	-6.03	(-8.77, -3.29)	2.61E-05	*XCL1*
rs8014986	14	A	0.23	199	-8.27	(-12.04, -4.50)	2.73E-05	*HHIPL1*
rs7873440	9	T	0.41	199	-6.69	(-9.8, -3.58)	3.80E-05	*RG9MTD3*
rs4989483	16	G	0.49	199	-5.51	(-8.07, -2.95)	3.81E-05	*FLJ32252*
rs7676378	4	T	0.25	199	-8.99	(-13.21, -4.77)	4.60E-05	*FAT4*
rs7340961	4	C	0.41	199	-6.08	(-8.95, -3.21)	4.94E-05	*AGPAT9*

C)

SNP	Chr.	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs2333021	14	A	0.37	270	6.74	(4.05, 9.43)	1.61E-06	*ZFYVE1*
rs4462385	12	A	0.25	270	8.01	(4.41, 11.61)	1.88E-05	*SLC16A7*
rs8019592	14	T	0.20	270	11.53	(6.30, 16.75)	2.19E-05	*RAD51L1*
rs2395015	7	A	0.29	270	8.33	(4.52, 12.14)	2.56E-05	*SMURF1*
rs5767477	22	G	0.37	270	5.55	(2.96, 8.13)	3.64E-05	*TBC1D22A*

* adjusted for 10 principal components, sex, and cohort of origin
^ adjusted for 10 principal components and sex

Appendix Table 4-7. Association of SNPs with binary tuberculin skin test induration (at and above 5mm versus below) in the imputed SLC25A48/IL9 region of the combined cohort in a A) dominant and B) additive genetic model; adjusting for 10 principal components, sex, and cohort of origin

A)

SNP	Chr.	Position	Minor Allele	Imputation Certainty	MAF	n	Odds Ratio	95% Confidence Interval	p value
rs17169187	5	135163758	C	0.99	0.24	469	0.25	(0.16, 0.40)	4.57E-09
rs17169180	5	135161055	C	0.99	0.24	469	0.26	(0.17, 0.42)	9.36E-09
rs13167664	5	135162467	G	0.99	0.24	469	0.26	(0.17, 0.42)	9.36E-09
rs35520957	5	135163307	T	0.99	0.24	469	0.26	(0.17, 0.42)	9.36E-09
rs877356	5	135161418	T	Not Imputed	0.24	469	0.27	(0.17, 0.42)	1.22E-08

B)

SNP	Chr.	Position	Minor Allele	Imputation Certainty	MAF	n	Odds Ratio	95% Confidence Interval	p value
rs17169187	5	135163758	C	0.99	0.24	469	0.32	(0.21, 0.48)	2.56E-08
rs17169180	5	135161055	C	0.99	0.24	469	0.33	(0.22, 0.49)	4.47E-08
rs13167664	5	135162467	G	0.99	0.24	469	0.33	(0.22, 0.49)	4.47E-08
rs35520957	5	135163307	T	0.99	0.24	469	0.33	(0.22, 0.49)	4.47E-08
rs877356	5	135161418	T	Not Imputed	0.24	469	0.33	(0.22, 0.49)	5.45E-08

Appendix Table 4-8. Association of SNPs with continuous tuberculin skin test induration in the imputed *SLC25A48/IL9* region of the combined cohort in a A) dominant and B) additive genetic model; adjusting for 10 principal components, sex, and cohort of origin

A)

SNP	Chr.	Position	Minor Allele	Imputation Certainty	MAF	n	Beta	95% Confidence Interval	p value
rs17169187	5	135163758	C	0.99	0.24	469	-4.29	(-5.69, -2.88)	4.58E-09
rs17169180	5	135161055	C	0.99	0.24	469	-4.16	(-5.56, -2.75)	1.35E-08
rs13167664	5	135162467	G	0.99	0.24	469	-4.16	(-5.56, -2.75)	1.35E-08
rs35520957	5	135163307	T	0.99	0.24	469	-4.16	(-5.56, -2.75)	1.35E-08
rs877356	5	135161418	T	Not Imputed	0.24	469	-4.14	(-5.55, -2.74)	1.45E-08

B)

SNP	Chr.	Position	Minor Allele	Imputation Certainty	MAF	n	Beta	95% Confidence Interval	p value
rs17169187	5	135163758	C	0.99	0.24	469	-3.43	(-4.62, -2.24)	2.84E-08
rs17169180	5	135161055	C	0.99	0.24	469	-3.34	(-4.53, -2.15)	6.65E-08
rs13167664	5	135162467	G	0.99	0.24	469	-3.34	(-4.53, -2.15)	6.65E-08
rs35520957	5	135163307	T	0.99	0.24	469	-3.34	(-4.53, -2.15)	6.65E-08
rs877356	5	135161418	T	Not Imputed	0.24	469	-3.34	(-4.53, -2.14)	6.95E-08

Appendix Table 4-9. Interferon gamma release assay results by TST case/control status in A) the Ugandan cohort and B) the Tanzanian cohort

A)

IGRA Antigen	Mean PPD Negative	Mean PPD Positive	p value
n	25	89	
MEDIA	45.27 (75.76)	43.07 (116.80)	0.93
ESAT-6	30.14 (34.32)	143.73 (498.86)	0.30
CXFT	101.91 (205.62)	1495 (422.05)	0.18
CFP10	69.62 (123.05)	190.69 (713.37)	0.41

B)

IGRA Antigen	Mean PPD Negative	Mean PPD Positive	p value
n	161	79	
PHA	18708.11 (27709.84)	22263.52 (35855.16)	0.40
MEDIA	184.85 (412.16)	227.66 (522.57)	0.49
ESAT-6	444.77 (1598.35)	4027.45 (9705.44)	8.60E-06
Ag85	335.01 (1273.74)	1387.03 (2772.05)	0.0001
MVS	220.40 (416.09)	525.32 (1494.13)	0.017
WCL	1366.25 (3454.75)	7711.33 (14517.40)	3.21E-07

Appendix Table 4-10. Single nucleotide polymorphisms associating with A) tuberculin skin test induration case/control status (< versus ≥ 5mm) and B) continuous tuberculin skin test induration using a dominant genetic model in the combined cohort, below a $5x10^{-5}$ p value after removing patients with possible false negative TST results; adjusted for 10 principal components, sex, and cohort of origin

A)

SNP	CHR	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Gene
rs877356	5	T	0.22	453	0.27	(0.17, 0.42)	1.77E-08	*SLC25A48/IL9*
rs7808481	7	A	0.22	453	2.64	(1.69, 4.16)	2.14E-05	*Loc340268*
rs697635	12	T	0.24	453	0.39	(0.25, 0.61)	3.14E-05	*ANKRD33*
rs2389096	13	T	0.24	452	2.53	(1.62, 3.95)	4.14E-05	*GPC6*
rs1880386	10	A	0.21	453	2.52	(1.62, 3.93)	4.51E-05	*GRID1*
rs9584956	13	G	0.24	453	2.43	(1.58, 3.73)	4.92E-05	*DOCK9*

B)

SNP	CHR	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs877356	5	T	0.22	453	-4.22	(-5.69, -2.78)	1.68E-08	*SLC25A48/IL9*
rs697635	12	T	0.24	453	-3.20	(-4.65, -1.75)	1.95E-05	*ANKRD33*
rs7239554	18	A	0.28	453	-3.19	(-4.64, -1.74)	1.97E-05	*C18orf10*
rs7808481	7	A	0.22	453	3.17	(1.70, 4.65)	3.04E-05	*Loc340268*
rs9920077	15	A	0.46	453	3.43	(1.83, 5.04)	3.37E-05	*KIAA1024*
rs12454816	18	A	0.23	453	3.11	(1.65, 4.57)	3.50E-05	*CDH20*
rs2389096	13	T	0.24	452	3.12	(1.65, 4.60)	3.95E-05	*GPC6*

Appendix Table 4-11. Single nucleotide polymorphisms associating with A) tuberculin skin test induration case/control status (< versus ≥ 5mm) and B) continuous tuberculin skin test induration using a dominant genetic model in the combined cohort, below a $5x10^{-5}$ p value after removing patients with possible false positive TST reaction to a childhood BCG vaccine; adjusted for 10 principal components, sex, and cohort of origin

A)

SNP	CHR	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Gene
rs877356	5	T	0.23	449	0.27	(0.17, 0.44)	5.58E-08	*SLC25A48/IL9*
rs1880386	10	A	0.22	449	2.59	(1.66, 4.04)	2.93E-05	*GRID1*
rs7239554	18	A	0.28	449	0.40	(0.26, 0.61)	3.16E-05	*C18orf10*

B)

SNP	CHR	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs877356	5	T	0.23	449	-3.86	(-5.30, -2.42)	2.25E-07	*SLC25A48/IL9*
rs6733728	2	C	0.38	449	-3.23	(-4.69, -1.76)	1.91E-05	*Loc402093*
rs12454816	18	A	0.23	449	3.16	(1.71, 4.61)	2.46E-05	*CDH20*
rs9345216	6	C	0.22	449	3.11	(1.65, 4.57)	3.60E-05	*Loc100129847*
rs7239554	18	A	0.28	449	-3.07	(-4.52, -1.62)	3.91E-05	*C18orf10*
rs6744638	2	G	0.44	449	-3.19	(-4.70, -1.69)	3.93E-05	*Loc402093*
rs697635	12	T	0.24	449	-3.04	(-4.48, -1.58)	4.90E-05	*ANKRD33*

Appendix Table 4-12. Single nucleotide polymorphisms associating with A) tuberculin skin test induration case/control status (< versus ≥ 5mm) and B) continuous tuberculin skin test induration using a dominant genetic model in the combined cohort, below a $5x10^{-5}$ p value after removing patients with possible false positive TST reaction to a childhood BCG vaccine and possible false negative TST reactions; adjusted for 10 principal components, sex, and cohort of origin

A)

SNP	CHR	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Gene
rs877356	5	T	0.23	433	0.27	(0.17, 0.44)	7.44E-08	*SLC25A48/IL9*
rs7239554	18	A	0.28	433	0.38	(0.25, 0.59)	1.87E-05	*C18orf10*
rs1880386	10	A	0.22	433	2.65	(1.68, 4.18)	2.67E-05	*GRID1*
rs697635	12	T	0.25	433	0.38	(0.24, 0.60)	3.25E-05	*ANKRD33*
rs12473869	2	T	0.23	433	2.64	(1.67, 4.18)	3.51E-05	*Loc100131048*
rs7326145	13	A	0.25	433	2.58	(1.64, 4.05)	3.84E-05	*COL4A2*
rs492479	12	C	0.23	433	0.39	(0.25, 0.61)	4.00E-05	*KSR2*
rs10263964	7	C	0.36	433	0.40	(0.25, 0.62)	4.77E-05	*CNTNAP2*

B)

SNP	CHR	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs877356	5	T	0.23	433	-3.95	(-5.42, -2.48)	2.39E-07	*SLC25A48/IL9*
rs12454816	18	A	0.23	433	3.40	(1.91, 4.88)	9.27E-06	*CDH20*
rs697635	12	T	0.25	433	-3.31	(-4.79, -1.83)	1.52E-05	*ANKRD33*
rs9345216	6	C	0.25	433	3.28	(1.78, 4.77)	2.14E-05	*Loc100129847*
rs7239554	18	A	0.28	433	-3.25	(-4.72, -1.76)	2.14E-05	*C18orf10*
rs6733728	2	C	0.38	433	-3.16	(-4.66, -1.67)	3.99E-05	*Loc402093*
rs9920077	15	A	0.46	433	3.44	(1.80, 5.07)	4.49E-05	*KIAA1024*
rs10263964	7	C	0.36	433	-3.12	(-4.61, -1.63)	4.75E-05	*CNTNAP2*

Appendix Table 4-13. Association of SNPs with A) case/control tuberculin skin test induration status (< versus ≥ 5mm) and B) continuous tuberculin skin test induration in the combined cohort using a dominant genetic model; adjusting for 10 principal components, sex, cohort of origin, and missing IGRA

A)

SNP	Chr.	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Gene
rs877356	5	T	0.23	469	0.27	(0.17, 0.43)	1.78E-08	*SLC25A48/IL9*
rs7808481	7	A	0.22	469	2.55	(1.64, 3.95)	3.00E-05	*Loc340268*
rs12781609	10	T	0.32	469	0.41	(0.27, 0.63)	4.37E-05	*C10orf93*
rs2389096	13	T	0.24	468	2.48	(1.60, 3.83)	4.49E-05	*GPC6*

B)

SNP	Chr.	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs877356	5	T	0.23	469	-4.09	(-5.50, -2.69)	2.12E-08	*SLC25A48/IL9*
rs6974557	7	T	0.26	469	-3.06	(-4.47, -1.66)	2.44E-05	*Loc100128056*
rs2389096	13	T	0.24	468	3.06	(1.62, 4.51)	3.68E-05	*GPC6*
rs7239554	18	A	0.28	469	-3.00	(-4.42, -1.58)	4.08E-05	*C18orf10*

Appendix Table 4-14. Association of SNPs with A) case/control tuberculin skin test induration status (< versus ≥ 5mm) and B) continuous tuberculin skin test induration using a dominant genetic model in the combined cohort, including patients with prior TB; adjusting for 10 principal components, sex, and cohort of origin

A)

SNP	Chr.	Minor Allele	MAF	n	Odds Ratio	95% Confidence Interval	p value	Gene
rs877356	5	T	0.23	481	0.27	(0.17, 0.43)	1.56E-08	*SLC25A48/IL9*
rs12781609	10	T	0.32	481	0.42	(0.27, 0.63)	3.82E-05	*C10orf93*
rs7808481	7	A	0.21	481	2.47	(1.60, 3.81)	4.06E-05	*Loc340268*

B)

SNP	Chr.	Minor Allele	MAF	n	Beta	95% Confidence Interval	p value	Gene
rs877356	5	T	0.23	481	-4.05	(-5.44, -2.67)	1.70E-08	*SLC25A48/IL9*
rs6733728	2	C	0.38	481	-2.99	(-4.40, -1.57)	4.19E-05	*Loc402093*
rs16827624	2	A	0.24	481	-2.92	(-4.31, -1.52)	4.84E-05	*Loc100131051*
rs10051419	5	C	0.49	480	3.41	(1.78, 5.04)	4.86E-05	*OR7H2P*

Appendix Table 4-15. Association of the rs877356-rs2069885 haplotype using additive genetic models for both SNPs with TST induration case/control status (< versus ≥ 5mm) in the *SLC25A48/IL9* region in A) the combined cohort, B) the Ugandan cohort and C) the Tanzanian cohort

A)

Haplotype	Case	Control	Ca-Freq	Co-Freq
C-A	43	17	0.09	0.04
C-G	360	301	0.74	0.67
T-A	7	12	0.01	0.03
T-G	76	120	0.16	0.27
Likelihood ratio chisq = 42.90 df = 3 p-value = 2.59E-09*				

B)

Haplotype	Case	Control	Ca-Freq	Co-Freq
C-A	26	3	0.09	0.03
C-G	216	60	0.72	0.61
T-A	2	4	0.01	0.04
T-G	56	31	0.19	0.32
Likelihood ratio chisq = 25.84 df = 3 p-value = 1.03E-05^				

C)

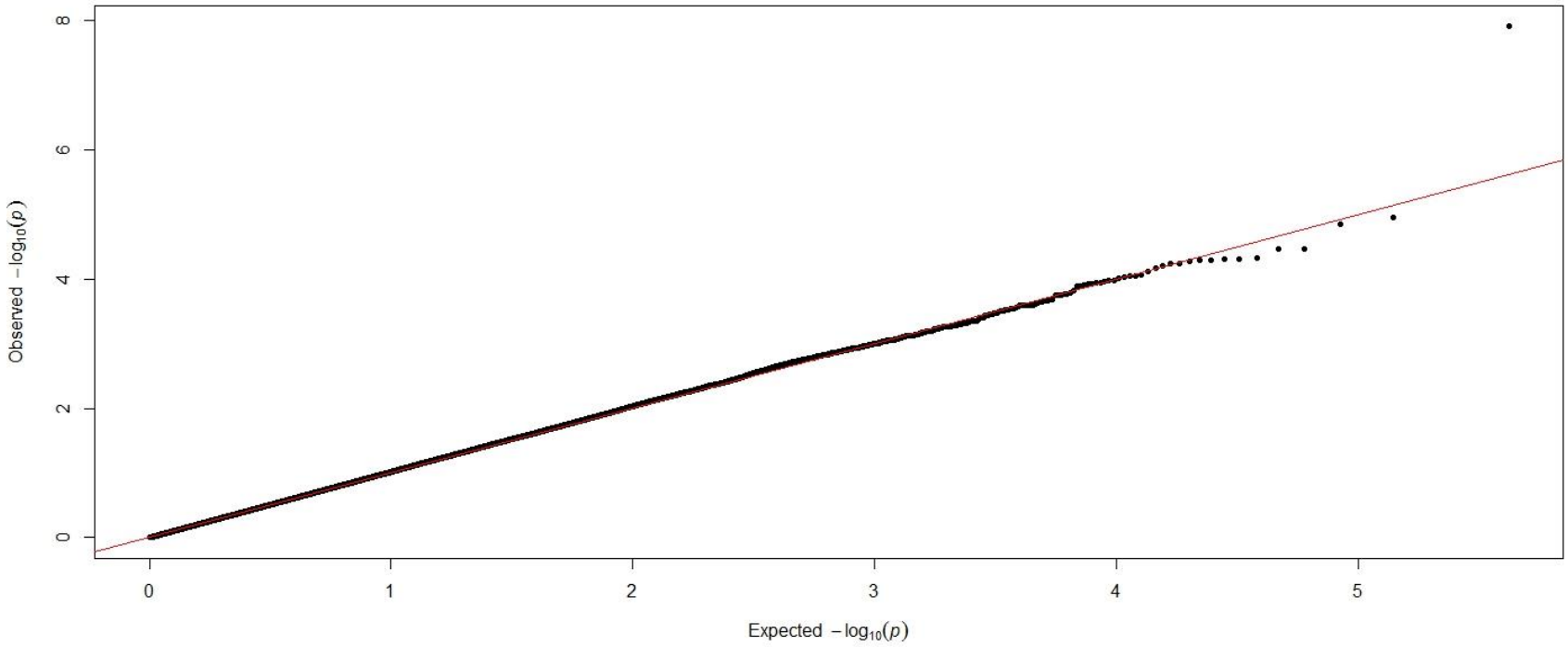
Haplotype	Case	Control	Ca-Freq	Co-Freq
C-A	18	14	0.10	0.04
C-G	143	241	0.74	0.67
T-A	4	8	0.01	0.03
T-G	21	89	0.16	0.27
Likelihood ratio chisq = 22.06 df = 3 p-value = 6.34E-05^				

* adjusted for 10 principal components, sex, and cohort of origin
^ adjusted for 10 principal components and sex

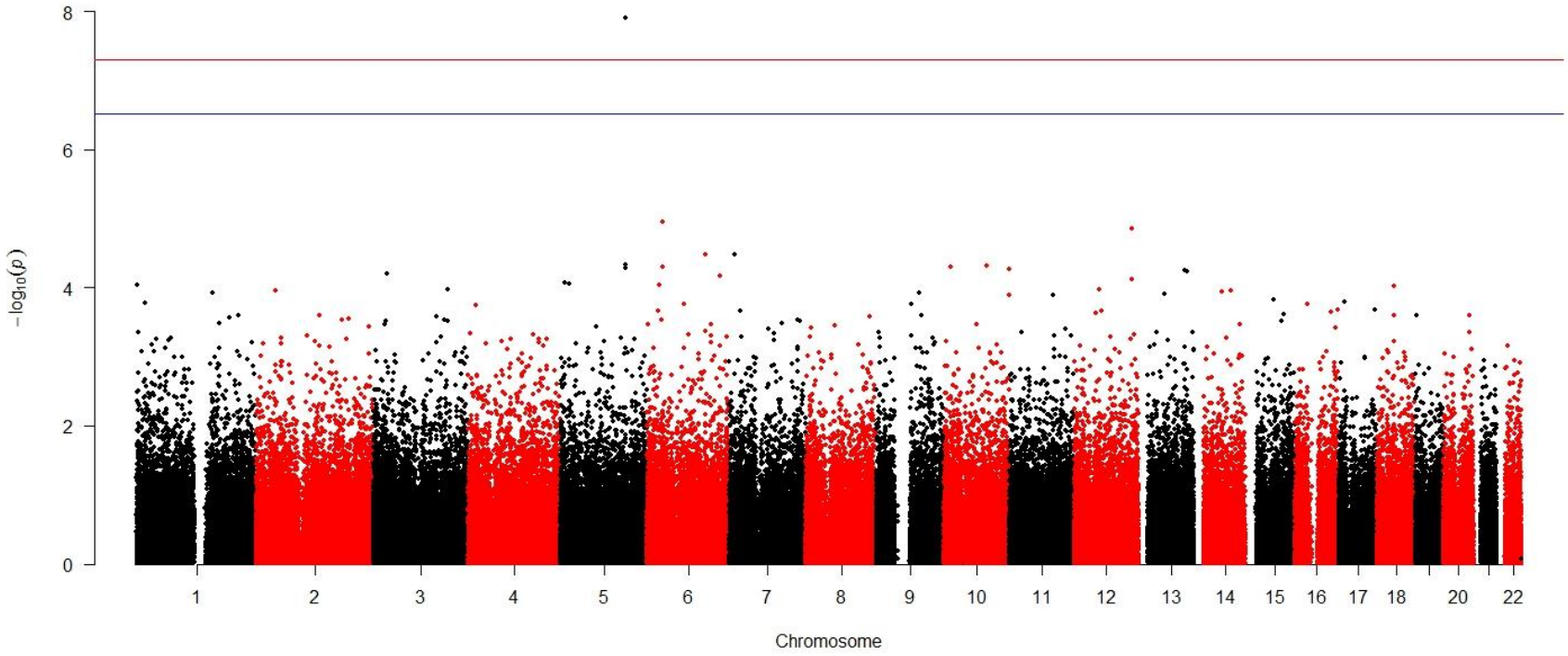
Appendix Table 4-16. Single nucleotide polymorphisms associating with TST induration case/control status (< versus ≥ 5mm) in the *GAS2* region in the combined cohort, adjusted for 10 principal components, sex, and cohort of origin

SNP	Position	Minor Allele	Odds Ratio	95% Confidence Interval	p value
rs141215155	22716709	C	0.32	(0.14, 0.72)	0.0061
rs147337656	22717086	A	0.32	(0.14, 0.72)	0.0061
rs140137885	22722819	G	0.32	(0.14, 0.72)	0.0061
rs143310525	22724201	C	0.32	(0.14, 0.72)	0.0061
rs147250499	22751877	T	0.32	(0.14, 0.72)	0.0061
rs149974430	22711780	A	0.35	(0.16, 0.77)	0.0097
rs146563332	22711993	C	0.35	(0.16, 0.77)	0.0097
rs113171697	22712670	A	0.52	(0.31, 0.87)	0.013
rs148539033	22822106	T	0.37	(0.17, 0.82)	0.013

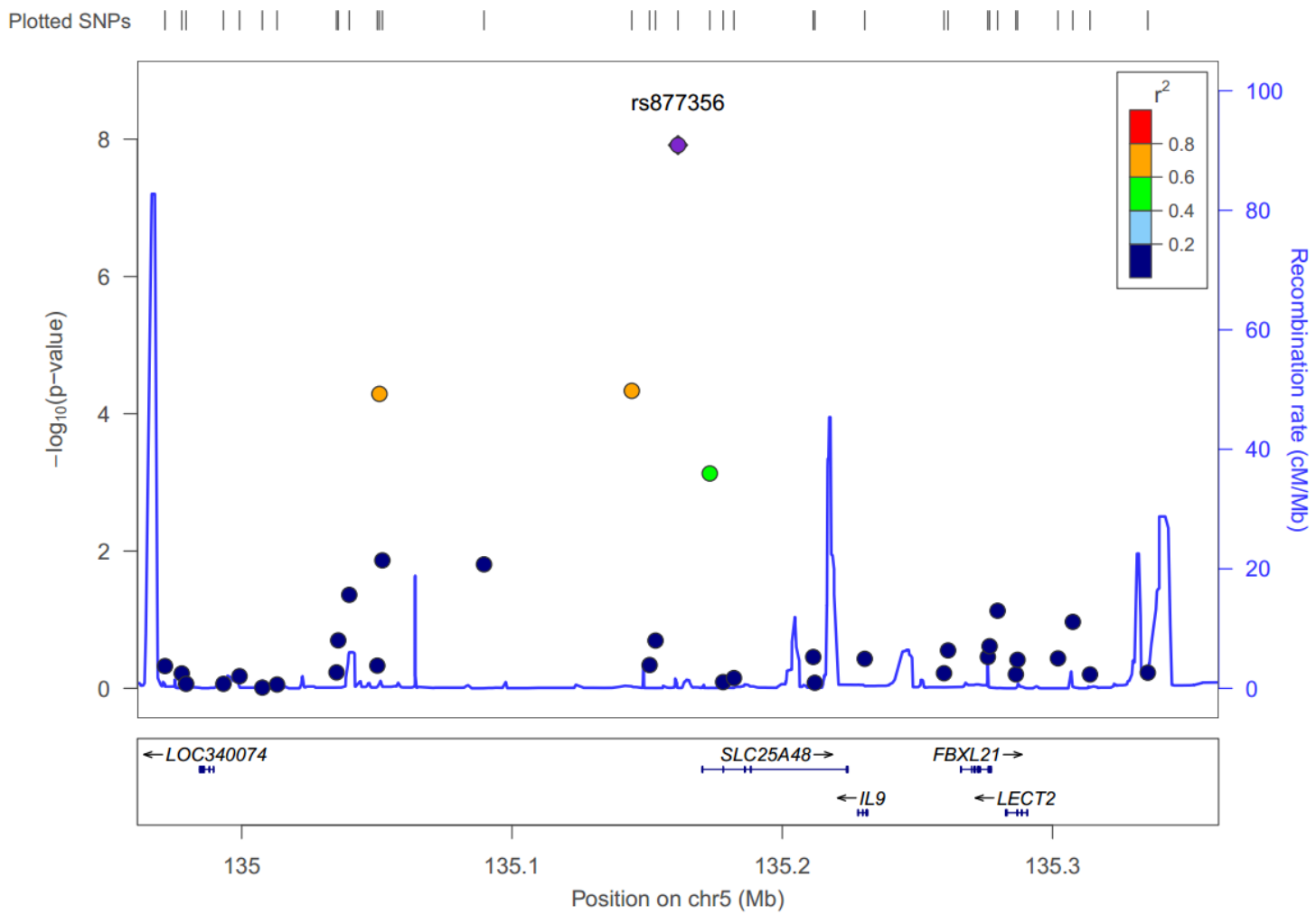
Appendix Figure 4-1. QQ plot of results from a logistic regression association of case/control tuberculin skin test induration status (< versus ≥ 5mm) with a dominant genetic model of available SNPs for the combined Ugandan and Tanzanian datasets



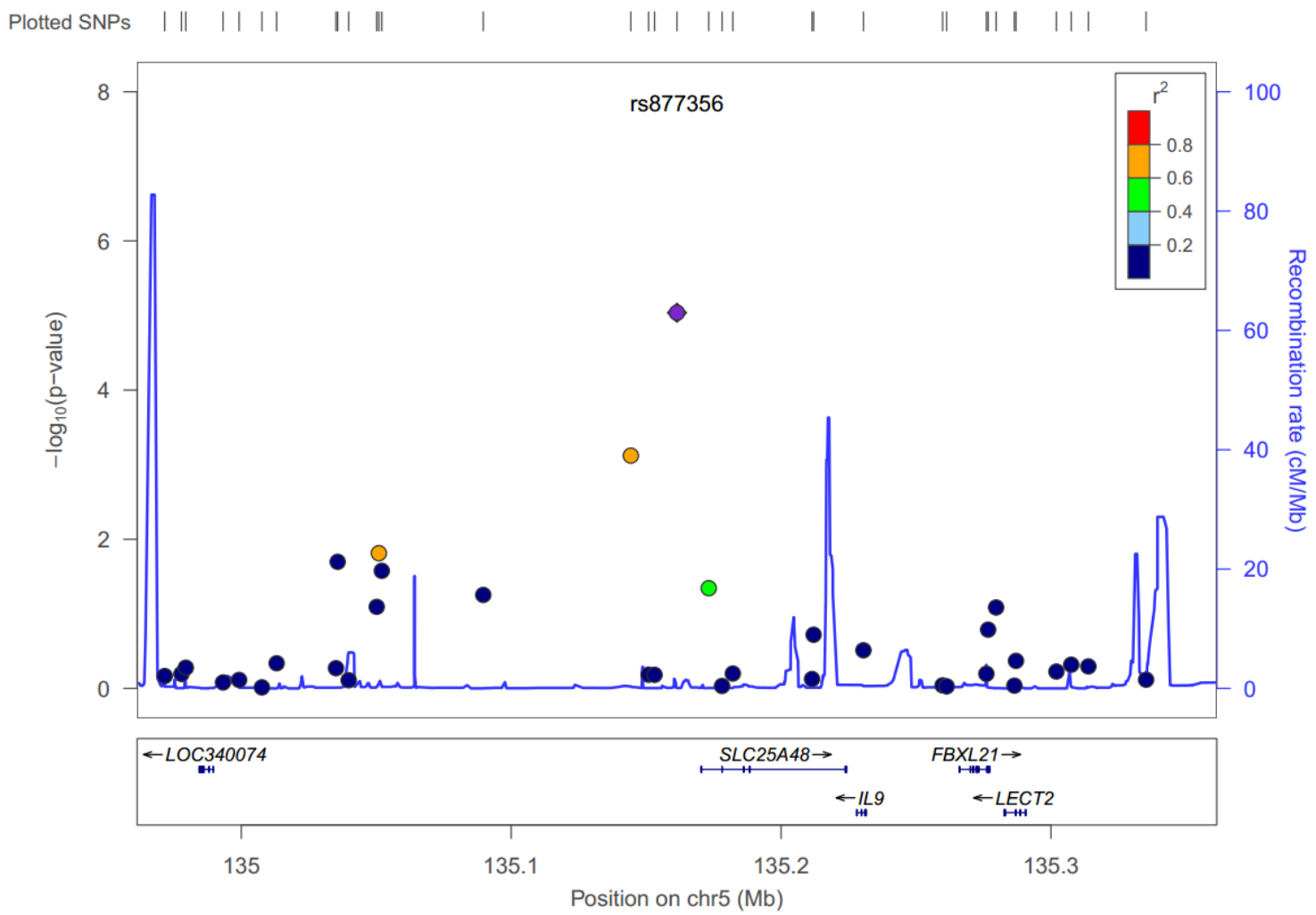
Appendix Figure 4-2. Manhattan plot of results from a logistic regression association of case/control tuberculin skin test induration status (< versus ≥ 5mm) with a dominant genetic model of available SNPs for the combined Ugandan and Tanzanian datasets; blue line – Bonferroni adjusted significance threshold, red line – genome-wide significance threshold



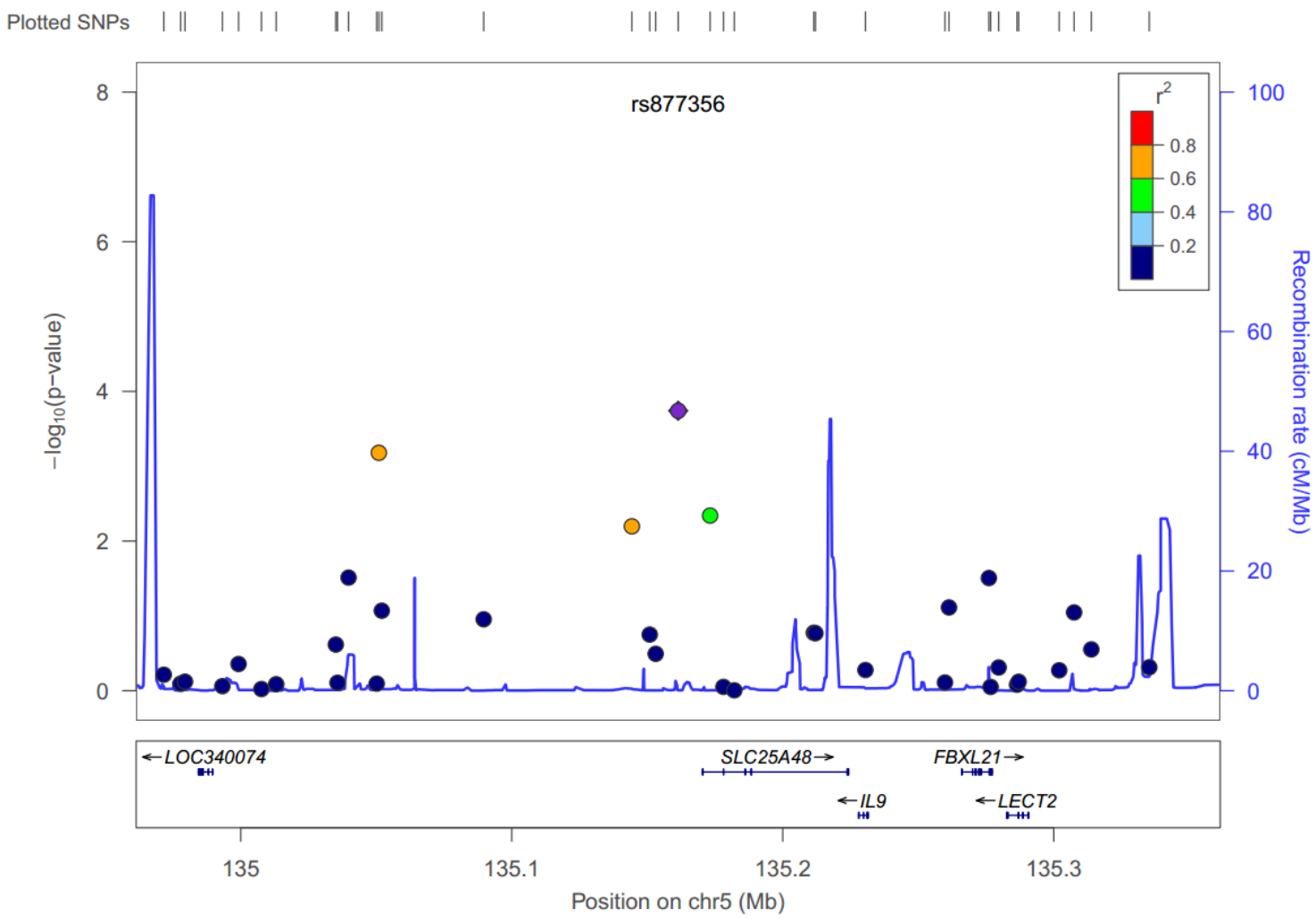
Appendix Figure 4-3. Locus zoom plot of results from a logistic regression of case/control tuberculin skin test induration status (< versus ≥ 5mm) with SNPs in the *SLC25A48/IL9* region using a dominant genetic model in the combined cohort, adjusted for 10 principal components, sex, and cohort of origin



Appendix Figure 4-4. Locus zoom plot of results from a logistic regression association of case/control tuberculin skin test induration status (< versus ≥ 5mm) with a dominant genetic model of SNPs in the *SLC25A48/IL9* region in the Ugandan cohort, adjusted for 10 principal components and sex

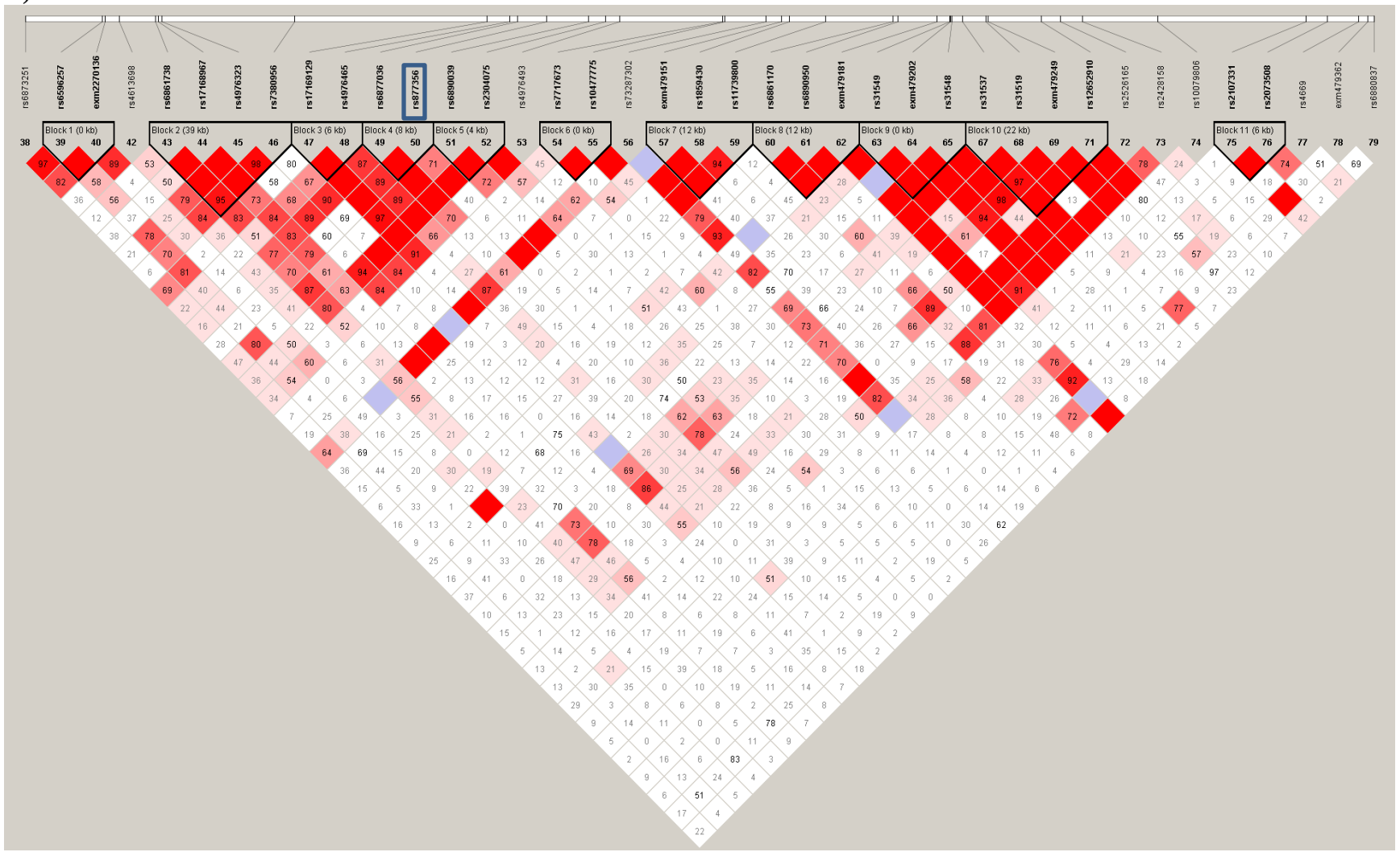


Appendix Figure 4-5. Locus zoom plot of results from a logistic regression association of case/control tuberculin skin test induration status (< versus ≥ 5mm) with a dominant genetic model of SNPs in the *SLC25A48/IL9* region in the Tanzanian cohort, adjusted for 10 principal components and sex

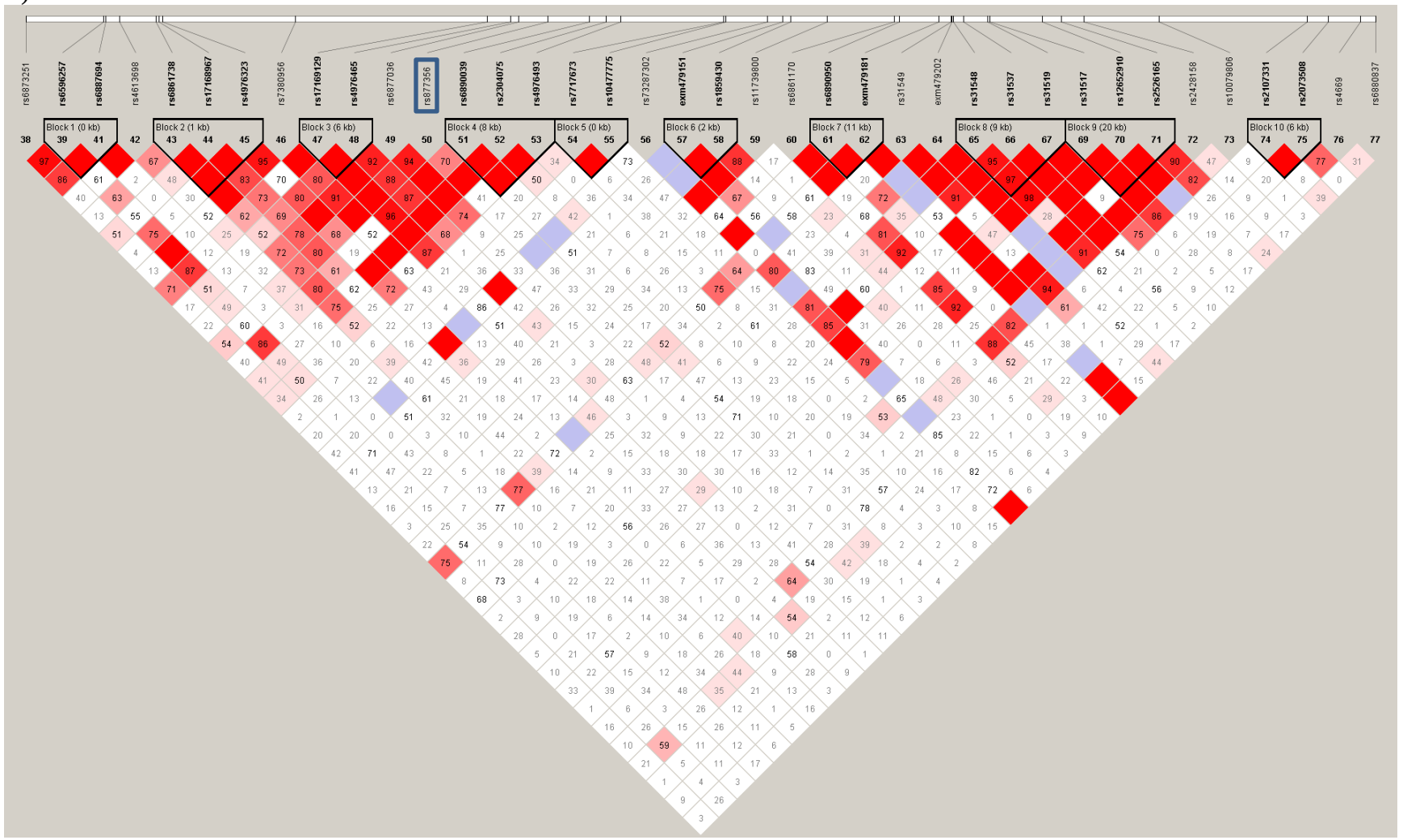


Appendix Figure 4-6. Haploview plots of the rs877356 region (384 kb range) using the D' metric in A) the Tanzanian cohort and B) the Ugandan cohort

A)

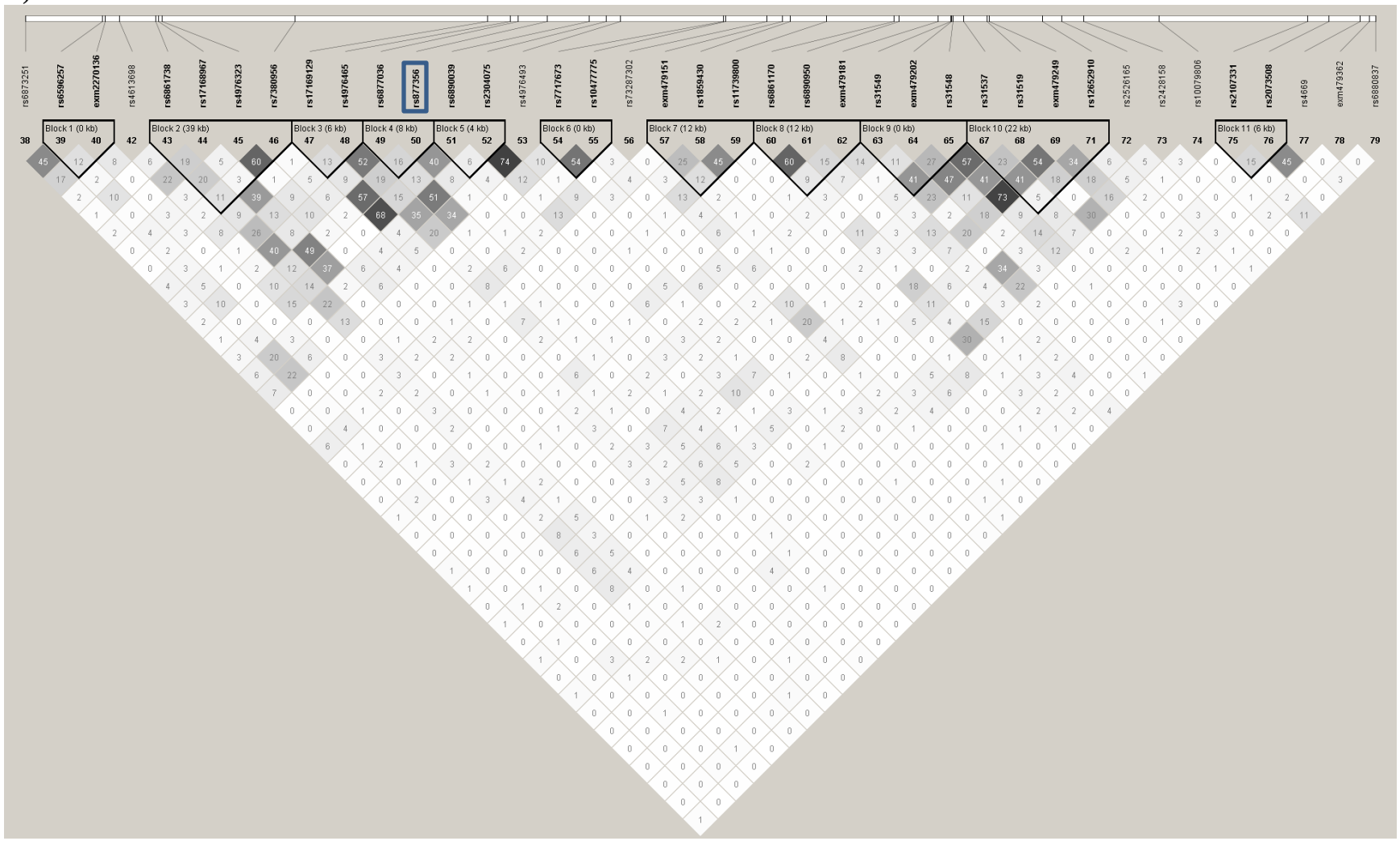


B)

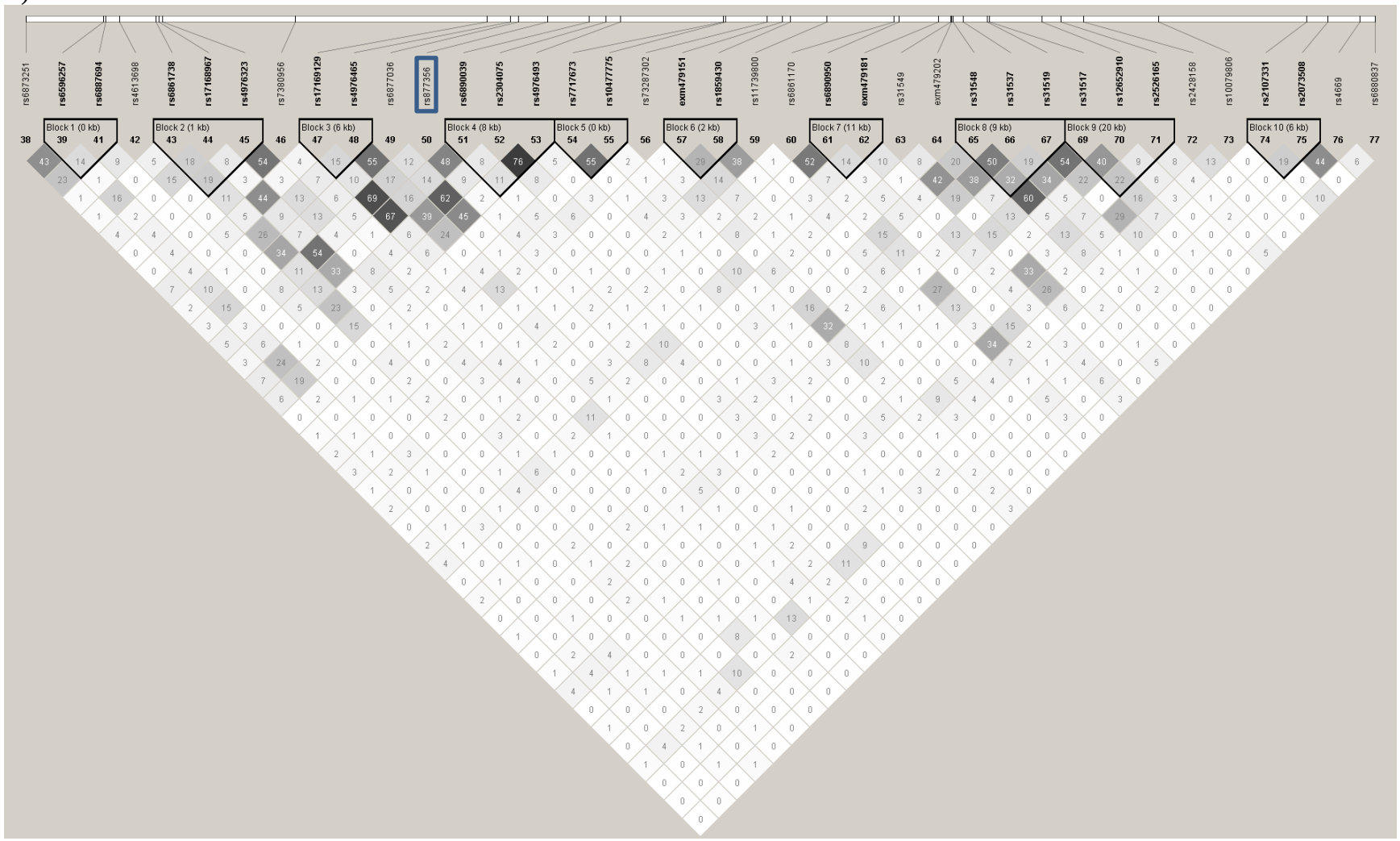

rs6873251
rs6596257
rs6887694
rs4613698
rs6861738
rs17168967
rs4976323
rs7380956
rs17169129
rs4976465
rs6877036
rs877356
rs6890039
rs2304075
rs4976493
rs7717673
rs10477775
rs73287302
exm479151
rs1859430
rs11739800
rs6861170
rs6890950
exm479181
rs31549
exm479202
rs31548
rs31537
rs31519
rs31517
rs12652910
rs2526165
rs2428158
rs10079606
rs2107331
rs2073508
rs4669
rs6880837
Block 1 (0 kb)
Block 2 (1 kb)
Block 3 (6 kb)
Block 4 (8 kb)
Block 5 (0 kb)
Block 6 (2 kb)
Block 7 (11 kb)
Block 8 (9 kb)
Block 9 (20 kb)
Block 10 (6 kb)

Appendix Figure 4-7. Haploview plots of the rs877356 region (384 kb region) using the r^2 metric in A) the Tanzanian cohort and B) the Ugandan cohort

A)



B)


rs6873251
rs6596257
rs6887694
rs4613698
rs6861738
rs17168967
rs4976323
rs7380956
rs17169129
rs4976465
rs6877036
rs877356
rs6890039
rs2304075
rs4976493
rs7717673
rs10477775
rs73287302
exm479151
rs1859430
rs11739800
rs6861170
rs6890950
exm479181
rs31549
exm479202
rs31548
rs31537
rs31519
rs31517
rs12652910
rs2526165
rs2428158
rs10079806
rs2107331
rs2073508
rs4669
rs6880837
Block 1 (0 kb)
Block 2 (1 kb)
Block 3 (6 kb)
Block 4 (8 kb)
Block 5 (0 kb)
Block 6 (2 kb)
Block 7 (11 kb)
Block 8 (9 kb)
Block 9 (20 kb)
Block 10 (6 kb)

Appendix Table 5-1. MDR metrics of the top model associating with TB disease using A) variants in or near *IL12B* (rs4921437), and *SLC6A3* (rs7737692) and cohort of origin as covariates and B) variants in or near *IL12B* (rs4921437), *TNF-α* (rs2242656), and *CR1* (rs2940252) and cohort of origin as covariates

A)

Cohort	Testing balanced accuracy	Cross-validation consistency
Combined	0.68	10/10

B)

Cohort	Testing balanced accuracy	Cross-validation consistency
Combined	0.66	10/10

Appendix Table 5-2. MDR metrics of top loci and interactions associating with TB disease for variants on chromosome 5 omitting rs4921437 from analyses in A) two-locus and B) three-locus analyses

A)

SNP1	SNP2	Training Balanced Accuracy	Testing Balanced Accuracy	Cross-Validation Consistency	Cohort
rs460000	rs7704367	0.63	0.48	3/10	Tanzania
rs460000	rs6869411	0.66	0.46	2/10	Uganda
rs460000	rs7704367	0.61	0.54	4/10	Combined*

*adjusted for cohort of origin

B)

SNP1	SNP2	SNP3	Training Balanced Accuracy	Testing Balanced Accuracy	Cross-Validation Consistency	Cohort
rs31489	rs460000	rs7704367	0.70	0.49	4/10	Tanzania
rs2853672	rs460000	rs4921227	0.73	0.44	3/10	Uganda
rs7737692	rs1295686	rs6869411	0.66	0.55	2/10	Combined*

*adjusted for cohort of origin

Appendix Table 5-3. MDR metrics of the top models for all available SNPs associating with TB disease omitting variant rs4921437 in A) two-locus and B) three-locus analyses

A)

SNP1	SNP2	Training Balanced Accuracy	Testing Balanced Accuracy	Cross-Validation Consistency	Cohort
rs6436917	rs36987	0.64	0.47	4/10	Tanzania
rs3129943	rs12788021	0.71	0.70	10/10	Uganda
rs4680367	rs7496458	0.61	0.54	4/10	Combined*

*adjusted for cohort of origin

B)

SNP1	SNP2	SNP3	Training Balanced Accuracy	Testing Balanced Accuracy	Cross-Validation Consistency	Cohort
rs16827657	rs7704367	rs4808737	0.69	0.41	1/10	Tanzania
rs6793453	rs1265761	rs12788021	0.76	0.68	4/10	Uganda
rs460000	rs6887695	rs372889	0.66	0.53	1/10	Combined*

*adjusted for cohort of origin

Appendix Table 5-4. MDR metrics of the top model associating with TB disease using A) variants near *IL9* (rs877356), and *SLC6A3* (rs931709) and cohort of origin as covariates and B) variants in or near *IL9* (rs877356), *RAB6C* (rs2521933), *SLC6A3* (rs17597967) and cohort of origin as covariates

A)

Cohort	Testing balanced accuracy	Cross-validation consistency
Combined	0.66	10/10

B)

Cohort	Testing balanced accuracy	Cross-validation consistency
Combined	0.67	10/10

Appendix Table 5-5. MDR metrics of top loci and interactions associating with MTB infection for variants on chromosome 5 omitting rs877356 from analyses in A) two-locus and B) three-locus analyses

A)

SNP1	SNP2	Training Balanced Accuracy	Testing Balanced Accuracy	Cross-Validation Consistency	Cohort
rs1835874	rs17169129	0.64	0.49	4/10	Tanzania
rs872015	rs6865443	0.67	0.45	3/10	Uganda
rs931709	rs17169129	0.61	0.56	7/10	Combined*

*adjusted for cohort of origin

B)

SNP1	SNP2	SNP3	Training Balanced Accuracy	Testing Balanced Accuracy	Cross-Validation Consistency	Cohort
rs401681	rs7705049	rs1835874	0.71	0.53	4/10	Tanzania
rs12523324	rs981883	rs2304075	0.75	0.44	2/10	Uganda
rs12515850	rs981883	rs17169129	0.66	0.63	9/10	Combined*

*adjusted for cohort of origin

Appendix Table 5-6. MDR metrics of the top models for all available SNPs associating with MTB infection omitting variant rs877356 in A) two-locus and B) three-locus analyses

A)

SNP1	SNP2	Training Balanced Accuracy	Testing Balanced Accuracy	Cross-Validation Consistency	Cohort
rs4848628	rs17005295	0.64	0.50	5/10	Tanzania
rs6865443	rs10833965	0.71	0.65	8/10	Uganda
rs2215813	rs17234274	0.63	0.59	6/10	Combined*

*adjusted for cohort of origin

B)

SNP1	SNP2	SNP3	Training Balanced Accuracy	Testing Balanced Accuracy	Cross-Validation Consistency	Cohort
rs4848628	rs17005295	rs27061	0.70	0.53	3/10	Tanzania
rs3094419	rs2521930	rs10833965	0.77	0.52	2/10	Uganda
rs2521920	rs2521933	rs6865443	0.68	0.58	2/10	Combined*

*adjusted for cohort of origin

REFERENCES

1. UNAIDS. UNAIDS report on the global AIDS epidemic. (2013).
2. Programme, W.H.O.G.T. Global tuberculosis report 2014. (2014).
3. Selwyn, P.A. *et al.* Prospective study of human immunodeficiency virus infection and pregnancy outcomes in intravenous drug users. *JAMA* **261**, 1289-94 (1989).
4. Di Perri, G. *et al.* Nosocomial epidemic of active tuberculosis among HIV-infected patients. *Lancet* **2**, 1502-4 (1989).
5. Vidal, S.M., Malo, D., Vogan, K., Skamene, E. & Gros, P. Natural resistance to infection with intracellular parasites: isolation of a candidate for Bcg. *Cell* **73**, 469-85 (1993).
6. Liu, J. *et al.* Identification of polymorphisms and sequence variants in the human homologue of the mouse natural resistance-associated macrophage protein gene. *Am J Hum Genet* **56**, 845-53 (1995).
7. Picard, C. *et al.* Inherited interleukin-12 deficiency: IL12B genotype and clinical phenotype of 13 patients from six kindreds. *Am J Hum Genet* **70**, 336-48 (2002).
8. Morris, G.A. *et al.* Interleukin 12B (IL12B) genetic variation and pulmonary tuberculosis: a study of cohorts from The Gambia, Guinea-Bissau, United States and Argentina. *PLoS One* **6**, e16656 (2011).
9. Ma, X. *et al.* Association between interleukin-8 gene alleles and human susceptibility to tuberculosis disease. *J Infect Dis* **188**, 349-55 (2003).
10. Thye, T. *et al.* Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat Genet* **42**, 739-41 (2010).
11. Thye, T. *et al.* Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nat Genet* **44**, 257-9 (2012).
12. Cobat, A. *et al.* Two loci control tuberculin skin test reactivity in an area hyperendemic for tuberculosis. *J Exp Med* **206**, 2583-91 (2009).
13. Stein, C.M. *et al.* Genome scan of M. tuberculosis infection and disease in Ugandans. *PLoS One* **3**, e4094 (2008).
14. Hall, N.B. *et al.* Polymorphisms in TICAM2 and IL1B are associated with TB. *Genes Immun* (2014).
15. Kaslow RA, M.J., Hill AVS. *Genetic Susceptibility to Infectious Diseases*, (Oxford University Press, 2008).
16. Sahiratmadja, E. *et al.* Association of polymorphisms in IL-12/IFN-gamma pathway genes with susceptibility to pulmonary tuberculosis in Indonesia. *Tuberculosis (Edinb)* **87**, 303-11 (2007).
17. von Reyn, C.F. *et al.* Prevention of tuberculosis in Bacille Calmette-Guerin-primed, HIV-infected adults boosted with an inactivated whole-cell mycobacterial vaccine. *AIDS* **24**, 675-85 (2010).
18. Ritchie, M.D. *et al.* Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Hum Genet* **69**, 138-47 (2001).
19. Hahn, L.W., Ritchie, M.D. & Moore, J.H. Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* **19**, 376-82 (2003).

20. Cambier, C.J., Falkow, S. & Ramakrishnan, L. Host evasion and exploitation schemes of Mycobacterium tuberculosis. *Cell* **159**, 1497-509 (2014).
21. Kumar V, A.A., Fausto N, Aster JC. *Robbins & Cotran Pathologic Basis of Disease*, (Saunders 2010).
22. Janeway CA, T.P. *Immunobiology: the Immune System in Health and Disease*, (Garland Science Publishing, 2005).
23. CDC. TB Elimination: the difference between latent TB infection and TB disease. (2011).
24. Jones, B.E. *et al.* Relationship of the manifestations of tuberculosis to CD4 cell counts in patients with human immunodeficiency virus infection. *Am Rev Respir Dis* **148**, 1292-7 (1993).
25. Perronne, C. *et al.* [Tuberculosis in patients infected with the human immunodeficiency virus. 30 cases]. *Presse Med* **17**, 1479-83 (1988).
26. Gourevitch, M.N., Hartel, D., Selwyn, P.A., Schoenbaum, E.E. & Klein, R.S. Effectiveness of isoniazid chemoprophylaxis for HIV-infected drug users at high risk for active tuberculosis. *AIDS* **13**, 2069-74 (1999).
27. Gold, C.Q.-T.
28. CDC. Interferon-Gamma Release Assays (IGRAs) - Blood Tests for TB Infection. (2011).
29. T.SPOT®.TB, O.
30. CDC. Testing for Tuberculosis.
31. CDC. Treatment of tuberculosis. (2003).
32. Modongo, C. *et al.* Successful MDR-TB treatment regimens including amikacin are associated with high rates of hearing loss. *BMC Infect Dis* **14**, 542 (2014).
33. Neyrolles, O. & Quintana-Murci, L. Sexual inequality in tuberculosis. *PLoS Med* **6**, e1000199 (2009).
34. Corbett, E.L. *et al.* The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Arch Intern Med* **163**, 1009-21 (2003).
35. Maurer, U. *et al.* The Wilms' tumor gene product (WT1) modulates the response to 1,25-dihydroxyvitamin D3 by induction of the vitamin D receptor. *J Biol Chem* **276**, 3727-32 (2001).
36. Sciesielski, L.K., Kirschner, K.M., Scholz, H. & Persson, A.B. Wilms' tumor protein Wt1 regulates the Interleukin-10 (IL-10) gene. *FEBS Lett* **584**, 4665-71 (2010).
37. Singh, S.P., Mehra, N.K., Dingley, H.B., Pande, J.N. & Vaidya, M.C. HLA-DR associated genetic control of pulmonary tuberculosis in north India. *Indian J Chest Dis Allied Sci* **25**, 252-8 (1983).
38. Malo, D., Vidal, S., Lieman, J.H., Ward, D.C. & Gros, P. Physical delineation of the minimal chromosomal segment encompassing the murine host resistance locus Bcg. *Genomics* **17**, 667-75 (1993).
39. Awomoyi, A.A. *et al.* Interleukin-10, polymorphism in SLC11A1 (formerly NRAMP1), and susceptibility to tuberculosis. *J Infect Dis* **186**, 1808-14 (2002).
40. Gao, P.S. *et al.* Genetic variants of NRAMP1 and active tuberculosis in Japanese populations. International Tuberculosis Genetics Team. *Clin Genet* **58**, 74-6 (2000).
41. Delgado, J.C., Baena, A., Thim, S. & Goldfeld, A.E. Ethnic-specific genetic associations with pulmonary tuberculosis. *J Infect Dis* **186**, 1463-8 (2002).

42. Ryu, S. *et al.* 3'UTR polymorphisms in the NRAMP1 gene are associated with susceptibility to tuberculosis in Koreans. *Int J Tuberc Lung Dis* **4**, 577-80 (2000).
43. Ma, X. *et al.* 5' dinucleotide repeat polymorphism of NRAMP1 and susceptibility to tuberculosis among Caucasian patients in Houston, Texas. *Int J Tuberc Lung Dis* **6**, 818-23 (2002).
44. Pan, H. *et al.* Ipr1 gene mediates innate immunity to tuberculosis. *Nature* **434**, 767-72 (2005).
45. Tosh, K. *et al.* Variants in the SP110 gene are associated with genetic susceptibility to tuberculosis in West Africa. *Proc Natl Acad Sci U S A* **103**, 10364-8 (2006).
46. Flores-Villanueva, P.O. *et al.* A functional promoter polymorphism in monocyte chemoattractant protein-1 is associated with increased susceptibility to pulmonary tuberculosis. *J Exp Med* **202**, 1649-58 (2005).
47. Bellamy, R. *et al.* Assessment of the interleukin 1 gene cluster and other candidate gene polymorphisms in host susceptibility to tuberculosis. *Tuber Lung Dis* **79**, 83-9 (1998).
48. Shin, H.D. *et al.* Common interleukin 10 polymorphism associated with decreased risk of tuberculosis. *Exp Mol Med* **37**, 128-32 (2005).
49. Abhimanyu *et al.* Differential serum cytokine levels are associated with cytokine gene polymorphisms in north Indians with active pulmonary tuberculosis. *Infect Genet Evol* **11**, 1015-22 (2011).
50. Seah, G.T. & Rook, G.A. High levels of mRNA encoding IL-4 in unstimulated peripheral blood mononuclear cells from tuberculosis patients revealed by quantitative nested reverse transcriptase-polymerase chain reaction; correlations with serum IgE levels. *Scand J Infect Dis* **33**, 106-9 (2001).
51. Olesen, R. *et al.* DC-SIGN (CD209), pentraxin 3 and vitamin D receptor gene variants associate with pulmonary tuberculosis risk in West Africans. *Genes Immun* **8**, 456-67 (2007).
52. Fitness, J. *et al.* Large-scale candidate gene study of tuberculosis susceptibility in the Karonga district of northern Malawi. *Am J Trop Med Hyg* **71**, 341-9 (2004).
53. Bellamy, R. *et al.* Tuberculosis and chronic hepatitis B virus infection in Africans and variation in the vitamin D receptor gene. *J Infect Dis* **179**, 721-4 (1999).
54. Barreiro, L.B. *et al.* Promoter variation in the DC-SIGN-encoding gene CD209 is associated with tuberculosis. *PLoS Med* **3**, e20 (2006).
55. Li, C.M. *et al.* Association of a polymorphism in the P2X7 gene with tuberculosis in a Gambian population. *J Infect Dis* **186**, 1458-62 (2002).
56. Garred, P. *et al.* Mannan-binding lectin in the sub-Saharan HIV and tuberculosis epidemics. *Scand J Immunol* **46**, 204-8 (1997).
57. Ogus, A.C. *et al.* The Arg753GLn polymorphism of the human toll-like receptor 2 gene in tuberculosis disease. *Eur Respir J* **23**, 219-23 (2004).
58. Velez, D.R. *et al.* Variants in toll-like receptors 2 and 9 influence susceptibility to pulmonary tuberculosis in Caucasians, African-Americans, and West Africans. *Hum Genet* **127**, 65-73 (2010).
59. Means, T.K. *et al.* Differential effects of a Toll-like receptor antagonist on Mycobacterium tuberculosis-induced macrophage responses. *J Immunol* **166**, 4074-82 (2001).
60. Khor, C.C. *et al.* A Mal functional variant is associated with protection against invasive pneumococcal disease, bacteremia, malaria and tuberculosis. *Nat Genet* **39**, 523-8 (2007).

61. Rose, D.N., Schechter, C.B. & Adler, J.J. Interpretation of the tuberculin skin test. *J Gen Intern Med* **10**, 635-42 (1995).
62. Dudbridge, F. Likelihood-based association analysis for nuclear families and unrelated subjects with missing genotype data. *Hum Hered* **66**, 87-98 (2008).
63. Jouanguy, E. *et al.* Partial interferon-gamma receptor 1 deficiency in a child with tuberculoid bacillus Calmette-Guerin infection and a sibling with clinical tuberculosis. *J Clin Invest* **100**, 2658-64 (1997).
64. Akahoshi, M. *et al.* Influence of interleukin-12 receptor beta1 polymorphisms on tuberculosis. *Hum Genet* **112**, 237-43 (2003).
65. Cervino, A.C. *et al.* Fine mapping of a putative tuberculosis-susceptibility locus on chromosome 15q11-13 in African families. *Hum Mol Genet* **11**, 1599-603 (2002).
66. Ben-Selma, W., Harizi, H. & Boukadida, J. Association of TNF-alpha and IL-10 polymorphisms with tuberculosis in Tunisian populations. *Microbes Infect* **13**, 837-43 (2011).
67. Kim, F. *et al.* Deficiencies of macronutrient intake among HIV-positive breastfeeding women in Dar es Salaam, Tanzania. *J Acquir Immune Defic Syndr* **67**, 569-72 (2014).
68. Stein, C.M., Hall, N.B., Malone, L.L. & Mupere, E. The household contact study design for genetic epidemiological studies of infectious diseases. *Front Genet* **4**, 61 (2013).
69. Blumberg, H.M. *et al.* American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America: treatment of tuberculosis. *Am J Respir Crit Care Med* **167**, 603-62 (2003).
70. Purcell, S. PLINK(v1.07).
71. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
72. Gauderman, W.J. Sample size requirements for matched case-control studies of gene-environment interaction. *Stat Med* **21**, 35-50 (2002).
73. StataCorp. Stata Statistical Software: Release 11. (StataCorp LP, College Station, TX, 2009).
74. RDC, T. R: A language and environment for statistical computing. *Vienna: R Foundation for Statistical Computing* (2007).
75. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* **28**, 3326-8 (2012).
76. Turner, S.D. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. *bioRxiv beta* (2014).
77. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-7 (2010).
78. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
79. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
80. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457-70 (2011).
81. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263-5 (2005).
82. International HapMap, C. The International HapMap Project. *Nature* **426**, 789-96 (2003).

83. International HapMap, C. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
84. Lahey, T. *et al.* Polyantigenic interferon-gamma responses are associated with protection from TB among HIV-infected adults with childhood BCG immunization. *PLoS One* **6**, e22074 (2011).
85. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).
86. McVean, G.A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581-4 (2004).
87. Fujita, P.A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* **39**, D876-82 (2011).
88. Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72 (2006).
89. Jarvis, J.P. *et al.* Patterns of ancestry, signatures of natural selection, and genetic association with stature in Western African pygmies. *PLoS Genet* **8**, e1002641 (2012).
90. Miller, S.A., Dykes, D.D. & Polesky, H.F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* **16**, 1215 (1988).
91. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74 (2012).
92. O'Donnell, C.J. *et al.* Genome-wide association study for subclinical atherosclerosis in major arterial territories in the NHLBI's Framingham Heart Study. *BMC Med Genet* **8 Suppl 1**, S4 (2007).
93. Nair, R.P. *et al.* Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat Genet* **41**, 199-204 (2009).
94. Yang, Q., Kathiresan, S., Lin, J.P., Tofler, G.H. & O'Donnell, C.J. Genome-wide association and linkage analyses of hemostatic factors and hematological phenotypes in the Framingham Heart Study. *BMC Med Genet* **8 Suppl 1**, S12 (2007).
95. Tso, H.W., Lau, Y.L., Tam, C.M., Wong, H.S. & Chiang, A.K. Associations between IL12B polymorphisms and tuberculosis in the Hong Kong Chinese population. *J Infect Dis* **190**, 913-9 (2004).
96. Velez Edwards, D.R. *et al.* MCP1 SNPs and pulmonary tuberculosis in cohorts from West Africa, the USA and Argentina: lack of association or epistasis with IL12B polymorphisms. *PLoS One* **7**, e32275 (2012).
97. Morahan, G. *et al.* Association of variants in the IL12B gene with leprosy and tuberculosis. *Tissue Antigens* **69 Suppl 1**, 234-6 (2007).
98. Swaminathan, B. *et al.* Autophagic marker MAP1LC3B expression levels are associated with carotid atherosclerosis symptomatology. *PLoS One* **9**, e115176 (2014).
99. Huffmeier, U. *et al.* Common variants at TRAF3IP2 are associated with susceptibility to psoriatic arthritis and psoriasis. *Nat Genet* **42**, 996-9 (2010).
100. Tsunemi, Y. *et al.* Interleukin-12 p40 gene (IL12B) 3'-untranslated region polymorphism is associated with susceptibility to atopic dermatitis and psoriasis vulgaris. *J Dermatol Sci* **30**, 161-6 (2002).
101. Sieburth, D. *et al.* Assignment of genes encoding a unique cytokine (IL12) composed of two unrelated subunits to chromosomes 3 and 5. *Genomics* **14**, 59-62 (1992).

102. Capon, F. *et al.* Sequence variants in the genes for the interleukin-23 receptor (IL23R) and its ligand (IL12B) confer protection against psoriasis. *Hum Genet* **122**, 201-6 (2007).
103. Chan, S.H. *et al.* Induction of interferon gamma production by natural killer cell stimulatory factor: characterization of the responder cells and synergy with other inducers. *J Exp Med* **173**, 869-79 (1991).
104. Gately, M.K. *et al.* Regulation of human lymphocyte proliferation by a heterodimeric cytokine, IL-12 (cytotoxic lymphocyte maturation factor). *J Immunol* **147**, 874-82 (1991).
105. Cooper, A.M., Magram, J., Ferrante, J. & Orme, I.M. Interleukin 12 (IL-12) is crucial to the development of protective immunity in mice intravenously infected with mycobacterium tuberculosis. *J Exp Med* **186**, 39-45 (1997).
106. Flynn, J.L. *et al.* IL-12 increases resistance of BALB/c mice to Mycobacterium tuberculosis infection. *J Immunol* **155**, 2515-24 (1995).
107. Holscher, C. *et al.* A protective and agonistic function of IL-12p40 in mycobacterial infection. *J Immunol* **167**, 6957-66 (2001).
108. de Jong, R. *et al.* Severe mycobacterial and Salmonella infections in interleukin-12 receptor-deficient patients. *Science* **280**, 1435-8 (1998).
109. Altare, F. *et al.* Impairment of mycobacterial immunity in human interleukin-12 receptor deficiency. *Science* **280**, 1432-5 (1998).
110. Jasenosky, L.D., Scriba, T.J., Hanekom, W.A. & Goldfeld, A.E. T cells and adaptive immunity to Mycobacterium tuberculosis in humans. *Immunol Rev* **264**, 74-87 (2015).
111. Gaffen, S.L., Jain, R., Garg, A.V. & Cua, D.J. The IL-23-IL-17 immune axis: from mechanisms to therapeutic testing. *Nat Rev Immunol* **14**, 585-600 (2014).
112. Gopal, R. *et al.* Unexpected role for IL-17 in protective immunity against hypervirulent Mycobacterium tuberculosis HN878 infection. *PLoS Pathog* **10**, e1004099 (2014).
113. Heintzman, N.D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108-12 (2009).
114. Creyghton, M.P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107**, 21931-6 (2010).
115. Sobota, R.S. *et al.* Addressing Population-Specific Multiple Testing Burdens in Genetic Association Studies. *Ann Hum Genet* (2015).
116. Li, J.Z. *et al.* Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**, 1100-4 (2008).
117. Rosenberg, N.A. *et al.* Genetic structure of human populations. *Science* **298**, 2381-5 (2002).
118. Scheinfeldt, L.B. & Tishkoff, S.A. Recent human adaptation: genomic approaches, interpretation and insights. *Nat Rev Genet* **14**, 692-702 (2013).
119. Raviglione, M.C., Snider, D.E., Jr. & Kochi, A. Global epidemiology of tuberculosis. Morbidity and mortality of a worldwide epidemic. *JAMA* **273**, 220-6 (1995).
120. Stewart, G.R., Robertson, B.D. & Young, D.B. Tuberculosis: a problem with persistence. *Nat Rev Microbiol* **1**, 97-105 (2003).
121. Rieder, H.L. Epidemiologic Basis of Tuberculosis Control. *International Union Against Tuberculosis and Lung Disease, Paris*, 162 (1999).
122. Ma, N. *et al.* Clinical and epidemiological characteristics of individuals resistant to M. tuberculosis infection in a longitudinal TB household contact study in Kampala, Uganda. *BMC Infect Dis* **14**, 352 (2014).

123. Vukmanovic-Stejic, M., Reed, J.R., Lacy, K.E., Rustin, M.H. & Akbar, A.N. Mantoux Test as a model for a secondary immune response in humans. *Immunol Lett* **107**, 93-101 (2006).
124. Sepulveda, R.L. *et al.* Evaluation of tuberculin reactivity in BCG-immunized siblings. *Am J Respir Crit Care Med* **149**, 620-4 (1994).
125. Mahan, C.S. *et al.* Innate and adaptive immune responses during acute M. tuberculosis infection in adult household contacts in Kampala, Uganda. *Am J Trop Med Hyg* **86**, 690-7 (2012).
126. Liu, X. *et al.* Genome-wide association study identifies candidate genes for Parkinson's disease in an Ashkenazi Jewish population. *BMC Med Genet* **12**, 104 (2011).
127. Palmieri, F. The mitochondrial transporter family SLC25: identification, properties and physiopathology. *Mol Aspects Med* **34**, 465-84 (2013).
128. Postma, D.S. *et al.* Genetic susceptibility to asthma--bronchial hyperresponsiveness coinherited with a major gene for atopy. *N Engl J Med* **333**, 894-900 (1995).
129. Hopp, R.J., Townley, R.G., Biven, R.E., Bewtra, A.K. & Nair, N.M. The presence of airway reactivity before the development of asthma. *Am Rev Respir Dis* **141**, 2-8 (1990).
130. Longo, G., Strinati, R., Poli, F. & Fumi, F. Genetic factors in nonspecific bronchial hyperreactivity. An epidemiologic study. *Am J Dis Child* **141**, 331-4 (1987).
131. Asthma, G.I.f. Guide for Asthma Management and Prevention. (2014).
132. Shimwela, M. *et al.* Asthma prevalence, knowledge, and perceptions among secondary school pupils in rural and urban coastal districts in Tanzania. *BMC Public Health* **14**, 387 (2014).
133. von Hertzen, L., Klaukka, T., Mattila, H. & Haahtela, T. Mycobacterium tuberculosis infection and the subsequent development of asthma and allergic conditions. *J Allergy Clin Immunol* **104**, 1211-4 (1999).
134. von Mutius, E. *et al.* International patterns of tuberculosis and the prevalence of symptoms of asthma, rhinitis, and eczema. *Thorax* **55**, 449-53 (2000).
135. Wells, W.F., Ratcliffe, H.L. & Grumb, C. On the mechanics of droplet nuclei infection; quantitative experimental air-borne tuberculosis in rabbits. *Am J Hyg* **47**, 11-28 (1948).
136. Cambier, C.J. *et al.* Mycobacteria manipulate macrophage recruitment through coordinated use of membrane lipids. *Nature* **505**, 218-22 (2014).
137. Charlson, E.S. *et al.* Topographical continuity of bacterial populations in the healthy human respiratory tract. *Am J Respir Crit Care Med* **184**, 957-63 (2011).
138. Goswami, R. & Kaplan, M.H. A brief history of IL-9. *J Immunol* **186**, 3283-8 (2011).
139. Temann, U.A., Laouar, Y., Eynon, E.E., Homer, R. & Flavell, R.A. IL9 leads to airway inflammation by inducing IL13 expression in airway epithelial cells. *Int Immunol* **19**, 1-10 (2007).
140. Uyttenhove, C., Simpson, R.J. & Van Snick, J. Functional and structural characterization of P40, a mouse glycoprotein with T-cell growth factor activity. *Proc Natl Acad Sci U S A* **85**, 6934-8 (1988).
141. Schmitt, E., Van Brandwijk, R., Van Snick, J., Siebold, B. & Rude, E. TCGF III/P40 is produced by naive murine CD4+ T cells but is not a general T cell growth factor. *Eur J Immunol* **19**, 2167-70 (1989).
142. Beriou, G. *et al.* TGF-beta induces IL-9 production from human Th17 cells. *J Immunol* **185**, 46-54 (2010).

143. Lu, L.F. *et al.* Mast cells are essential intermediaries in regulatory T-cell tolerance. *Nature* **442**, 997-1002 (2006).
144. Ye, Z.J. *et al.* Differentiation and recruitment of Th9 cells stimulated by pleural mesothelial cells in human Mycobacterium tuberculosis infection. *PLoS One* **7**, e31710 (2012).
145. Kaplan, M.H. Th9 cells: differentiation and disease. *Immunol Rev* **252**, 104-15 (2013).
146. Dugas, B. *et al.* Interleukin-9 potentiates the interleukin-4-induced immunoglobulin (IgG, IgM and IgE) production by normal human B lymphocytes. *Eur J Immunol* **23**, 1687-92 (1993).
147. Petit-Frere, C., Dugas, B., Braquet, P. & Mencia-Huerta, J.M. Interleukin-9 potentiates the interleukin-4-induced IgE and IgG1 release from murine B lymphocytes. *Immunology* **79**, 146-51 (1993).
148. Sears, M.R. *et al.* Relation between airway responsiveness and serum IgE in children with asthma and in apparently normal children. *N Engl J Med* **325**, 1067-71 (1991).
149. Donahue, R.E., Yang, Y.C. & Clark, S.C. Human P40 T-cell growth factor (interleukin-9) supports erythroid colony formation. *Blood* **75**, 2271-5 (1990).
150. Williams, D.E. *et al.* T-cell growth factor P40 promotes the proliferation of myeloid cell lines and enhances erythroid burst formation by normal murine bone marrow cells in vitro. *Blood* **76**, 906-11 (1990).
151. Longphre, M. *et al.* Allergen-induced IL-9 directly stimulates mucin transcription in respiratory epithelial cells. *J Clin Invest* **104**, 1375-82 (1999).
152. Yamasaki, A. *et al.* IL-9 induces CCL11 expression via STAT3 signalling in human airway smooth muscle cells. *PLoS One* **5**, e9178 (2010).
153. Siezen, C.L. *et al.* Genetic susceptibility to respiratory syncytial virus bronchiolitis in preterm children is associated with airway remodeling genes and innate immune genes. *Pediatr Infect Dis J* **28**, 333-5 (2009).
154. Janssen, R. *et al.* Genetic susceptibility to respiratory syncytial virus bronchiolitis is predominantly associated with innate immune genes. *J Infect Dis* **196**, 826-34 (2007).
155. Schneider, C., King, R.M. & Philipson, L. Genes specifically expressed at growth arrest of mammalian cells. *Cell* **54**, 787-93 (1988).
156. Keane, J. *et al.* Infection by Mycobacterium tuberculosis promotes human alveolar macrophage apoptosis. *Infect Immun* **65**, 298-304 (1997).
157. Vieira, A.R. *et al.* Medical sequencing of candidate genes for nonsyndromic cleft lip and palate. *PLoS Genet* **1**, e64 (2005).
158. Wetherill, L. *et al.* Association of substance dependence phenotypes in the COGA sample. *Addict Biol* (2014).
159. Olson, J.E. *et al.* Centrosome-related genes, genetic variation, and risk of breast cancer. *Breast Cancer Res Treat* **125**, 221-8 (2011).
160. Hatzimichael, E. *et al.* Study of specific genetic and epigenetic variables in multiple myeloma. *Leuk Lymphoma* **51**, 2270-4 (2010).
161. Cooke, G.S. & Hill, A.V. Genetics of susceptibility to human infectious disease. *Nat Rev Genet* **2**, 967-77 (2001).
162. Collins, R.L. *et al.* Multifactor dimensionality reduction reveals a three-locus epistatic interaction associated with susceptibility to pulmonary tuberculosis. *BioData Min* **6**, 4 (2013).

163. White, M.J. *et al.* Epiregulin (EREG) and human V-ATPase (TCIRG1): genetic variation, ethnicity and pulmonary tuberculosis susceptibility in Guinea-Bissau and The Gambia. *Genes Immun* **15**, 370-7 (2014).
164. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2015 update. *Nucleic Acids Res* **43**, D470-8 (2015).
165. Lin, P.L., Plessner, H.L., Voitenok, N.N. & Flynn, J.L. Tumor necrosis factor and tuberculosis. *J Investig Dermatol Symp Proc* **12**, 22-5 (2007).
166. Keane, J., Remold, H.G. & Kornfeld, H. Virulent Mycobacterium tuberculosis strains evade apoptosis of infected alveolar macrophages. *J Immunol* **164**, 2016-20 (2000).
167. Schlesinger, L.S., Bellinger-Kawahara, C.G., Payne, N.R. & Horwitz, M.A. Phagocytosis of Mycobacterium tuberculosis is mediated by human monocyte complement receptors and complement component C3. *J Immunol* **144**, 2771-80 (1990).
168. Hosmer D.W., L.S. (ed.) *Applied logistic regression*, (John Wiley & Sons, New York, 2000).
169. Campa, D. *et al.* Polymorphisms of dopamine receptor/transporter genes and risk of non-small cell lung cancer. *Lung Cancer* **56**, 17-23 (2007).
170. Stapleton, J.A., Sutherland, G. & O'Gara, C. Association between dopamine transporter genotypes and smoking cessation: a meta-analysis. *Addict Biol* **12**, 221-6 (2007).
171. Geluk, A. *et al.* Identification of HLA class II-restricted determinants of Mycobacterium tuberculosis-derived proteins by using HLA-transgenic, class II-deficient mice. *Proc Natl Acad Sci U S A* **95**, 10797-802 (1998).
172. Abel, B. *et al.* Toll-like receptor 4 expression is required to control chronic Mycobacterium tuberculosis infection in mice. *J Immunol* **169**, 3155-62 (2002).
173. Drennan, M.B. *et al.* Toll-like receptor 2-deficient mice succumb to Mycobacterium tuberculosis infection. *Am J Pathol* **164**, 49-57 (2004).
174. Kavelaars, A., Cobelens, P.M., Teunis, M.A. & Heijnen, C.J. Changes in innate and acquired immune responses in mice with targeted deletion of the dopamine transporter gene. *J Neuroimmunol* **161**, 162-8 (2005).
175. Tian, K., Wang, Y. & Xu, H. WTH3 is a direct target of the p53 protein. *Br J Cancer* **96**, 1579-86 (2007).