THE INFLUENCE OF CANDIDATE GENE POLYMORPHISMS IN TUBERCULOSIS AMONG SELECTED ETHIOPIAN POPULATIONS

Ephrem Mekonnen Gebeyehu

A Dissertation Submitted to
The Department of Microbial, Cellular and Molecular Biology

Presented in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy (Biology/Applied Genetics)

Addis Ababa University
Addis Ababa, Ethiopia

April, 2018

DECLARATION: PhD Candidate

I, Ephrem Mekonnen, declare that this thesis is my own work and that it has not been presented in other University, College or Institution, seeking for similar degree or other purposes. Where information has been derived from other sources, I confirm that all have been duly acknowledged.

Signed by PhD Candidate:

Name: Ephrem Mekonnen (M.Sc., Department of Microbial and Cellular Biology, Addis Ababa University)

Date:_____

Signature: _____

DECLARATION: Supervisor

This is to certify that this thesis entitled "A Genetic Epidemiological Investigation of the Influence of Candidate Gene Polymorphisms in Tuberculosis Progression Among Ethiopian Populations: FMO2, TICAM2, NOD1" is prepared and submitted to the Department of Microbial, Cellular, and Molecular Biology by Ephrem Mekonnen in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biology (Applied Genetics), that it complies with the regulations of the Graduate Studies of Addis Ababa University and meets the accepted standards with respect to originality and quality.

Signed by Research Supervisor:

Name: Endashaw Bekele (Professor, Department of Microbial and Cellular Biology, Addis Ababa University)

Date:_____

Signature:_____

ABSTRACT

Background: Tuberculosis (TB) is an ancient malady that remains a modern global health concern. The ancient relationship between *Mycobacterium tuberculosis* (*Mtb*) and *Homo sapiens* has evolved into a spectrum of co-existence pattern which, from the human perspective, ranges from a complete and fatal susceptibility to TB to a total resistance to infection and/or progression to disease. Essentially, therefore, infection by *Mtb*, although necessary, is not a sufficient cause for TB disease and numerous studies have demonstrated that this spectrum of host-pathogen interaction outcome is mediated in part by the genetic constitution of individuals that impact their potential for innate and adaptive immunity against TB. However, there is a conspicuous lack of replication of results and the hunt for novel associations in different populations continues unabated. It has been suggested that the lack of power in the investigative process to identify genetic risk factors to TB emanate mainly from the lack of precision in the 'Definition of the Phenotype'. Because of the complexity of TB, a precise and consistent definition of the disease is a major challenge and it has been difficult to provide reliable TB phenotype definition criteria amenable for genetic epidemiological analysis. There is also the possibility of unaccounted differences in the genetic architecture of the studied populations and the potentially differential or minor effects of either rare or common variants identified in the studies. The primary aim of this genetic epidemiological study was 1) to investigate the role of genetic variations within candidate genes towards susceptibility to TB by resequencing genes which previously showed an association signal in another Sub-Saharan African cohort (Ugandan population), as well as 2) testing an original candidate gene hypothesis in Ethiopian populations. The study focused on three innately expressed genes: NOD1 [Nucleotide-binding Oligmerization Domain containing 1] and TICAM2 [Toll/Interleukin-1 Receptor Domain-Containing Adaptor

Molecule 2], based on a recent finding in an East African population that indicated significant statistical association and another study demonstrating a biological plausibility of TICAM2-NOD1 synergistic action and, thus, were deemed to warrant an effort to replicate those findings in an independent population; and, FMO2 (Flavin-Containing-Monooxygenase-2) that demonstrates some curious characteristics vis-à-vis TB that come from some immunologic, pharmacogenetic, and population/evolutionary genetics observations. First, FMO2 is highly expressed in activated pulmonary-macrophages regulating oxidative-stress level, an essential mechanism of innate immunity against TB. In contrast, FMO2 is regarded as potentially deleterious because it causes adverse reactions to anti-TB drug treatment. Furthermore, FMO2 possesses a unique polymorphism, FMO2*1/FMO2*2 (rs6661174), with differential ethno-geographic distribution. The functional ancestral-variant, FMO2*1, is only found in African and some Hispanic populations with highest frequency in Sub-Saharan-Africa while Caucasians and Asians are homozygous for the dysfunctional derived-allele, FMO2*2. However, there are no reported investigations into the potential involvement of FMO2 in the pathogenesis of diseases exerting population-specific selective pressures.

Method: This study focused on finding a method of unraveling the TB-phenotype complexity by drawing TB-trait definitions closely based on its known natural progression stages from infection to disease onset. This ensures that intermediate stages of the disease are included as phenotypes of interest rather than just analyzing the final binary-trait outcome: 'Active-TB vs. No-active-TB' (presence/absence). An intermediate stage may be a distinct phenotype having its own immunogenetic profile that could otherwise be missed. TB cases and household controls (n=292) were ascertained from 3 different ethnic groups. Latent *Mtb* infection was determined

using Quantiferon to develop reliable TB progression phenotypes. Exonic regions of TICAM2, NOD1, and FMO2 genes were sequenced. Various statistical tests of association were done that accounted for possible confounding by sex, age, and population stratification.

Result: Multiple SNPs in FMO2, TICAM2 and NOD1 were associated with TB. Among the most significant findings were two SNPs in NOD1 achieving a study-wide significance threshold: rs751770147 [p=7.28x10-05] and chr7:30477156(T), a novel variant, [p=1.04x10-04]. Three SNPs in TICAM2 were nominally associated with TB, including rs2288384 [p=0.003]. Haplotype-based association tests supported the SNP-based results. The study also identified for the first time an association between FMO2 and TB both at the SNP and haplotype level. Two novel SNPs achieved a study-wide significance [chr1:171181877(A), p=3.15E-07, OR=4.644 and chr1:171165749(T), p=3.32E-06, OR=6.825] while several SNPs (twenty two) showed nominal signals of association. The pattern of association suggested a protective effect of FMO2 against both active and latent TB with distinct genetic variants underlying the TB-progression pathway. Haplotype-based tests confirmed the SNP-based results with a single haplotype bearing the ancestral-and-functional FMO2*1 "C" allele ("AGCTCTACAATCCCCTCGTTGCGC") explaining the overall association (haplotype-specific-p=0.000103). Strikingly, not only was FMO2*1 associated with reduced risk to "Active TB" (p=0.0118, OR=0.496) but it also does not co-segregate with the other 5'-3' flanking top high-TB-risk alleles.

Conclusion: The study design not only helped to replicate previous association signals of TICAM2 and NOD1 with TB but also identified novel genetic variants associated with TB in Ethiopian populations thus further validating the genes' involvement in TB pathogenesis. The

study also identified for the first time the association of FMO2 gene with TB and provided an evidence for the existence of an evolutionary adaptation to an ancient disease based on the ancestral FMO2*1 polymorphism. The study sheds light on the possible impact of host-pathogen co-evolution on the present differential ethno-geographic distribution of the FMO2*1 variant that coincides with the origin of both humans and *Mtb* in Sub-Saharan Africa. The novel discovery calls for a revision of the notion that FMO2*1 is "potentially deleterious". Rather, the study indicates that FMO2*1 is associated with reduced risk to TB progression acting in a haplotypic framework. The finding also puts into question the prudence of prescribing thiourea based anti-TB drug treatment regimens for populations harbouring high proportions of FMO2*1 without genetic screening. The study examined multiple ethnic groups in Ethiopia, and found that the association results are robust to population stratification. As Ethiopia is considered to be the origin of both humanity and *Mtb*, these findings are of particular significance for understanding *Mtb*-human co-evolution and the genetic underpinnings of TB in general.

ACKNOWLEDGMENTS

I am very grateful to Prof. Endashaw Bekele of Addis Ababa University (AAU), who has always been my source of inspiration and mentor, for the invaluable instructions, trainings and constant supervision of this work. I am highly grateful to Dr. Catherine M. Stein at Case Western Reserve University (CWRU), (USA) for the critical help in fully funding this research and contributing valuable suggestions to the study design.

I am highly grateful to Drs. Kifle Dagne and Kassahun Tesfaye at AAU for their helpful suggestions. I deeply appreciate Drs. Luca Pagani and Rosemery Ekong of University College London (UK) for their laboratory trainings.

I am highly indebted to Dr. Charles Rotimi for offering me a Pre-Doc Research Fellowship at the Center for Research on Genomics and Global Health at National Inistitutes of Health (NIH), (USA), and Drs. Adebowale Adeyemo, Amy Bently, and particularly to my friend Fasil Ayele Tekola, for their invaluable training in large scale genomic sequence data analysis during my fellowship. I am also very grateful to the Welcome Trust Foundation (UK) for granting me a scholarship to train in Genetic Epidemiology. I am grateful to the genomics laboratory staff at CWRU (USA), for their help in DNA sequencing and Dr. Keith Chervenak for his help in obtaining laboratory equipments, reagents and biological sample import permits. I am thankful to the Department of Microbial, Cellular, and Molecular Biology and the Institute of Biotechnology of Addis Ababa University, Faculty of Life Sciences, and all the staff, for providing me the opportunity to enroll for a PhD study and conduct this research. I am also thankful to the

Ethiopian Health and Nutrition Research Institute and Holetta Agricultural Research Institute, and all the staff, for allowing me to conduct some work in their laboratories.

I am extremely grateful to Alem Ketema Hospital in Merhabete, Adigrat Hospital in Adigrat, and Arbaminch Hospital in Arbaminch for their unreserved cooperation during sample collection and their hospitality. I am very thankful for all the staff who professionally supported me in appropriate sample ascertainment, recruitment, interview, and collection of blood-samples.

I am highly indebted to my entire family and especially to my wife Debrina and my daughter Lazary for their unfailing support, and, for their patience!

I would like to thank very much all the participants of the study who willingly agreed to be a part of this study. I applaud them for their high spirits of voluntarism. This study would not have been possible without their participation and I dedicate this work to them.

I. Table of Contents

List of figures

List of Tables

II. LIST OF ABBREVIATIONS

A1:	Allele 1 (minor allele)
A2:	Allele 2 (major allele)
Bonf./BONF :	Bonferroni
BP:	Physical position of base-pair
Chi-sqr CHISQ:	Chi-square test
CI:	Confidence interval
CHR:	Chromosome
CMH:	Cochran-Mantel-Haenszel test of allelic association
DNA:	Deoxyribonucleic acid
E:	Exponent
F_A:	Frequency in Affected individuals
F_U:	Frequency in Unaffected individuals
Freq.:	Frequency
FMO2:	Flavin Containing Monooxygenase 2
HWE:	Hardy-Weinberg equilibrium
IBD:	Identity-by-descent
IBS:	Identity-by-state
IGRA:	Interferon-gamma release assay
L95:	Lower bound of 95% confidence interval
U95:	Upper bound of 95% confidence interval
LD:	Linkage disequilibrium
log.reg:	logistic regression test
LTBI:	Latent TB Infection
MAF:	Minor Allele Frequency

MDR-TB:	Multi-Drug Resistant TB
MDS:	Multi-Dimensional Scaling
Mtb:	*Mycobacterium tuberculosis*
NMISS:	Number of non-missing genotypes
NOD1:	Nucleotide-binding Oligmerization Domain containing 1
OR:	Odds Ratio
P:	P-value
QC:	Quality Control
QFT:	QuantiFERON®-TB Gold In-Tube (blood test for LTBI based on cell-mediated interferon gamma cytokine release assay; IGRA kit)
SE:	Standard Error
SNP:	Single Nucleotide Polymorphism
TB:	Tuberculosis
TICAM2:	TIR (Toll/Interleukin-1 Receptor Domain)-Containing Adaptor Molecule 2
U95:	Upper bound 95% confidence interval
UQP-UQN: vs.	TB-Unaffected-QFT-Positive individuals with latent TB infection (cases) TB-Unaffected-QFT-Negative individuals with no latent TB infection (controls)
XDR-TB:	Extensively Drug Resistant TB

III. INTRODUCTION

Tuberculosis (TB): Disease description

TB is one of the oldest known infectious human diseases and it is an understatement to report that it a major cause of morbidity and mortality globally. Observations from different fields of studies ranging from anthropology to genomics show that few diseases have had an association with human beings as ancient and impactful as that of TB (Comas, et al, 2013). TB is lethal in two thirds of patients who do not receive proper treatment killing, on average, another three people every minute (WHO, 2016). The co-existence of *Mycobacterium tuberculosis* (*Mtb*) with humans for such a long time has provided the bacterium opportunity to evolve and develop mechanisms that enable it to evade both our natural immune responses and clinical treatment efforts (Galagan, 2014). Despite the widespread use of an attenuated live vaccine Bacille Calmette-Guérin (BCG) and several antibiotics like isoniazid, rifampin, ethambutol, and pyrazinamide, TB remains a major public health concern requiring improved treatment approaches and more specific and rapid diagnostics (Issar, 2003; Zhang, 2005). The problem is further compounded by the looming danger of the spread of drug-resistant *Mtb* strains [causing multidrug-resistant TB, (MDR-TB), extensively drug-resistant TB (XDR-TB), totally drug-resistant TB (TDR-TB)] that threaten to overcome the available tools of TB control (Ernst, 2007).

TB is a complex disease with heterogenous manifestations. Although primarily a pulmonary disease (PTB) that is initiated by the deposition of *Mtb* contained in aerosol droplets onto lung alveolar surfaces, it also affects several organ systems. And, from the point of infection, the

progression of the disease can have several outcomes, determined by both intrinsic factors such as the immuno-genetic constitution of individuals and extrinsic factors like the nutritional status of the individual and co-infections and treatments that suppress or compromise the host's immune and physiological systems (Issar, 2003).

Therefore, it is essential to differentiate infection from TB disease. Infection simply means the presence of *Mtb* in a host, whether it leads to active disease or not. A survey of the TB literature shows that there is a general agreement on the course of TB pathogenesis despite some difficulty in delineating where one stage ends and another begins mortality (Issar, 2003; Van Crevel, et al, 2002), (Ma, et al, 2014). A recent review of the progression and resolution of the disease defines TB at least into two stages (Figure-1) (Pai, et al, 2016): latent infection and active disease. In the first stage, *Mtb* is transmitted by inhalation into the lungs where it is phagocytosed by macrophages which are thought to be the predominant host cells for the majority of its infectious life cycle. Internalization by macrophages triggers an immune response, the recruitment of additional monocytes and, ultimately, the formation of a tuberculous granuloma (which lends the disease its modern name) that effectively contains the infected cells. The success of *Mtb* is partly due to its ability to survive in a granuloma for a long time in an asymptotic latent state. In the second stage, the granuloma disintegrates, *Mtb* reactivates, disseminates and the patient progresses to active disease. In other words, for active disease and transmission to occur *Mtb* containment in the granuloma must either fail due to changes in the host immune status or be overcome by the pathogen (Chao, et al, 2010; Galagan, 2014). However, there is variation in TB progression in that after the first exposure of uninfected individuals to infectious TB cases not all subjects exposed to TB become infected: 5-10% eliminate primary infection. And, in the second

stage, only 5-10% of *Mtb*-infected individuals progress to active TB disease. Generally, it is reported that 90-95% of people infected with *Mtb* will contain the initial infection without symptoms and develop a latent infection, while the remaining 5-10% will develop active TB disease after exposure, with symptoms such as cough, fever, weight loss and night sweats. Another manifestation of variation in TB is that it is characterized by a variable course of disease (incubation time, dormant infection, or latency), disease site, and severity. These observations have led to the hypothesis that these two stages may have different immunogenetic underpinnings. In addition, the pathogen may play a role in disease progression since some human-adapted strains in the Mycobacterium Tuberculosis Complex (MTBC) are reportedly more virulent than others, as defined by increased transmissibility, rapid pathogenesis, higher morbidity and mortality in infected individuals (reviewed by Brites, et al, 2015; Gagneux, 2012). It has also been shown that that distinct geographic regions of the world have variable distributions of mycobacterial lineages, with certain regions having a major lineage that is a minor contributor elsewhere (Ernst, et al, 2007; Filliol, et al, 2006; Firdessa, et al, 2013). For example, associations were found between mycobacterial genotypes and phenotypes in humans and that the transmission of phylogeographic lineages of *Mtb* is non-randomly associated with various ethnic populations

Therefore, although *Mtb* is necessary for TB, it is not sufficient to cause the disease. In other words, active TB disease is not exactly an inevitable outcome of exposure to, or infection by, *Mtb*. In recent decades the role of host genetics in susceptibility or resistance to infectious diseases such as TB has received more attention (Hill, 2006; Sirugo, et al 2008; Moller, et al, 2009). In this regard Africa, as the ultimate source of modern humans harboring higher genetic

variation than any other continent, has become a focus of studies of the patterns of genetic variation that are crucial to understanding how genes affect phenotypic variation, including disease predisposition to TB (Gurdasani, et al, 2014).

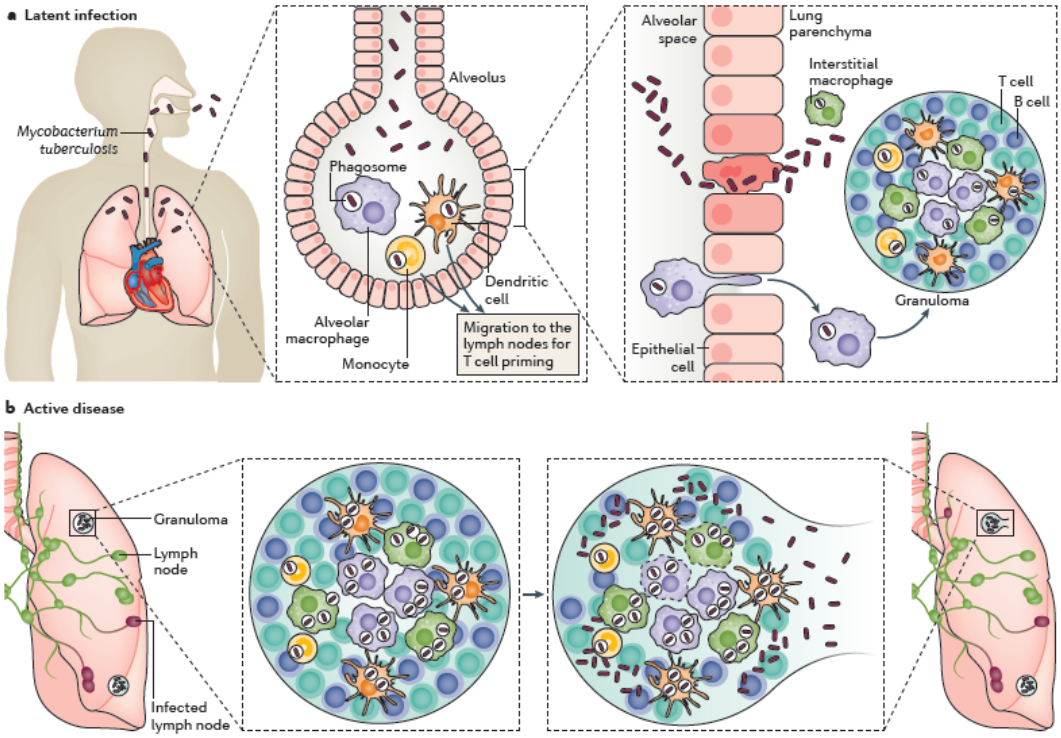


Figure 1: Summary of TB progression (Source: Pai, et al, 2016)

Infection begins when Mtb enters the lungs via inhalation, reaches the alveolar space and encounters the resident alveolar macrophages. If this first line of defence fails to eliminate the bacteria, Mtb invades the lung interstitial tissue, either by the bacteria directly infecting the alveolar epithelium or the infected alveolar macrophages migrating to the lung parenchyma. Subsequently, either dendritic cells or inflammatory monocytes transport Mtb to pulmonary lymph nodes for T cell priming. This event leads to the recruitment of immune cells, including T cells and B cells, to the lung parenchyma to form a granuloma. b | The bacteria replicate within the growing granuloma. If the bacterial load becomes too great, the granuloma will fail to contain the infection75 and bacteria will disseminate eventually to other organs, including the brain. At this phase, the bacteria can enter the bloodstream or re‑enter the respiratory tract to be released — the infected host is now infectious, symptomatic and is said to have active TB disease.

Finally, from a public health point of view, pulmonary TB is the most important form of the disease because of its infectious nature. By contrast, extrapulmonary TB is generally non-infectious and therefore has a lower public health priority (Gagneux, 2012).

Components of innate and adaptive immunity to TB

The constant battle against invading microorganisms necessitates an immune system capable of providing mechanisms both for the recognition and destruction of harmful microorganisms. Generally, upon encountering pathogenic microorganisms, an immuno-competent host activates two distinct effector mechanisms to ensure effective elimination of the pathogen: the innate and the adaptive immune systems (Philips, et al, 2011; Halff, 2013; Kawai, et al, 2010). Innate immunity provides the "first line of defense" and ensures a rapid but nonspecific and short-lived response, involving for instance the release of antimicrobial antigens, activation of the complement system, and the secretion of cytokines that induce inflammation and attract phagocytes. Innate immune cells also stimulate the activation of the long-lasting adaptive immune response by the presentation of pathogen-derived antigens leading to the generation of antibodies by B-cells and activation of antigen-specific cytotoxic T-lymphocytes. This results in the specific recognition and clearance of pathogens, the elimination of infected cells, and the development of immune memory.

After the uptake of *Mtb* in alveolar macrophages several possible scenarios may be envisaged (Van Crevel, et al, 2002; Philips, et al, 2011):

Pre-innate immune response: since *Mtb* captured in intracellular membrane-bound vesicles (endosome) are not exposed to cytosolic proteases and these vesicles do not always bind to lysosomes, *Mtb* is able to avoid degradation by the immune system. Therefore, in the first three 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.

Initiation of innate immune response: once infection is established, however, a focal nonspecific inflammatory response follows. This response is regulated by a network of proinflamatory (for example, TNF-α, IL-1β, IL-6, IL-12, IL-18, IFN-γ) and anti-inflammatory (for example, IL-10, TGFβ, IL-4) cytokines and chemokine (for example, IL-8, MCP-1, RANTES) responses. The activities of the innate host defense mechanism include, the cellular uptake of *Mtb*, which involves different cellular receptors and humoral factors. This initial response determines the local outgrowth of *Mtb* (sometimes dissemination) or containment of infection. Various receptors, including Toll-like and Nod-like receptors, play a crucial role in immune recognition of *Mtb*, which is the next step. The subsequent inflammatory response is regulated by production of pro- and anti-inflammatory cytokines and chemokines. The innate host response is also necessary for induction of adaptive immunity to TB in which phagocytic cells play a key role in antigen presentation and the initiation of T-cell immunity which follows. On the other hand, *Mtb* may be destroyed immediately, in which case no adaptive T-cell response is developed.

Adaptive immune response: Roughly three weeks following *Mtb* infection, the host starts to mount an adaptive response by inducing the differentiation of T cells into two main types of CD4 expressing T helper cells, TH1 and TH2. TH1 cells control infections by intracellular pathogens

while TH2 cells are responsible for extracellular pathogens and they induce B cells to produce antibodies. Upon *Mtb* 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 TH1 CD4 T cells. The mature TH1 cells mount an interferon gamma (IFN-γ) response and activate phagolysosomes in *Mtb* containing macrophages, and induce the production of nitric oxide and reactive oxygen species. These activated macrophages also use tumor necrosis factor (TNF) signaling in a localized inflammatory response that leads to the formation of granuloma. The granuloma consists of a central core of macrophages, organized into progressively necrotic multinucleated giant cells, and surrounded by CD4 T cells.

Generally, both innate and adaptive immunity mechanisms are involved in determining the outcome of TB infection and, as the outcome determines survival, the immunity genes behind TB susceptibility or resistance phenotypes are subjected to natural selection (Quintana-Murci, et al, 2007; Barreiro, et al, 2005). In this respect, different gene polymorphisms have been found which are associated with increased susceptibility to and severity of TB (Azad, et al, 2012) (Figure-2). Some of these polymorphisms are functional, but for many of these no functional (immunologic) changes have been demonstrated yet, and these associations need further confirmation and investigation.

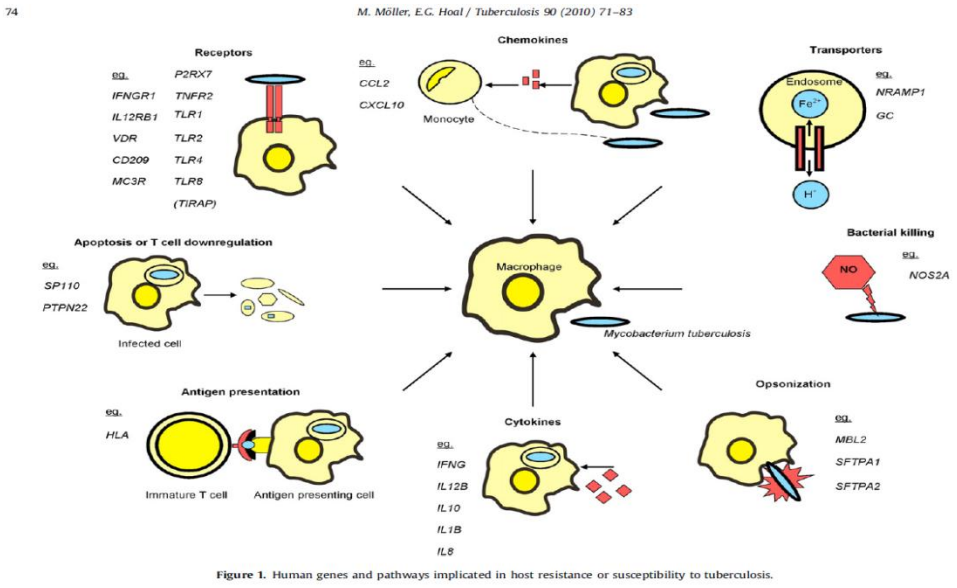


Figure 2: Human genes and pathways implicated in host resistance or susceptibility to TB (Source: Azad, et al, 2012)

Impact of TB-HIV co-infection on TB prognosis

Although individuals co-infected with TB and HIV (*Human Immunodeficiency Virus*) were excluded from analysis in this study, it is worthwhile to note its impact briefly. The factors that govern the development of tuberculosis disease are complex and various factors have been clearly associated with increased susceptibility to tuberculosis. HIV infection is one of the most important (Toossi, et al, 2001; Nahid, et al, 2007). The impact of HIV-TB co-infection can be assessed from several aspects related to the synergistic effect of the dual infection on TB pathogenesis and vice versa. In general, however, TB appears to be the first opportunistic infection in an HIV-infected person, while active TB has been shown to induce HIV virus-replication, thus accelerating the progression to AIDS. HIV-TB interaction dramatically increases the progression of both. Latent TB-infection in HIV-positive persons seems to reactivate at a rate of 10% per year (as opposed to 10% in a lifetime for HIV-negative persons).

The clinical presentation of TB may be altered in HIV-positive patients, especially in progressed stages of HIV-infection when immunity is considerably compromised. HIV-induced immunosuppression is a substantial contributor to TB reactivation. Smear-negative and extra-pulmonary forms of TB are then more common and X-ray abnormalities are atypical.

Although the underlying causal mechanisms and immunological aspects remain poorly understood, the synergy of HIV and TB takes several forms (Vermund, et al, 2007). HIV-induced immunosuppression increases the likelihood that quiescent TB will reactivate. In other words, the newly active TB patient is now infectious for TB whereas without the HIV co-infection, the patient might have remained uninfectious. TB itself up-modulates the host immune system; an activated T cell that is activated in response to infection from *Mtb* (or a number of other infections such as helminthes or herpeviruses) produces more HIV than a quiescent cell such that HIV expression increases in the face of co-infections. Furthermore, the cytokine profile of the tuberculous microenvironment is conducive to induction of HIV-1 replication. Development of TB during HIV-1 infection is associated with intense immune activation and an increase in HIV-1 enhancing proinflammatory cytokines, such as IL-1β, IL-6 and TNF-α.

Overview of TB diagnosis in Ethiopia

WHO has a recommended international strategy for control of TB called "DOTS" (Directly Observed Therapy, Short-course) which is implemented in Ethiopia. Among the main elements of DOTS are mechanisms for promptly diagnosing and treating people who have TB disease (Caws, et al, 2008). Diagnosis of TB is thought of as having two aspects, which are also both

stage-specific and purpose-specific: diagnosis of TB infection and diagnosis of active disease (Glassroth, 2005). The relative importance of the different characteristics of a diagnostic test depends upon the setting in which the test is to be performed and the intended use of the results (Steingart, et al, 2007). In Ethiopia, the diagnosis of TB is based on the recommendations of the National TB and Leprosy Control Programme (NTLC) which is revised and updated periodically (FMoH, 2011). Briefly, however, there are two diagnostic algorithms developed to detect PTB and lymph node TB (Mengistu, et al, 2005). Since other diagnostic services such as Mycobacterial cultures and pathological services are not available for routine purposes, the NTLC advocates adherence to these diagnostic algorithms: Patients presenting with symptoms suggestive of PTB who had productive cough for three weeks or more with at least two positive sputum smears or one positive smear and x-ray findings consistent with active PTB are classified as smear-positive PTB cases. Patients presenting with cough of three weeks or more with initial three negative smears and no clinical response to a course of broad-spectrum antibiotics, three negative smear results after a course of broad-spectrum antibiotics, x-ray findings consistent with active PTB and decided by a clinician to be treated with anti-TB chemotherapy are classified as smear-negative PTB cases. Patients presenting with dry cough of three weeks or more are diagnosed based on strong clinical evidence and x-ray findings consistent with active TB. Patients presenting with symptoms suggestive of TB other than the lungs, which do not respond to a course of broad-spectrum antibiotics and decided by a clinician to be treated with anti-TB chemotherapy are classified as EPTB cases. In children, TB is diagnosed if there are symptoms and signs suggestive of TB, contact history with a known TB patient and x-ray findings consistent with active TB. The diagnosis of smear-negative PTB primarily relies on patients'

clinical conditions, response to broad spectrum antibiotics and chest radiographic evidences. Until recently, the tuberculin skin test (TST) was the only available method for diagnosing LTBI.

TB epidemiology in Ethiopia

While epidemiological reports of TB (incidence/prevalence) from around the world paint a grim picture of human failure to cope with the disease which has been declared a global emergency by the WHO, there is an urgent need to find more effective ways of containment or eradication of TB especially in poorer regions of the world, like Ethiopia, where the disease strikes hard (Ernst, et al, 2007). Furthermore, with the uneven geographical distribution of TB there is a risk that it would be relegated to the status of the "poor-man's disease" and ignored by richer countries which have managed to control TB in their population, mainly by arresting transmission rather than treatment, and which might have the resources to contribute a lot.

According to a publication by the Federal Ministry of Health of Ethiopia (FMoHE, 2011) released around the beginning of this study (Table-1), the national routine surveillance of TB data revealed that during 2000/01-2009/10 more than 100,000 all forms of TB cases were detected every year and a total of 1,231,145 cases were reported, out of which, 392,319 cases (32%) wear smear positive pulmonary TB and 430,274 (35%) were extra-pulmonary TB. During this decade the number of cases detected increased annually showing an increment of 40% in 2009/10 as compared to 2000/01. Latent TB infection is also high in Ethiopia: one study reported that more than 65% of all Tuberculin Skin Tested (TST) Ethiopians are positive (Tegbaru, et al, 2006).

Table 1: Notification of new smear positive and all forms of TB cases by region: 2009/10 (Source: FMoHE, 2011)

Regions	New Smear Positive TB			New all forms of TB cases			All TB forms CDR (/100,000)
	Annual estimate (163/100,000)	Cases detected	CDR	Annual estimate (378/100,000)	Cases detected	CDR	
Tigray	7,573	2,112	28	4,646,197	10,812	62	233
Amhara	29,514	7,732	26	18,106,982	33,728	49	186
SNNPR	26,715	10,534	39	16,389,550	23,100	37	141
National	129,740	46,634	36	79,594,841	149,508	50	188
Estimates are based on 2009 WHO report							

TB-HIV co-infection remains a complicating factor. During the 1990s Ethiopia experienced a severe HIV epidemic. A TB epidemic has coincided with this development, and in 1999, Ethiopia was ranked number 8 among the 23 countries with the highest total TB burden. In 1997, about 30% of all new TB cases were believed to occur in HIV-positive individuals (Bruchfeld, et al, 2002). According to the 2014 WHO report, the prevalence and incidence of all forms of TB were 211 and 224 per 100,000 of the population, respectively. In 2013, TB mortality was estimated to be 32 per 100,000 excluding HIV related deaths. These figures were used to demonstrate that the Millennium Development Goal on reducing TB incidence rate has been achieved in Ethiopia. For example, by 2013 the national TB incidence rate had fallen to 224 per 100,000 as compared to 369 in 1990 and TB prevalence and mortality rates have also reduced by 50%. But, still, and according to the same WHO report, 22 high burden countries account for 80% of the of the global TB cases and Ethiopia has the tenth highest TB burden in the world and is also one of the 27 high MDR-TB burden countries.

Role of Mtb-human co-evolution in TB

In an insightful and inspirational comparative reviews (Gagneux, 2012; Mulugeta, et al, 2014; Brites, et al 2015, Comas, et al, 2013) of the accumulated knowledge on *Mtb* and human genetic diversity the authors elaborate on the possible patterns of co-evolution and its role in TB pathogenesis based on molecular phylogenetic and epidemiologic data. First, it is pointed out that Africa harbours the largest diversity of MTBC in the world since it has been discovered that Africa is the only region of the world that harbours all of the main human-adapted MTBC lineages. Second, this diversity and distribution parallels and follows the foot-steps of *Homo sapiens* as described in th "Out-of Africa-and-back" migration model. And, third, human-adapted MTBC exhibits a phylogeographic population structure with different lineages associated with different human populations, i.e., host-specific adaptation of MTBC lineages consistent with host–pathogen co-evolution. This scenario led to the postulate that human MTBC originated in Africa and accompanied the Out-of-Africa migrations of modern humans and, consequently, it is possible that the biology and the epidemiology of human TB have been shaped by the long-standing association between MTBC and its human host (Figure-3). Accordingly, TB has been characterized as a disease caused by a pathogen that has co-evolved to gain the capacity to survive within the human cells of the immune system that are designed to eradicate microbial pathogens, the macrophages.

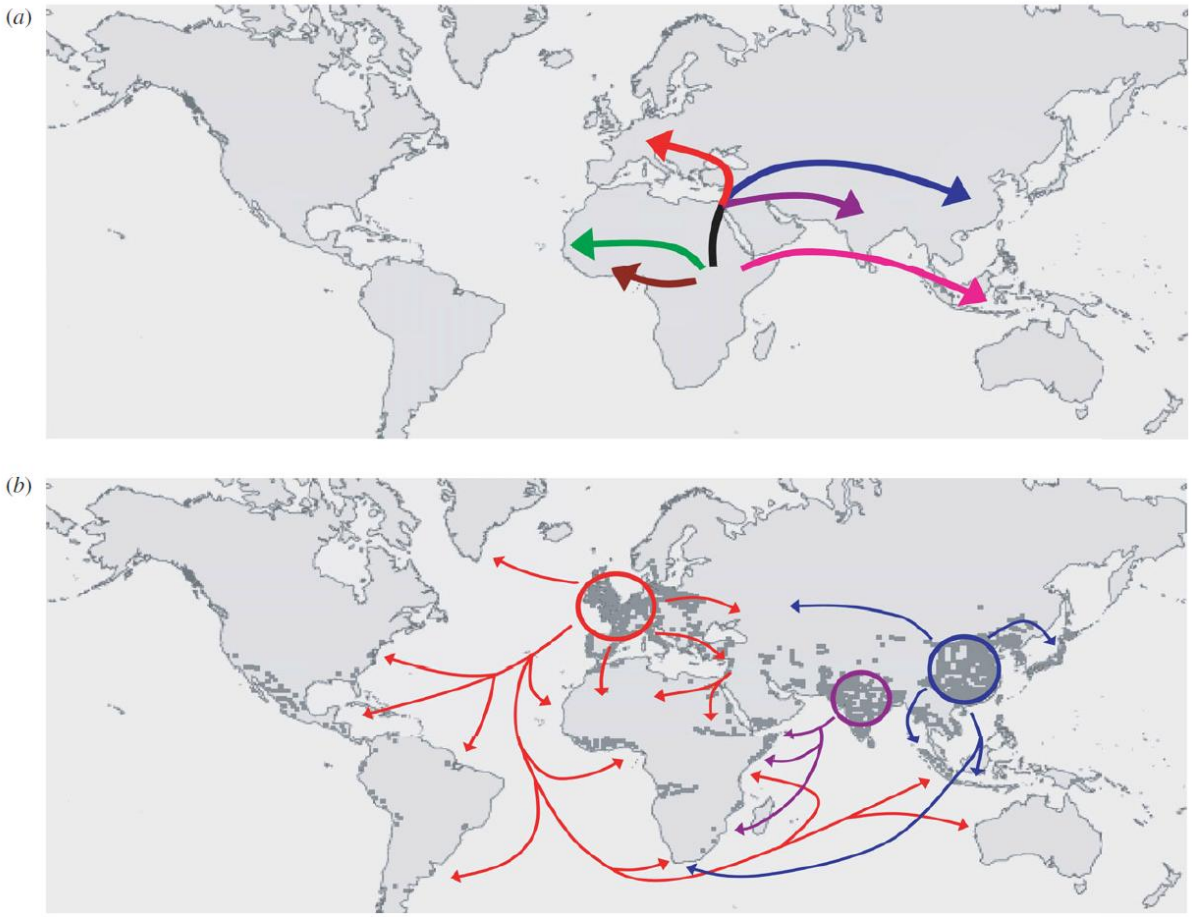


Figure 3: 'Out-of-and-back-to-Africa' scenario for the evolutionary history of human TB (Gagneux, 2012).

(a) MTBC originated in Africa and some lineages accompanied the Out-of-Africa migrations of modern humans. (b) The three evolutionarily 'modern' MTBC lineages seeded Europe, India and China, respectively, and expanded as a consequence of the sharp increases in human populations in these regions starting a few centuries ago (each dark grey dot corresponds to 1 million people). These lineages then spread throughout the world via exploration, trade and conquest.

The role of human genetic diversity in TB pathogenesis

One of the most important questions in TB research is why only approximately 10% of latently infected individuals develop active disease any time during their lifetime. If we knew the answer to this question, we would be in a much better position to design an effective TB vaccine. After several studies, the evidence for a human genetic component in the commonly observed inter-individual variation in susceptibility to TB is considered incontrovertible (Galagan, 2014; Moller, et al, 2010). However, the evidences suggest that, except for a few extreme and very rare cases of Mendelian disorders, TB is a classical example of a complex disease in which each individual TB-associated genetic locus contributes a small proportion of the observed variation in disease susceptibility. In other words, since TB is a complex disease, it is unlikely that any single factor with a major effect can adequately explain TB risk and, therefore, new candidate genes and genomic regions need to be explored. However, there are also evidences that suggest susceptibility to TB may depend on the ability of *Mtb* to modulate epigenetic (stable and heritable changes occuring in cells without change in DNA sequence by various modifications of nucleotides, associated histones and untranslated RNA pool) mediators and repress host immune response genes and presents a new challenge to understanding the function of epigenetics and its regulators in the patho-physiology of TB (Warner, et al, 2014; Bierne, et al, 2012), (Yadav, et al, 2015).

Overview of human genetic polymorphisms associated with TB

The application of genetic epidemiological tools to the study of the genetic aspects of TB susceptibility and pathogenesis is considered to have taken a turning point by the discovery of

polymorphisms in NRAMP1, now named SLC11A1, which were found to impact resistance/susceptibility to a spectrum of *Mtb* infections (Abel, et al, 2000). SLC11A1 is a transmembrane protein in the lysosomal and endosomal membranes that functions as a divalent ion pump removing them into the cytosol. It was postulated that in doing so, it deprives bacteria such as *Mtb* of essential elements such as iron, thereby inhibiting growth. Polymorphisms that impair proper function of SLC11A1 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. The methodology used to study SLC11A1 was based on the analysis of an extended pedigree rather than a large population: a model-based linkage method (classical LOD score) with both a recessive and a dominant mode of inheritance assumed; and, to model gene-environment interactions, individuals were assigned to risk (liability) classes on the basis of age, BCG vaccination, tuberculin skin-test results, and previous disease information. The specification of liability classes was considered to be crucial to obtaining evidence for linkage, which stresses the importance of properly accounting for the epidemiological context in this type of analysis.

In Africa, several genetic association or linkage analyses with TB have been carried out which reported positive, suggestive or no linkage/association results (Sirugo, et al, 2008; Azad, et al, 2012). Using the complementary approach of candidate gene analysis, case–control studies of West African samples have identified associations with variants in several genes; for example SLC11A1, (Bellamy, et al, 1998); vitamin D receptor (Bornman, et al, 2004; Lombard, et al, 2006); CD209 (DC-SIGN), PTX3 (Olesen, et al, 2007); and P2X7 genes (Li, et al, 2002) to mention a few, have all been associated with TB. In East Africa, a combined linkage and

association study of Ugandans has shown that IL10, interferon gamma receptor 1 (IFNGR1), and TNF alpha receptor 1 (TNFR1) variants are linked and associated to TB, but not with susceptibility to latent infection (Stein, et al, 2005). And, in the first genome-wide linkage scan for a major infectious disease in Africans, evidence of linkage was found on chromosomes Xq27 and 15q11 (Bellamy, et al, 2000), further supporting the conclusion that TB susceptibility loci exist. Another recent analysis of affected sibling pairs from South Africa (of mixed ancestry) and from Malawi, along with a case–control study in West Africans have identified two putative loci for susceptibility, one at 6p21-q23 and one at 20q13.31-33. At the latter locus, variation in the melanocortin 3 receptor (MC3R) and cathepsin Z (CTSZ) genes were implicated in the pathogenesis of TB (Cooke, et al, 2008). And a recent study of innate and adaptive immunity genes in TB identified TICAM2 and NOD1 among others (Hall, et al, 2015).

TB is a multifactorial infectious disease which has been shown to have a strong genetic component through twin studies, genome-wide linkage and association studies, and candidate gene association studies. Candidate gene studies have focused on genes that mediate both innate and adaptive immune pathways and particularly on pathogen pattern recognition receptor pathways and have added to our understanding of the multigenic basis of variable resistance to TB (Kawai, et al, 2010). However, since the immune response to *Mtb* infection, progression to disease, and the molecular signaling pattern and the extent of tissue involvement change over the course of disease it is unlikely that any single factor can adequately explain susceptibility to TB. Numerous genetic variants have been variably associated with susceptibility to TB (Hill, 2006; Sirugo, et al, 2008; Moller, et al 2009; Moller, et al 2010; Azad, et al, 2012). Although there are apparent variations in the reported associations that could be attributed to variable evolutionary

selection pressures as a result of long-term host-pathogen interactions (Caws, et al, 2008) in certain populations as well as differences in study designs, the overall conclusion is that polymorphisms involved in the initial innate recognition of the presence of pathogen and the induction of downstream innate and adaptive immune responses are critical for the efficient elimination of infection (Kawai, et al, 2010).

IV. GENETIC EPIDEMIOLOGICAL APPROACHES TO THE STUDY OF TUBERCULOSIS AS A COMPLEX DISEASE

Overview of genetic epidemiological methodologies

Genetic epidemiological studies are motivated by the potential existence of genetic components behind the numerous observations of differential susceptibility to infectious and non-infectious diseases between individuals or populations (Azad, et al, 2012). In Africa, where both high human genetic polymorphism and infectious diseases are still prevalent, genetic epidemiological investigations are crucial to improve health conditions (Hill, 2006; Sirugo, et al, 2008); and, if such studies discover genetic variants with strong effect sizes that impact susceptibility/resistance to disease, they could guide public health policy. In simple terms, genetic epidemiology (GE) is the study of genetic factors that determine the distributions and dynamics of diseases in populations. The primary objective of GE is finding genes and nucleotide variants underlying human diseases and characterizing them (Weiss, 1993; Thomas 2004). The most common differences in DNA sequence between individuals are single nucleotide polymorphisms (SNPs), that is, changes in a single DNA base pair. The study designs and statistical approaches used in genetic epidemiology can be grouped into generalized categories (Stein, 2011):

1. **Familial aggregation:** used to evaluate whether individuals with a family history of a given disease have a greater risk of disease than individuals without such a family history.
2. **Familial segregation:** used to determine the transmission pattern of a disease trait within families. It involves first, determining whether there is transmission from the parents to the children in a manner consistent with Mendelian laws; second, estimating the disease model;

third, and more generally, estimating parameters of a Mendelian model, including the disease allele frequency and penetrances.

3. **Familial co-segregation with a genetic marker:** used to locate a disease locus on the genome based on the familial co-segregation of a trait of interest with one or several genetic markers. It forms the basis of linkage analysis.
4. **Association with a genetic variant:** once the approximate location of a disease gene has been found by linkage analysis, genetic association analysis can be conducted in that region alone to pinpoint the gene and allelic variants involved. These studies take advantage of linkage disequilibrium (LD), i.e., the fact that throughout the genome, alleles at tightly linked loci are associated in the population; and, thus, the markers analyzed need not be functional, but may simply be in LD or be in 'indirect association' with the functional variant and serve as 'proxies'. This enables genome-wide association studies (GWAS) that encompass the entire genome and do not require a priori hypothesis about biological plausibility/function. Candidate-gene gene association studies involve far fewer SNPs than GWAS.

The results of GE studies of TB have mainly pointed to the extremes of a TB causal-spectrum. As discussed by (Abel, et al, 2000; Azad, et al, 2012) population-based studies focusing on candidate genes selected on the basis of their known or suspected role in innate or adaptive anti-mycobacterial immunity, have reported associations with TB. However, the reported predisposing alleles having only moderate effects and the molecular basis of the genetic control of TB remains to be validated. In contrast, there is clearly a causal relationship between certain rare Mendelian immunodeficiencies affecting T cells or phagocytes and severe TB. However, these rare mutations alone cannot account for the millions of TB cases reported annually

worldwide. There is, therefore, a gap between causal susceptibility in rare individuals and uncertain predisposition in the general population. However, it is possible that these two aspects of genetic predisposition to TB do not conflict but, rather, represent the two ends of a continuous causal spectrum. In this context, the study of NRAMP1 provided an opportunity to reconcile these two extreme poles of TB genetics: predisposition to tuberculosis was shown to be associated with a major gene effect that is neither fully monogenic nor polygenic. The major-gene control reported supports the hypothesis of a continuous spectrum in the genetic control of clinical TB, since it bridges the gap between simple Mendelian susceptibility and complex polygenic predisposition to TB. These insights into the genetic component of TB help towards advancing hypotheses to be tested by various GE-based study designs.

Literature review of TB-related investigations identified only one published study on genetic susceptibility to TB in Ethiopian populations (Malik, et al, 2005). The authors investigated polymorphisms in the SFTPA1 and SFTPA2 genes, both expressed in the lungs, in Ethiopian families and identified four polymorphisms that modify the risk of TB susceptibility

Problems associated with testing TB-related GE hypothesis

It has been hypothesized that, for populations that have not been extensively exposed to a lethal pathogen like *Mtb*, and before antibiotics were available, highly susceptible individuals who died early from the disease carried high-risk genotypes. Under this hypothesis, relatively rare susceptibility alleles (especially dominant ones) would have probably disappeared rapidly from the population.However, other alleles with milder effects may remain and predispose individuals

to TB in a less-pronounced manner (Abel. et al 2000; Weiss, 1993). This view of TB has major implications for understanding the biology of anti-mycobacterial immunity in general and has led to the formulation of some baseline models for the study of genetic influences on TB: rare Mendelian immune defects in particular patients with severe/uncommon clinical features; major-gene effects in certain specific kindred and populations with no ancestral history of *Mtb* exposure; and, more-common polymorphisms with less-pronounced effects in populations with a longer history of exposure to *Mtb*.

However, the testing of such hypotheses has been hindered by the complex nature of TB and a conclusive or comprehensive genetic model has not yet emerged. Some of the likely sources of the limitations for, and inconsistencies between, TB genetic studies are discussed in (Stein, 2007):

Disease heterogeneity: Most studies analyzed only the binary trait TB (presence, absence), which does not reflect the complexity of the disease trait. TB may be expressed with varying severity in a number of organ systems after a long and variable latency period and, therefore, phenotype definition is not trivial.

Shared environment: Most studies have failed to account for the influence of shared environment within households on TB risk, which may spuriously inflate estimates of heritability.

Co-morbidity: Most studies have not accounted for co-infections/-morbidities which increase the risk of TB, such as HIV.

Diagnosis: Previous genetic studies have also differed dramatically in TB diagnostic criteria and characterization of controls.

Genotype characterization: Few studies have examined the linkage disequilibrium (LD) structure within the genes of interest, and thus, were unable to account for untyped polymorphisms that may influence TB risk.

Generally, therefore, the major problem in testing these and other mechanistic or probabilistic hypotheses about the genetic of TB is the heterogeneous, complex, or multi-factorial nature of TB itself as well as disease diagnosis and definition. Moreover, there are studies with results consistent with the conclusion that unique environmental and natural selective factors may have resulted in the development of population-/ethnic-specific host genetic factors associated with TB susceptibility and resistance (Delgado, et al, 2002; Coussens, et al, 2013). Therefore, considering the longstanding host–pathogen interaction, it is important to account for the possibility that some degree of co-evolution (broadly defined as 'reciprocal, adaptive genetic changes in interacting host and pathogen species') might have occurred between MTBC and its human host in a an ethno-geographic-dependant manner (Gagneux, 2012). And, even with proper attention to TB phenotype definitions and *Mtb* strain effects, it should be realized that the standard genetic epidemiological methods based on nucleotide sequence data alone cannot be employed to assess the potential role of epigenetics in host immune responses.

V. HYPOTHESES AND OBJECTIVES OF THE PRESENT STUDY

Hypotheses

The primary hypothesis of this study was derived from an extensive review of the TB literature that revealed firm biological and statistical evidences for the existence of genetic polymorphisms that have a significant role in TB pathogenesis. Such polymorphisms potentially affect the expression, structure and function of immunity genes leading to variations in anti-*Mtb* immune responses. There is also a correspondingly overwhelming epidemiologic evidence for the existence of variation in TB progression: in some individuals primary infection is eliminated; in some individuals progression is arrested at the LTBI stage; and, in some individuals latency can transition to Active TB. The basic premises and the central hypothesis of this study can be summarized as:

Premise-1: There is a commonly observed inter-individual variation in the immune response to *Mtb* exposure and infection. In other words, not all individuals exposed to *Mtb* get infected, and not all *Mtb*-infected individuals progress to active TB.

Premise-2: Exposure to, and infection by, *Mtb* is a necessary but not a sufficient cause for TB disease.

Central Hypothesis: Polymorphisms in human immunity genes contribute to variation in the immune response to *Mtb* exposure and infection.

Reviews of the TB literature also revealed that while numerous immunogenetic polymorphisms have been demonstrated to have a significant role in TB pathogenesis, there has also been a distinctive failure to replicate these associations across different populations which, in turn, raises questions about the validity of the findings. This observation led to another hypothesis of

the present study that the reasons for the non-replication of TB genetic association results may be explained by the lack of precise definition of TB phenotypes that frustrates the discovery of underlying gene-disease associations or lead to spurious conclusions. Because of the complexity of TB, a precise and consistent definition of the disease is a major challenge and it has been difficult to provide reliable TB phenotype definition criteria amenable for genetic epidemiological analysis. Clearly defined traits can help increase power to detect disease-predisposing loci and thus more informative than studying a heterogeneous lump of a complex phenotype.

It is also hypothesized in this study that TB progression stages may represent distinct TB phenotypes with respective stage-specific immunogenetic risk profiles and, thus, formulating test-models with a clear definition of the TB trait based on its known natural history from exposure, to infection, and progression to active disease, can ensure that intermediate stages of the disease are included as phenotypes of interest. Furthermore, intermediate phenotypes may be more closely tied to the level of gene expression that could otherwise be missed.

General and specific objectives of the study

General Objective: The general objective of this study is to contribute knowledge in TB-genetics by, first, investigating the influence of polymorphisms in candidate innate immunity genes on TB susceptibility and, second, by studying whether TB progression has a stage-specific genetic risk profile.

Specific Objectives:

1. **Replication study:** to replicate a previous finding of TB genetic association signals (TICAM2 and NOD1 genes in Ugandan population) in an independent population (Ethiopian population); and, to test the validity of both the association signals and the study design
2. **Original study:** to test a novel hypothesis of TB-candidate gene (FMO2) association based on firm biological evidences
3. **Population genetics:** to assess if there are common or population-specific signatures of genetic association with TB-phenotypes; and, to genetically characterize and compare the Ethiopian populations studied with respect to the selected candidate genes (frequency and distribution of polymorphisms, patterns of linkage disequilibrium and haplotypes).

To summarize, the objective of the present research was to disentangle some of the inherent complexities in studying the genetic underpinnings of variations in the immune response to TB by putting forward a genetic epidemiological hypothesis based on an intermediate phenotype model.

VI. RATIONAL FOR SELECTION OF CANDIDATE GENES IN THIS STUDY

Description of TB candidate genes: TICAM2, NOD1 AND FMO2

In this study, three innate immunity genes were selected after extensive literature review on the genetics of TB, results of significant associations with TB as well as firm biological evidences of involvement in TB pathogenesis, including anti-TB drug treatment outcome: TICAM2 (Toll/Interleukin-1 Receptor Domain-Containing Adaptor Molecule 2), NOD1 (Nucleotide-binding Oligmerization Domain Containing 1), and FMO2 (Flavin-containing Monooxygenase 2).

Overview of the synergistic role of pattern recognition receptors in innate immunity: TICAM2 of TLRs and NOD1 of NLRs

Pattern recognition receptors (PRRs) recognize PAMPs (highly conserved microbial-specific motifs, pathogen-associated molecular patterns) and/or DAMPs (damage/danger-associated molecular patterns comprising signals that indicate the existence of internal errors, such as reactive oxygen species (ROS) or a change in intracellular ion levels, as well as molecules that are released by dying and injured cells such as ATP), with their corresponding specific set of innate germ-line-encoded PRRs. PAMPs are conserved structures essential for the survival of the pathogenic microorganism that are non-existent in the host. They constitute a wide variety of molecules such as bacterial cell wall components (e.g. lipopolysaccharides and peptidoglycans), pathogen-specific proteins, and nucleic acid structures derived from bacteria and viruses such as unmethylated CpG motifs and dsRNA respectively (Halff, 2013). PRRs can be found in the extracellular space, integrated in cellular membranes or in the cytosol. The innate immune system comprises several classes of PRRs. TLRs recognize microbes on the cell surface and

within endosomes, whereas NLRs sense microbial molecules in the cytosol. Among these, the Toll-like receptors (TLRs) and Nucleotide-binding Oligomerization domain-containing Receptors (NLRs) have been extensively investigated (Moreira, et al, 2012; Franchi, et al, 2008).

Currently, the field of study on the putative agonists/ligands, interaction partners, signaling pathways, and associated genetic polymorphisms of TLRs and NLRs is very dynamic, rapidly expanding and different studies may report contradicting outcomes. Thus, many aspects related to their role are not fully understood and much work remains to be done (Halff, 2013). In this regard, one study (Tada, et al, 2005) serves very well to demonstrate the synergistic activities of members of the TLRs and NRLs in the immune system. The researchers reviewed that Freund's complete adjuvant, which contains killed mycobacterial cells, has been widely used as a powerful adjuvant to induce cell-mediated immunity, represented by delayed-type hypersensitivity, as well as to enhance humoral immunity against test antigens in laboratory animals. And that a series of studies have identified peptidoglycan (PGN), arabinogalactan, and mycolic acid as being the mycobacterial component responsible for the unique adjuvant activity of Freund's complete adjuvant and thereafter the PGN moieties of various bacteria were revealed to also be active in this respect. Then, in the mid-1970s, the minimal essential structure of PGN for adjuvant activity was demonstrated to be muramyldipeptide (MDP; N-acetylmuramyl-Lalanyl-D-isoglutamine) by use of a chemically synthesized compound. MDP was also shown to reproduce various bioactivities of PGN by activating macrophages via TLR2 (Figure-4). The research authors also cite that dendritic cells (DCs) are initiators and modulators of immune responses. Peripheral DCs are characterized by the ability to capture and process antigens, their migration to lymphoid organs, and the expression of various costimulatory molecules for

antigen-specific lymphocyte activation. Cytokines secreted by DCs initiate and enhance both innate and acquired immunity. Many studies suggest that the activation of DCs by microbial components leads to the secretion of IL-12, which subsequently induces Th1 development and gamma interferon (IFN-ɣ) production by T cells. Various TLR agonists differentially modulate IL-12 production in DCs and are involved in determining the Th1/Th2 balance. In the peripheral blood, NOD2 is highly expressed on DCs as well as on monocytes and granulocytes. And, NOD1 and NOD2 were revealed to be receptors for MDP as well as demonstrated to recognize a PGN motif. With these background, the researchers employed various powerful synthetic NOD1, NOD2 and TLR adjuvants or agonists like MDP that mimic the bacterial peptidoglycan moiety to stimulate human dendritic cell (DC) cultures and examine whether cell-mediated immunity through T helper type 1 (Th1) responses, especially delayed-type hypersensitivity. The results showed that immature DCs derived from human monocytes expressed mRNAs for NOD1, NOD2, TLR2, TLR3, TLR4, and TLR9 and synergistically induced various cytokines and interferon production in DCs to initiate Th1-lineage immune responses. The study demonstrated that agonists of NOD1 and NOD2 in combination with TLR4 and TLR9 agonists synergistically induced IL-12 production in human DCs in culture and that the IL-12 thus generated promoted T cells to produce IFN-ɣ. These findings strongly suggest that the combinatory stimulation of DCs via the NOD pathway and the TLR pathway synergistically promotes Th1-lineage immune responses. In bacterium-host interactions, host cells should be stimulated with bacterial PGN fragments, namely, a NOD1 and/or NOD2 agonist(s), in addition to various TLR agonists. Therefore, the above synergism should generally occur in host-bacterium interactions particularly in facultative intracellular parasitic bacteria, represented by mycobacteria.

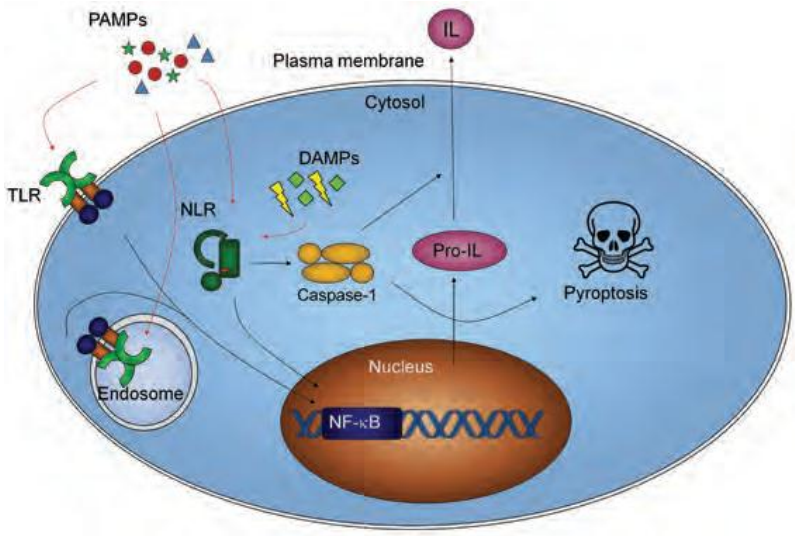


Figure 4: Synergistic signalling pathways induced by TLR and NLR activation, and their crosstalk (Halff, 2013).

Some TLRs and NLRs specifically recognize extracellular and intracellular molecules that are derived from bacteria or viruses, such as fragments of the bacterial cell wall or DNA/RNA. Others notice damage to the cell by recognizing molecules that only occur in damaged but not in healthy cells such molecules are proteins that leak from mitochondria into the cytosol when the cell is infected. As soon as the presence of such danger signals are detected, TLRs and NLRs cause the cell to either secrete molecules that help the immune system to recognize and clear the infection or they activate the protein caspase-1 that in turn induces a specific cell death mechanism called pyroptosis. The death and clearance of the infected cell ensures that the infection cannot spread any further.

TICAM2 of TLRs

Among the first described PRRs were the membrane-spanning TLRs, discovered as recently as in 1994, which sense PAMPs in the extracellular environment as well as in endosomes. TLRs are primarily expressed in immune cells, such as macrophages, monocytes, B-lymphocytes, and dendritic cells. TICAM2, also known as TRAM, is a member of the TLRs residing on cell membranes and, unlike NLRs, mediate innate extracellular recognition of microbes (Elson, 2007; Takeda, et al, 2005). TICAM2 was first discovered and named by a study (Seya, et al, 2005) that reported the identification of a protein that directly binds TLR4 and harbored a TIR (Toll-IL-1 receptor) domain that has significant homology to TICAM-1, and conferred heterophilic

dimerization with TICAM-1 and supports TLR 3, 4-mediated immune responses by facilitating the expression of IFN-inducible genes. Its physical structure and interaction properties has been elucidated recently (Enokizonoa, et al, 2013). Based on its functional and physical association analyses, it was demonstrated that TICAM-2 was an adapter that physically bridges TLR4 and TICAM-1 helping them to bind LPS (lipopolysaccharide) and functionally transmits LPS-TLR4 signaling to TICAM-1, which in turn activates interferonregulatory factor (IRF-3) followed by expression of IFN-ɣ. LPS is a constituent of the outer membrane of bacteria and its recognition as an agonist for TLR4 activates macrophages and dendritic cells (DCs) in the innate immune system and expresses many genes including Nuclear factor-B (NF-B)- and IRF-3/IFN-inducible genes. A definitive association between TB phenotypes and genetic polymorphisms in TICAM2 gene was discovered for the first time only recently by (Hall, et al, 2015) and was also shown to be associated with a candidate TB vaccine trial outcome (Matsumiya, et al, 2013).

NOD1 of NLRs

Only just over a decade ago, the cytosolic NLRs were discovered as a novel class of PRRs, complementing the function of TLRs by their intracellular localization. They were named after NOD1, the first family member to be discovered. NLRs are primarily expressed in immune cells and epithelial cells. Not only do they sense invasive PAMPs but also endogenous DAMPs. NOD1 is a member of the intracellular pattern recognition receptors, NLRs, which are important for the recognition of unique muropeptides of bacterial peptidoglycan fragments (derived predominantly from the cell walls of *Mtb*) or damage-associated molecular patterns that localize to the cytosol. Following microbial sensing, NOD1 directly recruits factors which initiate intracellular signaling cascade that lead to the activation of transcriptional responses culminating

in the expression of a subset of innate immunity genes involved in inflammation (e.g., production of proinflammatory cytokines and chemokines), antimicrobial mechanisms and autophagy in cells responsible for eliminating *Mtb* including epithelial cells, alveolar macrophages and monocyte-derived macrophages (Moreira, et al, 2012). Mutations in several NLR members were found to be associated with the development of inflammatory disorders. Further understanding of NLRs should provide new insights into the mechanisms of host defense and the pathogenesis of inflammatory diseases (Franchi, et al, 2008). Recent studies using human and murine models revealed that the resolution of respiratory infections proceeds in a NOD1-dependent manner and that their genetic polymorphisms are linked to disease susceptibility and severity. For example, a study by (Esmeralda, et al, 2014) investigated the presence of NOD1 in human alveolar macrophages and examined its involvement with inflammatory cytokines and the induction of antimicrobial autophagy. NOD1 was expressed in alveolar macrophages (AMs), monocyte-derived macrophages (MDMs) and to a lesser extent monocytes (MNs). NOD1 up-regulation led to a significant production of IL1β, IL6, IL8, and TNFα in AMs and MDMs. Autophagy activity determined by expression of proteins was also induced in a NOD1-dependent manner in AMs and MDMs but not in MNs. In addition, recruitment of NOD1 and autophagy proteins to the Mtb localization site was observed in infected AMs. AMs are responsible for microbial lung clearance by orchestrating inflammatory responses that stimulate epithelial lung cells to produce additional chemokines and antimicrobial peptides, which amplify innate responses and help recruit other cells, such as monocytes and neutrophils. AMs are also the cells responsible for eliminating *Mtb*. MNs and MDMs, as human monocytic cells, have been reported to initiate proinflammatory responses following NOD1 ligand recognition. The study concluded that NOD1 is involved in the innate responses of both alveolar and monocyte-derived

macrophages, where it is upregulated and induces pro-inflammatory cytokines, autophagy, and signaling the transcriptional induction of genes involved in immune responses with potential implications in the killing of *Mtb* in humans. Similar conclusions were made by another study (Lee, et al, 2014) that investigated the role of NOD1 with respect to cytokine production by bone marrow-derived macrophages in response to *Mtb* infection.

Overview of the immune effector function of FMO2 of FMOs

FMO2 is a member of a super family of monooxygenase genes. In humans, eleven distinct FMO genes exist: five encoding active oxygenases (FMO1–5) and six pseudogenes (Hernandez, et al, 2004). The former are expressed in a developmental-, sex-, and tissue-specific manner (Zhang, et al, 2006). FMO2 is the major isoform predominantly and highly expressed in human lung (aka, Pulmonary FMO). The lung plays an important role in the metabolism of inhaled foreign chemicals, environmental toxicants, carcinogens, and drugs as well as being the main port of entry, deposition and establishment of inhaled infectious pathogens like *Mtb* (Henderson, et al, 2008). Human FMO2 possesses an FMO2*1(C)/FMO2*2(T) polymorphism: (g.23238C>T, dbSNP #rs6661174). The ancestral FMO2*1(C) allele encodes for a full-length functionally active enzyme whilst the derived alternate allele, FMO2*2(T), produces a truncated polypeptide that is functionally inactive due to a single-nucleotide transition mutation that converts a glutamine codon to a premature TAG stop codon in exon 9 (Dolphin, et al, 1988).

FMO2 oxygenase activity, oxidative stress and anti-mycobacterial innate immunity

Several studies have demonstrated the essential role of modulating oxidative stress levels in the innate antimycobacterial immune defense as it affects *Mtb* survival, persistence and subsequent

reactivation (Verma, et al, 2014; Bhimrao, et al 2011; Akiibinu, et al, 2011; Oyedeji, et al, 2013; Gebrehiwot, et al, 2015). Oxygenases in activated macrophages induce oxidative stress through generation of hypoxic conditions and highly reactive oxidants such as reactive oxygen species (ROS). Oxidative stress arises when oxidant load exceeds the endogenous antioxidant capacity. The process occurs in two major stages:1) Oxygenase mediated oxygen uptake leading to oxygen depletion (hypoxia) and, 2) oxygenase mediated generation of oxidizing species leading to the production of cytotoxic free radicals. Hypoxia keeps the aerobic *Mtb* in the latent stage and prevents its proliferation. The free radicals damage almost every part of the target cell (both host and pathogen) through instability and fragmentation of DNAs, proteins and lipids; dysfunction of enzymes; impairment of membrane functions (decreased fluidity, inactivation of membrane-bound receptors, and increased permeability to ions). Although the cytotoxic response of phagocytes causes damage to host tissue (e.g. necrosis), the non-specificity of oxidants is an advantage since it prevents a pathogen from escaping this part of the immune response by mutation of a single molecular target. In this regard, human FMO2 has been demonstrated to regulate the level of oxidative stress by the generation of metabolites that enhance the release of ROS in the form of H_2O_2 Siddens eatl 2014). Furthermore, it has been shown that a marked difference exists in ROS leakage from common allelic FMO2 variants. Another source of evidence for the involvement of FMO2 in anti-TB immune response through its oxidative potential comes from studies of the role of pharmacogenomics in the treatment of TB.

FMO2 oxygenase activity, metabolism of anti-tubercular drugs and pharmacogenomics

Besides pharmacogenomic studies into the general influence of genetic variation in patient response to anti-tubercular drug treatments including the development of serious adverse events (Ramachandran, et al, 2012), several pharmacokinetic studies have focused particularly on the

role of FMO2 (Henderson, et al, 2008; Kreuger, et al, 2005; Palmer, et al, 2012). FMO2 substrates are wide-ranging including therapeutic drugs, dietary-derived compounds and environmental pollutants including thioureas, a widely used class of industrial and pharmaceutical compounds. FMO2, through the same basic oxygenase activity that produces immunity-related oxidative stress, metabolizes drug-related exogenous substrates susceptible to oxidation. For example, pharmacogenomic evidences have shed light on how FMO2*1 enzyme functions in relation to the metabolism of the major thiourea-containing anti-MDRTB [defined as TB caused by strains of *Mtb* tuberculosis that are resistant to at least isoniazid and rifampicin (WHO, 2010) drugs such as ethionamide and thiacetazone that result in the production of toxic intermediates (Henderson, et al, 2004; Marilyn, et al, 2008; Francois, et al 2009).

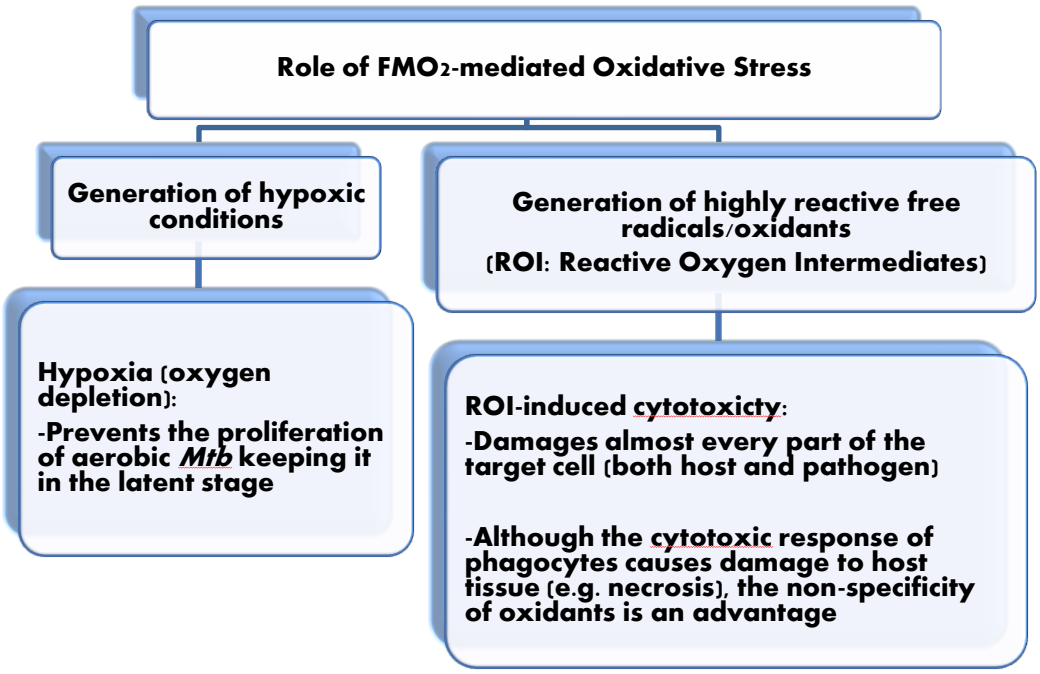
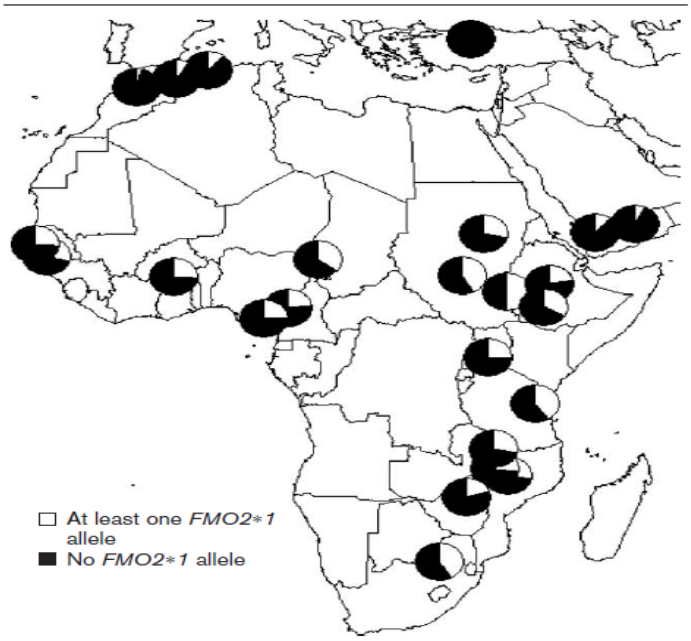


Figure 5: The dual role of oxidative stress driven by FMO2 encoded oxygenase enzyme

Ethnic differentiation in TB and FMO2

Several studies in different populations have identified genetic polymorphisms associated with the variable outcome of *Mtb* infection between individuals including African populations (Sirugo, et al, 2008). Studies in African populations, particularly sub-Saharan Africa, are important because both *Mtb* and humans are considered to have originated and co-evolved in this sub-continent (Brites, et al, 2015; Comas, et al, 2015; Gagneux, et al, 2012). Furthermore, studies have demonstrated the existence of ethnic-specific genetic associations with TB (Delgado, et al, 2002) as well as differentiation in anti-TB immune response profile between Africans and Europeans (Coussens, et al, 2013). A correspondingly distinctive ethno-geographic differentiation has been shown in the expression and distribution of FMO2 polymorphisms between African populations and those of non-recent African descent (Veeramah, et al, 2008; Whetstine, et al 2000; Krueger, et al, 2004). Particularly, all Europeans and Asians genotyped to date are homozygous for the dysfunctional FMO2*2 allele while, conversely, the functional FMO2*1 variant is found only in Africans (particularly in sub-Saharan Africa), recent African descendants and Hispanics (Figure-5).



Map showing the percentage of individuals with at least one *FMO2*1* allele in Africa and two nearby countries.

Figure 6: Differential distribution of FMO2*1/FMO2*2 (Veeramah, et al, 2008)

In general, the oxygenase activity of the FMO2*1 variant can be described as functioning in an antagonistic pleiotropy vis-à-vis TB: while possessing the FMO2.1 variant helps to fight *Mtb* infection by mounting innate immune responses via its involvement in the modulation of pulmonary oxidative stress level it also increases the risk of pulmonary toxicity by inducing the adverse metabolism of particular anti-TB drugs. Furthermore, the differential ethno-geographic distribution of FMO2*1 means that there would be a corresponding risk-benefit profile in various populations with regard to resistance to TB and susceptibility to adverse reactions to anti-TB drug treatment. Accordingly, the simultaneous prevalence in sub-Saharan Africa of both high endemic TB and a genetic risk factor for adverse anti-TB drug treatment has led some researchers to characterize FMO2*1 as a "potentially deleterious" variant (Veeramah, et al, 2008). In this regard, it was estimated that some 220 million individuals in sub-Saharan Africa may express a functional FMO2 enzyme and, therefore, potentially at risk of FMO2 mediated

toxicity. However, despite evidences suggesting the involvement of FMO2 in antimicrobial immune response, no studies were done to investigate the "potentially beneficial" aspect of the FMO2*1 variant with regard to TB pathogenesis. The significance of such a study would also be robust since FMOs in general have been shown to be not markedly regulated or inducted by environmental factors and one could predict inter-individual variability in FMO enzyme expression would be predominantly genetic in origin and it may be possible to more clearly delineat correlations (or lack thereof) between wild type FMO or particular FMO allelic variants and susceptibility to disease (Kreuger, et al, 2005).

VII. RATIONALE FOR SELECTION OF STUDY-POPULATIONS

Ethiopia: A 'model human population' for the study of the genetic profiles of diseases and therapeutics

Several anthropologic, linguistic, and genetic studies have confirmed that Ethiopia is the origin of human beings and harbours the highest genetic variation of all populations in the world to such an extent that diversity has been shown to decrease as geographic distance increases away from Ethiopia (Pagani, et al, 2012, 2015). Recent studies have also discovered that present day Ethiopian populations have the greatest proportion of novel genetic variation and that most of the novel variation appears to be unshared (private) and rare (Gurdasani, et al, 2014). Figure-6 shows: a, The overlap of SNPs between Zulu, Ugandan and Ethiopian individuals (subsampled to 100 samples each). b, The overlap of novel variants (those not in the 1000 Genomes Project phase I integrated call set, '1000G') between the three populations.

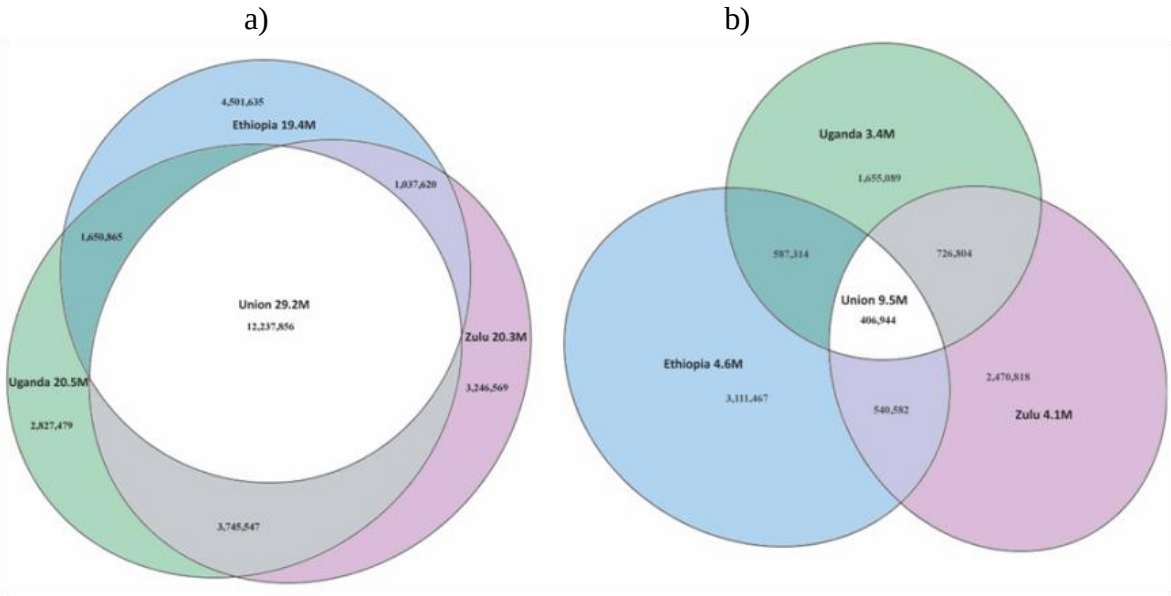


Figure 7: Allele sharing between sequenced populations in the African Genome Variation Project (Source: Gurdasani, et al, 2014)

Allele sharing between sequenced populations in theAGVP. a, The overlap of SNPs between 43WGS data from Zulu, Ugandan and Ethiopian individuals (subsampled to 100 samples each). b, The overlap of novel variants (those not in the 1000 Genomes Project phase I integrated call set, '1000G') between the three populations. There appear to be a large proportion of unshared (private) variants in each population: between 10% and 23% of the total number of variants in a given population. The proportion of novel variants was high, with Ethiopia showing the greatest proportion of novel variation. Most of the novel variation appears to be unshared and rare.

Therefore, it is reasonable to suggest that the tremendous genetic diversity existing in Ethiopia can be harnessed towards the understanding of the genetic basis of various infectious and non-infectious diseases that afflict humanity and their treatment. With respect to the current study of human genetic susceptibility to TB, Ethiopia is also 'well situated'. As referred to earlier, there is increasing evidence indicating the common origin in Ethiopia of MTBC and humans and that MTBC has been coevolving with anatomically modern humans for millenia. These conclusions are based on the observed congruence in their phylogeographies, the dating of major branching events, and the discovery of a unique lineage of MTBC localized only in Ethiopia. Therefore, studies of the impact of *Mtb*-human interactions would best be investigated in an Ethiopian setting.

VIII. MATERIALS AND METHODS

Ethical considerations

The research proposal received Ethical Clearance from the relevant institutions in Ethiopia:the Ethical Review Committee of Department of Biology at Addis Ababa University, andthe National Health Research Ethics Review Committee of the Federal Ministry of Science and Technology of Ethiopia. All participants were voluntarily recruited after due explanation of the objectives and risks of the research and after obtaining a signed informed consent form.

Selected study-populations

The study-populations were selected in a manner that traverses and represents some of the major ethno-geographic groups in Ethiopia: the North, Central and Southern ethno-geographies (Figure-7): Adigrat Hospital, Adigrat (mostly of the Tigre ethnic group, North Ethiopia), Alem Ketema Hospital, Merhabete (mostly of the Amhara ethnic group, Central Ethiopia), and Arbaminch Hospital, Arbaminch (mostly of the Gamo ethnic group, Southern Ethiopia). The three study sites were identified based on the FMoH's 2011 publication on TB epidemiology and distribution (incidence/prevalence) in Ethiopia; the availability of hospitals in the region capable of diagnosing and treating TB; geographical location; and financial resources available.

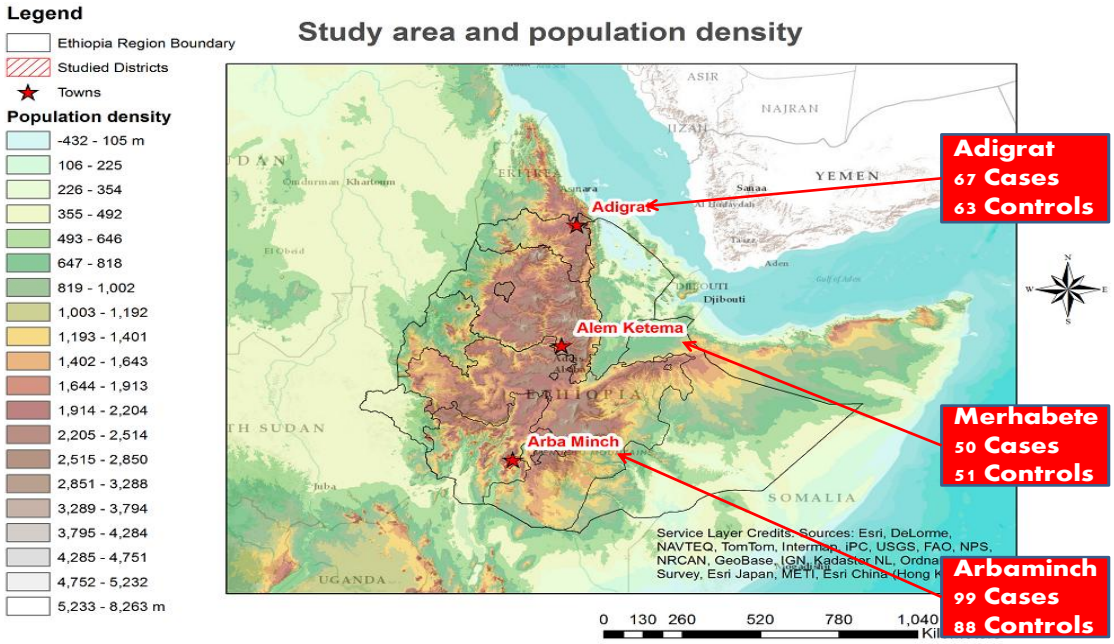


Figure 8: Selected study populations and sampling sites

Phenotyping (Clinical characterization)

As described before, because of the complexity of TB, a precise and consistent definition of the disease is a major challenge and it has been difficult to provide a reliable TB phenotype amenable for genetic analysis. Therefore, in this study, an effort has been made to ascertain samples using carefully defined phenotypic criteria.

Blood sample collection

Sampling was performed from 2011-2013. Questionnaires, asked by interpreters in the local languages when necessary, on demography (sex, age, ethnicity) and TB-related history were filled out before obtaining blood samples. 10ml blood was drawn into vacutainers from both cases and controls. An additional 3ml of blood was collected from controls only (1ml each drawn

directly into the Nil, Antigen and Mitogen QFT-assay tubes). All blood samples were drawn by qualified hospital personnel. Plasma and buffy-coat were extracted on sampling-site from all 10ml blood samples while the QFT-tubes were incubated right after blood being drawn.

Detection of active TB

Active pulmonary TB was diagnosed by hospital medical personnel as per the guidelines provided by the FMoH of Ethiopia (FMoHE, 2005) and all cases were undergoing anti-TB treatment during recruitment.

Detection of latent TB infection

In this study a QuantiFERON®-TB Gold In-Tube (QFT®, www.cellestis.com) kit for interferon-gamma release assay (IGRA) was used to detect LTBI. IGRA is an in vitro diagnostic method that indirectly tests for latent *Mtb* infection. The product manual describes that QFT is a test for Cell Mediated Immune (CMI) responses to *Mtb*-specific peptide antigens that simulate mycobacterial proteins. According to the manufacturer's description, these proteins, ESAT-6, CFP-10 and TB7.7(p4), are absent from all BCG strains and from most non-tuberculosis mycobacteria with the exception of M. kansasii, M. szulgai and M. marinum. Individuals infected with *Mtb* complex organisms usually have lymphocytes in their blood that recognize these and other mycobacterial antigens. This recognition process involves the generation and secretion of the cytokine, IFN-γ. The detection and subsequent quantification of IFN-γ forms the basis of this test. 1ml QFT-tubes were incubated and processed on site as described in the QuantiFERN-TB Gold In-Tube product protocol on randomly selected control samples; and,

ELISA was performed on the incubated blood samples at the Ethiopian Health and Nutrition Research Institute laboratories. Results of IGRA are provided in Supplementary Table-1.

Detection of HIV serostatus

Standard HIV test kit (obtained from the Federal Pharmaceutical Supplies Agency of Ethiopia) was used to determine HIV status of all samples at Department of Microbial, Cellular and Molecular Biology, Genetics Research Laboratory, of Addis Ababa University (Supplementary Table-2).

Genotyping/Exonic region sequencing

DNA extraction

Genomic DNA was extracted from all buffy-coat samples using QIAGENE (FlexiGene) DNA extraction kit and following the procedures described in the product manual, at the Genetics Research Laboratory of the Microbial, Cellular, and Molecular Biology Study Unit of Addis Ababa University. DNA quality and quantity were determined using Nanodrop , at the Holeta Agricultural Research Institute and, later, using Qubit at Case Western Reserve University (CWRU). DNA aliquots were shipped to CWRU laboratories for sequencing after obtaining an export permit from the National Health Research Ethics Review Committee of the Federal Ministry of Science and Technology of Ethiopia and an import permit from the Centers for Diseases Control of the USA. All remaining DNA samples (and plasma) were stored at AAU Genetics Research Laboratory.

DNA sequencing

After DNA quality and quantity analyses, 380 samples were selected for sequencing on four plates of 96 wells each. Samples of cases and controls were randomized to minimize sequencing bias. Coding (exonic) regions of the three candidate genes were targeted for this association study since polymorphisms in these regions can have a relatively more direct effect on the translated protein sequence and, therefore, are more likely to alter their function. Sequencing was performed using Illumina MiSeq technology and the Homosapiens/UCSC/hg19 human genome reference panel. MiSeq generated sequence data for each individual sample was generated in GVCF format (containing all sequence data) for each of the four plates with 96 wells and data was merged using PLINK software.

DNA sequence quality control (QC)

In this study, although all apriori attempts were made to avoid potential biases through careful collection of case and control groups and appropriate laboratory practices, a thorough assessment of data quality were undertaken. Recommended protocols (Purcell, et al, 2007; Anderson, et al, 2010) for DNA sequence QC were carried out before statistically testing for association. The steps involved the identification and removal of substandard DNA samples (per-individual QC) and markers (per-SNP QC) that could introduce bias and lead to either false-positive or false-negative associations. The QC protocol was applied to genotypes that passed the genotype calling algorithms used by the Illumina platform. Downstream QC thresholds were selected that maximize individual and marker sizes while ensuring appropriate QC for both. For the per-individual QC, individuals with less than 90% genotyping rate (i.e., individuals missing genotypes for more than 10% of the total markers) were removed. And, as for the per-marker

QC, markers with genotyping failure rate of less than 95% (i.e., markers genotyped in less than 95% of all samples) and, markers with HWE deviation value of $p<0.001$, were all removed. Tests for significant genotyping difference between cases and controls were all negative. Other QC measures including tests for evidence of population stratification and appropriate adjustment thereof were carried out and will be described in the relevant sections. The data cleaning procedure removed more than 99% of the total of almost 20,000 exonic nucleotide sequences of the 3 candidate genes: 0.8%, 0.3%, and 0.3% markers were left for analysis from the FMO2, TICAM2 and NOD1 candidate genes, respectively. This is not surprising since, on average 94%, of the sequences were monomorphic (i.e., loci with no variation) in the entire sampled population dataset and, hence, uninformative for association analysis. Each test-model and subsequent population-specific (within population) analysis were subjected to the same QC criteria. DNA sequence data before and after QC filtering is provided in Supplementary Tables 3-5. Descriptions of markers before and after the QC process are provided in Supplementary tables 3-5.

Setup of statistical tests for association

The association analysis was set up in such a way that allows the comparison of the genetics (genotype/alleles) of individuals with a specific case-phenotype with that of individuals with a specific control-phenotype. Case-control phenotypes wrere defined in increasingly restrictive or exclusive criteria. This set up helps to study the impact of phenotype definition and perform a sensitivity analysis by observing the trend of association test statistic across the datasets. In fact, this statistical test-model is the major strength of the present study: the fact that controls were

rigorously characterized in terms of not just being without symptomatic active pulmonary TB (No Active TB) but also with regard to being either latently infected (LTBI) or not latently infected (No LTBI).

Basic single SNP association analysis

Basic allelic test of association by comparing the minor allele frequency of individual SNPs between cases and controls. Both chi-square tests of Pearson's and Fisher's exact test were used.

Logistic regression analysis

Association statistic based on additive model, i.e., that each extra copy of the risk allele increases risk equally. This is an association analysis based on comparing allelic dosage [coded as 0 (minor allele absent), 1 (heterozygote), 2 (homozygote for the minor allele)] for significance differences between cases and controls using logistic regression model.

Covariate analysis

Covariates directly or indirectly influence disease processes and such factors may act as confounders and obscure the true relationship between the independent (in this case SNPs) and the dependent (in this case TB status) variables in a study and need to be accounted for. The framework of logistic regression also allows the inclusion of covariates, test their effect on the phenotype, and account for them. Sex and ethno-geographic categories (EGCs/sample collection sites: Merhabete vs. Adigrat; Merhabete vs. Arbaminch, Adigrat vs. Arbaminch) were included as covariates in a logistic regression model, and age in a linear regression model. The logistic regression test without covariate inclusion is an estimate of SNP-effect size referring to the effect

of each extra risk allele (minor allele, A1) based on the additive (allelic dosage: each allele is assumed to have equal effect) model.

Examining genotypic models

Tests for whether specific genotype configurations have specific risk profiles were performed using the dominant/recessive models for tests of dominance components and the tests for alternate genotypic models (i.e., instead of tests of additive and dominant components, explicitly test for heterozygote and homozygote effects). In other words, besides the aforementioned basic allelic and logistic regression association tests that compare frequencies of alleles in cases and controls and the additive test that examines the effect of each extra minor allele, respectively, tests to examine whether different genotypic configurations of the disease-associated SNPs have specific risk profiles were also conducted: genotypic (AA, Aa, aa), dominant (AA, Aa vs. aa), recessive (AA vs. Aa, aa), and heterozygotes vs. homozygote effects (AA/aa vs. Aa).

Test for population specific effects (heterogeneity test)

This test was done to check for possible population/EGC-specific SNP-phenotype associations or test whether there was a difference in the strength of association between EGCs.

Linkage disequilibrium (LD) estimation and haplotype-/LD-based association tests for independent effect

SNPs that were found to be significantly associated with a phenotype were tested to check if they have an effect independent of the haplotypic background formed by the remaining SNPs that also showed phenotype-association signal. The test procedure involves inference of haplotypes based

on the phenotype-associated SNPs, grouping similar haplotypes, and testing for differences between cases and controls.

Empirical assessment and visualization of population stratification

This was performed using multidimensional scaling (MDS) analysis based on identity-by-state (IBS) of SNPs to clusters individuals into homogeneous groups and performing to investigate population structure and identify stratification, if any. PLINK employs complete linkage agglomerative clustering, based on pair-wise SNP IBS distance. The procedure converts the proportion of SNPs IBS into a distance matrix and creates relatively homogeneous clusters that enable visualization of population stratification as well as statistical adjustments for population stratification.

Population-stratified single SNP association analysis

Stratified analysis was performed that conditions on or adjusts for potential stratification effects based on self-declared ethnicity and sampling site (EGC) as well as on algorithms that analyze for hidden population stratification based on empiric MDS-inferred homogenous clusters.

To summarize, the various tests described above serve not only to test for SNP-phenotype associations but also account for possible confounding. Confounding occurs when an apparent association between an exposure (genotype) and outcome (TB) is observed by the presence of other variables: confounders and effect modifiers. A confounder is linked with both the potential risk factor and the outcome but is not a causal factor itself while effect modification occurs when

a variable differentially modifies the observed effect of a risk factor (genotype) on disease status, i.e., the effect is real but the magnitude is different for different groups of individuals).

Methods of interpretation of association test results in this study

Association is a statistical estimate about the co-occurrence of two events: in GE these are alleles/genotypes and phenotypes (Thomas, 2004; Strachan, et al, 2011). Allele A is said to be associated with disease D if people who have D also have A significantly more often (or less often) than would be predicted from their respective frequencies in the population. The strength of the association is measured by the odds ratio (OR). ORs are used to compare the relative odds of the occurrence of the outcome of interest, given exposure to the variable of interest. An odds ratio (OR) is described as being a measure of association between an exposure (in this case the minor allele is taken as the effect allele) and an outcome (in this case TB phenotypes). The OR represents the odds that an outcome will occur given exposure to the effect allele, compared to the odds of the outcome occurring in the absence of that exposure, i.e., compared to exposure to the reference allele (the major allele, usually, the ancestral allele). In this study, the OR is used to determine whether exposure to the minor-allele of a particular SNP is a risk factor for a particular TB phenotype, and to compare the magnitude of the SNP-effect for that phenotype. [Note: OR=1: the minor allele confers no additional risk of TB; OR>1: minor allele associated with higher odds of TB, increased risk; OR<1: minor allele associated with lower odds of TB, reduced risk]. The confidence interval (CI) is used to estimate the precision of the OR; a large CI indicates a low level precision of the OR estimate, where as a small CI indicates a higher precision. In practice, the 95% CI is often used as a proxy for the presence of statistical

significance if it does not overlap the null OR value (e.g., OR=1, no effect difference between the alleles). Nevertheless, it would be inappropriate to interpret an OR with 95% CI that spans the null value as indicating evidence for lack of association between the exposure and outcome.

Population association tests are considered more powerful to detect weak signals of genetic susceptibility than other methods; however, association can have several possible causes, not all of which are genetic (Strachan T 2011), some of which are:

Direct causation: while the possession of a certain allele may be neither necessary nor sufficient, having the particular allele increases the likelihood of, or susceptibility to, a disease (risk allele).

Epistatic and pleiotropic effects: occurs when alleles within the same gene (pleitropy) or alleles at different loci/gene (epistsis) interact in a synergistically or antagonistically.

Population stratification: occurs when a certain population contains several genetically distinct subgroups, and both a specific allele and a specific disease trait happen to be particularly frequent in one subgroup.

Type I error: occurs when a large number of markers are interrogated for association with a disease. It is estimated that 5% of the results will be significant at p=0.05 level, even without true effect; these are false positives or type I errors and need to be corrected for multiple testing.

LD: occurs when a disease-associated allele marks or is linked to an ancestral chromosome segment that carries a sequence variant that actually causes susceptibility to the disease, i.e., the associated-site is acting as a proxy for the causal site.

In this study, both empirical and user-defined methods were employed to identify and account for possible spurious associations. Association analyses were performed within each test-models and SNP-phenotype associations were declared nominally significant at $p<0.05$ and deemed as achieving study-wide significance if they passed the Bonferroni-adjusted threshold in any one of the statistical test procedures. (Note: Bonferroni adjustment of p-values are strict corrections for multiple testing that can be made in two ways depending on whether reference is made to the p-value threshold for the entire test or the p-value for a particular SNP: a) for a 5% significance threshold, divide 0.05 with the total number of tested SNPs in a dataset and set the value as the corrected threshold, or, as was done here, b) multiply the p-value for each SNP by the number of tested SNPs in a dataset and stick with the 5% significance for each tested SNP.) Finally, different statistical analysis of associations may result in different statistic and cannot be expected to do so since the procedures differ. For example, the Fisher's exact test is based on counts while the Pearson test works on frequencies. Therefore, one test might fail to detect a significant association while the other could pick a signal. The Fisher's exact test is preferred for samples with small sizes. Logistic regression works on allelic dosage (coded 1 for Aa, 2 for aa, and 0 for AA) rather than allele counts and frequencies, and the differences might be particularly large for very rare alleles where the SNP is monomorphic in either cases or controls.

Statistical analyses software

PLINK (version 1.07), R (version 3.1.1), and SPSS (version 15.0) software were employed for statistical analysis and visualization.

IX. RESULTS AND DISCUSSION

Basic SNP-based association tests

The following section presents the results of the various association tests and since significant genetic associations were observed for both susceptibility and resistance to a disease state, the results are tabulated separately. Corresponding tables are provided that show details of statistical test results: values for significance tests, odds ratios of effect, and 95% confidence intervals are listed for the best results only while significance values are shown for each phenotype-associated SNP corresponding to the specific test. Noticeable difference were observed between the four test-models in the identity and number of significantly phenotype-associated SNPs detected as well as the pattern (strength and direction) of significant associations which will be discussed later in the 'Pattern of Significant Associations' section. All other genotypic test models (i.e., besides the basic allelic tests that compare frequencies of alleles in cases and controls and the additive test that examines the effect of each extra minor allele), intended to fit a dominant or recessive model for the minor allele were either non-significant or were not applicable for all SNPs as these model-based tests require a minimum number of (e.g., 5) observations per cell, a criteria not always met. Therefore, only the allelic/additive and trend (Cochran-Armitage) tests were viable. Descriptions minor allele frequencies for all phenotype-associated SNPs are provided in Supplementary Table 6.

Association test results in the 'Active TB vs. No Active TB' test-model dataset

In this test-model with active TB as the phenotype of interest, there were 153 active pulmonary TB cases and 139 controls with no symptoms of active TB. There were a total of 94 QC-passed SNPs available for analyses (58 SNPs in FMO2, 12 in TICAM2, and 24 in NOD1 genes). A total of 14 SNPs showed at least a nominal signal of being significantly associated with susceptibility to active TB: 8 SNPs in FMO2, 4 in NOD1, and 2 in TICAM2. Of these SNPs, study-wide significance was met by 4 SNPs: 2 in FMO2 (chr1:171181877, $p=3.15x10^{-07}$, OR=4.644, 95% CI 2.425-8.893; chr1:171165749, $p=3.32x10^{-06}$, OR=6.825, 95% CI 2.63-17.71), and 2 SNPs in NOD1 (chr7:30485722, $p=7.28x^{-05}$, OR=4.111, 95% CI 1.946-8.684; chr7:30477156, p=0.0001037, OR=16.66, 95% CI 2.214-125.3). On the other hand, multiple SNPs were found to be nominally associated with resistance to active TB: 12 SNPs in FMO2, 3 in NOD1, and 1 in TICAM2. P-values ranged from 0.0029 to 0.048 and OR= 0.1-0.5. Results are presented in Tables 2-3.

Table 2: Results of SNP-based association analysis in 'Active TB vs. No Active TB' test-model: Increased risk

Active TB vs. No Active TB: Genes/SNPs associated with susceptibility to active TB																														
Best results								Fisher				Pearson				Logistic reg.				Covariate					CMH				
Gene	SNP	A1	P	BONF.	OR	L95	U95	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Sex	Age	Mer-Adi	Mer-Arb	Adi-Arb	EGC	IBS	IBS-Mer	IBS-Adi	IBS-Arb
FMO2	chr1:171181877	A	3.15E-07	2.96E-05	4.6	2.4	8.9	3.15E-07			2.12E-06	6.03E-07			7.85E-06	6.48E-07			1.19E-04	7.53E-07	4.72E-07	2.51E-02	1.07E-05	3.97E-06	1.77E-06	2.16E-06			4.12E-05
	chr1:171165749	T	3.32E-06	3.12E-04	6.8	2.6	17.7	3.32E-06			1.90E-04	6.72E-06			4.11E-04	3.95E-05			5.05E-03	4.27E-05	3.11E-05	1.61E-02	1.61E-03	5.18E-04	1.63E-05	8.34E-06			9.46E-04
	chr1:171179939	G	2.45E-02		2.3	1.1	4.6			4.47E-02				3.34E-02				2.45E-02						2.86E-02	3.54E-02				
	chr1:171180021	G	2.45E-02		2.3	1.1	4.6			4.47E-02				3.34E-02				2.45E-02						2.86E-02	3.54E-02				
	chr1:171174312	A	3.51E-02		1.8	1.0	3.0					4.40E-02				4.98E-02				4.33E-02					3.51E-02				
	chr1:171178490	T	3.51E-02		1.8	1.0	3.0					4.40E-02				4.61E-02				3.99E-02	4.63E-02				3.51E-02				
	chr1:171168469	A	4.72E-02		3.4	0.9	12.4					4.72E-02																	
	chr1:171181150	A	4.72E-02		3.4	0.9	12.4					4.72E-02																	
NOD1	chr7:30485722	T	7.28E-05	6.84E-03	4.1	1.9	8.7	8.25E-05			2.58E-04	7.28E-05			4.40E-04	1.02E-04			1.72E-03	8.69E-05	7.57E-05		8.90E-04	2.32E-04	1.76E-04	9.69E-05			1.44E-03
	chr7:30477156	T	1.04E-04	9.54E-03	16.7	2.2	125.3	2.61E-04			1.04E-04	2.30E-04			2.78E-04	2.31E-04			4.17E-03	2.54E-04	1.73E-04		8.82E-04	2.21E-04	2.36E-04	4.85E-04			6.07E-04
	chr7:30490711	T	2.07E-03		3.5	1.5	7.7	2.29E-03				2.26E-03				2.44E-03				2.65E-03	2.08E-03		2.14E-03	2.38E-03	2.07E-03	6.63E-03			
	chr7:30491081	A	1.59E-02		4.5	1.2	16.8	2.12E-02				1.78E-02				3.73E-02				4.81E-02	4.50E-02					1.59E-02			
TICAM	chr5:114915999	A	1.91E-02		10.0	1.1	93.3	3.96E-02				2.72E-02														1.91E-02			
	chr5:114916090	G	2.30E-02		1.8	1.1	3.0				3.39E-02				2.54E-02				2.95E-02										2.30E-02

Table 3: Results of SNP-based association analysis in 'Active TB vs. No Active TB' test-model: decreased risk

Active TB vs. No Active TB: Genes/SNPs associated with resistance to active TB																													
Best results								Fisher				Pearson				Logistic reg.				Covariate					CMH				
Gene	SNP	A1	P	BONF.	OR	L95	U95	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Sex	Age	Mer-Adi	Mer-Arb	Adi-Arb	EGC	IBS	IBS-Mer	IBS-Adi	IBS-Arb
FMO2	chr1:17117977	G	5.08E-03		0.5	0.35	0.83	3.48E-02			2.84E-02	3.27E-02			2.47E-02	3.25E-02			2.48E-02	2.30E-02	2.73E-02			5.08E-03	1.54E-02				
	chr1:17118007	G	5.08E-03		0.5	0.35	0.83	3.48E-02			2.84E-02	3.27E-02			2.47E-02	3.25E-02			2.48E-02	2.30E-02	2.89E-02			5.08E-03	1.54E-02				
	chr1:17118020	C	5.08E-03		0.5	0.35	0.83	3.48E-02			2.84E-02	3.27E-02			2.47E-02	3.25E-02			2.48E-02	2.30E-02	2.77E-02			5.08E-03	1.54E-02				
	chr1:17117809	C	1.18E-02		0.3	0.88	1.00	1.51E-02				1.18E-02				1.70E-02				1.39E-02	1.74E-02			2.92E-02	2.34E-02				
	chr1:17117902	C	1.18E-02		0.3	0.88	1.00	1.51E-02				1.18E-02				1.70E-02				1.39E-02	1.74E-02			2.92E-02	2.34E-02				
	chr1:17117476	C	1.59E-02		0.2	0.86	1.00	2.05E-02				1.59E-02				1.93E-02				1.87E-02	2.08E-02			3.28E-02					
	chr1:17117324	C	1.72E-02		0.6	0.38	0.91	3.71E-02				2.97E-02				3.60E-02				2.59E-02	3.84E-02			2.10E-02	1.72E-02				
	chr1:17117469	A	1.88E-02		0.6	0.38	0.92	3.56E-02				2.92E-02				3.63E-02				2.71E-02	3.83E-02			2.37E-02	1.88E-02				
	chr1:17117482	A	1.88E-02		0.6	0.38	0.92	3.56E-02				2.92E-02				3.63E-02				2.71E-02	3.83E-02			2.37E-02	1.88E-02				
	chr1:17117687	A	1.88E-02		0.6	0.38	0.92	3.56E-02				2.92E-02				3.63E-02				2.71E-02	3.83E-02			2.37E-02	1.88E-02				
	chr1:17117785	T	2.98E-02		0.5	0.28	0.94	3.49E-02				3.10E-02				4.29E-02				2.98E-02	4.69E-02								
	chr1:17117947	T	4.90E-02		0.3	1.00	1.00					4.90E-02																	
NOD1	chr7:30464249	TG	1.20E-02		0.1	0.01	0.73																						1.20E-02
	chr7:30465424	C	1.20E-02		0.1	0.01	0.73																						1.20E-02
	chr7:30498962	C	2.85E-02		0.2	0.03	0.93								3.23E-02				4.79E-02					3.39E-02	4.41E-02				2.85E-02
TICAM2	chr5:11491602	A	2.98E-03		0.3	0.13	0.69	2.24E-02			8.38E-03	1.96E-02			6.25E-03	2.52E-02			1.35E-02	2.82E-02	2.72E-02		2.04E-02	3.40E-02	2.96E-02	1.55E-02			2.98E-03	

Association test results in the 'Active TB vs. No Latent TB (No LTBI)' test-model dataset

In this test-model the control phenotype was restricted to those individuals that tested negative to *Mtb*-specific interferon-gamma release assay (IGRA) (n=64), indicating the absence of latent infection despite being a household contact of active TB patients, and compared with active TB patients as cases (n=153). A total of 91 SNPs passed QC: 58 in FMO2, 12 in TICAM2, and 21 in NOD1. A total of 5 SNPs showed at least nominal significance association with susceptibility to active TB. Of these SNPs 3 were significant at the Bonferroni corrected level: 2 SNPs in FMO2 (chr1:171181877, p=3.19x10^{-06}, OR=8.729, 95% CI 2.675-28.48; chr1:171165749, p=6.58x10^{-05}, OR=15.88, 95% CI 2.149-117.3), and 1 SNP in NOD1 (chr7:30485722, p=0.0001597, OR=8.665, 95% CI 2.056-6.52). Furthermore, 8 SNPs were also found to be associated with resistance to active TB at the nominal level: 5 SNPs in FMO2 and 3 in NOD1. Results are presented in Tables 4-5.

Table 4: Results of SNP-based association analysis in 'Active TB vs. No LTBI' test-model: increased risk

Active TB vs. No LTBI: Genes/SNPs associated with susceptibility to active TB																														
Best results								Fisher				Pearson				Logistic reg.				Covariate					CMH					
Gene	SNP	A1	P	BONF.	OR	L95	U95	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Sex	Age	Mer-Adi	Mer-Arb	Adi-Arb	EGC	IBS	IBS-Mer	IBS-Adi	IBS-Arb
FMO2	chr1:171181877	A	3.19E-06	2.61E-04	8.7	2.7	28.5	3.19E-06		2.96E-02	4.31E-03	2.20E-05		2.70E-02	5.88E-03	1.12E-04		4.29E-02	1.30E-02	1.17E-04	7.75E-05	1.09E-02	2.63E-03	1.26E-03	9.94E-05	3.01E-05		3.28E-02	
	chr1:171165749	T	6.58E-05	5.40E-03	15.9	2.1	117.3	6.58E-05				3.13E-04				4.87E-03				4.84E-03	4.15E-03	4.08E-02		1.57E-02	7.37E-04	2.38E-04			
NOD1	chr7:30485722	T	1.60E-04	1.31E-02	8.7	2.1	36.5	1.60E-04			3.31E-02	4.70E-04			3.34E-02	2.04E-03			4.77E-02	2.03E-03	1.34E-03	4.08E-02	2.54E-02	7.91E-03	1.65E-03	3.84E-04		4.74E-02	
	chr7:30477156	T	8.84E-03		3.4	1.4	8.7	1.43E-02				1.50E-02			4.65E-02	1.31E-02				1.33E-02	8.84E-03		2.48E-02	4.12E-02	2.21E-02	1.82E-02			
	chr7:30490711	T	1.89E-02		3.7	1.2	11.1	2.73E-02				2.28E-02				2.40E-02				2.54E-02	1.89E-02		3.07E-02		3.00E-02	2.90E-02			

Table 5: Results of SNP-based association analysis in 'Active TB vs. No LTBI' test-model: decreased risk

Active TB vs. No LTBI: Genes/SNPs associated with resistance to active TB																												
Best results								Fisher				Pearson				Logistic reg.				Covariate					CMH				
Gene	SNP	A1	P	BONF.	OR	L95	U95	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Sex	Age	Mer-Adi	Mer-Arb	Adi-Arb	EGC	IBS	IBS-Mer	IBS-Adi	IBS-Arb
FMO2	chr1:171174762	C	1.42E-02		0.2	0.8	1.0	1.90E-02				1.42E-02				1.48E-02				1.57E-02	1.73E-02			3.37E-02					
	chr1:171178090	C	2.57E-02		0.5	0.3	0.9					4.06E-02				4.28E-02				4.11E-02	2.57E-02								
	chr1:171179025	C	2.57E-02		0.5	0.3	0.9					4.06E-02				4.28E-02				4.11E-02	2.57E-02								
	chr1:171180201	C	3.72E-02		0.6	0.3	1.0																	3.72E-02					
	chr1:171179477	T	4.04E-02		0.6	0.3	1.0															4.04E-02							
NOD1	chr7:30498962	C	1.87E-02		0.2	0.0	0.9					3.88E-02			1.92E-02				3.74E-02		4.85E-02			1.89E-02	1.87E-02	3.38E-02			
	chr7:30464932	G	2.46E-02		0.1	0.0	0.8					3.88E-02									4.92E-02			2.46E-02	4.84E-02				
TICAM2	chr5:114916028	A	3.08E-03		0.4	0.2	0.8	8.79E-03			1.39E-02	4.98E-03			7.95E-03	6.24E-03			1.43E-02	6.91E-03	6.86E-03		2.06E-02	1.64E-02	1.43E-02	3.08E-03			

Association test results in the 'Active TB vs. Latent TB (LTBI)' test-model dataset

In this test-model, in a further attempt at narrowing the control phenotype, controls were defined as those individuals that tested positive to IGRA indicating latent TB infection (n=70). These controls were compared with active TB patients (n-153) based on a total of 92 QC-passed SNPs: 56 SNPs in FMO2, 12 in TICAM2, and 24 in NOD1. A total of 10 SNPs showed a nominal significance of association with susceptibility to active TB: 4 SNPs in FMO2, 5 in NOD1, and 1 in TICAM. Of these, 2 SNPs in FMO2 gene were significant at study-wide level: chr1:171181877, p=0.000454, OR=5.904, 95% CI 2.188-15.93; chr1:171165749, p=0.0007443, OR=5.708, 95% CI 1.722-18.92. In this dataset also 5 SNPs were found to be nominally

associated with resistance to active TB. P-values ranged from 0.021to 0.039 and ORs from 0.16 to 0.58. Results are presented Tables 6-7.

Table 6: Results of SNP-based association analysis in 'Active TB vs. LTBI' test-model: increased risk

Active TB vs. LTBI: Genes/SNPs associated with susceptibility to active TB																											
Best resultss								Fisher				Pearson				Logistic reg.				Covariate					CMH				
Gene	SNP	A1	P	BONF.	OR	L95	U95	Combined	Merhabete	Adigrat	Arbaminoh	Combined	Merhabete	Adigrat	Arbaminoh	Combined	Merhabete	Adigrat	Arbaminoh	Sex	Age	Mer-Adi	Mer-Arb	Adi-Arb	EGC	IBS	IBS-Mer	IBS-Adi	IBS-Arb
FMO2	chr1:171181877	A	4.54E-04	4.18E-02	5.9	2.2	15.9	5.92E-04				9.32E-04				6.11E-04				6.02E-04	6.18E-04		6.52E-04	4.54E-04	1.12E-03	1.40E-03			
	chr1:171165749	T	7.44E-04	6.85E-02	5.7	1.7	18.9	7.44E-04				1.44E-03				2.86E-03				2.84E-03	2.87E-03			8.51E-03	1.68E-03	1.65E-03			
	chr1:171179939	G	2.23E-02		1.7	1.1	2.8																	2.23E-02					
	chr1:171180021	G	2.23E-02		1.7	1.1	2.8																	2.23E-02					
NOD1	chr7:30477156	T	2.35E-03		9.8	2.3	42.7	2.88E-03				3.31E-03				3.58E-03				3.95E-03	3.65E-03		1.36E-02	2.35E-03	3.38E-03				
	chr7:30485722	T	4.53E-03		4.9	1.6	14.6	9.07E-03				9.54E-03				8.72E-03				7.59E-03	8.91E-03		8.92E-03	4.53E-03	1.04E-02	1.23E-02			
	chr7:30490711	T	1.48E-02		6.3	1.4	27.9	3.35E-02				2.92E-02				2.86E-02				3.01E-02	2.92E-02		3.75E-02	1.48E-02	3.04E-02	4.09E-02			
	chr7:30491081	A	4.50E-02		6.3	0.8	48.3					4.50E-02														4.97E-02			
	chr7:30464872	A	4.51E-02		4.0	0.9	17.5																		4.51E-02				
TICAM2	chr5:114916090	G	4.55E-02		1.8	1.0	3.3					4.55E-02																	

Table 7: Results of SNP-based association analysis in 'Active TB vs. LTBI' test-model: decreased risk

Active TB vs. LTBI: Genes/SNPs associated with resistance to active TB																											
Best results								Fisher				Pearson				Logistic reg.				Covariate					CMH				
Gene	SNP	A1	P	BONF.	OR	L95	U95	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Sex	Age	Mer-Adi	Mer-Arb	Adi-Arb	EGC	IBS	IBS-Mer	IBS-Adi	IBS-Arb
FMO2	chr1:171168545	C	2.10E-02		0.2	0.0	0.9			0.0363				0.0223				0.025				0.025						0.021	
	chr1:171179779	G	2.40E-02		0.6	0.3	0.9				0.046				0.0403				0.0396	0.0423				0.024	0.0484				
	chr1:171180071	G	2.40E-02		0.6	0.3	0.9				0.046				0.0403				0.0396	0.0423				0.024	0.0484				
	chr1:171180201	C	3.83E-02		0.6	0.4	1.0																	0.0383					
	chr1:171154303	C	4.00E-02		0.2	1.0	1.0							0.04				0.0493										0.042	

Association test results in the 'Latent TB (LTBI) vs. No Latent TB (No LTBI)' test-model dataset

The last test-model was set up in such a way that the outcome of interest is latent TB infection as compared to no latent TB infection considered as the control phenotype as determined by the *Mtb*-specific IGRA. A total of 3 (2 in FMO2 and 1 in NOD1) and 5 SNPs (all in FMO2) were associated nominally with susceptibility to latent TB infection and resistance to LTBI, respectively. None of them survived Bonferroni correction. The p-values for both the susceptibility and resistance SNPs ranged from 0.02 to 0.04. Results are presented in Tables 8-9.

Table 8: Results of SNP-based association analysis in 'LTBI vs. No LTBI' test-model: increased risk

LTBI vs. No LTBI: Genes/SNPs associated with susceptibility to LTBI																													
Best results								Fisher				Pearson				Logistic reg.				Covariate					CMH				
Gene	SNP	A1	P	BONF.	OR	L95	U95	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Sex	Age	Mer-Adi	Mer-Arb	Adi-Arb	EGC	IBS	IBS-Mer	IBS-Adi	IBS-Arb
FMO2	chr1:171168545	C	2.21E-02		8.5	1.0	72.4			2.84E-02				2.21E-02				3.76E-02										2.88E-02	
	chr1:171181877	A	4.13E-02		5.4	1.1	27.4															4.13E-02							
NOD1	chr7:30469270	C	4.03E-02		1.9	1.0	3.7													4.03E-02						4.08E-02			

Table 9: Results of SNP-based association analysis in 'LTBI vs. No LTBI' test-model: decreased risk

LTBI vs. No LTBI: Genes/SNPs associated with resistance to LTBI																													
Best results								Fisher				Pearson				Logistic reg.				Covariate					CMH				
Gene	SNP	A1	P	BONF.	OR	L95	U95	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Combined	Merhabete	Adigrat	Arbaminch	Sex	Age	Mer-Adi	Mer-Arb	Adi-Arb	EGC	IBS	IBS-Mer	IBS-Adi	IBS-Arb
FMO2	chr1:171179287	T	2.55E-02		0.0	0.9	1.0		3.21E-02				2.55E-02														2.83E-02		
	chr1:171179670	G	4.23E-02		0.0	1.1	1.0						4.23E-02														4.91E-02		
	chr1:171179939	G	4.40E-02		0.3	0.1	1.0																						4.40E-02
	chr1:171180021	G	4.40E-02		0.3	0.1	1.0																						4.40E-02
	chr1:171179477	T	4.54E-02		0.1	1.0	1.0						4.54E-02																

Covariate analyses

Figure-8 displays the comparisons between regression coefficient values before and after covariate inclusion in the logistic regression model. Introduction of covariates in the logistic regression model in order to control for their potential effect can result in either an increase or

decrease of the statistical measures of SNP-effect (minor allele) on the phenotype mean and is also reflected in the changing values of the regression coefficient. A rule-of-thumb is that if the regression coefficient value tends to approach a value of zero ('Descent-to-Zero'), it suggests a possible confounding effect by the introduced covariate variable and, hence, needs to be accounted for. A descending/ascending line indicates a trend towards a decreasing/increasing effect: as an estimate of the effect-size, since it is a function (exponentiation) of the ORs. It is apparent from the chart that sex and age as covariates had no impact on SNP-effect: the corresponding reg. coef. plots almost perfectly fitted the logistic regression coefficient line. On the other hand, the incorporation of ethno-geographic covariates in the logistic regression model resulted in relatively more pronounced degrees of deviations in the ORs and p-values of some SNPs, particularly SNPs with only nominally significant associations as can be seen in the chart by the lines lying either below or above the logistic reg. coef. plot lines (thick red line). This is consistent with the observation that, where as some SNPs incurred loss of significance other non-significant SNPs gained slight significance after the inclusion of these covariates. Again, the effect of highly significant SNPs on the corresponding phenotypes remained mainly unaffected by covariate inclusions. It can be concluded that, in terms of OR, logistic reg. coef. and p-value changes, the inclusion of different covariates had different effects for different SNPs in the same test-model as well as different effects for the same SNP in different test-models.

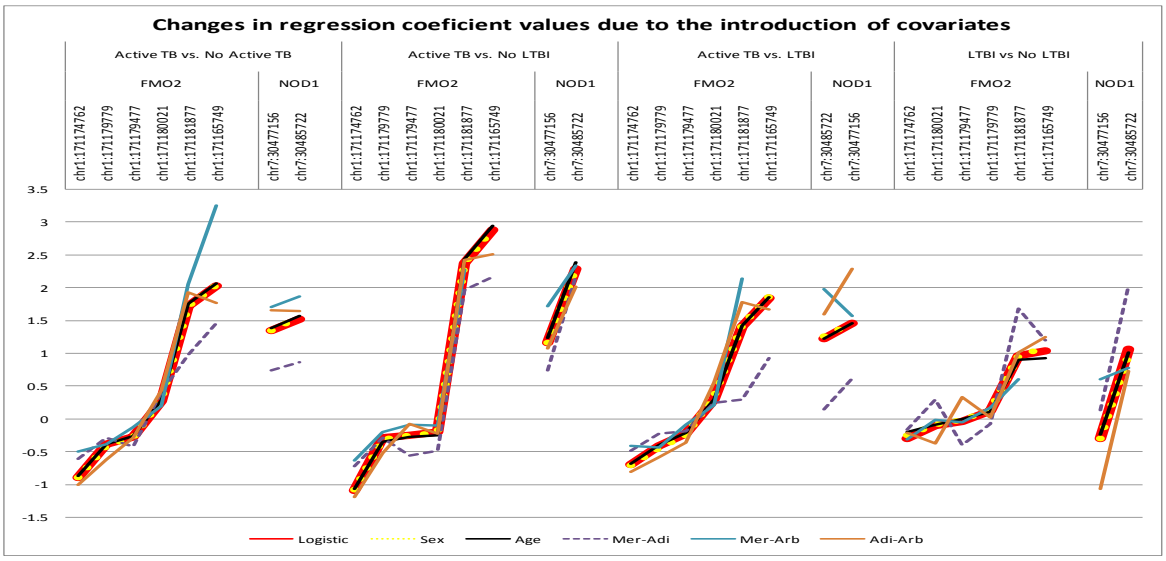


Figure 9: Covariate analysis: sex, age and EGC

Examining the regression coefficient is also important because the direction of the coefficient indicates whether the minor/risk allele (effect allele, A1) increases (positive association, OR>1/+ve beta) or decreases (negative association or protective effect, OR<1/-ve beta) risk, i.e., effect of the SNP on the phenotype, relative to the major allele (reference allele, A2). It was to demonstrate this that both susceptibility and resistance SNPs were included in the above chart.

The downside of covariate analysis, however, is that it comes at the cost of losing some degree of freedom and thus reduces statistical power to detect significant correlations between candidate SNPs and phenotype. This is especially true in studies with not so large sample sizes and, therefore, the observed changes in significance values (i.e., p-values for SNP-phenotype association after controlling for the covariates, as compared to the analysis without covariate inclusion), may not reflect the true effect of the covariate. In such cases it is advisable to closely

examine the pattern of changes in the regression coefficient values instead of the p-values themselves.

To conclude this part, the covariate analysis results suggest that it could be wise to perform a population-stratified test of association to account for possible confounding caused by ethno-geographic covariates, i.e. possible population stratification.

Analysis of patterns of significant associations

Consistency is an important concept in understanding cause-and-effect and a valuable criterion in interpreting association test results: "Consistency: the same association between a disease and a suspected causal agent should be found in studies of different populations. Failure to find consistency may be explained by differences in study designs; therefore, a causal hypothesis can be regarded as plausible only when there is a general consistency of findings from studies conducted in the same way" (Farmer R 1996).

While one might comment that 'consistency' as describe in the above quote is equivalent to 'replication', replications in 'different' populations might fail simply because the causative or associated genetic variants are rarely expected to be found uniformly distributed in different populations. The relevance of the quote to this study is that it is logical to assume that the greater the consistency or persistence of a given SNP-phenotype association test statistic (significant or not) across different test-models and statistical test procedures in the same study population, the more unlikely it is to be due to some constant error that somehow permeates every statistical

inquiry undertaken, i.e., the result is robust. On the other hand, the consistency (or specificity) of a given statistic across different test-models or datasets, may signify that the distinct clinical outcomes are regulated by similar (or different) genetic mechanisms. To summarize, analyzing patterns of association serves as a means to unravel distinct genetic risk profiles underlying different TB pathogenesis pathways: Is there any evidence for TB-progression stage-specific associations? While, in most cases, the number and identity of significantly associated SNPs differed between different test-models, four patterns of significant association can be discerned. Results are presented in Table 10.

Restricted association (A single SNP/gene can be associated with one, and only one, test-models): Test-model-restricted SNP-phenotype associations were observed regardless of the direction of association, i.e., susceptibility or resistance. This means that some SNPs in all the three candidate genes were only associated with only one test-model but not in others. However, it must be noted that this pattern is not observed at the gene level but only applies to particular variants/SNPs within the gene. This pattern could be an indication that the constructed test-models and the corresponding case-control phenotype definitions, or the pathologic pathways involved therein, may be distinct and controlled by different genetic mechanisms affected specifically by the associated SNPs. This is supported by the observation that none of the SNPs in this category were found to be associated in any other test-model. They were exclusively associated with one and only one dataset indicating a phenotype-specific action. For example the association of some SNPs from all three candidate genes was restricted to the 'Active TB vs. No Active TB' test-model; only one SNP in NOD1 was negatively associated in 'Active TB vs. No

LTBI' dataset; and a couple of SNPs in both FMO2 and NOD1 were specifically associated with TB-phenotypes in the other two test-models.

Consistent association (A single SNP/gene can be associated with multiple test-models): Multiple SNPs were shown to be significantly associated across test-models. In fact, all the top-four highly significant SNPs in this study (SNPs chr1:171165749 and chr1:171181877 in FMO2 and SNPs chr7:30477156 and chr7:30485722 in NOD1) were found to be associated with a TB phenotype in three test-models: Active TB vs. No Active TB, Active TB vs. No LTBI, and Active TB vs. LTBI. The single most consistently significant SNP of this study, chr1:171181877 in FMO2, was found to be significantly associated in all four test-models covering the spectrum of TB disease association, including being associated with the test-model LTBI vs. No LTBI. Full or partial persistence of an association result for a particular SNP across datasets may suggest either the case-control phenotype definitions in different datasets are broadly/loosely defined or, on the other hand, it may indicate that the phenotype definitions are perfectly distinct and valid but the consistently associated SNP may be involved in all TB progression pathways. This might be a case of pleiotropy at the SNP level where one locus influences two or more seemingly unrelated phenotypes. Synergistic pleiotropy both at the gene and SNP level is not uncommon in genetics and could parallel a gradient following TB disease pathology and progression from infection, LTBI, to active TB.

Correlated association (Multiple SNPs/genes can be concordantly associated with the same test-models): Some SNPs exhibited a correlation of significant association in the sense that two or more genes/SNPs seemed to be associated with the same phenotype concurrently in different

test-models. This may indicate that the phenotypes may be influenced by the correlated genes/SNPs acting in concert in a network of biological pathways that lead to the specific associated phenotype, a phenomemnon known as epistasis. On the other hand, the LD structure between such genes and variants may explain correlated association and will be discussed later in detail. In fact, moderate to strong LD (r2=0.2-1) was observed between some of the phenotype-associated SNPs. However, none of the top-four SNPs were even in moderate LD either between themselves or others.

Directional Association (Different SNPs in the same gene can have different TB-risk effect (Susceptibility/Resistance): Another highly apattern association pattern is the presence of both SNPs within a single gene that were associated in different directions, i.e., >1 ORs (+ve reg. coef., beta) and <1 ORs (-ve reg. coef). For example, five (different) SNPs in FMO2 seem to be associated with susceptibility (>1 OR) with active TB and seven other (different) SNPs of FMO2 were associated with resistance (<1 OR) to active TB within the same test-model.

Table 10: Pattern of SNP-phenotype associations

Pattern of SNP-phenotype associations across test models: Directional, Consistent, Correlated, Restricted

SNPs associated with susceptibility	FMO2									NOD1						TICAM2	
	chr1:171181877(A)	chr1:171165749(T)	chr1:171179939(G)	chr1:171180021(G)	chr1:171168469(A)	chr1:171174312(A)	chr1:171178490(T)	chr1:171181150(A)	chr1:171168545(C)	chr7:30477156(T)	chr7:30485722(T)	chr7:30490711(T)	chr7:30491081(A)	chr7:30464872(A)	chr7:30469270(C)	chr5:114916090(G)	chr5:114915999(A)
Active TB vs. No Active TB (14 SNPs)	x	x	x	x	x	x	x	x		x	x	x	x			x	x
Active TB vs. No LTBI (5 SNPs)	x	x								x	x	x					
Active TB vs. LTBI (10 SNPs)	x	x	x	x						x	x	x	x	x		x	
LTBI vs. No LTBI (3 SNPs)	x								x						x		

SNPs associated with resistance	FMO2																		NOD1				TICAM2
	chr1:171180201(C)	chr1:171179477(T)	chr1:171174762(C)	chr1:171178090(C)	chr1:171179025(C)	chr1:171179779(G)	chr1:171180071(G)	chr1:171173242(C)	chr1:171174691(A)	chr1:171174821(A)	chr1:171176879(A)	chr1:171177858(T)	chr1:171154303(C)	chr1:171168545(C)	chr1:171179670(G)	chr1:171179939(G)	chr1:171180021(G)	chr1:171179287(T)	chr7:30498962(C)	chr7:30464249(TG)	chr7:30465424(C)	chr7:30464932(G)	chr5:114916028(A)
Active TB vs. No Active TB (16 SNPs)	x	x	x	x	x	x	x	x	x	x	x	x							x	x	x		x
Active TB vs. No LTBI (8 SNPs)	x	x	x	x	x														x			x	x
Active TB vs. LTBI (5 SNPs)	x					x	x						x	x									
LTBI vs. No LTBI (5 SNPs)		x													x	x	x	x					

*=Significant association signal (p<0.05)

To summarize, besides pleiotropy and epistasis, the observed differences in the pattern of association (i.e., restricted association, consistent or persistent association, and correlated association) could be explained in terms of:

1. the extent of mutual exclusivity in case-control phenotype definitions between the various datasets
2. the effect of differences in case-control sample size between the various datasets
3. variation in allelic frequency distribution between the various datasets
4. variation in linkage disequilibrium pattern

Possible explanations for the observed patterns of associations

1. **Extent of mutual exclusivity in case-control phenotype definitions between the various datasets**

Active TB vs. No Active TB test-model dataset: This is the most inclusive and broadly or loosely defined dataset. Cases, labeled 'TB Affected', are symptomatic active TB patients and although the case-phenotype definition does not differentiate between the different TB types almost all of them are pulmonary TB (PTB) patients. [Note: This definition of case-phenotype also applies to Active Tb vs. No LTBI and Active TB vs. LTBI datasets.] Controls are 'TB Unaffected', unrelated household contacts of TB patients who were recruited mainly because they were considered to have been exposed to *Mtb* by virtue of the fact that they were care givers to TB patients and/or lived in close proximity with patients in a more or less overall TB endemic community. The control-phenotype definition in this test-model does not differentiate between individuals who have been exposed to *Mtb* and may harbour latent TB, individuals who have been exposed to *Mtb* and got infected but managed to eliminate the infection, and individuals who have been exposed but not infected. And, some studies have shown that a high proportion of tested individuals in Ethiopia test positive for TST which is an indication for the presence of a high rate of latent TB infection and/or transmission (Tegbaru, et al, 2006). In this study, tests for latent TB via the IGRA described before found a 40-57% latent TB infection.

Active TB vs. NO LTBI test-model dataset: Controls are individuals who tested negative for IGRA. These individuals are considered to have no latent TB although they were most likely to have been exposed to *Mtb*. Therefore, this control-phenotype definition excludes latent TB infection (LTBI).This control phenotype has been described in other studies as 'resistance to infection'.

Active TB vs. LTBI test-model dataset: Controls are individuals who tested positive for IGRA and are considered to have latent TB infection. Therefore, this control-phenotype definition excludes individuals who were exposed to *Mtb* but managed to eliminate the infection.

LTBI vs. No LTBI test-model dataset: This dataset is based entirely on IGRA results. Cases are individuals who do not manifest symptoms of active TB but tested positive for IGRA and are considered to have latent TB infection. Controls are individuals who were most likely exposed to *Mtb* but eliminated the infection and tested negative for IGRA. In other words, this dataset excludes active TB from cases and latent TB from controls.

2. **Effect of differences in case-control sample size and number of test-SNPs between the various datasets**

Sample size affects the power of statistical tests to detect significant associations. In this study the progressively stricter definition of case-control phenotypes comes at the cost of reduced sample size. This difference in sample size could explain in part the difference in the observed pattern of association. Furthermore, not all SNPs passed QC criteria (based on genotyping rate, HWE test, and MAF) in all four datasets. For example, differences in minor allele frequency of the test-SNPs, may explain the differences in the observed pattern of significant SNP-phenotype associations between the four datasets. A certain SNP might get enriched or reduced (in some extreme cases may also result in the removal altogether of below threshold markers) in terms of its frequency in one dataset as compared to the other datasets. However, there was not much difference in the total number of QC-passed test-SNPs in the four datasets which was 94, 91, 92, and 98.

Tests for heterogeneous associations: Do the observed associations vary between EGCs?

In the sections presented above, it was tried to assess and explain the possible causes of for the observed patterns of association test results between datasets. There may also be differences in association effect sizes between ethnic groups and the sexes. Therefore, the Breslow-Day test was used to test for between-population effect differences in association based on differences of odds ratio for association; a significant P-BD value indicates between-cluster heterogeneity in the odds ratios for the SNP-phenotype association (Table-11).

Table 11: Analysis for heterogeneous association between EGCs

Heterogenous association between ethno-geographic categories														
Test-model	Gene	SNP	BP	A1	MAF	A2	CHISQ	P	OR	SE	L95	U95	CHISQ_BD	P_BD
Active TB vs. No Active TB	NOD1	chr7:30490711	30490711	T	0.06507	G	9.484	0.002072	3.457	0.4108	1.545	7.733	6.674	0.03555
Active TB vs. LTBI	FMO2	chr1:171165749	171165749	T	0.08296	G	9.866	0.001684	5.891	0.6258	1.728	20.09	6.85E+00	0.03248
		chr1:171181877	171181877	A	0.1368	C	10.62	0.001117	3.448	0.3984	1.579	7.527	9.939	0.006948
	NOD1	chr7:30477156	30477156	T	0.09641	G	8.591	0.003378	3.901	0.4932	1.484	10.26	7.344	0.02542
		chr7:30485722	30485722	T	0.09641	G	6.564	0.0104	3.111	0.4596	1.264	7.659	7.999	0.01832

The heterogeneity test result suggested that, for some markers including the top-SNPs in this study, the effect was not equally present in all populations while there was no effect difference for the majority of the phenotype-associated SNPs. This result supported the previous results of regression analysis taking population clusters as covariates which may indicate population-specific effects. Differences both in sample size and allele frequencies can result in between-population effect differences.

Tests of allele frequency difference between EGCs

The results of the above covariate analysis and heterogeneity test raise the more basic question of whether allele frequencies differ between the EGCs. There were varying degrees of nominal differences in the minor allele frequency of some SNPs, including two of the top-SNPs, between the three EGCs (Table-12). Allele frequencies of all the phenotype-associated markers are provided in Supplementary Table-6.

Table 12: Analysis for allele frequency differences between EGCs

Phenotype-associated SNP allele frequency differences between EGC										
Gene	SNP	F_EGC1	F_EGC2	P	OR	SE	L95	U95	Compared EGC	BONF
FMO2	chr1:171165749	0.08	0.03297	0.03826	2.551	0.4665	1.022	6.364	Merhabete-Arbaminch	1
	chr1:171168469	0.04362	0.004902	0.009654	9.26	1.042	1.202	71.35	Adigrat-Arbaminch	1
		0.04362	0.01099	0.04618	4.105	0.7655	0.9157	18.4	Merhabete-Arbaminch	1
	chr1:171174762	0.02333	0.1127	3.12E-05	0.188	0.4419	0.07907	0.447	Adigrat-Arbaminch	0.03671
		0.02333	0.07527	0.006202	0.2935	0.4728	0.1162	0.7414	Merhabete-Arbaminch	1
	chr1:171177858	0.0604	0.1324	0.005577	0.4214	0.3191	0.2255	0.7876	Adigrat-Arbaminch	1
		0.1337	0.06989	0.03904	2.053	0.3542	1.026	4.111	Merhabete-Adigrat	1
	chr1:171179779	0.3733	0.2789	0.03123	1.54	0.201	1.038	2.284	Merhabete-Arbaminch	1
	chr1:171179939	0.45	0.5484	0.03493	0.6738	0.1876	0.4665	0.9731	Merhabete-Arbaminch	1
	chr1:171180021	0.4533	0.5484	0.0416	0.6829	0.1875	0.4729	0.9862	Merhabete-Arbaminch	1
	chr1:171180071	0.37	0.2796	0.04013	1.513	0.2025	1.018	2.251	Merhabete-Arbaminch	1
	chr1:171181150	0.04333	0.004902	0.009966	9.195	1.042	1.193	70.85	Adigrat-Arbaminch	1
		0.04333	0.005263	0.01372	8.561	1.042	1.111	65.99	Merhabete-Arbaminch	1
NOD1	chr7:30469270	0.2667	0.1374	0.000856	2.284	0.2518	1.394	3.741	Merhabete-Arbaminch	1
	chr7:30485722	0.09406	0.03846	0.03037	2.596	0.4546	1.065	6.327	Merhabete-Adigrat	1
		0.09732	0.03846	0.01754	2.695	0.4322	1.155	6.287	Merhabete-Arbaminch	1
	chr7:30491081	0.04392	0.005556	0.01629	8.223	1.042	1.066	63.4	Merhabete-Arbaminch	1

The highest number of nominal differences in phenotype-associated SNP allelic frequency was observed between Arbaminch and Merhabete followed by between Arbaminch and Adigrat, the least difference being between Merhabete and Adigrat. The relative differences may reflect differences in the demographic history, and migration that influence the population genetic backgrounds of the compared ethnic groups. For example, the minimal difference between Merhabete and Adigrat could be due to their common Semitic origin while the Arbaminch

samples are Omotic. However, such conclusions require the analysis of millions of SNPs and will be discussed later. Another source of such inter-ethnic differences in allele frequency could simply be a consequence of differential genotyping error. However, since samples were purposefully randomized on sequencing plates both in terms of case-control status and ethnicity, the latter explanation is not valid, and was supported by the genotype missingness test described in the QC section.

These apparent inter-ethnic difference in allele frequencies can be a case for concern since a signal of association may arise for an ancestrally informative SNP, not because of an association with disease risk, but because of allele frequency differences between the populations that differentially comprise the cases and controls. Therefore, it may be necessary to perform analysis within-EGCs (as will be explained later) or to include population as a covariate (as was done earlier). Moreover, efforts were made to remove or reduce the effect of population stratification through empirical analysis divergent ancestry rather than based on self-declared ethnicity.

Empirical assessment of population stratification

There have been recent population genetic studies that suggested the Ethiopian population not only harbours one of the highest genetic variation in the world (Gurdasani, et al; Pagani, et al, 2015) but also show evidence of stratification along ethnic/linguistic lines (Pagani, et al, 2012) which might be a challenge for genetic association analysis in Ethiopia. Some of the issues raised against including individuals from different ethno-geographic backgrounds in association studies are: it may not be possible to adequately control for population admixture or stratification;

assuming the same genes are involved in disease etiology within different populations, the same functional variants may either not exist or have different allele frequencies within the different populations and, therefore, the same tagSNPs may not detect associations within different populations or may not be tagged equally well in all populations; and, statistically, these situations can reduce the power of the study

However, in the absence of extensive DNA sequence database for the study populations, the assumptions cannot be taken for granted and have to be verified by studies like the present study. In this study, the following within and between population tests were performed in order to assess the extent of such variation and control for any possible stratification effect on the outcome of the association test statistic: test of allele frequency difference (Chi-square test), described above; population-specific test of association, i.e., within each EGC; stratified analysis of association based on self-declared ethnicity (EGC); stratified analysis based on assignment of individuals to homogeneous clusters based on empiric SNP data ,i.e., testing of identity-by-state (IBS) allelic similarity between all possible pairs of individuals (ancestry investigation); and, LD pattern and haplotype structure based analyses.

Population-specific tests of association

Obviously, the first option is, in as much as the actual ancestry of each individual can be inferred based on self-declared ethnicity, this information can be used to test for associations within specified EGCs. This is a simple statistical option to test for association between TB phenotypes and SNPs that does not depend on adjusting or controlling for possible population stratification

but, instead, directly tests for association within specific populations. Although the simplicity of the population-specific test is offset by small samples size, it resolves the issue of possible confounding or effect modification due to population heterogeneity by ensuring the analysis is done within supposedly homogenous populations based on sampling site.

The associations were mostly replicated in the Arbaminch population which had the largest sample size, one SNP in Adigrat and none in Merhabete. The results, while supporting the evidence for the association of the SNPs with TB susceptibility, they also showed that first, there is a loss of power to detect statistically significant associations due to a reduction in sample size, and concomitant allele frequency reduction, while conducting strata-specific data analysis. Second, examination of the observed changes in the Odds Ratios (differences in OR values between the ethno-geography-specific test results and the combined results) also revealed the possibility of confounding and/or effect modification incurred by the epidemiologic factor 'ethno-geography'. The results are presented in each of the association test result tables under the columns Fisher, Pearson and logistic regression.

Pair-wise IBS clustering and multi-dimensional scaling analysis

This was done by performing multidimensional scaling (MDS) analysis to investigate population structure, identify stratification, clustering individuals into homogeneous groups based on genetic data. PLINK employs complete linkage agglomerative clustering, based on pair-wise SNP identity-by-state (IBS) distance with some modifications to the clustering process (Purcell, et al, 2007):

I. **Based on pair-wise population concordance (PPC) test:** a significance test for whether two individuals belong to the same random-mating population and used to only merge clusters that do not contain individuals differing at a certain p-value (i.e. do not merge clusters that contain significantly different individuals, p=0.01). This method was used to analyze sampled populations from Merhabete, Adigrat and Arbaminch for signals of both within and between population stratification. The entire exonic SNP data of each population was used for the analysis.

II. **Based on phenotype/test-model:** in addition to the PPC test, a constraint was applied that ensures every cluster formed by the PPC procedure has at least one case and one control so that association test can be performed within each homogenous cluster. Only SNPs from each of the four phenotype-based datasets were included in the analysis.

R software was used to help visualize clusters on the N x N matrix of SNP IBS pair-wise distances. For example, plotting cluster 1 (C1) values against cluster (C2) will give a scatter plot in which each point is an individual; the two axes correspond to a reduced representation of the data in two dimensions, which can be useful for identifying any clustering. Standard classical (metric) multidimensional scaling is used.

Notes on cluster colouring: colouring based on default cluster solutions generated by empiric SNP data (genetic data); colouring in order to help identify population (EGC) membership (green=Merhabete; blue=Adigrat; red=Arbaminch). Each spot represents a single individual.

Only the first couple of MDS analysis components are presented in the MDS plots (Figures 10-15) since it was found that clustering was sufficiently formed even at the initial levels of component resolution stress, i.e., further displaying of components to show details of the analysis under progressively stringent structure resolution was unnecessary. Figure-10, Figure-11, Figure-12 and Figure-13 are for the Combined Population, Merhabete, Adigrat and Arbaminch, respectively. Clusters shown within Figures 14-15 were based on the 'Active TB vs. No Active TB' test-model, the largest dataset in the study.

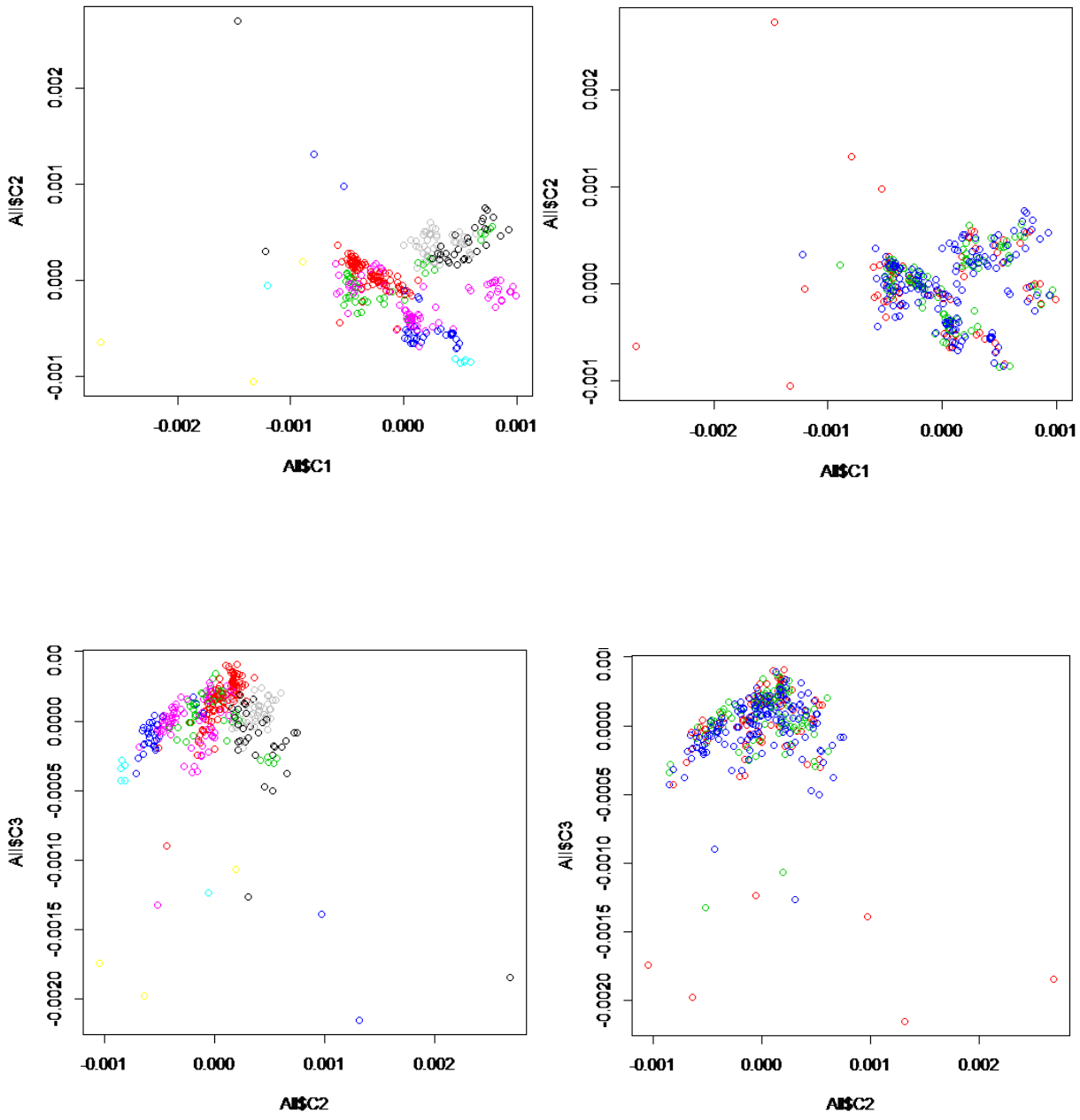


Figure 10: Plots of the first 2 components of multidimensional scaling analysis for the combined population

MDS plot of the combined population: colouring based on SNP data (left) and EGC membership (right)

Although the analysis of pair-wise clustering based on IBS typically requires whole-genome level data to give accurate results (Purcell, et al, 2007), the analyses for population stratification as visualized by the MDS-plots under various levels of MDS stress managed to generate clearly defined and identifiable clusters (clines/population splits) as well as a few outliers. The plot of the first two MDS components above, generated by more than 19,000 marker/nucleotide data of 347 individuals, distinguish at least four clusters which is an indication of the presence of population stratification. However, the corresponding MDS plots in which individuals were coloured based on their self-declared ethnicity (i.e., colour codes based on EGC-membership are superimposed on the clusters formed by empiric genetic data) shows that the stratification is not along the lines of ethnic background. In other words, within each cluster, there are found members of all EGCs. This is in sharp contrast to previous findings that suggested Ethiopian population stratification follows ethnic/linguistic lines (Pagani, et al, 2012). However, this difference in finding may be due to differences in the populations studied and number of markers analyzed. There are also differences in inference since the present population stratification analysis is based on exonic region nucleotide sequences rather than whole genome data. The phenotype/test-model-based MDS plots also identified clusters although the stratification does not follow EGC membership.

There is also evidence of within-population sub-structuring based on the same IBS-based MDS cluster analysis procedure applied to members of the same EGC. The level of stratification varies between populations: the Merhabete sample showed at least two clusters, Adigrat at least three, and Arbaminch at least four clusters. This gradient of population sub-structure could be interpreted as showing the level of intra-ethnic heterogeneity or population admixture with the

Merhabete sample being the most homogenous and the Arbaminch population possessing the highest level of heterogeneity or admixture although this can be affected by sample size differences. This heterogeneity pattern may not be a surprise as, for example, the sample from Arbaminch comes from a region known for its relatively high ethnic diversity in Ethiopia, the region officially named as the 'Southern Peoples, Nations, and Nationalities'. But, it should be emphasized that, sampling was deliberately carried out with every effort made to recruit both cases and controls from within a single ethnic group based on self-declared ethnicity: the Amhara ethnic group, in Merhabete, the Tigray in Adigrat, and the Gamo in Arbaminch.

On the other hand, the fact that the observed structuring in the combined population does not follow ethno-geographic lines and similar evidence of sub-structuring exists within in each individual EGC, may indicate the almost universal pattern that 'intra-ethnic' variation may be higher than 'inter-ethnic' variation. The later explanation has been found to be viable in several studies of population genetics. Moreover, the HWE test for random mating in a homogenous population was non-significant (at $p<0.001$) except for very few markers out of 19,530: in the combined sample, only 12 markers were significant in controls (13 in cases); Merhabete, 4 markers in controls (3 in cases); Adigrat, 3 markers in controls (2 in cases); and in the Arbaminch sample, 6 in controls (9 in cases).

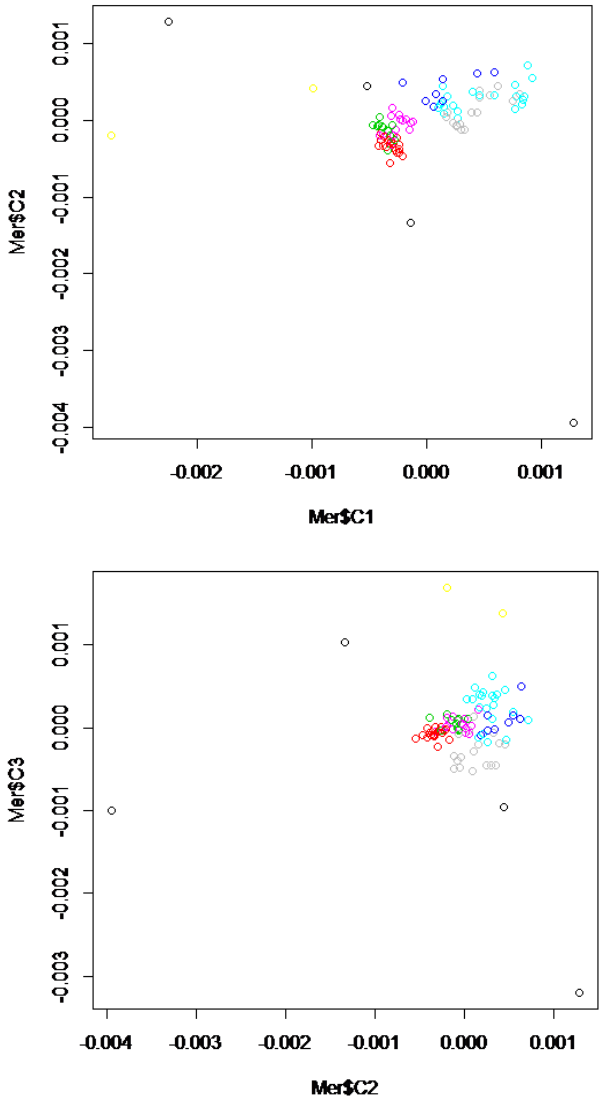


Figure 11: Plots of the first 2 components of multidimensional scaling analysis for Merhabete population

MDS plot of Merhabete population: colouring based on genetic data

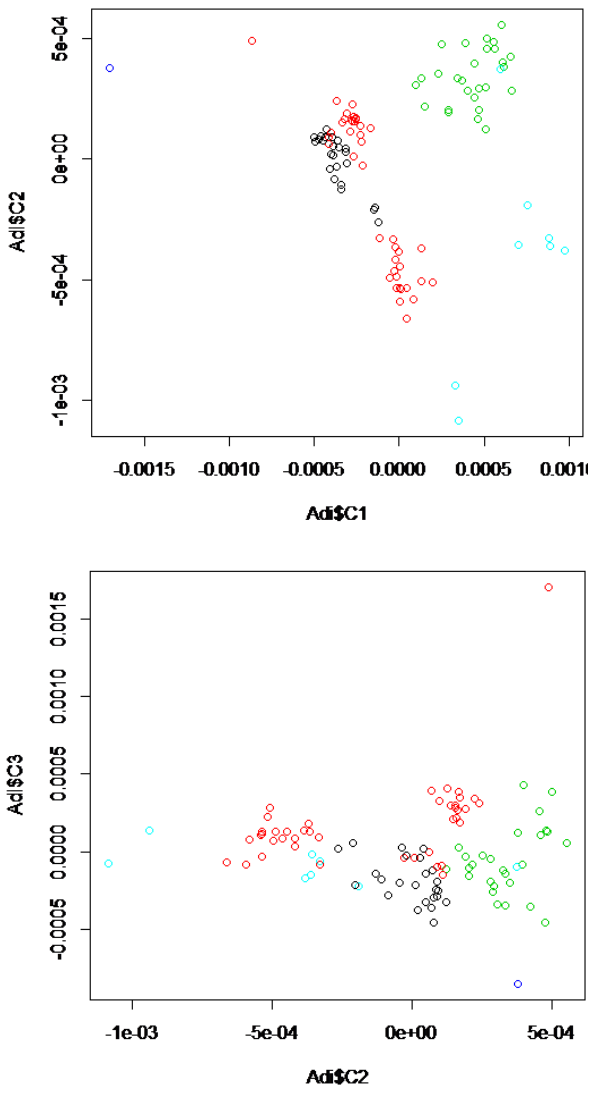


Figure 12: Plots of the first 2 components of multidimensional scaling analysis for Adigrat population

MDS plot of Adigrat population: colouring based on genetic data

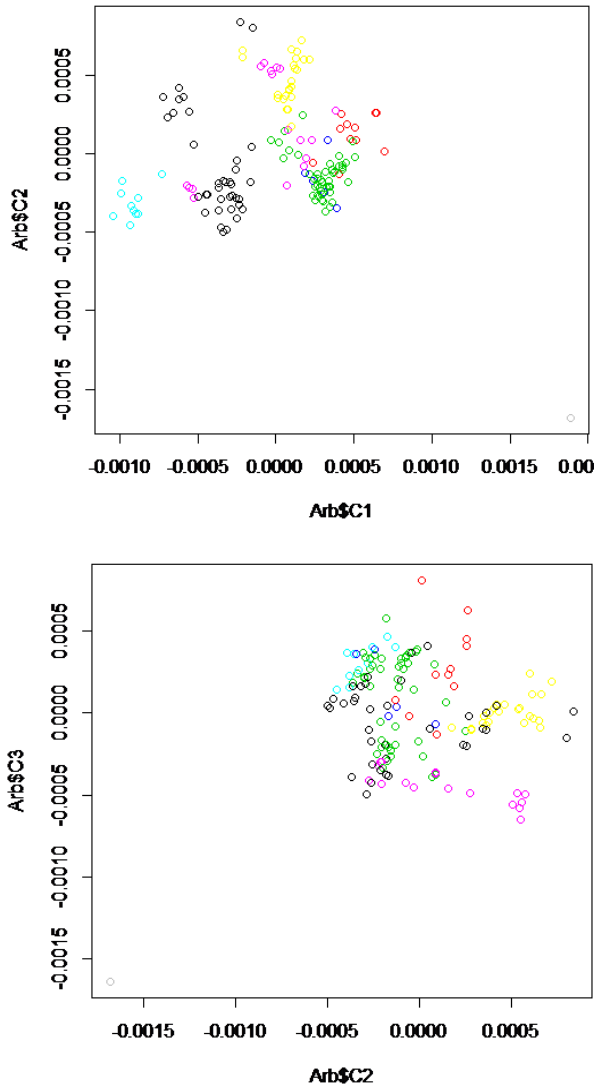


Figure 13: Plots of the first components of multidimensional scaling analysis for Arbaminch population

MDS plot of Arbaminch population: colouring based on genetic data

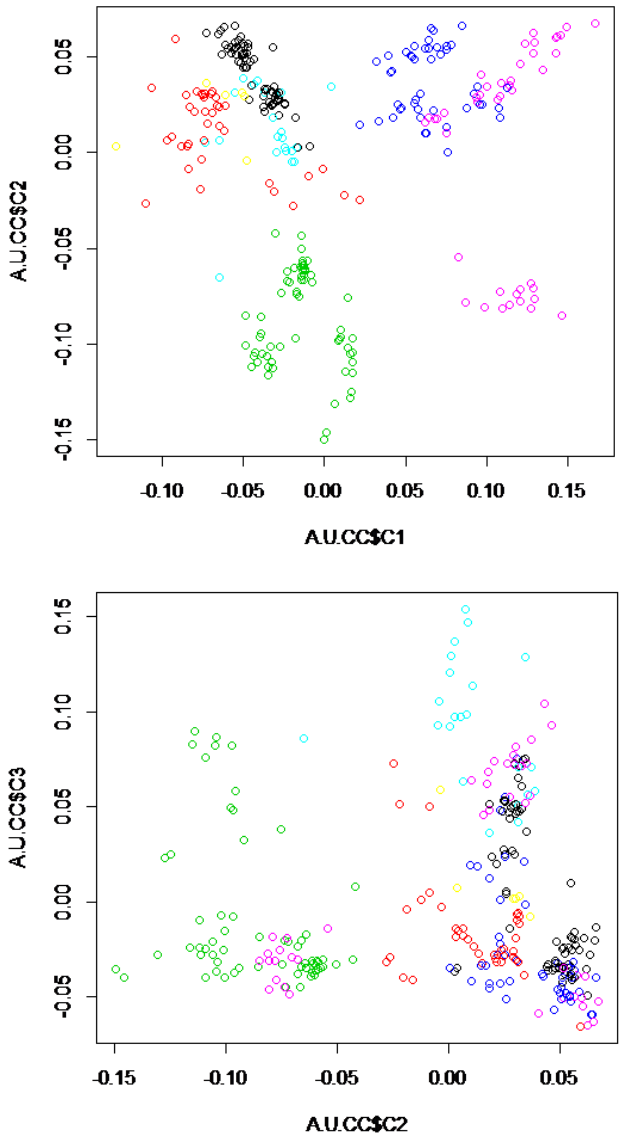


Figure 14: Plots of the first components of multidimensional scaling analysis for the 'Active TB vs. No Active TB' (Test-model 1)

MDS plot of Active TB Vs. No Active TB: colouring based on genetic data

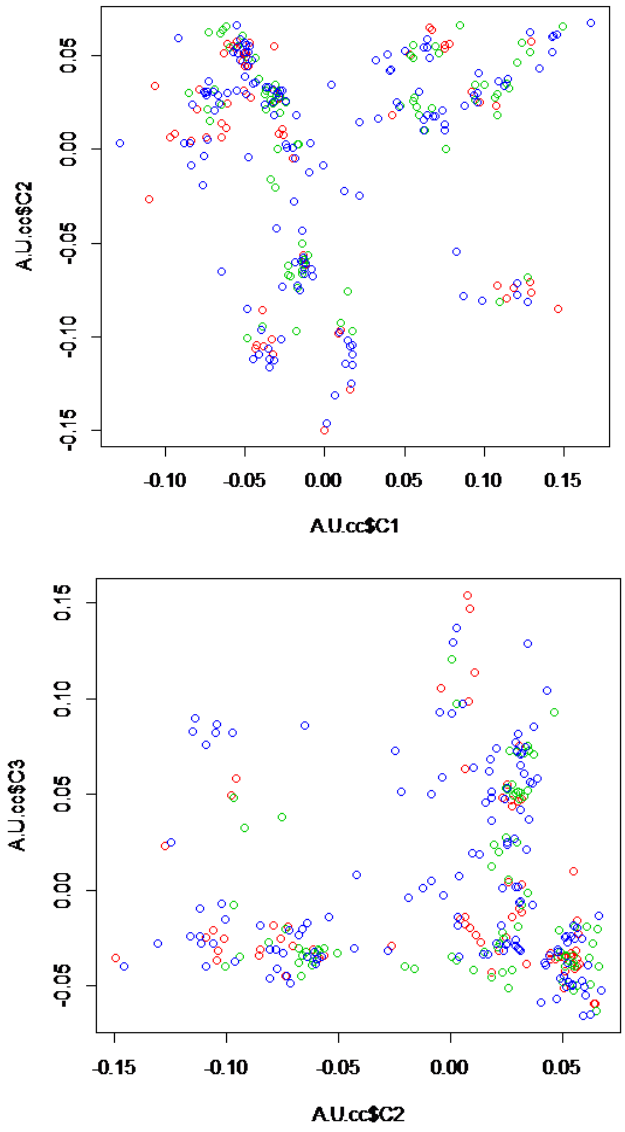


Figure 15: Plots of the first 2 components of multidimensional scaling analysis for the 'Active TB vs. No Active TB' (Test-model 2)

MDS plot of Active TB vs. No Active TB: colouring based on EGC membership

The main implication of the results of the MDS analysis of population structure described above is that statistical tests of SNP-phenotype association based on self-declared-ethnicity stratified analysis alone may not adequately remedy or account for heterogeneity. In other words, the level of within EGC heterogeneity may cast a reasonable doubt on the validity of the assumption of homogeneity upon which the basic statistical association analyses were based; the worst scenario being that the associated SNPs may just be ancestry-informative SNPs. Therefore, IBS-based stratified analysis (which generates empiric, genetic-data-based, relatively homogeneous clusters) was carried out along with the other stratified tests. And, as mentioned earlier, the stratification observed in the MDS plots, does not follow ethnic lines and, within each cluster, individuals from all three EGC populations are represented in approximate proportion to their respective sample size.

EGC and IBS based stratified tests of association

In light of the results of the various analyses for population stratification presented above (allele frequency comparison, IBS-based MDS analysis) and EGC-based covariate analysis, one concern might be that subtle population stratification might be biasing the association statistics. In such cases, the Cochran-Mantel-Haenszel (CMH) test can be employed: tests based on the self-declared ethnicities taken as *de facto* clusters and IBS clustering generated by genetic data. The CMH approach provides stratified tests of SNP-disease association based on an "average" odds ratio that controls for the potential confounding due to the cluster variable, i.e., test of whether the effect of a particular SNP varies between cases and controls whilst controlling for any possible stratification and provide a single statistical estimate of SNP-phenotype association

such as odds ratios and their p-values. P-values generated by the basic allelic tests of association (Fisher, Pearson, logistic regression) were compared with the population-specific and stratified tests of association. Overall, the values obtained after the EGC and empiric adjustment for potential population admixture using CMH test, the most significant SNPs were replicated with minimal p-value changes while the moderately significant SNPs became less significant and others more significant, i.e., the stratified tests identified some SNPs not picked by the other tests. For example, from the susceptibility category, the four top-SNPs of this study, two FMO2 SNPs (chr1:171165749 and chr1:171181877) and two NOD1 SNPs (chr7:30477156 and chr7:30485722) were still significant in the IBS-stratified tests of association. EGC-based stratified test of association not only identified all of the top-SNPs picked by the basic tests but also found other unidentified SNPs. For example, from the susceptibility category, eight SNPs were phenotype-associated (four each in FMO2 and NOD1) of which three were not identified by the previous basic allelic tests: two SNPs in FMO2 (chr1:171174312 and chr1:171178490) and 1 in NOD1 (chr7:30464872).

With regard to the important issue of how efficient were the tests at addressing the cryptic stratification observed, it is informative to note a statistic known as genomic inflation factor (GIF) which is the ratio of the median of the Chi-square statistic to the expected median value. It is assumed that if there were no population stratification, GIF would be 1. Therefore, if the GIFs are reduced after adjusting for stratification in the association tests, it is consistent with the idea that there was some substructure inflating the distribution of test statistics in the previous basic analyses such as the Pearson's Chi-squared test (Table-13).

Table 13: Comparison of genomic inflation factor

Comparison of genomic inflation factor					
Population	Test	Active TB vs. No Active TB	Active TB vs. No LTBI	Active TB vs. LTBI	LTBI vs. No LTBI
Combined	Person's Chi-squared	2.49857	2.16664	2.07412	1
	CMH: EGC-based	2.16437	1.41631	2.15561	1
	CMH: IBS-based	1	1.14626	1.09557	1
Merhabete	Person's Chi-squared	1	1	1	1
	CMH: IBS-based	1	1	1.0476	1.31579
Adigrat	Person's Chi-squared	1.81587	1.60836	1.20975	1
	CMH: IBS-based	1	1.01817	1	1
Arbaminch	Person's Chi-squared	2.35536	1.05442	2.45004	1
	CMH: IBS-based	1.18744	1.45295	1.14211	1.63568

Examining the changes in GIF showed that, association analysis based on pair-wise IBS clustering did a better job at reducing the GIF value to close to 1 even when compared with the GIF value calculated when self-declared ethnicity was used as a stratification variable. However, the Arbaminch population still showed a slight inflation as compared to the others. Therefore, it can be concluded that there was indeed some cryptic population structuring in the combined sample dataset as well as sub-structuring within each EGC sample dataset and, therefore, the statistics generated based on IBS stratified analysis are more reliable. Furthermore, it indicates that although there is cryptic population stratification, it has minimal distortive effect on the statistical findings of this study, at least with regard to the highly associated SNPs and it can be addressed by employing appropriate adjustments. It is also apparent that the extent of sub-population (within EGC) cryptic stratification is different between the three ethnic groups: the Merhabete sample needed much less adjustment for stratification (GIF is already 1 before adjustment and, therefore, may not even require adjustment for stratification) while the Arbaminch sample GIF was much reduced by IBS-based adjustment.

To summarize this section on stratified tests of association, the advantage of IBS-based association test is that it does not rely on trusting self-declared ethnicity, does not depend on visual inspection of MDS-plots, and the only option when there is no available information about ethnicity. Tests for SNP-disease association conditional on the clustering generated by IBS analysis, was employed both in the combined population and within each EGC. IBS-based clustering was performed in such a way that each cluster contains at least 1 case and 1 control, i.e., so that it is informative for association with a moderate threshold of 0.01 (do not merge individuals differing at $p<0.01$).

Analysis of LD patterns and haplotype structure

LD and haplotype structure analyses helps to better understand the association patterns observed in this study. It also helps to evaluate whether the effect of differing genetic background among populations with different ethnicities is reflected in distinct patterns of LD block and haplotype structure. LD is simply a co-occurrence/inheritance of alleles at different loci in a non-independent manner *en bloc*. In the analysis of the pair-wise LD pattern in the overall genetic dataset in this study (19,530 markers), the first thing that was apparent is that the observed LD pattern verses distance between markers follows the anticipated profile in that, given the amount of genetic data available, there was a slight but observable gradient of LD profile following the corresponding extent of inter-marker physical distance in the exonic regions (Figure 16). With almost contiguous exonic sequence data, it is reasonable to assume that adjacent markers would generally show strong LD, pair-wise correlation decreasing with distance and vice versa. Secondly, the proportion of neighbouring markers showing strong LD ($r2>=0.8$) was low with an

apparent density gap between 0.5 and 1. On the other hand, there were markers found tens of kbs apart with strong LD. In other words, LD strength did not necessarily correspond with shorter inter-marker distance. This LD profile was similar across the three candidate genes studied and across the three ethnic groups. Another observation was that, within the framework of overall similarity in LD profile between the three EGCs, there were slight but discernible population-specific LD patterns. For example, there were stretches of extended LD appearing at specific r2 values in one population but not in the others and the proportion of strong pair-wise LD being highest in the Merhabete sample, Arbaminch the least with Adigrat in the middle.

However, since the LD analysis was based only on exonic region sequences, it was not possible to check whether this LD pattern was characteristic of the entire geneic region. In order to compare LD patterns between exonic and intronic regions, an independent Ethiopian sequence dataset from different populations spanning both the intronic and exonic regions was obtained from (Pagani, et al, 2015) with comparable genotyping rate (GR) (Figure 17). Examination of LD patterns in Figure 17 revealed an LD gradient consistent with an inverse relationship between LD and physical distance between markers, particularly in FMO2 and NOD1 genes. Moreover, the correlation score was shown to be dispersed between $r^2=0$ and $r^2=1$ in a more or less continuous manner following the gradient. This pattern is in contrast to the LD pattern in the exonic regions (Figure 16) which showed distinct gaps in the correlation score almost as if there were directional selections for certain levels of LD strength. This difference in LD pattern between exonic and intronic regions might be a reflection of the functional constraints imposed on the exonic regions which cannot maintain randomized mutation and recombination events without losing their function. Therefore, just as the nucleotide sequences of functional regions

are known to be conserved, it might be the case that specific LD structures and patterns of exonic regions are also maintained across generations and populations. The common feature between exonic and intronic LD patterns is the observation that the proportion of low pairwise LD (r2<= 0.2) is higher.

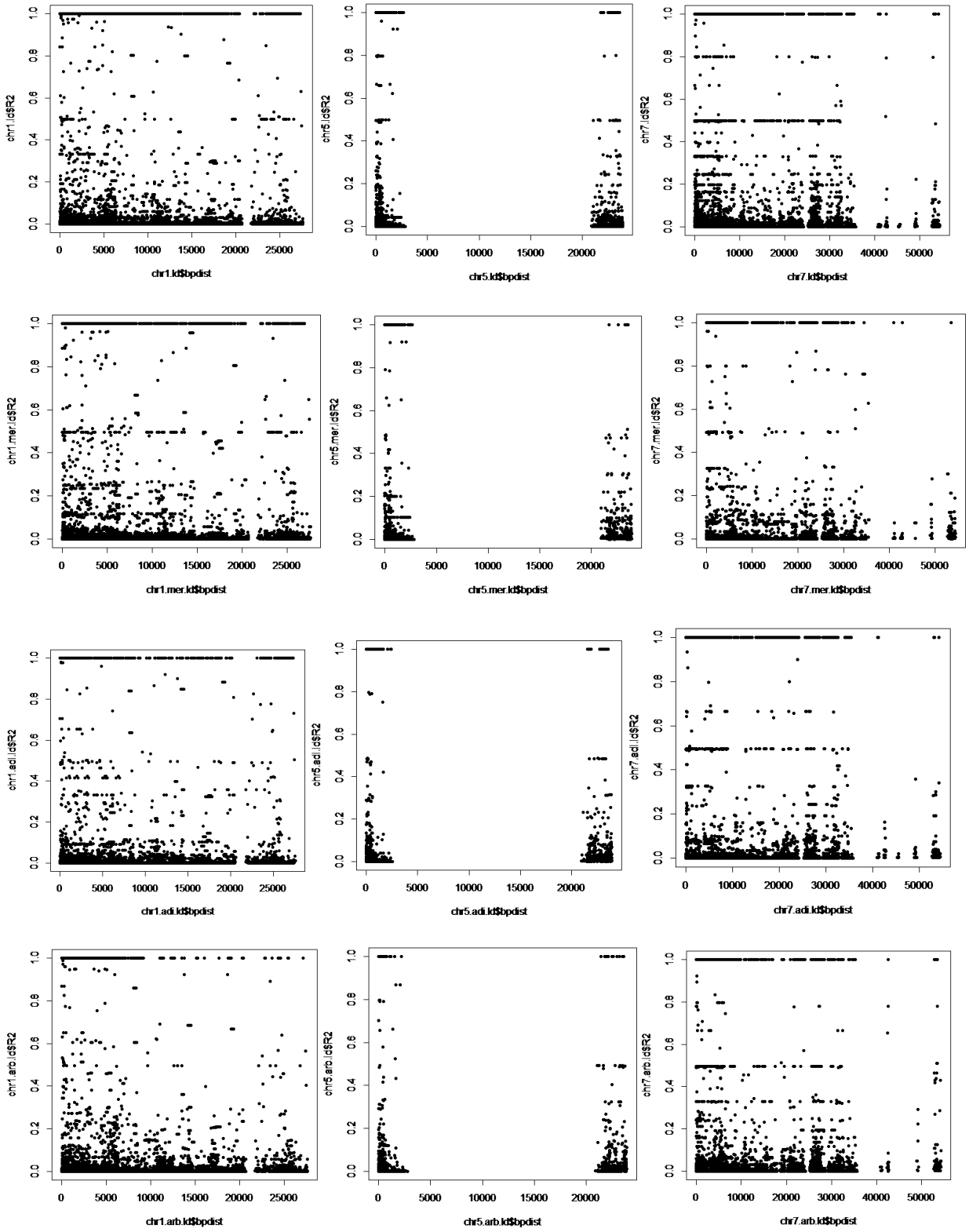


Figure 16: Comparison of linkage disequilibrium pattern in exonic regions between EGCs

(1st row: Combined; 2nd row: Merhabete; 3rd row: Adigrat; 4th row: Arbaminch)

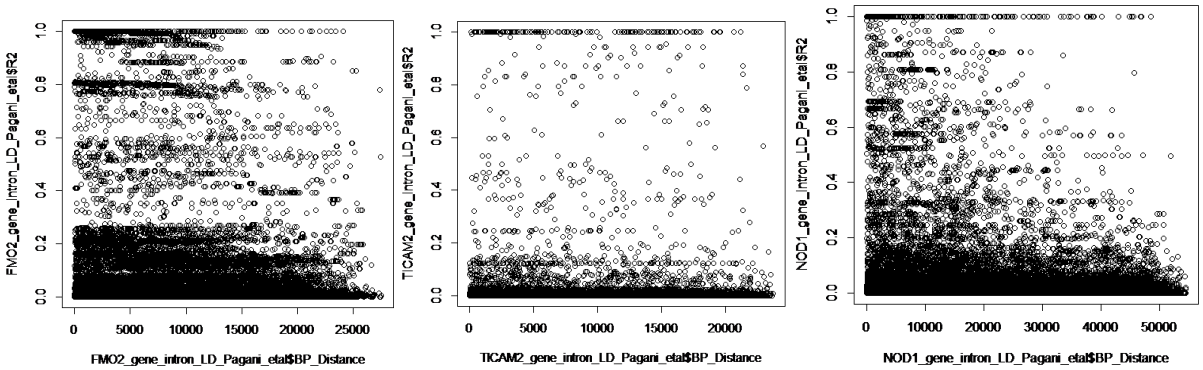


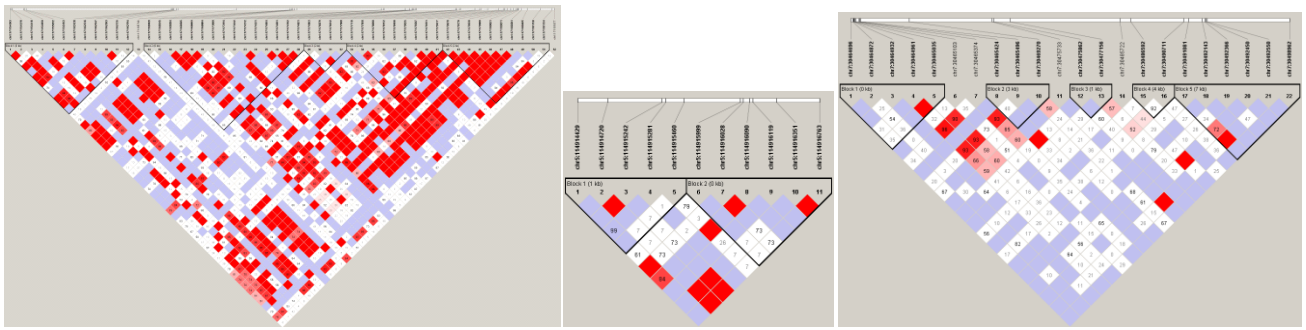
Figure 17: Comparison of linkage disequilibrium pattern in both exonic and intronic regions between candidate genes

Analyses of LD pattern and haplotype structure around disease associated genomic regions, and within the context of differing ethnic backgrounds and phenotype definitions, could facilitate inferring the evolutionary history of variants that increase TB risk. And since each population could have its own demographic and evolutionary history in which distinct allele frequencies, LD patterns, and haplotype structure can develop, population-specific htSNPs may need to be defined in order to identify optimal sets of markers for tagSNP-based association studies.

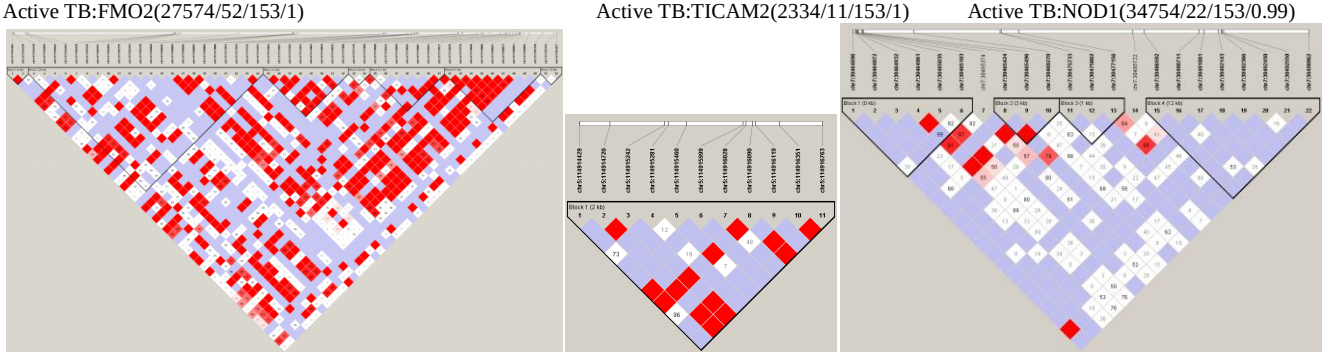
Therefore, in order to have a closer LD profile Haploview software was implemented on all QC-passed (keeping individuals only with greater than 10% genotyping rate, and SNPs with greater than 95% genotyping rate, greater than 1% minor allele frequency, and at HWE of $p< 0.001$). Plots 18-19 depict the overall LD-block profile of the three candidate genes the interpretation of which should be within the context of two factors that may affect LD pattern in opposite directions: first, the DNA sequences are of exonic or functional genomic regions where sequences/alleles are conserved by evolutionary forces that select for functionally important loci, and hence, LD is most likely to be conserved also; second, on the other hand, the samples

represent some of the ancient population history of Ethiopia, and hence, the strength and range of LD might be lessened (LD decay) by recombination events.

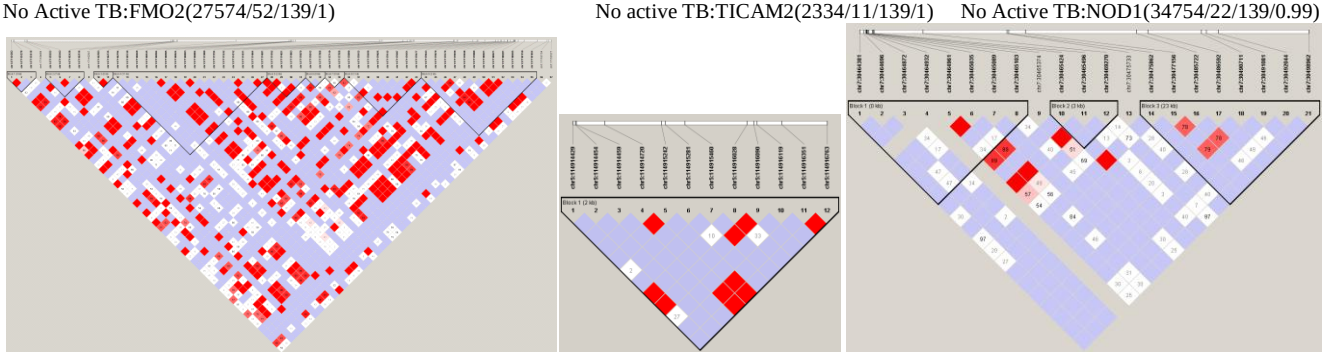
Notes on the LD-block definition and standard colouring scheme of Haploview: Blocks were defined by LD analysis function using the 'Four Gametes Rule' in Haploview whereby for each marker pair, the population frequencies of the 4 possible two-marker haplotypes are computed. If all 4 are observed with at least frequency 0.01, a recombination is deemed to have taken place. Blocks are formed by consecutive markers where only 3 gametes are observed. Colours: white (D' < 1, LOD < 2; recombination); shades of pink/red (D' < 1, LOD $\geq$ 2; moderate LD); blue (D' = 1, LOD < 2); bright red (D' = 1, LOD $\geq$ 2; strong LD). (LOD: a confidence estimate based on the logarithm of the odds of there being LD between two markers). The numbers within the squares represent pair-wise r-square values. Haploview cannot process indels and therefore they are not represented in the LD plots.



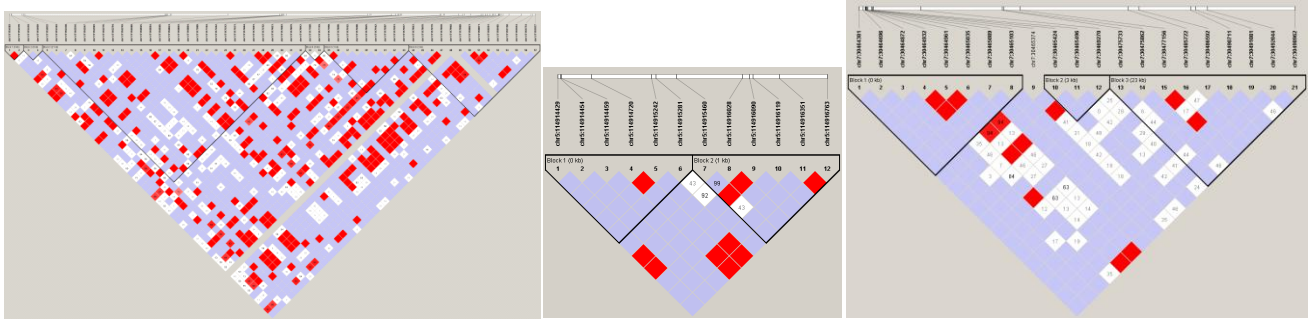
Active TB:FMO2(27574/52/153/1) Active TB:TICAM2(2334/11/153/1) Active TB:NOD1(34754/22/153/0.99)



No Active TB:FMO2(27574/52/139/1) No active TB:TICAM2(2334/11/139/1) No Active TB:NOD1(34754/22/139/0.99)



LTBI:FMO2(27574/57/70/1) LTBI:TICAM2(2334/12/70/1) LTBI:NOD1(34754/21/70/0.99)



No LTBI:FMO2(27574/57/64/1) No LTBI:TICAM2(2334/12/64/1) No LTBI:NOD1(34754/21/64/99)

Figure 18: Comparison of LD blocks between test-models

The descriptions at the bottom represent: Dataset, gene, nucleotide sequence span, #SNPs, #samples, and genotyping rate, respectively.

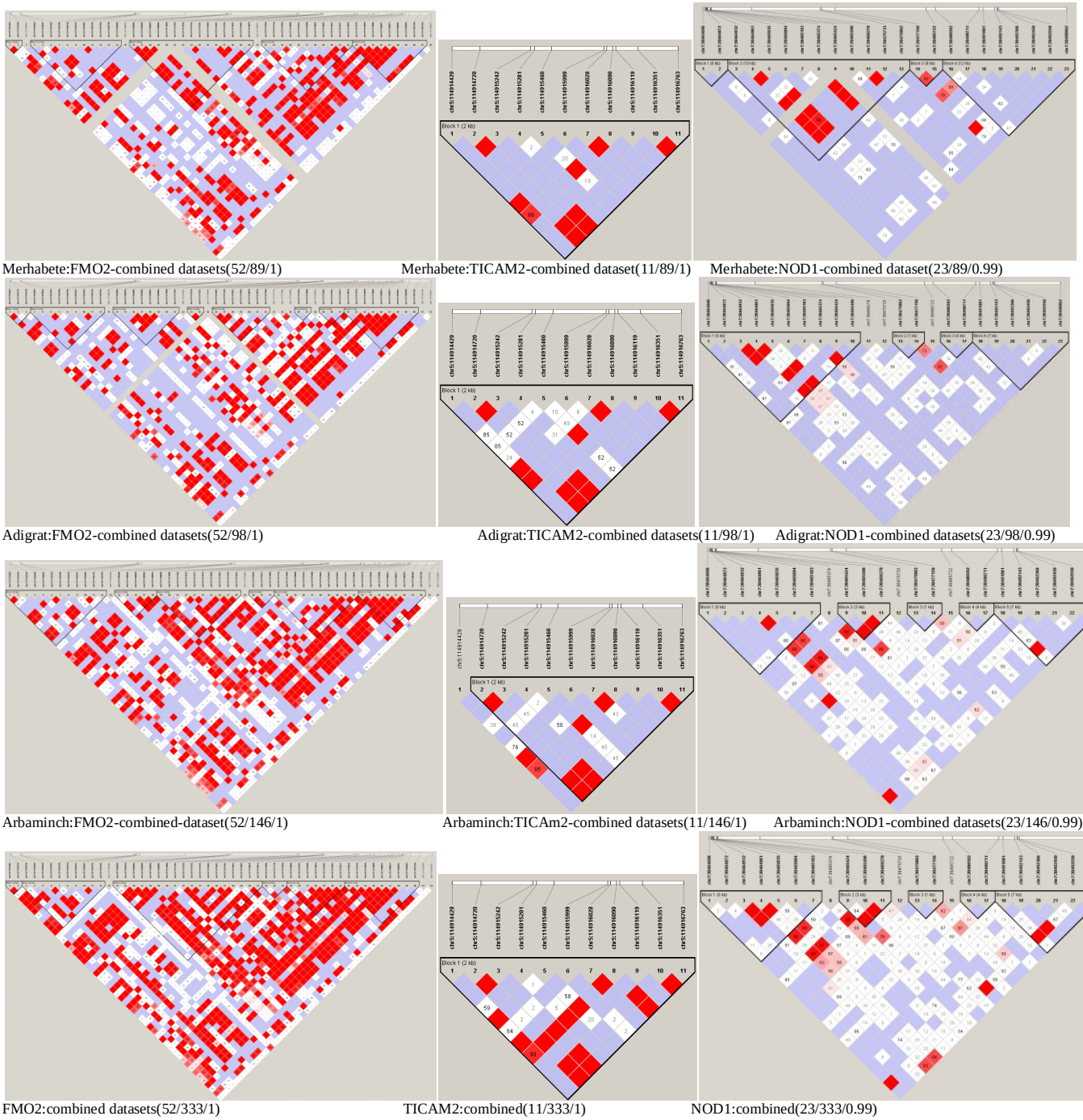


Figure 19: Comparison of LD blocks between EGCs

The descriptions at the bottom represent: Gene, EGC, #SNPs, #samples, and genotyping rate, respectively.

Examination of the LD block pattern also showed a high proportion of low LD with no obvious private, population- and/or TB phenotype-defining patterns. However, the three candidate genes showed inter-genic variation in LD strength as well as showing gene-specific recombination patterns with the FMO2 gene showing a higher proportion of strong LD and LD-blocks although the latter can be explained by the fact that FMO2 covers a much larger kb-span. However, it seems, for a functional (and, hence, conserved) genomic region, and even within the context of the ancient demographic history of Ethiopian populations, the exonic regions of the three candidate genes studied conspicuously lack strong LD between the test-SNPs. It is also noteworthy that almost all of the top-SNPs identified in this study are located in very low LD regions, i.e., regions of very high recombination hot-spots forming LD-block/haplotype boundaries. The implication of this finding is that, first, these SNPs are relatively ancestral and, second, they cannot be tagged easily by other SNPs and have to be genotyped themselves in future studies of the same populations. The inverse implication is that, since these top-phenotype-associated-SNPs are located right on top of recombination hot-spots, the association signal they produced represents a direct phenotype-SNP association rather than an indirect association signal via strong linkage with other SNPs (indirect association) at least within the bounds of this genetic dataset. This means that, if the SNPs turnout to be functional, they are most likely to be causal as well. However, and particularly for those SNPs lying at the outset or fringes of the dataset, there may still be LD with untyped neighbouring intronic SNPs and this requires imputation from other datasets or resequencing of flanking up-/down-stream regions.

Summary of pair-wise LD patterns between phenotype-associated SNPs

In the presence of strong LD between two or more SNPs, it is logical to expect that these SNPs could produce similar signals of association merely due to their physical proximity or correlated inheritance. Therefore, pair-wise genotypic correlation, r^2, was calculated based on genotypic allele counts and without phasing. With a scale of 0 to 1 (r^2=0, perfect equilibrium/independence, and r^2=1 perfect correlation), none of the top-four significantly phenotype-associated SNPs showed even moderately strong LD ($r^2 \geq 0.5$). Therefore, it is difficult to attribute a big role to LD for their pattern of association in the test-models. However, it can be seen that it could have an effect among the other SNPs that showed slightly significant associations (Table-14).

Table 14: Summary of LD between phenotype-associated SNPs

Summary of LD (r2≥2) between phenotype-associated SNPs									
	SNP-B								
SNP-A	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
chr1:171173242	chr1:171174762	chr1:171177858				chr1:171179025	chr1:171178090	chr1:171176879	chr1:171174691 chr1:171174821 chr1:171178490
chr1:171174691	chr1:171174762	chr1:171177858 chr1:171179477				chr1:171179025	chr1:171178090		chr1:171174821 chr1:171176879
chr1:171174762	chr1:171174821	chr1:171178090 chr1:171179025				chr1:171177858			
chr1:171174821		chr1:171177858 chr1:171179477	chr1:171179779			chr1:171179025	chr1:171178090		chr1:171176879
chr1:171176879		chr1:171177858 chr1:171179477 chr1:171179939	chr1:171179779			chr1:171179025	chr1:171178090		
chr1:171177858	chr1:171179477		chr1:171178090 chr1:171179025						
chr1:171178090	chr1:171179939 chr1:171180021		chr1:171179477 chr1:171179779 chr1:171180071						chr1:171179025
chr1:171179025	chr1:171179939 chr1:171180021	chr1:171179779 chr1:171180071 chr1:171180201	chr1:171179477						
chr1:171179287	chr1:171179939 chr1:171180021			chr1:171179477		chr1:171179670			
chr1:171179477			chr1:171179670	chr1:171179939 chr1:171180021					
chr1:171179779				chr1:171179939 chr1:171180021					chr1:171180071 chr1:171180201
chr1:171179939				chr1:171180071 chr1:171180201					chr1:171180021
chr1:171180021				chr1:171180071 171180071 chr1:171180201					
chr1:171180071									chr1:171180201
chr7:30464249		chr7:30485722 chr7:30490711							chr7:30465424

Summary of tagSNPs

An analysis to identify SNPs, from the entire genetic dataset, which tag the phenotype-associated SNPs (tagSNPs) was performed. In view of the observation that there is an overall lack of strong pair-wise LD in the current dataset, a minimum value of r^2=0.5 was set as being necessary to declare that one SNP tags another SNP. Figures 9-10 present results of this analysis per EGC for the three candidate genes.

It was noticeable that, as expected, there is both inter-genic and inter-SNP variation in the number and kb-span (distance between the tagged and tagging SNPs) of tagSNPs. The number of tagSNPs found ranged from zero tagSNPs to almost 20. Again, it can be seen that the top-SNPs of this study had few or no SNPs that tag them even at r^2=0.5. The distance between SNPs ranged from 5kb to more than 20kb.

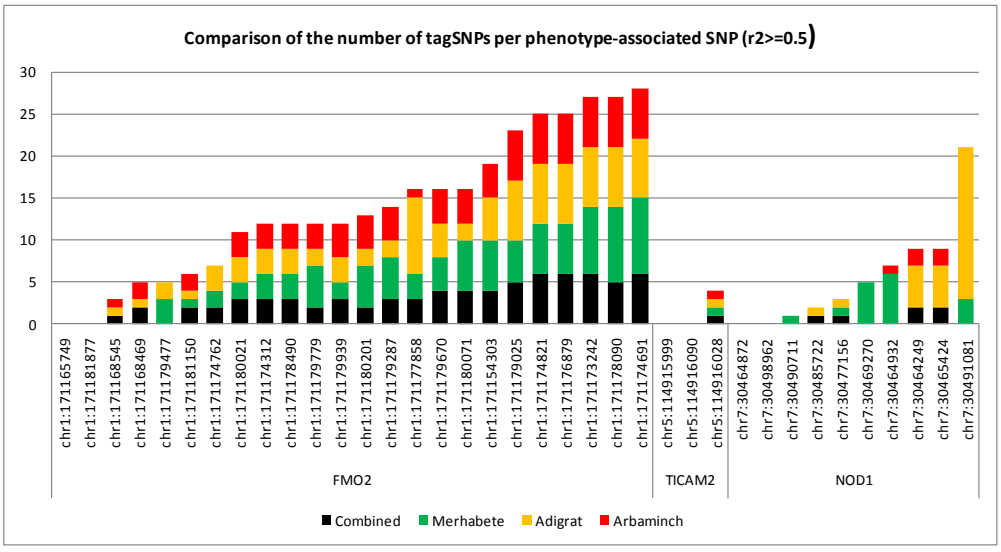


Figure 20: Comparison of number of tagSNPs per phenotype-associated SNPs

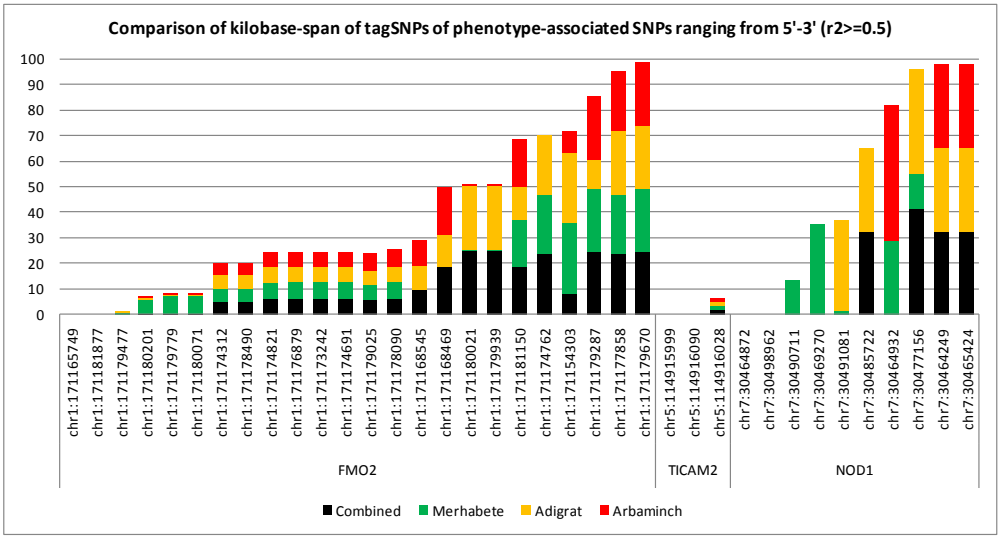


Figure 21: Comparison of distances (kb) between tagSNPs and tagged phenotype-associated SNPs

As has been demonstrated in the previous section on the analysis of the extent and pattern of LD, there were also inter-ethnic differences in both the number and kb-span of tagSNPs in a candidate-gene-specific manner. For example, the Merhabete population had the largest number of tagSNPs (90) for the phenotype-associated SNPs of the FMO2 gene located within an average of 9kb-span; the Adigrat population had the highest number of tagSNPs (30) for the NOD1 gene located within 17kb-span; and the Arbaminch population had the lowest number of tagSNPs for both the FMO2 and NOD1 genes. The TICAM2 gene had the lowest number and kb-span with equivalent values in all three EGCs (Figures 11-12). Although these variations might not be statistically significant, it is informative for studies in these populations that are designed on haplotype-tagging SNPs. For example, based on this dataset, it is relatively easy to find tagSNPs for FMO2 loci in the Merhabete population and in Adigrat population for the NOD1 gene; and, much difficult to do the same for the TICAM2 gene in all populations. It may also reflect

population-specific demographic (for example, population bottlenecks, migration) and evolutionary (for example, differential environmental or disease related selective pressures) history each EGC has undergone.

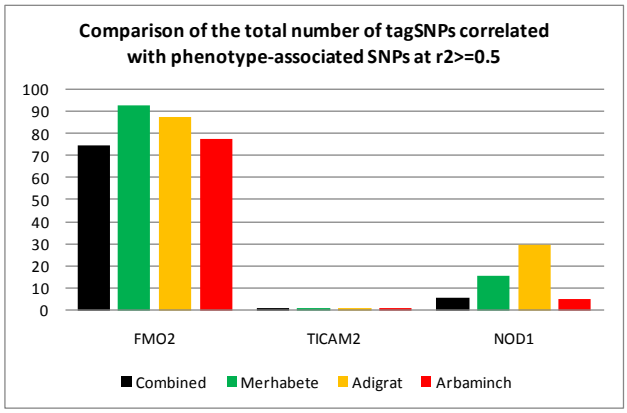


Figure 22: Comparison of the total number of tagSNPs between EGCs

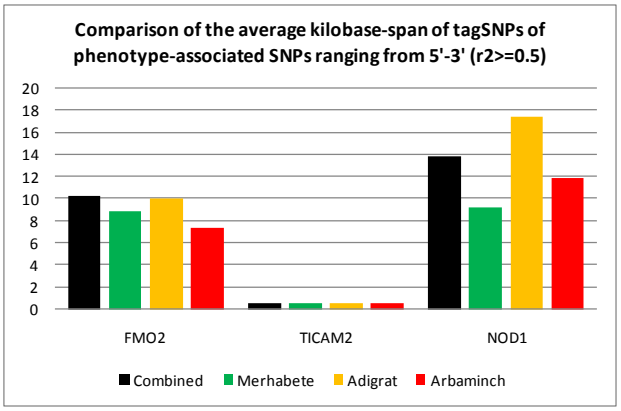


Figure 23: Comparison of the average distance (kb) of tagSNPs between EGCs

LD-/Haplotype-based association analysis

All the statistical tests of association described above were based on single nucleotide polymorphisms. And, although methods based on SNPs lead to significant results, methods based on haplotypes comprising multiple SNPs on the same inherited chromosome may provide additional power for mapping disease loci and also provide insight on factors influencing the dependency among genetic markers (Liu, et al, 2008). This analysis is based on haplotype inference among a specified set of SNPs to phase and test for association.

In this study, PLINK software was used perform haplotype-based association test which is a function designed to provide a representation of a single SNP association analysis in terms of the surrounding haplotypic background, i.e., association within a haplotype-based framework that track the reference SNPs. Specifically, all the phenotype-associated SNPs discovered were retested for association to see: whether the SNPs have independent haplotyic effects or test whether a set of SNPs explain an overall association statistic; and, whether specific haplotypes have association or a specific set of haplotypes explain an overall association test statistic

LD-based association can be thought of as a technical validation and refining of a single SNP association signal by framing the test within the haplotypic context of the flanking SNPs, grouping similar haplotypes, and testing for differences in the frequency of the various groups. For example, if a haplotype shows association results of a similar or greater magnitude, it could be taken as additional support for the original association of the reference SNP. And, although it is still possible that the association may be just due to chance, or due to population stratification, etc, the haplotype association would suggest that it is unlikely to be due to some technical

genotyping artefact that was specific to the reference SNP, as the same signal is observed from the haplotypic background.

In the current study, the haplotype-based association test was performed for all three candidate genes in all datasets and all haplotypes with >= 0.01 minor-haplotype frequencies were included in the test. In Tables 15-17, only results showing significant overall (omnibus) and/or haplotype-specific associations are presented.

The haplotype-based association test for TICAM2 identified two haplotypes in 'Active TB vs. No Active' and the omnibus test was also significant. Conditional test controlling for haplotype effects identified showed both explain the overall association.

Table 15: Results of haplotype-based association analysis for the TICAM2 gene

Conditional haplotype test results : TICAM2 Gene			
HAPLOTYPES: Active TB vs. No Active TB	FREQ	P: Haplotype-specific test	P: After controlling
AGG	0.0154	0.017	0.112
CAA	0.122	0.0225	0.0871
CGG	0.426	0.572	0.00757
CGA	0.437	0.708	0.00692
Likelihood ratio test: chi-square =			
df = 3			
p = 0.0178			

Three haplotypes, all explaining the significant omnibus test in the 'Active TB vs. No Active TB' and 'Active TB vs. No LTBI' datasets, were identified in NOD1. 5 haplotypes were significant in 'Active TB vs. LTBI' test-model but no single haplotype accounted for the significant overall test. A single haplotype was associated with LTBI in the 'LTBI vs. No LTBI' dataset.

Table 16: Results of haplotype-based association analysis in the NOD1 gene

Conditional haplotype test results : NOD1 Gene		
HAPLOTYPES: Active TB vs. No Active TB FREQ	**P: Haplotype-specific test**	**P: After controlling**
TGTATGGGGG 0.63	0.000711	0.867
TGTATTTGGG 0.0209	0.0138	0.201
TGTATGGGAG 0.0142	0.039	0.106
Likelihood ratio test: chi-square = 13.3		
df = 6		
p = 0.0381		
HAPLOTYPES: Active TB vs. No LTBI FREQ	**P: Haplotype-specific test**	**P: After controlling**
GTTGGGGG 0.614	0.0065	0.309
GTCTTTGG 0.0163	0.0168	0.19
GTTTGGGG 0.0116	0.0389	0.121
Likelihood ratio test: chi-square = 15.7		
df = 8		
p = 0.0471		
HAPLOTYPES: Active TB vs. LTBI FREQ	**P: Haplotype-specific test**	**P: After controlling**
TGATGGGAG 0.0167	0.00538	0.00353
TGACTTTGG 0.0121	0.013	0.00195
TGATTTGGG 0.0211	0.0317	0.00108
TGATTGGGG 0.0113	0.032	0.00108
TGATGGGGG 0.61	0.0357	0.001
Likelihood ratio test: chi-square = 32.3		
df = 10		
p = 0.000359		
LTBI vs. No LTBI FREQ	**P: Haplotype-specific test**	**P: After controlling**
TGTATTTTGG 0.0112	0.0466	0.172
Likelihood ratio test: chi-square = 13		
df = 7		
p = 0.0725		

Several haplotypes of the FMO2 gene were significantly associated ($p<0.05$) with TB-phenotypes: three haplotypes each with 'Active TB' in the 'Active TB vs. No Active TB', in 'Active TB vs. No LTBI', and in 'Active TB vs. LTBI' datasets (the omnibus tests were also significant). Two specific haplotypes were significant in the 'LTBI vs. No LTBI' test-model although the omnibus test was not significant.

Several haplotypes of the FMO2 gene were significantly associated ($p<0.05$) with TB-phenotypes: three haplotypes each with 'Active TB' in the 'Active TB vs. No Active TB', in 'Active TB vs. No LTBI', and in 'Active TB vs. LTBI' datasets (the omnibus tests were also significant). Two specific haplotypes were significant in the 'LTBI vs. No LTBI' test-model although the omnibus test was not significant.

Table 17: Results of haplotype-based association analysis in the FMO2 gene

Conditional haplotype test results : FMO2 Gene		
HAPLOTYPES: Active TB vs. No Active TB FREQ	**P: Haplotype-specific test**	**P: After controlling**
AGCTCTACAATCCCCTCGTTGCGC 0.0453	0.000103	0.0558
ATCTTTGGGGGTCTCCCAGGATGA 0.0259	0.0127	0.0038
CGCTTTGGGGGTCTCCCGTTGCGA 0.0132	0.0338	0.00216
Likelihood ratio test: chi-square = 38.4		
df = 15		
p = 0.000798		
HAPLOTYPES: Active TB vs No LTBI FREQ	**P: Haplotype-specific test**	**P: After controlling**
AGCTCTACAATCCCCTCGTTGCGC 0.0378	0.000136	0.0669
ATCTTAGGGGGTTTCCCAGGATGA 0.012	0.0243	0.00388
ATCTTTGGGGGTCTCCCAGGATGA 0.027	0.0381	0.00301
Likelihood ratio test: chi-square = 37.2		
df = 15		
p = 0.0012		
HAPLOTYPES: Active TB vs. LTBI FREQ	**P: Haplotype-specific test**	**P: After controlling**
AGCTCTACAATCCCCTCGTTGCGC 0.0378	0.000136	0.0669
ATCTTAGGGGGTTTCCCAGGATGA 0.012	0.0243	0.00388
ATCTTTGGGGGTCTCCCAGGATGA 0.027	0.0381	0.00301
Likelihood ratio test: chi-square = 37.2		
df = 15		
p = 0.0012		
HAPLOTYPES: LTBI vs. LTBI FREQ	**P: Haplotype-specific test**	**P: After controlling**
AGCTCTACAATCCCCTCGTTGCGC 0.0294	0.00703	0.442
ATCTTTGGGGGTCTCCCAGGATGA 0.0196	0.0345	0.262
Likelihood ratio test: chi-square = 21.4		
df = 15		
p = 0.125		

When specific haplotypes of FMO2 were controlled (i.e., effectively left out of the association test model) in order to see if they explain the overall association, one haplotype was found to explain the entire association (shown by the rise in p-value, p>0.05, while the rest remained significant). This is a remarkable and novel finding in that this particular haplotype contains the well known FMO2 isoform, FMO2*1 (C) allele, flanked by two of the top TB-susceptibility alleles consistently discovered in this study by the SNP based association tests described earlier: SNPs chr1:171165749 (risk allele-T) and chr1:171181877 (risk allele-A).

Table 18: The FMO2*1 allele does not segregate with susceptibility alleles

Haplotypes of FMO2 gene formed by phenotype-associated SNPs																								
1 171154303 chr1:171154303	1 171165749 chr1:171165749	1 171168469 chr1:171168469	1 171168545 chr1:171168545	1 171173242 chr1:171173242	1 171174312 chr1:171174312	1 171174691 chr1:171174691	1 171174762 chr1:171174762	1 171174821 chr1:171174821	1 171176879 chr1:171176879	1 171177858 chr1:171177858	1 171178090 chr1:171178090	1 171178490 chr1:171178490	1 171179025 chr1:171179025	1 171179287 chr1:171179287	1 171179477 chr1:171179477	1 171179670 chr1:171179670	1 171179779 chr1:171179779	1 171179939 chr1:171179939	1 171180021 chr1:171180021	1 171180071 chr1:171180071	1 171180201 chr1:171180201	1 171181150 chr1:171181150	1 171181877 chr1:171181877
0.199	0.0668	0.024	0.0411	0.173	0.118	0.17	0.0582	0.17	0.17	0.0822	0.135	0.118	0.135	0.185	0.32	0.152	0.33	0.488	0.488	0.33	0.33	0.024	0.111
C A	T G	A C	C T	C T	A T	A G	C G	A G	A G	T G	C T	T C	C T	T C	T C	G C	G A	G T	G T	G A	C T	A G	A C
A	G	C	T	C	T	A	C	A	A	T	C	C	C	C	T	C	G	T	T	G	C	G	C
A	G	C	T	C	T	A	G	A	A	G	C	C	C	C	T	C	G	T	T	G	C	G	C
A	G	C	T	C	T	A	G	A	A	T	C	C	C	C	T	C	G	T	T	G	C	G	C
A	G	C	T	T	T	G	G	G	G	G	T	C	T	C	C	C	A	G	G	A	T	G	A
A	G	C	T	T	T	G	G	G	G	G	T	C	T	T	T	G	A	T	T	A	T	G	A
C	G	C	T	T	T	G	G	G	G	G	T	C	T	C	C	C	G	T	T	G	C	G	A
A	G	A	T	T	T	G	G	G	G	G	T	C	T	C	C	C	A	G	G	A	T	A	C
A	G	C	T	C	T	A	G	A	A	G	T	C	T	C	C	C	G	T	T	G	C	G	C
A	G	C	T	T	A	G	G	G	G	G	T	T	T	C	C	C	A	G	G	A	T	G	C
A	G	C	T	T	T	G	G	G	G	G	T	C	T	C	C	C	A	G	G	A	T	G	C
A	G	C	T	T	T	G	G	G	G	G	T	C	T	C	C	C	G	T	T	G	C	G	C
A	G	C	T	T	T	G	G	G	G	G	T	C	T	T	T	G	A	T	T	A	T	G	C
C	G	C	C	T	T	G	G	G	G	G	T	C	T	T	T	C	A	T	T	A	T	G	C
C	G	C	C	T	T	G	G	G	G	G	T	C	T	T	T	G	A	T	T	A	T	G	C
C	G	C	T	T	A	G	G	G	G	G	T	T	T	C	C	C	A	G	G	A	T	G	C
C	G	C	T	T	T	G	G	G	G	G	T	C	T	C	C	C	G	T	T	G	C	G	C
A	T	C	T	T	A	G	G	G	G	G	T	T	T	C	C	C	A	G	G	A	T	G	A
A	T	C	T	T	T	G	G	G	G	G	T	C	T	C	C	C	A	G	G	A	T	G	A
C	T	C	T	T	T	G	G	G	G	G	T	C	T	C	C	C	G	T	T	G	C	G	A

A surprising, but ultimately logical, finding was that the 'C' allele never appears on the same haplotypic background in combination with its two flanking TB-risk alleles: the left-flanking 'T'

risk-allele and the right-flanking 'A' risk allele (Table-18). It only appears as a 'GCC' haplotype. The 'ultimately logical' part is the fact that in the SNP-based association tests discussed earlier, the 'C' allele (FMO2*1) was found to be negatively associated with Active TB, i.e., it has a protective effect against Active TB phenotype and, hence, the 'GCC' haplotype represents the 'protective haplotype' while the alternative 'TTA' allele represents the 'disease haplotype'. In general, there appears to be a dichotomy of '-C-/-T-' haplotypes associated with decreased and increased risk to Active TB. It can be concluded, therefore, that the protective effects of the FMO2 gene is most probably due to a haplotypic effect of the FMO2*A1 locus while the susceptibility SNPs are particularly located on recombination hotspots and seem to act independently.

To summarize this section, the LD-based association tests strongly support the findings obtained through the basic SNP-based association. All the moderate and strong signals of associations discovered by the allelic tests were replicated. This lends support to the robustness of the significant associations since, at the very least, the LD-based tests preclude possible technical genotyping artefacts that may have influenced the association statistic. The above results demonstrates the value of association analysis methods based on LD for detecting genetic variations that are responsible for complex human diseases such as TB.

LD block structure/Haplotype diversity of FMO2 and allelic/genotypic distribution of FMO2*1/FMO2*2:

In light of the novel findings regarding the FMO2 gene, it was deemed necessary to assess the FMO2 data and provide some contrasting descriptions of its population genetics in terms of the

LD/haplotype architecture of the exonic: number of LD blocks, LD-kb span, haplotype frequency and diversity including number of SNPs included per haplotype. Pair-wise LD between phenotype-associated SNPs only, and corresponding haplotypes and frequencies, was calculated for the combined and individual EGCs. The extent of LD was estimated in kilobase distance between the right-left flanking SNPs. [Note: Haplotype blocks were estimated following the default procedure in Haploview where by 95% CI are generated on D' and each model comparison is 'strong LD', 'inconclusive' or 'strong recombination'. A block is created if of 95% informative (i.e., non-conclusive') comparisons are 'strong LD'.

In simple terms, the extent of LD is a matter of mutation and recombination rates, where by a new mutation creates a new extended haplotype and recombination events decay or breakdown haplotypes in to shorter sizes and increase haplotype diversity or number of haplotypes. Therefore, the older a population is, the more opportunity for mutational events to occur, and the higher the chance for recombination to act upon and create haplotype diversity. The existence of haplotype block structure has serious implication for association-based studies and for mapping of disease genes/SNPs. On the one hand, if a causative allele occurs within a long block of LD, then it may be difficult to localize that causative gene at a fine scale by association mapping. On the other hand, if the diversity of haplotypes within blocks is low, then common disease genes may be mapped, at least to within a haplotype block by using fewer markers (Anderson, et al, 2003).

In the current genetic dataset, the number of LD blocks ranged from three in Arbaminch to seven in Adigrat populations, with Merhabete showing five blocks (Table-19). The longest haplotype

block was found in Merhabete with13 kb and with 9 SNPs, the longest haplotype in Arbaminch was 8 kb with 8 SNPs, and that of Adigrat 2kb with 6 SNPs included. The largest number of haplotype blocks per LD block was found in Arbaminch with 11 haplotypes, Merhabete showed 7, and Adigrat 5. The highest maximum frequency observed for any haplotype, 0.90, was found in Adigrat, the highest frequency in Merhabete was 0.80, and Arbaminch's highest haplotype frequency was 0.44.

Table 19: Comparison of the haplotype structure of FMO2 gene between EGCs

Comparison of haplotype structure in the FMO2 gene									
EGC	# Blocks	Start BP	End BP	Span KB	NSNPS	NHAP/BLOCK	Min.Hap. Freq	Max. Hap. Freq	Avg. Hap. Freq
		171180201	171180240	0.04	2	3	0.056	0.662	0.333
		171179989	171180071	0.083	3	4	0.051	0.488	0.250
	5	171178490	171179939	1.45	8	6	0.033	0.363	0.164
		171154303	171162735	8.433	8	8	0.010	0.342	0.125
Combined		171168585	171177119	8.535	13	12	0.011	0.329	0.082
		171154303	171154378	0.076	2	3	0.060	0.804	0.333
		171179477	171179670	0.194	2	3	0.138	0.702	0.333
	5	171179779	171180071	0.293	5	4	0.059	0.543	0.249
		171173242	171176879	3.638	7	6	0.011	0.694	0.166
Merhabete		171154959	171168804	13.846	9	7	0.016	0.391	0.143
		171154303	171154378	0.076	2	3	0.083	0.784	0.333
		171174312	171174531	0.22	2	2	0.098	0.902	0.500
		171168584	171168804	0.221	4	4	0.054	0.569	0.250
Adigrat	7	171179779	171180071	0.293	4	3	0.172	0.485	0.330
		171162438	171162735	0.298	3	2	0.142	0.858	0.500
		171180201	171181754	1.554	2	3	0.147	0.662	0.333
		171174691	171177068	2.378	6	5	0.020	0.672	0.200
		171179670	171180021	0.352	5	5	0.034	0.445	0.198
Arbaminch	3	171168584	171176912	8.329	12	11	0.010	0.290	0.090
		171154303	171162735	8.433	8	6	0.053	0.313	0.164

Comparison of the overall LD and haplotye structure between the EGCs revealed a general pattern: the higher the number of haplotypes a population has, the lower the LD block kb-span, the lower the number of haplotypes per LD block, the lower the number of SNPs per haplotype, the higher the haplotype frequencies, and vice versa (Figure-13).

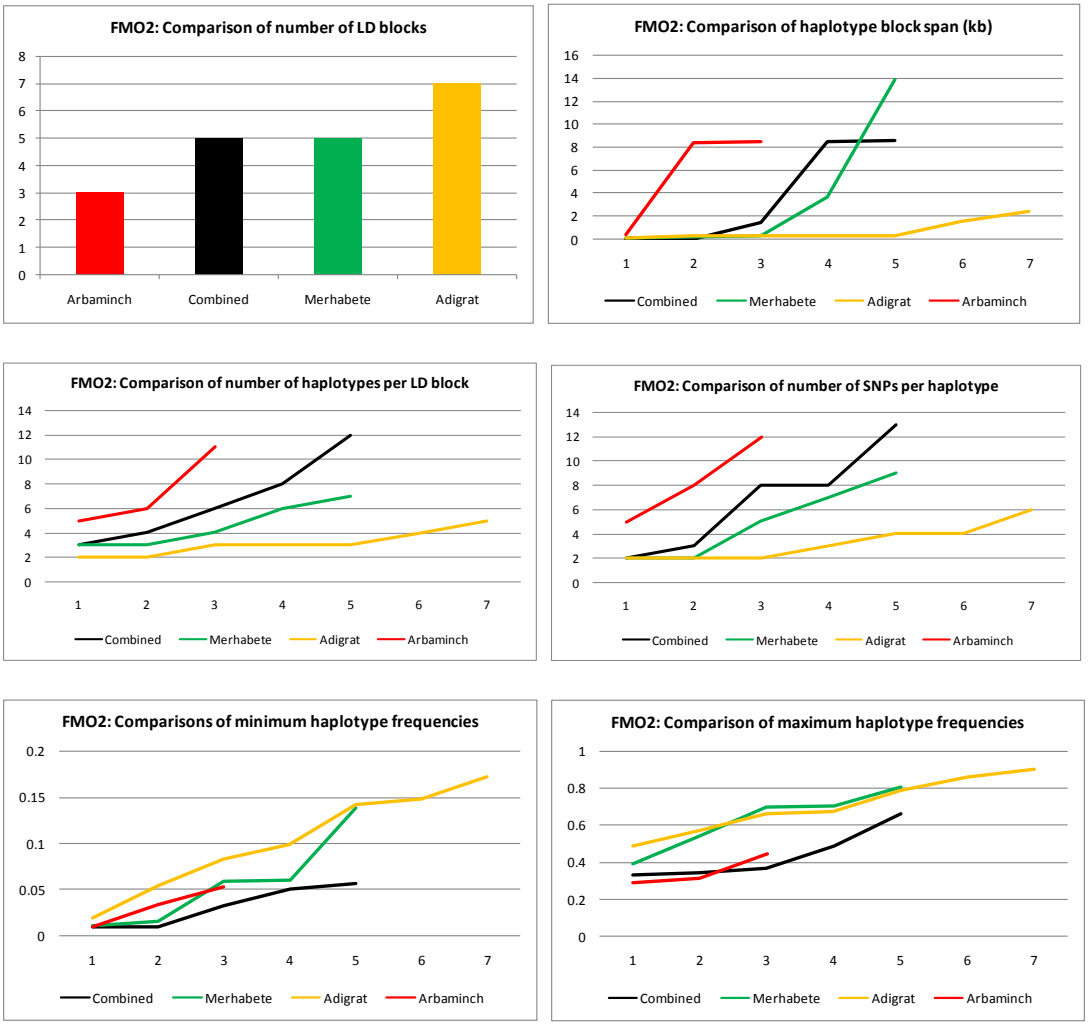


Figure 24: Comparisons of haplotype structure in FMO2

The FMO2*2 allele (T) was found to be the major allele in the present study while the ancestral FMO2*1 (C) allele remained a minor one (Table-20 and Figure-25). For example, 24% of all individuals genotyped in this study carried at least one of the potentially deleterious FMO2*1 ancestral allele. Another Ethiopian data obtained from a more recent genotyping effort (Pagani, et al, 2015) also revealed that the ancestral allele (C) was the minor allele (MAF=0.1205). There was also an apparent inter-ethnic difference (although not statistically significant) in the proportions of FMO2*1/FMO2*2 alleles: the Adigrat population had the highest proportion of

the 'C' allele (32%), Merhabete 23%, and Arbaminch had the least with 19%. A previous study (Krishna, et al, 2008) reported larger proportions with the Gambela population carrying as much as 49% of the FMO2*1 allele. On the other hand, the reverse is true in other populations: the FMO2*1 allele is reported to be virtually absent in Europeans and some Asian populations (Krueger, et al, 2005). The reasons for such inter-ethnic and global differences in the frequency and distribution FMO2*1 should also be another reason for curiosity.

Table 20: Comparisons of allelic and genotypic frequency distribution of the FMO2*1/2 locus between populations

Comparison of the allelic anf genotypic distribution of FMO2*1 locus									
		Genotype count			Genotype frequency				
Population	Total sample	CC	TC	TT	CC	TC	TT	MAF (C allele)	At least 1 C
ALL,PS	304	8	64	232	0.026	0.211	0.763	0.132	0.237
Merhabete,PS	74	1	16	57	0.014	0.216	0.770	0.122	0.230
Adigrat,PS	89	3	25	61	0.034	0.281	0.685	0.174	0.315
Arbaminch,PS	141	4	23	114	0.028	0.163	0.809	0.110	0.191
Gambella	106	5	47	54	0.047	0.443	0.509	0.269	0.491
Addis ababa	24	1	7	16	0.042	0.292	0.667	0.188	0.333
Borena, wollo	36	0	7	29	0.000	0.194	0.806	0.097	0.194
Dessie, wollo	26	1	6	19	0.038	0.231	0.731	0.154	0.269
Pagani, etal	220	4	45	171	0.018	0.205	0.777	0.121	0.223
FMO2*1=C ancestral, active allele; FMO2*2A, mutant, inactive allele. PS: present study									

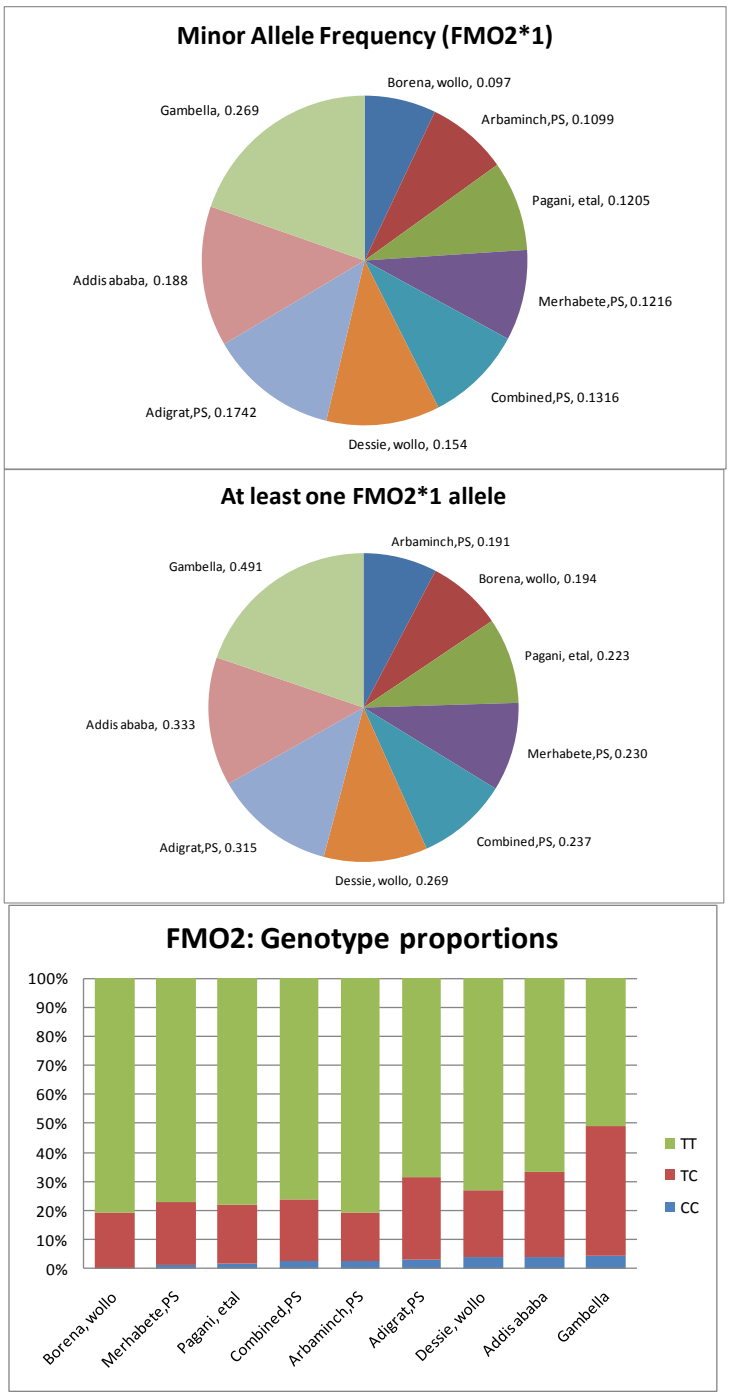


Figure 25: Descriptions allelic and genotypic frequency and distribution of the FMO2*1/FMO2*2 locus

This phenomenon of the ancestral allele being the minor allele is uncommon. Could this pattern of allelic distribution be explained by drift alone where by the frequency of a mutant allele/isoform increases at the expense of the ancestral allele, even to fixation (100%), due to

random selection and transmission? Or is there a positive selection for one allele and a negative one for the other? When one considers the fact that the FMO2*2 mutant, but now major, allele is dysfunctional, and thus neutral, it is tempting to ask if the ancestral, but now minor, allele which codes for a functional protein involved in metabolism has a deleterious effect and is thus undergoing a natural purgatory. For the later to happen, the FMO2*1 allele must be acting to cause or predispose to some highly penetrant, early-onset, and deleterious phenotype. The recent discovery of its association with adverse reactions to industrial and pharmaceutical chemicals is insufficient to fully account for its low frequency since these chemicals are relatively recent man-made substrates and have not had much time to act in a differential or population-specific manner.

In lieu of the purifying selection postulate described above, it is proposed here that one possible explanation for the differential distribution of the FMO2 alleles is in terms of a combination of human population genetics concepts: 'Out-of-Africa', 'founder effect', and 'drift'. In this scenario, what if the wild FMO2*1 allele mutated into the FMO2*2 allele before the OOA event; then the FMO2*2 allele was carried disproportionately (drift) by the OOA migrants; then the founder-effect will propagate and establish the FMO2*2 allele in the newly founded and increasing population. This will explain the absence of the FMO2*1 allele in non-African populations. What about the increased frequency of the FMO2*2 allele in African populations? This could happen by drift, especially since this allele is functionally neutral. On the other hand, the FMO2*1 allele need not be considered 'deleterious' (except, of course, vis-a-vis its interaction with some unnatural, recently-manufactured chemicals mentioned above) to explain its low frequency despite being an ancestral allele. To the contrary, as the current study demonstrates, it

could have a beneficial protective health effect that could explain its persistence in African populations and populations with recent African descent. And, the discovery of an ancestral allele that protects against TB is not surprising in TB-endemic populations that were being persistently challenged by an ancient disease for hundreds of centuries.

Summary of the study results

Summary of the association of TICAM2 and NOD1with TB

This study replicated the association of NOD1 and TICAM2 genes with TB in Ethiopian populations. Two variants were found in NOD1 that were significantly associated with TB risk, and additional variants that were nominally associated. Furthermore, one of the most significantly associated SNPs in NOD1 was novel. Though the study did not observe a statistically significant association between TB and TICAM2 variants after accounting for multiple testing, these nominally significant results still provide an independent replication of a previous report (Hall, et al, 2015). The study also found consistent evidence of association of NOD1 variants with active TB in both our primary and sensitivity test models, and two of these results were significant after multiple-test correction. There were no notable associations with LTBI as the case phenotype, except for a single nominally associated NOD1 SNP, though some of the individuals in this category may progress to active TB later in time by virtue of possessing any of the susceptibility variants. Although the relatively smaller sample size in this test-model may also have reduced the power to detect association signals it also captured one SNP in NOD1, rs17159043, which showed signals of association with TB in both the original study ($p=0.009$) and the current replication study ($p=0.040$ in LTBI v. No LTBI). Generally, the study provided an independent replication of association between TB and variants in TICAM2 and NOD1. By conducting QFT to assess latent infection status, the study was able to demonstrate that a higher proportion of these variants are associated with susceptibility to active TB disease, not a latent infection. Some variants in both genes were also associated with reduced risk to active TB.

Summary of the association of FMO2 with TB

This study identified for the first time an association between FMO2 genetic polymorphisms and TB progression phenotypes both at the SNP and haplotype levels. The study discovered multiple SNPs, including novel variants, associated with increased or decreased risk to TB in Ethiopian populations. The ancestral, functionally active, minor allele (FMO2*1), when compared with the mutant, functionally inactive, major allele (FMO2*2), was consistently found to be significantly associated (p=0.01, OR=0.3-0.6) with resistance to Active TB (in 'Active TB vs. No Active TB' and 'Active TB vs. No LTBI' genetic test-models) by SNP-based tests of association. These significant associations were confirmed by LD-/haplotype-based tests of association in all four test-models (p=0.000103-0.00703). These protective haplotypic effects were exclusively found on a haplotypic background consisting of the FMO2*1 (C) allele and not the the FMO2*2 allele. The protective haplotype never appears in allelic combinations that consist of alleles found to be highly associated with susceptibility to TB, i.e., the protective FMO1*1 allele does not segregate with susceptibility alleles. It is proposed here that the differential expression of FMO2 genetic variants in *Mtb*-infected cells may function by modulating biological pathways involved in the regulation of oxidative stress status or, in general, the redox profile/balance by acting directly on *Mtb*-derived components or in combination with other resident environmental (xenobiotic) substrates. The currently proposed mechanism of action of FMO2 is supported by the findings of numerous studies of the role of oxidative stress status of cells as it relates to TB pathogenesis and reported its association with TB progression. The novel discovery of this study revealed the "double-edged-sword" of FMO2 vis-à-vis TB: FMO2 has dual role in TB pathogenesis and anti-TB pharmacogenomics. The dual role of FMO2 in TB is demonstrable in that on the one hand FMO2 has been shown to metabolize some widely used anti-tubercular drugs, such as ethionamide and thiacetazone, with adverse toxic reactions while on the other hand, as this study

shows, FMO2 polymorphisms are associated with TB pathogenesis mainly by conferring reduced risk to TB.

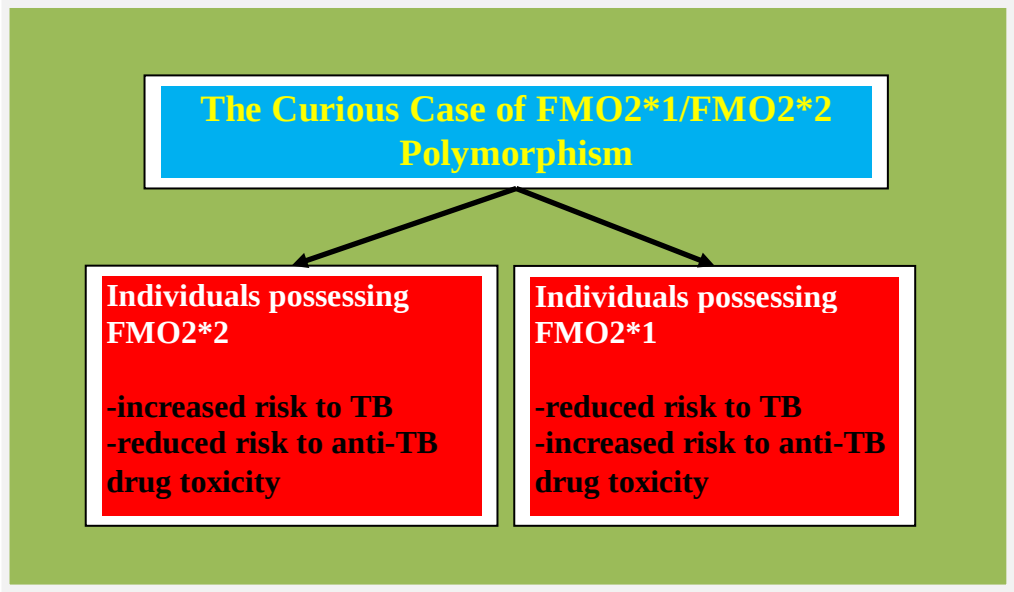


Figure 26: The "double-edged-sword" FMO2 vis-à-vis TB: TB pathogenesis and TB pharmacogenomics

X. GENE/SNP ANNOTATION

In this study, phenotype-associated SNPs were annotated using UCSC Genome Browser (GRCh37/hg19) on which the sequencing platform was designed. In this study, since strong LD (r2>0.8) was not observed for most of the highly significant SNPs, only the phenotype-associated SNPs were included in the query.

Most of the SNPs which showed highly significant association signals were novel (Table-21). Some SNPs were coding with non-synonymous, synonymous mutations while others were reported either as intronic or 3'/5' untranslated regions.

Table 21: Functional consequences of mutations

GENE	chr#:base_position	rs_ID	A1 (Minor allele)	A2 (Major allele)	PREDICTED FUNCTION (UCSC)
FMO2	chr1:171154303	rs28369794	C	A	UPSTREM GENE VARIANT
	chr1:171165749	NOVEL VARIANT	T	G	
	chr1:171168469	rs112884205	A	C	INTRON VARIANT
	chr1:171168545	rs2307492	C	T	INTRON VARIANT, MISSENSE VARIANT
	chr1:171173242	rs7517460	C	T	INTRON VARIANT
	chr1:171174312	rs16864177	A	T	INTRON VARIANT (IN STROG LD WITH A MISSENSE VARIANT)
	chr1:171174691	rs7536646	A	G	SYNONYMOUS VARIANT
	chr1:171174762	rs28369899	C	G	MISSENSE VARIANT
	chr1:171174821	rs7536745	A	G	INTRON VARIANT
	chr1:171176879	rs6671692	A	G	SYNONYMOUS VARIANT, INTRON VARIANT
	chr1:171177858	rs28369911	T	G	INTRON VARIANT
	chr1:171178090	rs6661174	C	T	STOP GAINED, 3' UTR VARIANT (NONSENSE)
	chr1:171178490	rs28369914	T	C	DOWNSTREAM GENE VARIANT, 3' UTR VARIANT
	chr1:171179025	rs6664553	C	T	3' UTR VARIANT
	chr1:171179287	rs7512785	T	C	3' UTR VARIANT
	chr1:171179477	rs7515157	T	C	3' UTR VARIANT
	chr1:171179670	rs28369918	G	C	3' UTR VARIANT
	chr1:171179779	rs73032526	G	A	3' UTR VARIANT
	chr1:171179939	rs3174837	T	G	3' UTR VARIANT, SPLICE REGION VARIANT, INTRON VARIANT
	chr1:171180021	rs6425286	T	G	3' UTR VARIANT
	chr1:171180071	rs6673781	G	A	3' UTR VARIANT
	chr1:171180201	rs6668231	C	T	3' UTR VARIANT (N STRONG LD WITH A MISSENSE VARIANT)
	chr1:171181150	rs113252377	A	G	3' UTR VARIANT
	chr1:171181877	NOVEL VARIANT	A	C	
TICAM2	chr5:114915999	NOVEL VARIANT	A	C	
	chr5:114916028	rs2288384	A	G	3' UTR VARIANT
	chr5:114916090	rs256996	A	G	3' UTR VARIANT
NOD1	chr7:30464249	NOVEL VARIANT	TG	T	
	chr7:30464872	rs5743374	A	G	3' UTR VARIANT
	chr7:30464932	rs112070346	G	T	3' UTR VARIANT
	chr7:30465424	rs5743370	C	A	INTRON VARIANT
	chr7:30469270	rs17159043	C	T	INTRON VARIANT
	chr7:30477156	NOVEL VARIANT	T	G	
	chr7:30485722	rs751770147	T	G	INTRON VARIANT
	chr7:30490711	rs543650951	T	G	INTRON VARIANT
	chr7:30491081	rs151170709	A	G	NC TRANCRIPT VARIANT, MISSENSE VARIANT
	chr7:30498962	NOVEL VARIANT	C	G	

Further annotation using SNPnexus annotation tool (http://www.snp-nexus.org), a web server for functional annotation of novel and publicly known genetic variants, showed that none of the SNPs themselves were reported as actually being directly associated with a disease phenotype, instead, they were indirectly implicated by association with the entire genic region, or a couple of SNPs there in, which are associated (Table-22). The associated diseases are rather wide-ranging from classes of immunity, metabolism, to psychosis. The distinct absence of previous reports of association with TB goes to show the point made earlier that either submissions to the databases searched were not made, such as the recent study which identified polymorphisms in the TICAM2 and NOD1 gene associated with TB (Hall, et al, 2015), or the associations identified in the current study are novel. That being said, there were some interesting reports of disease associations the biological pathways of which may be relevant to TB pathogenesis: infection (e.g., in TICAM2), lipoprotein metabolism (e.g., in FMO2 and NOD1), and leprosy, bacteremia, asthma and diabetes (e.g., in NOD1). Obviously, the immune system is relevant in the development of any infection, leprosy, bacteremia, and asthma. Leprosy is caused by a close relative of *Mtb*, *M. leprea* and shares much of its evolutionary and natural history of disease progression. The human immune system also reacts to both infections in a similar fashion. Inflammatory reactions leading to asthma have also been associated with the physical blockage or entrapment of *Mtb*. Human pathways involved in lipoprotein and trglyceride metabolism have been suggested to be manipulated by the *Mtb* bacteria for immune evasion and its own survival in infected macrophages and granulomas.

Table 22: Previously reported disease associations for the current phenotype-associated SNPs

Previously Reported Genetic Association of Complex Diseases and Disorders (GAD) for the Phenotype-Associated SNPs Identified in the Current Study							
SNP	Band	Effect, PS	Phenotype	Disease Class	Gene	Associated SNPs	Population
chr1:171165749:T/G:1	q24.3	S	{leukemia, myeloid}, {Chronic renal failure|Kidney Failure, Chronic}, {Lipoproteins}, {Hypothroidism}, {Hearing loss}	{Cancer}, {Renal}, {Metabolic}	FMO2, SLC39A1, MTF1	rs2072704, rs3748682	African American, Mexican, Africa-south of Sahara
chr1:171168469:A/C:1	q24.3	S					
chr1:171173242:C/T:1	q24.3	R					
chr1:171174312:A/T:1	q24.3	S					
chr1:171174691:A/G:1	q24.3	R					
chr1:171174762:C/G:1	q24.3	R					
chr1:171174821:A/G:1	q24.3	R					
chr1:171176879:A/G:1	q24.3	R					
chr1:171177858:T/G:1	q24.3	R					
chr1:171178090:C/T:1	q24.3	R					
chr1:171178490:T/C:1	q24.3	S					
chr1:171179025:C/T:1	q24.3	R					
chr1:171179287:T/C:1	q24.3	R					
chr1:171179477:T/C:1	q24.3	R					
chr1:171179670:G/C:1	q24.3	R					
chr1:171179779:G/A:1	q24.3	R					
chr1:171179939:G/T:1	q24.3	S/R					
chr1:171180021:G/T:1	q24.3	S/R					
chr1:171180071:G/A:1	q24.3	R					
chr1:171180201:C/T:1	q24.3	R					
chr1:171181150:A/G:1	q24.3	S					
chr1:171181877:A/C:1	q24.3	S	{Lipoproteins}, {Hypothyroidism}	Metabolic	SLC39A1, MTF1	rs2072704, rs3748682	
chr5:114915999:A/C:1	q22.3	S	{Cystitis|Pyelonephritis|Urinary Tract Infections}	Infection	TICAM2		
chr5:114916028:A/G:1	q22.3	R					
chr7:30464872:A/G:1	p14.3	S	{asthma, hay fever IgE}, {atopic dermatitis}, {Crohn's disease ulcerative colitis}, {sarcoidosis}, {inflammatory bowel disease}, {cholangitis, sclerosing}, {duodenal ulcer gastritis}, {dermatitis and eczema}, {atherosclerosis, coronary}, {Type 2 Diabetes|edema | rosiglitazone}, {Leprosy}, {Chlamydophilia infections, stroke}, {Asthma|Bronchial Hyperreactivity|Hypersensitivity, Immediate}, {Bacteremia}, {Leukemia, Lymphocytic, Chronic, B-Cell}, {Triglycerides}	Immune, Infection, Metabolic, Cardiovascular, cancer, Pharmacogenomic, Schizophrenia	NOD1	rs56500751, rs10487933	Turkish, European,-Scandinivian, Scottish, German, Hungary, Hungary, Spain, India
chr7:30464932:G/T:1	p14.3	R					
chr7:30477156:T/G:1	p14.3	S					
chr7:30485722:T/G:1	p14.3	S					
chr7:30490711:T/G:1	p14.3	S					
chr7:30491081:A/G:1	p14.3	S			CARD4, MLL3, EIF2AK1		
chr7:30498962:C/G:1	p14.3	R			NOD1,CARD4, MLL3,		

XI. CONCLUSIONS, STRENGTHS, LIMITATIONS AND RECOMMENDATIONS OF THE STUDY

This study conclusively achieved its stated general and specific objectives and significantly proved its hypotheses. The study proposed to study the influence of candidate gene polymorphisms in TB progression. The basic premise of the study was that since there is a commonly observed inter-individual variation in the immune response to *Mtb* exposure and infection in that not all individuals exposed to *Mtb* get infected, and not all *Mtb*-infected individuals progress to active TB, it follows that the mere exposure to, and infection by, *Mtb* is, while necessary, not a sufficient cause for TB disease. The hypothesis tested was that polymorphisms in human innate immunity genes contribute to variation in the immune response to *Mtb* exposure and infection. This hypothesis was tested by setting up a genetic epidemiological association study design based on three candidate genes: TICAM2, NOD1 and FMO2. The study also hypothesized that the reasons for the apparent non-replication of previous TB genetic association results may be explained by the lack of precise definition of TB phenotypes that frustrates the discovery of underlying gene-disease associations or lead to spurious conclusions. Therefore, this study focused on constructing precise and consistent criteria of TB definition amenable for genetic epidemiological analysis based on an intermediate phenotype model that closely reflects the natural history of TB progression from *Mtb*-exposure, latent infection to active disease. This design also enabled to test another hypothesis of the study that TB progression stages may represent distinct TB phenotypes with respective stage-specific immuno-genetic risk profiles. In general, the study aimed to replicate previous TB genetic association signals and to test a novel TB-candidate gene association.

The study successfully replicated for the first time a novel finding of an association between TB and TICAM2 and NOD1 genes in an Ugandan cohort (Hall, et al, 2015) in Ethiopian populations. These genes are receptor genes of the innate immunity system involved in the synergistic antigen recognition, presentation, and processing that lead to the priming of the adaptive immune system against *Mtb* infection. The finding of this study is significant because, first, it replicated a previous result in an independent population further confirming the involvement of TICAM2 and NOD1 in the immune response to TB; second, the study discovered novel TB-associated SNPs in TICAM2 and NOD1 genes in the sampled Ethiopian populations; and, third, it validated the efficacy of the intermediate-phenotype model of the genetic epidemiological study design used for the identification of TB-associated genetic variants. Although it is typical in replication studies to perform a meta-analysis, this was not done in this study for two reasons. First, this is only the first attempt at replication of these two genes. Second, rather than analyzing the same exact SNPs as in the original report, this study conducted exon sequencing with some flanking buffers. This alternative approach would make a meta-analysis with the data from the original work nearly impossible. Although this enabled the identification of novel variants associated with TB, the gaps between the exons prevented coverage of all the SNPs in the original study, i.e., the percentage of matching sequences between the two studies was low (6 SNPs in TICAM2 and 34 SNPs in NOD1) which were further reduced by QC filtering. This effectively precluded a comprehensive comparative assessment at finer levels beyond the overall gene level. Nonetheless, this study captured one SNP in NOD1, rs17159043, which showed signals of association with TB in both the original study ($p=0.009$) and the current replication study ($p=0.040$ in LTBI v. No LTBI).

The third candidate gene of this study, FMO2, is an immune effector gene that plays an essential role of modulating oxidative stress levels in the innate anti-mycobacterial immune defense system as it affects *Mtb* survival, persistence and subsequent reactivation. The high expression level of FMO2*1 in the lung and its pharmacogenetic effect in relation to its oxygenase-driven metabolism of some anti-TB drugs resulting in toxic intermediates have been well characterized. However, it has never been studied in relation to TB pathogenesis before and, thus, both the hypothesis and the discovery of association with TB in this study are entirely novel. FMO2's candidacy in the present research was, therefore, purely a consequence of a daring 'out-of-the-box' thinking: "What if FMO2, with a specific functionally active genetic variant that is highly prevalent in sub-Saharan Africa, including Ethiopia, and known to be associated with adverse reaction to an anti-TB drugs, is also involved with the pathogenesis of TB disease itself through its mediation of oxidative stress in activated immune cells?". The study identified for the first time an association between FMO2 and TB both at the SNP and haplotype level. The pattern of association suggested a protective effect of FMO2 against both active and latent TB with distinct genetic variants underlying the TB progression pathway. Haplotype-based tests confirmed the SNP-based results with a single haplotype bearing the ancestral-and functional FMO2*1 "C" allele ("AGCTCTACAAT**C**CCCTCGTTGCGC") explaining the overall association. A remarkable finding of this study was that not only was FMO2*1 nominally associated with reduced risk to "Active TB" but it also does not co-segregate with the 5'-3' flanking top high-TB-risk alleles identified in the study. The study provides an evidence for the existence of an evolutionary adaptation to an ancient disease based on an ancestral genetic variant acting in a haplotypic framework in Ethiopian populations. Coupled with the apparently high prevalence of the FMO2*1 variant responsible for the adverse toxic reactions to some anti-MDRTB drugs

among various populations of Ethiopia, the current finding has enormous public health implications. From the genomics aspect, FMO2 might represent a curious paradigm of pleiotropy in action: a locus with genetic polymorphisms that render individuals susceptible to a disease also harbours nearby variants that not only protect against the same disease but, if a particular drug type is administered to treat the very same disease, it may lead to adverse reactions. Furthermore, the protective effect of the ancestral allele FMO2*1 was found to be both at the allelic and haplotypic level, although the haplotypic effect is much more pronounced. The evolutionary implication of this finding is enormous and raises further questions. For example, the evolutionary history of the FMO2 gene at the FMO2*1/2 locus is curious: What is the basis of the current high disparity in the frequency and distribution of these alleles between populations? Now that this study has identified a dichotomous 'protective/disease' haplotypes involving this locus, is there a possibility of directional selection? If the ancestral allele has indeed a protective effect, how then did it become to be a minor allele restricted to certain ancestral human populations, particularly in sub-Saharan Africa? And, can its frequency and distribution be explained in conjunction with available evidences for the origin of both modern humans and *Mtb* in Africa and the 'Out-of-Africa-and-Back' concept of human migration? In light of, first, the high prevalence and wide distribution of the FMO2*1 variant amongst Ethiopian populations and, second, the increasing incidence and prevalence of MDRTB, the pharmaco-genetic implication of the FMO2 finding in this study cannot be overstated since it practically means: we can no longer go on prescribing the usual thiourea-containing anti-tubercular drug regimen (and dosage?) without regard to patients' genetic risk profiles. This is particularly important with the current emergence and spread of MDR and XDR strains of *Mtb* that has led to the increased use of such drugs worldwide. For instance, studies in Ethiopia

(Agonafir, et al, 2010; Abate, et al 2014) found that not only were both MDR-/XDR-TB present in Ethiopian patients but also resistant strains against ethionamide were the most prevalent types and indicated the need for increased efforts to expand diagnostic services, treatment and care (Fantahun, et al, 2014). A more directed clinical study into the role of FMO2 in TB pathogenesis and pharmacogenomics is in line with this efforts. For example, functional annotation of FMO2 polymorphisms might reveal the immunogenetic basis for its association with both TB disease and treatment outcome. In this regard, studies (Veeramah, et al, 2008; Coussens, et al, 2013) have demonstrated the existence of differential TB immunologic profiles at presentation, becoming even more marked following initiation of anti-mycobacterial therapy, between patients of African vs. Eurasian ancestry that associated with ethnic variation in host genotype. Therefore, it is recommended here that, biomarkers that can easily (at "point-of-care") be used to screen for such polymorphisms should be developed. It is imperative to replicate the current findings in independent populations as well as investigating the possibility of the FMO2 gene, particularly at the FMO2*1 locus, being associated with other disease pathogenesis that involve oxidative stress. Future studies focusing on functional annotation of the genes and SNPs that showed association signals are also necessary. Variations that might be functionally relevant, such as those leading to amino acid changes or splicing site alterations, found in known regulatory elements or conserved non-coding sequences are priority targets for further investigations. The selection of an optimal number of tagSNPs based on both potential functional effect and haplotype information could increase the sensitivity and efficiency of large-scale genotyping projects. However, the interpretation of candidate SNP sites requires the integration of genome annotation data, such as gene(s)/protein(s) structure, and related information about the splicing isoforms with the sequence information. This is an essential step to enable and facilitate

hypothesis generation for further experimentation and validation. Generally, the novel FMO2 finding demonstrates the intricacies in the spectrum of genetic associations with TB pathogenesis and treatment. From the evolutionary perspective, it is also informative to note that how changes in the environment, in this case the manufacturing, distribution and utilization of new drugs to treat an ancient disease, may create a pressure on a genetic architecture evolutionarily shaped to fight the same disease. In other words, it is not a matter of the FMO2*1 allele being naturally deleterious, rather the change in the human environment that is becoming an artificial risk. And how, ultimately, it is the combination of all the factors involved in the resolution of TB infection (some with minor, some with major effects) that determines the outcome. It is the hope of identifying such genetic factors, particularly those with major effects, which drives genetic epidemiological investigations like the current study.

Overall, SNP-based association tests identified multiple variants within each of the candidate genes at varying degrees of significance. Association tests under the framework of LD and haplotypes also confirmed the robustness of these findings as well as identifying significantly associated disease-/protective-haplotypes. The associations identified were both with TB susceptibility (increased risk effect) and resistance (decreased risk effect). Examination of the strength and pattern of statistically significant associations, in terms of being restricted, consistent, correlated or directional also helped to shed light on the possible biological pathways involved underlying TB progression and the possible stage-specific effects of the genes and their variants through pleiotropy and/or epistasis. By contrasting TB disease with no disease and LTBI v. no LTBI, the study design helped to unravel a pattern of SNP-phenotype associations to determine which stage of TB pathogenesis was associated with these variants. However, there is

some limitation in this design in that individuals who were labeled as controls (no active TB, no LTBI, or with LTBI) at the time of data collection might develop LTBI or active TB later on. Therefore, a prospective study with follow up cohorts could produce a better result although in a TB endemic setting it could be difficult to differentiate between active TB as a result of reactivation of latent TB or due to new infections. In addition to the careful phenotypic characterization of the control population, there are other aspects of this study that further enhance its validity. First, although other transcription-regulatory regions are also important, it focused on padded full exon resequencing of the two candidate genes of interest, enabling the discovery of novel genetic variants with a higher probability of translation. Second, it recruited individuals from multiple ethnic groups, and examined possible population stratification effects.

The study also has a population genetics component that compared and characterized the sampled Ethiopian populations using allelic frequencies, MDS, IBS/IBD, LD and haplotype structure that found cryptic evidences of population stratification. A significant finding, however, was that the stratification cannot be explained in terms of ethnic clustering alone, i.e., the stratification did not follow ethnic lines. This might be an indication that intra-ethnic variation was greater than inter-ethnic variation. This finding seems to be in contradiction with recent population genetic analysis of Ethiopian populations which concluded that not only stratification exists but it is along ethnic/linguistic/cultural lines (Pagani, et al, 2012). However, the reason for the apparent contradiction may be the difference in the source and generation of the genetic datasets used for analysis, i.e., genome-wide data vs. exone sequences. The observed pattern of relatively minimal inter-ethnic population differentiation in terms of rare single nucleotide variations is consistent with the nature of exonic regions which, by functional necessity or

constraints, do not tolerate random variation as in non-coding /functional regions of the genome. On the other hand, the observed differences in LD/haplotype structures suggest the possibility of population-specific genetic signatures that need to be taken into consideration. This suggestion was corroborated by the fact that, in this study, the efficacy of IBS-based stratified analysis of association (i.e., based on the genetic dataset alone) in statistically controlling for the observed cryptic population stratification was much more than the stratified test of association based on self-declared ethnicity and sampling site.

The study results also support the hypothesis that the Ethiopian setting could indirectly help to explore the possibility that *Mtb*-human co-evolution may result in differential TB genetic risk profiles across population categories. As Ethiopia is considered to be the origin of both humanity and *Mtb*, the findings of this study are important for understanding the possible impact of *Mtb*-human co-evolution in TB pathogenesis. For example, while the Ethiopian study replicated the Ugandan findings at the gene level, the associated SNPs in overlapping loci are not shared and some are novel. Therefore, while the Ugandan and Ethiopian findings indicate, on the one hand, a common signature of association at the gene level in the two populations, on the other hand, at the SNP level, the results suggest at the possible existence of population-specific signatures of association which may be driven by differential *Mtb*-human co-evolution. This phenomenon of population-specific adaptation is not uncommon and has been previously reported in Ethiopian populations with regard to genetic and physiological adaptations to high altitude (hypoxia) (Beall, et al, 2002; Scheinfeldt, et al, 2012).

This study is only the second genetic epidemiological investigation of TB in Ethiopia but unique in its study design: sampling, control-phenotype ascertainment and accounting for possible population stratification. The current study validates the importance of carefully designed genetic epidemiological studies especially in populations where there is tremendous ethno-geographic variation and, thus, supposedly challenging for such investigations. It clearly demonstrates that even modest scale association studies conducted in populations with high genetic diversity can capture novel variants and reveal a broader spectrum of disease causality. It is usually advised in genetic epidemiology that case-control association studies should avoid sampling from heterogeneous populations. In light of this maxim one might question the prudence of sampling from populations with different ethno-geographic backgrounds as was done in this study. This manner of sampling was deliberately opted for in this study for several reasons. First, there are statistical methodologies to adjust or account for possible population stratification and in this study various techniques were applied to identify and control for known stratifications and empirically for hidden stratification. The result indicated that population stratification has minimal distortive effect on the statistical findings of this study. Second, although the existence of high genetic variation among Ethiopian populations has been well documented and there have been some studies which documented evidence of stratification along ethno-linguistic lines, these studies were mainly based on non-autosomal (X/Y/mitochondrial DNA) or limited whole-genome data. However, since the extent of genetic differentiation in exonic regions verses other regions has not been characterized there is no conclusive evidence yet to warrant the application of the "Avoid heterogeneous population sampling!" truism in candidate gene studies. Third, it was one of the explicit objectives of the study to characterize the Ethiopian population with respect to the selected candidate gene regions and, therefore, the collection of genomic samples

from as diverse populations as possible will enrich the overall Ethiopian genomic data. And, lastly, from the practical point of view, sampling from any single site and population would not only have taken much longer time but also there is no easy way of ascertaining the ethnic background of any target population prior to the actual sample collection and downstream comparative genetic analysis particularly in the absence of an Ethiopian database of ancestry-informative-markers. The study results show that although there were slight evidences suggestive of the presence of population-specific genetic signatures, there was no strong indication of population stratification that would lead to a recommendation of prohibition of sampling from different ethno-geographic backgrounds and pooling data together particularly for studies based on exonic data of candidate genes. Therefore, the results of this study should encourage investigators who may be intimidated by the 'burden' of the high genetic diversity in African populations.

Sample size is often the "Achilles' heel" of scientific research and is commonly cited as one of the major "limitations of the study". The relatively small sample size of this study is a limitation as it affects statistical inference and declaration of significance. Particularly, the progressively stricter definition of case-control phenotypes across test-model constructs comes at the cost of reduced sample size. This difference in sample size could explain in part the difference in the observed pattern of SNP-phenotype associations between the test-models. Lastly, there were quite a few subjects that tested HIV positive or had indeterminate Quantiferon results, which eliminated them from some analyses, thus reducing power. Therefore, future research with larger sample size is warranted. However, it is also worth while pointing out that merely garnering larger sample sizes does not always guarantee the detection of strong signals of association. A

good example is a recent genome-wide study on TB genetics (Thye, et al, 2010) that was done using an impressive case-control size (2,237 cases and 3,122 controls from Ghana and the Gambia) and declared to result in 90% power to detect a significant association. However, this study conspicuously failed to identify significant association from a total of 354,607 autosomal SNPs except for a single SNP rs4331426 which, unfortunately, also happened to land in a 'gene-desert' region on chromosome 18q11.2. Therefore, the current relatively much modest study demonstrates the usefulness of population-based, candidate-gene studies with careful case-control definitions in identifying informative disease-associated genetic variants in a cost-effective manner.

Finally, it would be appropriate to share in the optimism of Ernst JD (Ernst, et al, 2007) and conclude by paraphrasing: "......since *Mtb* was first discovered, there have been major advances in understanding the pathogenesis of TB in terms of host responses and how the bacteria exploits and evades them. With advances in basic immunology, genetics, epidemiology/statistics and their application to the problems of TB, better diagnostic tests, novel therapies, and efficacious vaccines are expected. Millions have died of TB, but the future holds great promise for making TB a disease of the past". It is hoped the current study contributes some knowledge in this regard. It is also hoped the design and conduct of this study, with particular emphasis on disease trait definitions that incorporate intermediate-phenotypes, will serve as a genetic epidemiological model to future studies of not only TB but other infectious and non-infectious complex disease traits. This study also underscores the power of candidate-gene based studies in the era of genome-wide association studies even in a setting with high genetic diversity. The conduct of the study, interpretation of the results and any limitations should be informative to the efforts of the

Human Heredity and Health in Africa initiative (H3A, 2014) at enriching region-specific health-related genomic knowledge.

XII. REFERENCES

Abate D, Tedla Y, Meressa D, Ameni G. "Isoniazid and rifampicin resistance mutations and their effect on second-line anti-tuberculosis treatment". *Int J Tuberc Lung Dis.* (2014); 18(8):946–951.

Abel L, Casanova JL. "Genetic Predisposition to Clinical Tuberculosis: Bridging the Gap between Simple and Complex Inheritance". *Am J Hum Genet.* (2000); 67: 274–277.

Agonafir M, Lemma E, Wolde-Meskel D, Goshu S, Santhanam A, Girmachew F, Demissie D, Getahun M, Gebeyehu M, Soolingen DV. "Phenotypic and genotypic analysis of multidrug-resistant tuberculosis in Ethiopia". *Int J Tuberc Lung Dis.* (2010); 14(10):1259–1265.

Akiibinu MO, Ogunyemi EO, Shoyebo EO. "Levels of Oxidative Metabolites, Antioxidants and Neopterin in Nigerian Pulmonary Tuberculosis Patients". *Eur J Gen Med.* (2011); 8(3):213-218.

Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. "Data quality control in genetic case-control association studies". *Nat Protoc.* (2010); 5:1564-1573.

Anderson EC, November J. "Finding haplotype Block Boundaries by using the Minimum-Description-Length Principle". *Am J Hum Genet.* (2003); 73(2):336-54.

Azad AK, Sadee W, Schlesinger LS. "Innate Immune Gene Polymorphisms in Tuberculosis". Edited by H. L. Andrews-Polymenis". *Infect Immun.* (2012); 80:10.

Barreiro LB, Patin E, Neyrolles O, Cann HM, Gicquel B, Quintana-Murci L. "The Heritage of Pathogen Pressures and Ancient Demography in the Human Innate-Immunity CD209/CD209L Region." *Am J Hum Genet.* (2005); 77(5): 869–886.

Beall CM, Decker MJ, Brittenham GM, Kushner I, Amha G, Strohl KP. "An Ethiopian pattern of human adaptation to high-altitude hypoxia". *PNAS.* (2002); 99(26):17215-17218.

Bellamy R, Beyers N, McAdam KP, Ruwende C, Gie R, Samaai P, Bester D, Meyer M, Corrah T, Collin M, Camidge DR, Wilkinson D, Hoal-Van Helden E, Whittle HC, Amos W, van Helden P, Hill AV. "Genetic susceptibility to tuberculosis in Africans: a genome-wide scan". *Pro Natl Acad Sci USA.* (2000); 97:8005–8009.

Bellamy R, Ruwende C, Corrah T, McAdam KP, Whittle HC, Hill AV. "Variations in the NRAMP1 gene and susceptibility to tuberculosis in West Africans". *N Engl J Med.* (1998); 338:640–644.

Bhimrao DP, Adinath NS, Archana SK. "Effect of micronutrients supplementation on oxidative stress and antioxidant status in pulmonary tuberculosis". *Biomed Res.* (2011); 22(4):455-459.

Bierne H, Hamon M, Cossart P. "Epigenetics and Bacterial Infections". *Cold Spring Harb Perspect Med.* (2012); 2(12).

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

Brites D, Gagneux S. "Co-evolution of Mycobacterium tuberculosis and *Homo sapiens*". *Immunol Rev.* (2015); 264: 6–24.

Bruchfeld J, Getachew A, Palme IB, Bjorvatn B, Solomon G, Hoffner S, Lindquist L. "Molecular Epidemiology of Drug Resistance of Mycobaterium tuberculosis Isolates from Ethiopian Pulmonary Tuberculosis Patients without Human Immunodeficiency Virus Infection". *J Clin Microbiol.* (2002); 40(5).

Caws M, Thwaites G, Dunstan S, Hawn TR, Lan NTN, Thuong NTT, Stepniewska K, Huyen MNT, Bang ND, Loc TH, Gagneux S, Van Soolingen D, Kremer k, Van der Sande M, Small P, Anh PTH, Chinh NT. "The Influence of Host and Bacterial Genotype on the Development of Disseminated Disease with Mycobacterium tuberculosis". *PLoS Pathog.* (2008); 4(3).

Chao MC, Rubin EJ. "Letting Sleeping dos Lie: Does Dormancy Play a Role in Tuberculosis?". *Annu. Rev. Microbiol.* (2010); 64:293–311.

Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, Parkhill J, Malla B, Berg S, Thwaites G, Yeboah-Manu D, Bothamley G, Mei J, Wei1 L, Bentley S, Harris SR, Nieman S. "Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans". *Nature Genetics.* (2013); 45(10).

Cooke GS, Campbell SJ, Bennett S, Lienhardt C, McAdam KPWJ, Sow O, Gustafson P, Mwangulu F, van Helden P, Fine P, Hoal EG, Hill AV. "Cooke GS, Campbell SJ, Bennett S, Lienhardt C, McAdam KMapping of a novel susceptibility locus suggests a role for MC3R and CTSZ in human tuberculosis". *Am J Respir Crit Care Med.* (2008); 178(2):203-207.

Coussens AK, Wilkinson RJ, Nikolayevskyy V, Elkington PT, Hanifa Yasmeen, Islam K. "Ethnic Variation in Inflammatory Profile in Tuberculosis". *PLoS Pathog.* (2013); 9(7).

Delgado JC, Baena A, Thim S, Goldfeld AE. "Ethnic-Specific Genetic Associations with Pulmonary Tuberculosis". *J Infect Dis.* (2002); 186(10):1463-1468.

Dolphin CT, Beckett DJ, Janmohamed A, Cullingford TE, Smith RL, Shephard EA. "The flavin-containing monooxygenase 2 gene (FMO2) of humans, but not of other primates, encodes a truncated, nonfunctional protein". *J Biol Chem.* (1988); 273:30599-30607.

Elson, G. "Contribution of Toll-like receptors to innate immune responses to Gram-negative and Gram-positive bacteria". *Am J Hematol.* (2007);109:(4).

Enokizonoa Y, Kumetaa H, Funamib K, Horiuchia M, Sarmientoc J, Yamashitac K, Standleyc DM, Matsumotob M, Seyab T, Inagakia F. "Structures and interface mapping of the TIR domaincontaining adaptor molecules involved ininterferon signaling". *PNAS* (2013);110(49).

Ernst JD, Trevejo-Nunez G, Banaiee N. "Genomics and the evolution, pathogenesis, and diagnosis of tuberculosis". *J Clin Invest.* (2007); 117: 1738–1745.

Juarez E, Carranza C, Hernandez-Sanchez F, Loyola E, Escobedo D, Leon-Contreras JC, Hernandez-Pando R, Torres M, Sada E. "Nucleotide-oligomerizing domain-1 (NOD1) receptor activation receptor activation induces pro-inflammatory responses and autophagy in human alveolar macrophages". *BMC Pulm Med.* (2014); 14(152).

Fantahun B, Sack U, Rodloff AC. "Multidrug-resistant tuberculosis in Ethiopia: efforts to expand diagnostic services, treatment and care". *Antimicrob Resist Infect Control.* (2014); 3(31).

Federal Ministry of Health, Ethiopia. "Tuberculosis, Leprosy and TB/HIV Prevention and Control Programme (Manual)". (2005).

Federal Ministry of Health, Ethiopia. "Tuberculosis Prevention and Control Program. A special Issue for TB Day, 24th March 2011". (2011).

Filliol I, Motiwala AS, cavatore M, Qi W, hazbon MH, Valle MB, Fyfe J, Garcia-Garcia L, Rastogi n, Sola C, Zozio T, Guerrero MI, Leon CI, Crabtree J, Angiuoli S, Eisenach KD, Durmaz R, Joloba ML, Rendon A, Sifuentes-Osorino J, Leon AP, Cave MD. "Global Phylogeny of Mycobacterium tuberculosis Based on Single Nucleotide Polymorphism (SNP) Analysis: Insights into Tuberculosis Evolution, Phylogenetic Accuracy of Other DNA Fingerprinting Systems, and Recommendations for a Minimal Standard SNP Set". *J Bacteriol.* (2006); 188: 759–772.

Firdessa R, Berg S, Elena H, Schelling E, Gumi B, Girum E, Endalamaw G, Teklu K, Meseret H, Jemal H, Zinsstag J, Robertson BD, Gobena A, Lohan AJ, Loftus B, Comas I. "Mycobacterial Lineages Causing Pulmonary and Extrapulmonary Tuberculosis, Ethiopia". *Emerg Infect Dis.* (2013); 19(3).

Franchi L, Park JH, Shaw MH, Marina-Garcia N, Chen G, Kim YG, Núñez G. "Intracellular NOD-like receptors in innate immunity, infection and disease". *Cell Microbiol.* (2008); 10(1).

Francois AA, Nishida CR, Ortiz de Montellano PR, Phillips IR, Shephard EA. "Human flavin-containing monooxygenase 2.1 catalyzes oxygenation of the antitubercular drugs thiacetazone and ethionamide". *Drug Metab Dispos.* (2009); 37:178-186.

Farmer R, Miller D, Lawrence R. "Epidemiology and Public Health Medicine". *Blackwell Science, Fourth Edition.* (1996).

Gagneux, S. "Host–pathogen coevolution in human tuberculosis". *Phil. Trans. R. Soc. B* (2012); 367(1590):850-859.

Galagan, JE. "Genomic insights into tuberculosis". *Nature Reviews/Genetics.* (2014); 15:307-317.

Gebrehiwot G, Daniel S, Getnet Y, Menon MKC. "The Non-Enzymatic Antioxidant and Level of Oxidative Stress of Tuberculosis Patients in Selected Treatment Center in Addis Ababa Ethiopia". *Journal of Tuberculosis Research.* (2015); 3:63-71.

Glassroth, J. "Tuberculosis 2004: Challenges and Opportunities". *Trans Am Clin Climatol Assoc.* (2005); 116: 293–310.

Gurdasani D, Carstensen T, Fasil TA, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, Ritchie GRS, Xue Y, Asimit J, Nsubuga RN, Young EH, Pomilla C, Kivinen K, Rockett K, Kamali A, Doumatey AP, Asiki G, Seeley J, Sisay-Joof F, Jallow M, Tollman S, Ephrem M, Ekong R, Tamiru O, Bradman N, Bojang K, Ramsay M, Adeyemo A, Endashaw B, Motala A, Norris SA, Pirie F, Kaleebu P, Kwiatkowski1 D, Tyler-Smith C, Rotimi C, Zeggini E, Sandhu MS. "The African Genome Variation Project shapes medical genetics in Africa". *Nature.* (2014); 517(7534):327.

Halff, EF. "Structural and functional studies on Nod Like Receptors: insights into NAIP/NLRC4 inflammasome formation". *Uitgeverij BOXPress,'s-Hertogenbosch.* (2013).

Hall NB, Igo Jr. RP, Malone LL, Truitt B, Schnell A, Tao L, Okware B, Nsereko M, Chervenak K, Lancioni C, Hawn TR, Mayanja-Kizza H, Joloba ML, Boom WH,

Stein CM. "Polymorphisms in TICAM2 and IL1B are associated with TB." *Gens Immun.* (2015); 16: 127–133.

Henderson MC, Siddens LK, Morre JT, Krueger SK, Williams DE. "Metabolism of the anti-tuberculosis drug ethionamide by mouse and human FMO1, FMO2 and FMO3 and mouse and human lung microsomes". *Toxicol Appl Pharmacol.* (2008); 233(3):420-427.

Hernandez D, Janmohamed A, Chandan P, Phillips IR, Shephard EA. "Organization and evolution of the flavin-containing monooxygenase genes of human and mouse: identification of novel gene and pseudogene clusters". *Pharmacogenetics.* (2004); 14:117-130.

Hill, AV. "Aspects of genetic susceptibility to human infectious diseases". *Annu Rev Genet.* (2006); 40:469–486.

Issar, S. "Mycobacterium tuberculosis Pathogenesis and Molecular Determinants of Virulence". *Clin Microbiol Rev.* (2003); 16(3): 463–496.

Kawai T, Akira S. "The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors". *Nat immunol.* (2010); 1(5).

Krishna RV, Mark GT, Michael EW, Zeitlyn D, Tarekegn A, Bekele E, Mendellh NR, Shephard EA, Bradman N, Phillips IR. "The potentially deleterious functional variant flavin-containing monooxygenase 2*1 is at high frequency throughout sub-Saharan Africa". *Pharmacogenet Genomics.* (2008); 18(10): 877-886.

H3Africa Consortium. "Research capacity: enabling the genomic revolution in Africa". *Science.* (2014); 344: 1346–1348.

Kreuger SK, Williams DE. "Mamalian flavin-containing monooxygenases: structure/function, genetic polymorphisms and role in drug metabolism". *Pharmacol Ther.* (2005); 106(3).

Krueger SK, Siddens LK, Martin SR, Yu Z, Pereira CB, Cabacungan ET, Hines R N, Ardlie KG, Raucy JL, Williams DE. "Differences in FMO2*1 allelic frequency between Hispanics of Puerto Rican and Mexican descent". *Drug Metab Dispos.* (2004); 32:1337-1340.

Krueger SK, Siddens LK, Henderson MC, Andreasen EA, Tanguay RL, Pereira CB, Cabacungan ET, Hines RN, Ardlie KG, Williams DE. "Haplotype and functional analysis of four flavin-containing monooxygenase isoform 2 (FMO2) polymorphisms in Hispanics." *Pharmacogenet Genomics.* (2005); 15(4): 245–256.

Lee JY, Hwang EH, Kim DJ, Shin SJ, Park JH. "The role of nucleotide-binding oligomerization domain 1 (NOD1) in cytokine production by macrophages in response to Mycobacterium tuberculosis". *Conference: 2nd Annual Meeting of the International-Cytokine-and-Interferon-Society, Melbourne, Australia* (2014).

Li CM, Campbell SJ, Kumararatne DS, Bellamy R, Ruwende C, McAdam KP, Hill AV, Lammas DA. "Association of a polymorphism in the P2X7 gene with tuberculosis in a Gambian population." *J Infect Dis.* (2002); 186:1458–1462.

Liu N, Zhang K, Zhao H. "Haplotype-Association Analysis". *Adv Genet.* (2008); 60:335-405.

Lombard Z, Dalton DL, Venter PA, Williams RC, Bornman L. "Association of HLA-DR, -DQ, and vitamin D receptor alleles and haplotypes with tuberculosis in the Venda of South Africa". *Hum Immunol.* (2006); 67:643–654.

Ma N, Zalwago S, Malone LL, Nsereko M, Wampande EM, Thiel BA, Okware B, Igo RP, Joloba ML, Mupere E, Mayanja-Kizza H, Boom WH, Stein CM. "Clinical and epidemiological characteristics of individuals resistant to M. tuberculosis infection in a longitudinal TB Household contact study in Kampala, Uganda". *BMC Infect Dis.* (2014); 14(352).

Malik S., Greenwood CMT., Eguale T., Kifle A., Beyene J., Habte A., Tadesse A., Gebrexabher H., Britton S., Schurr E. "Variants of the SFTPA1 and SFTPA2 genes and susceptibility to tuberculosis in Ethiopia". *J Hum Genet.* (2006); 118(6): 752-759.

Marilyn C. Henderson, Lisbeth K. Siddens, Jeffrey T. Morre, Sharon k. Krueger, and David E. Williams. "Metabolism of the Anti-Tuberculosis Drug Ethionamide by Mouse and Human FMO1, FMO2 and FMO3 and Mouse and Lung Microsomes." *Toxicol Appl Pharmacol.* (2008); 233(3):420-427.

Matsumiya M, Stylianou E, Grifiths K, Lang Z, Meyer J, Harris SA, Rowland R, Minassian AM, Pathan AA, Fletcher H, McShane H. "Roles for Treg Expansion and HMGB1 Signaling through the TLR1-2-6 Axis in Determining the Magnitude of the Antigen-Specific Immune Response to MVA85A." *PLoS One.* (2013); 8(7):e67922.

Mengistu MM, Tesfaye WT, Madeley JR. "The quality of tuberculosis diagnosis in districts of Tigray region of northern Ethiopia." *Ethiop J Health Dev.* (2005); 19:13-20.

Moller M, Hoal EG. "Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis." *Tuberculosis(Edinb).* (2010); 90:71–83.

Moreira LM, Zamboni DS. "NOD1 and NOD2 Signaling in Infection and Inflamation." *Front Immunol.* (2012); 3(28).

Moses AO, Emmanuel OO, Ganiyu AO, Fidelis AA, Dickson AO. "Assessment of antioxidants and nutritional status of pulmonary tuberculosis patients in Nigeria." *Eur J Gen Med.* (2008); 5(4):208-211.

Mulugeta B, Gobena A, Bjune G, Couvin D, Rastogi N, Fekadu A. "Strain Diversity of Mycobacterium tuberculosis Isolates from Pulmonary Tuberculosis Patients in Afar Pastoral Region of Ethiopia." *BioMed Res Int.* (2014); 2014(238532).

Nahid P, Gonzalez LC, Rudoy I, Jong BD, Unger A, Kawamura LM, Osmond DH, Hopewell PC, Daley CL. "Treatment Outcomes of Patients with HIV and Tuberculosis." *Am J Respir Crit Care Med.* (2007); 175(11).

Olesen R, Wejse C, Velez DR, Bisseye C, Sodemann M, Aaby P, Rabna P, Worwui A, Chapman H, Diatta M, Adegbola RA, Hill PC, Ostergaard L, Williams SM, Sirugo G. "DC-SIGN (CD209), pentraxin 3 and vitamin D receptor gene variants associate with pulmonary tuberculosis risk in West Africans." *Genes Immun.* (2007). 8:456–467.

Oyedeji SO, Adesina AA, Oke OT, Oguntuase NR, Esan A. "Oxidative Stress and Lipid Profile Status in Pulmonary Tuberculosis Patients in South Western Nigeria". *J Med Sci.* (2013); 3:228-232.

Pagani L, Schiffels S, Gurdasani D, Danecek P, Scally A, Chen Y, Xue Y, Harber M, Ekong R, Tamiru O, Ephrem M, Luiselli D, Bradman N, Endashaw B, Zalloua P, Durbin R, Kivisi T. "Tracing the Route of Modern Humans out of Africa by

Using 225 HUman Genome Sequences From Ethiopians and Egyptians." *A J Hum Genet.* (2015); 96:1-5.

Pagani P, Kivisild T, Ayele T, Ekong R, Plaster C, Romero IG, Ayub Q, Mehdi SQ, Thomas MG, Luiselli D, Endashaw B, Bradman N, Balding DJ, Tyler-Smith C. "Ethiopian Genetic Diversity Reveals Linguistic Stratification and Complex Influences on the Ethiopian Gene Pool". *A J Hum Genet.* (2012); 91:83–96.

Pai M, Behr MA, Dowdy D, Dheda K, Divangahi M, Boehme CC, Ginsberg A, Swaminathan S, Spigelman M, Haileyesus G, Menzies D, Raviglione M. "Tuberculosis". *Nat Rev/Dis Primers.* (2016); 2(1).

Palmer AL, Leykam VL, Larkin A, Krueger SL, Phillips IR, Shephard EA. "Metabolism and Pharmaccokinetics of the Anti-Tuberculosis Drug Ethionamide in a Flavin-Contianing Monooxygenase Null Mouse". *Pharmaceuticals.* (2012); 5(11):1147-1159.

Pawar BD, Suryakar AN, Khandelwal AS. "Effect of micronutrients supplementation on oxidative stress and antioxidant status in pulmonary tuberculosis." *Biomed Res.* (2011); 22 (4): 455-459.

Philips JA, Ernst JD. "Tuberculosis Pathogenesis and Immunity". *Annu. Rev. Pathol. Mech. Dis.* (2012); 7:353–84.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC. "PLINK (1.07) Documentation: a toolset for whole-genome association and population-based linkage analysis". (2007); URL: http://pngu.mgh.harvard.edu/purcell/plink/.

Quintana-Murci L, Alcaïs A, Abel L, Casanova JL. "Immunology in natural selection: clinical, epidemiological and evolutionary genetics of infectious diseases". *Nat Immuno.* (2007); l 8:1165–1171.

Ramachandran G, Swaminathan S. "Role of pharmacogenomics in the treatment of tuberculosis: a review". *Pharmacogenome Pers Med.* (2012); 5:89-98.

Rice, J. "Definition of the Phenotype." *Adv Genet.* (2001).

Scheinfeldt LB, Soi S, Thompson S, Ranciaro A, Dawit W, Beggs W, Lambert C, Jarvis JP, Dawit A, Gurja B, Tishkoff SA. "Genetic adaptation to high altitude in the Ethiopian highlands". *Genome Biol.* (2012); 13(1).

Seya T, Oshiumi H, Sasai M, Akazawa T, Matusumoto M. "TICAM-1 and TICAM-2: toll-like receptor adapters that participate in induction of type 1 interferons". *Int J Biochem Cell Biol.* (2005); 37(3).

Siddens LK, Henderson MC, VanDyke JE, Williams DE, Krueger SK. "Characterization of mouse flavin-containing monooxygenase transcript levels in lung and liver, and activity of expressed isoforms". *Biochem Pharmacol.* (2008); 75(2).

Sirugo G, Hennig BJ, Adeyemo AA, Matimba A, Newport MJ, Ibrahim ME, Ryckman KK, Tacconelli A, Mariani-Costantini R, Novelli G, Soodyall H, Rotimi CN, Ramesar RS, Tishkov SA, Williams SM. "Genetic studies of African populations: an overview on disease susceptibility and response to vaccines and therapeutics". *Hum Genet.* (2008); 123:557–598.

Stein CM, Nshuti L, Chiunda AB, Boom WH, Elston RC, Mugerwa RD, Iyengar SK, Whalen CC. "Evidence for a Major Gene Infl uence on Tumor Necrosis Factor-

Expression in Tuberculosis: Path and Segregation Analysis". *Hum Hered.* (2005); 60(2):109-18.

Stein CM, Guwatudde D, Nakakeeto M, Peters P, Elston RC, Tiwari HK, Mugerwa R, Whalen CC. "Heritability analysis of cytokines as intermediaate phenotypes of tuberculosis." *J Infect Dis.* (2003); 187: 1679-1685.

Stein CM. "Epidemiology of Tuberculosis Susceptibility: Impact of Study Design". *PLoS Pathog.* (2011); 7(1).

Stein CM, Elston RC. "Finding genes underlying human disease". *Clin Genet.* (2009); 75(2).

Stein CM, Zalwango S, Chiunda AB, Millard C, Leontiev DV, Horvath AL, Cartier KC, Chervenak K, Boom WH, Elston RC, Mugerwa RD, Whalen CC, Iyengar SK. "Linkage and association analysis of candidate genes for TB and TNFalpha cytokine expression: evidence for association with IFNGR 1, IL-10, and TNF receptor 1 genes". *Hum Genet.* (2007); 121:663-673.

Steingart KR, Henry M, Laal S, Hopewell PC, Ramsay A, Menzies D, Cunningham J, Weldingh KPai M. " Commercial Serological Antibody Detection Tests for the Diagnosis of Pulmonary Tuberculosis: A Systematic Review". *PLoS Med.* (2007); 4(6).

Strachan T, Read A. "Human Molecular Genetics". *Garland Science, Taylor and Francis Group, LLC.* (2011).

Tada H, Aiba S, Shibata KI, Ohteki T, Takada H. "Synergistic Effect of Nod1 and Nod2 Agonists with Toll-Like Receptor Agonists on Human Dendritic Cells To Generate Interleukin-12 and T Helper Type 1 Cells". *Inf Immun.* (2005); 73(12).

Takeda K, Akir S. "Toll-like receptors in innate immunity". *J Soc Immunol.* (2005); 17(1).

Tegbaru B, Wolday D, Messele T, Legesse M, Mekonnen Y, Miedema F, Van Baarle D. "Tuberculin Skin Test Conversion and Reactivity Rates among Adults with and without Human Immunodeficiency Virus in Urban Settings in Ethiopia". *Clin Vaccine Immunol.* (2006); 13: 784–789.

Thomas, DC. "Statistical Methods in Genetic Epidemiology". *Oxf Univ Press.* (2004).

Thye T, Vannberg FO, Wong SH, Owusu-Dabo E, Osei I, Gyapong J, Sirugo G, Sisay-Joof F, Enimil A, Chinbuah MA, Floyd S, Warndorff DK, Sichali L, Malema S, Crampin AC, Ngwira B. "Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2". *Nat Genet.* (2010); 42(9).

Toossi Z, Mayanja-Kizza H, Hirsch CS, Edmonds KL, Spahlinger T, Hom DL, Aung H, Mugyenyi P, Ellner JJ, Whalen CW. "Impact of tuberculosis (TB) on HIV-1 activity in dually infected patients". *Clin Exp Immunol.* (2001); 123: 233–238.

Van Crevel R, Ottenhoff THM, Van der Meer JWM. "Innate Immunity to Mycobacterium tuberculosis". *Clin Microbiol Rev.* (2002); 15: 294–309.

Veeramah KR, Thomas MG, Weale ME, Zeitlyn D, Ayele T, Endashaw B, Mendell NR, Shephard EA, Bradman N, Phillips IR. "The potentially deleteriousfunctional variant flavin-containing monooxygenase 2*1 is at high frequency throughout sub-Saharan Africa". *Pharmacogenet genom.* (2008); 18:877-886.

Verma I, Jindal SK, Ganguly NK. "Studies in Respiratory Disorders: Oxidative Stress in Tuberculosis". *Springer.* (2014); pp. 101-114.

Vermund S, Yamamoto N. "Co-infection with human immunodeficiency virus and tuberculosis in Asia". *Tuberculosis.* (2007); 87: 18–25.

Warner DF, Mizrahi V. "Translating genomics research into control of tuberculosis: lessons learned and future prospects." *Genome Biol.* (2014); 15(514).

Weiss, KM. "Genetic Variation and Human Disease". *Camb. Univ. Press.* (1993).

Whetstine JR, Yueh MF, Hopp KA, McCarver DJ, Williams DE, Park CS, Kang JH, Cha YN, Dolphin CT, Shephard EA, Phillips IR, Hines RN. "Ethnic Differences in Human Flavin-Containing Monooxygenase 2 (FMO2) Polymorphisms: Detection of Expressed Protein in African-Americans". *Toxicol Appl Pharmacol.* (2000);168:216-224.

WHO. "World Health Organization Global TB Report". (2016)

WHO. "Multidrug and extensively drug-resistant TB (M/XDR-TB): 2010 Global Report on Surveillance and Response". (2010).

Yadav V, Dwivedi VP, Bhattacharya D, Mittal A, Moodley P, Das G. "Understanding the Host Epigenetics in Mycobacterium tuberculosis Infection". *J Genet Genome Res.* (2015); 2(1).

Yimer SA, Norheim G, Namouchi A, Ephrem DZ, Kinander W, Tønjum T, Shiferaw B, Mannsåker T, Bjune G, Abraham A, Holm-Hansenb C. "Mycobacterium tuberculosis Lineage 7 Strains Are Associated with Prolonged Patient Delay in Seeking Treatment for Pulmonary Tuberculosis in Amhara Region, Ethiopia". *J Clin Microbiol.* (2015); 53(4).

Zhang Y. "The Magic Bullets and Tuberculosis Drug Targets". *Annu Rev Pharmacol Toxicol.* (2005);45:529–64.

Zhang J, Cashman JR. "Quantitative Analysis of FMO Gene mRNA Levels in Human Tissues". *Drug Metabol Dispos.* (2006); 34(1):19-26.

XIII. Appendix-1: Supplementary tables

Supplementary Table-1: Summary of latent TB infection test result based on interferon-ɣ release assay

Summary of LTBI test result							
Sampling site	Indeterminate	%	Negative	%	Positive	%	Total
Adigrat	0	0.000	26	0.481	28	0.519	54
Arbaminch	7	0.106	21	0.318	38	0.576	66
Merhabete	1	0.021	28	0.583	19	0.396	48
Total	8	0.048	75	0.446	85	0.506	168

Supplementary Table-2: Summary of HIV serostatus test result

	Cases					Controls				
Sampling site	Negative	%	Positive	%	Total	Negative	%	Positive	%	Total
Adigrat	53	0.791	14	0.209	67	61	0.968	2	0.032	63
Arbaminch	90	0.909	9	0.091	99	88	1.000	0	0	88
Merhabete	34	0.680	16	0.320	50	46	0.902	5	0.098	51
Total	177	0.819	39	0.181	216	195	0.965	7	0.035	202

Supplementary Table-3: DNA sequence data in each EGC before quality control filtering

Raw Sample Description																				
Ethno-Geographic Category (Sampling Site)	n	Cases	Controls	Males	Females	Per candidate Gene														
						Total				FMO2				TICAM2				NOD1			
						#SNPs	GR	MM	%MM	#SNPs	GR	MM	%MM	#SNPs	GR	MM	%MM	#SNPs	GR	MM	%MM
Merhabete	95	49	46	54	41	19514	0.92	18842	0.966	7596	0.97	7404	0.975	3575	0.84	3475	0.972	8343	0.9	7963	0.954
Adigrat	102	51	51	55	47	19516	0.95	18963	0.972	7595	0.96	7493	0.987	3575	0.87	3475	0.972	8346	0.93	7995	0.958
Arbaminch	150	94	56	126	24	19518	0.94	18748	0.961	7597	0.99	7453	0.981	3576	0.86	3452	0.965	8345	0.93	7843	0.940
combined	347	194	153	235	112	19530	0.93	17887	0.916	7600	0.99	7284	0.958	3578	0.86	3326	0.930	8352	0.62	7277	0.871

Supplementary Table-4: DNA sequence data in each test-model before quality control filtering

Quality Control																				
Test-model	n	Cases	Controls	Males	Females	Before QC														
						Total				FMO2				TICAM2				NOD1			
						#SNPs	GR	MM	%MM	#SNPs	GR	MM	%MM	#SNPs	GR	MM	%MM	#SNPs	GR	MM	%MM
Active TB vs. No Active TB	292	153	139	206	86	19527	0.94	18023	0.923	7600	0.99	7294	0.960	3578	0.86	3343	0.934	8349	0.92	7386	0.885
Active TB vs.No LTBI	217	153	64	160	57	19525	0.94	18409	0.943	7599	0.99	7384	0.972	3578	0.86	3406	0.952	8348	0.92	7619	0.913
Active TB vs. LTBI	223	153	70	158	65	19523	0.94	18287	0.937	7598	0.99	7337	0.966	3577	0.86	3374	0.943	8348	0.92	7576	0.908
LTBI vs. No LTBI	134	70	64	88	46	19514	0.94	18639	0.955	7596	0.99	7377	0.971	3575	0.86	3432	0.960	8343	0.92	7830	0.939
Average						19522	0.94	18340	0.939	7598	0.99	7348	0.967	3577	0.86	3389	0.947	8347	0.92	7602.75	0.9108

Supplementary Table-5: DNA sequence data in each EGC after quality control filtering

Quality Control								
Test-model	After QC							
	Total		FMO2		TICAM2		NOD1	
	#SNPs	GR	#SNPs	GR	#SNPs	GR	#SNPs	GR
Active TB vs. No Active TB	94	0.99	58	1.000	12	1	24	0.99
Active TB vs.No LTBI	91	0.99	58	1.000	12	1	21	0.99
Active TB vs. LTBI	92	0.99	56	1.000	12	1	24	0.99
LTBI vs. No LTBI	98	0.99	62	1.000	13	1	23	0.99
Average	94	0.99	59	1.000	12	1	23	0.99

Supplementary Table-6: Minor allele frequency of TB-phenotype-associated SNPs

Minor allele frequencies of phenotype-associated SNPs in the different test-model datasets											
				Active TB vs No Active TB		Active TB vs. No LTBI		Active TB vs. LTBI		LTBI vs. No LTBI	
Gene	SNP	A1	A2	MAF	NCHROBS	MAF	NCHROBS	MAF	NCHROBS	MAF	NCHROBS
FMO2	chr1:171154303	C	A	0.1986	584	0.1912	434	0.1973	446	0.2201	268
	chr1:171165749	T	G	0.06678	584	0.08065	434	0.08296	446	0.01493	268
	chr1:171168469	A	C	0.02397	584	0.02995	434	0.02691	446	0.01119	268
	chr1:171168545	C	T	0.0411	584	0.03687	434	0.0426	446	0.04851	268
	chr1:171173242	C	T	0.1729	584	0.1567	434	0.1592	446	0.1978	268
	chr1:171174312	A	T	0.1182	584	0.129	434	0.1278	446	0.09328	268
	chr1:171174691	A	G	0.1695	584	0.1544	434	0.1547	446	0.194	268
	chr1:171174762	C	G	0.05822	584	0.053	434	0.04709	446	0.08209	268
	chr1:171174821	A	G	0.1695	584	0.1544	434	0.1547	446	0.194	268
	chr1:171176879	A	G	0.1695	584	0.1544	434	0.1547	446	0.194	268
	chr1:171177858	T	G	0.08219	584	0.07143	434	0.07175	446	0.1007	268
	chr1:171178090	C	T	0.1353	584	0.1221	434	0.1166	446	0.1604	268
	chr1:171178490	T	C	0.1182	584	0.1267	434	0.13	446	0.09328	268
	chr1:171179025	C	T	0.1353	584	0.1221	434	0.1166	446	0.1604	268
	chr1:171179287	T	C	0.1849	584	0.1843	434	0.1883	446	0.1866	268
	chr1:171179477	T	C	0.3202	584	0.3065	434	0.3049	446	0.347	268
	chr1:171179670	G	C	0.1524	584	0.1544	434	0.1502	446	0.1567	268
	chr1:171179779	G	A	0.3305	584	0.3088	434	0.3184	446	0.3657	268
	chr1:171179939	G	T	0.488	584	0.4931	434	0.4978	446	0.4552	268
	chr1:171180021	G	T	0.488	584	0.4931	434	0.4978	446	0.4552	268
	chr1:171180071	G	A	0.3305	584	0.3088	434	0.3184	446	0.3657	268
	chr1:171180201	C	T	0.3305	584	0.3111	434	0.3161	446	0.3657	268
	chr1:171181150	A	G	0.02397	584	0.02995	434	0.02691	446	0.01119	268
	chr1:171181877	A	C	0.1113	584	0.129	434	0.1368	446	0.04104	268
TICAM2	chr5:114915999	A	C	0.01541	584	0.01843	434	0.02018	446	0	0
	chr5:114916028	A	G	0.1216	584	0.1198	434	0.1031	446	0.1567	268
	chr5:114916090	G	A	0.4418	584	0.4447	434	0.4574	446	0.4216	268
NOD1	chr7:30464249	TG	T	0.01199	584	0	0	0.01121	446	0.01866	268
	chr7:30464872	A	G	0.04281	584	0.05069	434	0.04036	446	0.02985	268
	chr7:30464932	G	T	0.0137	584	0.01843	434	0	0	0.01866	268
	chr7:30465424	C	A	0.01199	584	0	0	0.01121	446	0.01866	268
	chr7:30469270	C	T	0.2089	584	0.1982	434	0.2265	446	0.1978	268
	chr7:30477156	T	G	0.0839	584	0.1014	434	0.09641	446	0.04104	268
	chr7:30485722	T	G	0.07877	584	0.08986	434	0.09641	446	0.02985	268
	chr7:30490711	T	G	0.06507	584	0.07604	434	0.07623	446	0.03358	268
	chr7:30491081	A	G	0.02759	580	0.03488	430	0.03167	442	0.01119	268
	chr7:30498962	C	G	0.01884	584	0.01843	434	0.01345	446	0.02985	268

XIV. Appendex-2: Research participant consent form (English version)

A PhD dissertation project, titled *"A Study of Genetic Influences on Tuberculosis Susceptibility and Progression in Ethiopia"* is planned to be undertaken by Ato Ephrem Mekonnen who is a PhD student at Addis Ababa University, Dept. of Biology. The main objective of the study is to analyze the possible impact of human genetic variation pertaining to TB-related immune response pathways. The research design is a population-based case-control study to be conducted in various parts of Ethiopia. Active PTB patients [cases] and their healthy matches (spouses/partners/household contacts) [controls] will be voluntarily recruited only after obtaining their informed consent. All phenotyping procedures [TB-related medical check-ups and drawing of 10 ml venous blood sample] will be carried out by qualified health personnel with strict adherence to safety and privacy regulations. All participants found to be positive for LTBI and/or HIV will be offered appropriate counseling and medication free of charge. Participants are free to withdraw from the study anytime they want without fear of deprivation of treatment benefits. Furthermore, all research participants' data will be held strictly confidential. Only those individuals who understand and sign this consent form will be recruited for participation in the study.

I fully understand the above message and agree to participate in the study.

_____ _____

Name *Signature/Date*