**Distribution Agreement**

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

______________________________ _______________

Matheus Fernandes Gyorfy Date

# A Genome-Wide Association Study of Resistance to Tuberculosis Infection in a Multi-Ancestry Brazilian Cohort

By

Matheus Fernandes Gyorfy
Degree to be awarded: Master of Public Health

Department of Epidemiology

Dr. Yan Sun
Committee Chair

# A Genome-Wide Association Study of Resistance to Tuberculosis Infection in a Multi-Ancestry Brazilian Cohort

By

Matheus Fernandes Gyorfy

B.S., Colorado State University, 2019
B.A, Colorado State University, 2019

Thesis Committee Chair: Dr. Yan Sun, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2023

# Abstract

A Genome-Wide Association Study of Resistance to Tuberculosis Infection in a Multi-Ancestry Brazilian Cohort

By Matheus Fernandes Gyorfy

**Background:** Tuberculosis (TB) impacts over a quarter of world's population. Although its global incidence rate has been steadily decreasing due to advancements in testing capabilities and novel drug regimens, the relationship between host genetic and molecular factors and infectious pathogen remains largely underexplored. Studies assessing resistance to *Mycobacterium tuberculosis* (*Mtb*) infection have been limited by a lack of consistent classification of infectious and exposure levels. **Methods:** Among household contacts of active TB patients, we used detailed measurements of exposure levels and TB infection to identify most likely resisters to *Mtb* infection. Using imputed single nucleotide polymorphisms (SNP) data from TOPMed imputation panel, we conducted a genome-wide association study (GWAS) of resistance to *Mtb* infection in 1,540 multi-ancestry Brazilian participants. **Results:** A total of 232 (15.1%) individuals whose Tuberculin Skin Test (TST) or Interferon-Gamma Release Assay (IGRA) results were negative and who experienced the highest-level exposure were categorized as resisters. This analysis identified SNPs significantly associated with resistance to *Mtb* infection, from four loci close to genes *PARD3B* (rs888091, OR = 2.93 [95%CI: 2.56, 3.30]; p = $9.99\times10^{-9}$]), *AC073987.1* (rs117179998, OR = 2.81 [95%CI: 2.46, 3.16]; p = $1.02\times10^{-8}$), *IQCA1* (rs35136956, OR = 2.10 [95%CI: 1.81, 2.39]; p = $3.88\times10^{-8}$) in chromosome 2, and *COL18A1* (rs80327334, OR = 2.15 [95%CI: 1.88, 2.42]; p = $4.69\times10^{-8}$) in chromosome 21. However, we observed substantial inflation of low p-values (inflation factor of 1.17) which can be caused by relatedness among household contacts. **Conclusion:** Our findings demonstrated the role of human genetic factors in the resistance to *Mtb* infection. In the future, we will address the global inflation by adjustment of relatedness of study participants. To further validate our results, we will conduct replication and meta-analysis using similar household contact cohorts from India and South Africa.

# A Genome-Wide Association Study of Resistance to Tuberculosis Infection in a Multi-Ancestry Brazilian Cohort

By

Matheus Fernandes Gyorfy

B.S., Colorado State University, 2019
B.A, Colorado State University, 2019

Thesis Committee Chair: Dr. Yan Sun, PhD

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Public Health
in Epidemiology
2023

## Acknowledgements

I would like to thank my mentor and thesis advisor Dr. Yan Sun for all of his guidance and mentorship throughout this process. I am beyond lucky to have had the opportunity to collaborate with him, and I look forward to continuing to build upon this work with future research projects in the many years to come. I would also like to thank all of the members in Dr. Sun’s lab for their kindness and willingness to help troubleshooting and orienting the analyses in the right direction. Lastly, I would like to thank my partner and family for keeping me sane and being active supporters of my life goals.

## Table of Contents

INTRODUCTION ........ 1

METHODS ........ 5

STUDY DESIGN AND POPULATION PHENOTYPE ........ 5
SAMPLES GENOTYPING, QUALITY CONTROL, AND IMPUTATION ........ 6
STATISTICAL, GRAPHICAL, AND ANNOTATIONS METHODS ........ 6
ETHICAL CONSIDERATIONS ........ 7

RESULTS ........ 8

PRIMARY ANALYSIS ........ 9
PRELIMINARY ANALYSIS ........ 9
PREVIOUSLY PUBLISHED LOCI ........ 10

DISCUSSION ........ 11

LIMITATIONS AND FUTURE WORK ........ 16

REFERENCES: ........ 18

TABLES AND FIGURES ........ 21

# Introduction

Tuberculosis (TB) is an airborne chronic infectious disease caused by the bacterium *Mycobacterium tuberculosis* (*Mtb*) and is the 13th leading cause of death in the world[1]. Although TB primarily affects the lungs, it can also impact the brain, the kidney, and the spine[2–4]. When infected individuals cough or sneeze, *Mtb* is spread through the air via droplets. Once these droplets are inhaled, *Mtb* is taken up by alveolar macrophages in the lungs, where it can replicate and cause disease[5]. However, not all individuals who are exposed to *Mtb* will develop active TB disease. According to current estimates, around 2 billion individuals are infected with *Mtb* worldwide, but the distribution of disease burden is heavily unbalanced on a global scale[1]. In 2021, eight countries in Africa and Asia constituted over 66% of global TB cases, and those same continents accounted for more than 80% of global TB deaths[1]. South America is yet another continent where TB incidence is considerably high; the World Health Organization has designated Brazil as a top 30 TB high burden country due to its TB incidence of 96,000 (the highest in the continent) which is equivalent to a rate of 45 cases per 100,000 individuals.[1] Even though this is the lowest rate among the assigned top 30 countries, it is still far from Brazil's health ministry's goal rate of 10 cases per 100,000[6] and from the current United States rate of 2.4 cases per 100,000[1]. There are many layers of complexity that serve as obstacles to the eradication efforts implemented. One barrier to fighting TB is the epidemiologic distinction between active TB disease and latent TB infection (LTBI) where the virus is present in a person's organism but remains dormant within granulomas found in the lungs[7]. Although 25% of the world's population is infected with *Mtb* (i.e., LTBI), but only around 10.6 million individuals have active TB[1]. This means that the great majority of TB positive individuals have LTBI and therefore cannot infect others, but could progress from LTBI into active TB due to various factors usually pertaining to weakened immune systems like co-infections with HIV, malnutrition, or genetic and epigenetic factors[8–10]. TB can be spread through the inhalation of *Mtb*-ridden droplets in the air, but the infection response varies immensely across individuals in terms of susceptibility and resistance. For centuries, there have been reports and studies observing people who demonstrate resistance to TB infection despite being highly and frequently exposed[9]. From a clinical perspective, this remarkable characteristic is critical to enhance our ability to treat and prevent *Mtb* infections, however, the molecular mechanisms behind this response require further investigations. The current knowledge behind human host response to mycobacterial infections indicates a deep involvement of the activity of interferon-gamma (IFN-$\gamma$), IL-6, IL-10, IL-12, and IL-23 during a type I cytokine response[11], which brings to light the plausibility of how human host genetics may influence an individual's capacity of being resistant to *Mtb* infection[12].

Host genomics plays a role in susceptibility and resistance to *Mtb* infection and progression to active TB, although the precise genetic factors and pathways that are involved in this process are still being elucidated. Previous genome-wide association studies have investigated the progression from LTBI to active TB (i.e., susceptibility). Genome-wide association studies (GWAS) are a powerful tool for investigating genetic variants and their relationship with various disease traits by analyzing millions of single nucleotide polymorphisms (SNP) across the genome of large numbers of individuals. This study design yields statistical associations between individuals' genotype and phenotype which may give insight into the biological pathways behind traits like disease resistance or susceptibility[13]. Importantly, to better investigate genetic architecture and improve the generalizability of study findings, it is critical that multiple ancestries are included within the study population[14]. For example, a GWAS of 833 TB cases and 1220 controls in a Han Chinese population discovered two risk loci (rs12437118 and rs6114027, corresponding to genes ESRRB and TGM6, respectively) significantly associated with susceptibility to TB. In this study, active TB cases were confirmed by sputum culture for *Mtb*, presence of acid-fast bacilli in sputum smear, and clinical presentation and radiological signs, but the degree to which individuals were exposed was never measured[15]. Another GWAS of active TB in Ghanaian and Gambian populations identified multiple loci associated with susceptibility to TB infection, although only one SNP presented statistically significant associations following replication studies (rs4331426). The active TB cases were identified by medical documentation, physical examination, sputum smears and culture of Mtb[16]. While these studies provide insight into the genetic mechanisms that may play a role in TB susceptibility, the role of host genetics as it pertains to resistance to *Mtb* infection remains largely underexplored[17–19]. A 2021 study discovered a locus at 10q26.2 (variant rs17155120) significantly associated with resistance to *Mtb* infection in a Southern Vietnamese cohort. The genetic association was replicated in a French multi-ethnic population and a South African cohort[20]. While these studies defined their TB resisters based on tuberculin skin test (TST) or Interferon Gamma Release Assay (IGRA) results, another study approached this issue by defining resistance based on the intensity of tuberculin reactivity in a Ugandan cohort. Although loci in chromosomal regions 2q21-2q24 and 5p13-5q22 demonstrated the strongest association with resistance to *Mtb* infection, they were not statistically significant after multiple testing correction[21,22]. The observed inconsistency in definitions of resisters has proven to be a major obstacle in connecting findings across different studies. Nevertheless, the cumulative evidence from these studies demonstrates the plausibility for human genetic factors that play a role in preventing some individuals from ever becoming infected with *Mtb*. In order to identify a robust genetic association across global populations with high TB burden, more and larger studies must be conducted where resisters are defined in a consistent manner and where cohorts include the diverse genetic ancestry.

In our study, we conducted a GWAS among 1,540 household-contacts of confirmed TB cases in a multi-ancestry Brazilian cohort. We performed two separate analyses with the goal of exploring how different definitions of the population resisters may impact the discovery of genetic variants associated with resistance to *Mtb* infection. Although both analyses followed the same model of Resistance Category ~ SNP + Age + Sex + PC1 through PC10, we used a stringent definition of the resistance category for the primary analysis. The strength of our study design is a product of the comprehensive definitions of TB resisters considering the exposure levels, measurement of *Mtb* infection, and infectivity of *Mtb* strains, and the diverse genetic ancestry of the study population.

# Methods

## Study Design and Population Phenotype

This study was a genome-wide association study (GWAS) that aimed to identify genetic variants associated with resistance to TB infection in a multi-racial population. Prior to data quality control exclusions, the study population consisted of 1,563 Brazilian individuals from various ethnic backgrounds. Inclusion criteria for the study were children or adults who were household contacts of confirmed active TB cases and were willing to provide informed consent. Exclusion criteria included individuals who were immunocompromised due to HIV infection or other conditions. Eight phenotypic categories representing a resistance hierarchy were created based on exposure level and TST/IGRA test results (**Table 1**).

For the primary analysis, we employed a strict definition of resistance where only those with the strongest evidence of negative IGRA or TST results (negative results for one or both tests in most recent assessment) and the highest level of exposure (contact reported sharing a bed or room with index case or spending 5 or more hours indoor with them) were classified as resisters. This is analogous to the resistance hierarchy of Resisters A through C being categorized as resisters, and Resisters D and E plus Unknown A and Infected A and B being categorized as infected. As for the preliminary analysis, where we used an alternative and less strict definition of resistance, resisters were categorized to be any individual with a negative TST or IGRA test (including discordant results) and with any level of exposure to TB. According to the resistance hierarchy, this is equivalent to the categories of Resisters A through E plus Unknown A all being defined as resisters, and categories Infected A and B were defined to be the infected individuals in the study.

## Samples Genotyping, Quality Control, and Imputation

1,563 blood samples received from Brazil were sent to Akesogen Inc. (Peachtree Corners, Georgia) for DNA extraction and genotyping (by Illumina Global Screening Array v3). Positive control and duplicated samples were included to ensure reproducibility. The genotype data of 1,542 samples which passed calling rate test (>95%) were merged with demographic and phenotype data for further quality control procedures. One sample was removed for missing phenotype, three for sex-mismatch, and seven for identical genomic profiles. Out of the 1,542 samples that passed the calling rate test, only 1,540 were included in the study due to two samples missing demographic information for variables within our model. The genotype data were uploaded to Imputation Server (https://imputation.biodatacatalyst.nhlbi.nih.gov/) for imputation with TOPMed Reference Panel using Minimac4. The imputed data, in GRCh38, were then filtered by imputation quality $R^2 > 0.3$ and used in GWAS analysis. We also applied a filter to exclude SNPs with a minor allele frequency (MAF) below 5% and Hardy-Weinberg Equilibrium (HWE) proportions below $1 \times 10^{-4}$.

## Statistical, Graphical, and Annotations Methods

Statistical analysis was performed using a general linear model (GLM) with principal component analysis (PCA), where the outcome variable was TB resistance (whose definition depended on the analysis) and the predictor variable was the genotype of each SNP. Covariates, including age, sex, and ancestry (top 10 principal components), were included in the model to account for potential confounding factors. Principal components (PCs) were calculated using plink2 and R (v4.0.3). Study samples' genotype data was merged with the 1000 Genome Project reference panel to better visualize ancestries, and non-autosomal chromosomes were not included in the analysis. The described model translates to Resistance Category ~ SNP + Age + Sex + PC1 + PC2 + PC3 + PC4 + PC5 + PC6 + PC7 + PC8 + PC9 + PC10.

A genome-wide significance (GWS) threshold of p-value < $5 \times 10^{-8}$ was used to identify significant associations between SNPs and TB resistance. GWS SNPs from previously published studies in different populations were replicated and compared to SNPs in our study. The genomic inflation factor ($\lambda$) was calculated by dividing the median of the observed $\chi^2$ test statistic by the expected median of the corresponding $\chi^2$ distribution. Post-analysis annotations were made using the MAGMA tool in FUMA's SNP2GENE function (v1.5.2). All graphs were made using R (v4.0.3), and tables were crafted using Microsoft Excel.

## Ethical Considerations

This study was approved by the institutional review board at Emory and at each participating site, and all participants provided written informed consent before participation in the study. All data were de-identified to protect participant confidentiality.

## Results

Household contacts were recruited from TB study clinics in three different cities in Brazil (Manaus, Salvador, Rio de Janeiro) via the Regional Prospective Observational Research in Tuberculosis (RePORT)-Brazil. After data quality control, there were 1,540 individuals included in both analyses. For the primary analysis, there were 1,308 *Mtb* infection cases and 232 resisters. Based on two-sample Z tests of proportions and unpaired T tests, the cases and resisters populations were statistically significantly by proportions of individuals who racially identified as black or mixed, positive IGRA and TST results, negative IGRA and TST results from over 90 days, all levels of infectiousness of the index case, and all levels of exposure level (**Table 2**).

For the secondary analysis, based on the previously described categories, there were 738 TB cases and 802 resisters. Based on Chi-squared tests nd unpaired T tests, the cases and resisters populations were statistically significantly different by sex, age, BMI, proportions of individuals who racially identified as white or black, all IGRA test results, positive TST results, negative TST results from over 90 days, missing TST results, all levels of infectiousness of the index case, and the low and missing levels of exposure (**Table 3**).

A PCA was conducted to account for the differences in ancestry within the study population. When combined with the reference panel from the 1000 Genomes Project, there was strong evidence to support that the study's goal of assessing a multi-racial population was achieved. The process of achieving genetic heterogeneity is critical to understand the true effect of gene variants in the impact or susceptibility to disease. Based on the results from the

PCA (**Figure 1**), the study participants included a genetically diverse and admixed population including African, European and native American ancestries.

## Primary Analysis

Upon completion of quality control, a total of 7,032,343 variants with MAF of 5% or more were used in the GWAS. The primary analysis used the strict phenotypic definition where individuals with negative TST or IGRA results were negative and with the highest-level exposure were categorized as resisters. We used a quantile-quantile (QQ) plot and calculated its corresponding inflation factor ($\lambda$) of 1.17 (**Figure 2**). This value indicates that the observed p-values are more significant than expected, which in this case is likely due to the high relatedness of the individuals in the study population. The results from this analysis (**Figure 3**) uncovered four statistically significant SNPs based on the $5\times10^{-8}$ significance threshold. The SNPs with highest significance were rs888091 (OR = 2.93, 95% CI 2.56 – 3.30, p = $9.99\times10^{-9}$), rs117179998 (OR = 2.81, 95% CI 2.46 – 3.16, p = $1.02\times10^{-8}$), rs35136956 (OR = 2.10, 95% CI 1.81 – 2.39, p = $3.88\times10^{-8}$) in chromosome 2, and rs80327334 (OR = 2.15, 95% CI 1.88 – 2.42, p = $4.69\times10^{-8}$) in chromosome 21 (**Table 4**).

## Preliminary Analysis

The preliminary analysis used a less strict and more inclusive definition of the resisters in comparison to the primary analysis. To analyze the distribution of p-values acquired from the analysis, we used a quantile-quantile (QQ) plot and calculated its corresponding inflation factor ($\lambda$) of 1.18 (**Figure 4**). Comparable to the primary analysis and as expected due to the high relatedness of the study population, this lambda value indicates that the observed p-values are more significant than expected. The results from this preliminary analysis (**Figure 5**) uncovered three SNPs that surpassed the predetermined GWS threshold of $5\times10^{-8}$. The SNPs with highest significance were rs11018572 (OR = 1.58, 95% CI 1.41 – 1.74, p = $2.16\times10^{-8}$) in chromosome 11, rs2369257 (OR = 0.43, 95% CI 0.12 – 0.73, p = $2.83\times10^{-8}$) in chromosome 19, and rs9274695 (OR = 1.56, 95% CI 1.40 – 1.72, p = $4.69\times10^{-8}$) in chromosome 6 (**Table 5**).

## Previously Published Loci

A literature review of previously conducted studies uncovered a list of SNPs associated with resistance or susceptibility to TB infection in other populations. Only loci with SNPs of p-values lower than $5 \times 10^{-8}$ were included in the list. With this restrictive threshold and the overall small number of studies specifically targeting resistance to TB infection, there were two loci of interest that met the required criteria, but neither locus were significantly associated with resistance to *Mtb* infection (p-value >0.05) in our primary analysis (**Table 6**).

# Discussion

In this GWAS of resistance to *Mtb* infection, we investigated two separate definitions of resisters based on the combination of exposure levels, infectivity of the Mtb strains and two measures of Mtb infection. We identified significant genetic association which may reveal the role of human genetic factors in the resistance to *Mtb* infection, and provide insights into potential targets for TB prevention and treatment.

Our primary analysis uncovered SNPs in four statistically significant loci within or near the following genes: *PARD3B* (rs888091, OR = 2.93, 95% CI 2.56 – 3.30, p = 9.99×10$^{-9}$), *AC073987*.1 (rs117179998, OR = 2.81, 95% CI 2.46 – 3.16, p = 1.02×10$^{-8}$), *IQCA1* (rs35136956, OR = 2.10, 95% CI 1.81 – 2.39, p = 3.88×10$^{-8}$) on chromosome 2, and *COL18A1* (rs80327334, OR = 2.15, 95% CI 1.88 – 2.42, p = 3.71×10$^{-8}$) on chromosome 21. The SNP rs888091 is located in an intergenic region of *PARD3B*, which is a previously published locus significantly associated with predisposition to TB in African populations, but not in a Chinese population[23]. The effect of this association within our study was found to be in the same direction of protection against TB as the previously published study in African populations. This similarity strengthens our evidence that this locus may require further functional and replicational investigations to better understand the cellular mechanisms of resistance to Mtb infection. Current knowledge indicates that the protein product of the *PARD3B* gene makes up part of a protein-containing complex predicted to be involved in establishing cell polarity by means of phosphatidylinositol binding activity[24]. This cell membrane maintenance mechanism could play a role in resistance to Mtb infection by potentially inhibiting the ability of the bacteria to bind to alveolar epithelial cells. The locus with the second highest association included sentinel SNP rs117179998 which is close to the *AC073987.1* gene which encodes a long non-coding RNA. This is a newly reported statistically significant locus in the literature and needs to be further replicated and investigated to understand the molecular functions. Another chromosome 2 locus (rs35136956) is located in the intronic region of gene *IQCA1*. Previous studies have shown that *IQCA1* may be associated with nucleotide-binding and catalytic cellular processes[25]. This may play a role in an organism's energy production as it relates to fighting a new infection, but whether these cellular mechanisms translate from animals to humans remains to be explored. In chromosome 21, we found a strong association in a locus (rs80327334) upstream to gene *COL18A1*, and it has been investigated for its role in anti-tuberculosis drug-induced hepatotoxicity. One study in a Western Chinese Han population found that individuals with mutations in the same locus have a decreased risk of anti-tuberculosis drug-induced hepatotoxicity[26].

The secondary analysis encompassed a very broad definition of what resistance entailed where having a negative TST or IGRA result at any point in the study (regardless of the timing of a positive TB test result from the index case) would qualify an individual as a resister, and only those with positive TST or IGRA results were categorized as *Mtb* infected. This analysis uncovered statistically significant associations with p-values of less than $5\times10^{-8}$ in three different loci near genes *NOX4* (rs11018572, OR = 1.58 [1.41 – 1.74]; $p = 2.16\times10^{-8}$), *ATCAY* (rs2369257, OR = 0.43 [0.12 – 0.73]; $p = 2.83\times10^{-8}$), and *HLA-DQB1* (rs9274695, OR = 1.56 [1.40 – 1.72]; $p = 4.69\times10^{-8}$). The rs11018572 SNP is found in chromosome 11 and its closest gene *NOX4* has been shown to be involved in tuberculous fibrosis by binding to microRNA molecules that regulate the production of proinflammatory cytokines and the deposition of the extracellular matrix, stiffness, and parenchymal scarring[27]. This locus was also found to be statistically significant in two other studies not directly related to TB. The first study demonstrated an association with insomnia[28]; given that insomnia is defined to be a difficulty in falling or remaining asleep, this locus could play a role in lung function which can impact an individual's quality of sleep. It is via this lung function mechanism that this locus could be associated with both resistance to TB and insomnia. The other study where this locus was found pertained to an association with cutaneous melanoma. Skin cancer has been shown to be more prevalent among individuals with high sensitivity to ultraviolet radiation in combination with pleiotropic genes like *CASP8*[29]. In a 2022 study, virulent *Mtb* was associated with an increase in *CASP8* expression which plays a role in activation and function of monocyte-derived macrophages[30]. Next, the SNP rs9274695 is located in *HLA-DQB1* region on chromosome 6. According to a meta-analysis from 2016, HLA class II genes like *HLA-DQB1* have demonstrated statistically significant protection against pulmonary TB (OR = 0.77; CI = 0.61 – 0.97)[31], which is in the same directionality of effect of protection as the one observed in our study. Additionally, this locus also has a strong association with traits like waist-to-hip ratio and hip circumference. Given this potential role in morphological traits, it is also possible that this locus is associated with chest cavity size leading to potential mechanisms that yield resistance against Mtb infection by means of inhalation potential[32]. Lastly, rs2369257 is found in chromosome 6 closest to the *ATCAY* gene. This gene is speculated to be important in the metastasis of metastatic pheochromocytomas and paragangliomas via TNF signaling pathways[33]. There is strong evidence demonstrating that drug regimens based on TNF inhibitors may pose a serious health risk to patients with LTBI and lead to the reactivation of TB disease[34].

Due to the specifications of loci that may provide resistance to TB infection and the stringent threshold of statistical significance, only two studies met the requirements providing ten potential loci of interest. The earliest study was done in 2017 while analyzing a population of Tanzanians and Ugandans. This study found that the locus in close proximity to genes *SLC25A48* and *IL9* (rs877356) showed a strong resistance effect over TB infection (OR = 0.27 [95%CI: 0.17 – 0.42]; p = $1.22\times10^{-8}$)[35]. By contrast, our primary analysis yielded non-statistically significant results (OR = 1.07 [95%CI: 0.77 – 1.37]; p = 0.59). This lack of replication in statistical significance across studies may be due to cohort differences. In the study investigating Tanzanians and Ugandans, only HIV+ individuals were included, whereas our study design did not take HIV seropositivity into account. It is possible that some loci's biological role is more strongly identifiable in studies accounting for coinfections. This location is still intriguing given that *IL9* may provide pleiotropic effects on organismal immunity by promoting IL4-mediated production of antibodies, including IgE which plays a key role in bronchial hyper-responsiveness[35].

The remaining locus was derived from a study analyzing resistance to TB infection in a Southern Vietnamese population and later validated in populations from France and South Africa (OR = 0.5 [95% CI: 0.45 – 0.55], p = $1.26\times10^{-9}$). This locus was found within chromosome 10 in intronic and upstream regions of the *C10orf90* gene[20], and our primary analysis did not find statistical significance within the same genomic region (OR = 0.96, 95% CI: 0.66 – 1.26, p = 0.79). There are multiple reasons that could be behind this lack of replication across studies, including ancestry-specific distributions and different case definitions. Although this locus was replicated in multiple populations, it is still possible that ancestry distributions for the three populations investigated in the study are different from that of our study population. Additionally, our study used case definitions that relied on testing results for TB and differences in exposure to the index cases, whereas this study only relied on TST or IGRA test results. The combination of these and other factors may explain why the locus was not statistically significant within our analysis results.

## Limitations and Future Work

While these results provide encouraging insights on genetic variants associated with resistance to TB infection, there are multiple steps that must be taken to properly adjust these results to accurately represent the study population. As previously mentioned, the QQ plots for both analyses (**Figures 2 & 4**) along with their respective lambda-values gave insight into the high inflation taking place within the test statistics. Given the relatedness of the study participants acquired by the study design of sampling household-contacts of TB index cases, the observed high inflation was expected and is likely not due to polygenicity[36,37]. We will address this problem by employing statistical and computational methods that account for the relatedness among sampled individuals. Proper adjustment for relatedness will reduce the false-positive associations within the results.

Another limitation revolves around the definition of what it means to be resistant to TB infection. In the literature, often times terms like "TB susceptibility" and "TB resistance" are used interchangeably even though the two have different epidemiological definitions. The term resistance requires that an individual does not become at all infected with the disease of interest, whereas susceptibility inherently implies differing levels of infection development given that the infection is acquirable. Studying TB further complicates this issue as studies utilize varying definitions when dealing with individuals with active TB, LTBI, or completely uninfected individuals. This lack of agreement in definitions can be problematic when establishing loci of interest and applying generalizability where the outcomes are not the same across studies[38]. An additional layer of complexity to this issue is how resistance is measured. Our study developed a resistance hierarchy (**Table 1**) that categorized five different types of resisters, one category for unknown (discordant) outcomes, and two categories for infected individuals. This level of depth in definition is necessary given the use of two different tests for present of *Mtb* plus the varying levels of exposure endured by the study participants. By developing multiple resister, unknown, and infected categories, we allow for multiple association analyses to be conducted by changing what we define as a resister within the constraints of the resistance hierarchy.

The importance of ancestry diversity cannot be overstated when it comes to applying the results from one study across different populations. As it was demonstrated by the PCA (**Figure 1**), this Brazilian cohort appears to be composed of multiple ethnicities on par with the goal of the study. Although this existing ancestry diversity strengthens the findings from this study, performing replications in other cohorts could bring light to some new insights or concretize the current findings. With this in mind, we will be replicating this study in a cohort in two other countries present in the WHO top 30 TB burden countries – India and South Africa. In 2021, the incidence rate of TB is 188 per 100,000 individuals in India and 513 per 100,000 individuals in South Africa[1]. These rates exceed that of Brazil by 3- to over 10-fold, and by including these countries we can gain better perspectives on genetically driven resistance to TB infection in countries with very high incidence rates.

# References:

1. Geneva: World Health Organization. Global Tuberculosis Report 2022 [Internet]. 2022. Available from: http://apps.who.int/bookorders.
2. Rajshekhar V. Surgery for brain tuberculosis: a review. Acta Neurochir (Wien). 2015;157(10):1665–78.
3. Romanowski K, Clark EG, Levin A, Cook VJ, Johnston JC. Tuberculosis and chronic kidney disease: an emerging global syndemic. Kidney Int. 2016;90(1):34–40.
4. Jain AK, Dhammi IK. Tuberculosis of the spine: A review. In: Clinical Orthopaedics and Related Research. Lippincott Williams and Wilkins; 2007. p. 39–49.
5. de Martino M, Lodi L, Galli L, Chiappini E. Immune Response to Mycobacterium tuberculosis: A Narrative Review. Front Pediatr 2019;7.
6. Kritski A, Dalcolmo MP, Mello FCQ, et al. The role of the Brazilian tuberculosis research network in national and international efforts to eliminate tuberculosis. Jornal Brasileiro de Pneumologia. 2018;44(2):77–81.
7. Rao M, Ippolito G, Mfinanga S, et al. Latent TB Infection (LTBI) – Mycobacterium tuberculosis pathogenesis and the dynamics of the granuloma battleground. International Journal of Infectious Diseases 2019;80:S58–61.
8. Narasimhan P, Wood J, Macintyre CR, Mathai D. Risk factors for tuberculosis. Pulm Med. 2013;
9. Abel L, Fellay J, Haas DW, et al. Genetics of human susceptibility to active and latent tuberculosis: present knowledge and future perspectives. Lancet Infect Dis. 2018;18(3):e64–75.
10. Abel L, El-Baghdadi J, Bousfiha AA, Casanova JL, Schurr E. Human genetics of tuberculosis: A long and winding road. Philosophical Transactions of the Royal Society B: Biological Sciences. 2014;369(1645).
11. Cottle LE. Mendelian susceptibility to mycobacterial disease. Clin Genet. 2011;79(1):17–22.
12. Kinnear C, Hoal EG, Schurz H, van Helden PD, Möller M. The role of human host genetics in tuberculosis resistance. Expert Rev Respir Med. 2017;11(9):721–37.
13. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. Nature Reviews Methods Primers. 2021;1(1).
14. Peterson RE, Kuchenbaecker K, Walters RK, et al. Genome-wide Association Studies in Ancestrally Diverse Populations: Opportunities, Methods, Pitfalls, and Recommendations. Cell. 2019;179(3):589–603.
15. Zheng R, Li Z, He F, et al. Genome-wide association study identifies two risk loci for tuberculosis in Han Chinese. Nat Commun 2018;9(1).
16. Thye T, Vannberg FO, Wong SH, et al. Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. Nat Genet 2010;42(9):739–41.
17. Kanabalan RD, Lee LJ, Lee TY, et al. Human tuberculosis and Mycobacterium tuberculosis complex: A review on genetic diversity, pathogenesis and omics approaches in host biomarkers discovery. Microbiol Res. 2021;246.

18. Jiao L, Song J, Chen H, et al. Genetic architecture of tuberculosis susceptibility: A comprehensive research synopsis, meta-analyses, and epidemiological evidence. Infection, Genetics and Evolution. 2022;104.
19. Cai L, Li Z, Guan X, et al. The Research Progress of Host Genes and Tuberculosis Susceptibility. Oxid Med Cell Longev. 2019;2019.
20. Quistrebert, Orlova, Kerner, et al. Genome-wide association study of resistance to Mycobacterium tuberculosis infection identifies a locus at 10q26.2 in three distinct populations. PLoS Genet 2021;17(3).
21. Igo RP, Hall NB, Malone LSL, et al. Fine-mapping analysis of a chromosome 2 region linked to resistance to Mycobacterium tuberculosis infection in Uganda reveals potential regulatory variants. Genes Immun 2019;20(6):473–83.
22. Stein CM, Zalwango S, Malone LSL, et al. Genome scan of M. tuberculosis infection and disease in Ugandans. PLoS One 2008;3(12).
23. Wang X, Tang NLS, Leung CC, et al. Association of polymorphisms in the Chr18q11.2 locus with tuberculosis in Chinese population. Hum Genet 2013;132(6):691–5.
24. Falker-Gieske C, Iffland H, Preuß S, et al. Meta-analyses of genome wide association studies in lines of laying hens divergently selected for feather pecking using imputed sequence level genotypes. BMC Genet 2020;21(1).
25. Huang Y, Li Y, Wang X, et al. An atlas of CNV maps in cattle, goat and sheep. Sci China Life Sci 2021;64(10):1747–64.
26. Cheng Y, Jiao L, Li W, et al. Collagen type XVIII alpha 1 chain (COL18A1) variants affect the risk of anti-tuberculosis drug-induced hepatotoxicity: A prospective study. J Clin Lab Anal 2021;35(2).
27. Woo SJ, Kim Y, Jung H, Lee JJ, Hong JY. MicroRNA 148a Suppresses Tuberculous Fibrosis by Targeting NOX4 and POLDIP2. Int J Mol Sci 2022;23(6).
28. Watanabe K, Jansen PR, Savage JE, et al. Genome-wide meta-analysis of insomnia prioritizes genes associated with metabolic and psychiatric pathways. Nat Genet 2022;54(8):1125–32.
29. Liyanage UE, MacGregor S, Bishop DT, et al. Multi-Trait Genetic Analysis Identifies Autoimmune Loci Associated with Cutaneous Melanoma. Journal of Investigative Dermatology 2022;142(6):1607–16.
30. Ramon-Luing LA, Olvera Y, Flores-Gonzalez J, et al. Diverse Cell Death Mechanisms Are Simultaneously Activated in Macrophages Infected by Virulent Mycobacterium tuberculosis. Pathogens 2022;11(5).
31. Oliveira-Cortez A, Melo AC, Chaves VE, Condino-Neto A, Camargos P. Do HLA class II genes protect against pulmonary tuberculosis? A systematic review and meta-analysis. European Journal of Clinical Microbiology and Infectious Diseases. 2016;35(10):1567–80.
32. Christakoudi S, Evangelou E, Riboli E, Tsilidis KK. GWAS of allometric body-shape indices in UK Biobank identifies loci suggesting associations with morphogenesis, organogenesis, adrenal cell renewal and cancer. Sci Rep 2021;11(1).
33. Zhang C, Kang Y, Yang Q, Article R. Joint application of multiple genes in the diagnosis and pathogenesis of metastatic peochormocytoma/paraganglioma

based on bioinformatics. 2022;Available from: https://doi.org/10.21203/rs.3.rs-1601822/v1

34. Robert M, Miossec P. Reactivation of latent tuberculosis with TNF inhibitors: critical role of the beta 2 chain of the IL-12 receptor. Cell Mol Immunol. 2021;18(7):1644–51.

35. Sobota RS, Stein CM, Kodaman N, et al. A chromosome 5q31.1 locus associates with tuberculin skin test reactivity in HIV-positive individuals from tuberculosis hyper-endemic regions in east Africa. PLoS Genet 2017;13(6).

36. Gross A, Tönjes A, Scholz M. On the impact of relatedness on SNP association analysis. BMC Genet 2017;18(1):104.

37. Yang J, Weedon MN, Purcell S, et al. Genomic inflation factors under polygenic inheritance. European Journal of Human Genetics 2011;19(7):807–12.

38. Gutierrez J, Kroon EE, Möller M, Stein CM. Phenotype Definition for “Resisters” to Mycobacterium tuberculosis Infection in the Literature—A Review and Recommendations. Front Immunol. 2021;12.

## Tables and Figures

| Hierarchy | QFT/TST (Final) | Contact Exposure Level | Index infectiousness | Notes |
|---|---|---|---|---|
| Resister A | QFT- (>90 days) & TST- (>90 days) | High* | High | |
| Resister B | QFT- (>90 days) OR TST- (>90 days) | High | High | No positive QFT/TST results |
| Resister C | QFT- (<90 days) OR TST- (<90 days) | High | High | No positive QFT/TST results |
| Resister D | QFT- or TST- (any time) | Low^ | Medium/Low | Either low contact exposure or low index infectiousness |
| Resister E | QFT- or TST- (any time) | Unknown | Unknown | Either unknown/missing contact exposure or index infectiousness |
| Unknown A | Discordant | Any/Unknown | Any/Unknown | QFT+ and TST-(any time), OR QFT-(any time) and TST+ |
| Infected B | QFT+ OR TST+ | Any/Unknown | Any/Unknown | |
| Infected A | QFT+ & TST+ | Any/Unknown | Any/Unknown | |

**Table 1.** Resistance hierarchy defining how study individuals were categorized based on test results and contact exposure level.

**"High" contact exposure defined as reported sharing a bed or room with index case or 5 or more hours indoor.*

*^"Low" contact exposure defined as answering the exposure questions and not reporting any of the "high" exposures (room or bed sharing, 5 or more hours indoors).*

This table was developed by Fay Willis at Emory's Rollins School of Public Health.

QFT: QuantiFERRON Test; also referred to as interferon-gamma release assay (IGRA)

TST: Tuberculin Skin Test

| | Mean(SD)/ N(%) | Mean(SD)/ N(%) | P-value |
|---|---|---|---|
| TB Status | TB+ (n = 1308) | TB- (n = 232) | --- |
| Sex - male | 530 (41%) | 89 (38%) | 0.54 |
| Age | 32.9 (19.3) | 30.9 (18.4) | 0.14 |
| BMI | 24.6 (6.6) | 25.4 (6.2) | 0.086 |
| Race - | -- | -- | -- |
| White | 252 (19%) | 44 (19%) | 0.91 |
| Black | 250 (19%) | 62 (27%) | 0.0078 |
| Mixed | 432 (60%) | 125 (54%) | < 0.0001 |
| QFT Final Test Results - | -- | -- | -- |
| Positive (anytime) | 741 (57%) | 0 (0%) | < 0.0001 |
| Negative (>90 days) | 556 (43%) | 227 (98%) | < 0.0001 |
| Negative (<90 days) | 11 (1%) | 5 (2%) | 0.069 |
| TST Final Test Results - | -- | -- | -- |
| Positive (anytime) | 44 (3%) | 0 (0%) | 0.0047 |
| Negative (>90 days) | 3 (0%)* | 4 (2%) | 0.0018 |
| Negative (<90 days) | 2 (0%)* | 2 (1%) | 0.050 |
| Missing | 1259 (96%) | 226 (97%) | 0.38 |
| Infectiousness of Index case - | -- | -- | -- |
| High | 727 (56%) | 232 (100%) | < 0.0001 |
| Medium | 322 (25%) | 0 (0%) | < 0.0001 |
| Low | 259 (20%) | 0 (0%) | < 0.0001 |
| Exposure Level - | -- | -- | -- |
| Same bed | 134 (10%) | 45 (19%) | < 0.0001 |
| Same room | 120 (9%) | 35 (15%) | 0.0058 |
| At least 5 hours indoors | 385 (29%) | 152 (66%) | < 0.0001 |
| None of the above | 152 (12%) | 0 (0%) | < 0.0001 |
| Missing | 517 (40%) | 0 (0%) | < 0.0001 |

**Table 2.** Study population stratified by TB infection status based on resister definition for primary analysis. Categorical variables (sex, race, QFT results, TST results, infectiousness of index case, and exposure level) are represented by number (N) and percentage of stratified groups. Continuous variables are represented by mean and standard deviations based on stratified groups.

*: Percentage values were rounded to nearest whole number which explains categories not adding to 100% in total.

Abbreviations: TB (Tuberculosis); SD (Standard Deviation); BMI (Body Mass Index); QFT (QuantiFERON Test, also referred to as interferon-gamma release assay [IGRA]); TST (Tuberculin Skin Test).

| | Mean (SD)/ N(%) | Mean (SD)/ N(%) | P-value |
|---|---|---|---|
| TB Status | TB+ (n = 738) | TB- (n = 802) | --- |
| Sex - male | 277 (35%) | 342 (43%) | 0.041 |
| Age | 35.8 (19.5) | 30.2 (18.5) | < 0.0001 |
| BMI | 25.2 (6.5) | 24.3 (6.5) | 0.0067 |
| Race - | -- | -- | -- |
| White | 116 (16%) | 180 (22%) | 0.0008 |
| Black | 181 (25%) | 131 (16%) | < 0.0001 |
| Mixed | 432 (59%) | 484 (60%) | 0.47 |
| QFT Final Test Results - | -- | -- | -- |
| Positive (anytime) | 738 (100%) | 3 (0%)* | < 0.0001 |
| Negative (>90 days) | 0 (0%) | 783 (98%) | < 0.0001 |
| Negative (<90 days) | 0 (0%) | 16 (2%) | 0.00012 |
| TST Final Test Results - | -- | -- | -- |
| Positive (anytime) | 36 (5%) | 8 (1%) | < 0.0001 |
| Negative (>90 days) | 0 (0%) | 7 (1%) | 0.011 |
| Negative (<90 days) | 0 (0%) | 4 (0%)* | 0.055 |
| Missing | 702 (95%) | 783 (98%) | 0.0080 |
| Infectiousness of Index case - | -- | -- | -- |
| High | 545 (74%) | 414 (52%) | < 0.0001 |
| Medium | 118 (16%) | 204 (25%) | < 0.0001 |
| Low | 75 (10%) | 184 (23%) | < 0.0001 |
| Exposure Level - | -- | -- | -- |
| Same bed | 87 (12%) | 92 (11%) | 0.85 |
| Same room | 67 (9%) | 88 (11%) | 0.22 |
| At least 5 hours indoors | 250 (34%) | 287 (36%) | 0.43 |
| None of the above | 42 (6%) | 110 (14%) | < 0.0001 |
| Missing | 292 (40%) | 225 (28%) | < 0.0001 |

**Table 3.** Study population stratified by TB infection status based on alternative resister definition for preliminary analysis. Categorical variables (sex, race, QFT results, TST results, infectiousness of index case, and exposure level) are represented by number (N) and percentage of stratified groups. Continuous variables are represented by mean and standard deviations based on stratified groups.

*: Percentage values were rounded to nearest whole number which explains categories not adding to 100% in total.

Abbreviations: TB (Tuberculosis); SD (Standard Deviation); BMI (Body Mass Index); QFT (QuantiFERON Test, also referred to as interferon-gamma release assay [IGRA]); TST (Tuberculin Skin Test).

| SNP | CHR | Position | Effect Allele | Ref Allele | MAF | OR | OR 95% CI | P-value | Genomic Region | Nearest Gene |
|---|---|---|---|---|---|---|---|---|---|---|
| rs888091 | 2 | 205627269 | T | A | 0.051 | 2.93 | (2.56 - 3.30) | $9.99\times10^{-9}$ | intergenic | PARD3B |
| rs117179998 | 2 | 103147519 | C | A | 0.052 | 2.81 | (2.46 - 3.16) | $1.02\times10^{-8}$ | ncRNA intronic | AC073987.1 |
| rs80327334 | 21 | 45404346 | C | T | 0.138 | 2.15 | (1.88 - 2.42) | $3.71\times10^{-8}$ | upstream | COL18A1 |
| rs35136956 | 2 | 236409294 | T | G | 0.105 | 2.1 | (1.81 - 2.39) | $3.88\times10^{-8}$ | intronic | IQCA1 |

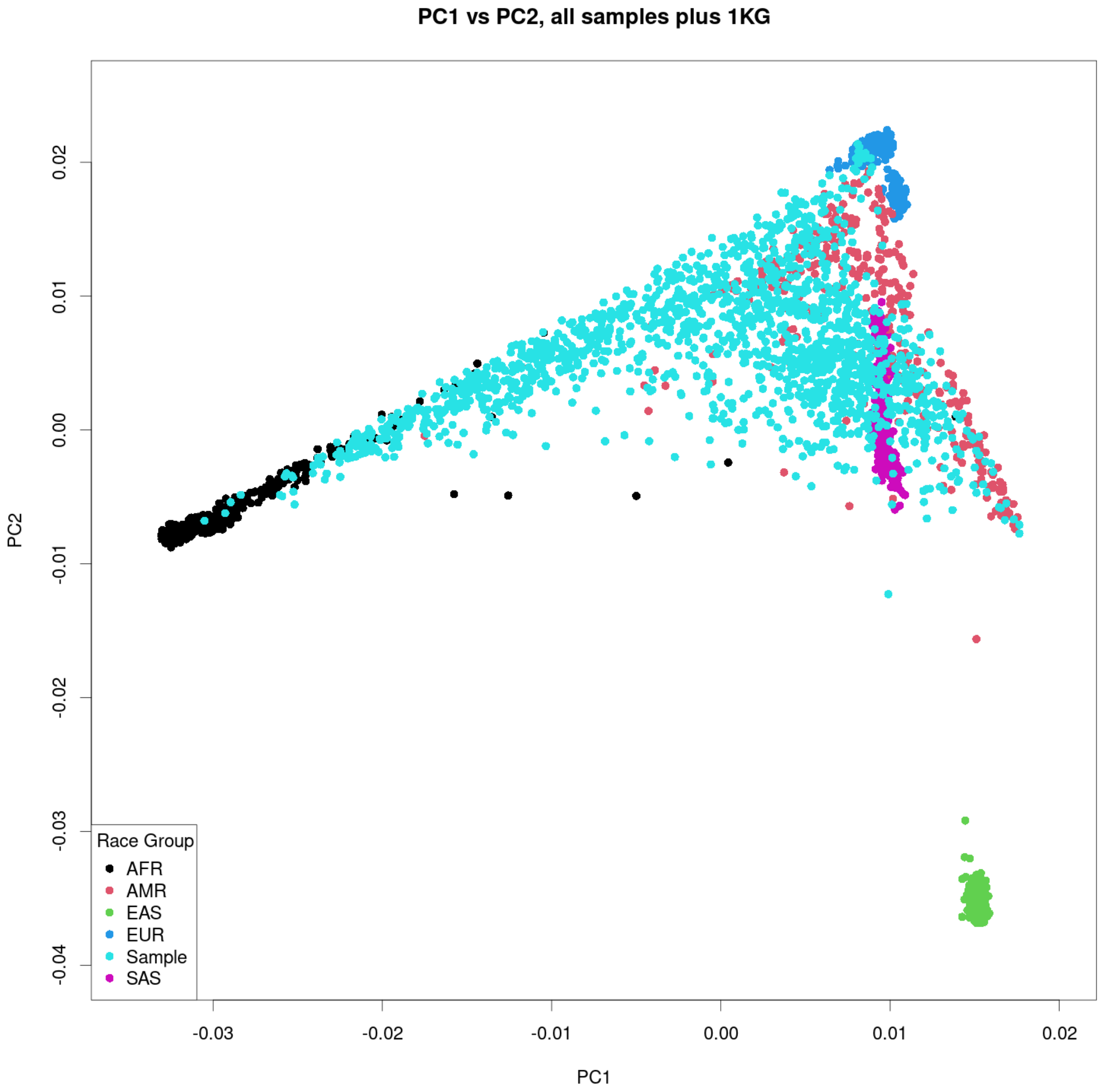
**Table 4.** List of statistically significant ($p < 5\times10^{-8}$) loci associated with resistance to TB infection using less strict resister definition from the second analysis. Abbreviations: SNP (Single-Nucleotide Polymorphism); CHR (Chromosome); MAF (Minor Allele Frequency); OR (Odds Ratio); CI (Confidence Interval).

| SNP | CHR | Position | Effect Allele | Ref Allele | MAF | OR | OR 95% CI | P-value | Genomic Region | Nearest Gene |
|---|---|---|---|---|---|---|---|---|---|---|
| rs11018572 | 11 | 89318712 | T | C/G | 0.376 | 1.58 | (1.41 - 1.74) | $2.15\times10^{-8}$ | intergenic | NOX4 |
| rs2369257 | 19 | 3879402 | G | A | 0.127 | 0.43 | (0.12 - 0.73) | $2.83\times10^{-8}$ | upstream | ATCAY |
| rs9274695 | 6 | 32669220 | T | C/G | 0.364 | 1.56 | (1.40 - 1.72) | $4.69\times10^{-8}$ | upstream | HLA-DQB1 |

**Table 5.** List of statistically significant ($p < 5\times10^{-8}$) loci associated with resistance to TB infection in preliminary analysis using the alternative and less strict resister definition. Abbreviations: SNP (Single-Nucleotide Polymorphism); CHR (Chromosome); MAF (Minor Allele Frequency); OR (Odds Ratio); CI (Confidence Interval).

| SNP | CHR | Nearest Gene | Effect Allele | OR | 95% CI | P-value | Study Population | Source | OR from Primary Analysis | 95% CI | P-value |
|---|---|---|---|---|---|---|---|---|---|---|---|
| rs17155120 | 10 | C10orf90 | T | 0.5 | (0.45 - 0.55) | $1.26\times10^{-9}$ | Vietnam, France, South Africa | Quistrebert et al., 2021[20] | 0.96 | (0.66 - 1.26) | 0.79 |
| rs877356 | 5 | SLC25A48/IL9 | T | 0.27 | (0.17 - 0.42) | $1.22\times10^{-8}$ | Tanzania, Uganda | Sobota et al., 2017[35] | 1.07 | (0.77 - 1.37) | 0.59 |

**Table 6.** List of previously published loci with corresponding genes, effect alleles, odds ratio (OR), 95% confidence intervals, P-values, and study populations. The last three columns (from left to right) indicate the association found within our study. Abbreviations: SNP (Single-Nucleotide Polymorphism); CHR (Chromosome); MAF (Minor Allele Frequency); OR (Odds Ratio); CI (Confidence Interval).



**Figures 1.** Plot comparing different sets of principal components acquired through PCA analysis. This figure shows Principal Component 1 (PC1) versus Principal Component 2 (PC2). Study samples are represented in light blue, and the remaining dots originated from the 1000 Genome Project where African (AFR) ancestry are represented in black, European (EUR) ancestry in dark blue, Native American ancestry (AMR) in red, East Asians (EAS) in green, and South Asians (SAS) in pink.

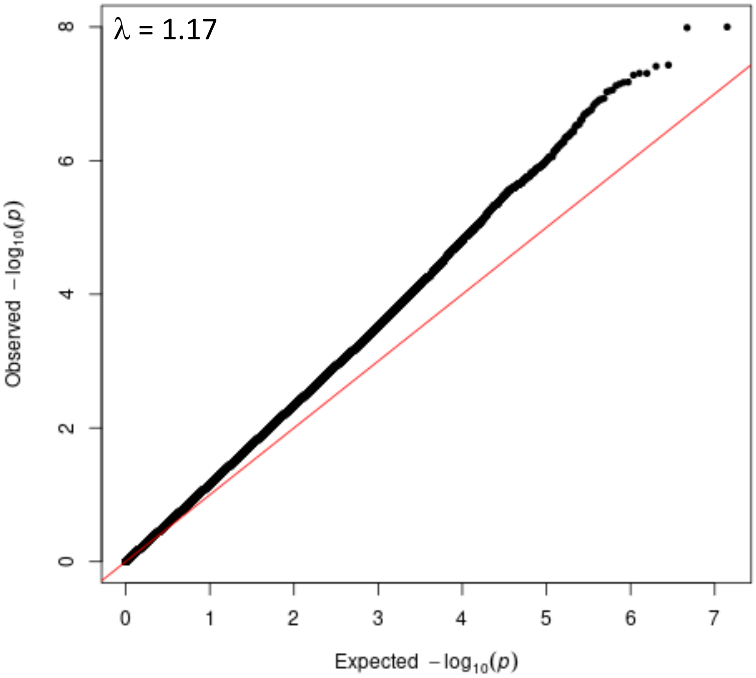


**Figure 2.** QQ-Plot showing observed p-values from the primary analysis against expected p-values scaled to -$\log_{10}$(p). The calculated inflation factor (λ) is 1.17 which indicates inflated distribution of low p-values.

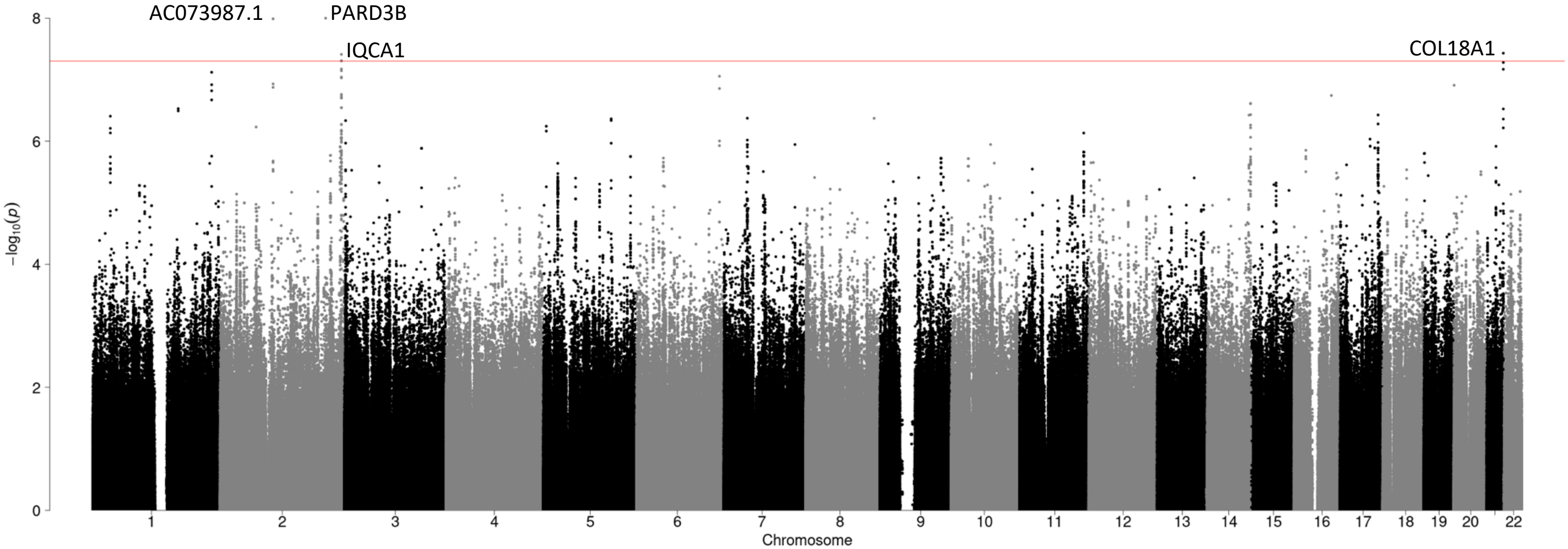


**Figure 3.** Manhattan plot of all primary analysis p-values logarithmically scaled on the y-axis and chromosomes 1 through 22 in the x-axis. Alternating colors were used to distinguish between adjacent chromosomes. Statistical significance threshold is represented by the red line. The genes closest to the four statistically significant loci are labeled.

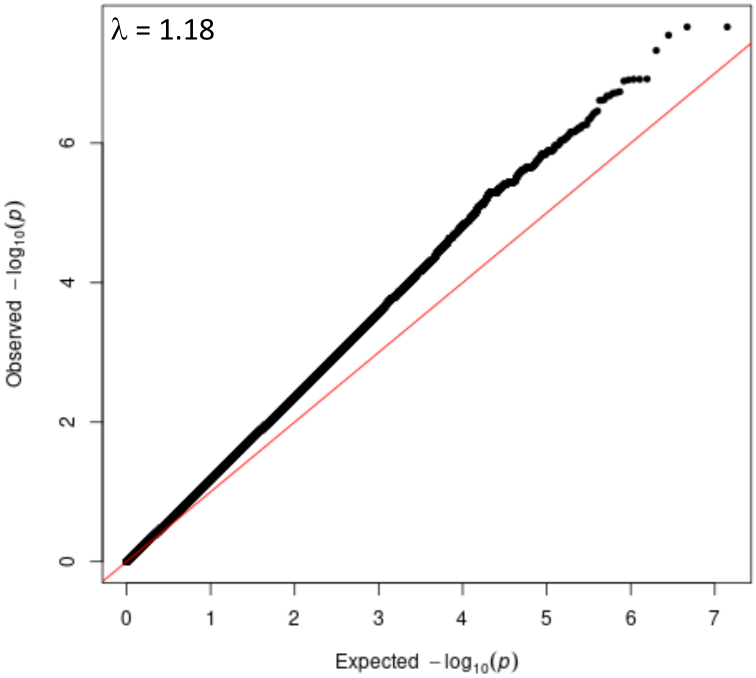


**Figure 4.** QQ-Plot showing observed p-values from the preliminary analysis against expected p-values scaled to -log$_{10}$(p). The calculated inflation factor (λ) is 1.18 which indicates inflated distribution of low p-values.

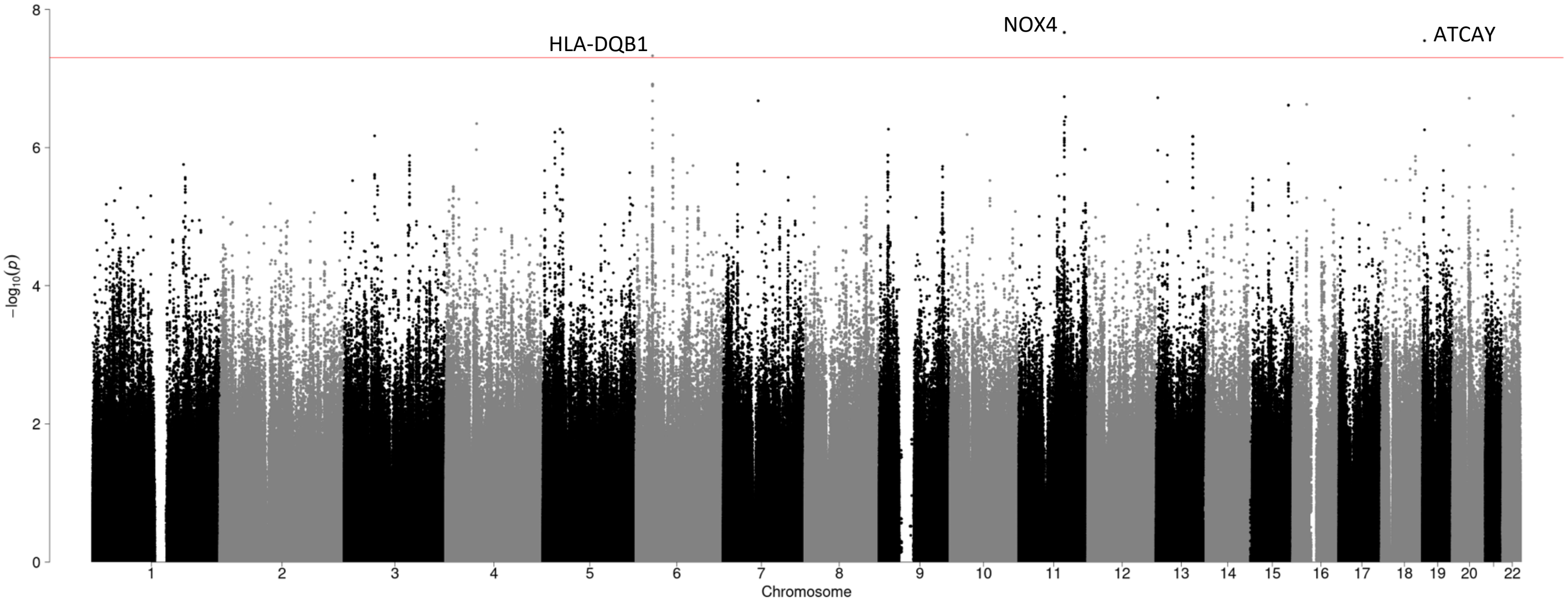


**Figure 5.** Manhattan plot of all primary analysis p-values logarithmically scaled on the y-axis and chromosomes 1 through 22 in the x-axis. Alternating colors were used to distinguish between adjacent chromosomes. Statistical significance threshold is represented by the red line. The genes closest to the three statistically significant loci are labeled.