Genome-Wide Associations Between Human Genotypes and *Mycobacterium tuberculosis* Clades Causing Disease

Stephanie Julia Pitts

Dissertation presented for the degree of Doctor of Philosophy in Human Genetics in the Faculty of Medicine and Health Sciences at Stellenbosch University

Supervisor: Prof Craig Kinnear

Faculty of Medicine and Health Sciences

Department of Biomedical Sciences

Co-supervisors:

Prof Marlo Möller, Prof Eileen Hoal

Prof Gian van der Spuy, Prof Gerard Tromp

April 2019

Declaration

By submitting this dissertation electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature: ……………………….. Date: **April 2019**

Abstract

The World Health Organization (WHO) declared tuberculosis (TB) to be a global health emergency in 1993, and despite decades of extensive biomedical research, it remains a major cause of morbidity and mortality around the world. A disease primarily affecting the lungs, TB manifests following infection with a pathogenic member of the *Mycobacterium tuberculosis* (*M. tb)* Complex (MTBC) such as *M. africanum* and *M. tb*, although infection alone is not sufficient for disease. Each member of the MTBC consists of several strains (or clades), with variable virulence and disease-causing mechanisms. *M. africanum* is the main cause of TB in West African countries including Ghana, while *M. tb* is responsible for TB cases in most other parts of the world, with stratification of clades by geographical location.

TB is a multifactorial disease, influenced by environmental factors, bacterial virulence, and the genetic susceptibility of the host. While the genetic susceptibility of the host to the tuberculous disease has been extensively studied using genome-wide association studies and candidate gene studies, no method currently exists to perform an association analysis between the genetic architecture of the host and the susceptibility to the many clades of *M. tb* or *M. africanum* causing disease.

Two geographically distinct cohorts were included in this study: a cohort of 947 participants self-identifying as belonging to the five-way admixed South African Coloured (SAC) population with paired infecting *M. tb* isolate information was used to establish the protocol for performing the association analysis, while a second cohort consisting of 3 311 participants recruited in Ghana was used to validate this method. The method developed includes quality control filters on both the host genotype data, and the infecting isolate database. Thereafter, haplotype phasing and genotype imputation of several reference panels was performed to increase the number of single nucleotide polymorphisms (SNPs) available for association testing. An assessment of imputation quality scores revealed the best imputation reference panel for the study cohort and a multinomial logistic regression (MLR) analysis was performed to assess potential associations between host genotypes and infecting bacterial clades of multiple classes.

Here, we demonstrated that the African Genome Resource (used via the Sanger Imputation Server) produced the highest quality of imputed genotype data for the SAC cohort, while the 1000 Genomes Phase 3 reference panel was the best reference panel for the Ghanaian cohort. MLR was performed while controlling for covariates including age, sex, and ancestry proportions. After genotype imputation, 445 SAC - and 1 272 Ghanaian participants passed quality control and were tested for association to five- and six infecting superclades, respectively. Models of association revealed no SNPs reaching genome-wide significance for the SAC cohort, while 32 SNPs met the GWAS cut-off of 5 x 10^{-8} for the Ghanaian cohort. For the Ghanaian cohort, the risk allele of SNP rs551641937 (g.62385889G>A), located on chromosome 15, was determined to increase the risk of TB caused by the EAI/AFRI superclade by 276 times, when compared to the LAMCAM reference superclade. The emphasis of the dissertation was to perform an association analysis using host genotype and pathogen data and finding the best reference panel for imputing each of the two datasets was a secondary aim. This study demonstrates the first method successfully testing host-genotype associations with multiple clades of *M. tb* isolates causing disease.

Opsomming

Tuberkulose (TB) is in 1993 as 'n globale gesondheidsprobleem deur die Wêreld Gesondheidsorganisasie verklaar. Ondanks dekades se omvattende biomediese navorsing, bly TB 'n hoofoorsaak van sterftes wêreldwyd. TB is 'n siekte wat hoofsaaklik die longe affekteer en manifesteer na infeksie met 'n patogeniese lid van die *Mycobacterium tuberculosis*-kompleks (MTBK), naamlik *M. africanum* en *M. tb*. Elke lid van die MTBK bestaan uit verskeie stamme (of klade) wat verskil in virulensie. *M. africanum* is die hoofoorsaak van TB in lande in Wes-Afrika insluitend Ghana, terwyl *Mycobacterium tuberculosis (M. tb)* verantwoordelik is vir TB gevalle in die meeste ander dele van die wêreld, met klades wat gegroepeer kan word volgens hulle geografiese ligging.

TB is 'n komplekse siekte met verskeie faktore wat dit beïnvloed, insluitend omgewingsfaktore, bakteriële virulensie en die genetiese vatbaarheid van die gasheer. Verskeie studies, insluitend genoom-wye assosiasie studies (GWAS) en kandidaat studies, is al uitgevoer om die genetiese vatbaarheid van die gasheer vir TB te ondersoek. Tot dusver is daar geen metode om assosiasies te analiseer tussen die genetiese struktuur van die gasheer en die vatbaarheid tot enige van die verskeie klades van *M. tb* of *M. africanum*.

Die studie het gebruik gemaak van twee kohorte in verskillende geografiese areas: 'n groep van 947 deelnemers wat hulself geïdentifiseer het as deel van die Suid Afrikaanse Kleurling (SAK) populasie, en 'n tweede groep met 3 311 deelnemers vanaf Ghana. Die SAK groep, afkomstig van vyf voorvaderlike populasies, met ooreenstemmende *M. tb* isolaat informasie was gebruik om die protokol vir gasheer genotipe-tot-infeksie klade te ontwikkel. 'n Tweede groep vanaf Ghana was ingesluit om die metode te valideer. Die metode sluit kwaliteitskontrole filters in vir beide die gasheer genotipe data, asook vir die infeksie isolaat databasis. Die volgende stap was haplotipe fasering en genotipe imputasie. Dit was uitgevoer met verskeie verwysings panele om die hoeveelheid enkel-nukleotied polimorfismes (ENP) beskikbaar vir assosiasie toetse te vermeerder. Die kwaliteit van imputasie was bepaal deur die beste verwysing paneel elke kohort te kieswaarna multinomiale logistieke regressie (MLR) analiese gebruik was om potensiale assosiasies tussen die gasheer genotipe en infekterende bakteriële klades van veelvuldige klasse te bepaal.

Hierdie studie demonstreer dat die Afrika Genoom Hulpbron (gebruik deur die Sanger Imputasie Bediener) die beste kwaliteit imputasies gegee het vir genotipe data vir die SAK populasie, terwyl die 1000 Genome Fase 3 verwysings paneel die beste was vir die Ghana kohort. MLR analise het ouderdom, geslag en genetiese afkoms in ag geneem. Na genotipe imputasie, het 445 SAK en 1 272 Ghana deelnemers die kwaliteits kontrole stappe geslaag en is afsonderlik getoets vir moontlike assosiasies met vyf of ses infekterende superklades, onderskeidelik. Die modelle van assosiasie het nie enige ENP in die SAK populasie uitgelig wat genoom-wyd statisties betekenisvol was nie, maar daar was egter 32 ENP's wat 'n waarskynlikheids waarde kleiner as 5 x 10^{-8} gehad het vir die Ghana kohort. Daar is gevind dat een van die ENKs, rs551641937 (g.62385889G>A) geleë op kromosoom 15, die risiko van TB in verband met die EAI/AFRI super-klade 276 keer verhoog in vergelyking met die LAMCAM super-klade. Die klem van die verhandeling was om 'n assosiasie-analise uit te voer met behulp van gasheergenotipe en patogeen data en die vind van die beste verwysingspaneel om elkeen van die twee datastelle toe te pas, was 'n sekondêre doelwit. Hierdie studie demonstreer die eerste metode wat suksesvol gebruik was om te toets vir assosiasies tussen gasheer genotipe en veelvuldige klades wat TB veroorsaak.

Acknowledgements

There are many people who have played a significant role in my academic successes thus far. I'd like to take this opportunity to firstly thank my PhD supervisor, Prof Craig Kinnear for an unimaginable amount of support - from the start of my funding applications for this degree, to the very end of my thesis submission. I will be forever grateful for the advice and insight you have shared with me. Before I joined the MAGIC lab, I had very little knowledge of Tuberculosis and Human Genetics, but I leave the lab much richer in knowledge and appreciation for both these research fields. Thank you for believing in my capabilities when I doubted whether I was "good enough" and for always finding a way to ensure that our working environment was a "happy place"- it really made it easy to come to work each day. Lastly, I think your successes as a supervisor might not only be measured by the number of students who graduate to stay in your lab, but also by how many students you have given wings, to help them fly.

Next, I'd like to thank Prof Marlo Möller. Thank you for walking with me every step of the way through this PhD. Together, you and Craig have led a team of amazing supervisors who have all helped in their own way to see this project through. Prof Eileen Hoal - without your vision to start this research group, I would not be here. Thank you for your vision all those years ago, and for your vision today. I aspire to always have the same passion for Science and for tackling the medical problems society faces, as you do. Gian, thank you for your time. You were always available to help when I simply could not "figure out the code". Prof Gerard Tromp, thank you for always pushing me to learn beyond my existing boundaries. From the beginning of this adventure, you have shown me what it means to think logically, while being a free-thinker when exploring data. Thank you for always having an open door. Without any reservation, I can confirm that this has turned out to be a "Dream Team" of supervisors. Albeit many, each supervisor played an essential role in the success of this project, and it would not have seen fruition without your knowledge, patience, and advice.

I will be forever grateful to the members of the MAGIC lab and TB Host Genetics Research group during my time at Stellenbosch University. Coming from another Institution, I was welcomed with open arms and I cannot thank each of you enough for your friendship and support. In particular, I'd like to thank Anél and Keren for the moments of insanity which we spilt over coffee. A special thanks is extended to Haiko for his willingness to collaborate on analyses and publications. As there are too many more to mention, I owe much of my sanity to the members of this amazing research group. Thank you.

Next, I'd like to thank Prof Paul van Helden and Prof Gerhard Walzl, for both in their own time, leading the Division. Thank you for assisting whenever I needed funding for conference and courses. I hope that through these attendances, I have represented our University, Division, and the Host Genetics Research Group well.

I'd like to also thank Prof Rob Warren for his collaboration with our research group through his provision of TB strain data for the South African cohort, and for assisting with funding for conference attendance. My appreciation is also extended to Dr Anzaan Dippenaar and Dr Lizma Streicher for always being willing to share their knowledge and insight with regards to the TB strains included in this study. Without your help, attempting to unravel the connections with human genetics components would have been significantly more difficult.

Thank you to Dr Thorsten Thye and Prof Stefan Niemann for their provision of patient genotype data and strain data for the Ghanaian cohort. Your willingness to collaborate with our research group is highly appreciated. Here I'd also like to thank the participants of both study cohorts who essentially were the cornerstone of this study. Your contributions to medical research cannot be appreciated enough.

I'd also like to thank the National Research Foundation (NRF), and the South African Medical Research Council (SAMRC), who funded the first, and last two years of my PhD degree, respectively. This research was partially funded by the South African government through the South African Medical Research Council. The content is solely the responsibility of the authors and does not necessarily represent the official views of the South African Medical Research Council. The South African Tuberculosis Bioinformatics Initiative (SATBBI) is also thanked for the contributions of this initiative to my attendance at the SASBI/SAGS conference. This work was also supported by the National Research Foundation of South Africa (grant number 93460) to Eileen Hoal and by a Strategic Health Innovation Partnership grant from the South African Medical Research Council and Department of Science and Technology/South African Tuberculosis Bioinformatics Initiative (SATBBI, Gerhard Walzl) to Gerard Tromp.

Lastly, I would like to thank my family and friends for their unwavering support over the last three years. The last year was particularly tough, with many changes occurring simultaneously, but your constant support and concern for me played a large role in my success with completing this thesis.

Ereshia, Farren, and Jody: You have been sources of strength and motivation. Thank you for your concern and help whenever I needed it. To Jodie, thank you for the coffees and for checking in on me; I hope to be as good a friend to you as you have been to me. To aunty Charmaine and uncle Malcolm, I now get to call you “Mom” and “Dad”. Thank you for being there for me through each of my degrees, for always expressing concern for me, and always making sure I have enough to eat. I hope to always make you proud.

To my husband, Francuois Müller, it will take me a lifetime to thank you for all the support you have given me thus far. You have seen me through all my degrees, kept me focussed, and kept me grounded. Thank you for always inspiring me to do what I love, and to strive to become better at what I do. You have always been a source of inspiration for helping others, and for solving the problems which our communities face - Thank you for always being you.

To my mom and dad, I dedicate this thesis to you. You have been with me through it all - the Extra Maths classes, the late study nights. I can only strive to find more ways in which to say “Thank you”. You have always inspired me to work hard, and be better than I was yesterday. Thank you Mommy, for always asking “Do you enjoy this work?” - it truly has played a big role in leading me to my current path, and in building the firm academic foundation which I stand on today. I hope to continue my journey in Science as someone who perseveres through the tough times, and always seeks to improve the lives of others.

Table of Contents

List of Abbreviations

1000GP3	1000 Genomes Project phase 3
AFB	Acid-fast bacilli
AGR	African Genome Resource
AIDS	Acquired immunodeficiency virus
AIMs	Ancestry Informative Markers
BAL	Bronchoalveolar lavage
BCG	Bacille Calmette-Guérin
BMI	Body mass index
CAAPA	Consortium on Asthma among African-ancestry Populations in the Americas
CFP-10	Culture filtrate protein-10
ELISA	Enzyme-linked immunosorbent assay
ESAT-6	Early secretory antigenic target 6kd
ETB	Ethambutol
GWAS	Genome-wide association studies
HDT	Host-directed therapy
HIV	Human immunodeficiency virus
HLA	Human leukocyte antigen
HWE	Hardy-Weinberg equilibrium
IFN-γ	Interferon gamma
IGRA	Interferon-gamma release assay
INH	Isoniazid
LAM	Latin American-Mediterranean
LCC	Low-Copy clade
LDL	Low density lipoprotein
LJ	Lowenstein-Jensen
LPA	Line probe assay
LTBI	Latent TB infection
MAF	Minor allele frequency
M. africanum 1	*M. africanum* West-African 1
M. africanum 2	*M. africanum* West-African 2
MBL	Mannose binding lectin

MDR	Multi-drug resistant
MGIT	mycobacterium growth indicator tube
M. tb	*Mycobacterium tuberculosis*
MTBC	*Mycobacterium tuberculosis* Complex
NAA	Nucleic acid amplification
PC	Principal component
PCA	Principal Components Analysis
PheWAS	Phenome-wide association study
PZA	Pyrazinamide
QFT	Quantiferon®-TB Gold In-Tube
RARA	Retinoic acid receptor alpha
RIF	Rifampicin
Rsq	R-squared
SAC	South African Coloured
SNP	Single nucleotide polymorphism
TB	Tuberculosis
TDR	Total drug resistant
TLR	Toll-like receptor
TNF	Tumour-necrosis factor
TST	Tuberculin skin test
VDR	Vitamin D receptor
WHO	World Health Organization
XDR	Extensively drug-resistant

List of Figures

List of Tables

CHAPTER ONE

1. Introduction

1.1. Background

The World Health Organization (WHO) declared tuberculosis (TB) to be a global health emergency in 1993, and despite decades of extensive biomedical research, it remains a major cause of morbidity and mortality around the world (WHO 2017b). A disease primarily affecting the lungs, TB manifests following infection with a pathogenic member of the *Mycobacterium tuberculosis* (*M. tb)* Complex (MTBC). The MTBC consists of five species of mycobacteria, namely *M. africanum*, *M. canetti*, and *M. tb* (all pathogenic in humans), *M. microti*, a pathogen affecting mainly rodents, and *M. bovis*, which has adapted to cause disease in both humans and animals (Frothingham 1995).

In this dissertation, reference will be made to clades which can be defined as the name given to a family of strains. For example, the modern Lineage 2 consists of the Beijing clade determined by means of a defined spoligotyping pattern. The Beijing clade, however, consists of numerous strains. The term 'clade' may also be used interchangeably with 'sub-lineage', albeit that the latter term is commonly used in reference to numeric annotations of sub-members (Coll et al. 2014). The term "superclades" is used to describe the grouping of clades using a SNP-based phylogenetic tree.

To date, several *M. tb* clades (Beijing, Haarlem, etc.) have been described and classified as belonging to one of seven MTBC lineages (Blouin et al. 2012). While the two clades of *M. africanum*, namely West African-1 and West African-2, are localised to Ghana and surrounding West African countries, there is a distinct spread of the seven *M. tb* lineages around the world (Gagneux et al. 2006; Chihota et al. 2018). Despite being derived from a common mycobacterial ancestor, different members of *M. tb* have been shown to cause a range of clinical symptoms and may exhibit varying responses to drug therapy (Gutierrez et al. 2005; van der Spuy et al. 2009). The response of the human host to infection with *M. tb* is also known to vary greatly and may be attributed to genetic susceptibility factors of the host (Brosch et al. 2002; Coscolla and Gagneux 2010).

TB is a complex disease and despite many genetic studies aiming to identify susceptibility genes for TB, these investigations have had limited success and the factors predisposing to the disease remain largely unknown. Furthermore, outcomes of infection with a member of the MTBC depends on a number of factors, including the virulence of the bacterium, environmental conditions, and the susceptibility of the host to developing the disease. Numerous studies have investigated the associations between the host genotypes and TB, as a disease of interest (Herb et al. 2007; Thye et al. 2010; Bellamy et al. 1998). To this end, a number of host genetic factors have been identified to modulate susceptibility to the disease. However, to the limit of our current knowledge, no studies have been conducted to investigate the potential genome-wide associations between human genotypes and different members of the MTBC.

Following diagnosis with active TB, patients are admitted to a standard treatment program consisting of four first-line drugs (Isoniazid (INH), Rifampicin (RIF), Ethambutol (ETB), and Pyrazinamide (PZA)) taken daily for two months, followed by a four-month treatment with RIF and INH. To reduce the chances of further transmission of *M. tb*, treatment is started as soon as a case of TB is confirmed in adults. However, this is usually initiated prior to any drug-susceptibility testing, or genotyping of the infecting strain. This clinical practice, along with incorrect prescription of anti-TB drugs and patient non-compliance has led to the development of numerous multi-drug resistant, and extensively-drug resistant TB strains (Cohen and Murray 2004). In recent years, host-directed therapies (HDT) have been proposed as an adjunctive to traditional antimycobacterial therapy. In contrast to antimicrobials, which directly target the pathogen, HDTs aim to target the host's immune system in an effort to curb the progression of the disease, thereby avoiding the development of resistance to antimycobacterial drugs.

Different members of the MTBC have shown to dominate different regions of the world, and simultaneously have varying degrees of virulence affecting their ability to cause disease in humans (Gagneux 2012). However, geographical separation of the MTBC lineages are potentially also being driven by population-specific host-genetic factors influencing susceptibility to infection. Thoroughly understanding the genetic factors within the host driving *M. tb* strain diversity within a study population may offer insight into the disease mechanisms, and possibilities for host-directed therapies to combat the epidemic.

1.2. Problem identification

Numerous studies have reported a role for host genetic components in susceptibility to TB (Bellamy 1998; Moller and Hoal 2010; Kinnear et al. 2017). These predisposing genetic factors may explain why only a small proportion of immunocompetent individuals who have been infected with *M. tb* progress to develop active disease (Bloom and Murray 1992). In addition to this, investigations of host genetic factors involved in susceptibility to different strains of *M. tb* have recently gained traction (Salie et al. 2014; Kamgue Sidze et al. 2013; Intemann et al. 2009; Brown et al. 2010). Therefore, it is tempting to speculate that if certain genetic markers conferring susceptibility to particular *M. tb* strains are common within a given population, this may in part explain the variable success rate amongst different strains within the community (Hanekom et al. 2007).

Most studies investigating strain-specific genetic susceptibility to TB have used candidate gene study designs, while one recent study has used a genome-wide association analysis of susceptibility to different TB strains (Omae et al. 2017). In order to improve our understanding of the genetic susceptibility to TB clades, this study leveraged genome-wide genotyping data from the host and pathogen data to perform a genome-wide screen for *M. tb* clade-specific genetic associations in cohorts originating from two distinct populations.

1.3. Aims

The aim of this project was to investigate the association(s) between host genetic factors and the *M. tb* clade infecting the study participants.

1.4. Objectives

The objectives of the study were to:

1. Develop a method for performing a genotype-to-strain association analysis using the South African dataset as a test cohort as follows:
    1.1. Match genotyped study participants to their infecting *M. tb* isolates
    1.2. Define *M. tb* clade and superclade groupings using a SNP-based phylogenetic tree
    1.3. Perform a preliminary assessment using PCA
    1.4. Obtain high-quality imputed genotype data through testing of multiple reference panels on the study dataset
    1.5. Perform an association analysis using high-quality imputed host genotype data and *M. tb* data
2. Replicate the method on a second dataset, namely a dataset from Ghana

CHAPTER TWO

2. Literature Review

2.1. Prevalence

As the ninth leading cause of death, and the leading cause of death by an infectious disease, TB remains a global health concern, outranking death caused by HIV/AIDS alone (WHO 2017b). According to the 2016 WHO TB report, 10.4 million incidence cases of TB were recorded worldwide of which 90% of those infected were adults, and 65% were male (WHO 2017b). Despite advancements in TB control and treatment, an increase of 200 000 TB cases from the previous year was reported globally, along with an estimated 1.3 million deaths due to TB amongst HIV-negative individuals (WHO 2017b). Since the current study focusses on two independent cohorts from South Africa and Ghana, the paragraphs to follow will describe the TB epidemics in both countries.

In South Africa, TB remains a serious health concern, with the 2015 WHO report ranking South Africa sixth out of 22 high-burden countries. For most high-burden countries, the TB incidence rates range between 150 and 300 cases per 100 000 individuals annually. However, in 2014, South Africa saw a rate of more than 500 TB cases per 100 000 individuals, ranking second in incidence after Mozambique (WHO 2014a). The 2017 WHO TB country profile reported a marked decline in TB incidence for South Africa for 2016 (Table 1).

Table 1: Statistics for South Africa, and Ghana for the 2016 reporting year

	South Africa	**Ghana**
Population (in millions)	56	28
TB Incidence	180 000	34 000
Mortality (excludes HIV+TB)	23 000	10 000
New MDR cases	3.4%	1.5%
Reference	WHO 2017c	WHO 2017a

2.2. Socioeconomic and environmental risk factors

The increasing rates of TB globally can be attributed to a number of factors (Figure 1). Since the emergence of HIV/AIDs, the health sector has been burdened by a marked increase in TB incidence (Bloom and Murray 1992). Other co-morbidities increasing risk for developing active TB include diabetes (Jiménez-Corona et al. 2013; Prada-Medina et al. 2017; Restrepo et al. 2018) and autoimmune diseases such as rheumatoid arthritis (Gómez-Reino et al. 2003; Lim et al. 2017), liver cirrhosis (Lin et al. 2014), and cancer with some forms of the disease increasing the risk of developing active TB nine-fold (Cheng et al. 2016).

High humidity, poor house ventilation, and close contact with active TB patients have also been associated with increased risk of developing the active form of the disease (Pratiwi 2016; Seddon et al. 2013). Individuals living in highly-populated communities such as nursing homes (Stead et al. 1990), shelters, and jails (Coker et al. 2006) have also been shown to be at increased risk for contracting *M. tb.* Smoking, diabetes, and nutritional status are also known risk factors for developing the disease (Ramaliba et al. 2017; Cegielski et al. 2012). Low body mass index (BMI), reduced subcutaneous fat tissue, or low skeletal muscle volume have been reported to increase TB risk in individuals with normal nutritional status (Cegielski et al. 2012).

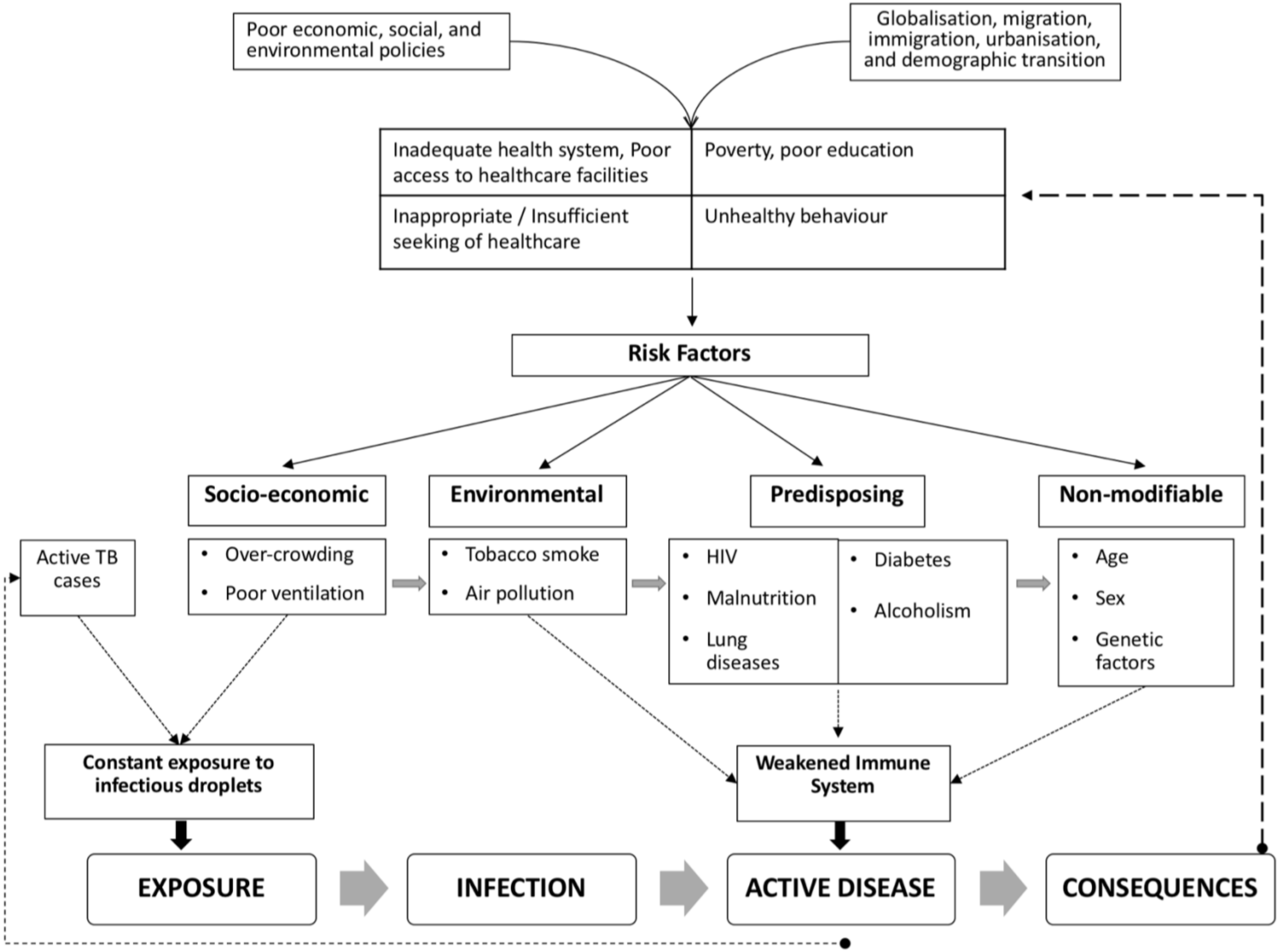


Figure 1: Risk factors driving the TB epidemic (adapted from Lönnroth et al. (2009))

2.3. Transmission, acquisition, and response of the human host to infection

TB is an airborne disease transmitted when aerosolised droplets containing the infectious *M. tb* bacillus are expelled from an infected person through coughing or sneezing and inhaled by an uninfected individual. Between one and 200 *M. tb* bacilli have been shown to cause disease in an exposed individual (Sakamoto 2012). However, as few as 10 bacilli have been shown to cause an infection in previously unexposed individuals (Behr et al. 1999). As a non-motile bacterium, *M. tb* requires a suitable host for successful transmission. To this end, *M. tb* has been hypothesised to have co-evolved with humans over thousands of years and as such, humans have become the ideal host (Comas et al. 2013; Hoal et al. 2017).

The response of the human immune system to infection with *M. tb* is highly variable, influenced by numerous factors including the presence of pre-existing infections, the virulence of the bacterium, environmental and host genetic factors. Macrophages are immune cells central to the protective response of the host against foreign pathogens such as bacteria, fungi, and viruses. Following inhalation of *M. tb* into the lungs, pattern recognition receptors such as toll-like receptors (TLR) present on the cell surface of alveolar macrophages recognise and target the *M. tb* for a process of degradation known as phagocytosis (Thoma-Uszynski 2001). The combined efforts of these inflammatory molecules and immune cells results in the successful phagocytosis of the bacteria and formation of a well-organised collection of immune cells and degrading bacteria known as the granuloma (Davis and Ramakrishnan 2009).

The containment of the bacteria within the macrophage elicits the activation and release of various proinflammatory cytokines such as interferon-gamma (IFN-γ), and several forms of tumour-necrosis factor (TNF), further promoting the recruitment of leukocytes, monocytes, and neutrophils to the site of infection. Dendritic cells, like macrophages, play crucial roles in immunity against TB, and following the phagocytosis of *M. tb,* migrate to regional lymph nodes to promote the recruitment of lymphocytes to the site of infection, and the subsequent release of IFN-γ (Stein et al. 2003). This cytokine promotes the induction of the autophagy process in which the *M. tb*-containing macrophage is targeted for degradation by means of p47 GTPase activity (Gutierrez et al. 2004).

The formation of the granuloma is beneficial to both the host and the bacterium. For the host, the granuloma signifies the successful capture of a pathogen. The bacteria, however, also find the granuloma to be a safe location as they are protected from the pro-inflammatory cytokine, IFN-γ (Sakamoto 2012). Subsequent to engulfing a foreign organism, the macrophage forms a phagosome, which is able to fuse with the lysosome, creating an acidic phagolysosome that facilitates the degradation of the pathogen. However, *M. tb* are able to modify this acidification process by interfering with the production of the vacuolar proton ATPase and subsequently are able to survive within the phagosome before escaping into the cytosol for rapid replication (Sturgill-Koszycki et al. 1994). The relationship between *M. tb* and the human host is complex, seen as a complicated game of tug-of-war and co-evolution (Brites and Gagneux 2015; Comas et al. 2013; Hoal et al. 2017).

2.4. Clinical presentation, Diagnosis and Treatment of TB

In 1993, the WHO declared TB to be a Global Emergency and in September 2000, included TB treatment strategies as a global health priority in the Millennium Development Goals (WHO 2014b). Since then, extensive TB research has led to the development of numerous treatments, resulting in the decline of TB incidence in many developed countries.

The clinical manifestation of *M. tb* infection can be classified by its anatomical involvement as either pulmonary TB, with primary involvement of the lungs, or as extrapulmonary TB, which manifests in various organs and tissues outside of the lung, including the brain, gastrointestinal tract, lymph nodes, skin, and joints (WHO 2013). TB is also classified according to the disease state of the host as either active TB, or latent TB infection, described in more detail below (Al-Orainey 2009).

The Bacille Calmette-Guérin (BCG) vaccine was developed between 1906 and 1919 through numerous rounds of passaging of *M. bovis* on a growth surface consisting of ox-bile and potato slices soaked in glycerol (Sakamoto 2012). However, despite the widespread use of the BCG vaccine in children, TB incidence has continued to increase in many developing countries (WHO 2017b). Although BCG is commonly used to prevent TB in children, it has been unreliable in preventing disease in adults (Brosch et al. 2007).

2.4.1. Latent TB

After infection with *M. tb*, the majority of individuals will remain asymptomatic and contain the bacterium, and enter a stage termed latent TB infection (LTBI). These individuals do not exhibit the clinical symptoms of the disease due to the mycobacteria not being in an actively replicating state, but remain at risk for developing the disease through endogenous reactivation (Vynnycky and Fine 2000). This large reservoir of individuals harbouring "dormant" mycobacteria have been hypothesised to, with sudden sufficient immunosuppression, become a significant source for active TB cases (Lin and Flynn 2010).

LTBI is at present inferred from measures of acquired anti-mycobacterial immunity, such as a tuberculin skin test (TST) and/or interferon gamma release assay (IGRA). The TST (also known as the Mantoux test) is used to test for exposure to *M. tb* antigens. An intracutaneous injection of 0.1 ml of tuberculin is administered, followed by visual inspection and the measurement of induration by a clinician (Ayub et al. 2004; Nayak and Acharjya 2012). This visual measurement of induration is highly subjective, leading to large margins of error, variability in the interpretation and can result in both false-positive and false-negative diagnosis. False-positive results have been identified in patients with previous exposure to non-tuberculous mycobacteria as well as individuals with prior vaccination with BCG, particularly when administered for the first time at school-going age, or as multiple booster shots (Farhat et al. 2006).

The IGRA is an *in vitro* assay, used to quantitatively evaluate the response of the host's cell-mediated immunity to *M. tb* bacilli. Although useful in confirming the results of a TST, the IGRA, like the TST, cannot differentiate between latent infection with *M. tb* and reactivity due to vaccination with BCG (Mandalakas et al. 2008). However, unlike the TST, an IGRA assay does not require a follow-up assessment, and has demonstrated a high degree of specificity in regions experiencing low TB incidence (Sester et al. 2011). Two commercially available IGRAs are the QuantiFERON®-TB Gold In-Tube (QFT) assay, and the T-SPOT.*TB* assay. Both of these assays are enzyme-linked immunosorbent assay (ELISA)-based and use the mycobacterial antigens, early secretory antigenic target 6kD (ESAT-6) and culture filtrate proteins (CFP-10), to induce an immunological reaction (Horvat 2015). While the QuantiFERON® assay directly measures the amount of IFN-γ produced, the T-SPOT.*TB* assay measures the number of IFN-γ-producing T-cells. (Pai et al. 2014).

In low incidence countries such as the United States, the decision to test for latent TB is preceded by the decision to treat if the test outcome is positive (Schluger and Burzynski 2010). However, this approach is not highly favoured as it comes coupled with social consequences such as stigma (Daftary et al. 2017), effects on health due to administering of anti-TB drugs with unfavourable side-effects, and a financial strain on both the patient and the health sector (van’t Hoog et al. 2014). Owing to the high cost of screening a large proportion of the population for a low yield of patients requiring TB treatment, most countries have taken the decision to only screen for TB in high-risk individuals, preventing a generalised treatment for latent TB infection (van’t Hoog et al. 2014). However, the risk for initiating an outbreak of active TB as harboured by latently infected individuals may be considered sufficient to justify testing individuals with latent TB residing in a high-burden setting. By treating individuals latently infected with *M. tb*, the number of active TB cases may be minimised and the epidemic curbed (Lin and Flynn 2010).

2.4.2. Active TB

Active pulmonary TB is diagnosed in patients presenting with symptoms of infection with *M. tb* which include fever, a persistent cough, and significant weight loss, as well as being culture-positive for the *M. tb* bacteria, while patients with latent TB do not have any clinical signs of the disease (Cohen et al. 1996). Diagnosis of active TB can generally be placed into three categories: 1) Radiological methods, 2) Smear microscopy and culture, and 3) Molecular methods. Radiological examination may be conducted using the standard chest X-ray or chest computed tomography. Both technologies are able to provide a visual assessment of internal lung structures, but are insufficient as a sole means of diagnosing active TB due to the chest X-ray being able to visualise cavitary lesions in some, but not all patients with active TB (Krysl et al. 1994). Thus, active TB still needs to be confirmed through examination of a sputum sample for the presence of *M. tb* bacilli.

Sputum smear microscopy is the most widely-used and accessible method for the detection of *M. tb* in patients suspected of having TB. However, this method requires the patient to be present at the healthcare facility over several consecutive days to provide the multiple sputum samples needed for testing, and is thus a costly and inconvenient way of being tested for TB (Parsons et al. 2011). To obtain a sample for staining, individuals suspected of having active TB are required to produce sufficient sputum for staining of the acid-fast bacilli (AFB) using Ziehl-Neelsen staining (Ryu 2015). This has been hard to achieve in children, and HIV-positive individuals, as they do not usually produce a sufficient sample and even when more invasive methods such as bronchoalveolar lavage (BAL) or gastric aspiration, are used, 95% of cases in children were AFB smear-negative, and the diagnosis inconclusive (Starke and Taylor-Watts 1989).

Although direct examination of sputum samples using microscopy techniques is useful for identifying the presence of *M. tb*, the method is unable to differentiate between drug-susceptible and drug-resistant strains. To achieve this, *M. tb* can be cultured on various forms of culture media, such as Lowenstein-Jensen (LJ) media in the solid, slant, or broth form. On solid media, culturing of *M. tb* can take between two to four weeks for samples that were *M. tb* positive during microscopic examination, while microscopy-negative sputum samples may take up to two months for successful culture of the bacterium (Ryu 2015). This lengthy time for culture drastically impedes rapid diagnosis of patients with active TB, delaying treatment and enabling the transmission of the pathogen during a highly infectious stage. The culturing of *M. tb* on LJ media is however no longer common practise, and has been replaced with the mycobacterium growth indicator tube (MGIT) system which detects the growth of mycobacteria in culture via fluorescence of an oxygen sensor (Tortoli et al. 1999).

Lastly, three molecular methods are commercially available for TB testing. Nucleic acid amplification (NAA) is able to detect the presence of *M. tb* weeks before a diagnosis is confirmed by culture. Despite its rapidity, currently used NAA tests are not recommended in cases where evidence for TB is low as the positive predictive value for the test has been shown to be less than 50% in such cases (American Thoracic Society 2000). The line probe assay (LPA) is another molecular diagnostic technique which rapidly tests for drug susceptibility in *M. tb* (WHO 2008). The test assesses the potential survival of the bacterium in response to two first-line anti-TB drugs, INH and RIF, by testing for the presence of a wild type or mutant allele which confer these drug-resistance capabilities. The GeneXpert MTB/RIF Ultra assay developed by Cepheid is an automated NAA test (WHO 2013). The cartridge-based assay is able to offer a TB diagnosis, as well as RIF resistance status, as early as two hours after sample collection. A benefit that this technique offers over all others is that the GeneXpert cartridges are preloaded with all the reagents necessary for the assay, thus requiring very little hands-on time, and subsequently optimising TB diagnosis (WHO 2013).

Four first-line TB drugs currently in use are EMB, INH, PZA, and RIF. The TB treatment strategy consists of a two-month schedule consisting of all four drugs, followed by four months treatment with RIF and INH (WHO 2013). To prevent the emergence of new antimicrobial-resistant strains, patients are required to diligently complete the treatment regimen consisting of the four drugs taken daily for a period of six to nine months. Though effective in curbing the emergence of MDR- and extensively drug-resistant (XDR) strains, the chemotherapeutic nature of the drugs has been shown to be toxic to patients, perhaps contributing to a decrease in patient compliance (Gülbay et al. 2006).

Patients presenting with microbial resistance to INH and RIF are classified as MDR and may need up to two years of treatment with fluoroquinolones and aminoglycosides to completely eradicate *M. tb*. Resistance to first-line anti-TB drugs, fluoroquinolones, and aminoglycosides in patients is classified as XDR cases and requires treatment with drugs that are much more expensive than first- and second-line drugs, have shown to produce more negative side-effects, and have been associated with more instances of poor patient outcomes (Pietersen et al. 2014). Total drug resistant (TDR) strains of *M. tb* are identified by resistance to all first-line and second-line anti-TB drugs and have been confirmed in Iran, India, and South Africa (Migliori et al. 2007; Udwadia et al. 2012; Velayati et al. 2009; Klopper et al. 2013). In 2013, a study identified several patients infected with an atypical Beijing genotype clone in South Africa, notably developing resistance to all first-line drugs, fluoroquinolones, and aminoglycosides, amongst other second-line drug therapies (Klopper et al. 2013). This observation is of particular significance due to South Africa experiencing a high burden of TB caused by members of the Beijing genotype (van der Spuy et al. 2009; Chihota et al. 2018).

2.5. Genetics of TB susceptibility

2.5.1. Studies investigating genetic susceptibility to TB

In addition to socio-economic and environmental factors, and the presence of predisposing diseases, the genetic make-up of the human host has also been shown to play a significant role in determining susceptibility to a disease. Numerous studies have unequivocally shown associations between genomic loci and susceptibility to infectious diseases such as malaria (Rockett et al. 2014), HIV (Pastinen et al. 1998) and TB (Herb et al. 2007; Thye et al. 2010; Bellamy et al. 1998).

Some of the earliest events alluding to a human genetic contribution to TB susceptibility were tragic events in history and claimed many lives. The Qu'Appelle population of the Sasketchewan province in Canada were heavily impacted following their first exposure to *M. tb* which resulted in an annual loss of up to 10% of the population (Motulsky 1960). During the first three generations following the arrival of the bacterium to the community, more than half of the families had succumbed to the disease. After this initial period during which most of the susceptible individuals had died, the annual death rate as caused by TB was reduced to 0.2%, possibly owing to a selection against the susceptibility genes for TB within the population (Motulsky 1960).

The Lübeck disaster was another tragic event which provided early evidence for genetic components playing a role in susceptibility to *M. tb* infection. Instead of receiving the attenuated strain for vaccination, a total of 252 infants were accidentally injected with a BCG vaccine contaminated with virulent *M. tb*. The infants displayed variable responses to the bacterium where 108 of the infants had signs of TB and survived, while 67 infants died as a result of developing active TB (Rieder 2003; "The Lübeck Catastrophe: A General Review" 1931). Besides the possibility of genetic components being responsible for the variable response to infection with *M. tb*, the vaccine vials administered to the infants also contained variable dosages of the infectious bacterium. This became evident when infants receiving lower dosages of the bacterium were recorded to have a wide range of clinical phenotypes, while those who received a higher dosage were highly susceptible to developing TB, indicating the apparent ability of the innate immune system to control infection caused by low doses of *M. tb* (Fox et al. 2016).

Early studies involving twins provided evidence for genetic components modulating susceptibility to TB. A higher degree of concordance for disease was found in monozygotic twins compared to dizygotic twins (Kallmann and Reisner 1943). Although these observations were substantiated during a reanalysis of the Prophit study (Comstock 1978), this study did not consider the possible confounding effects of environmental factors. Thus, results from a comparison of hereditary factors with environmental factors concluded that environmental factors such as the number of bacilli during transmission were of more significance than genetic factors of the host in determining progression to disease (van der Eijk et al. 2007).

Numerous studies have proceeded to investigate the risk of disease amongst individuals living in close proximity to TB patients. An early observation of TB in families reported that individuals who were spouses to TB patients, and came from a family with a history of TB themselves, were at increased risk for developing the disease than spouses with no family history of TB (Puffer 1944). Another study of TB cases in a nursing home in the USA revealed that individuals with African ancestry were more likely than those of European ancestry to be infected with *M. tb*, even when they were living in the same environment ((Stead et al. 1990) as reviewed in (Kinnear et al. 2017)).

A number of approaches have been used to investigate the observed differences in the genetic susceptibility of the human host to *M. tb*. These include genome-wide linkage analyses, candidate-gene association studies, and genome-wide association studies, and will be discussed in the paragraphs to follow (reviewed in (Möller et al. 2010; Abel et al. 2017; Kinnear et al. 2017)).

2.5.2. Whole-Genome Linkage Studies

Linkage studies interrogating the whole genome aim to trace the inheritance of chromosomal regions harbouring putative susceptibility genes and have proven to be highly successful when examining monogenic diseases (Ferreira 2004), while associations found for complex diseases such as TB have been difficult to replicate (reviewed in (Altmüller et al. 2001)). In a TB linkage study conducted by Jamieson and colleagues, four genes located on chromosome 17q showed individual effects associated with modifying susceptibility to TB in a cohort of Brazilian patients (Jamieson et al. 2004). Another study conducted on an extended Aboriginal Canadian family of 81 members reported linkage between a TB-susceptibility locus and *D2S424*, a gene in close proximity to the natural resistance associated macrophage protein-coding gene (*NRAMP1)*, while there was no association with the human leukocyte antigen (*HLA*) class of genes - a complex well-known to play a role in the progression of TB (Greenwood et al. 2000).

A genome-wide linkage analysis of 92 sibling-pairs from The Gambia and South Africa revealed suggestive evidence of linkage to loci on 15q and Xq, and TB (Bellamy et al. 2000). A linkage study conducted in a Ugandan population identified chromosomal regions 2q21, 2q24, 5p13, and 5q22 as being associated with TST negativity (Stein et al. 2008), while in a South African cohort, reactivity to the TST test was linked to the *TST1* chromosomal region 11p14 and *TST2* located on chromosome 5p15 (Cobat et al. 2009), and was replicated in a French cohort (Cobat et al. 2015). Two loci on chromosomes 3q and 8q were associated with modulating the production of IFN-γ via the ESAT-6 pathway (Jabot-Hanin et al. 2016), while in a Peruvian population, variants on chromosome 3q23 were shown to be associated with early progression to active TB (Luo et al. 2018).

2.5.3. Candidate-gene studies

Early candidate-gene association studies have been successful in providing new clues to TB susceptibility as the study design allows for the investigation of associations between genes pre-selected on the basis of their biological importance to disease mechanisms, as well as to the disease of interest. The method compares allelic and genotyping frequencies of a specific genetic marker between a group of unrelated cases and controls. One caveat to this method, however, is that it requires the genotype distribution of a particular marker in the control cohort to be in “Hardy-Weinberg Equilibrium” (HWE) (Schaid and Sommert 1993). HWE is a feature of a population where genotype and haplotype frequencies remain constant from one generation to the next in the absence of migration, natural selection, assortative mating, or mutation (Wigginton et al. 2005). Despite these factors being difficult to control for, most populations generally appear to adhere to the expected allele frequencies, and deviations from HWE at a particular locus may be suggestive of genotyping errors, extensive population structure due to admixture, or may appear in the affected individuals, thereby indicating an association between the marker and the disease under study (Wigginton et al. 2005).

Numerous genes found throughout the genome have been shown to play a role in susceptibility to TB. Genes encoding a number of proteins such as the HLAs, *NRAMP1*, mannose binding lectin (*MBL*), IFN-gamma, and Vitamin D Receptor (*VDR*) have been associated with variations in susceptibility to TB (Bellamy et al., 1998; Søborg et al., 2003; Yim and Selvaraj 2010). Amongst the genes evaluated are many with key roles in the functioning of the immune system such as those belonging to the *HLA* complex, *NRAMP1 (SLC11A1)*, and IFN-γ.

A large cohort of 1 916 sputum-positive pulmonary TB patients from Ghana were genotyped for the *ALOX5* g.760G>A variant and individuals who were heterozygous for the polymorphism were found to be at increased risk for developing TB. Furthermore, harbouring this exonic variant had a greater association (OR= 1.70; (95% CI: 1.2–2.6)) with infection caused by *M. africanum* West African-2 (Herb et al. 2007). Modelling a recessive mode of inheritance, a protective association (OR= 0.60; (95% CI: 0.4–0.9)) was identified between the occurrence of TB and the *MBL2* G57E variant in a cohort of Ghanaian patients (Thye et al. 2011). TB patients belonging to the Ewe ethnic group were significantly more likely to be infected with *M. africanum* (OR= 3.02; (95% CI: 1.67–5.47)) and further stratification by lineage revealed that the association was strongly driven by infection with members of *M. africanum* West Africa 1 (Asante-Poku et al. 2015).

Using a candidate gene approach, polymorphisms in the *CCL2* and *NOS2A* genes were investigated for more than 800 cases and controls belonging to the South African Coloured (SAC) population and the T allele of one SNP, rs8078340, was found to be significantly associated with having TB (OR=1.4; 95% CI: 1.1–1.8) (Möller et al. 2009). A recent study by Hong and others investigated a Korean cohort of 46 cases and 1 313 controls for genome-wide associations to TB. Although the authors were unable to identify novel SNPs significantly associated with the disease, the study was able to replicate associations between pulmonary TB and ten SNPs in, or in close proximity to a number of immune-related genes, as previously identified in another Korean cohort (Hong et al. 2017).

A limitation of the candidate-gene study design, however, is that it requires an *a priori* hypothesis regarding genes to target in the association analysis. To tackle this limitation, genome-wide association studies (GWAS) have become a popular alternative for identifying genetic associations with disease. Through genotyping of many common genetic variants, GWA studies enable a global interrogation of an individual's genome for associations to disease, without the limitation of predefined candidate genes (Hirschhorn and Daly 2005).

2.5.4. Genome-wide Association Studies (GWAS)

GWA studies aim to identify SNPs that differ in frequency between disease cases and well-matched controls. To do so, participants are genotyped at 100s of 1000s to millions of pre-selected SNPs spanning the entire genome. As GWAS are hypothesis-generating, careful consideration is required when selecting variants to be included on the genotyping array. GWA studies focus on a notion of "common diseases harbour common variants" and thus generally focus on variants with frequencies greater than 5% in the population (Reich and Lander 2001). Given an adequate sample size, GWAS have greater power to detect genetic associations with small effects, compared to linkage studies (Risch and Merikangas 1996).

The first TB GWAS reported an association between a region found on chromosome 18q11.2 and TB susceptibility in a case-control study of TB patients from The Gambia and Ghana (Thye et al. 2010). Following this, 13 other TB GWAS have been performed (Table 2) and hold promise for refining the methods used to identify TB-related genetic associations (Uren et al. 2017).

In a modification of the traditional GWAS study design, Daya and others sought to identify interactions between gene pairs which may influence susceptibility to TB in the SAC population. The *IL23R-ATG4C*, *GRIK1-GRIK3*, *and NRG1-NRG3* gene pairs were found to potentially be involved in susceptibility to TB. Various models of these three gene pairs were successfully validated in a secondary dataset from The Gambia (Daya et al. 2015).

Table 2: Results of previous TB GWAS studies as summarised by Uren et al. 2017

Population	Variant /Gene	Number of Cases	Number of Controls	Reference
Ghana The Gambia	rs4331426 (gene desert)	921 1 316	1 740 1 382	(Thye et al. 2010)
Black, White, Asian from USA	rs4893980 (PDE11A) rs10488286 (KCND2) rs2026414 (PCDH15) rs10487416 (unknown gene)	48	57	(Oki et al. 2011)
Thai and Japanese	Intergenic region between HSPEP1-MAFB	620	1 524	(Mahasirimongkol et al. 2012)
Indonesia	rs1418267 (TXNDC4) rs2273061 (JAG1) rs4461087 (DYNLRB2) rs1051787 (EBF1) rs10497744, rs1020941 (TMEFF2) rs188872 (CCL17) rs10245298 (HAUS6) rs6985962 (PENK)	108	115	(Png et al. 2012)
Ghana The Gambia	rs2057178 (WT1, intergenic)	2 127 1 207	5 636 1 349	(Thye et al. 2012)

Russia		1 025	983	
Indonesia		4 441	5 874	
South African Coloured	rs2057178, rs11031728 (WT1, intergenic) rs10916338, rs1925714 (RNF187) rs6676375 (PLD5) rs1075309 (SOX11) rs958617 (CNOT6L) rs1727757 (ZFPM2) rs2505675 (LOC100508120) rs1934954 (CYP2C8) rs12283022, rs12294076 (DYNC2H1) rs7105967, rs7947821 (DCUN1D5) rs6538140 (E2F7) rs1900442 (VWA8) rs17175227 (SMOC1) rs40363 (NAA60) rs2837857 (DSCAM) rs451390 (C2CD2) rs3218255 (IL2RB)	642	91	(Chimusa et al. 2014)
Russia	rs4733781, rs10956514, rs1017281, rs1469288, rs17285138, rs2033059,	5 530	5 607	(Curtis et al. 2015)

	rs12680942 (ASAP1)			
Morocco	rs358793 (Intergenic) rs17590261 (Intergenic) rs6786408 (FOXP1) rs916943 (AGMO)	556	650	(Grant et al. 2016)
Uganda and Tanzania	rs4921437 (IL-12)	267	314	(Sobota et al. 2016)
Iceland	rs557011, rs9271378 (located between HLADQA1 and HLA-DRB1) rs9272785 (HLA-DQA1)	8 162	277 643	(Sveinbjornsson et al. 2016)

2.5.5. Controlling Population Stratification in a GWAS

Some of the limitations of GWAS studies include insufficient sample sizes, poor definitions of case and control groups, and controlling for population stratification. In addition, GWAS have inherent statistical challenges due to the large number of variants being examined within a large cohort under study. Thus, one of the major limitations of GWAS is the possibility of false-negative or false-positive associations being detected, which may be mitigated by statistically correcting for multiple testing (Visscher et al. 2012).

Population genetic variation, commonly termed "population stratification" results from the admixture of different founder populations, and is evidenced by differences in allele frequencies in subpopulations (Cardon and Palmer 2003). If not corrected for during the statistical analyses, population stratification may result in false associations with the disease of interest (Oetjens et al. 2016; Daya et al. 2013).

A number of software programs have been developed which allow for the quantification of population stratification for inclusion as a covariate in the statistical analysis phase of GWAS and candidate gene association studies. These software programs include among others, EIGENSTRAT (Price et al. 2006), ADMIXTURE (Alexander et al. 2009), RFMix (Maples et al. 2013), and STRUCTURE (Pritchard et al. 2000). The inclusion of principal components calculated from the genotype data is also a valid method for correcting for population stratification (Price et al. 2006).

Population stratification resulting from admixture is an important factor which has been shown to confound the results reported in many genome-based association studies. As most published GWAS have been performed on populations originating from Europe, replication of GWAS results obtained from these populations in other distinct populations have proven difficult due to population stratification confounding results of association analyses (Need and Goldstein 2009). It has therefore not been established as to what degree European GWAS data can be used to infer the population structure and allelic frequencies in cohorts belonging to historically older populations such as those originating from Africa (Martin et al. 2017). To correct for the confounding effect of variable ancestry, ancestry proportions need to be derived for the population of interest using software such as those described and included as covariates in the statistical analyses (Pearson and Manolio 2008).

The SAC population is a highly admixed population historically residing in the Western Cape Province of South Africa. In this province, the SAC population makes up a significant proportion of the local population; the 2011 Census reported the "Coloured" population comprising 42.4% of the population residing in the Western Cape (Strategic Development Information and GIS Department and City of Cape Town 2012). An analysis of the population substructure of the SAC population, showed that the five-way admixed population has genetic contributions from the Khoesan Africans, non-Khoesan Africans, European, South and East-Asian populations (De Wit et al. 2010). These ancestry proportions can be accurately determined using a set of 96 ancestry informative markers (AIMs) in individuals belonging to the SAC population, and furthermore serves to adjust for confounding when included as covariates in an association analysis (Daya et al. 2013).

The software package, PROXYANC, was also developed to provide a platform for identifying the best reference populations for local ancestral contributions within a five-way admixed population such as the SAC (Chimusa et al. 2013). Selecting a reference population for a complex admixed population such as the SAC is not an easy task, and poorly selected references may negatively impact the statistical power to detect an association. PROXYANC serves to improve the selection of appropriate ancestral populations in admixture mapping studies and imputation of missing data in admixed genotypes (Chimusa et al. 2013).

The SAC population presents with a unique genotype composition, with individuals predominantly located within an environment of high TB incidence. The combination of complex admixture and observed high TB incidence may contribute to the TB burden seen in the Western Cape region of South Africa. It also serves as an ideal highly admixed population for studying the genetic susceptibility to the many *M. tb* clades causing disease in this region, presenting an opportunity for unique studies of association.

Chimusa and colleagues performed a GWAS on a cohort belonging to the SAC population in which genotypes were imputed with the HapMap3 release 2 and 1000 Genomes Project (1000GP) reference populations (Chimusa et al. 2014). Using a mixed model approach, the authors aimed to replicate TB susceptibility loci identified from previous studies, in this population (Chimusa et al. 2014). The authors were able to replicate the susceptibility locus (rs2057178) located in the *WT1* gene on chromosome 11, as identified in cohorts of TB patients in The Gambia, Indonesia, and Russia (Thye et al. 2012). This study by Chimusa described evidence that disease associations may be stratified by ethnicity and strongly influenced by ancestry.

A recent post-GWAS analysis was performed to predict regulatory variants potentially associated with susceptibility to TB. Using bioinformatics analyses incorporating RegulomeDB and Ensembl's Variant Effect Predictor, the authors fine-mapped six statistically significant novel intronic polymorphisms not identified by previous GWAS studies performed in the same cohort (Uren et al. 2017). The study showed that beyond the results of a GWAS, further bioinformatic analyses hold potential for identifying novel disease variants *in silico*.

To address the TB epidemic, multiple avenues of research need to be conducted (Stein 2011). To this end, numerous studies have focussed on finding genetic associations with the disease and have provided promising candidates for further investigation as TB biomarkers and targets for host-directed therapies, such as the granuloma, induction of autophagy, and anti-inflammatory responses (Kolloli and Subbian 2017). One area of TB research sparsely investigated is the potential associations between the human genome and different strains of *M. tb* circulating within high-risk communities. Understanding the biological mechanisms driving this complex relationship between host and pathogen will improve our existing knowledge of the disease and provide further avenues of exploration for improving the lives of millions affected by TB.

2.6. Genetic associations with different *M. tb* strains

2.6.1. Origin and epidemiology

Several species within the MTBC are known pathogens and include *M. tb, M. africanum, M. bovis, M. canetti, and M. microti*. and despite sharing 99.9% nucleotide similarity, have shown stark variations in their host of choice, level of pathogenicity, and phenotype (Brosch et al. 2002). The MTBC has been described as consisting of eight lineages: seven lineages deriving from a common ancestor are adapted to cause disease in humans, and one lineage is known to cause TB in animals (Comas et al. 2013; Wirth et al. 2008). Phylogenetic studies have demonstrated that based on the presence or absence of a gene deletion, and the fact that horizontal gene transfer is unheard of in members of the MTBC, the lineages can be classified as 'modern' or 'ancient' (Brosch et al. 2002). Strains considered to have evolved more recently and are therefore considered more 'modern', are classified as such based on the deletion of the

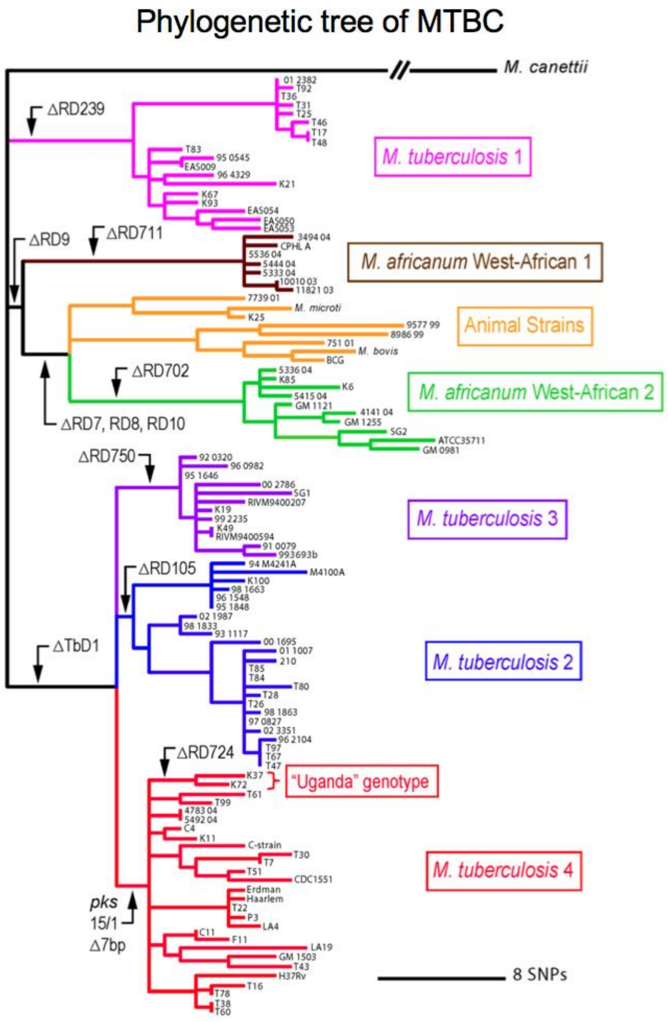
TbD1 region and presence of the RD9 region, whereas strains in which the TbD1 region had not been deleted, but were lacking the RD9 region are considered to be evolutionarily ‘ancient’ strains (Brosch et al. 2002; Gutierrez et al. 2005).

The seven human-adapted MTBC lineages comprise five members belonging to the *M. tb sensu stricto* species as well as two members belonging to the *M. africanum* species. Several human-adapted strains have been classified as belonging to one of three ‘modern’ lineages based on a deletion of the TbD1 gene. Lineages 2, 3, and 4 geographically dominate most of East-Asian, East-African, European and American countries, respectively (Figure 2).

The animal-adapted species, namely *M. bovis, M. canetti, and M. microti*, as well as human-adapted strains belonging to Lineages 1, 5, 6, and 7, have been classified as ‘ancient’ members of the MTBC. Lineage 1, also known as the Indo-Oceanic lineage, is found mainly in populations residing in the Indian Ocean and the Philippines (Figure 2). Lineages 5 and 6 are composed of the *M. africanum* West-Africa 1 and West-Africa 2 strain families, respectively (Figure 2). The geographical isolation of these two lineages to West-African countries may possibly be explained by the strains having more time than the ‘modern’ lineages to establish itself as a pathogen causing stable disease in a sympatric population which has co-evolved to survive infection with the bacterium. Lineage 5 and 6 have also been established as having high genomic variability, possibly contributing to drug-resistance mechanisms, immunogenic effects, and protein secretion (Ates et al. 2018; Otchere et al. 2018).

A novel lineage, Lineage 7, was recently defined as a predominant lineage in Ethiopia and other countries in the Horn of Africa (Firdessa et al. 2013). A total of 36 isolates obtained from patients with lymph node or pulmonary TB presented with previously undefined spoligotyping patterns. Although the isolates had intact TbD1 deletions similar to that of the modern lineages, they had other distinguishing features which led to their classification as belonging to the novel Lineage 7, which localises phylogenetically between the ancient Indo-Oceanic lineage 1, and modern TB lineages 2, 3, and 4 (Blouin et al. 2012; Firdessa et al. 2013). Strains belonging to Lineage 7 reportedly grew slower than other strains of *M. tb* and patients infected with this lineage took a median of three weeks from the onset of symptoms to report to their health care provider, with the authors hypothesising that Lineage 7 strains caused milder symptoms than the other, more virulent strains (Yimer et al. 2015). This intermediary effect of the bacterium on the host may possibly be explained by their phylogenetic localisation between the ancient and modern strains.

In contrast to the known pathogenicity of the ‘ancient’ strains (Gonzalo-Asensio et al. 2018), the evolutionarily ‘modern’ strains such as members of the Beijing family (lineage 2) and Haarlem family (lineage 4) have been described as hypervirulent and capable of causing severe disease in susceptible populations (Brites and Gagneux 2015; Hoal et al. 2017). Members of these two lineages have been a major cause for concern as they have been implicated in drug-susceptible and drug-resistant outbreaks of the disease.



Lineage	Spoligotyping	Geographical location
M. tb. Lineage 1	*EAI*	East Africa, India
M. af. Lineage 5	*WA1*	Eastern part of West Africa
M. af. Lineage 6	*WA2*	Western part of West Africa
M. tb. Lineage 3	*CAS*	East Africa, North India, Pakistan
M. tb. Lineage 2	*Beijing*	South Africa, East Asia, Russia
M. tb. Lineage 4	*Haarlem, LAM, X, T*	Europe, Americas, Throughout Africa, Australia

Figure 2: The phylogeny of members of the MTBC as derived using Maximum Parsimony, as published in de Jong et al. (2010), adapted from Hershberg et al. (2008). The figure is further adapted using the phylogeography as published in Gagneux and Small (2007) with Spoligotyping annotation from Brudey et al. (2006). The recently described Lineage 7 is not annotated on this phylogenetic tree.

Global TB outbreaks have been noticeably dominated by specific lineages in defined geographical regions (Gagneux and Small 2007). This observation has been substantiated by the dominance of all six *M. tb* lineages across the African continent, while in Asia, *M. tb* lineages 5 and 6 are largely absent (Figure 3). Continental separation of MTBC lineages has been observed, with Lineage 4 dominating most of Europe, the Americas and Australia. East Asia is dominated by four of the six lineages, whereas all six lineages can be found within the African continent (Gagneux and Small 2007; Chihota et al. 2018).

Advancements in technology and accessibility to international modes of travel have greatly increased the transmission of many communicable diseases, including TB (Knobler et al. 2006; Zumla et al. 2016). However, the geographical dominance of *M. tb* lineages in different parts of the world contradicts the idea that all strains have equal access to different populations across the globe. Thus, it has been strongly suggested that the genetic make-up of the host may play a significant role in defining the susceptibility to strains found to infect a given population (For Reviews see: (Abel et al. 2017; Meyer and Thye 2014; Moller and Hoal 2010)).

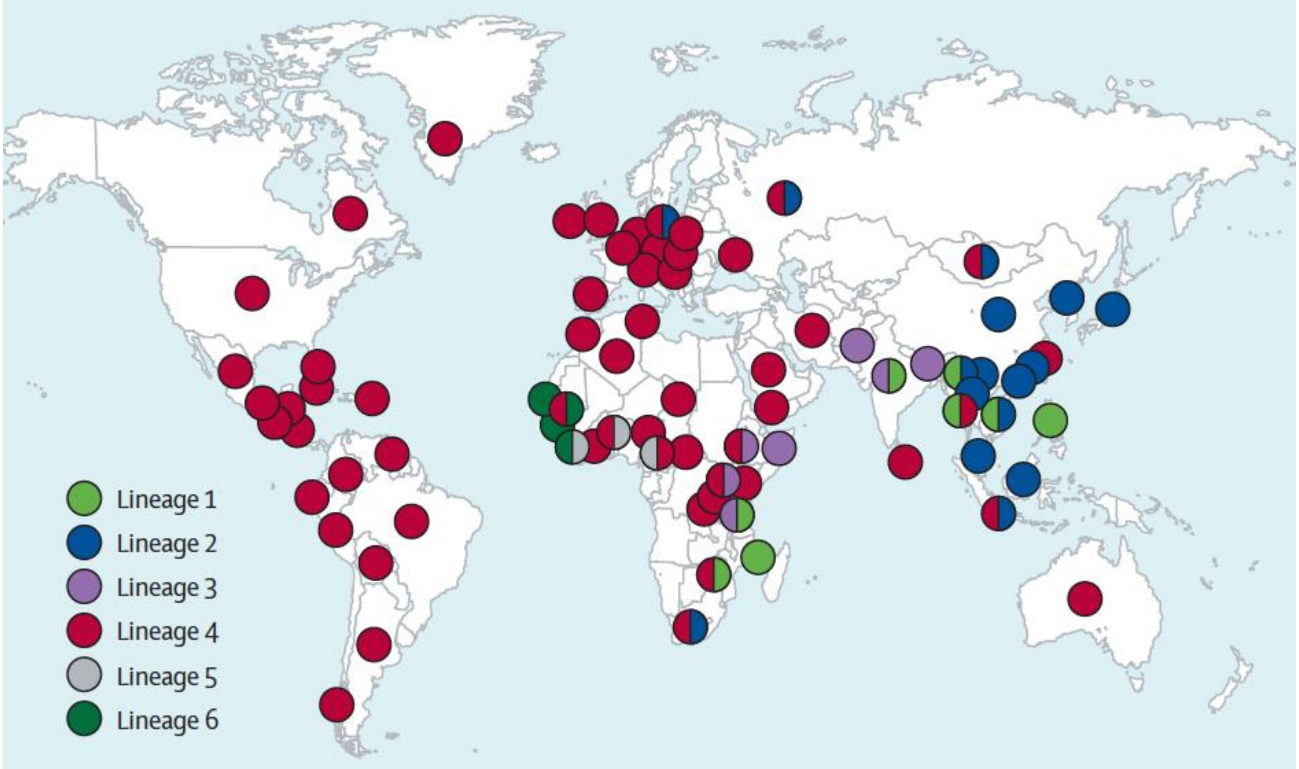


Figure 3: Global distribution of dominant M. tb phyla as originally published by Gagneux and Small (2007). Southern Africa is dominated by lineages 2 and 4 comprising Beijing, Haarlem, LAM, and T, whilst West Africa is dominated by lineages 5 and 6, namely WA1 and WA2, with interplay from Lineage 4.

2.6.1.1. South Africa

South Africa is dominated by lineages 2 and 4 comprising the Beijing, Haarlem, LAM, and T clades (Figure 3). A recent study by Sekati and others investigated the drug-susceptibility patterns and spoligotype distributions of *M. tb* amongst 104 clinical isolates from children in several provinces in South Africa including Gauteng, the North West, Limpopo, and Mpumalanga. The study revealed 21 strains from 93 isolates which could be classified according to their spoligotype signature. Furthermore, members of the Beijing strain family were reported as dominant strains in all four provinces, followed by the T, and Latin American-Mediterranean (LAM) subtypes (Sekati et al. 2015). Another study by Hove and others investigated the spoligotype distribution of *M. tb* isolates from adults in the Soshanguve township in Pretoria, Gauteng. From 89 *M. tb* positive isolates, 75 were grouped into clusters. The Beijing clade formed the largest group of 28% of the isolates, followed by the LAM3, LAM4, and LAM9 strain family members making up 13%, 4%, and 3%, respectively (Hove et al. 2012).

For the Western Cape Province of South Africa, the most recent investigation of *M. tb* clade distribution described 2 727 cases of TB for 2 150 patients in which both drug-susceptible, and drug-resistant *M. tb* cases were investigated. Of the participants recruited, 1 737 were infected by a drug-susceptible *M. tb* strain. Eight clades could be derived for the collection of strains evaluated, of which the dominant clades were Beijing, LAM, Low-Copy clade (LCC), Haarlem, and lastly, Quebec (van der Spuy et al. 2009).

Using non-linear regression analysis, the numbers of annual TB cases attributed to infection with the Haarlem, Quebec, LCC, and LAM families did not increase significantly over the study period from January 1993 to December 2004. In contrast, the Beijing strain produced a logistic increase over time, demonstrating a significant increase in the number of TB infections caused by this strain (van der Spuy et al. 2009). Unlike the West-African region, Southern Africa has not shown a predominance for TB caused by *M. africanum* (Demers et al. 2010).

2.6.1.2. Ghana

West Africa is dominated by lineages 5 and 6, namely WA1 and WA2, with interplay with Lineage 4 (Figure 3). In contrast to the dominance of *M. tb* strain families in South Africa, Ghana's pulmonary TB epidemic is largely driven by infection with *M. africanum*, with some disease cases resulting from infection with *M. tb*. First identified in Senegal in 1968 by Castets and colleagues, *M. africanum* was classified as type 1 and type 2 intermediary species between *M. tb* and *M. bovis* (de Jong et al. 2010). However, as a result of genomic deletion studies, *M. africanum* type 2 was reclassified as *M. tb* "Uganda sublineage" as it was found to be a sub-lineage of *M. tb* lineage 4 (Niemann et al. 2002).

M. africanum, originally known as *M. africanum* type 1, is further classified into two distinct lineages, *M. africanum* West African 1 (*M. africanum* 1) and *M. africanum* West African 2 (*M. africanum* 2), each occupying specific geographical niches in West Africa (Figure 4). Spoligotyping of *M. tb* isolates belonging to Lineage 3 revealed the Ghana, Haarlem, Uganda I, Uganda II, LAM, New_1, and H37Rv_like sub-lineages to be circulating in Ghana (Asante-Poku et al. 2016). The distribution of *M. tb* clades dominating TB infections in South Africa is unknown in Ghana.

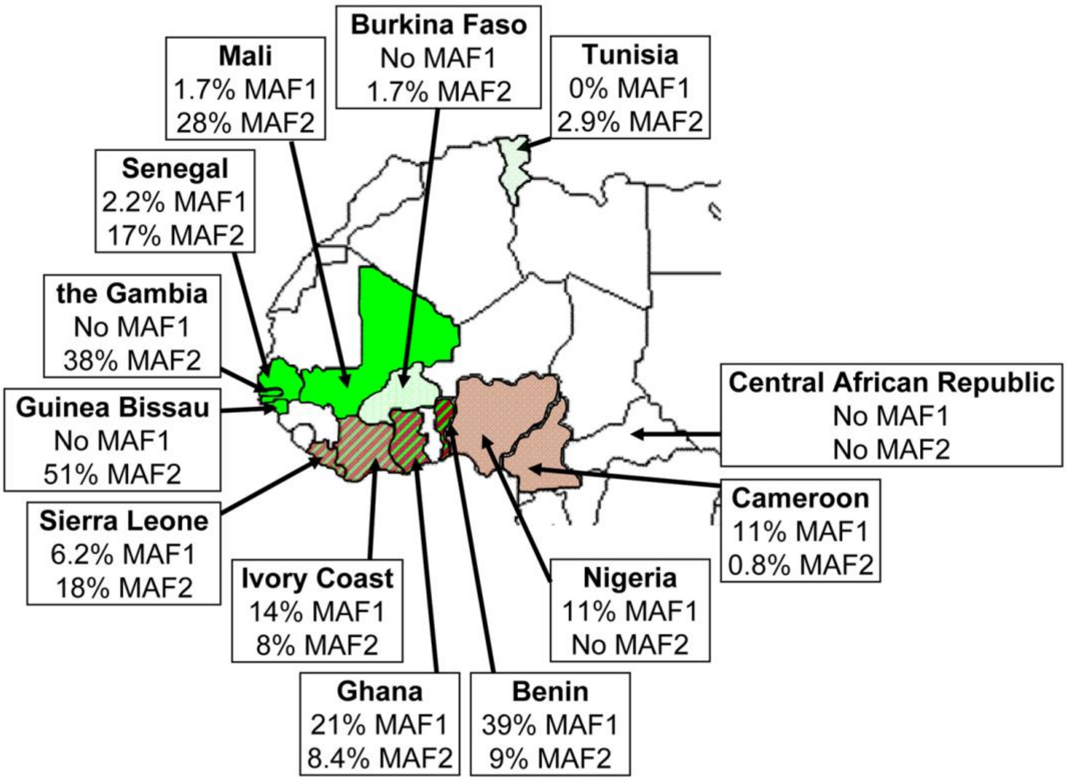


Figure 4: Prevalence of M. africanum in West Africa as originally published in de Jong et al. (2010).

2.6.2. Investigating the genetic susceptibility of the host to different *M. tb* strains

A number of studies performed to date have linked various disease phenotypes to host genotypes through SNP association testing. The number of phenotype definitions are endless and may be related to patient response to disease such as disease states, measured physiological parameters, or may be classified according to the infecting pathogen, such as different strains.

Susceptibility of the human host to infection with different strains of *M. tb* has only in recent years gained some attention. The first study to investigate host genotype-bacterial strain interactions was performed by Caws *et al.* on a cohort of 237 Vietnamese adults with pulmonary TB. Significant associations were found between genotypes with the C allele (T597C), of the toll-like receptor-2 gene (*TLR2*) and the bacterial strain and individuals with this allele were more likely to develop TB caused by mycobacteria belonging to the East-Asian/Beijing strain family (OR=1.57 [95% CI 1.15-2.15]) (Caws et al. 2008). However, the study design was of the nature that genes of interest were pre-selected based on previous TB susceptibility studies.

HLA types are known to be important in the immune response to pathogens. In a candidate-gene study, *HLA* types were investigated for associations between the *HLA* alleles of individuals self-identifying as belonging to the SAC population, and the *M. tb* strain responsible for their active TB (Salie et al. 2014). Individuals with the *HLA-B27* allele were shown to be at decreased risk for having an additional infection by a Beijing strain, when they had multiple episodes of infection with a Beijing strain. Specific HLA types were also found to be associated with disease caused by the different strains investigated (Salie et al. 2014)

While comprehensive methods have been developed for performing traditional case-control GWAS, research methods for analysing multiple phenotype-to-genotype associations for infectious diseases is still in its infancy. Genotype-first GWAS aim to test for associations between thousands of variants and a single trait of interest, such as a disease or health-related trait, in a case-control study design. A variation of this GWAS is the phenotype-first GWAS (PheWAS) in which the participants are classified according to phenotype. A phenome scan is used to test for association between genotypes and a comprehensive list of phenotypes. Variations in phenome scans include environment-wide association studies (EnWAS) and Mendelian randomization-PheWAS (Millard et al. 2017).

A phenome-wide association study (PheWAS) is a novel algorithm scanning for associations between targeted genotypes and numerous phenotypes (Denny et al. 2010). The pairing of genetic data with electronic medical records provides a platform for investigating phenome-wide associations to specific genetic markers of interest. In the proof of concept study, a cohort of 6 005 individuals of European-American ancestry were genotyped for five SNPs which have been reported in previous studies to be associated with seven common non-infectious diseases. In this study, the authors were able to replicate four out of seven known SNP-disease associations with significant *P*-values. The algorithm was also able to identify 19 novel statistical associations between the genotyped SNPs and the diseases queried, at a *P*-value less than 0.01 (Denny et al. 2010).

A follow-up study included 13 835 individuals of European ancestry who were analysed for associations between 3 144 genotyped SNPs and 1 358 phenotypes. The study replicated 66% of previously reported GWAS associations with sufficient statistical power and reported 63 associations with potential pleiotropy (Denny et al. 2013). Although this study demonstrated the importance of investigating genotype-phenotype associations, it focussed on a targeted set of SNPs with *a priori* association to disease.

A recent study aimed to identify genome-wide associations with TB onset, stratifying by infecting *M. tb* lineage, and the age at onset in a cohort of Thai participants (Omae et al. 2017). The study initially attempted to identify age-related associations between five *M. tb* lineages and two age-stratified groups of TB participants, namely 219 young cases (under the age of 45), and 467 old cases (over the age of 45). To reduce the complexity of the association tests, the *M. tb* lineages were tested as one lineage versus a collective of all other lineages. After applying Bonferroni corrections, the authors were unable to locate SNPs with genome-wide significance in either of the age-groups for any of the five *M. tb* lineages. However, when reducing the five lineages to two groups consisting only of ‘Beijing’ and ‘non-Beijing’ cases, and testing for age-related association to TB, the authors identified a single SNP on chromosome 1p13, rs1418425, reported to have a significant association to non-Beijing infected cases classified in the “old” age category (P = 1.58 x10^{-07} OR=1.62 [95% CI 1.35-1.93]). The authors were able to replicate the SNP in two independent cohorts, further demonstrating the importance of performing GWAS with a specific focus on pathogen lineage (Omae et al. 2017).

As presented in this literature study, numerous pieces of evidence for genetic associations to disease have been demonstrated since the advent of TB research. However, only one recent study has attempted to investigate the association between the human genome and different members of the MTBC causing TB. By not placing the pathogenic nature of the MTBC members in a single biological basket, we endeavour to untangle some of the complexities driving this scourge of humanity.

CHAPTER THREE

3. Materials and Methods Chapter Overview:

Chapter 3 of this dissertation describes the materials to develop the bioinformatics-based methods that enabled us to perform a genome-wide association analysis between human genotypes and the infecting *M. tb* clade responsible for the disease status of study participants (Figure 5). This study relied heavily on the provision of data from paired patient samples (blood and sputum) in order for the computational methods to be developed. Thus, the genotype data and *M. tb* isolate data used in this study were generated as part of previous studies and archived for further analysis. Sections 3.1.1 to 3.1.2 were performed in these previous studies and are described for background purposes. To reduce redundancy in explanation, subsequent steps are presented simultaneously for both cohorts.

Two cohorts are described in the present study. The paired data for the first cohort, the SAC cohort, is housed at Stellenbosch University and was readily available for method development described in this dissertation. A second cohort, from Ghana, was obtained once the method had been established, and tested the ability to perform an association analysis with the given method.

All subsequent methods described herein were developed for the current dissertation. All R scripts were written by the author of this dissertation, unless otherwise stated, and duly cited. All R scripts were written using R version 3.4.0 “You Stupid Darkness” (R Core Team 2017), and associated packages are cited in-text. Ethics approval for the current study was granted by the Health Research Ethics Committee at Stellenbosch University (Project number: S17/02/037).

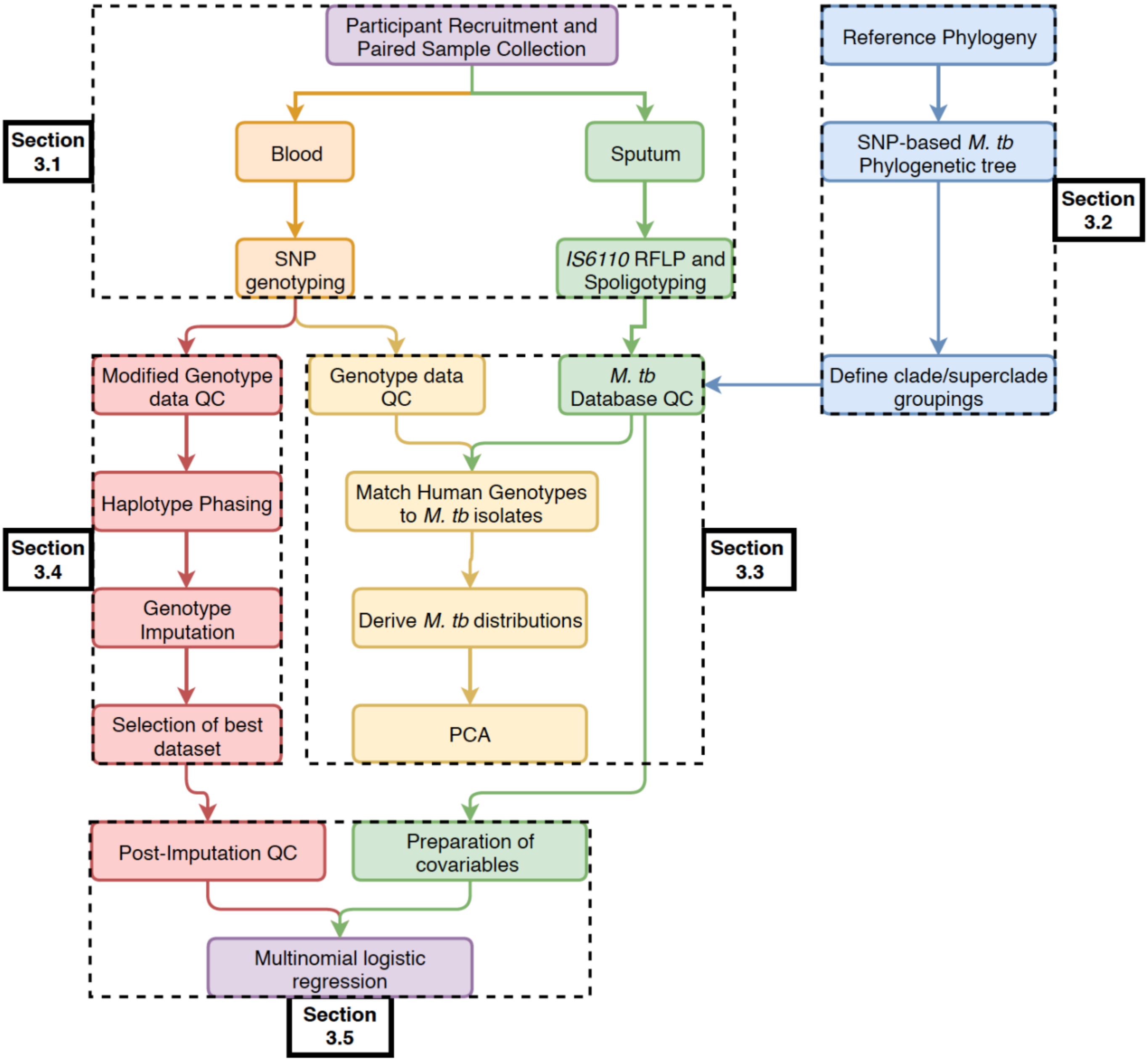

Participant Recruitment and Paired Sample Collection
Section 3.1
Blood
Sputum
SNP genotyping
IS6110 RFLP and Spoligotyping
Reference Phylogeny
SNP-based *M. tb* Phylogenetic tree
Section 3.2
Define clade/superclade groupings
Modified Genotype data QC
Genotype data QC
M. tb Database QC
Haplotype Phasing
Match Human Genotypes to *M. tb* isolates
Section 3.4
Section 3.3
Genotype Imputation
Derive *M. tb* distributions
Selection of best dataset
PCA
Post-Imputation QC
Preparation of covariables
Multinomial logistic regression
Section 3.5

Figure 5: Workflow designed to enable an association analysis of genome-wide host SNP markers to multiple classes of infecting M. tb isolates.

3.1. Participant recruitment and sample collection

Sections 3.1.1 to 3.1.2 below describe the collection of blood and sputum samples from patients recruited into the SAC or Ghanaian cohort. At the time of recruitment for the SAC cohort, two separate studies contributed data - one study focussed on collecting blood for genotyping, while a second, independent study was concerned with collecting sputum for an epidemiological study. However, not all patients had both blood and sputum samples collected during the sampling period. Thus, it was necessary to match patients who had been genotyped to records of their sputum collection in a second database. The details of this process is described in the paragraphs that follow.

3.1.1. Participant recruitment

3.1.1.1. SAC cohort

For the SAC cohort, study participants were recruited as part of a previous TB epidemiology study in the Western Cape Province of South Africa during the period of January 1993 through December 2004 (Project number: 95/072) from suburbs where the TB incidence was high (28.9 % in 2005) and the prevalence of HIV was a low 2% (Kritzinger et al. 2009; Shisana et al. 2012). All study participants self-identified as belonging to the SAC population, were HIV-negative, and provided written informed consent. All subsequent research was conducted according to the principles expressed in the Declaration of Helsinki (WHO 2001).

Blood and sputum samples were collected from each study participant as approved by the Health Research Ethics Committee at Stellenbosch University (Project numbers: S17/01/013 and 95/072). DNA was extracted from blood samples using the Nucleon BACC Genomic DNA extraction kit (Illumina, Buckinghamshire, UK) and genotyped on two platforms: the GeneChip Human Mapping 500K SNP array (Affymetrix, subsidiary of Thermo Fisher Scientific, California, United States) and the Infinium Multi-Ethnic Genotyping Array (MEGA) (Illumina, California, United States).

3.1.1.2. Ghanaian cohort

For the Ghanaian cohort, study participants were enrolled in Ghana, West Africa, between September 2001 and July 2004. All cases were HIV-negative and confirmed to have pulmonary TB by two independent radiologists (Thye et al. 2012). Ethics approval for the sample collection was granted by the Committee on Human Research, Publications and Ethics, School of Medicine Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana, and the Ethics Committee of the Ghana Health Service in Accra, Ghana (Thye et al. 2012).

3.1.2. Sample processing

3.1.2.1. SNP Genotyping

3.1.2.1.1. SAC cohort

DNA was extracted followed by SNP genotyping using the Affymetrix 500K array. Genotype-calling was performed using the Affymetrix Power Tools pipeline (V1.10.0) as previously described (De Wit et al. 2010). For a subset of individuals in the SAC cohort, DNA was genotyped using the Multi-Ethnic Genotyping Array (MEGA) from Illumina. Following standard genotyping quality control (QC), ancestry proportions for the SAC cohort on the Affymetrix 500k array were estimated using the unsupervised algorithm implemented in ADMIXTURE (Daya et al. 2013). For the genotype data obtained from the MEGA array, genotype-calling was performed using Genome studio v2.04 and following standard genotype QC, ancestry proportions were estimated (Schurz et al. 2018) using the ADMIXTURE software (Alexander et al. 2009).

3.1.2.1.2. Ghanaian cohort

DNA extracted from blood samples was genotyped using the Affymetrix SNP 6.0 array at the Affymetrix Services Laboratory in California, and at ATLAS Biolabs GmbH in Berlin. Genotypes were called using the Birdseed version 2 algorithm and using the Eigenstrat software, ancestry proportions in the form of principal components were derived (Thye et al. 2012). For both cohorts included in this study, genotype data was made available in PLINK (Purcell and Chang, n.d.; Chang et al. 2015) format as described in Table 3.

Table 3: PLINK file formats

PLINK file type	Column Descriptions
PED	Family ID, Individual ID, Paternal ID, Maternal ID, Sex, Phenotype
MAP	Chromosome, SNP ID, Genetic distance, Physical position
BED	Binary PED file
BIM	Chromosome, SNP ID, Genetic distance, Physical Position, Allele 1, Allele 2
FAM	Family ID, Individual ID, Paternal ID, Maternal ID, Sex, Phenotype

3.1.2.2. Spoligotyping // IS*6110* RFLP records

3.1.2.2.1. SAC cohort

For each infection, sputum samples were collected from participants for bacterial culture on LJ media. *M. tb* isolates were identified using spoligotyping and IS*6110* Restriction Fragment Length Polymorphism (RFLP) methods as previously described (van der Spuy et al. 2009). All *M. tb* infection data were entered onto manually-curated departmental databases for archiving.

3.1.2.2.2. Ghanaian cohort

M. tb isolates extracted from sputum samples were cultured on LJ media at the Kumasi Centre for Collaborative Research (Owusu-Dabo et al. 2006). Strains were identified by IS*6110* RFLP and spoligotyping (Supply et al. 2006).

3.2. Defining *M. tb* clades and superclades

M. tb clades were grouped into superclades by clustering closely-related clades near a point of divergence on a SNP-based phylogenetic tree of *M. tb* (Figure 6) (Dippenaar 2014). This was performed to reduce the number of clades of low frequency into superclades with higher frequency, as low frequency groups are known to induce an unfavourable collinearity effect on logistic regression models (Bergtold et al. 2011). Clades remaining unchanged were also referred to as "superclades" after clustering. Clades present in the Strain database, but not represented on the phylogenetic tree, were kept as distinct groups and not clustered with any of the members on the existing tree, except when a suitable reference phylogeny was found. After clustering, superclades with a frequency less than ten in the study dataset were excluded from subsequent analyses.

3.3. Preliminary assessment of paired data

A preliminary assessment using PCA was conducted to determine whether there was potentially a relationship between the host genotypes and the occurrence of specific infecting *M. tb* clades. Following standard genotyping QC, a PCA was performed using the infecting clade data and genotype data that had passed through several QC filters.

3.3.1. Genotype QC

To prepare the patient genotypes for a preliminary analysis of the association with the *M. tb* clades and superclades, variants lacking chromosome and/or base pair annotation were excluded from the PLINK files for all participants. This was followed by iterative QC filters for a maximum of 10% individual genotype missingness (--mind), 2% SNP genotype missingness (--geno), and 5% SNP minor allele frequency (MAF; --maf) using PLINK. These iterations alternated with a "sex" check and identification of duplicate samples at a level of first-cousin relatedness (pi-HAT > 0.125). A Perl script written by Yekai Xiong (Chang et al. 2014; Gao et al. 2015) was used to identify one individual per related pair who needed to be removed. This was followed by a test for excess heterozygosity in PLINK to identify individuals with genotyping heterozygosity four standard deviations from the mean. These individuals were removed from the dataset. PCA was performed to identify outliers in the dataset and any separation of the genotype data based on *M. tb* clades and superclades.

3.3.2. *M. tb* database quality control

3.3.2.1. SAC cohort

Two archived in-house databases were used to retrieve information of the infecting *M. tb* clade for the SAC cohort. One database contained patient records and is referred to as the "Patients database" and is abbreviated "P", while a second database contained spoligotyping records for study participants from which *M. tb* cultures had been derived (Table 4). As the *M. tb* database consisted of multiple levels of information for the bacterium, it is herein referred to as the "*M. tb* database" and is abbreviated "S" for "Strain".

Table 4: Database sources used to complete this objective

Data source	Column Descriptions	Number of observations
Patients database (P)	Case ID, PatientID, Date of Birth, Gender, Clinical Disease	3 937
M. tb database (S)	Patient ID, Culture Date, SAWC, IS6110_3, Family, Family Group, Type, Description, Infection, Reactivation, Evolved	4 583
SAC cohort	Family ID, Individual ID, Sex, Phenotype (extracted from PLINK .fam file)	947
link	Case ID, Sample ID	4 408

Bash shell scripting was used to extract the Family ID, Individual ID, Sex, and Phenotype columns for the genotyped individuals and an R script was written to link the genotyped study participants to their corresponding *M. tb* infections. Once *M. tb* clade-matched study participants who were also genotyped were identified, QC was performed on the patient records as follows: the first QC measure was to identify any individuals who may have had duplicate records as a result of being genotyped more than once. If any individual had been genotyped more than once, the raw genotype-calling quality scores of each sample was compared, followed by the removal of the sample with the poorer quality score from the dataset. The second QC measure was to identify and correct incorrect assignments of infection number.

3.3.2.2. Ghanaian cohort

Host genotype data and bacterial strain data were generously provided by Dr Thorsten Thye and Prof Stefan Niemann, and was assessed using the same QC filters as those used on the SAC dataset.

3.3.3. Derive *M. tb* distributions

M. tb clade and superclade distributions were derived for subsets of the study cohort. If records for multiple infections were available, *M. tb* distributions were derived for the first infection recorded for each patient present in the database, as well as any subsequent infections. Distributions of *M. tb* isolates were also computed for the subsets of study participants having only one infection recorded in the PS database, and those having multiple infections.

It is possible for study participants to have been sampled multiple times during the recruitment period and a number of possibilities exist which may confound the representation of an *M. tb* infection. We acknowledge that these individuals may have gone to different clinics during the sample collection period or may have failed to return to the clinic. Thus, it is possible that a subsequent *M. tb* infection may be misrepresented with regards to the true order of events or may not be recorded in the Strain database at all. For this reason, we took the pragmatic approach of analysing the first recorded infection for each patient during the sample collection period. Frequency distributions of clades and superclades were reported for the first- and second infection records, where available, as well as for study participants having only one recorded infection in the database, and two recorded infections in the database.

3.3.4. PCA of first recorded infection

Genotypes for the samples passing the QC filters were converted to the genomic data structure (gds) format using the **gdsfmt** (Zheng et al. 2012) package in R. PCA was performed on the genotype data using the **snpgdsPCA** function available within the **SNPRelate** package (Zheng et al. 2012). Eigenvalues generated from the PCA were plotted in a scree plot and visually inspected to identify at which eigenvalue the variation in the data became least explicable. PCA plots were generated by selecting paired eigenvalue columns and plotting them combined with an overlay of either the clade, or superclade.

3.4. Generate high-quality imputed genotype data

Prior to performing an association analysis, a critical step is to obtain high quality genotype data by filtering for SNP genotype missingness, sample missingness and minor allele frequency. Additionally, variants lacking chromosome or base pair information were updated using the 1000 Genomes Phase 3[1] (1000GP3) reference panel and the dbSNP database. To maximise the number of variants tested in the association analysis, the cleaned genotype data were imputed using three different reference panels to determine which panel served the best for the given dataset in imputing missing variants.

3.4.1. Modified Data QC for Imputation

3.4.1.1. Update variants lacking chromosome or base pair position information

Instead of discarding unannotated SNPs as was done in 3.3.1, chromosome and base pair position annotations were retrieved for SNPs lacking this information. To do this, the SNP overlap between the 1000GP3 reference panel and the study dataset was calculated. Chromosome and base pair position annotations were retrieved from the 1000GP3 and only SNPs also found to be validated by laboratory methods in the dbSNP[2] database were retained and subsequently updated for their missing information. SNPs found in the 1000GP3, but not validated according to dbSNP were removed from the study dataset.

3.4.1.2. Examine the data for ambiguity in alignment to reference data

When genotype data are generated, the reference panel used in the array design may be an outdated version of the most recently updated human reference panel (Wang et al. 2017). In regions where the reference panel has been updated with “strand flips” or “SNP flips”, the original genotype data may not be completely oriented in the same way as the reference genome, and the process of genotype imputation may be erroneous. Thus, to facilitate accurate genotype imputation and meta-analyses of datasets genotyped on different platforms, it is essential that study data be oriented to the strand direction of the reference data (Deelen et al. 2014).

[1] ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/phase3/
[2] ftp://ftp.ncbi.nih.gov/snp/

A previous researcher in our group, Dr Michelle Daya, used LiftOver[1] to convert the Affymetrix genotyping data for the SAC cohort from build 36 of the human genome reference (which was used in the array design) to build 37[2]. LiftOver updates SNP IDs (rs numbers) and/or genomic co-ordinates (base pair positions) between the assembly on which the input genotyping data was generated, aligned to the desired (generally, more recent) reference genome assembly.

Using Genotype Harmonizer[3] (Deelen et al. 2014), genotyping data were aligned to match the strand direction used in the 1000GP3 reference dataset. The study genotypes were entered into Genotype Harmonizer using default parameters along with flags to update the study variant IDs, match the study reference alleles, and to keep variants not in the reference dataset. The linkage disequilibrium (LD) checker was kept off as default, and a maximum MAF of 0.05 was included as a back-up filter. In instances where insufficient variants were in LD and the MAF of a variant was less than or equal to the specified value in both study data as in the reference, the minor allele was used as a backup for alignment.

3.4.1.3. Genotype QC

Following alignment of alleles to the 1000GP3 reference panel with Genotype Harmonizer, genotyping QC was performed using PLINK and consisted of iterative filters for 10% sample missingness, 2% SNP genotyping missingness, and a 5% MAF filter, until no more SNPs or samples were removed. These three filters alternated with a “sex” check, and check for too little, or excessive genotyping heterozygosity. Too little genotyping heterozygosity amongst participants in the sample may be an indication of inbreeding, whereas excessive heterozygosity may be indicative of sample contamination (Anderson et al. 2010). Related individuals were identified but not removed from the dataset at this stage because increased sample sizes, regardless of relatedness, have been shown to improve phasing and imputation accuracy (Deelen et al. 2014).

[1] http://genome.ucsc.edu/cgi-bin/hgLiftOver
[2] http://hgdownload.cse.ucsc.edu/goldenPath/hg19/chromosomes/
[3] http://www.molgenis.org/downloads/GenotypeHarmonizer/GenotypeHarmonizer-1.4.20-dist.tar.gz

3.4.2. Data preparation, Haplotype phasing, and Genotype imputation

3.4.2.1. In-house (IH) workflow

Haplotype phasing and genotype imputation was performed using five workflows as shown in Table 5. For the In-House workflow (abbreviated as IH-1000GP3), genotypes in PLINK format passing QC filters were phased per chromosome using ShapeITv2 (Delaneau et al. 2008). The PHASER mode (version 2.r837) of the ShapeITv2 algorithm was kept at a default of 35 Markov Chain Montecarlo (MCMC) iterations and the effective population size was specified as 15 000. Due to the admixed nature of the SAC population, the effective population size for the cohort was determined as an intermediate value of those suggested on the ShapeITv2[1] website. The window-based model used 100 states per window, approximately 2 Mb in size. When phasing the X chromosome, males and females were split and heterozygous haploid variants were identified and removed using PLINK.

Table 5: Haplotype phasing and Genotype Imputation workflows

Workflow name	Reference Panel	Abbreviation used	Phasing software	Imputation software
In-house	1000GP3	IH-1000GP3	ShapeITv2	IMPUTE2
Michigan Imputation Server	1000GP3	MIS-1000GP3	ShapeITv2	Minimac3
Michigan Imputation Server	CAAPA*	MIS-CAAPA	ShapeITv2	Minimac3
Sanger Imputation Server	1000GP3	SIS-1000GP3	ShapeITv2	PBWT$^{\lozenge}$
Sanger Imputation Server	AGR**	SIS-AGR	ShapeITv2	PBWT

* CAAPA: Consortium on Asthma among African-ancestry Populations in the Americas

**AGR: African Genome Resource

$^{\lozenge}$ PBWT: Positional Burrows-Wheeler Transformation

For each chromosome, a .samples and .haps file is produced following the haplotype phasing step. The .samples file is similar to the PLINK fam file in that it contains seven columns specifying the Individual ID in two columns, a column specifying the proportion of missing data, a paternal and maternal ID, “sex”, and phenotype. The .haps file is a tab-delimited file in SNP-major format where each line contains a chromosome number, SNP ID, SNP position, encoding for allele A, and encoding for allele B.

[1] http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html#gettingstarted

IMPUTE2[1] version 2.3.2 (Howie et al. 2009) was used for the imputation of 1000GP3 SNP markers into the study dataset. IMPUTE2 describes Panel 0 as being the phased reference haplotypes, and Panel 2 as being the phased study haplotypes. The genetic map corresponding to the chromosome being imputed was used as the source of genetic recombination rates[2]. IMPUTE2 documentation recommends dividing each chromosome into approximately 5 Mb pair chunks. Thus, for each chromosome, the last base pair position was divided by 5 Mb in order to determine how many chunks needed to be specified per chromosome. The default buffer region of 250 kb was used to include a specified number of SNPs on each side of the analysis region using the **-int** flag. Using a buffer region is recommended by IMPUTE2 as it prevents the deterioration of imputation quality near the edges of the chunk being analysed (Howie et al. 2009).

The input genotype calling threshold was kept at the default of 0.9 and the MCMC iterations were also kept at default values, along with 500 Hidden Markov Model (HMM) states for imputation. Imputation using IMPUTE2 produces six files for each chunk of the chromosomes imputed. The format of these files are shown in Table 6. A concatenated INFO file was generated for each chromosome comprising of the info files per chunk. All genotype files obtained per chromosome were concatenated into a single .gen genotype file.

Table 6: Standard output file format from IMPUTE2

Output file	File Format
genotype file (no extension)	snp_id, rs_id. position, a0, a1, three genotype probabilities for SNP per sample
_info	snp_id, rs_id, position, a0, a1, exp_freq_a1, info, certainty, type, info_type0, concord_type0, r2_type0
info_by_sample	concord_type0, r2_type0
_summary	A summary of the input files, output files, and processing performed.
_warnings	Any potential errors in the processing
_samples	Individual ID in two columns, missingness proportion, paternal ID, maternal ID, “sex”, and phenotype

[1] http://mathgen.stats.ox.ac.uk/impute/impute_v2.html
[2] https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html

3.4.2.2. Michigan Imputation Server (MIS) workflow

In preparation for upload to the Michigan Imputation Server[1] (Das et al. 2016), genotypes were converted to variant call format (VCF) format using PLINK, sorted by ascending chromosomal order, and checked for concordance to internal MIS specifications using the checkvcf.py[2] script. As required by the MIS, individual chromosomes were compressed using **bgzip** and indexed using **tabix**. Both packages are found within BCFtools[3]. The individual chromosomes were uploaded in an unphased state and ShapeITv2 was specified for the haplotype phasing step followed by imputation using the MIS' tool of preference, Minimac3 (Das et al. 2016). Per the requirements of the software, Chromosome 23 genotypes in PLINK format were converted to be annotated as chromosome "X".

For the first analysis of the study data on the MIS, phasing was preceded by a QC step in which the "mixed" population option was selected when specifying the desired workflow, followed by imputation with the 1000GP3 reference panel. In the second independent analysis on the MIS, the Consortium on Asthma among African-ancestry Populations in the Americas (CAAPA)[4] reference panel was specified along with a mandatory selection of the "African-American" population for the QC. It is important to note that the CAAPA reference does not have reference haplotype data to facilitate imputation of the X chromosome.

3.4.2.3. Sanger Imputation Server (SIS) workflow

In contrast to the MIS, the SIS[5] required a single concatenated and compressed file in VCF format. As advised on the SIS help page, a file check was performed on the concatenated VCF file using the BCFtools **norm -ce** flag to ensure that the format met the specifications required by the SIS. The SIS strictly does not accept a VCF file failing on the alignment to the human reference human_g1k_v37.fasta[6] and thus if errors were returned, the **fixref** plugin available from BCFtools was used. Haplotype phasing was performed on the server using ShapeIT2, followed by genotype imputation of the available reference panels using the Positional Burrows-Wheeler Transformation (PBWT) algorithm (Durbin 2014). The first analysis of the

[1] https://imputationserver.sph.umich.edu/start.html#!pages/home
[2] https://github.com/zhanxw/checkVCF
[3] https://github.com/samtools/bcftools/releases/download/1.9/bcftools-1.9.tar.bz2
[4] https://www.caapa-project.org/
[5] https://imputation.sanger.ac.uk/
[6] ftp://ftp.ncbi.nlm.nih.gov/1000genomes/ftp/technical/.../human_g1k_v37.fasta.gz

study dataset on the SIS was performed using the 1000GP3 reference panel and the second independent imputation made use of the African Genome Resource (AGR) reference panel.

3.4.3. Selection of high-quality imputed genotype data

Both the IH workflow using IMPUTE2, and the SIS workflows using the PBWT algorithm produce an INFO metric, while the MIS workflow using Minimac3 produces an r-squared (Rsq) quality metric. Both quality metrics are an estimation of the squared correlation between the true, unobserved genotypes, and the imputed genotypes (Browning and Browning 2009). The quality metric ranges from 0 to 1, where values near 1 indicate a high certainty of the imputed SNP whereas a value near 0 is a SNP imputed with low certainty. Negative values indicate high uncertainty in the imputation and a value of -1 is assigned when the metric cannot be calculated for a particular marker (Marchini and Howie 2010).

For each of the imputation analyses, the INFO or Rsq values were extracted for all SNPs and the mean quality scores were calculated per MAF bin specified as 0-5%, 5-10%, 10-20%, 20-30%, 30-40%, and 40-50%. In the case of the SIS, the MAF needed to first be calculated using VCFtools[1]. There is no universal cut-off for filtering on the quality score metric, but generally a value between 0.3 and 0.5 is an accepted cut-off (Marchini and Howie 2010). Thus, SNPs with an INFO or Rsq value greater than 0.45 were selected for calculating SNP density for each workflow. Plots were also generated to compare imputation quality scores across the defined MAF bins.

3.5. Perform an association analysis using high-quality imputed host genotype data and *M. tb* superclade data

The final step in the method was to perform an association analysis using the patient genotype and infecting *M. tb* data. To do so, imputed genotypes were assessed to extract only SNPs imputed with high certainty and to determine which reference panel imputed the study dataset with the highest quality. The association analysis was then performed using a multinomial logistic regression framework implemented in SNPTest.

[1] http://vcftools.sourceforge.net/

3.5.1. Preparation of genotypes: Post-imputation QC

The dataset, which had the highest imputation scores per MAF bin, was filtered to exclude monomorphic variants by selecting dosage values greater than zero and less than two. SNPs with a quality score of greater than 0.45 were prioritised for the association analysis and filtered iteratively for a maximum of 10% individual genotype missingness, 2% SNP genotype missingness, and 5% SNP MAF using PLINK. Related individuals identified prior to imputation were removed followed by a second round of iterative filters for SNP- and sample missingness and MAF. Samples which were *M. tb* clade-matched were extracted from the remaining samples which had passed all QC filters.

If the IH dataset in gen/sample format was selected as the best dataset, the first step was to obtain a list of SNPs with high quality. To do so, the concatenated INFO file for each chromosome was assessed in R to obtain the subset of all the SNPS from which monomorphic SNPs and INDELS had been removed, as well as SNPs not meeting the INFO score cut-off of 0.45. Thereafter, the genotype file for each chromosome was filtered using the **subset** mode in GTOOL version 0.7.5[1] to retain their corresponding SNPs of high quality.

The filtered genotype files were then merged into a single genotype file using the **merge** mode in GTOOL. The merged genotype and sample files were copied, and the copy was converted to PLINK PED/MAP format using the **-G** flag in GTOOL, incorporating a genotype calling threshold of 0.7. Following conversion of the PLINK PED/MAP file to BED/BIM/FAM format, the PLINK genotype files were filtered iteratively for a maximum of 10% individual genotype missingness, 2% SNP genotype missingness, and 5% SNP MAF using PLINK. Related individuals identified prior to imputation were removed followed by a second round of iterative filters for SNP- and sample missingness and MAF. Trailing whitespace was removed from the .gen file to meet the column specifications of SNPTEST[2] v2.5.2 (Marchini 2010) and samples which were *M. tb* clade-matched were extracted from the remaining samples which had passed all QC filters.

[1] http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html

[2] https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html

3.5.2. Preparation of covariables file

All available covariables were obtained for the cohort under study. These included, where available, sex, and age at time of active TB and subsequent recruitment into the study. The first sex-specific GWAS for TB was recently performed, and although no genome-wide significant associations were detected, the results of the study showed strong evidence of possible sex-specific associations for risk of developing TB (Schurz et al. 2018). In light of this finding, we found it critical to include sex as a covariable in this analysis to reduce the risk of bias in the association test.

As an individual ages, so does one's general immunity decrease, and subsequently, susceptibility to infectious illnesses, including TB, increases (reviewed in Wang 2012). Thus, to mitigate the bias of age influencing the association analysis, age at the time of TB illness and recruitment into the study was included as a covariable.

To correct for differences in ethnicity amongst participants, either ancestry proportions or principal components were calculated and included as covariables. SNPTEST is unable to include covariables when the variance in the values provided is "too small" as indicated here https://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=oxstatgen;f8f2270b.1207 . As SNPTEST is still under development, it has not yet been established, or recorded in the software manual, to what degree of variance covariable data will not be accepted for inclusion in the logistic regression. For developing the method, it was established through trial and error that if the variance was below 0.001, these covariables could not be included. Additionally, for any samples where any of the covariables were not available, the samples were removed and not included in the regression analysis.

3.5.3. Multinomial logistic regression

For the association analysis, SNPTEST was used to perform the multinomial logistic regression (MLR) analysis using an additive genetic model. As opposed to a linear regression model which assesses the relationship between a continuous dependent (response) variable and the independent (predictor) variable(s), a standard logistic regression is performed when the dependent variable is dichotomous. An extension of the standard logistic regression, the multinomial logistic regression, is appropriate for testing the relationship between one independent variable (e.g. individual SNP genotypes) and the dependent variable which has more than two levels (e.g. multiple clades or superclades of *M. tb*). This method also allows for the adjustment of potential confounding factors. For both cohorts investigated in this study, we were able to adjust for potential confounders such as ethnicity, sex, and age by including these as covariables in the model.

SNPTEST tests for association using frequentist statistical methods. Two discrete variables, namely sex and superclade, as well as continuous variables, namely age at TB onset and ancestry proportions or principal components were included in the analysis. The phenotype tested was specified as the *M. tb* superclade. Thus, the MLR model specified was the occurrence of the *M. tb* superclade as a function of the baseline covariables given, as well as the genotypes supplied. The standard genome-wide significance cut-off of alpha = 5×10^{-8} was used when reporting significance of SNPs (The International HapMap Consortium 2005; Pe'er et al. 2008). Odd's ratios for the multiple phenotypes tested are calculated against a baseline phenotype by setting the odds of that phenotype occurring, given the genotype, to 1. A baseline phenotype may be specified by the user, or it will be determined internally by taking the first phenotype to appear alphabetically. For this study, the baseline phenotype was specified as the dominant superclade in the cohort, or a common superclade of intermediate frequency if more than one cohort is being studied.

SNPs with a Likelihood Ratio Threshold (LRT) p-value of less than 5×10^{-4} were selected and analysed in R and odds ratios were calculated from the beta values generated by SNPTEST. SNPs with a standard error greater than 1.5 for their odds ratios were excluded and SNPs with an LRT p-value less than 1×10^{-6} were prioritised for further investigation. As this was a methods-development study, these thresholds were selected pragmatically for the completion of the method.

3.5.4. Gene annotations of selected SNPs

The Variant Effect Predictor (VEP) Tool (McLaren et al. 2016) was used to retrieve gene annotations for the SNPs of interest.

CHAPTER FOUR

4. Results

4.1. Participant recruitment and sample collection

4.1.1. Participant recruitment

A total of 947 study participants were recruited for the SAC cohort, while 3 311 participants were recruited for the Ghanaian cohort. All participants provided blood and sputum samples for SNP genotyping, and *M. tb* isolate identification, respectively.

4.1.2. Sample processing

4.1.2.1. SNP genotyping

4.1.2.1.1. SAC cohort

A total of 947 participants recruited into the study were genotyped for 500 000 SNPs on the Affymetrix 500K SNP array. Of the 500 000 SNPs on the array, 397 337 (79.4%) were successfully genotyped for all samples. Of the 947 samples genotyped, 853 were TB cases and 516 were male. For the subset of cases used in this cohort, 55% of the individuals recruited were male (Table 7). Ancestry proportions for the five-way admixed SAC cohort genotyped on the Affymetrix array indicated that the study participants were predominantly of isiXhosa ancestry (33%), followed by Khomani (31%), European (16%), South-Asian (13%), and East-Asian ancestry (7%) (Daya et al. 2013; Chimusa et al. 2013; De Wit et al. 2010).

Table 7: Summary of patient recruitment for the SAC and Ghanaian cohorts

	SAC	**Ghana**
Cases	853	1 359
Male	516	2 087
Female	431	1 224
Cases + Male	469 (55%)	933 (69%)
Cases + Female	384 (45%)	426 (31%)
Total number of participants	947	3 311
Total number of variants	397 337	783 338

4.1.2.1.2. Ghanaian cohort

For the Ghanaian cohort, 3 311 participants were genotyped for 906 600 SNPs on the Affymetrix SNP 6.0 array. Of the available SNPs, all samples were successfully genotyped for 783 338 variants. Of the samples included in this study, 1 359 were TB cases and 69% of the participants were male (Table 7). PCA revealed that the study participants had contributing ethnicities from the Akan, Ga-Adangbe, Exe, and several other ethnic groups in northern Ghana (Thye et al. 2012).

4.1.2.2. Spoligotyping // IS*6110* RFLP records

4.1.2.2.1. SAC cohort

A Patients-Strain database (PS) was created by linking each patient's record in the Patients database to its corresponding *M. tb* isolate record from the Strain database. The PS database could not be directly linked to the genotyping file as there were no overlapping columns between the two databases (Table 4). Thus, using the Case ID in an intermediary file, Individual IDs from the FAM file were matched to the Patient IDs in the PS database. The Individual ID was consistent with the Sample ID column in the linking file and yielded a dataset of 527 genotyped study participants matched on Sample ID to 609 *M. tb* records (Figure 7).

4.1.2.2.2. Ghanaian cohort

Of the 3 311 TB study participants genotyped, 1 318 of the 1 359 cases had corresponding *M. tb* clade information and no samples were found to be genotyped more than once.

4.2. Defining *M. tb* clades and superclades

Using the SNP-based phylogeny for *M. tb* (Dippenaar 2014), *M. tb* clades were grouped into superclades by clustering closely-related clades near a point of divergence as indicated by the dashed red line and coloured dots in Figure 6. This point of divergence was chosen to reduce the number of clades of low frequency into superclades with higher frequency.

Clustering based on the phylogenetic tree reduced 12 distinct clades into five closely-related superclades. The East-African Indian (EAI) and *M. africanum* clades (green bracket) merged into the "EAI_afri" superclade, as did the CAS and Beijing clades into the "BeijingCAS" (red bracket). The Low-copy Clades (LCC), Pre-Haarlem, Haarlem-like, and Haarlem clades merged into one superclade designated as “HaarlemsLCC” (orange bracket), while the Quebec, LAM, T, and Lineage 7 clades remained unchanged and are indicated by black brackets Figure 6).

The SAC cohort contained seven of the 12 clades on the phylogenetic tree, namely Beijing, CAS (represented as CAS1 in the Strain database), Haarlem, Haarlem-Like, LCC, T and Quebec. A clade denoted as “Other” was also present in the SAC cohort but does not appear on the phylogenetic tree (Figure 6) and consisted of 35 “Family” classifications, and thus was kept as a distinct member during the grouping strategy. The “T” clade was excluded from subsequent analysis due to low frequency in the cohort after clustering into superclades.

The Ghanaian cohort contained 12 clade annotations obtained from spoligotyping. The afri-181 and afri-438 clades were represented by *M. africanum* on the phylogenetic tree and were subsequently grouped with EAI at the point of divergence, and named as the EAI_afri superclade. Beijing and CAS were merged, as were Haarlem and X into a “HaarlemX” superclade. T and U clades were clustered as indicated on Figure 6. The “Ghana-2” clade was kept as a distinct superclade, while LAM and CAM were grouped based on the similarity in their spoligotyping patterns illustrated in Stucki *et al.* 2016.

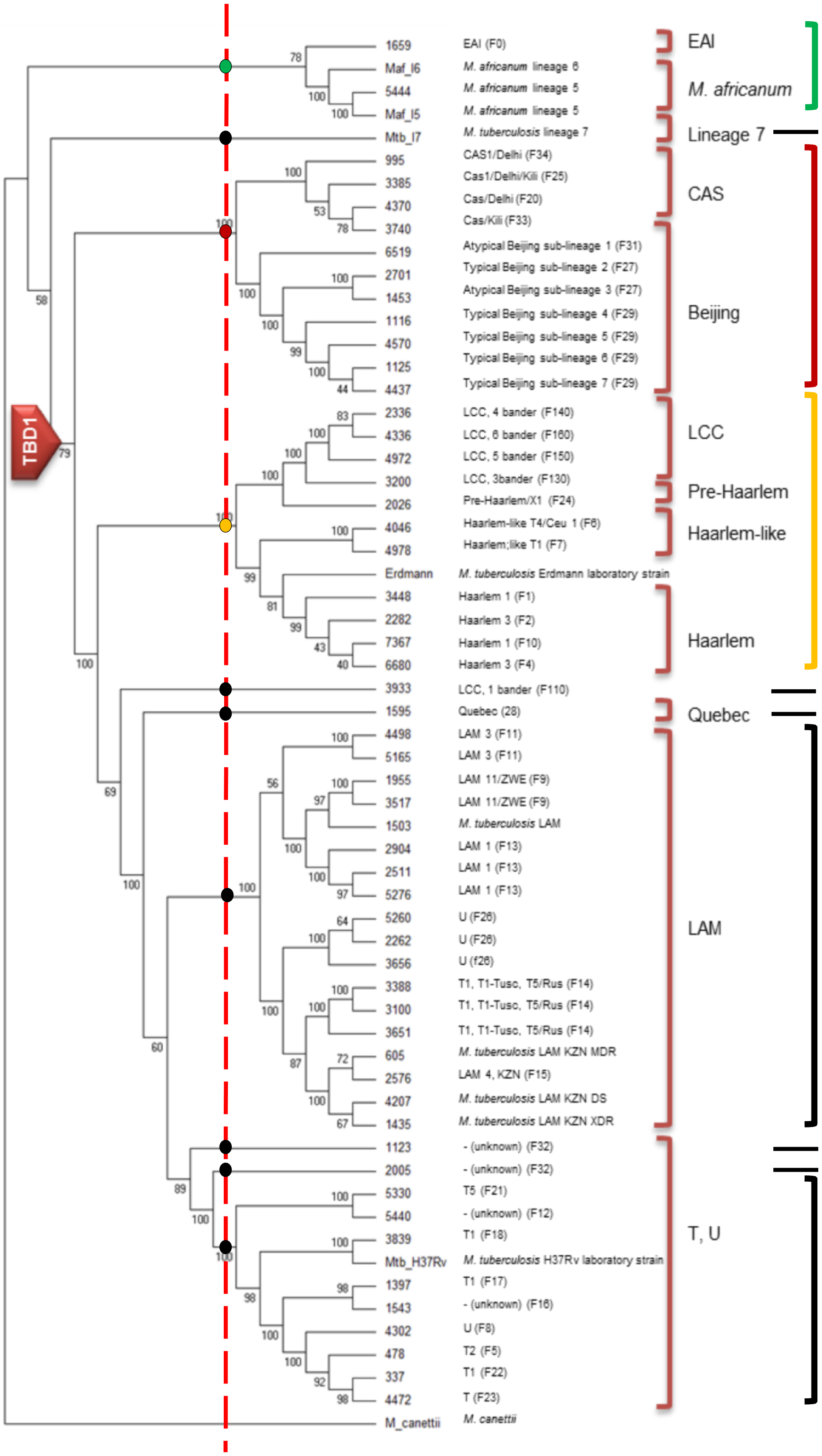

TBD1
1659 EAI (F0)
Maf_I6 M. africanum lineage 6
5444 M. africanum lineage 5
Maf_I5 M. africanum lineage 5
Mtb_I7 M. tuberculosis lineage 7
995 CAS1/Delhi (F34)
3385 Cas1/Delhi/Kili (F25)
4370 Cas/Delhi (F20)
3740 Cas/Kili (F33)
6519 Atypical Beijing sub-lineage 1 (F31)
2701 Typical Beijing sub-lineage 2 (F27)
1453 Atypical Beijing sub-lineage 3 (F27)
1116 Typical Beijing sub-lineage 4 (F29)
4570 Typical Beijing sub-lineage 5 (F29)
1125 Typical Beijing sub-lineage 6 (F29)
4437 Typical Beijing sub-lineage 7 (F29)
2336 LCC, 4 bander (F140)
4336 LCC, 6 bander (F160)
4972 LCC, 5 bander (F150)
3200 LCC, 3bander (F130)
2026 Pre-Haarlem/X1 (F24)
4046 Haarlem-like T4/Ceu 1 (F6)
4978 Haarlem;like T1 (F7)
Erdmann M. tuberculosis Erdmann laboratory strain
3448 Haarlem 1 (F1)
2282 Haarlem 3 (F2)
7367 Haarlem 1 (F10)
6680 Haarlem 3 (F4)
3933 LCC, 1 bander (F110)
1595 Quebec (28)
4498 LAM 3 (F11)
5165 LAM 3 (F11)
1955 LAM 11/ZWE (F9)
3517 LAM 11/ZWE (F9)
1503 M. tuberculosis LAM
2904 LAM 1 (F13)
2511 LAM 1 (F13)
5276 LAM 1 (F13)
5260 U (F26)
2262 U (F26)
3656 U (f26)
3388 T1, T1-Tusc, T5/Rus (F14)
3100 T1, T1-Tusc, T5/Rus (F14)
3651 T1, T1-Tusc, T5/Rus (F14)
605 M. tuberculosis LAM KZN MDR
2576 LAM 4, KZN (F15)
4207 M. tuberculosis LAM KZN DS
1435 M. tuberculosis LAM KZN XDR
1123 - (unknown) (F32)
2005 - (unknown) (F32)
5330 T5 (F21)
5440 - (unknown) (F12)
3839 T1 (F18)
Mtb_H37Rv M. tuberculosis H37Rv laboratory strain
1397 T1 (F17)
1543 - (unknown) (F16)
4302 U (F8)
478 T2 (F5)
337 T1 (F22)
4472 T (F23)
M_canettii M. canettii
EAI
M. africanum
Lineage 7
CAS
Beijing
LCC
Pre-Haarlem
Haarlem-like
Haarlem
Quebec
LAM
T, U

Figure 6: Clustering of M. tb clades produced seven distinct superclades.

4.3. Preliminary assessment of paired data

4.3.1. Genotype QC

4.3.1.1. SAC cohort

For genotyping QC, a total of 397 337 variants and 947 individuals were loaded from the genotype files in PLINK format. Three hundred and eighty one variants were excluded from the genotyping data set because they had no assigned chromosome or base pair position. Additionally, iterations of the filters for individual missingness (--mind), SNP missingness (--geno), and MAF (--maf) removed 11 individuals, 96 876 variants, and 31 516 variants from the dataset, respectively.

The "sex" check flagged 14 "problem" individuals in which either the "sex" in the PED file did not match the genotypes in the MAP file, or ambiguity was detected between the SNP data and pedigree data, and were subsequently removed from the dataset. Duplicate samples were identified and a pairwise analysis of Identity-By-Descent (IBD) showed that the dataset consisted of a number of related individuals. A pi-HAT threshold of 0.125 was chosen to filter at the level of first-cousin relatedness, resulting in 189 pairs of related individuals. In order to obtain the maximum set of unrelated individuals, the Perl script by Yekai Xiong identified 137 individuals who needed to be removed. The remaining 785 individuals were filtered for --mind/--geno/--maf resulting in 4 273 variants being filtered out, and leaving 264 291 variants for 785 unrelated individuals. The test for excess genotyping heterozygosity resulted in four individuals being removed, followed by the removal of 58 variants not meeting the MAF threshold, leaving 781 individuals and 264 233 variants.

Using a file containing the Sample IDs for the 525 participants having first infection records in the PS database, the genotype dataset of 781 unrelated individuals was filtered to extract only the genotypes of those participants having *M. tb* clade information. From this subset of samples passing QC, 439 of the 525 individuals having a record for their first infection were retained and assessed using PCA. It is likely due to the iterative filters for poor genotyping, sample missingness, and relatedness, that individuals having clade information were removed.

4.3.1.2. Ghanaian cohort

For the Ghanaian cohort, genotype data was available in PLINK format for a total of 783 338 variants for 3 311 samples. No variants required annotation of chromosome number or base pair position. Samples were filtered iteratively for 10% individual genotype missingness, 2% SNP genotype missingness and 5% MAF, resulting in no individuals, 73 322 variants, and 32 711 variants being removed, respectively. The "sex" check removed 24 samples, while an additional 93 samples were identified as being closely related to at least one other sample in the cohort. A test for excess heterozygosity resulted in an additional 47 samples being removed, leaving a dataset consisting of 3 147 samples and 677 305 variants. Of the 1 315 clade-matched samples, 1 273 samples were present in the dataset passing QC and were analysed using PCA.

4.3.2. *M. tb* database quality control

4.3.2.1. SAC cohort

The first QC measure involved assessing the dataframe for any duplicated genotypes based on the number of Sample IDs matched to strain information in the database. This analysis showed that three individuals were genotyped twice (Figure 7). Genotyping quality scores for each duplicated genotype were examined and the sample genotyped with the poorer quality score was excluded from further analysis.

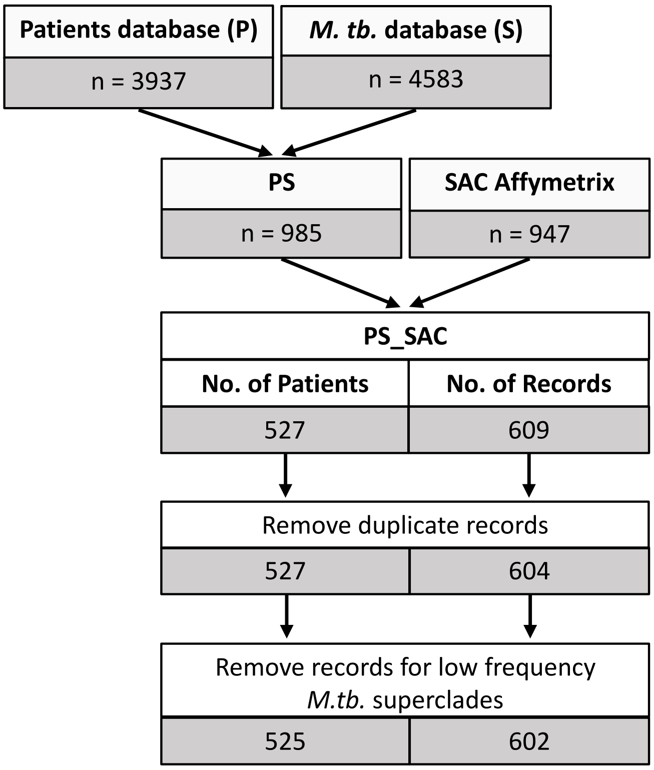


Figure 7: Filtering of Patient-Strain database yielded 525 genotyped study participants from the SAC cohort with information for the M. tb infection.

The second QC measure was to identify study participants having an incorrect infection number assigned to their record(s). This filter identified five individuals with incorrect infection numbers assigned. Generating a table of the number of records per Patient ID showed that a total of 459 participants had only one *M. tb* record, while one individual had four records in the PS dataframe (Table 8).

Table 8: Number of genotyped study participants with single- and multiple infections in the database

Number of records per Patient ID in PS database	Number of Study participants
1	459
2	60
3	7
4	1
Total number of study participants	**527**

4.3.2.2. Ghanaian cohort

For the Ghanaian cohort, no infection number was provided with this dataset, and thus all records were considered as the first infection for participants in this cohort.

4.3.3. Derive *M. tb* distributions

4.3.3.1. SAC cohort

Frequency distributions for clades and superclades were derived for a number of subsets of the genotyped and clade-matched study participants. *M. tb* clade- and superclade distributions for the first recorded infection in the SAC cohort are shown in Figure 8, and Figure 9, respectively. The distribution of *M. tb* clades and superclades for the second recorded infection is shown in Figure 10, and Figure 11, respectively. The first recorded infection in the database was dominated by the LAM clade and closely followed by Beijing, whereas the second infection was dominated by the Beijing clade and followed by LAM. There were no cases of HaarlemLike reported for the second infection. The frequency distribution of superclades showed similar frequencies for LAM, BeijingCAS1, and HaarlemsLCC for the first recorded infection (Figure 9). As seen in Figure 9, LAM dominated the superclade distribution in the first infection, but was only third in abundance in the second infection, dominated by the BeijingCAS1 superclade (Figure 11).

The *M. tb* clade- and superclade distribution amongst participants having one-, and two infections in the database is shown in Figure 31. LAM, followed by the Beijing clade dominated the frequency distribution for participants having only one recorded infection in the database. In contrast, Beijing was followed by LAM in the clade distribution amongst participants with two infections recorded in the database, while the distribution of the remaining six clades followed the same pattern in both subsets (Figure 31 **A** and **C**). After grouping the clades into superclades, the frequency distribution of LAM, HaarlemsLCC, and BeijingCAS1 dominated both subsets of participants having one- and two recorded infections while Other and Quebec superclades were in least abundance (Figure 31 **B** and **D**).

For the participants having two infections, an interesting observation was made when stratifying the 120 records by infection number. The dominant clade for the first of two infections was LAM followed by Beijing and Haarlem (Figure 32 **A** in 7.1), whereas for the same cohort of 60 participants, records for the second of two infections were dominated by Beijing, followed by LCC and LAM (Figure 32 **C**). Frequency distributions for the same cohort based on superclade showed a combination of both distributions (Figure 32 **B** and **D**) where the first infection of two was dominated by the HaarlemsLCC superclade, followed by LAM, and then Beijing.

4.3.3.2. Ghanaian cohort

Frequency distributions were derived for clade and superclade definitions used in this cohort and are shown in Figure 12, and Figure 13, respectively. After grouping the clades into superclade categories, the EAI_afri and LAM_CAM superclades dominated the distribution in the cohort with both groupings having more than 400 records (Figure 13). The Haarlem_X, and T_U superclades had a frequency of approximately 160, and 250 in the cohort, respectively, while the BeijingCAS and Ghana-2 superclades were in least abundance (Figure 13). All *M. tb* clade information available for this cohort was for one recorded infection; no data were available regarding previous infections or multiple infections.

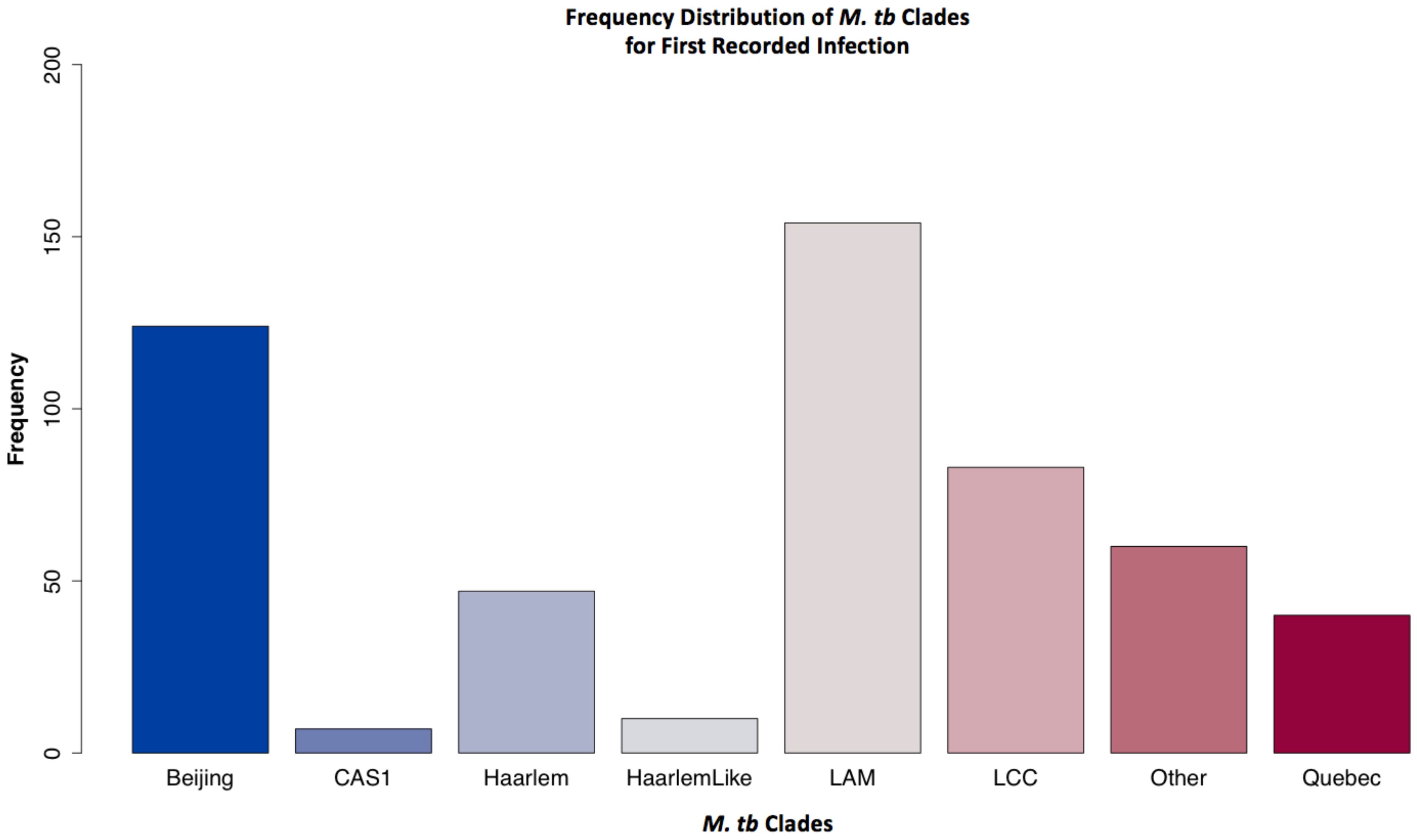


Figure 8: Frequency distributions of M. tb Clades for the First infection records (n = 525) in the SAC cohort

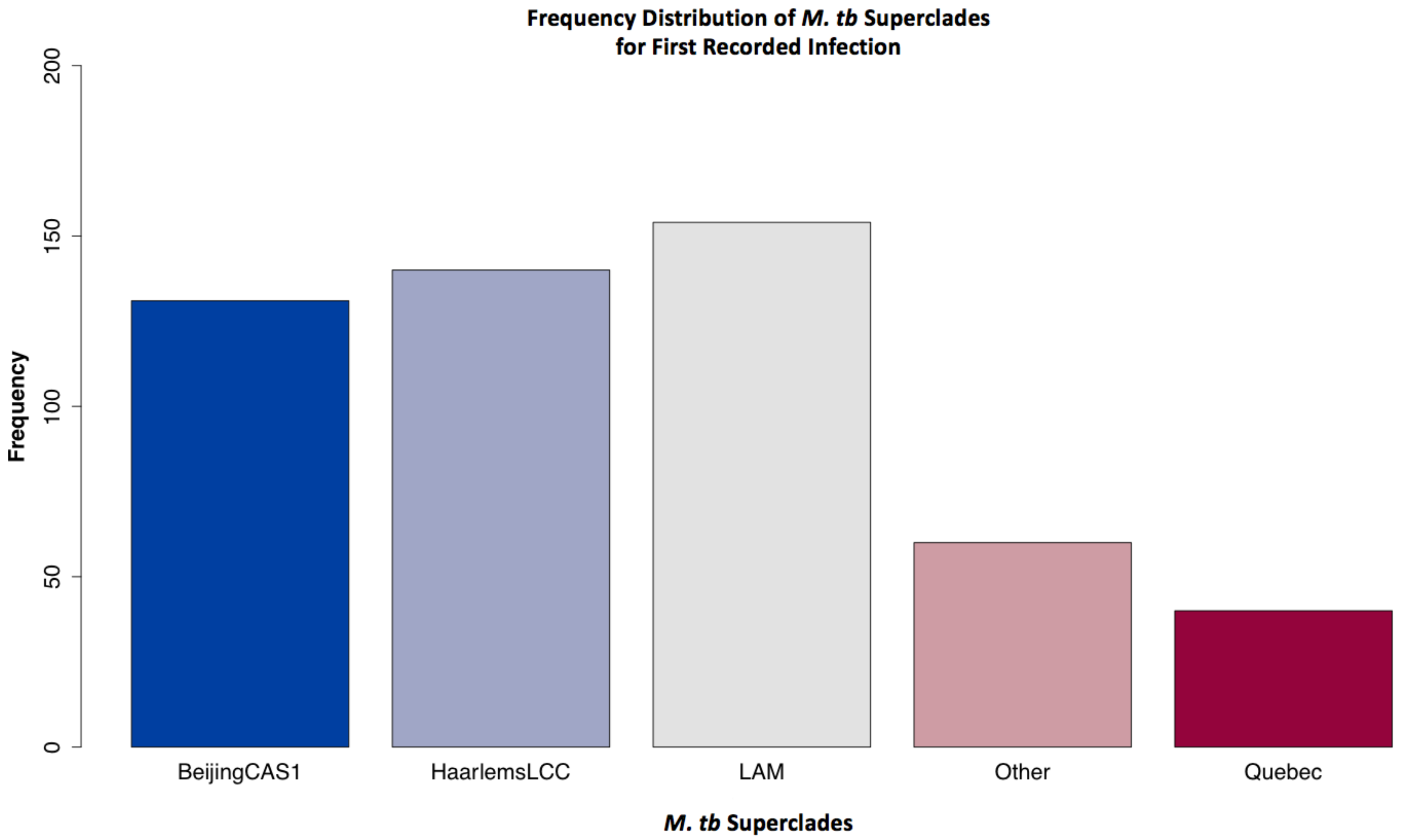


Figure 9: Frequency distributions of M. tb Superclades for the First infection records (n = 525) in the SAC cohort

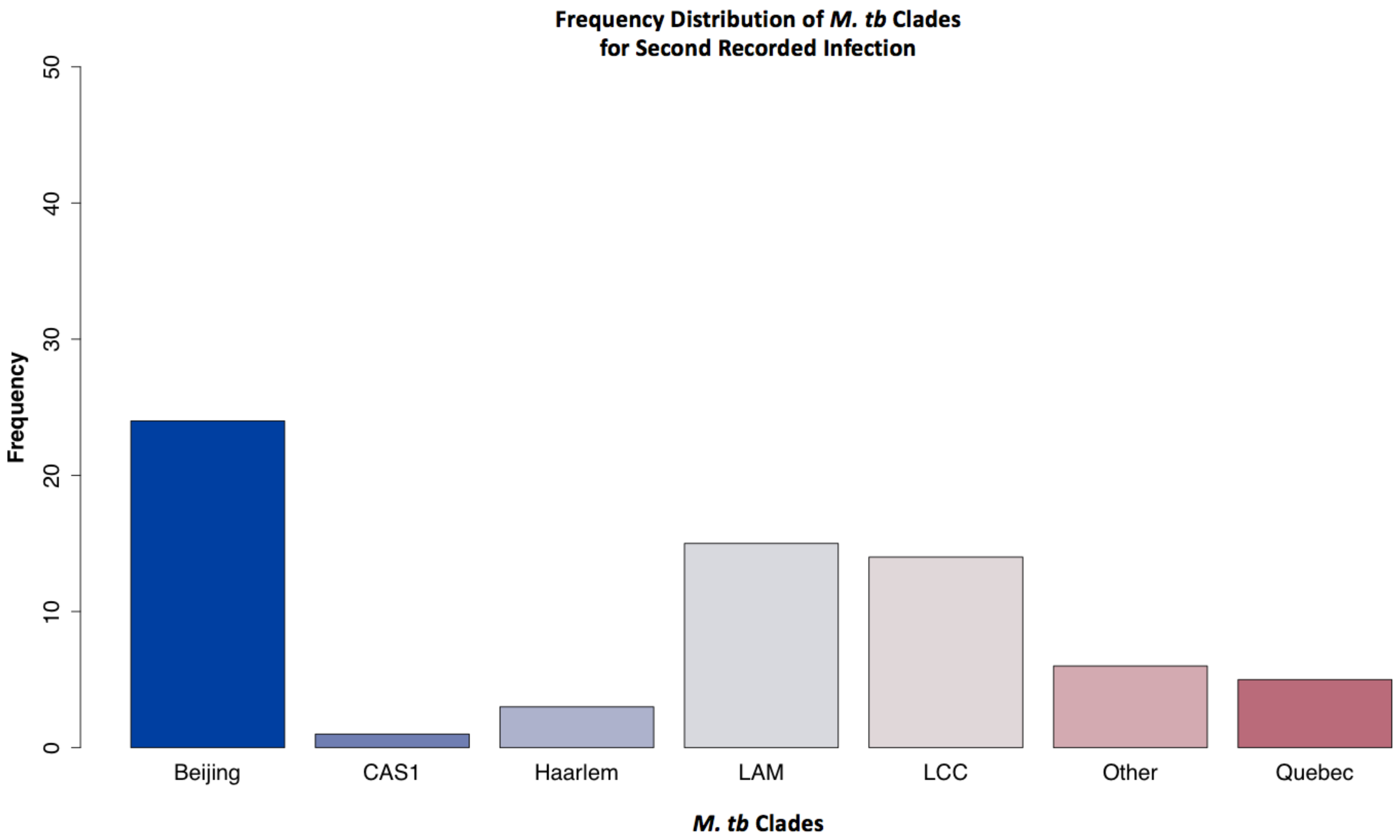


Figure 10: Frequency distributions of M. tb Clades for the Second infection records (n = 60) in the SAC cohort

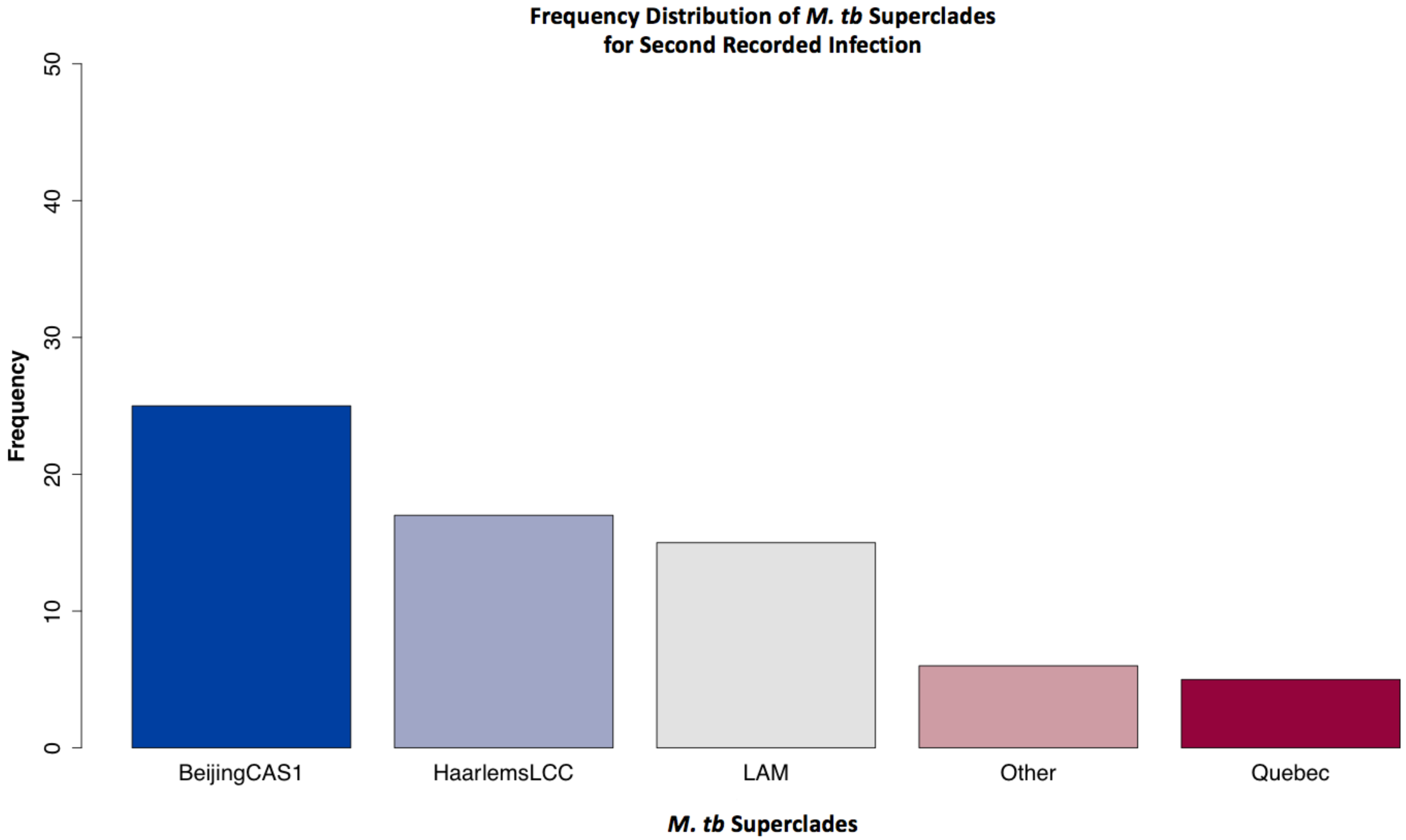


Figure 11: Frequency distributions of M. tb Superclades for the Second infection records (n = 60) in the SAC cohort

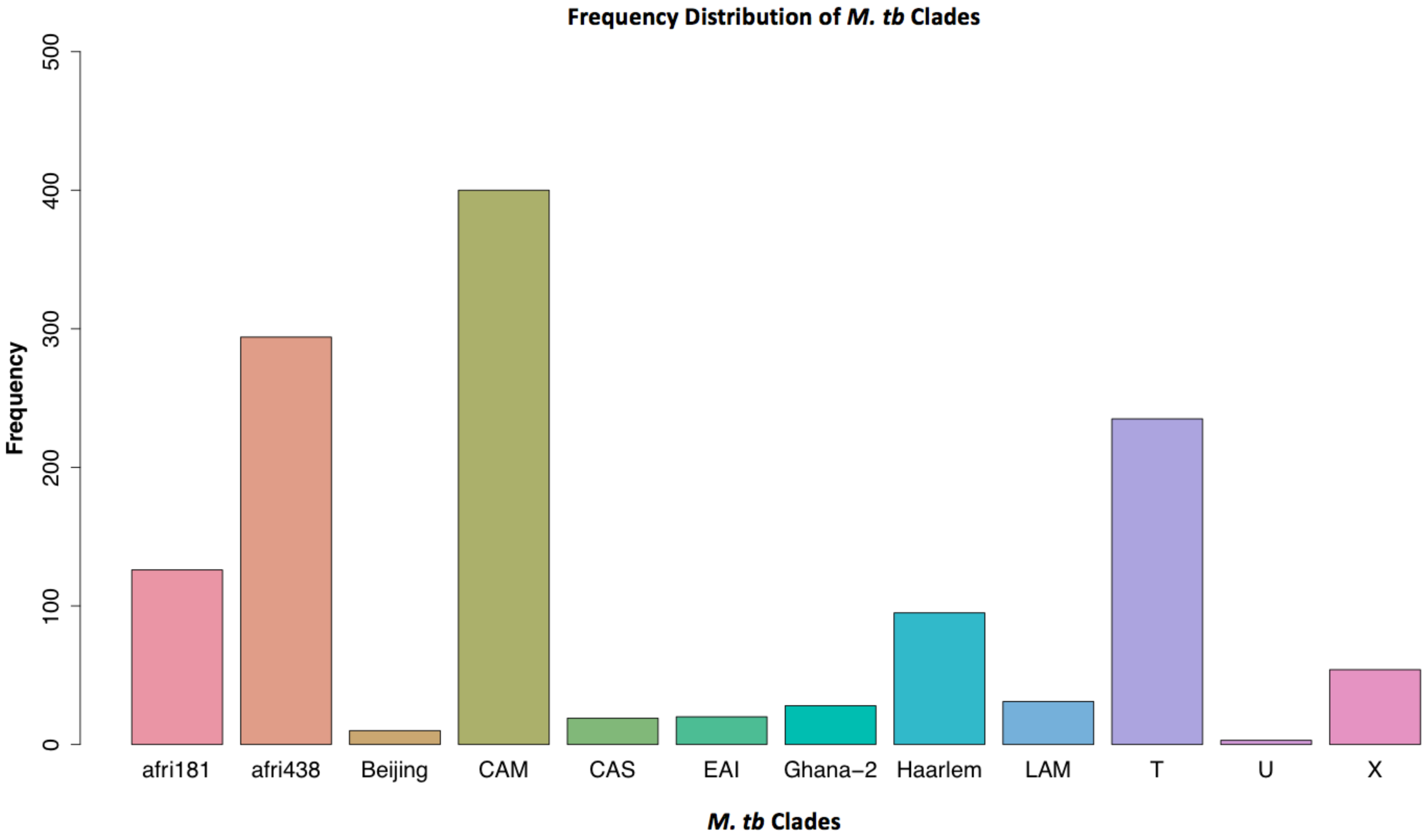


Figure 12: Frequency distributions of M. tb Clades (n= 1 318) in the Ghanaian cohort

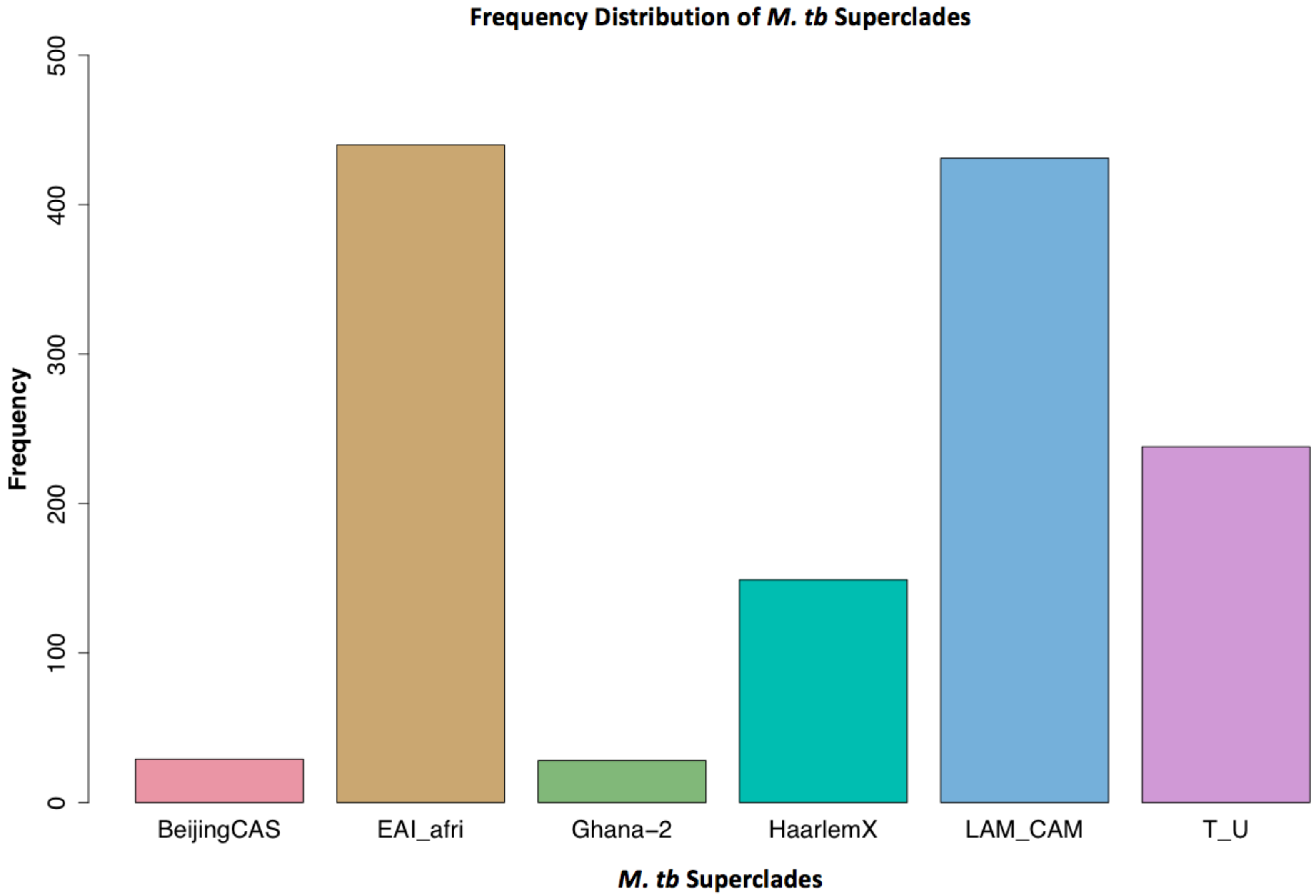


Figure 13: Frequency distributions of M. tb Superclades (n= 1 318) in the Ghanaian cohort

4.3.4. PCA of first recorded infection

4.3.4.1. SAC cohort

After loading the compressed PLINK files for PCA, 7 072 non-autosomal SNPs were excluded, and the working space consisted of 439 samples and 257 161 SNPs. A scree plot consisting of 32 eigenvalues was plotted (Figure 14) and showed that 10% of the variance in the data could be explained by the principal components generated. A cut-off was chosen at the point of Principal Component (PC) 3, as the variation in the data became least explicable after the third PCPC (Figure 14).

Clade, and superclade distributions and PCA plots for the 439 samples that passed genotyping QC were generated and are shown in Figure 15 and Figure 16, respectively. The overlay of either clade or superclade data for the first recorded infection showed no separation of the genotypes into distinct clusters based on *M. tb* designations. Clade frequency distributions for the SAC cohort showed that the first recorded infection was mostly caused by LAM and that there were few infections caused by CAS1 or HaarlemLike (Figure 15 **A**). No extreme outliers were evident from the PCAs generated for the SAC cohort.

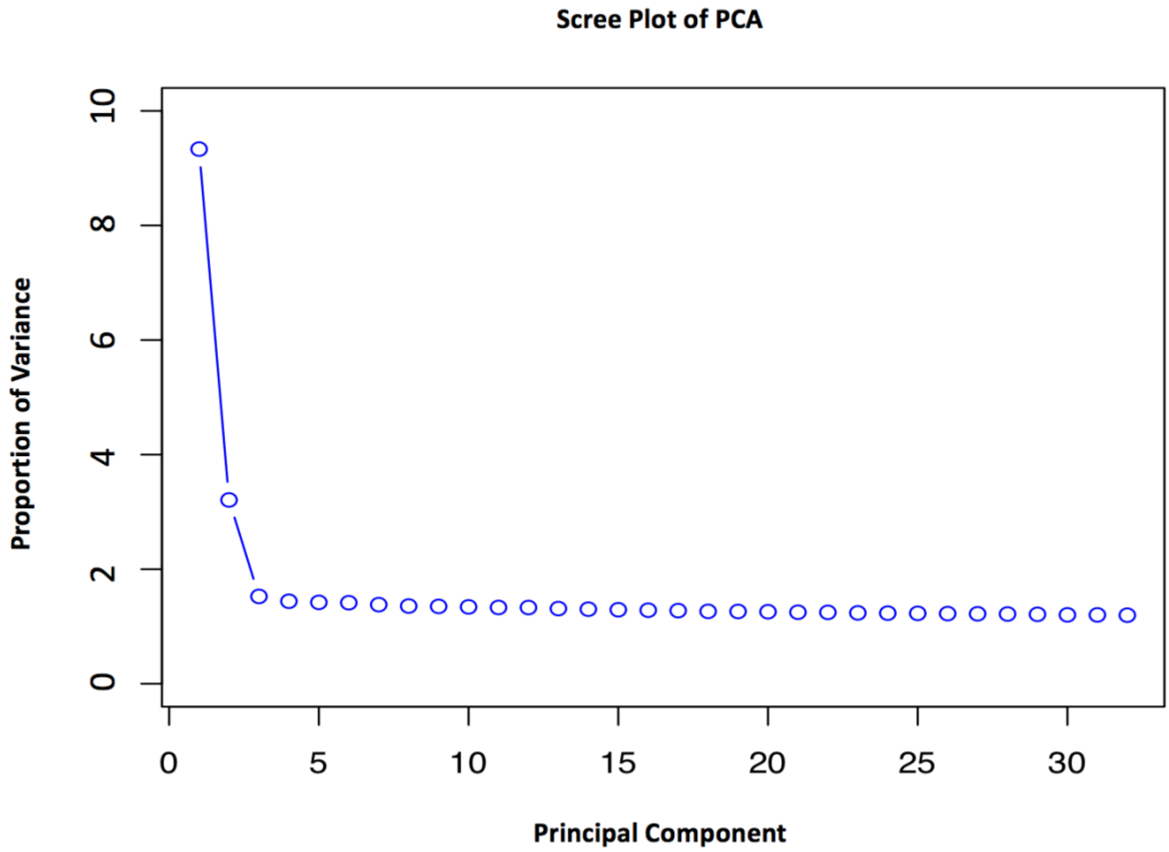


Figure 14: Scree plot for the First recorded infection in the SAC cohort passing genotyping QC (n = 439).

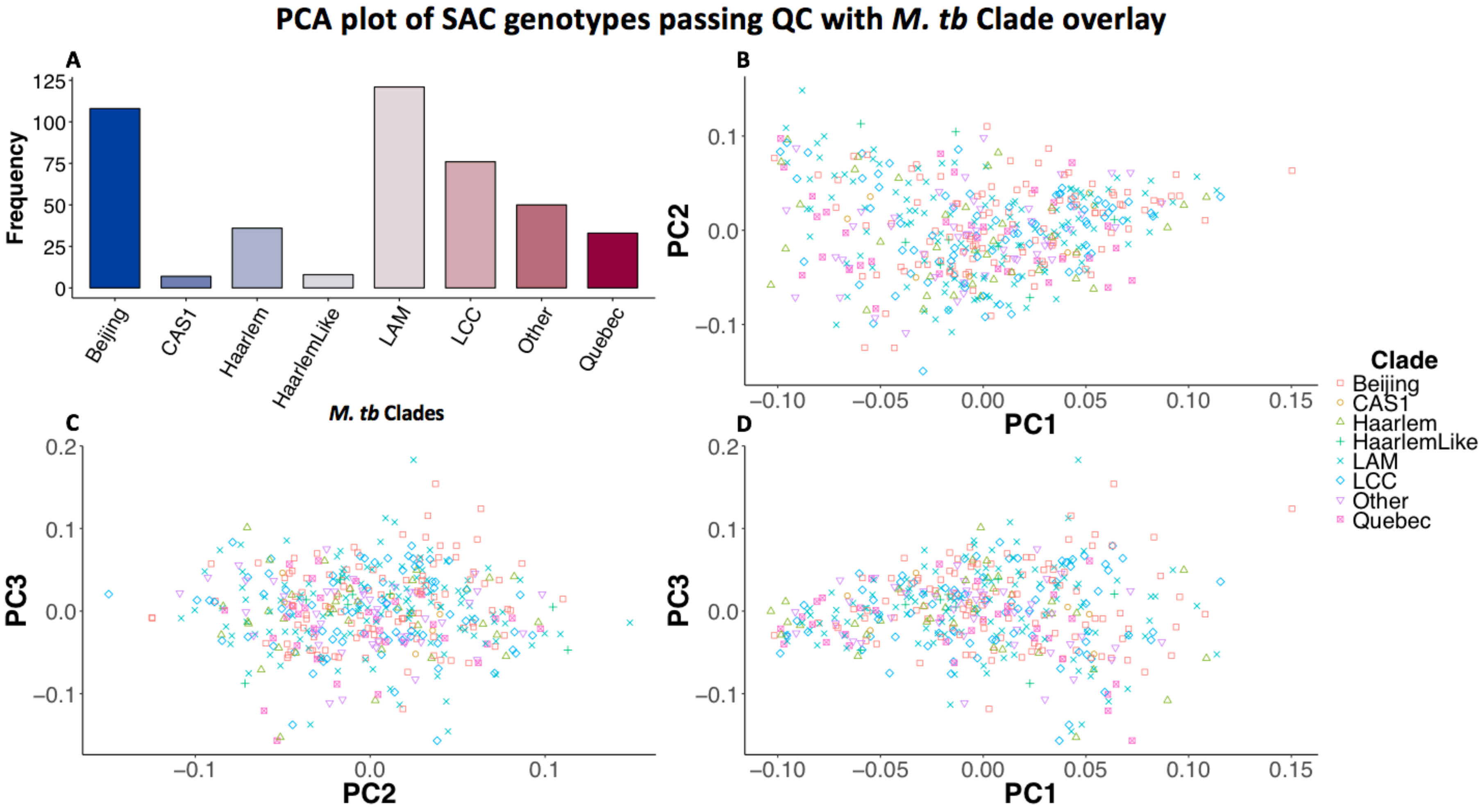


Figure 15: Clade distribution (A) and PCA of 439 genotyping records with the first infection recorded in the SAC cohort. Figures B, C, and D show PC1 and PC2, PC2 and PC3, and PC1 and PC3, respectively.

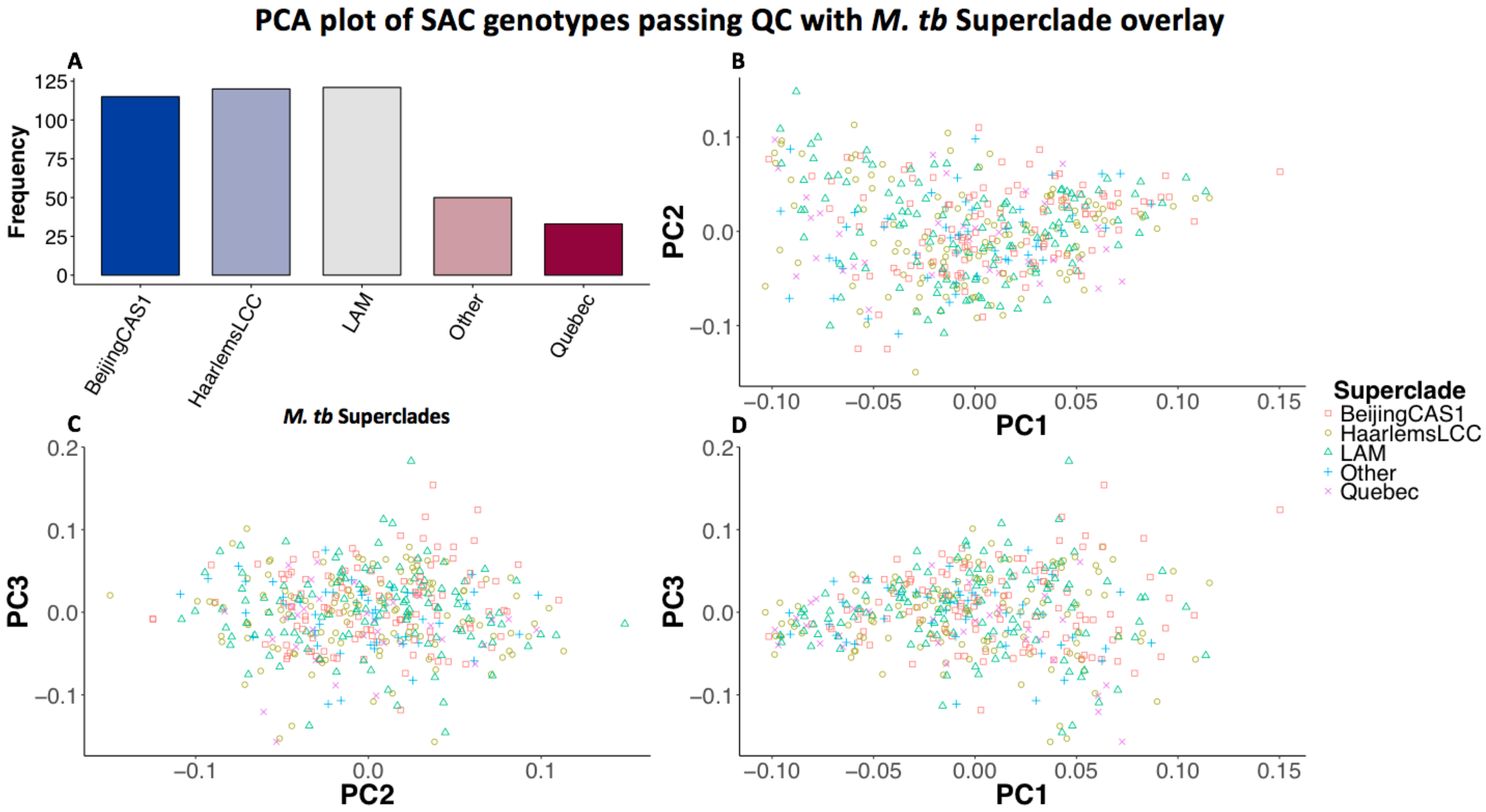


Figure 16: Superclade distribution (A) and PCA of 439 genotyping records with the first infection recorded in the SAC cohort. Figures B, C, and D show PC1 and PC2, PC2 and PC3, and PC1 and PC3, respectively.

4.3.4.2. Ghanaian cohort

No non-autosomal SNPs were detected and thus the working space consisted of 1 273 samples having 677 405 variants. A scree plot consisting of 32 eigenvalues was generated and showed that a very small proportion (2%) of the observed variance could be explained by the first three principal components generated (Figure 17). PC plots were generated for the first three principal components. The clade and superclade distributions for the 1 273 samples passing the QC filters was determined and is shown in Figure 18 **A** and Figure 19 **A**. As with the SAC dataset, the overlay of clade and superclade information showed no separation of the genotypes into distinct clusters as shown in the PC plots in Figure 18 and Figure 19. The LAMCAM superclade dominated the first recorded infection, while the Ghana-2 superclade was least abundant in this cohort. Furthermore, no extreme outliers were evident from the PCA.

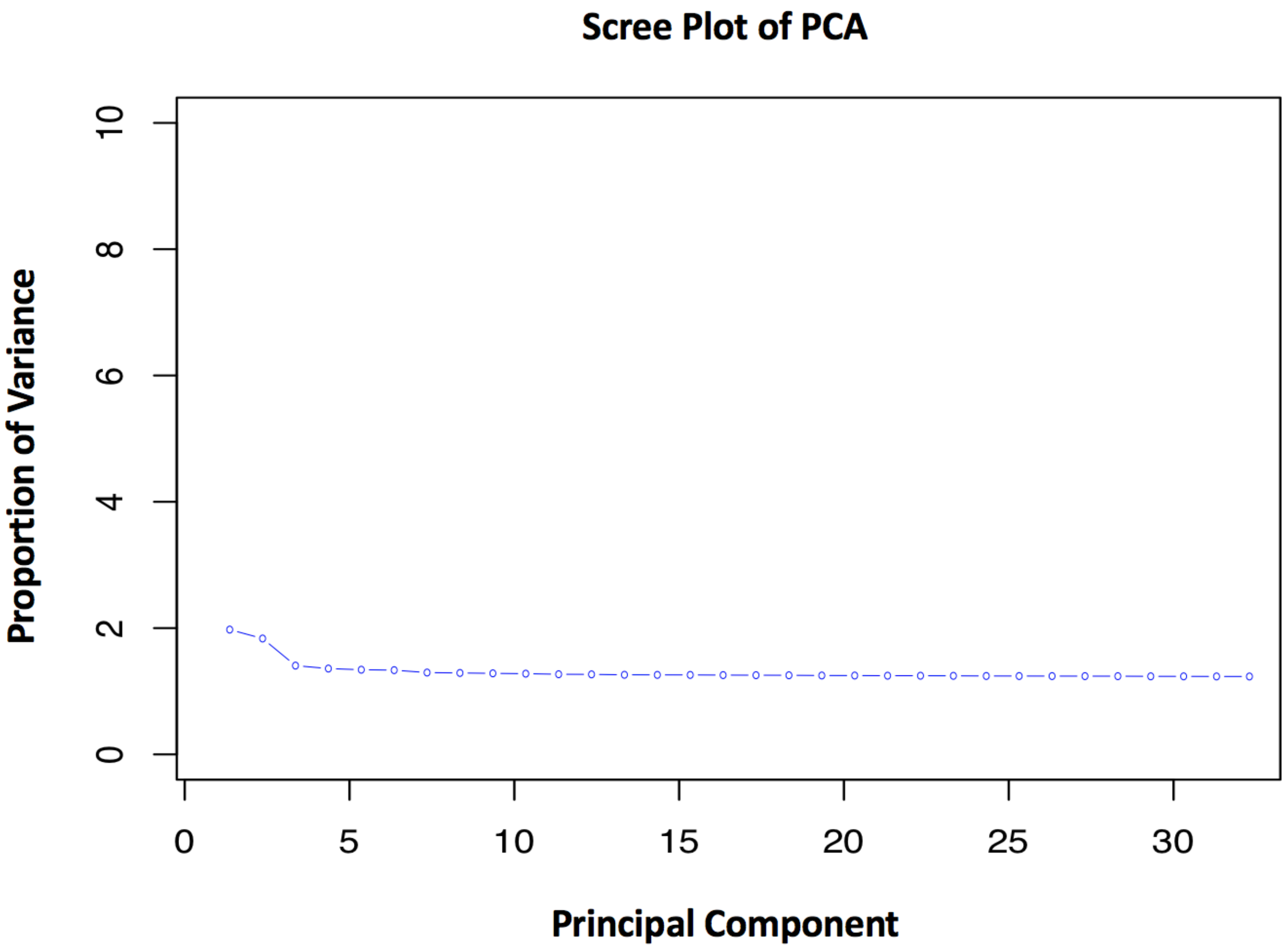


Figure 17: Scree plot for the First recorded infection in the Ghanaian cohort passing genotyping QC (n = 1 273). The PCA was able to account for a maximum of 2% of the variation in the patient genotypes.

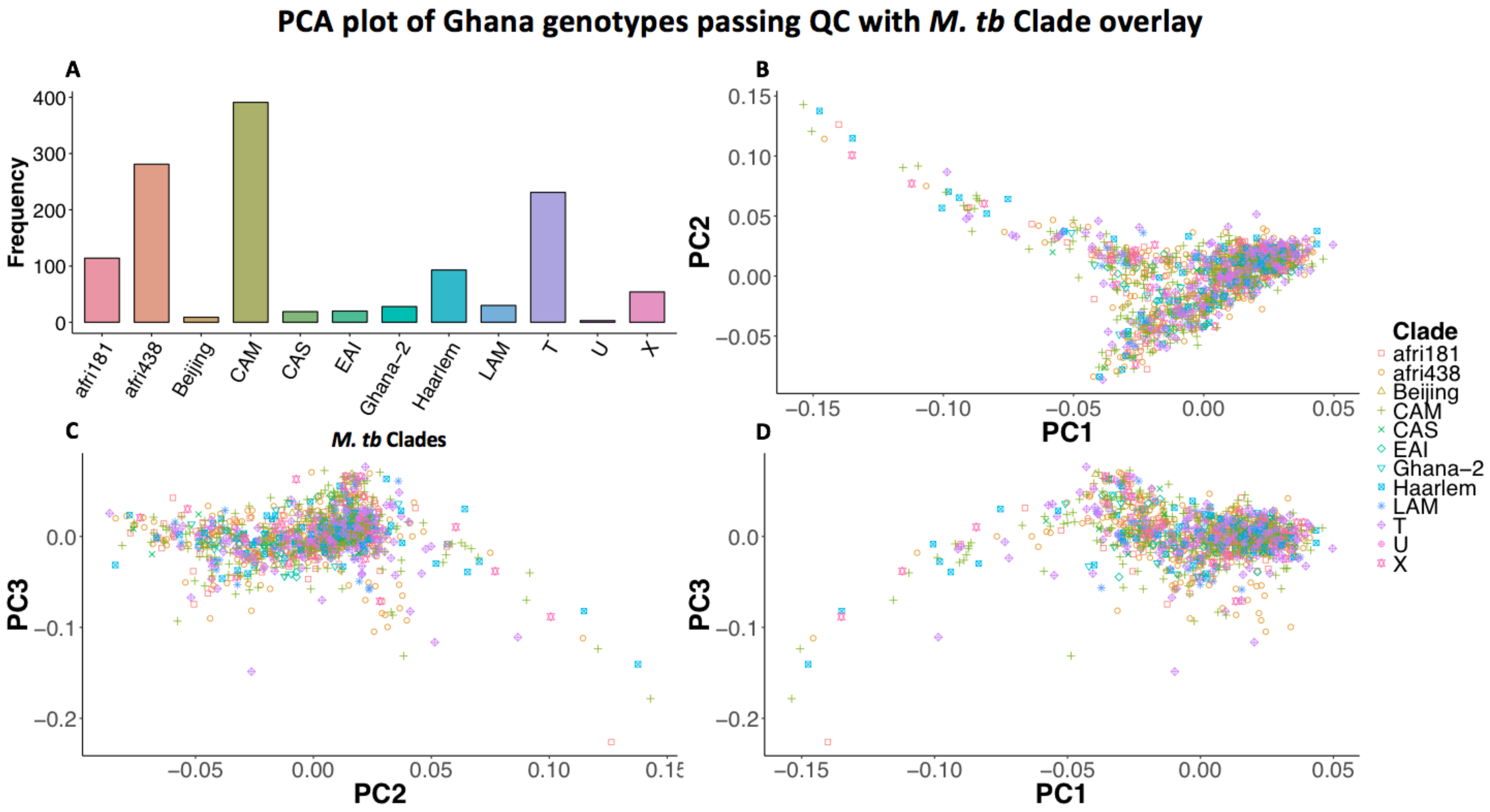


Figure 18: Clade distribution (A) and PCA of 1 273 genotyping records with the first infection recorded for the Ghanaian cohort. Figures B, C, and D show PC1 and PC2, PC1 and PC3, and PC2 and PC3, respectively.

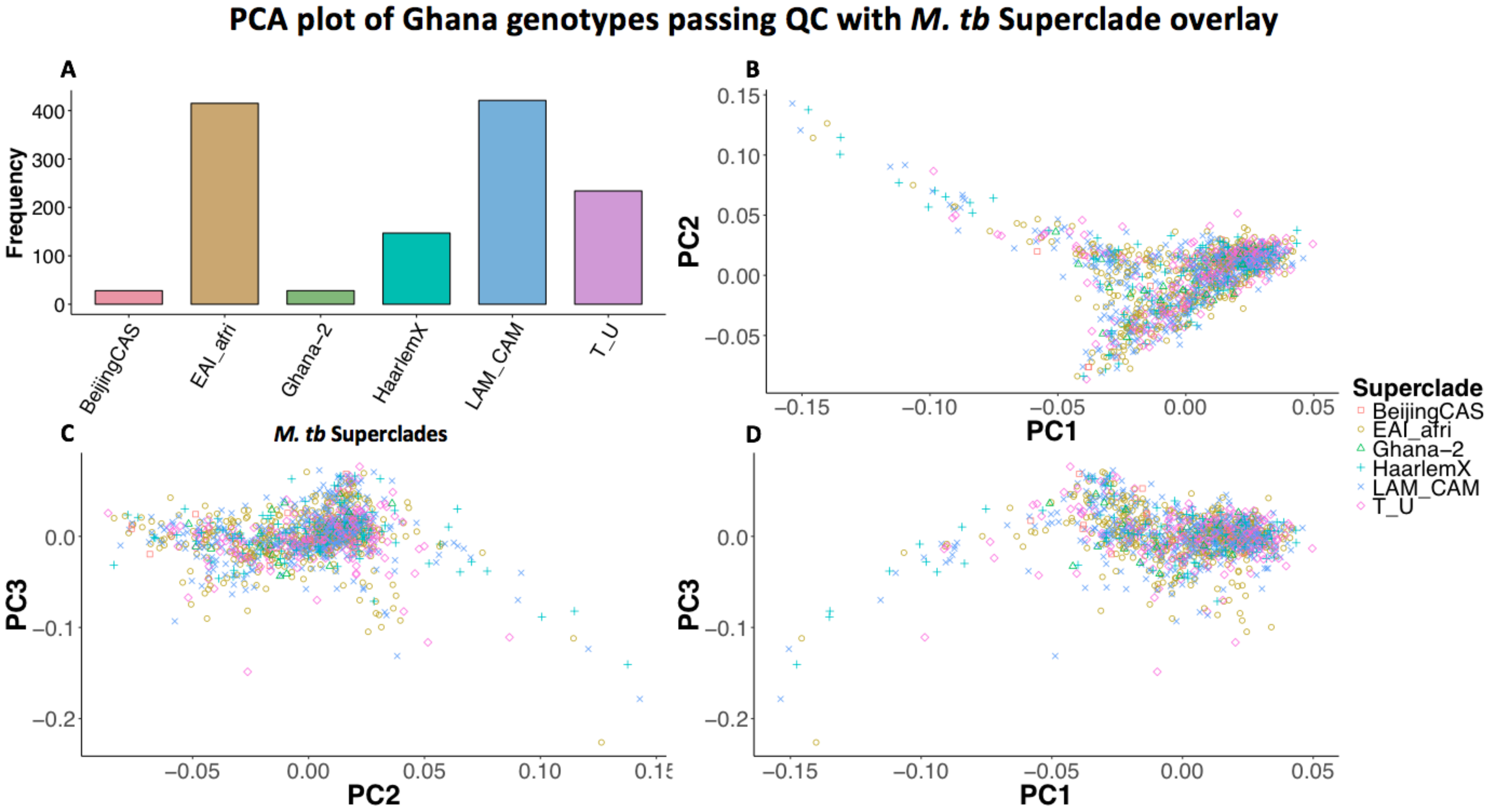


Figure 19: Superclade distribution (A) and PCA of 1 273 genotyping records with the first infection recorded for the Ghanaian cohort. Figures B, C, and D show PC1 and PC2, PC2 and PC3, and PC1 and PC3, respectively.

4.4. Generate high-quality imputed genotype data

4.4.1. Modified Data QC for Imputation

To prepare the data for imputation, a modified QC method was designed as shown in Figure 5. General genotyping QC steps were modified to include the updating of variants lacking base pair or chromosome information using the dbSNP database, and checking for SNP and strand flips using Genotype Harmonizer. These steps ensured that the data being imputed were correctly aligned to the most widely-used reference panel, the 1000GP3, and that the maximum number of the originally genotyped variants were assessed in the genotyping QC steps that followed. The results of this QC method are shown in Figure 20 for the SAC dataset and Figure 21 for the Ghanaian dataset.

The initial SAC dataset comprised of 947 participants genotyped for 397 337 variants. After pre-imputation QC, imputation, and post-imputation QC, the SAC dataset consisted of 445 clade-matched participants with 7 145 406 variants after imputation with the AGR reference panel yielding the highest quality imputed genotypes (Figure 20).These 7 145 406 variants for 445 participants were the evaluated for association using a multinomial logistic regression.

The Ghanaian dataset comprised of 3 311 participants with genotypes for 783 338 variants. The 1000GP3 reference panel imputed using our In-House method yielded the highest quality imputed genotypes across several MAF bins. After QC, the dataset consisted of 5 275 890 variants for 1 272 clade-matched individuals, which were then assessed for association using a multinomial logistic regression (Figure 21). The subsections to follow describe the filtering of both datasets in greater detail.

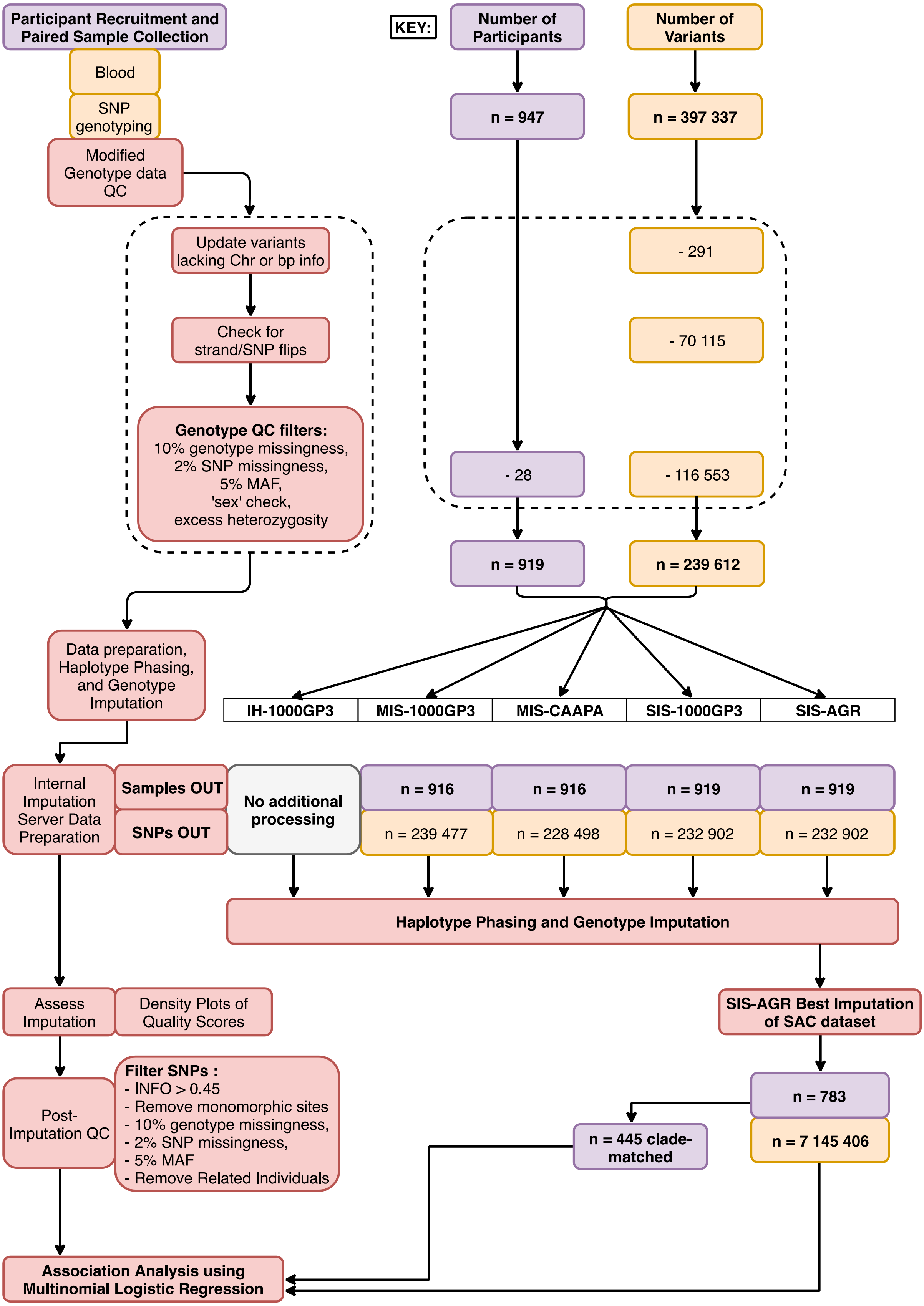

Participant Recruitment and Paired Sample Collection
Blood
SNP genotyping
Modified Genotype data QC
Update variants lacking Chr or bp info
Check for strand/SNP flips
Genotype QC filters:
10% genotype missingness,
2% SNP missingness,
5% MAF,
'sex' check,
excess heterozygosity
KEY:
Number of Participants
Number of Variants
n = 947
n = 397 337
- 291
- 70 115
- 28
- 116 553
n = 919
n = 239 612
Data preparation, Haplotype Phasing, and Genotype Imputation
IH-1000GP3
MIS-1000GP3
MIS-CAAPA
SIS-1000GP3
SIS-AGR
Internal Imputation Server Data Preparation
Samples OUT
SNPs OUT
No additional processing
n = 916
n = 916
n = 919
n = 919
n = 239 477
n = 228 498
n = 232 902
n = 232 902
Haplotype Phasing and Genotype Imputation
Assess Imputation
Density Plots of Quality Scores
Post-Imputation QC
Filter SNPs :
- INFO > 0.45
- Remove monomorphic sites
- 10% genotype missingness,
- 2% SNP missingness,
- 5% MAF
- Remove Related Individuals
SIS-AGR Best Imputation of SAC dataset
n = 783
n = 7 145 406
n = 445 clade-matched
Association Analysis using Multinomial Logistic Regression

Figure 20: Results of filtering the SAC dataset using the modified genotyping QC method prior to imputation

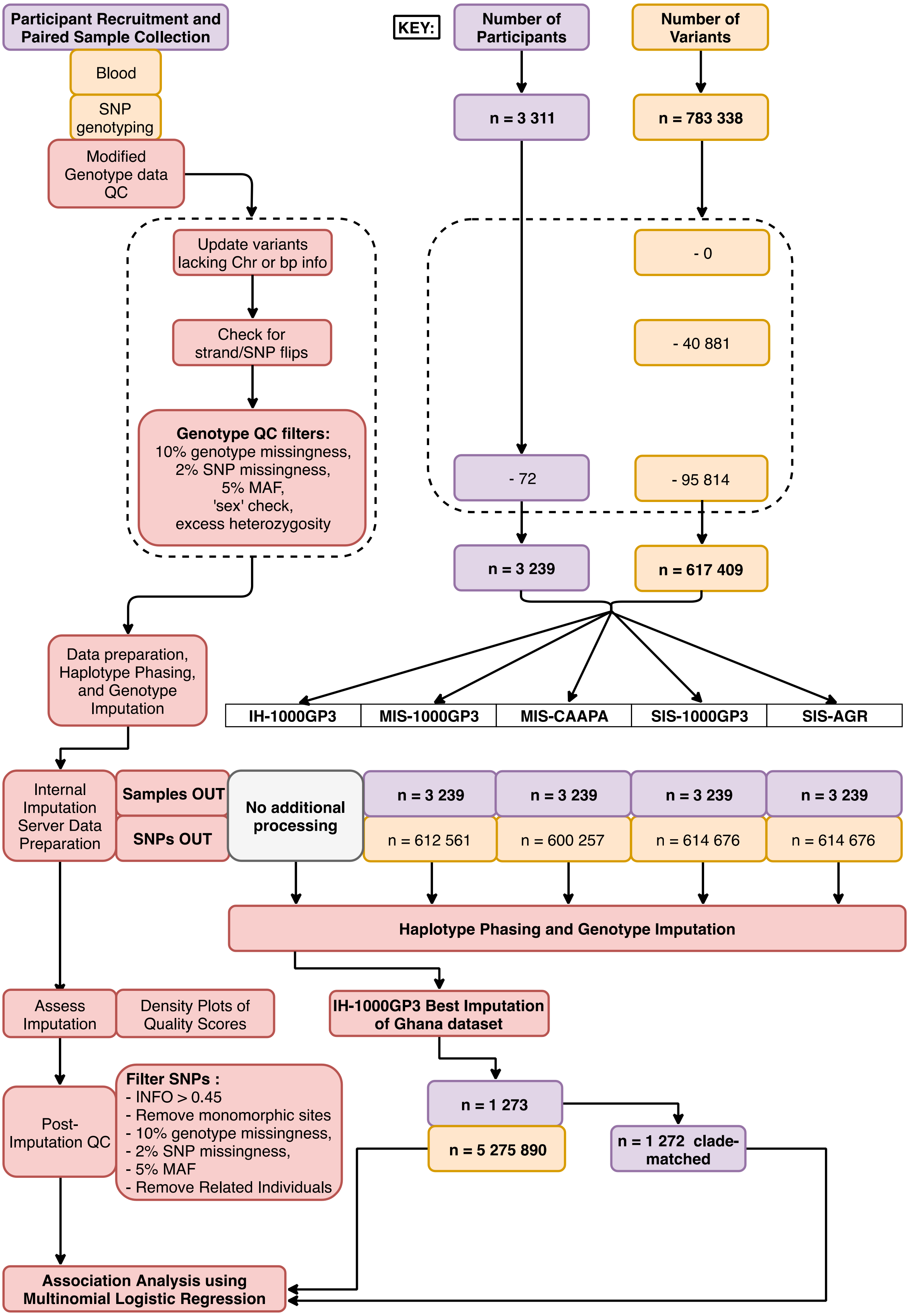

Participant Recruitment and Paired Sample Collection
Blood
SNP genotyping
Modified Genotype data QC
Update variants lacking Chr or bp info
Check for strand/SNP flips
Genotype QC filters:
10% genotype missingness,
2% SNP missingness,
5% MAF,
'sex' check,
excess heterozygosity
KEY:
Number of Participants
Number of Variants
n = 3 311
n = 783 338
- 0
- 40 881
- 72
- 95 814
n = 3 239
n = 617 409
Data preparation, Haplotype Phasing, and Genotype Imputation
IH-1000GP3
MIS-1000GP3
MIS-CAAPA
SIS-1000GP3
SIS-AGR
Internal Imputation Server Data Preparation
Samples OUT
SNPs OUT
No additional processing
n = 3 239
n = 3 239
n = 3 239
n = 3 239
n = 612 561
n = 600 257
n = 614 676
n = 614 676
Haplotype Phasing and Genotype Imputation
Assess Imputation
Density Plots of Quality Scores
IH-1000GP3 Best Imputation of Ghana dataset
Post-Imputation QC
Filter SNPs :
- INFO > 0.45
- Remove monomorphic sites
- 10% genotype missingness,
- 2% SNP missingness,
- 5% MAF
- Remove Related Individuals
n = 1 273
n = 5 275 890
n = 1 272 clade-matched
Association Analysis using Multinomial Logistic Regression

Figure 21: Results of filtering the Ghana dataset using the modified genotyping QC method prior to imputation

4.4.1.1. Update variants lacking chromosome or base pair position information

4.4.1.1.1. SAC cohort

A total of 381 SNPs from the genotype files did not have a chromosome or base pair position assigned. Instead of immediately excluding these SNPs, as was done in 3.3.1, chromosome number and base pair positions were retrieved, where possible, using the 1000GP3 reference in conjunction with validation of the information in the dbSNP database (Figure 20). The percentage overlap between the study data (397 046 SNPs) and the 1000GP3 was calculated to be 97.26 %.

Of the 381 SNPs, 91 were found in the 1000GP3. All 91 SNPs found in the 1000GP3 were also found in the dbSNP database. However, one SNP, rs41516445 located on chromosome 5 was flagged in dbSNP as not being validated by laboratory methods, and was thus excluded from the list of SNPs to update. The 90 variants for which chromosome number and base pair position could be retrieved in the 1000GP3, and were verified in dbSNP, were successfully updated in the PLINK genotype files. The 291 SNPs for which information could not be updated were then written to file and excluded using PLINK, yielding a BIM file containing 397 046 variants with a genotyping rate of 98.6 % for 947 samples (Figure 22).

4.4.1.1.2. Ghanaian cohort

No variants in the Ghanaian genotype files lacked chromosome or base pair information. The percentage overlap between the genotype data (783 338 SNPs) and the 1000GP3 reference panel was calculated to be 99.48 %.

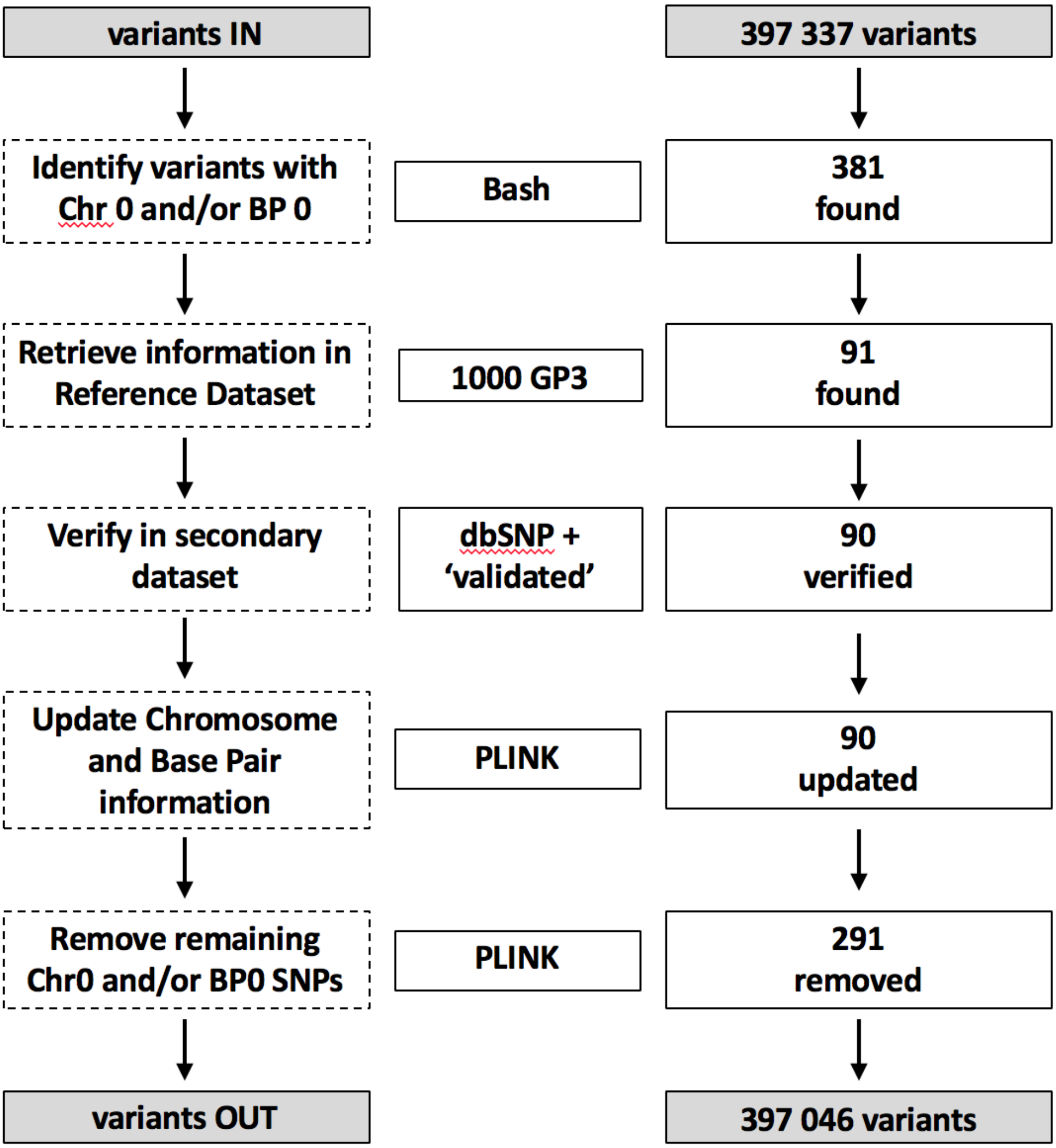

variants IN
Identify variants with Chr 0 and/or BP 0
Retrieve information in Reference Dataset
Verify in secondary dataset
Update Chromosome and Base Pair information
Remove remaining Chr0 and/or BP0 SNPs
variants OUT
Bash
1000 GP3
dbSNP + 'validated'
PLINK
PLINK
397 337 variants
381 found
91 found
90 verified
90 updated
291 removed
397 046 variants

Figure 22: Schema of SNP annotation update workflow for SAC cohort

4.4.1.2. Examine the data for ambiguity in alignment to reference data

4.4.1.2.1. SAC cohort

A total of 947 samples with 397 046 SNPs were analysed using Genotype Harmonizer. Of the SNPs loaded, the first iteration aligned 329 524 non-A/T and non-G/C (i.e. non-ambiguous) SNPs, and swapped 317 SNPs to match the allele order of the 1000GP3. A total of 21 out of 63 429 ambiguous SNPs were swapped based on their LD patterns. A total of 40 881 variants were excluded during the alignment phase: 2 786 due to the study variant being found in the reference dataset, but the alleles not being comparable, and 38 095 variants being excluded due to there not being enough non-ambiguous SNPs in LD to assess the strand based on LD. No non-biallelic SNPs were found, leaving 947 samples with 356 165 variants.

4.4.1.2.2. Ghanaian cohort

The genotype files of 3 311 samples with 783 338 SNPs were loaded into Genotype Harmonizer. Of the SNPs loaded, the first iteration aligned 656 121 non-A/T and non-G/C SNPs, and swapped 376 SNPs to match the allele order of the 1000GP3. A total of 36 out of 118 868 ambiguous SNPs were swapped based on their LD patterns, and a total of 70 115 variants were excluded during the alignment phase. No non-biallelic SNPs were found, leaving 3 311 samples with 713 223 variants.

4.4.1.3. Genotype QC

4.4.1.3.1. SAC cohort

A total of 28 individuals were removed - 11 due to individual genotype missingness, 13 due to ambiguity in the “sex” assignment, and 4 due to excess heterozygosity in their genotypes. Furthermore, 116 553 variants were removed during the iterative process of filtering for SNP missingness and MAF, leaving a dataset consisting of 919 individuals and 239 612 variants that passed QC with a genotyping rate of 99.39 %.

4.4.1.3.2. Ghanaian cohort

A total of 72 individuals were removed - 24 due to ambiguity in the “sex” assignment, and 48 individuals removed due to excess heterozygosity. Furthermore, 95 814 variants were removed during the iterative process of filtering for SNPSNP missingness and MAF, leaving a dataset consisting of 3 239 individuals and 617 409 variants with a genotyping rate of 99.39 %.

4.4.2. Data preparation, Haplotype phasing, and Genotype imputation

4.4.2.1. In-house (IH) workflow

No additional data preparation was required for the IH workflow using ShapeIT2 and IMPUTE2. For the IH workflow, a data summary is not produced as a standard output for imputation using IMPUTE2 software.

4.4.2.2. Michigan Imputation Server workflow

The MIS produces a data summary as shown for both cohorts in Table 9.

Table 9: Summary of data pre-processing on the Michigan Imputation Server

	SAC		**Ghana**	
	MIS-1000GP3	**MIS-CAAPA**	**MIS-1000GP3**	**MIS-CAAPA**
Chromosomes	1-23	1-22	1-22	1-22
Samples IN	**919**	**919**	**3 239**	**3 239**
No. of SNPs IN	**239 612**	**233 309**	**617 409**	**617 409**
No. "sex" detected and therefore filtered:	3	0	0	0
Alternative allele frequency > 0.5 sites:	166	0	0	0
Reference Overlap	99.49	97.94	99.55	97.23
Match	164 643	153 863	420 805	412 495
Allele Switch	74 116	68 912	176 073	172 432
Strand flip	1	0	0	0
Strand flip and allele switch	0	0	1	0
A/T, C/G genotypes	6 090	5 723	15 683	15 330
Filtered sites:				
Filter flag set	0	0	0	0
Invalid alleles	0	0	0	0
Duplicated sites	0	0	0	0
NonSNP sites	0	0	0	0
Monomorphic sites	0	0	0	0
Allele mismatch	54	8	2 051	25
SNPs call rate < 90%:	0	0	0	0
Excluded sites in total:	55	8	2 052	25
Remaining sites in total (before imputation)	**239 477**	**228 498**	**612 561**	**600 257**
Samples OUT	**916**	**916**	**3 239**	**3 239**

4.4.2.3. Sanger Imputation Server workflow

4.4.2.3.1. SAC cohort

The Sanger Imputation Server (SIS) does not allow for any mismatches between the reference dataset and the study dataset. Per the recommendations of the tool, the first analysis of the dataset through the fixref plugin indicated that 32.7% of the SNPs were mismatched to the human reference genome (human_g1k_v37.fasta). After aligning alleles to the human reference, a second run of the fixref plugin indicated 993 SNPs (0.4%) in the study dataset were unresolved. A third alignment of the study alleles to the human reference, flagged 406 mismatch SNPs (0.2%). These 406 SNPs were removed from the dataset, leaving a total of 232 902 SNPs completely aligned to the orientation of the reference dataset. This data was uploaded to the SIS for imputation with the 1000GP3, and AGR reference panels.

4.4.2.3.2. Ghanaian cohort

The first analysis of the dataset through the fixref plugin reported that 31.4% of the SNPs were mismatched to the human reference (human_g1k_v37.fasta). After aligning SNPs, a second run of the plugin indicated 0.4% of the SNPs in the study dataset were mismatched to the reference. A third attempt was made to align the alleles but when the number of reference mismatches remained unchanged, 2 453 SNPs were removed from the dataset, leaving a total of 614 676 SNPs completely aligned to the reference dataset. This data was uploaded to the SIS for imputation in two independent analyses: one using the 1000GP3 reference panel, and the second using the AGR reference panel.

4.4.3. Selection of high-quality imputed genotype data

4.4.3.1. Comparison of SNP density obtained from each workflow

SNP densities were calculated for the proportion of SNPs with a quality score metric greater than 0.45. For the SAC cohort, the SIS workflow using the AGR resource imputed the highest proportion of SNPs (Table 10, Figure 23 **E**, Figure 24 **D**), whereas for the Ghanaian cohort, the MIS workflow using the CAAPA resource imputed the greatest proportion of SNPs with a quality metric greater than 0.45 (Table 10, Figure 25 **D**, Figure 26 **D)**.

Table 10: Percentage proportion of SNPs with a quality metric greater than 0.45

	SAC		**GHANA**	
	Chr 1	**Chr X**	**Chr 1**	**Chr 22**
IH-1000GP3[1]	39	29	38	39
MIS-1000GP3[2]	32	18	49	45
MIS-CAAPA[3]	22	-	56	50
SIS-1000GP3[4]	36	32	41	40
SIS-AGR[5]	43	40	38	36

[1] IH-1000GP3: In-House workflow using 1000GP3 reference panel
[2] MIS-1000GP3: Michigan Imputation Server workflow using 1000GP3 reference panel
[3] MIS-CAAPA: Michigan Imputation Server workflow using CAAPA reference panel
[4] SIS-1000GP3: Sanger Imputation Server workflow using 1000GP3 reference panel
[5] SIS-AGR: Sanger Imputation Server workflow using AGR reference panel

4.4.3.1.1. SAC cohort

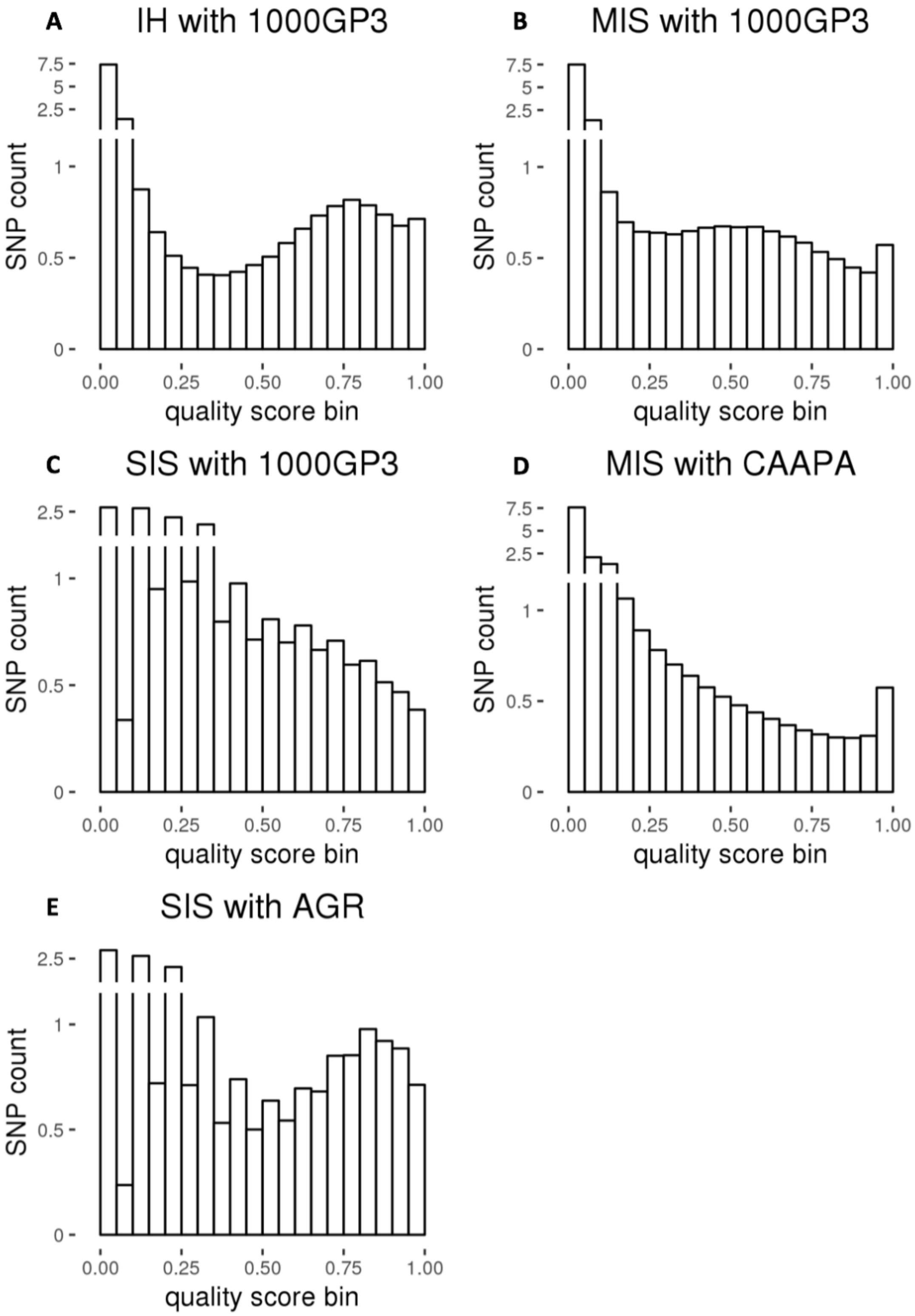


Figure 23: SNP Density plots for Chromosome 1 of the SAC cohort post imputation using the five workflows: (A)IH with 1000GP3, (B)MIS with 1000GP3, (C)SIS with 1000GP3, (D)MIS with CAAPA, and (E)SIS with AGR

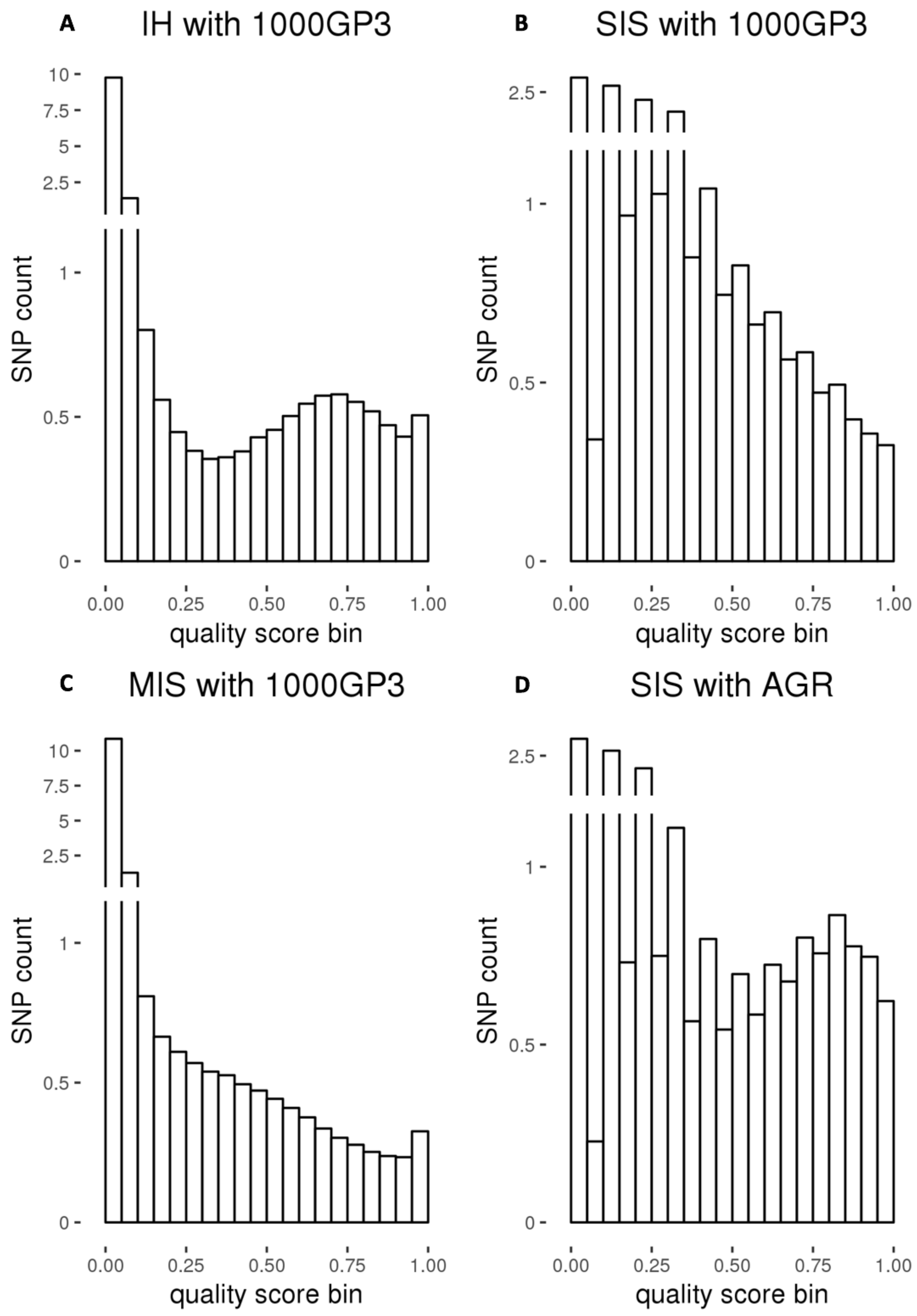


Figure 24: SNP Density plots for Chromosome X of the SAC cohort post imputation using four workflows: (A)IH with 1000GP3, (B)SIS with 1000GP3, (C)MIS with 1000GP3, (D)SIS with AGR. The MIS CAAPA workflow does not facilitate imputation of the X chromosome.

4.4.3.1.2. Ghanaian cohort

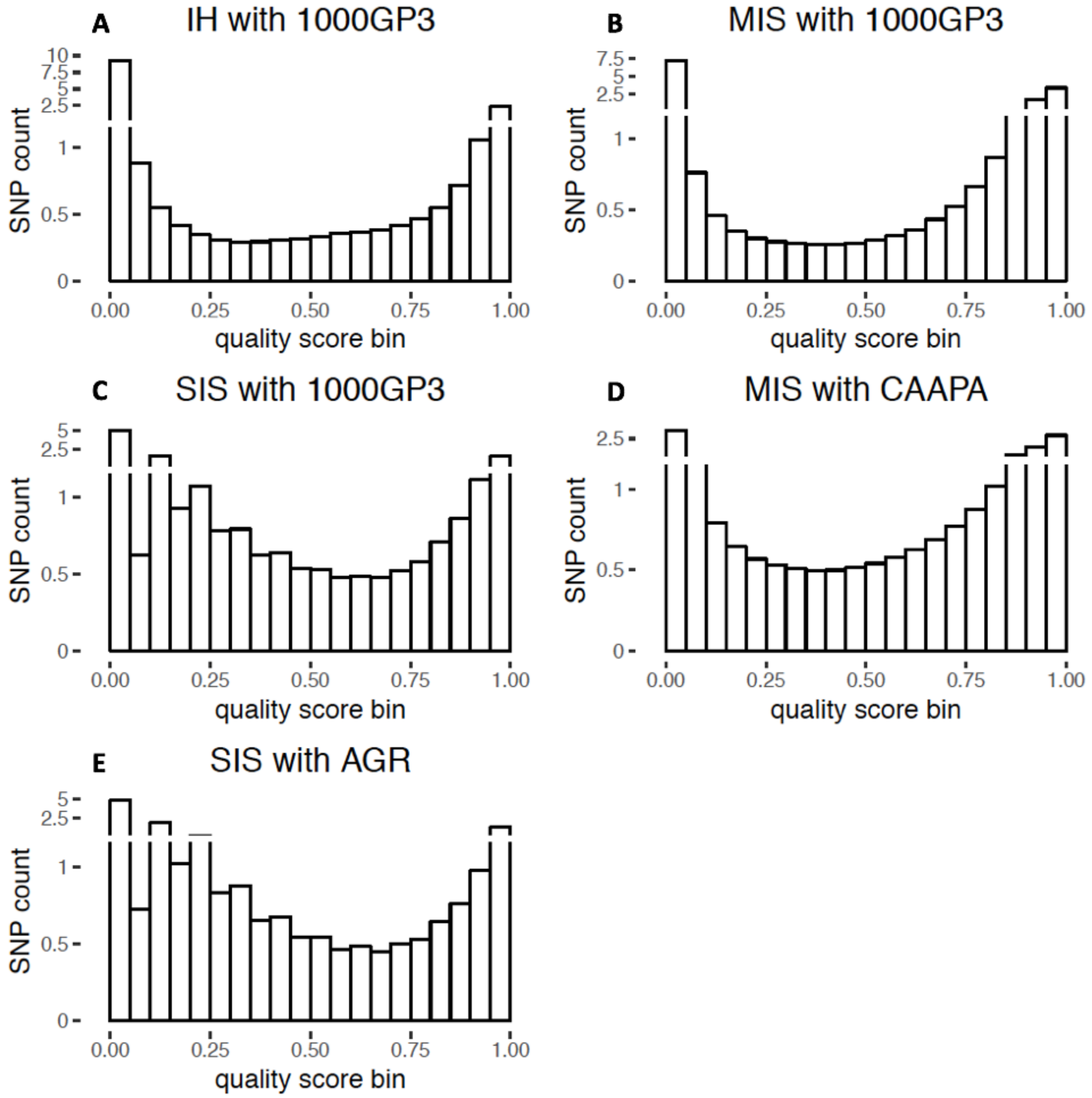


Figure 25: SNP Density plots for Chromosome 1 of the Ghanaian cohort post imputation using the five workflows: (A)IH with 1000GP3, (B)MIS with 1000GP3, (C)SIS with 1000GP3, (D)MIS with CAAPA, (E)SIS with AGR.

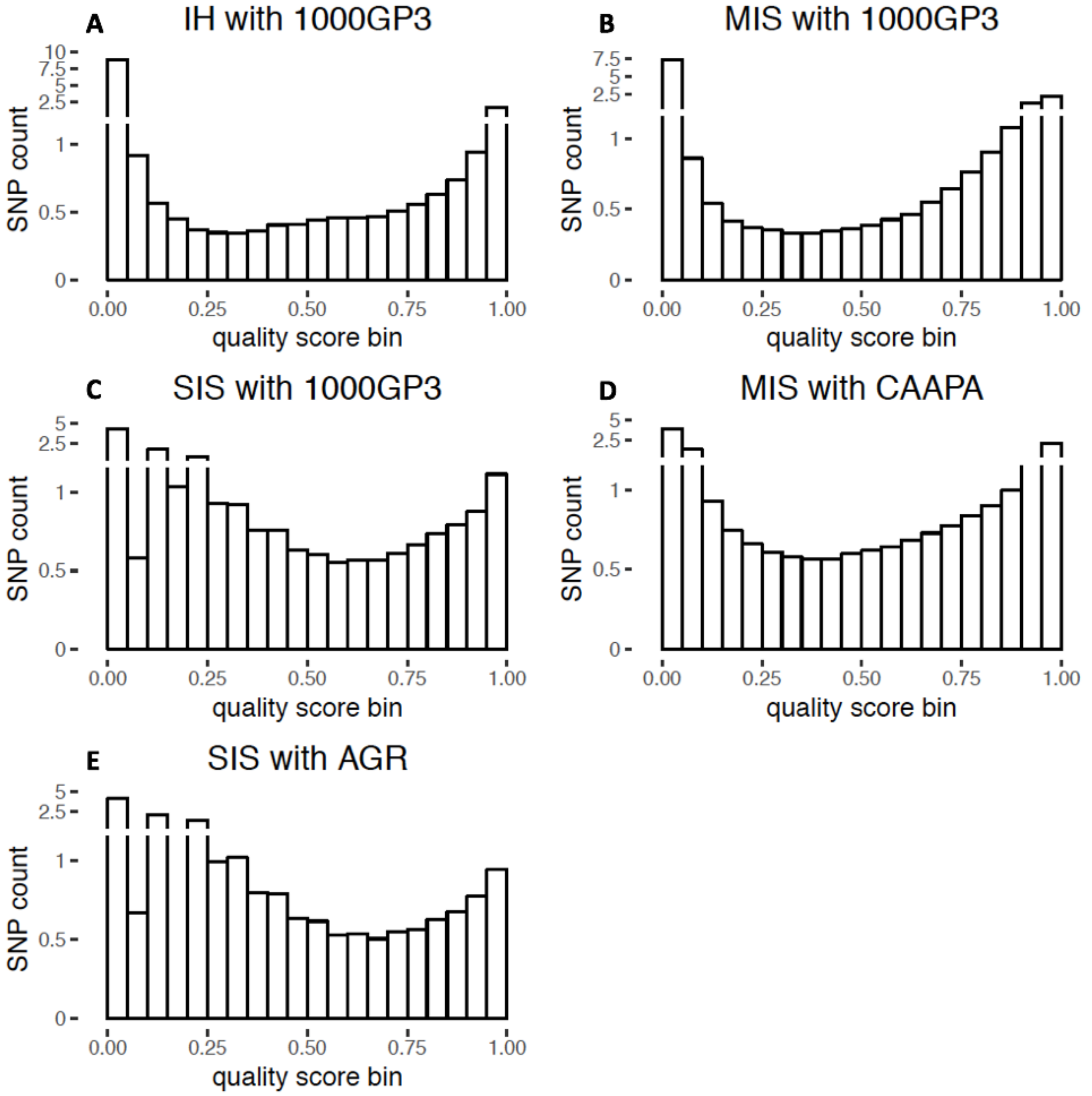


Figure 26: SNP density plots for Chromosome 22 of the Ghanaian cohort post imputation using the five workflows: (A)IH with 1000GP3, (B)MIS with 1000GP3, (C)SIS with 1000GP3, (D)MIS with CAAPA, (E)SIS with AGR.

4.4.3.2. Comparison of quality scores across MAF bins

When plotting the median quality scores obtained per MAF bin for each of the five tools, the SIS workflow using the AGR resource was confirmed to have produced the highest quality of imputed data for the SAC cohort whereas the 1000GP3 reference panel using any of the workflows produced the highest quality for the Ghanaian cohort.

4.4.3.2.1. SAC cohort

For both chromosomes 1 and X, imputation using either the 1000GP3 or the CAAPA resource with the MIS performed the worst for the SAC dataset (Figure 27 **A** and **B**) with the maximum median quality score only reaching 0.82 at a MAF of 50%. In comparison, the SIS-AGR workflow outperformed all other workflows, and the result correlated with the AGR imputing the highest SNP density for chromosome 1 (Figure 23 **E**) and chromosome X (Figure 24 **E**).

4.4.3.2.2. Ghanaian cohort

For the Ghanaian cohort, despite the CAAPA resource imputing the greatest SNP density (Figure 26 **D**), the IH workflow using the 1000GP3 reference panel imputed the highest quality of SNPs per MAF bin but was very closely followed by the other workflows and reference panels from the 20-30% MAF bin upwards (Figure 28 **A** and **B**).

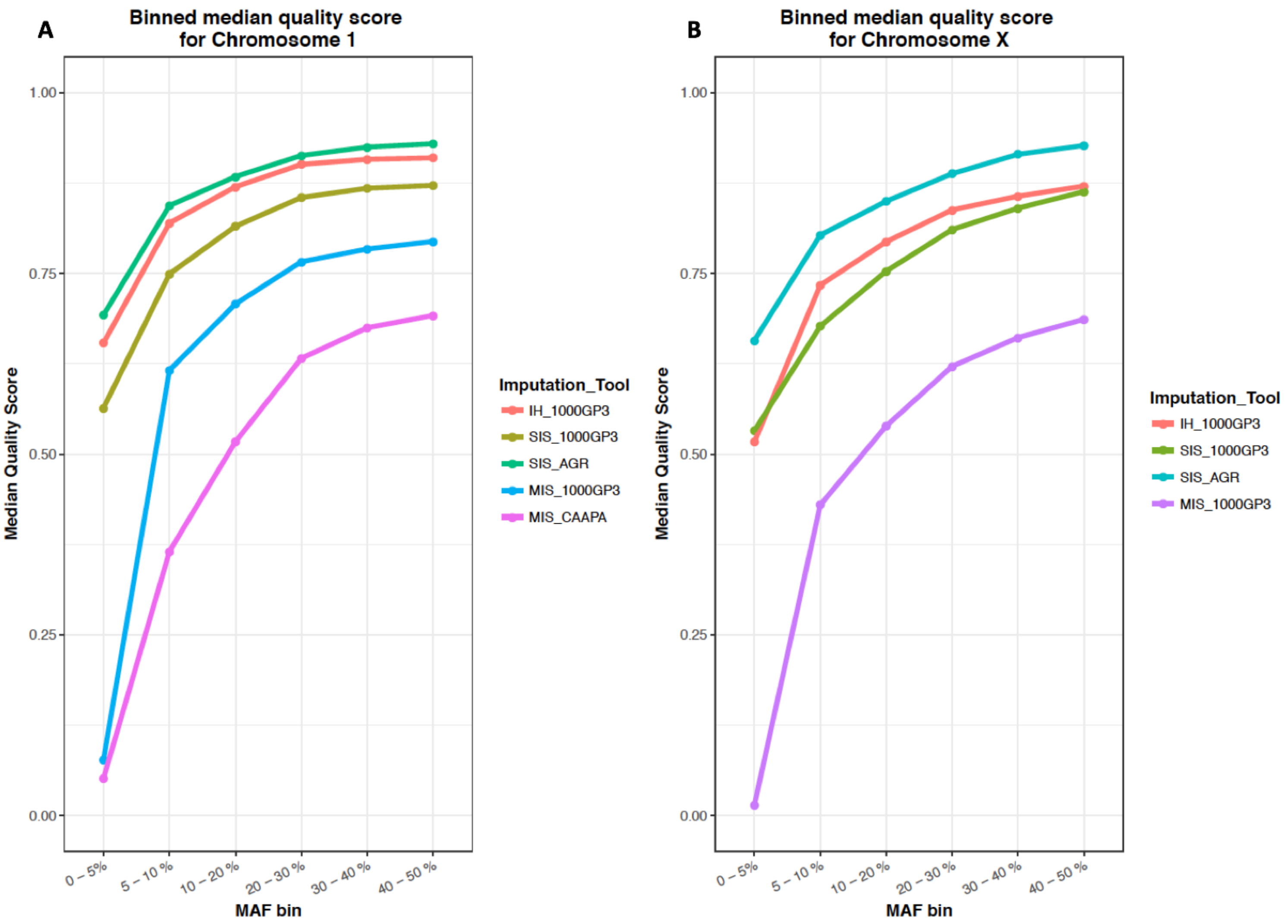


Figure 27: Median quality scores across MAF bins for Chromosome 1 and X for the SAC cohort, using the five workflows for chromosome 1 and four workflows for the X chromosome

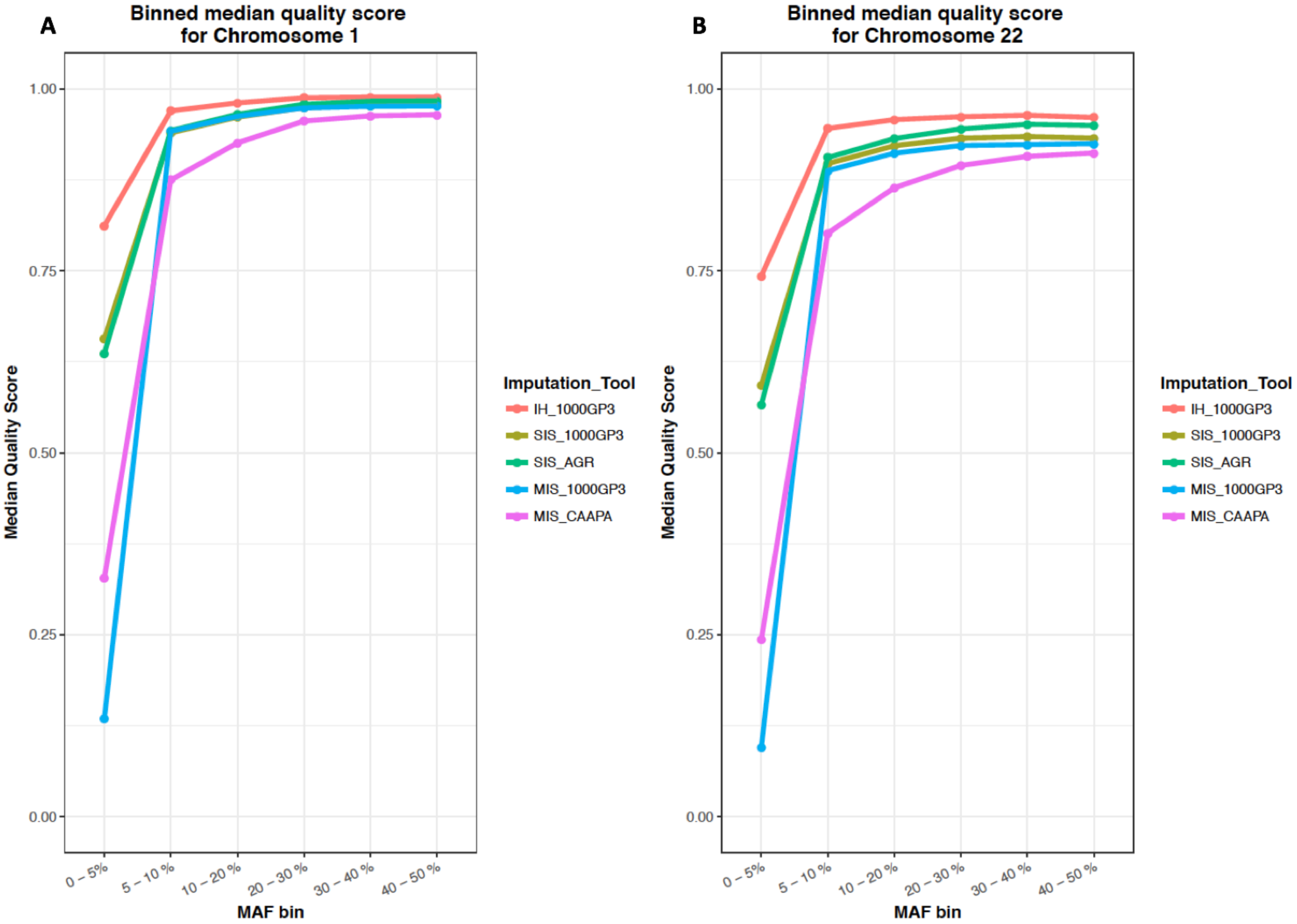


Figure 28: Median quality scores across MAF bins for Chromosome 1 and 22 for the Ghanaian cohort, using the five workflows

4.5. Perform an association analysis using high-quality imputed host genotype data and *M. tb* superclade data

4.5.1. Preparation of genotypes: Post-imputation QC

4.5.1.1. SAC cohort

The data obtained from the SIS, imputed with the AGR reference panel was selected as the dataset with the highest imputation quality across all those assessed. After removal of monomorphic sites and filtering on an INFO score of 0.45, 28 566 283 SNPs for 919 samples remained. After removing the 136 related individuals identified prior to imputation, and filtering for SNP- and sample missingness, and MAF, a dataset of 7 145 406 variants for 783 participants remained. Of the 525 clade-matched participants, 445 were extracted from the dataset of samples passing QC and used in the association analysis.

4.5.1.2. Ghanaian cohort

The In-house dataset, imputed with the 1000GP3 reference panel was selected as the best dataset. After filtering out monomorphic SNPs, INDELS, and variants not reaching the INFO score cut-off, 25 968 622 SNPs remained for 3 239 samples. After filtering for MAF, SNP- and sample missingness, and removal of 93 related individuals identified pre-imputation, the dataset comprised of 5 275 890 variants for 1 273 clade-matched samples. One sample was removed due to there not being covariable data for that sample.

4.5.2. Preparation of covariables file

4.5.2.1. SAC cohort

For the SAC cohort, the covariables age, “sex”, and ethnicity in the form of ancestry proportions were available for all samples. Available ancestry proportions were for the European, African, San, South-Asian, and East-Asian ancestries. Of the 445 samples which were clade-matched and passed QC after imputation, ancestry proportions were available for 357 samples as generated previously using the Affymetrix genotype data for this cohort and ADMIXTURE software. For the remaining 88 samples, ancestry proportions were obtained from a run of ADMIXTURE using genotype data from the MEGA array.

The East-Asian ancestry being the smallest contributing ancestry proportion was not included as a covariable in the analysis. Variances were calculated for each of the four remaining ancestry proportions and determined to be 0.027 (San), 0.035 (African), 0.014 (European), and 0.009 (South Asian). As the variances were greater than the minimum cut-off (determined through trial and error) of 0.001, they were included as covariables in the analysis. No "missing" proportion was available for this cohort, and thus a column denoting this data as "NA" was added to meet the specifications of SNPTEST.

4.5.2.2. Ghanaian cohort

For the Ghanaian cohort, age, "sex", and ethnicity in the form of principal components was available as covariables. One sample passing QC filters did not have one of the covariables and was thus excluded from the dataset leaving 1 272 samples for the association analysis. The variance in the PCs provided for the cohort was calculated to be 0.0002 (PC1), 0.0003 (PC2), and 0.0004 (PC3) and thus determined to be insufficient for inclusion in the analysis as covariables.

4.5.3. Multinomial logistic regression

4.5.3.1. SAC cohort

A multinomial logistic regression was run on the 445 clade-matched samples using SNPTEST under an additive model. All results were reported using the LAM superclade as the baseline. A total of 4 631 SNPs had an LRT p-value less than 0.0005 and eleven SNPs had an LRT p-value less than 1 x 10^{-6}. A single SNP, rs9389610, located on chromosome 6, had a p-value of 1.60 x 10^{-7}. Odds ratios are reported in Table 11 and standard errors of the odds ratios are shown in Figure 29.

Individuals with the A allele of this SNP were twice as likely to be infected with a member of the BeijingCAS1 superclade, than to be infected with either the HaarlemsLCC or LAM superclades. Results from the logistic regression at this SNP also reported that individuals with the A allele were only slightly more at risk of being infected with a member of the 'Other' superclade, when compared to the BeijingCAS1 superclade, and were very unlikely to be infected with a member of the Quebec superclade.

For the four SNPs located on chromosome 5, individuals with the risk allele were two to three times more at risk of being infected with the HaarlemsLCC or BeijingCAS1 superclade as compared to the reference LAM superclade. The risk allele also doubled the chances of being infected with the Quebec superclade while halving the risk of being infected with a member of the “Other” superclade (Table 11).

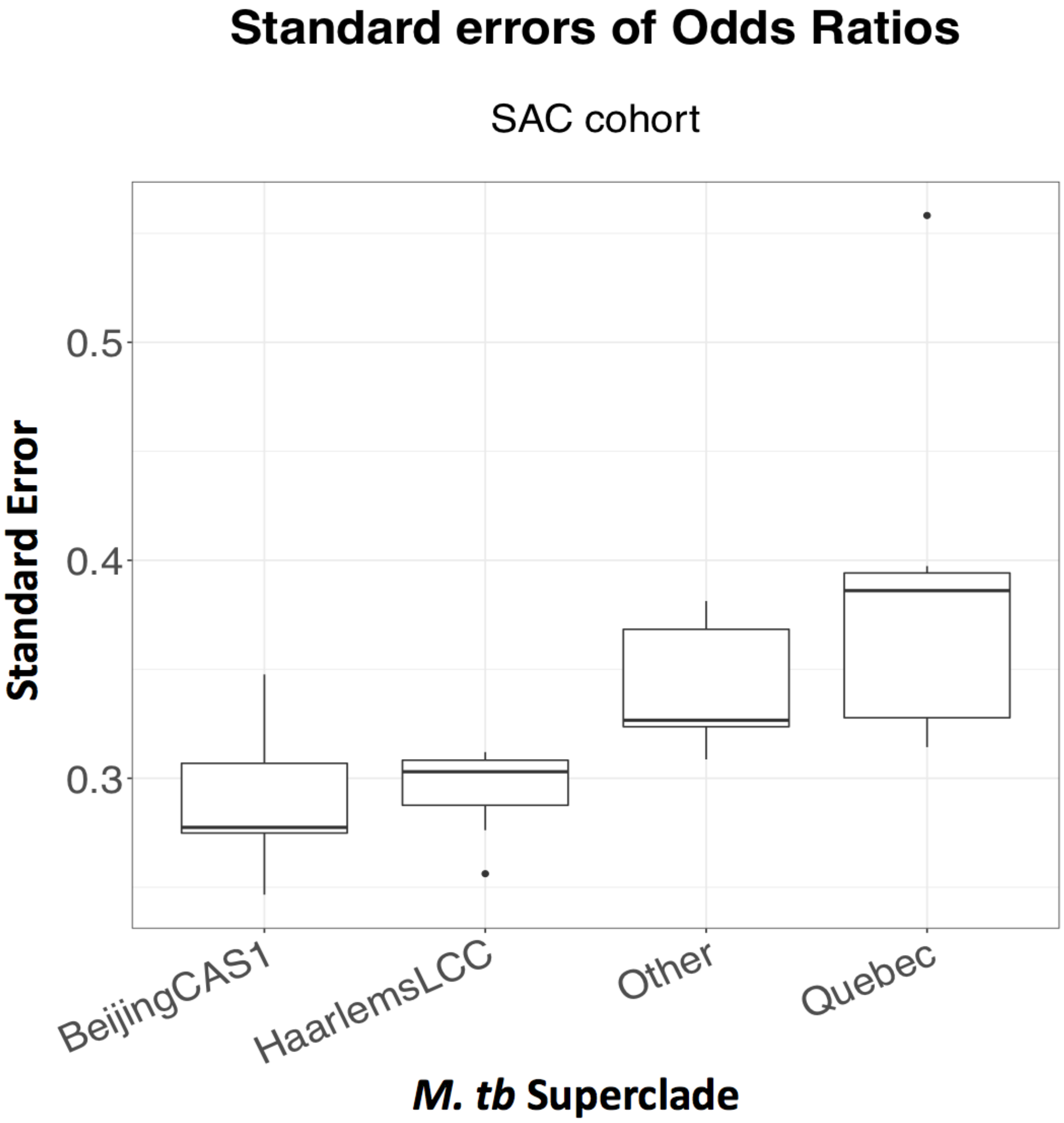


Figure 29: Standard errors of odds ratios calculated for each superclade against the reference LAM superclade for the SAC cohort.

For the six SNPs located on chromosome 17, the risk allele was shown to double the risk of being infected with a member of the LAM superclade than with the HaarlemsLCC superclade. Individuals with the risk allele of these six SNPs were also equally at risk of being infected with a member of the BeijingCAS1 or LAM superclade and were twice as likely to be infected with the Other superclade when compared to the BeijingCAS1 superclade (Table 11).

Table 11: Top 11 SNPs identified by MLR to be associated with different M. tb superclades in the SAC cohort

Chr	SNP ID	Reference allele	Risk allele	OR LAM (Reference)	BeijingCAS1 OR(95%CI)	HaarlemsLCC OR(95%CI)	Other OR(95%CI)	Quebec OR(95%CI)	LRT_p_value
5	rs17458866	C	T	1	0.34 (0.19-0.61)	0.44 (0.26-0.76)	0.46 (0.23-0.95)	1.99 (1.07-3.68)	10e-07
5	rs13355101	G	A	1	0.31 (0.17-0.57)	0.46 (0.27-0.79)	0.47 (0.23-0.96)	2 (1.08-3.7)	6.43e-07
5	rs12518239	C	A	1	0.29 (0.15-0.56)	0.39 (0.22-0.7)	0.51 (0.25-1.05)	1.91 (1.01-3.62)	9.41e-07
5	rs28769614	C	T	1	0.27 (0.13-0.53)	0.37 (0.2-0.68)	0.48 (0.23-1.01)	1.92 (1-3.66)	3.03e-07
6	rs9389610	G	A	1	2.19 (1.35-3.55)	1.07 (0.64-1.76)	2.78 (1.52-5.08)	0.25 (0.08-0.73)	1.60e-07
17	rs78022196	G	A	1	1.04 (0.6-1.8)	0.59 (0.32-1.09)	1.96 (1.03-3.73)	5.31 (2.44-11.57)	5.13e-07
17	rs72843143	C	T	1	0.98 (0.57-1.69)	0.57 (0.31-1.04)	1.89 (1-3.58)	4.81 (2.26-10.25)	8.18e-07
17	rs8071332	A	G	1	0.94 (0.55-1.6)	0.59 (0.33-1.06)	2.03 (1.08-3.81)	4.77 (2.22-10.28)	6.54e-07
17	rs10438776	T	C	1	0.99 (0.58-1.71)	0.55 (0.3-1.01)	1.94 (1.02-3.69)	4.99 (2.3-10.85)	3.93e-07
17	rs17682747	G	A	1	0.98 (0.57-1.68)	0.56 (0.31-1.03)	1.9 (1.01-3.58)	4.85 (2.27-10.33)	5.93e-07
17	rs7208461	T	C	1	0.94 (0.55-1.61)	0.54 (0.3-0.98)	1.77 (0.94-3.34)	4.68 (2.17-10.1)	8.64e-07

4.5.3.2. Ghanaian cohort

A multinomial logistic regression was run on the 1 272 clade-matched samples using SNPTEST under an additive model. All results were reported using the LAM_CAM superclade as the baseline. A total of 32 SNPs had an LRT p-value less than 1 x 10^{-6} (Table 12). Standard errors for the odds ratios are shown in Figure 30.

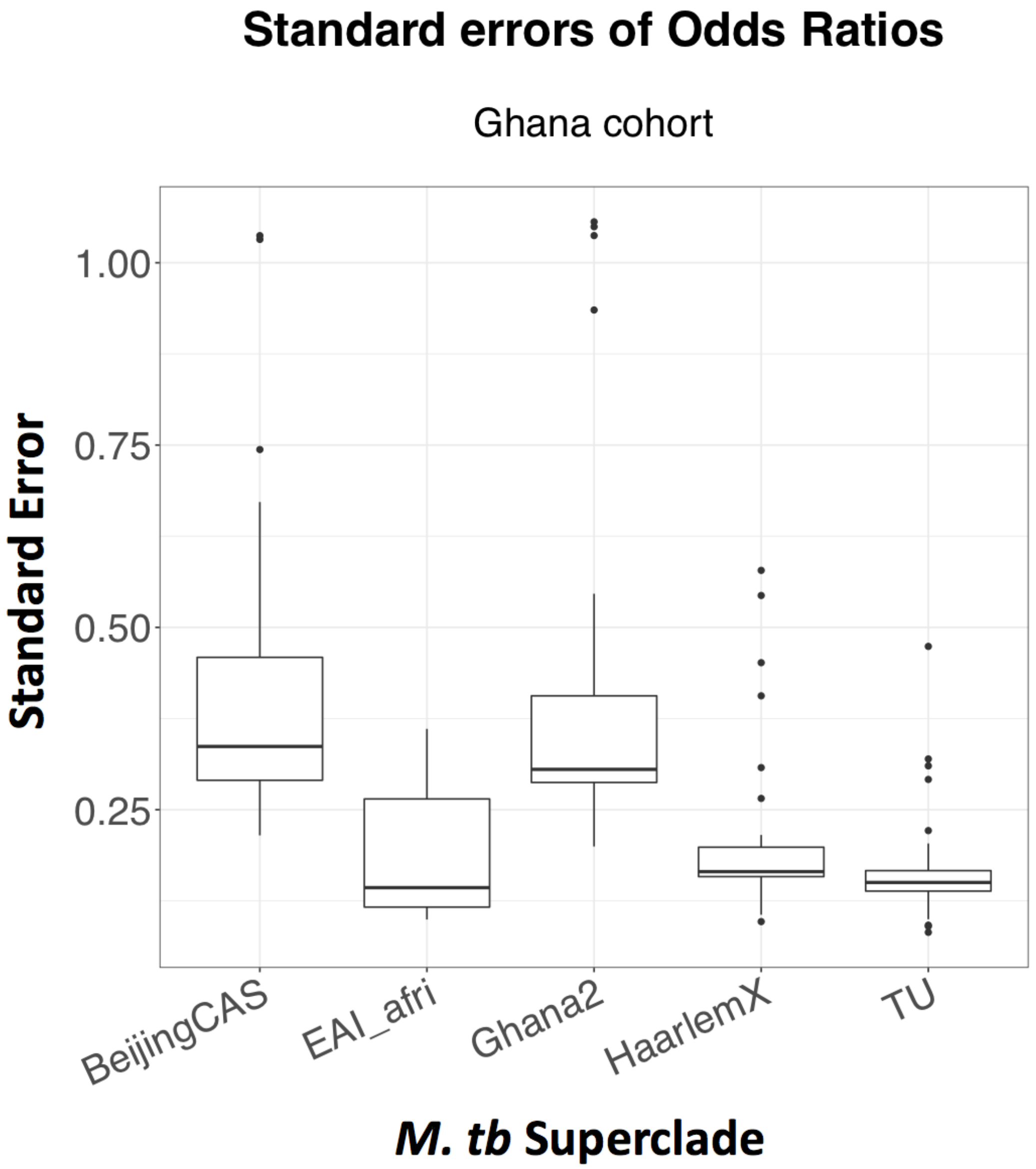


Figure 30: Standard errors of odds ratios calculated for each superclade against the reference LAM superclade for the Ghanaian cohort

Table 12: Top 32 SNPs identified by MLR to be associated with different M. tb superclades in the Ghanaian cohort

Chr	SNP ID	Reference	Risk	OR_LAMCAM	BeijingCAS	EAI_afri	Ghana2	OR_HaarlemX	OR_TU	LRT
		allele	allele	(Reference)	OR	OR	OR	OR	OR	p-value
					(95%CI)	(95%CI)	(95%CI)	(95%CI)	(95%CI)	
6	rs529920	A	G	1	0.4 (0.22-0.7)	0.69 (0.57-0.84)	1.04 (0.6-1.8)	1.1 (0.84-1.44)	1.24 (0.98-1.56)	1.86e-07
12	rs73418916	A	G	1	0.3 (0.11-0.81)	34.15 (20-58.33)	1.38 (0.64-2.96)	0.76 (0.51-1.12)	0.88 (0.63-1.22)	2.31e-97
12	rs138396290	T	C	1	0.32 (0.13-0.77)	4.59 (3.63-5.82)	1.31 (0.73-2.37)	0.87 (0.63-1.19)	0.91 (0.7-1.19)	3.32e-62
12	rs75717431	T	C	1	2.46 (1.4-4.35)	0.98 (0.78-1.24)	3.3 (1.88-5.82)	0.95 (0.69-1.31)	0.59 (0.44-0.8)	8.55e-09
12	rs77428482	G	A	1	2.56 (1.45-4.52)	1.01 (0.81-1.28)	3.2 (1.81-5.64)	0.95 (0.69-1.32)	0.6 (0.45-0.81)	1.93e-08
12	rs77562721	G	A	1	2.5 (1.42-4.41)	1.04 (0.82-1.3)	3.12 (1.77-5.5)	0.93 (0.68-1.29)	0.6 (0.45-0.81)	2.53e-08
12	rs41524146	C	G	1	2.38 (1.36-4.18)	1 (0.8-1.26)	3.19 (1.82-5.58)	0.97 (0.71-1.34)	0.61 (0.45-0.81)	2.62e-08
12	rs7299395	G	A	1	2.16 (1.22-3.82)	0.99 (0.8-1.24)	3.12 (1.77-5.52)	0.93 (0.69-1.27)	0.64 (0.48-0.84)	2.50e-07
12	rs74550821	G	A	1	2.59 (1.47-4.57)	1.05 (0.84-1.32)	3.2 (1.82-5.63)	1.01 (0.74-1.39)	0.63 (0.47-0.84)	2.89e-08
12	rs144335343	C	T	1	2.33	1.08	3.63	1	0.69	9.15e-08

					(1.32-4.11)	(0.87-1.35)	(2.05-6.41)	(0.73-1.36)	(0.52-0.92)	
12	rs6582329	A	T	1	2.58	1.07	3.42	1.04	0.67	6.00e-08
					(1.46-4.57)	(0.85-1.34)	(1.94-6.02)	(0.76-1.43)	(0.5-0.9)	
12	rs12296167	T	G	1	2.42	2.03	3.33	1.03	1.41	2.02e-09
					(1.25-4.7)	(1.6-2.59)	(1.71-6.48)	(0.74-1.42)	(1.07-1.85)	
12	rs544003050	A	G	1	0.5	4.72	1.42	0.75	0.74	4.17e-40
					(0.24-1.05)	(3.47-6.43)	(0.68-2.98)	(0.52-1.07)	(0.54-1)	
12	rs58262822	C	G	1	0.58	8.5	0.75	0.81	0.93	1.96e-124
					(0.34-0.99)	(6.29-11.48)	(0.47-1.2)	(0.65-1.01)	(0.78-1.11)	
12	rs11108508	T	C	1	1.09	6.61	1.33	0.71	1.11	1.04e-112
					(0.56-2.09)	(5.26-8.32)	(0.73-2.44)	(0.48-1.05)	(0.85-1.46)	
12	rs41472447	A	G	1	2.56	0.97	2.94	0.92	0.59	5.70e-09
					(1.48-4.41)	(0.78-1.21)	(1.71-5.08)	(0.68-1.26)	(0.44-0.79)	
13	rs549053537	A	T	1	0.63	5.51	1.23	0.86	0.91	8.12e-79
					(0.41-0.96)	(4.12-7.39)	(0.83-1.82)	(0.71-1.04)	(0.78-1.07)	
13	rs73497904	C	G	1	1.34	65.39	1.2	0.96	1.23	2.23e-244
					(0.69-2.59)	(40.67-105.13)	(0.6-2.39)	(0.67-1.39)	(0.92-1.65)	
13	rs9524738	G	C	1	1.88	0.31	1.32	1.34	1.16	2.67e-52
					(0.86-4.14)	(0.26-0.38)	(0.68-2.55)	(0.97-1.84)	(0.9-1.5)	
15	rs551641937	G	A	1	0.52	275.89	0.52	0.39	1.11	3.61e-236
					(0.07-3.97)	(152.64-498.69)	(0.07-3.98)	(0.14-1.14)	(0.59-2.08)	

15	rs35799802	C	T	1	0.76 (0.36-1.62)	3.08 (2.29-4.13)	0.89 (0.42-1.88)	1.09 (0.75-1.59)	1.19 (0.87-1.64)	3.94e-15
15	rs55747528	C	T	1	0.53 (0.24-1.15)	28.58 (15.28-53.45)	1.47 (0.66-3.25)	0.72 (0.49-1.05)	0.79 (0.57-1.08)	1.49e-69
16	rs577800201	C	T	1	0.31 (0.07-1.33)	93.05 (54.17-159.85)	0.87 (0.32-2.36)	0.67 (0.4-1.13)	1.01 (0.68-1.5)	1.20e-167
16	rs187181146	C	T	1	0.64 (0.35-1.17)	8.65 (6.61-11.31)	0.94 (0.59-1.49)	0.83 (0.66-1.06)	0.92 (0.76-1.12)	9.77e-152
16	rs35868343	G	A	1	0.72 (0.46-1.11)	10.65 (7.14-15.87)	1.01 (0.66-1.55)	0.91 (0.74-1.12)	0.92 (0.77-1.1)	1.90e-93
17	rs143309838	G	A	1	0.57 (0.28-1.17)	9.38 (7.2-12.23)	0.93 (0.57-1.53)	0.84 (0.65-1.08)	0.99 (0.81-1.21)	2.08e-165
17	rs144224512	C	G	1	0.36 (0.05-2.69)	220.49 (118.29-411.01)	0.5 (0.08-3.13)	0.42 (0.17-1.01)	0.74 (0.4-1.36)	1.21e-215
17	rs374315920	C	T	1	3.52 (0.94-13.13)	548.96 (270.65-1113.45)	1.11 (0.14-8.79)	0.85 (0.27-2.63)	0.93 (0.37-2.36)	1.68e-255
17	rs77139740	A	G	1	0.34 (0.12-1.01)	44.94 (26.85-75.2)	0.97 (0.43-2.22)	0.73 (0.48-1.12)	0.89 (0.63-1.26)	3.74e-120

17	rs77641928	G	A	1	1.24 (0.37-4.17)	57.29 (33.69-97.42)	0.46 (0.06-3.62)	0.67 (0.3-1.49)	0.94 (0.53-1.67)	4.98e-282
22	rs553728019	A	C	1	0.51 (0.23-1.09)	23.66 (12.1-46.25)	1.05 (0.47-2.33)	0.76 (0.52-1.11)	0.79 (0.57-1.09)	2.83e-55
22	rs60153275	C	T	1	0.93 (0.32-2.71)	61.01 (36.66-101.54)	0.93 (0.32-2.71)	0.7 (0.38-1.28)	0.98 (0.63-1.51)	7.19e-281

4.5.4. Gene annotations of selected SNPs

4.5.4.1. SAC cohort

The VEP tool was used to retrieve gene annotations for SNPs of interest. While these SNPs are unlikely to have a direct effect on the gene expression itself, the SNP may be in LD with other nearby SNPs which do have a direct effect on the gene. For the SAC cohort, the most significantly associated SNP was rs9389610 (g.139039029G>A), located on chromosome 6. This SNP is an imputed SNP and its two closest directly genotyped SNPs were rs4896385 (g.139011266G>T), and rs7742202 (g.139074280A>G). The rs4896385 SNP located in *NHSL-1*, while the rs7742202 SNP is located in *GVQW2*. Neither of these genes have been previously shown to be involved in the pathogenesis of TB. The four SNPs located on chromosome 5 (Table 11) were annotated to the StAR Related Lipid Transfer Domain Containing 4 gene *(STARD4)* using the VEP tool, while the six SNPs located on chromosome 17 were annotated to the *TANC2* gene.

4.5.4.2. Ghanaian cohort

Given the prevalence of TB caused by *M. africanum* strains in West-African countries, including Ghana, the most significant results for this cohort were that of several SNPs located on chromosomes 15, 16, and 17. The risk allele of one SNP located on chromosome 15, rs551641937 (g.62385889G>A), was determined by the MLR to increase the risk of TB caused by the EAI/AFRI superclade by 276 times, when compared to the LAMCAM reference superclade. The VEP tool, however, did not have any gene annotation listed for this SNP in the database.

The risk allele of the SNP rs577800201 (g.20476046C>T) was shown to increase the risk of being infected with the EAI/AFRI superclade by 93 times, compared to the LAMCAM superclade, and was annotated by the VEP to map to *ACSM2A*. Lastly, the risk allele of SNP rs374315920 (g.38496435C>T) located on chromosome 17, was found to increase an individual's risk of being infected with the EAI/AFRI superclade by more than 500 times, as compared to the LAMCAM reference superclade, and the MLR specific for this SNP was highly significant with an LRT p-value of 1.68x10^{-255}. This SNP was also annotated by the VEP to lie within the retinoic acid receptor alpha (RARA) gene.

CHAPTER FIVE

5. DISCUSSION

5.1. Overview

TB is a highly infectious disease affecting millions of people each year. The genetic susceptibility of the host to develop the disease has been extensively studied using a number of study designs, including linkage analysis, candidate gene, and GWAS. Furthermore, a number of selected genes have been investigated for their contribution to genetic susceptibility of the host to different strains of *M. tb*. To date, no workflow has been established to perform a genome-wide scan of genetic markers affecting susceptibility to different *M. tb* strains.

The aim of the present study was to use one cohort to develop a bioinformatics workflow which facilitates a test for association of genome-wide SNP markers to multiple strains causing TB. Once established, a second cohort was used to independently verify the method developed. The study included host genotype and pathogen isolate data from two ethnically distinct cohorts and using several reference panels, imputation was performed to obtain a significantly larger number of genotypes than previously available in the raw genotype dataset.

After imputation of the study dataset, imputation quality was assessed using marker density, and quality score in MAF categories as indicators. Following the selection of the study dataset that had been imputed with the highest accuracy, an association analysis between the host genotypes and the infecting *M. tb* superclade was performed using the MLR functionality available within SNPTEST.

The MLR functionality within SNPTEST enabled the genome-wide investigation of genetic markers for association to a number of *M. tb* superclades which were defined by clustering of clades using a SNP-based phylogenetic tree. Results showed that none of the genotype-*M. tb* relationships passed the GWAS p-value cut-off of 5×10^{-8} for the SAC cohort. For the Ghanaian cohort however, 32 SNPs passed the GWAS cut-off and may be considered as potential targets for further investigation of HDTs suitable for individuals of West-African ancestry.

5.2. Review of the method

5.2.1. Paired sample collection and SNP genotyping

The foundation of the association analysis method developed in this study lies in the significance of paired sample collection. As with a standard case-control GWAS in which genetic associations to a disease are evaluated, whole blood was necessary for the extraction of sufficient volumes of DNA from which genotypes of common SNPs could be derived. In addition to whole blood, the collection of sputum from the same participants provided the necessary bacterial samples for the determination of the infecting *M. tb* isolate. The blood and sputum samples were collected as part of separate studies, and data were archived for future analyses. In the present study, two independently collected, yet biologically related, datasets were brought together in a novel way to explore the relationship between the affected human host and the pathogenic bacterium, for two geographically distinct cohorts.

For the SAC cohort, we had access to an extensive database of *M. tb* isolates. Of the 527 genotyped participants with *M. tb* isolate information, 68 individuals had records for multiple infections. While it is possible for participants to have been sampled multiple times throughout the recruitment period, we acknowledge that a number of confounding factors may have affected the correct sequential representation of an *M. tb* infection: participants may have gone to different clinics during the sample collection period, or may simply have failed to return to the clinic. Thus, subsequent *M. tb* infections may have been incorrectly represented with regards to the true order of events or may not have been recorded at all. In light of this, the decision to only analyse the ‘first recorded infection’ available for each patient in the cohort was taken, and may have minimised the confounding effect of these uncontrollable events.

5.2.2. Imputation performance

The more admixed a population is, the greater the heterogeneity in their haplotype block structure. This genetic complexity requires large reference panels with suitable ancestry to facilitate accurate genotype imputation. While several reference panels exist to facilitate genotype imputation, most of these panels focussed on representing populations of European ancestry, and little representation was made for African populations. Therefore, the present study focussed on evaluating the quality of imputation attainable for the five-way admixed SAC population and the Ghanaian cohort using the 1000GP3, AGR, and CAAPA reference panels.

The five-way admixed SAC population contains genetic contributions from Bantu-speaking Africans, Europeans, KhoeSan, and South- and East-Asians (Daya et al. 2013; De Wit et al. 2010). While imputation has previously been performed on this population, it was done so using the 1000 Genomes Phase 1 (The 1000 Genomes Project Consortium 2012) and the HapMap3 release 2 (The International HapMap 3 Consortium 2010) reference panels, the latter has since been deprecated and in addition, represented mostly individuals of European ancestry (Chimusa et al. 2014). To assess recently released reference panels which promised to be of greater suitability for the admixed African ancestry of the SAC, we tested the AGR, and the CAAPA reference panels, as well as the most recent release of the 1000 Genomes reference panel for imputation quality.

Haplotype phasing followed by genotype imputation was carried out using five methods which were compared for their imputation quality prior to association analysis (Table 5). Regarding ease of use, the Michigan Imputation Server was more user-friendly than the Sanger Imputation Server, as the latter required uploading of data via Globus[1] and numerous authentications via email. However, given that the AGR is not yet publicly available, the SIS is a valuable tool to consider for genotype imputation using this reference panel.

The 1000GP3 Version 5 (Sudmant et al. 2015) is publicly accessible, consists of 2 504 samples and covers 81 027 987 autosomal markers as well as 3 209 655 markers for the X chromosome. Spanning 26 populations across the world, the 1000GP3 offers one of the most diverse reference panels to have been compiled to date by including samples sourced from African, American, European, South- and East-Asian countries. Continental African populations contributing to the reference panel include individuals from the Esan ethnic group in Nigeria, Luhya in Kenya, the Yoruban people, as well as participants from The Gambia. Additionally, African-American participants were recruited in the United States and Barbados. The 1000GP3 panel was used for genotype imputation in the IH, MIS, and SIS methods used in this study.

The AGR is comprised of 4 956 samples and covers 93 421 145 sites across the autosomes and 3 583 058 sites on the X chromosome. Approximately half of the resource is comprised of samples from the 1000GP3, while around 2 000 samples were sourced from regions in the East-African country of Uganda. Around 100 samples were sourced from several regions in

[1] https://www.globus.org/globus-connect-server

Ethiopia, as well as from Egypt, the Zulu people in South Africa, and the Nama/Khoesan people in Namibia. With twice the number of samples, and the addition of a tremendous amount of new genetic material of African ancestry, the AGR promised to be the best resource for imputing untyped genetic markers in both the SAC and Ghanaian cohorts, when compared to the 1000GP3 reference panel.

The CAAPA resource (Mathias et al. 2016) consists of 883 individuals, amounting to approximately one third of the samples on the 1000GP3 and just over a fifth of the number of samples on the AGR reference panel. The resource includes individuals self-reporting as having African ancestry and were recruited from nine cities in the United States, four populations in the Caribbean, four in Central- and South America, and two populations representing West Africa. In contrast to the 1000GP3 and the AGR reference panels which facilitated imputation of the X chromosome, imputation of the X chromosome with the CAAPA resource via the Michigan Imputation Server was not possible.

Three of the five workflows performed imputation of the 1000GP3 reference panel, while two methods made use of the AGR, or the CAAPA reference panel. The differences between the three methods using the 1000GP3 was the imputation software used, as well as additional strict QC filters that were imposed on the study dataset by the MIS and SIS methods. For all five workflows, SHAPEIT2 was chosen for haplotype phasing to maintain consistency in the haplotype phasing software. Although some studies have reported that pre-phasing reduces imputation accuracy (Roshyara et al. 2016), it is known to significantly speed up the computationally intensive process of genotype imputation (Kanterakis et al. 2015; B. Howie et al. 2012).

For the IH method, imputation was performed using the IMPUTE2 software, while the MIS uses minimac3, and the SIS uses the PBWT algorithm. IMPUTE2 reports an information (INFO) metric while the minimac software and PBWT algorithm produce an Rsq metric. Both forms of quality metrices range from 0 to 1 with higher values indicating better imputation qualities.

Our results show that the SAC dataset was imputed with the highest quality using the AGR and this outcome was only a slight improvement on the imputation of the 1000GP3 using the IH method (Figure 27). In contrast to the results obtained using the AGR, imputation of either the

1000GP3 or the CAAPA panels performed by the MIS had the lowest imputation quality scores for the SAC.

Where the SIS tool with AGR reached a median quality score of 0.93 for SNPs with an MAF ranging between 40 and 50% (Figure 27 **A**), the MIS only managed to impute the CAAPA and 1000GP3 to a median quality score of 0.69 and 0.79, respectively. Although the cut-off for imputation quality ranges from a score of 0.3 to 0.5, the trend of having lower quality scores across the MAF bins further emphasises the selection of the dataset imputed with the highest quality, which was the dataset imputed with the AGR. The MIS also performed significantly worse than the other tools when imputing rare SNPs with an MAF below 5%, as the median quality score obtained did not even reach 0.1 which is far lower than the chosen cut-off of 0.45. This was likely in part due to the CAAPA reference panel being the most inappropriate resource for imputation of the admixed SAC as the individuals contained on the panel are of African ancestry but reside in cities in the United States of America (Mathias et al. 2016), and thus may have recent admixture from American ancestors. Unlike the CAAPA resource, both the 1000GP3 and the AGR contain haplotypes of individuals from continental African countries and were therefore better representatives for the ancestry proportions present in the SAC population, making either of these two reference panels better than the CAAPA resource for imputation of the SAC cohort.

The difference in imputation quality between the IH-1000GP3 and MIS-1000GP3 methods may have also been as a result of the imputation tool used, as SHAPEIT was used in both methods for haplotype phasing, and the same version of the 1000GP3 reference panel was used. Thus, it is possible that the imputation tool, which differed in the two methods, influenced the imputation quality achieved, although a recent simulation study has reported similar accuracies for imputation of common variants using a range of reference panels and comparing the performance of minimac3, IMPUTE2 and Beagle 4.1 (Das et al. 2016). A comparison of the three most commonly used imputation tools, minimac, BEAGLE and IMPUTE2, found that minimac and IMPUTE2 outperformed usage with BEAGLE (Liu et al. 2015). Another study found that minimac3, which is based on the same mathematical formula as minimac, performed only slightly better than IMPUTE2 (Das et al. 2016).

When evaluating the Ghanaian cohort, we found that although imputation of the 1000GP3 reference panel with the IH method performed the best, there was very little difference in the median quality scores for the different workflows seen for SNPs with an MAF of 10-50% (Figure 28). For rare variants (MAF 0-5%) however, the IH method outperformed all others with a median quality score above 0.75, whereas both analyses with the MIS produced a median score below the cut-off of 0.45. Thus, for the Ghanaian cohort, all reference panels and methods tested could be considered viable options for imputing common variants with an MAF of 10-50%, but should be considered carefully for variants with an MAF below 10%.

In contrast to the AGR which contains no individuals recruited from West-African countries, the CAAPA resource contains 88 individuals recruited from the West-African country of Nigeria (Mathias et al. 2016). We found that the CAAPA resource performed well when imputing SNPs with a MAF above 10% (). From an MAF of 20-50%, the CAAPA resource performed similarly to the other four resources and may thus be considered suitable for imputing cohorts of West-African ancestry, such as the Ghanaian cohort used in this study (Figure 28).

5.2.3. Multinomial Logistic Regression

To explore the relationship between the genetics of the host and the infecting bacterium, we used a logistic regression modelling approach. To make use of the MLR function in SNPTEST, it was necessary to define the reference superclade against which beta values for each other superclade would be calculated. For this method, the LAM superclade was selected as the reference due to it having an intermediate to high frequency in both cohorts (Figure 16 **A** and Figure 19 **A**).

5.2.4. SNP-based phylogenetic clustering of *M. tb* clades

To reduce the number of phenotype groups of low frequencies being tested in the multinomial logistic regression, the many *M. tb* strains were clustered into “clades” and furthermore into “superclades”. For a logistic regression, it is generally recommended that a minimum of ten samples be included per group (in this case, superclade) being tested for association. In this study, superclades which did not have the minimum of ten samples were excluded, and this exclusion was only necessary for the SAC cohort.

For the SAC cohort, eight clades were reduced to five superclades (Figure 8 and Figure 9), and for the Ghanaian cohort, 12 clades were grouped into six superclades (Figure 12 and Figure

13). For the SAC cohort, there was an unbalanced contribution of Beijing and CAS1 when grouping as BeijingCAS1 (Figure 8). Although the Haarlem and HaarlemLike frequencies were noticeably different, their biological similarity (Figure 2) may have compensated for this difference, and was subsequently grouped with LCC which occurred at a similar frequency to Haarlem (Figure 9). The LAM, Other, and Quebec clades were reclassified as superclades after clustering, with the latter two superclades having the lowest frequencies (n=60 and n=40, respectively) in the SAC cohort. Furthermore, prior to grouping into superclades, the clade distribution of the selected samples included in this study correlated with the distribution described in the most recent epidemiological study performed on the same population and thus demonstrated that the SAC study cohort included in the association analysis was a suitable representation of the strain distribution seen at the time of sample collection (van der Spuy et al. 2009).

For the Ghanaian cohort, only one of the 12 clades - the U clade - had a frequency of below 10 (Figure 12), and was clustered with the T clade (Figure 13) which had a significantly larger number of cases (n=235). A largely unbalanced frequency was also observed when clustering LAM and CAM cases, with the latter clade having a frequency 13 times more than LAM. This was also seen when clustering the two *M. africanum* clades, “afri181” and “afri438”, with EAI (Figure 13). The frequency of the Haarlem clade was twice as large as the X clade, while the Beijing and CAS clades had a similar frequency prior to grouping (Figure 12). This grouping strategy, and the frequencies of the contributing clades need to be carefully considered when interpreting the odds ratios obtained from the multinomial logistic regression, as they may inflate the effect predicted for the smaller contributing clade.

The genome-wide significance cut-off of 5 x 10^{-8} is the widely accepted threshold for GWAS studies and aims to prevent inflating the occurrence of type I errors on a genome-wide scale (Durbin 2014). Unless the sample size is large enough to obtain sufficient power for the phenotype being tested, this very low threshold reduces the probability of correctly identifying SNPs having a small effect size (Stringer et al. 2011). When assessing the standard errors of the odds ratios obtained for the MLR, we found that the smallest superclades in both cohorts had the largest variation in their standard errors (Figure 29 and Figure 30). This further demonstrated the need for sufficient sample sizes in each phenotype (superclade) being tested for genetic associations, as well as justifying the clustering of clades into superclades.

5.2.5. Human genome-wide associations with *M. tb* superclades

Although none of the SNPs passed the GWAS p-value cut-off of 5 x 10^{-8} in the SAC cohort, the most significant associations are reported in Table 11. It is likely due to the small sample size, with a subsequent reduction in statistical power, that none of the SNPs passed the p-value cut-off for the SAC cohort. In contrast, 30 SNPs were reported to be significantly associated with the superclade phenotypes tested for the Ghanaian cohort (Table 12). None of the SNPs with an LRT p-value less than 0.0005 in either cohort were found in the other, demonstrating the potentially population-specific association of SNPs with the different *M. tb* superclades, which has been previously shown in the investigation of TLRs and their association with cases of TB in populations of different ethnicities (Schurz et al. 2015).

For the Ghanaian cohort, 30 SNPs with significant LRT p-values for their MLRs were identified as being associated with the *M. tb* superclades investigated (Table 12). Nine of the SNPs located on chromosome 12 were mapped to *PDZRN4*. For these nine SNPs, the risk allele increased the chances of individuals being infected with the BeijingCAS superclades 2.5 times, and in the region of 3 times for the Ghana2 superclade, while the risk allele halved the chances of being infected with the TU superclade. Due to the low frequencies of the BeijingCAS and Ghana2 superclades observed for this cohort (Figure 13), it is possible that these odds ratios had been inflated by small sample sizes. Thus, the best interpretation for this cohort is that individuals with the risk alleles for these SNPs were at equal risk for having TB caused by the LAMCAM, EAI/AFRI, and HaarlemX superclades.

Sparse literature exists describing the direct influence of *M. tb* infection on the STARD4 and RARA genes, and no studies have investigated the outcomes of infection with different *M. tb* strains on these genes. The literature currently available are described in the paragraphs to follow.

5.2.5.1. STARD4

Cholesterol is essential for the maintenance of mammalian cell walls, and is a precursor for vitamin D, bile acids, and steroid hormones (Reviewed in Cruz et al. 2013). While cholesterol can be produced endogenously from acetate, the presence of excess concentrations of low density lipoprotein (LDL) cholesterol is known to be a serious health risk (Colpo 2005). The STARD4 gene encodes the StAR-related lipid transfer protein which plays a crucial role in the transmembrane trafficking of lipids (such as cholesterol) - an important energy source (Soccio et al. 2002).

Differential transcriptional responses have been observed when mouse bone marrow derived macrophages were infected with two strains of *M. tb*. Infection with the HN878 strain was reported to have upregulated more host genes involved in lipid metabolism including the *STARD4* gene, than infection with CDC1551 (Koo et al. 2012). The H37Rv *M. tb* strain reportedly induced a significantly upregulated expression of 386 genes in response to a lipid-rich environment, with those genes being implicated in efflux systems, sulphur reduction processes, and the capturing of iron (Aguilar-Ayala et al. 2017).

Infection of macrophages with pathogens such as *M. tb* stimulates the process of lipid droplet formation (Daniel et al. 2011). It has been hypothesised that *M. tb* initiates this process in an attempt to secure a reliable source of carbon to fuel bacterial growth (Brzostek et al. 2009). In addition to lipids being an energy source, the accumulation of cholesterol in the bacterial cell wall was shown to drastically reduce the permeability of the cell wall, subsequently reducing the penetrating capability of the anti-TB drug RIF (Brzostek et al. 2009). However, a recent study has contradicted the notion that lipid droplet formation is a bacteria-driven process. Instead, it was proposed that the formation of lipids is an immune system-activated process, and does not occur as a result of direct stimulation by *M. tb*, but rather via the IFN-γ, H1F-α-dependent pathway of the host immune system (Knight et al. 2018).

5.2.5.2. RARA

All-trans retinoic acid, the active form of Vitamin A, plays an essential role in the normal functioning of the adaptive and innate immune systems. The oral administration of retinoic acid to rats resulted in inhibition of the *M. tb* growth, following *in vitro* infection (Yamada et al. 2007).

The results of this study have highlighted several SNPs which possibly significantly increased the risk of individuals with Ghanaian ethnicity to being infected with the endemic TB strain of *M. africanum*. Given the burden of disease, and the dominance of *M. africanum* strains in Ghana, it may be a worthwhile expedition to explore the functional effect of these SNPs on the biological processes described.

5.3. Method validation in secondary cohort

The SAC cohort used to develop the method provided a platform of numerous challenges, enabling a thorough examination of the data and potential problems which may be encountered in other datasets. These challenges including dealing with missing data, incorrect infection assignments, and multiple infections. The SAC cohort is also a highly admixed population, not adequately represented in most of the currently publicly available reference panels. This degree of admixture provided another challenge which enabled the investigation of the suitability of different reference panels for imputation in the study dataset. Once the study cohort could be determined as having being imputed with a high quality, it was deemed sufficient for the association test using multinomial logistic regression. Although sample size likely impeded our statistical power to determine significant genome-wide associations to *M. tb* strains in the SAC dataset, we were able to demonstrate the use of the method and obtain SNPs near GWAS significance.

Replication of the method in the Ghanaian dataset was remarkably easier as only one infection was recorded in the database provided, with no missing data, making the Strain database QC easier in this dataset than in the SAC dataset. Furthermore, the principal components provided for the cohort to account for differences in ethnic contributions showed that the Ghanaian cohort was significantly less admixed than the SAC cohort. Individuals from West-Africa are also well-represented in the 1000GP3 and thus, it was not surprising that the 1000GP3 reference panel provided the highest quality of imputed genotype data.

Where whole-genome sequencing is not possible due to cost, genotype imputation of haplotypes from reference panels is a valuable solution to inferring untyped SNP markers to be used in GWAS. The use of the MIS and SIS aided the development of this method tremendously as the computational burden is split between an in-house server and two off-site servers. Both off-site servers are publicly accessible and require no payment for imputation service, making it a feasible option for resource-restricted institutions and research groups.

5.4. Limitations of the method

Obtaining a suitable sample size is a problem inherent of GWA studies making use of logistic regression modelling. Furthermore, the many phenotypes being analysed in this study, demanded a sufficient number of cases for each class. The frequency of each *M. tb* clade however, is dependent not only on the host, but is also affected by the virulence of the bacterium. Thus, with all these considered, the sample sizes included in the MLR should be sufficient for inclusion in the analysis but will likely also reflect the distribution in the population.

Another limitation of the study is that at the time of analysis, the AGR reference panel was not publicly available for download to a local machine. Thus, its use in this study could only be facilitated via the SIS, a freely accessible online imputation server. Through this, we were able to obtain high-quality imputed data for the SAC, but it was necessary to be mindful when drawing comparisons as the other workflows made use of different imputation software.

5.5. Recommendations for future studies

This study would not have been possible were it not for the collection of paired samples of blood and sputum from study participants. Thus, future studies will require that both samples be collected from participants in order to perform this association analysis. In order to prevent the reduction in statistical power of the association test, we chose to exclude low frequency *M. tb* cases at the superclade level. However, this may have brought in a weakness in interpretation of the odds ratios derived from the model. Therefore, in further applications of this method, it may be advisable to exclude low frequency clades prior to clustering, as opposed to the superclades used in this study. Lastly, Bayesian analysis of this data should be explored as it allows for the inclusion of priors such as strain prevalence which could not be included using the current method. Future studies may also assess imputation accuracy obtained from other combinations of software and reference panels, such as the AGR when it becomes publicly available, along with the IMPUTE2 software. The candidate genes identified in both cohorts might also be explored as possible targets for HDTs.

5.6. Concluding remarks

The main output of this study is the method developed which facilitates the association analysis of genome-wide SNP markers to the *M. tb* clades causing disease in humans. From the study we were able to identify the AGR as the best resource for imputing the admixed SAC population, as well as the 1000GP3 as the best resource for imputing cohorts of West-African ancestry. In contrast to another study that aimed to perform a genome-wide association analysis with Beijing- and non-Beijing *M. tb* clades, we developed a method that could assess the potential associations between SNPs and more than two phenotype classes. Furthermore, the SNPs obtained from the MLR showed that with a large cohort (as in the Ghanaian cohort), we were able to identify SNPs with highly significant LRT p-values for association to *M. tb* superclade.

With the current trajectory of the TB epidemic, novel methods are needed to augment current anti-TB therapies and combat the disease. This study provides the groundwork for future genome-wide association studies wishing to investigate the relationship between the host and the many strains of *M. tb* causing disease. Furthermore, the SNPs identified in this study may be evaluated in functional studies to assess their viability as targets for host-directed therapies.

6. References

Abel, Laurent, Jacques Fellay, David W Haas, Erwin Schurr, Geetha Srikrishna, Michael Urbanowski, Nimisha Chaturvedi, Sudha Srinivasan, Daniel H Johnson, and William R Bishai. 2017. "Genetics of Human Susceptibility to Active and Latent Tuberculosis: Present Knowledge and Future Perspectives." *The Lancet Infectious Diseases*. https://doi.org/10.1016/S1473-3099(17)30623-0.

Aguilar-Ayala, Diana A., Laurentijn Tilleman, Filip Van Nieuwerburgh, Dieter Deforce, Juan Carlos Palomino, Peter Vandamme, Jorge A. Gonzalez-Y-Merchand, and Anandi Martin. 2017. "The Transcriptome of Mycobacterium Tuberculosis in a Lipid-Rich Dormancy Model through RNAseq Analysis." *Scientific Reports* 7 (1). https://doi.org/10.1038/s41598-017-17751-x.

Alexander, David H, John Novembre, and Kenneth Lange. 2009. "Fast Model-Based Estimation of Ancestry in Unrelated Individuals." *Genome Research*, 1655–64. https://doi.org/10.1101/gr.094052.109.vidual.

Al-Orainey, Ibrahim O. 2009. "Diagnosis of Latent Tuberculosis: Can We Do Better?" *Annals of Thoracic Medicine* 4 (1): 5.

Altmüller, Janine, Lyle J. Palmer, Guido Fischer, Hagen Scherb, and Matthias Wjst. 2001. "Genomewide Scans of Complex Human Diseases: True Linkage Is Hard to Find." *The American Journal of Human Genetics* 69 (5): 936–50. https://doi.org/10.1086/324069.

American Thoracic Society. 2000. "Diagnostic Standards and Classification of Tuberculosis in Adults and Children." *American Journal of Respiratory and Critical Care Medicine* 161 (4): 1376–95. https://doi.org/10.1164/ajrccm.161.4.16141.

Anderson, Carl A., Fredrik H Pettersson, Geraldine M Clarke, Lon R Cardon, Andrew P. Morris, and Krina T. Zondervan. 2010. "Data Quality Control in Genetic Case-Control Association Studies." *Nature Protocols* 5 (9): 1564–73. https://doi.org/10.1038/nprot.2010.116.

Asante-Poku, Adwoa, Isaac Darko Otchere, Stephen Osei-Wusu, Esther Sarpong, Akosua Baddoo, Audrey Forson, Clement Laryea, et al. 2016. "Molecular Epidemiology of Mycobacterium Africanum in Ghana." *BMC Infectious Diseases* 16. https://doi.org/10.1186/s12879-016-1725-6.

Asante-Poku, Adwoa, Dorothy Yeboah-Manu, Isaac Darko Otchere, Samuel Y. Aboagye, David Stucki, Jan Hattendorf, Sonia Borrell, Julia Feldmann, Emelia Danso, and Sebastien Gagneux. 2015. "Mycobacterium Africanum Is Associated with Patient Ethnicity in Ghana." *PLoS Neglected Tropical Diseases* 9 (1). https://doi.org/10.1371/journal.pntd.0003370.

Ates, Louis S, Anzaan Dippenaar, Fadel Sayes, Alexandre Pawlik, Christiane Bouchier, Laurence Ma, Robin M Warren, et al. 2018. "Unexpected Genomic and Phenotypic Diversity of Mycobacterium Africanum Lineage 5 Affects Drug Resistance, Protein Secretion, and Immunogenicity." Edited by Richard Cordaux. *Genome Biology and Evolution* 10 (8): 1858–74. https://doi.org/10.1093/gbe/evy145.

Ayub, Asad, Steven H. Yale, Kurt D. Reed, Rana M. Nasser, and Steven R. Gilbert. 2004. "Testing for Latent Tuberculosis." *Clinical Medicine and Research* 2 (3): 191–94.

Behr, M. A., S. A. Warren, H. Salamon, P. C. Hopewell, A. Ponce De Leon, C. L. Daley, and P. M. Small. 1999. "Transmission of Mycobacterium Tuberculosis from Patients Smear-Negative for Acid-Fast Bacilli." *Lancet* 353 (9151): 444–49. https://doi.org/10.1016/S0140-6736(98)03406-0.

Bellamy, R. 1998. “Genetic Susceptibility to Tuberculosis in Human Populations.” *Thorax* 53 (7): 588–93.

Bellamy, Richard, Nulda Beyers, Keith P.W.J. McAdam, Cyril Ruwende, Robert Gie, Priscilla Samaai, Danite Bester, et al. 2000. “Genetic Susceptibility to Tuberculosis in Africans: A Genome-Wide Scan.” *Proceedings of the National Academy of Sciences of the United States of America* 97 (14): 8005–9. https://doi.org/10.1073/pnas.140201897.

Bellamy, Richard, Cyril Ruwende, Tumani Corrah, Keith P.W.J. McAdam, Hilton C. Whittle, and Adrian V.S Hill. 1998. “Variations in the NRAMP1 Gene and Susceptibility of Tuberculosis in West Africans.” *New England Journal of Medicine*, 640–44. https://doi.org/10.1056/NEJM199803053381002.

Bergtold, James S., Elizabeth A. Yeager, and Allen Featherstone. 2011. “Sample Size and Robustness of Inferences from Logistic Regression in the Presence of Nonlinearity and Multicollinearity.” In . Pittsburgh, Pennsylvania.

Bloom, Barry R., and Christopher J. L. Murray. 1992. “Tuberculosis: Commentary on a Reemergent Killer.” *Science* 257 (5073): 1055–64.

Blouin, Yann, Yolande Hauck, Charles Soler, Michel Fabre, Rithy Vong, Céline Dehan, Géraldine Cazajous, et al. 2012. “Significance of the Identification in the Horn of Africa of an Exceptionally Deep Branching Mycobacterium Tuberculosis Clade.” Edited by Igor Mokrousov. *PLoS ONE* 7 (12): e52841. https://doi.org/10.1371/journal.pone.0052841.

Brites, Daniela, and Sebastien Gagneux. 2015. “Co-Evolution of Mycobacterium Tuberculosis and Homo Sapiens.” *Immunological Reviews* 264 (1): 6–24. https://doi.org/10.1111/imr.12264.

Brosch, Roland, Stephen V. Gordon, Thierry Garnier, Karin Eiglmeier, Wafa Frigui, Philippe Valenti, Sandrine Dos Santos, et al. 2007. “Genome Plasticity of BCG and Impact on Vaccine Efficacy.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (13): 5596–5601. https://doi.org/10.1073/pnas.0700869104.

Brosch, Roland, Stephen V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, and K. Kremer. 2002. “A New Evolutionary Scenario for the Mycobacterium Tuberculosis Complex.” *Proceedings of the National Academy of Sciences* 99 (6): 3684–3689.

Brown, Timothy, Vladyslav Nikolayevskyy, Preya Velji, and Francis Drobniewski. 2010. “Associations between Mycobacterium Tuberculosis Strains and Phenotypes.” *Emerging Infectious Diseases* 16 (2): 272–80. https://doi.org/10.3201/eid1602.091032.

Browning, Brian L., and Sharon R. Browning. 2009. “A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals.” *American Journal of Human Genetics* 84 (2): 210–23. https://doi.org/10.1016/j.ajhg.2009.01.005.

Brzostek, A., J. Pawelczyk, A. Rumijowska-Galewicz, B. Dziadek, and J. Dziadek. 2009. “Mycobacterium Tuberculosis Is Able To Accumulate and Utilize Cholesterol.” *Journal of Bacteriology* 191 (21): 6584–91. https://doi.org/10.1128/JB.00488-09.

Cardon, Lon R, and Lyle J Palmer. 2003. “Population Stratification and Spurious Allelic Association.” *The Lancet* 361 (9357): 598–604. https://doi.org/10.1016/S0140-6736(03)12520-2.

Caws, Maxine, Guy Thwaites, Sarah Dunstan, Thomas R Hawn, Nguyen Thi Ngoc Lan, Nguyen Thuy Thuong Thuong, Kasia Stepniewska, et al. 2008. “The Influence of Host and Bacterial Genotype on the Development of Disseminated Disease with

Mycobacterium Tuberculosis.” *PLoS Pathogens* 4 (3): 1–9. https://doi.org/10.1371/journal.ppat.1000034.

Cegielski, J. Peter, Lenore Arab, and Joan Cornoni-Huntley. 2012. “Nutritional Risk Factors for Tuberculosis Among Adults in the United States, 1971–1992.” *American Journal of Epidemiology* 176 (5): 409–22. https://doi.org/10.1093/aje/kws007.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets.” *GigaScience* 4 (1). https://doi.org/10.1186/s13742-015-0047-8.

Chang, Diana, Feng Gao, Andrea Slavney, Li Ma, Yedael Y. Waldman, Aaron J. Sams, Paul Billing-Ross, Aviv Madar, Richard Spritz, and Alon Keinan. 2014. “Accounting for EXentricities: Analysis of the X Chromosome in GWAS Reveals X-Linked Genes Implicated in Autoimmune Diseases.” Edited by Tanja Zeller. *PLoS ONE* 9 (12): e113684. https://doi.org/10.1371/journal.pone.0113684.

Cheng, Matthew P., Claire Nour Abou Chakra, Cedric P Yansouni, Sonya Cnossen, Ian Shrier, Dick Menzies, and Christina Greenaway. 2016. “Risk of Active Tuberculosis in Patients with Cancer: A Systematic Review and Meta-Analysis.” *Clinical Infectious Diseases*, ciw838. https://doi.org/10.1093/cid/ciw838.

Chihota, Violet N., Antoinette Niehaus, Elizabeth M. Streicher, Xia Wang, Samantha L. Sampson, Peter Mason, Gunilla Källenius, et al. 2018. “Geospatial Distribution of Mycobacterium Tuberculosis Genotypes in Africa.” Edited by Ana Paula Arez. *PLOS ONE* 13 (8): e0200632. https://doi.org/10.1371/journal.pone.0200632.

Chimusa, Emile R, Michelle Daya, Marlo Möller, Raj Ramesar, Brenna M Henn, Paul D Van Helden, Nicola J Mulder, and Eileen G Hoal. 2013. “Determining Ancestry Proportions in Complex Admixture Scenarios in South Africa Using a Novel Proxy Ancestry Selection Method.” *PLoS ONE* 8 (9). https://doi.org/10.1371/journal.pone.0073971.

Chimusa, Emile R., Noah Zaitlen, Michelle Daya, Marlo Möller, Paul D van Helden, J. Mulder Nicola, Alkes L. Price, and Eileen G. Hoal. 2014. “Genome-Wide Association Study of Ancestry-Specific TB Risk in the South African Coloured Population.” *Human Molecular Genetics* 23 (3): 796–809. https://doi.org/10.1093/hmg/ddt462.

Cobat, A., Caroline J. Gallant, Leah Simkin, Gillian F. Black, Kim Stanley, Jane Hughes, T. Mark Doherty, et al. 2009. “Two Loci Control Tuberculin Skin Test Reactivity in an Area Hyperendemic for Tuberculosis.” *The Journal of Experimental Medicine* 206 (12): 2583–91. https://doi.org/10.1084/jem.20090892.

Cobat, A., Christine Poirier, Eileen Hoal, Anne Boland-Auge, France de La Rocque, François Corrard, Ghislain Grange, et al. 2015. “Tuberculin Skin Test Negativity Is Under Tight Genetic Control of Chromosomal Region 11p14-15 in Settings With Different Tuberculosis Endemicities.” *Journal of Infectious Diseases* 211 (2): 317–21. https://doi.org/10.1093/infdis/jiu446.

Cohen, Robert, Shirin Muzaffar, Jose Capellan, Hasan Azar, and Mustafa Chinikamwala. 1996. “The Validity of Classic Symptoms and Chest Radiographic Configuration in Predicting Pulmonary Tuberculosis.” *Chest* 109 (2): 420–23.

Cohen, T., and M. Murray. 2004. “Modeling Epidemics of Multidrug-Resistant M. Tuberculosis of Heterogeneous Fitness.” *Nature Medicine* 10 (10): 1117–21. https://doi.org/10.1038/nm1110.

Coker, Richard, Martin McKee, Rifat Atun, Boika Dimitrova, Ekaterina Dodonova, Sergei Kuznetsov, and Francis Drobniewski. 2006. “Risk Factors for Pulmonary Tuberculosis in Russia: Case-Control Study.” *BMJ : British Medical Journal* 332 (7533): 85–87. https://doi.org/10.1136/bmj.38684.687940.80.

Coll, Francesc, Ruth McNerney, José Afonso Guerra-Assunção, Judith R. Glynn, João Perdigão, Miguel Viveiros, Isabel Portugal, Arnab Pain, Nigel Martin, and Taane G. Clark. 2014. "A Robust SNP Barcode for Typing Mycobacterium Tuberculosis Complex Strains." *Nature Communications* 5: 4812. https://doi.org/10.1038/ncomms5812.

Colpo, Anthony. 2005. "LDL Cholesterol: 'Bad' Cholesterol, or Bad Science?" *Journal of American Physicians and Surgeons* 10 (3): 7.

Comas, Iñaki, Mireia Coscolla, Tao Luo, Sonia Borrell, Kathryn E Holt, Midori Kato-Maeda, Julian Parkhill, et al. 2013. "Out-of-Africa Migration and Neolithic Co-Expansion of Mycobacterium Tuberculosis with Modern Humans." *Nature Genetics* 45 (10): 1176–82. https://doi.org/10.1038/ng.2744.

Comstock, G.W. 1978. "Tuberculosis in Twins: A Re-Analysis of the Prophit Survey." *American Review of Respiratory Disease* 117 (4): 621–24.

Coscolla, Mireilla, and Sebastien Gagneux. 2010. "Does M. Tuberculosis Genomic Diversity Explain Disease Diversity?" *Drug Discovery Today. Disease Mechanisms* 7 (1): e43–59. https://doi.org/10.1016/j.ddmec.2010.09.004.

Cruz, Pedro M. R., Huanbiao Mo, Walter J. McConathy, Nirupama Sabnis, and Andras G. Lacko. 2013. "The Role of Cholesterol Metabolism and Cholesterol Transport in Carcinogenesis: A Review of Scientific Findings, Relevant to Future Cancer Therapeutics." *Frontiers in Pharmacology* 4. https://doi.org/10.3389/fphar.2013.00119.

Curtis, James, Yang Luo, Helen L Zenner, Delphine Cuchet-Lourenço, Changxin Wu, Kitty Lo, Mailis Maes, et al. 2015. "Susceptibility to Tuberculosis Is Associated with Variants in the ASAP1 Gene Encoding a Regulator of Dendritic Cell Migration." *Nature Genetics* 47 (5): 523–27. https://doi.org/10.1038/ng.3248.

Daftary, Amrita, Mike Frick, Nandita Venkatesan, and Madhukar Pai. 2017. "Fighting TB Stigma: We Need to Apply Lessons Learnt from HIV Activism." *BMJ Global Health* 2 (4): e000515. https://doi.org/10.1136/bmjgh-2017-000515.

Daniel, Jaiyanth, Hédia Maamar, Chirajyoti Deb, Tatiana D. Sirakova, and Pappachan E. Kolattukudy. 2011. "Mycobacterium Tuberculosis Uses Host Triacylglycerol to Accumulate Lipid Droplets and Acquires a Dormancy-Like Phenotype in Lipid-Loaded Macrophages." Edited by Vojo Deretic. *PLoS Pathogens* 7 (6): e1002093. https://doi.org/10.1371/journal.ppat.1002093.

Das, Sayantan, Lukas Forer, Sebastian Schönherr, Carlo Sidore, Adam E Locke, Alan Kwong, Scott I Vrieze, et al. 2016. "Next-Generation Genotype Imputation Service and Methods." *Nature Genetics* 48 (10): 1284–87. https://doi.org/10.1038/ng.3656.

Davis, J. Muse, and Lalita Ramakrishnan. 2009. "The Role of the Granuloma in Expansion and Dissemination of Early Tuberculous Infection." *Cell* 136 (1): 37–49. https://doi.org/10.1016/j.cell.2008.11.014.

Daya, Michelle, Lize Van Der Merwe, Ushma Galal, Marlo Möller, Muneeb Salie, Emile R. Chimusa, Joshua M. Galanter, et al. 2013. "A Panel of Ancestry Informative Markers for the Complex Five-Way Admixed South African Coloured Population." *PLoS ONE* 8 (12): 12–12. https://doi.org/10.1371/journal.pone.0082224.

Daya, Michelle, Lize Van Der Merwe, Paul D. Van Helden, Marlo Möller, and Eileen G. Hoal. 2015. "Investigating the Role of Gene-Gene Interactions in TB Susceptibility." *PLoS ONE* 10 (4): 1–25. https://doi.org/10.1371/journal.pone.0123970.

De Wit, Erika, Wayne Delport, Chimusa E. Rugamika, Ayton Meintjes, Marlo Moller, Paul D. Van Helden, Cathal Seoighe, and Eileen G. Hoal. 2010. "Genome-Wide Analysis of the Structure of the South African Coloured Population in the Western Cape." *Human Genetics* 128 (2): 145–53. https://doi.org/10.1007/s00439-010-0836-1.

Deelen, Patrick, Marc Jan Bonder, K. Joeri van der Velde, Harm-Jan Westra, Erwin Winder, Dennis Hendriksen, Lude Franke, and Morris A. Swertz. 2014. "Genotype Harmonizer: Automatic Strand Alignment and Format Conversion for Genotype Data Integration." *BMC Research Notes* 7 (1): 901.

Deelen, Patrick, Androniki Menelaou, Elisabeth M van Leeuwen, Alexandros Kanterakis, Freerk van Dijk, Carolina Medina-Gomez, Laurent C Francioli, et al. 2014. "Improved Imputation Quality of Low-Frequency and Rare Variants in European Samples Using the 'Genome of The Netherlands.'" *European Journal of Human Genetics* 22 (11): 1321–26. https://doi.org/10.1038/ejhg.2014.19.

Delaneau, Olivier, Cédric Coulonges, and Jean-François Zagury. 2008. "Shape-IT: New Rapid and Accurate Algorithm for Haplotype Inference." *BMC Bioinformatics* 9: 540. https://doi.org/10.1186/1471-2105-9-540.

Demers, Anne-Marie, Serge Mostowy, David Coetzee, Robin Warren, Paul van Helden, and Marcel A. Behr. 2010. "Mycobacterium Africanum Is Not a Major Cause of Human Tuberculosis in Cape Town, South Africa." *Tuberculosis* 90 (2): 143–44. https://doi.org/10.1016/j.tube.2010.02.004.

Denny, Joshua C, Lisa Bastarache, Marylyn D Ritchie, Robert J Carroll, Raquel Zink, Jonathan D Mosley, Julie R Field, et al. 2013. "Systematic Comparison of Phenome-Wide Association Study of Electronic Medical Record Data and Genome-Wide Association Study Data." *Nature Biotechnology* 31 (12): 1102–11. https://doi.org/10.1038/nbt.2749.

Denny, Joshua C., Marylyn D. Ritchie, Melissa A. Basford, Jill M. Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R. Masys, Dan M. Roden, and Dana C. Crawford. 2010. "PheWAS: Demonstrating the Feasibility of a Phenome-Wide Scan to Discover Gene–Disease Associations." *Bioinformatics* 26 (9): 1205–10. https://doi.org/10.1093/bioinformatics/btq126.

Dippenaar, Anzaan. 2014. "A Phylogenomic-and Proteomic Investigation into the Evolution and Biological Characteristics of the Members of the Group 2 Latin-American Mediterranean (LAM) Genotype of Mycobacterium Tuberculosis." PhD Thesis, Stellenbosch: Stellenbosch University.

Durbin, R. 2014. "Efficient Haplotype Matching and Storage Using the Positional Burrows-Wheeler Transform (PBWT)." *Bioinformatics* 30 (9): 1266–72. https://doi.org/10.1093/bioinformatics/btu014.

Eijk, Ellen A. van der, Esther van de Vosse, Jan P. Vandenbroucke, and Jaap T. van Dissel. 2007. "Heredity versus Environment in Tuberculosis in Twins." *American Journal of Respiratory and Critical Care Medicine* 176 (12): 1281–88. https://doi.org/10.1164/rccm.200703-435OC.

Fadista, João, Alisa K Manning, Jose C Florez, and Leif Groop. 2016. "The (in)Famous GWAS P-Value Threshold Revisited and Updated for Low-Frequency Variants." *European Journal of Human Genetics* 24 (8): 1202–5. https://doi.org/10.1038/ejhg.2015.269.

Farhat, M, C Greenaway, M Pai, and D Menzies. 2006. "False-Positive Tuberculin Skin Tests: What Is the Absolute Effect of BCG and Non-Tuberculous Mycobacteria?," 13.

Ferreira, M A. 2004. "Linkage Analysis: Principles and Methods for the Analysis of Human Quantitative Traits." *Twin.Res.* 7 (5): 513–30. https://doi.org/10.1375/1369052042335223.

Firdessa, Rebuma, Stefan Berg, Elena Hailu, Esther Schelling, Balako Gumi, Girume Erenso, Endalamaw Gadisa, et al. 2013. "Mycobacterial Lineages Causing Pulmonary and Extrapulmonary Tuberculosis, Ethiopia." *Emerging Infectious Diseases* 19 (3): 460–63. https://doi.org/10.3201/eid1903.120256.

Fox, Gregory J., Marianna Orlova, and Erwin Schurr. 2016. “Tuberculosis in Newborns: The Lessons of the ‘Lübeck Disaster’ (1929–1933).” Edited by James B. Bliska. *PLOS Pathogens* 12 (1): e1005271. https://doi.org/10.1371/journal.ppat.1005271.

Frothingham, R. 1995. “Differentiation of Strains in Mycobacterium Tuberculosis Complex by DNA Sequence Polymorphisms , Including Rapid Identification of M . Bovis BCG . These Include : Differentiation of Strains in Mycobacterium Tuberculosis Complex by DNA Sequence Polymorphis.” *Journal of Clinical Microbiology* 33 (4): 840–44.

Gagneux, Sebastien. 2012. “Host-Pathogen Coevolution in Human Tuberculosis.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 367 (1590): 850–59. https://doi.org/10.1098/rstb.2011.0316.

Gagneux, Sebastien, Kathryn DeRiemer, Tran Van, Midori Kato-Maeda, Bouke C de Jong, Sujatha Narayanan, Mark Nicol, et al. 2006. “Variable Host-Pathogen Compatibility in Mycobacterium Tuberculosis.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (8): 2869–73. https://doi.org/10.1073/pnas.0511240103.

Gagneux, Sebastien, and Peter M. Small. 2007. “Global Phylogeography of Mycobacterium Tuberculosis and Implications for Tuberculosis Product Development.” *Lancet Infectious Diseases* 7 (5): 328–37. https://doi.org/10.1016/S1473-3099(07)70108-1.

Gao, Feng, Diana Chang, Arjun Biddanda, Li Ma, Yingjie Guo, Zilu Zhou, and Alon Keinan. 2015. “XWAS: A Software Toolset for Genetic Data Analysis and Association Studies of the X Chromosome.” *Journal of Heredity* 106 (5): 666–71. https://doi.org/10.1093/jhered/esv059.

Gómez-Reino, Juan J., Loreto Carmona, Vicente Rodríguez Valverde, Emilio Martín Mola, and Maria Dolores Montero. 2003. “Treatment of Rheumatoid Arthritis with Tumor Necrosis Factor Inhibitors May Predispose to Significant Increase in Tuberculosis Risk: A Multicenter Active-Surveillance Report: Safety of TNF-Inhibitor Therapy.” *Arthritis & Rheumatism* 48 (8): 2122–27. https://doi.org/10.1002/art.11137.

Gonzalo-Asensio, Jesús, Irene Pérez, Nacho Aguiló, Santiago Uranga, Ana Picó, Carlos Lampreave, Alberto Cebollada, Isabel Otal, Sofía Samper, and Carlos Martín. 2018. “New Insights into the Transposition Mechanisms of IS6110 and Its Dynamic Distribution between Mycobacterium Tuberculosis Complex Lineages.” Edited by Carmen Buchrieser. *PLOS Genetics* 14 (4): e1007282. https://doi.org/10.1371/journal.pgen.1007282.

Grant, A. V., A. Sabri, A. Abid, I. Abderrahmani Rhorfi, M. Benkirane, H. Souhi, H. Naji Amrani, et al. 2016. “A Genome-Wide Association Study of Pulmonary Tuberculosis in Morocco.” *Human Genetics* 135 (3): 299–307. https://doi.org/10.1007/s00439-016-1633-2.

Greenwood, Celia M.T., T. Mary Fujiwara, Lucy J. Boothroyd, Mark A. Miller, Danielle Frappier, E. Anne Fanning, Erwin Schurr, and Kenneth Morgan. 2000. “Linkage of Tuberculosis to Chromosome 2q35 Loci, Including NRAMP1, in a Large Aboriginal Canadian Family.” *The American Journal of Human Genetics* 67 (2): 405–16. https://doi.org/10.1086/303012.

Gülbay, Banu Eriş, Özlem Ural Gürkan, Öznur Akkoca Yıldız, Zeynep Pınar Önen, Ferda Öner Erkekol, Ayşe Baççıoğlu, and Turan Acıcan. 2006. “Side Effects Due to Primary Antituberculosis Drugs during the Initial Phase of Therapy in 1149 Hospitalized Patients for Tuberculosis.” *Respiratory Medicine* 100 (10): 1834–42. https://doi.org/10.1016/j.rmed.2006.01.014.

Gutierrez, Maximiliano G., Sylvain Brisse, Roland Brosch, Michel Fabre, Bahia Omaïs, Magali Marmiesse, Philip Supply, and Veronique Vincent. 2005. “Ancient Origin and

Gene Mosaicism of the Progenitor of Mycobacterium Tuberculosis." *PLoS Pathogens* 1 (1): e5. https://doi.org/10.1371/journal.ppat.0010005.

Gutierrez, Maximiliano G., Sharon S. Master, Sudha B. Singh, Gregory A. Taylor, Maria I. Colombo, and Vojo Deretic. 2004. "Autophagy Is a Defense Mechanism Inhibiting BCG and Mycobacterium Tuberculosis Survival in Infected Macrophages." *Cell* 119 (6): 753–66. https://doi.org/10.1016/j.cell.2004.11.038.

Hanekom, M., G. D. Van Der Spuy, E. Streicher, S. L. Ndabambi, C. R E McEvoy, M. Kidd, N. Beyers, T. C. Victor, P. D. Van Helden, and R. M. Warren. 2007. "A Recently Evolved Sublineage of the Mycobacterium Tuberculosis Beijing Strain Family Is Associated with an Increased Ability to Spread and Cause Disease." *Journal of Clinical Microbiology* 45 (5): 1483–90. https://doi.org/10.1128/JCM.02191-06.

Herb, F., T. Thye, S. Niemann, E. N.L. Browne, M. A. Chinbuah, J. Gyapong, I. Osei, et al. 2007. "ALOX5 Variants Associated with Susceptibility to Human Pulmonary Tuberculosis." *Human Molecular Genetics* 17 (7): 1052–60. https://doi.org/10.1093/hmg/ddm378.

Hirschhorn, Joel N., and Mark J. Daly. 2005. "Genome-Wide Association Studies for Common Diseases and Complex Traits." *Nature Reviews Genetics* 6 (2): 95–108. https://doi.org/10.1038/nrg1521.

Hoal, Eileen G., Anzaan Dippenaar, Craig Kinnear, Paul D. van Helden, and Marlo Möller. 2017. "The Arms Race between Man and Mycobacterium Tuberculosis : Time to Regroup." *Infection, Genetics and Evolution*. https://doi.org/10.1016/j.meegid.2017.08.021.

Hong, Eun Pyo, Min Jin Go, Hyung-Lae Kim, and Ji Wan Park. 2017. "Risk Prediction of Pulmonary Tuberculosis Using Genetic and Conventional Risk Factors in Adult Korean Population." Edited by Sunil K. Ahuja. *PLOS ONE* 12 (3): e0174642. https://doi.org/10.1371/journal.pone.0174642.

Hoog, Anna H van't, Ikushi Onozaki, and Knut Lonnroth. 2014. "Choosing Algorithms for TB Screening: A Modelling Study to Compare Yield, Predictive Value and Diagnostic Burden." *BMC Infectious Diseases* 14 (1). https://doi.org/10.1186/1471-2334-14-532.

Horvat, Rebecca T. 2015. "Gamma Interferon Assays Used in the Diagnosis of Tuberculosis." Edited by C. J. Papasian. *Clinical and Vaccine Immunology* 22 (8): 845–49. https://doi.org/10.1128/CVI.00199-15.

Hove, P, J Molepo, S Dube, and M Nchabeleng. 2012. "Genotypic Diversity of Mycobacterium Tuberculosis in Pretoria." *South African Journal of Epidemiology and Infection* 27 (2): 77–83.

Howie, Bryan, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R. Abecasis. 2012. "Fast and Accurate Genotype Imputation in Genome-Wide Association Studies through Pre-Phasing." *Nature Genetics* 44 (8): 955–59. https://doi.org/10.1038/ng.2354.

Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. 2009. "A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies." *PLOS Genetics* 5 (6): e1000529. https://doi.org/10.1371/journal.pgen.1000529.

Intemann, Christopher D., Thorsten Thye, Stefan Niemann, Edmund N. L. Browne, Margaret Amanua Chinbuah, Anthony Enimil, John Gyapong, et al. 2009. "Autophagy Gene Variant IRGM −261T Contributes to Protection from Tuberculosis Caused by Mycobacterium Tuberculosis but Not by M. Africanum Strains." Edited by William Bishai. *PLoS Pathogens* 5 (9): e1000577. https://doi.org/10.1371/journal.ppat.1000577.

Jabot-Hanin, Fabienne, Aurélie Cobat, Jacqueline Feinberg, Ghislain Grange, Natascha Remus, Christine Poirier, Anne Boland-Auge, et al. 2016. "Major Loci on Chromosomes 8q and 3q Control Interferon γ Production Triggered by Bacillus Calmette-Guerin and 6-KDa Early Secretory Antigen Target, Respectively, in Various Populations." *Journal of Infectious Diseases* 213 (7): 1173–79. https://doi.org/10.1093/infdis/jiv757.

Jamieson, S, E N Miller, G F Black, C S Peacock, H J Cordell, J M M Howson, M-A Shaw, et al. 2004. "Evidence for a Cluster of Genes on Chromosome 17q11–Q21 Controlling Susceptibility to Tuberculosis and Leprosy in Brazilians." *Genes & Immunity* 5 (1): 46–57. https://doi.org/10.1038/sj.gene.6364029.

Jiménez-Corona, María Eugenia, Luis Pablo Cruz-Hervert, Lourdes García-García, Leticia Ferreyra-Reyes, Guadalupe Delgado-Sánchez, Miriam Bobadilla-del-Valle, Sergio Canizales-Quintero, et al. 2013. "Association of Diabetes and Tuberculosis: Impact on Treatment and Post-Treatment Outcomes." *Thorax* 68 (3): 214–20. https://doi.org/10.1136/thoraxjnl-2012-201756.

Jong, Bouke C. de, Martin Antonio, and Sebastien Gagneux. 2010. "Mycobacterium Africanum—Review of an Important Cause of Human Tuberculosis in West Africa." *PLoS Neglected Tropical Diseases* 4 (9). https://doi.org/10.1371/journal.pntd.0000744.

Kallmann, F.J., and D. Reisner. 1943. "Twin Studies on the Significance of Genetic Factors in Tuberculosis." *American Review of Tuberculosis and Pulmonary Diseases* 6: 549–71.

Kamgue Sidze, Larissa, Emmanuel Mouafo Tekwu, Christopher Kuaban, Jean-Paul Assam Assam, Jean-Claude Tedom, Stefan Niemann, Matthias Frank, and Véronique N. Penlap Beng. 2013. "Estimates of Genetic Variability of Mycobacterium Tuberculosis Complex and Its Association with Drug Resistance in Cameroon." *Advances in Infectious Diseases* 03 (01): 55–59. https://doi.org/10.4236/aid.2013.31007.

Kanterakis, Alexandros, Patrick Deelen, Freerk van Dijk, Heorhiy Byelas, Martijn Dijkstra, and Morris A Swertz. 2015. "Molgenis-Impute: Imputation Pipeline in a Box." *BMC Research Notes* 8 (1). https://doi.org/10.1186/s13104-015-1309-3.

Kinnear, Craig, Eileen G. Hoal, Haiko Schurz, Paul D. Van Helden, and Marlo Moller. 2017. "The Role of Human Host Genetics in Tuberculosis Resistance." *Expert Review of Respiratory Medicine* 11 (9): 721–37. https://doi.org/10.1080/17476348.2017.1354700.

Klopper, Marisa, Robin Mark Warren, Cindy Hayes, Nicolaas Claudius Gey van Pittius, Elizabeth Maria Streicher, Borna Müller, Frederick Adriaan Sirgel, et al. 2013. "Emergence and Spread of Extensively and Totally Drug-Resistant Tuberculosis, South Africa." *Emerging Infectious Diseases* 19 (3): 449–55. https://doi.org/10.3201//EID1903.120246.

Knight, Matthew, Jonathan Braverman, Kaleb Asfaha, Karsten Gronert, and Sarah Stanley. 2018. "Lipid Droplet Formation in Mycobacterium Tuberculosis Infected Macrophages Requires IFN-γ/HIF-1α Signaling and Supports Host Defense." Edited by Padmini Salgame. *PLOS Pathogens* 14 (1): e1006874. https://doi.org/10.1371/journal.ppat.1006874.

Knobler, Stacey, Adel A. F Mahmoud, Stanley M Lemon, Institute of Medicine (U.S.), Forum on Microbial Threats, Institute of Medicine (U.S.), and Board on Global Health. 2006. *The Impact of Globalization on Infectious Disease Emergence and Control: Exploring the Consequences and Opportunities : Workshop Summary*. Washington, DC: National Academies Press. http://public.eblib.com/choice/publicfullrecord.aspx?p=3378073.

Kolloli, Afsal, and Selvakumar Subbian. 2017. “Host-Directed Therapeutic Strategies for Tuberculosis.” *Frontiers in Medicine* 4. https://doi.org/10.3389/fmed.2017.00171.

Koo, Mi-Sun, Selvakumar Subbian, and Gilla Kaplan. 2012. “Strain Specific Transcriptional Response in Mycobacterium Tuberculosis Infected Macrophages.” *Cell Communication and Signaling* 10 (1): 2. https://doi.org/10.1186/1478-811X-10-2.

Kritzinger, Fiona E, Saskia Den Boon, Suzanne Verver, Donald A Enarson, Carl J Lombard, Martien W Borgdorff, Robert P Gie, and Nulda Beyers. 2009. “No Decrease in Annual Risk of Tuberculosis Infection in Endemic Area in Cape Town , South Africa” 14 (2): 136–42. https://doi.org/10.1111/j.1365-3156.2008.02213.x.

Krysl, J, M Korzeniewska-Kosela, N.L. Müller, and JM FitzGerald. 1994. “Radiologic Features of Pulmonary Tuberculosis: An Assessment of 188 Cases.” *Canadian Association of Radiologists Journal* 45 (2): 101–7.

Lim, Chong Hong, Hsin-Hua Chen, Yi-Hsing Chen, Der-Yuan Chen, Wen-Nan Huang, Jaw-Ji Tsai, Tsu-Yi Hsieh, et al. 2017. “The Risk of Tuberculosis Disease in Rheumatoid Arthritis Patients on Biologics and Targeted Therapy: A 15-Year Real World Experience in Taiwan.” Edited by Miguel Santin. *PLOS ONE* 12 (6): e0178035. https://doi.org/10.1371/journal.pone.0178035.

Lin, P. L., and J. L. Flynn. 2010. “Understanding Latent Tuberculosis: A Moving Target.” *The Journal of Immunology* 185 (1): 15–22. https://doi.org/10.4049/jimmunol.0903856.

Lin, Yi-Ting, Ping-Hsun Wu, Chun-Yu Lin, Ming-Yen Lin, Hung-Yi Chuang, Jee-Fu Huang, Ming-Lung Yu, and Wan-Long Chuang. 2014. “Cirrhosis as a Risk Factor for Tuberculosis Infection—A Nationwide Longitudinal Study in Taiwan.” *American Journal of Epidemiology* 180 (1): 103–10. https://doi.org/10.1093/aje/kwu095.

Liu, Qian, Elizabeth T. Cirulli, Yujun Han, Song Yao, Song Liu, and Qianqian Zhu. 2015. “Systematic Assessment of Imputation Performance Using the 1000 Genomes Reference Panels.” *Briefings in Bioinformatics* 16 (4): 549–62. https://doi.org/10.1093/bib/bbu035.

Luo, Yang, Sara Suliman, Samira Asgari, Tiffany Amariuta, Roger Calderon, Leonid Lecca, Segunda R. Leon, et al. 2018. “Progression of Recent Mycobacterium Tuberculosis Exposure to Active Tuberculosis Is a Highly Heritable Complex Trait Driven by 3q23 in Peruvians.” *BioRxiv*, 28. http://dx.doi.org/10.1101/401984.

Mahasirimongkol, Surakameth, Hideki Yanai, Taisei Mushiroda, Watoo Promphittayarat, Sukanya Wattanapokayakit, Jurairat Phromjai, Rika Yuliwulandari, et al. 2012. “Genome-Wide Association Studies of Tuberculosis in Asians Identify Distinct at-Risk Locus for Young Tuberculosis.” *Journal of Human Genetics* 57 (6): 363–67. https://doi.org/10.1038/jhg.2012.35.

Mandalakas, A M, A C Hesseling, N N Chegou, H L Kirchner, X Zhu, B J Marais, G F Black, N Beyers, and G Walzl. 2008. “High Level of Discordant IGRA Results in HIV-Infected Adults and Children,” 7.

Maples, Brian K., Simon Gravel, Eimear E. Kenny, and Carlos D. Bustamante. 2013. “RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference.” *The American Journal of Human Genetics* 93 (2): 278–88. https://doi.org/10.1016/j.ajhg.2013.06.020.

Marchini, Jonathan. 2010. “SNPTEST v2 Technical Details,” 10.

Marchini, Jonathan, and Bryan Howie. 2010. “Genotype Imputation for Genome-Wide Association Studies.” *Nature Reviews Genetics* 11 (7): 499–511. https://doi.org/10.1038/nrg2796.

Martin, Alicia R., Christopher R. Gignoux, Raymond K. Walters, Genevieve L. Wojcik, Benjamin M. Neale, Simon Gravel, Mark J. Daly, Carlos D. Bustamante, and Eimear

E. Kenny. 2017. “Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations.” *The American Journal of Human Genetics* 100 (4): 635–49. https://doi.org/10.1016/j.ajhg.2017.03.004.

Mathias, Rasika Ann, Margaret A. Taub, Christopher R. Gignoux, Wenqing Fu, Shaila Musharoff, Timothy D. O’Connor, Candelaria Vergara, et al. 2016. “A Continuum of Admixture in the Western Hemisphere Revealed by the African Diaspora Genome.” *Nature Communications* 7. https://doi.org/10.1038/ncomms12522.

McLaren, William, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R. S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. 2016. “The Ensembl Variant Effect Predictor.” *Genome Biology* 17 (1). https://doi.org/10.1186/s13059-016-0974-4.

Meyer, Christian G., and Thorsten Thye. 2014. “Host Genetic Studies in Adult Pulmonary Tuberculosis.” *Seminars in Immunology* 26 (6): 445–53. https://doi.org/10.1016/j.smim.2014.09.005.

Migliori, G. B., G. De Iaco, G. Besozzi, R. Centis, and D. M. Cirillo. 2007. “First Tuberculosis Cases in Italy Resistant to All Tested Drugs.” *Weekly Releases (1997–2007)* 12 (20): 3194. https://doi.org/10.2807/esw.12.20.03194-en.

Millard, Louise AC, Neil M. Davies, Tom R. Gaunt, George Davey Smith, and Kate Tilling. 2017. “Software Application Profile: PHESANT: A Tool for Performing Automated Phenome Scans in UK Biobank.” *International Journal of Epidemiology*.

Möller, Marlo, Erika De Wit, and Eileen G. Hoal. 2010. “Past, Present and Future Directions in Human Genetic Susceptibility to Tuberculosis.” *FEMS Immunology & Medical Microbiology* 58 (1): 3–26. https://doi.org/10.1111/j.1574-695X.2009.00600.x.

Moller, Marlo, and Eileen G. Hoal. 2010. “Current Findings, Challenges and Novel Approaches in Human Genetic Susceptibility to Tuberculosis.” *Tuberculosis* 90 (2): 71–83. https://doi.org/10.1016/j.tube.2010.02.002.

Möller, Marlo, Almut Nebel, Ruta Valentonyte, Paul D. van Helden, Stefan Schreiber, and Eileen G. Hoal. 2009. “Investigation of Chromosome 17 Candidate Genes in Susceptibility to TB in a South African Population.” *Tuberculosis* 89 (2): 189–94. https://doi.org/10.1016/j.tube.2008.10.001.

Motulsky, A. G. 1960. “Metabolic Polymorphisms and the Role of Infectious Diseases in Human Evolution.” *Human Biology* 32 (1): 28–62.

Nayak, Surajit, and Basanti Acharjya. 2012. “Mantoux Test and Its Interpretation.” *Indian Dermatology Online Journal* 3 (1): 2. https://doi.org/10.4103/2229-5178.93479.

Need, Anna C., and David B. Goldstein. 2009. “Next Generation Disparities in Human Genomics: Concerns and Remedies.” *Trends in Genetics* 25 (11): 489–94. https://doi.org/10.1016/j.tig.2009.09.012.

Niemann, S., S. Rüsch-Gerdes, M. L. Joloba, C. C. Whalen, D. Guwatudde, J. J. Ellner, K. Eisenach, et al. 2002. “Mycobacterium Africanum Subtype II Is Associated with Two Distinct Genotypes and Is a Major Cause of Human Tuberculosis in Kampala, Uganda.” *Journal of Clinical Microbiology* 40 (9): 3398–3405. https://doi.org/10.1128/JCM.40.9.3398-3405.2002.

Oetjens, Matthew T., Kristin Brown-Gentry, Robert Goodloe, Holli H. Dilks, and Dana C. Crawford. 2016. “Population Stratification in the Context of Diverse Epidemiologic Surveys Sans Genome-Wide Data.” *Frontiers in Genetics* 7. https://doi.org/10.3389/fgene.2016.00076.

Oki, Noffisat O, Alison A Motsinger-Reif, Paulo RZ Antas, Shawn Levy, Steven M Holland, and Timothy R Sterling. 2011. “Novel Human Genetic Variants Associated with Extrapulmonary Tuberculosis: A Pilot Genome Wide Association Study.” *BMC Research Notes* 4 (1): 28. https://doi.org/10.1186/1756-0500-4-28.

Omae, Yosuke, Licht Toyo-oka, Hideki Yanai, Supalert Nedsuwan, Sukanya Wattanapokayakit, Nusara Satproedprai, Nat Smittipat, et al. 2017. “Pathogen Lineage-Based Genome-Wide Association Study Identified CD53 as Susceptible Locus in Tuberculosis.” *Journal of Human Genetics*. https://doi.org/10.1038/jhg.2017.82.

Otchere, Isaac Darko, Mireia Coscollá, Leonor Sánchez-Busó, Adwoa Asante-Poku, Daniela Brites, Chloe Loiseau, Conor Meehan, et al. 2018. “Comparative Genomics of Mycobacterium Africanum Lineage 5 and Lineage 6 from Ghana Suggests Distinct Ecological Niches.” *Scientific Reports* 8 (1). https://doi.org/10.1038/s41598-018-29620-2.

Owusu-Dabo, Ellis, Ohene Adjei, Christian G. Meyer, Rolf D. Horstmann, Anthony Enimil, Thomas F. Kruppa, Frank Bonsu, et al. 2006. “Mycobacterium Tuberculosis Drug Resistance, Ghana.” *Emerging Infectious Diseases* 12 (7): 1170–72. https://doi.org/10.3201/eid1207.051028.

Pai, Madhukar, Claudia M. Denkinger, Sandra V. Kik, Molebogeng X. Rangaka, Alice Zwerling, Olivia Oxlade, John Z. Metcalfe, et al. 2014. “Gamma Interferon Release Assays for Detection of Mycobacterium Tuberculosis Infection.” *Clinical Microbiology Reviews* 27 (1): 3–20. https://doi.org/10.1128/CMR.00034-13.

Parsons, Linda M., Ákos Somoskövi, Cristina Gutierrez, Evan Lee, C. N. Paramasivan, Alash’le Abimiku, Steven Spector, Giorgio Roscigno, and John Nkengasong. 2011. “Laboratory Diagnosis of Tuberculosis in Resource-Poor Countries: Challenges and Opportunities.” *Clinical Microbiology Reviews* 24 (2): 314–50. https://doi.org/10.1128/CMR.00059-10.

Pastinen, Tomi, Kirsi Liitsola, Paavo Niini, Mika Salminen, and Ann-Christine SyväNen. 1998. “Contribution of the CCR5 and MBL Genes to Susceptibility to HIV Type 1 Infection in the Finnish Population.” *AIDS Research and Human Retroviruses* 14 (8): 695–98. https://doi.org/10.1089/aid.1998.14.695.

Pearson, Thomas A., and Teri A. Manolio. 2008. “How to Interpret a Genome-Wide Association Study.” *Jama* 299 (11): 1335–1344.

Pe’er, Itsik, Roman Yelensky, David Altshuler, and Mark J. Daly. 2008. “Estimation of the Multiple Testing Burden for Genomewide Association Studies of Nearly All Common Variants.” *Genetic Epidemiology* 32 (4): 381–85. https://doi.org/10.1002/gepi.20303.

Pietersen, Elize, Elisa Ignatius, Elizabeth M Streicher, Barbara Mastrapa, Xavier Padanilam, Anil Pooran, Motasim Badri, et al. 2014. “Long-Term Outcomes of Patients with Extensively Drug-Resistant Tuberculosis in South Africa: A Cohort Study.” *The Lancet* 383 (9924): 1230–39. https://doi.org/10.1016/S0140-6736(13)62675-6.

Png, Eileen, Bachti Alisjahbana, Edhyana Sahiratmadja, Sangkot Marzuki, Ron Nelwan, Yanina Balabanova, Vladyslav Nikolayevskyy, et al. 2012. “A Genome Wide Association Study of Pulmonary Tuberculosis Susceptibility in Indonesians.” *BMC Medical Genetics* 13 (1). https://doi.org/10.1186/1471-2350-13-5.

Prada-Medina, Cesar A., Kiyoshi F. Fukutani, Nathella Pavan Kumar, Leonardo Gil-Santana, Subash Babu, Flávio Lichtenstein, Kim West, et al. 2017. “Systems Immunology of Diabetes-Tuberculosis Comorbidity Reveals Signatures of Disease Complications.” *Scientific Reports* 7 (1). https://doi.org/10.1038/s41598-017-01767-4.

Pratiwi, Rita Dian. 2016. “Socio-Economic and Environmenys Risk Factors of Tuberculosis in Wonosobo, Central Java, Indonesia.” In , 89. Graduate Studies in Public Health, Graduate Program, Sebelas Maret University Jl. Ir Sutami 36A, Surakarta 57126. Telp/Fax: (0271) 632 450 ext.208 First website:http//:s2ikm.pasca.uns.ac.id Second website: www.theicph.com. Email: theicph2016@gmail.com. https://doi.org/10.26911/theicph.2016.027.

Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38: 904.

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly. 2000. "Inference of Population Structure Using Multilocus Genotype Data." *Genetics* 155: 945–59.

Puffer, Ruth Rice. 1944. *Familial Susceptibility to Tuberculosis. Its Importance as a Public Health Problem.*

Purcell, Shaun, and Christopher Chang. n.d. *PLINK 1.9*. www.cog-genomics.org/plink/1.9/.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*. 2017. Vienna, Austria. https://www.R project.org/.

Ramaliba, T.M., T.G. Tshitangano, H.A. Akinsola, and M. Thendele. 2017. "Tuberculosis Risk Factors in Lephalale Local Municipality of Limpopo Province, South Africa." *South African Family Practice* 59 (5): 182–87. https://doi.org/10.1080/20786190.2017.1304734.

Reich, David E., and Eric S. Lander. 2001. "On the Allelic Spectrum of Human Disease." *TRENDS in Genetics* 17 (9): 502–510.

Restrepo, Blanca I., Léanie Kleynhans, Alejandra B. Salinas, Bassent Abdelbary, Happy Tshivhula, Genesis P. Aguillón-Durán, Carine Kunsevi-Kilola, et al. 2018. "Diabetes Screen during Tuberculosis Contact Investigations Highlights Opportunity for New Diabetes Diagnosis and Reveals Metabolic Differences between Ethnic Groups." *Tuberculosis* 113: 10–18. https://doi.org/10.1016/j.tube.2018.08.007.

Rieder, H L. 2003. "Clarification of the Luebeck Infant Tuberculosis." *Pneumologie* 57 (402–405): 4.

Risch, Neil, and Kathleen Merikangas. 1996. "The Future of Genetic Studies of Complex Human Diseases." *Science* 273: 1516–17.

Roshyara, Nab Raj, Katrin Horn, Holger Kirsten, Peter Ahnert, and Markus Scholz. 2016. "Comparing Performance of Modern Genotype Imputation Methods in Different Ethnicities." *Scientific Reports* 6 (1). https://doi.org/10.1038/srep34386.

Ryu, Yon Ju. 2015. "Diagnosis of Pulmonary Tuberculosis: Recent Advances and Diagnostic Algorithms." *Tuberculosis and Respiratory Diseases* 78 (2): 64–71. https://doi.org/10.4046/trd.2015.78.2.64.

Sakamoto, K. 2012. "The Pathology of Mycobacterium Tuberculosis Infection." *Veterinary Pathology* 3 (2): 872–80. https://doi.org/10.1177/0300985811429313.

Salie, Muneeb, Lize Van Der Merwe, Marlo Moller, Michelle Daya, Gian D. Van Der Spuy, Paul D. Van Helden, Maureen P. Martin, et al. 2014. "Associations between Human Leukocyte Antigen Class i Variants and the Mycobacterium Tuberculosis Subtypes Causing Disease." *Journal of Infectious Diseases* 209 (2): 216–23. https://doi.org/10.1093/infdis/jit443.

Schaid, D J, and S S Sommert. 1993. "Genotype Relative Risks: Methods for Design and Analysis of Candidate-Gene Association Studies." *American Journal of Human Genetics* 53: 1114–26.

Schluger, Neil W., and Joseph Burzynski. 2010. "Recent Advances in Testing for Latent TB." *Chest* 138 (6): 1456–63. https://doi.org/10.1378/chest.10-0366.

Schurz, Haiko, Michelle Daya, Marlo Möller, Eileen G. Hoal, and Muneeb Salie. 2015. "TLR1, 2, 4, 6 and 9 Variants Associated with Tuberculosis Susceptibility: A Systematic Review and Meta-Analysis." Edited by Graham R. Wallace. *PLOS ONE* 10 (10): e0139711. https://doi.org/10.1371/journal.pone.0139711.

Schurz, Haiko, Craig J Kinnear, Christopher R Gignoux, Genevieve L Wojcik, Paul D van Helden, Gerard C Tromp, Brenna M Henn, Eileen G Hoal, and Marlo Moller. 2018. "A Sex-Stratified Genome-Wide Association Study of Tuberculosis Using a Multi-Ethnic Genotyping Array." *BioRxiv*. https://doi.org/10.1101/405571.

Seddon, James A, Anneke C Hesseling, Peter Godfrey-Faussett, Katherine Fielding, and H Simon Schaaf. 2013. "Risk Factors for Infection and Disease in Child Contacts of Multidrug-Resistant Tuberculosis: A Cross-Sectional Study." *BMC Infectious Diseases* 13 (1). https://doi.org/10.1186/1471-2334-13-392.

Sekati, E M, J Molepo, and M Nchabeleng. 2015. "Molecular Characterisation and Associated Drug Susceptibility Patterns of Mycobacterium Tuberculosis Isolates from South African Children Molecular Characterisation and Associated Drug Susceptibility Patterns of Mycobacterium Tuberculosis Isolates from So." *Southern African Journal of Infectious Diseases* 30 (1): 11–16. https://doi.org/10.1080/23120053.2015.1103955.

Sester, M., G. Sotgiu, C. Lange, C. Giehl, E. Girardi, G. B. Migliori, A. Bossink, et al. 2011. "Interferon- Release Assays for the Diagnosis of Active Tuberculosis: A Systematic Review and Meta-Analysis." *European Respiratory Journal* 37 (1): 100–111. https://doi.org/10.1183/09031936.00114810.

Shisana, O., T Rhele, L.C. Simbayi, K. Zuma, S. Jooste, N. Zungu, D. Labadarios, and D. Onoya. 2012. *South African National HIV Prevalence, Incidence and Behaviour Survey, 2012*.

Søborg, Christian, Hans O. Madsen, Åse B. Andersen, Troels Lillebaek, Axel Kok-Jensen, and Peter Garred. 2003. "Mannose-Binding Lectin Polymorphisms in Clinical Tuberculosis." *The Journal of Infectious Diseases* 188 (5): 777–82. https://doi.org/10.1086/377183.

Sobota, Rafal S., Catherine M. Stein, Nuri Kodaman, Laura B. Scheinfeldt, Isaac Maro, Wendy Wieland-Alter, Robert P. Igo, et al. 2016. "A Locus at 5q33.3 Confers Resistance to Tuberculosis in Highly Susceptible Individuals." *The American Journal of Human Genetics* 98 (3): 514–24. https://doi.org/10.1016/j.ajhg.2016.01.015.

Soccio, Raymond E., R. M. Adams, M. J. Romanowski, E. Sehayek, S. K. Burley, and J. L. Breslow. 2002. "The Cholesterol-Regulated StarD4 Gene Encodes a StAR-Related Lipid Transfer Protein with Two Closely Related Homologues, StarD5 and StarD6." *Proceedings of the National Academy of Sciences* 99 (10): 6943–48. https://doi.org/10.1073/pnas.052143799.

Spuy, G. D. van der, K. Kremer, S.L. Ndabambi, N. Beyers, R. Dunbar, B.J. Marais, P.D. van Helden, and R.M. Warren. 2009. "Changing Mycobacterium Tuberculosis Population Highlights Clade-Specific Pathogenic Characteristics." *Tuberculosis* 89 (2): 120–25. https://doi.org/10.1016/j.tube.2008.09.003.

Starke, Jeffrey R., and Kym T. Taylor-Watts. 1989. "Tuberculosis in the Pediatric Population of Houston, Texas." *Pediatrics* 84 (1): 28–35.

Stead, William W., John W. Senner, William T. Reddick, and John P. Lofgren. 1990. "Racial Differences in Susceptibility to Infection by Mycobacterium Tuberculosis." *New England Journal of Medicine* 322 (7): 422–27. https://doi.org/10.1056/NEJM199002153220702.

Stein, Catherine M. 2011. "Genetic Epidemiology of Tuberculosis Susceptibility: Impact of Study Design." *PLoS Pathogen* 7 (1): 1–8. https://doi.org/10.1371/journal.

Stein, Catherine M, David Guwatudde, Margaret Nakakeeto, Pierre Peters, Robert C Elston, Hemant K Tiwari, Roy Mugerwa, and Christopher C Whalen. 2003. "Heritability Analysis of Cytokines as Intermediate Phenotypes of Tuberculosis." *The Journal of Infectious Diseases* 187 (11): 1679–85. https://doi.org/10.1086/375249.

Stein, Catherine M, Sarah Zalwango, LaShaunda L. Malone, Sungho Won, Harriet Mayanja-Kizza, Roy D. Mugerwa, Dmitry V. Leontiev, et al. 2008. "Genome Scan of M. Tuberculosis Infection and Disease in Ugandans." Edited by Madhukar Pai. *PLoS ONE* 3 (12): e4094. https://doi.org/10.1371/journal.pone.0004094.

Strategic Development Information and GIS Department, and City of Cape Town. 2012. "Census 2011 - City of Cape Town."

Stringer, Sven, Naomi R. Wray, René S. Kahn, and Eske M. Derks. 2011. "Underestimated Effect Sizes in GWAS: Fundamental Limitations of Single SNP Analysis for Dichotomous Phenotypes." Edited by Nicholas John Timpson. *PLoS ONE* 6 (11): e27964. https://doi.org/10.1371/journal.pone.0027964.

Stucki, David, Daniela Brites, Leïla Jeljeli, Mireia Coscolla, Qingyun Liu, Andrej Trauner, Lukas Fenner, et al. 2016. "Mycobacterium Tuberculosis Lineage 4 Comprises Globally Distributed and Geographically Restricted Sublineages." *Nature Genetics* 48 (12): 1535–43. https://doi.org/10.1038/ng.3704.

Sturgill-Koszycki, S, P. Schlesinger, P Chakraborty, P. Haddix, H. Collins, A. Fok, R. Allen, S. Gluck, J Heuser, and D. Russell. 1994. "Lack of Acidification in Mycobacterium Phagosomes Produced by Exclusion of the Vesicular Proton-ATPase." *Science* 263 (5147): 678–81. https://doi.org/10.1126/science.8303277.

Sudmant, Peter H., Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, Kai Ye, et al. 2015. "An Integrated Map of Structural Variation in 2,504 Human Genomes." *Nature* 526 (7571): 75–81. https://doi.org/10.1038/nature15394.

Supply, Philip, Caroline Allix, Sarah Lesjean, Mara Cardoso-Oelemann, S. Rusch-Gerdes, Eve Willery, Evgueni Savine, et al. 2006. "Proposal for Standardization of Optimized Mycobacterial Interspersed Repetitive Unit-Variable-Number Tandem Repeat Typing of Mycobacterium Tuberculosis." *Journal of Clinical Microbiology* 44 (12): 4498–4510. https://doi.org/10.1128/JCM.01392-06.

Sveinbjornsson, Gardar, Daniel F Gudbjartsson, Bjarni V Halldorsson, Karl G Kristinsson, Magnus Gottfredsson, Jeffrey C Barrett, Larus J Gudmundsson, et al. 2016. "HLA Class II Sequence Variants Influence Tuberculosis Risk in Populations of European Ancestry." *Nature Genetics* 48 (3): 318–22. https://doi.org/10.1038/ng.3498.

The 1000 Genomes Project Consortium. 2012. "An Integrated Map of Genetic Variation from 1,092 Human Genomes." *Nature* 491 (7422): 56–65. https://doi.org/10.1038/nature11632.

The International HapMap 3 Consortium. 2010. "Integrating Common and Rare Genetic Variation in Diverse Human Populations." *Nature* 467 (7311): 52–58. https://doi.org/10.1038/nature09298.

The International HapMap Consortium. 2005. "A Haplotype Map of the Human Genome." *Nature* 437 (7063): 1299–1320. https://doi.org/10.1038/nature04226.

"The Lubeck Catastrophe: A General Review." 1931. *The British Medical Journal* 1 (3674): 96–98.

Thoma-Uszynski, S. 2001. "Induction of Direct Antimicrobial Activity Through Mammalian Toll-Like Receptors." *Science* 291 (5508): 1544–47. https://doi.org/10.1126/science.291.5508.1544.

Thye, Thorsten, Stefan Niemann, Kerstin Walter, Susanne Homolka, Christopher D. Intemann, Margaret Amanua Chinbuah, Anthony Enimil, et al. 2011. "Variant G57E of Mannose Binding Lectin Associated with Protection against Tuberculosis Caused by Mycobacterium Africanum but Not by M. Tuberculosis." Edited by Tanya Parish. *PLoS ONE* 6 (6): e20908. https://doi.org/10.1371/journal.pone.0020908.

Thye, Thorsten, Ellis Owusu-Dabo, Fredrik O Vannberg, Reinout Van Crevel, James Curtis, Edhyana Sahiratmadja, Yanina Balabanova, et al. 2012. “Common Variants at 11p13 Are Associated with Susceptibility to Tuberculosis.” *Nature Genetics* 44 (3): 257–59. https://doi.org/10.1038/ng.1080.

Thye, Thorsten, Fredrik O. Vannberg, Sunny H. Wong, Ellis Owusu-Dabo, Ivy Osei, John Gyapong, Giorgio Sirugo, et al. 2010. “Genome-Wide Association Analyses Identifies a Susceptibility Locus for Tuberculosis on Chromosome 18q11.2.” *Nature Genetics* 42 (9): 739–41. https://doi.org/10.1038/ng.639.

Tortoli, Enrico, Paola Cichero, Claudio Piersimoni, M Tullia Simonetti, Giampietro Gesu, and Domenico Nista. 1999. “Use of BACTEC MGIT 960 for Recovery of Mycobacteria from Clinical Specimens: Multicenter Study.” *J. CLIN. MICROBIOL.* 37: 5.

Udwadia, Zarir F., Rohit A. Amale, Kanchan K. Ajbani, and Camilla Rodrigues. 2012. “Totally Drug-Resistant Tuberculosis in India.” *Clinical Infectious Diseases* 54 (4): 579–581.

Uren, Caitlin, Brenna M. Henn, Andre Franke, Michael Wittig, Paul D. van Helden, Eileen G. Hoal, and Marlo Möller. 2017. “A Post-GWAS Analysis of Predicted Regulatory Variants and Tuberculosis Susceptibility.” *PLoS ONE* 12 (4). https://doi.org/10.1371/journal.pone.0174738.

Velayati, Ali Akbar, Mohammad Reza Masjedi, Parissa Farnia, Payam Tabarsi, Jalladein Ghanavi, Abol Hassan ZiaZarifi, and Sven Eric Hoffner. 2009. “Emergence of New Forms of Totally Drug-Resistant Tuberculosis Bacilli: Super Extensively Drug-Resistant Tuberculosis or Totally Drug-Resistant Strains in Iran.” *Chest* 136 (2): 420–425.

Visscher, Peter M., Matthew A. Brown, Mark I. McCarthy, and Jian Yang. 2012. “Five Years of GWAS Discovery.” *American Journal of Human Genetics* 90 (1): 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029.

Vynnycky, E., and Paul E. M. Fine. 2000. “Lifetime Risks, Incubation Period, and Serial Interval of Tuberculosis.” *American Journal of Epidemiology* 152 (3): 247–63. https://doi.org/10.1093/aje/152.3.247.

Wang, Jing, David C. Samuels, Yu Shyr, and Yan Guo. 2017. “StrandScript: Evaluation of Illumina Genotyping Array Design and Strand Correction.” *Bioinformatics* 33 (15): 2399–2401. https://doi.org/10.1093/bioinformatics/btx186.

Wang, Shu-Hua. 2012. “The Influence of Increasing Age on Susceptibility of the Elderly to Tuberculosis.” *Open Longevity Science* 6 (1): 73–82. https://doi.org/10.2174/1876326X01206010073.

WHO. 2001. “World Medical Association Declaration of Helsinki.” Bulletin of the World Health Organization. http://www.who.int/bulletin/archives/79%284%29373.pdf.

WHO. 2008. “Molecular Line Probe Assays for Rapid Screening of Patients at Risk of Multidrug-Resistant Tuberculosis (MDR-TB).” http://www.who.int/tb/laboratory/lpa_policy.pdf.

WHO. 2013. “World Health Organization. Definitions and Reporting Framework for Tuberculosis – 2013 Revision.” www.who.int/iris/bitstream/10665/79199/1/9789241505345_eng.pdf.

WHO. 2014a. “2015 Global Tuberculosis Report.” *2015 Global Tuberculosis Report*.

WHO. 2014b. “WHO | MDG 6: Combat HIV/AIDS, Malaria and Other Diseases.” WHO. World Health Organization. 2014. http://www.who.int/topics/millennium_development_goals/diseases/en/.

WHO 2017a. “Ghana Tuberculosis Profile.” TB burden estimates and country-reported TB data.

https://extranet.who.int/sree/Reports?op=Replet&name=/WHO_HQ_Reports/G2/PROD/EXT/TBCountryProfile&ISO2=GH&outtype=pdf.

WHO 2017b. *GLOBAL TUBERCULOSIS REPORT 2017*. S.l.: WHO.

WHO. 2017c. "South Africa Tuberculosis Profile." TB burden estimates and country-reported TB data. https://extranet.who.int/sree/Reports?op=Replet&name=/WHO_HQ_Reports/G2/PROD/EXT/TBCountryProfile&ISO2=ZA&outtype=PDF.

Wigginton, Janis E, David J Cutler, and Goncalo R Abecasis. 2005. "A Note on Exact Tests of Hardy-Weinberg Equilibrium." *American Journal of Human Genetics* 76 (5): 887–93. https://doi.org/10.1086/429864.

Wirth, Thierry, Falk Hildebrand, Caroline Allix-Béguec, Florian Wölbeling, Tanja Kubica, Kristin Kremer, Dick van Soolingen, et al. 2008. "Origin, Spread and Demography of the Mycobacterium Tuberculosis Complex." Edited by Mark Achtman. *PLoS Pathogens* 4 (9): e1000160. https://doi.org/10.1371/journal.ppat.1000160.
Human Genetics 128 (2): 145–53. https://doi.org/10.1007/s00439-010-0836-1.

Yamada, Hiroyuki, Satoru Mizuno, A. Catharine Ross, and Isamu Sugawara. 2007. "Retinoic Acid Therapy Attenuates the Severity of Tuberculosis While Altering Lymphocyte and Macrophage Numbers and Cytokine Expression in Rats Infected with Mycobacterium Tuberculosis." *The Journal of Nutrition* 137 (12): 2696–2700. https://doi.org/10.1093/jn/137.12.2696.

Yim, Jae Joon, and Paramasivam Selvaraj. 2010. "Genetic Susceptibility in Tuberculosis." *Respirology* 15 (2): 241–56. https://doi.org/10.1111/j.1440-1843.2009.01690.x.

Yimer, Solomon A., Gunnstein Norheim, Amine Namouchi, Ephrem D. Zegeye, Wibeke Kinander, Tone Tønjum, Shiferaw Bekele, et al. 2015. "Mycobacterium Tuberculosis Lineage 7 Strains Are Associated with Prolonged Patient Delay in Seeking Treatment for Pulmonary Tuberculosis in Amhara Region, Ethiopia." Edited by G. A. Land. *Journal of Clinical Microbiology* 53 (4): 1301–9. https://doi.org/10.1128/JCM.03566-14.

Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir. 2012. "A High-Performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data." *Bioinformatics* 28 (24): 3326–28. https://doi.org/10.1093/bioinformatics/bts606.

Zumla, Alimuddin, Abdulaziz Bin Saeed, Badriah Alotaibi, Saber Yezli, Osman Dar, Kingsley Bieh, Matthew Bates, et al. 2016. "Tuberculosis and Mass Gatherings—Opportunities for Defining Burden, Transmission Risk, and the Optimal Surveillance, Prevention, and Control Measures at the Annual Hajj Pilgrimage." *International Journal of Infectious Diseases* 47: 86–91. https://doi.org/10.1016/j.ijid.2016.02.003.

7. Appendix I: Extended *M. tb* clade and superclade distributions in the SAC cohort having multiple infection records

7.1. Distributions of Clades and Superclades for One and Two infection participants in the SAC cohort

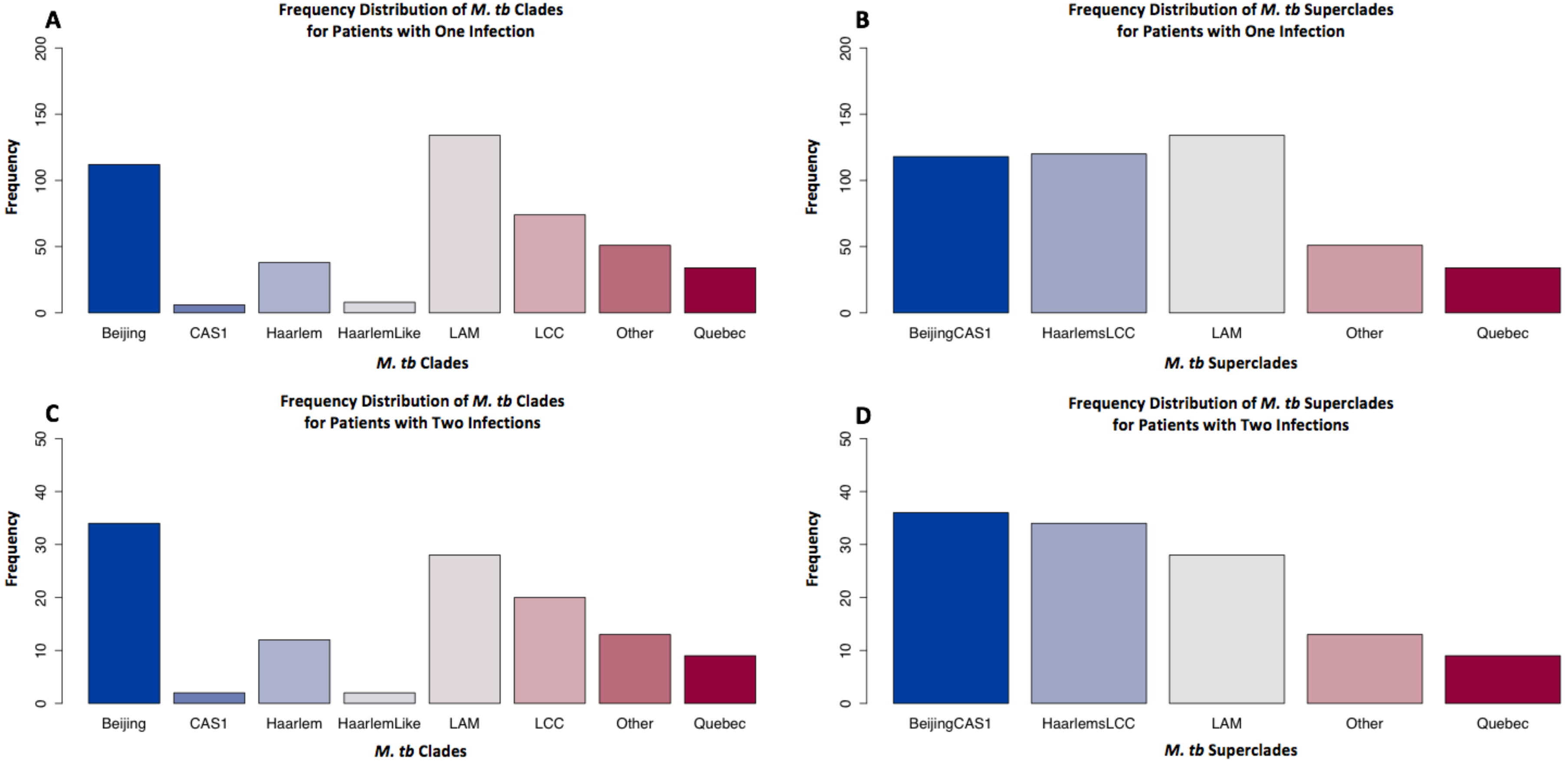


Figure 31: Frequency distributions for records of participants having one infection (n = 459) and the two infections (120 records for 60 participants) in the SAC cohort matched to M. tb clade. Figures A and B show the distribution for participants with one infection while figures C and D show the distribution for participants having two infections.

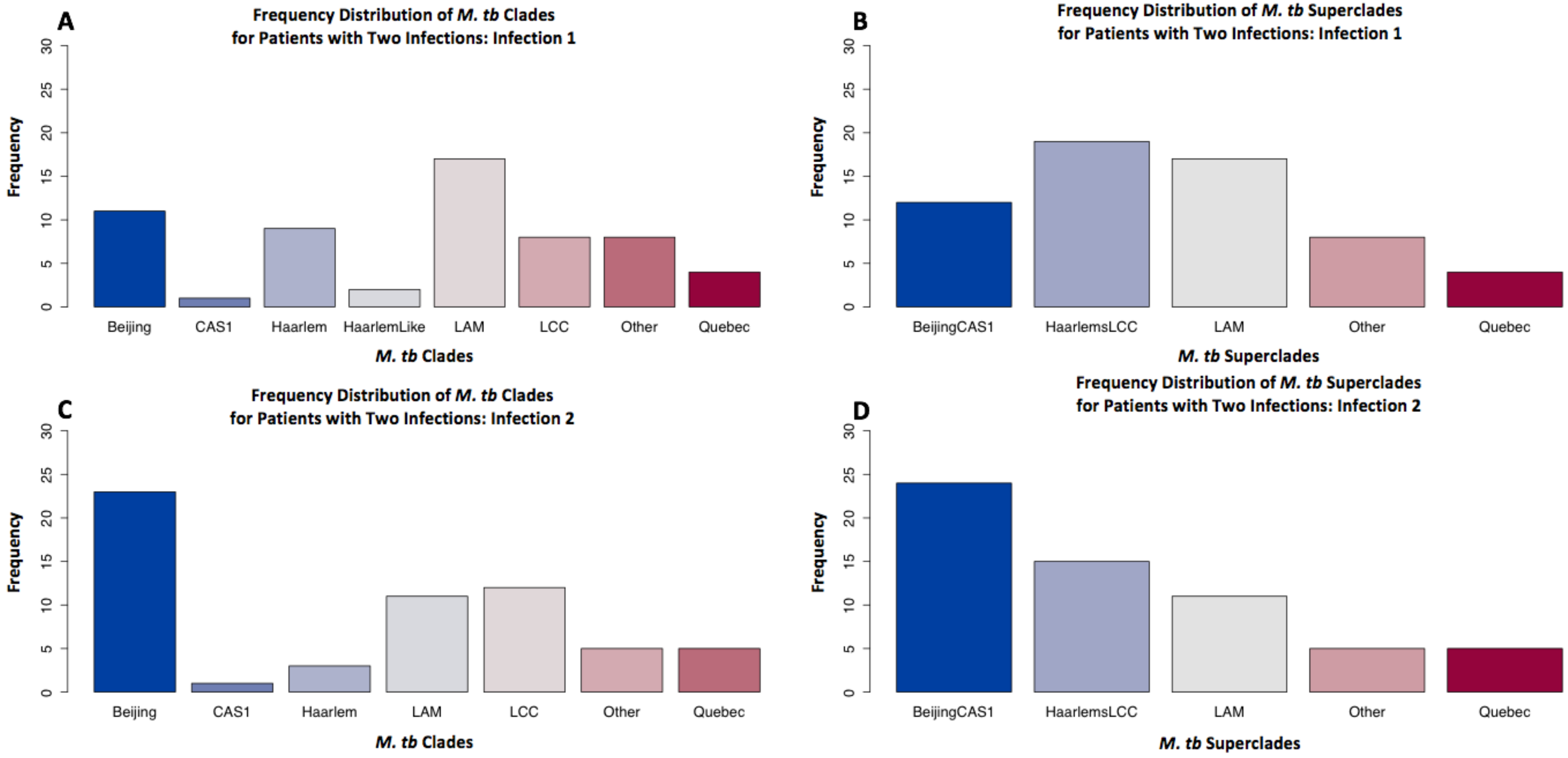


Figure 32: Frequency distributions for records for each of the infections listed for the genotyped participants having two infections recorded in the database. Figures A and B show the clade- and superclade distributions for the first infection, while figures C and D show the clade and superclade distributions for the second of the two infections.