Sex-bias and tuberculosis susceptibility: Bioinformatic and Biostatistical evaluation of trans-ethnic genomic datasets

by

Haiko Schurz

Dissertation presented for the degree of Doctor of Philosophy in Human Genetics in

the

Faculty of Medicine and Health Sciences at

Stellenbosch University

Supervisor: Dr. Marlo Möller

Co-supervisors: Prof. Eileen Hoal

Prof. Gerard Tromp

Dr. Craig Kinnear

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.

This dissertation includes no original papers published in peer reviewed journals or books and five unpublished publications. The development and writing of the papers (published and unpublished) were the principal responsibility of myself and for each of the cases where this is not the case a declaration is included in the dissertation indication the nature and extent of the contributions of co-authors.

Date: 19 September 2018

Declaration by the candidate:

Regarding chapter 2,3,4,5 and 6, the nature and scope of my contribution were as follows:

Chapter 2: The X chromosome and sex-specific effects in infectious disease susceptibility.

- First author
- Conceived the review
- Writing of manuscript

Chapter 3: Autosomal and X chromosome markers confirm strong sex-biased admixture in the South African Coloured population.

- First author
- Analysis and interpretation of data
- Writing of manuscript

Chapter 4: A sex-stratified genome-wide association study of tuberculosis using a multi-ethnic genotyping array.

- First author
- Analysis and interpretation of data
- Writing of manuscript

Chapter 5: Evaluating the accuracy of imputation in the 5-way admixed South African Coloured population.

- Co-first author
- Consulting on phasing, imputation and quality assessment
- Analysed accuracy of imputation
- Writing manuscript

Chapter 6: X-linked trans-ethnic meta-analysis reveals Tuberculosis susceptibility variants.

- First author
- Analysis and interpretation of data
- Writing manuscript

For all chapter the candidates and co-authors contributions are listed in Addendum A and B. For all chapters the candidate did the majority of the work and the extend of contribution was at least 90%, except for chapter 5 where the candidate was a co-first author and did 50% of the work.

Abstract

Approximately 25% of the world's population is infected with *Mycobacterium tuberculosis* (*M.tuberculosis*). Progression to active tuberculosis (TB) is influenced by the infecting strain of *M.tuberculosis*, the environment and the genetic makeup of the host. Globally, the incidence rate for TB in males is nearly twice as high compared to females, indicating that biological sex of an individual also contributes to TB susceptibility. While environmental factors and sex hormones influence the immune system and affect the male bias, they do not fully account for it. This suggests that the X chromosome and the unique biology regulating X-linked gene expression in females could significantly influence progression to active TB.

The X chromosome contains nearly 200 genes that are involved in the immune system. This clearly links the X chromosome to both the innate and humoral immune response and could explain why females have a more robust immune response against infections. X-linked genes have also been implicated in TB susceptibility, but these have not been conclusively linked to disease. Population specific effects could further contribute to the impact of the X chromosome on disease progression especially for populations that experienced sex-biased admixture.

Here we investigated the five way admixed South African Coloured (SAC) population that has sex-specific genetic contributions from Bantu-speaking African, European, KhoeSan and South and East Asian populations. We showed that global ancestry inference could be used to detect the presence of sex-biased admixture and that this correlates with previous results indicating a KhoeSan female bias and a European and Bantu-speaking African male bias.

We used SAC genome-wide association (GWAS) data and analysed the autosomes and X chromosome in a sex-stratified and combined manner, revealing sex-specific effects on both the autosome and X chromosome. A genome-wide interaction analysis also revealed significant interactions highlighting the need for epistatic and sex-stratified analysis in complex diseases.

X chromosome data from the International Tuberculosis Host Genetic Consortium (ITHGC) was available to conduct a large trans-ethnic X-linked meta-analysis of TB susceptibility. The meta-analysis included imputed GWAS data from two Chinese, one Russian, a Ghanaian and Gambian and two SAC cohorts (23229 samples). We optimised imputation in our SAC data and showed that even diverse African populations can be imputed with great accuracy. The meta-analysis revealed novel X-linked genes associated with TB susceptibility. These genes were located in genomic regions on the X chromosome previously associated with TB susceptibility. Results from the meta-analysis also further confirmed the presence of both sex-specific and population specific effects.

Our work highlights the importance of not only conducting sex-stratified analysis to elucidate sex-specific effects, but also to plan the study accordingly. Due to the strong impact of population specific

effects, extremely large meta-analysis will be needed to fully elucidate global and population specific susceptibility variants. While the X chromosome has been mostly neglected in the past, tools for its analysis are now readily available. Our findings support the mandatory inclusion of the X chromosome in large-scale genetic studies.

Opsomming

Ongeveer 25% van die wêreld se bevolking is geïnfekteer met Mikobacterium tuberculosis. Die ontwikkeling van aktiewe tuberkulose (TB) word beïnvloed deur die bakterium, die omgewing en die genetiese komponent van die gasheer. Die siekte tas twee keer soveel mans as vroue aan, wat aandui dat biologiese geslag ook bydra tot TB vatbaarheid. Alhoewel beide omgewingsfaktore en geslagshormone die immuunstelsel sowel as die TB geslagsvooroordeel beïnvloed, is dit nie ten volle daarvoor verantwoordelik nie. Dit dui daarop dat die X-chromosoom en die unieke biologie wat X-gekoppelde geenuitdrukking in vroue reguleer aansienlik kan bydra tot die ontwikkeling van aktiewe TB.

Die X-chromosoom bevat byna 200 gene wat by die immuunstelsel betrokke is. Hierdie gene verbind die X-chromosoom aan beide die aangebore en humorale immuunrespons en kan verduidelik waarom vroue 'n sterker immuunrespons teen infeksies het. X-gekoppelde gene is ook betrek by TB vatbaarheid, maar dit is nie voldoende verbind aan die siekte nie. Bevolkingspesifieke effekte kan verder bydra tot die impak van die X-chromosoom op die ontwikkeling van die siekte, veral vir bevolkings waar genetiese vermenging met geslagsvooroordeel plaas gevind het.

Hierdie tesis ondersoek Suid-Afrikaanse individue (SAC) wat geslags-spesifieke genetiese vermenging het van Bantoe-sprekende Afrikane-, Europese, KhoeSan- en Suid- en Oos-Asiatiese bevolkings. Ons wys dat globale vermenging inferensie gebruik kan word om die teenwoordigheid van geslagsvooroordeel te bepaal en dat dit korreleer met vorige resultate wat dui op 'n vroulike KhoeSan vooroordeel en 'n manlike Europese en Bantoe-sprekende Afrika vooroordeel.

Ons gebruik SAC genoom-wye assosiasie (GWAS) data en ontleed die outosoom- en X-chromosoom op 'n geslags-gestratifiseerde en gekombineerde wyse, wat geslags-spesifieke effekte op beide die autosome en X-chromosoom openbaar. 'n Genoom-wye interaksie-analise het ook betekenisvolle interaksies aangedui wat die nut van epistatiese en geslags-gestratifiseerde analise in komplekse siektes beklemtoon.

X chromosoom data van die Internasionale Tuberkulose gasheer genetiese konsortium (ITHGC) was beskikbaar om 'n groot transetniese X-gekoppelde meta-analise van TB-vatbaarheid uit te voer. Die meta-analise het toegepaste GWAS data van twee Chinese, een Russiese, 'n Ghanese en Gambiese en twee SAC versamelings (23229 monsters) ingesluit. Ons het toerekening in ons SAC-data optimiseer en toon dat selfs diverse Afrika-bevolkings met groot akkuraatheid toegereken kan word. Die meta-analise onthul nuwe X-gekoppelde gene wat verband hou met TB-vatbaarheid. Hierdie gene word gevind in genomiese streke op die X-chromosoom wat voorheen geassosieer is met TB-vatbaarheid. Resultate van die meta-analise bevestig verder die teenwoordigheid van beide gespesifiseerde en populasie spesifieke effekte.

Ons werk beklemtoon die belangrikheid van nie net geslags-gestratifiseerde analise om geslags-spesifieke effekte te verduidelik nie, maar ook om die studie dienooreenkomstig te beplan. As gevolg van die sterk impak van bevolkingspesifieke effekte sal uiters groot meta-analise nodig wees om globale en populasie spesifieke vatbaarheidsvariante ten volle te verklaar. Alhoewel die X-chromosoom in die verlede meestal verwaarloos tydens genetiese analise, is die gereedskap daarvoor nou beskikbaar. Ons bevindinge ondersteun die verpligte insluiting van die X-chromosoom in grootskaalse genetiese studies.

Acknowledgements

I would like to thank the following people and institutions for their contribution to this work:

My PhD supervisor Dr. Marlo Möller, MSc supervisor Dr. Muneeb Salie and principle investigator Prof Eileen Hoal. When I started my honours in the TB Host Genetics group my intention was to only stay for the one year, so thank you for showing me the value of a PhD and convincing me to stay on. I have gained a tremendous amount of knowledge in the past few years and discovered my true passion, Computational biology. Also, my other co-supervisors, Prof. Gerard Tromp for all the help on statistics, programming and especially language. Dr. Craig Kinnear for just making the lab a fun place to be and always having good advice and feedback on hand.

Prof. Paul van Helden, our former Head of department thank you for always showing a keen interest in my work and progress thereof on almost a daily basis, your presence will be severely missed in the years to come. To our current head of department Prof Gerhard Walzl thank you for the academic and financial support.

The TB Host Genetics group and everyone else in the Magic lab, thank you for making it a fun and calm work environment where everyone felt welcome, accepted and appreciated. Also, to the members of the rest of the Division of Molecular Biology and Human Genetics, thank you for the support and feedback.

Our collaborators, Dr. Chris Gignoux and Dr. Genevieve Wojcik for all the help with the GWAS data. Dr. Brenna Henn and Dr. Meng Lin for allowing us access to the reference data for admixture analysis as well as helpful advice on population genetics in general. A special thank you also to Dr. Vivek Naranbhai for getting us access to the International Tuberculosis Host Genetics Consortium TB GWAS data, without which this PhD thesis would not have been possible.

I would like to acknowledge and thank the study participants for their contribution and participation. 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. This work was also supported by the National Research Foundation (NRF) of South Africa (grant number 93460) to Eileen Hoal. This work was also supported 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. A big thank you to the National Research Foundation for my personal funding throughout the years of my postgraduate studies. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NRF.

Finally, I would like to thank my family, friends and girlfriend for all the support and encouragement over the years, this would not have been possible without you.

Table of Contents

List of Abbreviations

1000G:	1000 Genomes Phase 3 data
95CI:	95% confidence interval
ACTRT1:	Actin related protein T1
AGR:	African Genome Resource
AGVP:	African Genome Variation Project
AIDS:	Acquired immune deficiency syndrome
ARMCX1:	Armadillo repeat containing X-linked 1
ARSF:	Arylsulfatase F
ASNS:	Asparagine synthetase
ATP2C1:	ATPase secretory pathway Ca2+ transporting 1
ATRX:	ATRX, chromatin remodeler
BCG:	Bacillus Calmette–Guérin
BRS3:	Bombesin receptor subtype 3
C5orf64:	Chromosome 5 open reading frame 64
CAAPA:	Consortium on Asthma among African ancestry populations in the Americas
CDC:	Centre for disease control and prevention
CFAP54:	Cilia and flagella associated protein 54
CI:	Confidence interval
CIITA:	Class II major histocompatibility complex transactivator
CNV:	Copy number variation
CXorf51B:	Chromosome X open reading frame 51B
CYBB:	Cytochrome b-245 beta polypeptide
DIAPH2:	Diaphanous related formin 2
DNA:	Deoxyribonucleic acid
DPF3:	Double PHD fingers 3
DROSHA:	Drosha ribonuclease III
ER:	Estrogen receptors
EPTB:	Extra pulmonary TB
FE:	Fixed effects
FRMPD4:	FERM and PDZ domain containing 4
FRMPD4:	FERM and PDZ domain containing 4

GABRA3:	Gamma-aminobutyric acid A receptor subunit alpha 3
GRAMD2B:	GRAM domain containing 2B
GWAS:	Genome-wide association study
HIV:	Human immunodeficiency virus
HRC:	Haplotype reference consortium
HUWE1:	HECT, UBA & WWE domain containing 1
HWE:	Hardy-Weinberg equilibrium
IFN-γ:	Interferon gamma
IL-10:	Interleukin 10
IL:	Interleukin
IL1RAPL1:	Interleukin 1 receptor accessory protein like 1
INH:	Isoniazid
ITHGC:	International Tuberculosis Host Genetics Consortium
L1:	Long interspersed nuclear elements
LD:	Linkage disequilibrium
LINC00400:	Long intergenic non-protein coding RNA 400
LINC02153:	Long intergenic non-protein coding RNA 2153
LINC02246:	Long intergenic non-protein coding RNA 2246
lncRNA:	long non-coding RNA
MAF:	Minor allele frequency
MDR:	Multidrug-resistant
MEGA:	Multi-ethnic genotyping array
MHC:	Major histocompatibility complex
MIR514A1:	MicroRNA 514a-1
miRNA:	micro RNA
MIS:	Michigan imputation server
mRNA:	messenger RNA
MSMD:	Mendelian susceptibility to mycobacterial diseases
MTND6P12:	MT-ND6 pseudogene 12
mtRNA:	Mitochondrial DNA
NADPH:	Nicotinamide adenine dinucleotide phosphate hydrogen
ncRNA:	Non-coding RNA
NCS1:	neuronal calcium sensor 1
NEMO:	NF-kB essential modulator
NF-kB:	Nuclear factor kappa-light-chain-enhancer of activated B cells

NGS:	Next generation sequencing
NHS:	NHS actin remodeling regulator
NTM:	Neurotrimin
OMIM:	Online Mendelian inheritance in man
OR:	Odds ratio
P_comb:	Combined p-value using Stouffers method
P_Diff:	P-value for sex-differentiation test
PAGE4:	Prostate associated gene 4
PAR:	Pseudoautosomal region
PAS:	Para-aminosalicylic acid
PBMC:	Peripheral blood mononuclear cell
PBWT:	Positional Burrows-Wheeler transformation
PCSK6:	Proprotein convertase subtilisin/kexin type 6
PGC:	Primordial germ cells
PRC:	Polycomb repressive complex
PRR:	Pattern recognition receptors
pTB:	Pulmonary Tuberculosis
PTC:	Premature termination codon
PTCHD1-AS:	PTCHD1 antisense RNA
QC:	Quality control
RE:	Random effects
RIF:	Rifampicin
RN7SKP120:	RNA, 7SK small nuclear pseudogene 120
RNA:	Ribonucleic acid
RNF125:	Ring finger protein 125
RNF126:	Ring finger protein 126
RNU6-974P:	RNA, U6 small nuclear 974, pseudogene
RTN4RL1:	Reticulon 4 receptor like 1
SAC:	South African Coloured
SALL2:	Spalt like transcription factor 2
SIS:	Sanger imputation server
SNP:	Single nucleotide polymorphism
SNP:	Single nucleotide polymorphism
SPANXN2:	SPANX family member N2
SRPX:	Sushi repeat containing protein X-linked
TB:	Tuberculosis

TB:	Tuberculosis
TBL1X:	Transducin beta like 1 X-linked
TENT4A:	Terminal nucleotidyltransferase 4A
Th1:	T helper cells
TLR:	Toll-like receptor
TLR:	Toll-like receptor
TNFS5:	encodes CD40 ligand DNA
TNFα:	Tumor necrosis factor alpha
TNFα:	Tumor necrosis factor alpha
TST:	Tuberkulin skin test
UPF3B:	UPF3B, regulator of nonsense mediated mRNA decay
URI1:	URI1, prefoldin like chaperone
VCF:	Variant call format
WHO:	World Health Organisation
WT1:	Wilms tumor 1
Xce:	X controlling element
XCI:	X chromosome inactivation
XDR:	Extensively drug-resistant
XIC:	X inactivation centre
Xist:	X inactivation specific transcript
Xite:	X-inactivation intergenic transcription element
XR:	X-linked recessive
XWAS:	X chromosome wide association study

List of Figures

List of supplementary Figures

List of Tables

List of supplementary Tables

1 General Introduction

1.1 Brief history of Tuberculosis

Mycobacterium tuberculosis (*M.tuberculosis*), the causative agent of tuberculosis (TB), interacted with the human host for most of human history, during which *M. tuberculosis* has perfectly adapted to its environment (1). Early evidence of *M. tuberculosis* infection was found in skeletal remains from the Iron Age (400-230 BC) and exhumed mummies from Egypt (2–5). At first, TB was thought to be a hereditary disease as it seemed to cluster within families, but in 1699 Gaspart Laurant Bale discovered that tubercles were the cause of TB and for the first time the infectious nature of TB was proposed (6,7). Soon after this discovery Europe was hit by the “Great White Plague”, the biggest TB epidemic in history which lasted over 200 years from the early 18th century to the late 19th century (8). During this time advances were made in understanding the disease. In 1720 Benjamin Marten expanded on the discovery by Gaspart Laurant and proposed that TB is caused by minute living creatures residing in the tubercles, which can be transmitted through close contact (9). In 1839 Johan Lucas Schönlein termed the disease Tuberculosis, after the tubercles that caused it (8). In 1882, Robert Koch identified *M. tuberculosis* as the TB-causing agent. (10–12). Koch also identified the tuberculin compound, which led to the development of the Tuberculin Skin Test (TST) that is used as a proxy for *M. tuberculosis* infection, but not active disease (13–16). The discovery of X-ray technology by Wilhelm Conrad von Röntgen in 1895 allowed for an alternative detection tool for detecting active TB, a tool still in use (17).

Following the discovery of these detection methods, further progress was made against the disease in the 20th century. In 1908 Albert Calmette and Camille Guérin created the *Bacille Calmette-Guerin* (BCG) strain from an attenuated strain of *M. tuberculosis* (10,18). The BCG vaccine against *M. tuberculosis* was first administered to humans in 1921 and administered on a large scale to children in Europe after the second world war (1945-1948) (19). Meta-analysis on the efficacy of BCG vaccination revealed a protective efficacy of 19% against infection and 58% protection against developing active TB, but only in children (20). During the second world war (1943) the first antibiotic, streptomycin, was developed which was soon followed by the development of the second anti-TB drug, para-aminosalicylic acid (PAS) in 1947 (21–23). Then in 1960, following the discovery of isoniazid (INH) in 1952, TB was considered a 100% curable disease, a thought that persisted until the first outbreak of drug resistant TB in the United States of America (USA) in the 1970’s (8,22,24). Drug resistant *M. tuberculosis,* remains to this day the biggest hurdle to eradicating the disease (25). A genetic nomenclature of scientific terms and concepts used in this thesis is given in Table 1.1 below.

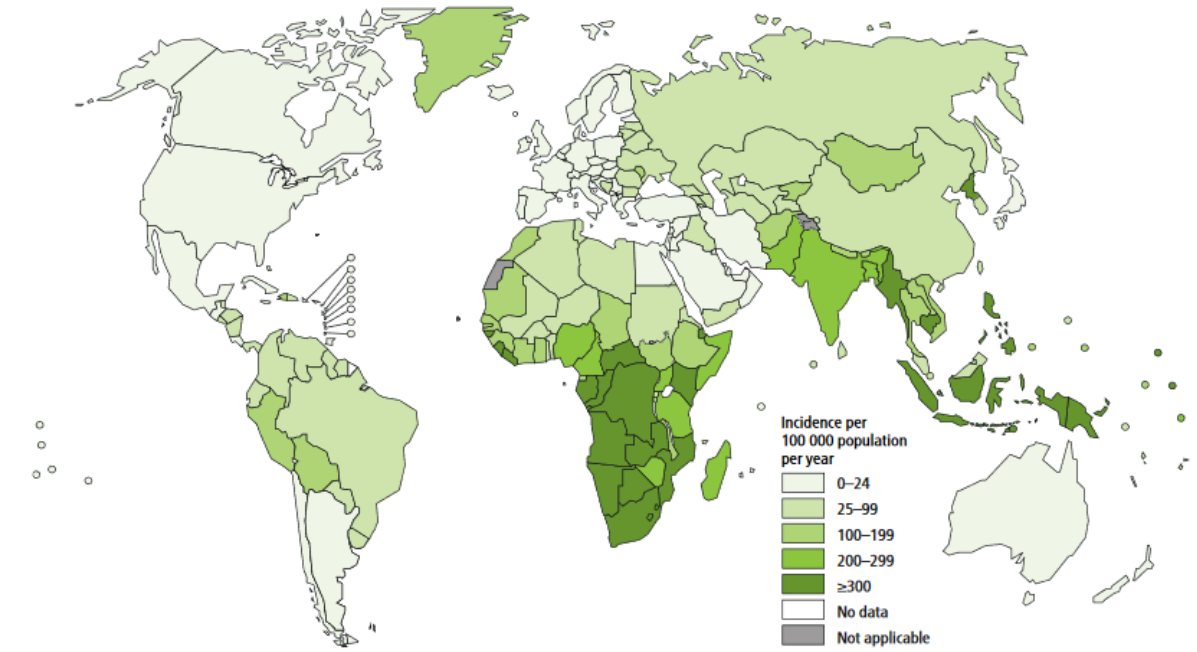
Table 1.1: Genetic nomenclature

Term	Definition
95% Confidence interval	Range of values so defined that there is a specified probability that the value of a parameter lies within it. (95%CI)
Allele	Refers to a particular form occurring at a locus. In humans, each individual will have a maximum of two forms, since it is possible that their parents differed from each other at this locus and that the individual inherited a different form from each parent.
Allelic association	Statistical analysis to determine whether disease is associated with particular allelic variants at one or more loci, usually by comparing marker allele frequencies between a disease group and a control group.
Association by LD	Association with a marker in strong LD with the causal locus. Also known as indirect association.
Bias	Inclination or prejudice for or against one person or group.
Candidate gene	Genetic association studies that focus on associations between genetic variation within pre-specified genes of interest and phenotypes or disease states.
Effective population size	The number of individuals in a population who contribute offspring to the next generation.
False positive associations	Association due to confounding by stratification between cases and controls. Also known as confounded association.
Fine-mapping	Determine the genetic variant (or variants) responsible for complex traits, given evidence of an association of a genomic region with a trait.
Genotype	A description of the two alleles at a given locus.
Genome-wide association study	Observational study of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait. (GWAS)
Haplotype	A section of DNA on a single chromosome where certain alleles of different markers tend to be inherited as a unit.
Heterogeneous	There are two types of genetic heterogeneity: allelic heterogeneity, which occurs when a similar phenotype is produced by different alleles within the same gene; and locus heterogeneity, which occurs when a similar phenotype is produced by mutations at different loci.
Heterozygosity	A measure of genetic diversity within a population.
Hardy-Weinberg equilibrium	The Hardy–Weinberg principle, also known as the Hardy–Weinberg equilibrium, model, theorem, or law, states that allele and genotype frequencies in a population will remain constant from generation to generation in the absence of other evolutionary influences. (HWE)
Imputation	Statistical inference of unobserved genotypes.
Lineage	Direct descent from an ancestor; ancestry or pedigree.
Linkage analysis	Statistical analysis to localise genes and markers with respect to each other in the genome, based on recombination frequency. Linkage analysis can also be used to map a disease phenotype in relation to polymorphic markers.
Linkage disequilibrium	The non-random association of alleles at different loci. (LD)
Locus	A specific position in DNA.
Minor allele frequency	Frequency at which the second most common allele occurs in a given population. (MAF)
Missingness	Degree of missing data from a set.
Odds ratio	How strongly the presence or absence of an allele associated with the presence or absence of disease in a given population. (OR)
Pseudo autosomal region	Regions of the X and Y chromosome that pair and recombine during meiosis. (PAR)

Term	Definition
Phasing	Phasing, or haplotype estimation, refers to the statistical estimation of haplotypes from genotype data.
Polymorphism	Variations in DNA that originated during evolution as a result of mutations. Indicates that a locus has more than one form in the population. Should occur with a frequency of greater than 1% in the population to be classified as such.
Population admixture	The recent combination of two or more previously distinct populations. Unknown admixture may result in spurious genetic associations in a case-control study design.
Recombination	The rearrangement of genetic material, especially by crossing over in chromosomes or by the artificial joining of segments of DNA from different organisms.
Sex	The anatomy of an individual's reproductive system, and secondary sex characteristics.
Sex-specific	An effect that is localised to only one sex.
Sex-stratified	Splitting by sex and separating males and females.
Single Nucleotide Polymorphism	A variation type in DNA where a single nucleotide has more than one allele in a population. Coding SNPs are found in genes and non-coding SNPs are present in promoters, introns, or intergenic regions. Synonymous SNPs will not result in an amino acid change in the protein, while non-synonymous SNPs will lead to a change of the amino acid at that position. (SNPs)
SNP/Genotyping microarray	Array technology that allows large numbers (up to a million) of SNPs to be genotyped on a single array typically the size of a microscope slide.
True association	Association with the true causal variant. Also known as direct association.
X chromosome inactivation	The process through which one X chromosome is inactivated in females. (XCI)
X inactivation centre	Locus on the X chromosome responsible for regulating XCI. (XIC)
X chromosome wide association study	Similar to GWAS except that XWAS specifically targets the X chromosome genetic variants within GWAS data. (XWAS)

1.2 Epidemiology of tuberculosis

In 1993 the World Health Organisation (WHO) estimated that approximately 25% of the world's population was latently infected with *M. tuberculosis* and declared TB a global health problem (26). While TB is the leading cause of death by a single infectious agent worldwide, the TB incidence has declined globally by approximately 1.4% between 2000 and 2016 (27). TB deaths have decreased by 37% since the year 2000 and notification rates are slowly decreasing. While these are inspiring statistics, the incidence rate will need to reduce by a further 3-4% a year if the WHO goals for combating TB and reducing the number of deaths by 75% (compared to the 2015 death rate) are to be met by 2025 (27). While TB is mostly under control in developed countries it is still a serious problem for developing countries (28–30).



.

Figure 1.1: Global TB incidence rates for 2016 *(27)*.

South Africa has an extremely high TB burden at an incidence rate of 781 per 100 000 individuals (Figure 1.1) (27). The high TB prevalence in South Africa is further compounded by the high HIV prevalence (12.6%), since *M. tuberculosis* takes advantage of the compromised immune system of HIV positive individuals (31,32). Drug resistant TB, often the result of poor adherence to treatment, further exacerbates the disease burden.

1.3 M. tuberculosis, transmission, infection and symptoms

M. tuberculosis is a rod shaped, aerobic, non-spore forming, acid-fast bacilli, with a thick lipid cell wall (33). This thick lipid wall, consisting of mycolic acid (fatty acid) and arabinogalactan (peptidoglycan bound polysaccharide), acts as a barrier and a key component of *M. tuberculosis* success as it increases virulence and aids in developing drug resistance, evading the host immune response and surviving in macrophages (34).

Spread of *M. tuberculosis* and transmission of the disease is due to small airborne droplet nuclei that contain viable bacteria (35). Sneezing, coughing or talking in close proximity can disperse these droplet nuclei from an infected individual to the surroundings (Figure 1.2) (36,37). Upon inhalation of these droplets the hosts cell mediated (innate) immune response is triggered to fend of the invading *M. tuberculosis* bacilli (38). Infection begins when the *M. tuberculosis* bacilli are phagocytosed by alveolar macrophages or dendritic cells, where they will either be destroyed or survive depending on the virulence of the infecting strain and the efficacy of the host immune response (Figure 1.2) (38,39).

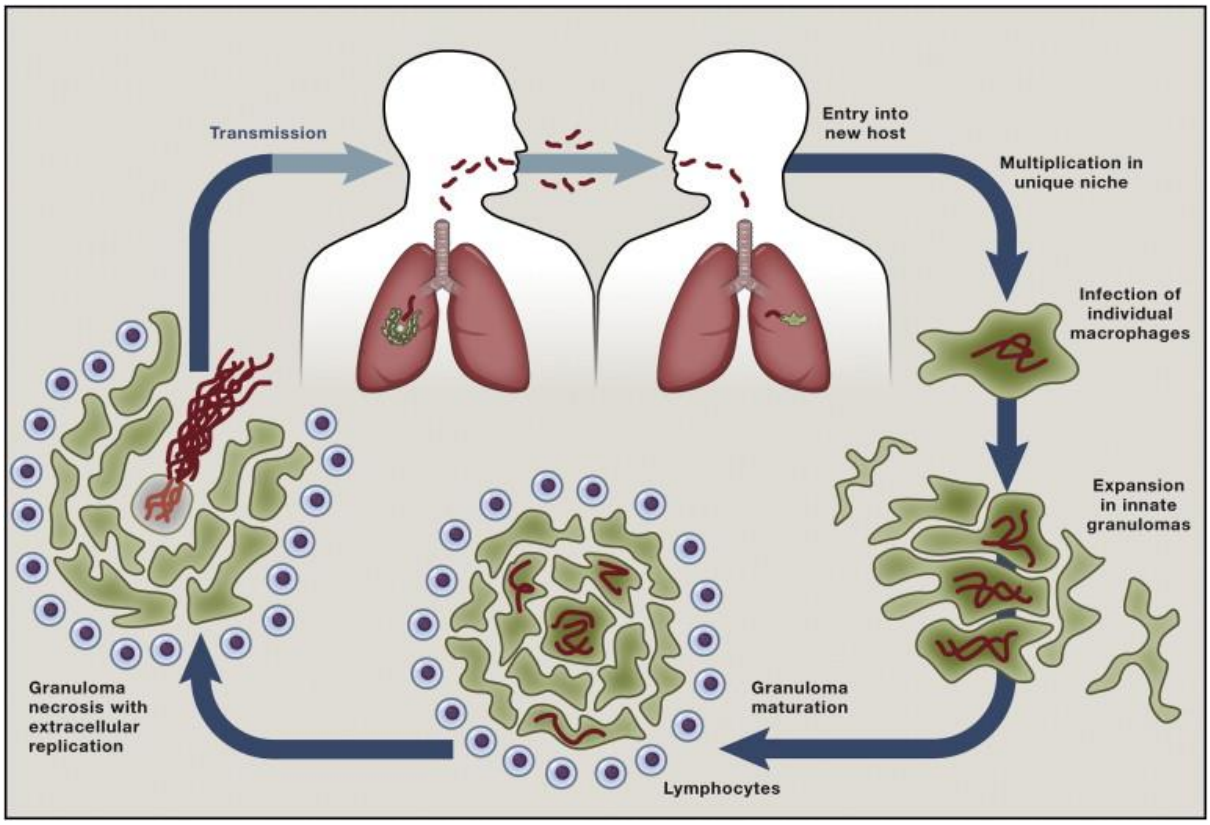


Figure 1.2: Mechanisms by which TB is transmitted between individuals *(40)*.

Failure to kill the bacilli results in the formation of granulomas, a collection of macrophages, T and B lymphocytes and fibroblasts, forming a structure designed to stop the spread of the *M. tuberculosis* (Figure 1.2). Following formation of the granuloma the T-lymphocytes secrete cytokines to promote the killing of *M. tuberculosis.* Here the bacilli can be cleared, or if the immune system is unable to clear it the bacilli can enter a seemingly dormant state (latent infection) (38). However, if the immune system cannot control the infection, due to a compromised immune system as a result of HIV coinfection for example, then the bacteria will multiply and cause active disease.

If the *M. tuberculosis* infection is contained in the lungs it is referred to as pulmonary TB (PTB), but the granulomas can also rupture (Figure 1.2), and the bacteria can spread to and infect other parts of the body resulting in extra pulmonary TB (EPTB). Symptoms of specifically PTB include chronic and persistent coughing, loss of appetite and weight, fever, night sweats and haemoptysis (41), while EPTB presents with a great variety of symptoms and is difficult to diagnose (42).

1.4 Diagnosis and treatment

Once an individual presents with TB symptoms quick and accurate diagnosis is vital for ensuring the best treatment outcome. Several techniques to diagnose active TB are available with varying degrees of sensitivity and accuracy (41). Sputum microscopy, liquid or solid media culturing, chest X-ray, histopathological biopsy, and nucleic acid amplification test are some examples. There are also more

rapid tests such as the GeneXpert assay that can detect *M. tuberculosis* in a sample and the updated version of the GeneXpert can even detect Rifampicin (RIF) resistant *M. tuberculosis* strains (43).

Once diagnosed, an appropriate treatment regimen must be set in place which can last at least six months and can have serious side effects (44,45). The first line treatment against drug susceptible *M. tuberculosis* consists of a four-drug regime including INH, RIF, pyrazinamide and ethambutol and can have high cure rates (45,46). However, due to the long treatment time and side effects of the drugs, treatment adherence and completion is a serious problem which has led to the development of drug resistant *M. tuberculosis* (47).

1.5 Tuberculosis is a complex disease

Upon infection, progression to active TB depends on several factors including the strain of infecting *M. tuberculosis,* the host's environment and behaviour and finally the genetic makeup of the host (48). An individual's socio-economic standing can influence TB susceptibility, as individuals living in overcrowded conditions are more likely to contract TB (48). Nutrition is also important as inadequate nutrition can alter the immune response and increase risk of developing TB (49–51). Finally, multiple meta-analysis and systematic reviews show that smoking and alcohol consumption considerably increase TB susceptibility (52–57).

While these behavioural and environmental factors do influence TB susceptibility there is also very clear evidence that host genetics play a significant role in TB susceptibility. The first piece of evidence is that only 5-10% of infected individuals globally develop active TB. An example supporting the genetic contribution is the 1926 Lübeck incident in Germany, where 251 new-borns were vaccinated with live *M. tuberculosis* instead of BCG. Of the new-borns seventy-seven died, 47 developed latent TB and 127 got sick but recovered (58). This event demonstrates how individual immune responses differ in efficacy, demonstrating the importance of host genetics which underlies the immune system.

Several twin studies have also shown that monozygotic twins have higher concordance for the disease than dizygotic twins (59–61). Moreover, studies showed of adopted children showed that they were more likely to die from TB disease if their biological parents had died from it, clearly indicating a genetic component to TB susceptibility and that environmental factors are not the only influential factors (62–65). Population specific effects have been observed for disease susceptibility and are also indicators of the role of host genetics. Europeans have been challenged by TB since before the 18th century and thus, through a process of natural selection, host genetic markers may have been selected that made the European population less susceptible to TB (64,66). Subsequently, when the Europeans brought TB to Africa through colonialization the bacteria were spread to African populations who had not been challenged by European TB strains before and thus had not built up any genetic resistance, making them more susceptible to TB (1,64,67,68).

Since the discovery that host genetics influences TB susceptibility several methods have been used to determine which genes or genetic loci are associated with disease susceptibility. These approaches included family and linkage studies to find areas of the genome that segregate more often in affected individuals compared to control individuals (69,70). Heritability analysis to determine if TB susceptibility is a heritable trait and which genetic loci are responsible for this have been conducted and suggest a clear heritable component in TB susceptibility (71–73). Animal studies have also been used to prove the genetic contribution to TB susceptibility and homologous susceptibility genes have been mapped in humans (6,74). Finally, a number of single nucleotide polymorphism (SNP)- based association studies have been conducted in the form of candidate gene association studies as well as genome wide association studies (GWAS) to identify SNPs and genes associated with TB susceptibility (1).

1.6 Association studies

The human genome consists of approximately 3.2 billion base pairs. Adenine (A), thymine (T), guanine (G) and cytosine (C) nucleotides are linked in a specific sequence, creating the blueprint for life. The vast majority of these genomic sites (99%) are identical with all humans carrying the same base residue at both chromosomal homologs (75). The remainder of the nucleotides can vary within the population and explain much of the diversity observed in humans. Approximately 10 to 15 million of these variable base residues or single SNPs have been identified in humans. SNPs are used as biological markers to detect genes that are associated with disease.(76,77).

Initially, candidate gene association studies were done to correlate SNPs with a particular phenotype or disease (77). Based on prior knowledge of biology and pathogenicity of a disease, candidate genes were identified and SNPs within these genes were tested to determine their association with a specific trait or disease. While this is an attractive study design, candidate gene associations studies have had limited success, with results often not being replicated in independent studies and associations being generally weak (77). The biggest limitation of candidate studies is that they require prior knowledge of disease pathogenesis to identify genes of interest and thus require a hypothesis. However, this hypothesis is limited to one gene and does not consider the rest of the genome and its potential interaction with the gene of interest.

Genome-wide association studies (GWAS) managed to bypass this limitation as advances in technology allowed for simultaneous genotyping of thousands of SNPs across the entire genome. This advantage of SNP microarrays meant that prior knowledge of gene or pathogen function was no longer required. Studies are thus not limited to the effect of just one gene and could analyse genome wide genetic contributions to a single trait or phenotype without a hypothesis, i.e. hypothesis free testing. This shift towards a hypothesis free data driven science allows for an unbiased analysis of the variation across the genome at a fraction of the time or cost of candidate gene studies (78).

The goal of GWAS is to detect associations between variants and traits in samples from the population, with the aim of understanding which genes are involved in a disease or will lead to improved treatment or preventative strategies (79). The potential for GWAS to succeed depends on various factors including how many loci affecting a certain trait segregate in the population, the sample size and power required to detect associations, the joint distribution of effect sizes and minor allele frequencies (MAF) of the SNPs and the panel of SNPs used to genotype the population (79). Correct phenotyping is also of vital importance as variations in the phenotype can introduce heterogeneity, which will reduce the power to detect informative associations and could lead to spurious results.

In order to conduct informative GWAS studies SNP microarrays have been developed that can genotype between 100 000 and 2.5 million variants simultaneously. Furthermore, as even the densest genotyping array captures only a fraction of the variation in the human genome, SNP arrays were designed in such a way that SNPs not directly genotyped can be inferred based on the pattern of available genotypes (79). This process, called imputation, exploits the phenomenon of linkage disequilibrium (LD) and is based on the correlation between SNPs in the current human genome (79). This correlation is a result of historical evolutionary forces and when alleles at two or more loci appear together more often than would be expected by chance then these variants are said the be in LD (75). Two SNPs in high LD can thus serve as proxies for one another and due to the correlation, genotyping one can give almost complete info on the other SNPs genotype.

LD patterns need to be elucidated by mathematically, quantifying them in fully sequenced or densely genotyped reference populations (75). By comparing regions of LD between an appropriate reference population and the study population missing genotypes can be inferred into the study data (80). This increases the number of variants available for testing in GWAS, which in turn increases the likelihood of detecting significant associations. Datasets for imputation are now widely available and represent a diverse set of populations with dense genotyping, perfectly suited for imputation.

These reference datasets include amongst others, the 1000 Genomes phase 3 data (1000G) (81), the Human Genome Diversity Project (82), Haplotype Reference Consortium (HRC) (83) and the HapMap consortium (84). These reference datasets at first, as with the SNP arrays, were tailored for European populations making them less than ideal for studying or imputing more diverse or admixed populations like Africans, Hispanics, or in our case the 5-way admixed South African Coloured (SAC) population (85–87). More recently however increased number of individuals from more diverse populations were added to current reference datasets and additional databases were also initiated that focused more specifically on African populations, which were greatly underrepresented. The Consortium on Asthma among African ancestry populations in the Americas (CAAPA) (88), the African Genome variation project (AGVP) (89) and the African Genome Resource[1] (AGR, not publicly

[1] https://imputation.sanger.ac.uk/

available) are three resources which have recently become a viable option for accurate imputation of African populations.

Another advantage of LD and imputation is that it allows us to conduct meta-analysis by maximising the SNP overlap between different GWAS datasets (90). Many of the available SNP arrays genotype for different markers and at varying density across the genome and thus overlap between different genotyping platforms can vary significantly (91). Meta-analysis increases the power to detect associations by increasing the sample size through the combination of different studies on the same trait or disease (77). By increasing the number of samples and overlapping variants, the power of meta-analysis is further amplified and can allow for identification of population specific and globally associated variants.

However, LD and imputation does have its drawbacks. Increasing the number of variants increases the multiple test correction burden. By adjusting the significance threshold based on the number of independent tests performed, the number of false positive associations can be controlled (92). This multiple testing burden can be further compounded by small sample sizes and low power, making it impossible to detect small effects or link rare variants with a trait or disease. Furthermore, due to LD, spurious associations can survive multiple test correction if the associated variant is not actually the causal one, but merely linked to the causal variant (75). This is especially true for rare variants, which are often in LD with other non-causal rare variants and a single causative rare locus could be linked with multiple false positive associations (93). Teasing out which of the two variants in LD is the actual causal variant is a process called fine mapping and can be very difficult to determine through GWAS alone (93). To resolve this, further investigation is required in a population with lower LD, in which the causal variant might be identified, or functional studies can be conducted on the effect of the variant *in vitro* or *in vivo* (94).

The effects of LD and sample size are two of the limitations of GWAS. Other limitations include statistical and computational issues e.g. analysing and storing large amounts of data (75,76). As the number of markers on the array increased so did the statistical challenge of analysing them. Also, many traits or phenotypes are polygenic and associated with many alleles of small effects that collectively contribute to the phenotype (93). Furthermore, there are also interactions between different loci in the genome (epistasis) and between genes and the environment, both of which can influence the power of a GWAS (77).

The design of SNP arrays also requires prior knowledge of the genome in order to choose the most informative markers for genotyping (95). Knowledge of these markers was limited when GWAS were introduced and could have, at least initially, limited the applicability of chosen markers. While these limitations were daunting, instead of limiting SNP microarray technologies they instead spurred advances in other areas of research. New computational and statistical methods were developed to

clean and analyse these large datasets. For example, the concept of determining false discovery rate (FDR) was introduced in order to lower the impact of conventional multiple test correction and methods to determine how accurate statistical testing was were also introduced in the form of quantile-quantile (QQ) plots and Manhattan plots (93).

Improved analysis techniques also led to an increased number of SNPs associated with a certain trait or disease, which in turn led to research to understand the link between these variants and the phenotypic effect. Elucidating biological functions allowed for more relevant markers to be added to SNP arrays resulting in a cycle of progress where progress in one field (SNP arrays) fuelled progress in the other two (computational methods and biological understanding), resulting in advances that would not have been possible independently (75). By dealing with these limitations instead of succumbing to them, GWAS has found many applications in many research areas and contexts and experienced great success over the past decade. GWAS has mapped causal variants (96), LD patterns (97), introduced phasing (98) and inferred demography, ancestry and evolution (85,99–103).

Despite these successes there is one area of research where GWAS has made a significant impact but has not yet delivered what was expected of it and that is the analysis of complex diseases. Complex diseases are acquired diseases where susceptibility depends on both the environment, behaviour and hereditary factors and does not follow a classic pattern of inheritance (77). Instead complex disease susceptibility is more likely a result of a combined effect of many common variants with small effects sizes (polygenic) as well as environmental factors and gene-gene interactions (77,93). Omnigenic effects could also influence susceptibility to complex diseases, but the omnigenic model has not been fully explored and there are some concerns regarding its validity and thus it will not be considered here (104,105). While the environment and behaviour play a large role, there is a proven genetic component in many complex diseases and by choosing samples carefully (in the case of a case control association study) the environmental and behavioural impact can be controlled for, allowing us to focus on the genetic factor.

Sample size and multiple test correction is perhaps the largest stumbling block for GWAS in complex diseases as multiple genes with small effects will require tremendous power to identify and could be lost due to multiple test correction. Increasing the sample size through conducting meta-analysis would thus greatly benefit the analysis of complex diseases by increasing the power. Furthermore, many genes identified in GWAS for complex diseases have no known biological effect, due to incomplete knowledge of gene function and thus their role in disease susceptibility cannot be elucidated (77). Despite these limitations GWAS in complex diseases, such as inflammatory bowel disease, diabetes and Alzheimers, have successfully identified novel pathogenic mechanisms and therapeutic targets (106,107).

Another area of success of GWAS in complex disorders and traits is its relationship to the concept of missing heritability. Heritability is the proportion of phenotypic variance that can be explained by genetic variance and missing heritability is thus observed heritability that is not, or not yet explained by genetic variance (108). GWAS for height and susceptibility to schizophrenia for example elucidated a great portion of the missing heritability. In 2009, 40 SNPs were significantly associated with height explaining 5% of heritability, by 2014 the number of significant associations increased to roughly 700 SNPs, explaining about 20% of heritability (109,110). Similarly, the number of variants significantly associated with schizophrenia increased from one SNP in 2009 to 108 SNPs in 2014 (111,112).

Elucidating the missing heritability and identifying all loci that contribute towards it, is both a function of genome coverage (how much genetic diversity can be captured) and sample size. As the sample size and genomic coverage of arrays increases more and more variants will be identified explaining an ever-increasing amount of the missing heritability, but as the effects of these variants are likely to get smaller the sample sizes will constantly need to increase in order to capture all variation in the genome (79). Here imputation and meta-analysis can be of great benefit to increase the genomic coverage and sample size and power to detect associations in order to fully elucidate complex diseases.

1.7 GWAS and the X chromosome

Since the introduction of GWAS, the sample sizes have been steadily increasing and so is the amount of data available for imputation and meta-analysis. The computational methods and statistics to deal with this ever-increasing amount of data has also improved over time allowing for better quality and more informative analysis. However, one aspect of GWAS and its influence on traits, phenotypes and diseases has until very recently been ignored: the analysis of the X chromosome and the impact of X-linked genes on a trait or disease and its potential contribution to the missing heritability (113).

Most genotyping array contain probes for X-linked variants, making this data readily available, yet it has been consistently excluded from most GWAS for several reasons. Firstly, compared to the autosome the X chromosome has fewer markers and suffers from lower genotyping accuracy. Secondly, X-linked variants are unique as they are diploid for females and haploid for males, which pose a statistical challenge. Comparing or combining diploid and haploid markers for association analysis is challenging and certain statistical techniques, such as the Hardy–Weinberg equilibrium testing (HWE), cannot be calculated for haploid loci (113). This means that analysing X chromosome genotyping data cannot be done in the same way that autosomal genotyping data is analysed.

The X chromosome requires separate quality control and analysing in a sex-stratified manner (114,115). Stratifying the data by sex pose a problem as it reduces the sample size and thus power to detect associations. This, combined with the fact that male haploid genotypes further lower the power and the number of X-linked markers is low, means that many studies are not powerful enough

to detect X-linked loci (113). Also, the special analysis techniques required might be too daunting for researches with limited specialised knowledge in bioinformatics and statistics, requiring someone with expertise in these fields. This statistical complexity is further compounded by the fact that in females one X chromosome is randomly inactivated in each cell of the body to equalise dosage of gene expression between males and females. This process of X chromosome inactivation (XCI) and associated processes such as genes that escape inactivation and skewed inactivation will further influence phenotype and complexity of X-linked analysis, especially as the processes of XCI are not yet fully understood (113,116,117).

Although the X chromosome has been ignored in the past due to its statistical and biological complexity it is clearly involved in diseases susceptibility. The Online Mendelian inheritance in man (OMIM) catalogue of human genes and genetic disorders suggests that about 7% of phenotypes with known molecular basis (autoimmune disorders, cognitive and behavioural conditions) are caused by X-linked genes (113). Furthermore, the X chromosome contains approximately 5% of the genes in human genome, and many of these have been shown to be involved in the immune function (118,119). The X chromosome also contains the highest density of regulatory miRNA molecules, providing further evidence for the involvement of the X chromosome in biological functions and possibly disease susceptibility (118).

Even though evidence for the involvement of the X chromosome in immune functions is strong and tools to impute and analyse X-linked variants have improved it is still being ignored in most GWAS studies (114,115,120,121). This is evident when looking at the records of published GWAS. To date there are 3420 publications, in which 62652 unique SNP-trait associations have been identified are recorded in the online GWAS catalogue[2]. Of these associations only 385 SNPs (0.6%) were X-linked and only 157 (0.25%) of them reached genome wide significance (p-value < $5e^{-8}$) (122). This highlights the extent to which the X chromosome has been ignored in the past and it is vital for this to change.

If most of the GWAS published to date had X chromosome data but excluded it from analysis, then there is a tremendous amount of unexplored data that is already available for analysis. As tools to analyse the X chromosome are now available there is no reason not to explore it. Only by starting to include the X chromosome in GWAS analysis will its impact on immune functions and sex-bias be elucidated. Identifying novel X-linked variants will spur research and advances in functional annotation of X-linked genes, increasing understanding of biological function and thus enable us to increase coverage of X chromosomal markers on SNP arrays. As use of X-linked analysis tools increase they too will improve resulting in improved data which could elucidate sex specific immune responses and explain a portion of missing heritability (113). Analysis of the involvement of the X

[2] https://www.ebi.ac.uk/gwas/

chromosome is especially important for complex diseases that present with a strong sex-bias, such as TB, discussed below.

1.8 Tuberculosis and the X chromosome

Tuberculosis presents with a strong male sex-bias and globally the incidence rate is nearly twice as high in males compared to females (Figure 1.3) (27). Yet, while multiple candidate gene association studies and GWAS have been done for TB susceptibility in diverse populations, none specifically investigated the X chromosome or any alternative cause for the sex-bias (72,85,123–131). Some candidate gene studies have identified X-linked genes associated with TB susceptibility such as *Toll-like receptor 8* (*TLR8)* and some GWAS studies analysed X-linked variants (96,123,126,132–138). The fact that females have a more robust immune response towards infections indicates that X-linked variants could easily influence this male bias (139–141).

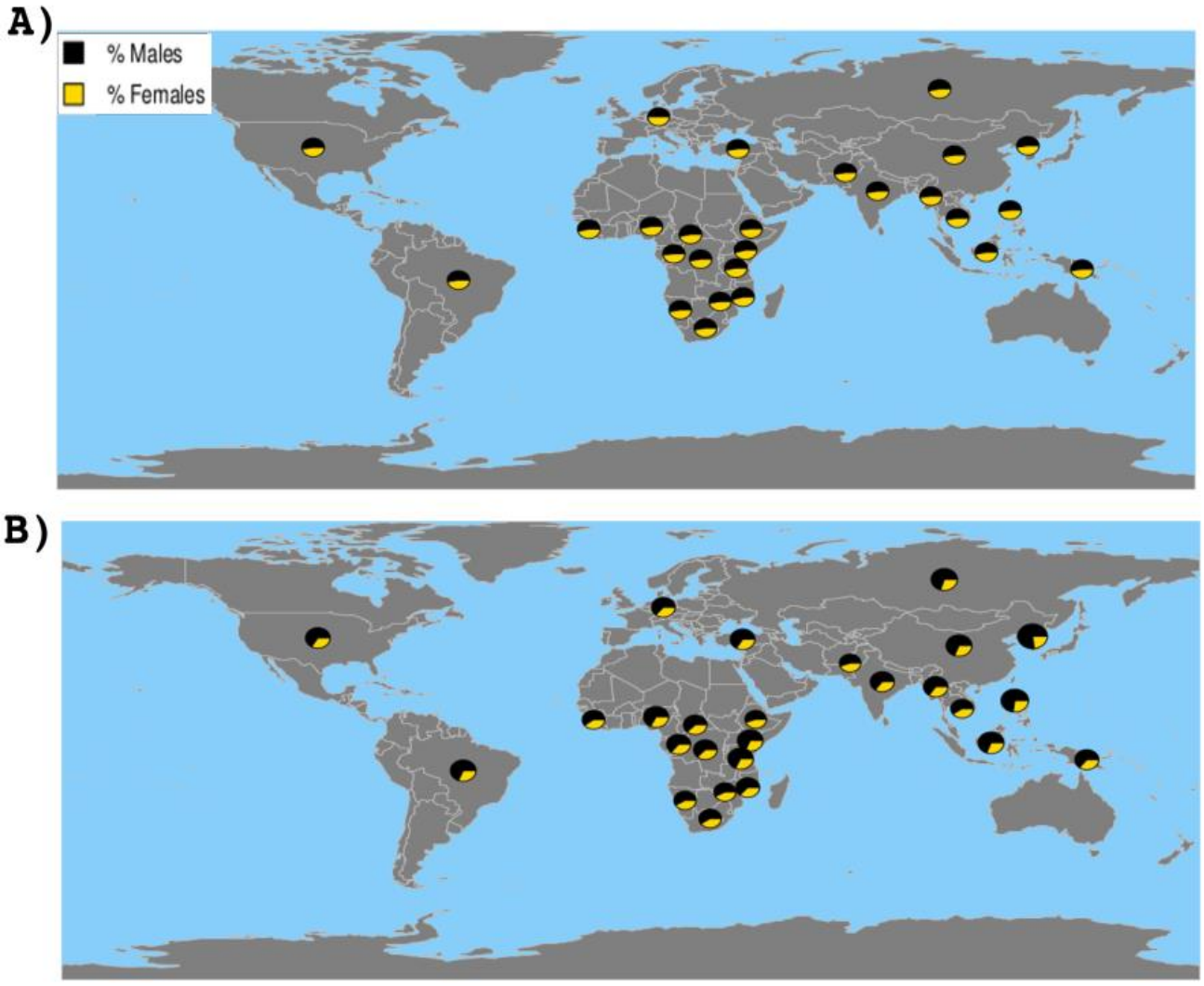


Figure 1.3: Worldwide male to female ratios of TB incidence for children under the age of 14 (A) and children and adults over the age of 14 (B). World map image obtained from the R packages 'rworldmap' and 'rworldxtra' and the data was obtained from the 2017 WHO TB report (27,142).

Figure 1.3A shows TB incidences in children under the age of 14 (pre-pubertal) for 29 high burden countries as well as the European Union and the United States of America. While incidence rates in some countries were the same between males and females, most have a slight male bias and no country had a female bias for incidence rates in young children (Figure 1.3A). Comparing these incidence rates to those from individuals over the age of 14 (Figure 1.3B) results in a clear increase in the male sex-bias across all countries, even those that initially had equal incidence rates between the sexes (Figure 1.3B).

Since the male bias is present before puberty, but then significantly increases with age (Figure 1.3), indicates that sex hormones are likely to have a definite impact on this sex-bias. Furthermore, the fact that this male sex-bias presents globally (Figure 1.3) suggests that the bias can also not be fully explained by environmental, behavioural and socioeconomic factors or the strain of *M. tuberculosis,* although they do influence the sex-bias (140,143,144).

Another influential factor in TB susceptibility is the influence of population specific effects (99,145). The effect of ethnicity is especially important when considering admixed populations as not correcting for ancestral components in statistical analysis can confound results (99,138). Furthermore, admixture events can be sex-biased with respect to the male to female ratio of founder contributions and this sex-bias in admixture leads to different ancestral distribution on the autosome compared to the X chromosome (146,147). These differences in ancestral distribution could further influence the sex-bias of TB disease through population specific effects amplified by certain ancestral components on the X chromosome of admixed individuals.

Based on the evidence that the X chromosome and X-linked genes are involved in TB susceptibility and potentially influence the male bias it presents with, which could be further amplified by sex-bias admixture events, leads us to hypothesise that:

- X-linked genes and their genetic interactions and functional mechanisms will elucidate the sex-bias of TB disease, which could be further compounded by sex-biased admixture events.
    - **Aims and Objectives:**
        1. Test for the presence of sex-biased admixture in the SAC population.
            i. Infer global ancestry on the autosome and X chromosome separately and test for significant differences in their distribution.
            ii. Compare to previous results to confirm the viability of using global admixture components to infer sex-biased admixture.
        2. To identify TB susceptibility loci responsible for the sex-bias observed in TB susceptibility by conducting a TB GWAS and meta-analysis including data from multiple ethnic backgrounds.

i. Do association testing of the X chromosome in all the individual TB GWAS datasets
ii. Determine the accuracy and optimise imputation in the admixed SAC population.
iii. Impute X chromosome genotypes in all available TB GWAS data.
iv. Perform a trans-ethnic X chromosome, TB meta-analysis to determine sex specific and/or combined effects

3. Investigate gene-gene interaction to determine how X-linked genes interact with each other and the rest of the genome.
   i. Carry out a genome-wide epistatic analysis (gene-gene interactions) on the SAC data

1.9 Structure of thesis

Each chapter for this thesis is structured for potential publication. All chapters are in the format of a journal article, but whether submitted for publication or not, they have been edited to have the same format and referencing style (Vancouver) for the sake of consistency throughout this thesis. Supplementary material was inserted at the end of the respective chapters and one overall bibliography is given at the end of the thesis.

Chapter 2: The X chromosome and sex-specific effects in infectious disease susceptibility.

This review sets the tone for the thesis and highlights the involvement and importance of the X chromosome in immune functions and potential impact on sex-bias of infectious diseases, specifically TB. This review has been submitted to the journal of Human Genomics and is currently under review.

Chapter 3: Autosomal and X chromosome markers confirm strong sex-biased admixture in the South African Coloured population.

This paper proves that global ancestry inference on the autosome and X chromosome can be accurately determined and used to infer presence of sex-bias in the SAC population.

Chapter 4: A sex-stratified genome-wide association study of tuberculosis using a multi-ethnic genotyping array.

This TB GWAS in the SAC population is the first GWAS to conduct not only X chromosome specific analysis, but also sex stratified tests on the autosome. This chapter also reports on the first genome wide epistatic analysis performed for a TB susceptibility study. Results for this chapter indicate strong sex specific effects, highlighting the need for sex stratified and X-linked analysis. This chapter has been submitted to the journal Frontiers in Genetics and is currently under review

Chapter 5: Evaluating the accuracy of imputation in the 5-way admixed South African Coloured population.

This methodological paper assesses the accuracy and aims to maximise the quality of imputation in the 5-way admixed SAC population on both the autosome and X chromosome.

Chapter 6: X-linked trans-ethnic meta-analysis reveals Tuberculosis susceptibility variants

This chapter reports on the first ever X chromosome specific trans-ethnic meta-analysis for TB to identify global and population specific X-linked susceptibility loci.

.

2 The X chromosome and sex-specific effects in infectious disease susceptibility

Haiko Schurz [1,3*], Muneeb Salie [2], Gerard Tromp [1,3,4], Eileen G. Hoal [1], Craig J. Kinnear [1], Marlo Möller [1]

[1] DST-NRF Centre of Excellence for Biomedical Tuberculosis Research; South African Medical Research Council Centre for Tuberculosis Research; Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

[2] Department of Genetics, St. Jude Children's Research Hospital, Memphis, TN, 38105, USA

[3] South African Tuberculosis Bioinformatics Initiative (SATBBI), Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

[4] Centre for Bioinformatics and Computational Biology, Stellenbosch University, Cape Town, South Africa.

[*] Corresponding author: haiko@sun.ac.za

2.1 Abstract

The X chromosome and X-linked variants have largely been ignored in genome-wide and candidate association studies of infectious diseases due to the complexity of statistical analysis of the X chromosome. This exclusion is significant, since the X chromosome contains a high density of immune-related genes and regulatory elements that are extensively involved in both the innate and adaptive immune responses. Many diseases present with a clear sex bias and, apart from the influence of sex hormones and socioeconomic and behavioural factors, the X chromosome, X-linked genes and X chromosome inactivation mechanisms contribute to this difference. Females are functional mosaics for X-linked genes due to X chromosome inactivation and this, combined with other X chromosome inactivation mechanisms such as genes that escape silencing and skewed inactivation, could contribute to an immunological advantage for females in many infections. In this review we discuss the involvement of the X chromosome and X-inactivation in immunity and address its role in sexual dimorphism of infectious diseases using tuberculosis susceptibility as an example, in which male sex bias is clear, yet not fully explored.

2.2 Key words

Tuberculosis, Sex-bias, X chromosome, Host genetics, Susceptibility

2.3 Introduction

The human sex chromosomes are genomic structures that distinguish males and females on the chromosomal level. The XY sex-determination system is present in humans and females have two X chromosomes, while males have one Y and one X chromosome (148). These chromosomes evolved approximately 180 million years ago from ordinary autosomes (149). Recombination during male meiosis was suppressed and over time, resulting in vast levels of divergence between the human sex chromosomes, with the exception of the pseudoautosomal regions (PAR1 and PAR2) located at the termini of the X and Y chromosomes (150). Over 800 protein coding and 600 non-coding genes are distributed over the nearly 155 million base pairs of the X chromosome (151). Until recently the X chromosome has largely been excluded from candidate gene and genome-wide association studies (GWAS) due to the statistical complexity of analysing and comparing the haploid male to diploid female data, but analysis tools have now been developed to incorporate this chromosome.

Gao *et al.* (115) developed a toolset for X chromosome data analysis and association studies that can be used for quality control and analysis of X chromosome GWAS data. Other software using genotyping data, but not specifically focused on the X chromosome, have also included the option to analyse X-linked genotypes. PLINK version 1.9, a software to conduct association testing using genotyping data incorporated different models to analyse the X chromosome (120). Impute2 and shapeit2 are programs designed to impute and phase genotyping data respectively and until recently imputation and phasing was not possible for the X chromosome thus excluding this chromosome from

downstream analyses(152,153). The ability to increase the amount of genotyping data through imputation and including the X chromosome in statistical analysis allows for X-linked meta-analysis and could help elucidate sexual dimorphism. Admixture analysis uses an individual's genomic data to determine ancestry by comparing allele frequencies to those of reference populations. Until recently this analysis was inaccurate for haploid genotypes and thus overestimated X-linked ancestral components in males. However, inclusion of haploid specific ancestry inference in the ADMIXTURE v1.3.0 software now allows for X-linked global ancestry inference (154). These ancestral components can now be included as covariates in X-linked association testing to improve the quality of the results. The software RFMIX also incorporated the option of assigning local ancestry on the X chromosome (155), allowing the comparison of autosomal and X-linked ancestral distributions, which could be indicate sex-biased admixture (156–158).

The development of these tools is especially significant for diseases in which a sex bias is present. Human males are more susceptible to many diseases, including bacterial infections, while females are more likely to develop autoimmunity (140). This sex bias is not only due to socioeconomic and behavioural factors, such as the underreporting of female cases and/or access to healthcare, but may also in part be due to biological sex differences as determined by the X chromosome and X chromosome inactivation (XCI) (159). XCI is the process through which one X chromosome is inactivated to balance dosage of gene expression between XX females and XY males. XCI is established early during embryonic development and is maintained almost indefinitely. As males are haploid for the X chromosome it has been suggested that any damaging genetic variants on the X chromosome will have a more pronounced immunological consequence in males than in females, thereby introducing sex-based differences and influencing the sex bias of a disease. In contrast, females, who are functional mosaics for X-linked genes, may have less severe consequences, further compounded by the process of skewed XCI and genes escaping silencing (160). This review will focus on the involvement of the X chromosome and XCI in immunity and will address sexual dimorphism in infectious diseases using tuberculosis (TB) susceptibility as an example, in which sex bias is clear, yet not fully explored.

2.4 X-chromosome, the immune system and sex hormones

Many X-linked genes are involved in the innate and adaptive immune system (161). This includes pattern recognition receptors (PRRs) such as toll-like receptor (*TLR7)* and *TLR8* as well as *IRAK1*, a key regulatory molecule in the TLR dependent signalling pathway (162). A number of transcriptional and translational control effectors functioning downstream of activated cytokine receptors are also located on the X chromosome (144). For example, NF-kB essential modulator (NEMO) modulates NF-kB expression, a transcription regulator that is involved in multiple immune pathways (163). Furthermore, it is not only X-linked genes that could influence the sex bias, but also X-linked control mechanisms like non-coding micro RNA (miRNA). The X chromosome contains approximately 10%

of the total genomic miRNA (164), which is involved in the regulation of gene expression by supressing mRNA translation or triggering mRNA degradation. Locations of immune related genes and key miRNA regions are indicated in Figure 2.1.

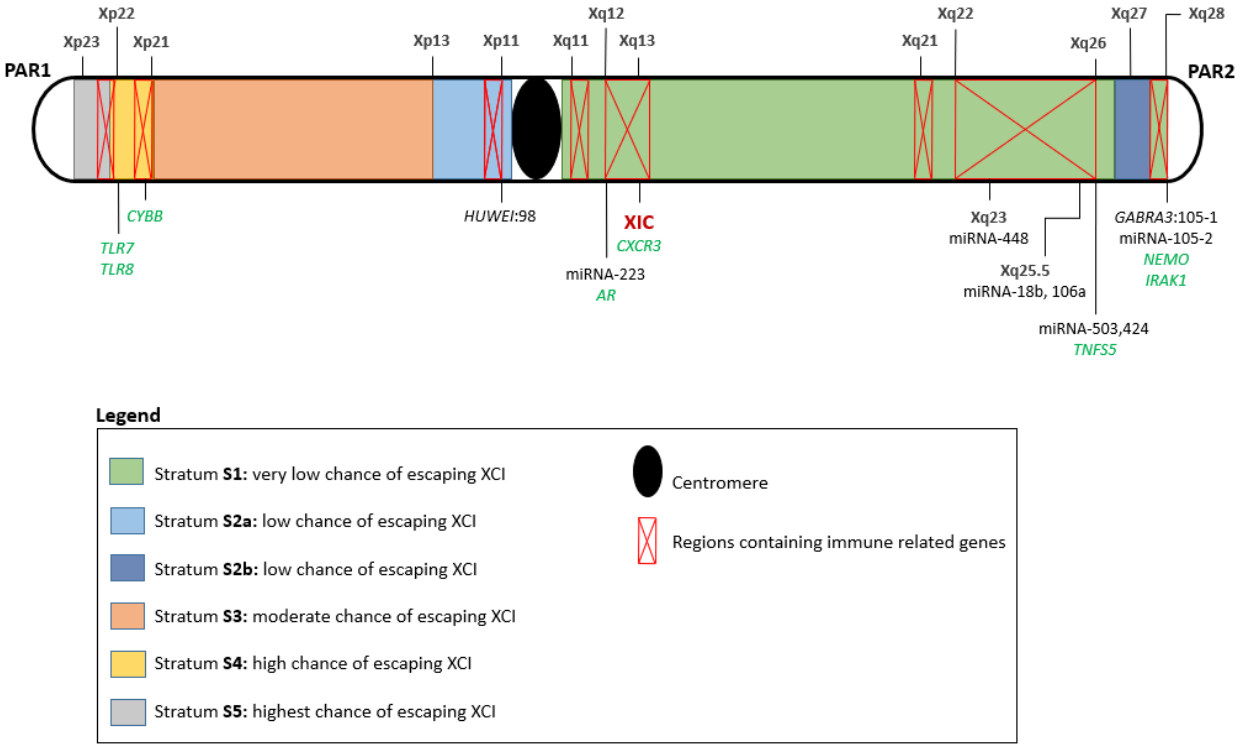


Figure 2.1: Illustration of the X chromosome indicating the five different strata and chances of genes escaping inactivation within each stratum. Regions lined in red contains the highest densities of immune associated genes while genes discussed in this review are indicated in green. Genes that contain intragenic miRNA are indicated in black followed by the miRNA number. XIC: X chromosome inactivation centre containing *XIST, XACT* genes; PAR: pseudoautosomal region; *TLR8*: Toll-like receptor 8; *TLR7:* Toll-like receptor 7; *CYBB:* cytochrome b-245, beta polypeptide; *AR:* Androgen receptor; *CXCR3:* C-X-C motif chemokine receptor 3; *TNFS5*: encodes CD40 ligand; *NEMO:* NF-kB essential modulator; *IRAK1:* Interleukin-1 receptor associated kinase 1; *HUWE1:* HECT, UBA & WWE domain containing 1; *GABRA3:* Gamma-aminobutyric acid A receptor subunit alpha 3.

The androgen receptor, a sex hormone receptor that inhibits antibody production is also coded on the X chromosome, showing that even the effect of sex hormones can be amplified by the X-linked sex hormone receptor genes (144). Sex hormones are involved in the immune response, and multiple immune related cells, including T-cells, B-cells, natural killer cells, macrophages and dendritic cells express estrogen receptors (ER-alpha and ER-beta), indicating that immune related cells are partly controlled by the female sex steroid hormone estrogen (144,165,166). In humans it is evident that females have increased resistance against microbial infections, which suggests that females have a more vigorous immune defence against most invading pathogens (167–170). Females also have higher antibody responses and more adverse reactions in response to a number of vaccines (144). Estrogen acts as an immune activator while testosterone acts as an immune suppressor (144,171).

Testosterone has been shown to have an inhibitory effect on the immune system through upregulation of anti-inflammatory cytokines (IL-10), while estrogen enhances the immune system by upregulating pro-inflammatory cytokines (TNFα) (172). In line with these hormone functions, it has been observed that for some diseases the male bias becomes apparent only after sexual maturation (ages 15-16 years) and female progression to disease and mortality rates are altered during their reproductive years (173). However, sex-based differences in immune responses exist between pre-pubertal girls and boys as well as post-menopausal women and elderly men, indicating that sex bias is present without the involvement of hormones (144). These differences could be attributed to the complexity of studying the impact of hormones on disease susceptibility while using different experimental designs between studies (140). Sex hormones also vary with age and physiological state of the individual and can regulate transcription of many genes involved in the development and maturation of immune cells. They also influence the regulation and modulation of the immune response and immune signalling pathways (174). Although both sex-hormones and the X chromosome affect the immune system, the effects of these two factors are likely independent of each other (140).

2.5 X chromosome inactivation

Females carry both a maternal and paternal X chromosome, while males carry only a maternal copy. In order to regulate dosage expression of X-linked genes, one X chromosome is inactivated in females, resulting in them being functional mosaics for X-linked genes (164). XCI is initiated in early fetal development and either the maternal or paternal X chromosome is randomly silenced in XX cells. This is maintained through epigenetic mechanisms in subsequent cellular divisions to ensure balanced expression X-linked genes in females (175).

XCI developed as a response to gene loss in the Y chromosome during the evolutionary development of the human sex chromosomes (150). Mammalian sex chromosomes developed from a pair of autosomes approximately 300 myr ago (176). Several large-scale chromosomal inversions on the Y chromosome led to disruption of homology between the sex chromosomes, suppressing recombination and resulting in Y chromosomal gene loss in the inverted chromosomal region (150). These inversions on the Y chromosome are referred to as strata as indicated in figure 1. Following gene loss on the Y chromosome X-linked gene expression needed to be increased in males to control the dosage of gene expression from the single X chromosome. In female's upregulation of X-linked genes would disrupt dosage compensation as they have two X chromosomes and as a result one X chromosome is inactivated. However gene expression is upregulated on the active X chromosome in order to regulate dosage (177,178). XCI is a vital mechanism in females as many X-linked genes are extremely dosage sensitive and any disruption of the dosage compensation mechanism could have severe developmental and health consequences (177).

Mary Lyon first proposed the XCI hypothesis based on her observations in mice (179) and since then significant progress has been made in elucidating the XCI mechanism in mice. The XCI mechanism

in humans is still unclear and beyond the scope of this review but discussed elsewhere (177,180,181,181–186). Briefly human XCI is thought to be controlled by the X inactivation centre (XIC), an X-linked locus located at Xq13 (Figure 2.1) and containing multiple protein and RNA coding genes potentially involved in the XCI mechanism (187). The two main long noncoding RNAs identified thus far are the X-inactivation specific transcript (*XIST*), responsible for silencing and the X active specific transcript (*XACT*) which keeps the X chromosome active (188–191). The exact mechanisms of how these lncRNAs determine the state of a X chromosome is still unclear and it has also been proposed that a third regulatory element, potentially coded by a gene on chromosome 19 is also involved in the XCI process (177). Hypotheses about the lncRNAs as well as an autosomal regulatory element are discussed in detail elsewhere (177,187,192–195). While the exact mechanisms are unclear, the importance of these lncRNAs have been validated as SNPs or mutations in the XIC can have severe effects on XCI, by disrupting dosage compensation, which could impact on female development and health (177,196). In fact, evidence of the effect of XCI can be seen in tumorigenesis and noncongenital diseases, where loss of XCI control has led to tissue instability and decreased defence against diseases (139,197,198), including autoimmune diseases (199).

While disruption of XCI could be detrimental to females as it disrupts dosage compensation, the mosaic nature as a result of XCI could give them a distinct advantage over males (140,181). Deleterious X-linked mutations have large effects and could lead to death or disease in males due to them being haploid for X-linked genes. In females however, random inactivation leads to a mosaic makeup where about half of the cell population expresses the mutant allele while the other half expresses the wild type allele. This heterozygous expression means the wild type allele can compensate for the mutant allele and lessen the impact or penetrance of this allele in females compared to males (177). This mosaic advantage in heterozygous females can be further compounded by non-random or skewed inactivation and genes that escape silencing.

2.6 Escaping X inactivation and skewed or non-random inactivation

While the XCI process in humans is not yet fully understood, studies of human aneuploidy indicate that in a diploid human cell there is always just one active X chromosome in either sex (177,181). In Turner syndrome individuals have only one sex chromosome (one X chromosome, X0) which is kept active, while in males with Klinefelter syndrome (XXY) one X chromosome is silenced (177). This suggests that the human XCI mechanisms protects one X chromosome while inactivating all others.

However, some X-linked genes have Y homologs (most of them situated on the distal end of Xp and PAR regions) and thus two copies are present in males and females. To maintain dosage balance between the sexes these XY genes escape silencing. Most genes that escape silencing are located in the Xp region and are often depleted in repressive marks associated with XCI and enriched for markers associated with active gene transcription (200). These regions that escape inactivation carry features associated with active chromatin (201). This suggests that genes that escape silencing are

subjected to a regional bias, which correlates with the theory that distal genes in younger strata (regions on the X chromosome that differentiated from the Y chromosome last and contain more XY genes than older strata) have a higher chance of escaping inactivation.

More recent evidence extrapolated on the idea of regional bias in escape from inactivation and showed that the chance of genes escaping silencing is also dependent on a gene to gene specific bias (176). This is supported by the fact that approximately 15-20 % of X-linked genes outside of PAR also escape silencing even though they are subject to less regional bias. Naqvi *et al.* (176) classified X-linked genes into 3 classes, namely X-linked genes with a surviving homolog (class 1) and X-linked genes without a surviving homolog that are either subject to XCI (class 2) or escape silencing (class 3) (176). These three classes of X-linked genes differ based on dosage sensitivities. Class 1 genes were most dosage sensitive and expression required strict regulation, while class 2 genes had intermediate dosage sensitivity while class 3 genes that escaped silencing had the lowest dosage sensitivity (176). This suggests that genes that escape silencing are subjected to regional bias and the chance of escape depends on the sensitivity of that gene to changes in dosage. While defects in the XCI mechanism could disrupt the XCI pattern of dosage sensitive genes and be detrimental to the health and development of females, genes that are less sensitive to dosage could escape resulting in altered gene expression between the sexes and potentially contribute towards a sex-specific phenotype, which could contribute to sex biases in disease susceptibility (140,161,202).

Random inactivation ideally leads to a balanced mosaic of X-linked genes in females. However, this balance can be disrupted, especially in heterozygous females carrying deleterious mutations on one or both X chromosomes, or if the XCI mechanism is defective, leading to a skewed inactivation pattern. Skewed inactivation is the process by which one X chromosome is preferentially silenced in over 75% of cells. If a cell has a deleterious mutation on the active X chromosome it could alter the viability of the cell and even lead to cell death, suggesting that these mutations could lead to positive or negative selection of a specific active X chromosome (203,204). The extent of this selection pressure is correlated with three factors. Firstly, the viability of the cell which will be determined by the active X chromosome. If cells with an active X chromosome with a detrimental gene die, then only cells with the viable gene will propagate. This depends on the type of mutation (synonymous or non-synonymous) and its effect on gene function. Second, the gene function can influence the skewing if it is tissue-specific while a constitutively expressed gene could affect the skewing on a global scale. Finally, genes escaping inactivation can also influence selection as they will influence the penetrance of the mutated gene (205). While cell viability combined with XCI can skew inactivation patterns, other aspects can also lead to non-random inactivation. Defects in the XCI mechanism can also lead to skewed inactivation and SNPs in the *XIST* gene correlates with skewing. Plenge *et al.* (196) showed that skewed inactivation profiles in multiple females occurred due to a C to G transversion in the promoter region of the *XIST* gene (196). However, some females with this transversion still had nearly

random inactivation suggesting that the transversion alone is not enough to skew inactivation and some other defect compounding the effects is likely present as well.

Other factors that can result in skewed inactivation is reduced number of embryonic cells at the onset of XCI and age. The lower the number of cells at the onset of XCI the higher the chance of observing non-random inactivation and any bottleneck during development that limits the number of cells can lead to skewed inactivation (205). Age has also been correlated with degree of skewing which seems to increase in older women (206–209). The exact reason why skewing increases with age is unclear, but it could be as a result of stochastic loss and genetic selection of subtle SNPs, gradually increasing their penetrance over time due to increased skewing in the XCI pattern (206,207,210,211). The causes of skewed XCI discussed here suggest that this process is genetically determined (203) and can give females an advantage by protecting them from deleterious mutations and their effects. However, skewed inactivation patterns have also been observed in numerous tumours and cancer types (200,212). This suggests that the combined impact of XCI, genes that escape silencing and skewing can lead to sex-specific phenotypes and potentially affect disease and developmental bias between the sexes.

2.7 X chromosome and infectious disease susceptibility

It is well documented that females have a stronger innate and humoral immune response than males and are thus less susceptible to many bacterial, fungal, parasitic and viral infections, while being more prone to developing an autoimmune disease or malignancies (Table 2.1, (168)). However as not every microorganism elicits a sex-differentiated response it has been proposed that the invading organisms and how they interact with the host are important contributing factors to whether or not the host immune response will differ between the sexes (213).

Table 2.1: Sex bias of selected bacterial, fungal, parasitic and viral infections.

Infection	Organism	Disease	Bias	Reference
Bacterial	*Treponema pallidum*	syphilis	male	(214–216)
	Borrelia burgdorferi	Lyme disease	Male (age)	(217,218)
	Vibrio vulnificus	Infection	Male	(219)
	Staphylococcus aureus	Infection	Male	(220,221)
	Pseudomonas aeruginosa	Infection	Male	(220,221)
	Escherichia coli	Bacteraemias	Female	(220,221)
Fungal	*Cryptococcus neoformans*	fungal meningitis	Male	(222–224)
	Candida albicans	onychomycosis	Female	(225–230)

Infection	Organism	Disease	Bias	Reference
	Paracoccidioidal brasiliensis	Infect mucosal membranes	Male	(231)
Parasitic	*Schistosoma*	Schistosomiasis	Male	(232–234)
	Leishmania	Leishmaniasis	Male	(232–234)
	Taenia	Tapeworm	Female	(232–234)
Viral	*Influenza A*	Influenza	Male	(141,235–237)
	Hepatitis C	Hepatitis	Male	(238,239)

Many infections exhibit sex biased incidence rates and many of them present with a male bias (Table 2.1). While age and sex hormones contribute, as in the case of Lyme disease and Hepatitis, these factors do not fully account for this (217,218,238,239). This suggests that the X chromosome and XCI may contribute to this bias. Supporting evidence from this can be taken from the mouse four core genotype (FCG) model. In this model the sex chromosome complement of the mice (XX or XY) does not relate to the gonadal sex, allowing for both XX males and females as well as XY males and females (240). This allows the study of the phenotypic effect based on sex complement, with and without the influence of sex hormones. Studies using the FCG model have identified differences in behaviour, gene expression, disease susceptibility that were solely due to sex chromosome complement and independent of sex hormones (240).

While the FCG is only a model it can still provide useful information and shows that sex chromosome complements, X-linked genes and XCI can severely impact sex-differences in phenotype. Recent studies in female T and B cells could explain the enhanced female immune response to infection. XCI in female lymphocytes is predisposed to become partially reactivated, allowing genes to escape silencing leading to overexpression of immune related genes (193,241). Female T-cells had biallelic expression of *CD40LG, CXCR3* and *TLR7*. The same was observed for B-cells where biallelic expression and increased transcription of X-linked immune-related genes was observed (241). Furthermore, in both T and B-cells the *XIST* RNA pattern was dispersed and the inactivated X chromosome lacked typical heterochromatic modifications usually associated with the inactive X chromosome (241).

These studies in female lymphocytes provide mechanistic evidence for enhanced female immunity to infectious diseases and the involvement of X-linked genes and XCI. The enhanced immune response and increased expression of immune related genes could also explain why females are more prone to developing autoimmune disorders (140,168,241,242).

2.8 X chromosome and tuberculosis

TB, caused by the bacterium *Mycobacterium tuberculosis*, is the leading cause of death due to a single infectious agent worldwide. Approximately one quarter of the world's population is infected with the bacterium, but only 5-15% will develop active TB (Houben and Dodd, 2016). The severity of this pandemic is exacerbated by the emergence of multidrug-resistant and extensively drug-resistant (MDR and XDR) *M. tuberculosis* strains. Although vital to the affected individual, it is clear that antimycobacterial treatment alone will not eradicate this disease. Host-directed therapy is emerging as a complementary approach to reduce the global TB burden, but will require an improved understanding of the host immune response and the genetic mechanisms that underlie it (244). To date, variants of genes involved in both the innate and adaptive immune responses have been associated with TB (reviewed by (245)). However, these investigations have been largely aimed at the autosome, while excluding the X chromosome. Given the high density of immune related genes on the X chromosome (144) and the fact that TB presents with a clear sex bias across populations, this is a serious oversight (246).

In most countries the TB notification rate is twice as high in HIV negative males than in HIV negative females (246). This ratio ranged from 1.56 to 2.73 and while it differs between countries, it was clear that more men than women are affected regardless of ethnicity or geographical location. Epidemiological data has shown that males and females differ in infection prevalence, varying rates of progression, differences in incidence of clinical disease and mortality rates due to TB (247). The cause of this male sex bias is not fully understood, but may include socioeconomic and behavioural factors, such as the underreporting of female cases and/or access to healthcare (166,248–250). However, these differences in case reporting may influence the bias but cannot explain the consistent global trend for male bias in TB (165). In a large meta-analysis including 29 surveys from 14 countries, a strong male bias was found in both TB notifications and prevalence and it was concluded that access to healthcare is not a confounding factor (251). This was replicated by Salim *et al.* (248) who conducted a survey of 223 936 individuals in Bangladesh and identified 7 001 TB suspects at a female to male ratio of 0.52:1. Sputum was obtained from these individuals and 64 positive TB cases were identified at a female to male ratio of 0.33:1. These observed ratios did not differ much and were in fact lower than the female to male ratio observed through diagnosis in clinics which stood at 0.42:1. The authors concluded that reduced access of women to health care facilities does not significantly influence the bias seen (248). In a study conducted in Syria, men and women did not have different knowledge or attitudes towards TB, but women reported more barriers to seeking health care. They were more likely to comply with treatment and had higher treatment success rates compared to men which could influence the bias when it comes to TB mortality (252). Furthermore, men seem to engage in more "high risk" TB activities, including traveling, smoking, going to bars, and hazardous careers (e.g. mining) (165). In high burden countries more men than women engage in smoking and it has

been suggested that smoking may explain up to one third of the gender bias observed in TB (253). Alcohol consumption could have a similar effect. However other risk factors, specifically HIV infection and proximity to household contacts appear to have a female bias, which suggests that although behaviour may influence the bias it is not sufficient to fully explain the existing sex bias in TB (165). Another contributing factor may be the influence of sex hormones on the immune system (discussed in section 2).

Females have been shown to have a more robust immune system (as described in section 5) and this is in part mediated by sex hormones that control development and maturation of immune cells (T-cells, macrophages, neutrophils) involved in combating TB. Type 1 T helper cells (Th1) are affected differently by male and female sex hormones. Testosterone upregulates IL-10 while down regulating IFN-γ (254), and estrogen increases IFN-γ, TNFα and IL-12 production while supressing production of IL-10 (255). Macrophages, which play a central role in controlling TB through active killing of mycobacteria, are also influenced by sex hormones. The female hormone estradiol has been shown to enhance macrophage activation (173), while testosterone down regulates macrophage activation by decreasing expression of TLR4, a vital receptor for detecting *M. tuberculosis* and initiating the innate immune response (167). Neutrophils have recently garnered interest with regards to their role in protection against TB and have been proposed to be the predominantly infected phagocytic cell type in pulmonary tuberculosis (pTB) (256). Neutrophil recruitment to areas of infection needs to be balanced as under and over recruitment of neutrophils can have a detrimental effect on tissue pathology (257). In response to trauma, testosterone decreases neutrophil activation while estrogen increases it, but the effect of this on TB is unknown and requires further investigation (258). As neutrophil recruitment needs to be balanced to avoid under or over recruitment to sites of infection it stands to reason that the regulation of this recruiting mechanism is of vital importance. In fact, miRNA-223 (Xq12, figure 1), previously identified to be involved in the immune response by Pinheiro *et al.* (259) can limit recruitment of neutrophils by down regulating chemokine (C-X-C motif) ligand 2 (CXCL2) and chemokine (C-C motif) ligand 3 (CCL3). Mice with a miRNA-223 knockout were more susceptible to *M. tuberculosis,* due to excessive neutrophil accumulation in the lungs which subsequently led to tissue damage (260). Given that miRNA-233 is X-linked, is subject to the effects of skewed inactivation or may escape silencing, it could be differentially expressed between males and females. Up regulation due to escape from silencing or preferential expression of one gene copy due to skewed inactivation could down regulate recruitment and thus the pathological accumulation of neutrophils leading to a sex bias in TB susceptibility. Clearly these factors do not fully explain the male bias associated with TB disease development, suggesting that the host genotype, specifically the X chromosome, may also contribute.

The third possible reason for the sex bias in TB susceptibility is linked to the X chromosome where skewed inactivation or genes escaping silencing could give females an enhanced immune response

against *M. tuberculosis.* Some of the earliest evidence of this X-linked genetic contribution to sex bias in TB susceptibility came from the “Lübeck Disaster” in 1929. Bacillus Calmette–Guérin (BCG) vaccine accidentally contaminated with *M. tuberculosis* was administered to 251 neonates. 173 of these children developed signs of active TB but recovered, while 72 died and during follow-up male children were more likely to have poor outcomes than females (261). Evidence from studies of Mendelian susceptibility to mycobacterial disease (MSMD) also supports the influence of the X chromosome to disease susceptibility. MSMD is a rare congenital syndrome that results in the predisposition to diseases caused by non-virulent mycobacteria, BCG vaccines and environmental mycobacteria known not to be disease causing in humans (262). MSMD is classified into two types, where autosomal MSMD is linked to defects in 5 autosomal genes (*IFNGR1, IFNGR2, STAT1, IL12RB1* and *IL12B*) involved in the interleukin 12/23 dependant interferon γ (IFN-γ) mediated immune response (263). On the other hand X-linked recessive (XR)-MSMD is less well understood (163). Several genetic defects have been proposed to cause XR-MSMD, and based on the genes involved, XR-MSMD can be further subdivided into two types, XR-MSMD type 1 and XR-MSMD type 2. Type 1 XR-MSMD is caused by mutations in the leucine zipper domain of the NF-kB essential modulator (*NEMO*) gene, which selectively impairs the CD40 and NF-kB/c-Rel-mediated induction of IL-12 production by monocytes and monocyte derived dendritic cells (262). Predisposition of type 2 XR-MSMD is increased by mutations in two regions on the X chromosome, Xp11.4-Xp21.2 (129 known genes) and Xq25-Xq26.3 (70 known genes). These regions may cause XR-MSMD independent of *NEMO* and Bustamante *et al.* (264) proposed that variants in the cytochrome b-245 beta polypeptide (*CYBB)* gene could predispose to XR-MSMD-2 due to their selective effect on macrophages. *CYBB* encodes the gp91 protein, which is an essential component of the NADPH oxidase complex and severely affects respiratory burst in macrophages, thereby impeding their function and predisposing to XR-MSMD-2. *NEMO* and *CYBB* are both X-linked genes that affect immune related cells and as such can alter susceptibility to TB. XR-MSMD, like TB, shows a sex bias and affects more males than females which can be attributed to females carrying two X chromosomes. If one of the X chromosomes carries a defective *NEMO* or *CYBB* gene random XCI can result in the functional gene product still being expressed and reducing the risk of disease. Skewed inactivation or escape from silencing could further increase the observed sex bias as *NEMO* and *CYBB* have a low (stratum S1) and high (stratum S4) chance of escaping inactivation (figure 1). However, TB in immunocompetent individuals is a multigenic disease linked to variants in multiple genes that have a cumulative effect on disease susceptibility and is even further complicated by gene-gene interactions.

The first genome-wide linkage analysis of TB susceptibility identified the chromosome Xq26 region as containing susceptibility genes, but did not specifically investigate sex bias (265). Although no specific genes could be identified, the CD40 ligand encoded by the *TNFSF5* gene at Xq26.3 showed promise (figure 1), but requires further investigation (265). A study by Campbell *et al.* (266) on 121

TB cases and their parents identified a *TNFSF5* (a CD40 ligand) variant (-726) to be associated with TB susceptibility in males. However, they failed to replicate this association in a West African cohort of 1200 individuals.

More recently, sex-specific associations with genetic variants in the X-linked toll-like receptor (*TLR*) 8 gene (Table 2), which encodes a pattern recognition receptor, were identified (132,134–138). Davila *et al.* (135) identified four variants in *TLR8* (rs3764879, rs3788935, rs3761624 and rs3764880) that were significantly associated with TB susceptibility in Indonesian males, but not females. These findings were validated in a male only cohort from Russia and all four variants were again significantly associated with TB susceptibility in males (267). A second study conducted in a paediatric Turkish cohort showed a significant association between rs3764880 and TB susceptibility in males but not females and rs3764879 showed no significant association in this cohort (268). Hashemi-Shahri *et al.* (136) also investigated the influence of rs3764880 on TB susceptibility in a cohort from Iran but found no association in either males or females. Significant associations were found for both males and females in a Pakistani cohort for rs3764880, but males were more strongly associated ($p=0.0013$ for females and $p<0.0001$ for males) (269). Salie *et al.* (138) was the first to identify an association between rs3761624 and TB disease in females only ($p<0.001$ for females and $p=0.164$ for males). Two SNPs, namely rs3764879 and rs3764880, were also investigated in this South African Coloured (SAC) population and were significantly associated in both males and females, but with opposite effects. Finally, Lai *et al.* (137) showed that rs3764879 was significantly associated with TB in males but not females. The conflicting results of these studies investigating *TLR8* may be explained by cohort size, ethnicity, *M. tuberculosis* strain and environmental factors.

It is clear that the X chromosome and XCI (section 5) is significantly involved in TB susceptibility and the male sex-bias and future studies will need to focus on elucidating these effects. Fully understanding the sex-biased nature of TB will allow for medication tailored to a specific sex, which could improve treatment outcome, decrease the global TB burden and stem the tide of emerging drug resistant *M. tuberculosis* strains.

Table 2.2: *TLR8* association studies from different populations.

Study	Cohort	Case	Control	SNP	Allele	Gender	OR*	95% CI*	P-value
Davila et al. (135)	Indonesia	77	49	rs3764879	C	Male	1.9	1.2-2.7	0.012
		76	74	rs3764879	C	Female	1.1	0.8-1.7	0.44
		76	51	rs3761624	A	Male	1.8	1.2-2.8	0.007
		76	74	rs3761624	A	Female	1.1	0.8-1.7	0.44
		76	50	rs3788935	A	Male	1.8	1.2-2.7	0.017
		76	74	rs3788935	A	Female	1.1	0.8-1.7	0.44

Study	Cohort	Case	Control	SNP	Allele	Gender	OR*	95% CI*	P-value
		76	51	rs3764880	A	Male	1.8	1.2-2.9	0.007
		76	74	rs3764880	A	Female	1.1	0.8-1.7	0.44
	Russia	1067	994	rs3764879	C	Male	1.2	1.02-1.48	0.03
		1069	997	rs3788935	A	Male	1.2	1.02-1.48	0.03
		1070	1000	rs3761624	A	Male	1.2	1.01-1.46	0.04
		1069	997	rs3764880	A	Male	1.2	1.02-1.48	0.03
Dalgic et al. (134)	Turkish children	72	62	rs3764880	A	Male	0.43	0.16-0.72	0.007
		156	124	rs3764880	A	Female	NS	NS	NS
		72	62	rs3764879	C	Male	NS	NS	NS
		156	124	rs3764879	C	Female	NS	NS	NS
Hashemi-Shahri *et al.* (136)	Iran	77	62	rs3764880	G	Male	1.15	0.84-1.59	0.80
		196	166	rs3764880	G	Female	1.15	0.75-1.75	0.51
Bukhari et al. (132)	Pakistan	45	22	rs3764880	A	Male	/	/	<0.0001
		58	65	rs3764880	A	Female	0.363	0.199-0.660	0.0013
Salie et al. (138)	SAC	204	99	rs3761624	A	Male	/	/	0.164
		217	336	rs3761624	A	Female	1.54	1.19-1.99	<0.001
		205	99	rs3764879	C	Male	0.72	0.55-0.93	0.013
		220	334	rs3764879	C	Female	1.41	1.08-1.83	0.011
		1887	81	rs3764880	A	Male	0.75	0.57-0.98	0.036
		199	306	rs3764880	A	Female	1.42	1.09-1.87	0.011
Lai *et al.* (137)	Chinese	96	146	rs3764879	C	Male	4.04	1.82-8.99	<0.001
		40	97	rs3764879	C	Female	5.05	0.44-57.38	0.191

*OR: Odds ratio; 95% CI: 95% confidence interval

2.9 Discussion and concluding remarks

It is clear that sex-specific effects contribute to infectious disease susceptibility and females have a major immunological advantage over males. Understanding the origin of sex bias could guide treatment by allowing sex-specific diagnostic and treatment regimes, thereby decreasing time to initiation of treatment as well as increasing treatment success of diseases with sex differences. The X chromosome may contribute to the missing heritability or contain biomarkers that could be used as diagnostic tools. As analytical tools are now available to fully include the X chromosome in genetic analyses, it is clear that the X chromosome should not be ignored. Importantly, due to the haploid

nature of males the power to detect a significant association will be halved when compared to a female cohort of similar size and this could have an effect on the results of sex-stratified analysis. Thus, care must be taken when analysing results, and a non-significant association in one sex does not imply that that specific sex is not affected by the variant but could simply be as a result of insufficient power to detect a sex specific association.

While socioeconomic and behavioural factors as well as sex hormones do influence sex bias, these factors do not fully account for it, which leads to the conclusion that the X chromosome itself is likely to greatly influence the immune response and sex bias in disease susceptibility. The X chromosome contains multiple immune-related genes and immune regulatory elements as well as the XIC that regulates X chromosome inactivation. It is therefore clear that the X chromosome is involved in the immune response and genes that escape inactivation or are preferentially inactivated could influence the dosage of X-linked gene expression between the sexes and as such could further influence the sex bias in disease. It is thus of vital importance that the XCI mechanisms be further investigated to understand all the regulatory elements involved and the contribution to sex bias. Furthermore, the role of the X chromosome in the innate and adaptive immune response should be extensively investigated to determine how it contributes and differs between the sexes. Elucidating the function of the X chromosome and including it in biological studies and analyses could improve the understanding of complex diseases such as TB.

2.10 Declaration

2.10.1 Competing interest

The authors report no conflict of interest.

2.10.2 Funding

This research was partially funded by the South African government through the South African Medical Research Council and the National Research Foundation of South Africa.

2.11 Acknowledgements

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. This work was also supported by the National Research Foundation of South Africa. This work was also supported 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, GW) to GT.

3 Global ancestry inference on the autosome and X chromosome identifies sex-biased admixture in a highly admixed population.

Haiko Schurz [1,2], Caitlin Uren[1], Meng Lin[3] , Craig J Kinnear [1], Paul D van Helden [1], Gerard Tromp [1,2,4], Brenna Henn [5], Eileen G Hoal [1], Marlo Möller [1]

[1] DST-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

[2] South African Tuberculosis Bioinformatics Initiative (SATBBI), Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

[3] Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033

[4] Centre for Bioinformatics and Computational Biology, Stellenbosch University, Cape Town, South Africa.

[5] Department of Anthropology, and the UC Davis Genome Center, University of California, Davis, CA, 95616.

3.1 Abstract

Recently admixed populations provide us with a unique opportunity to study population history and fine-map disease loci. The South African Coloured population is a complex admixed group with at least 5 ancestral populations (Bantu-speaking African, KhoeSan, European, South and East Asian). Previous studies suggested the presence of sex-biased admixture within this population, as well as others in South Africa, however, the evidence for sex-biased admixture is limited and under-investigated. Determining sex-biased admixture can inform recent population dynamics and it is vital to include this as a confounder during X-linked association studies of diseases that present with a sex-bias.

Significant sex-biased admixture for Bantu-speaking African, KhoeSan and European ancestry was identified using global admixture inference. The results presented here correlate with previous results based on mtDNA and Y chromosome markers, revealing a female bias for the KhoeSan and a male bias for the European and Bantu-speaking African ancestral components. The Asian components did not present a strong bias in this admixed population.

Here we show that global ancestry inference on the autosome and X chromosome can successfully be used to quickly and accurately identify the presence of sex-biased admixture in a highly admixed population, without the need for mtDNA and Y chromosome data.

3.2 Introduction

Genetic admixture is the process by which new genetic lineages are introduced into a population through the interbreeding of two or more previously isolated populations (270). Analysing the distribution of ancestral components in recently admixed individuals allows the study of population history and natural selection and can even be used to fine-map disease-causing variants (94,99). Previous investigations into sex-bias were done by comparing admixed and ancestral lineages using genetic material inherited from only one sex such as mitochondrial DNA (mtDNA) in females and the Y chromosome in males. By identifying the source populations of lineages present in the mtDNA and Y chromosomes of an admixed group, sex-biased admixture can be elucidated. A female bias is present when lineages from a specific ancestral population are increased in the mtDNA compared to the Y chromosome (271). This type of analysis has been used to identify sex-biased admixture in Icelanders (272), Tibeto-Burman (273), Native Americans (274,275), African Americans (276,277) and the SAC population (278).

An alternative method of investigating sex-biased admixture is to analyse the ancestral distribution on the X chromosome compared to the autosome. This method is dependent on the X chromosome spending 2/3 of its time in females while undergoing recombination and 1/3 of its time in males without recombination. In contrast, autosomes always experiences recombination and as a result the X chromosome retains longer linkage disequilibrium blocks than the autosome, which can be used to

investigate sex-bias (146,147). In a 2-way admixed population, if the mean ancestral component of population 1 is higher on the X chromosome than the autosome then this is an indication of female sex-bias from ancestral population 1 and male bias from ancestral population 2. It could however also indicate a female bias from both population 1 and 2, but a higher contribution from population 1. Alternatively it could indicate a male bias from both population 1 and 2 with more influx from population 2 (279). As a result of these alternative interpretations, historical evidence should always be used to corroborate the findings. This method quickly and effectively indicates sex-biased admixture and was previously implemented successfully in a three-way admixed African-American and Hispanic population (156–158). Computer simulations indicate that the method has high accuracy in populations that experienced a single admixture event, but it is less accurate if the population experienced continuous admixture (14). The SAC population is characterised by recent, continuous admixture (85,278), so it is unclear how well this method will perform. Furthermore, this method has not been tested on a complex admixed population such as the SAC.

The SAC population arose from relatively recent admixture resulting from the early encounters of European and Bantu-speaking African males with KhoeSan females at the Cape of Good Hope (South Africa) approximately 350 years ago (278,280). Ancestral contributions from the KhoeSan, Bantu-speaking African, European and South and East Asian populations are present in this group(278,280). Previous work using mtDNA and Y chromosome markers has shown that the ancestral distribution among SAC individuals is characterised by sex-biased admixture (278). The KhoeSan and South and East Asian ancestral components were female biased, while the European and Bantu-speaking African components were male biased (278).

We performed sex-biased admixture analysis in the SAC by investigating the distribution of ancestral components between the autosome and X chromosome. The results were compared to previous mtDNA and Y chromosome data to assess how accurately the admixture method calculates sex-bias in a 5-way admixed population.

3.3 Methods

3.3.1 Genotyping data

DNA was extracted from blood samples of SAC individual's as reported previously (281). Ethics approval was obtained from the Health Research Ethics Committee of Stellenbosch University (project registration number S17/01/013, 95/072 and NO6/07/132) before participant recruitment. Written informed consent was obtained from all study participants prior to blood collection. In total 800 samples from the SAC were genotyped on the Illumina Multi ethnic genotyping array (MEGA, Illumina, Miami, USA). The reference populations used to infer ancestry were Europeans (Utah Residents with Northern and Western European Ancestry) and South Asian (Gujarati Indians in Houston, Texas and Pathan of Punjab) obtained from the 1000 Genomes Phase 3 data (81) and East Asian (Han Chinese

in Beijing, China and Japanese, Tokyo, Japan), African (Luhya in Webuye, Kenya, Bantu-speaking African, Yoruba from Nigeria) and KhoeSan (Nama/Khomani) (103,282). The Zcall software was used to recall rare genotypes (MAF < 5%) for all datasets before aligning them to the 1000 Genomes Phase 3 reference panel and removing all ambiguous variants (81,283). PLINK v 1.07 (284) was used to check for sex concordance prior to merging all datasets. The merged data was filtered for Hardy Weinberg Equilibrium (<0.05), minor allele frequency (<0.03) and individual and SNP missingness (>10%) using PLINK (v1.07). Finally, all variants on the X chromosome that were heterozygous in males were removed and the merged dataset was checked for ambiguous variants using PLINK (v1.07) and snpflip[3] version v0.0.6 respectively.

3.3.2 Admixture analysis

For the admixture analysis the software ADMIXTURE 1.3 was used (154). Due to the level of relatedness in the SAC and the limited number of individuals per reference population, the SAC individuals were split into 20 running groups. A running group is a set of unrelated individuals for which admixture is to be inferred. Due to the limited number of reference individuals not all SAC data can be run simultaneously and thus needs to be split up. Each running group contains on average 42 unrelated SAC individuals, matching the number of individuals per reference population. Relatedness for the SAC was determined from the genotyping data using the software KING (version 2.1.4) (285). For each running group, Admixture was inferred five times at random seed values for both the autosome and X chromosome separately. For the X chromosome the ADMIXTURE software was run in haploid mode for males (--haploid=”male:23”) in order to ensure accurate admixture inference for haploid genotypes (286). The values for the five runs were then averaged for each individual before the results were analysed. ADMIXTURE was run at K=5 as we were interested in the 5 main SAC ancestral components. Pong (version 1.4.7) was used to visualise the admixture results across all runs (287).

3.3.3 Sex-bias analysis

Sex-bias was analysed by comparing the distribution of each ancestral component between the autosome and X chromosome. The data was assessed for normality (Figure S3.1) upon which a Wilcoxon signed rank test (a paired test for skewed data) was implemented, using the R programming environment (version 3.2.4 (142)), to determine significant differences in ancestral distribution. Since 5 ancestral populations were investigated, the Bonferroni correction for multiple tests was applied for the 5 tests performed, at a family-wise alpha threshold of 0.05. The results were also compared to previous results on sex-biased admixture in the SAC using mtDNA and Y chromosome markers in order to determine how the X chromosome-based sex-bias analysis compares.

[3] https://github.com/biocore-ntnu/snpflip

3.4 Results

Following quality control and merging of the data, a total of 558,213 autosomal and 13,399 X chromosome variants overlapped between all the datasets. The results for the global ancestry inference are shown in Figure 3.1 and 3.2 and the mean and standard deviations of each ancestral component are shown in Table 3.1 for both the autosome and X chromosome. On average the SAC individuals possess 28% Bantu-speaking African, 20.5% European, 27.1% KhoeSan, 14.7% South Asian and 9.6% East Asian ancestry on the autosome. These mean autosomal ancestral components reflect previous results of admixture analysis in the SAC population (85,99) ensuring accuracy of the inference. As shown in Table 3.1 and Figure 3.2 the mean for the Bantu-speaking African (p-value = $7.16e^{-5}$) and European (p-value = $1.47e^{-31}$) component is significantly larger on the autosome, following multiple test correction, suggesting a male bias. Contributions from the KhoeSan population (p-value = $1.78e^{-20}$) present with a significantly higher mean on the X chromosome compared to the autosome, indicating a female bias. The East Asian component passed the significance threshold (p-value = 0.038) following multiple testing correction (female bias), while the South Asian component did not reach statistical significance (p-value = 0.32), but the mean values for the South Asian components were slightly higher in females compared to males.

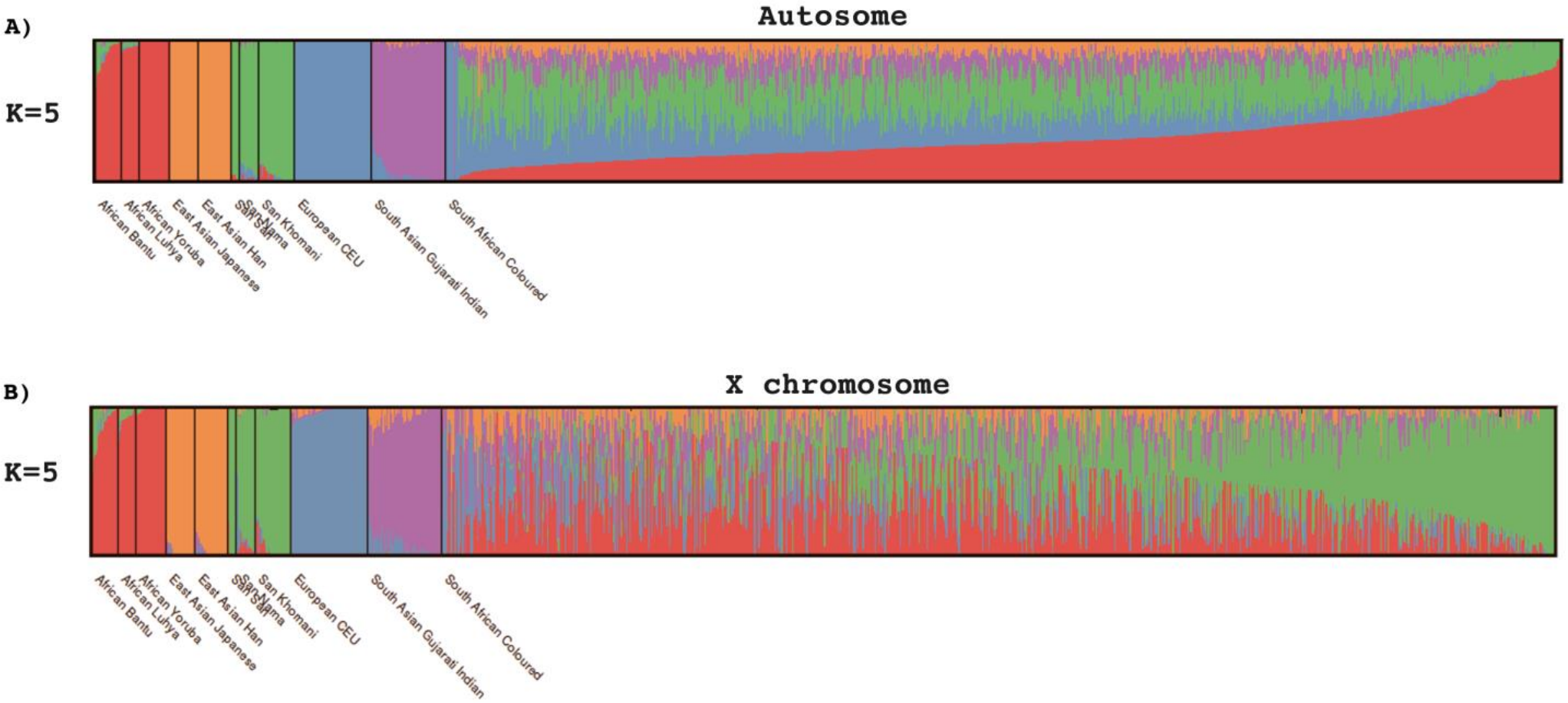


Figure 3.1: Admixture plot for all SAC and reference individuals for the autosome (A) and X chromosome (B). Each column represents one individual and all individuals are aligned between (A) and (B).

Table 3.1: Sex-biased distribution of each ancestral component.

Population	Median		Mean		Bias	p-value adjusted
	Autosome	X chromosome	Autosome	X chromosome		
African	0.236	0.203	0.280	0.267	Male	$7.16e^{-5}$
European	0.153	0.051	0.205	0.162	Male	$1.47e^{-31}$
KhoeSan	0.261	0.294	0.271	0.328	Female	$1.78e^{-20}$
South Asian	0.112	0.061	0.147	0.148	Female	$3.20e^{-1}$
East Asian	0.055	0.028	0.096	0.099	Female	$3.80e^{-2}$

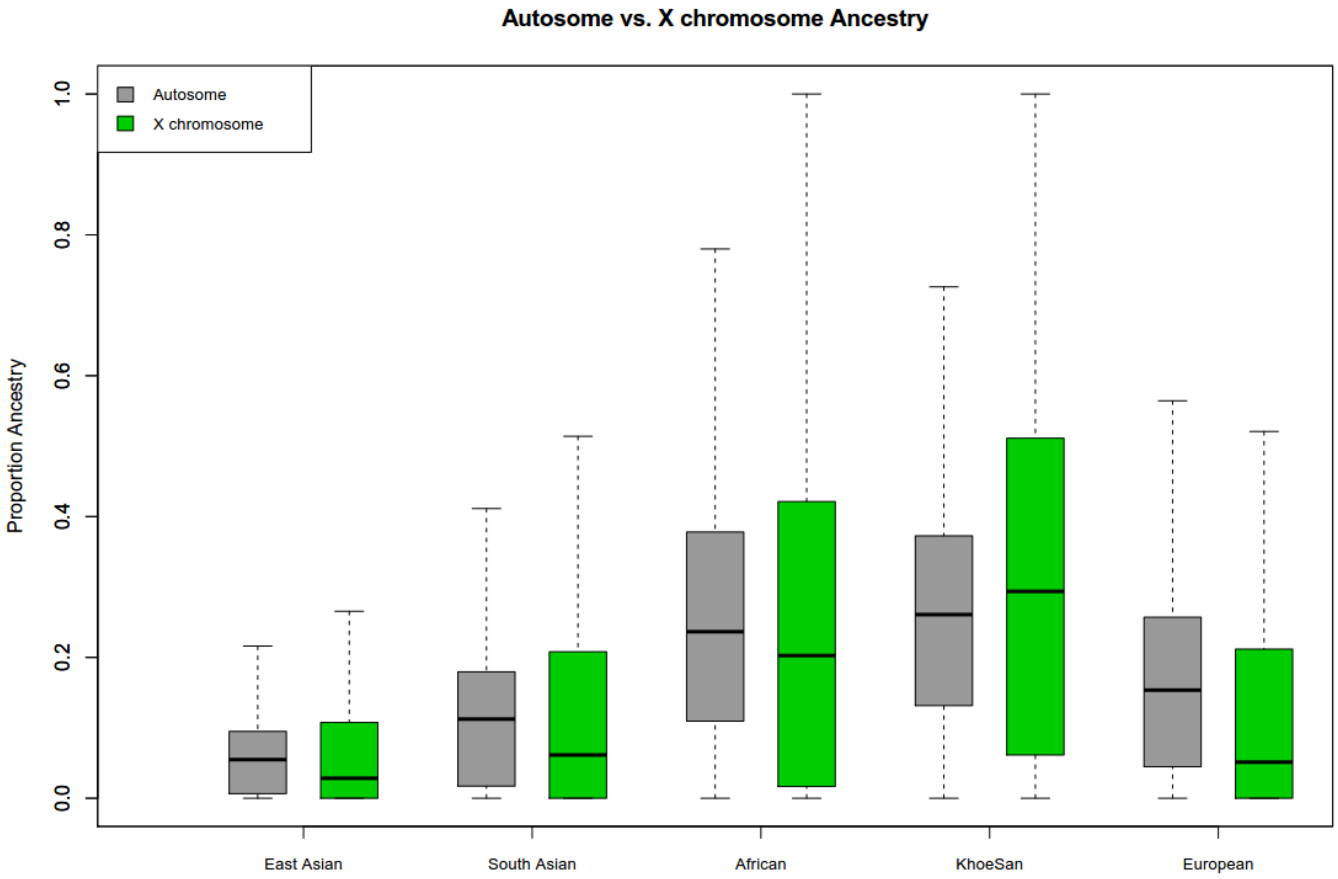


Figure 3.2: Boxplot of Autosomal (grey) and X chromosome (green) ancestral components indicating median values (thick black line), first and third quartile (box) and range (whiskers).

3.5 Discussion

We used global ancestry inference on the autosome and X chromosome to detect sex-biased admixture in the 5-way admixed SAC population. Previous sex-bias analysis of this population using mtDNA and Y chromosome markers indicated that the SAC population has a strong female KhoeSan bias with 60% of mtDNA and between 5.3% - 20% Y chromosome lineages originating from the KhoeSan population. The Bantu-speaking African and the European lineages indicated a male bias with a 19% mtDNA and at least 24% Y chromosome lineages from Bantu-speaking African populations and a 4.6% mtDNA and 32.5% Y chromosome lineages from European populations. Finally, for the South and South East Asian components previous results indicate only slight

differences between maternal and paternal contributions (278). For the South Asian component 8.7% mtDNA and 9.6% Y chromosome lineages were identified, while the South East Asian component had 7.6% mtDNA and 7.4% Y chromosome lineages, suggesting no sex-bias for the Asian components (278).

Our results show concordance with previous findings, which indicate a significant male bias from Bantu-speaking African and European ancestral populations and a significant female bias for the KhoeSan component (Table 3.1). For the South and East Asian component, the differences between maternal and paternal contributions were not significant in previous studies, whereas here we identified a significant East Asian female bias in the SAC population. The South Asian component was not significantly different between the autosome and X chromosome and indicated a marginal female bias in our data while Quintana-Murci *et al.* (2010) showed a marginal male bias (based on mean values, but no statistical significance). The discrepancy in the Asian component between the two studies can be explained. It could simply be a result of the variation in ancestral distribution observed when ancestry is inferred for different individuals of the same population. As the differences for the Asian ancestral component between the studies is small, it could be explained by this variation in ancestral distribution. Furthermore, Quintana-Murci *et al.* (278) analysed only 20 SAC individuals using 64 mtDNA markers, 46 Y chromosome repeats and 14 Y chromosome tandem repeats. This could result in an inaccurate presentation of population-wide ancestral components which could affect the accuracy of the results. This accuracy could be further affected by the limited number of mtDNA and Y chromosome markers used to detect ancestral lineages. For this study the SNP density, number of individuals per reference population and SAC sample size was considerably larger, leading to more precise admixture inference. The increased precision could explain why a significant difference for the East Asian component and a marginal female bias for the South Asian component is identified in our study. However, increased precision in admixture inference does not imply increased accuracy for determining sex-biased admixture. In order to fully quantify the accuracy of the admixture method further studies using mtDNA and Y chromosome data from more individuals and with more typed markers are required, especially to fully elucidate the contributions from Asian populations.

Despite the slight discrepancy in the results for the Asian components this analysis reveals that global ancestry inference can be used to accurately infer sex-bias in a 5-way admixed population. However, the accuracy of the results will depend on the precision of admixture inference and it is thus vital to ensure that appropriate reference populations are used (85,288). One potential limitation in this analysis was the fact that the SAC population experienced continuous admixture. Simulations that compared autosomal and X chromosome ancestral distributions lacked accuracy in populations that experienced continuous admixture (279). However, this was not the case for this study as it was possible to precisely determine sex-biased contributions from Bantu-speaking African, KhoeSan and

European populations, but not for Asians. This could be due to the smaller Asian genetic contribution to the SAC population and not necessarily due to continuous admixture events.

Determining sex-biased admixture can inform on recent population dynamics and socio-cultural factors associated with the founding of emerging populations. Global admixture inference can be used to accurately infer sex-biased admixture in a 5-way admixed population and significant bias for Bantu-speaking African, KhoeSan and European ancestry were identified in this study.

3.6 Acknowledgements

We would like to acknowledge and thank the study participants for their contribution and participation. 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. This work was also supported by the National Research Foundation of South Africa (grant number 93460) to EH. This work was also supported by South African Tuberculosis Bioinformatics Initiative (SATBBI), a Strategic Health Innovation Partnership grant from the South African Medical Research Council and Department of Science and Technology to GT.

The authors report no conflict of interest.

3.7 Supplementary material

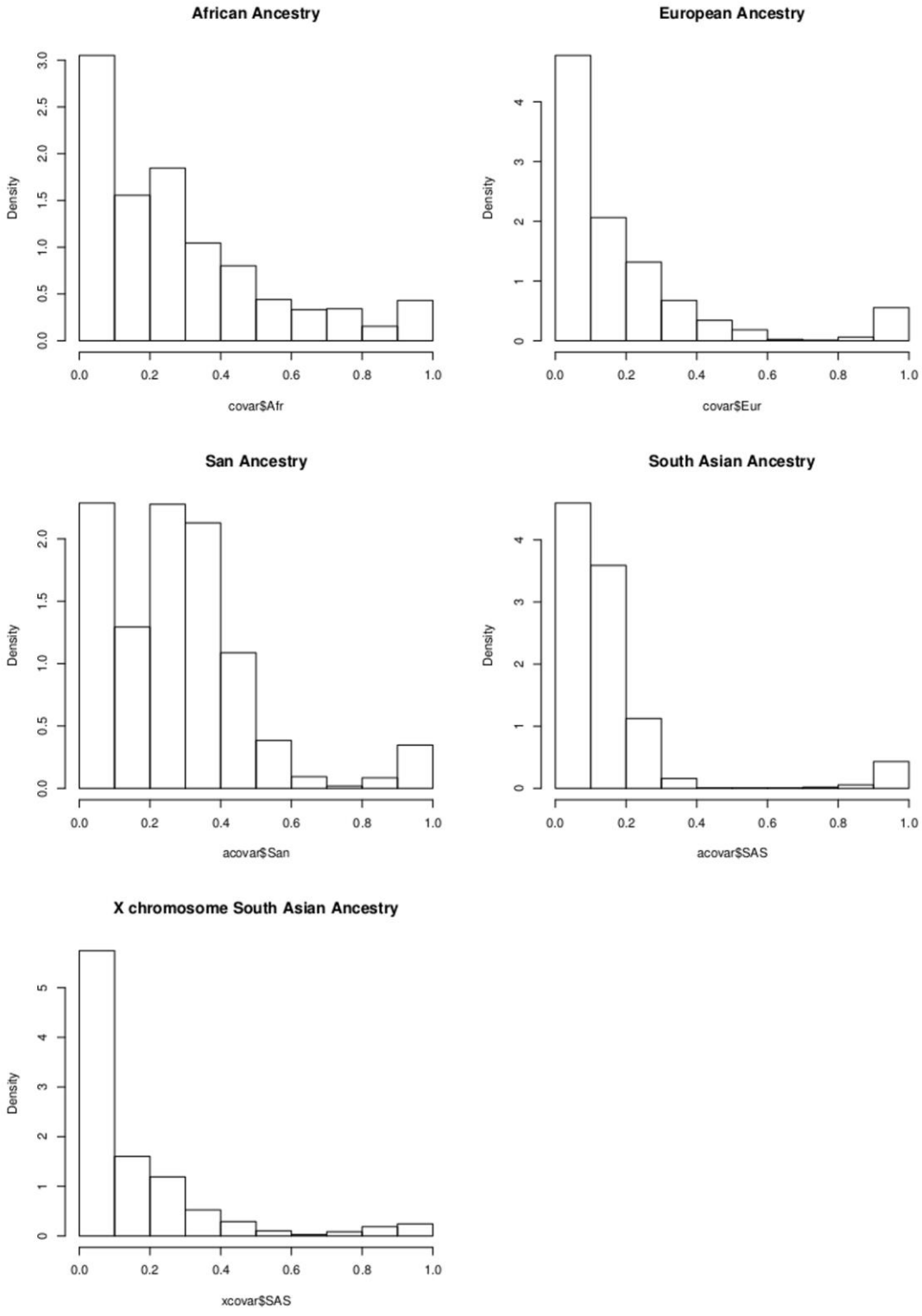


Figure S3.1: Histograms of the five ancestral components indicating data that is not normally distributed.

4 A sex-stratified genome-wide association study of tuberculosis using a multi-ethnic genotyping array

Haiko Schurz [1,2*], Craig J Kinnear [1], Chris Gignoux [3], Genevieve Wojcik [4], Paul D van Helden [1], Gerard Tromp [1,2,5], Brenna Henn [6], Eileen G Hoal [1], Marlo Möller [1]

[1] DST-NRF Centre of Excellence for Biomedical Tuberculosis Research; South African Medical Research Council Centre for Tuberculosis Research; Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

[2] South African Tuberculosis Bioinformatics Initiative (SATBBI), Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

[3] Colorado Center for Personalized Medicine and Department of Biostatistics and Informatics, University of Colorado, Anschutz Medical Campus, Aurora, CO, 80045.

[4] Department of Genetics, Stanford University, Stanford, CA, 94305.

[5] Centre for Bioinformatics and Computational Biology, Stellenbosch University, Cape Town, South Africa.

[6] Department of Anthropology, and the UC Davis Genome Center, University of California, Davis, CA, 95616.

* Corresponding author: haiko@sun.ac.za

4.1 Abstract

Tuberculosis (TB), caused by Mycobacterium tuberculosis, is a complex disease with a known human genetic component. Males seem to be more affected than females and in most countries the TB notification rate is twice as high in males as in females. While socio-economic status, behaviour and sex hormones influence the male bias they do not fully account for it. Males have only one copy of the X chromosome, while diploid females are subject to X chromosome inactivation. In addition, the X chromosome codes for many immune-related genes, supporting the hypothesis that X-linked genes could contribute to TB susceptibility in a sex-biased manner. We report the first TB susceptibility genome-wide association study (GWAS) with a specific focus on sex-stratified autosomal analysis and the X chromosome. A total of 810 individuals (410 cases and 405 controls) from an admixed South African population were genotyped using the Illumina Multi Ethnic Genotyping Array, specifically designed as a suitable platform for diverse and admixed populations. Association testing was done on the autosome (827386 variants) and X chromosome (20939 variants) in a sex stratified and combined manner. SNP association testing was not statistically significant using a stringent cut-off for significance but revealed likely candidate genes that warrant further investigation. A genome wide interaction analysis detected 16 significant interactions. Finally, the results highlight the importance of sex-stratified analysis as strong sex-specific effects were identified on both the autosome and X chromosome.

4.2 Key words

Tuberculosis, GWAS, Sex-bias, Host genetics, X chromosome, sex-stratified, susceptibility

4.3 Introduction

Tuberculosis (TB) caused by *Mycobacterium tuberculosis* (*M. tuberculosis*) is a global health epidemic and the leading cause of death due to a single infectious agent (246). In 2016 1.3 million TB deaths were reported in HIV negative individuals and an additional 374000 deaths related to TB/HIV co-infection were recorded. The majority of these deaths occurred in southeast Asian and African countries (246). TB is a complex disease, influenced by environmental and behavioural factors such as socio-economic status and smoking, as well as definite human genetic components. The contribution of the host genes to disease has been highlighted by numerous investigations, including animal (74), twin (59,289,290), linkage (70,265) and candidate gene association studies (291). More recently genome-wide association studies (GWAS) in diverse populations have been done (123–131,133).

Interestingly another influential factor in TB disease development is an individual's biological sex, which has been largely ignored in past TB studies and was usually only used as a covariate for adjusting association testing statistics. In 2016, males comprised 65% of the 10.4 million recorded TB cases, indicating that the TB notification rate is nearly twice as high in males as in females (WHO,

2017). While socio-economic and behavioural factors do influence this ratio, it does not fully explain the observed sex-bias (140). Another factor that influences sex-bias is the effect that sex hormones (estrogen and testosterone) have on the immune system. Estrogen is an immune activator, upregulating pro-inflammatory cytokines (TNFα), while testosterone is an immune suppressor, upregulating anti-inflammatory cytokines (IL-10) (171). This could explain why men are more susceptible to infectious diseases compared to females (140). However, as sex-based differences in immune responses differ even between pre-pubertal boys and girls, as well as between post-menopausal women and elderly men, it shows that sex hormones do not fully explain the sex-bias (144). Thus, it has been proposed that the X chromosome and X-linked genes directly contribute to the observed sex-bias.

There are approximately 1500 genes on the X chromosome, many of which are involved in the adaptive or innate immune system (161). Since females have two X chromosomes, one requires silencing in order to equalise dosage of gene expression to that of men who only have one X chromosome. This silencing occurs randomly in each cell, making females functional mosaics for X linked genes and giving them a major immunological advantage over males (140). As males are haploid for X-linked genes any damaging polymorphisms or mutations on the X chromosome will have a more pronounced immunological effect in males than in mosaic females, thereby influencing the sex-bias (160).

To date, eleven GWAS investigating susceptibility to clinical TB have been published (123–131,133,292). There has not been significant overlap between the 11 published TB GWAS, but it seems that replication is more likely when populations with similar genetic backgrounds are compared: the *WT1* locus was associated with disease in populations from West and South Africa (130,133). Critically, genotyping microarrays that did not fully accommodate African genetic diversity were used in these studies (123,124,130,131,133). It is therefore possible that unique African-specific susceptibility variants were not tagged by these initial arrays, since linkage disequilibrium (LD) blocks are shorter in African populations (293). Moreover, none of the GWAS included or examined the X chromosome or sex-stratified analysis of the autosomes as was done in an asthma cohort (294). Genetic differences between asthmatic males and females were identified on the autosome, with certain alleles having opposite effects between the sexes. Candidate gene association studies provide independent confirmation of the involvement of the X chromosome in TB susceptibility, through the association of X-linked *TLR8* susceptibility variants with active TB. Davila *et al.* (295) investigated 4 *TLR8* variants (rs3761624, rs3788935, rs3674879, rs3764880) in an Indonesian cohort and showed that all variants conferred susceptibility to TB in males but not females. The results for males were validated in male Russian individuals (295). These results were validated for rs3764880 in Turkish children, but no significant association was found for rs3764879 (296). Hashemi-Shahri *et*

al. (136) found no significant *TLR8* associations in an Iranian population, while rs3764880 was significantly associated with TB susceptibility in both males and females in a Pakistani cohort (132). In admixed South African Coloured (SAC) individuals rs3764879 and rs3764880 were significantly associated in both males and females, while rs3761624 was only significantly associated in females (138). Interestingly, in this cohort opposite effects were consistently found between the sexes for the same allele in all investigated *TLR8* variants (138), echoing the asthma findings of Mersha et al (294). Finally, in a Chinese cohort rs3764879 was significantly associated with TB disease in males but not females. While many of these variants did not reach genome wide significance they still provide evidence of the involvement of X-linked genes in TB susceptibility.

We report the first TB susceptibility GWAS with a specific focus on sex-stratified autosomal analysis and the X chromosome to elucidate the male sex-bias. Individuals from the unique five-way admixed South African Coloured (SAC) population, with ancestral contributions from Bantu-speaking African, KhoeSan, European, South and East Asian groups were genotyped in this study (86,297). These genetic contributions are due to both the complex colonisation history of South Africa and the country's importance as a refreshment station on major trade routes during the fifteenth to nineteenth century (103,298). This is therefore the first GWAS in the SAC that uses an array (Illumina Multi Ethnic Genotyping Array, see Section 2.2) specifically designed to detect variants in the 4 most commonly studied populations, making it the most suitable platform for diverse and admixed populations at the time of genotyping.

4.4 Materials and methods

4.4.1 Study population

Study participants were recruited from two suburbs in the Cape Town metropole of the Western Cape. These suburbs were chosen for its high TB incidence and low HIV prevalence (2%) at the time of sampling (1995-2005) (299). Approximately 98% of the residents in these suburbs self-identify as SAC and have similar socio-economic status, which reduces confounding bias in the association testing (133). The cohort consists of 420 pulmonary TB (pTB) cases, bacteriologically confirmed to be culture and/or smear positive and 419 healthy controls from the same suburbs. Approximately 80% of individuals over the age of 15 years from these suburbs have a positive tuberculin skin test (TST) , indicating exposure to *M. tuberculosis* (300). All study participants were over 18 years of age and HIV negative.

Approval was obtained from the Health Research Ethics Committee of Stellenbosch University (project registration number S17/01/013 and 95/072) before participant recruitment. Written informed consent was obtained from all study participants prior to blood collection. DNA was extracted from the blood samples using the Nucleon BACC Genomic DNA extraction kit (Illumina, Buckinghamshire, UK). DNA concentration and purity were checked using the NanoDrop® ND-1000 Spectrophotometer

and NanoDrop® v3.0.1 software (Inqaba Biotechnology, Pretoria, SA). The study adhered to the ethical guidelines as set out in the “Declaration of Helsinki, 2013 (301).

4.4.2 Genotyping

Genotyping was done using the Illumina Multi-ethnic genotyping array (MEGA) (Illumina, Miami, USA) which contains 1.7 million markers from various ethnicities making it highly suitable for diverse and admixed populations. The array is based on novel variants identified by the Consortium on Asthma among African ancestry populations in the Americas (CAAPA), the Illumina human core content for European and Asian populations as well as multi-ethnic exome content from African, Asian and European populations. The array also contains ancestry informative markers specific to the SAC population. While the KhoeSan population is not highly represented on the array, which could lead to a certain level of ascertainment bias, at the time of genotyping it was the most suitable platform for this diverse and admixed populations. Genome studio v2.04 (Illumina, Miami, USA) was used for SNP calling to calculate intensity scores and to call common variants (MAF >= 5%), followed by analysis with zCall to recall rare genotypes (MAF <5%) (283).

4.4.3 Genotyping quality control

Quality control (QC) of the genotyping data was done using the XWAS version 2.0 software and QC pipeline to filter out low quality samples and SNPs (115,302). Data were screened for sex concordance, relatedness (up to third degree of relatedness) and population stratification (as determined by principal component analysis). Genotypes for males and females were filtered separately in order to maintain inherent differences between the sexes. SNPs were removed from the analysis if missingness correlated with phenotype (threshold = 0.01) as well as individual and SNP missingness (greater than 10%), minor allele frequency (less than 1%) and Hardy–Weinberg equilibrium (HWE) in controls (threshold = 0.01). Filtering continued iteratively until no additional variants or individuals were removed. Overlapping markers between the sexes were merged into a single dataset. X chromosome genotypes were extracted and variants were removed if the MAF or missingness was significantly different between the sexes (threshold = 0.01). A flow diagram explaining quality control steps and association testing of the data is shown in Figure S4.1.

4.4.4 Admixture

The SAC population is a 5-way admixed population with ancestral contributions from Bantu-speaking African populations, KhoeSan, Europeans and South and East Asians. To avoid confounding during association testing the ancestral components are included as covariates (99). Admixture was estimated for the autosome (chromosome 1-22) and the X chromosome separately using the software ADMIXTURE (v1.3) (303) and reference genotyping data for 5 ancestral populations. The reference populations used to infer ancestry were European (CEU) and South Asian (Gujarati Indians in Houston, Texas and Pathan of Punjab) extracted from the 1000 Genomes Phase 3 data (81), East

Asian (Han Chinese in Beijing, China), African (Luhya in Webuye, Kenya, Bantu-speaking African, Yoruba from Nigeria) and San (Nama/Khomani) (103,282). Due to the limited number of individuals availablefor each reference population the SAC data had to be divided into 21 groups to equal the number of individuals per reference population. The number of individuals per reference population and admixed population has to be kept consistent in order to maximise the accuracy of the admixture results by not over-representing one particular population in the analysis. Therefore, admixture inference was done separately for each of the 21 SAC groups, referred to as running groups. Each running group was analysed five times at different random seed values. The results for each individual were averaged across the five runs in order to obtain the most accurate ancestry estimations (286). Four ancestral components (African, San, European and South Asian(297)) were included as covariates in the logistic regression association testing with the smallest component (East Asian) excluded in order to avoid complete separation of the data.

4.5 Association analysis

4.5.1 SNP based association analysis

Autosomal TB association testing was done with sex-stratified and combined datasets using the additive model in PLINK (version 1.7[4]) (304) in order to detect sex-based differences. TB association testing for the X chromosome were done separately in males and females using XWAS (version 2) and the results were combined using Stouffers method in order to obtain a combined association statistic (115,302). A sex-differentiated test was conducted for the X chromosome using the XWAS software to test for significant differences in genetic effects between males and females. SNP based association testing (sex-stratified or not) compares the frequency of alleles between cases and controls to determine if a specific allele co-occurs with a phenotype (TB) more often than would be expected by chance. The sex-differentiation test on the other hand compares the effect size (OR) of a variant between the sexes to determine if a variant has a different effect on risk between the sexes. The sex-differentiation test is explained in more detail by Chang *et al.* (302). X chromosome inactivation states were also included in the association testing as covariates using a method developed by Wang *et al.* (305). To include inactivation states in the association analysis the most likely state was determined for each SNP. A variant can either be inactivated, or it can be skewed towards the deleterious or normal allele or the variant can escape inactivation. To determine which of the four states is most probable the likelihood ratio for each one was calculated and the inactivation state that maximised the likelihood ratio was applied to the SNP in question. This was done for each variant as inactivation states vary along the X chromosome (for a detailed description see Wang *et al.* (305)). Ancestry, sex and age were included in the analyses as covariates where applicable. Information on other risk factors known to influence TB susceptibility such as smoking, and alcohol

[4]http://pngu.mgh.harvard.edu/purcell/plink/

consumption was not available for this study cohort and could not be included as covariates. Multiple testing correction was done using the SimpleM method (306), which adjusts the significance threshold based on the number of SNPs that explains 95% of the variance in the study cohort. This method is less conservative than Bonferroni correction and is a close approximation of permutation results in a fraction of the time. For the autosome the genome-wide significance threshold was set to $5.0e^{-8}$ (307).

4.5.2 Gene based association analysis

Gene-based association testing groups SNPs together and thus decreases the multiple testing burden and increase power to detect an association. Gene-based association testing was done using the XWAS v2 scripts, which were implemented using the Python[5] (version 2.7.10) and R programming environment (version 3.2.4, (142)) and R packages corpcor and mvtnorm. Reference files for the known canonical genes on the X chromosome for human genome build 37 were included in the XWAS v2 software package and used to group variants and p-values by gene (115,302). Bonferroni correction was used to adjust for multiple testing instead of SimpleM, as all genes, unlike SNPs, are independent of each other in the context of association testing and as such the multiple test correction cannot be less than the number of genes tested.

4.5.3 Interaction analysis

Genome-wide SNP interaction analysis was done using CASSI[6] (v2.51). A joint effects model was implemented for a rapid overview of interactions of all variants across the genome (autosome and X chromosome). Variants from significant interactions were reanalysed using a logistic regression approach with covariate correction, which would not be feasible for a genome-wide interaction analysis as it would be too computationally intensive. As there is no consensus on the significance threshold for genome wide interaction analysis Bonferroni correction was used in order to avoid potential inflation of false positive results.

4.6 Results

4.6.1 Cohort summary

In total 410 TB cases and 405 healthy controls passed the sex-stratified QC procedure. General summary statistics for the cohort, including mean and standard deviation of age and global ancestry as well as the ratio of males to females in both cases and controls are shown in Table 4.1. Clear differences were observed between TB cases and controls for both age and ancestry, justifying the inclusion as covariates. Ancestral distributions were compared using the Wilcoxon signed-rank test and were shown to significantly differ (unpublished results) between the autosome and X chromosome (Figure 4.1). Y chromosome and mitochondrial haplogroup analysis also revealed strong sex biased admixture in the SAC population, with a strong female KhoeSan and male Bantu-

[5] http://www.python.org
[6] https://www.staff.ncl.ac.uk/richard.howey/cassi/using.html

speaking African and European bias (278). As sex biased ancestry has been shown to reflect in strong differences between the autosomal and X chromosome ancestral components they were included as covariates in the respective analyses (156–158).

Table 4.1: SAC sample characteristics showing case/control and sex distribution, mean and standard deviation of age and global ancestral components

Group	Number	Female (%)	Age	San	African	European	South Asian	East Asian
TB cases	410	242 (59)	36.32 ± 11.04	33.89 ± 18.83	29.11 ± 19.80	17.10 ± 16.81	12.95 ± 10.82	7.08 ± 7.22
Controls	405	223 (55)	30.55 ± 12.91	33.75 ± 19.59	29.92 ± 20.46	16.12± 15.81	13.26 ± 12.38	7.04 ± 7.44

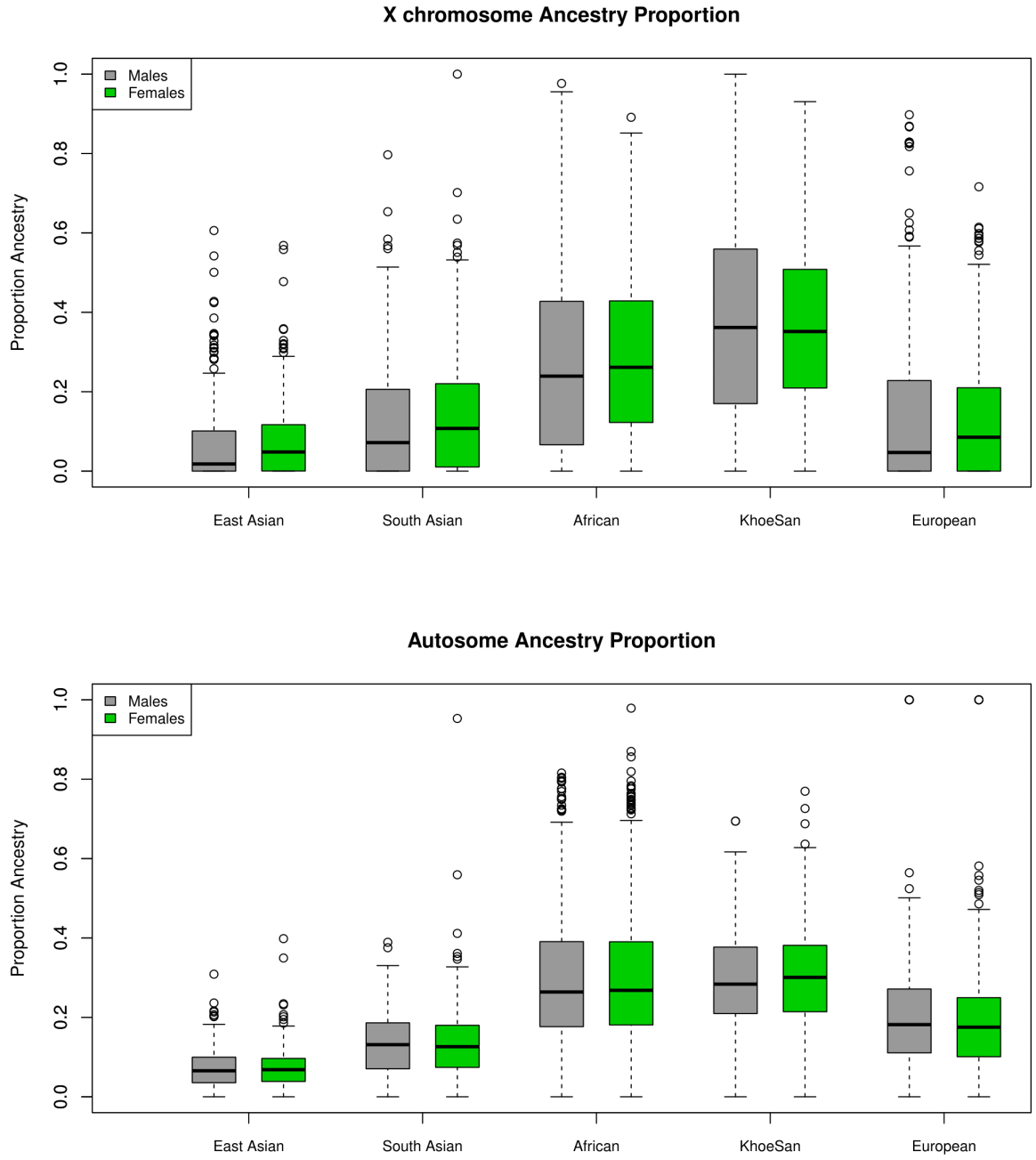


Figure 4.1: Ancestral distribution on the X chromosome and autosome for males and females.

4.6.2 Association testing results

4.6.2.1 SNP based

The top results for the autosomal association testing are shown in Table 4.2 and Figure S4.2, with the QQ-plot indicating no constraints on the analysis or inflation of the results (Figure S4.2). Following multiple test correction, no significant associations were identified for the combined or sex-stratified analysis, but it is important to note that the top associations differed between the sex-stratified and combined analyses as well as for males and females (Table 4.2). The top hit for the combined

autosomal association test was rs17410035 (OR= 0.4, p-value = $1.5e^{-6}$, Table 4.2), located in the 3'-UTR of the *DROSHA* gene, which encodes a type 3 RNase. This RNase is involved in miRNA processing and miRNA biogenesis (308). Although little evidence exists that rs17410035 has an impact on *DROSHA* gene expression or miRNA biogenesis (which could affect gene expression) it has been associated with increased colon cancer (OR= 1.22, p-value = 0.014) (308) and cancer of the head and neck (OR= 2.28, p-value = 0.016) (309). When the rs17410035 SNP interacts with other variants (rs3792830, rs3732360) it can further increase the risk for cancer of the head and neck (309), which illustrates the importance of doing interaction analysis. For the autosomal sex-stratified analysis the variant with the lowest p-value in males was rs11960504 (OR = 2.8, p-value = $7.21e^{-6}$, Table 4.2) located downstream of the *GRAMD2B* gene, a gene for which no information is available. The top hit in the females was rs2894967 (OR = 2.17, p-value = $4.77e^{-6}$) located upstream of the *TENT4A* gene, a gene coding for a DNA polymerase shown to be involved in DNA repair (310). Closer inspection of the data revealed that the effects between the sexes were in the same direction for all top hits in the combined analysis, whereas all variants identified in the sex-stratified analysis had effects in opposite directions between the sexes, or one sex had no effect, indicating that even on the autosome strong sex specific effects are prominent.

Table 4.2: Top associations for the combined and sex-stratified autosomal association testing.

Chr	SNP	A1	Location	Gene	Group	OR*	95CI*	P-value
5	rs17410035	T	5'UTR	*DROSHA*	Combined	0.404	0.28-0.58	$1.50e^{-6}$
5	rs1501847	G	5'UTR	*C5orf64*	Combined	1.708	1.37-2.14	$2.64e^{-6}$
7	rs2665441	C	3'UTR	*ASNS*	Combined	1.681	1.34-2.10	$5.51e^{-6}$
9	rs1662230	G	5'UTR	*RN7SKP120*	Combined	2.278	1.58-3.27	$8.91e^{-6}$
12	rs199911028	G	Intronic	*CFAP54*	Combined	2.966	1.89-4.67	$2.58e^{-6}$
15	rs142644068	C	Intronic	*PCSK6*	Combined	0.132	0.06-0.30	$1.56e^{-6}$
5	rs11960504	T	3'UTR	*GRAMD2B*	Male	2.801	1.79-4.39	$7.21e^{-6}$
13	rs9315991	A	Intronic	*LINC00400*	Male	0.394	0.27-0.58	$2.03e^{-5}$
14	rs8016621	A	Intronic	*SALL2*	Male	0.252	0.14-0.46	$5.91e^{-6}$
5	rs2894967	C	5'UTR	*TENT4A*	Female	2.173	1.56-2.90	$4.77e^{-6}$
9	rs10819610	T	Intronic	*NCS1*	Female	0.514	0.39-0.67	$1.55e^{-6}$
14	rs7152005	T	Intronic	*DPF3*	Female	2.13	1.66-2.90	$1.52e^{-6}$
21	rs2150367	T	Intronic	*LINC02246*	Female	0.502	0.38-0.67	$1.56e^{-6}$

*OR: Odds ratio; 95% CI: 95% confidence interval

For the X chromosome specific association testing a sex-stratified test was conducted and the results were then combined using Stouffers method, which provided a good fit between expected and observed p-values (QQ-plot Figure 4.3) (115,302). The simpleM method indicated that of the 20939

X-linked variants 17600 explained 95% of the variance in the data resulting in a significance threshold of $2.8e^{-6}$ (0.05/17600). No statistically significant associations with TB susceptibility were identified in either sex-stratified or the combined analysis (Table 4.3 and Figure 4.3). The lowest p-value association for the X-linked combined (p-value = $2.62e^{-5}$) and females (OR = 1.83, p-value = $1.06e^{-4}$) only analysis was the same variant, rs768568, located in the *TBL1X* gene. For the males the top hit was rs12011358 (OR = 0.37, p-value = $1.25e^{-4}$) located in the *MTND6P12* gene. Both of these genes have not been previously associated with TB susceptibility and *MTND6P12* is a pseudogene with unknown expression patterns or function. Variants in *TBL1X* have been shown to influence prostate cancer (311) and central hypothyroidism (312) susceptibility. *TBL1X* is a regulator of nuclear factor kappa-light-chain-enhancer of activated B cells (NF-kB) and is thus involved in the immune system which could impact TB susceptibility.

The method of modelling X chromosome inactivation states, developed by Wang *et al.* (305), was also incorporated into the X-linked association testing, but no significant observations were observed. Although the p-values were generally lower than for the Stouffer method, the QQ-plot revealed that including estimations of X chromosome inactivation states inflated the p-values and increased the chance of type 1 errors and these results were therefore discounted (Table S4.1 and Figure S4.3).

Table 4.3: Most significant X-linked associations, using Stouffers method to combine p-values.

Chr	SNP (Location)	A1	Gene	Male			Female			P combined
				OR*	95CI*	P	OR*	95CI*	P	
X	rs768568 (Intron)	C	*TBL1X*	1.69	1.0-2.86	$5.07\ e^{-2}$	1.83	1.35-2.49	$1.06\ e^{-4}$	$2.62e^{-5}$
X	rs12011358 (5'UTR)	T	*MTND6P12*	0.37	0.22-0.62	$1.25\ e^{-4}$	0.72	0.53-0.96	$2.72\ e^{-2}$	$2.82e^{-5}$
X	rs930631 (3'UTR)	T	*MIR514A1*	0.48	0.29-0.79	$3.66\ e^{-3}$	0.67	0.49-0.90	$7.74\ e^{-3}$	$8.94e^{-5}$

*OR: Odds ratio; 95% CI: 95% confidence interval

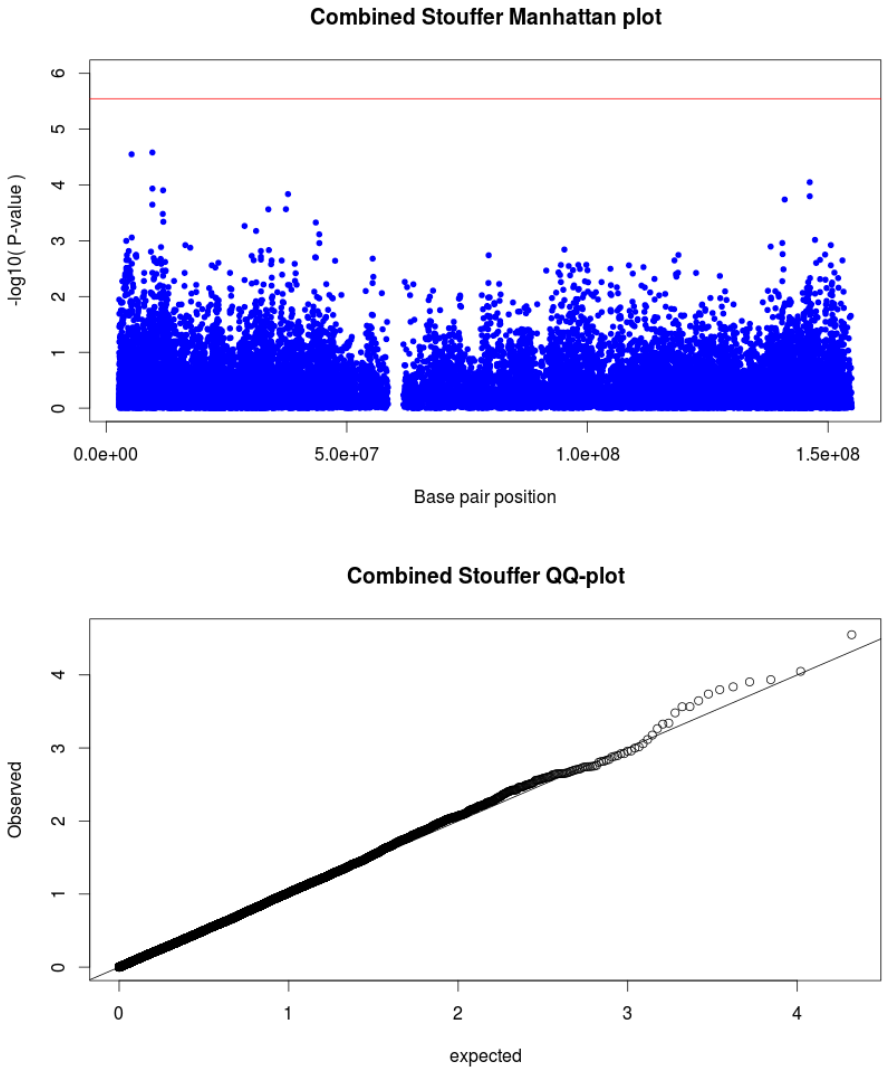


Figure 4.2: Manhattan plot (above) for X-linked associations with significance threshold indicated (red line). QQ-plot (below) shows good correlation between expected and observed p-values.

The sex differentiation test did not result in any significant associations (Table 4.4), with the association with the lowest p-value was located in a pseudogene, *RNU6-974P* (p-value = $8.33e^{-5}$). The second lowest p-value was for a variant upstream of the *SRPX* (p-value = $2.18e^{-4}$) gene which has previously been shown to have a tumour suppressor function in prostate carcinomas (313). Whether these variants are associated with TB susceptibility or influence sex-bias is unclear, but the vastly opposite effects between the sexes are noteworthy. When comparing the OR for the sex differentiation test it is clear that variants can have major sex specific effects again highlighting the need for sex-stratified analysis (Table 4.4).

Table 4.4: Sex-differentiation analysis results.

Chr	SNP	A1	Location	Gene	Male			Female			P-Diff
					OR*	95CI*	P	OR*	95CI*	P	
X	rs145407087	C	3'UTR	*RNU6-974P*	0.427	0.12-1.54	0.193	7.147	1.95-26.19	0.003	$8.33e^{-5}$
X	rs5917743	C	5'UTR	*SRPX*	0.300	0.05-1.89	0.203	15.04	1.81-124.7	0.012	$2.18e^{-4}$
X	rs1337567	C	5'UTR	*DIAPH2*	1.45	0.88-2.39	0.146	0.571	0.42-0.78	$3.49e^{-4}$	$6.73e^{-4}$

*OR: Odds ratio; 95% CI: 95% confidence interval

4.6.2.2 Gene based

The X chromosome gene-based analysis, in which 1105 X-linked genes were analysed did not show any significant associations using a Bonferroni-adjusted significance threshold of $4.5e^{-5}$ (Table 4.5). The top hit for the combined analysis was in the chromosome X open reading frame 51B (*CXorf51B)* (p-value = $=1.28e^{-4}$) coding for an uncharacterised protein (LOC100133053). The top hit for males was in an RNA coding region that interacts with Piwi proteins (*DQ590189.1,* p-value = $1.7e^{-3}$), a subfamily of Argonaute proteins. While Piwi proteins are involved in germline stem cell maintenance and meiosis the function of the Piwi interacting RNA molecules are unknown (314). For females the top hit was the *ARMCX1* gene (p-value = $6.07e^{-4}$) a tumour suppressor gene involved in cell proliferation and apoptosis of breast cancer cells. While this gene has not been previously implicated in TB susceptibility, *M. tuberculosis* has been shown to affect apoptosis pathways in order to evade the host immune response, suggesting that *ARMCX1* could affect TB susceptibility (315). While not significant the analysis again reveals strong sex specific effects and the sex-stratified and combined analysis gave three different results (Table 4.5).

Table 4.5: X chromosome gene-based association results.

Chr	Gene	Group	P-value
X	*CXorf51B*	Combined	$1.28e^{-4}$
X	*DQ580189.1*	Male	$1.7e^{-3}$
X	*ARMCX1*	Female	$6.07e^{-4}$

4.6.2.3 Interaction analysis

A genome-wide interaction analysis was performed using the software Cassie. In total 1893973105 interactions were analysed and following a Bonferroni correction for the number of interactions performed the significance threshold was set to $2.6e^{-11}$. For the joint effects model, 18 interactions passed the significance threshold (Table S4.2). The interaction with lowest p-value was between rs1823897 upstream of the *ARSF* gene and rs7064174 in the *FRMPD4* gene (p-value = $7.23e^{-14}$), two

genes for which not much information is available, and it is unclear how they could be involved in TB susceptibility. The top 450 associations from the joint effects model were then retested using logistic regression and the same covariates as the SNP based association testing. No significant interactions (threshold of $2.6e^{-11}$) were observed in the logistic regression model (Table 4.6), but as Bonferroni correction is very conservative the top interactions should still be considered as they reach the significance level for SNP based GWAS.

Among the top hits in the logistic regression analysis (Table 4.6) some could impact TB susceptibility as they are involved in immune functions. The top hit interaction was between rs2631914 (*LINCO2153),* which is upregulated in people with major depressive disorder (316) and rs8067702 (*RTN4RL1),* previously associated with congenital heart disease, microcephaly and mild intellectual disability (317). While this interaction is not very informative in the context of TB three other interactions were identified that could impact TB susceptibility (Table 4.6).

The first interaction of interest is between *RNF125* gene (rs35996537) and *URI1* (rs1118924), involved in downregulation of CD^{4+}/CD^{38-} T-cells and PBMCs in HIV-1 positive individuals and NF-kB/CSN2/Snail pathway, activated by TNFα respectively (318,319). Second the interaction between rs386560079 (*ATP2C1),* which is involved in regulation of intracellular Ca^{2+}/Mn^{2+} concentrations through the Golgi apparatus (320) and rs6498130 (*CIITA).* Variants in the *CIITA* gene reduce the expression of *MHC class II* proteins and receptors resulting in an immune privilege phenotype (321). The final interaction of interest is between rs12286374 (*NTM),* which is mainly expressed in the brain and promotes neurite outgrowth and adhesion (322) and rs2040739 (*RNF126)* a ring type E3 ligase involved in the Protein B kinase pathway which has been previously implicated in glucose metabolism, apoptosis, cell proliferation and transcription (323). While none of these genes have previously been implicated in TB susceptibility the fact that some of them are involved in immune functions suggests a role in TB susceptibility

Table *4.6:* Logistic regression interaction analysis with covariate adjustment.

Chr1	SNP1	Location	Gene1	Chr2	SNP2	Location	Gene2	P
8	rs2631914	5'UTR	*LINC02153*	17	rs8067702	3'UTR	*RTN4RL1*	$1.73e^{-9}$
2	rs6756958	Intronic	*GALNT5*	4	rs201376793	5'UTR	*RNU6ATAC 13P*	$1.92e^{-9}$
4	rs882773	5'UTR	*HMX1*	18	rs9303903	3'UTR	*METTL4*	$5.76e^{-9}$
12	rs1798087	5'UTR	*TSPAN1*	13	rs2091337	3'UTR	*LOC105370 290*	$5.8e^{-9}$
1	rs7517749	Intronic	*RGS7*	23	rs5907910	3'UTR	*SPANXA2-OT1*	$1.14e^{-8}$
7	rs757808	5'UTR	*KIAA0087*	8	rs12676973	3'UTR	*FUT10*	$1.77e^{-8}$

Chr1	SNP1	Location	Gene1	Chr2	SNP2	Location	Gene2	P
4	rs1919904	5'UTR	*TMPRSS11E*	11	rs10769029	Intronic	*ALX4*	$1.95e^{-8}$
5	rs10040477	5'UTR	*LINC02148*	12	rs1918193	intronic	*SYT1*	$2.82e^{-8}$
1	rs6694239	5'UTR	*TNR*	2	rs985256	intronic	*SPATS2L*	$3.22e^{-8}$
12	rs7975477	Intronic	*MGAT4C*	20	rs6123951	intronic	*PHACTR3*	$3.29e^{-8}$
18	rs35996537	3'UTR	*RNF125*	19	rs1118924	Intronic	*URI1*	$3.82e^{-8}$
5	rs10040477	5'UTR	*LINC02148*	12	rs1918195	Intronic	*SYT1*	$3.87e^{-8}$
3	rs386560079	Intron	*ATP2C1*	16	rs6498130	Intronic	*CIITA*	$3.94e^{-8}$
11	rs4237591	3'UTR	*CNTN5*	14	rs11850085	Intronic	*SLC8A3*	$4.54e^{-8}$
1	rs1411276	Intronic	*TGFBR3*	4	rs1972127	Intronic	*PRKG2*	$4.86e^{-8}$
12	rs7962106	5'UTR	*AVPR1A*	18	rs200219001	Intronic	*LDLRAD4*	$4.94e^{-8}$
14	rs242402	Intronic	*PELI2*	19	rs2459744	5'UTR	*SBK3*	$5.22e^{-8}$
11	rs12286374	5'UTR	*NTM*	19	rs2040739	Intronic	*RNF126*	$5.49e^{-8}$

4.7 Discussion

In this GWAS we investigated TB susceptibility in the admixed SAC population, with a specific focus on sex-bias and the X chromosome. A sex-stratified QC protocol was applied to the data in order to conserve inherent differences between the sexes and all statistical analysis were conducted in a sex-stratified and combined dataset in order to fully assess the impact of sex on TB susceptibility and the male sex-bias it presents with. We found no significant associations on the autosome or X chromosome for both the sex-stratified and combined SNP and gene-based association testing. A few significant interactions were identified, but the impact of these on TB susceptibility is unclear and will require further investigation to validate and functionally verify.

For the combined autosomal SNP based association testing the only potential variant of interest is rs17410035 located in the *DROSHA* gene (Table 4.2) which is potentially involved in miRNA biogenesis and could impact TB susceptibility if immune related regulatory miRNA is affected. For the X-linked association testing the association with the lowest p-value in males was in an uninformative pseudogene, while the female and combined analysis revealed the same variant, rs768568 located in the *TBL1X* gene (Table 4.3). The TBL1X protein has been shown to be a co-activator of NF-kB mediated transcription of cytokine coding genes, but the mechanism of activation is unclear (311). NF-kB is a vital component of the proinflammatory signalling pathway and is involved in multiple immune pathways including TLRs (324), which have previously been shown to influence TB susceptibility (291). Based on this one could extrapolate that variants in the *TBL1X* gene could affect activation and proinflammatory signalling of NF-kB, which could have a direct effect on the immune

system and thus TB susceptibility. The direction of effect for this variant was the same in males and females (Table 4.3), but was less significant in males probably due to loss of power when analysing haploid genotypes. For the variants identified in the sex differentiated analysis it is unclear how they could influence TB susceptibility as the top hit is located in a pseudogene. However, the sex differentiated test did reveal just how big the difference in effects can be between the sexes for a specific variant (Table 4.4). If these variants with opposite effects are not analysed in a sex-stratified way then the effects would cancel each other out and any information on sex specific effects would be lost. The X-linked gene-based association test revealed no significant associations despite having more power than the SNP based association testing. A possible reason for this could be that Bonferroni correction was used and as this is very conservative possible associations could have been missed. When looking at the associations with the lowest p-value (Table 4.5) however it is unclear how the identified genes could be implicated in TB susceptibility.

The joint effects interaction analysis revealed several significant interactions, but as association results have been previously shown to be severely influenced by admixture (325) only the results for the logistic regression analysis will be discussed here. A few variants were identified in the logistic interaction analysis that could impact TB susceptibility (Table 4.6). *URI1* (rs1118924) is activated by TNFα and is involved in the NF-kB/CSN2/Snail pathway, *CIITA* (rs6498130) impacts expression of MHC class II proteins and receptors and rs35996537 (*RNF125)* and rs2040739 *(RNF126)* are both E3 ubiquitin ligase proteins which affect a multitude of cellular functions, such as apoptosis (323) and protein degradation (326). NF-kB, TNFα, MHC class II, E3 ligases, apoptosis and T-cells have all been implicated in TB susceptibility and could collectively contribute by influencing the immune response (315,326–332). As TB is a complex disease all potential influential factors need to be considered and as such the interaction analysis cannot be ignored. Shortcomings of the interaction analysis are that they are very computationally intensive and suffer from a massive multiple test correction burden. Future research should thus focus on ways to prioritise variants for interaction analysis to decrease computation time as well as have sufficient sample size to minimise multiple test correction burden.

A previous GWAS in the SAC population found a significant association with TB susceptibility in the *WT1* gene (rs2057178, OR = 0.62, p-value = $2.71e^{-6}$) (333). This association did not reach genome-wide significance in our study (OR = 0.75, p-value = 0.049). At the time of the GWAS by Chimusa *et al*. (333) there were few African and KhoeSan (only 6 KhoeSan) individuals in the reference data used for imputation and the accuracy of imputation in this population was not known. As the identified variant (rs2057178) was imputed into the data it should have been validated in the SAC population using an appropriate genotyping approach. Secondly although the variant reached a significance threshold for the number of variants tested it did not reach genome wide significance threshold of $5.0e^{-8}$ (307). Finally, the GWAS performed by Chimusa *et al.* (333) only contained 91 control

individuals compared to 642 cases, which could affect the power of the study. Chimusa *et al.* (333) were unable to replicate previous associations identified in the X-linked *TLR8* gene (295). The two *TLR8* variants in our data, rs3764880 (OR = 1.73, p-value = $3.1e^{-4}$) and rs3761624 (OR = 1.70, p-value = $3.94e^{-4}$) also did not show significant associations. While the haploid genotypes in males contributes to this, a second influential factor could be admixture. Chimusa *et al.* (333) did not perform X chromosome specific admixture analysis, which could affect association testing of X-linked genes. Furthermore, only six KhoeSan reference individuals were available, which could affect the accuracy of admixture inference and severely affect the results. For our study 307 KhoeSan individuals were available, improving the admixture inference and could explain why stronger effects (higher OR) were detected for the *TLR8* variants when compared to Chimusa *et al.* (333). It is also important to note that using global ancestry components as covariates does not correct for ancestry at any specific locus and as a result each locus in this population could have up to five different ancestries. This could greatly reduce power and contribute to the lack of replication between studies. In order to address this, future studies could incorporate local ancestry inference into the analysis in order to determine the number of ancestries at a locus of interest. Other candidate genes identified in previous GWAS studies were also separately analysed here, but associations did not replicate (Online supplementary material).

We did not find any significant associations with TB susceptibility but highlight the need for sex-stratified analysis. Closer inspection of the data revealed that a large number of SNPs with opposite direction of effects for not only the X chromosome, but the autosome too. Sex specific effects has previously been reported for autosomal variants associated with pulmonary function in asthma (334). In the SAC population these opposite effects have previously been observed for X-linked variants in the *TLR8* gene (138) and the same is observed in this study. Sex-stratified analysis should therefore be included in association studies and incorporated in the study design. This can be done by keeping the male to female ratio balanced in the cases and controls. It would also be prudent to do the power calculation for the males and females separately. This will ensure sufficient power for sex-stratified analysis and could elucidate informative sex specific effects. This study was done in a 5-way admixed population. As was observed for the interaction analysis including admixture components significantly changes the association results. Furthermore it was observed (unpublished results) that the ancestral distribution between the X chromosome and autosome are different (Figure 4.1), which is an indication of sex-biased admixture (279,286) and highlights the importance of including X chromosome admixture components for X-linked and sex-bias analysis. It is important to note here that the ancestral components in the SAC present with a very wide range (Figure 4.2) and all this variability could affect the power of association studies. It is therefore desirable to increase the sample size when analysing admixed individuals. Alternatively, a meta-analysis can be conducted, including data from all 5 ancestral populations, or local ancestry inference could be included in the analysis.

In conclusion, while no significant associations were identified this study shows the importance of conducting sex-stratified analysis. This analysis should be incorporated during the study design phase to ensure sufficient power and allow the inclusion of covariates with sex specific effects (in this case admixture components). The sex-stratified analysis revealed that the effect of certain variants can differ between males and females, not only for the X chromosome but also for the autosome. TB is a complex disease with most genetic associations that do not replicate across different populations, which complicates the elucidation of the genetic impact on disease susceptibility. By including sex-stratified analysis and identifying sex specific effects and the cause for the male bias we can adjust treatment according to sex and potentially improve treatment outcome and survival.

4.8 Acknowledgements

We would like to acknowledge and thank the study participants for their contribution and participation. 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. This work was also supported by the National Research Foundation of South Africa (grant number 93460) to EH. This work was also supported 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, GW) to GT.

4.9 Author contribution

HS, MM, CK, GT conceived the idea for this study. CG, GW, BH did the calling and QC of the raw genotyping data. HS did the analysis and wrote first draft. BH Assisted with admixture analysis. All authors contributed to writing and proofreading for approval of the final manuscript.

4.10 Conflict of interest

The authors report no conflict of interest

4.11 Supplementary material

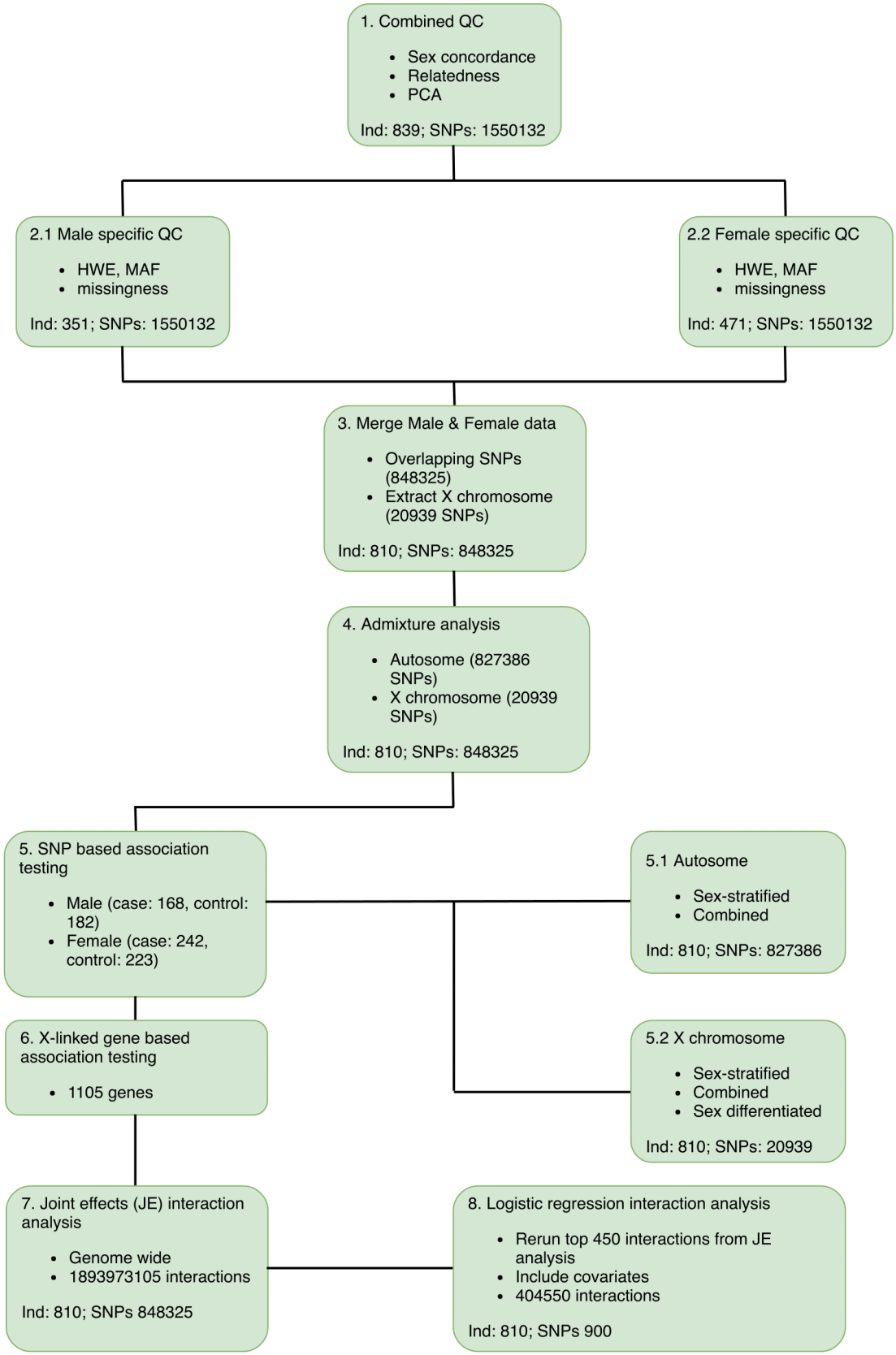


Figure S4.1: Flow diagram of data QC and association testing.

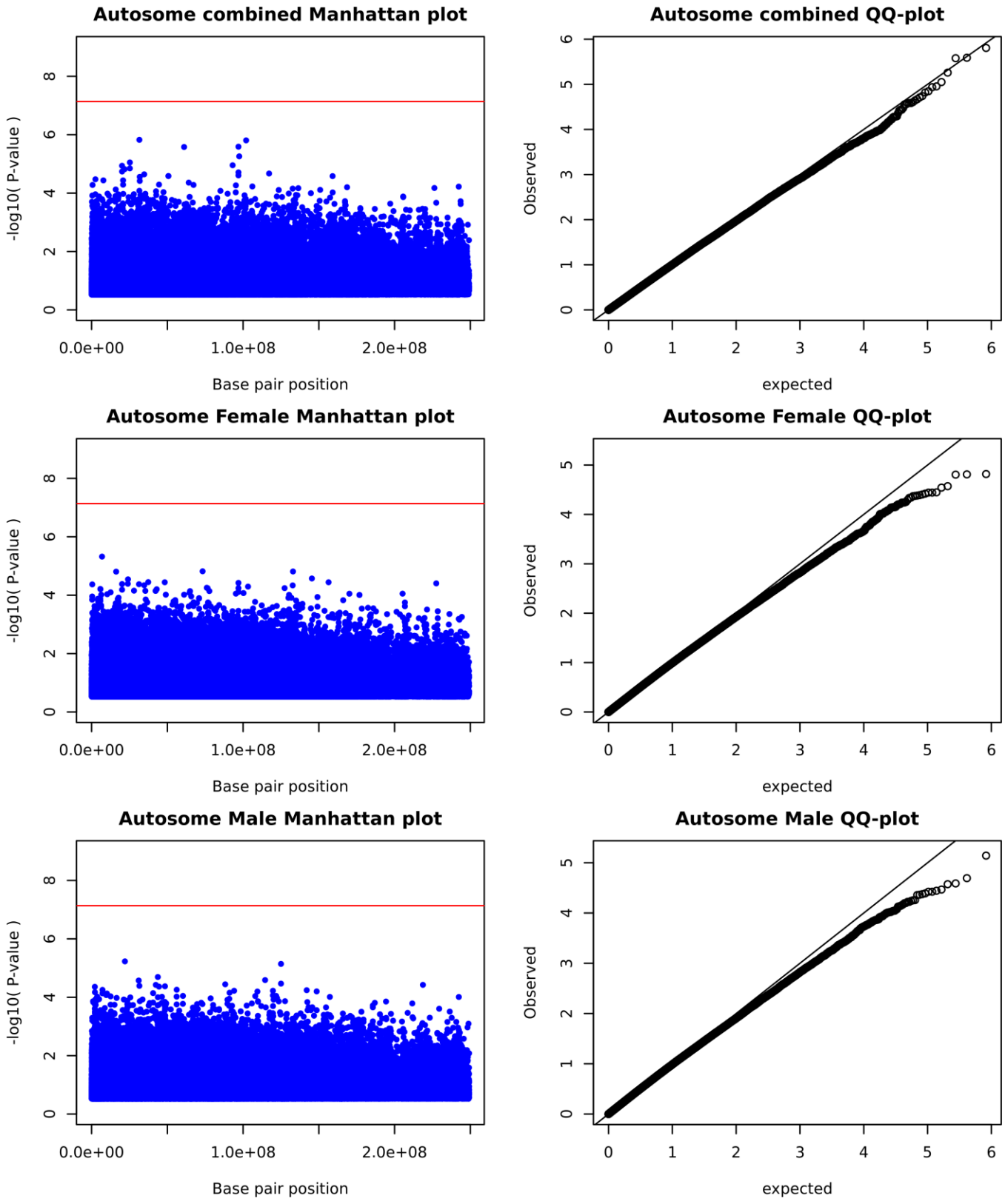


Figure S4.2: Manhattan plot and QQ plot for sex-stratified and combined analysis on the Autosome. Red line indicates significance threshold ($5e^{-8}$).

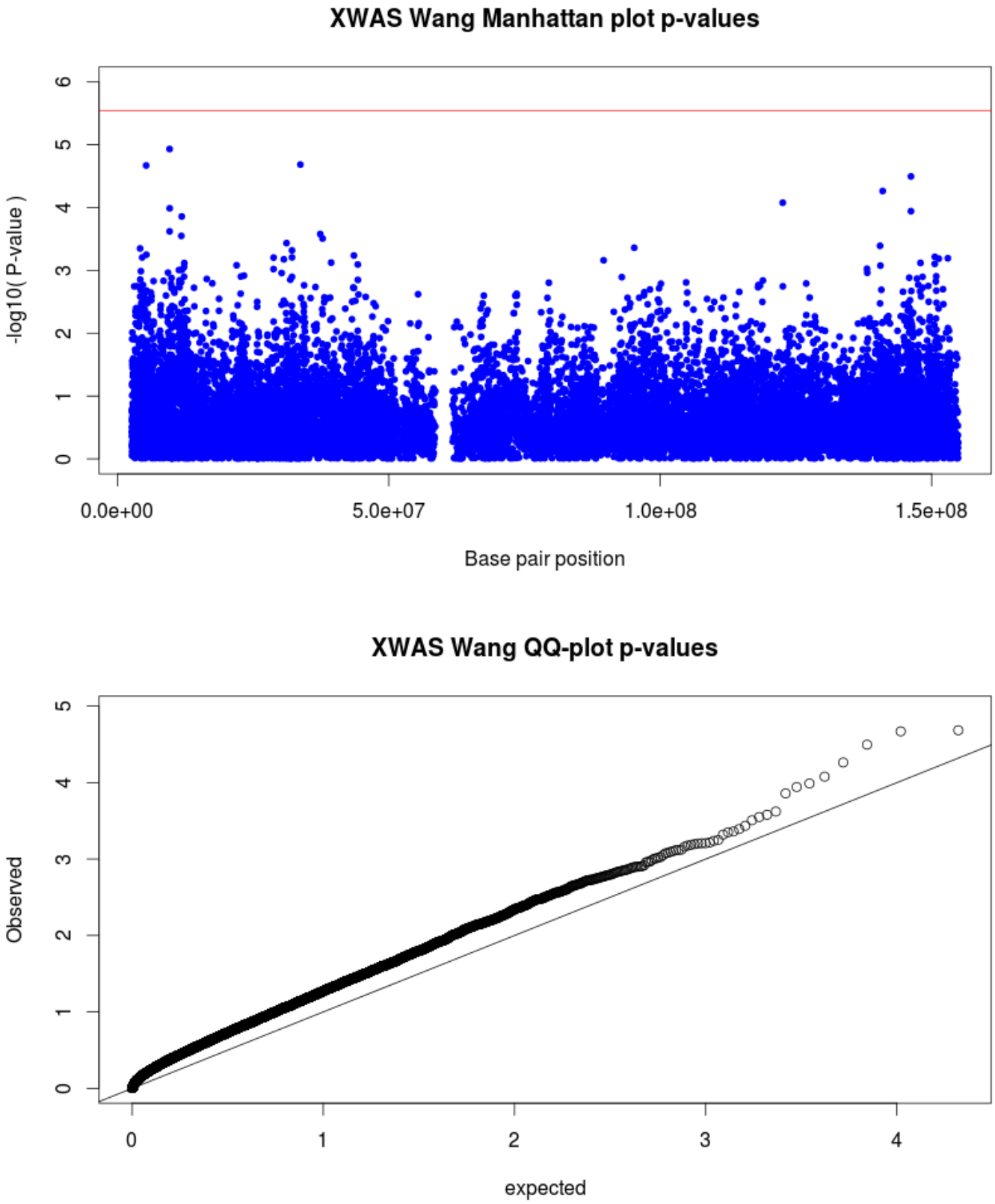


Figure S4.3: Manhattan and QQ-plot for X-linked SNP association testing including modelling for inactivation states, the red line indicates significance threshold of $2.8e^{-6}$. QQ-plot indicates inflated p-values and potential increase in type 1 errors.

Table S4.1: Most significant results for X-linked SNP association testing including modelling of X chromosome inactivation states.

SNP	Gene	Location	Model	OR	95CI-L	95CI-H	p
rs768568	TBL1X	Intron	Escape_of_XCI	0.56	0.43	0.72	$1.17e^{-5}$
rs6631824	DMD	5'UTR	Random_XCI	0.69	0.59	0.82	$2.07e^{-5}$
rs12011358	NLGN4X	3'UTR	Skewed_XCI_to_risk_allele	1.51	1.25	1.82	$2.14e^{-5}$
rs930631	MIR506	3'UTR	Random_XCI	1.41	1.20	1.65	$3.19e^{-5}$
rs176024	MAGEC3	5'UTR	Escape_of_XCI	0.56	0.43	0.74	$5.45e^{-5}$
23:9959944	SHROOM2	5'UTR	Skewed_XCI_to_risk_allele	0.21	0.07	0.49	$1.32e^{-3}$
rs386827412	GRIA3	Intron	Skewed_XCI_to_risk_allele	0.64	0.51	0.80	$8.35e^{-5}$
rs17340554	DMD	Intron	Skewed_XCI_to_risk_allele	0.35	0.19	0.61	$3.67e^{-4}$
rs5933749	TBL1X	Intron	Escape_of_XCI	1.72	1.31	2.27	$1.03e^{-4}$

Table S4.2: Results for genome wide interaction analysis using the joint effects model and no adjustment for covariates.

Chr1	SNP1	Gene1	Chr2	SNP2	Gene2	P-value
23	rs1823897	↑* ARSF	23	rs7064174	FRMPD4	$7.23e^{-14}$
23	rs426247	↑ ARSE	23	rs68046754	PCDH11X	$1.96e^{-12}$
5	rs2112508	LOC107986418	10	rs1194709	↑ RPL31P44	$2.81e^{-12}$
7	rs1347075	LOC107986770	8	rs958374	↑ LOC157273	$3.49e^{-12}$
23	rs5991619	↓* MAOA	23	rs73535318	RNU6-555P	$5.87e^{-12}$
13	rs7991005	GPC6	16	rs2287072	LPCAT2	$5.95e^{-12}$
23	Rs11094800	↓ NLGN4X	23	rs68046754	PCDH11X	$7.20e^{-12}$
5	rs7341174	LINC02147	12	rs1918191	SYT1	$7.63e^{-12}$
23	rs68046754	PCDH11X	23	rs6528958	↓ MAGEC2	$8.37e^{-12}$
1	rs6694239	↓ PAPPA2	2	rs985256	SPATS2L	$1.12e^{-11}$
23	rs4824843	↑ FUNDC1	23	rs58762927	LOC107986770	$1.29e^{-11}$
23	rs2170314	↑ DIAPH2	23	rs4633188	↑ DIAPH2	$1.32e^{-11}$
5	rs2112508	LOC107986418	10	rs1194716	↑ RPL31P44	$1.33e^{-11}$
23	rs5972637	DMD	23	rs73535318	RNY6-555P	$1.65e^{-11}$
23	rs1352015	EDA2R	23	rs68046754	PCDH11X	$1.75e^{-11}$
23	rs5916341	NLGN4X	23	rs68046754	PCDH11X	$2.50e^{-11}$
23	rs3072699	NLGN4X	23	rs68046754	PCDH11X	$2.50e^{-11}$
23	rs1823897	↑ ARSE	23	rs7053176	↑ CXCR3	$2.59e^{-11}$

↑*: Upstream 5'UTR; ↓*: Downstream 3'UTR

5 Evaluating the accuracy of imputation methods in a five-way admixed population.

Haiko Schurz[1,2,†] , Stephanie Pitts[1,2,†], Paul D. van Helden[1], Gerard Tromp[1,2], Eileen G Hoal[1], Craig J Kinnear[1]*, Marlo Möller[1]*

[1] DST-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

[2] South African Tuberculosis Bioinformatics Initiative (SATBBI), Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

† Co-first authors

* Co-senior authors

5.1 Abstract

Genotype imputation is a powerful tool for increasing statistical power in an association analysis. Meta-analysis of multiple study datasets also requires a substantial overlap of SNPs for a successful association analysis, which can be achieved by imputation. Quality of imputed datasets is largely dependent on the software used, as well as the reference populations chosen. The accuracy of imputation of available reference populations has not been tested for the five-way admixed South African Coloured (SAC) population. In this study, five methods were tested for imputation accuracy and quality, comprising three imputation software packages and three reference panels. We show that the African Genome Resource is the best reference panel for the SAC population, with implementation via the freely accessible Sanger Imputation Server.

5.2 Introduction

Over the past decade, genotyping technologies for genome-wide association studies (GWAS) have allowed for extensive and rapid genotyping of common variants (335–337). Commercial single nucleotide polymorphism (SNP) genotyping arrays contain between 300 000 and 2.5 million markers, but none have complete coverage of the human genome. Genotype imputation can be used to improve both coverage and power of a GWAS by inferring the alleles of un-genotyped SNPs based on the linkage disequilibrium (LD) patterns derived from directly genotyped markers and comparing them to a suitable reference population (338–340). These imputed variants can then be used for association testing, to improve fine-mapping of a target region, or to conduct a meta-analysis.

Meta-analysis is a powerful and commonly used technique, but if the study data were generated using different platforms, there may be a reduction in statistical power due to minimal overlap between the genotyped markers. To overcome this reduction in power, imputation may be used to increase the marker overlap between datasets, thereby improving the power of a meta-analysis (90,338,341,342).

Imputation is dependent on the adequate matching of haplotypes based on LD and thus it is essential that the reference population is genetically similar to the population being imputed. Numerous reference datasets are freely available online and can be used for imputation via suitable imputation software. These include amongst others, the 1000 Genomes phase 3 data (1000G, 10), the Human Genome Diversity Project (82), Haplotype Reference Consortium (HRC, 12) and the HapMap consortium (84). Most of the above-mentioned reference panels focus mainly on the European population and representation of African populations and admixed populations containing African ancestry is limited.

African and admixed populations are more heterogeneous in their haplotype block structure and, as such, would benefit from a larger reference dataset incorporating more genetic diversity (337). Reference datasets of this nature would increase the chances that an observed haplotype is present in the reference data, thereby greatly improving the imputation accuracy for African and admixed

individuals with African ancestry. Fortunately, recent years have seen a substantial increase in the representation of African populations in the 1000G data (81) and additional databases focusing on representing African populations have been established. The Consortium on Asthma among African ancestry populations in the Americas (CAAPA, Mathias et al., 2016) is available for download from dbGap with Accession ID:phs001123.v1.p1 and the African Genome variation project (AGVP) (89) as well as the African Genome Resource (AGR) (not publicly available[7]) are three resources which have recently become a viable option for accurate imputation of African populations.

Apart from choice of reference panel, the software used also affects the imputation accuracy (341). Many imputation software packages are freely available and have been previously tested and validated for accuracy, including Impute2 (80), Beagle (343), MaCH, MaCH-Minimac and MaCH-Admix (344). These imputation software packages were evaluated in African and African-American populations using different reference panels and produced varying degrees of imputation quality and accuracy (341,344).

Huang *et al.* (345) tested imputation accuracy in 29 populations using the HapMap reference and showed that the highest imputation accuracy was achieved for European populations, followed by East-Asian, Central- and South-Asian, American, Oceanian, Middle-Eastern and African populations. An additional finding from this study was that combining multiple reference populations resulted in improved imputation accuracy for any population analysed (345). While more appropriate reference panels are now available, which would increase the accuracy of imputation in African individuals, these results indicate that there are difficulties when imputing populations for which there is a limited number of reference individuals.

Imputation accuracy has previously been assessed for African populations (341,344,345) and for populations with two- or three-way admixture, with results reaching over 75% accuracy (346). In the present study, we assessed the accuracy of imputation in the five-way admixed South African Coloured (SAC) population. The SAC population contains genetic contributions from Bantu-speaking Africans, KhoeSan, Europeans, and South- and East-Asians (86,87). While imputation in this population has been conducted previously and the resulting data used for association analyses (333), the accuracy of imputation in this highly admixed population is yet to be evaluated.

Here we tested the quality and accuracy of imputation in the SAC population using different imputation software and reference panels and show that the Sanger Imputation Server using the AGR reference panel produced the highest quality and accuracy in imputed data. An in-house method using IMPUTE2 and 1000G reference panel imputed more variants than Sanger (AGR) but at a slightly reduced quality and accuracy.

[7] https://imputation.sanger.ac.uk/

5.3 Methods

5.3.1 SAC data

Two sources of data for the SAC cohort were available, namely genotypes obtained using the Affymetrix 500k array containing 500 000 SNP markers (Affymetrix, California, USA) and the Illumina (Illumina, California, USA) multi-ethnic genotyping array (MEGA) with 1.7 million markers. Approval was obtained from the Health Research Ethics Committee of Stellenbosch University (project registration number S17/01/013, S17/02/037 and 95/072) before participant recruitment and written informed consent was obtained from all study participants prior to blood collection.

Genotype datum obtained using the MEGA array was subjected to iterative quality control (QC) using PLINK v1.9 (120,304) as previously described (281), with the exception of related individuals not being removed. Individuals with more than 10% missing information and SNPs with more than 2% missingness were removed, as well as any variants with a minor allele frequency (MAF) below 5% as well as loci with excessive heterozygosity.

These QC steps were iterated until no additional variants or individuals were removed, and concluded with a sex-concordance check to remove individuals with incorrect sex information. Genotype Harmoniser version 1.4.15 (347) was used to strand align the two datasets to the 1000 Genomes Phase 3 reference panel (human genome build 37, (81)), update SNP IDs and remove any variants not in the reference panel. For the strand alignment a minimum LD value of 0.3 with at least three flanking variants was required for alignment. A secondary MAF alignment was also used at a threshold of 5%. Finally, the minimum posterior probability was set to 0.4.

5.3.2 Phasing and imputation

Three different methods were used for phasing and imputation to assess which performed best for our admixed population. The first was an in-house method where the Affymetrix data (PLINK files) were phased with SHAPEIT v2 (348), using the default effective population size of 15 000. Imputation was then done using IMPUTE2 v2.3.2 (80) and the 1000G Phase 3 reference panel (81), with default parameters except for the effective population size, which again was set to 15 000.

The second method made use of the Sanger Imputation server (SIS[8]). Genotypes from the Affymetrix 500k array in PLINK file format were converted to Variant Call Format (VCF) using PLINK v1.9 and then uploaded to the server where phasing was performed using SHAPEITv2.r790 (348) followed by imputation using the Positional Burrows-Wheeler Transformation (PBWT) algorithm (349). Imputation was done in two separate runs: the first run used the 1000G Phase 3 reference panel for imputation, and the second run made use of the African Genome Resource panel.

[8] https://imputation.sanger.ac.uk/

The third method made use of the Michigan Imputation server (MIS, (350)). PLINK files were converted to VCF using PLINK v1.9 and uploaded to the server for two imputation runs, both of which were run on the QC and imputation mode. SHAPEITv2.r790 was used for haplotype phasing in both runs followed by imputation using the Minimac3 algorithm (350). For the first run the mixed population option was used for the QC and haplotype phasing was done followed by imputation with the 1000G Phase 3 reference panel. For the second imputation run, it was mandatory for the African-American population to be selected for QC when imputing with the CAAPA reference panel.

Although haplotype pre-phasing has been shown to decrease imputation accuracy slightly it was used in this study for consistency between the methods (the Michigan server did not have an option to not phase data) and to increase the speed of imputation (80).

For all imputation runs, the reference panels included all available populations since using an all-inclusive reference panel is known to improve imputation accuracy (345). Of the five variations of imputation performed, only the MIS (CAAPA) run was incapable of performing imputation on the X chromosome. Results for the X chromosome have however been included for the other four imputation runs since the accuracy of X-linked imputation has not been evaluated previously.

5.3.3 QC of imputed data

Imputed data were returned from the imputation software in one of two formats: either in the form of a VCF file, or in Impute2 (gen/sample) format and based on the format, one of two QC procedures was employed to convert the imputed data from genotype probabilities to actual genotypes. Data output from the two procedures were compared and showed complete overlap and can thus be used interchangeably.

Procedure 1: For the in-house imputation done using Impute2, a gen/sample output file was obtained and converted to a PLINK file using GTOOL[9] version 0.7.5. R version 3.2.4 was used to identify INDELS, which were then removed using GTOOL (142). This was performed in order to more accurately assign SNP IDs and allele information when genotypes were called using GTOOL. The genotype calling threshold was set to 0.7, which was determined to have the best ratio of imputation accuracy and number of imputed variants (Figure S5.1). Once genotypes were called, the resulting ped/map PLINK files were converted to bed/bim/fam PLINK files and all variants with no-call alleles were removed.

Procedure 2: For the imputation done using the two online servers, VCF files were returned. The VCF files were converted to PLINK ped/map files using a genotype calling threshold of 0.7 (PLINK command: --vcf-min-gp command) and coding all no-call alleles as N (PLINK command: --output-

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

missing-genotype N). INDELS and SNPs with no-call alleles were removed and the files were converted to PLINK bed format (bed/bim/fam).

5.3.4 Imputation quality and accuracy

To assess imputation quality, the INFO score (in the case of IMPUTE2) and the comparable r-squared value (for PBWT and Minimac3) were used. These values range from 0 to 1, where a higher value indicates increased quality of an imputed SNP. Mean INFO score and r-squared values were plotted against MAF in order to determine how quality was affected by MAF and to assess which imputation method had the best quality at a given MAF.

Imputation accuracy was assessed by extracting the overlapping individuals from the MEGA and imputed Affymetrix data and using PLINK to remove any variants that overlapped between the two platforms prior to imputation. The analysis was performed per chromosome and for each SNP the alleles were compared between the imputed Affymetrix data and the MEGA data. If both alleles of a SNP matched it would be considered a complete match (or a flip match if alleles were correct but strand swopped). If only one allele matched it was considered a half match and if no alleles matched it was considered a no-match. For each chromosome the total number of imputed variants was recorded and their distribution by MAF was plotted to determine how the number of variants correlated with MAF between the different imputation methods.

To determine the imputation accuracy, the SNP overlap between the MEGA and imputed Affymetrix data was assessed. Within this overlap the number of SNPs that were complete-, flip-, half- or non-matched were recorded along with their average INFO score or r-squared value. Since SNPs that are flipped can be flipped to align a reference, or a different dataset if a meta-analysis is planned, the flipped SNPs were considered matches for the purposes of calculating imputation accuracy. Accuracy was calculated by comparing the proportion of SNPs in the overlap that were complete (or flipped) matches to the number of overlapping SNPs. This gave an indication of accuracy and error rate within the overlapping region and should be a good indication of overall imputation accuracy. These calculations were performed for the autosomes and the X chromosome separately in order to determine how accurately and with what quality the imputation software imputed X-linked variants compared to autosomal variants.

5.4 Results

5.4.1 Genotyping data

After QC and strand alignment there were 919 individuals and 239 612 variants with a genotyping rate of 99.39% in the Affymetrix 500k dataset, and 771 Individuals with 1 491 347 variants in the MEGA data with a genotyping rate of 99.43%. A total of 325 individuals were genotyped on both the Affymetrix and MEGA array and 43 140 SNP markers overlapped between the two platforms.

Overlapping individuals were extracted and overlapping variants (prior to imputation) were removed from both datasets.

5.4.2 Imputation

The two imputation methods that performed best were the in-house IMPUTE2 with 1000G reference panel and the Sanger imputation server with the AGR reference panel. The in-house method had the most imputed variants across both the autosomes (60 438 387) and X chromosome (2 574 793), followed by SIS (AGR) (52 088 766 autosomal and 1 638 163 X-linked variants), while the SIS with 1000G reference panel had slightly fewer imputed variants than with the AGR panel (50418390 autosomal and 1679254 X-linked variants). The Michigan imputation server had only about half as many imputed variants as the other methods, for either reference panel (Table 5.1). The number of imputed variants that did not reach the genotype calling threshold (0.7) was lowest in the in-house method followed by the Michigan server results, and SIS (1000G) and SIS (AGR) had the highest percentage of variants not reaching genotype calling threshold (Table 5.1). When imputed Affymetrix variants were compared to the MEGA genotypes, the SIS (AGR) data had the highest accuracy (within the overlapping region) on both the autosomes (89.29%) and X chromosome (90.18%). The imputation accuracy for the in-house and SIS (1000G) method was very similar, with the in-house method having a slightly lower genome wide error rate. The accuracy of the Michigan server was good on the autosomes (~83%) but lacking for the X chromosome (~70%) (Table 5.2). The SIS (AGR) imputed the least X-linked variants, but at the highest accuracy, whereas the in-house method had twice as many X-linked variants as Sanger with only a 2.48% drop in accuracy (Table 5.2 and 5.3).

Table 5.1: Number of imputed variants and variants overlapping with MEGA as well as the percentage of calls that did not reach the genotype calling threshold (0.7). Imputed number of SNPs is given in millions and Overlapping number is given per ten thousand.

Method	Reference	Autosomes		X chromosome		% No calls
		Imputed[1]	Overlap[2]	Imputed[1]	Overlap[2]	
In-house	1000G	57.8	72.1	2.5	1.6	25.46
SIS	1000G	48.7	46.7	1.7	1.0	35.89
	AGR	50.5	60.5	1.6	1.5	44.18
MIS	1000G	28.6	47.7	1.3	1.1	35.22
	CAAPA	16.9	34.3	NA	NA	43.40

[1]Number of SNPs in millions
[2]Number of SNPs per ten thousand

Table 5.2: Genome wide error rate and accuracy of imputation on the autosomes and X chromosome.

Method	Reference	Accuracy in overlap (%)		GW Error rate in overlap (%)
		Autosomes	X chromosome	
In-house	1000G	88.00	87.70	12.00
SIS	1000G	87.15	88.23	12.85
SIS	AGR	89.29	90.18	10.70
MIS	1000G	83.26	69.87	16.74
MIS	CAAPA	62.45	NA	37.55

Table 5.3: Number of SNPs and accompanying average info score for the three categories, within the MEGA overlapping region.

Method	Reference	Autosomes						X chromosome					
		Total		Half		No		Total		Half		No	
In-house	1000G	632[1]	0.78	38[1]	0.36	48[1]	0.89	35[1]	0.73	2.7[1]	0.37	2.1[1]	0.83
SIS	1000G	407	0.79	25	0.46	35	0.87	8.9	0.8	0.5	0.56	0.7	0.88
	AGR	541	0.79	23	0.5	42	0.89	12.9	0.83	0.6	0.6	0.8	0.89
MIS	1000G	400	0.69	45	0.11	33	0.83	19.5	0.57	7.1	0.08	1.3	0.70
	CAAPA	214	0.68	105	0.03	24	0.76	NA					

[1]Number of SNPs in thousands

For the autosomes and X chromosome, the SIS (AGR) had the best imputation quality across all MAF ranges, closely followed by the in-house method where quality was second to SIS (1000G) only for low MAF (0-1%) variants on the X chromosome (Figure 5.1). The Michigan server produced the lowest quality imputation (Figure 5.1 and Table 5.3). The mean quality score was comparable across all autosomal chromosomes and thus only chromosome 1 is shown to represent the autosome and for comparison to the X chromosome (Figure 5.1). Figure 5.2 confirms that SIS (AGR) and the in-house method had the best imputation quality since more SNPs were imputed at high quality for both autosome and X chromosome. Since SIS (AGR) has the largest number of imputed genotypes not reaching the calling threshold a trade-off between quality and number of variants exists between SIS (AGR) and the in-house method.

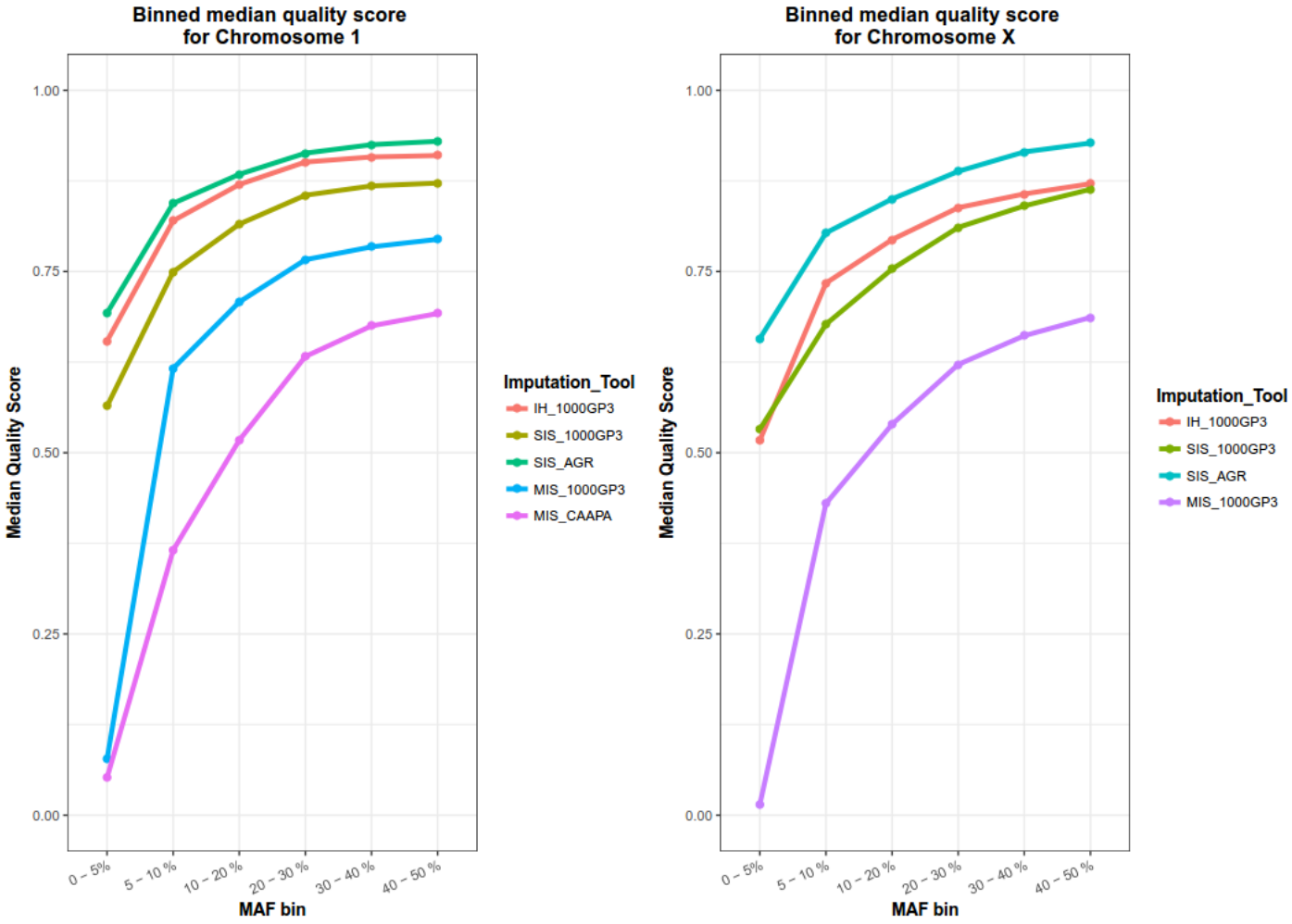


Figure 5.1: Mean quality score for all variants in a certain MAF range for all imputed datasets. In-house (IH), Sanger Imputation server (SIS) and Michigan Imputation Server (MIS).

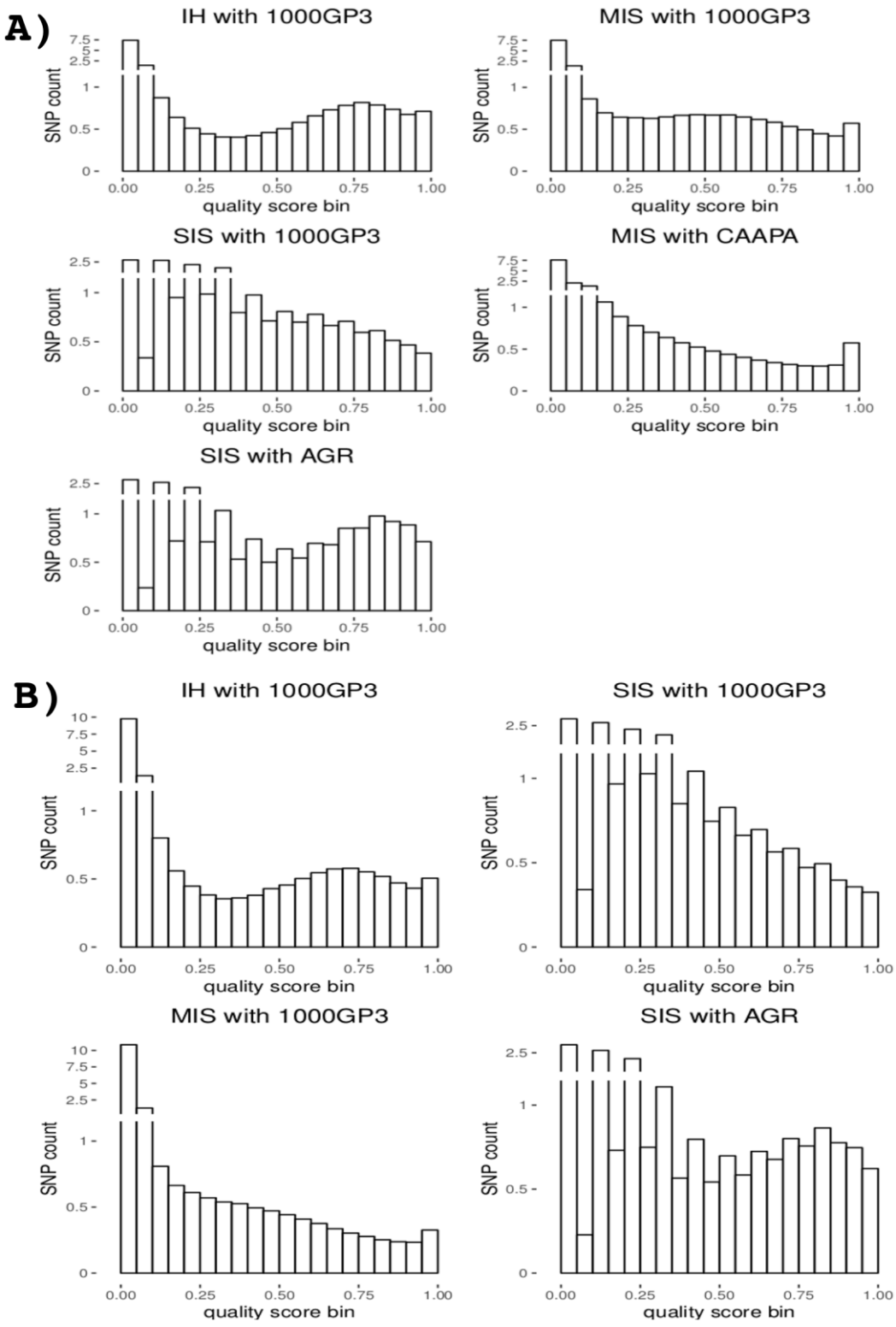


Figure 5.2: Distribution of the number of imputed SNPs by quality score for A) chromosome 1 and B) the X chromosome.

5.5 Discussion

Imputation accuracy was previously evaluated in African and three-way admixed populations, but we have performed the first evaluation in a five-way admixed population. The imputation accuracy in African-American individuals (considered to be three-way admixed) ranges from 78% (340) to 89% (80). Bantu-speaking Southern African individuals have been imputed with an accuracy of about 95% and even African San individuals had an imputation accuracy of 89% (345). In the present study, the SIS (AGR) and the in-house imputation method had similar accuracies (89% and 88% respectively, Table 5.2) compared to previous results from African and admixed populations. It should however be noted, that the clear majority of non-matching variants were ambiguous (Imputed genotype A/T and MEGA genotype G/C, or vice versa) and the majority of half-matched variants were imputed as monomorphic (data not shown). These ambiguous variants were imputed at high quality (Table 5.3) and were not removed when filtering on quality score but could be removed or aligned to a reference allele using appropriate software (such as Genotype Harmonizer). However, removal of these ambiguous variants is not mandatory. When analysing a single dataset, the ambiguous variants of interest can be compared to a relevant reference genome and then flipped. This is especially useful when conducting a meta-analysis since these variants will then be comparable even though they originate from different data sets. If these ambiguous variants are considered to be correctly imputed, then the accuracy of imputation with the SIS (AGR) increases to 96% while the accuracy of the in-house imputation method increases to 94%. Accuracy and quality can be further improved by removing half-matching variants by applying a quality score and MAF filter.

Since four of the five methods were capable of imputing X-linked variants, and since the quality and accuracy of X chromosome imputation has not been previously tested, we included it for this analysis. The X chromosome had only slightly lower or higher imputation quality for all imputation runs when compared to the autosomes, indicating that X chromosome imputation can be performed with confidence (Table 5.2 and 5.3). Although not specifically analysed here, the quality of imputation at low MAF should also be noted: the imputation quality for rare variants was unexpected as large reference panels with the correct populations are required to accurately impute rare variants (351,352) (Figure 5.1).

The biggest limitation for imputation in the five-way admixed population is the lack of a suitable reference panel. Imputation in the San population has been shown to have the lowest imputation accuracy (89%) compared to other African populations (345), which could be due to a lack of applicable reference individuals. Since the main ancestral component in the SAC population is KhoeSan (86) this could affect the accuracy and quality of imputation in this population. However, this has improved due to the addition of KhoeSan individuals to the reference panel.

In conclusion, we have shown that imputation of the SAC population is feasible and produces quality data on both the autosomes and X chromosome. While the SIS (AGR) imputation had the best quality

and accuracy, the in-house method using Impute2 and 1000G Phase 3 also produced imputed data of a high standard and had the highest number of imputed variants. This method may prove especially useful in the case of a meta-analysis where one wishes to maximise SNP overlap between datasets. As the number of applicable reference populations and individuals grows, imputation accuracy will improve for African and admixed populations, but it remains the gold-standard to Sanger sequence a variant of interest to confirm that the imputed variant is present in the population prior to conducting further research.

5.6 Acknowledgements

We would like to acknowledge and thank the study participants for their contribution and participation. 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. This work was also supported by the National Research Foundation of South Africa (grant number 93460) to EH 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, GW) to GT.

The authors report no conflict of interest.

5.7 Supplementary material

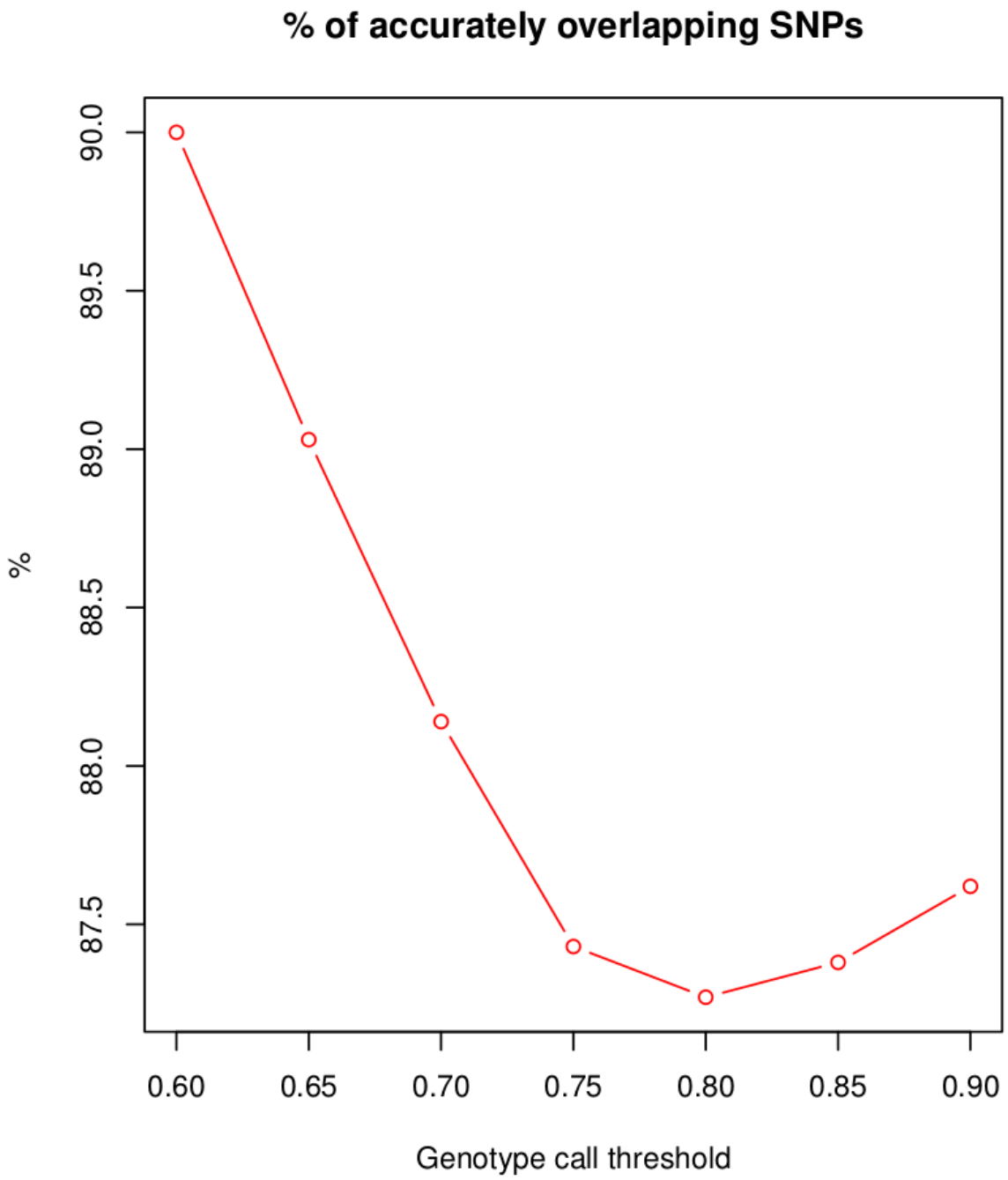


Figure S5.1: Percentage of overlapping variants that match between the imputed and MEGA data for different genotype calling thresholds.

6 X-linked trans-ethnic meta-analysis reveals Tuberculosis susceptibility variants.

Haiko Schurz [1,2], Craig J Kinnear [1], Paul D van Helden [1], Gerard Tromp [1,2,3], Eileen G Hoal [1], Marlo Möller [1]

[1] DST-NRF Centre of Excellence for Biomedical Tuberculosis Research, South African Medical Research Council Centre for Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

[2] South African Tuberculosis Bioinformatics Initiative (SATBBI), Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.

[3] Centre for Bioinformatics and Computational Biology, Stellenbosch University, Cape Town, South Africa.

6.1 Abstract

Globally Tuberculosis (TB) presents with a clear male bias that cannot be completely accounted for by environment, behaviour, socioeconomic factors or the impact of sex-hormones on the immune system. This suggests that genetic and biological differences, specifically relating to the X chromosome, further influence the male sex bias. The X chromosome has been shown to be heavily implicated in immune function and yet has largely been ignored in previous association studies. Here we report the first X chromosome specific association study on TB susceptibility. We identified X-linked TB susceptibility variants using seven genotyping datasets and 23229 individuals from different ethnic backgrounds. A sex-stratified and combined meta-analysis was conducted using the XWAS software and genomic regions previously associated with TB susceptibility were reproduced in this study. While significant associations were identified, the genes that they are located in have not previously been implicated in TB susceptibility.

6.2 Introduction

The X chromosome has been estimated to encode approximately 1500 of the 20000 protein coding genes in the human genome (119) and has the highest density of regulatory miRNA (118). This means that the X chromosome codes for over 5% of proteins and 10% of miRNA respectively and even though many of these are involved in immune functions the X chromosome has generally been ignored in previous association studies (353). According to the GWAS catalogue, 62652 unique SNP-trait associations have been identified by 3420 publications and of these associations only 385 SNPs were located on the X chromosome with 157 reaching genome wide significance (p-value < $5e^{-8}$) (122). This indicates the extent to which the X chromosome has been ignored in GWAS even though it represents a significant portion of the genome. In the past this was due to the analysis complexities introduced to GWAS by the X chromosome due to the haploid nature of males, but recently tools have been developed to analyse this chromosome specifically (114,115). Females have a more robust immune response against infections, which is partially attributable to the X chromosome and further influenced by the processes of X chromosome inactivation (XCI) and genes that escape silencing (139,140,197,235). Given the role of the X chromosome in sex biased immune responses it should not be excluded from statistical analysis, especially for infectious diseases, which often presents with a male bias, and autoimmune diseases, which often presents with a female bias (168).

Tuberculosis (TB), caused by *Mycobacterium tuberculosis* (*M. tuberculosis*), presents with a strong male bias with the global reported incidence rate being nearly twice as high in males compared to females (27). This male bias has been shown to be influenced, but not fully explained, by socioeconomic and behavioural factors, access to healthcare and the impact of sex hormones (estrogen and testosterone) on the immune system (140,143,144). As these factors do not fully explain the sex bias we hypothesise that X-linked genes and the process of XCI could further clarify the phenomenon. Thirteen published GWAS have investigated TB susceptibility (72,123–127,129–

131,133,281,292), but only six of these included the X chromosome in their analysis (72,96,123,126,133,281) and only one of the studies focused specifically on the X chromosome (281). Our previous study revealed strong sex specific effects on both the autosome and X chromosome, suggesting that sex-stratified QC and association testing should be done when analysing the X chromosome and autosomes (281). These sex-stratified analyses are of vital importance as sex specific effects of variants will be lost in a combined analysis which will negate the opposite directions of effects between the sexes. As a result of this, vital information concerning TB susceptibility and sex bias is lost, highlighting the need for sex-stratified analysis.

Apart from ignoring the X chromosome, previous studies have also validated poorly across populations. This is because ancestry has a major impact on TB susceptibility which has been shown in meta-analysis and admixture studies (291,354). Here we report the first X chromosome specific sex-stratified meta-analysis to identify TB susceptibility loci and elucidate its male sex bias, using GWAS data from European, Asian and African cohorts. Trans-ethnic meta-analyses have the advantage of giving us an overview of both population specific and global susceptibility variants and helping us to understand how ethnicity and sex influence TB susceptibility.

6.3 Methods

6.3.1 Study cohorts

Data for this meta-analysis was obtained from the International Tuberculosis Host Genetics Consortium (ITHGC) (96,123,128,131,355) and published TB GWAS (133,281). The ITHGC consists of 14 datasets, but eight cohorts had no X chromosome data or too few X-linked variants for imputation and were excluded from the analysis. In total two Asian, one European, two African and two South African Coloured (SAC) cohorts were included in this meta-analysis (Table 6.1), with a total of 11632 cases, 11597 controls and a 1.85:1 male to female ratio. Approval for the study was obtained from the Health Research Ethics Committee of Stellenbosch University (project registration number S17/01/013 and 95/072).

Table 6.1: Genotyping platform and number of samples for each cohort prior to quality control and imputation.

Country	Ethnicity	Cases	Controls	Females (%)	Platform	Source
China (1)	Asian	483	587	28.19	Affymetrix SNP Array 6.0	(128)
China (2)	Asian	1290	1145	48.74	Human OmniZhonghua-8 chips.	(128)
Russia	European	5914	6022	29.32	Affymetrix Genome-Wide Human SNP Array 6.0	(123)

Country	Ethnicity	Cases	Controls	Females (%)	Platform	Source
Gambia	African	1316	1382	39.53	Affymetrix GeneChip 500K	(96)
Ghana	African	1359	1952	36.97	Affymetrix SNP Array 6.0	(131)
SACA	SAC	850	90	44.66	Affymetrix 500k	(133)
SACM	SAC	420	419	55.78	Illumina MEGA	(281)

6.3.2 Quality control and Imputation

Quality control (QC) prior to imputation was performed as described previously (281), in order to obtain high quality genotypes for imputation. Briefly, the XWAS software (v3.0) and pipeline was used to implement a sex-stratified QC on all data, removing individuals without sex information and relatedness (up to 3rd degree) (114,115). As the datasets had varying coverage of the X chromosome due to the different genotyping platforms used (Table 6.1), imputation was done to increase the number of overlapping variants for the meta-analysis. An in-house method employing Impute2 (356) and the full 1000 Genomes Phase 3 reference panel (81) was used for imputation as previously described (**Chapter 5**), however data was not phased prior to imputation. Phasing was avoided in order to increase imputation accuracy (356). Imputed variants were filtered at an info score of 0.45 and all non-binary variants were removed before converting the data to plink format using GTOOL version 0.7.5 (350) at a genotype calling threshold of 0.7. Finally, the data underwent another sex-stratified QC procedure, using the same parameters as prior to imputation except that sex concordance and relatedness was not checked again. Overlapping variants between all datasets were extracted for the meta-analysis and aligned to the same reference allele where possible using the Plink v1.9 reference-allele command (357).

6.3.3 Meta-analysis

The meta-analysis was conducted by combining the association test results for each individual dataset using the XWAS (v3.0) software (114,115). A sex-stratified logistic regression test was performed for each dataset using the first five principle components as covariates for the ITHGC data and the four main ancestral components (Bantu-speaking African, KhoeSan, European and South Asian) for the SAC. The SAC was subject to sex biased admixture events and as a result there are significant differences between the ancestral components of the autosome and X chromosome. Therefore, the Admixture software (version 1.3) was used to determine X chromosome specific ancestral components (using the –haploid flag), as described previously (**Chapter 3**). Association statistics for males and females were combined for the sex-stratified meta-analysis to determine sex specific effects. In addition, all male and female results were also combined into one meta-analysis to determine effects that are independent of sex. This was done in an overall and population stratified

manner by testing all seven datasets together, or by grouping the two Asian or African cohorts (SAC, Ghana and Gambia). The Chi-squared based Q statistic was used to assess the heterogeneity between included studies and for a heterogeneity *p*-value > 0.1 the fixed effects (FE) model was implemented and for *p*-values ≤ 0.1 the random effects (RE) model was used to calculate pooled odds ratio (OR) and p-values (358). Genome wide significance threshold for the meta-analysis was set to $5e^{-8}$ (307).

6.4 Results

6.4.1 Cohort summary

Summary statistics of all the datasets following imputation and QC are shown in Table 6.2. In total 20255 individuals passed QC (10026 cases and 10229 controls) of which 35% were female. Between the different datasets 69983 variants overlapped and aligned to the same reference allele.

Table 6.2: Cohort summary and number of overlapping variants post Imputation and QC.

Cohort	Sample size	Females	Cases	Cases Female	Control	Control Female	SNPs before imputation	SNPs after imputation
China1	990	275	434	138	556	137	23907	180228
China2	2264	1103	1203	551	1061	552	16995	190719
Gambia	2541	1026	1229	353	1312	673	9305	339775
Ghana	3002	1112	1269	400	1733	712	32870	514299
SACA	596	256	577	242	19	14	9419	250838
SACM	815	465	410	242	405	223	24974	352167
Russia	10047	2861	4904	1293	5143	1568	28147	231499

6.4.2 Individual association results

Five genome wide significant variants were identified for the individual datasets (Table 6.3). Manhattan and QQ-plot for these results are given in supplementary Figures S6.1 and S6.2. One SNP (rs4465088) is located downstream of the actin related protein T1 gene (*ACTRT1)* in the Ghanaian females (OR = 4.73, p-value = $4.53e^{-18}$) and in the combined analysis (p-value = $1.60e^{-16}$) (fishers' method for combining p-values). *ACTRT1* encodes for a protein related to cytoskeletal beta-actin and while it is also involved in spermatid formation this gene has been shown to have tumour suppressive properties. The tumour suppressive properties were identified in basal cell carcinomas which result from the aberrant activation of the Hedgehog signalling pathway (359). While these genes have not been previously implicated in TB susceptibility the sonic hedgehog pathway (homolog of the hedgehog pathway) is implicated in mycobacterial immune evasion mediated through exploitation of regulatory T-cells (360).

Four significant variants were identified in the Russian cohort, three of which were significant in males (rs6610096, rs7888114, rs2563800), one in females (rs6610096) and two for the combined test statistic (rs6610096, rs5975736). The rs6610096 variant, located in the prostate associated gene 4 (*PAGE4*) was significant in male, females and the combined analysis. *PAGE4* is a member of the cancer/testis X antigens and while its function is poorly understood it is a DNA binding protein and silencing of *PAGE4 in vitro* caused cell death by apoptosis, indicating that *PAGE4* has an anti-apoptotic function (361). Expression of *PAGE4* is also inversely correlated with cancer progression and dysregulation of *PAGE4* modulates androgen receptor signalling and promotes progression to advanced prostate cancer (362). The rs2563800 variant, significantly associated in males and the combined analysis, is located in a non-coding RNA (ncRNA) *LOC105373293*, with no known function. *FRMPD4,* a gene previously implicated in schizophrenia, was only significantly associated in males (rs7888114), while rs5975736, located in the Bombesin Receptor Subtype 3 (*BRS3)* gene was associated in the combined dataset and had the same direction of effect between the sexes (363). *BRS3* is involved in energy homeostasis and upregulates glucose-stimulated insulin secretion in human pancreatic islet cells and could have a potential role in treatment of obesity and diabetes mellitus (364,365).

Table 6.3: Results for X chromosome association testing of individual datasets.

SNP	Gene (locus)	Position	A1	Male			Female			P comb	Cohort
				OR*	P-value	95% CI*	OR	P-value	95%CI		
rs4465088	ACTRT1 (Xq25)	3'UTR	G	0.94	$8.58e^{-1}$	0.46-1.91	4.73	$4.53e^{-18}$	3.42-6.55	$1.60e^{-16}$	Ghanaian
rs6610096	PAGE4 (Xp11.23)	5'UTR	A	1.53	$9.95e^{-14}$	1.37-1.72	1.70	$8.95e^{-16}$	1.49-1.94	$5.84\ e^{-27}$	Russia
rs7888114	FRMPD4 (Xp22.2)	Intron	C	19.29	$9.72e^{-9}$	7.05-52.8	3.14	$1.18e^{-3}$	1.57-6.31	$3.00\ e^{-10}$	Russia
rs2563800	LOC105373293 (Xq21.31)	Intron	A	1.57	$1.08e^{-8}$	1.35-1.83	1.04	0.659	0.87-1.25	$1.41\ e^{-7}$	Russia
rs5975736	BRS3 (Xq26.3)	3'UTR	A	2.49	$6.1e^{-7}$	1.74-3.56	2.29	$3.43e^{-5}$	3.52-49.2	$5.36\ e^{-10}$	Russia

*OR: Odds ratio; 95% CI: 95% confidence interval

6.4.3 Meta-analysis results

Both a genome wide significance threshold ($5e^{-8}$) and a Bonferroni threshold ($7.14e^{-7}$) were implemented to correct for the number of variants tested in the meta-analysis (0.05/69983). As TB is a complex disease and variants are unlikely to have large effect sizes, lowering the significance threshold can help to reduce the loss of valuable information in the form of false negative results. In total four variants reached significance in the sex-stratified population specific analysis, while no significant associations were identified in the combined (male and female) and sex-stratified analysis including all datasets (Table 6.4 and 6.5). The quantile-quantile (QQ) and Manhattan plots for these association tests are given in supplementary Figures S6.3-S6.5. The most significantly associated variant for the non-sex-stratified analysis including all datasets was rs79720685 (OR = 0.83, p-value = $3.06e^{-5}$) located in the interleukin 1 receptor accessory protein like 1 (*IL1RAPL1*) gene (Table 6.4). Variants in this gene have been associated with cardiovascular disease (366), autism (367) and presented with a male sex bias in a previous XWAS of asthma susceptibility in children (353). Variants in *IL1RAPL1* downregulate interleukin (IL) 13 which has a negative impact on the IL-1R pathway, a potential regulator of inflammation and a critical component of the host innate immune response against infections (368). The impact of this gene on TB susceptibility is unclear but given its role in the immune response it may contribute to the disease. The top hit for the combined meta-analysis in the Chinese cohorts was related to spermatogenesis and thus not informative in the context of TB susceptibility (369). For the African cohorts the combined analysis revealed the variant with the lowest p-valur to be in the actin remodelling regulator *NHS* gene, previously associated with Nanca-Horan Syndrome (a congenital cataract disease), dental abnormalities, brachymetacarpia (an abnormal shortness of the metacarpal bones) and mental retardation (370–373).

Table 6.4: Meta-analysis results for the combined analysis.

SNP	Gene (locus)	Position	A1	N	P-value	OR*	Q	Model	cohort
rs79720685	IL1RAPL1 (Xp21.3-21.2)	Intron	T	14	$3.06e^{-5}$	0.83	0.37	Fixed	All
rs58085560	SPANXN2 (Xq27.3)	3'UTR	C	4	$2.85e^{-5}$	1.32	0.86	Fixed	China
rs5909376	NHS (Xp22.2-22.13)	Intron	T	8	$4.11e^{-5}$	1.22	0.63	Fixed	African SAC

*OR: Odds ratio

For the sex-stratified meta-analysis the variants with the lowest p-values in the analysis including all cohorts for males was rs753468 (OR_M = 0.84, p-value = $3.21e^{-5}$), located in the *ATRX* gene and for females rs7053675 (OR_F = 0.83, p-value = $5.56e^{-6}$) located in the *PTCHD1-AS* gene (Table 6.5). The chromatin remodeller *ATRX* plays a role in supressing deleterious DNA secondary

structures that form a transcribed telomeric repeat, and loss of function of the *ATRX* gene can increase DNA damage and stall replication and homology-directed repair (374). This suggests that *ATRX* is involved in essential biological processes and it has been previously implicated in intellectual disability and osteosarcoma (375). Disruptions of the *PTCHD1-AS* gene are prevalent in ~1% of Autism spectrum disorders and intellectual disabilities (376). For the Chinese cohorts the sex-stratified analysis revealed 3 genome wide significant associations (rs1726176, rs5939510, rs1726203) in the long intergenic non-protein coding RNA 1546 (*LINCO1546*) in males, while no significant associations were identified in females. The close proximity of these variants in *LINC01546* suggest LD between the variants and it is unclear if any of the variants influence TB susceptibility as no functional information is available for the *LINC01546* locus. Finally, in the African cohorts one variant, in the *UPF3B* gene, reached significance in females after Bonferroni correction (rs2428212, OR_F = 2.03, p-value = $4.72e^{-7}$) but not in males. *UPF3B* is a regulator of nonsense mediated mRNA decay (NMD) and rapidly breaks down aberrant mRNA with a premature termination codon (PTC). *UPF3B* is involved in a central step in RNA surveillance by regulating crosstalk between the NMD pathway and the PTC-bound ribosome complex (377). Mutations in the *UPF3B* gene disrupt the NMD pathway, which is critical for neuronal development and can cause various forms of intellectual disability (378–380). While this gene has not previously been implicated in TB susceptibility, it could influence disease by altering RNA regulation linked to host defence against TB.

Table 6.5: Meta-analysis results for the sex-stratified analysis.

SNP	Gene (locus)	Position	A1	N	Male		Female		Q (M/F)	Model	Cohort
					OR*	P	OR	P			
rs753468	ATRX (Xq21.1)	Intron	C	7	0.84	$3.21e^{-5}$	1.06	$1.62e^{-1}$	0.42/0.66	Fixed	All
rs7053675	PTCHD1-AS (Xp22.11)	Intron	A	7	1.05	$2.11e^{-1}$	0.83	$5.56e^{-6}$	0.15/0.01	Fixed	All
rs1726176	LINC01546 (Xp22.33)	3'UTR	A	7	1.82	$4.20e^{-8}$	0.97	$7.19e^{-1}$	0.80/0.75	Fixed	China
rs5939510	LINC01546 (Xp22.33)	3'UTR	A	7	1.82	$4.89e^{-8}$	0.98	$7.50e^{-1}$	0.79/0.74	Fixed	China
rs1726203	LINC01546 (Xp22.33)	3'UTR	G	7	1.81	$6.37e^{-8}$	0.97	$6.92e^{-1}$	0.86/0.69	Fixed	China
rs2428212	UPF3B (Xq24)	Intron	C	7	0.76	$2.28e^{-1}$	2.03	$4.72e^{-7}$	0.95/0.15	Fixed	African SAC

*OR: Odds ratio

6.5 Discussion

Here we report the first X chromosome specific meta-analysis to investigate human genetic susceptibility to TB and the observed male bias. While no susceptibility variants reached significance in the meta-analysis including all cohorts, four sex-stratified variants were identified in the population stratified analysis, three variants for males in the Chinese cohorts and one variant in females in the

African cohorts (Table 6.5). Analysis of the individual data also revealed five significant associations in the Ghanaian and Russian datasets (Table 6.3). While most of the genes identified in this study have not been previously implicated in TB susceptibility a few could potentially be involved in mechanisms associated with the disease. *ACTR1, IL1RAPL1, ATRX* and *UPF3B* are involved in cellular functions that could be linked to host defence against TB. *ACTR1* and *IL1RAPL1* are involved in immune pathways via the sonic hedgehog signalling pathway and IL-1R pathway respectively, and mutations can affect T-cell regulation (360) and host immune response to infection (368), both of which are involved in host defence against TB. *ATRX* and *UPF3B* are involved in essential biological processes by monitoring and controlling aberrant RNA that could negatively impact transcription and RNA regulation (374,377). The role of these genes in immune function and RNA regulation suggests that they could impact TB susceptibility, but further investigation is required to elucidate the functional mechanisms underlying our statistical findings. Previous TB susceptibility studies investigating X-linked genes have identified several *TLR8* variants, but these associations were not replicated in this study (132,134–138). Possible reasons for this could be the impact of population specific effects and more stringent significance thresholds. However, while the exact variants and genes previously identified with TB susceptibility on the X chromosome were not replicated the genomic regions where these genes are located did replicate. In a linkage study the genomic region Xq, specifically Xq26, was associated with TB susceptibility in an African cohort (69). Indirect evidence of TB susceptibility loci on the X chromosome has also been provided from studies in Mendelian susceptibility to mycobacterial diseases (MSMD) where two X-linked regions, Xp11.4-Xp21.2 and Xq25-26.3 have been associated with MSMD (381,382). All these genomic regions were validated in this study, as well as the genomic locus where *TLR8* is situated (Xp22). We identified significant associations at Xp11.23, Xp21-Xp22.33, Xq21 and Xq24-Xq27.3, which overlap with previously associated genomic regions. This suggests that these genomic regions are implicated in TB susceptibility, but further research and fine-mapping is required to elucidate which genes and variants in this region contribute to the phenotype. Many X-linked genes have not been fully characterised, and their functions are still unclear; a recurring theme in our study. As a result, the impact on TB susceptibility cannot be elucidated using bioinformatic analysis alone and functional analysis of the genes is required. This is a major limitation in XWAS as many significant associations cannot be fully elucidated without functional verification, due to the lack of information about X-linked genes and their involvement in biological mechanisms. The impact of a variant on gene function can be tested by introducing the variant into an appropriate cell line (in vitro) or animal model (in vivo) using genomic editing methods such as CRISPR (383,384). Infection studies using these edited cell lines or animals can be done to determine the function of the variant.

A second conclusion that can be drawn from these results is that there are strong population specific effects influencing TB susceptibility. The fact that no significant associations were identified, even

though the study has more power when all data is included, supports this hypothesis. Previous studies in admixed populations have also shown increased susceptibility for some ancestral components over others (354). While larger studies may identify global susceptibility variants the impact of ethnicity on TB susceptibility cannot be ignored and population stratified analysis must be performed to elucidate the full complexity of TB disease. In the era of personalised medicine this will also be of vital importance as medication can be tailored and targeted for specific population groups and sexes (385). We have shown in a previous XWAS in the SAC population (281) that strong sex specific effects exist, and this is mirrored in this study. When comparing the OR of males and females (Table 6.3 and 6.5) it is clear that many sex-specific variants have opposite directions of effect or a negligible effect in one of the sexes. We found 6 significant associations in males but only 3 in females and none in the combined analysis, making a strong case for sex specific effects.

We suggest that during the planning of TB susceptibility studies, power should be determined based on the sample size of one sex, to maintain enough statistical power for sex-stratified analysis. Furthermore, care should be taken during sample selection to minimise population specific effects and subsequent larger and more powerful trans-ethnic meta-analysis will need to be performed to identify global susceptibility variants. Identifying population specific, sex-specific and global susceptibility variants can elucidate some of the complexity of TB pathogenesis and eventually allow for tailored or even preventative treatment.

6.6 Acknowledgement

We would like to acknowledge and thank the study participants for their contribution and participation. 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. This work was also supported by the National Research Foundation of South Africa (grant number 93460) to EH. This work was also supported by South African Tuberculosis Bioinformatics Initiative (SATBBI), a Strategic Health Innovation Partnership grant from the South African Medical Research Council and Department of Science and Technology to GT. We would also like to thank and acknowledge Dr Vivek Naranbhai and the International Tuberculosis Host Genetic consortium for allowing us access to their TB GWAS data.

The authors report no conflict of interest.

6.7 Supplementary material

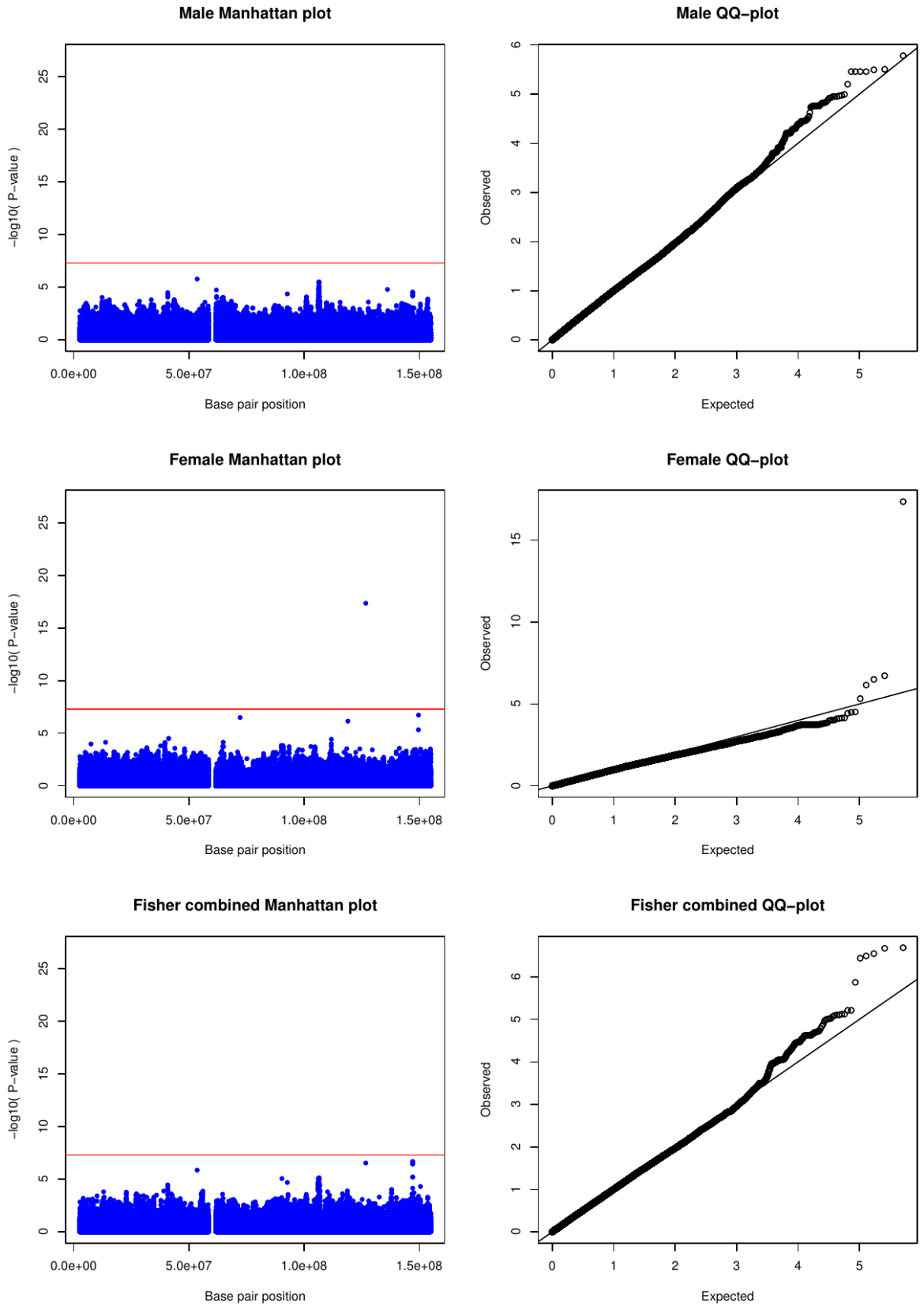


Figure S6.1: Manhattan and QQ-plot for X-linked SNP association testing of the Ghanaian cohort.

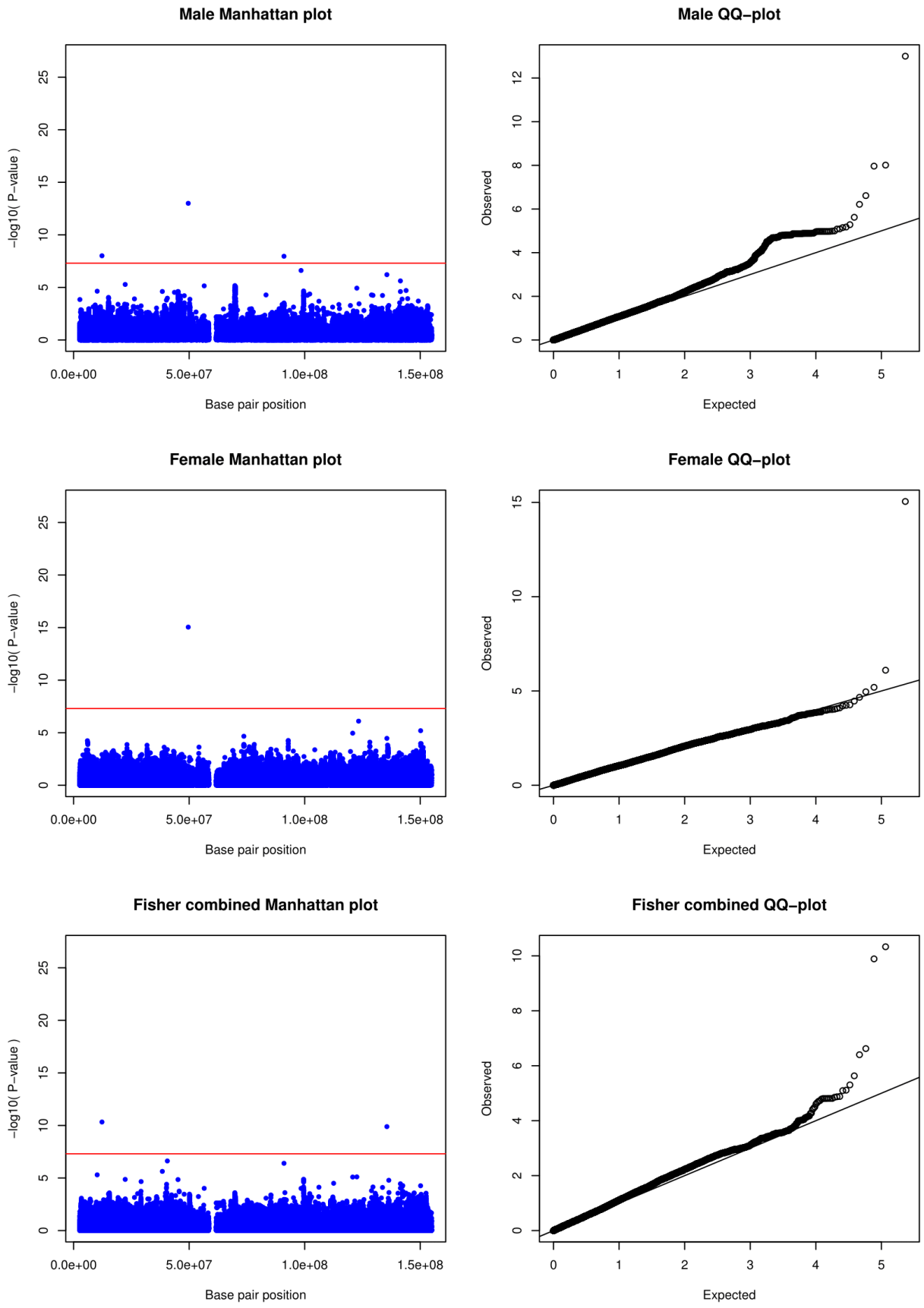


Figure S6.2: Manhattan and QQ-plot for X-linked SNP association testing of the Russian cohort.

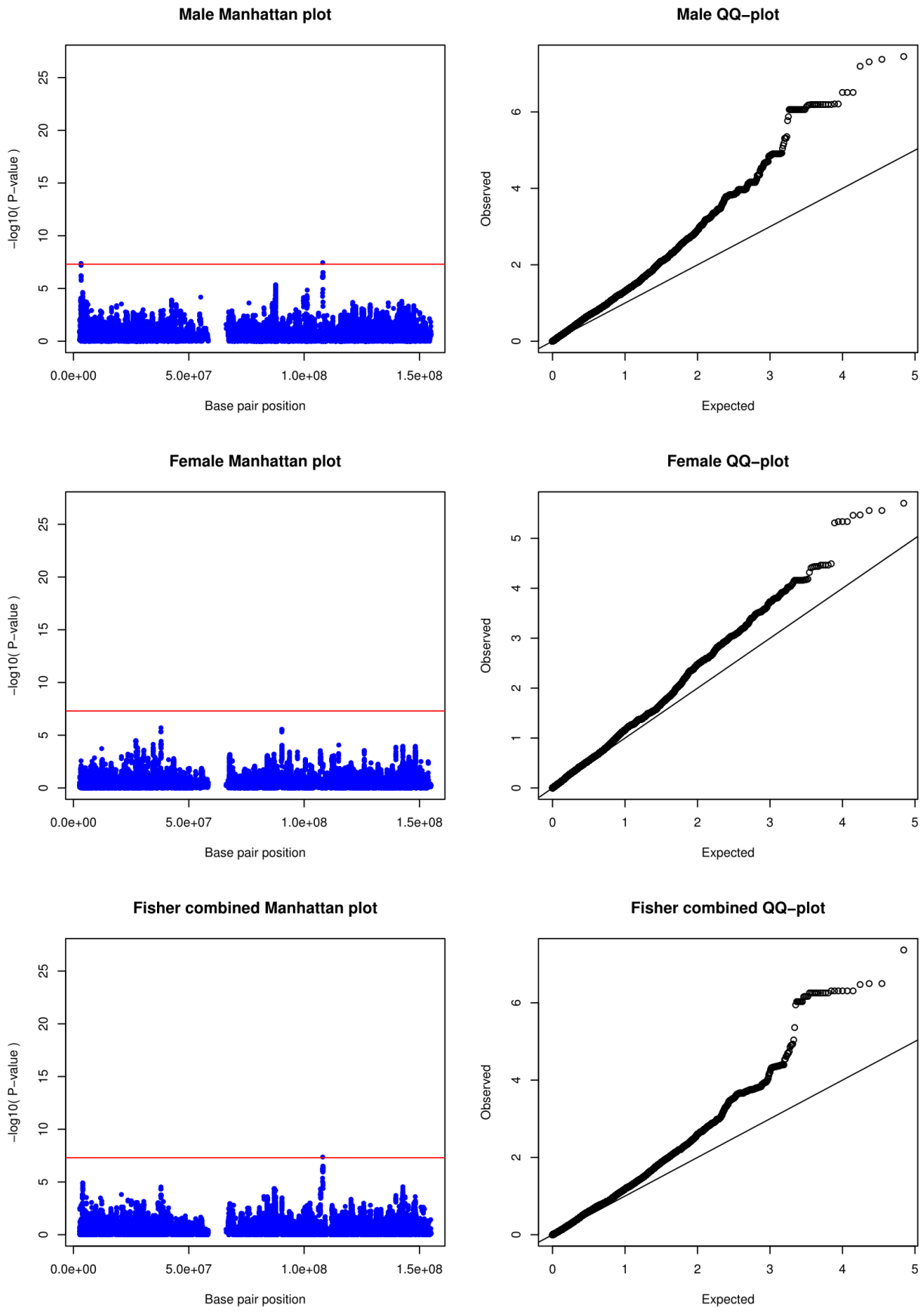


Figure S6.3: Manhattan and QQ-plot for the X-linked meta-analysis of the Chinese cohorts.

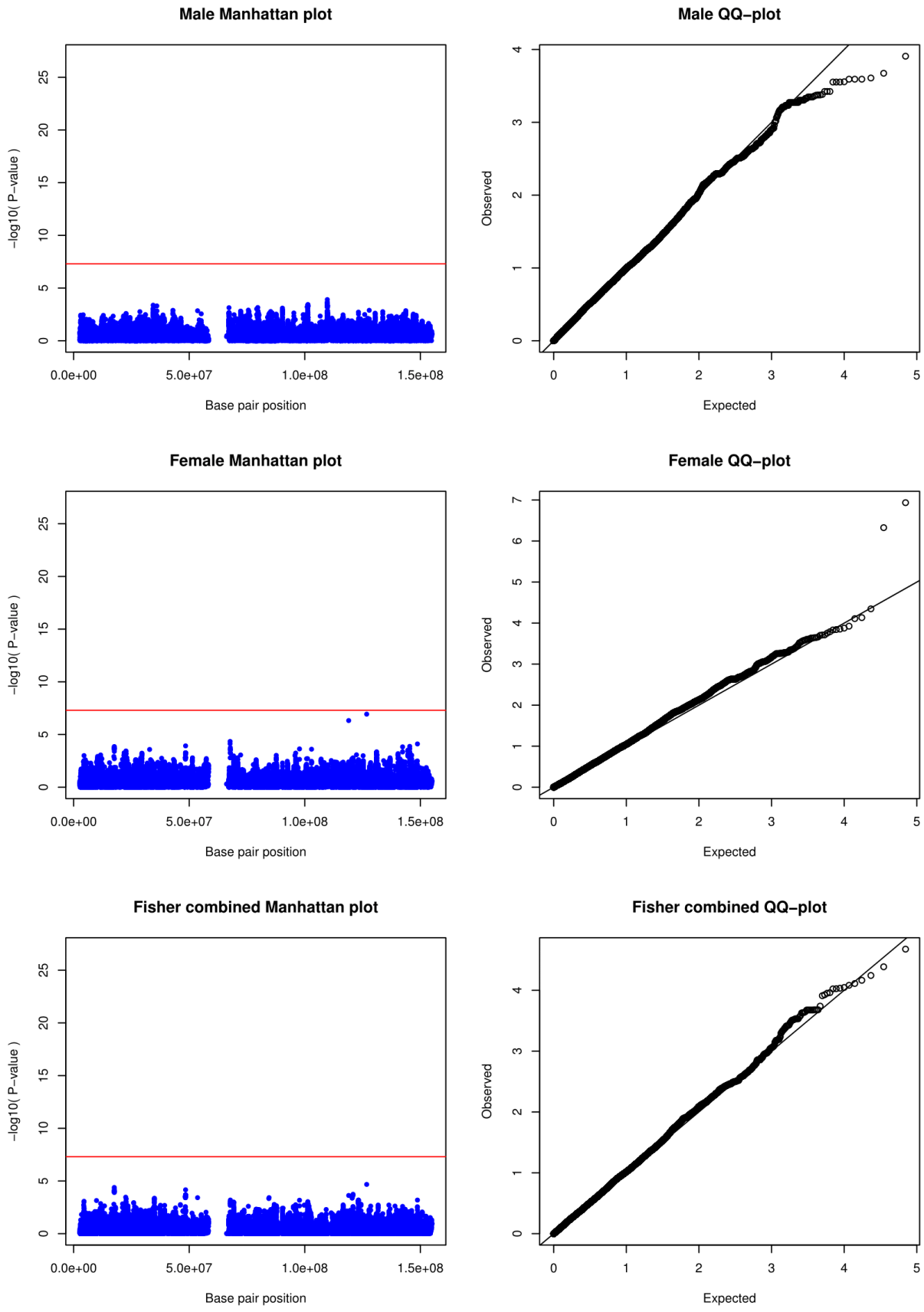


Figure S6.4: Manhattan and QQ-plot for the X-linked meta-analysis of the African cohorts, including the Gambian, Ghanaian and SAC data.

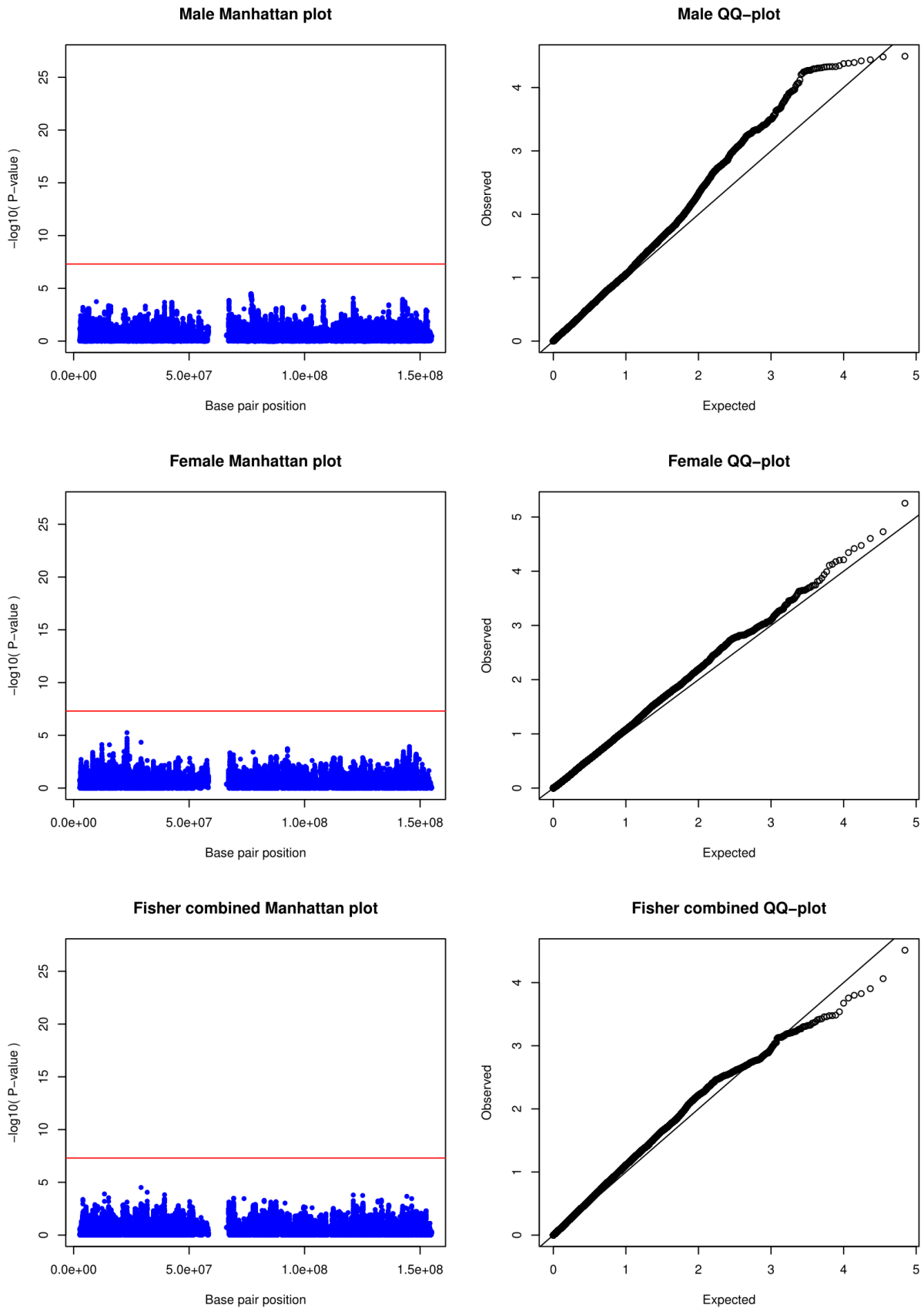


Figure S6.5: Manhattan and QQ-plot for the X-linked meta-analysis including all cohorts.

7 General discussion and conclusion

7.1 Summary

Sex-bias in diseases is present from infancy and persists across age groups, suggesting that it is independent of sex hormones, behaviour and environment and thus must have a host genetic contributing factor. Incidence rates for many infectious diseases (bacterial, fungal, viral and parasitic) persistently presents with a male sex-bias regardless of age, while auto-immune diseases present with a female bias (140,212,386,387). Significant sex differences in pharmacokinetics and pharmacodynamics of drugs have also been reported for men, women and pregnant females (388). It is thus important to understand the interplay of disease susceptibility factors and the genes on the sex chromosomes to discover novel pathways that will inform the understanding of disease, its sex-bias and the search for solutions.

The work presented in this thesis investigates the genetic contribution to the observed male sex-bias of TB. Globally the TB notification rate is nearly twice as high in males compared to females (27). We hypothesise that this bias is in part caused by X-linked genes and the unique biology of the X chromosome. Females are diploid for X-linked genes and males are haploid, which leads to the random inactivation of one X chromosome in females (148). This random inactivation makes females functional mosaics for X-linked genes and could give them an immunological advantage, driving the sex-bias. Flaws in the inactivation process such as genes that escape silencing and skewed inactivation could further influence the observed sex-bias.

Based on the possible involvement of the X chromosome, we reviewed (**Chapter 2**) past literature to determine the involvement of the X chromosome and X-linked genes in the immune system. The X chromosome contains a wealth of genomic information and a large number of X-linked genes have immune-related functions (118,119). The process of X chromosome inactivation (XCI) was also discussed along with the mechanisms and causes for genes escaping silencing and skewed inactivation and how this could influence disease susceptibility. As behaviour, socioeconomic factors and sex hormones also influence the male sex-bias in TB these factors were considered and it was concluded that while they influence the sex-bias they do not fully explain it, further supporting the role of the X chromosome (140,143,144). As females have been shown to have a more robust immune response against infection, the role of the X chromosome and sex-bias in infectious diseases, caused by bacteria, fungi, parasites and viruses is discussed (144). The role of the X chromosome was also discussed in the context of TB and evidence strongly suggestive of the role of the X chromosome in TB susceptibility and the observed male bias was presented.

Admixed populations, such as the South African Coloured population, add another level of complexity to X-linked analysis. Admixture can be sex-biased and result in significantly different ancestral distributions on the autosome compared to the X chromosome. Thus, the presence of sex-biased

admixture and its influence on TB susceptibility needs to be investigated **(Chapter 3).** A previous study has elucidated the presence of sex-biased admixture events in the history of the SAC using mtDNA and Y chromosomal markers (389). In this chapter an alternative method for determining sex-bias using global ancestry proportions inferred separately for the X chromosome and autosome was implemented (146,147). We show that there are significant differences in global admixture proportions between the autosome and X chromosome in the SAC population. The Bantu-speaking African and the European ancestral component present with a male bias, while the KhoeSan and Asian ancestral components present with a female bias. These results correlate perfectly with previous results (389) and highlight the importance of using X chromosome and autosomal ancestral components as covariates for association testing of the X chromosome and autosome respectively. The results also validate the global ancestry approach for determining the presence of sex-bias in this admixed population.

Using the ancestral components inferred in **Chapter 3**, a sex-stratified GWAS was performed in the SAC population using the Illumina MEGA array, a genotyping array specifically tailored for diverse populations (**Chapter 4**). A sex-stratified autosomal and X chromosome specific QC procedure was implemented followed by sex-stratified and combined association testing on the autosome and X chromosome (114,115). A gene-based association test as well as a sex differentiation test was also performed for X-linked genes. We also conducted the first genome wide interaction analysis using a joint effects model, followed by interaction analysis of the top 450 associations using a logistic regression model, correcting for age and ancestry. While the SNP based association analysis did not reveal any significant results, the interaction analysis identified genes that showed promise as potential future candidate genes. Finally, this GWAS showed that strong sex specific effects are evident for both autosomal and X-linked genes, highlighting the importance of conducting sex-stratified analysis.

Due to the lack of power in the GWAS study (**Chapter 4**) and the availability of additional GWAS data through our collaboration with the International Tuberculosis Host Genetics Consortium (ITHGC) an X-linked multi-ethnic meta-analysis was done to investigate TB susceptibility and its sex-bias (**Chapter 6**). In order to ensure maximal SNP overlap between all the datasets and increase power, we needed to impute the data. While multiple methods for imputation are readily available, none have been tested or optimised for our 5-way admixed SAC population. To address this an analysis was performed to assess the quality and accuracy of imputation in the SAC using various software and reference datasets (**Chapter 5**). Results from this analysis indicated that the SAC population can be imputed on both the X chromosome and autosome with adequate quality and accuracy. Determining imputation performance in this population is an important step as the SA population presents with a much more complex admixture pattern than previously investigated admixed populations such as African Americans and Hispanics. Furthermore, the KhoeSan and African founder populations of the

SA are under-investigated, and less data is available for imputation in these populations compared to European and Asian populations. Thus, knowing that we can accurately impute the SA population allows us to include these datasets in meta-analysis and increase power to detect associations by increasing the number of variants to be tested (**Chapter 6**).

In total seven datasets were included for this first ever meta-analysis investigating X-linked variants in TB susceptibility. Two Chinese, one Russian, a Gambian and Ghanaian and two SAC cohorts were included with 69983 overlapping variants left for analysis following imputation. All datasets went through a sex-stratified QC procedure (as discussed in **Chapter 4**) and again a sex-stratified and combined X-linked SNP based association test was done. Firstly, association testing was done on all individual datasets, revealing five novel associations in the Ghanaian and Russian cohorts. While no significant associations were identified in the combined and sex-stratified meta-analysis including all datasets, two significant associations were discovered in Chinese males when the meta-analysis was stratified by population. The genes identified in this meta-analysis have not been previously associated with TB and while some have limited functional information available a few potential candidate genes (*ACTR1, IL1RAPL1, ATRX* and *UPF3B)* that could influence TB susceptibility were identified but require further investigation. While these genes have not been previously associated with disease progression, the genomic regions in which they are located are known TB susceptibility loci (Xp11.23, Xp21-Xp22.33, Xq21 and Xq24-Xq27.3) (132,138,265,381,382,390). The meta-analysis also indicated the strength of population specific effects in TB susceptibility and this could explain why no significant associations were identified in the analysis including all cohorts. Furthermore, as for the GWAS (**Chapter 4**), the meta-analysis revealed sex-specific effects that could potentially influence the male sex-bias in TB susceptibility.

Overall these studies highlight the importance of conducting sex-stratified analysis and including appropriate covariates for X-linked association testing, due to the presence of sex-biased admixture. Furthermore, imputation in diverse populations was shown to have good quality and accuracy allowing for their inclusion in meta-analysis. Results from the association testing highlight the strong impact of population specific and sex-specific effects and a few potential candidate genes were identified, which warrant further investigation to elucidate their impact on TB susceptibility and the male sex-bias it presents with.

7.2 Limitations and future work

While the work conducted in this thesis revealed novel insights some clear limitations are evident and need to be addressed in future studies.

The first limitation is the power to detect significant associations, which is based on the sample size of the study. The complexity of TB, population specific effects and the sample size reduction due to splitting the data for sex-stratified analysis also negatively impacts on the results. TB is a complex

disease and as with most complex diseases effects of individual variants are likely to be small (77,79). Furthermore, there is unlikely to be a single causative variant and it is more likely that multiple variants with small effect collectively influence TB susceptibility. This polygenic nature means that each individual is likely to carry a number of variants that increase susceptibility and a number of variants that decrease susceptibility (79). This means that each individual could potentially carry a unique set of alleles that collectively influence susceptibility and while software has been developed to detect these interactions (gene-gene interaction analysis) they require immensely powerful studies in order to overcome the multiple testing burden (79,93) (as in **Chapter 4**). Furthermore, as GWAS are designed based on the analysis of common variants the impact of rare variants on disease susceptibility cannot be elucidated (79).

The second factor influencing power is the reduction of sample size due to the sex-stratified analysis. Due to the genetic nature of the X chromosome males and females need to be analysed separately, which reduces sample size and power. Furthermore, the fact that males are haploid for X-linked genes further reduces power compared to their female counterparts. As this thesis has shown, strong sex specific effects are present on not only the X chromosome but autosome as well (**Chapter 4 and 6**), suggesting that sex-stratified analyses are vital to fully elucidate TB susceptibility. This issue of reduced power in sex-stratified analysis can be rectified by doubling the sample size and ensuring that sufficient power to detect associations is available in one sex of the study. However, sample collection is difficult and expensive and thus not always feasible. Alternative methods to increase sample size, such as meta-analysis, need to be explored.

The third factor influencing power, especially in a meta-analysis, is population specific effects. Ancestry has been shown to influence TB susceptibility and genetic associations have replicated poorly across different populations (1,99,333,391). Africans are more susceptible to TB, while Europeans are more resistant as a result of longer historic pressure by the disease (64). These population specific effects can affect the meta-analysis if specific ethnic groups have different directions of effect for a particular SNP, which would reduce the effect of that variant in a multi-ethnic meta-analysis. The same variant could be significantly associated with disease in a population stratified analysis or GWAS, which is what we observed for our meta-analysis (**Chapter 6**). Despite this, meta-analyses are valuable for the study of TB susceptibility as they could identify associated variants across all populations but will likely require more power than currently available.

When analysing the results of the GWAS and meta-analysis (**Chapter 4 and 6**) another limitation comes to light and that is the lack of functional characterisation and annotation of X-linked genes. Many of the associations identified in this study are in uncharacterised genes or genes with very limited functional information. It is thus difficult to decipher how an associated gene could impact TB susceptibility without functional verification. This lack of information for X-linked genes is a direct result

of the X chromosome being ignored in past association studies, a practice that needs to change if the function and impact of X-linked variants are to be deciphered.

Other limiting factors, not specific to this thesis, but rather to GWAS in general, could also have affected the outcome of the analysis. SNP arrays require prior knowledge of the genome to design probes and incomplete annotation can influence array effectiveness (95). This is particularly true for the long-ignored X chromosome. SNP arrays are also limited to non-repetitive sequences in the genome which complicates analysis of related genes and alternatively spliced transcripts and copy number variations (CNVs) (95). The PCR based amplification methods for microarray genotyping can also introduce biases that influence the results. The specific strain of *M. tuberculosis* and its virulence also influences TB susceptibility, but strain information was not available for this thesis and as such could not be taken into account.

Future work should thus focus on reducing the impact of the limitations mentioned above. Larger sample sizes will reduce the power loss in sex-stratified analysis and allow for detection of small effect sizes. Also, as the number of TB GWAS studies increases, more data will be available to conduct powerful meta-analyses to identify both general and population specific associations. While these steps can improve the outcome of GWAS and GWAS-based meta-analysis they will not negate the limitations of the SNP microarrays in general. One way to overcome these limitations is by conducting next generation sequencing (NGS) instead. NGS requires no prior knowledge of the genome and can capture more genetic variation than GWAS. NGS is also not limited to non-repetitive sequences and can easily annotate related genes, alternative splicing and CNVs. PCR amplification is also reduced or eliminated in NGS and thus no amplification-based biases can be introduced. Finally, sequencing data is a powerful detection tool for rare variants which cannot be detected by GWAS but could influence disease susceptibility.

While NGS could address many of the problems facing GWAS there are limitations to conducting large NGS studies. Sequencing is still expensive compared to GWAS and a definite limiting factor especially in a resource constrained setting. Furthermore, NGS data is extremely large and requires a massive amount of storage space and computational power to analyse, further increasing the cost associated with NGS analysis. As a result, GWAS is still an invaluable analytic tool especially when large sample sizes are to be analysed, which is the case when conducting a pTB GWAS study. More severe phenotypes of TB are likely caused by rare variants and thus large sample sizes will be needed when doing GWAS, or NGS must be done to capture these rare variants in smaller sample sizes. GWAS can also serve as a screening tool to collect *a priori* evidence before conducting an expensive NGS analysis. While the price of NGS will continue to decrease, SNP arrays are still more robust than sequencing and will continue to bring valuable insight to disease aetiology (79). Furthermore, some analysis tools originally developed for GWAS have already been adapted and used in NGS data

analysis. Analytical tools and computational methodologies developed for GWAS represent a significant scientific advancement that will ultimately aid in the analysis of NGS data, which is generally more difficult to analyse (75).

The fact that NGS analysis adapted tools from GWAS brings to light another detrimental factor caused by excluding the X chromosome. If analytical tools for the X chromosome are not developed and improved then they will not be adopted for general use by the scientific community, thus hindering advancement of X-linked analysis. This will also translate to a lack of analytical tools and knowledge of how to analyse X-linked NGS data and could cause the X chromosome to be excluded from NGS studies as it was for GWAS. X chromosomal analysis should be encouraged and become an indispensable part of data analysis. This alone will lead to methodological advances and allow full utilisation of both GWAS and NGS data.

7.3 Conclusion

Since the advent of GWAS, SNP arrays have improved, as did computational methodologies and the amount of publicly available data, allowing for analyses such as ancestry inference, imputation and meta-analysis. The fact that the X chromosome was mostly excluded in GWAS caused the number of X chromosomal analysis tools and functional X-linked gene annotation to be far lower compared to the autosome. In future, focus will need to shift from data analysis to functional annotation to link GWAS results to biological function and fully understand disease aetiology. Furthermore, X chromosome analysis tools must be incorporated in GWAS pipelines as this is the only way to get the methods generally accepted and used by the scientific community.

In conclusion, this thesis highlights the need for sex-stratified and X-linked analysis and the methods presented here should encourage other researchers to include the X chromosome in any future analysis. Only by including the X chromosome in GWAS analysis will we improve the X-linked analysis tools and elucidate the role of the X chromosome in disease susceptibility. This study revealed strong sex and population specific effects that need to be accounted for in the study design to retain sufficient power for sex-stratified analysis. Future, more powerful meta-analysis will also need to be done to identify general and population specific susceptibility loci. This thesis thus presents clear evidence for the involvement of the X chromosome in TB susceptibility and the male sex-bias and future studies will need to focus on elucidating these effects. Fully understanding the sex-biased nature of TB will allow for medication tailored to a specific sex, which could improve treatment outcome. Furthermore, identifying population specific and globally associated variants could also lead to population specific and general treatment regimes and ultimately improve the global health status.

8 References

1. Kinnear C, Hoal EG, Schurz H, van Helden PD, Möller M. The role of human host genetics in tuberculosis resistance. Expert Rev Respir Med. 2017 Sep;11(9):721–37.

2. Crubézy E, Ludes B, Poveda J-D, Clayton J, Crouau-Roy B, Montagnon D. Identification of Mycobacterium DNA in an Egyptian Pott's disease of 5400 years old. Comptes Rendus Académie Sci - Ser III - Sci Vie. 1998 Nov 1;321(11):941–51.

3. Nerlich AG, Haas CJ, Zink A, Szeimies U, Hagedorn HG. Molecular evidence for tuberculosis in an ancient Egyptian mummy. Lancet Lond Engl. 1997 Nov 8;350(9088):1404.

4. Taylor PR, Martinez-Pomares L, Stacey M, Lin H-H, Brown GD, Gordon S. Macrophage receptors and immune recognition. Annu Rev Immunol. 2005;23:901–44.

5. Zimmerman MR. Pulmonary and osseous tuberculosis in an Egyptian mummy. Bull N Y Acad Med. 1979 Jun;55(6):604–8.

6. Bellamy R. Genetic susceptibility to tuberculosis in human populations. Thorax. 1998 Jul 1;53(7):588–93.

7. Herzog H. History of tuberculosis. Respir Int Rev Thorac Dis. 1998;65(1):5–15.

8. Leao SC, Portaels F. History. In: Tuberculosis 2007: From Basic Science to Patient Care. 2007. p. 25–49.

9. Doetsch RN. Benjamin Marten and his "New Theory of Consumptions." Microbiol Rev. 1978 Sep;42(3):521–8.

10. Daniel TM. Selman Abraham Waksman and the discovery of streptomycin. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2005 Feb;9(2):120–2.

11. Daniel TM. Captain of Death: the story of Tuberculosis. University of Rochester Press. New York; 1997.

12. Koch, R. Die aetiologie der tuberculose, a translation by Berna Pinner and Max Pinner with an introduction by Allen K. Krause. Am Rev Tuberc. 1932;25:285–323.

13. Koch R. Classics in infectious diseases. The etiology of tuberculosis: Robert Koch. Berlin, Germany 1882. Rev Infect Dis. 1982 Dec;4(6):1270–4.

14. Pirquet CV. FREQUENCY OF TUBERCULOSIS IN CHILDHOOD. J Am Med Assoc. 1909 Feb 27;LII(9):675–8.

15. von Pirquet, C. Die allergieprobe zur diagnose der tuberkulose in kindesalter. Wiener Medizinische Wochenschrift. 1907;28:1369–74.

16. von Pirquet, C. Der diagnostische wert der kutanen tuberkulinreaktion bei der tuberkulose des kindesalters auf grund von 100 sektionen. Wien Klin Wchnschr. 1907;20:1123–8.

17. Haas LF. Wilhelm Conrad Von Röntgen (1845-1923). J Neurol Neurosurg Psychiatry. 2001 Jan;70(1):126.

18. Sakula A. BCG: who were Calmette and Guérin? Thorax. 1983 Nov;38(11):806–12.

19. Comstock GW. The International Tuberculosis Campaign: a pioneering venture in mass vaccination and research. Clin Infect Dis Off Publ Infect Dis Soc Am. 1994 Sep;19(3):528–40.

20. Roy A, Eisenhut M, Harris RJ, Rodrigues LC, Sridhar S, Habermann S, et al. Effect of BCG vaccination against Mycobacterium tuberculosis infection in children: systematic review and meta-analysis. BMJ. 2014 Aug 5;349:g4643.

21. Lehmann J. TWENTY YEARS AFTERWARD HISTORICAL NOTES ON THE DISCOVERY OF THE ANTITUBERCULOSIS EFFECT OF PARAAMINOSALICYLIC ACID (PAS) AND THE FIRST CLINICAL TRIALS. Am Rev Respir Dis. 1964 Dec;90:953–6.

22. Ryan F. The forgotten plague: how the battle against tuberculosis was won--and lost. Little, Brown; 1994. 492 p.

23. Schatz A, Bugie E, Waksman SA. Streptomycin, a substance exhibiting antibiotic activity against gram-positive and gram-negative bacteria. 1944. Clin Orthop. 2005 Aug;(437):3–6.

24. Iseman MD. Tailoring a time-bomb. Inadvertent genetic engineering. Am Rev Respir Dis. 1985 Oct;132(4):735–6.

25. Chaisson RE, Nuermberger EL. Confronting multidrug-resistant tuberculosis. N Engl J Med. 2012 Jun 7;366(23):2223–4.

26. Houben RMGJ, Dodd PJ. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. PLOS Med. 2016 Oct 25;13(10):e1002152.

27. WHO | Global tuberculosis report 2017 [Internet]. WHO. [cited 2017 Nov 7]. Available from: http://www.who.int/tb/publications/global_report/en/

28. Johnson MD, Decker CF. Tuberculosis and HIV Infection. DisMon. 2006 Nov;52(11–12):420–7.

29. von Reyn CF, Kimambo S, Mtei L, Arbeit RD, Maro I, Bakari M, et al. Disseminated tuberculosis in human immunodeficiency virus infection: ineffective immunity, polyclonal disease and high mortality. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2011 Aug;15(8):1087–92.

30. Whalen C, Horsburgh CR, Hom D, Lahart C, Simberkoff M, Ellner J. Accelerated course of human immunodeficiency virus infection after tuberculosis. Am J Respir Crit Care Med. 1995 Jan;151(1):129–35.

31. Publication | Statistics South Africa [Internet]. [cited 2018 Aug 20]. Available from: http://www.statssa.gov.za/?page_id=1854&PPN=P0302&SCH=7048

32. Statistical release. 2017;22.

33. Knechel NA. Tuberculosis: pathophysiology, clinical features, and diagnosis. Crit Care Nurse. 2009 Apr;29(2):34–43.

34. Rajni, Rao N, Meena LS. Biosynthesis and Virulent Behavior of Lipids Produced by Mycobacterium tuberculosis: LAM and Cord Factor: An Overview. Biotechnol Res Int [Internet]. 2010 Dec 19 [cited 2018 Sep 17];2011. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3039431/

35. RILEY RL, MILLS CC, O'GRADY F, SULTAN LU, WITTSTADT F, SHIVPURI DN. Infectiousness of air from a tuberculosis ward. Ultraviolet irradiation of infected air: comparative infectiousness of different patients. Am Rev Respir Dis. 1962 Apr;85:511–25.

36. Fernstrom A, Goldblatt M. Aerobiology and Its Role in the Transmission of Infectious Diseases [Internet]. Journal of Pathogens. 2013 [cited 2018 Aug 25]. Available from: https://www.hindawi.com/journals/jpath/2013/493960/

37. Turner RD, Bothamley GH. Cough and the Transmission of Tuberculosis. J Infect Dis. 2015 May 1;211(9):1367–72.

38. Lerner TR, Borel S, Gutierrez MG. The innate immune response in human tuberculosis. Cell Microbiol. 2015 Sep;17(9):1277–85.

39. O'Garra A, Redford PS, McNab FW, Bloom CI, Wilkinson RJ, Berry MPR. The immune response in tuberculosis. Annu Rev Immunol. 2013;31:475–527.

40. Cambier CJ, Falkow S, Ramakrishnan L. Host Evasion and Exploitation Schemes of Mycobacterium tuberculosis. Cell. 2014 Dec 18;159(7):1497–509.

41. Lawn SD, Zumla AI. Tuberculosis. Lancet. 2011 Jul 2;378(9785):57–72.

42. Caws M, Thwaites G, Dunstan S, Hawn TR, Lan NTN, Thuong NTT, et al. The influence of host and bacterial genotype on the development of disseminated disease with Mycobacterium tuberculosis. PLoS Pathog. 2008 Mar;4(3):e1000034.

43. Boehme CC, Nabeta P, Hillemann D, Nicol MP, Shenai S, Krapp F, et al. Rapid molecular detection of tuberculosis and rifampin resistance. N Engl J Med. 2010 Sep 9;363(11):1005–15.

44. Abdool Karim SS, Naidoo K, Grobler A, Padayatchi N, Baxter C, Gray AL, et al. Integration of antiretroviral therapy with tuberculosis treatment. N Engl J Med. 2011 Oct 20;365(16):1492–501.

45. Zumla A, Raviglione M, Hafner R, von Reyn CF. Tuberculosis. N Engl J Med. 2013 Feb 21;368(8):745–55.

46. Combs DL, O'Brien RJ, Geiter LJ. USPHS Tuberculosis Short-Course Chemotherapy Trial 21: effectiveness, toxicity, and acceptability. The report of final results. Ann Intern Med. 1990 Mar 15;112(6):397–406.

47. Centers for Disease Control and Prevention (CDC). Emergence of Mycobacterium tuberculosis with extensive resistance to second-line drugs--worldwide, 2000-2004. MMWR MorbMortalWklyRep. 2006 Mar 24;55(11):301–5.

48. Narasimhan P, Wood J, MacIntyre CR, Mathai D. Risk Factors for Tuberculosis. Pulm Med [Internet]. 2013 [cited 2015 Sep 1];2013. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583136/

49. Cegielski JP, Kohlmeier L, Cornoni-Huntley J. Malnutrition and tuberculosis in a nationally representative cohort of adults in the United States, 1971–1987. In: Proceedings of the 44th Annual Meeting, American Society of Tropical Medicine and Hygiene. San Antonio, Texas, USA; 1995. p. 152.

50. Cegielski JP, McMurray DN. The relationship between malnutrition and tuberculosis: evidence from studies in humans and experimental animals. IntJTubercLung Dis. 2004 Mar;8(3):286–98.

51. Comstock GW, Palmer CE. Long-term results of BCG vaccination in the southern United States. Am Rev Respir Dis. 1966 Feb;93(2):171–83.

52. Arcavi L, Benowitz NL. Cigarette smoking and infection. Arch Intern Med. 2004 Nov 8;164(20):2206–16.

53. Bates MN, Khalakdina A, Pai M, Chang L, Lessa F, Smith KR. Risk of tuberculosis from exposure to tobacco smoke: a systematic review and meta-analysis. Arch Intern Med. 2007 Feb 26;167(4):335–42.

54. Maurya V, Vijayan VK, Shah A. Smoking and tuberculosis: an association overlooked. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2002 Nov;6(11):942–51.

55. Pai M, Mohan A, Dheda K, Leung CC, Yew WW, Christopher DJ, et al. Lethal interaction: the colliding epidemics of tobacco and tuberculosis. Expert Rev Anti Infect Ther. 2007 Jun;5(3):385–91.

56. Slama K, Chiang C-Y, Enarson DA, Hassmiller K, Fanning A, Gupta P, et al. Tobacco and tuberculosis: a qualitative systematic review and meta-analysis. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2007 Oct;11(10):1049–61.

57. Yanbaeva DG, Dentener MA, Creutzberg EC, Wesseling G, Wouters EFM. Systemic effects of smoking. Chest. 2007 May;131(5):1557–66.

58. Fox GJ, Orlova M, Schurr E. Tuberculosis in Newborns: The Lessons of the “Lübeck Disaster” (1929–1933). PLoS Pathog [Internet]. 2016 Jan 21 [cited 2018 Aug 25];12(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4721647/

59. Comstock GW. Tuberculosis in twins: a re-analysis of the Prophit survey. AmRevRespirDis. 1978 Apr;117(4):621–4.

60. Kallmann FJ, Reisner D. Twin studies on the significance of genetic factors in tuberculosis. AmRevTuberc. 1943;47:549–547.

61. Simonds B. Tuberculosis in twins. Pitman Medical Publishing Company; 2004.

62. Bellamy R. Susceptibility to mycobacterial infections: the importance of host genetics. Genes Immun. 2003 Jan;4(1):4–11.

63. Hill AV. Genetics and genomics of infectious disease susceptibility. BrMedBull. 1999;55(2):401–13.

64. Hoal EG. Human genetic susceptibility to tuberculosis and other mycobacterial diseases. IUBMBLife. 2002 Apr;53(4–5):225–9.

65. Möller M, Hoal EG. Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. Tuberculosis(Edinb). 2010 Mar 3;90(2):71–83.

66. van der Eijk EA, van de Vosse E, Vandenbroucke JP, van Dissel JT. Heredity versus environment in tuberculosis in twins: the 1950s United Kingdom Prophit Survey Simonds and Comstock revisited. Am J Respir Crit Care Med. 2007 Dec 15;176(12):1281–8.

67. Hoal EG, Dippenaar A, Kinnear C, van Helden PD, Möller M. The arms race between man and Mycobacterium tuberculosis: Time to regroup. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2017 Aug 23;

68. Lipsitch M, Sousa AO. Historical intensity of natural selection for resistance to tuberculosis. Genetics. 2002 Aug;161(4):1599–607.

69. Bellamy R. Identifying genetic susceptibility factors for tuberculosis in Africans: a combined approach using a candidate gene study and a genome-wide screen. Clin Sci. 2000 Mar;98(3):245–50.

70. Greenwood CM, Fujiwara TM, Boothroyd LJ, Miller MA, Frappier D, Fanning EA, et al. Linkage of tuberculosis to chromosome 2q35 loci, including NRAMP1, in a large aboriginal Canadian family. Am J Hum Genet. 2000 Aug;67(2):405–16.

71. Cobat A, Gallant CJ, Simkin L, Black GF, Stanley K, Hughes J, et al. High Heritability of Antimycobacterial Immunity in an Area of Hyperendemicity for Tuberculosis Disease. J Infect Dis. 2010 Jan 1;201(1):15–9.

72. Luo Y. Progression of recent Mycobacterium tuberculosis exposure to active tuberculosis is a highly heritable complex trait driven by 3q23 in Peruvians. :28.

73. Stein CM, Guwatudde D, Nakakeeto M, Peters P, Elston RC, Tiwari HK, et al. Heritability Analysis of Cytokines as Intermediate Phenotypes of Tuberculosis. J Infect Dis. 2003 Jun 1;187(11):1679–85.

74. Pan H, Yan B-S, Rojas M, Shebzukhov YV, Zhou H, Kobzik L, et al. Ipr1 gene mediates innate immunity to tuberculosis. Nature. 2005 Apr 7;434(7034):767–72.

75. LaFramboise T. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. Nucleic Acids Res. 2009 Jul;37(13):4181–93.

76. Lamy P, Grove J, Wiuf C. A review of software for microarray genotyping. Hum Genomics. 2011 May 1;5(4):304–9.

77. Riancho JA. Genome-wide association studies (GWAS) in complex diseases: advantages and limitations. Reumatol Clin. 2012 Apr;8(2):56–7.

78. Beiko R. Bioinformatics: Hypothesis Free—Or Hypotheses Freed? BioScience. 2014 Sep 1;64(9):844–5.

79. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017 Jul 6;101(1):5–22.

80. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009 Jun;5(6):e1000529.

81. Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, et al. An integrated map of structural variation in 2,504 human genomes. Nature. 2015 Oct 1;526(7571):75–81.

82. Cavalli-Sforza LL. The Human Genome Diversity Project: past, present and future. 2005;8.

83. the Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet. 2016 Oct;48(10):1279–83.

84. Consortium TIH 3. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 Sep 2;467(7311):52–8.

85. Chimusa ER, Daya M, Möller M, Ramesar R, Henn BM, van Helden PD, et al. Determining Ancestry Proportions in Complex Admixture Scenarios in South Africa Using a Novel Proxy Ancestry Selection Method. PLoS ONE. 2013 Sep 16;8(9):e73971.

86. Daya M, Merwe L van der, Galal U, Möller M, Salie M, Chimusa ER, et al. A Panel of Ancestry Informative Markers for the Complex Five-Way Admixed South African Coloured Population. PLOS ONE. 2013 Dec 20;8(12):e82224.

87. de Wit E, Delport W, Rugamika CE, Meintjes A, Möller M, van Helden PD, et al. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. Hum Genet. 2010 Aug;128(2):145–53.

88. Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, et al. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. Nat Commun. 2016 Oct 11;7:12522.

89. Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, et al. The African Genome Variation Project shapes medical genetics in Africa. Nature. 2015 Jan 15;517(7534):327–32.

90. Anderson CA, Pettersson FH, Barrett JC, Zhuang JJ, Ragoussis J, Cardon LR, et al. Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. Am J Hum Genet. 2008 Jul;83(1):112–9.

91. Jiao S, Hsu L, Hutter CM, Peters U. The use of imputed values in the meta-analysis of genome-wide association studies. Genet Epidemiol. 2011 Nov;35(7):597–605.

92. Johnson RC, Nelson GW, Troyer JL, Lautenberger JA, Kessing BD, Winkler CA, et al. Accounting for multiple comparisons in a genome-wide association study (GWAS). BMC Genomics. 2010 Dec 22;11:724.

93. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013 Jul 22;9:29.

94. Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M. Genome-wide association studies in diverse populations. Nat Rev Genet. 2010 May;11(5):356–66.

95. Hurd PJ, Nelson CJ. Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief Funct Genomic Proteomic. 2009 May;8(3):174–83.

96. Consortium TWTCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature. 2007 Jun;447(7145):661.

97. Pe'er I, de Bakker PIW, Maller J, Yelensky R, Altshuler D, Daly MJ. Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet. 2006 May 21;38:663.

98. Niu T. Algorithms for inferring haplotypes. Genet Epidemiol. 2004 Dec;27(4):334–47.

99. Daya M, van der Merwe L, Gignoux CR, van Helden PD, Möller M, Hoal EG. Using multi-way admixture mapping to elucidate TB susceptibility in the South African Coloured population. BMC Genomics. 2014;15:1021.

100. Koed K, Wiuf C, Christensen L-L, Wikman FP, Zieger K, Møller K, et al. High-density single nucleotide polymorphism array defines novel stage and location-dependent allelic imbalances in human bladder tumors. Cancer Res. 2005 Jan 1;65(1):34–45.

101. Neafsey DE, Schaffner SF, Volkman SK, Park D, Montgomery P, Milner DA, et al. Genome-wide SNP genotyping highlights the role of natural selection in Plasmodium falciparum population divergence. Genome Biol. 2008;9(12):R171.

102. Price AL, Butler J, Patterson N, Capelli C, Pascali VL, Scarnicci F, et al. Discerning the ancestry of European Americans in genetic association studies. PLoS Genet. 2008 Jan;4(1):e236.

103. Uren C, Kim M, Martin AR, Bobo D, Gignoux CR, Helden PD van, et al. Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. Genetics. 2016 Sep 1;204(1):303–14.

104. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. Cell. 2017 Jun 15;169(7):1177–86.

105. Wray NR, Wijmenga C, Sullivan PF, Yang J, Visscher PM. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. Cell. 2018 Jun 14;173(7):1573–80.

106. Bossini-Castillo L, Martín J-E, Díaz-Gallo LM, Rueda B, Martín J. Genética de la esclerodermia. Reumatol Clínica. 2010 Sep 1;6:12–5.

107. Kingsmore SF, Lindquist IE, Mudge J, Gessler DD, Beavis WD. Genome-wide association studies: progress and potential for drug discovery and development. Nat Rev Drug Discov. 2008 Mar;7(3):221–30.

108. Maher B. Personal genomes: The case of the missing heritability. Nature. 2008 Nov 6;456(7218):18–21.

109. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. 2009 Oct;461(7265):747–53.

110. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014 Oct 5;46:1173.

111. Consortium SWG of the PG, Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, et al. Biological insights from 108 schizophrenia-associated genetic loci. Nature. 2014 Jul;511(7510):421–7.

112. Consortium TIS. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009 Aug;460(7256):748–52.

113. Wise AL, Gyi L, Manolio TA. eXclusion: Toward Integrating the X Chromosome in Genome-wide Association Analyses. Am J Hum Genet. 2013 May 2;92(5):643–7.

114. Chang D, Gao F, Slavney A, Ma L, Waldman YY, Sams AJ, et al. Accounting for eXentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. PloS One. 2014;9(12):e113684.

115. Gao F, Chang D, Biddanda A, Ma L, Guo Y, Zhou Z, et al. XWAS: A Software Toolset for Genetic Data Analysis and Association Studies of the X Chromosome. J Hered. 2015 Oct;106(5):666–71.

116. Ma L, Hoffman G, Keinan A. X-inactivation informs variance-based testing for X-linked association of a quantitative trait. BMC Genomics [Internet]. 2015 Dec [cited 2018 Aug 12];16(1). Available from: http://www.biomedcentral.com/1471-2164/16/241

117. Wang J, Yu R, Shete S. X-Chromosome Genetic Association Test Accounting for X-Inactivation, Skewed X-Inactivation, and Escape from X-Inactivation. Genet Epidemiol [Internet]. 2014 [cited 2015 May 2];38. Available from: http://www.readcube.com/articles/10.1002%2Fgepi.21814

118. Bianchi I, Lleo A, Gershwin ME, Invernizzi P. The X chromosome and immune associated genes. J Autoimmun. 2012 May;38(2–3):J187-192.

119. Brooks WH. X chromosome inactivation and autoimmunity. Clin Rev Allergy Immunol. 2010 Aug;39(1):20–9.

120. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.

121. Marchini J, Howie B, Myers S, McVean G, Donnelly P. A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet. 2007 Jul;39(7):906–13.

122. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017 Jan 4;45(Database issue):D896–901.

123. Curtis J, Luo Y, Zenner HL, Cuchet-Lourenço D, Wu C, Lo K, et al. Susceptibility to tuberculosis is associated with variants in the ASAP1 gene encoding a regulator of dendritic cell migration. Nat Genet. 2015 May;47(5):523–7.

124. Grant AV, Sabri A, Abid A, Abderrahmani Rhorfi I, Benkirane M, Souhi H, et al. A genome-wide association study of pulmonary tuberculosis in Morocco. Hum Genet. 2016 Mar;135(3):299–307.

125. Mahasirimongkol S, Yanai H, Mushiroda T, Promphittayarat W, Wattanapokayakit S, Phromjai J, et al. Genome-wide association studies of tuberculosis in Asians identify distinct at-risk locus for young tuberculosis. J Hum Genet. 2012 Jun;57(6):363–7.

126. Oki NO, Motsinger-Reif AA, Antas PR, Levy S, Holland SM, Sterling TR. Novel human genetic variants associated with extrapulmonary tuberculosis: a pilot genome wide association study. BMC Res Notes. 2011 Jan 31;4:28.

127. Png E, Alisjahbana B, Sahiratmadja E, Marzuki S, Nelwan R, Balabanova Y, et al. A genome wide association study of pulmonary tuberculosis susceptibility in Indonesians. BMC Med Genet. 2012;13:5.

128. Qi H, Zhang Y-B, Sun L, Chen C, Xu B, Xu F, et al. Discovery of susceptibility loci associated with tuberculosis in Han Chinese. Hum Mol Genet. 2017 Dec 1;26(23):4752–63.

129. Sobota RS, Stein CM, Kodaman N, Scheinfeldt LB, Maro I, Wieland-Alter W, et al. A Locus at 5q33.3 Confers Resistance to Tuberculosis in Highly Susceptible Individuals. Am J Hum Genet. 2016 Mar 3;98(3):514–24.

130. Thye T, Owusu-Dabo E, Vannberg FO, van Crevel R, Curtis J, Sahiratmadja E, et al. Common variants at 11p13 are associated with susceptibility to tuberculosis. Nat Genet. 2012 Mar;44(3):257–9.

131. Thye T, Vannberg FO, Wong SH, Owusu-Dabo E, Osei I, Gyapong J, et al. Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. NatGenet. 2010 Sep;42(9):739–41.

132. Bukhari M, Aslam MA, Khan A, Iram Q, Akbar A, Naz AG, et al. TLR8 gene polymorphism and association in bacterial load in southern Punjab of Pakistan: an association study with pulmonary tuberculosis. Int J Immunogenet. 2015 Feb;42(1):46–51.

133. Chimusa ER, Zaitlen N, Daya M, Möller M, Helden PD van, Mulder NJ, et al. Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. Hum Mol Genet. 2014 Feb 1;23(3):796–809.

134. Dalgic N, Tekin D, Kayaalti Z, Cakir E, Soylemezoglu T, Sancar M. Relationship between toll-like receptor 8 gene polymorphisms and pediatric pulmonary tuberculosis. Dis Markers. 2011;31(1):33–8.

135. Davila S, Hibberd ML, Hari Dass R, Wong HEE, Sahiratmadja E, Bonnard C, et al. Genetic association and expression studies indicate a role of toll-like receptor 8 in pulmonary tuberculosis. PLoS Genet. 2008 Oct;4(10):e1000218.

136. Hashemi-Shahri SM, Taheri M, Gadari A, Naderi M, Bahari G, Hashemi M. Association Between TLR8 and TLR9 Gene Polymorphisms and Pulmonary Tuberculosis. Gene Cell Tissue [Internet]. 2014 Apr [cited 2015 Jan 30];1(1). Available from: http://genecelltissue.com/18316.abstract

137. Lai Y-F, Lin T-M, Wang C-H, Su P-Y, Wu J-T, Lin M-C, et al. Functional polymorphisms of the TLR7 and TLR8 genes contribute to Mycobacterium tuberculosis infection. Tuberc Edinb Scotl. 2016 May;98:125–31.

138. Salie M, Daya M, Lucas LA, Warren RM, van der Spuy GD, van Helden PD, et al. Association of toll-like receptors with susceptibility to tuberculosis suggests sex-specific effects of TLR8 polymorphisms. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2015 Aug;34:221–9.

139. Agrelo R, Wutz A. ConteXt of change—X inactivation and disease. EMBO Mol Med. 2010 Jan;2(1):6–15.

140. Jaillon S, Berthenet K, Garlanda C. Sexual Dimorphism in Innate Immunity. Clin Rev Allergy Immunol. 2017 Sep 30;

141. Siddiqui RA, Sauermann U, Altmüller J, Fritzer E, Nothnagel M, Dalibor N, et al. X Chromosomal Variation Is Associated with Slow Progression to AIDS in HIV-1-Infected Women. Am J Hum Genet. 2009 Aug 14;85(2):228–39.

142. R Development Core Team. R: A language and environment for statistical computing. www.r-project.org [Internet]. R foundation for statistical computing, Vienna, Austria; 2013. Available from: http;//www.R-project.org

143. Cutolo M, Capellino S, Sulli A, Serioli B, Secchi ME, Villaggio B, et al. Estrogens and autoimmune diseases. Ann N Y Acad Sci. 2006 Nov;1089:538–47.

144. Klein SL, Marriott I, Fish EN. Sex-based differences in immune function and responses to vaccination. Trans R Soc Trop Med Hyg. 2015 Jan;109(1):9–15.

145. Chimusa ER, Zaitlen N, Daya M, Möller M, van Helden PD, Mulder NJ, et al. Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. Hum Mol Genet. 2013 Sep 20;

146. Cox MP, Karafet TM, Lansing JS, Sudoyo H, Hammer MF. Autosomal and X-linked single nucleotide polymorphisms reveal a steep Asian-Melanesian ancestry cline in eastern Indonesia and a sex bias in admixture rates. Proc Biol Sci. 2010 May 22;277(1687):1589–96.

147. Pereira V, Tomas C, Sanchez JJ, Syndercombe-Court D, Amorim A, Gusmão L, et al. The peopling of Greenland: further insights from the analysis of genetic diversity using autosomal and X-chromosomal markers. Eur J Hum Genet EJHG. 2015 Feb;23(2):245–51.

148. Balaton BP, Dixon-McDougall T, Peeters SB, Brown CJ. The eXceptional nature of the X chromosome. Hum Mol Genet. 2018 Apr 26;

149. Hughes JF, Page DC. The Biology and Evolution of Mammalian Y Chromosomes. Annu Rev Genet. 2015;49:507–27.

150. Lahn BT, Page DC. Four evolutionary strata on the human X chromosome. Science. 1999 Oct 29;286(5441):964–7.

151. Chromosome X: 1-1 - Chromosome summary - Homo sapiens - Ensembl genome browser 88 [Internet]. [cited 2018 Oct 22]. Available from: http://mar2017.archive.ensembl.org/Homo_sapiens/Location/Chromosome?r=X

152. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009 Jun;5(6):e1000529.

153. Delaneau O, Coulonges C, Zagury J-F. Shape-IT: new rapid and accurate algorithm for haplotype inference. BMC Bioinformatics. 2008 Dec 16;9:540.

154. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res [Internet]. 2009 Jul 31 [cited 2015 Oct 7]; Available from: http://genome.cshlp.org/content/early/2009/07/31/gr.094052.109

155. Maples BK, Gravel S, Kenny EE, Bustamante CD. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. Am J Hum Genet. 2013 Aug 8;93(2):278–88.

156. Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. Proc Natl Acad Sci. 2010 Jan 12;107(2):786–91.

157. Bryc K, Velez C, Karafet T, Moreno-Estrada A, Reynolds A, Auton A, et al. Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. Proc Natl Acad Sci U S A. 2010 May 11;107(Suppl 2):8954–61.

158. Wang S, Ray N, Rojas W, Parra MV, Bedoya G, Gallo C, et al. Geographic Patterns of Genome Admixture in Latin American Mestizos. PLOS Genet. 2008 Mar 21;4(3):e1000037.

159. Washburn TC, Medearis DN, Childs B. SEX DIFFERENCES IN SUSCEPTIBILITY TO INFECTIONS. Pediatrics. 1965 Jan;35:57–64.

160. Abramowitz LK, Olivier-Van Stichelen S, Hanover JA. Chromosome imbalance as a driver of sex disparity in disease. J Genomics. 2014;2:77–88.

161. Brooks WH. X chromosome inactivation and autoimmunity. Clin Rev Allergy Immunol. 2010 Aug;39(1):20–9.

162. Kawai T, Akira S. TLR signaling. Cell Death Differ. 2006 May;13(5):816–25.

163. Bustamante J, Picard C, Boisson-Dupuis S, Abel L, Casanova J-L. Genetic lessons learned from X-linked Mendelian susceptibility to mycobacterial diseases. Ann N Y Acad Sci. 2011 Dec;1246:92–101.

164. Bianchi I, Lleo A, Gershwin ME, Invernizzi P. The X chromosome and immune associated genes. J Autoimmun. 2012 May;38(2–3):J187-192.

165. Nhamoyebonde S, Leslie A. Biological differences between the sexes and susceptibility to tuberculosis. J Infect Dis. 2014 Jul 15;209 Suppl 3:S100-106.

166. Borgdorff MW, Nagelkerke NJ, Dye C, Nunn P. Gender and tuberculosis: a comparison of prevalence surveys with notification data to explore sex differences in case detection. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2000 Feb;4(2):123–32.

167. Fish EN. The X-files in immunity: sex-based differences predispose immune responses. Nat Rev Immunol. 2008 Sep;8(9):737–44.

168. Klein SL, Flanagan KL. Sex differences in immune responses. Nat Rev Immunol. 2016 Oct;16(10):626.

169. Libert C, Dejager L, Pinheiro I. The X chromosome in immune functions: when a chromosome makes the difference. Nat Rev Immunol. 2010 Aug;10(8):594.

170. vom Steeg LG, Klein SL. SeXX Matters in Infectious Disease Pathogenesis. PLoS Pathog. 2016 Feb;12(2):e1005374.

171. Cutolo M, Capellino S, Sulli A, Serioli B, Secchi ME, Villaggio B, et al. Estrogens and autoimmune diseases. Ann N Y Acad Sci. 2006 Nov;1089:538–47.

172. Cutolo M, Capellino S, Sulli A, Serioli B, Secchi ME, Villaggio B, et al. Estrogens and autoimmune diseases. Ann N Y Acad Sci. 2006 Nov;1089:538–47.

173. Neyrolles O, Quintana-Murci L. Sexual inequality in tuberculosis. PLoSMed. 2009 Dec;6(12):e1000199.

174. van Lunzen J, Altfeld M. Sex differences in infectious diseases-common but neglected. J Infect Dis. 2014 Jul 15;209 Suppl 3:S79-80.

175. Brockdorff N. Chromosome silencing mechanisms in X-chromosome inactivation: unknown unknowns. Dev Camb Engl. 2011 Dec;138(23):5057–65.

176. Naqvi S, Bellott DW, Lin KS, Page DC. Conserved microRNA targeting reveals preexisting gene dosage sensitivities that shaped amniote sex chromosome evolution. Genome Res. 2018 Feb 15;gr.230433.117.

177. Migeon BR. Choosing the Active X: The Human Version of X Inactivation. Trends Genet TIG. 2017;33(12):899–909.

178. Lyon MF. Gene Action in the X-chromosome of the Mouse (Mus musculus L.). Nature. 1961 Apr;190(4773):372–3.

179. Lyon MF. Possible Mechanisms of X Chromosome Inactivation. Nature. 1971 Aug 25;232(34):229–32.

180. X-chromosome inactivation and its implications for human disease | bioRxiv [Internet]. [cited 2018 Oct 22]. Available from: https://www.biorxiv.org/content/early/2017/03/07/076950

181. Cantone I, Fisher AG. Human X chromosome inactivation and reactivation: implications for cell reprogramming and disease. Philos Trans R Soc Lond B Biol Sci. 2017 Nov 5;372(1733).

182. Orstavik KH. X chromosome inactivation in clinical practice. Hum Genet. 2009 Sep;126(3):363–73.

183. Avner P, Heard E. X-chromosome inactivation: counting, choice and initiation. Nat Rev Genet. 2001 Jan;2(1):59–67.

184. Peeters SB, Korecki AJ, Simpson EM, Brown CJ. Human cis-acting elements regulating escape from X-chromosome inactivation function in mouse. Hum Mol Genet. 2018 Apr 1;27(7):1252–62.

185. Berletch JB, Yang F, Disteche CM. Escape from X inactivation in mice and humans. Genome Biol. 2010;11(6):213.

186. Moreira de Mello JC, Fernandes GR, Vibranovski MD, Pereira LV. Early X chromosome inactivation during human preimplantation development revealed by single-cell RNA-sequencing. Sci Rep [Internet]. 2017 Dec [cited 2018 Oct 22];7(1). Available from: http://www.nature.com/articles/s41598-017-11044-z

187. Brown CJ, Ballabio A, Rupert JL, Lafreniere RG, Grompe M, Tonlorenzi R, et al. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. Nature. 1991 Jan 3;349(6304):38–44.

188. Vallot C, Ouimette J-F, Makhlouf M, Féraud O, Pontis J, Côme J, et al. Erosion of X Chromosome Inactivation in Human Pluripotent Cells Initiates with XACT Coating and Depends on a Specific Heterochromatin Landscape. Cell Stem Cell. 2015 May 7;16(5):533–46.

189. Vallot C, Huret C, Lesecque Y, Resch A, Oudrhiri N, Bennaceur-Griscelli A, et al. *XACT*, a long noncoding transcript coating the active X chromosome in human pluripotent cells. Nat Genet. 2013 Mar;45(3):239–41.

190. Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. Cell. 2016 May 5;165(4):1012–26.

191. Okamoto I, Patrat C, Thépot D, Peynot N, Fauque P, Daniel N, et al. Eutherian mammals use diverse strategies to initiate X-chromosome inactivation during development. Nature. 2011 Apr;472(7343):370–4.

192. Migeon BR, Beer MA, Bjornsson HT. Embryonic loss of human females with partial trisomy 19 identifies region critical for the single active X. Wutz A, editor. PLOS ONE. 2017 Apr 12;12(4):e0170403.

193. Syrett CM, Sindhava V, Hodawadekar S, Myles A, Liang G, Zhang Y, et al. Loss of Xist RNA from the inactive X during B cell development is restored in a dynamic YY1-dependent two-step process in activated B cells. Chadwick BP, editor. PLOS Genet. 2017 Oct 9;13(10):e1007050.

194. Vallot C, Patrat C, Collier AJ, Huret C, Casanova M, Liyakat Ali TM, et al. XACT Noncoding RNA Competes with XIST in the Control of X Chromosome Activity during Human Early Development. Cell Stem Cell. 2017 05;20(1):102–11.

195. Horvath JE, Sheedy CB, Merrett SL, Diallo AB, Swofford DL, NISC Comparative Sequencing Program null, et al. Comparative analysis of the primate X-inactivation center region and reconstruction of the ancestral primate XIST locus. Genome Res. 2011 Jun;21(6):850–62.

196. Plenge RM, Hendrich BD, Schwartz C, Arena JF, Naumova A, Sapienza C, et al. A promoter mutation in the XIST gene in two unrelated families with skewed X-chromosome inactivation. Nat Genet. 1997 Nov;17(3):353–6.

197. Chaligné R, Popova T, Mendoza-Parra M-A, Saleem M-AM, Gentien D, Ban K, et al. The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer. Genome Res. 2015 Apr;25(4):488–503.

198. Van der Meulen J, Sanghvi V, Mavrakis K, Durinck K, Fang F, Matthijssens F, et al. The H3K27me3 demethylase UTX is a gender-specific tumor suppressor in T-cell acute lymphoblastic leukemia. Blood. 2015 Jan 1;125(1):13–21.

199. Invernizzi P, Pasini S, Selmi C, Gershwin ME, Podda M. Female predominance and X chromosome defects in autoimmune diseases. J Autoimmun. 2009 Aug;33(1):12–6.

200. Berletch JB, Yang F, Xu J, Carrel L, Disteche CM. Genes that escape from X inactivation. Hum Genet. 2011 Aug;130(2):237–45.

201. Brown CJ, Greally JM. A stain upon the silence: genes escaping X inactivation. Trends Genet TIG. 2003 Aug;19(8):432–8.

202. Dunford A, Weinstock DM, Savova V, Schumacher SE, Cleary JP, Yoda A, et al. Tumor suppressor genes that escape from X-inactivation contribute to cancer sex bias. Nat Genet. 2017 Jan;49(1):10–6.

203. Renault NKE, Pritchett SM, Howell RE, Greer WL, Sapienza C, Ørstavik KH, et al. Human X-chromosome inactivation pattern distributions fit a model of genetically influenced choice better than models of completely random choice. Eur J Hum Genet EJHG. 2013 Dec;21(12):1396–402.

204. Ruttum MS, Lewandowski MF, Bateman JB. Affected females in X-linked congenital stationary night blindness. Ophthalmology. 1992 May;99(5):747–52.

205. Brown CJ, Robinson WP. The causes and consequences of random and non-random X chromosome inactivation in humans. Clin Genet. 2000 Nov;58(5):353–63.

206. Busque L, Mio R, Mattioli J, Brais E, Blais N, Lalonde Y, et al. Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. Blood. 1996 Jul 1;88(1):59–65.

207. Wareham KA, Lyon MF, Glenister PH, Williams ED. Age related reactivation of an X-linked gene. Nature. 1987 Jul 25;327(6124):725–7.

208. Sharp A, Robinson D, Jacobs P. Age- and tissue-specific variation of X chromosome inactivation ratios in normal women. Hum Genet. 2000 Oct;107(4):343–9.

209. Migeon BR, Axelman J, Beggs AH. Effect of ageing on reactivation of the human X-linked HPRT locus. Nature. 1988 Sep;335(6185):93–6.

210. Gale RE, Fielding AK, Harrison CN, Linch DC. Acquired skewing of X-chromosome inactivation patterns in myeloid cells of the elderly suggests stochastic clonal loss with age. Br J Haematol. 1997 Sep;98(3):512–9.

211. Vickers MA. Assessment of mechanism of acquired skewed X inactivation by analysis of twins. Blood. 2001 Mar 1;97(5):1274–81.

212. Vallot C, Ouimette J-F, Rougeulle C. Establishment of X chromosome inactivation and epigenomic features of the inactive X depend on cellular contexts. BioEssays News Rev Mol Cell Dev Biol. 2016 Sep;38(9):869–80.

213. McClelland EE, Smith JM. Gender specific differences in the immune response to infection. Arch Immunol Ther Exp (Warsz). 2011 Jun;59(3):203–13.

214. Marcus U, Bremer V, Hamouda O. Syphilis surveillance and trends of the syphilis epidemic in Germany since the mid-90s. Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull. 2004 Dec;9(12):11–4.

215. Pope V, Larsen SA, Rice RJ, Goforth SN, Parham CE, Fears MB. Flow cytometric analysis of peripheral blood lymphocyte immunophenotypes in persons infected with Treponema pallidum. Clin Diagn Lab Immunol. 1994 Jan 1;1(1):121–4.

216. Righarts AA, Simms I, Wallace L, Solomou M, Fenton KA. Syphilis surveillance and epidemiology in the United Kingdom. Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull. 2004 Dec;9(12):21–5.

217. Jarefors S, Bennet L, You E, Forsberg P, Ekerfelt C, Berglund J, et al. Lyme borreliosis reinfection: might it be explained by a gender difference in immune response? Immunology. 2006 Jun;118(2):224–35.

218. Schwartz AM, Hinckley AF, Mead PS, Hook SA, Kugeler KJ. Surveillance for Lyme Disease—United States, 2008–2015. MMWR Surveill Summ. 2017;66(22):1.

219. Kuo Chou T-N, Chao W-N, Yang C, Wong R-H, Ueng K-C, Chen S-C. Predictors of mortality in skin and soft-tissue infections caused by Vibrio vulnificus. World J Surg. 2010 Jul;34(7):1669–75.

220. Allard C, Carignan A, Bergevin M, Boulais I, Tremblay V, Robichaud P, et al. Secular changes in incidence and mortality associated with Staphylococcus aureus bacteraemia in Quebec, Canada, 1991–2005. Clin Microbiol Infect. 2008 May;14(5):421–8.

221. Laupland KB, Gregson DB, Church DL, Ross T, Pitout JDD. Incidence, risk factors and outcomes of Escherichia coli bloodstream infections in a large Canadian region. Clin Microbiol Infect. 2008 Nov;14(11):1041–7.

222. Aguiar PADF de, Pedroso R dos S, Borges AS, Moreira T de A, Araújo LB de, Röder DVD de B. The epidemiology of cryptococcosis and the characterization of Cryptococcus neoformans isolated in a Brazilian University Hospital. Rev Inst Med Trop São Paulo [Internet]. 2017 [cited 2018 Jul 1];59(0). Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0036-46652017005000208&lng=en&tlng=en

223. Amornkul PN, Hu DJ, Tansuphasawadikul S, Lee S, Eampokalap B, Likanonsakul S, et al. Human immunodeficiency virus type 1 subtype and other factors associated with extrapulmonary Cryptococcosis among patients in Thailand with AIDS. AIDS Res Hum Retroviruses. 2003 Feb;19(2):85–90.

224. Micol R, Lortholary O, Sar B, Laureillard D, Ngeth C, Dousset J-P, et al. Prevalence, Determinants of Positivity, and Clinical Utility of Cryptococcal Antigenemia in Cambodian HIV-Infected Patients. JAIDS J Acquir Immune Defic Syndr. 2007 Aug 15;45(5):555.

225. Li S, Yu X, Wu W, Chen DZ, Xiao M, Huang X. The opportunistic human fungal pathogen Candida albicans promotes the growth and proliferation of commensal Escherichia coli through an iron-responsive pathway. Microbiol Res. 2018 Mar;207:232–9.

226. Ruiz-Herrera J, Victoria Elorza M, Valentín E, Sentandreu R. Molecular organization of the cell wall of Candida albicans and its relation to pathogenicity. FEMS Yeast Res. 2006 Jan 1;6(1):14–29.

227. Ellabib MS, Agaj M, Khalifa Z, Kavanagh K. Yeasts of the genus Candida are the dominant cause of onychomycosis in Libyan women but not men: results of a 2-year surveillance study. Br J Dermatol. 2002 Jun;146(6):1038–41.

228. Shi W, Mei X, Gao F, Huo K, Shen L, Qin H, et al. Analysis of genital Candida albicans infection by rapid microsatellite markers genotyping. Chin Med J (Engl). 2007 Jun 5;120(11):975–80.

229. White S, Larsen B. Candida albicans morphogenesis is influenced by estrogen. Cell Mol Life Sci CMLS. 1997 Oct 1;53(9):744–9.

230. Zhang X, Essmann M, Burt ET, Larsen B. Estrogen effects on Candida albicans: a potential virulence-regulating mechanism. J Infect Dis. 2000;181(4):1441–1446.

231. Restrepo A, Benard G, Castro CC de, Agudelo CA, Tobón AM. Pulmonary Paracoccidioidomycosis. Semin Respir Crit Care Med. 2008 Apr;29(02):182–97.

232. Kelvin EA, Carpio A, Bagiella E, Leslie D, Leon P, Andrews H, et al. The association of host age and gender with inflammation around neurocysticercosis cysts. Ann Trop Med Parasitol. 2009 Sep;103(6):487–99.

233. Lezama-Davila CM, Oghumu S, Satoskar AR, Isaac-Marquez AP. Sex-associated Susceptibility in Humans with Chiclero's Ulcer: Resistance in Females is Associated with Increased Serum-levels of GM-CSF. Scand J Immunol. 65(2):210–1.

234. Sady H, Al-Mekhlafi HM, Mahdy MAK, Lim YAL, Mahmud R, Surin J. Prevalence and Associated Factors of Schistosomiasis among Children in Yemen: Implications for an Effective Control Programme. PLoS Negl Trop Dis [Internet]. 2013 Aug 22;7(8). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3749985/

235. Anastos K, Gange SJ, Lau B, Weiser B, Detels R, Giorgi JV, et al. Association of race and gender with HIV-1 RNA levels and immunologic progression. J Acquir Immune Defic Syndr 1999. 2000 Jul 1;24(3):218–26.

236. Anejo-Okopi J, Abah IO, Barshep Y, Ebonyi AO, Daniyam C, Isa SE, et al. Demographic and clinical correlates of HIV-1 RNA levels in antiretroviral therapy-naive adults attending a tertiary hospital in Jos, Nigeria. J Virus Erad. 3(1):51–5.

237. Meier A, Chang JJ, Chan ES, Pollard RB, Sidhu HK, Kulkarni S, et al. Sex differences in the Toll-like receptor–mediated response of plasmacytoid dendritic cells to HIV-1. Nat Med. 2009 Aug;15(8):955–9.

238. Network EPHCV. A significant sex—but not elective cesarean section—effect on mother-to-child transmission of hepatitis C virus infection. J Infect Dis. 2005;192(11):1872–1879.

239. Schott E, Witt H, Hinrichsen H, Neumann K, Weich V, Bergk A, et al. Gender-dependent association of CTLA4 polymorphisms with resolution of hepatitis C virus infection. J Hepatol. 2007 Mar 1;46(3):372–80.

240. Arnold AP, Chen X. What does the "four core genotypes" mouse model tell us about sex differences in the brain and other tissues? Front Neuroendocrinol. 2009 Jan;30(1):1–9.

241. Wang J, Syrett CM, Kramer MC, Basu A, Atchison ML, Anguera MC. Unusual maintenance of X chromosome inactivation predisposes female lymphocytes for increased expression from the inactive X. Proc Natl Acad Sci U S A. 2016 Apr 5;113(14):E2029-2038.

242. Invernizzi P, Pasini S, Selmi C, Gershwin ME, Podda M. Female predominance and X chromosome defects in autoimmune diseases. J Autoimmun. 2009 Aug;33(1):12–6.

243. Houben RMGJ, Dodd PJ. The Global Burden of Latent Tuberculosis Infection: A Re-estimation Using Mathematical Modelling. PLoS Med. 2016 Oct;13(10):e1002152.

244. Kaufmann SHE, Dorhoi A, Hotchkiss RS, Bartenschlager R. Host-directed therapies for bacterial and viral infections. Nat Rev Drug Discov. 2018 Jan;17(1):35–56.

245. Kinnear C, Hoal EG, Schurz H, van Helden PD, Möller M. The role of human host genetics in tuberculosis resistance. Expert Rev Respir Med. 2017 Sep;11(9):721–37.

246. WHO | Global tuberculosis report 2017 [Internet]. WHO. [cited 2018 Jan 19]. Available from: http://www.who.int/tb/publications/global_report/en/

247. Holmes CB, Hausler H, Nunn P. A review of sex differences in the epidemiology of tuberculosis. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 1998 Feb;2(2):96–104.

248. Hamid Salim MA, Declercq E, Van Deun A, Saki K a. R. Gender differences in tuberculosis: a prevalence survey done in Bangladesh. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2004 Aug;8(8):952–7.

249. Pinzan CF, Ruas LP, Casabona-Fortunato AS, Carvalho FC, Roque-Barreira M-C. Immunological Basis for the Gender Differences in Murine Paracoccidioides brasiliensis Infection. PLoS ONE. 2010 May 21;5(5):e10757.

250. Lotter H, Helk E, Bernin H, Jacobs T, Prehn C, Adamski J, et al. Testosterone Increases Susceptibility to Amebic Liver Abscess in Mice and Mediates Inhibition of IFNγ Secretion in Natural Killer T Cells. PLoS ONE. 2013 Feb 12;8(2):e55694.

251. Borgdorff MW, Nagelkerke NJ, Dye C, Nunn P. Gender and tuberculosis: a comparison of prevalence surveys with notification data to explore sex differences in case detection. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2000 Feb;4(2):123–32.

252. Bashour H, Mamaree F. Gender differences and tuberculosis in the Syrian Arab Republic: patients' attitudes, compliance and outcomes. East Mediterr Health J Rev Santé Méditerranée Orient Al-Majallah Al-Ṣiḥḥīyah Li-Sharq Al-Mutawassiṭ. 2003 Jul;9(4):757–68.

253. Watkins RE, Plant AJ. Does smoking explain sex differences in the global tuberculosis epidemic? Epidemiol Infect. 2006 Apr;134(2):333–9.

254. Pinzan CF, Ruas LP, Casabona-Fortunato AS, Carvalho FC, Roque-Barreira M-C. Immunological Basis for the Gender Differences in Murine Paracoccidioides brasiliensis Infection. PLoS ONE. 2010 May 21;5(5):e10757.

255. Lotter H, Helk E, Bernin H, Jacobs T, Prehn C, Adamski J, et al. Testosterone Increases Susceptibility to Amebic Liver Abscess in Mice and Mediates Inhibition of IFNγ Secretion in Natural Killer T Cells. PLoS ONE. 2013 Feb 12;8(2):e55694.

256. Eum S-Y, Kong J-H, Hong M-S, Lee Y-J, Kim J-H, Hwang S-H, et al. Neutrophils are the predominant infected phagocytic cells in the airways of patients with active pulmonary TB. Chest. 2010 Jan;137(1):122–8.

257. Martineau AR, Newton SM, Wilkinson KA, Kampmann B, Hall BM, Nawroly N, et al. Neutrophil-mediated innate immune resistance to mycobacteria. J Clin Invest. 2007 Jul 2;117(7):1988–94.

258. Deitch EA, Ananthakrishnan P, Cohen DB, Xu DZ, Feketeova E, Hauser CJ. Neutrophil activation is modulated by sex hormones after trauma-hemorrhagic shock and burn injuries. Am J Physiol Heart Circ Physiol. 2006 Sep;291(3):H1456-1465.

259. Pinheiro I, Dejager L, Libert C. X-chromosome-located microRNAs in immunity: Might they explain male/female differences? BioEssays. 2011 Nov 1;33(11):791–802.

260. Dorhoi A, Iannaccone M, Farinacci M, Faé KC, Schreiber J, Moura-Alves P, et al. MicroRNA-223 controls susceptibility to tuberculosis by regulating lung neutrophil recruitment. J Clin Invest. 2013 Nov;123(11):4836–48.

261. Fox GJ, Orlova M, Schurr E. Tuberculosis in Newborns: The Lessons of the "Lübeck Disaster" (1929-1933). PLoS Pathog. 2016 Jan;12(1):e1005271.

262. Filipe-Santos O, Bustamante J, Haverkamp MH, Vinolo E, Ku C-L, Puel A, et al. X-linked susceptibility to mycobacteria is caused by mutations in NEMO impairing CD40-dependent IL-12 production. J Exp Med. 2006 Jul 10;203(7):1745–59.

263. Bustamante J, Picard C, Fieschi C, Filipe-Santos O, Feinberg J, Perronne C, et al. A novel X-linked recessive form of Mendelian susceptibility to mycobaterial disease. J Med Genet. 2007 Feb;44(2):e65.

264. Bustamante J, Picard C, Boisson-Dupuis S, Abel L, Casanova J-L. Genetic lessons learned from X-linked Mendelian susceptibility to mycobacterial diseases. Ann N Y Acad Sci. 2011 Dec;1246:92–101.

265. Bellamy R, Beyers N, McAdam KP, Ruwende C, Gie R, Samaai P, et al. Genetic susceptibility to tuberculosis in Africans: a genome-wide scan. ProcNatlAcadSciUSA. 2000 Jul 5;97(14):8005–9.

266. Campbell SJ, Sabeti P, Fielding K, Sillah J, Bah B, Gustafson P, et al. Variants of the CD40 ligand gene are not associated with increased susceptibility to tuberculosis in West Africa. Immunogenetics. 2003 Oct;55(7):502–7.

267. Davila S, Hibberd ML, Hari Dass R, Wong HEE, Sahiratmadja E, Bonnard C, et al. Genetic association and expression studies indicate a role of toll-like receptor 8 in pulmonary tuberculosis. PLoS Genet. 2008 Oct;4(10):e1000218.

268. Dalgic N, Tekin D, Kayaalti Z, Cakir E, Soylemezoglu T, Sancar M. Relationship between toll-like receptor 8 gene polymorphisms and pediatric pulmonary tuberculosis. Dis Markers. 2011;31(1):33–8.

269. Bukhari M, Aslam MA, Khan A, Iram Q, Akbar A, Naz AG, et al. TLR8 gene polymorphism and association in bacterial load in southern Punjab of Pakistan: an association study with pulmonary tuberculosis. Int J Immunogenet. 2015 Feb;42(1):46–51.

270. Seldin MF, Pasaniuc B, Price AL. New approaches to disease mapping in admixed populations. Nat Rev Genet. 2011 Jun 28;12(8):523–8.

271. Rasteiro R, Bouttier P-A, Sousa VC, Chikhi L. Investigating sex-biased migration during the Neolithic transition in Europe, using an explicit spatial simulation framework. Proc Biol Sci. 2012 Jun 22;279(1737):2409–16.

272. Helgason A, Hickey E, Goodacre S, Bosnes V, Stefánsson K, Ward R, et al. mtDna and the islands of the North Atlantic: estimating the proportions of Norse and Gaelic ancestry. Am J Hum Genet. 2001 Mar;68(3):723–37.

273. Wen B, Xie X, Gao S, Li H, Shi H, Song X, et al. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. Am J Hum Genet. 2004 May;74(5):856–65.

274. Carvajal-Carmona LG, Soto ID, Pineda N, Ortíz-Barrientos D, Duque C, Ospina-Duque J, et al. Strong Amerind/White Sex Bias and a Possible Sephardic Contribution among the Founders of a Population in Northwest Colombia. Am J Hum Genet. 2000 Nov;67(5):1287–95.

275. Mesa NR, Mondragón MC, Soto ID, Parra MV, Duque C, Ortíz-Barrientos D, et al. Autosomal, mtDNA, and Y-chromosome diversity in Amerinds: pre- and post-Columbian patterns of gene flow in South America. Am J Hum Genet. 2000 Nov;67(5):1277–86.

276. Parra EJ, Kittles RA, Argyropoulos G, Pfaff CL, Hiester K, Bonilla C, et al. Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. Am J Phys Anthropol. 2001 Jan;114(1):18–29.

277. Sans M, Weimer TA, Franco MHLP, Salzano FM, Bentancor N, Alvarez I, et al. Unequal contributions of male and female gene pools from parental populations in the African descendants of the city of Melo, Uruguay. Am J Phys Anthropol. 2002 May;118(1):33–44.

278. Quintana-Murci L, Harmant C, Quach H, Balanovsky O, Zaporozhchenko V, Bormans C, et al. Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture. Am J Hum Genet. 2010 Apr 9;86(4):611–20.

279. Goldberg A, Rosenberg NA. Beyond 2/3 and 1/3: The Complex Signatures of Sex-Biased Admixture on the X Chromosome. Genetics. 2015 Sep;201(1):263–79.

280. de Wit E, Delport W, Rugamika CE, Meintjes A, Möller M, van Helden PD, et al. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. Hum Genet. 2010 Aug;128(2):145–53.

281. Schurz H, Kinnear CJ, Gignoux CR, Wojcik GL, Helden PD van, Tromp GC, et al. A sex-stratified genome-wide association study of tuberculosis using a multi-ethnic genotyping array. bioRxiv. 2018 Aug 31;405571.

282. Martin AR, Lin M, Granka JM, Myrick JW, Liu X, Sockell A, et al. An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. Cell. 2017 Nov 30;171(6):1340-1353.e14.

283. Goldstein JI, Crenshaw A, Carey J, Grant GB, Maguire J, Fromer M, et al. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. Bioinforma Oxf Engl. 2012 Oct 1;28(19):2543–5.

284. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

285. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinforma Oxf Engl. 2010 Nov 15;26(22):2867–73.

286. Shringarpure SS, Bustamante CD, Lange K, Alexander DH. Efficient analysis of large datasets and sex bias with ADMIXTURE. BMC Bioinformatics. 2016 May 23;17:218.

287. Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. pong: fast analysis and visualization of latent clusters in population genetic data. Bioinformatics. 2016 Sep 15;32(18):2817–23.

288. Daya M, Merwe L van der, Galal U, Möller M, Salie M, Chimusa ER, et al. A Panel of Ancestry Informative Markers for the Complex Five-Way Admixed South African Coloured Population. PLOS ONE. 2013 Dec 20;8(12):e82224.

289. Flynn JL. Lessons from experimental Mycobacterium tuberculosis infections. MicrobesInfect. 2006 Apr;8(4):1179–88.

290. Sorensen TI, Nielsen GG, Andersen PK, Teasdale TW. Genetic and environmental influences on premature death in adult adoptees. NEnglJMed. 1988 Mar 24;318(12):727–32.

291. Schurz H, Daya M, Möller M, Hoal EG, Salie M. TLR1, 2, 4, 6 and 9 Variants Associated with Tuberculosis Susceptibility: A Systematic Review and Meta-Analysis. PloS One. 2015;10(10):e0139711.

292. Omae Y, Toyo-Oka L, Yanai H, Nedsuwan S, Wattanapokayakit S, Satproedprai N, et al. Pathogen lineage-based genome-wide association study identified CD53 as susceptible locus in tuberculosis. J Hum Genet. 2017 Dec;62(12):1015–22.

293. Campbell MC, Tishkoff SA. AFRICAN GENETIC DIVERSITY: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. Annu Rev Genomics Hum Genet. 2008;9:403–33.

294. Mersha TB, Martin LJ, Biagini Myers JM, Kovacic MB, He H, Lindsey M, et al. Genomic architecture of asthma differs by sex. Genomics. 2015 Jul 1;106(1):15–22.

295. Davila S, Hibberd ML, Hari DR, Wong HE, Sahiratmadja E, Bonnard C, et al. Genetic association and expression studies indicate a role of toll-like receptor 8 in pulmonary tuberculosis. PLoSGenet. 2008 Oct;4(10):e1000218.

296. Dalgic N, Tekin D, Kayaalti Z, Cakir E, Soylemezoglu T, Sancar M. Relationship between toll-like receptor 8 gene polymorphisms and pediatric pulmonary tuberculosis. Dis Markers. 2011 Jan 1;31(1):33–8.

297. Chimusa ER, Daya M, Möller M, Ramesar R, Henn BM, van Helden PD, et al. Determining ancestry proportions in complex admixture scenarios in South Africa using a novel proxy ancestry selection method. PloS One. 2013;8(9):e73971.

298. de Wit E, Delport W, Rugamika CE, Meintjes A, Möller M, van Helden PD, et al. Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. Hum Genet. 2010 Aug;128(2):145–53.

299. Kritzinger FE, den BS, Verver S, Enarson DA, Lombard CJ, Borgdorff MW, et al. No decrease in annual risk of tuberculosis infection in endemic area in Cape Town, South Africa. TropMedIntHealth. 2009 Feb;14(2):136–42.

300. Gallant CJ, Cobat A, Simkin L, Black GF, Stanley K, Hughes J, et al. Impact of age and sex on mycobacterial immunity in an area of high tuberculosis incidence. Int J Tuberc Lung Dis. 2010 Aug 1;14(8):952–9.

301. WMA - The World Medical Association-WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects [Internet]. [cited 2018 Aug 12]. Available from: https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/

302. Chang D, Gao F, Slavney A, Ma L, Waldman YY, Sams AJ, et al. Accounting for eXentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. PloS One. 2014;9(12):e113684.

303. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Sep;19(9):1655–64.

304. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.

305. Wang J, Yu R, Shete S. X-Chromosome Genetic Association Test Accounting for X-Inactivation, Skewed X-Inactivation, and Escape from X-Inactivation. Genet Epidemiol [Internet]. 2014 [cited 2015 May 2];38. Available from: http://www.readcube.com/articles/10.1002%2Fgepi.21814

306. Gao X, Becker LC, Becker DM, Starmer JD, Province MA. Avoiding the high Bonferroni penalty in genome-wide association studies. Genet Epidemiol. 2010 Jan;34(1):100–5.

307. Panagiotou OA, Ioannidis JPA, Genome-Wide Significance Project. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. Int J Epidemiol. 2012 Feb;41(1):273–86.

308. Mullany LE, Herrick JS, Wolff RK, Buas MF, Slattery ML. Impact of polymorphisms in microRNA biogenesis genes on colon cancer risk and microRNA expression levels: a population-based, case-control study. BMC Med Genomics. 2016 23;9(1):21.

309. Zhang X, Yang H, Lee JJ, Kim E, Lippman SM, Khuri FR, et al. MicroRNA-related genetic variations as predictors for risk of second primary tumor and/or recurrence in patients with early-stage head and neck cancer. Carcinogenesis. 2010 Dec;31(12):2118–23.

310. Ogami K, Cho R, Hoshino S. Molecular cloning and characterization of a novel isoform of the non-canonical poly(A) polymerase PAPD7. Biochem Biophys Res Commun. 2013 Mar 1;432(1):135–40.

311. Park S-Y, Na Y, Lee M-H, Seo J-S, Lee Y-H, Choi K-C, et al. SUMOylation of TBL1 and TBLR1 promotes androgen-independent prostate cancer cell growth. Oncotarget. 2016 Jul 5;7(27):41110–22.

312. Heinen CA, Losekoot M, Sun Y, Watson PJ, Fairall L, Joustra SD, et al. Mutations in TBL1X Are Associated With Central Hypothyroidism. J Clin Endocrinol Metab. 2016;101(12):4564–73.

313. Kim CJ, Shimakage M, Kushima R, Mukaisho K-I, Shinka T, Okada Y, et al. Down-regulation of drs mRNA in human prostate carcinomas. Hum Pathol. 2003 Jul;34(7):654–7.

314. Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature. 2006 Jul 13;442(7099):199–202.

315. Parandhaman DK, Narayanan S. Cell death paradigms in the pathogenesis of Mycobacterium tuberculosis infection. Front Cell Infect Microbiol [Internet]. 2014 Mar 5 [cited 2018 Jul 30];4. Available from: http://journal.frontiersin.org/article/10.3389/fcimb.2014.00031/abstract

316. Cui X, Sun X, Niu W, Kong L, He M, Zhong A, et al. Long Non-Coding RNA: Potential Diagnostic and Therapeutic Biomarker for Major Depressive Disorder. Med Sci Monit Int Med J Exp Clin Res. 2016 Dec 31;22:5240–8.

317. Tang M, Yang Y-F, Xie L, Chen J-L, Zhang W-Z, Wang J, et al. Duplication of 10q22.3-q23.3 encompassing BMPR1A and NGR3 associated with congenital heart disease, microcephaly, and mild intellectual disability. Am J Med Genet A. 2015 Dec;167A(12):3174–9.

318. Shoji-Kawata S, Zhong Q, Kameoka M, Iwabu Y, Sapsutthipas S, Luftig RB, et al. The RING finger ubiquitin ligase RNF125/TRAC-1 down-modulates HIV-1 replication in primary human peripheral blood mononuclear cells. Virology. 2007 Nov 10;368(1):191–204.

319. Zhou W, Wang Q, Xu Y, Jiang J, Guo J, Yu H, et al. RMP promotes epithelial-mesenchymal transition through NF-κB/CSN2/Snail pathway in hepatocellular carcinoma. Oncotarget. 2017 Jun 20;8(25):40373–88.

320. Deng H, Xiao H. The role of the <Emphasis Type="Italic">ATP2C1</Emphasis> gene in Hailey–Hailey disease. Cell Mol Life Sci. 2017 Oct 1;74(20):3687–96.

321. Mottok A, Woolcock B, Chan FC, Tong KM, Chong L, Farinha P, et al. Genomic Alterations in CIITA Are Frequent in Primary Mediastinal Large B Cell Lymphoma and Are Associated with Diminished MHC Class II Expression. Cell Rep. 2015 Nov 17;13(7):1418–31.

322. Maruani A, Huguet G, Beggiato A, ElMaleh M, Toro R, Leblond CS, et al. 11q24.2-25 micro-rearrangements in autism spectrum disorders: Relation to brain structures. Am J Med Genet A. 2015 Dec;167A(12):3019–30.

323. Song G, Ouyang G, Bao S. The activation of Akt/PKB signaling pathway and cell survival. J Cell Mol Med. 2005 Mar;9(1):59–71.

324. Lawrence T. The Nuclear Factor NF-κB Pathway in Inflammation. Cold Spring Harb Perspect Biol [Internet]. 2009 Dec;1(6). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2882124/

325. Daya M, van der Merwe L, van Helden PD, Möller M, Hoal EG. Investigating the Role of Gene-Gene Interactions in TB Susceptibility. PloS One. 2014;10(4):e0123970.

326. Shin YJ, Park SK, Jung YJ, Kim YN, Kim KS, Park OK, et al. Nanobody-targeted E3-ubiquitin ligase complex degrades nuclear proteins. Sci Rep. 2015 Sep 16;5:14269.

327. Bai X, Feldman NE, Chmura K, Ovrutsky AR, Su W-L, Griffin L, et al. Inhibition of Nuclear Factor-Kappa B Activation Decreases Survival of Mycobacterium tuberculosis in Human Macrophages. PLOS ONE. 2013 Apr 25;8(4):e61925.

328. Fallahi-Sichani M, Kirschner DE, Linderman JJ. NF-κB Signaling Dynamics Play a Key Role in Infection Control in Tuberculosis. Front Physiol. 2012;3:170.

329. Franco LH, Nair VR, Scharn CR, Xavier RJ, Torrealba JR, Shiloh MU, et al. The Ubiquitin Ligase Smurf1 Functions in Selective Autophagy of Mycobacterium tuberculosis and Anti-tuberculous Host Defense. Cell Host Microbe. 2017 13;22(3):421–3.

330. Hirsch CS, Johnson JL, Okwera A, Kanost RA, Wu M, Peters P, et al. Mechanisms of Apoptosis of T-Cells in Human Tuberculosis. J Clin Immunol. 2005 Jul 1;25(4):353–64.

331. Hirsch CS, Toossi Z, Vanham G, Johnson JL, Peters P, Okwera A, et al. Apoptosis and T Cell Hyporesponsiveness in Pulmonary Tuberculosis. J Infect Dis. 1999 Apr 1;179(4):945–53.

332. Torres M, Ramachandra L, Rojas RE, Bobadilla K, Thomas J, Canaday DH, et al. Role of phagosomes and major histocompatibility complex class II (MHC-II) compartment in MHC-II antigen processing of Mycobacterium tuberculosis in human macrophages. Infect Immun. 2006 Mar;74(3):1621–30.

333. Chimusa ER, Zaitlen N, Daya M, Möller M, van Helden PD, Mulder NJ, et al. Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. Hum Mol Genet. 2014 Feb 1;23(3):796–809.

334. Berhane K, McCONNELL R, Gilliland F, Islam T, James Gauderman W, Avol E, et al. Sex-specific Effects of Asthma on Pulmonary Function in Children. Am J Respir Crit Care Med. 2000 Nov 1;162(5):1723–30.

335. Ding C, Jin S. High-throughput methods for SNP genotyping. Methods Mol Biol Clifton NJ. 2009;578:245–54.

336. Ragoussis J. Genotyping technologies for genetic research. Annu Rev Genomics Hum Genet. 2009;10:117–33.

337. Vergara C, Parker MM, Franco L, Cho MH, Valencia-Duarte AV, Beaty TH, et al. Genotype imputation performance of three reference panels using African ancestry individuals. Hum Genet. 2018 Apr;137(4):281–92.

338. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010 Jul;11(7):499–511.

339. Pei Y-F, Zhang L, Li J, Deng H-W. Analyses and Comparison of Imputation-Based Association Methods. PLOS ONE. 2010 May 26;5(5):e10827.

340. Malhotra A, Kobes S, Bogardus C, Knowler WC, Baier LJ, Hanson RL. Assessing accuracy of genotype imputation in American Indians. PloS One. 2014;9(7):e102544.

341. Hancock DB, Levy JL, Gaddis NC, Bierut LJ, Saccone NL, Page GP, et al. Assessment of genotype imputation performance using 1000 Genomes in African American studies. PloS One. 2012;7(11):e50610.

342. McRae AF. Analysis of Genome-Wide Association Data. Methods Mol Biol Clifton NJ. 2017;1526:161–73.

343. Verma SS, de Andrade M, Tromp G, Kuivaniemi H, Pugh E, Namjou-Khales B, et al. Imputation and quality control steps for combining multiple genome-wide datasets. Front Genet [Internet]. 2014 [cited 2018 Jun 26];5. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2014.00370/full

344. Roshyara NR, Horn K, Kirsten H, Ahnert P, Scholz M. Comparing performance of modern genotype imputation methods in different ethnicities. Sci Rep. 2016 Oct 4;6:34386.

345. Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, Rosenberg NA, et al. Genotype-imputation accuracy across worldwide human populations. Am J Hum Genet. 2009 Feb;84(2):235–50.

346. Nelson SC, Stilp AM, Papanicolaou GJ, Taylor KD, Rotter JI, Thornton TA, et al. Improved imputation accuracy in Hispanic/Latino populations with larger and more diverse reference panels: applications in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Hum Mol Genet. 2016 Aug 1;25(15):3245–54.

347. Deelen P, Bonder MJ, van der Velde KJ, Westra H-J, Winder E, Hendriksen D, et al. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Res Notes. 2014 Dec 11;7:901.

348. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nat Methods. 2012 Feb;9(2):179–81.

349. Durbin R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). Bioinformatics. 2014 May 1;30(9):1266–72.

350. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. Nat Genet. 2016 Oct;48(10):1284–7.

351. Kim YJ, Lee J, Kim B-J, T2D-Genes Consortium, Park T. A new strategy for enhancing imputation quality of rare variants from next-generation sequencing data via combining SNP and exome chip data. BMC Genomics. 2015 Dec 29;16:1109.

352. Zheng H-F, Rong J-J, Liu M, Han F, Zhang X-W, Richards JB, et al. Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. PloS One. 2015;10(1):e0116487.

353. Marques CR, Costa GN, da Silva TM, Oliveira P, Cruz AA, Alcantara-Neves NM, et al. Suggestive association between variants in IL1RAPL and asthma symptoms in Latin American children. Eur J Hum Genet EJHG. 2017;25(4):439–45.

354. Daya M, van der Merwe L, van Helden PD, Möller M, Hoal EG. The role of ancestry in TB susceptibility of an admixed South African population. Tuberc Edinb Scotl. 2014 Jul;94(4):413–20.

355. Naranbhai V. The Role of Host Genetics (and Genomics) in Tuberculosis. Microbiol Spectr. 2016 Oct;4(5).

356. Howie BN, Donnelly P, Marchini J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLOS Genet. 2009 Jun 19;5(6):e1000529.

357. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience. 2015;4:7.

358. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002 Jun 15;21(11):1539–58.

359. Bal E, Park H-S, Belaid-Choucair Z, Kayserili H, Naville M, Madrange M, et al. Mutations in ACTRT1 and its enhancer RNA elements lead to aberrant activation of Hedgehog signaling in inherited and sporadic basal cell carcinomas. Nat Med. 2017 Oct;23(10):1226–33.

360. Holla S, Stephen-Victor E, Prakhar P, Sharma M, Saha C, Udupa V, et al. Mycobacteria-responsive sonic hedgehog signaling mediates programmed death-ligand 1- and prostaglandin E2-induced regulatory T cell expansion. Sci Rep. 2016 Apr 15;6:24193.

361. Zeng Y, He Y, Yang F, Mooney SM, Getzenberg RH, Orban J, et al. The cancer/testis antigen prostate-associated gene 4 (PAGE4) is a highly intrinsically disordered protein. J Biol Chem. 2011 Apr 22;286(16):13985–94.

362. Sampson N, Ruiz C, Zenzmaier C, Bubendorf L, Berger P. PAGE4 positivity is associated with attenuated AR signaling and predicts patient survival in hormone-naive prostate cancer. Am J Pathol. 2012 Oct;181(4):1443–54.

363. Matosin N, Green MJ, Andrews JL, Newell KA, Fernandez-Enright F. Possibility of a sex-specific role for a genetic variant in FRMPD4 in schizophrenia, but not cognitive function. Neuroreport. 2016 Jan 6;27(1):33–8.

364. Feng Y, Guan X-M, Li J, Metzger JM, Zhu Y, Juhl K, et al. Bombesin receptor subtype-3 (BRS-3) regulates glucose-stimulated insulin secretion in pancreatic islets across multiple species. Endocrinology. 2011 Nov;152(11):4106–15.

365. Ramos-Álvarez I, Martín-Duce A, Moreno-Villegas Z, Sanz R, Aparicio C, Portal-Núñez S, et al. Bombesin receptor subtype-3 (BRS-3), a novel candidate as therapeutic molecular target in obesity and diabetes. Mol Cell Endocrinol. 2013 Mar 10;367(1–2):109–15.

366. Barker CJ, Illies C, Gaboardi GC, Berggren P-O. Inositol pyrophosphates: structure, enzymology and function. Cell Mol Life Sci CMLS. 2009 Dec;66(24):3851–71.

367. Chung R-H, Ma D, Wang K, Hedges DJ, Jaworski JM, Gilbert JR, et al. An X chromosome-wide association study in autism families identifies TBL1X as a novel autism spectrum disorder candidate gene in males. Mol Autism. 2011 Nov 4;2(1):18.

368. Victora CG, Barros FC. Cohort profile: the 1982 Pelotas (Brazil) birth cohort study. Int J Epidemiol. 2006 Apr;35(2):237–42.

369. Westbrook VA, Schoppee PD, Vanage GR, Klotz KL, Diekman AB, Flickinger CJ, et al. Hominoid-specific SPANXA/D genes demonstrate differential expression in individuals and protein localization to a

distinct nuclear envelope domain during spermatid morphogenesis. Mol Hum Reprod. 2006 Nov;12(11):703–16.

370. Bixler D, Higgins M, Hartsfield J. The Nance-Horan syndrome: a rare X-linked ocular-dental trait with expression in heterozygous females. Clin Genet. 1984 Jul;26(1):30–5.

371. Horan MB, Billson FA. X-LINKED CATARACT AND HUTCHINSONIAN TEETH. J Paediatr Child Health. 2008 Mar 10;10(2):98–102.

372. Nance WE, Warburg M, Bixler D, Helveston EM. Congenital X-linked cataract, dental anomalies and brachymetacarpalia. Birth Defects Orig Artic Ser. 1974;10(4):285–91.

373. Subasioglu A, Savas S, Kucukyilmaz E, Kesim S, Yagci A, Dundar M. Genetic background of supernumerary teeth. Eur J Dent. 2015 Mar;9(1):153–8.

374. Nguyen DT, Voon HPJ, Xella B, Scott C, Clynes D, Babbs C, et al. The chromatin remodelling factor ATRX suppresses R-loops in transcribed telomeric repeats. EMBO Rep. 2017 Jun;18(6):914–28.

375. Ji J, Quindipan C, Parham D, Shen L, Ruble D, Bootwalla M, et al. Inherited germline ATRX mutation in two brothers with ATR-X syndrome and osteosarcoma. Am J Med Genet A. 2017 May;173(5):1390–5.

376. Noor A, Whibley A, Marshall CR, Gianakopoulos PJ, Piton A, Carson AR, et al. Disruption at the PTCHD1 Locus on Xp22.11 in Autism spectrum disorder and intellectual disability. Sci Transl Med. 2010 Sep 15;2(49):49ra68.

377. Neu-Yilik G, Raimondeau E, Eliseev B, Yeramala L, Amthor B, Deniaud A, et al. Dual function of UPF3B in early and late translation termination. EMBO J. 2017 16;36(20):2968–86.

378. Alrahbeni T, Sartor F, Anderson J, Miedzybrodzka Z, McCaig C, Müller B. Full UPF3B function is critical for neuronal differentiation of neural stem cells. Mol Brain [Internet]. 2015 Dec [cited 2018 Jul 27];8(1). Available from: http://www.molecularbrain.com/content/8/1/33

379. Nguyen LS, Jolly L, Shoubridge C, Chan WK, Huang L, Laumonnier F, et al. Transcriptome profiling of UPF3B/NMD-deficient lymphoblastoid cells from patients with various forms of intellectual disability. Mol Psychiatry. 2012 Nov;17(11):1103–15.

380. Xu X, Zhang L, Tong P, Xun G, Su W, Xiong Z, et al. Exome sequencing identifies UPF3B as the causative gene for a Chinese non-syndrome mental retardation pedigree. Clin Genet. 2013 Jun;83(6):560–4.

381. Bustamante J, Picard C, Fieschi C, Filipe-Santos O, Feinberg J, Perronne C, et al. A novel X-linked recessive form of Mendelian susceptibility to mycobaterial disease. J Med Genet. 2007 Feb;44(2):e65.

382. Filipe-Santos O, Bustamante J, Haverkamp MH, Vinolo E, Ku C-L, Puel A, et al. X-linked susceptibility to mycobacteria is caused by mutations in NEMO impairing CD40-dependent IL-12 production. J Exp Med. 2006 Jul 10;203(7):1745–59.

383. Gross MM, Strezoska Ž, Kelley ML. A CRISPR-Cas9 gene engineering work- flow: generating functional knockouts using Dharmacon™ Edit-R™ Cas9 and synthetic crRNA and tracrRNA. :7.

384. Shalem O, Sanjana NE, Zhang F. High-throughput functional genomics using CRISPR-Cas9. Nat Rev Genet. 2015 May;16(5):299–311.

385. Collins FS, Varmus H. A New Initiative on Precision Medicine. N Engl J Med. 2015 Feb 26;372(9):793–5.

386. Navarro P, Pichard S, Ciaudo C, Avner P, Rougeulle C. Tsix transcription across the Xist gene alters chromatin conformation without affecting Xist transcription: implications for X-chromosome inactivation. Genes Dev. 2005 Jun 15;19(12):1474–84.

387. Chu C, Zhang QC, da Rocha ST, Flynn RA, Bharadwaj M, Calabrese JM, et al. Systematic discovery of Xist RNA binding proteins. Cell. 2015 Apr 9;161(2):404–16.

388. Soldin OP, Mattison DR. Sex differences in pharmacokinetics and pharmacodynamics. Clin Pharmacokinet. 2009;48(3):143–57.

389. Quintana-Murci L, Harmant C, Quach H, Balanovsky O, Zaporozhchenko V, Bormans C, et al. Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture. AmJHumGenet. 2010 Apr 9;86(4):611–20.

390. Davila S, Hibberd ML, Hari Dass R, Wong HEE, Sahiratmadja E, Bonnard C, et al. Genetic association and expression studies indicate a role of toll-like receptor 8 in pulmonary tuberculosis. PLoS Genet. 2008 Oct;4(10):e1000218.

391. Schurz H, Daya M, Möller M, Hoal EG, Salie M. TLR1 , 2 , 4 , 6 and 9 Variants Associated with Tuberculosis Susceptibility: A Systematic Review and Meta-Analysis. PLOS ONE. 2015 Oct 2;10(10):e0139711.

9 Addendum A: Author extent of contribution

The general introduction (Chapter 1) and discussion (Chapter 7) is the candidates own work and was not co-authored. For chapter 2-6 the following authors contributed to:

Chapter 2: The X chromosome and sex-specific effects in infectious disease susceptibility.

Name	e-mail address	Nature of contribution	Extent of contribution (%)
Haiko Schurz	_____	Conceived review/ wrote first draft	90
Muneeb Salie	_____	Conceived review/ writing and proofreading	3
Gerard Tromp	_____	Writing and proofreading	1
Eileen G. Hoal	_____	Writing and proofreading	1.5
Craig J. Kinnear	_____	Writing and proofreading	1.5
Marlo Möller	_____	Conceived review/ writing and proofreading	3

Signature of candidate: Date: 19 September 2018

Chapter 3: Global ancestry inference on the autosome and X chromosome identifies sex-biased admixture in a highly admixed population.

Name	e-mail address	Nature of contribution	Extent of contribution (%)
Haiko Schurz	_____	Conceived idea/ Analysis and interpretation of data/ wrote first draft	90
Caitlin Uren	_____	Consult on admixture analysis/ writing and proofreading	1.5
Meng Lin	_____	Reference data preparation/ Consult on admixture analysis/ writing and proofreading	1.5
Craig J. Kinnear	_____	Writing and proofreading	1
Paul D van Helden	_____	Writing and proofreading	1

Name	e-mail address	Nature of contribution	Extent of contribution (%)
Gerard Tromp	_____	Consult on statistical analysis/ writing and proofreading	1
Brenna Henn	_____	Reference data preparation/ Consult on admixture analysis/ writing and proofreading	1
Eileen G. Hoal	_____	Writing and proofreading	1
Marlo Möller	_____	Conceived idea/ Writing and proofreading	2

Signature of candidate: Date: 19 September 2018

Chapter 4: A sex-stratified genome-wide association study of tuberculosis using a multi-ethnic genotyping array.

Name	e-mail address	Nature of contribution	Extent of contribution (%)
Haiko Schurz	_____	Conceived idea/ Analysis and interpretation of data/ wrote first draft	90
Craig J. Kinnear	_____	Conceived idea/ writing and proofreading	1.5
Chris Gignoux	_____	Calling and QC of raw genotypes/ writing and proofreading	1
Genevieve Wojcik	_____	Calling and QC of raw genotypes/ writing and proofreading	1
Paul D van Helden	_____	Writing and proofreading	1
Gerard Tromp	_____	Conceived idea/ Consult on statistical analysis/ writing and proofreading	1.5
Brenna Henn	_____	Calling and QC of raw genotypes/ Consult on	1

Name	e-mail address	Nature of contribution	Extent of contribution (%)
		admixture analysis/ writing and proofreading	
Eileen G. Hoal	_____	Writing and proofreading	1
Marlo Möller	_____	Conceived idea/ writing and proofreading	2

Signature of candidate: Date: 19 September 2018

Chapter 5: Evaluating the accuracy of imputation methods in a five-way admixed population.

Name	e-mail address	Nature of contribution	Extent of contribution (%)
Haiko Schurz	_____	Conceived idea/ Consulting on phasing, imputation and quality assessment/ analysed imputation accuracy/ wrote first draft	50
Stephanie Pitts	_____	Conceived idea/ Phasing/ imputation/ quality assessment/ writing and proofreading	40
Paul D van Helden	_____	Writing and proofreading	1
Gerard Tromp	_____	Consult on data analysis/ writing and proofreading	3
Eileen G. Hoal	_____	Writing and proofreading	1
Craig J. Kinnear	_____	Conceived idea/ writing and proofreading	1
Marlo Möller	_____	Conceived idea/ writing and proofreading	4

Signature of candidate: Date: 19 September 2018

Chapter 6: X-linked trans-ethnic meta-analysis reveals Tuberculosis susceptibility variants.

Name	e-mail address	Nature of contribution	Extent of contribution (%)
Haiko Schurz	_____	Conceived idea/ Analysis and interpretation of data/ wrote first draft	90
Craig J. Kinnear	_____	Writing and proofreading	1
Paul D van Helden	_____	Writing and proofreading	1
Gerard Tromp	_____	Conceived idea/ Consult on data analysis/ writing and proofreading	2
Eileen G. Hoal	_____	Writing and proofreading	1
Vivek Naranbhai	_____	Conceived idea/ Access to ITHGC data/ writing and proofreading	1
Marlo Möller	_____	Conceived idea/ writing and proofreading	4

Signature of candidate: Date: 19 September 2018

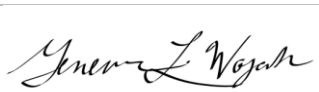
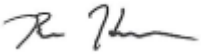
10 Addendum B: Declaration by co-authors

The undersigned hereby confirm that:

1. The declaration above accurately reflects the nature and extent of the contributions of the candidate and the co-authors to **chapters 2-6**

2. No other authors contributed to **chapters 2-6** besides those specified above, and

3. Potential conflicts of interest have been revealed to all interested parties and that the necessary arrangements have been made to use the material in **chapters 2-6** of this dissertation.

Name	Signature	Institutional affiliation	Date
Muneeb Salie		Department of Genetics, St. Jude Children's Research Hospital, Memphis, TN, 38105, USA	31/08/2018
Meng Lin		Center for Genetic Epidemiology, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033	31/08/2018
Chris Gignoux		Colorado Center for Personalized Medicine and Department of Biostatistics and Informatics, University of Colorado, Anschutz Medical Campus, Aurora, CO, 80045.	31/08/2018
Genevieve Wojcik		Department of Genetics, Stanford University, Stanford, CA, 94305.	06/09/2018
Brenna Henn		Department of Anthropology, and the UC Davis Genome Center, University of California, Davis, CA, 95616.	06/09/2018

Name	Signature	Institutional affiliation	Date
Caitlin Uren		DST-NRF Centre of Excellence for Biomedical Tuberculosis Research; South African Medical Research Council Centre for Tuberculosis Research; Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa.	10/09/2018
Stephanie Pitts			31/08/2018
Paul D van Helden			31/08/2018
Craig J. Kinnear			31/08/2018
Gerard Tromp			31/08/2018
Eileen G. Hoal			31/08/2018
Marlo Möller			31/08/2018

11 Addendum C: Ethics approval certificate

UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

Ethics Letter

05-Apr-2018

HREC Reference #: S17/01/013

Title: Sex-bias and tuberculosis susceptibility: Bioinformatic and Biostatiscal evaluation of trans-ethnic genomic datasets

Dear Mr Haiko Schurz,

Your request for extension/annual renewal of ethics approval dated 28 February 2018 refers.

The Health Research Ethics Committee reviewed and approved the annual progress report you submitted through an expedited review process.

The approval of the research project is extended for a further year.

Approval Date: 05 April 2018

Expiry Date: 04 April 2019

Kindly be reminded to submit progress reports two (2) months before expiry date.

Where to submit any documentation

Kindly note that the HREC uses an electronic ethics review management system, *Infonetica*, to manage ethics applications and ethics review process. To submit any documentation to HREC, please click on the following link: https://applyethics.sun.ac.za. Please note that you will first need to **"update"** the form in order to submit any annual progress report. Once the form has been updated, you can **"create a sub-form"** by selecting **"HREC Progress/Final Report".** You should also sign the form electronically in order for the form to be submitted for review.

Please remember to use your **Ethics Reference Number (S17/01/013)** on any documents or correspondence with the HREC concerning your research protocol and when you log into the *Infonetica* system.

National Health Research Ethics Council (NHREC) Registration Numbers: REC-130408-012 for HREC1 and REC-230208-010 for HREC2

Federal Wide Assurance Number: 00001372

Fakulteit Geneeskunde en Gesondheidswetenskappe
•
Faculty of Medicine and Health Sciences

Afdeling Navorsingsontwikkeling en -Steun • Research Development and Support Division

Posbus/PO Box 241 • Cape Town 8000 • Suid-Afrika/South Africa
Tel: +27 (0) 21 938 9677

UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

Institutional Review Board (IRB) Number: IRB0005240 for HREC1
Institutional Review Board (IRB) Number: IRB0005239 for HREC2

The Health Research Ethics Committee complies with the SA National Health Act No. 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 Part 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki and the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles, Structures and Processes 2015 (Department of Health).

Yours sincerely,

[redacted]

HREC Coordinator,
Health Research Ethics Committee 2.

STELLENBOSCH UNIVERSITY
Health Research Ethics Committee
05 APR 2018
STELLENBOSCH UNIVERSITEIT
Gesondheidsnavorsing Etiekkomitee

Fakulteit Geneeskunde en Gesondheidswetenskappe
•
Faculty of Medicine and Health Sciences

Afdeling Navorsingsontwikkeling en -Steun • Research Development and Support Division

Posbus/PO Box 241 • Cape Town 8000 • Suid-Afrika/South Africa
Tel: +27 (0) 21 938 9677