Using bioinformatics and biostatistics to elucidate susceptibility to tuberculosis in an admixed population

by

Michelle Daya

Dissertation presented for the degree of Doctor of Medical Science in the Faculty of Medicine and Health Sciences at Stellenbosch University

Department of Biomedical Sciences,
Stellenbosch University,
PO Box 19063, Francie van Zijl Drive, Tygerberg 7505, South Africa.

Promoters:

Prof. Eileen Hoal
Prof. Lize van der Merwe

March 2015

Declaration

By submitting this dissertation, 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This dissertation includes two original papers published in peer-reviewed journals and two unpublished publications. The development and writing of the papers (published and unpublished) were the principal responsibility of myself and, for each of the cases where this is not the case, a declaration is included in the dissertation indicating the nature and extent of the contributions of co-authors.

Signature: .....
M. Daya

Date: 2015/2/17 .....

Declaration by the candidate:

With regard to Research Article 1, the nature and scope of my contribution were as follows:

Nature of contribution	**Extent of contribution (%)**
Conceived idea Modified Galanter et al.'s python script for AIM selection Data analyses Wrote manuscript	75%

The following co-authors have contributed to Research Article 1:

Name	**E-mail**	**Nature of contribution**	**Extent of contribution (%)**
L van der Merwe U Galal	lizestats@gmail.com ushma.galal@uct.az.za	Critical assessment of analyses and manuscript.	10%
M Möller M Salie P D van Helden E G Hoal	marlom@sun.ac.za msalie@sun.ac.za pvh@sun.ac.za egvh@sun.ac.za	Conceived idea. Assessment of manuscript. Contributed reagents. Laboratory work.	5%
E Chimusa	echimusa@gmail.com	Determined best source populations.	5%
B M Henn C R Gignoux J M Galanter	brenna.henn@stonybrook.edu cgignoux@stanford.edu joshua.galanter@ucsf.edu	Assessment of python script changes. Identifying admixed San individuals.	5%

Declaration with signatures in possession of candidate and supervisor.

Declaration by the candidate:

With regard to Research Article 2, the nature and scope of my contribution were as follows:

Nature of contribution	**Extent of contribution (%)**
Conceived idea Data analyses Wrote manuscript	85%

The following co-authors have contributed to Research Article 2:

Name	**E-mail**	**Nature of contribution**	**Extent of contribution (%)**
L van der Merwe	lizestats@gmail.com	Critical assessment of analyses and manuscript.	10%
M Möller P D van Helden E G Hoal	marlom@sun.ac.za pvh@sun.ac.za egvh@sun.ac.za	Conceived idea. Assessment of manuscript. Contributed reagents. Laboratory work.	5%

Declaration with signatures in possession of candidate and supervisor.

Declaration by the candidate:

With regard to Research Article 3, the nature and scope of my contribution were as follows:

Nature of contribution	**Extent of contribution (%)**
Conceived idea Data analyses Wrote manuscript	80%

The following co-authors have contributed to Research Article 3:

Name	**E-mail**	**Nature of contribution**	**Extent of contribution (%)**
L van der Merwe	lizestats@gmail.com	Critical assessment of analyses and manuscript.	15%
M Möller P D van Helden E G Hoal	marlom@sun.ac.za pvh@sun.ac.za egvh@sun.ac.za	Conceived idea. Assessment of manuscript. Contributed reagents. Laboratory work.	5%

Declaration with signatures in possession of candidate and supervisor.

Declaration by the candidate:

With regard to Research Article 4, the nature and scope of my contribution were as follows:

Nature of contribution	**Extent of contribution (%)**
Conceived idea Data analyses Wrote manuscript	85%

The following co-authors have contributed to Research Article 4:

Name	**E-mail**	**Nature of contribution**	**Extent of contribution (%)**
L van der Merwe C R Gignoux	lizestats@gmail.com cgignoux@stanford.edu	Critical assessment of analyses and manuscript.	10%
M Möller P D van Helden E G Hoal	marlom@sun.ac.za pvh@sun.ac.za egvh@sun.ac.za	Conceived idea. Assessment of manuscript. Contributed reagents. Laboratory work.	5%

Declaration with signatures in possession of candidate and supervisor.

Abstract

Using bioinformatics and biostatistics to elucidate susceptibility to tuberculosis in an admixed population

M. Daya

Department of Biomedical Sciences,
Stellenbosch University,
PO Box 19063, Francie van Zijl Drive, Tygerberg 7505, South Africa.

Dissertation: Ph.D. (MedSc)

February 2015

Tuberculosis is the second leading cause of mortality from infectious disease worldwide. One of the factors involved in developing disease is the genetics of the host, yet the field of TB susceptibility genetics has not yielded the insights that were expected. The admixed South African Coloured population is the largest demographic in metropolitan areas of Cape Town that have some of the highest reported incidences of TB worldwide. The DST/NRF Centre for Biomedical TB Research at Stellenbosch University has studied a cohort of individuals from these communities for many years, in the hope of discovering novel TB susceptibility genes which may at least partly explain the exceptional rate of TB in this community. The large genotypic data sets collected by the centre are invaluable resources to aid understanding of the population genetics of the population, and to generate new hypotheses regarding genetic factors that may underpin progression to disease. Novel applications of statistical methods are used in this dissertation with a view to addressing these goals, and are presented in four research studies.

An important consideration in genetic association studies of the South African Coloured population is admixture, which may confound association results. This is the subject of the first two studies. The first study describes the development of a panel of ancestry informative markers that are tailored to the complex admixture that occurred in the population. The panel can be used as a cost-effective alternative to genome-wide data to correct for the confounding effect of admixture. In the second study, the panel is used to demonstrate the importance of adjustment for ancestry in TB susceptibility genetic association studies of the South African Coloured population. A previous study identified associations between ancestry and having TB in the population, but a limited number of controls were used in that study. Ancestry informative markers were therefore used to examine the previous finding, and the substantial effect that ancestry has in the development of TB was confirmed.

New hypotheses regarding genetic factors in TB susceptibility are generated in the third and fourth studies. The South African Coloured population received contributions from diverse source populations that may differ in their genetic susceptibility to TB, and the group is

therefore ideally suited to the discovery of TB susceptibility genetic variants and their probable ethnic origins. Genome-wide admixture mapping was used in the third study to identify regions of the genome that may harbour such variants. The study identified a number of novel candidate TB susceptibility genes, and provided further substantiating evidence for the role of genetic loci previously implicated in the disease. The fourth study investigated the role of gene-gene interactions (epistasis), an oft-cited explanation for the missing heritability of complex disease, in the South African Coloured TB case-control cohort. A number of interesting gene-gene pairs that may jointly modify the odds of having TB were identified, and some of these findings were validated in an independent TB case-control cohort from The Gambia.

Opsomming

Afrikaanse Titel

("Die gebruik van bioinformatika en biostatistiek om genetiese vatbaarheid vir tuberkulose te ondersoek in 'n bevolking met vermengde herkoms")

M. Daya

Departement Biomediese Wetenskappe,
Universiteit van Stellenbosch,
Posbus 19063, Francie van Zijl Weg, Tygerberg 7505, Suid Afrika.

Proefskrif: Ph.D. (MedSc)

Februarie 2015

Tuberkulose (TB) is wêreldwyd die tweede grootste oorsaak van sterftes weens aansteeklike siektes . Een van die faktore wat betrokke is by die ontwikkeling van die siekte is die genetika van die gasheer, maar die studie van genetiese vatbaarheid vir TB het nog nie voldoende antwoorde opgelewer nie. Die Suid-Afrikaanse Kleurlinge is die grootste bevolkingsgroep in stedelike gebiede van Kaapstad met van die hoogste voorkoms van TB wêreldwyd. Die DST/NRF Sentrum vir Biomediese TB-navorsing aan die Universiteit van Stellenbosch het oor baie jare 'n groep individue uit hierdie gemeenskappe gewerf en bestudeer, met die doel om nuwe TB-vatbaarheid-gene te ontdek om die buitengewone voorkoms van TB in die gemeenskap te verklaar. Die groot genotipiese datastelle wat deur die sentrum ingesamel is, is belangrike hulpbronne om die bevolkingsgroep se genetiese samestelling beter te verstaan, en om nuwe hipoteses te skep rakende genetiese faktore wat oorskakeling na siekte kan bewerkstellig. Nuwe toepassings van bestaande statistiese metodes is gebruik om hierdie doelwitte teweeg te bring, en word as vier studies in die verhandeling aangebied.

'n Belangrike oorweging in studies van die genetiese verband met siekte in die Suid-Afrikaanse Kleurling bevolkingsgroep, is die vermenging van herkoms, wat resultate van genetiese studies kan beïnvloed. Dit is die onderwerp van die eerste twee studies. Die eerste studie beskryf die ontwikkeling van 'n paneel van genetiese merkers, wat die herkoms van hierdie bevolkingsgroep beskryf. Die paneel kan gebruik word as 'n koste-effektiewe alternatief tot genoom-wye data om statistiese modelle aan te pas vir die teenwoordigheid van verskillende bevolkingsgroepe in mense se herkoms. Die paneel word in die tweede studie gebruik om die belangrikheid van sodanige aanpassings te demonstreer. 'n Vorige studie het verbande tussen die kleurlinge se herkoms en vatbaarheid vir TB geïdentifiseer, maar 'n beperkte aantal kontrole-pasiënte is in daardie studie gebruik. Herkoms insiggewende merkers is gebruik om die vorige bevindings te ondersoek, en die aansienlike invloed wat herkoms het in die ontwikkeling van TB was bevestig.

Nuwe hipoteses rakende genetiese faktore wat vatbaarheid vir TB beïnvloed, word in die derde en vierde studies geskep. Die Suid-Afrikaanse Kleurling bevolkingsgroep het genetiese

bydraes van verskeie bevolkings ontvang, wat mag verskil in hul genetiese vatbaarheid vir TB. Die groep kan dus gebruik word om nuwe TB vatbaarheid genetiese variante en hul waarskynlike etniese oorsprong te ontdek. Kartering van genoom-wye herkoms-vermenging is in die derde studie gebruik om areas van die genoom te identifiseer wat moontlik sulke variante huisves. Die studie het 'n aantal nuwe kandidaat TB vatbaarheid gene geïdentifiseer, en ook areas van die genoom verneem as betrokke by TB bevestig. Die vierde studie ondersoek die rol van geen-geen interaksies, 'n algemene verduideliking vir die onverklaarde oorerflikheid van komplekse siektes. 'n Aantal interessante geen-geen pare is geïdentifiseer, en 'n paar van hierdie bevindinge is bevestig in 'n groep van Die Gambië.

Acknowledgements

I would like to express my sincere gratitude to the following people and organisations.

DAAD, the NRF and the DST/NRF centre for providing the funding and resources that was necessary for me to do this work.

My supervisors, Eileen and Lize: Eileen, thanks for giving me this opportunity, for your wisdom, energy, good advice, support, and the way that you have of inspiring people (not just me). Lize, baie dankie vir jou eerlikheid, onderskraging, ondersteuning, aanmoediging, aandag aan detail, en baie tyd wat jy aan my spandeer het. Ek het baie by jou geleer, en hoop ek leer nog baie by jou in die toekoms. Eendag as ek groot is, wil ek soos jy wees.

Marlo, baie dankie dat jy altyd gereed is om te help, en al die handige gesprekke, oor werk en andersins! Ek is bly jy het my oorreed om aan te gaan met 'n PhD.

To all the people who shared office space with me, thanks for the friendly atmosphere, many laughs and trips for essential beverages.

To Ronel and Branden, thanks for your encouragement and good advice! Your support meant a lot to me. Also to the Daya family, mummy, 2N (&B), 2S (&A) for your unfailing love and good company.

Lezanne, baie dankie vir ons koffies, en jou baie gebede.

Aan Hom waarin ek lewe en beweeg en is, baie, baie dankie.

Dedications

This thesis is dedicated to Marc, without whose love and support I would not have been able to do this, and to my parents, who have always encouraged me to do what I love.

Contents

List of Figures

List of Tables

List of Abbreviations

(In alphabetical order)

AIM	ancestry informative marker
ANOVA	analysis of variance
ATG4C gene	autophagy related 4C, cysteine peptidase gene
B7-H5 gene	B7 homolog 5 gene
BCG	Bacille de Calmette et Guérin
CADM2 gene	cell adhesion molecule 2 gene
CADM3 gene	cell adhesion molecule 3 gene
CD4+	cluster of differentiation 4 plus
CHST11 gene	carbohydrate (chondroitin 4) sulfotransferase 11 gene
CHSY3 gene	chondroitin sulfate synthase 3 gene
CSPG gene	chondroitin sulfate proteoglycan gene
CTSZ gene	cathepsin Z gene
DNA	deoxyribonucleic acid
GADD45A gene	growth arrest and DNA-damage-inducible, alpha gene
GRIK1 gene	glutamate receptor 1 gene
GRIK2 gene	glutamate receptor 2 gene
HapMap	haplotype map
HGDP	human genome diversity project
HIV	human immunodeficiency virus
HWE	Hardy-Weinberg equilibrium
IL12RB1 gene	interleukin 12 receptor, beta 1 gene
IL23R gene	interleukin-23 receptor gene
ISG15 gene	ISG15 ubiquitin-like modifier gene
LAI	local ancestry inference
LD	linkage disequilibrium
LSBL	locus specific branch length
MAF	minor allele frequency
mtDNA	mitochondrial DNA
M. tuberculosis	*Mycobacterium tuberculosis*
NELL1 gene	neural epidermal growth factor-like 1 gene
NF-κB gene	nuclear factor kappa-light-chain-enhancer of activated B cells gene
NLRC5 gene	NLR family, CARD domain containing 5 gene
NOD2 gene	nucleotide-binding oligomerization domain containing 2 gene
NRG1 gene	neuregulin 1 gene
NRG2 gene	neuregulin 2 gene
OSM gene	oncostatin M gene
PCR	polymerase chain reaction
SAC	South African Coloured
SES	socioeconomic status
SFTPD gene	surfactant protein D gene
SNP	single nucleotide polymorphism
SP-D	surfactant protein D
TB	tuberculosis
Th17	T helper 17
TLR8 gene	toll-like receptor 8 gene
TRPV1 gene	transient receptor potential cation channel subfamily V member 1 gene
USA	United States of America
WHO	World Health Organization
WTCCC	Welcome Trust Case Control Consortium

Chapter 1

Introduction

1.1 The differences in our genes

Human molecular genetic variation was first described in 1919 by Hirzfeld and Hirzfeld with the discovery of the ABO gene that encodes the ABO blood groups (sugar antigens). Subsequent to the ABO blood groups, various blood group protein antigens were identified, which facilitated the further analysis of genetic variation. With the discovery of the structure of Deoxyribonucleic acid (DNA) by Watson and Crick in 1953, the development of polymerase chain reaction (PCR) in the 1980's and automated DNA sequencing in the 1990's, it became possible to study genetic variation of humans across the genome [1].

Analysis of variance (ANOVA) is a statistical technique that is used to divide the total variation of a set of observations into components that can be ascribed to different sources of variation. If for example observations are classified into different groups, the variation can be attributed to differences within groups and differences between groups. Using this idea, a number of population genetic studies have found that the largest proportion of human genetic variation can be attributed to variation within world-wide human populations. Only 5-15% can be attributed to variation between populations [2; 3; 4; 5; 6; 1]. Consequently, if a locus is selected at random, two individuals from different populations may appear to be more similar genetically than if they were from the same population [7]. However, using the aggregate properties of loci that vary between populations, statistical techniques such as principal component analysis can be used to cluster individuals into population groups. These groups often correspond to self reported ancestry, geography and language [8; 9; 1; 10; 11; 12].

What causes genetic variation between individuals and population groups? The root of genetic variation is DNA mutations that are passed on to offspring. If a mutation is found in multiple individuals, it is referred to as a genetic polymorphism. Genetic drift and natural selection may result in polymorphisms occurring in populations at different frequencies. Genetic drift occurs when a finite or small number of individuals form a new population. The founders of the new population will lack some of the mutations present in the parent population, and possess others at different frequencies, which will result in the new population having mutation (allele) frequencies that are significantly different from those in the parent population. This effect is particularly strong if the founding population is small, as the available pool of mutations or alleles is unlikely to be a good representation of the original population. Different populations are also exposed to varying environmental threats. Natural selection occurs when genetic polymorphisms affect survival, such as deleterious variants, those polymorphisms that confer protection against pathogens found in a particular environment, or variants that change reproduction prospects. Polymorphisms that increase the probability of survival to reproductive age will become more frequent in a population, whereas those that decrease the probability

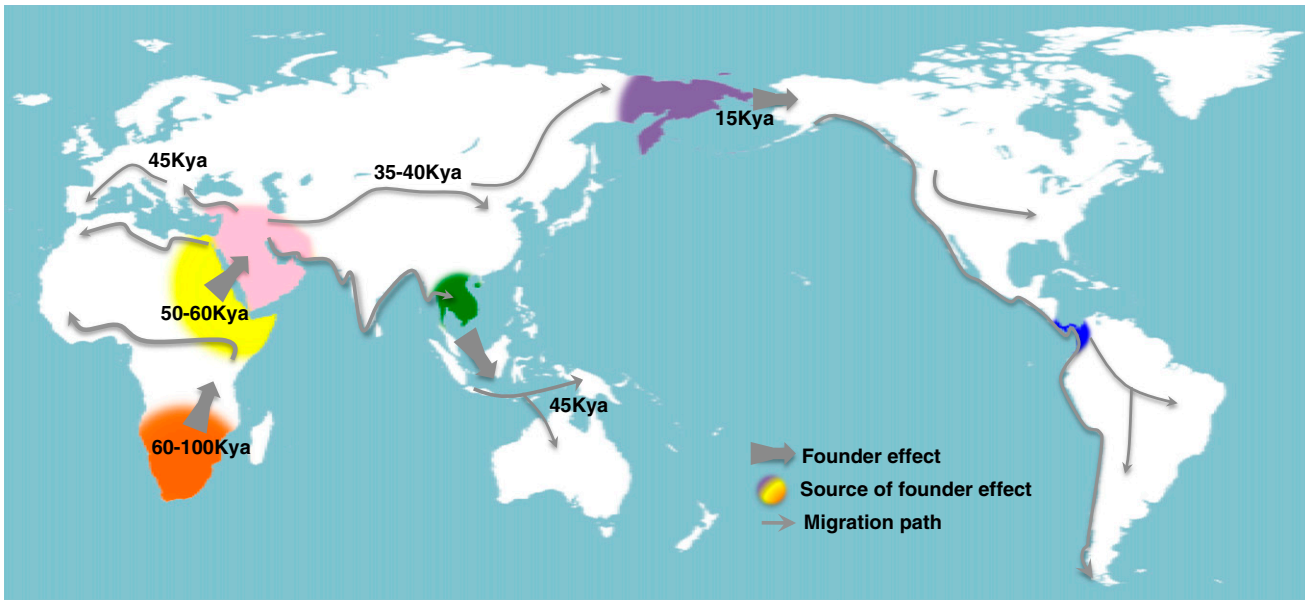


Figure 1.1: Migration routes of modern humans [20]

of survival will become less frequent.

The analysis of genetic data suggests that African populations are the oldest in the world. The oldest mitochondrial DNA (mtDNA) and Y-chromosome lineages are found in Africa [13; 14; 15; 16]. African populations also show the lowest levels of linkage disequilibrium (LD) and highest levels of genetic and phenotypic diversity compared to other continental populations [15; 8; 1; 17; 16]. Lower levels of LD are indicative of older population age as LD decays over time. Younger populations have reduced genetic diversity as gene flow is reduced when a new population is formed from a small subset of the original population. A strong correlation has been found between loss of genetic diversity and distance from Africa [9; 15; 18], suggesting that modern humans originated in Africa. Originally eastern Africa was proposed as the region of origin for modern humans[1], but a more recent hypothesis, substantiated by genetic data, proposes a southern African root[15; 12]. From southern or eastern Africa, humans migrated and expanded throughout the world, as illustrated in figure 1.1. This "out of Africa" hypothesis is supported by multiple lines of evidence. Archaeological and fossil data is generally thought to show an eastern African origin of modern humans [1]. Alternatively, inscribed ostrich egg shells and worked bone awls that were found in southern Africa as well as climactic evidence indicates that southern Africa might have been more hospitable than eastern Africa at the time of origin [15]. Concordance between genetic and language trees [19; 20; 21; 1] and the loss of phonemic diversity with decreased distance from Africa [22; 20] further supports conclusions drawn from genetic data. Finally, a malarial parasite and human gut bacterium show a remarkably similar pattern of worldwide DNA variation compared to humans [20], suggesting the same out of Africa migration as their hosts.

In the above discussion we established the foundation of genetic variation between humans. Note that although the aggregate properties of genetic variation can be used to cluster individuals into population groups, human genetic variation is continuous in nature, which implies that the concept of 'race' is a misnomer [3; 18]. The discussion will now focus on new populations that are formed from previously separated population groups.

1.2 An interesting combination

In population genetics, the term "admixture" refers to the phenomenon where two or more previously isolated population groups produce offspring. Population geneticists are often interested in quantifying admixture that occurred in the distant past, such as European and East Asian admixture in the western Chinese Uyghurs [23], and admixture that occurred between the Khoe-San and Bantu speaking populations during the East African Bantu migration into southern Africa [9]. Ancient admixture between modern humans and Neanderthals as well as Denisovans has also been reported [24; 25]. Recent admixture, occurring in the last few centuries between source populations with a reasonable genetic distance between them, is more relevant in medical research. Proportions of ancestry received from source populations typically vary largely between individuals from the same recently admixed population [26; 27; 28]. Examples of such admixed populations include African Americans, African Caribbeans, Latin Americans such as Mexicans, Cubans and Puerto Ricans, Anglo Indians and mixed ancestry populations in southern Africa [27].

Admixture mapping is a novel technique that is used to map regions of the genome to disease outcome. The technique is applicable when the source populations of a recently admixed population differ in their genetic susceptibility to a disease [29; 30]. Disease causing variants in the at-risk source population will be harboured in certain regions of the genome, and admixture mapping is based on the premise that such regions will be inherited more frequently by affected admixed individuals. The goal of admixture mapping is to identify these regions. It has been used for disease gene discovery of hypertension, multiple sclerosis, asthma, prostate cancer and kidney disease in African Americans [29; 31; 28; 32; 33; 34] and asthma and breast cancer in Latin Americans [35; 36]. Prior to the advent of genome-wide genotyping micro-array technologies, linkage analysis was used to identify chromosomal segments shared by affected individuals in a family, thus investigating disease susceptibility at a genome-wide level. Admixture mapping was thought to have more statistical power than linkage analysis for identifying disease-harbouring genomic regions [29; 30; 12]. Admixture mapping also provided greater resolution by identifying shorter genomic regions, thereby narrowing the search for causal variants. Initially, ancestry informative markers (AIMs) - those markers with different allele frequencies in the source populations of the admixed population - were used for admixture mapping. Genome-wide data can now be genotyped cost-effectively and has become the preferred alternative in admixture mapping studies, as the identification of genomic regions with more shared ancestry than the norm would not be limited to the genomic positions of a particular AIM set [28; 29; 30].

Recent admixture can also be a source of confounding in genetic association studies. In case-control genetic studies, if cases have a different proportion of ancestry from a source population compared to controls, associations found may be related to ancestry rather than disease, or the effect of real associations may be masked due to differences in ancestry [37; 38]. A number of statistical techniques can be used to solve this problem, if a large set of random markers, AIMs or genome-wide data is available. These techniques include genomic control and the use of principal components to estimate ancestry proportions, based on genetic data. These are then used as covariates in statistical models, to adjust for ancestry.

The predominant population group in the Western Cape, South Africa, is the admixed group known as the South African Coloured (SAC). Historical records and genetic data tell a consistent story about the heritage of this unique population. A number of studies showed that they received genetic contributions from click-speaking Africans, Bantu-speaking Africans, European, South and East Asians [26; 39; 9; 40; 41]. From historical records, we know that the SAC have their roots in the click-speaking Khoe-San who were indigenous to the Western

Cape at the time that the Dutch East India Company established its refreshment station in 1652 [26; 42; 43; 44; 45]. Slaves were brought in from the Indian subcontinent, east coast of Africa, Madagascar and Indonesia [26; 46; 45], as well as some political exiles from Indonesia and Malaysia [44; 26]. Men outnumbered women in the early Cape society and mixed liaisons between European settlers and KhoeKhoe women, and between the KhoeKhoe and slaves, were common [26; 45; 44; 47; 46]. The establishment of mission stations from the mid-1700s onwards further facilitated the integration of European, African (particularly Xhosa) and Khoe-San ancestries [26; 47; 44]. The "free black" Chinese who formed 9% of the Cape Town population in the early 1800s may also have contributed to the genetic material of the SAC, albeit in small proportions [26; 45; 47; 44; 39].

The SAC is the largest demographic in the metropolitan areas of Cape Town that have some of the highest reported incidences of tuberculosis (TB) worldwide, despite extensive Bacille de Calmette et Guérin (BCG) vaccination and low prevalence of the human immunodeficiency virus (HIV) [48]. As the group received contributions from diverse source populations that may differ in their genetic susceptibility to TB, the group is ideally suited to the discovery of TB susceptibility genetic variants and their probable ethnic origins. The next section explores the history and pathogenesis of TB, and how the ancestry and genetics of the human host may influence susceptibility to the disease.

1.3 An ancient scourge

Tuberculosis (TB) is caused by the bacillus *Mycobacterium tuberculosis* (*M. tuberculosis*) and is the second leading cause of mortality from infectious disease worldwide [49]. In 2012, 8.6 million new infections and 1.2 million deaths were reported [49] and in South Africa, TB is the fourth leading cause of mortality [50]. High income countries (most western European countries, Canada, USA, Japan, Australia and New Zealand) have an incidence of fewer than 10 per 100 000 (figure 1.2), whereas the incidence is a 1000 per 100 000 in South Africa and Swaziland, the highest incidence rate in the world [49].

M. tuberculosis and its human host have a long shared history. Evidence of Pott's disease, a presentation of extra pulmonary TB, has been found in Egyptian and South American mummies, and *M. tuberculosis* DNA has also been recovered from mummified tissues [51; 52; 53; 54; 55; 56]. Until recently it was believed that *M. tuberculosis* developed from bovine TB with the onset of animal domestication [57], about 10 000 years ago. Comparative genomics have however shown that *M. tuberculosis* is older than its animal counterparts [58; 59; 60; 61]. It has been postulated that *M. tuberculosis* originated in Africa, and that some lineages accompanied their human hosts during the out of Africa migration [59; 60]. Three of the modern *M. tuberculosis* lineages are thought to have evolved separately in Europe, India and China [59]. Co-evolution between modern lineages of the pathogen and its human hosts therefore seems likely. Modern lineages that evolved in Eurasia then spread throughout the world and back into Africa during the colonisation period [59; 62; 60].

Pulmonary TB is the most common form of TB and affects the lungs. Infectious *M. tuberculosis* bacilli can remain airborne for several hours as droplet nuclei, and when the nuclei are inhaled into the lungs, the bacilli are engulfed by alveolar macrophages. Infected macrophages collect to form granulomas which can contain the bacilli for many years, even indefinitely (latent infection). Loss of vascularisation and increased caseous necrosis of granulomas mark progression towards TB, until cavities are formed and granulomas collapse into the lungs (active infection). Infectious bacilli are then released into the airways again through coughing [63].

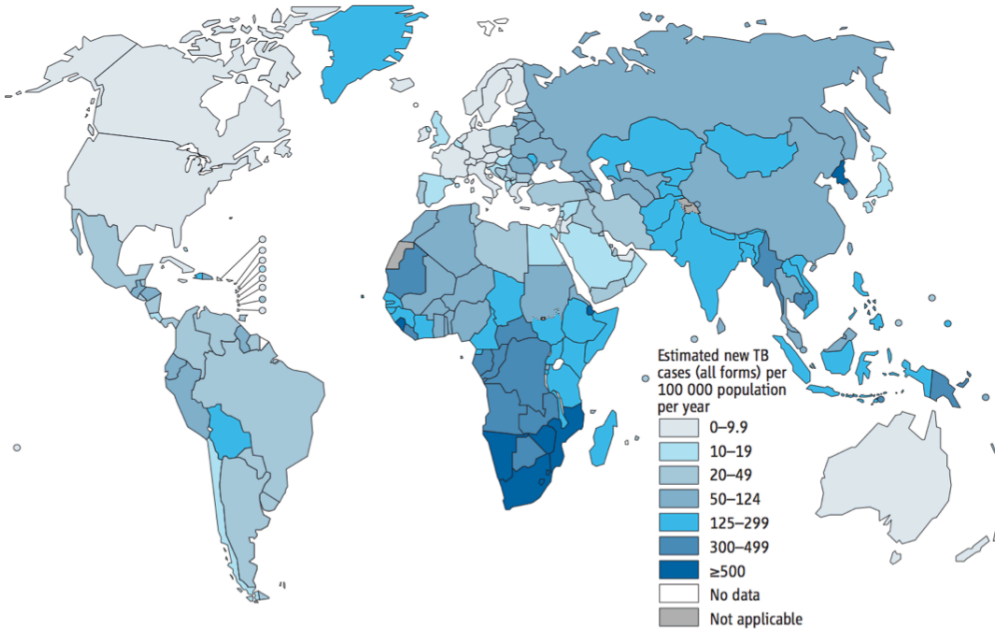


Figure 1.2: Estimated TB incidence rates worldwide [49]

Progression to active TB relies on three factors: the virulence of the bacilli, the environment and the host immune response. Environmental factors include age, sex, nutrition, smoking, alcohol use, socio-economic conditions and overcrowding [64; 49]. HIV infection severely compromises the immune response to TB, and explains a large proportion of the high TB incidence in South Africa [49]. Host genetics are also thought to play an important role in progression to disease, and may explain why only 5-10% of immunocompetent individuals progress to active TB [65; 66]. Since the causative agents of infectious disease are external, it is easy to underestimate the contribution of host genetics to infectious disease, but a large number of polymorphisms in the human genome are probably those affecting the immune response as a result of so-called selection [67; 68]. In support of the importance of immune response genes, a Danish adoption study concluded that the heritability of genetic susceptibility to infectious disease may be larger than that of cancer and cardiovascular disease [69]. Before the discovery of the bacterium, it was thought that TB was heritable, as the disease occurred more frequently in some families. A tragic incident in 1926 demonstrated that the bacterium is not solely responsible for developing disease, and that host genetics may indeed be important in combatting this pathogen. A group of children was immunised mistakenly with a virulent strain of *M. tuberculosis*. All children received the same dosage, but had disparate outcomes: 77 died, 127 showed radiological evidence of infection, and 47 showed no signs of disease [70].

The link between genetics and susceptibility to TB has been further established by adoption and twin studies, studies of the heritability of anti-mycobacterial phenotypes as well as animal models [71; 72; 73; 74; 75; 76; 77; 78; 79; 80; 81; 69; 82; 83; 84; 85]. The disparate prevalence of TB between some population groups has also been thought to support this argument [86]. Based on a large study of 165 racially integrated nursing homes in Arkansas (U.S.A), Stead et al. showed that Europeans are less susceptible to TB infection compared to individuals of African ancestry [87]. This finding was corroborated by Nahid et al., who used a cohort from coronary artery risk study to demonstrate that African Americans had twice the risk of TB infection compared to Caucasian Americans, after adjusting for socio-economic and

other factors [88]. Contrary to these findings, in a primary school outbreak of TB, Hoge et al. showed that Caucasian and African American students were equally likely to be infected [89]. All these studies however primarily measured the risk of TB infection, rather than active disease. Interestingly, African American students in the primary school study were more likely to have abnormal chest radiographs, and also had larger Mantoux tuberculin skin test (TST) reactions.

The differential TB infection rate and TST reactions observed between Caucasians and African American individuals indicate that the immune response to TB may differ between ethnicities. These differences in immune response may explain why BCG vaccination is effective in some populations, but not in others [63]. In vitro experiments demonstrated that although phagocytosis of bacilli appear to be more effective in African American compared to Caucasians, the growth rate of bacilli in macrophages is higher in African Americans [90]. In 2013, Coussens showed that variation in the inflammatory immune response between individuals may reflect ethnic genetic differences, which may be a result of different selective pressures on populations that migrated out of Africa versus those that remained [91].

The apparent higher resistance of Europeans to TB could possibly be explained by many centuries of exposure to the disease in densely populated European settlements [87], affording natural selection the opportunity to reduce the frequency of susceptibility genetic variants in subsequent generations [86]. In the last few centuries, up to a half of mortalities in Europe and in North America were caused by TB [92; 59]. TB resistance in Ashkenazi Jews has also been linked to the high rates of TB mortality in European ghettos [93; 86]. High mortality in populations that are newly exposed to the pathogen provides further substantiation of the theory of selective pressures on TB resistance. When the Qu'Appelle Indians were first exposed to TB, the initial mortality rate was as high as 10% [94]. Historical records from the 19th and early 20th century also indicate that high rates of TB were observed in African populations who had contact with Europeans, but not in those populations with limited contact [86]. A high prevalence of active TB in the Yanomami Indians was reported relatively recently (6.4% in a 1997 field study) [95]. First contact between this population and Europeans occurred in the 1960s, and the researchers found no evidence of confounding factors such as HIV, malnutrition or alcoholism in the community. This study also reported large dissimilarities between the TB immune response of the Yanomami and their European contemporaries.

Even though the genetics of the human host plays an important role in the pathogenesis of TB, much of the modern TB epidemic can be ascribed to socio-economic conditions and the HIV pandemic. Disentangling the effect of the environment and HIV from ethnicity is not straightforward; when measuring the effect that ethnicity has on TB susceptibility, rates of infection between different population groups cannot simply be compared to one another [96; 97; 88]. An admixed population such as the SAC is however ideally suited to investigate the link between ancestry and TB susceptibility. As the SAC received genetic contributions from various source populations that may well differ in their risk of developing disease, cases and controls recruited from communities with uniform socio-economic conditions and low HIV prevalence represent a unique opportunity to investigate the link between ancestry and TB susceptibility. To answer this question, genetic data from such a study group can be used to infer contributions of ancestry from source populations for each individual, and these estimates can be compared between cases and controls.

1.4 What does the data say?

TB is a complex disease and it is likely that many genetic variants exert small to moderate effects on disease outcome. To better understand the relationship between host genetics and TB susceptibility, the DST/NRF Centre for Biomedical TB Research at Stellenbosch University recruited a large cohort of SAC individuals from an epidemiological field setting in high TB burden metropolitan areas of Cape Town, South Africa. Samples were collected between 1994 and 2007 and comprise 955 TB cases and 521 control samples. The sample bank was used to perform a number of candidate gene association studies between 2003 and 2013. With a view to performing a case-only admixture mapping study, 959 samples were also genotyped in 2008 on the Affymetrix GeneChip Human Mapping 500K Array Set (864 cases and 95 controls). The same micro-array platform was used in a Welcome Trust Case Control Consortium (WTCCC) study of TB susceptibility in The Gambia (1498 cases and 1496 controls). Permission was obtained from the consortium to use the Gambian data set for one of the research topics presented in this dissertation (Research Article 3).

Central to the design of genome-wide micro-array platforms is the use of so-called haplotype tagging single nucleotide polymorphisms (SNPs). Proximal genetic material is inherited together from one parent during meiosis, resulting in correlation between markers at a population level. This phenomenon is known as linkage disequilibrium (LD), and segments of genetic material that are inherited together are referred to as haplotypes. This correlation between markers is leveraged to reduce the number of variants that require genotyping; variants are selected based on their ability to "tag" other variants in proximal genetic regions. Tests for association between such tag variants and a phenotype of interest therefore seldom reveal causal associations, and fine-mapping of the genetic region surrounding associated tag variants is required for discovery of the causal variant(s). Many of the variants genotyped in the TB candidate gene association studies described above were in fact selected as tag variants, although some were also selected based on their putative functional effects.

The data sets described above are invaluable resources to aid understanding of the population genetics of the SAC and genetic susceptibility to TB, and can be used to generate new hypotheses regarding genetic variants involved in the immune response of the human host. Better understanding of the genetic underpinnings and causal mechanisms that underlie the immune response to TB may well be crucial in the development of new vaccines and drug interventions to combat the disease. Bioinformatics can be defined as the use of information science and/or technology to understand biological data (http://www.bioinformatics.org/wiki/Bioinformatics). This body of work applies various statistical techniques and software tools to the SAC and Gambian data sets with the view of achieving these goals. The relationship between ancestry and TB susceptibility and the confounding effect that ancestry may have in genetic association studies of the SAC is explored. New hypotheses regarding genes that may be involved the immune response to TB is generated by means of admixture mapping and the identification of genes that may individually or jointly modify the odds of having TB.

1.5 Avenues of investigation

The data sets described in the previous section were used for the following four areas of research:

1. Statistical models can be used to test genetic association with TB, while adjusting for differences between case and control ancestry, but genome-wide data or ancestry informative markers (AIMs) are required in order to adjust for ancestry. The large SAC genome-wide data set provided a unique opportunity for the development of a panel of AIMs that are

tailored to the complex five-way admixture that occurred in the SAC. Chapter 3 (Research Article 1) describes the development of such a panel of AIMs. The panel can be used as a cost effective alternative to genome-wide data for reducing false positive findings resulting from ignoring admixture in genetic association studies of the population. Principal components derived from additive allelic coding of the genotypes in the AIMs panel can be added to models as covariates in order to adjust for the confounding effect of ancestry.

2. In a previous genome-wide TB case-control study of the SAC, Chimusa et al. found a positive correlation between African San ancestry and TB susceptibility, and negative correlations with European and Asian ancestries [98]. Since genome-wide data was available for only a small number of controls in the Chimusa et al. study, chapter 5 (Research Article 2) endeavours to validate this finding by genotyping the panel of AIMs identified in chapter 3 in additional individuals. The effect of adjusting for ancestry in candidate gene TB association studies of the SAC is also investigated.

3. One of the commonly posited explanations for the missing heritability of complex disease is gene-gene interactions, also referred to as epistasis. Chapter 6 (Research Article 3) investigates the role of gene-gene interactions in genetic susceptibility to TB, using the SAC data sets for discovery, and the Gambian data set for validation.

4. The SAC is ideally suited to the discovery of tuberculosis susceptibility genetic variants and their probable ethnic origins, but previous attempts at finding such variants using genome-wide admixture mapping were hampered by the inaccuracy of inferring ancestry along the genomes of individuals. Chapter 7 (Research Article 4) uses a novel algorithm implemented in the software package RFMix to address this problem, and uses the inferred ancestry to identify regions that may harbour TB susceptibility genes.

Before commencement of the above research topics, a number of data exploration and quality control steps were first applied to the micro-array data sets, to ensure that the data sets are of reasonable quality. These steps are described in chapter 2. Chapter 3 (Research Article 1) describes the development of a panel of AIMs for the SAC [99]. The panel was genotyped in a large number of samples from the SAC sample bank, and the quality control steps that were applied to the resultant genotypic data are described in chapter 4 [100]. After cleaning of the AIMs genotypic data, the AIMs were used to investigate the association between TB susceptibility and ancestry in chapter 5 (Research Article 2), and to investigate the role of gene-gene interactions in chapter 6 (Research Article 3). The admixture mapping study is documented in chapter 7 (Research Article 4). The dissertation concludes with a discussion in chapter 8.

Chapter 2

Data exploration and quality control

Low quality microarray data may result in the estimation of spurious results, therefore it is important to control their quality by filtering out possibly inaccurate samples and SNPs. This needs to be balanced against removing data unnecessarily, which would result in loss of statistical power. This chapter describes the quality control steps and rationale that were applied to the SAC and Gambian genome-wide data sets, using a combination of recommendations suggested by Laurie et al., the Welcome Trust Case Control Consortium (WTCCC), Miyagawa et al. and Ziegler et al [101; 102; 103; 104].

Checks were done in the order that they are described, and if a sample or SNP was removed, it was no longer present in the data set used for subsequent checks.

Software

The software package PLINK was used for filtering and calculation of test statistics [105] . The freely available R programming environment was used to summarize and report these statistics [106].

2.1 Quality control of the SAC data set

Some of the steps described below involved the checking of records in the DST/NRF Centre for Biomedical TB Research's TB case-control database, which is referred to simply as "the database".

2.1.1 Genotype calling

The SNP calling from the Affymetrix genotyping probe intensity (CEL) files was done in a previous study [26]. This resulted in some samples and SNPs being discarded. The accuracy of the process was determined by comparing the called genotypes of 9 HapMap samples with their known genotypes; the accuracy was greater than 99%. In addition, there were 4 duplicate SAC samples; concordance between the duplicate samples was greater than 97%.

After this process, 397 337 SNPs were retained for 959 individuals (864 cases and 95 controls).

2.1.2 Ambiguous Sex

As recommended by Laurie et al. and the Genetic Association Information Network (GAIN), the recorded sex of individuals was compared to the sex calculated by the calling algorithm,

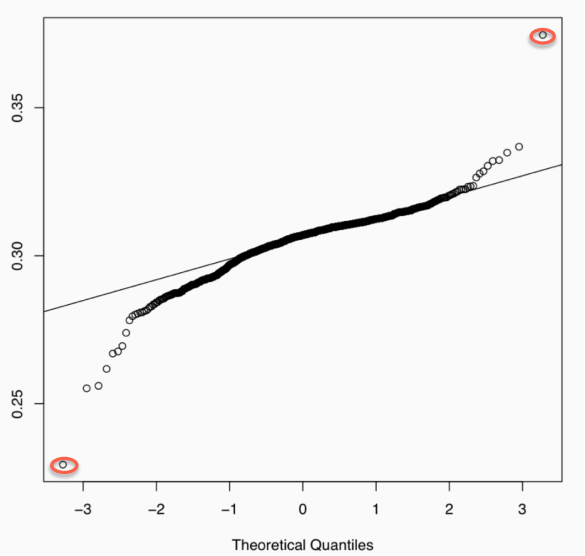


Figure 2.1: The distribution of the proportion of heterozygous genotypes per individual in the SAC cohort. A normal QQ plot of the proportion of heterozygous genotypes per individual is shown in this figure. Theoretical quantiles are shown on the x-axis, and the observed proportions of heterozygous genotypes are shown on the y-axis. The two outliers that were identified are circled in red.

based on genotypic probe intensity data [101; 107]. Twelve mismatches were found. For two of these individuals, the gender recorded in the database appeared to be wrong, based on the first names of the individuals, and the database records were corrected. The remaining 10 individuals were removed (9 cases and 1 control; due to uncertainty regarding the quality of the recorded data, the case/control status of these individuals was deemed questionable). Sex was not recorded for 69 individuals in the database, so the sex calculated by the calling algorithm was assigned to these individuals.

2.1.3 Heterozygosity Deviation

The WTCCC excluded individuals with less than 23% or more than 30% heterozygous genotypes, as this may be indicative of erroneous genotyping [102]. This may be appropriate for European populations, but not for the SAC. The mean percentage of heterozygous genotypes per SNP per individual in the SAC cohort was 30.5, with a standard deviation of 0.97%, and 291 individuals (30.4% of the cohort) had more than 31% heterozygotes genotypes. As a result of the diversity of the source populations from which the SAC received their genetic contributions, the relatively high levels of heterozygosity found in the cohort is not surprising. Deviation from average heterozygosity across all SNPs may indicate DNA contamination [104]. Ziegler et al. recommend removing individuals with a heterozygosity that exceeds 3 standard deviations from the mean over all SNPs for all individuals [104].

The Q-Q plot in figure 2.1 shows quantiles of the proportion of heterozygous genotypes per individual compared to standard normal quantiles. Although eight individuals had proportions less than 3 standard deviations from the mean, and three individuals had proportions that exceeded 3 standard deviations from the mean, and all of them clearly deviated from the standard normal distribution, only the two clear outliers were removed from the data set.

2.1.4 Frequency Filters

A high rate of missing SNPs may be indicative of genotyping quality problems, since failure to call a SNP is often non-random [101]. It is therefore necessary to filter samples with a high rate of missing SNPs, as well as SNPs with a high missing rate.

Laurie et al. recommends removing samples with a missing rate greater than 2% [101]. The WTCCC and Ziegler et al. recommends removing samples with a missing rate greater than 3% [102; 104]. (Ziegler et.al. also state that this number is sometimes lowered to 10%.) In our data set, the thresholds would result in the removal of 173 samples (using a 2% threshold) or 99 samples (using a 3% threshold). The threshold was therefore set at 5%, resulting in the removal of only 40 cases and 1 control.

Large study group sizes are required to detect an association between a SNP with a low minor allele frequency (MAF) and disease outcome [101]. Historically SNPs with a low MAF are also more difficult to call; the minor allele would mostly be present in heterozygous genotypes as opposed to homozygotes, and it would therefore be more difficult to distinguish between homozygotes and heterozygotes. This depends on number of individuals (study group size): a variant with a frequency of 1% is for example more difficult to call in 100 samples than in 10 000 samples. For these reasons, SNPs with a low MAF are removed in most studies. A missing rate threshold from 1 to 5 percent and a MAF threshold of 1 to 5 percent are commonly recommended [102; 101; 103]. The least stringent thresholds were applied, in order to retain more SNPs. SNPs that had a missing rate greater than 5% (71 SNPs) and a minor allele frequency less than 1% (5861 SNPs) were removed.

Deviation from Hardy-Weinberg equilibrium (HWE) might be the result of genotyping errors. In a recently admixed population deviation from HWE could be expected, but since the SAC admixture occurred several generations ago, this is unlikely. Departure from HWE in controls was also an uncommon occurrence in SAC candidate gene TB association studies carried out by the DST/NRF Centre for Biomedical TB Research. Laurie et al. recommends removing SNPs for which the HWE test in controls have a p-value of 1×10^{-6} or less, the WTCCC recommends a p-value of 5.7×10^{-7} or less. Miygawa et al. and Ziegler et al. [104] recommends a more stringent critical p-value of 1×10^{-4} or less [101; 102; 103; 104]. This cutoff resulted in the removal of only 528 SNPs that violate HWE.

The steps described above resulted in a total of 6 450 SNPs being removed, leaving 390 887 SNPs.

2.1.5 Outliers and Cryptic Relatedness

Laurie et al. recommends removing individuals who may have a different ethnicity compared to the majority of the study cohort, based on analysis of their genotypic data [101]. In order to identify such outliers, individuals were clustered based on how many alleles are shared identical by state (IBS) between all pairs of individuals. For a particular SNP, a pair of individuals may either be IBS 0 (no alleles in common, so two different homozygotes), IBS 1 (1 allele in common, so one heterozygote and any homozygote) or IBS 2 (both alleles in common, either both heterozygotes or two of the same homozygotes). The IBS clustering algorithm calculates the ratio of the proportion of IBS 0 genotypes vs. the proportion of heterozygotes IBS 2 genotypes for each possible pairing of individuals. The proportion of pairings for which the calculated ratio is significantly different from the expected ratio of 1:2 is reported as an individual's proportion of so-called *significantly different others*. An individual with a high proportion of *different others* would be an outlier and possibly of different ethnicity. The algorithm assumes uncorrelated SNPs, and was therefore applied to an autosomal subset of

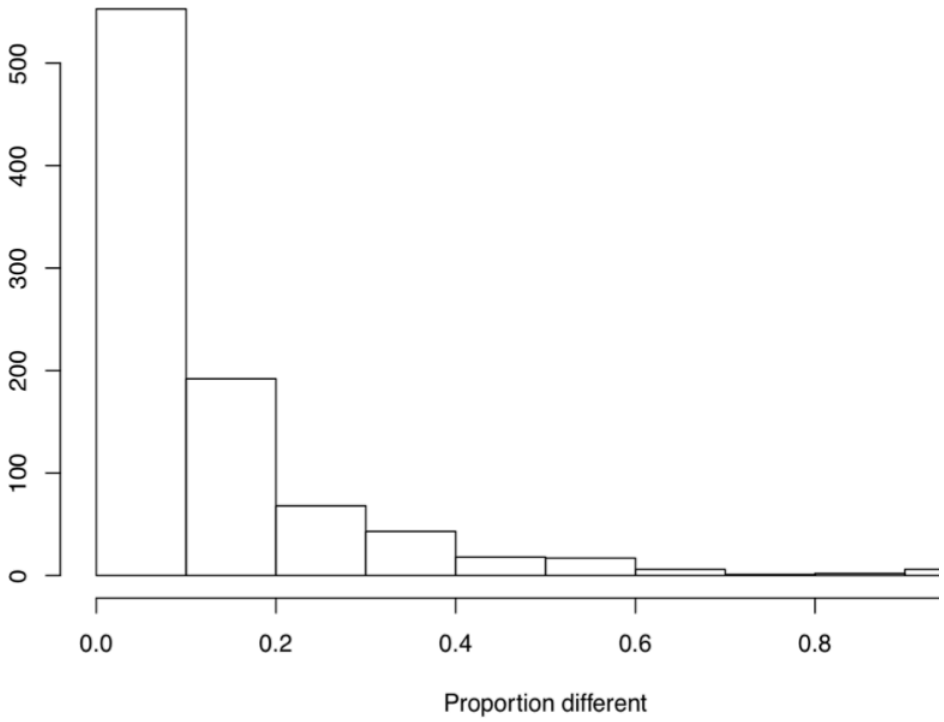


Figure 2.2: The distribution of the proportion of significantly different others in the SAC cohort. A histogram of the estimated proportion of significantly different others are depicted in this figure.

SNPs that were filtered according to linkage disequilibrium (LD) criteria as follows. Identify all pairs of SNPs with an LD r^2 value of more than 0.15, found in a window of 100 SNPs, and remove one SNP of each pair. Then slide the window 25 SNPs further and again remove SNPs so that no pair is in LD of $r^2 > 0.15$. Repeat the removal in a sliding window incremented by 25 SNPs in each iteration, yielding a set of uncorrelated SNPs. The distribution of the estimated proportion of significantly different others are shown in figure 2.2. Fifteen individuals exceeding a proportion of 0.6 were removed from the data set.

The number of IBS alleles between pairs of individuals were also used to estimate the coefficient of relatedness for each pair of individuals in the data set. The estimate is calculated as follows: $p_2 + \frac{1}{2}p_1$, where p_i denotes the proportion of alleles shared IBS $i, i = 1, 2$. Possible familial relationships identified in this manner are summarized in table 2.1. Three pairs had a coefficient of relatedness of 0.97 or higher - these matched to database records that had the same case identifier, but different sample identifiers. The duplicate individuals were removed from the data set. The number of full siblings/parent offspring pairs that were identified appears to be excessive (115 pairs), but this was also consistent with database records.

Two data sets were generated based on the above analysis. The first contained 888 individuals, including possible relative pairs, which can be used in analyses that do not assume a cohort of unrelated individuals. The second data set comprised 733 unrelated individuals (no first, second or third degree relatives).

2.1.6 Missing SNPs Associated With Disease

Association between the number of missing calls and phenotype may be caused by experimental differences between cases and controls, e.g. DNA extraction and date, genotyping batch, etc [101]. The unrelated data set was used to test for association between the number of missing

Table 2.1: Relative pairs in the SAC cohort. This table summarizes the number of relative pairs, identified based on alleles shared identical by state, per degree of relatedness.

Relationship	Degree of Relation	Coefficient Interval	Nr. Relationships
Monozygotic Twins	0	0.97 - 1.00	3
Full Siblings/Parent Offspring	1	0.43 - 0.59	115
Double First Cousins/Uncle Nephew/Half Siblings/Grandparent Grandchild	2	0.22 - 0.40	43
First Cousins	3	0.10 - 0.20	53
Total			214

calls and case-control status. The tests were done separately for each SNP. None of the association tests had p-values of 5.7×10^{-7} or less (the alpha level used in most genome-wide association studies), but 85 tests had p-values less than 1×10^{-4}. To determine whether this number is acceptable, tests for association between genotypes and case-control status were also done per SNP. The number of association tests with small p-values were much larger: 397 tests had p-values of 1×10^{-4} or less, and 14 reached genome-wide significance. A trend of association between missing calls and phenotype is therefore unlikely.

2.1.7 Sex Homozygosity

The software package PLINK allows for the assessment of sex using Wright's inbreeding coefficient as an estimate of homozygosity. Since ambiguous sex was already investigated using genotypic probe intensity data (section 2.1.2), which is considered to be superior to checking the homozygosity of the X-chromosome [101], this check was used to assess the quality of X chromosome data (rather than assessing the sex of individuals per se). Eleven females were more homozygous than expected (inbreeding coefficients larger than 0.2). To establish whether these individuals also had a greater proportion of autosomal homozygosity, which would explain their larger proportion of X chromosome homozygosity, inbreeding coefficients were estimated per individual using their autosomal SNPs. The mean autosomal inbreeding coefficient was calculated as 0.0041 with a standard deviation of 0.0203. The eleven females all had autosomal inbreeding coefficients that fell within the two standard deviation confidence interval. The X chromosome SNPs for these female individuals were therefore excluded from the data set.

2.2 Quality control of the Gambian data set

The quality control steps and rationale applied to the SAC data set was also used for quality control of to the WTCCC Gambian data set, but in a different order, as described.

2.2.1 Genotype calling

The WTCCC samples were genotyped on the 500k Affymetrix SNP chip, and genotype calling is described in the Supplementary Materials of Thye et al. [108].

Genotype calls for 2994 individuals (1498 cases and 1496 controls) and 500 568 SNPs, including the probability of the correctness of each SNP call, were obtained from the WTCCC. SNPs with a call probability less than 95% were set to missing. SNPs and samples with low average call probabilities will therefore have high rates of missing genotypes, and will be removed as part of the frequency filters described below.

2.2.2 Ambiguous Sex

Since genotypic probe intensity data was not available in the WTCCC data, the accuracy of the recorded sex of individuals could not be evaluated.

2.2.3 Frequency Filters

The same thresholds used in the SAC data set for removing low quality samples and SNPs were applied to the Gambian data set (maximum missing rate per sample of 5%, maximum missing rate in SNPs 5%, minimum MAF of 1%, minimum HWE p-value of 0.0001). After removing samples and SNPs that violated these criteria, 2 639 individuals (1337 cases and 1302 controls) and 405 348 SNPs remained in the data set.

2.2.4 Heterozygosity Deviation

In the SAC data, the heterozygosity check was done prior to sample and SNP filtering, as the genotype calling was already accurate. The Gambian data contained genotype calling inaccuracies that had to be filtered out before performing the heterozygosity check. The steps were therefore done in a different order, fixing the genotype calling first by filtering out SNPs and samples with large proportions of missing data.

The mean percentage of heterozygous genotypes in the Gambian data set is 30.19, with a standard deviation of 0.6%. There were 48 individuals with a proportion of heterozygous genotypes less than 3 standard deviations from the mean and 6 individuals with a proportion greater than 3 standard deviations from the mean. Only individuals with excess heterozygosity were removed, for the following reasons. The minimum percentage heterozygote genotypes observed in an individual was 26%, well above the minimum threshold of 23% that was used by the WTCCC [102]. An excess of heterozygosity, rather than lack of heterozygosity, is indicative of DNA contamination [104]. The original genome-wide association study that used this data set also removed the 6 individuals with excess heterozygosity, and only removed one individual with less than 23% heterozygous genotypes [108] (the latter individual was already filtered out by the preceding frequency filters, which were more stringent than the filters used in the original study).

2.2.5 Outliers and Cryptic Relatedness

IBS clustering was used to identify individuals with a large proportion of significant others, and as for the SAC data cleaning, a threshold of 0.6 was applied. This resulted in the removal of 126 individuals from the data set. The number of alleles shared IBS was also used to estimate the coefficient of relatedness per all pairs of individuals that remained in the data set. The type and number of relationships identified are summarized in table 2.2.

As in the case for the original genome-wide association study, a large number of duplicate samples and first degree relative pairs were identified. Duplicate samples were excluded, and two data sets were generated, one containing all individuals (1 206 cases and 1 222 controls), and one containing no first or second degree relatives (1 156 cases and 1 206 controls).

2.2.6 Missing SNPs Associated With Disease

To investigate potential batch effects, tests for association between the number of missing calls and case-control status were done as for the SAC data set. A disproportionate number of associations were found: 1205 of the tests had p-values of genome-wide significance, compared

Table 2.2: Relative pairs in the Gambian cohort. This table summarizes the number of relative pairs, identified based on alleles shared identical by state, per degree of relatedness.

Relationship	Degree of Relation	Coefficient Interval	Nr. Relationships
Monozygotic Twins	0	0.98 - 1.00	79
Full Siblings/Parent Offspring	1	0.41 - 0.57	42
Double First Cousins/Uncle	2	0.20 - 0.35	36
Nephew/Half Siblings/Grandparent Grandchild			
First Cousins	3	0.10 - 0.19	754
Total			911

to 123 genotype association tests that reached this threshold. This is suggestive of batch effects, i.e. differences in the lab procedures used to genotype cases and controls [101]. Unfortunately, no information is available that may explain these differences. The effect that the differences may have in association testing can however be mitigated by including principal components derived from the genetic data as covariates in statistical models [109].

2.2.7 Sex Homozygosity

The quality of X chromosome data were assessed using the same procedure applied to the SAC data set. Six females were identified as having excess homozygosity on their X chromosomes. Only one of these females had high autosomal homozygosity compared to the rest of the cohort. The X chromosome SNPs of the other five females were set to missing.

Chapter 3

Research Article 1

A panel of ancestry informative markers for the complex five-way admixed South African Coloured population

Michelle Daya[1], Lize van der Merwe[1,2,3], Ushma Galal[2], Marlo Möller[1], Muneeb Salie[1], Emile R. Chimusa[4], Joshua M. Galanter[5], Paul D. van Helden[1], Brenna M. Henn[6], Chris R. Gignoux[5], Eileen Hoal[1,*]

1 Molecular Biology and Human Genetics, MRC Centre for Molecular and Cellular Biology and the DST/NRF Centre of Excellence for Biomedical TB Research, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa
2 Biostatistics Unit, Medical Research Council, Tygerberg, South Africa
3 Statistics Department, University of Western Cape, Cape Town, South Africa
4 Computational Biology Group, Department of Clinical Laboratory Sciences, Institute of Infectious Disease and Molecular Medicine, University of Cape Town, Medical School, Cape Town, South Africa
5 University of California San Francisco, San Francisco, California, USA
6 Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York, USA
∗ E-mail: Corresponding egvh@sun.ac.za

The collective term for people of mixed ancestry in southern Africa is "Coloured" and is recognized and used officially in South Africa. Whilst we acknowledge that in some cultures this term may have acquired a derogatory connotation, this is certainly not intended here.

This article was published in December 2013 in the journal PLOS ONE [99]. Section, figure, table and reference numbering presented here is different to the published version.

3.1 Abstract

Admixture is a well known confounder in genetic association studies. If genome-wide data is not available, as would be the case for candidate gene studies, ancestry informative markers (AIMs) are required in order to adjust for admixture. The predominant population group in the Western Cape, South Africa, is the admixed group known as the South African Coloured (SAC). A small set of AIMs that is optimized to distinguish between the five source populations of this population (African San, African non-San, European, South Asian and East Asian) will enable researchers to cost-effectively reduce false-positive findings resulting from ignoring admixture in genetic association studies of the population. Using genome-wide data to find SNPs with large allele frequency differences between the source populations of the SAC, as quantified by Rosenberg et. al's I_n-statistic, we developed a panel of AIMs by experimenting with various

selection strategies. Subsets of different sizes were evaluated by measuring the correlation between ancestry proportions estimated by each AIM subset with ancestry proportions estimated using genome-wide data. We show that a panel of 96 AIMs can be used to assess ancestry proportions and to adjust for the confounding effect of the complex five-way admixture that occurred in the South African Coloured population.

3.2 Introduction

The predominant population group in the Western Cape, South Africa, is the admixed group known as the South African Coloured (SAC). The SAC had their origins in the diverse groups in the early days of Cape history, including European settlers from 1652, the slaves they brought in from Indonesia, India and other parts of Africa, local Bantu-speakers, and the indigenous Khoe-San. They therefore constitute a complex combination of continental populations [26]. Genetic variation between humans can be ascribed to differences between individuals within populations (85-90%) and to differences between populations (10-15%) [2; 3; 4; 5]. As humans migrated out of Africa, genetic drift or adaptation resulted in different frequencies of genetic variants in the resultant populations. It is often possible to cluster individuals into population groups that correspond to their self-reported ancestry because of these differences [110]. Admixture occurs when two or more previously separated population groups produce offspring, and it is a well-known confounder in genetic association studies [111; 112; 113]. In case-control genetic studies, if cases have a different proportion of ancestry from a source population compared to controls, associations found may be related to ancestry rather than disease [37]. It is therefore important to incorporate ancestry in regression models used in genetic association studies of admixed populations. Given genome-wide markers for individuals from an admixed population, principal components or ancestry proportions estimated by solving a multinomial model can be used as covariates to adjust for admixture. However, obtaining genome-wide markers in small follow-up or candidate gene association studies may be prohibitively expensive. Ancestry informative markers (AIMs) are those polymorphisms with the greatest difference in frequency between populations. AIMs can be used as a cost-effective alternative to genome-wide data, if the markers have different allele frequencies in the source populations of the admixed population.

Panels of AIMs have been drawn up for specific populations and purposes. Kosoy et al. set out to find AIMs to determine continental origin and admixture proportions for populations common in America [114]. A list of 128 SNPs were produced by considering the effect of a SNP for distinguishing ancestry independently of the contribution of other SNPs in the data set. This list was later reduced to 93 SNPs [115]. To distinguish between three populations, Galanter et al. [116] used the locus specific branch length (LSBL) of a SNP statistic measured between each pair of three populations [117]. The LSBL was calculated per SNP to develop a panel of AIMs for a diverse set of admixed populations in the Americas that has African, European and Native American ancestry. These AIMs are equally informative for each of the source ancestries, and the panel was shown to provide accurate ancestry proportion estimates by comparing with robust estimates inferred from genome-wide data. SNPs may also be selected by evaluating their combined effect using a performance function. Lao et al. [118] used an asymptotic approximation of the I_n-statistic calculated for multiple markers as a performance function [119; 120]. Lao showed that only ten SNPs are required to distinguish the continental ancestry of non-admixed individuals from Eurasia, Africa, America and East Asia. Paschou et al. selected SNPs with the highest loadings summed across the top principal components [121]. This study found that 14 SNPs can differentiate continental ancestry, 100 SNPs differentiate the intra-continental ancestry of the Chinese and Japanese populations, and 200 AIMs were necessary for the admixed Puerto Rican population.

A number of studies showed that the SAC received genetic contributions from click-speaking Africans (African San), Bantu-speaking Africans (African non-San), European, South and East Asians [26; 39; 9; 40; 41]. The large cohort of SAC individuals used in this paper represents the same population used in the genome-wide analysis performed by De Wit et al. [26] and Chimusa et al. [39]. De Wit et al. found that the cohort received large proportions of ancestry from African San, African non-San and European populations, and a smaller proportion of Asian ancestry. The Asian ancestry was most closely related to a Gujarati Indian population, followed by low levels of ancestry from East Asia. Similar proportions of ancestry were found by Quantana-Murci et al. [41] and Chimusa et al. [39]. These findings are consistent with historical records. Men outnumbered women in the early Cape Society and mixed liaisons were common [45; 44; 47; 46; 26]. The establishment of mission stations from the mid 1700s onwards further facilitated the integration of European, African (particularly Xhosa) and Khoe-San ancestries [122; 47; 26]. A large proportion of imported slaves originated from Bengal [45; 26]. Bengalis are genetically similar to the Gujarati Indians [9] used to represent the South Asian component in the De Wit and Chimusa studies. The small East Asian ancestry component may be ascribed to the "free black" Chinese who formed 9% of the Cape Town population in the early 1800s [47; 45; 122; 26]. This is more plausible than Indonesian ancestry, since the majority of the cohort are not Muslim and therefore unlikely to form part of the group known as the Cape Malay [26].

Sets of AIMs published by a number of studies [114; 115; 123; 124; 125; 5; 118; 116] are not suited to the SAC, since the Khoe-San was not considered as a separate population, or an insufficient number of Khoe-San individuals were used. Complex admixture models such as the five-way admixture that occurred in the SAC, with different levels of genetic distance between source populations, were also not considered. We therefore developed a panel of AIMs tailored to the SAC and assessed its accuracy compared to genome-wide data. Although all the methods discussed above select markers that are informative of ancestry, we also set out to ensure that the selected marker set is reasonably small and as efficient as possible in predicting ancestry. Preliminary investigations indicated that the method introduced by Galanter et al. [116] had the greatest chance of success, and we therefore adapted this method to allow more than three source populations.

3.3 Materials and methods

Our first step in selecting AIMs was to obtain genome-wide data from populations that are representative of the founding groups of the SAC. Using this data and various different methods to select AIMs, we then set out to find SNPs where the allele frequencies are the most differentiated between the various source populations.

Since the purpose of the AIMs is to adjust for the effects of admixture in genetic studies of the SAC, we assessed the accuracy of various candidate AIM panels by measuring the correlation between ancestry proportions estimated for a large study group of admixed individuals using AIMs and proportions estimated using genome-wide data. We used this information to select a final panel of AIMs of reasonable size.

Finally, we assessed whether the selected panel can be applied to four small South African Coloured study groups from different geographical locations, by measuring the correlation between AIM and genome-wide estimated ancestry proportions.

3.3.1 Ethics statement

Approval from the Ethics Committee of the Faculty of Health Sciences, Stellenbosch University (project registration numbers 95/072 and NO6/07/132), was obtained for the Cape Town study group presented in this study. Blood samples for DNA were collected with written informed consent. Sampling and DNA consent from the ǂKhomani San and individuals who self-identified as "Coloured" in Upington, South Africa and neighboring villages occurred in 2011 and 2012. Institutional Review Board (IRB) approval was obtained from Stanford University and Stellenbosch University (project registration number N11/07/210). ǂKhomani N|u-speaking individuals, local community leaders, traditional leaders, non-profit organizations and a legal counselor were all consulted regarding the aims of this research, prior to collection of DNA, and regular feedback was given to the community. This research was conducted according to the principles expressed in the Declaration of Helsinki.

3.3.2 Data

Genome-wide data were obtained from a large study group of individuals who self-identified as South African Coloured and who resided in the Cape Town suburbs of Ravensmead and Uitsig. DNA samples collected from the study group were genotyped on the Affymetrix GeneChip Human Mapping 500K Array Set. More details regarding the sampling and study site are described by De Wit et al [26]. After SNP calling, SNPs that failed a missing threshold of 5%, a minor allele frequency threshold of 1% or a HWE test with an alpha level of 0.0001 were removed. Outliers, related individuals and individuals with a genotyping rate of less than 95% were then removed, resulting in a data set of 733 individuals.

Genome-wide data of four small admixed study groups from different geographical locations were obtained as follows. The first group came from a ǂKhomani San community in the region of Upington in the Northern Cape, where DNA samples were collected from 21 unrelated individuals who either self-identified as Coloured or had at least one parent who self-identified as Coloured. The samples were genotyped on the Illumina 550K and Illumina OmniExpress (700K) platforms. SNPs that failed a missing threshold of 5% and a minor allele frequency threshold of 0.5% were removed from the data set. Data published by Schlebusch et al. [126] was used for the remaining groups. This data includes three admixed study groups of 20 individuals each. Two of the study groups comprise Coloured individuals from Colesberg in the Northern Cape and Wellington in the Western Cape, respectively. The third study group comprises 20 individuals from the community known as the Karretjie people in the Colesberg region. High proportions of Khoe-San ancestry are present in the Karretjie people [126], and it is thought that they also have European and Bantu ancestry. The DNA samples were genotyped on the Illumina Omni 2.5M SNP chip. The non-imputed data set was used, and no additional SNP quality control steps were performed.

The populations described in Table S1 of Chimusa et al. [39] were considered as potential source populations for the SAC. Principal component and ancestry proportion analysis were used to identify populations with relatively high levels of admixture (see Figures S3-S6 of Chimusa et al.), thereby ensuring that only non-admixed source populations were used for AIM selection. Consequently some of the southern and eastern African populations were excluded from subsequent analysis. Individuals in the Khoe-San data sets that showed relatively high levels of admixture were also removed. The HGDP Melanesian and Papua-New Guinean populations were additionally considered as potential source populations in order to have a comprehensive list, but were excluded since the populations did not appear to be closely related to the Cape Town study group (see supplementary figure 3.6.1), which fits with the historical evidence. The Khoe-San data set used to represent the Ju|'hoansi population was obtained

from a private data access committee (contact corresponding author). The data set represents the same group analyzed by Schlebusch et al. [126], but was genotyped on the Affymetrix genotyping platform instead of the OmniExpress platform, which overlaps better with SNPs in the other source population data sets that were considered.

Chimusa developed a novel algorithm that identifies the best populations to use as proxy source populations for a multi-way admixed population. This algorithm, as described by Bensmail [127], was used to guide selection of the best populations from the candidate proxy source populations identified by the preliminary investigation. The algorithm leverages the idea that LD is created between genetic loci when admixture occurs between previously isolated populations. A score statistic is calculated per candidate reference population, by measuring the correlation between the LD in the admixed population and the allele frequency difference between the candidate reference population paired with another reference population, for all such possible pairs. The results of the algorithm are summarized in supplementary table 3.6.1. The top scoring groups per source population were then used to represent the source populations of the SAC. Ideally only the top one or two scoring populations should be selected as reference populations, but this would have resulted in small sample sizes for the African San and African non-San data sets. Consequently all the African San and the top 8 African non-San populations were selected. The Pakistan South Asian population was not used as we did not have historical evidence to support the use of this population. The HapMap CHB Chinese was also excluded since the group appeared to be very similar to the HapMap CHD Chinese. The final source population data set is summarized in table 3.1. Supplementary figure 3.6.2 is a map representing the geographic locations of the source populations of the SAC used in this study, as well as the admixed SAC study groups.

AIMs were selected from the set of SNPs found in all of the source population data sets and the Cape Town study group data set. When estimating ancestry proportions of an admixed study group using genome-wide data, SNPs that were not found in all of the source population data sets were first removed, after which SNPs were filtered according to a linkage disequilibrium (LD) threshold. This was done as increased LD found in admixed populations may bias ancestry proportion estimation. Supplementary table 3.6.2 presents information on the thresholds applied and number of SNPs used for genome-wide ancestry proportion estimation.

3.3.3 Selecting ancestry informative markers

Rosenberg's I_n-statistic [119] is a measure of the informativeness of a genetic marker in determining an individual' s ancestry, for any number of potential source populations. It is often used to select AIMs, as markers with large allele frequency differences between populations will also have a large I_n-statistic. Galanter et al. selected SNPs based on the LSBL of this statistic, such that the total LSBL calculated for each of the source populations of admixed Latin Americans are equivalent [116].

The LSBL can however only be calculated for three populations and could therefore not be applied to the five source populations of the SAC. We therefore modified their approach to first select a proportion of SNPs according to the I_n-statistic calculated across all of the source populations, and to then select additional SNPs by balancing the total I_n-statistic between all pairs of source populations, as described below.

Rosenberg's I_n-statistic is defined as follows. For a SNP with alleles $\{A, a\}$ let p_A be the frequency of allele A calculated across all the individuals and let p_a be the frequency of allele a across all the individuals, for that marker. Let K be the number of populations represented by the individuals. Let p_{iA} be the frequency of allele A in population i and let p_{ia} be the frequency of allele a in population i. The informativeness of assignment of a SNP is given by

Table 3.1: Source population data. Data sets used to represent the five source populations of the South African Coloured population. The sample size reflects the group size after relative pairs have been removed. Henn et al. [15] merged the Juu San data from the Human Genome Diversity Project (HGDP) and Schuster et al. [128] and the African non-San data from Bryc et al [8].

Source population	Group	Description	Source	Platform	Size
African San (san)	kho	‡Khomani San from Northern Cape, South Africa	Henn 2011	Illumina 550K	14
	bus	Juu San from South Namibia	Henn 2011	Illumina 650K & 1M	9
	khs	Ju|'hoansi San from North Namibia	Private	Affymetrix 6.0	22
African non-San (afr)	brong	Ghana	Henn 2011	Affymetrix 500K	8
	kongo	Atlantic coast of Congo	Henn 2011	Affymetrix 500K	9
	igbo	Southeastern Nigeria	Henn 2011	Affymetrix 500K	15
	fang	Equatorial Guinea	Henn 2011	Affymetrix 500K	15
	bulala	Central Chad	Henn 2011	Affymetrix 500K	15
	mada	West Cameroon	Henn 2011	Affymetrix 500K	12
	hausa	West Nigeria	Henn 2011	Affymetrix 500K	12
	bamoun	West Cameroon	Henn 2011	Affymetrix 500K	18
European (eur)	CEU	Utah residents with Northern and Western European ancestry, USA	HapMap3	Release 3	111
	TSI	Italians from Italy	HapMap3	Release 3	102
South Asian (sas)	GIH	Gujarati Indians from Houston, Texas, USA	HapMap3	Release 3	97
East Asian (eas)	CHD	Chinese Metropolitan Denver, Colorado, USA	HapMap3	Release 3	106
	JPT	Japanese from Tokyo, Japan	HapMap3	Release 3	113

$$I_n = -p_A \ln(p_A) + \frac{1}{K}\sum_{i=1}^{K} p_{iA}\ln(p_{iA}) - p_a \ln(p_a) + \frac{1}{K}\sum_{i=1}^{K} p_{ia}\ln(p_{ia})$$

where $0\ln(0)$ is defined as 0.

It is similar to a log-likelihood ratio, where the ratio is the likelihood that an allele is assigned to one of the populations ($\frac{1}{K}\sum_{i=1}^{K} p_{iA}\ln(p_{iA}) + \frac{1}{K}\sum_{i=1}^{K} p_{ia}\ln(p_{ia})$), versus the likelihood that the allele is assigned to the average population ($-p_A\ln(p_A) - p_a\ln(p_a)$)

The allele frequency of each SNP in the data set was calculated, for each source population, and for the population groups included in a source population (for example the East Asian source population comprises the HapMap Japanese and Chinese study groups). SNPs were discarded if they were heterogeneous in these subgroups, based on a Chi-squared test that has a null hypothesis of equal allele frequencies in the subgroups. SNPs were then selected according to the I_n-statistic calculated across all the source populations, and the I_n-statistic calculated between pairs of populations. Checks were performed before a SNP was accepted as an AIM, to determine whether the SNP was already in the list of AIMs, or was in linkage disequilibrium with any of the SNPs in the list ($r^2 > 0.1$), or was located close to any of the SNPs (measured in number of base pairs).

SNPs were selected as follows. The I_n-statistic was calculated for all SNPs, across all the source populations, and used to select SNPs with the highest values. This multiple population I_n-statistic may however be skewed towards populations that are more differentiated (i.e.

SNPs from less differentiated populations will contribute less to the statistic and will therefore have a smaller probability of being selected as an informative marker). Additional SNPs were therefore selected by calculating the I_n-statistic of each SNP for each pair of populations, and then selecting SNPs by balancing the total pairwise I_n-statistic. For example, for five source populations there are $\binom{5}{2} = 10$ pairs of populations. The pair with the smallest total I_n-statistic was identified (initially, the total of all pairs are set to zero and are therefore tied) and the SNP with the highest I_n-statistic for the identified pair was selected as an AIM. In the case of a tie(s), the SNP with the highest I_n-statistic for the tied pair(s) was selected. If the SNP was accepted, its I_n-statistic value for the relevant pair was added to the pair's total I_n-statistic. This process was repeated until the required number of AIMs were accepted.

We generated panels of AIMs of sizes 25, 50, 75, ..., 500 using this approach, and experimented with including versus excluding SNPs that are heterogeneous in the populations that constitute a source population, different minimum distances between SNPs and selecting different proportions of markers (0, 0.1, 0.25, 0.5 and 1) using the multiple population I_n-statistic. We also experimented with selecting markers using the implementations provided by Lao et al. [118] and Paschou et al. [121].

3.3.4 Assessing ancestry informative marker panels

Let G be a matrix of genotypes for each of the n individuals in the data set, F be a matrix of variant allele frequencies for each of the k source populations, and Q be a matrix of k ancestry proportions for each of the n individuals. Ancestry proportions can be estimated by maximizing the likelihood function $L(Q, F|G)$.

A strong correlation between ancestry proportions estimated using AIMs for a particular ancestry and ancestry proportions estimated using genome-wide data for the same ancestry would show that the AIMs are informative for that ancestry, even though the number of markers used in the estimation has been much reduced from genome-wide data. We therefore estimated the ancestry proportions of individuals from a combined genome-wide data set composed of both the source population data sets and the Cape Town admixed study group, and identified ancestries as follows. The mean ancestry proportion was calculated for each of the k possible ancestries, per source population (using only individuals from that particular source population). The ancestry of a particular source population was then identified by determining which of the k possible ancestries had the largest mean ancestry proportion for that population. The same procedure was used for combined AIM data sets. The correlation between ancestry proportions estimated using the genome-wide data set and proportions estimated using each AIM data set was then calculated per ancestry, using individuals from the admixed study group.

3.3.5 Software

We modified the Python script provided by Galanter et al. [116] to support more than three source populations. Lao provided us with a Java implementation of his method and we ported the Paschou MATLAB implementation to R [121]. We used PROXYANC to select the best proxy ancestral populations. PLINK [105] was used for quality control filtering, LD filtering and to calculate allele frequencies per population. ADMIXTURE's unsupervised algorithm was used to estimate ancestry proportions [129] and the EIGENSTRAT smartpca program was used for principal component analysis [109]. Statistical analyses were performed using R.

The python script we used to select AIMs is available at `http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0082224` as supplementary file 1. PROXYANC is found at `http://www.cbio.uct.ac.za/proxyanc/software.html`.

3.4 Results

The correlation between ancestry proportions estimated using AIMs and proportions estimated using genome-wide data was calculated for AIM sets of increasing size (25, 50, ..., 500 SNPs) for different combinations of parameter settings.

For investigating the effect of heterogeneity between subgroups of a source population (the subgroups are summarized under the Population Group heading of table 3.1), we used a minimum distance of 100 000 base pairs between SNPs. We selected different proportions of markers using the multiple population I_n-statistic while the remaining SNPs were selected using the pairwise I_n-statistic. The difference between the correlation calculated using a AIM set selected from all markers versus the correlation of a AIM set of the same size selected from a marker set containing no heterogeneous SNPs was measured. A positive difference indicates that the AIM set selected from all markers has a higher correlation. Figure 3.1 depicts the magnitude and direction of the differences measured for the different AIM set sizes and multiple population I_n-statistic parameter settings. Since 390 of the 400 differences are positive, we ignored heterogeneity in subsequent AIM selections.

Figure 3.2 shows the differences between correlations estimated using a minimum distance of 100 000 versus a 1 000 000 base pairs between SNPs for different AIM set sizes and multiple population I_n-statistic parameter settings. A positive difference indicates that the 100 000 base pair distance has a larger correlation. Although the differences are small and the number of positive differences are not much larger than the number of negative differences, the magnitude of the positive differences are greater compared to the negative differences, except for one of the multiple population I_n-statistic parameter settings. For this reason, we used a minimum distance of a 100 000 base pairs between markers in our subsequent AIM selections.

A proportion of 0, 0.1, 0.25, 0.5 and 1 markers per set were selected using the multiple population I_n-statistic while the remaining SNPs were selected using the pairwise I_n-statistic. Selecting all markers using the multiple population statistic (i.e. a proportion of 1) resulted in the ambiguous classification of the source populations for smaller AIM sets; at least 200 SNPs were required for classifying the source populations correctly. Figure 3.3 shows the correlation per source population for AIM sets of increasing size for the first four multiple population I_n-statistic parameter settings. The figure shows that the optimal estimated proportions in terms of cost vs. benefit are obtained using approximately 100 SNPs - incremental improvement in accuracy of estimation using more markers is smaller after this point. Selecting all SNPs by balancing the total pairwise I_n-statistic appears to be slightly better compared to selecting some of the SNPs using the multiple population I_n-statistic and we therefore used this parameter setting for selecting the final panel of AIMs.

As it is conceivable that future cost reductions may render the cost of genotyping additional SNPs irrelevant, supplementary table 3.6.3 presents a panel of 2000 ordered AIMs that were selected using the criteria described above. This large panel can potentially also be used for local ancestry inference. It is currently possible to genotype 96 SNPs cost-effectively on a number of platforms, such as the BeadXpress system, and we therefore evaluated the first 96 SNPs (roughly the optimal number of markers) as our primary panel of AIMs. We also evaluated a panel with 24 additional SNPs, since this slightly larger set of 120 SNPs provides a 3.54% and 5.15% increase in correlation for the estimated African San and South Asian ancestry proportions respectively. This larger marker set can be genotyped using technologies such as Sequenom plexes and Taqman assays, and the results of its evaluation are summarized by supplementary tables and figures. As expected, for both the 96 and 120 SNP panels the number of AIMs selected per population pair is inversely proportional to the genetic distance between the two populations (supplementary table 3.6.4).

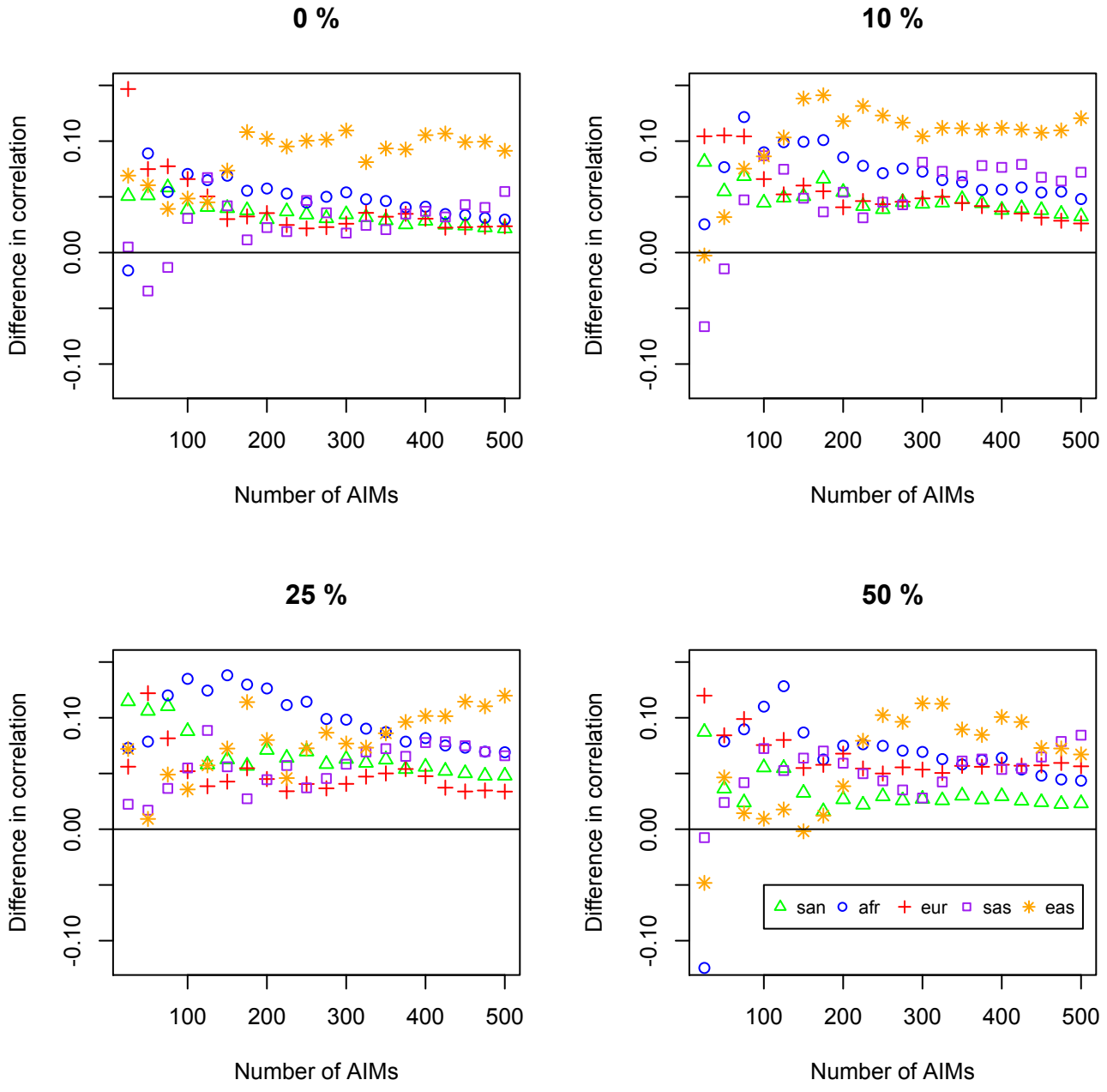


Figure 3.1: Scatter plots of the difference in correlation coefficients against the number of AIMs used in the calculation of the correlations, when ignoring heterogeneity versus removing heterogeneous SNPs. Both correlations are between ancestry proportions estimated from genome-wide data and ancestry proportions estimated using a set of AIMs selected from the genome-wide data. The difference is between the AIMs selected from all the genome-wide SNPs and those selected from genome-wide SNPs from which markers that are heterogeneous in subgroups of the source populations have been removed. The percentage of SNPs selected using the multiple I_n-statistic (the remainder were selected using the pairwise I_n-statistic) are shown for each plot. SNPs were selected with a minimum distance of 100 000 base pairs between them.

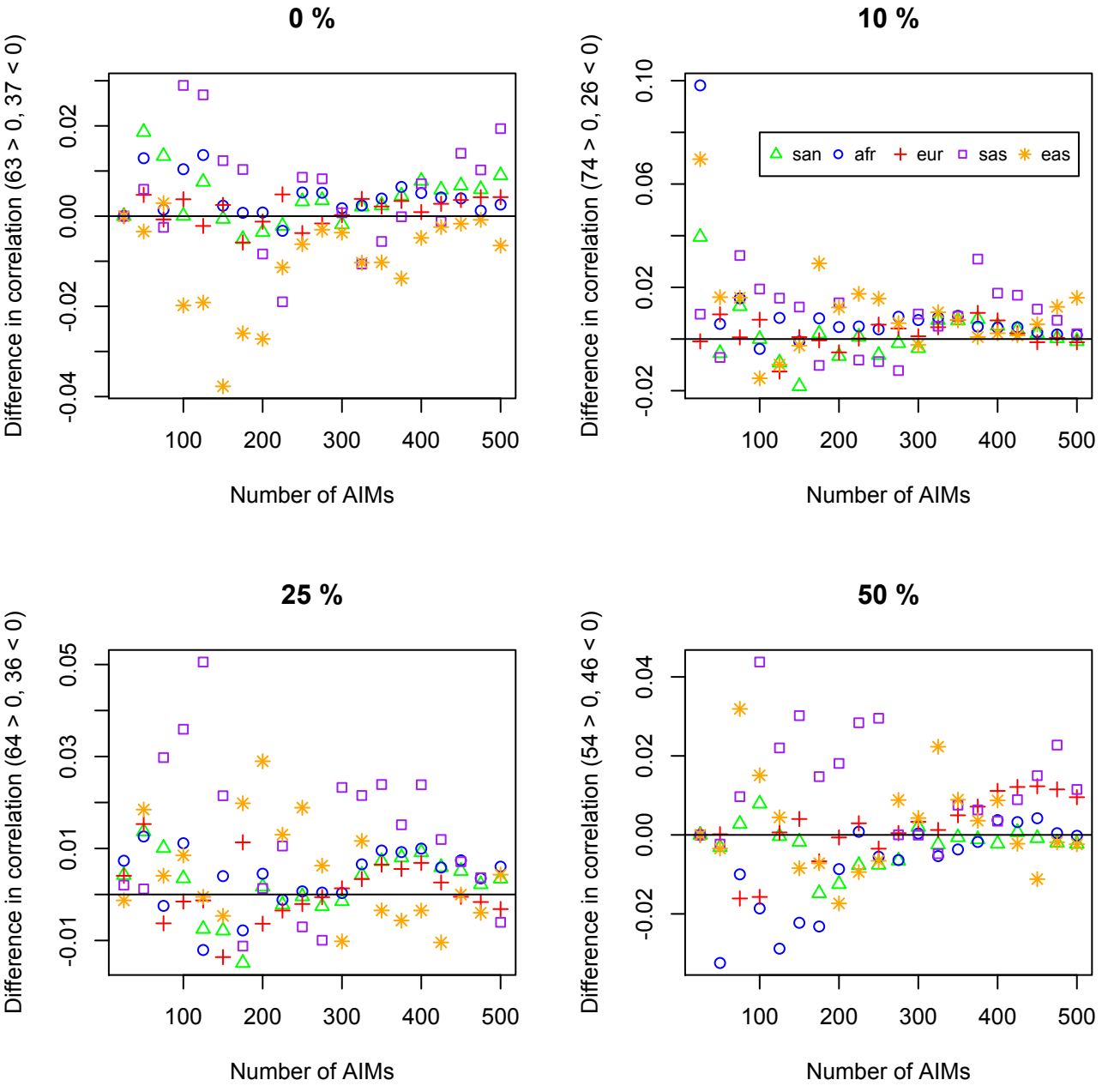


Figure 3.2: Scatter plots of the difference in correlation coefficients against the number of AIMs used in the calculation of the correlations, when using a minimum distance of 100 000 base pairs between SNPs versus a 1 000 000 base pairs. Both correlations are between ancestry proportions estimated from genome-wide data and ancestry proportions estimated using a set of AIMs selected from the genome-wide data. The difference is between the AIMs selected so that there is a minimum distance of 1 000 000 base pairs between them and those selected with a minimum distance of 100 000 base pairs between them. AIM sets were selected from all the genome-wide SNPs. The percentage of SNPs selected using the multiple I_n-statistic (the remainder were selected using the pairwise I_n-statistic) are shown for each plot.

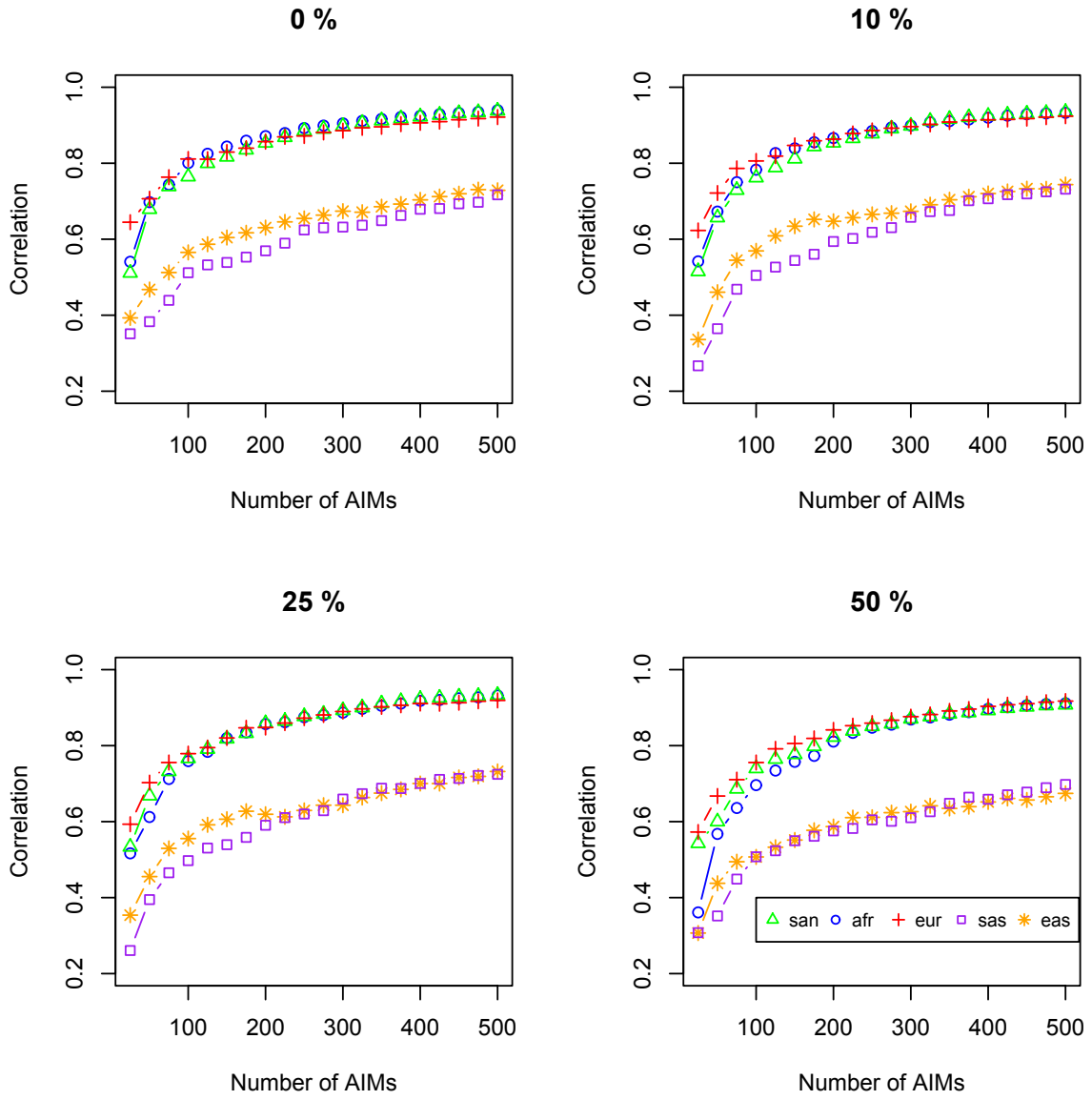


Figure 3.3: Admixture proportion correlation versus number of AIMs in set. Correlation between admixture proportions estimated using AIMs and proportions estimated using genome-wide data, using AIM sets of increasing size (increments of 25) for the Cape Town study group (n=733). A proportion of the SNPs in each set of AIMs were selected using the multiple I_n-statistic, indicated in each panel as a percentage, while the remaining SNPs were selected using the pairwise I_n-statistic, as described in the Methods section.

Table 3.2: Correlation and RSME of 96 and 120 AIMs. Correlation and RSME between ancestry proportions estimated using the 96 and 120 AIM panels respectively and proportions estimated using genome-wide data, for the Cape Town study group (n=733).

	96 panel		120 panel	
Ancestry	**Correlation**	**RSME**	**Correlation**	**RSME**
African San	0.7565	0.0684	0.7905	0.0621
African non-San	0.7930	0.0774	0.8160	0.0719
European	0.8019	0.0554	0.8150	0.0535
South Asian	0.4808	0.0658	0.5283	0.0625
East Asian	0.5665	0.0560	0.5822	0.0522

Table 3.2 summarizes the correlation and RSME for the 96 and 120 AIMs. Figure 3.4 shows Bland Altman plots per ancestral population of the difference between the genome-wide and AIMs estimated proportions versus the genome-wide estimated proportions for each individual (for the 96 AIMs panel). The figure suggests that there are no systematic differences in the ancestry estimation.

As large study groups may require fewer markers to differentiate ancestries [130], the ability of the AIMs to estimate ancestry proportions of a smaller group of South African Coloured individuals were evaluated using permutation testing. 100 individuals were randomly selected from the total of 733 and their ancestry proportions were estimated. The correlation with the genome-wide ancestry proportions for those individuals was then calculated. This process was repeated a 100 times. Figure 3.5 gives boxplots of the correlation coefficients calculated for each permutation. The red diamonds in the figure are the correlation coefficients calculated using all 733 individuals; this shows that the AIMs perform well for a smaller group of individuals.

Markers used to estimate the ancestry proportions of an admixed population can only perform well if they can also distinguish between the source populations of the admixed population. Figure 3.6 is a barplot of the estimated ancestry proportions for the combined data set, using AIMs and using genome-wide data for the estimation. It shows that for most of the source population individuals, the largest proportion of ancestry is correctly assigned to the relevant population group using AIMs, albeit less well when compared to using genome-wide data. The first three principal components formed using the AIMs for the source population data are depicted in supplementary figure 3.6.3, which also suggests that the AIMs can be used to group the five source populations, although the clusters are wider compared to genome-wide data. Fifty-one percent of the variance in the data is explained by the first three components.

Figure 3.7 is a histogram of the number of AIMs found on each chromosome, showing that the panel is representative of the entire genome, and that more markers are generally found on the larger chromosomes. This is important since ancestry proportions estimated from markers that are localized to only one part of the genome may differ substantially from an admixed individual's true ancestry proportions across their entire genome. The position of the markers on each chromosome is represented in supplementary figure 3.6.4.

Figure 3.8 depicts boxplots of ancestry proportions estimated using genome-wide data and proportions estimated using AIMs per source population. It shows that the distribution of proportions estimated using AIMs are similar to proportions estimated using genome-wide data, especially for the median ancestry proportions, while the variation of the proportions is only slightly inflated when using AIMs.

To assess the accuracy of the application of the panel to Coloured groups sampled from different geographic locations, we selected markers from the additional Coloured data sets

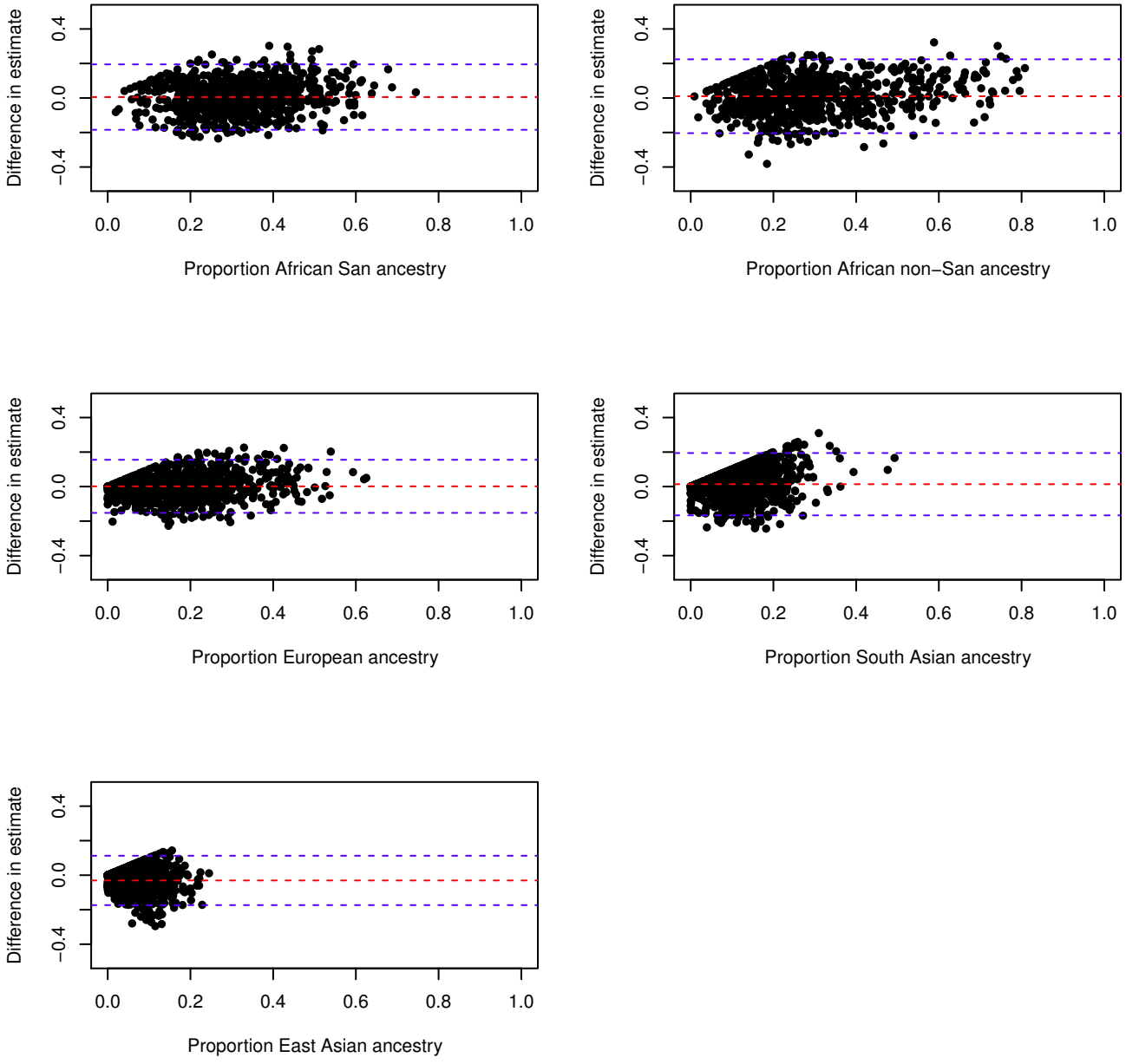


Figure 3.4: Bland Altman plots of differences between ancestry proportion estimates. Bland Altman plots per ancestral population of the difference between the genome-wide and AIMs estimated proportions (y-axis) versus the genome-wide estimated proportions (x-axis) for each individual, using 96 AIMs. Each panel respresents the ancestry proportions of one of the source populations of the SAC.

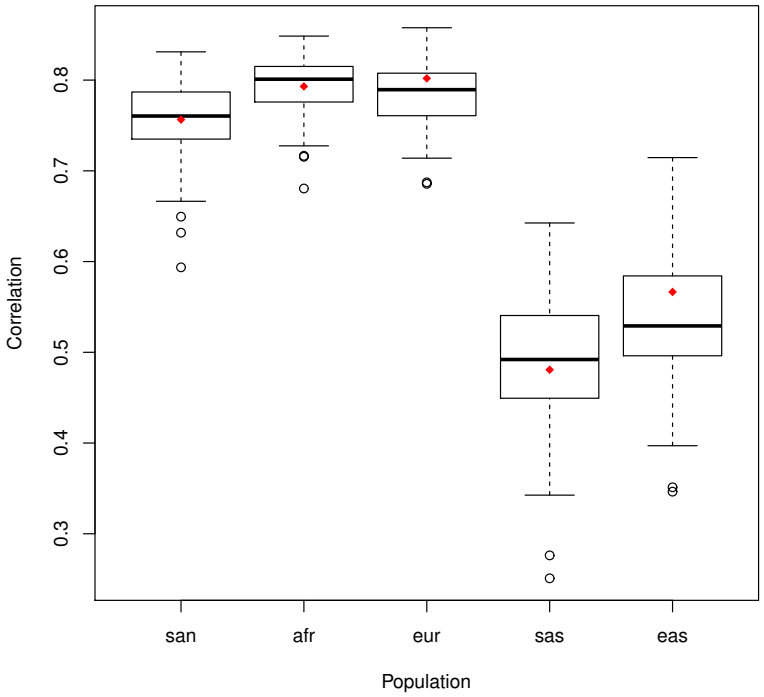


Figure 3.5: Boxplot of permutation correlation. A boxplot of correlation coefficients calculated in 100 permutations per source population, each permutation comprising a random draw of 100 individuals from the Cape Town study group (n=733). The correlation was measured between admixture proportions estimated using the panel of 96 AIMs and proportions estimated using genome-wide data. The red diamonds represent the correlation coefficients calculated using the entire study group.

described in Materials and Methods that overlapped with the 120-SNP panel. 76 overlapping SNPs were found in the Upington data set and 84 SNPs were found in the Schlebusch data sets. The number of markers per ancestry pair for each set is shown in figure 3.9. Table 3.3 summarizes the correlations between ancestry proportions estimated using the overlapping AIMs and genome-wide data for each study group. This shows that the markers perform well for each of the groups, considering the reduced size of the AIM panel, possible non-optimal number of markers per ancestry pair and the small group size. Figure 3.10 depicts boxplots of ancestry proportions estimated using genome-wide data versus proportions estimated using AIMs per source population. The figure illustrates that the distribution of the proportions estimated using AIMs are comparable to the distribution of genome-wide proportions for all the groups. The median and interquartile range of the ancestry proportion estimates inferred from genome-wide data and AIMs are also presented in table 3.4, for all the study groups.

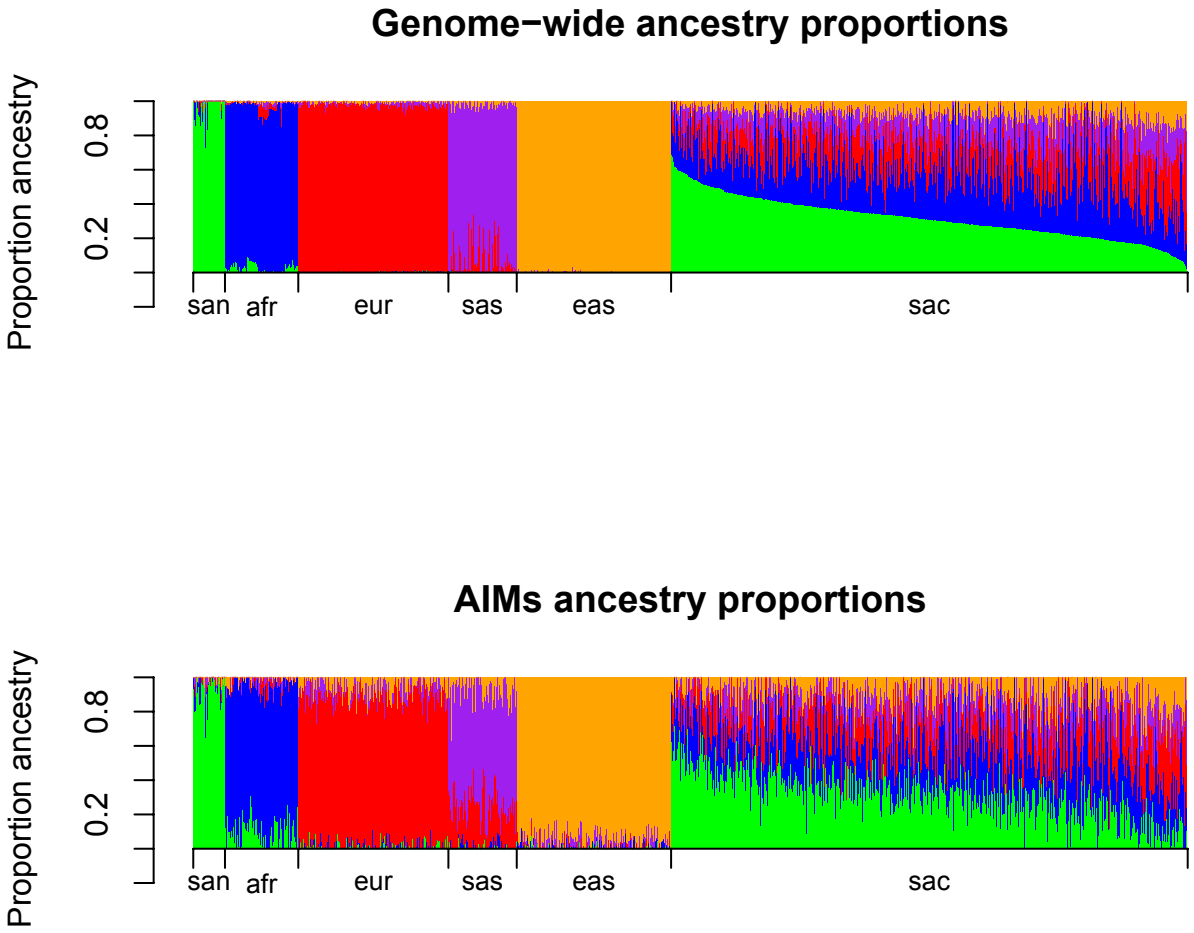


Figure 3.6: Barplots of ancestry proportions estimated using genome-wide data and using AIMs. In the first panel ancestry proportions were estimated using genome-wide data. The admixed study group (sac) is ordered by proportions of African San, African non-San, European, South Asian and East Asian ancestry. In the second panel ancestry proportions were estimated using 96 AIMs. Individuals appear in the same order as in the first panel.

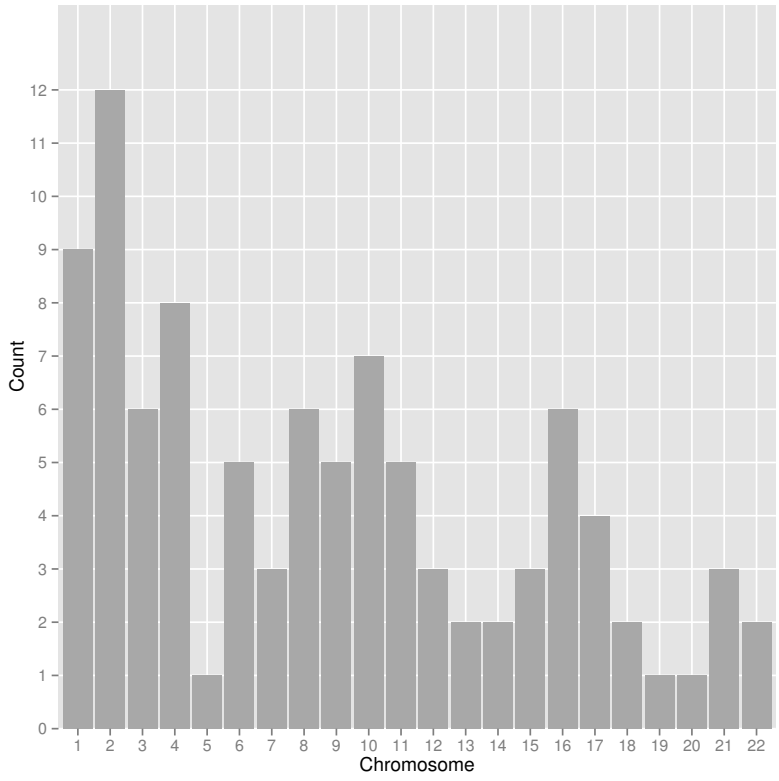


Figure 3.7: Histogram of the number of AIMs on each chromosome. Histogram that represents the number of markers in the panel of 96 AIMs per chromosome.

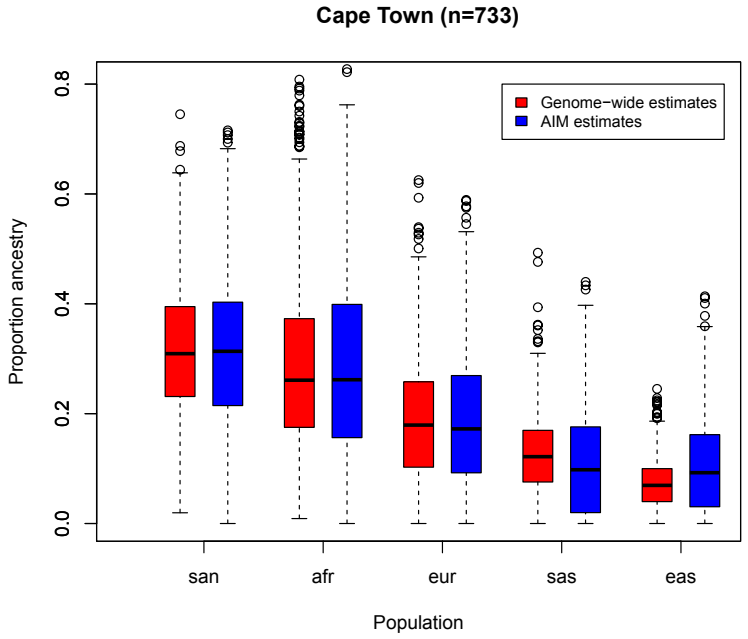


Figure 3.8: Boxplots of ancestry proportions of the Cape Town study group. Boxplots of ancestry proportions estimated using genome-wide data and proportions estimated using the panel of 96 AIMs are shown in this figure per source population, for the Cape Town study group (n=733).

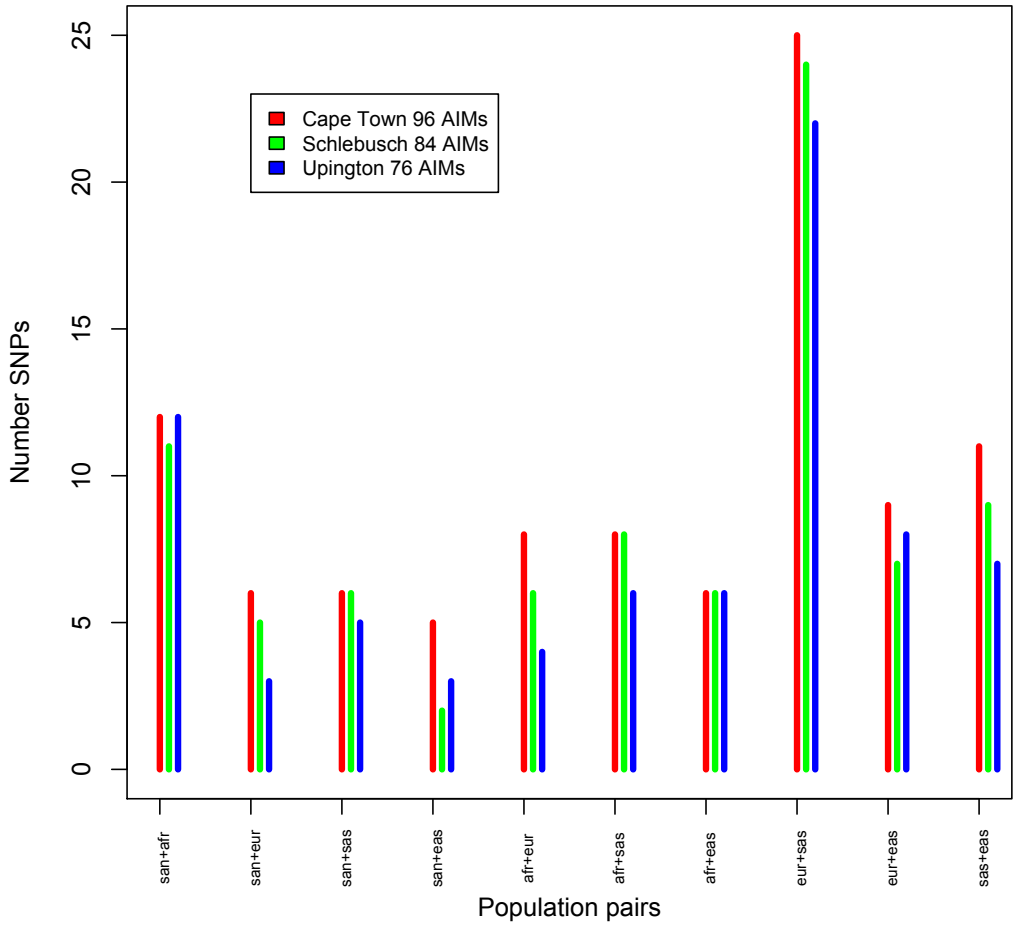


Figure 3.9: Number AIMs found in admixed study groups per population pair. The number of AIMs per source population pair found in the different admixed study group data sets.

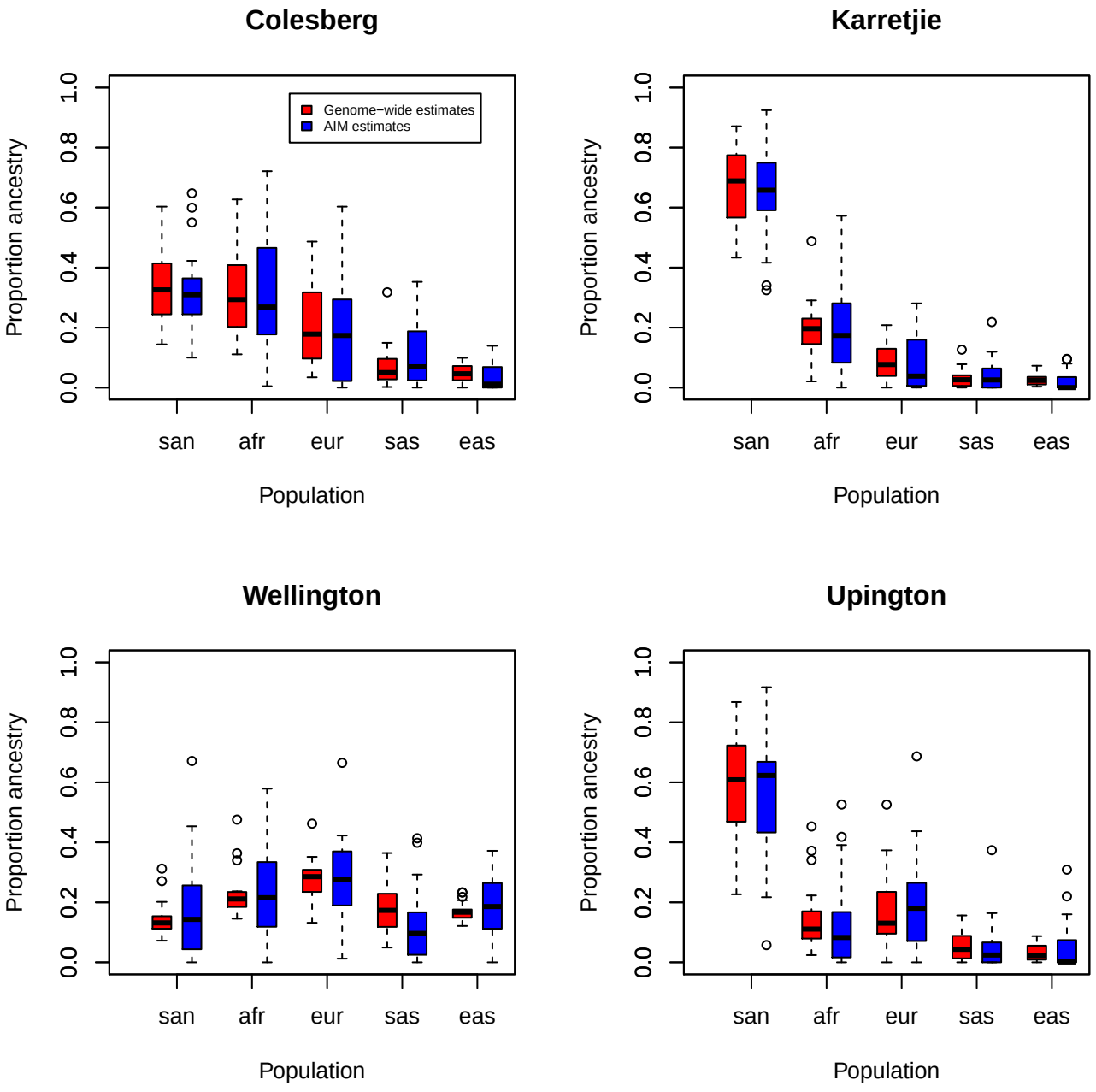


Figure 3.10: Boxplot of ancestry proportions of small admixed study groups. The distribution of ancestry proportions estimated using genome-wide data and proportions estimated using AIMs are shown in this figure for the small admixed study groups, per source population. The Colesberg, Karretjie and Wellington study groups are each comprised of 20 individuals and 84 AIMs were used to estimate ancestry proportions. The Upington study group comprises 21 individuals and 76 AIMs were used to estimate ancestry proportions.

Table 3.3: Correlation for different admixed study groups. Correlation between ancestry proportions estimated using AIMs and proportions estimated using genome-wide data, for small admixed study groups from different geographic locations. The number of AIMs reflects the number of markers in the 120 panel that were found in the genome-wide data sets of the study groups.

Study group	Number AIMs	African San	African non-San	European	South Asian	East Asian
Colesberg (n=20)	84	0.7661	0.8437	0.8996	0.4675	0.4731
Karretjie (n=20)	84	0.8436	0.7007	0.7724	0.5590	0.1815
Wellington (n=20)	84	0.7252	0.7102	0.8008	0.6783	0.3311
Upington (n=21)	76	0.8747	0.6304	0.8739	0.3777	0.3426

Table 3.4: Ancestry proportion distribution. Median and IQR of the ancestry proportions estimated using genome-wide data and AIMs, per admixed study group.

Study group	Data set	African San	African non-San	European	South Asian	East Asian
Cape Town (n=733)	Chip	0.31 (0.23-0.39)	0.26 (0.18-0.37)	0.18 (0.10-0.26)	0.12 (0.08-0.17)	0.07 (0.04-0.10)
	96 AIMs	0.31 (0.21-0.40)	0.26 (0.16-0.40)	0.17 (0.09-0.27)	0.10 (0.02-0.18)	0.09 (0.03-0.16)
	120 AIMs	0.31 (0.22-0.40)	0.27 (0.16-0.39)	0.17 (0.09-0.27)	0.11 (0.03-0.19)	0.08 (0.03-0.15)
Colesberg (n=20)	Chip	0.33 (0.25-0.40)	0.29 (0.21-0.40)	0.18 (0.10-0.29)	0.05 (0.03-0.09)	0.05 (0.02-0.07)
	84 AIMs	0.31 (0.24-0.35)	0.27 (0.18-0.46)	0.17 (0.03-0.29)	0.07 (0.03-0.19)	0.01 (0.00-0.05)
Karretjie (n=20)	Chip	0.69 (0.57-0.77)	0.20 (0.15-0.23)	0.08 (0.04-0.12)	0.03 (0.01-0.04)	0.02 (0.01-0.04)
	84 AIMs	0.66 (0.59-0.74)	0.17 (0.08-0.27)	0.04 (0.01-0.16)	0.03 (0.00-0.06)	0.00 (0.00-0.02)
Wellington (n=20)	Chip	0.13 (0.12-0.15)	0.21 (0.19-0.23)	0.29 (0.24-0.31)	0.17 (0.12-0.23)	0.17 (0.15-0.18)
	84 AIMs	0.14 (0.04-0.25)	0.22 (0.14-0.33)	0.28 (0.19-0.37)	0.10 (0.03-0.16)	0.19 (0.11-0.26)
Upington (n=21)	Chip	0.61 (0.47-0.72)	0.11 (0.08-0.17)	0.13 (0.10-0.23)	0.04 (0.01-0.09)	0.02 (0.01-0.06)
	76 AIMs	0.62 (0.43-0.67)	0.08 (0.02-0.17)	0.18 (0.07-0.26)	0.02 (0.00-0.07)	0.00 (0.00-0.07)

Supplementary tables 3.6.5 and 3.6.6 present correlations achieved by AIM sets of sizes 88, 194 and 314 AIMs for the Galanter et. al. study [116] and our large SAC study group, as well as sets of 500 and 2000 AIMs for five-way admixture in the SAC. The tables can be used to compare correlations in this study to those obtained by Galanter et al. As expected, the more complex five-way admixture modelling does not yield correlations that are quite as high as the Galanter study for sets of the same size, but this is easily rectified by including additional markers. In addition, when using only the markers that were selected to distinguish the African San, African non-San and European populations and using a simpler three-way admixture model, the correlations are comparable.

We also evaluated AIM panels selected by Lao et al.'s [118] and Paschou et al.'s methods [121], but could not find a smaller set of markers that resulted in stronger correlation between AIM and genome-wide estimated ancestry proportions.

3.5 Discussion

We report the development of a panel of AIMs for the South African Coloured population that enables researchers working with this population to assess population ancestry proportions and correct for substructure. The SAC has a complex history of admixture [26; 41] and has been used in many genetic association studies [131; 132; 133; 134; 135; 136; 137; 138; 66; 139; 140; 141]. Such candidate gene association studies investigate variants that are often not available in micro-array data. Obtaining genome-wide markers to then simply adjust for admixture may be prohibitively expensive. A viable cost-effective alternative is the genotyping of AIMs. To date, none of the published lists of AIMs have been developed or adequately assessed for distinguishing the ancestries of the SAC, which received genetic contributions from five source populations. Wacholder et al. has argued that confounding due to admixture is minimal for more than three source populations, and that the effect of admixture decrease as the number of strata increases [142]. This study was however limited to U.S. citizens with admixed European ancestry. Studies of multi-way admixed populations formed from different continental populations, that display larger differences in allele frequencies compared to intra-continental populations, may still suffer from the confounding effect of admixture. As an illustration, in a genome-wide tuberculosis (TB) case-control study of the SAC (642 cases and 91 controls), Chimusa et al. found a statistically significant positive correlation between the proportion of African San ancestry and TB susceptibility, and significant negative correlations when regarding European, East Asian and South Asian ancestries [98]. We therefore developed a panel of 96 AIMs for the SAC, by selecting SNPs that can distinguish between all pairs of source populations, as measured by Rosenberg's I_n-statistic. The AIMs can be used to adjust for the confounding effect of admixture in genetic association studies of the SAC. The correlation between AIMs and genome-wide estimated ancestry proportions may not be sufficient to suggest confidence in ancestry proportions estimated by AIMs at an individual level. However, when the entire study group is considered, the distribution of ancestry proportions are comparable. The panel therefore also has value for inferences about ancestry proportions at the population level. Although we focused on the ability of a small panel of AIMs to adjust for admixture, the entire set of 2000 AIMs can potentially be used to infer local ancestry. Note that accurate local ancestry inference in complex multi-way admixed populations such as the SAC, which has more than three source populations, is currently an unsolved problem. Whilst existing methods may achieve good accuracy on average, inference at particular regions, e.g. regions where the modeled and true ancestral populations differ due to selection, is still problematic.

We have used ancestry proportions estimated using genome-wide data as our gold standard against which to compare proportions estimated using AIMs. However, genome-wide estimated

proportions are by no means perfect. Accuracy will vary depending on the choice and number of source populations used. We have therefore taken care to select the best source populations for which genome-wide data is available while taking into account that sample sizes should be reasonable.

Excluding SNPs based on heterogeneity between subgroups of a source population, for example excluding SNPs that are heterogeneous in the three different Khoe-San groups, results in the exclusion of SNPs that can also distinguish source populations. This feature was introduced by Galanter et al. to ensure that their panel of AIMs can be applied to diverse American admixed populations, which may have received genetic contributions from different Native American populations [116]. Since this scenario does not apply to the SAC, and using this criterion results in a lower overall correlation between ancestry proportions estimated using AIMs and proportions estimated using genome-wide data, we ignored heterogeneity between subgroups in our final selection of AIMs.

The ability of the AIMs to distinguish South Asian and East Asian ancestries is markedly lower compared to the African San, African non-San and European ancestries. This could potentially be explained if the groups used as proxies for the South and East Asian source populations are not ideal representations of these ancestries in the SAC, although we have attempted to use the best reference groups for which genome-wide data were available. In addition, the genetic distance between South Asians and Europeans is relatively small compared to the genetic distance between other pairs of populations, and it is therefore more difficult to distinguish. Alternatively, the lower correlation of the Asian ancestries could be ascribed to the small proportions observed in our study groups. In the Galanter et. al. study, ancestry estimates for source populations that contributed less to the admixed population also had a relatively low correlation [116]. Due to these reasons, a much larger panel of AIMs would be required to improve the ability to distinguish the Asian ancestries. As the genetic contribution of the Asian ancestries to the SAC is relatively small, and because South Asians and Europeans are genetically similar, confounding due to the Asian ancestries are likely to be trivial in association studies. The list of AIMs presented in our study does state which source population pair each marker has been selected for. Markers selected for pairs that include the Asian ancestries can therefore easily be excluded, especially when a small panel is required. It is however our opinion that it is important to consider the Asian ancestries, since ignoring them would result in a less accurate overall estimation of ancestry.

The AIMs were selected from a set of markers that were successfully genotyped on the Affymetrix 500K chip for the admixed Cape Town study group, and that overlapped with source population data sets used in this study. The source population data sets were genotyped on a number of different microarray chips, including Illumina chips. It is therefore likely that the markers will also be genotyped successfully by other technologies, such as custom designed genotyping chips, the BeadXpress system, Sequenom plexes and Taqman assays.

According to the 2011 South African census, the majority of individuals who self-identify as South African Coloured reside in the Western Cape province [143]. The Cape Town study group of admixed individuals, recruited from the suburbs of Ravensmead and Uitsig in the Western Cape and who self-identified as South African Coloured, was used to assess the accuracy of the AIMs panel. We therefore believe that our panel of AIMs is applicable to the majority of individuals constituting this population group. We have also shown that the AIMs perform well for other Coloured groups residing in the Western Cape and the Northern Cape. These groups may be genetically distinct from one another due to genetic drift and different dates and levels of admixture between the different source populations. Since we have shown that the AIMs can distinguish the ancestries of the different admixed groups, the panel can also be used to correct for stratification when a study group has not been sampled from a relatively

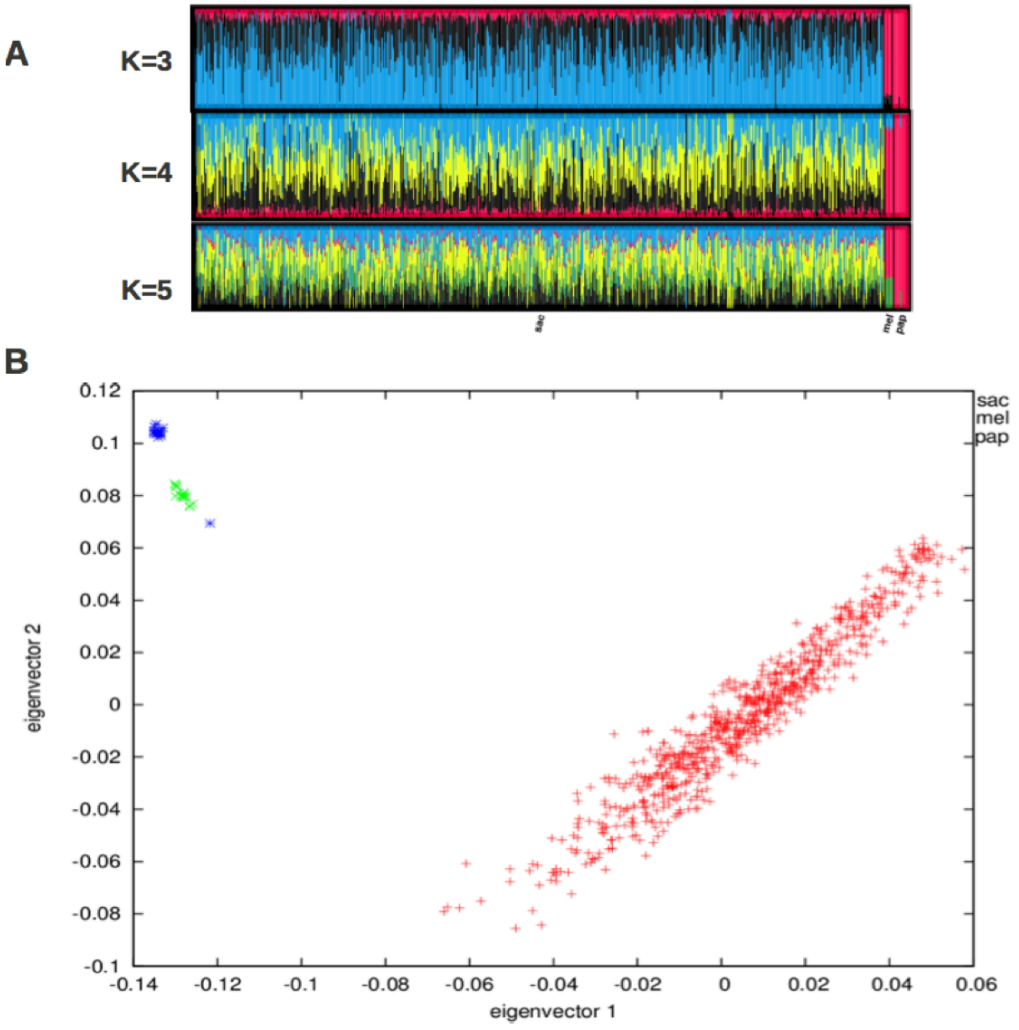
homogeneous admixed population. This is important as recent migration might introduce additional unknown heterogeneity into communities. It remains to be seen how well the AIMs perform in other Southern African mixed ancestry groups, such as the Cape Malay, a group which may have retained some distinction from the general South African Coloured population, groups living in the Eastern Cape and the Basters who reside mainly in Namibia. We have not been able to assess the accuracy of the panel for such groups due to the lack of availability of genome-wide data. It is, however, likely that the AIMs will also be applicable to these groups, since they were formed from the same source populations, or subsets of the same source populations. Consequently, the cost of studies regarding the overall genetic make-up of other Coloured groups can be much reduced. Based on our recent experience in Southern Africa, genotyping 120 AIMs were five times more cost-effective using Sequenom plexes compared to the most cost-efficient micro-array chips, which is particularly relevant when sample sizes are large. This is especially important in the light of limited access to research funding in Southern Africa. Although the cost of micro-array genotyping continues to decline, this also holds true for platforms designed for smaller marker sets, making it difficult to speculate on when the cost reduction will become a moot point.

In summary, we have developed a panel of 96 AIMs that is tailored to the complex five-way admixture that occurred in the South African Coloured population. This panel can be used as a cost effective alternative to genome-wide data for reducing false positive findings resulting from ignoring admixture in genetic association studies of the population.

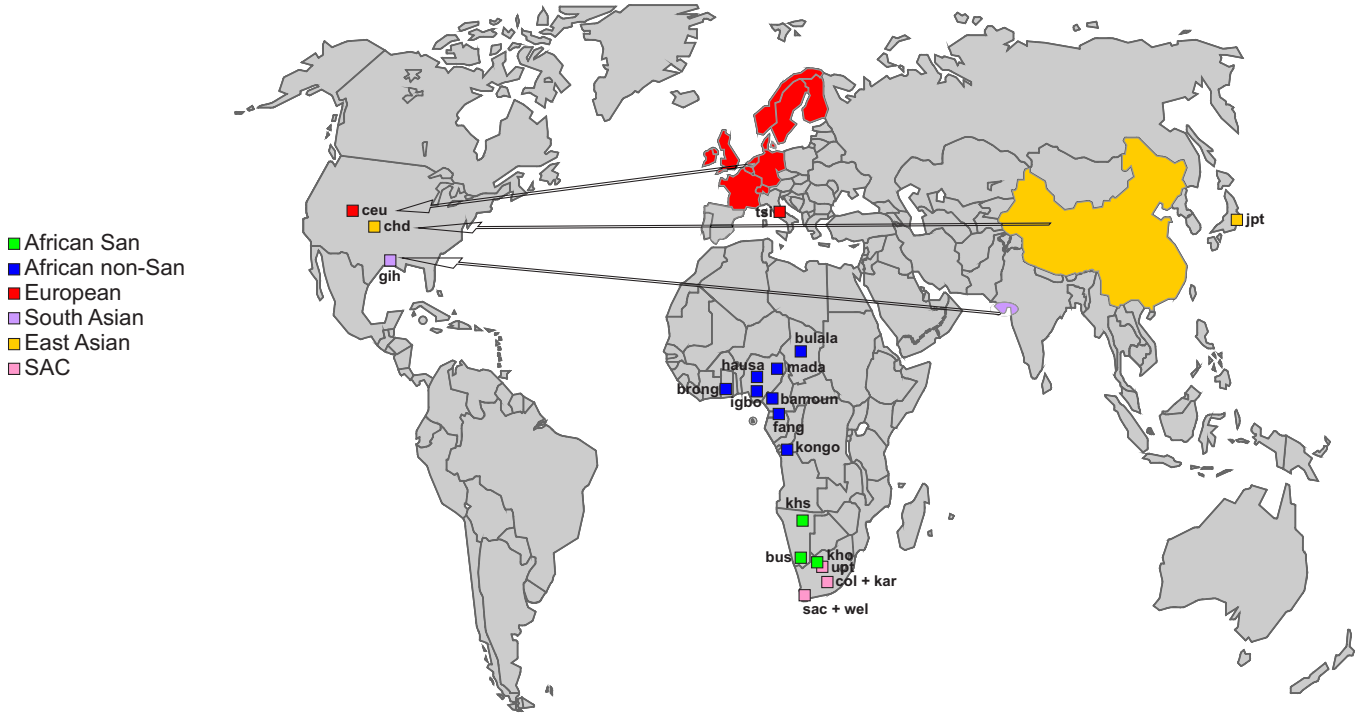
Acknowledgments

We thank all participants and field workers in this study.

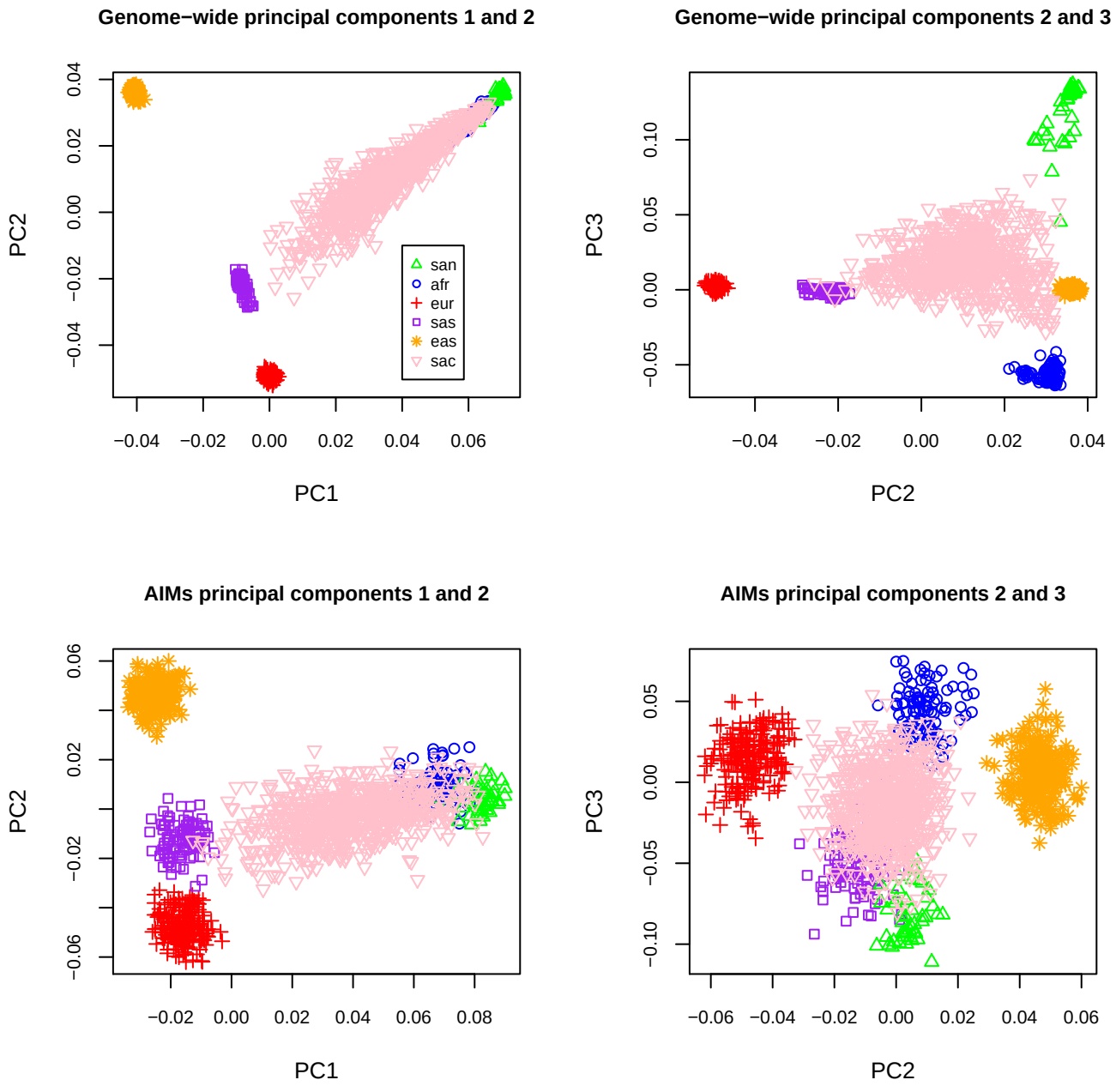
3.6 Supplementary figures and tables



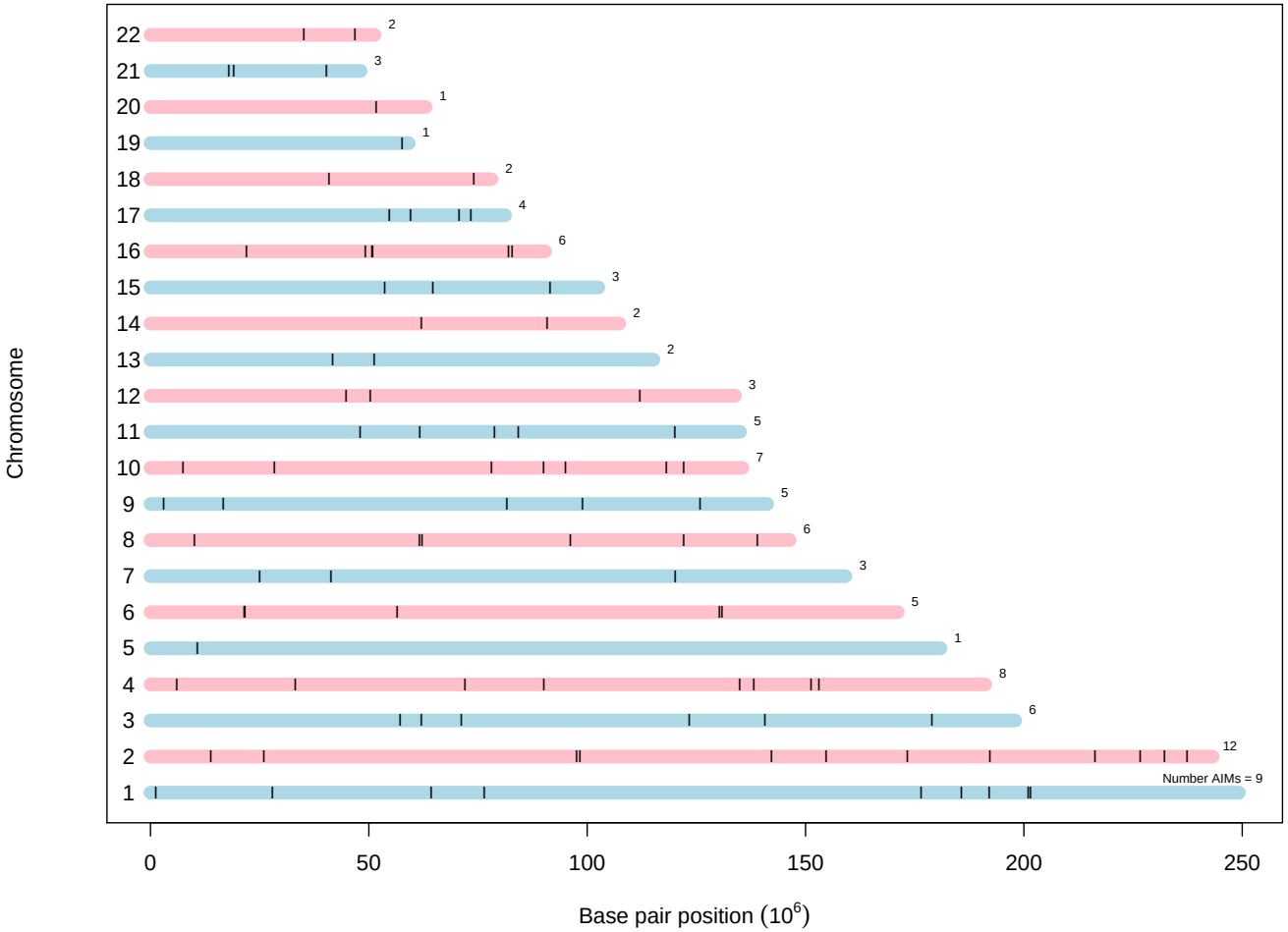
Supplementary Figure 3.6.1: Ancestry proportion and principal component analysis (PCA) of the SAC and the Oceania HGDP populations. (A) The proportion of each individual's ancestry. (B) The first and second eigenvectors of the PCA of the combined populations.



Supplementary Figure 3.6.2: World map with source and admixed populations. Abbreviations used for the source populations correspond to table 3.1. The admixed populations are indicated as follows: Cape Town = cpt, Colesberg = col, Karretjie = kar, Wellington = wel, Upington = upt. The *ceu*, *chd* and *gih* HapMap populations received ancestry from continents that differ from their sampling locations. Their approximate area of origin is in solid colour, with migration shown by arrows.



Supplementary Figure 3.6.3: Principal components formed using genome-wide data and AIMs. The first two panels show principal components 1 and 2 and 2 and 3 respectively, inferred from the source population genome-wide data. Similarly, panels 3 and 4 shows principal components inferred from 96 AIMs. Each data point represents the score of an individual for a principal component. The legend shows which source population each individual belongs to.



Supplementary Figure 3.6.4: Base pair position of AIMs per chromosome. The figure shows the position in number of base pairs of each of the 96 AIMs per chromosome.

Supplementary Table 3.6.1: Proxy ancestry scores. The results of the PROXYANC algorithm ordered by the magnitude of the score, per source population.

Candidate proxy	Score	Description	Source
		African San	
kho	163	Khoe-San South Africa	Henn 2011
bus	156	Khoe-San South Namibia	Henn 2011
khs	127	Khoe-San Namibia	Private[a]
		African non-San	
brong	899	Ghana	Henn 2011
kongo	809	Atlantic coast of Congo	Henn 2011
igbo	807	South Eastern Nigeria	Henn 2011
fang	668	Equatorial Bantu	Henn 2011
bulala	565	Central Chad	Henn 2011
mada	482	Cameroon	Henn 2011
hausa	449	West Africa	Henn 2011
bamoun	438	Cameroon	Henn 2011
yri	186	Yoruba in Ibadan	HapMap3
yor	160	Yoruba in Ibadan	Henn 2011
fulani	118	West-central Africa	Henn 2011
		European	
ceu	252	Northern European	HapMap3
tsi	198	Italy	HGDP
fre	165	French-france	HGDP
bas	162	Basque-France	HGDP
rus	154	Russian-russia	HGDP
sar	152	Sardinian-Italy	HGDP
		South Asian	
gih	191	Gujarati Indians	HapMap3
han	159	Pathan-Pakistan	HGDP
		East Asian	
chd	226	Chinese in Denver	HapMap3
chb	205	Han Chinese in Beijing	HapMap3
jpt	191	Japanese in Tokyo	HapMap3
jap	159	Japanese-Japan	HGDP
han	159	Pathan-Pakistan	HGDP
mia	153	Miao-China	HGDP
she	149	She-China	HGDP
dai	147	Dai-China	HGDP

[a] Private data access committee

Supplementary Table 3.6.2: The number of markers used for genome-wide ancestry proportion estimation per admixed study group. After the set of SNPs that overlap with all the source population data sets were found, a LD filter was applied to each admixed study group, using a window size of 50 SNPs and a shift size of 10 SNPs. Only the remaining SNPs were used for ancestry proportion estimation.

Study group	Number overlapping markers	Number remaining markers	r^2 threshold
Cape Town (n=733)	50 286	33 125	0.1
Colesberg (n=20)	29 914	14 662	0.3
Karretjie (n=20)	29 914	13 883	0.3
Wellington (n=20)	29 914	15 277	0.3
Upington (n=21)	30 466	16 195	0.3

Supplementary Table 3.6.3: 2000 AIMs. The top 2000 markers selected by our algorithm as AIMs for the South African Coloured population are found in *table_ s3.xls* at `http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0082224`. The table presents information on the marker location, allele frequency and population pair that a marker was selected for. The list is ordered according to marker selection, i.e. the panel of 96 AIMs evaluated are the first 96 markers in the table.

Supplementary Table 3.6.4: Number markers selected per source population pair. The number of markers selected per pair of source populations, for the panels of 96 and 120 AIMs. The number of markers selected are inversely proportional to the genetic distance between the populations that constitute the pair, as measured by Fst.

Population pair	96 panel	120 panel	Fst[a]
African San + African non-San	12	15	0.0841
African San + European	6	7	0.1951
African San + South Asian	6	7	0.1901
African San + East Asian	5	7	0.2331
African non-San + East Asian	6	8	0.1749
African non-San + European	8	9	0.1395
African non-San + South Asian	8	10	0.1297
European + South Asian	25	32	0.0338
European + East Asian	9	11	0.1090
South Asian + East Asian	11	14	0.0760

[a] Calculated using genome-wide data listed in table 1 and the R hierfstat package

Supplementary Table 3.6.5: Correlation obtained by Galanter et al. Correlation between ancestry proportions estimated using 88, 194 and 314 AIMs and proportions estimated using genome-wide data, for two of the admixed study groups in the Galanter et al. study.

		Correlation		
Study group	Nr AIMs	American	European	African
Mexico City Mexicans	314	0.985	0.980	0.748
	194	0.975	0.970	0.693
	88	0.960	0.943	0.592
GALA Puerto Ricans	314	0.735	0.943	0.959
	194	0.656	0.922	0.944
	88	0.520	0.849	0.877

Supplementary Table 3.6.6: Correlation obtained in the Cape Town study group for comparision to the Galanter et al. study. Correlation between ancestry proportions estimated using 88, 194 and 314 AIMs and proportions estimated using genome-wide data, for a 5-way and 3-way admixture model. Correlations for AIM sets of sizes 500 and 2000 is also given for the 5-way admixture model.

		Correlation				
Model	**Nr AIMs**	**African San**	**African non-San**	**European**	**South Asian**	**East Asian**
SAC 5-way admixture	2000	0.977	0.979	0.975	0.884	0.893
	500	0.938	0.939	0.922	0.717	0.730
	314	0.902	0.909	0.892	0.637	0.675
	194	0.845	0.866	0.852	0.563	0.631
	88	0.751	0.779	0.785	0.467	0.563
SAC 3-way admixture	314	0.937	0.944	0.975	-	-
	194	0.916	0.921	0.961	-	-
	88	0.858	0.863	0.930	-	-

Chapter 4

Quality control of the AIMs genotyping

Additional information on the quality control applied to the AIMs genotyping as well as issues that were considered in estimating ancestry proportions in a combined data set, genotyped on both the Sequenom and Affymetrix platforms, are described in this chapter.

4.1 Quality control of AIMs genotyping

This section describes quality control steps performed for the AIMs that were genotyped using the Sequenom platform and a Taqman assay. Sample and SNP quality was first evaluated, after which some samples and SNPs were removed from the data set.

The majority of the AIMs were genotyped using the Sequenom platform, therefore the data set is referred to as the Sequenom data set. The first 120 SNPs listed in the AIM set published by Daya et. al [99] were genotyped in 919 samples. For one of these samples, genotyping failed for all of the SNPs (genotyping failure refers to the situation where a SNP was not successfully called, i.e. a "missing" SNP). One of the SNPs, rs37268, failed genotyping in all samples. This sample and SNP were excluded from the data set prior to analysis. The initial Sequenom data set therefore comprised 918 samples and 119 SNPs.

The genome-wide Affymetrix data set, described by Chimusa et al. [98], contains genotypes for the same AIM set of 119 SNPs. To evaluate the concordance between samples and SNPs genotyped on the two platforms, some of the Affymetrix data set samples were re-genotyped using the Sequenom platform. Sample quality was further evaluated by summarizing the number of SNPs that failed genotyping in each sample. SNP quality was further evaluated by investigating the number of genotyping failures per SNP, as well as Hardy-Weinberg equilibrium (HWE).

4.1.1 Evaluation of sample quality

4.1.1.1 Missing genotypes per sample

When a SNP is not successfully called for a sample, its genotype is labeled as missing. A sample with a large number of missing SNPs may be indicative of low DNA quality for that sample. Depending on the extent of the genotyping failures, such samples may need to be excluded from the data set. Table 4.1 summarizes the number of missing genotypes found in the 918 samples, for 119 SNPs. More than half of the samples had zero missing genotypes. At least 100 SNPs were genotyped successfully in 825 of the samples. Most samples therefore had relatively few missing genotypes. However, 26 of the samples had 20 or less genotyped SNPs.

Table 4.1: Summary of the number of missing genotypes observed in samples, for the Sequenom data set (n=918 samples). The cumulative number of samples is shown in the third column.

Nr missing genotypes	Nr samples	Total
0	512	
1-10	311	823
11-20	2	825
21-30	4	829
31-40	56	885
41-50	1	886
51-60	0	886
61-70	6	892
71-80	0	892
81-90	0	892
91-100	0	892
101-110	1	893
111-119	25	918

Table 4.2: Summary of the number of genotype mismatches between the Sequenom and Affymetrix data sets observed in samples (n=315 samples). The cumulative number of samples is shown in the third column.

Nr mismathces	Nr samples	Total
0	177	
1	111	288
2	24	312
3	3	315

4.1.1.2 Concordance with Affymetrix data

The Affymetrix data set and Sequenom data set had 316 samples in common. There were 31 strand orientation mismatches in the set of 119 AIMs; the Affymetrix genotypes were flipped where necessary to match the Sequenom data. The number of mismatches was counted for each sample: if a SNP was successfully called (genotyped) on both platforms and the genotypes did not correspond, this was counted as a mismatch. The error rate per sample was calculated as the number of mismatches divided by the number of SNPs that were genotyped on both platforms, for that sample.

One of the samples had 73 mismatches in genotype. The most likely explanation for this large number of discrepancies is mislabeling of the physical sample. The mean error rate before and after removal of this sample was 0.0066 and 0.0047 respectively. After removal of the sample, the maximum number of errors that were found in a sample was 3, for 3 of the samples. Table 4.2 summarizes the number of samples with 0, 1, 2 and 3 errors after removal of the problematic sample (leaving 315 samples). The table shows that there is a high concordance between the two genotyping platforms.

4.1.2 Evaluation of SNP calling accuracy

4.1.2.1 Missing genotypes per SNP

The number of genotypes that were not called successfully for a SNP is an indication of the difficulty of genotyping that SNP. The genotyping accuracy of a SNP with a large number of

Table 4.3: Summary of the number of missing genotypes observed in SNPs, for the Sequenom data set (n=918 samples). The cumulative number of SNPs is shown in the third column.

Nr missing genotypes	Nr SNPs	Total
0	0	
1-10	0	0
11-20	2	2
21-30	13	15
31-40	55	70
41-50	14	84
51-60	3	87
61-70	2	89
71-80	15	104
81-90	6	110
91-100	3	113
101-150	6	119
151-918	0	119

Table 4.4: Summary of the number of missing genotypes observed in SNPs, for the Sequenom data set, after removing samples with 20 or less SNPs (n=825 samples). The cumulative number of SNPs is shown in the third column.

Nr missing genotypes	Nr SNPs	Total
0	35	
1-10	67	102
11-20	10	112
21-30	3	115
31-40	2	117
41-50	0	117
51-60	1	118
61-70	0	118
71-80	1	119
81-825	0	119

missing genotypes may therefore be lower compared to SNPs with a high call rate.

Table 4.3 summarizes the number of missing genotypes for the 119 SNPs when using all the samples (n=918 samples). As low quality samples may inflate the number of missing genotypes found in a SNP (i.e. the missing rate is due to low quality samples rather than the difficulty of calling the SNP), table 4.4 shows the number of missing genotypes for the 119 SNPs when excluding samples with more than 20 missing genotypes (n=825 samples). In this reduced data set, 35 of the SNPs were successfully genotyped in all samples, and 115 SNPs were successfully genotyped in at least 795 samples.

4.1.2.2 Concordance with Affymetrix data

The number of mismatches in genotype between the Affymetrix and Sequenom data sets was calculated for each SNP after removal of the sample with 73 mismatches in genotype between the Affymetrix and Sequenom data sets (leaving n=315 samples, see 4.1.1.2). The error rate per SNP was calculated as the number of mismatches divided by the number of samples genotyped successfully on both platforms for that SNP. The mean error rate is 0.0047. Table 4.5 summarizes the number of mismatches found in SNPs. 84 SNPs had 0 mismatches in genotype and 107 SNPs had 2 or less mismatches in genotype.

Table 4.5: Summary of the number of genotype mismatches between the Sequenom and Affymetrix data sets observed in SNPs (n=315 samples). The cumulative number of SNPs is shown in the third column.

Nr mismathces	Nr SNPs	Total
0	84	84
1	18	102
2	5	107
3	4	111
4	1	112
5	3	115
7	1	116
10	1	117
13	1	118
79	1	119

4.1.2.3 HWE

A SNP is said to be in HWE if the SNP's genotype frequencies indicate that the alleles that constitute the genotype are statistically independent. In a population where random mating occurs, SNPs are expected to be in HWE. HWE is usually tested in controls and not cases (cases may be genetically predisposed to the disease, and the assumption regarding random mating may therefore not hold). If a SNP fails a statistical test for HWE, this may indicate that genotypes have been incorrectly called in a relatively large number of samples.

After excluding 434 TB cases, HWE were tested in 484 control samples. 13 of the 119 SNPs failed the HWE test at an alpha level of 0.05. 6 of these SNPs failed the test at an alpha level of 0.01.

4.1.3 Removing low quality samples and SNPs from the data set

After evaluating the data as described above, we proceeded to remove low quality samples and SNPs from the data set. The resultant data set will be used to estimate admixture proportions for each of the individuals in the data set, in order to incorporate ancestry in statistical models. Since some correct information is more informative than no information, and our data evaluation were indicative of high genotyping accuracy, we took care to not remove samples and SNPs unnecessarily.

4.1.3.1 Sample removal

The sample that had a large number of mismatches between the Affymetrix and Sequenom data sets was removed from the data set. 26 samples for which only 20 or less SNPs were genotyped were also removed. The resultant data set was therefore composed of 891 samples.

4.1.3.2 SNP removal based on combined evidence of inaccuracy

A large proportion of missing genotypes, low concordance between the Affymetrix and Sequenom data sets and small HWE test p-values were used in conjunction to identify SNPs for removal from the data set. Table 4.6 presents information on SNPs that may have a low level of genotyping accuracy by any one of the three criteria. The table also provides additional information on HWE p-values calculated using the Affymetrix data set, as departure from HWE in both the Sequenom data set and Affymetrix data set may imply that the departure is due to population admixture or stratification rather than genotyping error. This is plausible as

the markers were specifically selected to identify admixture. Departure from HWE is therefore more likely compared to randomly selected markers. Based on this decision table, 3 SNPs were removed from the data set, leaving 116 SNPs.

4.1.3.3 Assessment of quality thresholds

To determine whether higher quality thresholds should be used for removing samples with a large proportion of missing genotypes, and whether additional questionable SNPs should be removed, we assessed the correlation between ancestry proportions estimated using the Affymetrix genome-wide data and the Sequenom AIMs, for samples found in both data sets. Ancestry proportions were estimated and correlations were calculated using the methodology described by Daya et al [99]. Ancestry proportions were estimated using the entire unrelated Affymetrix study group (n=732) plus source populations, and the entire Sequenom study group (n=891) plus source populations. As only unrelated individuals were used in the Affymetrix data set, the number of overlapping samples that were available for measuring correlation was reduced from 315 to 262. The correlation when using the thresholds described previously (with a resultant data set of 891 samples and 116 SNPs), when removing more samples with large proportions of missing genotypes, and when removing 3 additional questionable SNPs are shown in Table 4.7. The table shows that the correlation is effectively unchanged, which means that removing additional samples and SNPs that is potentially of low quality has a negligible effect on ancestry proportion estimation. This is most likely due to their potential negative effect in the maximum likelihood ancestry estimation being cancelled out by other high quality samples and SNPs. Additional samples and SNPs were therefore not removed from the Sequenom data set.

Table 4.6: Decision table for removal of SNPs with low-quality genotyping from the Sequenom data set. SNPs that had 3 or more mismatches between the Affymetrix and Sequenom data sets, more than 5% missing genotypes (after excluding samples with more than 20 missing genotypes) or HWE p-values smaller than 0.05 in Sequenom controls, are presented in this table.

			HWE P-value (Controls)	
SNP	**Proportion missing**	**Nr mismatches**	**Sequenom**	**Affymetrix**
	n=825	n=315	n=484	n=91
rs10127540	0.0024	5	0.3430	0.4638
rs10242455	0.0000	0	0.0074	0.0792
rs10493578*	0.0218	79	0.0000	0.7293
rs11197672	0.0061	0	0.0013	0.2827
rs1337775	0.0000	0	0.0057	1.0000
rs1468920	0.0036	3	0.6881	0.3552
rs1544396	0.0000	5	0.6883	1.0000
rs1545805	0.0012	0	0.0036	0.5268
rs16838138	0.0012	1	0.0274	0.1320
rs1689467*	0.0024	13	0.0463	1.0000
rs1800007	0.0194	0	0.0258	0.8045
rs2294654*	0.0424	10	0.0012	0.2385
rs2554832	0.0024	0	0.0292	1.0000
rs2579785	0.0073	3	0.2613	0.5102
rs2849266**	0.0012	4	0.0429	0.1967
rs4841295	0.0000	3	0.8072	0.2219
rs6607302**	0.0024	7	0.0652	0.8142
rs7165405	0.0073	1	0.0160	0.0355
rs7244148**	0.0727	3	0.0625	1.0000
rs7584977	0.0230	5	0.6403	0.6748
rs7601	0.0012	0	0.0259	0.3596

∗ Removed

∗∗ Questionable

Table 4.7: Correlation between ancestry proportions of samples genotyped on both platforms. Ancestry proportions were estimated using the entire unrelated Affymetrix study group (n=732) plus source populations, and the entire Sequenom study group (n=891) plus source populations. The correlation is between ancestry proportions estimated using the Affymetrix genome-wide data and proportions estimated using the Sequenom AIMs for samples that are contained in both data sets.

	Correlation (Difference)				
	African San	**African non-San**	**European**	**South Asian**	**East Asian**
All (n=262)	0.816	0.818	0.812	0.551	0.573
Samples excluded* (n=249)	0.823 (+0.007)	0.827 (+0.009)	0.814 (+0.002)	0.552 (+0.001)	0.585 (+0.012)
SNPs excluded** (n=262)	0.818 (+0.002)	0.821 (+0.003)	0.817 (+0.005)	0.538 (−0.013)	0.568 (−0.006)

∗ Samples with > 20 missing genotypes were excluded from the Sequenom data set before ancestry estimation

∗∗ Three additional questionable SNPs in table 4.6 were excluded from the Sequenom data set before ancestry estimation

4.1.4 Software

PLINK was used for strand flipping and HWE p-value and missing proportion calculations [105]. Other statistical analyses were done using the R programming environment [106].

4.2 Ancestry proportions in a combined data set

In order to adjust for admixture in TB association studies, ancestry proportions were estimated using AIMs that were genotyped either on the Sequenom or the Affymetrix platform. We highlight two issues that warrant consideration. Firstly, when combining the Affymetrix and Sequenom data sets, the resultant data set contains related individuals, but the multinomial model used to estimate ancestry proportions assumes independence, i.e. unrelated individuals. Secondly, a systematic difference between the Affymetrix versus Sequenom AIMs may bias results when ancestry proportions, estimated using AIMs genotyped by the two different platforms, are used in the same statistical model.

4.2.1 Combining the Sequenom and Affymetrix data sets

The single sample with high discordance between the data sets were excluded from the combined data set (see 4.1.1.2). For samples that were genotyped on both platforms, the Sequenom genotypes were used, since the Sequenom genotyping accuracy are believed to be higher than that of a micro-array platform. However, Affymetrix genotypes were used instead of the Sequenom genotypes when more than 5% of the Sequenom genotypes were missing.

All the individuals in the Sequenom data set are unrelated according to information in our database. Some individuals are however related in the combined data set. Additional cryptic relatedness were also discovered using the genome-wide Affymetrix data. Pairs of individuals that are cryptically related were identified by the proportion of their genome that they share IBS (identical by state). Pairs that have a kinship coefficient of 0.2 or higher was classified as related to each other. To satisfy the assumption of independent samples, ancestry proportions were therefore estimated in separate batches that contained only unrelated individuals.

4.2.2 Does the genotyping platform matter?

Bias due to a systematic difference between the Affymetrix versus Sequenom AIMs are unlikely due to the high concordance achieved between the two platforms (see 4.1.1.2). Since a larger proportion of cases were genotyped on the Affymetrix platform, this was formally evaluated by testing for association between platform and ancestry proportions.

Ancestry proportions were estimated for the 316 samples that were found in both the Affymetrix and Sequenom data sets, separately for each platform, using the methodology described by [99]. A mixed effects logistic regression model (generalized linear model with a binomial family and logit link) was used with the genotyping platform as outcome; African San, African non-San, European and South Asian ancestry proportions as fixed effects; and sample ID as random effect. P-values for the ancestries were 0.848, 0.883, 0.847 and 0.946 respectively, indicating that an association between the ancestry proportions and genotyping platform is unlikely. Therefore, when the ancestry proportions in a statistical model were estimated using AIMs genotyped by the two different platforms, an additional covariate for the genotyping platform is not required.

4.2.3 Software

IBS calculations were done using PLINK [105]. ADMIXTURE [129] was used to estimate ancestry proportions. The freely available R programming environment was used for statistical analysis [106]. Mixed-effect logistic regression models were fitted using the *glmer()* function found in the *lme4* R package [144].

Chapter 5

Research Article 2

The role of ancestry in TB susceptibility of an admixed South African population

Michelle Daya[1], Lize van der Merwe[1,2,3], Paul D. van Helden[1], Marlo Möller[1], Eileen Hoal[1,*]

1 Molecular Biology and Human Genetics, MRC Centre for Molecular and Cellular Biology and the DST/NRF Centre of Excellence for Biomedical TB Research, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa
∗ E-mail: Corresponding egvh@sun.ac.za
The collective term for people of mixed ancestry in southern Africa is "Coloured" and is recognized and used officially in South Africa. Whilst we acknowledge that in some cultures this term may have acquired a derogatory connotation, this is certainly not intended here.

This article was published in July 2014 in the journal Tuberculosis [100]. Section, figure, table and reference numbering presented here is different to the published version.

5.1 Abstract

Genetic susceptibility to tuberculosis (TB) has been well established and this, taken together with variation in susceptibility observed between different geographic and ethnic populations, implies that susceptibility to TB may in part be affected by ethnicity. In a previous genome-wide TB case-control study (642 cases and 91 controls) of the admixed South African Coloured population (SAC), we found a positive correlation between African San ancestry and TB susceptibility, and negative correlations with European and Asian ancestries. Since genome-wide data was available for only a small number of controls in the previous study, we endeavored to validate this finding by genotyping a panel of ancestry informative markers (AIMs) in additional individuals, yielding a data set of 918 cases and 507 controls. Ancestry proportions were estimated using the AIMs for each of the source populations of the SAC (African San, African non-San, European, South Asian and East Asian). Using logistic regression models to test for association between TB and ancestry, we confirmed the substantial effect of ancestry on TB susceptibility. We also investigated the effect of adjusting for ancestry in candidate gene TB association studies of the SAC. We report a polymorphism that is no longer significantly associated with TB after adjustment for ancestry, a polymorphism that is significantly associated with TB only after adjustment for ancestry, and a polymorphism where the association significance remains unchanged. By comparing the allele frequencies of these polymorphisms in the source populations of the SAC, we demonstrate that association results are likely to be affected by adjustment for ancestry if allele frequencies differ markedly in the source populations of the SAC.

5.2 Introduction

According to the World Health Organization (WHO), the highest burden of TB is carried by Asia and Africa [49] and in the USA, there is a marked contrast in incidence between ethnicities [145]. A large proportion of ethnic disparity in TB susceptibility can be attributed to socioeconomic factors [96; 97; 146] and the human immunodeficiency virus (HIV) epidemic [147]. The remaining difference, albeit small, could possibly be explained by genetic differences in susceptibility between population groups, since many investigations have shown that genetic factors are involved in the disease [66]. It is thought that certain population groups are more or less susceptible to TB infections, based on the history of their exposure to the disease and the development of resistance due to natural selection. Based on a large study of 165 racially integrated nursing homes in Arkansas (USA), Stead et al. showed that Europeans are less susceptible to TB infection compared to individuals of African ancestry [87]. This study was however limited due to its inability to control for all behavioral differences between the two groups. Nevertheless, the apparent higher resistance of Europeans to TB could possibly be explained by many centuries of exposure to the disease in densely populated European settlements [76].

The predominant population group in the Western Cape, South Africa, is the five-way admixed group (African San, African non-San, European, South Asian and East Asian) known as the South African Coloured (SAC) [26; 39]. Our SAC study participants are ideally suited to test if an association exists between ancestry and TB susceptibility, as they received genetic contributions from both African and European populations, who differ in TB rates, and come from the same high-TB communities with the same socioeconomic status (SES). In a genome-wide TB case-control study of the group (642 cases and 91 controls), Chimusa et al. [98] found a positive correlation between the proportion of African San ancestry and TB susceptibility, and negative correlations with European, South Asian and East Asian ancestries. Due to the small number of controls in the Chimusa et al. study, we endeavored to validate this finding by genotyping a panel of ancestry informative markers (AIMs), described by Daya et al. [99], in additional cases and controls. The selected AIMs were tailored to the SAC, as other panels of AIMs described in the literature did not adequately incorporate African San ancestry, one of the main ancestral components of the SAC. The complex five-way admixture that occurred in the SAC, with dissimilar genetic distances between source populations, was also not adequately modeled in other panels. Daya et al. evaluated various AIM selection strategies, and by comparing ancestry proportions estimated using different panels of AIMs to the gold standard of genome-wide estimated proportions, ensured that the selected panel of AIMs is best suited to ancestry inference in this population.

A previous TB susceptibility study of the SAC stated that no significant population stratification was found in the cohort [148]. This was based on the comparable allele frequency distributions between cases and controls of 25 randomly selected and uncorrelated SNP markers, implying that adjustment for ancestry is not necessary. However, as a result of the complex admixture that occurred in the SAC, a larger number of markers is ideally required to distinguish allele frequency differences that are due to ancestry. In addition, AIMs have greater power to discern ancestries compared to randomly selected markers [115]. Using the panel of AIMs, we therefore also investigate the effect of adjusting for ancestry in candidate gene association studies of susceptibility to TB in the SAC.

5.3 Materials and methods

5.3.1 Sample collection and ethical approval

A large study group of South African Coloured individuals was recruited from the Cape Town suburbs of Ravensmead and Uitsig (the same community recruited in the Chimusa et al. study). This district was selected due to its high TB incidence, uniform SES and low prevalence of HIV [148]. TB patients were identified through bacteriological confirmation (smear positive and/or culture positive). Healthy individuals that had no previous history of TB were selected as controls. The controls were living under the same conditions as TB patients, including SES status and availability of health facilities. HIV positive individuals were excluded from the study.

Approval for the study was obtained from the Ethics Committee of the Faculty of Health Sciences, Stellenbosch University (project registration numbers 95/072, NO6/07/132 and N11/07/210). Blood samples for DNA were collected with written informed consent. The research was conducted according to the principles expressed in the Declaration of Helsinki.

5.3.2 Genotyping, quality control and ancestry proportion estimation

Our sample bank comprises 955 case and 521 control samples, collected between 1994 and 2007. 969 samples were genotyped on the Affymetrix GeneChip Human Mapping 500K Array Set in 2008, 888 of which passed quality control criteria (of which a subset of 733 unrelated individuals were used by Chimusa et al. to test for association between TB susceptibility and ancestry) [98]. The sample bank was also used to perform a number of candidate gene studies between 2003 and 2013, summarized in supplementary table 5.6.1.

120 AIMs were selected to distinguish the source populations of the SAC [99] and genotyped in 918 samples. The genotyping was performed at the Institute for Clinical Molecular Biology at the Christian-Albrechts University in Kiel, Germany, using the Sequenom iPLEX platform (114 SNPs) and TaqMan assays (6 SNPs). 4 SNPs and 27 samples failed our quality control criteria and were removed from the data set. The remaining 116 AIMs were also available in 888 samples genotyped on the Affymetrix platform. 316 samples overlapped between the Affymetrix and Sequenom data sets, and were used to assess the concordance between the platforms. One of these samples had 73 discordant genotypes between the two platforms, which were most likely due to mislabeling of the physical sample. After excluding this sample, the mean proportion of discordant genotypes was 0.0047. In total, AIMs were available for 1425 samples (918 Sequenom samples + 888 Affymetrix samples - 316 overlapping samples - 27 samples that failed quality control - 38 samples with either missing sex or age information).

Ancestry proportions were estimated for the combined Sequenom and Affymetrix AIM data set, jointly for each of the five source populations of the SAC, using the methodology and source populations described by Daya et al [99]. As some of the samples were collected from families, the combined data set contained individuals that were related to one another. Since the multinomial model used to estimate ancestry proportions assumes independence, ancestry proportions were estimated in separate batches, each batch comprised of unrelated individuals.

We re-examined a number of previous candidate gene studies performed by our research group (236 SNPs of 78 genes genotyped in 11 published and 18 unpublished studies), now adjusting for ancestry. An unrelated subset of individuals from the combined data set described above was used in each of these studies. We report the genotype model results of rs2243639, rs2569190 and rs34069356 in order to illustrate the possible effects of ancestry adjustment.

These polymorphisms were genotyped using a Taqman assay, a PCR-RFLP method [149], and the SNPlex Genotyping System [141], respectively.

5.3.3 Statistical analysis

Mixed-effects logistic regression models (generalized linear models with a binomial family and logit link) were used to test for association between TB susceptibility and ancestry. A reduced data set that excluded samples used in the Chimusa et al. [98] study (n=696) was modeled, as well as a complete data set of all samples for which AIMs are available (n=1425). Models were fitted for the individual effect of each of the source ancestries of the SAC, as well as as a combined model that included all the ancestries. Age and sex were adjusted for by including them in the models as fixed effects. A family identifier was specified as a random effect to adjust for relatedness between groups of individuals (251 of the 1425 individuals could be grouped into 101 families).

Logistic regression models were used to test for association between TB susceptibility and genotype, adjusting for age and sex, and then adjusting for age, sex and ancestry. (Mixed-effects models were not used, since the individuals were unrelated.) The possibility of genotyping errors were assessed by evaluating Hardy-Weinberg equilibrium (HWE) in controls (exact test).

In this study, associations corresponding to a p-value of 0.05 or less were considered significant. The Bonferroni correction for multiple testing was not used as it may be over-conservative when several genetic associations are tested in the same group [150]. Most multiple testing correction methods may be unsuitable when there is a priori evidence that genes are associated with a phenotype [151; 152]. In addition, the reported genotypic association tests are intended to illustrate the confounding effect that ancestry may have on TB genetic susceptibility studies in the SAC, rather than quantifying the strength of the relationship between specific loci and TB susceptibility.

Genotype frequencies in individuals with low, medium and high African San ancestry were graphically contrasted between cases and controls. The low San ancestry group represented those individuals that fell within the first quartile of San ancestry. Similarly, the high San ancestry group fell within the fourth quartile of San ancestry. The medium San ancestry group fell within the second and third quartiles.

5.3.4 Software

The freely available R programming environment (from www.r-project.org) was used for statistical analysis [106]. Mixed-effects logistic regression models were fitted using the *glmer()* function from the *lme4* R package [144]. HWE was evaluated using the *HWE.exact()* function from the *genetics* R package [153; 154]. ADMIXTURE [129] was used to estimate ancestry proportions.

5.4 Results

5.4.1 Association between TB susceptibility and ancestry

Figure 5.1 is a box plot of the ancestry proportions in TB cases and controls of the five source populations of the SAC (n=1425). The figure shows that African ancestry, especially San ancestry, is higher in cases compared to controls. European and Asian ancestries are lower in cases compared to controls.

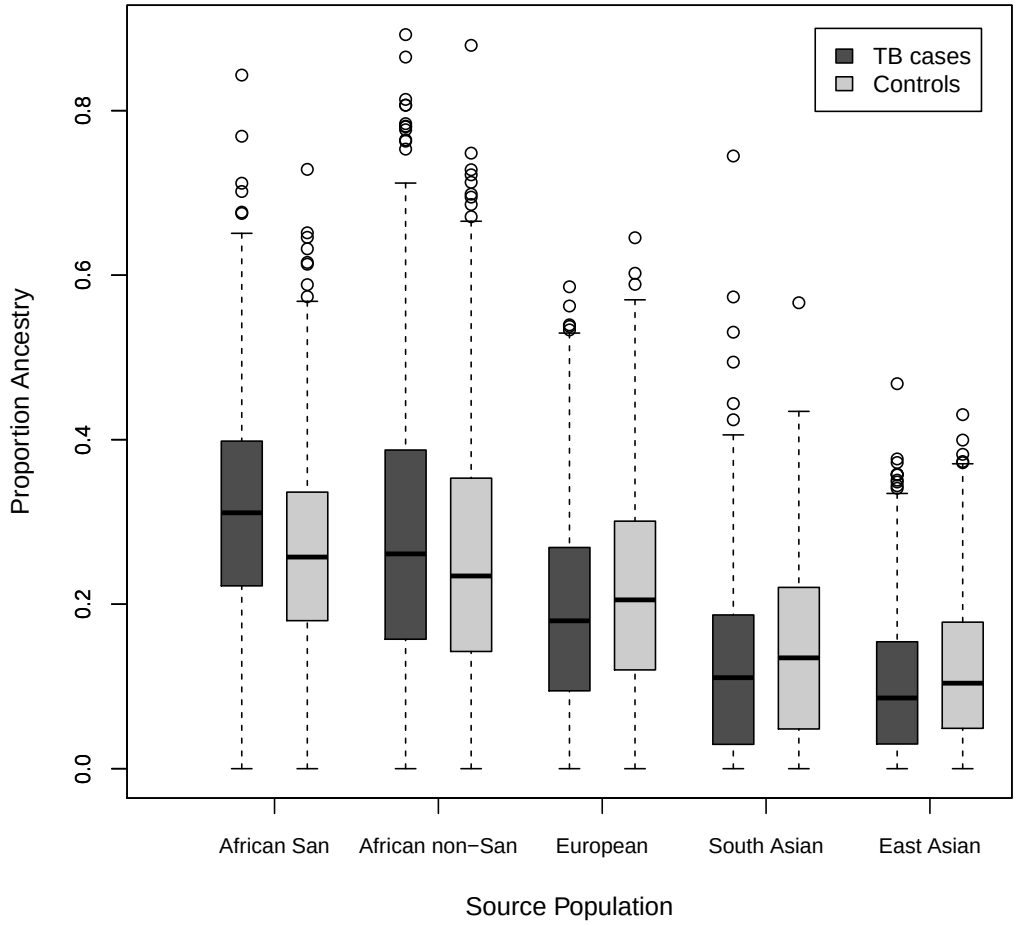


Figure 5.1: Box plot of ancestry proportions. Ancestry proportions of the five source populations of the SAC in cases (n=918) and controls (n=507).

Tables 5.1 and 5.2 summarize the age, sex, ancestry proportions and results of the TB susceptibility models for the complete data set (n=1425) and for a reduced data set that excludes samples used by Chimusa et al (n=696). All of the five source ancestries are individually associated with TB in the complete data set. African San and African non-San ancestry are associated with increased susceptibility to TB (p-value=1.06×10^{-11} and p-value=3.00×10^{-2} respectively), while European, South Asian and East Asian ancestries are protective (p-value=2.49×10^{-6}, p-value=1.15×10^{-4} and p-value=8.42×10^{-4} respectively). African San ancestry is the most significant association: Each 10% increase of African San ancestry increases the odds of having TB by 38%. In the reduced data set, the direction of association is the same compared to the complete data set, but only the African San and European associations are significant (p-value=6.25×10^{-5} and p-value=8.56×10^{-3} respectively). This is likely due to the smaller sample size of the reduced data set. The direction of association found in both data sets is the same as in the Chimusa et al. study. In the complete data set, a higher level of significance was attained. The higher level of significance is likely due to the larger sample size of the complete data set compared to the Chimusa et al. study. Contrary to the Chimusa et al. study, African non-San ancestry was also found to be significantly associated with TB, albeit less so compared to the other ancestries (p-value=3.00×10^{-2} in this study, p-value=1.09×10^{-1} in the Chimusa et al. study).

Table 5.1: Age, sex, ancestry proportions and TB susceptibility modeling results for the complete data set. Each of the source population ancestries of the SAC were tested individually for association with TB, adjusting for age and sex (individual models). All of the ancestries were also tested together in a combined model. The estimated odds ratio reflects the odds of having TB for each 0.1 increase in ancestry proportion.

	TB cases	Controls	P-value		OR [95% CI]	
	(n=918)	(n=507)	Individual	Combined	Individual	Combined
Nr males (freq)	493 (0.54)	122 (0.24)		<0.0001		
Age (mean ± SD)	36.00 ± 12.79	32.48 ± 10.69		<0.0001		
African San [IQR]	0.31 [0.22-0.40]	0.26 [0.18-0.34]	<0.0001	<0.0001	1.38 [1.26-1.52]	1.44 [1.24-1.68]
African non-San [IQR]	0.26 [0.16-0.39]	0.23 [0.14-0.35]	0.0300	0.0463	1.08 [1.01-1.16]	1.16 [1.00-1.33]
European [IQR]	0.18 [0.09-0.27]	0.21 [0.12-0.30]	<0.0001	0.8228	0.79 [0.72-0.87]	0.98 [0.82-1.17]
South Asian [IQR]	0.11 [0.03-0.19]	0.13 [0.05-0.22]	0.0001	0.7326	0.80 [0.71-0.90]	0.97 [0.82-1.15]
East Asian [IQR]	0.09 [0.03-0.15]	0.10 [0.05-0.18]	0.0008		0.80 [0.70-0.91]	

Table 5.2: Age, sex, ancestry proportions and TB susceptibility modeling results for the reduced data set. Each of the source population ancestries of the SAC were tested individually for association with TB, adjusting for age and sex (individual models). All of the ancestries were also tested together in a combined model. The estimated odds ratio reflects the odds of having TB for each 0.1 increase in ancestry proportion.

	TB cases	Controls	P-value		OR [95% CI]	
	(n=280)	(n=416)	Individual	Combined	Individual	Combined
Nr males (freq)	133 (0.48)	77 (0.19)		<0.0001		
Age (mean ± SD)	34.43 ± 15.28	32.7 ± 11.64		0.1079		
African San [IQR]	0.31 [0.20-0.39]	0.26 [0.18-0.33]	0.0001	0.0227	1.32 [1.15-1.51]	1.28 [1.04-1.59]
African non-San [IQR]	0.23 [0.14-0.35]	0.23 [0.15-0.34]	0.9860	0.9665	1.00 [0.90-1.11]	1.00 [0.82-1.23]
European [IQR]	0.20 [0.11-0.28]	0.20 [0.12-0.30]	0.0086	0.3627	0.83 [0.72-0.95]	0.89 [0.70-1.14]
South Asian [IQR]	0.12 [0.05-0.19]	0.13 [0.05-0.21]	0.4020	0.9958	0.93 [0.80-1.09]	1.00 [0.78-1.28]
East Asian [IQR]	0.09 [0.03-0.17]	0.10 [0.05-0.18]	0.2750		0.90 [0.75-1.09]	

Not all of the ancestries are independently associated with TB susceptibility in the combined models. Only the African San and African non-San associations reach significance in the complete data set (p-value=3.31×10^{-6} and p-value=4.63×10^{-2} respectively). In the reduced data set, only the African San association achieves significance (p-value=2.27×10^{-2}).

Our rationale for testing ancestries individually and in combination is as follows. All ancestries were tested together in a combined model to identify the ancestries that explained most of the variation in the data. An admixed individual's ancestry is by nature interdependent, as a larger proportion of ancestry from one source population means lower proportions of ancestry from another source population. Since African San ancestry is most strongly associated with TB susceptibility, and higher African ancestry implies lower European and Asian ancestries, the latter ancestries are no longer associated when African ancestries are also incorporated into the same model. This does not necessarily mean that European and Asian ancestries are not protective; rather, it shows that these ancestries are inversely proportional to African ancestry, and that their effect has already been encapsulated in the model by the African ancestries. To quantify the potential relationship between each ancestry and TB susceptibility, ancestries were therefore also tested individually in separate models.

Note that the reported p-values were not adjusted for multiple testing. A total of 12 ancestry association tests were performed (6 in the complete data set and 6 in the reduced data set). This would yield a Bonferroni significance threshold of 0.0042 (0.05 divided by 12), if one does not consider the genotype association tests that were also performed.

5.4.2 Adjusting for ancestry in candidate gene studies

We re-evaluated some of the previous candidate gene studies performed in our group, now adjusting for ancestry. At a critical significance level of 0.05, after adjusting for ancestry, 16 genotype associations were no longer significant, whilst 5 became significant. 16 and 199 genotype associations remained significant and non-significant, respectively. We report the estimates from three specific models to illustrate the possible effects of ancestry adjustment: a model where the estimated effect of the association increased markedly (negative confounding due to ancestry), another model where the estimate decreased (positive confounding), and a third model where the estimate remained the same.

Surfactant protein D (SP-D) is expressed by the *SFTPD* gene and has been shown to decrease the uptake and growth of *Mycobacterium tuberculosis* (*M. tuberculosis*) in macrophages [155]. One of the SNPs we report, rs2243639, is located in exon 5 of the *SFTPD* gene. We also report rs2569190, a promoter polymorphism that regulates the expression of the CD14 (name derived from the cluster of differentiation group of cell surface marker proteins) molecule. The CD14 molecule is expressed on the cell surface of macrophages, monocytes and granulocytes [156; 157]. Lastly we report rs34069356, located in the 3'UTR region of the cathepsin Z (*CTSZ*) gene. The *CTSZ* gene is expressed in macrophages and monocytes [141].

Table 5.3 summarizes results for rs2243639, rs2569190 and rs34069356. The *SFTPD* rs2243639 TT genotype is not significantly associated with TB susceptibility before adjusting for ancestry, but is significantly associated after adjustment. The opposite effect is observed for the *CD14* rs2569190 TT genotype. The *CTSZ* rs34069356 CT genotype association remains highly significant (p-value=1.25×10^{-7} and p-value=1.15×10^{-6} before and after adjustment for ancestry, respectively).

Table 5.3: The association between TB susceptibility and three different SNPs. The effect of adjusting versus not adjusting for ancestry is shown. The tests are first adjusted for age and sex only, and then adjusted for age, sex and ancestry. The estimated odds ratios reflect the odds of having TB when carrying the genotype, compared to the most common genotype.

	Controls		Cases	P-value		OR [95% CI]	
Gene (SNP)	**Count (Freq)**	**HWE[a]**	**Count (Freq)**	**Unadj[b]**	**Adj[c]**	**Unadj[b]**	**Adj[c]**
SFTPD (rs2243639)		0.6630		0.1797	0.0261		
Typed	342		338				
C/C	263 (0.77)		256 (0.76)			1.00	
C/T	74 (0.22)		72 (0.21)			0.92 [0.63-1.35]	1.13 [0.76-1.68]
T/T	5 (0.01)		10 (0.03)			2.62 [0.87-7.92]	4.75 [1.45-15.55]
C	600 (0.88)		584 (0.86)				
T	84 (0.12)		92 (0.14)				
CD14 (rs2569190)		0.6459		0.0380	0.2803		
Typed	321		341				
C/C	149 (0.46)		188 (0.55)			1.00	
C/T	137 (0.43)		131 (0.38)			0.74 [0.53-1.04]	0.85 [0.6-1.2]
T/T	35 (0.11)		22 (0.06)			0.51 [0.28-0.93]	0.63 [0.34-1.16]
C	435 (0.68)		507 (0.74)				
T	207 (0.32)		175 (0.26)				
CTSZ (rs34069356)		1.00		<0.0001	<0.0001		
Typed	265		313				
C/C	249 (0.94)		241 (0.77)			1.00	
C/T	16 (0.06)		72 (0.23)			4.38 [2.42-7.93]	4.53 [2.46-8.32]
C	514 (0.97)		554 (0.89)				
T	16 (0.03)		72 (0.11)				

[a] Hardy-Weinberg equilibrium p-value

[b] Not adjusted for ancestry (only adjusted for age and sex)

[c] Adjusted for ancestry (and adjusted for age and sex)

In an ideal case-control study, case and control groups should be similar to each other with respect to all factors that might be related to disease. In our study group, age, sex and ancestry differ significantly between cases and controls, which necessitates statistical adjustment for these factors. Figure 5.2 depicts changes in the estimated odds of having TB, when carrying a particular genotype compared to a reference genotype, as progressively more covariates are introduced into the logistic regression model. The estimated effect increased for the *SFTPD* rs2243639 TT versus CC genotype, while the effect decreased for the *CD14* rs2569190 TT versus CC genotype. The effect of the *CTSZ* rs34069356 CT versus CC genotype remained at the same level.

As an illustration of the conditions under which adjustment for ancestry may change the results of association tests, samples were stratified according to their proportion of African San ancestry, as this ancestry is most strongly associated with TB susceptibility. Figure 5.3 depicts the frequency of the reported variant genotypes in cases versus controls, for samples with low, medium and high proportions of African San ancestry. The figure reflects the pattern that results in the changes in odds ratio when adjusting for ancestry. Genotype frequency differences between cases and controls with medium proportions of African San ancestry exemplify the effect of the genotype on TB susceptibility when not adjusting for ancestry. The *SFTPD* rs2243639 TT genotype frequency does not appear to differ between cases and controls with medium proportions of African San ancestry, yet a clear difference is observed in both the low and high African San ancestry groups. The *CD14* rs2569190 TT genotype frequency difference between cases and controls with medium and high proportions of African San ancestry is almost non-existent in the low African San ancestry group, thereby reducing the apparent effect that is observed when not considering ancestry in statistical models. The *CTSZ* rs34069356 CT genotype frequency difference between cases and controls is large, with the highest frequency observed in cases, for each of the San ancestry stratifications.

The impact of adjusting for ancestry for the three SNPs can further be illustrated by comparing allele frequencies in the source populations of the SAC. This is depicted in Figure 5.4. Allele frequencies for the *SFTPD* and *CD14* SNPs differ markedly in the five source populations, whilst the *CTSZ* rs34069356 variant allele has a low frequency in all populations for which its allele frequency is known. The particular *SFTPD* and *CD14* allele inherited is therefore likely to be considerably affected by ancestry, whilst the same cannot be said of the *CTSZ* allele, due to its low minor allele frequency in all source populations.

5.5 Discussion

We have shown that African San ancestry increases susceptibility to TB by using a panel of AIMs to estimate the ancestry proportions of a study group of individuals from the South African Coloured population. Our study confirms the findings of Chimusa et al. [98], which used only a limited number of controls. African non-San ancestry may also increase TB susceptibility, although the association we found was relatively weak. European and Asian ancestry appears to be protective, but this result possibly reflects lower African ancestry, which would decrease susceptibility.

It has been postulated that *M. tuberculosis* originated in Africa and that the pathogen co-evolved with its hosts as humans migrated throughout the world [59; 62; 60]. Ancient lineages are thought to have spread by land, whilst modern lineages propagated via ocean travel during the colonization era [60]. In a previous study of *M. tuberculosis* strains found in the Ravensmead and Uitsig communities, 84% of the 1921 strains that were successfully classified were of East Asian and Euro-American lineage [158], both of which are modern [60]. Until recently, Southern

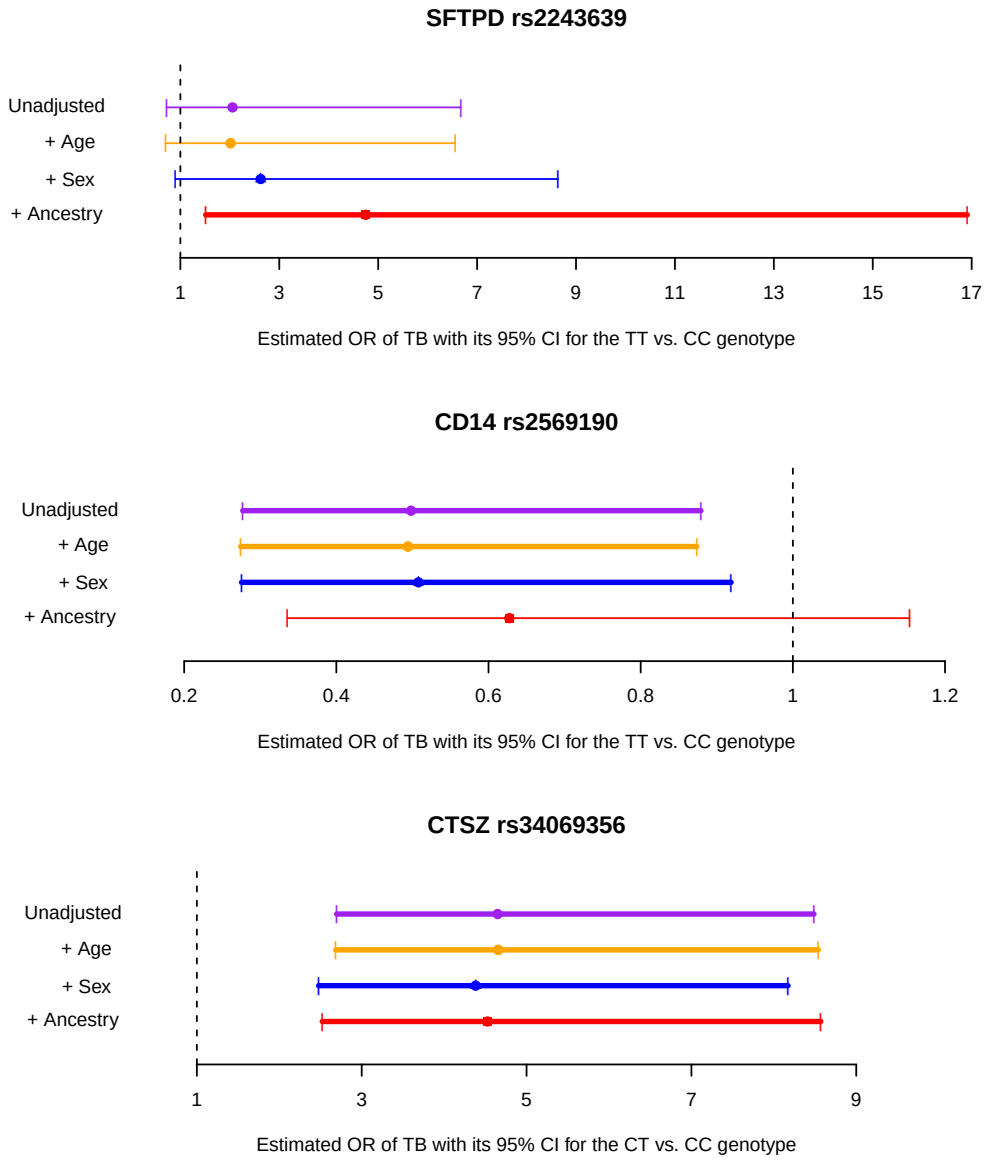


Figure 5.2: Comparison of TB case-control odds ratios with the addition of covariates. This figure depicts the estimated odds ratios of having TB, with their 95% confidence intervals, for selected genotypes. The odds ratios were estimated from models with no covariates, models including age, models including age and sex, and models including age, sex and ancestry. Confidence intervals that do not include one are statistically significant and are indicated by bold lines.

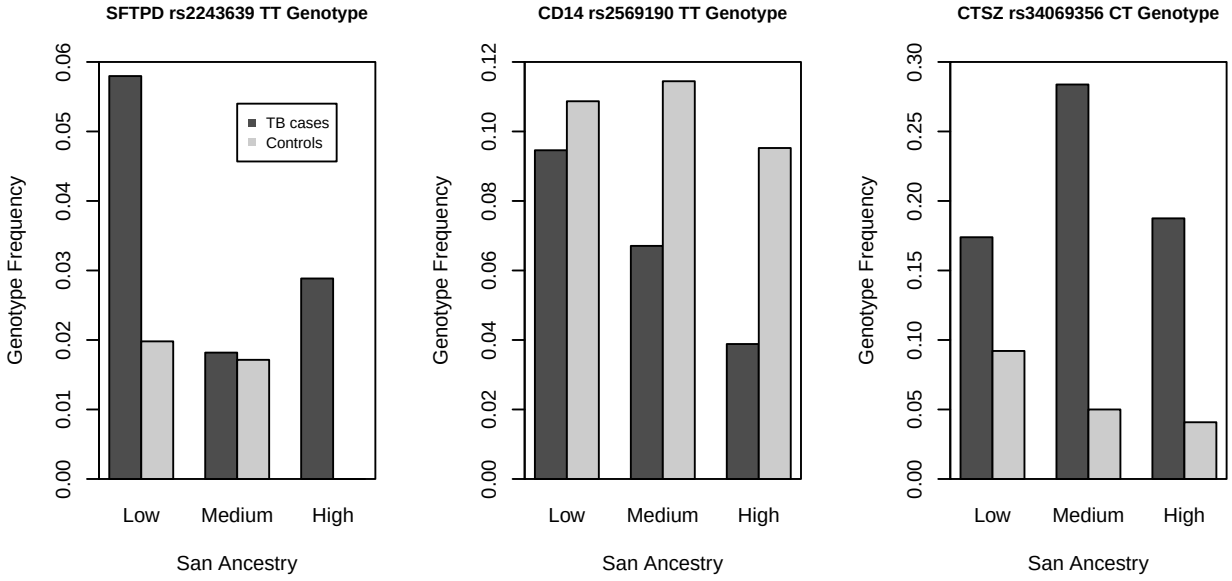


Figure 5.3: Genotype frequency stratified by African San ancestry and case/control status. Frequencies for the same genotypes shown in figure 5.2 are depicted in individuals with low, medium and high San ancestry. Frequency differences between cases and controls are contrasted in this figure for each of these ancestry groups. The low San ancestry group represents those individuals that fall within the first quartile of San ancestry. Similarly, the high San ancestry group falls within the fourth quartile of San ancestry. The medium San ancestry group falls within the second and third quartiles.

African populations were not exposed to modern strains of *M. tuberculosis* [76], resulting in little opportunity to develop resistance. This could possibly explain the strong association between African San ancestry and TB susceptibility we observed in our study group. Could differences in susceptibility between ethnicities be ascribed to the origin of infecting strains? Whilst this may be possible, using worldwide human population data, our recent investigations into the relationship between common HLA class I alleles in a population and the most prevalent strains in that population, showed that the relationship is far from simple [159].

We also illustrate the importance of adjusting for ancestry in an admixed population when such a group is used in a case-control study. Association results may be confounded by ancestry if cases and controls have different ancestry proportions from a particular source population [37]. We report two illustrations of this phenomenon in a TB association study of the SAC, one where the genotype association is significant after adjusting for ancestry but not before, and one where the genotype association is no longer significant. We also report a model where the significance of the genotype association remains largely unchanged. We demonstrate that association results are likely to be affected by adjustment for ancestry if allele frequencies differ markedly in the source populations of the admixed population. The mechanism resulting in allele frequency differences in the source populations of the SAC, e.g. genetic drift or natural selection, can however not be incorporated in disease association models. A non-significant association result that would have been significant before adjusting for ancestry indicates that differences in allele frequency between cases and controls are due to ancestry. This could be due to selection involving TB susceptibility, genetic drift or other selection pressures. The possibility of such a polymorphism being involved in progress to TB can therefore not be ruled out. Approaches such as admixture mapping are required to quantify the probability of the polymorphism playing a role in disease, based on statistics that measure whether a particular ancestry is over-represented at a chromosomal segment. By adjusting for ancestry, we can merely be more confident that significant associations are real and not artefacts due to

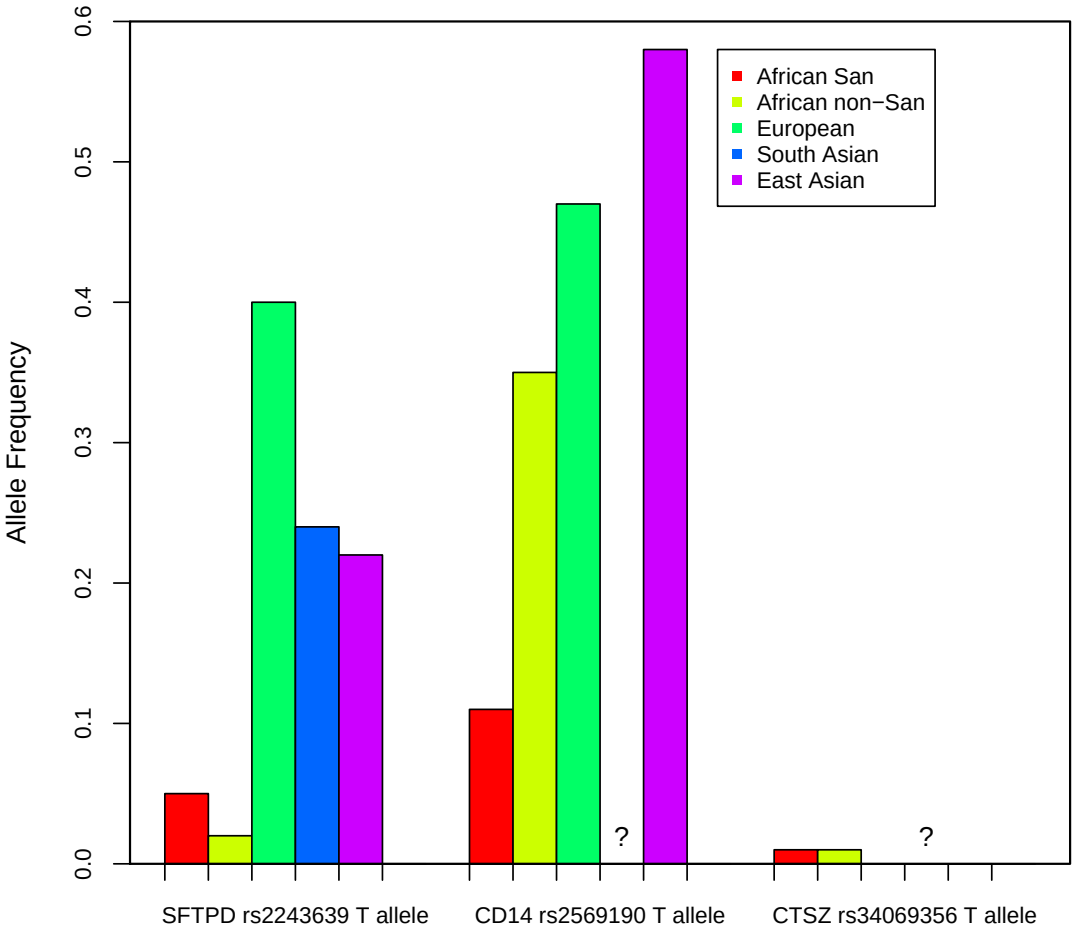


Figure 5.4: Allele frequencies in the source populations of the SAC. Variant allele frequencies of the *SFTPD* rs2243639, *CD14* rs2569190 and *CTSZ* rs34069356 SNPs are depicted in this figure. The *CTSZ* rs34069356 T allele has a frequency of 0.01 in the Khomani San (Brenna Henn, personal communication). The other African San allele frequencies were calculated using data published by Schlebusch et al. [126] (Khomani, Ju|'hoansi, Nama, !Xun populations). The South Asian rs2243639 allele frequency was calculated using the HapMap3 *GIH* population. Allele frequencies for the other source populations were obtained from the 1000 Genomes project. Unknown allele frequencies are indicated by question marks.

differences in ancestry.

Our findings have the following limitations. We have estimated ancestry proportions using AIMs, which are less robust compared to proportions estimated using genome-wide data, and have not been able to adjust our analysis for socioeconmic factors. Any bias introduced by using AIMs rather than genome-wide data is likely to result in less sensitive detection of population stratification, i.e. the association between ancestry and TB susceptibility may be stronger than reported, whereas the association between genotypes and TB susceptibility may be stronger than reported in the case of positive confounding, or less strong in the case of negative confounding. We believe that our results are valid despite this, since we have previously shown that the distribution of ancestry proportions estimated using the selected AIMs, versus proportions estimated using genome-wide data, is comparable in SAC groups of reasonable size [99]. In addition, our findings are concordant with those of Chimusa et al. [98], who used genome-wide data to estimate ancestry. Regarding the bias that may be present in genetic association studies, the panel of AIMs we used in this study were selected using a strategy similar to the one used by Galanter et al., who found that a panel of only 22 AIMs resulted in a marked decrease of false-positive findings in a type 2 diabetes study in Mexicans [116]. Although we have not been able to adjust our analysis for socioeconomic factors, we note that cases and controls were recruited from the same community that largely share the same SES and environmental risk factors for TB.

In conclusion, we have provided further evidence of association between ancestry and TB susceptibility and have illustrated the importance of adjusting for admixture in genetic TB susceptibility studies of the South African Coloured population.

Acknowledgements

We thank Alicia Martin for her help in obtaining Khomani San allele frequencies and Ben Viljoen and Corné de Kok for their work in the laboratory. We also thank the developers of the open source software we used in our analysis.

5.6 Supplementary tables

Supplementary Table 5.6.1: TB susceptibility candidate gene association studies. The table summarizes the total number of samples that were successfully genotyped in each study, how many samples have complete information (age, gender and ancestry), and how many samples overlapped with the Affymetrix data set.

			TB cases			Controls		
Study	**Year**	**Genes**	**Total**	**Compl**	**Affy**	**Total**	**Compl**	**Affy**
Rossouw et al. [160]	2003	IFNG	393	302	176	286	231	36
Hoal et al. [48]	2004	NRAMP1, NRAMP2	429	324	183	436	288	49
Babb et al. [161]	2007	SP110	379	334	271	304	268	45
Barreiro et al. [137]	2007	DCSIGN	339	286	208	227	202	32
Moller, unpublished [162], Moller et al. [163; 164; 136; 138; 139; 141]	2007-2011	ATG16L1, BTNL2, CARD15, CCL2, CTLA4, CTSZ, FCRL3, FZD5, IL10, IL12B, IL12RB1, IL12RB2, IL18, IL1RN, IL23R, IL4, IL6ST, INSIG2, MDR1, MHC2TA, MS4A2, NELL1, NOS2A, PADI4, PPARG, PTGER4, PTPN22, RUNX1, SH2D1A, SLC22A4, SLC22A5, SOCS3, TEX264, TLR2, TLR4, TNF, TNFRSF1A, TNFRSF1B, TNFSF15, WNT5A	781	584	412	703	390	78
Adams et al. [141]	2011	MC3R	439	386	289	505	410	78
Babb, unpublished [165]	2007	CCR5, RANTES, SDF1	312	229	183	228	178	42
De Wit, unpublished [166]	2009	IFNGR1, IL8, RANTES	397	303	183	289	231	45
Salie, unpublished [167]	2010	ANAXA11, CADM1, CADM2, CADM3, NCAM2	382	341	270	398	344	73
Lucas, unpublished [168]	2011	TLR8, TLR9	479	439	291	496	447	85
Wagman, unpublished [169]	2011	MARCO, SFTPD	383	343	272	395	343	73
Bruiners, unpublished [170]	2012	C1QA, C1QB	472	433	287	477	431	81
Unpublished	1999	IL1RA	193	156	83	158	99	20
Unpublished	2001	OPN5, OPN6	206	148	71	124	64	12
Unpublished	2002	SFTPD	161	124	62	144	75	14
Unpublished	2003	IL8	214	156	176	149	112	37
Unpublished	2003	IFNGR1, IFNGR2	373	275	182	348	226	45
Unpublished	2004	CO2REGION, HS3ST4	557	441	341	475	292	50
Unpublished	2007	FOXP3	502	414	288	513	351	75
Unpublished	2008	P2X7, TLR1	494	416	299	525	401	76
Unpublished	2010	CD14	387	341	252	406	321	55
Unpublished	2012	APOE	435	410	270	443	407	80
Unpublished	2013	IRGM, ISG15, NLRC5, NLRP3, NOD2	427	407	279	439	403	83

Chapter 6

Research Article 3

Investigating the role of gene-gene interactions in TB susceptibility

Michelle Daya[1], Lize van der Merwe[1], Paul D. van Helden[1], Marlo Möller[1], Eileen G. Hoal[1,*]

1 Molecular Biology and Human Genetics, MRC Centre for Molecular and Cellular Biology and the DST/NRF Centre of Excellence for Biomedical TB Research, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa
∗ E-mail: Corresponding egvh@sun.ac.za

This article has been submitted to the journal PLOS ONE.

6.1 Abstract

Tuberculosis (TB) is the second leading cause of mortality from infectious disease worldwide. One of the factors involved in developing disease is the genetics of the host, yet the field of TB susceptibility genetics has not yielded the answers that were expected. A commonly posited explanation for the missing heritability of complex disease is gene-gene interactions, also referred to as epistasis. In this study we investigate the role of gene-gene interactions in genetic susceptibility to TB using a cohort recruited from a high TB incidence community from Cape Town, South Africa. Our discovery data set incorporates genotypes from a large a number of candidate gene studies as well as genome-wide data. After limiting our search space to pairs of putative TB susceptibility genes, as well as pairs of genes that have been curated in online databases as potential interactors, we use statistical modelling to identify pairs of interacting SNPs. We attempt to validate the top models identified in our discovery data set using an independent genome-wide TB case-control data set from The Gambia. A number of models were successfully validated, indicating that interplay between the *NRG1* - *NRG3*, *GRIK1* - *GRIK3* and *IL23R* - *ATG4C* gene pairs may modify susceptibility to TB. Gene pairs involved in the NF-κB pathway were also identified in the discovery data set (*SFTPD* - *NOD2*, *ISG15* - *TLR8* and *NLRC5* - *IL12RB1*), but could not be tested in the Gambian study group due to lack of overlapping data.

6.2 Introduction

Tuberculosis (TB) is a serious global health problem, with 8.6 million new infections and 1.2 million deaths reported in 2012 [49]. In South Africa, it is the fourth leading cause of mortality [50]. The South African Coloured population (SAC) is the largest demographic in metropolitan

areas of Cape Town that have some of the highest reported incidences of TB worldwide, despite extensive BCG vaccination and low HIV prevalence [48].

Although up to a third of the world's population has latent TB infection [49], only about 10% of immunocompetent individuals progress to disease. Many studies have established that host genetic factors are involved in the disease [66]. As is the case for other complex diseases, only a small proportion of the posited heritability has been found [171; 172; 173]. The results of TB association studies are furthermore often inconsistent between studies [174; 175].

One of the common explanations for the missing heritability of complex disease is gene-gene interactions, a.k.a. epistasis [176; 171; 177; 178]. It has also been postulated that failure to validate genetic associations in independent studies may be ascribed to epistasis [179]. Epistasis can be defined as the effect of a genetic locus on a phenotype being modified by one or more other loci. The term was first used by Bateson based on his experiments with flower color in pea plants, showing that the effects of one gene can be masked by another gene [180; 181]. A similar term, "epistacy", was later coined by Fisher, referring to the interaction term in regression models that attempts to encapsulate the relationship between two genetic loci and an outcome variable [38]. Based on context, the term epistasis can thus either refer to biological interaction, where effects are mutually dependent and describe a state of nature, or statistical interaction, alluding to the interaction term of two or more variables in a regression model [38; 178; 152]. The notion of biological interaction has often been demonstrated experimentally in model organisms such as yeast, bacteria and animal models [182; 183; 184; 185; 186; 187; 188; 189; 190; 191; 177], but this has been less successfully demonstrated in humans [190; 68]. It should also be noted that absence of detectable statistical interaction does not necessarily imply lack of biological interaction [178; 152].

The immune system is complex and comprises many intricate elements, thus progression to active TB may be elucidated by identifying the interplay of gene products in the host defence against TB infection. Only a small number of TB susceptibility gene-gene interaction studies have been published to date [192; 140; 193; 194; 195; 196; 197] and these were limited to a small number of candidate genes. In this study, we use a large sample bank of TB case and control samples collected from SAC individuals residing in areas of high TB incidence to detect gene-gene interactions that may underlie TB susceptibility. The data set constitutes genotype data collected from a large number of candidate gene studies (76 genes and 214 SNPs), as well as a large micro-array (chip) data set (388 654 SNPs, 642 cases and 91 controls). We also incorporate correction for ancestry.

A large variety of software packages have been developed to detect gene-gene interactions [198; 199; 200; 201; 202; 203; 204; 205; 206; 105; 207; 208; 209; 210; 211; 212; 213]. Approaches implemented in these and other packages can be be broadly classed as "traditional" regression based approaches, Bayesian frameworks, testing for allelic association, machine learning and pathway or network based approaches. In traditional regression based methods, interactions are identified by a linear model with phenotype as outcome variable and genotypes as predictor variables. These models includes interaction term(s) which measures the departure of two or more loci from additivity. Interaction models can also be identified using Bayesian frameworks. A prior distribution for the unknown parameter(s), such as the number of predictors to use in a regression model, or the type of effect markers have on the phenotype (no, main or interaction) is specified. The posterior distribution of the parameter(s) is then estimated using simulation techniques such as Markov chain Monte Carlo (MCMC). Due to the large dimensionality of especially genome-wide data sets, an initial filtering step is sometimes employed prior to testing for association using regression. A simple technique employed initially was one proposed by Marchini et al. [214], where tests for interaction are limited to loci that are marginally associated with the phenotype. A more recent strategy is to limit association testing to loci based on

curated biological knowledge [215; 216; 217; 218]. Another popular approach is the use of test statistics that can be computed efficiently [198; 213]. A particularly intuitive test statistic that measures interaction and that can be computed efficiently is a test for allelic association [209], which can be computed in cases only, or used to test for differing allelic association between cases and controls. Machine learning and data mining techniques use computationally efficient algorithms to identify a set of variables that can be used to predict or classify an outcome. These techniques are especially useful for identifying multiple predictors, and often use the notion of training and testing data sets to first train and then test models on different subsets of data. Pathway and network based approaches have also recently become popular, and describes complex networks of interactions that may affect a phenotype [200; 190; 203]. Graph theory is used to find subnetworks of genes that represent a common pathway and that are enriched for association with the outcome of interest. In this way underlying disease pathways are identified, rather than specific variants that may be interacting.

With the advent of new genome editing technologies such as CRISPR, detecting pairs of SNPs rather than pathways or networks lends itself to experimental validation. Given the size of our study group, limiting our search to pairs of SNPs only is also appropriate. Our SAC case-control data set comprises sample sets that were genotyped in a number of different studies which did not always overlap well, resulting in a relatively sparse data set. Age, gender and ancestry also differ between cases and controls [100]. We therefore used statistical modelling rather than data mining techniques to identify interactions between pairs of SNPs. Statistical modelling allows for the adjustment of known confounders, and utilizes all available data for each test, without requiring imputation or other complex strategies to deal with missing data. We also limit our search space to pairs of genes that have been identified as TB susceptibility candidate genes, and pairs of genes that have been curated in online databases as potential interactors, a strategy that has previously been used successfully [216; 217]. Finally we attempt to validate our findings in an independent Gambian TB case-control data set.

6.3 Subjects and methods

6.3.1 Sample collection and ethical approval

Individuals from the Ravensmead and Uitsig suburbs in Cape Town, who self-identified as South African Coloured, were recruited to participate in this study. These suburbs have a homogenous socio-economic environment, low prevalence of HIV and high incidence of TB [48]. TB patients were diagnosed using bacterial confirmation (smear positive/culture positive). Healthy individuals with no prior history of TB were selected as controls. All participants were HIV negative. Our previous study of healthy children and young adults from the control community found that 80% of children older than 15 years had positive tuberculin skin tests (TST), an indication of latent infection with *Mycobacterium tuberculosis* (*M. tuberculosis*) [219]. The majority of the control population is therefore TST positive, and with the average age of the controls in this study being 31 years, we estimate a TST positivity of 80% or above. These healthy individuals had no previous history of TB disease or treatment and were unrelated to all others included in the study.

This study was approved by the Ethics Committee of the Faculty of Health Sciences, Stellenbosch University (project registration numbers 95/072, NO6/07/132 and N11/07/210). Blood samples for DNA were collected with written informed consent. The research was conducted according to the principles expressed in the Declaration of Helsinki.

Sample collection and ethical approval of a Gambian tuberculosis study group, obtained from the Welcome Trust Case Control Consortium (WTCCC), are described by Thye et al

[220].

6.3.2 Genotyping and quality control

A total of 955 case and 521 control SAC samples were collected between 1994 and 2007. The samples were used to perform a number of candidate gene studies, using unrelated individuals, summarized in supplementary table 6.7.1. Single nucleotide polymorphisms (SNPs) that were genotyped in these studies were used in the present study, and SNPs with a minor allele frequency (MAF) lower than 0.01 and a Hardy Weinberg equilibrium (HWE) p-value (exact test) lower than 0.01 were discarded, leaving 214 SNPs from 76 genes. The SAC sample bank was also used to genotype 969 samples on the Affymetrix GeneChip Human Mapping 500K Array set (Affymetrix 500K chip set). A total of 642 cases, 91 controls and 388 654 SNPs were retained in the data set after SNP calling, [26], quality control and removal of related individuals [98]. The data set was also aligned to the Genome Reference Consortium Human genome build 37 (GRCh37). The SAC candidate gene and Affymetrix data sets were then combined, and used to identify pairs of SNPs that jointly modify the odds of having TB.

The Gambian tuberculosis data set was used to validate the top interaction models found in the SAC data set. A total of 1 498 cases and 1 496 controls was genotyped on the Affymetrix 500K chip set (more detail can be found in Thye et al [108], Supplementary Methods). After SNP quality control (removal of SNPs with calling probability < 0.95, HWE p-value < 0.0001, MAF < 0.01, missing rate > 0.05) and alignment to GRCh37, 402 856 SNPs remained in the data set. Individuals with excess heterozygosity, outlying individuals and related individuals (degree of relatedness ≤ 2, according to identity by state estimates) were also removed. The final data set were composed of 1 156 cases and 1 206 controls.

6.3.3 Limiting the search space based on biological evidence

We limited tests for interaction to SNP pairs of genes that have been identified as TB susceptibility candidate genes (which we refer to as candidate gene SNP pairs), and pairs of genes that are known to interact based on experimental evidence, or that are found in the same biological pathway, ontological category or protein family (which we refer to as biofilter SNP pairs, after the software program used to identify the pairs). A total of 76 candidate genes previously genotyped by our group, as well as 33 additional tuberculosis and pulmonary tuberculosis candidate genes curated in the HGV&TB database based on literature reviews (http://genome.igib.res.in/cgi-bin/hgvtb/inter.cgi), with at least one SNP genotyped in the SAC chip data, was used to generate 5 886 candidate gene-gene pairs, composed of 1 278 unique candidate gene SNPs. Seventeen of these SNPs were genotyped on both the Affymetrix chip and another platform, and the strand orientation of 10 of these SNPs were flipped when combining the data sets. Duplicate genotypes were available for 4 686 SNPs, of which 280 genotypes mismatched (error rate of 0.06). The mismatched genotypes were discarded. Another 2 438 interacting gene-gene pairs were identified, comprised of 28 936 unique SNPs. After discarding SNP pairs with less than 60 genotypes available for either cases or controls, 854 703 candidate gene SNP pairs and 1 040 161 interaction SNP pairs were identified for testing.

6.3.4 Statistical analyses

Logistic regression was used to identify pairs of SNPs that jointly modify the odds of having TB. The genotypes of SNPs were encoded as factor variables, and SNPs on chromosome X of male individuals were encoded as homozygotes. Covariates, the main effects of each SNP and

an interaction term were included in each model (Case/Control ∼ Covariates + SNP1 + SNP2 + SNP1×SNP2).

The p-value of the interaction term was used to detect and report the significance of interactions (4 degrees of freedom test). Reported p-values were not corrected for multiple testing. To aid interpretation of the results, the nature of the association is illustrated by graphs of the observed genotype combination proportions in the data, as interaction effects such as odds ratios are difficult to describe and interpret. Furthermore, reliable estimates of odds ratios could often not be calculated, as some of the genotype combinations include zero counts. Graphs of allele combination frequencies in cases and controls are also provided. An expectation-maximisation (EM) algorithm was used to infer allele combinations per subject. The particular algorithm was originally designed to infer haplotypes, but does not assume the physical coupling of SNPs, and is therefore also appropriate for estimating allele combination frequencies. We note that the only uncertainty in inferring these allele pairs is double heterozygotes. The logits of the possible genotype combinations are also illustrated. This demonstrates the differing direction or magnitude that a SNP has on the odds of having disease, depending on the genotype of the second SNP; non-parallel lines being indicative of an interaction effect. The effects were estimated by absorbing the marginal effects of the SNPs into the SNP × SNP interaction term, and adjusting for the covariates included in the model by averaging over them [221].

After the top interacting pairs of SNPs were identified, the individual effects of each of the identified SNPs were tested separately in the SAC and Gambian cohorts using logistic regression. SNPs were encoded as factor variables and covariates were included in each of the models (Case/Control ∼ Covariates + SNP).

Allelic interaction of the identified top SNP pairs was also tested in the SAC cohort. SNPs were encoded as numeric variables, according to the number of copies of the rare variant, as follows: 0, 1 or 2 copies of the rare variant for additive encoding, 0 or 1 for dominant encoding, with 1 representing heterozygotes and rare homozygotes, and 0 or 1 for recessive encoding, with 1 representing rare homozygotes. Each of the nine possible allelic encoding combinations were then tested for each of the identified top SNP pairs.

Age and gender are differentially distributed in the SAC TB cases and controls and gender is differentially distributed between the Gambian TB cases and controls (supplementary table 6.7.2, age not available for the Gambian data). Age and gender were therefore included as covariates in the SAC study group models, and gender was included as a covariate in the Gambian study group models.

Previous work has shown that TB cases have a higher proportion of African ancestry compared to controls in the SAC study group [100; 98], necessitating adjustment for ancestry. Ancestry proportions for each of the 5 source ancestries of the SAC (African San, African non-San, European, South Asian and East Asian) were estimated using a panel of 116 AIMs, as described previously [100]. Ancestry proportions were estimated in a similar manner but using genome-wide data, for those individuals that were also genotyped on the Affymetrix chip. These ancestry proportions were included as covariates in the SAC study group models.

Quality control of the Gambian data set revealed that missing genotypes were associated with having TB for a relatively large proportion of SNPs, which may be indicative of batch effects [101]. As this can be mitigated by the inclusion of principal components in statistical models, principal components were used to adjust the analysis, rather than ancestry proportions as was done for the SAC cohort. Principal components would adjust the models for both differences in ancestry and batch effects between cases and controls [109]. Principal component analysis of the Gambian study group showed associations between having TB and principal

components 1, 2, 5, 6, 8, 9 and 10 (p-values < 0.05). These principal components were included as covariates in the Gambian study group models.

6.3.5 Validation

Statistical modelling was used to identify gene pairs that most likely jointly modify the odds of having TB, and not to quantify the achieved level of statistical significance. Due to the large number of tests done in the SAC study group, and the limited size of the study group (especially the limited number of controls that were available for many of the tests), none of the interaction associations would be statistically significant if adjusted for multiple testing. In addition, many of the multiple testing methods that have been suggested in the literature have severe shortcomings. The straightforward Bonferroni adjustment is too stringent when several genetic associations are tested in the same study group due to correlation (LD) between markers [150; 222]. Alternative methods of correcting for multiple tests were also not feasible for this study. Firstly, roughly 2 million tests were done in differing subsets of individuals from the same study group, which complicates the use of multiple testing correction methods that do not rely on the simple adjustment of p-values by for example dividing by the number of tests done. Bayesian methods require a priori probability of association, which is not known. Due to the large number of tests that were done, permutation testing is also not feasible. Permutation testing is also inappropriate in the context of gene-gene interactions, as permutation based methods do not account for correlation between genotypes [223]. Furthermore, a large proportion of the tests were done on an unbalanced number of cases and controls, which may result in biased permutation-based calculation of p-values. A method to determine the number of effective independent tests when testing pairs of SNPs for interaction in a genome-wide context has also been proposed [216; 224]. This number of effective tests can then be used in a Bonferroni adjustment or to control the false discovery rate. The method does however not take into account that a gene may be tested in multiple gene-pair models, and the accuracy of the original method was evaluated using permutation testing, which may be inappropriate for interaction tests. Due to these reasons an appropriate alpha level was not determined, and we simply selected the top 20 unique gene pairs for validation in the Gambian study group. A similar strategy has been suggested by Kerr [225], albeit in the context of unbalanced microarray gene expression data. The selected models would be the most likely true positives, if any exist.

As patterns of linkage disequilibrium (LD) differ between populations, tag SNPs of causal variants may vary between the SAC and Gambian populations. A SNP associated in the SAC study group points to a region of LD, and any SNP within this region may be the causal SNP [226]. The 20 models that were selected for validation were therefore tested using all possible combinations of SNP pairs found in the region of the SNP tested in the SAC study group. Using a strategy similar to that of Shriner et al. and Ramos et al. [226; 227], SNPs used for validation of a SNP tested in the SAC study group was selected based on the following criteria: the SNPs were found in the same gene region, within 250 000 base pair positions of the SNP, and having a pairwise LD r^2 value of at least 0.3 with the SNP in SAC controls. Although some of the SNPs genotyped in candidate gene studies were selected for their putative functional effects, we note that all the variants in the top twenty models that were genotyped in candidate gene studies were originally selected as they were variants in a gene of interest, and not for their functional effects per se (should this not have been the case, imputation of the exact SNPs for validation purposes would be the preferred strategy).

After selecting SNP pairs to test using this strategy, a resulting total of 245 regression models were fitted to the Gambian study group. P-values smaller than 0.05 were described as statistically significant.

6.3.6 Software

Version information, web URLs and important parameter settings of the software packages used in this study are summarized in supplementary table 6.7.3.

PLINK was used for quality control of the SAC chip data set and Gambian chip data set [105]. The SAC and Gambian chip data sets were aligned to GRCh37 using a script and Affymetrix SNP information files available at
http://www.well.ox.ac.uk/ wrayner/strand/.

ADMIXTURE was used to estimate ancestry proportions of the SAC study group [129]. For the Gambian study group, Eigenstrat was used to infer the top 10 principal components and test for association between these principal components and disease outcome [109]. Prior to estimating ancestry proportions and inferring principal components in the SAC chip data set and Gambian data set, PLINK was used to remove SNPs from the data set that were in LD, as this may lead to biased inference.

Biofilter was used to generate SNP pair combinations of genes that are known to interact based on experimental evidence, or that are found in the same biological pathway, ontological category or protein family. Only those combinations having three or more sources were used for testing interaction in the SAC chip data set. Biofilter was also used to find SNPs within gene regions that are available in the Gambian data set for validation of the top SAC gene-gene models.

The freely available R programming environment was used for statistical analyses, quality control of the SAC candidate genes and graphing [106]. The R *genetics* package was used to test for HWE in the SAC candidate genes and was also used to calculate pairwise LD r^2 and D' values [154]. The R *haplo.stats* package was used to estimate allele combination frequencies in cases and controls [228]. The adjusted logits of the genotype combinations were estimated using the *effects* package [221]. Figures were created using the R *ggplot2* package [229].

6.4 Results

The top 20 unique gene pair models discovered in the SAC cohort are summarized in table 6.1. These models were identified using logistic regression, that tests whether the effect of a SNP on disease outcome is modified by the effect of another SNP, after taking into account (adjusting for) the main effects of the two SNPs. When encountering the same gene-gene model but with differing SNP pairs, only the gene-gene model with the smallest p-value is shown (4 models were excluded for this reason). SNP pairs and p-values of the corresponding highest scoring Gambian models are also reported in the table. As no suitable SNPs were available for some of the genes, some of the models could not be tested in the Gambian data set.

Table 6.1: Top twenty interaction models. This table summarizes the top twenty interaction models identified in the SAC study group. P-values reflect the overall significance of the association between the genotype combinations and having TB, after adjusting for the main effects of the SNPs and covariates. A model of type C indicates a candidate gene pair, and a model of type B indicates a biofilter gene pair. These models were validated in the Gambian study group set using multiple SNPs found within the same gene regions, and the SNP pairs and p-values of the highest scoring Gambian models are reported. For some of the models, no SNPs were available in the Gambian data set for one or both of the genes (blank entries).

				SAC				Gambian			
	Gene 1	**Gene 2**	**Type**	**Nr cases**	**Nr controls**	**SNP 1**	**SNP 2**	**SNP 1**	**SNP 2**	**P-value**	**Nr tests**
Model 1	NRG1	NRG3	B	634	87	rs16879814	rs11191757	rs16879814	rs2224109	**0.0389**	$1 \times 12 = 12$
Model 2	GRIK1	GRIK3	B	620	90	rs465555	rs3738085	rs460583	rs476894	**0.0476**	$5 \times 11 = 55$
Model 3	SFTPD	NOD2	C	216	65	rs1923537	rs748855				
Model 4	IL23R	ATG4C	C	613	85	rs10489628	rs11208029	rs10489628	rs11208029	**0.0350**	$1 \times 3 = 3$
Model 5	FUT8	B4GALT1	B	627	90	rs17102844	rs12342831	rs9323464	rs10758189	0.1399	$4 \times 7 = 28$
Model 6	EXT1	EXT2	B	626	91	rs6469713	rs903509				
Model 7	ISG15	TLR8	C	271	321	rs15842	rs3761624				
Model 8	NCAM2	IRF8	C	620	87	rs8134735	rs8054065	rs8132838	rs147968	0.0794	$4 \times 2 = 8$
Model 9	ANK1	ANK3	B	606	91	rs2102360	rs2393618				
Model 10	NELL1	NOS2	C	639	91	rs1377741	rs2297516	rs1377741	rs2314809	0.4098	$1 \times 2 = 2$
Model 11	CADM3	SLC22A4	C	224	67	rs16841729	rs13179900				
Model 12	ANK2	ANK3	B	636	90	rs1354679	rs10821731	rs1354679	rs10761481	0.1544	$5 \times 18 = 90$
Model 13	NELL1	CADM2	C	625	89	rs4614448	rs17024414	rs4614448	rs17024876	**0.0329**	$3 \times 7 = 21$
Model 14	NLRC5	IL12RB1	C	231	245	rs289726	rs393548				
Model 15	PLCB1	PLCE1	B	633	91	rs708914	rs4918082	rs1703634	rs4918082	0.3165	$2 \times 1 = 2$
Model 16	C1QA	TMEFF2	C	263	79	rs12033074	rs4077949				
Model 17	NELL1	CADM3	C	621	84	rs11025887	rs862991	rs12577018	rs862991	0.2107	$2 \times 1 = 2$
Model 18	PDE2A	PDE4B	B	626	87	rs171021	rs536025	rs3781931	rs17423910	0.1169	$2 \times 4 = 8$
Model 19	CHST11	CHSY3	B	623	87	rs17036205	rs32225	rs17036205	rs244745	**0.0401**	$1 \times 10 = 10$
Model 20	SLC22A4	ALOX5	C	629	91	rs2306772	rs3740107	rs3792880	rs3780909	0.1117	$4 \times 1 = 4$

The effects of each of the SNPs in table 6.1 were also tested individually in the relevant cohorts, and these single SNP association results are reported in table 6.7.4. Only two of the SNPs are individually associated with having TB in the SAC cohort (rs15842 and rs3740107 of models 7 and 20), but with a much lower level of significance than the interaction effect of the models (single SNP p-values of 1.49×10^{-2} and 1.64×10^{-2}, interaction p-values of 6.23×10^{-6} and 1.37×10^{-5}, respectively). Only one of the genes reported in table 6.1, GRIK1, was identified by the top 36 single SNP associations from a previous genome-wide association study of the cohort [98]. By evaluating combinations of genes, a number of genes were identified that may play a role in TB pathogenesis, which would not have been evident if their effects were assessed individually.

Interaction effects observed in the SAC study group are illustrated in figures 6.1-6.3 for validated models (p-value < 0.05 in the Gambian data set), as well as models that could not be validated due to lack of data, but that have interesting functional interpretations. Note that due to the differing SNP pairs used in the validation, as well as different allele frequencies and LD patterns in the two cohorts, the trend observed in a "validated" Gambian model may not necessarily reflect that of the corresponding SAC model, and we use the term here to imply that there is evidence in both cohorts that the gene pair jointly modifies the odds of having TB. Figures 6.1 and 6.2 show the frequencies of the genotype and allele combinations in cases and controls. As per the definition of interaction, the allele combination graphs demonstrate the reversal of effects in cases and controls, e.g. if the SNP 1 allele 1 - SNP 2 allele 1 combination has a lower frequency in controls compared to cases, then the SNP 1 allele 1 - SNP 2 allele 2 combination has a higher frequency in controls compared to cases, i.e. the effect of allele 1 of SNP 1 is modified by the SNP 2 allele. Figure 6.3 depicts the joint effect that genotype combinations have on the odds of having TB, after adjustment for covariates; non-parallel lines being indicative of interaction effects. For example, model 7 in figure 6.3 shows that compared to the CT-AG genotype combination, the CT-GG combination increases the odds of having TB, whereas compared to the TT-AG combination, the TT-GG combination decreases the odds of having TB. Put another way, depending on whether the first SNP has one or two copies of the rare allele T, the effect of having two instead of one copies of the rare allele G for that SNP may increase or decrease the odds of having disease. The frequencies and effects in the SAC study group for the remaining top models are depicted similarly in supplementary figures 6.7.1-6.7.3, and figures 6.4-6.6 show the frequencies and effects of the validated Gambian models. Below we highlight models that were validated in the Gambian data set as well as three models that could not be tested, but that have interesting functional effects.

The *NRG1* - *NRG3* (Neuregulin 1 and 3) interaction effect observed in the SAC study group (model 1, p-value 8.32×10^{-7}) was also detected in the Gambian study group (p-value 0.0389). The SAC and WTCCC *NRG1* SNP is the same, and the Gambian *NRG3* SNP is located 66 235 base pairs upstream from the SAC *NRG3* SNP. In both the SAC and Gambian study groups, compared to the GG-AC/GG-AG combination, the GG-AA combination decreases the odds of having TB and the AG-AA combination increases the odds of having TB. The same pattern is thus observed, albeit using a different SNP in the second gene in the Gambian study group. Studies investigating the link between NRG1 and schizophrenia have demonstrated that NRG1 has a functional effect on the immune system by influencing immune cell adhesion [230] and the concentration of autoantibodies and pro-inflammatory cytokines in plasma [231]. Gene-gene interaction between *NRG1* and *NRG3* has also been observed in a schizophrenia study [232], and according to the NCBI BioSystems database, NRG3 may also be involved in the immune system.

An interaction between the *GRIK1* and *GRIK3* (glutamate receptor 1 and 2) genes was also detected in both study groups (model 2, SAC p-value 1.62×10^{-6} and Gambian p-value

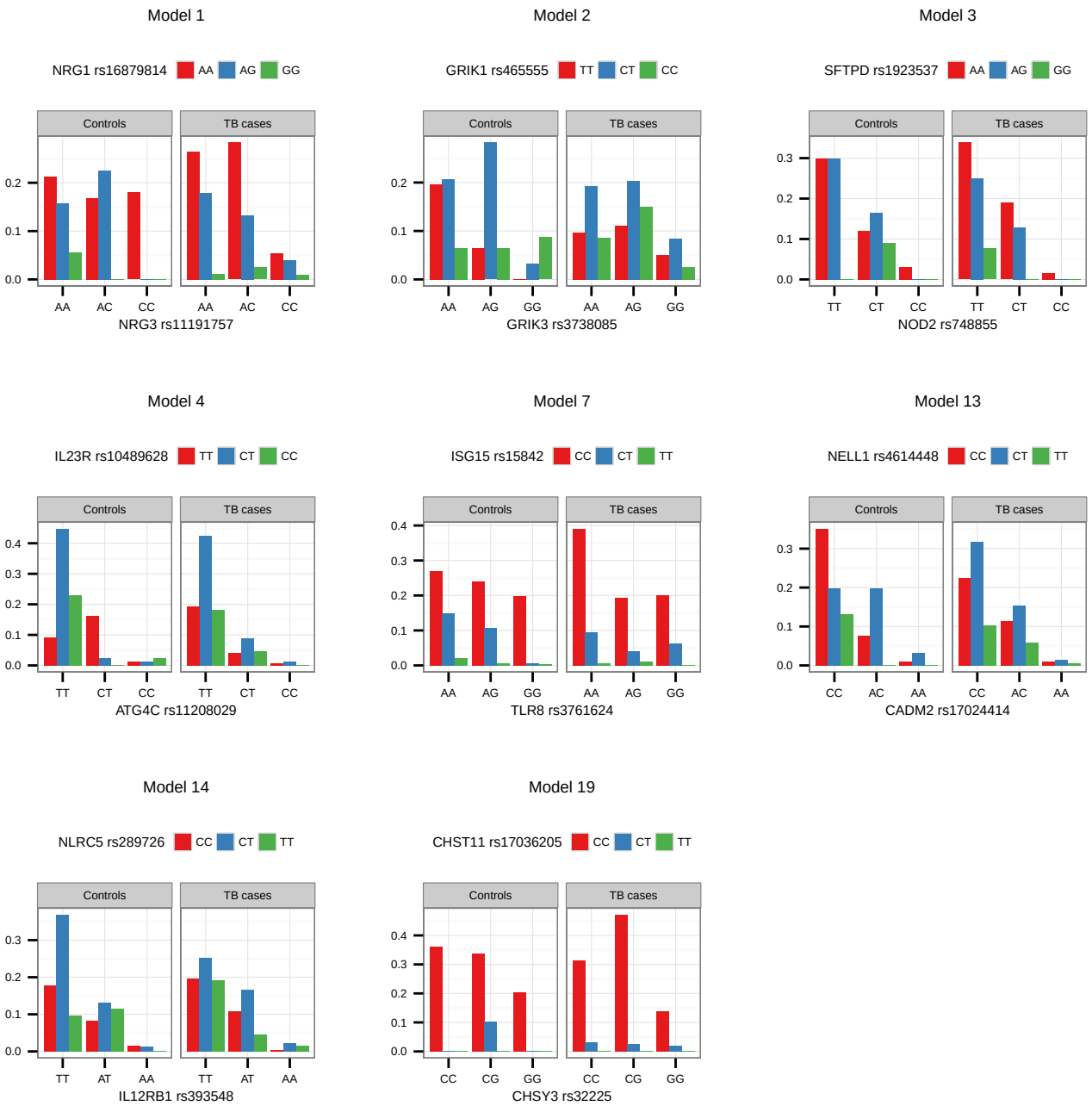


Figure 6.1: Genotype combination proportions in the SAC study group. The observed proportions of the nine possible SNP pair genotype combinations from models 1, 2, 3, 4, 7, 13, 14 and 19 are depicted in this figure, per cases and controls. Genotypes are ordered according to minor allele frequency, with the wildtype homozygote appearing first, and the rare homozygote appearing last.

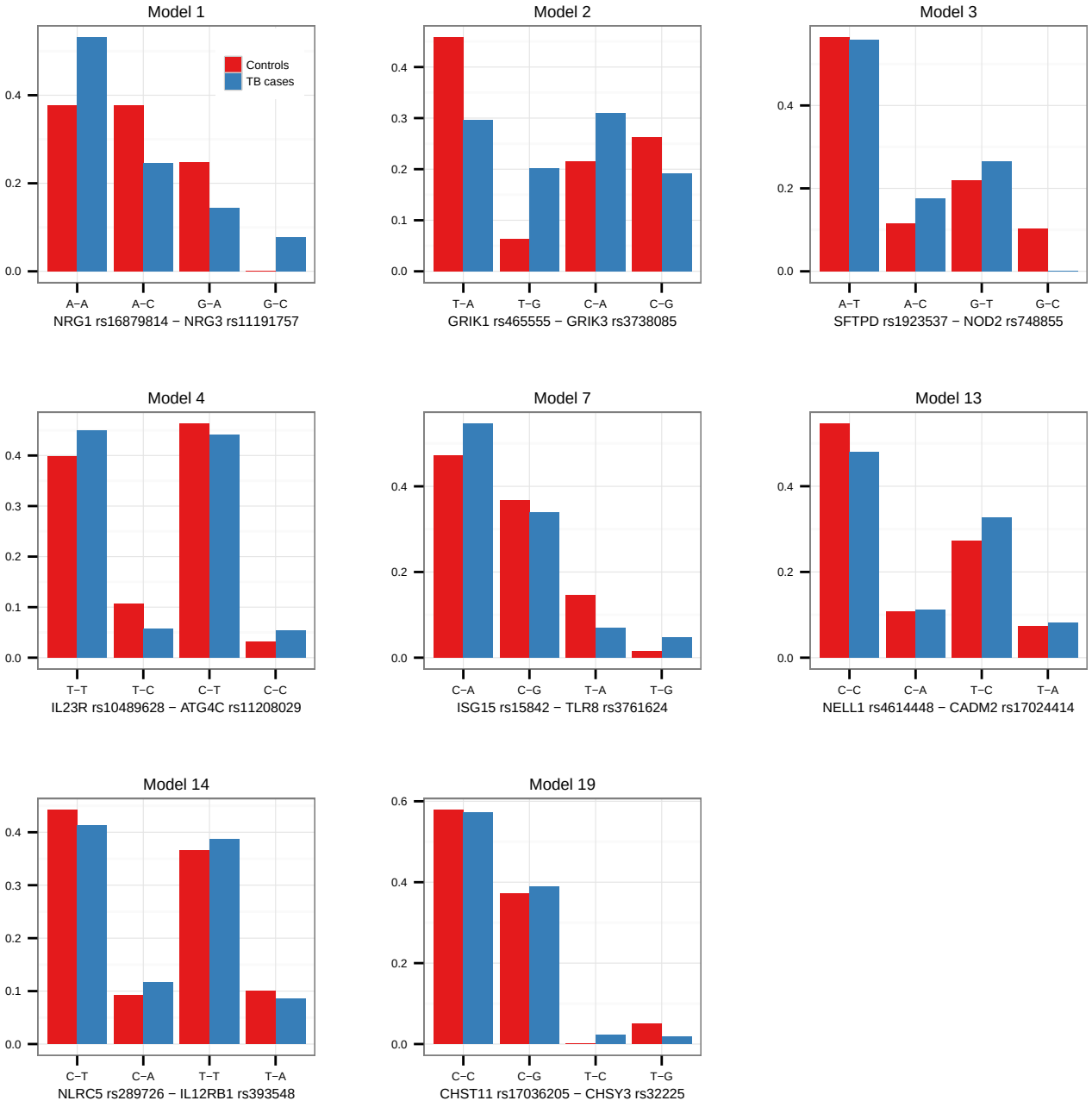


Figure 6.2: Allele combination frequencies in the SAC study group. The frequencies of the four possible SNP pair allele combinations from models 1, 2, 3, 4, 7, 13, 14 and 19 are depicted in this figure, per cases and controls. The frequencies were estimated using an EM-algorithm.

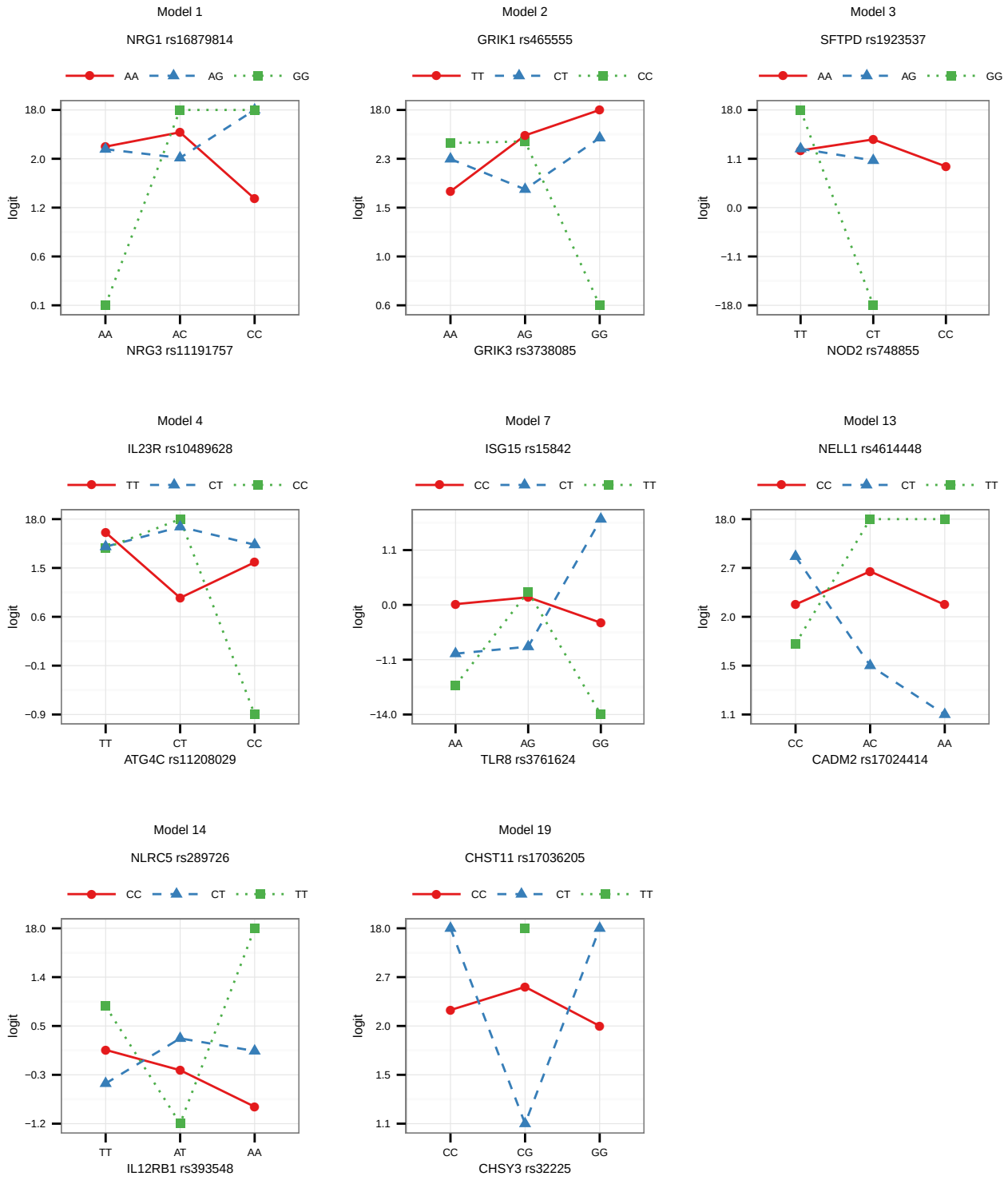


Figure 6.3: Effects in the SAC study group. The logits of genotype combinations from models 1, 2, 3, 4, 7, 13, 14 and 19 are depicted in this figure. Genotypes are ordered according to minor allele frequency, with the wildtype homozygote appearing first, and the rare homozygote appearing last. Non-parallel lines are indicative of interaction effects. The effects were estimated by absorbing the marginal effects of the SNPs into the SNP × SNP interaction term, and adjusting for the covariates included in the model by averaging over them.

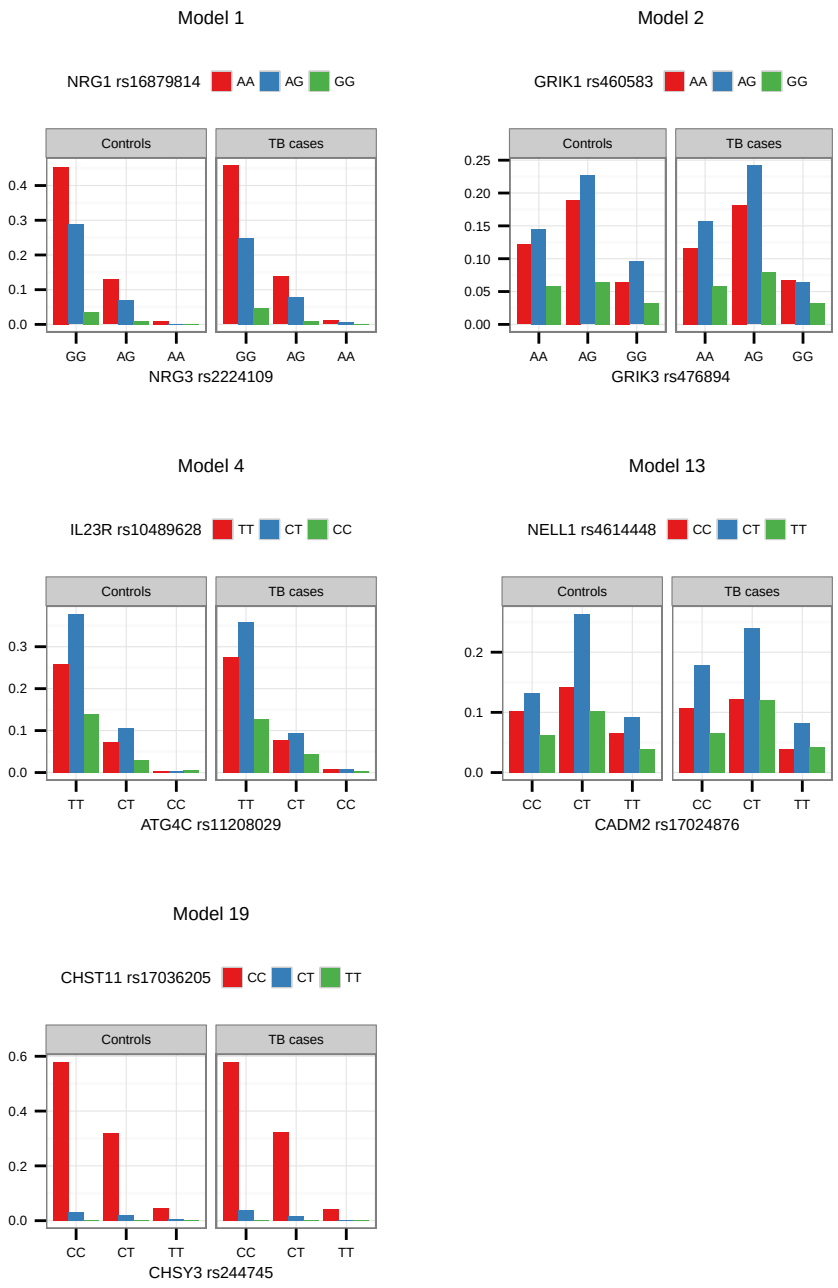


Figure 6.4: Genotype combination proportions in the Gambian study group. The observed proportions of the nine possible SNP pair genotype combinations from models 1, 2, 4, 13 and 19 are depicted in this figure, per cases and controls. Genotypes are ordered according to minor allele frequency, with the wildtype homozygote appearing first, and the rare homozygote appearing last.

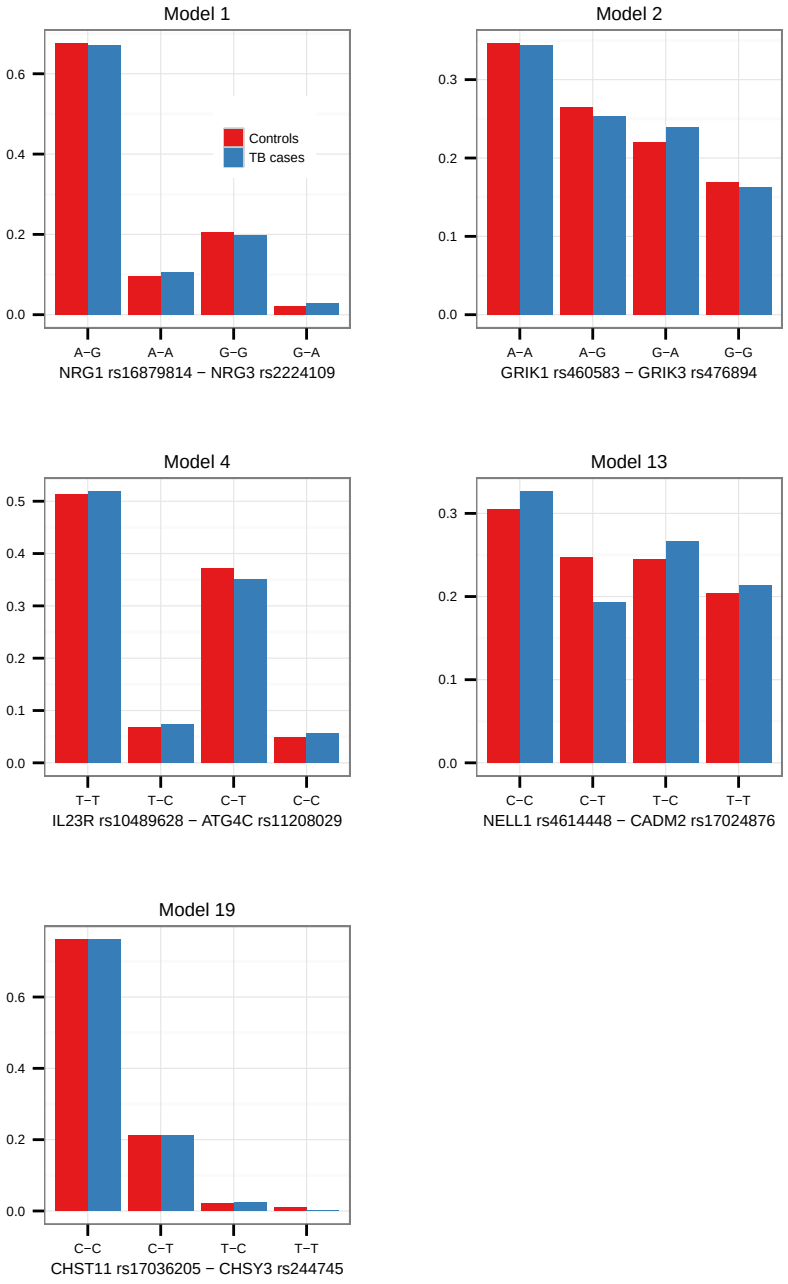


Figure 6.5: Allele combination frequencies in the Gambian study group. The frequencies of the four possible SNP pair allele combinations from models 1, 2, 4, 13 and 19 are depicted in this figure, per cases and controls. The frequencies were estimated using an EM-algorithm.

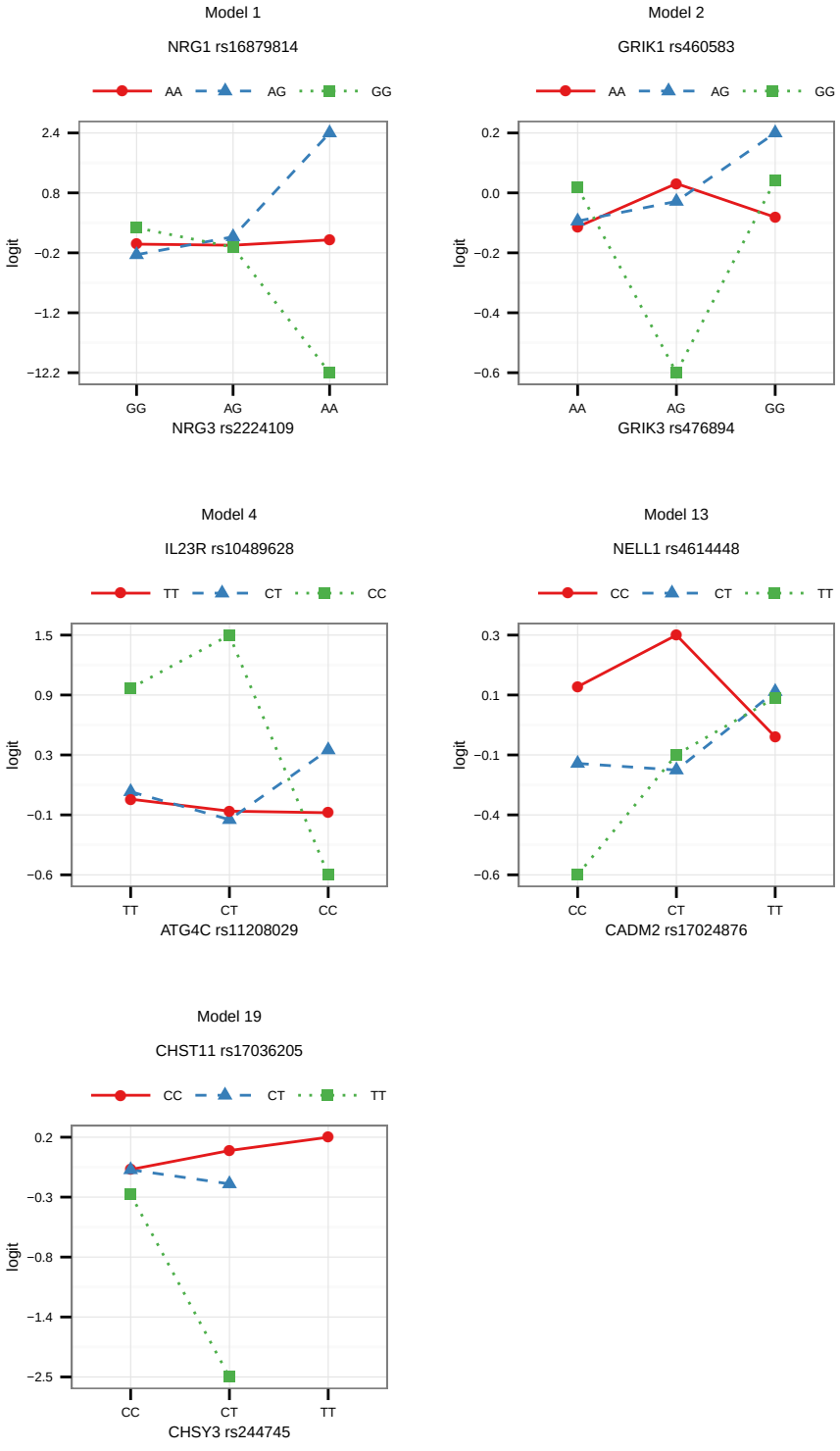


Figure 6.6: Effects in the Gambian study group. The logits of genotype combinations from models 1, 2, 4, 13 and 19 are depicted in this figure. Genotypes are ordered according to minor allele frequency, with the wildtype homozygote appearing first, and the rare homozygote appearing last. Non-parallel lines are indicative of interaction effects. The effects were estimated by absorbing the marginal effects of the SNPs into the SNP × SNP interaction term, and adjusting for the covariates included in the model by averaging over them.

0.0476). The TT-GG genotype combination was observed only in cases (5%) and the T-G allelic combination is more frequent in cases compared to controls (21% vs. 7%). The Gambian *GRIK1* SNP is located 3 478 base pairs downstream from the SAC *GRIK1* SNP and the Gambian *GRIK3* SNP is located 1 286 base pairs downstream from the SAC *GRIK3* SNP. *GRIK1* has been associated with susceptibility to diabetes [233], and according to T1DBase (a database focused on the genetics and genomics of type 1 diabetes susceptibility, http://www.t1dbase.org), *GRIK3* is also a putative diabetes susceptibility gene. Having diabetes increases susceptibility to TB [234], and this may explain the *GRIK1* - *GRIK3* interaction association we observed in the data.

Another model that was observed in both study groups is the interaction between *IL23R* (interleukin-23 receptor) and *ATG4C* (autophagy related 4C, cysteine peptidase) (model 4, SAC p-value 2.18×10^{-6} and Gambian p-value 0.0350). T helper 17 (Th17) cells are subsets of activated CD4+ (cluster of differentiation 4 plus) T cells that mediates the recruitment of macrophages to infected tissues. The Th17 response to *M. tuberculosis* infection is largely dependent on interleukin-23 [235]. ATG4C is thought to play a role in autophagy [236] and is up-regulated when TRPV1 (transient receptor potential cation channel subfamily V member 1) channels are expressed on CD4+ T cells [237; 238]. The *IL23R* and *ATG4C* gene products may therefore both be involved in the Th17 response to *M. tuberculosis*. The same SNPs are used in both models, with the CC-CC genotype combination decreasing the odds of having TB in both cohorts, compared to the CC-CT and CC-TT combinations. Both of the SNPs are located on chromosome 1p31 and are 6 centimorgans (4 451 385 base pairs) apart. Linkage disequilibrium between the SNPs is high in SAC controls ($D' = 0.5451$) but not in SAC cases ($D' = 0.0136$), and low in both Gambian controls and cases ($D' = 0.0011$ and $D' = 0.0386$ respectively).

Interaction between the *NELL1* (neural epidermal growth factor-like 1) and *CADM2* (cell adhesion molecule 2) genes is also evident in both the SAC and WTCCC study groups (model 13, SAC p-value 1.14×10^{-5} and Gambian p-value 0.0329), as well as interaction between the *NELL1* and *CADM3* (cell adhesion molecule 3) genes, although the latter was not validated in the Gambian study group (model 17, SAC p-value 1.26×10^{-5}). The same trend between the effects of the heterozygote genotype combinations is observed in both study groups for model 13, compared to the pairing with the wildtype homozygote genotype (CC) of the second SNP in the pair. A large degree of homology exists between the *CADM1*, *CADM2* and *CADM3* genes [239; 240] and CADM1 has been shown to affect the expression of interleukin-22 [241]. *NELL1* is expressed in pre-B cell development [242] and has been associated with inflammatory bowl disease, a complex auto-immune disorder [243]. The link between the interplay of these genes and TB susceptibility is however not clear. We also note that in the Gambian cohort, the NELL1 single SNP association signal is stronger than the interaction effect (single SNP p-value of 0.0022, see table 6.7.4, vs. interaction p-value of 0.0329).

The *CHST11* - *CHSY3* (carbohydrate (chondroitin 4) sulfotransferase 11 - chondroitin sulfate synthase 3) gene pair interaction was also detected in both study groups (model 19, SAC p-value 1.34×10^{-5} and Gambian p-value 0.0401). Uhlin et al. showed that expression of *CSPG* (chondroitin sulfate proteoglycan) decreased when monocyte-derived macrophages are treated with interferon-gamma [244]. CSPG is composed of a protein core and a chondroitin sulfate side chain. According to the NCBI BioSystems database, both the *CHST11* and *CHSY3* genes are involved in the chondroitin sulfate pathway. The CT-CG genotype combination has a higher frequency in SAC controls compared to cases (10% vs. 3%). The *CHST11* SNP is the same in the SAC and Gambian models, and the *CHSY3* Gambian SNP is 8 683 base pairs upstream from the SAC *CHSY3* SNP. The CT-TT combination was observed in 7 Gambian controls and in 1 case.

NF-κB (nuclear factor kappa-light-chain-enhancer of activated B cells) signalling plays in an important role in the host defense against *M. tuberculosis* infection [245]. Both the *SFTPD* (surfactant protein D) and *NOD2* (nucleotide-binding oligomerization domain containing 2) genes are involved in this pathway [246; 247]. Interaction between these genes was identified in the SAC study group (model 3), but no suitable SNPs were available for validation in the Gambian data set. The GG-CT genotype combination is present only in SAC controls (9%), whereas the GG-TT genotype combination is present only in SAC cases (8%).

The *ISG15* (ISG15 ubiquitin-like modifier) and *TLR8* (Toll-like receptor 8) gene products may also affect NF-κB signalling. ISG15 stimulates interferon-gamma production [248] which in turn activates NF-κB signalling [249]. It has also been postulated that TLR8 activates NF-κB signalling [250]. This gene pair showed interaction in the SAC study group (model 7) but could not be validated in the Gambian study group due to lack of suitable SNPs. The CT-GG genotype combination occurs in 7% of SAC cases, but in only 1% of controls.

Another model of interest that could not be validated in the Gambian cohort due to lack of suitable SNPs is the interaction between *NLRC5* (NLR family, CARD domain containing 5) and *IL12RB1* (interleukin 12 receptor, beta 1) genes (model 14). The interferon-gamma/interleukin-12 pathway is an important component of the immune defense against mycobacterial infections [251]. IL12RB1 is a receptor of interleukin-12, and it has been shown that the *NLRC5* promoter region is responsive to interferon-gamma, which implies that NLRC5 may function as a molecular switch of interferon-gamma activation [252]. The TT-AT genotype combination has a higher frequency in SAC controls compared to cases (12% vs. 5%), whereas the TT-TT combination has a higher frequency in SAC cases compared to controls (19% vs. 10%).

Finally, we explored whether allelic encoding of the SNPs may better explain the interactions detected in the SAC cohort. Table 6.7.5 summarizes the p-values of the four degrees of freedom genotypic tests for interaction that was used to select the top 20 models, as well as the p-values of the corresponding allelic models that attained the highest level of significance. It is evident from these results that dominant/recessive effects may in some cases better encapsulate the interaction effects observed in the data, and this is depicted in figure 6.7.4. We note that for all of the five models that were successfully validated in the Gambian cohort, the genotypic test for interaction achieved the highest level of significance.

6.5 Discussion

The South Africal Coloured population is an ideal cohort for the discovery of TB susceptibility genetic variants, since they received genetic contributions from diverse source populations that may differ in their susceptibility to TB. Seldin et al. has argued that it is important to study the role of complex disease epistasis in such admixed populations, and that this may well uncover novel interactions that are not detectable in the source populations [253]. In this study we used SAC genome-wide data as well as genotypes from a large number of candidate gene studies to discover genetic variants that may jointly modify the odds of having TB. We limited our search space to biologically plausible gene pair models and used statistical modelling to detect interactions, allowing us to adjust for known differences between cases and controls (age, gender and ancestry). Our study does however have a number of limitations, which we discuss below.

Genotypes available for testing the gene pair models were limited to SNPs that were genotyped on the Affymetrix 500K SNP chip as well as candidate gene studies performed in our group. The Affymetrix chip was originally designed based on LD patterns in European popu-

lations, and as a result the proportion of variants that are tagged in African populations may be much reduced [254].

Minor allele frequencies of the SNPs representing genes in the top models are in general quite different between the SAC and Gambian cohorts (table 6.7.4). Of the five models that were successfully validated, only one of the models was validated using exactly the same SNPs, three models were validated with one SNP in common, and one model was validated with completely different SNPs. Patterns of LD are also likely to differ between the SAC and Gambian cohorts, and according to NCBI, none of the SNPs in the top result set have functional effects, implying that the SNPs may all be tagging causative variants. Due to these factors, it is difficult to compare the effect sizes between the two cohorts directly. Indeed, two studies of the association between rs1024611 and TB susceptibility found that the association was statistically significant, but that the G allele of the SNP was protective in the one population, and increased susceptibility in the other population. The true causal variant that rs1024611 was in LD with was later identified, which may explain the opposite effects observed in the two populations [66]. The complexity of disentangling such different effects would be exacerbated in the context of interaction modelling. This could be alleviated to some degree if a higher density of markers was available, which would better capture causative variants. A denser marker panel could be imputed, but in our opinion, this exercise would likely be error-prone. Additional uncertainty would be introduced through imputation, and the proportion of genotype inaccuracies could potentially be large. Imputation relies on linkage disequilibrium between markers, which may not be captured accurately by the Affymetrix 500K SNP for our study cohorts, as a result of the chip's European-centric design. In addition, the San has contributed a large amount of genetic material to the SAC [26; 39; 99; 9; 40; 41], and due to the lack of large high density reference panels for the San, this may contribute to additional inaccuracies in imputation of the SAC data set.

It is difficult to quantify the precise levels of significance of our results, due to the large number of tests in the SAC data set, and the limitations of methods available to correct for multiple testing. If we were to use a multiple testing correction similar to the one used by Emily et al. [216], despite its limitations, we would have to show that the number of effective independent tests was 60 000, for the topmost model to be significant at a Bonferroni adjusted alpha level of 0.05 ($0.05/60000 = 8.33 \times 10^{-7}$, p-value of model 1 was 8.33×10^{-7}). Given that roughly 2 million models were tested, and that Emily et al. found that the number of effective independent tests was approximately six times less than the actual number of tests, it is unlikely that we would be able to demonstrate this (a 33 times reduction would be required). The SAC genome-wide data set was originally genotyped with a view to perform a case-only admixture mapping study, and for this reason, a limited number of controls was available for many of the two SNP interaction tests. Whilst the available group sizes are sufficient to detect two-SNP interactions, it is unlikely that any of the results would achieve statistical significance. For validation models fitted to the Gambian data set, we describe tests with p-values < 0.05 as statistically significant. As 20 gene pair tests were done (ranging between 2 to 90 SNP pair tests per gene pair, see table 6.1), none of these results would survive correction for multiple testing, and we note that the results should be interpreted with caution. We do however argue that both the SAC and Gambian data sets do not merely constitute random data, and that our results may contain actual associations that should not be dismissed [152]. Given the complex nature of the immune system defence against TB and the role that gene-gene interactions might play in this, it is plausible that some of our top results represent real biological phenomena, that may be worthy of further investigation.

Seven of the top models identified in the SAC study group could not be tested in the Gambian study group due to the absence of suitable SNPs. Eight of the models were not successfully

validated. These results could be false positives, but their validation failure could also be ascribed to a number of other reasons. Differing patterns of LD between the SAC and Gambian populations and lack of SNP coverage by the Affymetrix 500K SNP chip, which our SNP selection strategy could not fully compensate for, could result in unsuccessful validation. The *M. tuberculosis* genome varies substantially across geographic regions [59], including between South and West Africa, and it has been hypothesised that interactions between host and pathogen differ between population groups [65; 59; 174]. Due to the heterogeneity of the source populations that contributed to the formation of the SAC, it is also possible that some interactions involved in TB susceptibility are unique to the SAC [253]. In spite of these limitations, five of the top twenty models were indeed validated in an independent Gambian case-control data set, although the levels of significance of the validation models were not very small (p-value < 0.05 but > 0.01). These models indicate that TB susceptibility is modified by interplay between the *NRG1* - *NRG3*, *GRIK1* - *GRIK3* and *IL23R* - *ATG4C* gene pairs, and the fact that the validation population is ethnically very different could imply that the interactions found have universal relevance.

The frequencies and effects are depicted graphically to aid interpretation, but as the SNPs used in the models are tag variants that may not have causative functional effects, the biological implications of the models are not yet fully understood. Validation in other populations, fine-mapping of the causal variants and functional studies will be required to elucidate our findings.

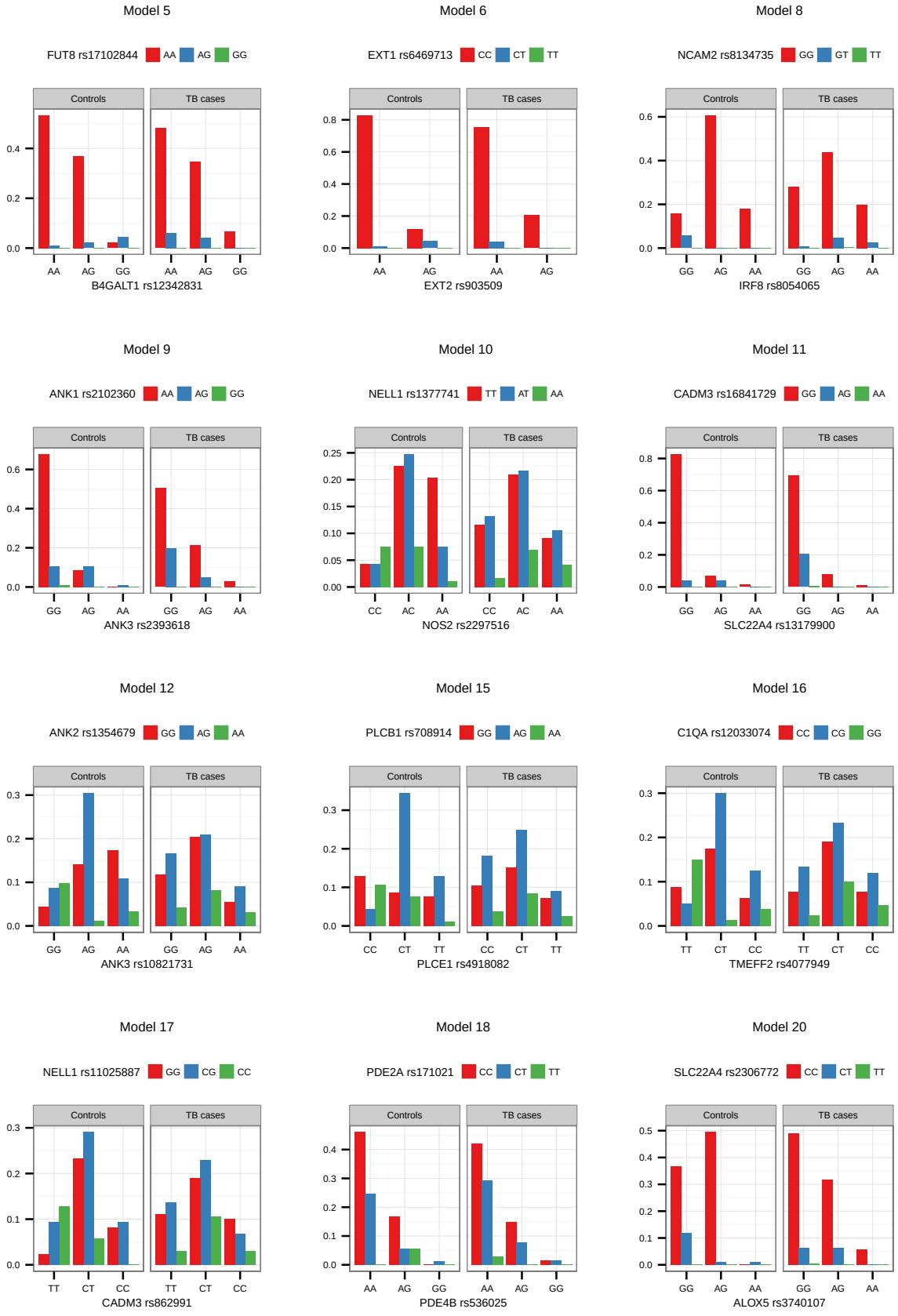
6.6 Conclusion

In this study we investigated the role of gene-gene interactions in TB susceptibility in the South African Coloured population. To our knowledge, in terms of number of genetic loci considered, this is the largest study of gene-gene interactions and TB susceptibility that has been reported to date. We report a number of interesting results, five of which were validated in an independent cohort from the Gambia.

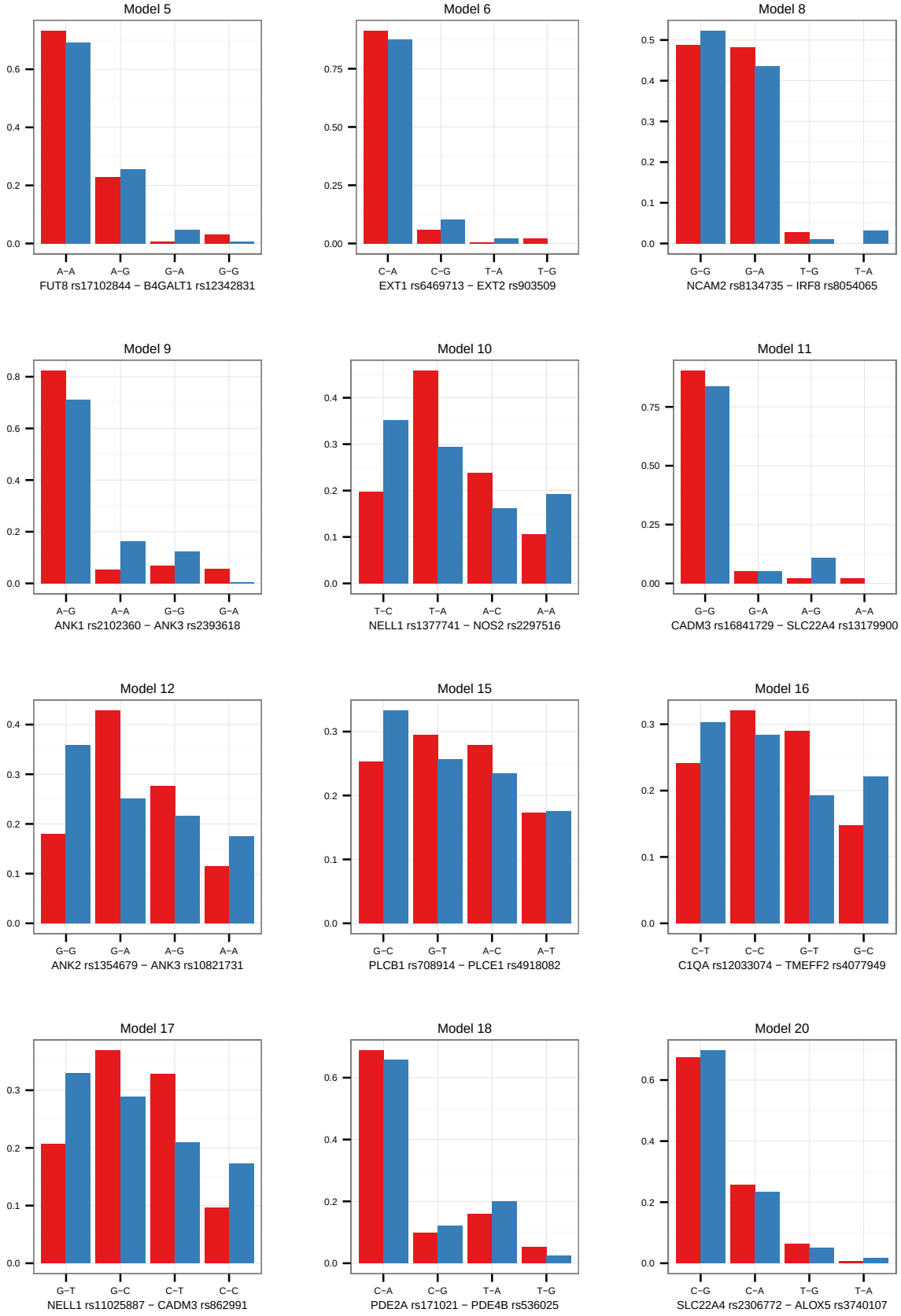
Acknowledgements

We thank all participants and field workers in this study. We also thank the developers of the open source software we used in our analyses.

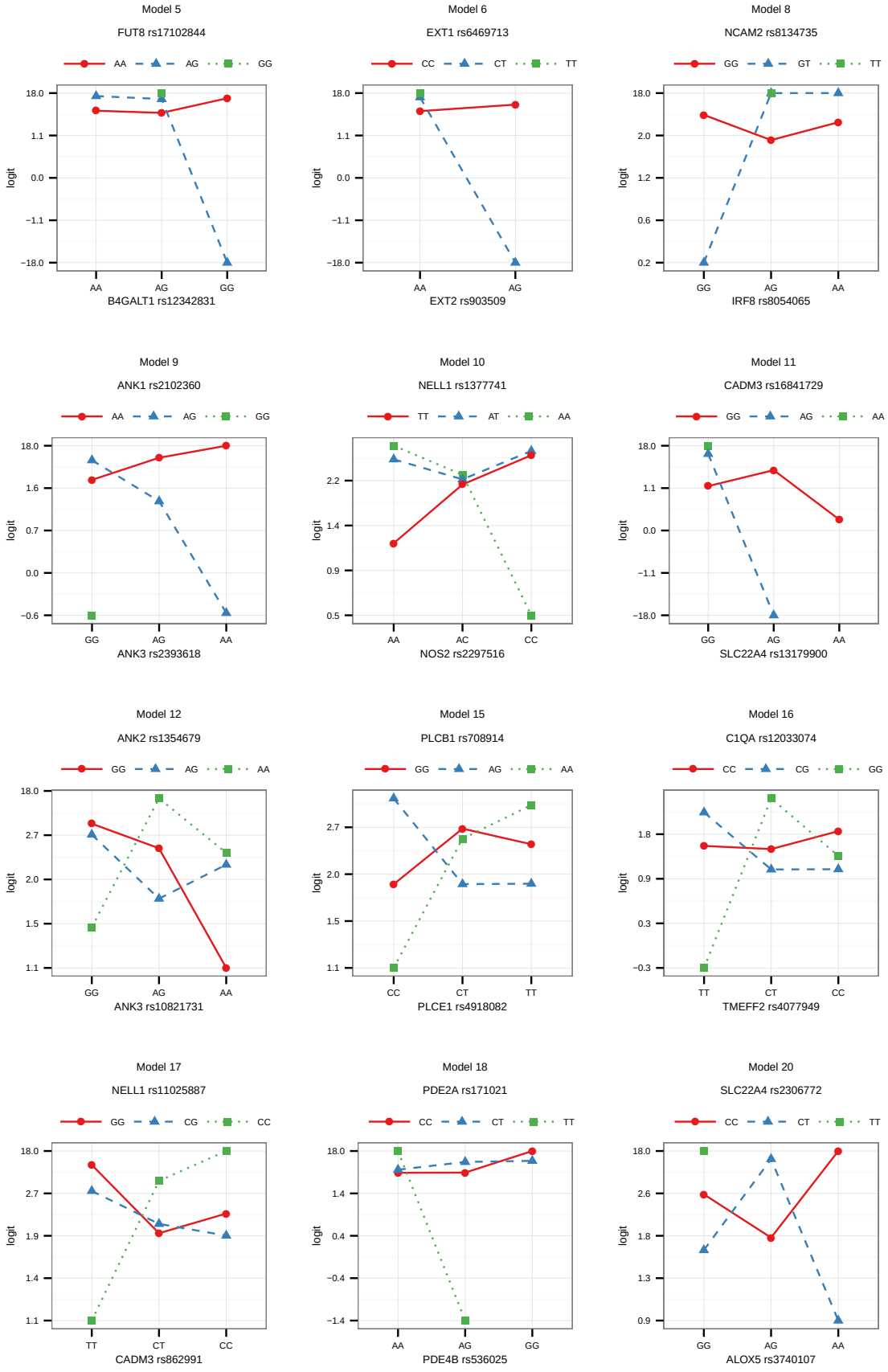
6.7 Supplementary figures and tables



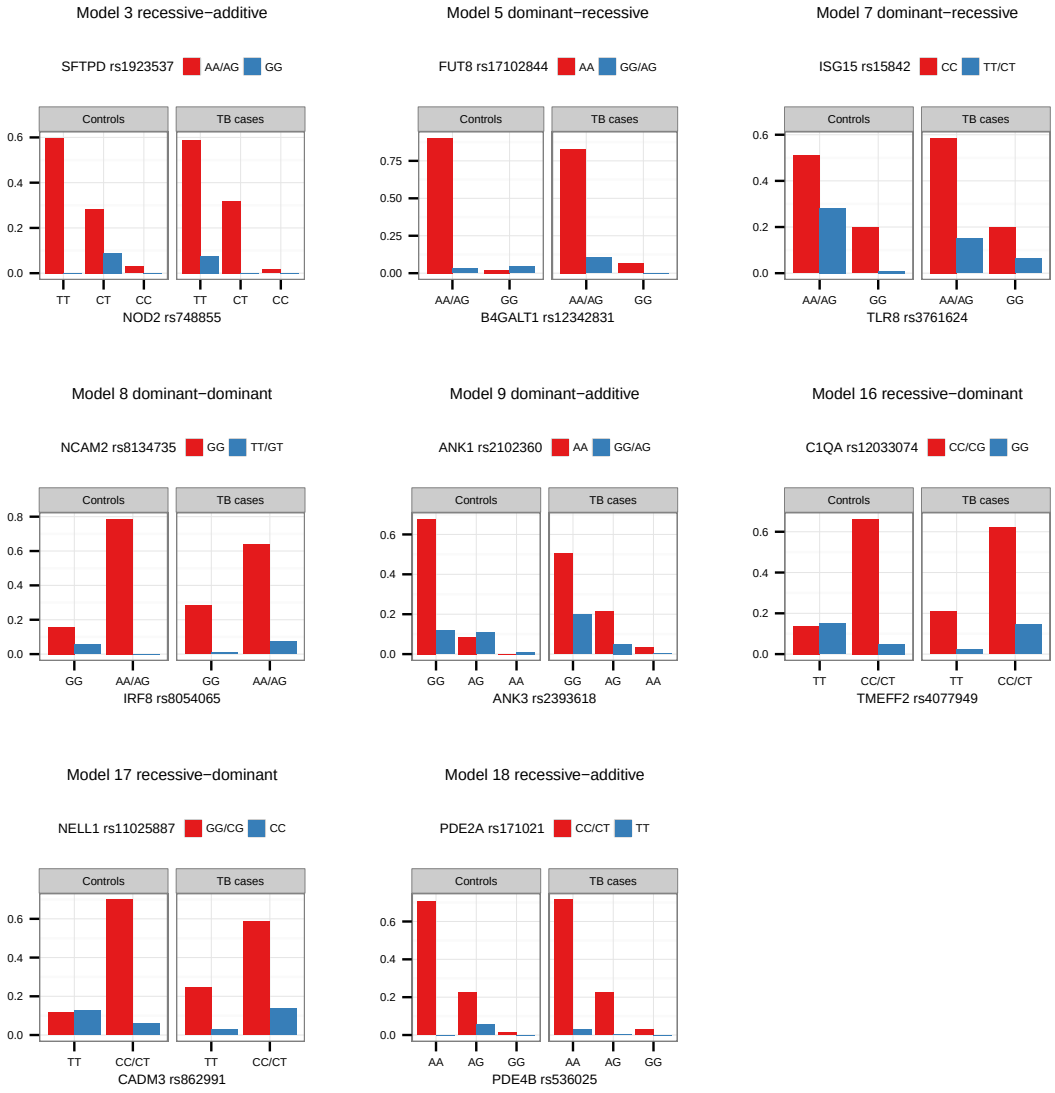
Supplementary Figure 6.7.1: Genotype combination proportions in the SAC study group. The observed proportions of the nine possible SNP pair genotype combinations from models 5, 6, 8, 9, 10, 11, 12, 15, 16, 17, 18 and 20 are depicted in this figure, per cases and controls. Genotypes are ordered according to minor allele frequency, with the wildtype homozygote appearing first, and the rare homozygote appearing last.



Supplementary Figure 6.7.2: Allele combination frequencies in the SAC study group. The frequencies of the four possible SNP pair allele combinations from models 5, 6, 8, 9, 10, 11, 12, 15, 16, 17, 18 and 20 are depicted in this figure, per cases and controls. The frequencies were estimated using an EM-algorithm.



Supplementary Figure 6.7.3: Effects in the SAC study group. The logits of genotype combinations from models 5, 6, 8, 9, 10, 11, 12, 15, 16, 17, 18 and 20 are depicted in this figure. Genotypes are ordered according to minor allele frequency, with the wildtype homozygote appearing first, and the rare homozygote appearing last. Non-parallel lines are indicative of interaction effects. The effects were estimated by absorbing the marginal effects of the SNPs into the SNP × SNP interaction term, and adjusting for the covariates included in the model by averaging over them.



Supplementary Figure 6.7.4: Dominant/recessive combination proportions in the SAC study group. The observed proportions of SNP pair genotype combinations from models 3, 5, 7, 8, 9, 16, 17 and 18 are depicted in this figure, per cases and controls. Recessive/dominant effects in these models may better explain the interactions observed in the cohort (smaller p-values were achieved compared to the genotypic models, and the best models with 1 or more recessive or dominant encodings listed in table 6.7.4 are presented in this figure). Rare homozygotes and heterozygotes are combined to represent dominant encoding of alleles, and wild type homozygotes and heterozygotes are combined to represent recessive encoding of alleles. For dominant and recessive allelic encodings of SNPs, the last genotype presented therefore reflect an encoding of 1.

Table 6.7.1: TB susceptibility candidate gene association studies. The table summarizes the total number of samples that were successfully genotyped in each candidate gene study and how many samples have complete confounder information (age, gender and ancestry).

			TB cases		Controls	
Study	**Year**	**Genes**	**Total**	**Complete**	**Total**	**Complete**
Rossouw et al. [160]	2003	IFNG	393	302	286	231
Hoal et al. [48]	2004	SLC11A1, SLC11A2	429	324	436	288
Babb et al. [161]	2007	SP110	379	334	304	268
Barreiro et al. [137]	2007	CD209	339	286	227	202
Möller, unpublished [162], Möller et al. [163; 164; 136; 138; 139; 141]	2007-2011	ATG16L1, BTNL2, NOD2, CCL2, CTLA4, CTSZ, FCRL3, FZD5, IL10, IL12B, IL12RB1, IL12RB2, IL18, IL1RN, IL23R, IL4, IL6ST, INSIG2, ABCB1, CIITA, MS4A2, NELL1, NOS2, P ADI4, PPARG, PTGER4, PTPN22, RUNX1, SH2D1A, SLC22A4, SLC22A5, SOCS3, TEX264, TLR2, TLR4, TNF, TNFRSF1A, TNFRSF1B, TNFSF15, WNT5A	781	584	703	390
Adams et al. [141]	2011	MC3R	439	386	505	410
Babb, unpublished [165]	2007	CCR5, CCL5, CXCL12	312	229	228	178
De Wit, unpublished [166]	2009	IFNGR1, IL8, RANTES	397	303	289	231
Salie, unpublished [167]	2010	ANXA11, CADM1, CADM2, CADM3, NCAM2	382	341	398	344
Lucas, unpublished [168]	2011	TLR8, TLR9	479	439	496	447
Wagman, unpublished [169]	2011	MARCO, SFTPD	383	343	395	343
Bruiners, unpublished [170]	2012	C1QA, C1QB	472	433	477	431
Unpublished	2001	SPP1	206	148	124	64
Unpublished	2002	SFTPD	161	124	144	75
Unpublished	2003	IL8	214	156	149	112
Unpublished	2003	IFNGR1, IFNGR2	373	275	348	226
Unpublished	2004	CO2REGION, HS3ST4	557	441	475	292
Unpublished	2007	FOXP3	502	414	513	351
Unpublished	2008	P2RX7, TLR1	494	416	525	401
Unpublished	2010	CD14	387	341	406	321
Unpublished	2012	APOE	435	410	443	407
Unpublished	2013	IRGM, ISG15, NLRC5, NLRP3, NOD2	427	407	439	403

Table 6.7.2: Age and gender in the tuberculosis study groups. P-values were calculated using logistic regression.

			Age		Gender	
		Sample size	Mean ± SD	P-value	Nr (Prop)	P-value
SAC candidate gene data	TB cases	918	36.00 ± 12.79	< 0.0001	493 (0.54)	< 0.0001
	Controls	507	32.48 ± 10.69		122 (0.24)	
SAC chip data	TB cases	642	36.69 ± 11.50	< 0.0001	361 (0.56)	0.6210
	Controls	91	31.47 ± 4.09		45 (0.49)	
Gambian data	TB cases	1156			823 (0.72)	< 0.0001
	Controls	1206			574 (0.48)	

Table 6.7.3: Software used in this study. A summary listing web URLs, version information and important parameter settings of software used in this study.

Program	Web URL	Version	Parameters
PLINK	http://pngu.mgh.harvard.edu/∼purcell/plink/	v1.07	*–indep-pairwise 50 10 0.1* was used for LD filtering; a custom version was used for the *fast-epistasis* tests
Welcome Trust strand and build scripts	http://www.well.ox.ac.uk/∼wrayner/strand/		The Affy-NSP-STY-b37.58-v4 strand and position files were used
ADMIXTURE	http://www.genetics.ucla.edu/software/admixture/	1.21	*K=5*
EIGENSOFT	http://genetics.med.harvard.edu/reich/Reich_Lab/Software.html	5.0.1	
Biofilter	http://ritchielab.psu.edu/software/biofilter-download	2.1.0	LOKI database was built on 5 Dec 2013
R	http://www.r-project.org	3.1.0	
effects R package	http://cran.r-project.org/web/packages/effects/index.html	3.0-1	The *effect()* function was used
genetics R package	http://cran.r-project.org/web/packages/genetics/index.html	1.3.8.1	The *HWE.exact()* and *LD* functions were used
ggplot2 R package	http://cran.r-project.org/web/packages/ggplot2/index.html	1.0.0	
haplo.stats R package	http://cran.r-project.org/web/packages/haplo.stats/index.html	1.6.8	The *group.bin()* function was used

Table 6.7.4: Single SNP summary of the top model SNPs in the SAC and Gambian cohorts. This table provides a summary of each SNP's individual minor allele frequency (MAF) and association with having TB.

			SAC			Gambian		
	Gene	**Region**	**SNP**	**MAF**	**P-value**	**SNP**	**MAF**	**P-value**
Model 1	NRG1	8p12	rs16879814	0.22	0.8200	rs16879814	0.23	0.3952
	NRG3	10q23	rs11191757	0.33	0.1287	rs2224109	0.13	0.3365
Model 2	GRIK1	21q22	rs465555	0.49	0.4642	rs460583	0.42	0.1397
	GRIK3	1p34	rs3738085	0.39	0.8017	rs476894	0.40	0.4022
Model 3	SFTPD	10q22	rs1923537	0.28	0.8429			
	NOD2	16q12	rs748855	0.18	0.4041			
Model 4	IL23R	1p31	rs10489628	0.49	0.7888	rs10489628	0.12	0.1323
	ATG4C	1p31	rs11208029	0.11	0.3534	rs11208029	0.41	0.4644
Model 5	FUT8	14q23	rs17102844	0.05	0.3336	rs9323464	0.47	0.0140
	B4GALT1	9p13	rs12342831	0.25	0.8881	rs10758189	0.18	0.0358
Model 6	EXT1	8q24	rs6469713	0.02	0.8728			
	EXT2	11p11	rs903509	0.10	0.6867			
Model 7	ISG15	1p36	rs15842	0.14	0.0149			
	TLR8	Xp22	rs3761624	0.38	0.5738			
Model 8	NCAM2	21q21	rs8134735	0.04	0.7718	rs8132838	0.23	0.2967
	IRF8	16q24	rs8054065	0.47	0.1589	rs147968	0.18	0.1974
Model 9	ANK1	8p11	rs2102360	0.13	0.2941			
	ANK3	10q21	rs2393618	0.16	0.3319			
Model 10	NELL1	11p15	rs1377741	0.35	0.0798	rs1377741	0.39	0.1964
	NOS2	17q11	rs2297516	0.50	0.2571	rs2314809	0.46	0.3252
Model 11	CADM3	1q21	rs16841729	0.10	0.0562			
	SLC22A4	5q31	rs13179900	0.06	0.7064			
Model 12	ANK2	4q25	rs1354679	0.39	0.3830	rs1354679	0.31	0.7538
	ANK3	10q21	rs10821731	0.45	0.1023	rs10761481	0.43	0.8734
Model 13	NELL1	11p15	rs4614448	0.40	0.7614	rs4614448	0.43	0.0022
	CADM2	3p12	rs17024414	0.19	0.4179	rs17024876	0.46	0.2754
Model 14	NLRC5	16q13	rs289726	0.48	0.3474			
	IL12RB1	19p13	rs393548	0.19	0.4955			
Model 15	PLCB1	20p12	rs708914	0.42	0.7643	rs1703634	0.45	0.4542
	PLCE1	10q23	rs4918082	0.44	0.8491	rs4918082	0.48	0.7421
Model 16	C1QA	1p36	rs12033074	0.42	0.7034			
	TMEFF2	2q32	rs4077949	0.50	0.9403			
Model 17	NELL1	11p15	rs11025887	0.39	0.9114	rs12577018	0.47	0.2634
	CADM3	1q21	rs862991	0.45	0.6270	rs862991	0.31	0.6527
Model 18	PDE2A	11q13	rs171021	0.23	0.5357	rs3781931	0.16	0.9916
	PDE4B	1p31	rs536025	0.14	0.2809	rs17423910	0.05	0.4467
Model 19	CHST11	12q23	rs17036205	0.04	0.9026	rs17036205	0.22	0.2632
	CHSY3	5q23	rs32225	0.41	0.8596	rs244745	0.03	0.9337
Model 20	SLC22A4	5q31	rs2306772	0.07	0.6799	rs3792880	0.13	0.2623
	ALOX5	10q11	rs3740107	0.25	0.0164	rs3780909	0.49	0.3145

Table 6.7.5: P-values of the top models in the SAC cohort. The genotypic model p-values, which were used to select the top 20 models, are presented in this table. The p-values of the corresponding allelic interaction models that achieved the smallest p-values are also shown.

		Genotypic	Allelic	
	Model	**P-value**	**Best model**	**P-value**
1	NRG1_rs16879814 - NRG3_rs11191757	8.32×10^{-7}	dominant-recessive	8.80×10^{-6}
2	GRIK1_rs465555 - GRIK3_rs3738085	1.62×10^{-6}	additive-recessive	4.78×10^{-6}
3	SFTPD_rs1923537 - NOD2_rs748855	1.89×10^{-6}	recessive-additive	5.57×10^{-7}
4	IL23R_rs10489628 - ATG4C_rs11208029	2.18×10^{-6}	dominant-dominant	8.53×10^{-6}
5	FUT8_rs17102844 - B4GALT1_rs12342831	2.54×10^{-6}	dominant-recessive	4.18×10^{-7}
6	EXT1_rs6469713 - EXT2_rs903509	2.67×10^{-6}	additive-additive	2.49×10^{-6}
7	ISG15_rs15842 - TLR8_rs3761624	6.23×10^{-6}	dominant-recessive	1.15×10^{-6}
8	NCAM2_rs8134735 - IRF8_rs8054065	8.06×10^{-6}	dominant-dominant	1.32×10^{-6}
9	ANK1_rs2102360 - ANK3_rs2393618	8.96×10^{-6}	dominant-additive	3.20×10^{-6}
10	NELL1_rs1377741 - NOS2_rs2297516	8.98×10^{-6}	additive-additive	1.09×10^{-6}
11	CADM3_rs16841729 - SLC22A4_rs13179900	9.62×10^{-6}	dominant-dominant	1.38×10^{-5}
12	ANK2_rs1354679 - ANK3_rs10821731	1.14×10^{-5}	recessive-dominant	3.79×10^{-5}
13	NELL1_rs4614448 - CADM2_rs17024414	1.14×10^{-5}	recessive-dominant	1.20×10^{-4}
14	NLRC5_rs289726 - IL12RB1_rs393548	1.16×10^{-5}	recessive-dominant	1.26×10^{-4}
15	PLCB1_rs708914 - PLCE1_rs4918082	1.20×10^{-5}	recessive-dominant	7.17×10^{-4}
16	C1QA_rs12033074 - TMEFF2_rs4077949	1.25×10^{-5}	recessive-dominant	6.91×10^{-6}
17	NELL1_rs11025887 - CADM3_rs862991	1.26×10^{-5}	recessive-dominant	1.53×10^{-6}
18	PDE2A_rs171021 - PDE4B_rs536025	1.29×10^{-5}	recessive-additive	7.80×10^{-7}
19	CHST11_rs17036205 - CHSY3_rs32225	1.34×10^{-5}	dominant-dominant	1.18×10^{-3}
20	SLC22A4_rs2306772 - ALOX5_rs3740107	1.37×10^{-5}	dominant-dominant	8.29×10^{-4}

Chapter 7

Research Article 4

Using multi-way admixture mapping to elucidate TB susceptibility in the South African Coloured population

Michelle Daya[1], Lize van der Merwe[1], Christopher R. Gignoux[2], Paul D. van Helden[1], Marlo Möller[1], Eileen G. Hoal[1,*]

1 Molecular Biology and Human Genetics, MRC Centre for Molecular and Cellular Biology and the DST/NRF Centre of Excellence for Biomedical TB Research, Faculty of Medicine and Health Sciences, Stellenbosch University, Tygerberg, South Africa
2 Department of Genetics, Stanford University, Stanford, California, United States of America
∗ E-mail: Corresponding egvh@sun.ac.za

This article was accepted by the journal BMC Genomics (November 2014).

7.1 Abstract

Background

The admixed South African Coloured population is ideally suited to the discovery of tuberculosis susceptibility genetic variants and their probable ethnic origins, but previous attempts at finding such variants using genome-wide admixture mapping were hampered by the inaccuracy of local ancestry inference. In this study, we infer local ancestry using the novel algorithm implemented in RFMix, with the emphasis on identifying regions of excess San or Bantu ancestry, which we hypothesize may harbour TB susceptibility genes.

Results

Using simulated data, we demonstrate reasonable accuracy of local ancestry inference by RFMix, with a tendency towards miss-calling San ancestry as Bantu. Regions with either excess San ancestry or excess African (San or Bantu) ancestry are less likely to be affected by this bias, and we therefore proceeded to identify such regions, found in cases but not in controls (642 cases and 91 controls). A number of promising regions were found (overall p-values of 7.19×10^{-5} for San ancestry and $< 2.00 \times 10^{-16}$ for African ancestry), including chromosomes 15q15 and 17q22, which are close to genomic regions previously implicated in TB. Promising immune-related susceptibility genes such as the *GADD45A*, *OSM* and *B7-H5* genes are also harboured in the identified regions.

Conclusion

Admixture mapping is feasible in the South African Coloured population and a number of novel TB susceptibility genomic regions were uncovered.

7.2 Background

The South African Coloured population (SAC) is a so-called admixed population that derived its origins from the diverse population groups that settled in the early Cape colony, including the indigenous San, early European settlers, slaves that were imported from Indonesia, India and other parts of Africa, and South African Bantu-speakers who later migrated to the area. Previous genetic research has shown that the SAC received ancestry contributions from click-speaking Africans (San), Bantu-speaking Africans, Europeans and South and East Asians, which is consistent with the historical records [26; 39; 9; 40; 41]. A high degree of heterogeneity in ancestral contributions between SAC individuals has also been illustrated previously [26; 99; 39]. The admixture that occurred in the SAC is therefore complex, constituting a number of different source ancestries, with dissimilar genetic distances between them.

Our study group of SAC individuals was recruited from metropolitan areas in Cape Town that have some of the highest reported incidences of tuberculosis (TB) worldwide, despite extensive BCG vaccination and low prevalence of HIV [148]. As the group received contributions from diverse source populations that may differ in their genetic susceptibility to TB, the group is ideally suited to the discovery of TB susceptibility genetic variants and their probable ethnic origins. Our previous work has shown that African ancestry in this group is associated with higher risk of TB infection, whereas European and Asian ancestries are protective [100; 98]. Areas of the genome with African ancestry that is much higher than the norm in a group of TB cases may therefore harbour genetic variants that increase the risk of developing TB. The process of finding such areas is known as admixture mapping, and this technique relies on the accurate inference of what is known as local ancestry per individual across their genome [32].

When admixture occurs between two or more population groups that were previously isolated, recombination events result in chromosomes that are a mosaic of blocks of ancestry deriving from different source populations. Given genetic data of an admixed individual and their source populations, statistical techniques can be used to determine the bounds of these segments and to assign the most probable source ancestries to them. These techniques rely on the probability of recombination events to distinguish the bounds of segments, and differences in allele and haplotype frequencies between source populations for classification of the ancestry of segments. The process is known as local ancestry inference (LAI).

In a previous study, Chimusa et. al concluded that accurate multi-way LAI was not feasible in the SAC using the LAI algorithms available at the time [98]. In this study, we re-evaluate this position using the novel LAI algorithm implemented in the RFMix software package, focusing on the classification of San and Bantu ancestry. These ancestries are of particular interest as Southern African populations were not exposed to modern strains of *Mycobacterium tuberculosis* (*M. tuberculosis*), the most prevalent in our SAC study group [158], until the recent past [76]. The relative lack of exposure of the SAC and Bantu populations to modern strains of *M. tuberculosis* could possibly have resulted in decreased resistance to developing the disease, especially in densely populated areas with low socio-economic conditions. Supporting this argument, a significant positive association between San ancestry and TB susceptibility in the SAC was found by Chimusa et. al. [98]. The association was confirmed in an independent sample in a later study, which also found a positive association with Bantu ancestry, although it was relatively weak [100]. Although each of the non-African ancestry components of the SAC

Table 7.1: Percentage of miss-called ancestry. This table summarizes the percentage of SNPs that were miss-called by LAMP-LD and RFMix per each of the six possible miss-call categories. The known ancestry of a simulated data set of 1500 SAC chromosomes was compared to the ancestry called by the software program (chromosome 1). The median percentage of miss-called SNPs across all SNPs as well as the median percentage of miss-called SNPs across SNPs of that source ancestry are shown. San ancestry can for example be miss-called as either Bantu or non-African ancestry. The median percentage of all SNPs that were miss-called as such are shown in the second and third columns of the first two rows, and the median percentage of San SNPs that were miss-called as such are shown in the fourth and fifth columns of the first two rows. The mean proportion of San, Bantu and non-African ancestry in the simulated data set was 0.3342, 0.2772 and 0.3885 respectively. The difference in number of SNPs miss-called by RFMix, compared to the corresponding number of SNPs miss-called by LAMP-LD, were significant with p-values $< 2 \times 10^{-16}$ for each of the six possible miss-call categories.

	Percentage of Total [IQR]		Percentage of Ancestry [IQR]	
Type of miss-call	**LAMP-LD**	**RFMix**	**LAMP-LD**	**RFMix**
San as Bantu	4.10 [1.65-9.76]	1.95 [0.88-3.55]	13.28 [5.12-35.16]	6.19 [3.13-10.45]
San as non-African	1.11 [0.05-9.58]	0.27 [0.11-0.66]	3.12 [0.18-36.09]	0.89 [0.36-2.05]
Bantu as San	0.08 [0.00-5.49]	0.04 [0.00-0.13]	0.32 [0.00-21.54]	0.14 [0.00-0.49]
Bantu as non-African	0.45 [0.02-8.63]	0.09 [0.02-0.23]	2.36 [0.07-33.44]	0.37 [0.09-0.92]
non-African as San	0.14 [0.00-8.31]	0.09 [0.03-0.19]	0.42 [0.00-25.87]	0.25 [0.08-0.54]
non-African as Bantu	0.95 [0.14-9.66]	0.18 [0.07-0.33]	3.09 [0.35-28.68]	0.47 [0.20-0.92]

(European, South Asian and East Asian) were negatively associated with TB susceptibility when tested in individual models, these associations were no longer significant when all five ancestry components were tested together.

In this study, we first explore the accuracy of LAI in the SAC using RFMix and compare its performance to other algorithms. After quantifying this using simulated data, we proceed to identify regions with excess San or Bantu ancestry found in TB cases but not in controls, and hypothesize that these regions may contain genes that affect TB susceptibility.

7.3 Results

7.3.1 Evaluating multi-way LAI accuracy using simulated data

Chimusa et al. previously evaluated the accuracy of LAI in the SAC using various software programs and found that LAMP-LD performed best [98]. As RFMix was not available at that time, our first step was to compare the accuracies of LAMP-LD and RFMix using a simulated data set of 1500 SAC chromosomes (chromosome 1). LAI was run five-way, but since only San and Bantu genome-wide ancestry is independently associated with TB susceptibility [100; 98], ancestry of SNPs that were called as European, South Asian or East Asian were labelled as non-African. The percentage of SNPs for which the called ancestry matched the known ancestry was 69.43% and 96.43% for LAMP-LD and RFMix respectively. The percentage of SNPs per type of miss-called ancestry is summarized in table 7.1, which shows that RFMix offers a significant improvement in the calling of local ancestry, especially when distinguishing San and Bantu ancestry.

Histograms of the difference between the mean ancestry called by RFMix per chromosome and the known mean ancestry in the simulated data set are shown in supplementary figure 7.7.1. Ancestry called by RFMix was on average 2.71% lower than the known San ancestry, whilst Bantu ancestry was on average 2.45% higher. This discrepancy can be ascribed to the relatively large proportion of San SNPs that were miss-called as Bantu. Non-African ancestry calls compared well to known values (on average only 0.26% higher).

Could chromosomal segments with large deviations from the mean ancestry be the result

Table 7.2: Correlation between the number of miss-called ancestry segments and deviation in ancestry. This table summarizes the correlation between the number of ancestry miss-calls that occurred at a segment of ancestry, per each of the six possible miss-call categories, and the deviation in local ancestry of the segment. Miss-called ancestry was identified by comparing the known ancestry of a simulated data set of 1500 SAC chromosomes to the ancestry called by RFMix (chromosome 1). Deviations in ancestry were calculated by subtracting the overall mean RFMix ancestry from the local ancestry of each segment, for each of the three source ancestries (San, Bantu, non-African).

	Deviation		
Number miss-called	**San**	**Bantu**	**non-African**
San as Bantu	−0.83	+0.89	+0.05
San as non-African	−0.39	+0.19	+0.43
Bantu as San	+0.01	+0.09	−0.17
Bantu as non-African	−0.02	+0.13	−0.21
non-African as San	+0.01	+0.06	−0.14
non-African as Bantu	−0.11	+0.24	−0.21

of LAI errors? To answer this question, the simulated chromosome 1 data set was divided into segments by determining the positions where ancestry switches occur (see *Materials and Methods - Delineating called ancestry segments*), yielding 1077 segments. Deviation from the overall RFMix mean ancestry was calculated for each of the ancestries, for each of the segments. Table 7.2 summarizes the correlation between the number of miss-called ancestry segments and deviation in ancestry. Segments with lack of San ancestry are associated with San segments that are miss-called as Bantu or non-African, whereas segments with excess Bantu ancestry are associated with Bantu segments miss-called as San. The relationship between number of errors in segments and deviation in ancestry are further depicted in supplementary figures 7.7.2, 7.7.3 and 7.7.4. Although errors appear to be distributed fairly evenly across the entire chromosome, segments with lack of San ancestry generally have more errors. The exception is a large number of errors that occurred around the centromere (supplementary figure 7.7.4), likely due to the dearth of SNPs in this region.

The negative effect that short tracts of ancestry and a large degree of admixture could have on LAI accuracy was explored next. Supplementary figure 7.7.5 shows the distribution of the length of tracts of ancestry in the simulated data and the proportion of SNPs with miss-called ancestry per tract. Tracts that were completely miss-called (all the SNPs in the tract were assigned incorrect ancestry by the LAI) occurred far more frequently in very short tracts of ancestry, and longer tracts of ancestry correlated with a smaller proportion of miss-called SNPs (r=-0.2906, p-value$< 2.00 \times 10^{-16}$). Supplementary figure 7.7.6 shows that there is a positive correlation between the number of tracts of ancestry on a chromosome and the number of miss-called SNPs for that chromosome (r=0.1847, p-value=2.54×10^{-13}), indicating that inferring local ancestry may be more error-prone for chromosomes with a large degree of admixture.

7.3.2 Local ancestry across the genome

We proceeded to run LAI on our study group of 733 unrelated SAC individuals using RFMix (642 TB cases and 91 controls). Figure 7.1 depicts local ancestry across the genome for cases and controls. The mean genome-wide San ancestry calculated from the local ancestry estimates was 0.2304 and 0.1847 for cases and controls respectively, the mean Bantu ancestry was 0.3792 (cases) and 0.3391 (controls), and the mean non-African ancestry was 0.3904 (cases) and 0.4761 (controls). Mean San ancestry was on average 12.10% lower than corresponding ADMIXTURE estimates and Bantu ancestry was on average 10.80% higher, whilst the mean non-African ancestry was comparable (only 1.31% higher). The differences in ancestry estimates are illustrated in supplementary figure 7.7.7. We speculate that the large discrepancy

between San and Bantu estimates can in part be ascribed to some of the San ancestry in the SAC being contributed by southern African Bantu populations such as the Xhosa [26; 39]. Admixture between the San and these populations likely occured during the Bantu expansion [46]. The relative older age of these admixture events would result in short tracts of San ancestry, which are harder to distinguish with genotype array data [255], also evident from our simulated data (supplementary figure 7.7.5). The distributions of called San and European tract lengths are comparable, supporting our conclusion that very short tracts of San ancestry may not have been identified, whereas the Bantu tract lengths are longer (supplementary figure 7.7.8). Xhosa admixture into the nascent SAC population likely occurred later than admixture between the San and Europeans [26], helping to explain the longer Bantu tract lengths.

RFMix output was divided into 13 860 segments of local ancestry and local ancestry deviations were calculated. Histograms of the local ancestry deviations in cases and controls are depicted in supplementary figure 7.7.9 and boxplots of the deviations are shown in supplementary figure 7.7.10. Similar to our findings in simulated data, the figures suggest that the tails of the deviation distributions are biased towards lack of San ancestry and excess Bantu and non-African ancestry.

7.3.3 Regions with excess San or Bantu ancestry in cases relative to controls

Previous work has shown that San and Bantu ancestry is associated with increased susceptibility to TB in the SAC, and that the non-African ancestry components (European, South Asian and East Asian) are protective. The non-African ancestry components were however not associated with TB susceptibility after adjustment for the other ancestry components, whereas the San and Bantu components remained significant. The positive association between San ancestry and TB susceptibility was highly significant, whereas association with Bantu ancestry was relatively weak (p-value=1.06×10^{-11} and p-value=3.00×10^{-2} respectively) [100]. We were therefore interested in finding regions of the genome with excess San or Bantu ancestry in cases, but not in controls. From our simulations and analysis of the distribution of local ancestry deviations in our study group, it is evident that excess Bantu ancestry may be enriched with miss-called ancestry. As this is not the case for excess San ancestry, or excess San or Bantu ancestry (i.e. lack of non-African ancestry), we used two joint models to test for differences in ancestry between cases and controls per ancestry segment. One model tested for difference in San ancestry, and the other tested for differences in African (San or Bantu) ancestry. Only segments with ancestry two standard deviations above the genome-wide mean in cases were included in the models (110 San segments and 238 African segments), and age, gender and genome-wide ancestry were adjusted for, yielding p-values of 7.19×10^{-5} (San) and $< 2.00 \times 10^{-16}$ (Bantu). Regions that differ significantly between cases and controls are summarized in tables 7.3 and 7.4. The regions are comprised of a number of contiguous ancestry segments, and individual p-values for these sub-regions are summarized in supplementary tables 7.7.1 and 7.7.2, ranging between 0.0465 and 0.0005. Some of the segments in the African analysis also have large differences in San ancestry (e.g. some of the chromosome 10 and 17 segments have 7% or more San ancestry in TB cases compared to controls, which is comparable to the differences detected in the San ancestry model). These segments were however not detected in the San analysis as the mean proportion of San ancestry in TB cases fell just short of the mean plus two standard deviation cut-off for inclusion in the San ancestry model (mean + 2SD = 0.2820). On chromosome 17, 10 segments with the smallest p-values gave estimated TB case-control odds ratios of having African ancestry versus any other ancestry ranging from 1.77 (95% CI:1.27-2.46) to 1.61 (95% CI: 1.16-2.24). (The odds ratio for each segment is the estimated odds of having TB versus not having TB for an African segment, compared to the odds for a non-African segment.)

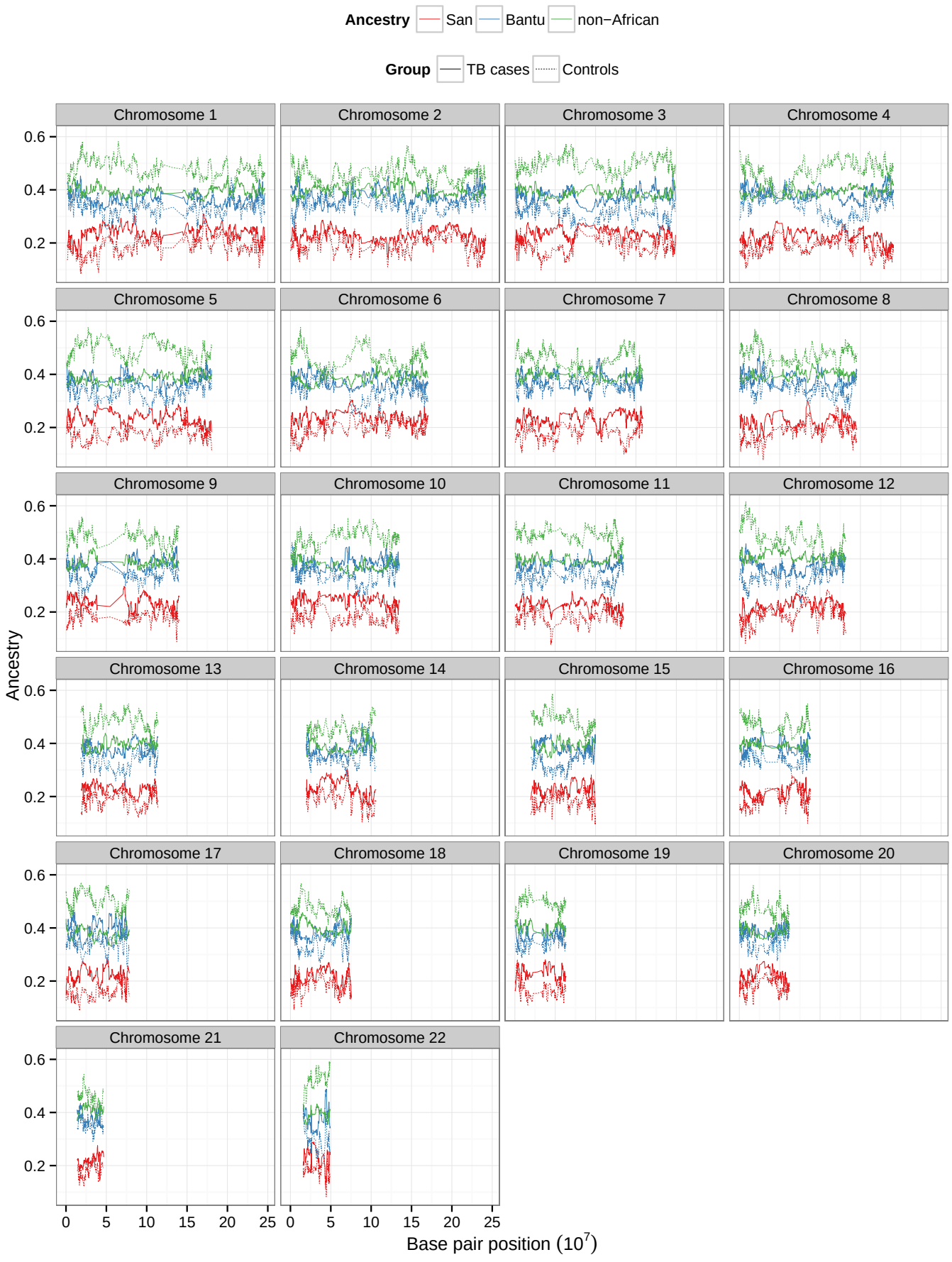


Figure 7.1: Mean local ancestry across the genome. The mean local ancestry estimates of TB cases and controls are shown per genomic position, for each of the source ancestries. Each panel represents a separate chromosome.

Table 7.3: Regions of the genome with excess San ancestry in TB cases, but not in controls. This table summarizes regions of the genome with excess San ancestry, found in TB cases but not in controls, after adjusting for age, gender and genome-wide San ancestry. Ancestry segments that are associated with increased San ancestry in cases compared to controls were identified and contiguous segments were merged. P-values for each of the individual ancestry segments are available in supplementary table 1. The mean RFMix genome-wide San ancestry estimates are 0.2304 and 0.1847 for cases and controls respectively, and the standard deviation of San local ancestry deviations is 0.0258 and 0.0321 in cases and controls respectively. Only regions of 500 000 base pairs or longer are shown (two short regions on chromosome 5 were excluded).

Region	Begin-end SNP	Length (Nr SNPs)	Mean San ancestry TB Cases	Control	Genes
1p31	rs12144711-rs7554551	671230 (123)	0.2902	0.1615	GADD45A, GNG12, DIRAS32
9q21	rs2309428-rs1847503	2080640 (323)	0.2909	0.1609	FAM189A, APBA1, PTAR1, C9orf135, MAMDC2, SMC5, KLF9
22q12	rs16986925-rs6006426	1290997 (152)	0.2850	0.1745	C22orf31, KREMEN1, EWSR1, RHBDD3, EMID1, AP1B1, RASL10A, GAS2L1, NEFH, RFPL1, NF2, NIPSNAP1, THOC5, UQCR10, CABP7, ZMAT5, ASCC2, MTMR3, HORMAD2, LIF, OSM

Table 7.4: Regions of the genome with excess African ancestry in TB cases, but not in controls. This table summarizes regions of the genome with excess African ancestry (San or Bantu), found in TB cases but not in controls, after adjusting for age, gender and genome-wide African ancestry. Ancestry segments that are associated with increased African ancestry in cases compared to controls were identified and contiguous segments were merged. P-values for each of the individual ancestry segments are available in supplementary table 7.7.2. The mean RFMix genome-wide African ancestry estimates are 0.6096 and 0.5238 for cases and controls respectively, and the standard deviation of local ancestry deviations is 0.0187 and 0.0336 in cases and controls respectively. Only regions of 500 000 base pairs or longer and that contain protein coding genes are shown (one short region on chromosome 5, one short region on chromosome 6, and four short regions on chromosome 10 were excluded).

Region	Begin-end SNP	Length (Nr SNPs)	Mean African ancestry TB Cases	Control	Genes
5q11	rs26090-rs1382907	739064 (70)	0.6480	0.4615	ISL1
10q22	rs827299-rs7083934	6243529 (693)	0.6607	0.5030	UNC5B, SLC29A3, CDH23, C10orf105, PSAP, B7-H5, CHST3, SPOCK2, ASCC1, DDIT4, DNAJB12, MICU1, MCU, OIT3, PLA2G12B, P4HA1, NUDT13, FAM149B1, DNAJC9, MRPS16, TTC18, ANXA7, PPP3CB, MSS51, MYOZ1, AGAP5, SYNPO2L, CAMK2G, NDST2, SEC24C, ZSWIM8, FUT11, CHCHD1, PLAU, C10orf55, VCL, AP3M1, ADK, KAT6B, DUPD1, SAMD8, DUSP13, VDAC2, COMTD1, ZNF503
15q15	rs1712435-rs16966424	2669916 (182)	0.6511	0.4963	PLA2G4D, VPS39, GANC, TMEM87A, CAPN3, SNAP23, ZNF106, HAUS2, LRRC57, TTBK2, CDAN1, UBR1, EPB42, TMEM62, TGM5, TGM7, TP53BP1, LCMT2, ZSCAN29, TUBGCP4, ADAL, CKMT1B, MAP1A, PPIP5K1, STRC, CKMT1A, CATSPER2, PDIA3, MFAP1, SERF2, HYPK, ELL3, SERINC4, WDR76, FRMD5, CASC4, CTDSPL2, EIF3J, SPG11, PATL2
17q22	rs7210845-rs9908090	5200677 (479)	0.6579	0.4698	ANKFN1, NOG, C17orf67, TRIM25, DGKE, COIL, SCPEP1, AKAP1, MSI2, CCDC182, MRPS23, CUEDC1, SRSF1, VEZF1, DYNLL2, EPX, OR4D1, MKS1, OR4D2, LPO, MPO, BZRAP1, SUPT4H1, RNF43, HSF5, SEPT4, MTMR4, C17orf47, TEX14, RAD51C, PPM1E, TRIM37, SKA2, GDPD1, SMG8, PRR11, CLTC, DHX40, PTRH2, VMP1, RPS6KB1, TUBD1, RNFT1, HEATR6, CA4, USP32, SCARNA20, RPL32P32, C17orf64, APPBP2, PPM1D

7.4 Discussion

Previously it was thought that admixture mapping in the SAC was not feasible using available LAI methods, due to the complex five-way admixture that occurred in this population [98]. Using the novel LAI algorithm implemented in RFMix, and by focusing on finding regions with excess San or excess San or Bantu ancestry in TB cases relative to controls, we have demonstrated that this technique can be applied to our SAC TB susceptibility study group. By limiting our tests for associations to only these directional ancestry components, and by using only two models to test for excess ancestry in TB cases relative to controls, we also reduced our study's vulnerability to false positive association signals that can arise as a result of multiple testing.

Based on their putative role in TB pathogenesis, we identified a number of genes in regions of excess ancestry that are convincing candidates for future studies. *GADD45A* (growth arrest and DNA-damage-inducible, alpha) is located on chromosome 1p31, a region with excess San ancestry in TB cases, and encodes a stress sensor protein that is involved in the regulation of myeloid cell innate immune function. Salerno et al. demonstrated that mice lacking Gadd45a are not able to recruit granulocytes and macrophages to the intraperitoneal cavity after administration of lipopolysaccharide (LPS) [256]. The recruitment of immune cells to the site of mycobacterial infection is crucial during TB. The oncostatin M gene (*OSM*) is also located in a region with excess San ancestry in TB cases (chromosome 22q12). Friendland et al. found that *M. tuberculosis* infection of monocytes resulted in prostaglandin-dependent OSM secretion, which synergized with tumour necrosis factor-α to drive fibroblast matrix metalloproteinases-1/-3 secretion [257].

Our analysis also found that chromosome 15q15 may contain an African TB susceptibility locus. Chromosome 15q was previously identified as containing TB susceptibility genes in a linkage study using African families [258] and fine-mapping localised the region to 15q11 [259]. Chromosome 17q22 also showed evidence of having excess African ancestry in TB cases. It is known that chromosome 17q11–q21 may contain a cluster of susceptibility genes for diseases caused by intramacrophage pathogens, such as *M. tuberculosis* [136]. Another gene with excess African TB ancestry was *B7-H5* (B7 homolog 5), located on chromosome 10q22. Its protein product plays a key regulatory role in T cell growth and cytokine production [260] and is expressed in macrophages.

Two of the 36 SNPs that were found to be moderately associated with having TB in a previous genome-wide association study (GWAS) are located in or proximal to regions identified in this study [98] (rs6694316 on 1p32 which is 16 cM way from the first SNP in the identified 1p31 region, that harbours the *GADD45A* gene, and rs4745272 on 9q21). Considering that the GWAS used the same SAC data set, we note that a number of new putative susceptibility regions were identified by our admixture mapping study, which would not have been evident based solely on single SNP association statistics.

7.4.1 Study limitations

Due to the correlation between segments of ancestry, especially contiguous segments of ancestry, p-values that accurately quantify the statistical significance of our results could not be estimated in this study. Regions of excess ancestry in TB cases, having ancestry more than 2 standard deviations from the mean, were first identified. P-values were not estimated in this step due to the correlation between segments of ancestry. The identified regions would be the most probable regions to harbour ancestry-related TB susceptibility variants, and a statistical model was then used to determine which of the regions had excess ancestry in TB cases relative to

controls. The estimated p-values from this model do not account for the fact that the variables were selected by first performing a case-only analysis. The statistical model also contained correlated variables, due to contiguous segments with similar ancestry. Correlation between a model's independent variables may result in inflated variance, and estimated p-values may therefore be inflated (biased towards type II errors). We therefore note that the reported p-values should not be interpreted as a strict quantification of the statistical significance of the results, but that the p-values rather indicate which regions are most likely to harbour ancestry related TB susceptibility variants.

RFMix estimates of San ancestry were much lower than ADMIXTURE estimates. This may be explained by the small group size of our San source population as well as short tracts of San ancestry inherited via Southern African Bantu populations, and results in bias towards local ancestry deviations with lack of San ancestry. We speculate that roughly 50 San individuals would be required to alleviate the small group size problem (the RFMix authors used 30 reference individuals each from Native American, European and African populations in their simulations, but as the genetic distance between the San and Bantu populations is smaller than the genetic distances between the populations used by the RFMix authors, a larger number of reference individuals would be required to distinguish San and Bantu ancestry). We have worked around these limitations by focusing our attention on finding regions of excess San or excess San or Bantu ancestry, which are less likely to be affected by spurious deviation in local ancestry, and this approach also supports our hypothesis regarding the directionality of these ancestries and TB susceptibility. The possibility of short tracts of San ancestry harbouring TB susceptibility genes are also less likely compared to longer tracts, as the former are less likely to overlap in a group of individuals. Miss-identification of these short tracts as Bantu ancestry has therefore probably not resulted in loss of information regarding TB susceptibility.

The small size of the San source population also limited the number of ancestral chromosomes that were used in the generation of our simulated data set. As a result, the bias seen in the simulated local ancestry deviation distributions may have been exacerbated. Due to lack of accurate historical records, our simulation also did not take into account such complexities as the timing of admixture events, and potential inaccuracies of the source (reference) populations. Pasanuic et al. [261] recently evaluated the accuracy of multi-way LAI in a group of nuclear Latino families by determining whether the local ancestry of offspring is congruous with Mendelian inheritance (it is for example implausible that a child has European ancestry at a locus if neither parent has European ancestry at that locus). Multi-way LAI accuracy of several algorithms was shown to be much lower compared to reported accuracies calculated from simplified simulation data sets. Despite these issues, we have been able to use our simulated data set to demonstrate the relatively superior LAI accuracy of RFMix, and to explore the direction of potential local ancestry deviation bias.

Our SAC study group samples were genotyped with the aim of performing a case-only admixture mapping study, and as a result only a small number of controls were genotyped. As case-only admixture mapping has subsequently (and correctly) been described as inappropriate for multi-way admixed populations due to artefactual ancestry deviations arising from inaccuracies in LAI[253], we first identified regions of excess ancestry found in cases only, and then validated these findings by testing for excess ancestry in cases relative to controls. Despite the small number of controls that were available, bias in local ancestry inferences was still controlled, and a number of novel regions that contain highly plausible candidate TB susceptibility genes were uncovered by our study.

7.5 Conclusion

The genetics of the South African Coloured population is arguably one of the most challenging and interesting examples found in present day multi-way admixed human populations. This is the first study to apply genome-wide admixture mapping to this highly complex group. We have demonstrated that admixture mapping is feasible in the South African Coloured population, a result which may be useful for other researchers that either study this population, or other populations with complex admixture. We have identified a number of novel candidate TB susceptibility genomic regions, as well as providing evidence to validate genetic loci previously implicated.

7.6 Materials and methods

7.6.1 Sample collection and ethics approval

Individuals residing in the Cape Town suburbs of Ravensmead and Uitsig, and who self-identified as South African Coloured, were recruited to participate in this study. These suburbs have a low prevalence of HIV but a high incidence of TB, as well as a relatively homogenous socio-economic environment [148]. Bacterial confirmation (smear positive/culture positive) was used to identify TB patients. Healthy individuals with no prior history of TB were selected as controls. All the participants in this study were HIV negative. Our previous study of healthy children and young adults from the control community found that 80% of children older than 15 years had positive tuberculin skin tests (TST), an indication of latent infection with *M. tuberculosis* [219]. The majority of the control population is therefore TST positive, and with the average age of the controls in this study being 31 years, we estimate a TST positivity of 80% or above. These healthy individuals had no previous history of TB disease or treatment and were unrelated to all others included in the study.

This study was approved by the Ethics Committee of the Faculty of Health Sciences, Stellenbosch University (project registration numbers 95/072, NO6/07/132 and N11/07/210). Blood samples for DNA were collected with written informed consent. The research was conducted according to the principles expressed in the Declaration of Helsinki.

7.6.2 Software

Web URLs, version information and parameter settings of the programs used in this study are summarized in supplementary table 3.

SHAPEIT [262] was used for phasing the SAC data set as well as the San source population data. A python script developed by J. Morrison was used to produce the ancestry break points of the simulated data set and to assign ancestry to segments along the chromosome. LAMP-LD [263] and RFMix [264] were used to infer local ancestry of the simulated data, whereas RFMix was used for inference of local ancestry of the SAC study group. ADMIXTURE [129] was used to estimate genome-wide ancestry of the SAC study group. PLINK [105] was used for merging admixed and source population data sets, in order to create input files required by ADMIXTURE. PLINK was also used to identify related individuals and filter SNPs according to quality control criteria. Prior to ADMIXTURE estimation and identification of related individuals, SNPs that were in LD were identified using PLINK and discarded, leaving 87 648 SNPs (see supplementary table 3 for PLINK parameter settings). Biofilter was used to identify genes that fall within specific regions of the genome. The R programming environment [106] was used for statistical analysis and the *ggplot2* [229] and *hexbin* [265] R packages was used to

Table 7.5: Source population data. Data sets used to represent the source populations of the South African Coloured population. The sample size reflects the group size after relative pairs have been removed.

Population	Group	Description	Source	Platform	Size
San		Ju|'hoansi San from North Namibia	Private	Affymetrix 6.0	21
Bantu	YRI	Yoruba in Ibadan, Nigeria	HapMap3	Release 2	112
non-African	CEU	Utah residents with Northern and Western European ancestry, USA	HapMap3	Release 2	112
	GIH	Gujarati Indians from Houston, Texas, USA	HapMap3	Release 2	88
	JPT+CHB	Japanese in Tokyo, Japan and Han Chinese in Beijing, China	HapMap3	Release 2	170

create the figures. The R *hierfstat* package was used to estimate F_{ST} (fixation index) between the source populations of the SAC, using the same data set that was created for ADMIXTURE estimation [266].

7.6.3 Genotyping and quality control

969 individuals from the SAC study group were genotyped on the Affymetrix GeneChip Human Mapping 500K Array Set. After SNP calling [26], quality control and the removal of related individuals [98], the data set comprised 381 530 autosomal SNPs of 642 TB cases and 91 controls.

Source (a.k.a. reference) populations used to infer the ancestry of SAC individuals are summarized in table 7.5 and the genetic distances (F_{ST}) between these source populations are summarized in supplementary table 7.7.4 . Populations used to represent the San were obtained from a private data access committee (contact corresponding author). The data set represents the same group analyzed by Schlebusch et al. [126], but was genotyped on the Affymetrix genotyping platform instead of the OmniExpress platform, which overlaps better with SNPs in the SAC study group data set. Phased HapMap3 Release 2 data was used to represent the remaining source populations. Pairwise IBS clustering was used to identify related individuals and only unrelated individuals (coefficient of relatedness < 0.5) were retained in the data sets. The San data set was filtered to remove SNPs that were not in Hardy-Weinberg equilibrium (p-value threshold of 0.0001) or had a large proportion of missing data (at least a 75% call rate).

7.6.4 Phasing

As local ancestry inference requires phased input data, the SAC and San data sets were phased using a Markov model to estimate haplotypes, implemented in the SHAPEIT software. The genetic map used by the HapMap project to phase the HapMap data sets was utilized for phasing the SAC and San data sets (NCBI build 36 release 22, obtained from ftp://ftp.hapmap.org/hapmap/recombination/2008-03_rel22_B36/rates/). The genotypes of 733 SAC individuals were phased using 381 530 autosomal SNPs. The San data set of 21 individuals were phased using 866 382 autosomal SNPs.

7.6.5 Combining data sets

After quality control and phasing of the SAC and San source population data sets, the data sets were reduced to contain only those SNPs present in all the data sets (328 866 autosomal SNPs). Where required, strand and reference alleles of the SAC and San data sets were flipped to match the HapMap data sets. The centimorgan (cM) genomic positions used by PCAdmix and RFMix to determine ancestry windows were calculated using the NCBI build 36 release 22 genomic map (obtained from ftp://ftp.hapmap.org/hapmap/recombination/2008-03_rel22_B36/rates/). The base pair positions of SNPs were obtained from the HapMap data, and in the case where an exact base pair position match was not found in the genomic map, the base pair position of the SNP was converted to cM by using a weighted average of the cM positions of the two SNPs closest to it.

7.6.6 Simulation

Using an approach similar to that of Pasanuic et al. [261] and Price et al. [267], the ancestry breakpoints of 1500 admixed chromosomes (chromosome 1) were first generated. Recombination positions and thus breaks in ancestry on each chromosome were generated using a random walk from base pair position zero to the end of the chromosome, with ancestry crossovers occurring as a Poisson process. The rate of the Poission process was set to 10 (the assumed number of generations since admixture) times a recombination rate of 10^{-8}. The average number of breakpoints per chromosome was 35. After determining the breakpoints, segments of ancestry on each chromosome were assigned as San, YRI, CEU, GIH and JPT+CHB using proportions 0.33, 0.28, 0.19, 0.13 and 0.07 respectively, corresponding to the genome-wide ancestry estimates of Chimusa et al.[98]. The ancestry assignments were based on draws from a uniform(0,1) distribution, and determining whether a draw falls between the interval [0.00, 0.33) (San), [0.33, 0.61) (YRI), [0.61, 0.80) (CEU), [0.80, 0.93) and [0.93, 1] (JPT+CHB).

Segments of ancestry assigned in this manner were then used to construct 1500 admixed chromosomes (chromosome 1). 10 source population chromosomes were selected randomly from each of the source data sets. The remaining chromosomes were set aside to use as input source populations for LAI. Admixed chromosomes were constructed by randomly copying segments of ancestry from the selected source population chromosomes, corresponding to the ancestry assigned to the segment. As a simple example, consider an admixed chromosome with San ancestry for SNPs at positions 1 to 100 on the chromosome, and CEU ancestry for SNPs at positions 101 to 200. The admixed chromosome would be constructed by randomly selecting a San chromosome and copying the SNPs at positions 1 to 100, followed by randomly selecting a CEU chromosome and copying the SNPs at positions 101 to 200.

Note that a limited number of source population chromosomes were used to simulate the admixed chromosomes. As a result, the data set does not contain independent observations. Unlike most LAI algorithms, the global ancestry proportion estimation algorithm implemented in the software program ADMIXTURE assumes independent observations. Taken together with the limited number of SNPs that are available compared to genome-wide data (only chromosome 1 SNPs are available), the simulated chromosome 1 data set is not suited to the estimation of chromosome-wide ancestry using ADMIXTURE, and RFMix chromosome-wide estimates were used instead.

7.6.7 Labelling ancestry

LAMP-LD, RFMix and ADMIXTURE were run using five source populations (table 7.5), but since only San and Bantu genome-wide ancestry is independently associated with TB

susceptibility [100; 98], European, South Asian and East Asian ancestries were merged and labelled as non-African after inference was performed.

7.6.8 Delineating called ancestry segments

RFMix labels the called ancestry of each SNP along a chromosome and does not identify windows of ancestry across a chromosome in its output. Segments of ancestry were therefore identified by determining the SNP positions where a switch in ancestry occurs for at least one chromosome. Each ancestry segment starts at such a position, and ends one SNP before the next ancestry switch position.

7.6.9 Calculating genome-wide ancestry using local ancestry

The genome-wide ancestry of each of the 733 SAC individuals was calculated using local ancestry called by RFMix, as follows: The number of SNPs labeled as a particular ancestry was counted per individual for each of the individual's 22×2 chromosomes. This total count was divided by the total number of SNPs across the genome, yielding an estimate for the individual's genome-wide ancestry.

7.6.10 Calculating local ancestry deviation

After delineating the called ancestry of the 1500 simulated chromosomes into 1077 segments, the mean San, Bantu and non-African ancestry of each segment was calculated. The local ancestry deviation of each segment was then calculated, seperately for each of the three ancestries, by subtracting the overall RFMix mean ancestry from the mean ancestry of the segment.

In the same way, after delineating the called ancestry of $733 \times 2 \times 22$ chromosomes in the SAC study group into 13 860 segments, the mean ancestry of each segment was calculated for each of the three source ancestries. The local ancestry deviation of each segment was calculated by subtracting the overall RFMix mean genome-wide ancestry from the mean ancestry of the segment, for each of the ancestries.

7.6.11 Correlation between miss-called ancestry and deviation in ancestry in simulated data

After local ancestry deviations were calculated for each of the segments identified in the simulated data set, and segments with miss-called ancestry were identified by comparing ancestry called by RFMix to the known ancestry of each segment, the number of segments with miss-called ancestries were calculated per identified segment. This was done for each of the six possible pairs of miss-called ancestry (San miss-called as Bantu, San miss-called as non-African, Bantu miss-called as San, Bantu miss-called as non-African, non-African miss-called as San and non-African miss-called as Bantu). A large number of miss-calls occurred in the segment of ancestry that spans the centromere, with a standardized local ancestry deviation Z-value of -0.7109, 0.7052, and 0.1518 for San, Bantu and non-African ancestry respectively. After discarding this outlying segment, Pearson's correlation coefficient was calculated between the number of miss-called segments and local ancestry deviation, for each of the six possible pairs of miss-called ancestry and San, Bantu and non-African local ancestry deviation combinations.

7.6.12 Calculating ancestry inference accuracy using simulated data

Accuracy of LAI using the simulated data set was calculated per chromosome as follows. The ancestry assigned to each SNP in the data set by the simulation process was compared to the ancestry assigned to the SNP by LAI. The proportion of SNPs that the LAI correctly assigned to each chromosome was then calculated. For each of the three considered ancestries (San, Bantu and non-African), the proportion of SNPs that were miss-called for each of the other two ancestries was also calculated. The miss-called proportions were calculated across all SNPs on the chromosome, as well as per number of SNPs of that particular ancestry.

7.6.13 Relationships between miss-called ancestry, tract length and degree of admixture

The lengths of tracts of ancestry in the simulated data set were calculated in terms of the number of SNPs that constitute a track. Each track's corresponding proportion of miss-called SNPs were calculated by comparing the ancestry assigned by RFMix to each SNP with the known ancestry of the SNP, and then dividing the number of miss-called SNPs by the length of the tract. Pearson's correlation coefficient was calculated to quantify the relationship between the length of a track and the proportion of errors on a track.

The number of tracks of ancestry present in each simulated chromosome was counted, and was used to represent a chromosome's degree of admixture. The number of miss-called SNPs was counted per chromosome, by comparing the ancestry assigned by RFMix to each SNP with the known ancestry of the SNP. Pearson's correlation coefficient was calculated to quantify the relationship between the number of tracks of ancestry of a chromosome and the number of miss-called SNPs on a chromosome.

7.6.14 Statistical analyses

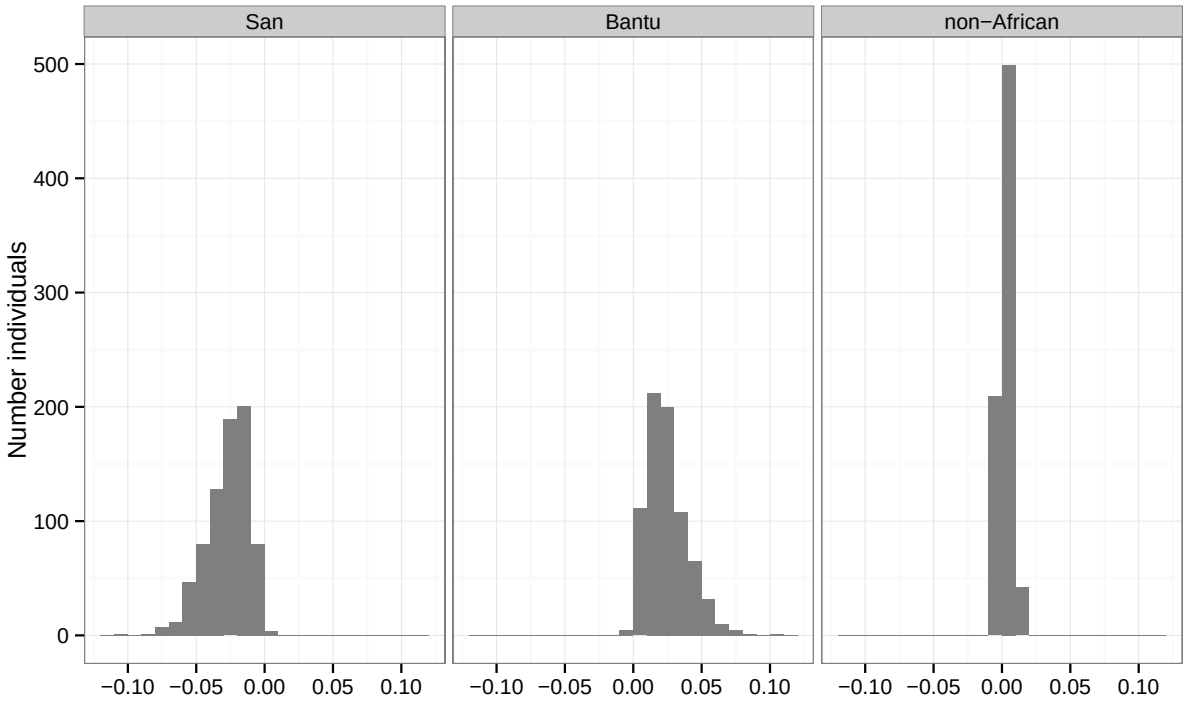
A single linear mixed-effects model was used to compare the number of SNPs that were miss-called by LAMP-LD vs. RFMix in the simulated data set. The proportion of SNPs that were miss-called for a particular pair of ancestries (the proportion of San SNPs miss-called as Bantu, the proportion of San SNPs miss-called as non-African, etc.) was log transformed for analysis, since the proportions were positively skewed. The interaction between software program (LAMP-LD or RFMix) and pair of ancestries was tested, while chromosome identifiers were specified as random effect, in order to adjust for the correlation between between different miss-call proportions on the same chromosome.

Two joint logistic regression models were used to test whether San and African ancestry differs between TB cases and controls (the statistical modelling term "joint" means that all the effects were estimated jointly in a single model, instead of doing an individual test for each combination of ancestry and segment). Joint modelling has the advantage of providing estimates that are adjusted for all other predictors in the same model. Case-control status was specified as outcome variable and the interaction of segment identifier and presence or absence of the particular ancestry (San/not-San or African/not-African) was specified as predictor. This statistical interaction between a specific ancestry and a segment identifier can be described as a predictor combining an ancestry indicator with the segment identifier. That means that the odds ratio for each segment is the estimated odds of having TB versus not having TB for a San segment, compared to the odds for a non-San segment, and the interpretation is analogous for African ancestry. Since age, gender and genome-wide ancestry distributions differ between TB cases and controls, the models were also adjusted for these variables.

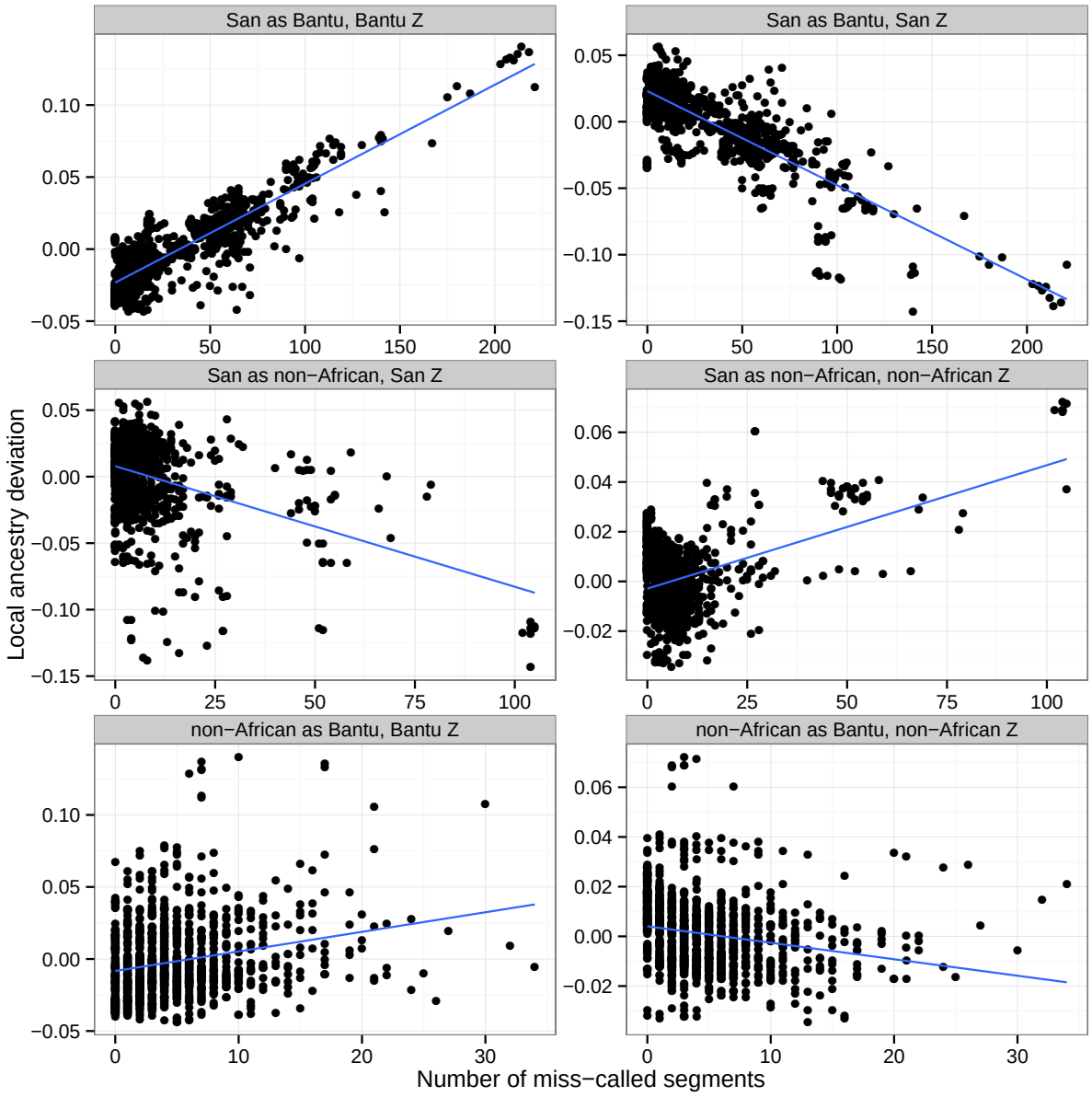
Acknowledgements

We thank all participants and field workers in this study. We also thank the developers of the open source software we used in our analyses.

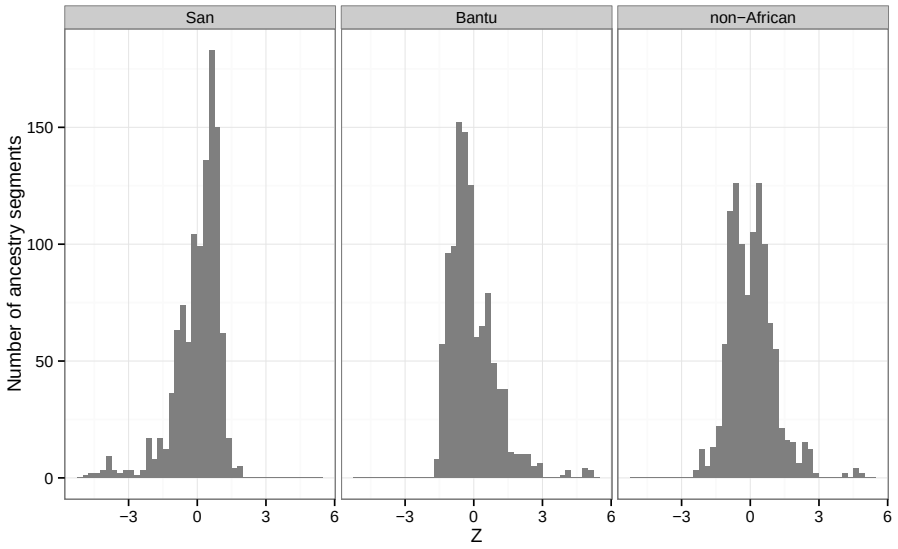
7.7 Supplementary figures and tables



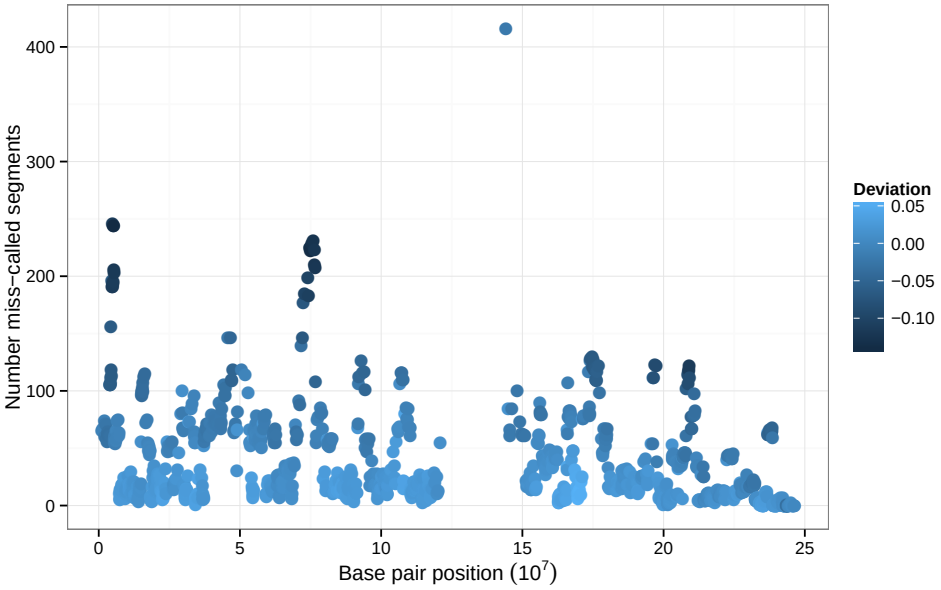
Supplementary Figure 7.7.1: Difference between ancestry called by RFMix and known ancestry per individual. The known ancestry of a simulated data set of 750 SAC individuals is compared to the ancestry calledy by RFMix per individual (chromosome 1). Histograms of the difference between the called mean ancestry and known mean ancestry of each individual are shown, per each of the three source ancestries.



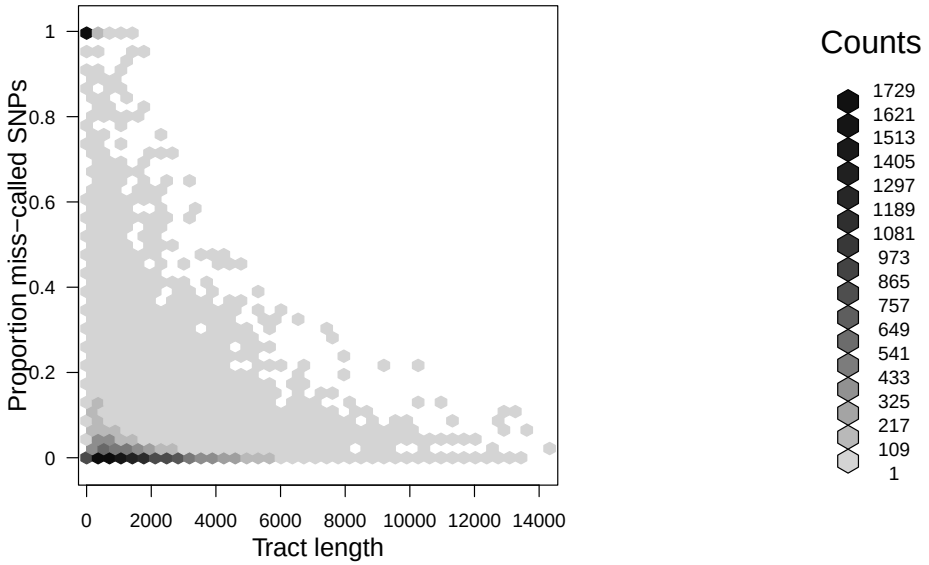
Supplementary Figure 7.7.2: Scatterplots of the number of miss-called ancestry segments against deviation in ancestry in simulated data. Miss-called ancestry was identified by comparing the known ancestry of a simulated data set of 1500 SAC chromosomes to the ancestry called by RFMix (chromosome 1). Deviations in ancestry were calculated by subtracting the overall RFMix mean ancestry from the local mean ancestry of each segment.



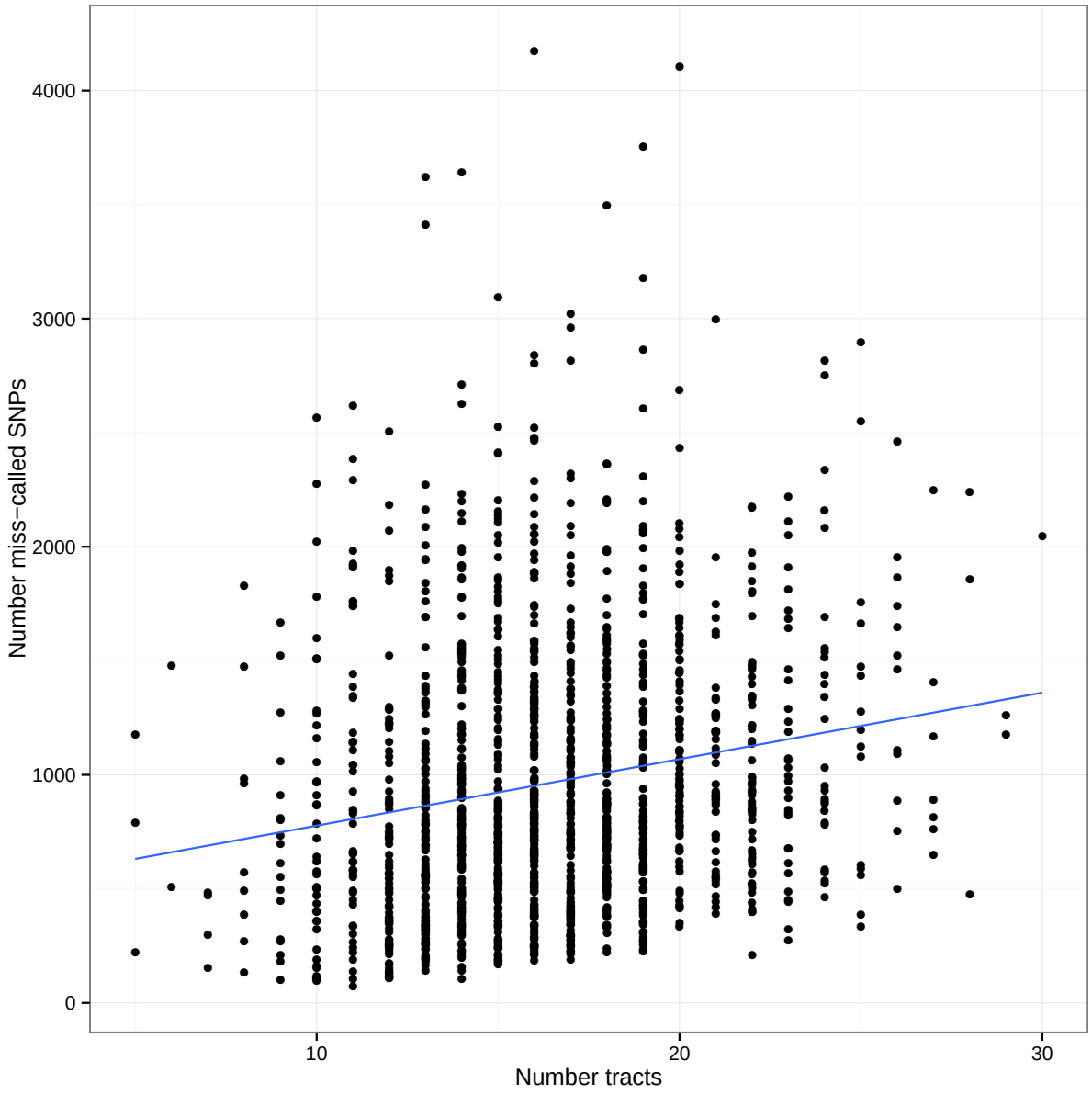
Supplementary Figure 7.7.3: Local ancestry deviations in simulated data. Histograms of local ancestry deviations in the simulated data set are shown in this figure, for each of the source ancestries. The deviation of each segment was calculated by subtracting the overall RFMix mean ancestry from the local mean ancestry of the segment (chromosome 1). Standardized deviation scores are shown at the bottom of the horizontal axis.



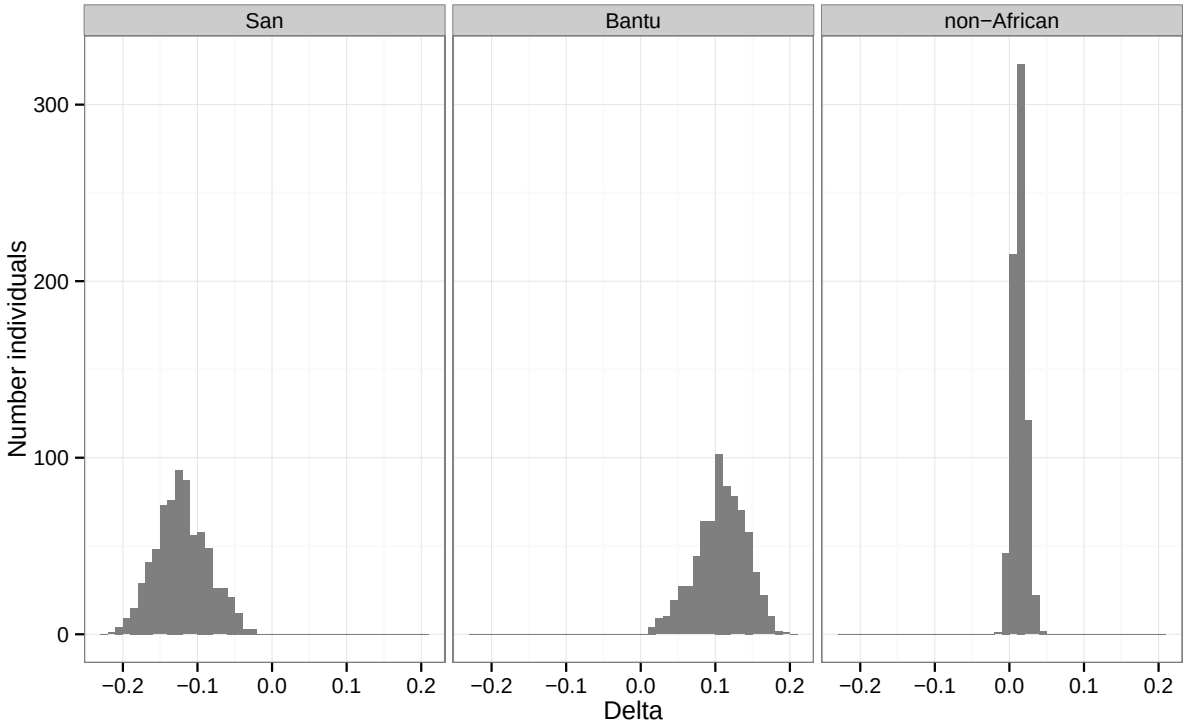
Supplementary Figure 7.7.4: Distribution of miss-called San ancestry segments in simulated data. The figure shows the base pair positions of San ancestry segments that were miss-called by RFmix to have Bantu or non-African ancestry, and the number of segments that were miss-called at a position, in a simulated data set of 1500 SAC haplotypes (chromosome 1). Data points are shaded according to deviation from the RFMix overall mean San ancestry, where darker shades indicate lower San ancestry compared to the mean.



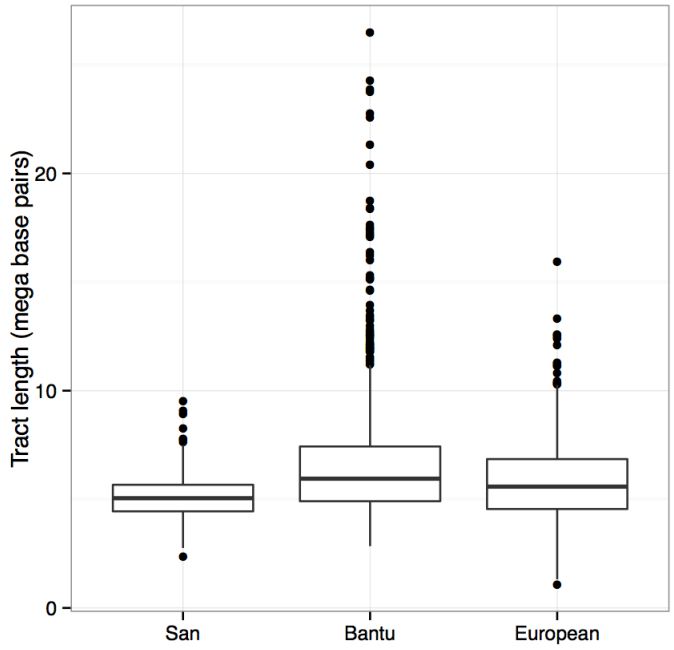
Supplementary Figure 7.7.5: Distribution of the length of tracts of ancestry and the proportion of SNPs with miss-called ancestry per tract in the simulated data. The lengths of tracts of ancestry in a simulated data set of 1500 SAC chromosomes (chromosome 1) were calculated in terms of the number of SNPs that constitute a track, and are shown on the x-axis. The proportion of SNPs that were miss-called were calculated per track by comparing the ancestry assigned by RFMix to each SNP with the known ancestry of the SNP, and is shown on the y-axis (number miss-called SNPs divided by the length of the tract). Hexagons denote one or more observations; the darker the shading, the more observations are represented.



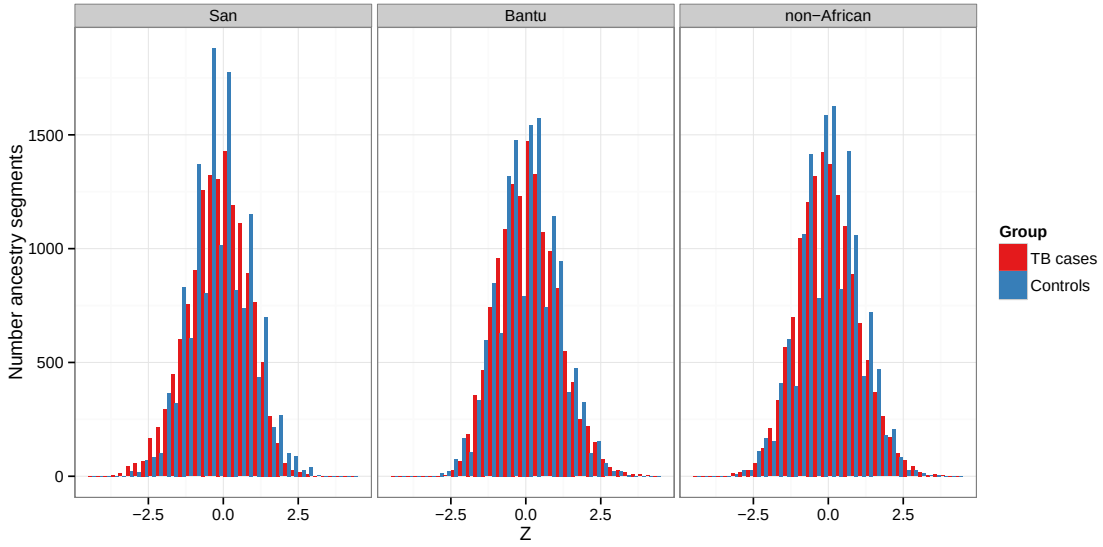
Supplementary Figure 7.7.6: Scatterplot of the number of tracts of ancestry on a chromosome and the number of miss-called SNPs for that chromosome. The number of tracts of ancestry in a simulated data set of 1500 SAC chromosomes (chromosome 1) were counted per chromosome and is shown on the x-axis. The corresponding number of miss-called SNPs for each simulated chromosome was determined by comparing the ancestry assigned by RFMix to each SNP with the known ancestry of the SNP, and is shown on the y-axis. Each data point therefore represents a single simulated chromosome, with its number of ancestry tracts read from the x-axis, and its number of miss-called SNPs read from the y-axis.



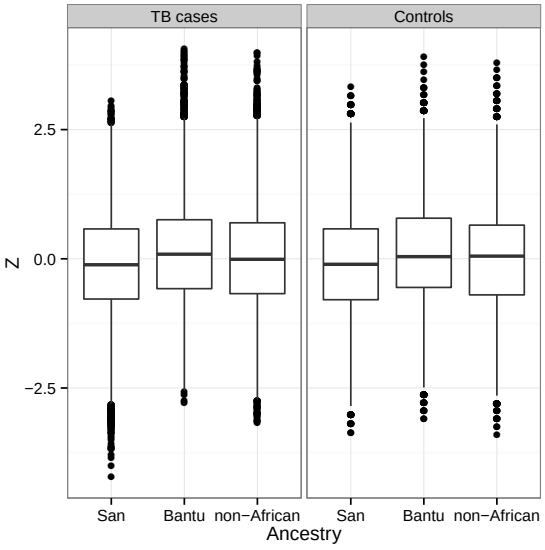
Supplementary Figure 7.7.7: Difference between RFMix and ADMIXTURE estimates of genome-wide ancestry in the SAC study group. The difference between the genome-wide ancestries estimated by RFMix and ADMIXTURE in a study group of 733 SAC individuals are shown in this figure. Histograms of the difference between each individual's RFMix and ADMIXTURE ancestry estimate are shown, per each of the three source ancestries.



Supplementary Figure 7.7.8: Boxplots of ancestry tract lengths in the SAC study group. The distribution of the mean San, Bantu and European tract lengths of each of the 733 individuals in the SAC study group are depicted in this figure.



Supplementary Figure 7.7.9: Histograms of local ancestry deviations in the SAC study group. Histograms of the deviations of local ancestry in the SAC study group (642 TB cases and 91 controls) are shown in this figure, for each of the source ancestries. The deviation of each segment was calculated by subtracting the mean RFMix genome-wide ancestry from the mean local ancestry of the segment, seperately for cases and controls. Standardized deviation scores are shown at the bottom of the horizontal axis.



Supplementary Figure 7.7.10: Boxplots of local ancestry deviations in the SAC study group. Boxplots of the standardized deviations of local ancestry in the SAC study group (642 TB cases and 91 controls) are shown in this figure, for each of the source ancestries. The deviation of each segment was calculated by subtracting the mean RFMix genome-wide ancestry from the mean local ancestry of the segment, separately for cases and controls. The local ancestry deviations were then standardized by dividing by the standard deviation of the local ancestry deviations.

Supplementary Table 7.7.1: Statistical significance of regions of the genome with excess San ancestry in TB cases relative to controls. This table summarizes regions of the genome with excess San ancestry, found in TB cases but not in controls, after adjusting for age, gender and genome-wide San ancestry. Segments were labeled according to their position on the chromosome; contiguous segments of ancestry therefore have contiguous segment identifiers.

Region	Segment ID	Begin-end SNP	Length (Nr SNPs)	Mean San ancestry TB Cases	Controls	P-value
1p31	375	rs12144711-rs10789239	107674 (24)	0.2897	0.1995	0.0135
1p31	376	rs4655567-rs4548410	254010 (44)	0.2928	0.2160	0.0326
1p31	377	rs12025677-rs11209202	131840 (20)	0.2936	0.2160	0.0319
1p31	378	rs10889741-rs6691251	88184 (9)	0.2889	0.2105	0.0269
1p31	379	rs2566762-rs7554551	82567 (26)	0.2858	0.1940	0.0099
5p13	114	rs10513153-rs1445823	130346 (35)	0.2827	0.2160	0.0238
5p13	235	rs16904004-rs6870368	115695 (17)	0.2843	0.2160	0.0465
9q21	269	rs2309428-rs6559488	131678 (20)	0.2858	0.2050	0.0231
9q21	270	rs11138342-rs11139997	353460 (40)	0.2889	0.2105	0.0319
9q21	271	rs10511968-rs11140836	172263 (28)	0.2850	0.2050	0.0294
9q21	272	rs11140862-rs7875663	573992 (84)	0.2967	0.2050	0.0138
9q21	273	rs6560137-rs7350298	302822 (55)	0.2952	0.2050	0.0203
9q21	274	rs1028879-rs7041925	179239 (37)	0.2913	0.2105	0.0339
9q21	275	rs2909293-rs1847503	332682 (59)	0.2936	0.2050	0.0222
22q12	93	rs16986925-rs5762996	143883 (42)	0.2882	0.2215	0.0326
22q12	94	rs132275-rs2301290	135145 (10)	0.2874	0.2215	0.0358
22q12	96	rs2857641-rs6006426	612310 (65)	0.2827	0.2215	0.0355

Supplementary Table 7.7.2: Statistical significance of regions of the genome with excess African (San or Bantu) ancestry in TB cases relative to controls. This table summarizes regions of the genome with excess African ancestry, found in TB cases but not in controls, after adjusting for age, gender and genome-wide African ancestry. Segments were labeled according to their position on the chromosome; contiguous segments of ancestry therefore have contiguous segment identifiers.

Region	Segment ID	Begin-end SNP	Length (Nr SNPs)	Mean San ancestry		Mean Bantu ancestry		P-value
				TB Cases	Controls	TB Cases	Controls	
5q11	244	rs1450660-rs1822824	303696 (33)	0.2702	0.2105	0.3770	0.3423	0.0081
5q11	250	rs26090-rs1382907	739064 (70)	0.2726	0.1885	0.3754	0.3588	0.0049
6q15	402	rs11969733-rs285612	217975 (24)	0.2375	0.2050	0.4104	0.3478	0.0091
6q15	403	rs16882779-rs790604	24779 (5)	0.2375	0.2050	0.4112	0.3478	0.0087
10q22	368	rs827299-rs1338638	57072 (21)	0.2298	0.1995	0.4361	0.3972	0.0351
10q22	369	rs1338637-rs12264572	94894 (12)	0.2227	0.1995	0.4424	0.3917	0.0223
10q22	370	rs7076330-rs10999736	138544 (22)	0.2196	0.2050	0.4439	0.3863	0.0297
10q22	371	rs16928536-rs3740458	63196 (16)	0.2204	0.2050	0.4463	0.3753	0.0126
10q22	372	rs7075861-rs1417207	95418 (14)	0.2196	0.2215	0.4463	0.3643	0.0190
10q22	373	rs10999804-rs2394797	57794 (12)	0.2243	0.2325	0.4439	0.3588	0.0189
10q22	374	rs7088556-rs17634834	32303 (4)	0.2266	0.2380	0.4416	0.3533	0.0189
10q22	376	rs10509336-rs7094749	111700 (19)	0.2562	0.2215	0.4073	0.3698	0.0222
10q22	377	rs10999960-rs9415039	64778 (6)	0.2656	0.2325	0.4027	0.3698	0.0310
10q22	378	rs10762477-rs7090957	176032 (23)	0.2695	0.2325	0.3964	0.3698	0.0309
10q22	379	rs10509339-rs10509767	338118 (38)	0.2625	0.2380	0.3988	0.3643	0.0314
10q22	380	rs10762505-rs3740293	1359930 (71)	0.2648	0.2380	0.3972	0.3698	0.0346
10q22	381	rs1004059-rs11000831	327517 (17)	0.2625	0.2215	0.3964	0.3698	0.0125
10q22	382	rs10824049-rs10824259	1080040 (72)	0.2609	0.2270	0.4003	0.3643	0.0094
10q22	383	rs10762651-rs7088635	321064 (29)	0.2601	0.2160	0.4050	0.3753	0.0086
10q22	384	rs4612741-rs2133705	336875 (35)	0.2570	0.2160	0.4089	0.3753	0.0080
10q22	385	rs1124372-rs9415136	131915 (31)	0.2531	0.2105	0.4097	0.3753	0.0070
10q22	386	rs4746341-rs17445672	222280 (41)	0.2555	0.2050	0.3972	0.3753	0.0110
10q22	387	rs16932945-rs1992012	126121 (23)	0.2586	0.1995	0.3902	0.3643	0.0038
10q22	388	rs17376389-rs2637266	276265 (75)	0.2625	0.1940	0.3847	0.3753	0.0068
10q22	389	rs1907323-rs4980117	360363 (60)	0.2632	0.1940	0.3863	0.3698	0.0033
10q22	391	rs2395453-rs7083934	200022 (30)	0.2508	0.1940	0.3964	0.3863	0.0162
10q22	395	rs1877998-rs11815134	113578 (12)	0.2702	0.2380	0.3863	0.3643	0.0410
10q25	508	rs3014204-rs17115877	237854 (22)	0.2562	0.1940	0.3910	0.3533	0.0042
10q25	514	rs10884128-rs10509806	507163 (60)	0.2702	0.2160	0.3777	0.3478	0.0213
10q25	542	rs10506868-rs11196030	82164 (14)	0.2586	0.2050	0.3972	0.3808	0.0370
15q15	136	rs1712435-rs677845	110381 (21)	0.2305	0.2435	0.4229	0.3423	0.0289
15q15	137	rs588695-rs493177	1088795 (90)	0.2290	0.2270	0.4221	0.3533	0.0293
15q15	138	rs574065-rs16966424	1455279 (71)	0.2274	0.2270	0.4213	0.3533	0.0341
17q22	296	rs7210845-rs9894332	55957 (9)	0.2648	0.2105	0.3847	0.3313	0.0082
17q22	297	rs17759236-rs9891519	269818 (44)	0.2765	0.2160	0.3832	0.3368	0.0090
17q22	298	rs929585-rs7208587	160574 (28)	0.2741	0.2160	0.3832	0.3368	0.0101
17q22	299	rs17760268-rs4793823	128914 (29)	0.2757	0.2105	0.3793	0.3423	0.0118
17q22	300	rs10491158-rs8069500	74109 (18)	0.2819	0.2270	0.3793	0.3203	0.0038
17q22	301	rs3914804-rs17820808	23232 (7)	0.2827	0.2270	0.3793	0.3203	0.0034
17q22	302	rs4794665-rs2525997	120913 (16)	0.2780	0.2270	0.3855	0.3148	0.0015
17q22	303	rs205499-rs11079268	128522 (9)	0.2765	0.2160	0.3801	0.3203	0.0013
17q22	304	rs7214685-rs8071417	45763 (5)	0.2687	0.1940	0.3871	0.3313	0.0005
17q22	305	rs721427-rs9652852	116730 (21)	0.2586	0.1995	0.4003	0.3313	0.0009
17q22	306	rs16969033-rs4794718	83871 (15)	0.2547	0.1995	0.4003	0.3368	0.0022
17q22	307	rs8071867-rs12949540	148635 (33)	0.2562	0.1940	0.4019	0.3423	0.0015
17q22	308	rs12601123-rs4793550	51316 (11)	0.2516	0.1885	0.4003	0.3643	0.0094
17q22	309	rs2111016-rs1024819	26322 (5)	0.2523	0.1830	0.4019	0.3588	0.0038
17q22	311	rs3744089-rs2190759	92443 (14)	0.2375	0.1885	0.4143	0.3698	0.0240
17q22	312	rs4793565-rs203257	44871 (10)	0.2368	0.1830	0.4151	0.3698	0.0180
17q22	313	rs10515149-rs4793574	81338 (16)	0.2360	0.1775	0.4213	0.3917	0.0285
17q22	314	rs9894704-rs2586083	77312 (13)	0.2079	0.1665	0.4517	0.3972	0.0158
17q22	315	rs7207440-rs16942637	6718 (3)	0.2048	0.1665	0.4533	0.4027	0.0291
17q22	316	rs2585842-rs2109248	107102 (9)	0.2048	0.1720	0.4548	0.4027	0.0369
17q22	317	rs7211774-rs2070107	165218 (18)	0.2072	0.1775	0.4556	0.4027	0.0423
17q22	318	rs13414-rs41346650	1007740 (40)	0.2095	0.1720	0.4541	0.4082	0.0372
17q22	319	rs1868916-rs10515177	723115 (40)	0.2235	0.1720	0.4439	0.4082	0.0274
17q22	320	rs9303417-rs9890799	527113 (31)	0.2150	0.1940	0.4541	0.3917	0.0334
17q22	321	rs11655927-rs9908090	494518 (30)	0.2150	0.1885	0.4517	0.3917	0.0252

Supplementary Table 7.7.3: Software used in this study. A summary listing web URLs, version information and important parameter settings of software used in this study.

Program	Web URL	Version	Parameters
SHAPEIT	https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html	v2.r727	NCBI build 36 release 22 was used as genetic map
admixture_sim.py	http://students.washington.edu/jeanm5/		See Simulation subsection of Materials and Methods
LAMP-LD	http://lamp.icsi.berkeley.edu/lamp/lampld/	v1.0	*win-size=20 nr-founders=25*
RFMix	https://sites.google.com/site/rfmixlocalancestryinference/	v1.0.2	A window size of 0.2 cM and 10 generations were used
ADMIXTURE	http://www.genetics.ucla.edu/software/admixture/	1.21	*K=5*
PLINK	http://pngu.mgh.harvard.edu/∼purcell/plink/	v1.07	*–indep-pairwise 50 10 0.1* was used for LD filtering
Biofilter	http://ritchielab.psu.edu/software/biofilter-download/	2.1.0	LOKI database was built on 5 Dec 2013
R	www.r-project.org	3.1.0	*cor.test()* was used to estimate Pearson's correlation coefficient
ggplot2 R package	http://cran.r-project.org/web/packages/ggplot2/index.html	2.1.0.0	
hexbin R package	http://cran.r-project.org/web/packages/hexbin/index.html	1.27.0	Used to create supplementary figure 5
hierfstat R package	http://cran.r-project.org/web/packages/hierfstat/index.html	0.04-10	The *wc()* function was used to estimate pairwise F_{ST}
lme4 R package	http://cran.r-project.org/web/packages/lme4/index.html	1.1-6	The *lmer()* function was used
lmerTest R package	http://cran.r-project.org/web/packages/lmerTest/index.html	1.1-6	Used for obtaining *lmer* p-values

Supplementary Table 7.7.4: Genetic distances between the source populations of the SAC. Pairwise F_{ST} (fixation index) values between each pair of SAC source populations are summarized in this table. F_{ST} was estimated using autosomal SNPs from the source populations described in table 5.

	YRI	**CEU**	**GIH**	**JPT+CHB**
SAN	0.0918	0.185	0.173	0.216
YRI		0.132	0.119	0.162
CEU			0.034	0.108
GIH				0.074

Chapter 8

Discussion

8.1 Motivation

Tuberculosis (TB) is the second leading cause of mortality from infectious disease worldwide and was declared a global emergency by the World Health Organization in 1993 [49]. The field of TB susceptibility genetics has not yielded the answers or new insights that were expected, and it is crucial that a number of different approaches are brought to bear on this problem. The admixed South African Coloured (SAC) population is the largest demographic in metropolitan areas of Cape Town that have some of the highest reported incidences of TB worldwide, despite extensive BCG vaccination and low HIV prevalence [48]. The DST/NRF Centre for Biomedical TB Research at Stellenbosch University has studied a cohort of the SAC from these communities for many years, in the hope of discovering novel TB susceptibility genes which may at least partly explain the exceptional rate of TB in this community. The ultimate aim of the research presented in this dissertation is to generate new hypotheses regarding genetic variants that may underlie susceptibility to TB, using the large SAC genotypic data sets available.

8.2 Research highlights

Research Article 1 presented a panel of ancestry informative markers (AIMs) that are tailored to the unique and complex admixture that occurred in the SAC [99]. Panels of AIMs generated in other studies are not suited to the SAC; these studies largely ignored Khoe-San ancestry, which forms a large component of ancestry in the SAC. The panel developed in Research Article 1 accounts for the five postulated source ancestries of the SAC and the different genetic distances between them. Adjusting statistical models for the confounding effect that admixture may have on genetic association studies of the SAC is crucial, and the panel presented can be used as a cost-effective alternative to genome-wide data to achieve this goal. The panel has already proved invaluable in candidate gene association studies carried out by our centre, and has also been requested and used by a number of researchers, who study other disease mechanisms in the SAC. The python script developed and used to select the panel of AIMs has also been requested by a research group in South America, who study a four-way admixed population.

A previous genome-wide TB case-control study of the SAC showed that African San ancestry correlates positively with TB susceptibility, and showed negative correlations between European and Asian ancestries and TB susceptibility. This study was however hampered by a limited number of controls. Research Article 2 endeavours to validate this finding by using the panel of AIMs presented in Research Article 1 [100]. The substantial effect that ancestry has on TB susceptibility was confirmed. The article also explored the confounding effect that ancestry has on studies of genetic association with

TB susceptibility in the SAC, and demonstrated that association results are likely to be affected by adjustment for ancestry if allele frequencies differ markedly in the source populations of an admixed population. The ancestry proportions estimated in this study have been loaded to the DST/NRF TB research centre's database, and in conjunction with a set of R scripts developed during the preparation of this article, can readily be used by the centre's researchers for the analysis of future TB candidate gene association studies.

One of the commonly posited explanations for the missing heritability of complex disease is gene-gene interactions. Research Article 3 aimed to identify gene-gene interactions that may elucidate genetic susceptibility to TB. This study was able to utilize candidate gene association genotypic data in the discovery data set, as the panel of AIMs developed and used in Research Articles 1 and 2 enabled adjustment for ancestry in the statistical analysis. Similar to Research Article 2, the scripts used in preparation of this article can be used by our centre's researchers to test for interaction between pairs of SNPs in future candidate gene association studies. Few gene-gene interactions have ever been identified in the field of TB susceptibility, partly due to the larger study groups required to identify associations, and to lack of consensus on the optimal method of analysis. Careful consideration was given to the most appropriate methods to use for the analyses of gene-gene interactions in the SAC TB cohort. Two statistical methods were presented in Research Article 3, the one illustrating genotypic effects and the other allelic effects, reinforcing each other's results. Supporting evidence of the role of five of the gene-gene interaction models identified in the SAC discovery data set was found in an independent cohort from The Gambia. Gene pairs involved in the NF-κB pathway were also identified in the discovery data set, but could not be tested in the Gambian study group due to the lack of suitable genotypic data. Since the number of controls in many of the identified SAC models is small, future work includes additional genotyping of these variants in the SAC control cohort. Validation in other population groups, fine-mapping of causal variants and functional studies are also required to substantiate the joint role that the identified gene pairs may have on modifying susceptibility to TB.

A central requirement for admixture mapping is that the source populations of the admixed population differ in their risk of developing disease. Research Article 2 provided evidence that African ancestry, especially San ancestry, increases susceptibility to TB. In Research Article 4, genome-wide multi-way admixture mapping was used to identify regions of the genome in the SAC cohort that may harbour TB susceptibility genetic variants. This is the first admixture mapping study to be reported in the field of TB host genetics, and the technique is applied to one of the arguably most complex examples of current day admixed human populations. A number of promising regions were found, including some close to genomic regions previously implicated in TB host genetics, and regions containing immune-related susceptibility genes. Future fine-mapping and functional studies of the identified regions are however required to establish possible causal disease mechanisms.

8.2.1 Novel analyses

The articles discussed here presented novel analysis methods and novel applications of existing statistical methods to address problems specific to assessing genetic effects on susceptibility to TB in the SAC case-control cohort. These methods can however be used in studies of many other diseases and in other population groups, and are as follows:

1. The algorithm developed to select a panel of AIMs can be applied to any admixed population, regardless of the number of populations contributing to its formation,

whereas the previous incarnation of the algorithm was applicable only to three-way admixed populations.

2. The estimation of ancestry proportions using the panel of AIMs, and how these estimates can be used to adjust for ancestry in statistical models, was clearly documented. This will be useful to other researchers interested in adjusting their analysis for ancestry.

3. Mixed effect models were used to assess the association between ancestry and TB susceptibility, allowing for the inclusion of additional individuals, who would otherwise have required exclusion due to familial relationships in the cohort.

4. The gene-gene interaction study describes a convincing approach to identify pairs of genes that may jointly modify disease outcome. By limiting tests to pairs of genes based on biological plausibility, exposure to false positive associations was somewhat reduced. The use of statistical modelling to detect interactions was also computationally possible due to the reduced number of tests. The analyses could therefore be adjusted for known confounders, which is especially important in the SAC case-control cohort.

5. Interaction effects are difficult to describe and interpret. To aid interpretation, the joint effect that SNP pairs may have on TB susceptibility was illustrated in two ways. The first type of illustration compared observed proportions of genotype combinations between cases and controls, as well as the effect that the combinations have on the odds of having TB, the latter being adjusted for confounders. The second illustration compared estimated frequencies of allele combinations between cases and controls, using an expectation-maximization algorithm to resolve uncertainties.

6. One of the limitations in the admixture mapping study was bias that was introduced due to the difficulty of distinguishing between San and Bantu ancestry. By examining the direction of the bias, and how this relates to what is known about the directionality of ancestry in TB susceptibility, it was still possible to use appropriate analyses that are unlikely to be affected by this bias.

7. In the admixture mapping study, rather than testing associations between genomic regions and disease outcome separately for each genomic region, which may necessitate the use of complicated methods to correct for multiple testing [35], the regions were tested together in a single statistical model. This joint model allows for the estimation of specific regional effects, with simultaneous adjustment for all other effects, which obviates the need for multiple testing correction.

8.2.2 Limitations

Over and above the limitations that were discussed in each research article, the following limitations in the respective research articles are worth highlighting.

The panel of AIMs that were developed in Research Article 1 represent at least one marker from each of the 22 autosomal chromosomes. While the number of markers selected per chromosome are generally proportional to the size of the chromosome, there are some exceptions, which is not ideal. For example, chromosome 2 is relatively enriched for selected AIMs, whilst only one AIM was selected on chromosome 5 (figure 3.7). A similar phenomenon was observed, but on different chromosomes, when a different set of source populations were experimented with (not described in the article), yielding

a different overlapping set of markers that AIMs could be selected from. Ancestry informative markers are likely to be distributed randomly across the genome, and human chromosomes are unlikely to differ in genetic diversity. The position of the genome where this marker is found would be random, and especially when selecting a small number of markers, by chance some chromosomes will be over and under represented.

Research Article 2 investigated the association between having TB and ancestry and showed that African ancestry (San and non-San) increased the odds of having TB, whereas Asian and European ancestry were protective, when testing each ancestry separately in individual models. When all the ancestries were included in a combined model, only the African ancestries remained significant. Causality can however not be inferred from the statistical modelling. Higher African ancestry necessarily implies lower European and Asian ancestries, and this does not mean that European and Asian ancestries are not protective, it simply means that this information is already contained within the African ancestries. Both these statements are therefore possibly true: African ancestry increases susceptibility to TB, whereas European and Asian ancestries are protective, and the result of the combined statistical model does not invalidate the latter statement. It is also unclear why African ancestry may increase susceptibility to TB. Although *mycobacterium tuberculosis* most likely has its roots in Africa, where on the continent the pathogen has its origins, and which African population groups were exposed to it, are still open questions. As Africans were not exposed to modern strains of TB until the recent colonisation of Southern Africa, another possible explanation is that the observed increased susceptibility to TB relates to the mostly modern strains of TB found in the Western Cape today.

Perhaps the biggest criticism that can be lodged against the gene-gene interaction study presented in Research Article 3 is not adjusting the results for multiple testing. However, the study does not claim that the results in the South African Coloured cohort, which involved 2 million statistical tests, are statistically significant. Statistical modelling was used solely to identify the top gene pair models in this cohort that are most likely to represent real effects. Permutation testing is the gold standard for correcting for multiple tests, but it would be computationally expensive to perform for 2 million tests, and this is also not appropriate in the context of gene-gene interactions, as permutation testing does not account for the correlation between markers. Other methods that control the family-wise error rate, such as the Hochberg correction, as well as methods that control the false discovery the rate, such as the Benjami-Hocberg correction as well as Storey and Robert's q-value method, rely on the ordering of p-values or test statistics [268; 269; 270]. Hypotheses that correspond to the top results in the ordered list are then identified based on thresholds that are a function of the desired false positive rate and position of the hypothesis in the ordered list. Since the gene-gene interaction study used a data set that is a combination of different studies, that in a lot of cases involved different subsets of samples, the use of these methods may be inappropriate. Furthermore, the methods also do not completely account for the large amount of correlation that exists between genetic variants and multiple gene-gene interaction tests. The arbitrary cut-off of the top 20 gene-gene pairs used in the study is therefore arguably no better or worse than a cut-off that would have been determined by these other methods.

The admixture mapping study presented in Research Article 4 evaluated the accuracy of local ancestry inference as it pertains to three source ancestries, San, Bantu and so-called non-African ancestry, the latter representing European and South and East Asian ancestries. The software program RFMix's local ancestry inference of the European and Asian ancestries were fairly accurate, and only introduced another 3% of overall error (measured using the simulation data). Whilst this result may be interesting to other researchers, this level of detail was not presented in the study, as differences in

ancestry within the non-African ancestry component are unlikely to produce insight into TB susceptibility, which was the main focus and aim of the study. Fine-mapping of the regions that were identified in this study should also still be done. The identified regions should ideally be genotyped at a higher density, possibly including additional controls, after which association tests should be used to identify gene regions with peaks of association. Sequencing of these genes and subsequent identification of variants with functional effects may then lead to the discovery of causal variants.

8.3 Concluding remarks

Taken together, the work presented in this thesis illustrates the importance of meticulous analysis of the data in terms of correcting for as many confounding variables as possible (such as the very common confounder of admixed ancestry). Replication of genetic association studies are often unsuccessful, and one of the reasons for this is the presence of confounders [271]. The panel of AIMs developed in Research Article 1 is a valuable tool not only for investigating susceptibility to TB in the SAC, but will also enable researchers who study other diseases in the SAC to correct for ancestry, and therefore have greater confidence in the validity of their research findings.

Genetic input from a group that may be very susceptible to TB may yield important clues in the quest for susceptibility or resistance variants. The large SAC TB data sets are invaluable resources in this quest and can be used to generate new data-driven hypotheses regarding disease mechanism. The classical scientific method requires formulation of a hypothesis, which is then tested by experimentation. Formulation of such hypotheses usually requires the researcher to predict or "guess" the factors which may be at play. This might be based on informed opinions and literature, but can still be subjective and prone to bias [272]. Generating new hypotheses based on data is an important strategy for avoiding this type of bias, and this was the ultimate goal of Research Articles 3 and 4, using the SAC data sets at hand.

Both the admixture mapping and gene-gene interaction studies yielded results that could potentially have been predicted from literature. Nevertheless, the studies have gone a long way towards narrowing the search from the multitude of immune-related genetic factors, and the interactions between them, that may be involved in the immune response to TB. Non-immune pathways that have an indirect but important role in the immune response to TB also need to be considered as potential mediators in disease progression. Identifying such pathways and their relative importance may be hard to do based on literature alone. Even if possible, experimental verification is still needed. The gene-gene interaction study showed that neururegulin, glutamate receptors and the chondroitin sulfate pathway may play a role in TB, although these are not traditionally classed as immune system components. In the admixture mapping study, a number of genes were identified in the chromosome 15 region, and these have been associated with anemia. Some of the identified regions also harbour genes that may be involved in glucose metabolism and development of diabetes (e.g. the *GNG12*, *ISL1* and *HSF5* genes). It is well known that iron is essential in the growth and metabolism of *Mycobacterium tuberculosis*, and that diabetes is a risk factor for developing TB. The findings of the admixture mapping study may therefore provide direction for investigating how these factors may influence the development of TB, based on the genetics that underpin them.

It has been argued that recurrent TB infections are higher than what would be expected based on incidence rates, suggesting that the susceptibility of some individuals cannot simply be overcome with the medications and vaccines available today [65; 273]. Insight into what genetic mechanisms drive this susceptibility may therefore be essential

for effective drug and vaccine development to combat this disease, not only in the South African Coloured community, but in the many affected communities worldwide. It is my hope that some of the work presented here contributes to this important endeavour.

Glossary

admixture	Occurs when two or more previously separated populations produce offspring.
allele	One of two or more forms of a genetic locus. For organsims with two or more chromosomes, this would be the form of the genetic locus on one of the chromosomes.
ancestry informative markers	Polymorphisms with large allele frequency differences between populations.
autosome	Any chromosome that is not a sex chromosome.
bacilli	A taxonomic class of bacteria.
base pair	Two chemical bases bonded to each other to form a rung of the DNA ladder.
DNA	The main component of chromosomes that transfers genetic characteristics of organisms.
DNA microarray	A gene chip used to simultaneously genotype multiple regions of a genome.
epistasis	A phenomenon where the effect of a genetic locus on a phenotype is modified by one or more other loci.
genome	The complete set of genetic material of an organism.
genotyping	The process of determining DNA sequence values of an individual.
haplotype	A specific combination of closely located alleles on the same chromosome, usually inherited from the same parent.
heterozygosity	More than one allele is observed at a site across homologous chromosomes.
homozygosity	A single allele is observed at a site across homologous chromosomes.
HWE	If the alleles at a single site between two homologous chromosomes are independent in a group of individuals, the site is said to be in Hardy-Weinberg equilibrium (HWE).
indel	A type of genetic variant that includes insertions, deletions and combinations thereof.
linkage disequilibrium	If the alleles at two or more genetic loci are correlated in a group of individuals, the loci are said to be in linkage disequilibrium. This phenomenon occurs when alleles are not inherited independently.
local ancestry	A chromosomal region in an admixed individual received from a particular source population.
marker	A DNA sequence with a known location on a chromosome.

microsatellite	Repetitive DNA sequences, usually several base pairs in length.
mtDNA	DNA found in cell mitochondria. mtDNA is maternally inherited.
mutation	a permanent change to DNA
nucleotide	The basic structural unit of nucleic acids such as DNA and RNA.
phenotype	The observable characteristics of an individual.
polymorphism	A difference in DNA sequence between members of the same species.
population stratifaction	Occurs when a study group of individuals represents multiple population groups rather than one homogeneous population.
PCR	A molecular biology technology that produces thousands of copies of a piece of DNA.
SNP	A type of genetic variation where a single nucleotide differs between members of the same species.
strand orientation mismatch	Occurs when one genotyping platform uses the forward DNA strand for genotyping and another platform uses the reverse strand.
wild type	The most common form of a polymorphism observed in a population.

Bibliography

[1] Cavalli-Sforza, L.L. and Feldman, M.W.: The application of molecular genetic approaches to the study of human evolution. *Nature Genetics*, vol. 33, pp. 266–275, 2003.

[2] Lewontin, R.C.: The apportionment of human diversity. *Evolutionary Biology*, vol. 6, pp. 381–398, 1972.

[3] Barbujani, G., Magagni, A., Minch, E. and Cavalli-Sforza, L.L.: An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences*, vol. 94, no. 9, pp. 4516–4519, 1997.

[4] Jorde, L.B., Bamshad, M.J., Watkins, W.S., Zenger, R., Fraley, A.E., Krakowiak, P.A., Carpenter, K.D., Soodyall, H., Jenkins, T. and Rogers, A.R.: Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. *The American Journal of Human Genetics*, vol. 57, no. 3, p. 523, 1995.

[5] Bamshad, M., Wooding, S., Watkins, W., Ostler, C., Batzer, M. and Jorde, L.: Human population genetic structure and inference of group membership. *The American Journal of Human Genetics*, vol. 72, no. 3, pp. 578–589, 2003.

[6] Rosenberg, N., Pritchard, J., Weber, J., Cann, H., Kidd, K., Zhivotovsky, L. and Feldman, M.: Genetic structure of human populations. *Science*, vol. 298, no. 5602, p. 2381, 2002.

[7] Witherspoon, D.J., Wooding, S., Rogers, A.R., Marchani, E.E., Watkins, W.S., Batzer, M.A. and Jorde, L.B.: Genetic similarities within and between human populations. *Genetics*, vol. 176, no. 1, pp. 351–359, 2007.

[8] Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo, J.-M., Wambebe, C. and Tishkoff, S.A.: Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proceedings of the National Academy of Sciences*, vol. 107, no. 2, pp. 786–791, 2010.

[9] Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.-M. and Doumbo, O.: The genetic structure and history of Africans and African Americans. *Science*, vol. 324, no. 5930, pp. 1035–1044, 2009.

[10] Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S. and Nelson, M.R.: Genes mirror geography within Europe. *Nature*, vol. 456, no. 7218, pp. 98–101, 2008.

[11] Elhaik, E., Tatarinova, T., Chebotarev, D., Piras, I.S., Calò, C.M., De Montis, A., Atzori, M., Marini, M., Tofanelli, S. and Francalacci, P.: Geographic population structure analysis of worldwide human populations infers their biogeographical origins. *Nature Communications*, vol. 5, 2014.

[12] Cooper, R.S., Tayo, B. and Zhu, X.: Genome-wide association studies: implications for multiethnic samples. *Human Molecular Genetics*, vol. 17, no. R2, pp. R151–R155, 2008.

[13] Tishkoff, S.A., Gonder, M.K., Henn, B.M., Mortensen, H., Knight, A., Gignoux, C., Fernandopulle, N., Lema, G., Nyambo, T.B. and Ramakrishnan, U.: History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Molecular Biology and Evolution*, vol. 24, no. 10, p. 2180–2195, 2007.

[14] Behar, D.M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S. and Comas, D.: The dawn of human matrilineal diversity. *The American Journal of Human Genetics*, vol. 82, no. 5, pp. 1130–1140, 2008.

[15] Henn, B., Gignoux, C., Jobin, M., Granka, J., Macpherson, J., Kidd, J., Rodríguez-Botigué, L., Ramachandran, S., Hon, L., Brisbin, A., Lin, A.A., Underhill, P.A., Comas, D., Kidd, K.K., Norman, P.J., Parham, P., Bustamante, C.D., Mountain, J.L. and Feldmane, M.W.: Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences*, vol. 108, no. 13, p. 5154, 2011.

[16] Gomez, F., Hirbo, J. and Tishkoff, S.A.: Genetic variation and adaptation in Africa: Implications for human evolution and disease. *Cold Spring Harbor Perspectives in Biology*, vol. 6, no. 7, p. a008524, 2014.

[17] 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. and McVean, G.A.: A map of human genome variation from population scale sequencing. *Nature*, vol. 467, no. 7319, pp. 1061–1073, 2010.

[18] Prugnolle, F., Manica, A. and Balloux, F.: Geography predicts neutral genetic diversity of human populations. *Current Biology*, vol. 15, no. 5, pp. R159–R160, 2005.

[19] Cavalli-Sforza, L.L., Minch, E. and Mountain, J.L.: Coevolution of genes and languages revisited. *Proceedings of the National Academy of Sciences*, vol. 89, no. 12, pp. 5620–5624, 1992.

[20] Henn, B.M., Cavalli-Sforza, L.L. and Feldman, M.W.: The great human expansion. *Proceedings of the National Academy of Sciences*, vol. 109, no. 44, pp. 17758–17764, 2012.

[21] Cavalli-Sforza, L.L., Piazza, A., Menozzi, P. and Mountain, J.: Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences*, vol. 85, no. 16, p. 6002–6006, 1988.

[22] Atkinson, Q.D.: Phonemic diversity supports a serial founder effect model of language expansion from Africa. *Science*, vol. 332, no. 6027, pp. 346–349, 2011.

[23] Xu, S. and Jin, L.: A genome-wide analysis of admixture in Uyghurs and a high-density admixture map for disease-gene discovery. *The American Journal of Human Genetics*, vol. 83, no. 3, pp. 322–336, 2008.

[24] Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Z., Gusic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. and Pääbo, S.: A draft sequence of the Neandertal genome. *Science*, vol. 328, no. 5979, pp. 710–722, 2010.

[25] Stoneking, M. and Krause, J.: Learning about human population history from ancient and modern genomes. *Nature Reviews Genetics*, vol. 12, no. 9, pp. 603–614, 2011.

[26] De Wit, E., Delport, W., Rugamika, C.E., Meintjes, A., Möller, M., van Helden, P.D., Seoighe, C. and Hoal, E.G.: Genome-wide analysis of the structure of the South African Coloured Population in the Western Cape. *Human Genetics*, vol. 128, no. 2, pp. 145–153, 2010.

[27] Halder, I. and Shriver, M.D.: Measuring and using admixture to study the genetics of complex diseases. *Human Genomics*, vol. 1, no. 1, p. 52r62, 2003.

[28] Shriner, D., Adeyemo, A., Ramos, E., Chen, G. and Rotimi, C.N.: Mapping of disease-associated variants in admixed populations. *Genome Biology*, vol. 12, no. 5, p. 223, 2011.

[29] Reich, D., Patterson, N., De Jager, P.L., McDonald, G.J., Waliszewska, A., Tandon, A., Lincoln, R.R., DeLoa, C., Fruhan, S.A., Cabre, P., Bera, O., Semana, G., Kelly, M.A., Francis, D.A., Ardlie, K., Khan, O., Cree, B.A.C., Hauser, S.L., Oksenberg, J.R. and Hafler, D.A.: A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nature Genetics*, vol. 37, no. 10, pp. 1113–1118, 2005.

[30] Smith, M.W. and O'Brien, S.J.: Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nature Reviews Genetics*, vol. 6, no. 8, pp. 623–632, 2005.

[31] Reich, D. and Patterson, N.: Will admixture mapping work to find disease genes? *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 360, no. 1460, pp. 1605–1607, 2005.

[32] Winkler, C.A., Nelson, G.W. and Smith, M.W.: Admixture mapping comes of age. *Annual Review of Genomics and Human Genetics*, vol. 11, pp. 65–89, 2010.

[33] Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S. and Manolio, T.A.: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, vol. 106, no. 23, pp. 9362–9367, 2009.

[34] Freedman, M.L., Haiman, C.A., Patterson, N., McDonald, G.J., Tandon, A., Waliszewska, A., Penney, K., Steen, R.G., Ardlie, K., John, E.M., Oakley-Girvan, I., Whittemore, A.S., Cooney, K.A., Ingles, S.A., Altshuler, D., Henderson, B.E. and Reich, D.: Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proceedings of the National Academy of Sciences*, vol. 103, no. 38, pp. 14068–14073, 2006.

[35] Torgerson, D.G., Gignoux, C.R., Galanter, J.M., Drake, K.A., Roth, L.A., Eng, C., Huntsman, S., Torres, R., Avila, P.C., Chapela, R., Ford, J.G., Rodríguez-Santana, J.R., Rodrïguez-Cintrön, W., Hernandez, R.D. and Burchard, E.G.: Case-control admixture mapping in Latino populations enriches for known asthma-associated genes. *Journal of Allergy and Clinical Immunology*, vol. 130, no. 1, pp. 76–82, 2012.

[36] Fejerman, L., Chen, G.K., Eng, C., Huntsman, S., Hu, D., Williams, A., Pasaniuc, B., John, E.M., Via, M., Gignoux, C., Ingles, S., Monroe, K.R., Kolonel, L.N., Torres-Mejía, G., Pérez-Stable, E.J., Burchard, E.G., Henderson, B.E., Haiman, C.A. and Ziv, E.: Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas. *Human Molecular Genetics*, vol. 21, no. 8, pp. 1907–1917, 2012.

[37] Parra, E.J., Below, J.E., Krithika, S., Valladares, A., Barta, J.L., Cox, N.J., Hanis, C.L., Wacher, N., Garcia-Mena, J., Hu, P., Shriver, M.D., Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium, Kumate, J., McKeigue, P.M., Escobedo, J. and Cruz, M.: Genome-wide association study of type 2 diabetes in a sample from Mexico City and a meta-analysis of a Mexican-American sample from Starr County, Texas. *Diabetologia*, vol. 54, no. 8, pp. 2038–2046, 2011.

[38] Cordell, H.J. and Clayton, D.G.: Genetic association studies. *The Lancet*, vol. 366, no. 9491, pp. 1121–1131, 2005.

[39] Chimusa, E.R., Daya, M., Möller, M., Ramesar, R., Henn, B.M., van Helden, P.D., Mulder, N.J. and Hoal, E.G.: Determining ancestry proportions in complex admixture scenarios in South Africa using a novel proxy ancestry selection method. *PLoS ONE*, vol. 8, no. 9, p. e73971, September 2013.

[40] Patterson, N., Petersen, D.C., Van Der Ross, R.E., Sudoyo, H., Glashoff, R.H., Marzuki, S., Reich, D. and Hayes, V.M.: Genetic structure of a unique admixed population: implications for medical research. *Human Molecular Genetics*, vol. 19, no. 3, pp. 411–419, 2010.

[41] Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Zaporozhchenko, V., Bormans, C., van Helden, P.D., Hoal, E.G. and Behar, D.M.: Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture. *The American Journal of Human Genetics*, vol. 86, no. 4, pp. 611–620, 2010.

[42] Boonzaier, E.: *The Cape herders: a history of the Khoikhoi of southern Africa*. New Africa Books, 1996.

[43] Elphick, R.: *Khoikhoi and the founding of White South Africa*. Ravan Press, Johannesburg, 1985.

[44] Mountain, A.: *The First People of the Cape, 1st edn.* David Phillips Publishers, Cape Town, 2003.

[45] Shell, R.: *Children of Bondage*. Witwatersrand University Press, Johannesburg, 1994.

[46] Nurse GT, Weiner JS, J.T.: *The peoples of Southern Africa and their affinities*. Clarendon Press, Oxford, 1985.

[47] Keegan, T.: *Colonial South Africa and the origins of the racial order*. University of Virginia Press, 1996.

[48] Hoal, E.G., Lewis, L.A., Jamieson, S.E., Tanzer, F., Rossouw, M., Victor, T., Hillerman, R., Beyers, N., Blackwell, J.M. and Van Helden, P.D.: SLC11a1 (NRAMP1) but not SLC11a2 (NRAMP2) polymorphisms are associated with susceptibility to tuberculosis in a high-incidence community in South Africa. *The International Journal of Tuberculosis and Lung Disease*, vol. 8, no. 12, pp. 1464–1471, 2004.

[49] WHO: Global tuberculosis report. Report 2013.11, World Health Organization (WHO), 2013.

[50] MRC: Revised burden of disease estimates for the comparative risk factor assessment, South Africa 2000. Report June 2006, Medical Research Unit, South Africa, 2006.

[51] Arriaza, B.T., Salo, W., Aufderheide, A.C. and Holcomb, T.A.: Pre-Columbian tuberculosis in Northern Chile: Molecular and skeletal evidence. *American Journal of Physical Anthropology*, vol. 98, no. 1, pp. 37–45, 1995.

[52] Cave, A.J.E. and Demonstrator, A.: The evidence for the incidence of tuberculosis in ancient Egypt. *British Journal of Tuberculosis*, vol. 33, no. 3, pp. 142–152, 1939.

[53] Daniel, T.M.: The history of tuberculosis. *Respiratory Medicine*, vol. 100, no. 11, pp. 1862–1870, 2006.

[54] Daniel, T.M.: The origins and precolonial epidemiology of tuberculosis in the Americas: can we figure them out? *The International Journal of Tuberculosis and Lung Disease*, vol. 4, no. 5, pp. 395–400, 2000.

[55] Salo, W.L., Aufderheide, A.C., Buikstra, J. and Holcomb, T.A.: Identification of Mycobacterium tuberculosis DNA in a pre-Columbian Peruvian mummy. *Proceedings of the National Academy of Sciences*, vol. 91, no. 6, pp. 2091–2094, 1994.

[56] Zimmerman, M.R.: Pulmonary and osseous tuberculosis in an Egyptian mummy. *Bulletin of the New York Academy of Medicine*, vol. 55, no. 6, p. 604, 1979.

[57] Smith, N.H., Hewinson, R.G., Kremer, K., Brosch, R. and Gordon, S.V.: Myths and misconceptions: the origin and evolution of Mycobacterium tuberculosis. *Nature Reviews Microbiology*, vol. 7, no. 7, pp. 537–544, 2009.

[58] Brosch, R., Gordon, S.V., Marmiesse, M., Brodin, P., Buchrieser, C., Eiglmeier, K., Garnier, T., Gutierrez, C., Hewinson, G., Kremer, K., Parsons, L.M., Pym, A.S., Samper, S., van Soolingen, D. and Cole, S.T.: A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proceedings of the National Academy of Sciences*, vol. 99, no. 6, pp. 3684–3689, 2002.

[59] Gagneux, S., DeRiemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K. and Gutierrez, M.C.: Variable host-pathogen compatibility in Mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, vol. 103, no. 8, pp. 2869–2873, 2006.

[60] Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S., Homolka, S., Roach, J.C., Kremer, K., Petrov, D.A. and Feldman, M.W.: High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography. *PLoS Biology*, vol. 6, no. 12, p. e311, 2008.

[61] Mostowy, S., Cousins, D., Brinkman, J., Aranaz, A. and Behr, M.A.: Genomic deletions suggest a phylogeny for the Mycobacterium tuberculosis complex. *Journal of Infectious Diseases*, vol. 186, no. 1, pp. 74–80, 2002.

[62] Wirth, T., Hildebrand, F., Allix-Béguec, C., Wölbeling, F., Kubica, T., Kremer, K., van Soolingen, D., Rüsch-Gerdes, S., Locht, C. and Brisse, S.: Origin, spread and demography of the Mycobacterium tuberculosis complex. *PLoS Pathogens*, vol. 4, no. 9, p. e1000160, 2008.

[63] Russell, D.G., Barry, C.E. and Flynn, J.L.: Tuberculosis: what we don't know can, and does, hurt us. *Science*, vol. 328, no. 5980, p. 852–856, 2010.

[64] Lienhardt, C., Bennett, S., Del Prete, G., Bah-Sow, O., Newport, M., Gustafson, P., Manneh, K., Gomes, V., Hill, A. and McAdam, K.: Investigation of environmental and host-related risk factors for tuberculosis in Africa. I. Methodological aspects of a combined design. *American Journal of Epidemiology*, vol. 155, no. 11, p. 1066–1073, 2002.

[65] Abel, L., El-Baghdadi, J., Bousfiha, A.A., Casanova, J.-L. and Schurr, E.: Human genetics of tuberculosis: a long and winding road. *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1645, p. 20130428, 2014.

[66] Möller, M. and Hoal, E.G.: Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. *Tuberculosis*, vol. 90, no. 2, pp. 71–83, 2010.

[67] Newport, M.J., Goetghebuer, T. and Marchant, A.: Hunting for immune response regulatory genes: vaccination studies in infant twins. *Expert Review of Vaccines*, vol. 4, no. 5, pp. 739–746, 2005.

[68] Barreiro, L.B. and Quintana-Murci, L.: From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nature Reviews Genetics*, vol. 11, no. 1, pp. 17–30, 2010.

[69] Sørensen, T.I., Nielsen, G.G., Andersen, P.K. and Teasdale, T.W.: Genetic and environmental influences on premature death in adult adoptees. *New England Journal of Medicine*, vol. 318, no. 12, p. 727–732, 1988.

[70] Rieder, H.L.: Clarification of the Luebeck infant tuberculosis. *Pneumologie (Stuttgart, Germany)*, vol. 57, no. 7, p. 402, 2003.

[71] Cobat, A., Gallant, C.J., Simkin, L., Black, G.F., Stanley, K., Hughes, J., Doherty, T.M., Hanekom, W.A., Eley, B., Beyers, N., Jaïs, J.-P., van Helden, P., Abel, L., Hoal, E.G., Alcaös, A. and Schurr, E.: High heritability of antimycobacterial immunity in an area of hyperendemicity for tuberculosis disease. *Journal of Infectious Diseases*, vol. 201, no. 1, pp. 15–19, 2010.

[72] Comstock, G.: Tuberculosis in twins: a re-analysis of the Prophit survey. *American Review of Respiratory Disease*, vol. 117, no. 4, pp. 621–624, 1978.

[73] Cooper, A.M.: Cell-mediated immune responses in tuberculosis. *Annual Review of Immunology*, vol. 27, pp. 393–422, 2009.

[74] Di Pietrantonio, T. and Schurr, E.: Mouse models for the genetic study of tuberculosis susceptibility. *Briefings in Functional Genomics & Proteomics*, vol. 4, no. 3, pp. 277–292, 2005.

[75] Diehl, K., von Verschuer, O., Arzt, G., von Verschuer, O., Médecin, G., von Verschuer, O. and Physician, G.: *Der Erbeinfluss bei der Tuberkulose*. G. Fischer, 1936.

[76] Dubos, R.J. and Dubos, J.: *The white plague: tuberculosis, man, and society*. Rutgers Univ Pr, 1952.

[77] Fortin, A., Abel, L., Casanova, J.L. and Gros, P.: Host genetics of mycobacterial diseases in mice and men: forward genetic studies of BCG-osis and tuberculosis. *The Annual Review of Genomics and Human Genetics*, vol. 8, pp. 163–192, 2007.

[78] Harvald, B. and Hauge, M.: Hereditary factors elucidated by twin studies. *Genetics and the Epidemiology of Chronic Diseases. Washington, DC: Department of Health, Education and Welfare*, pp. 61–76, 1965.

[79] Kallmann, F. and Reisner, D.: Twin studies on the significance of genetic factors in tuberculosis. *American Review of Tuberculosis*, vol. 47, pp. 549–574, 1942.

[80] O'Garra, A., Redford, P.S., McNab, F.W., Bloom, C.I., Wilkinson, R.J. and Berry, M.P.: The immune response in tuberculosis. *Annual Review of Immunology*, vol. 31, pp. 475–527, 2013.

[81] Simonds, B.: Tuberculosis in twins. *Tuberculosis in Twins.*, 1963.

[82] Stein, C.M., Guwatudde, D., Nakakeeto, M., Peters, P., Elston, R.C., Tiwari, H.K., Mugerwa, R. and Whalen, C.C.: Heritability analysis of cytokines as intermediate phenotypes of tuberculosis. *Journal of Infectious Diseases*, vol. 187, no. 11, pp. 1679–1685, 2003.

[83] Stein, C.M., Nshuti, L., Chiunda, A.B., Boom, W.H., Elston, R.C., Mugerwa, R.D., Iyengar, S.K. and Whalen, C.C.: Evidence for a major gene influence on tumor necrosis factor-α expression in tuberculosis: Path and segregation analysis. *Human Heredity*, vol. 60, no. 2, pp. 109–118, 2005.

[84] Uehlinger, E. and Künsch, M.: Über zwillingstuberkulose. *Lung*, vol. 92, no. 4, pp. 275–370, 1938.

[85] Vidal, S.M., Malo, D., Vogan, K., Skamene, E. and Gros, P.: Natural resistance to infection with intracellular parasites: Isolation of a candidate for *BCG*. *Cell*, vol. 73, no. 3, pp. 469–485, 1993.

[86] Stead, W.W.: Genetics and resistance to Tuberculosis could resistance be enhanced by genetic engineering? *Annals of Internal Medicine*, vol. 116, no. 11, pp. 937–941, 1992.

[87] Stead, W.W., Senner, J.W., Reddick, W.T. and Lofgren, J.P.: Racial differences in susceptibility to infection by Mycobacterium tuberculosis. *New England Journal of Medicine*, vol. 322, no. 7, 1990.

[88] Nahid, P., Horne, D.J., Jarlsberg, L.G., Reiner, A.P., Osmond, D., Hopewell, P.C. and Bibbins-Domingo, K.: Racial differences in tuberculosis infection in United States communities: The coronary artery risk development in young adults study. *Clinical Infectious Diseases*, vol. 53, no. 3, pp. 291–294, 2011.

[89] Hoge, C.W., Fisher, L., Donnell, H.D., Dodson, D.R., Tomlinson, G.V., Breiman, R.F., Bloch, A.B. and Good, R.C.: Risk factors for transmission of Mycobacterium tuberculosis in a primary school outbreak: lack of racial difference in susceptibility to infection. *American Journal of Epidemiology*, vol. 139, no. 5, pp. 520–530, 1994.

[90] Crowle, A.J. and Elkins, N.: Relative permissiveness of macrophages from black and white people for virulent tubercle bacilli. *Infection and Immunity*, vol. 58, no. 3, pp. 632–638, 1990.

[91] Coussens, A.K., Wilkinson, R.J., Nikolayevskyy, V., Elkington, P.T., Hanifa, Y., Islam, K., Timms, P.M., Bothamley, G.H., Claxton, A.P. and Packe, G.E.: Ethnic variation in inflammatory profile in tuberculosis. *PLoS Pathogens*, vol. 9, no. 7, p. e1003468, 2013.

[92] Wilson, L.G.: The historical decline of tuberculosis in Europe and America: its causes and significance. *Journal of the History of Medicine and Allied Sciences*, vol. 45, no. 3, pp. 366–396, 1990.

[93] O'Brien, S.J.: Ghetto legacy. *Current Biology*, vol. 1, no. 4, pp. 209–211, 1991.

[94] Motulsky, A.G.: Metabolic polymorphisms and the role of infectious diseases in human evolution. *Human Biology*, vol. 32, p. 28, 1960.

[95] Sousa, A.O., Salem, J.I., Lee, F.K., Verçosa, M.C., Cruaud, P., Bloom, B.R., Lagrange, P.H. and David, H.L.: An epidemic of tuberculosis with a high rate of tuberculin anergy among a population previously unexposed to tuberculosis, the Yanomami Indians of the Brazilian Amazon. *Proceedings of the National Academy of Sciences*, vol. 94, no. 24, pp. 13227–13232, 1997.

[96] Cantwell, M.F., McKenna, M.T., McCray, E. and Onorato, I.M.: Tuberculosis and race/ethnicity in the United States: impact of socioeconomic status. *American Journal of Respiratory and Critical Care Medicine*, vol. 157, no. 4, pp. 1016–1020, 1998.

[97] Elender, F., Bentham, G. and Langford, I.: Tuberculosis mortality in England and Wales during 1982–1992: its association with poverty, ethnicity and AIDS. *Social Science & Medicine*, vol. 46, no. 6, pp. 673–681, 1998.

[98] Chimusa, E.R., Zaitlen, N., Daya, M., Möller, M., van Helden, P.D., Mulder, N.J., Price, A.L. and Hoal, E.G.: Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. *Human Molecular Genetics*, vol. 23, no. 3, pp. 796–809, 2013.

[99] Daya, M., van der Merwe, L., Galal, U., Möller, M., Salie, M., Chimusa, E.R., Galanter, J.M., van Helden, P.D., Henn, B.M. and Gignoux, C.R.: A panel of ancestry informative markers for the complex five-way admixed South African Coloured population. *PLoS ONE*, vol. 8, no. 12, p. e82224, 2013.

[100] Daya, M., van der Merwe, L., van Helden, P.D., Möller, M. and Hoal, E.G.: The role of ancestry in TB susceptibility of an admixed South African population. *Tuberculosis*, vol. 94, no. 4, pp. 413–420, 2014.

[101] Laurie, C.C., Doheny, K.F., Mirel, D.B., Pugh, E.W., Bierut, L.J., Bhangale, T., Boehm, F., Caporaso, N.E., Cornelis, M.C., Edenberg, H.J., Gabriel, S.B., Harris, E.L., Hu, F.B., Jacobs, K.B., Kraft, P., Landi, M.T., Lumley, T., Manolio, T.A., McHugh, C., Painter, I., Paschall, J., Rice, J.P., Rice, K.M., Zheng, X., Weir, B.S. and GENEVA Investigators: Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, vol. 34, no. 6, pp. 591–602, 2010.

[102] Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, vol. 447, no. 7145, pp. 661–678, June 2007.

[103] Miyagawa, T., Nishida, N., Ohashi, J., Kimura, R., Fujimoto, A., Kawashima, M., Koike, A., Sasaki, T., Tanii, H., Otowa, T., Momose, Y., Nakahara, Y., Gotoh, J., Okazaki, Y., Tsuji, S. and Tokunaga, K.: Appropriate data cleaning methods for genome-wide association study. *Journal of Human Genetics*, vol. 53, no. 10, pp. 886–893, 2008.

[104] Ziegler, A., König, I. and Thompson, J.: Biostatistical aspects of genome-wide association studies. *Biometrical Journal*, vol. 50, no. 1, pp. 8–28, 2008.

[105] Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J. and Sham, P.C.: PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, September 2007.

[106] R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. ISBN 3-900051-07-0.
Available at: `http://www.R-project.org/`

[107] New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nature Genetics*, vol. 39, no. 9, pp. 1045–1051, September 2007.

[108] Thye, T., Vannberg, F.O., Wong, S.H., Owusu-Dabo, E., Osei, I., Gyapong, J., Sirugo, G., Sisay-Joof, F., Enimil, A., Chinbuah, M.A., Floy, S., Warndorff, D.K., Sichal, L., Malem, S., Crampin, A.C., Ngwir, B., Teo, Y.Y., Smal, K., Rocket, K., Kwiatkowsk, D., Fine, P.E., Hill, P.C., Newpor, M., Lienhard, C., Adegbola, R.A., Corra, T., Ziegle, A., Morris, A.P., Meyer, C.G., Horstmann, R.D. and S Hill, A.V.: Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11. 2. *Nature Genetics*, vol. 42, no. 9, pp. 739–741, 2010.

[109] Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D.: Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[110] Tang, H., Quertermous, T., Rodriguez, B., Kardia, S.L.R., Zhu, X., Brown, A., Pankow, J.S., Province, M.A., Hunt, S.C. and Boerwinkle, E.: Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *The American Journal of Human Genetics*, vol. 76, no. 2, p. 268, 2005.

[111] Lander, E.S. and Schork, N.J.: Genetic dissection of complex traits. *Science*, vol. 265, pp. 2037–2048, 1994.

[112] Marchini, J., Cardon, L.R., Phillips, M.S. and Donnelly, P.: The effects of human population structure on large genetic association studies. *Nature Genetics*, vol. 36, no. 5, pp. 512–517, 2004.

[113] Wu, C., DeWan, A., Hoh, J. and Wang, Z.: A comparison of association methods correcting for population stratification in case-control studies. *Annals of Human Genetics*, vol. 75, no. 3, pp. 418–427, 2011.

[114] Kosoy, R., Nassir, R., Tian, C., White, P.A., Butler, L.M., Silva, G., Kittles, R., Alarcon-Riquelme, M.E., Gregersen, P.K., Belmont, J.W., De La Vega, F.M. and Seldin, M.F.: Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Human Mutation*, vol. 30, no. 1, pp. 69–78, January 2009. ISSN 1098-1004.

[115] Nassir, R., Kosoy, R., Tian, C., White, P.A., Butler, L.M., Silva, G., Kittles, R., Alarcon-Riquelme, M.E., Gregersen, P.K., Belmont, J.W., De La Vega, F.M. and Seldin, M.F.: An ancestry informative marker set for determining continental origin: validation and extension using human genome diversity panels. *BMC Genetics*, vol. 10, no. 1, p. 39, 2009.

[116] Galanter, J.M., Fernandez-Lopez, J.C., Gignoux, C.R., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., Hidalgo-Miranda, A., Contreras, A.V., Figueroa, L.U., Raska, P., Jimenez-Sanchez, G., Silva Zolezzi, I., Torres, M., Ponte, C.R., Ruiz, Y., Salas, A., Nguyen, E., Eng, C., Borjas, L., Zabala, W., Barreto, G., Rondón González, F., Ibarra, A., Taboada, P., Porras, L., Moreno, F., Bigham, A., Gutierrez, G., Brutsaert, T., León-Velarde, F., Moore, L.G., Vargas, E., Cruz, M., Escobedo, J., Rodriguez-Santana, J., Rodriguez-Cintrón, W., Chapela, R., Ford, J.G., Bustamante, C., Seminara, D., Shriver, M., Ziv, E., Gonzalez Burchard, E., Haile, R., Parra, E., Carracedo, A. and for the LACE Consortium: Development of a panel of genome-wide ancestry informative markers to study admixture throughout the Americas. *PLoS Genetics*, vol. 8, no. 3, p. e1002554, March 2012.

[117] Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M. and Jones, K.W.: The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. *Human Genomics*, vol. 1, no. 4, pp. 274–286, 2004.

[118] Lao, O., Duijn, K., Kersbergen, P., Knijff, P. and Kayser, M.: Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *The American Journal of Human Genetics*, vol. 78, no. 4, pp. 680–690, 2006.

[119] Rosenberg, N., Li, L., Ward, R. and Pritchard, J.: Informativeness of genetic markers for inference of ancestry. *The American Journal of Human Genetics*, vol. 73, no. 6, pp. 1402–1422, 2003.

[120] Rosenberg, N.A.: Algorithms for selecting informative marker panels for population assignment. *Journal of Computational Biology*, vol. 12, no. 9, pp. 1183–1201, November 2005. ISSN 1066-5277, 1557-8666.

[121] Paschou, P., Ziv, E., Burchard, E.G., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M.W. and Drineas, P.: PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genetics*, vol. 3, no. 9, p. e160, 2007.

[122] Mountain, A.: *An unsung heritage*. David Phillips Publishers, Cape Town, 2004.

[123] Halder, I., Shriver, M., Thomas, M., Fernandez, J.R. and Frudakis, T.: A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *Human Mutation*, vol. 29, no. 5, pp. 648–658, May 2008.

[124] Collins-Schramm, H.E., Hanson, R.L., Knowler, W.C., Silva, G., Seldin, M.F., Chima, B., Morii, T., Wah, K., Figueroa, Y., Criswell, L.A. and Belmont, J.W.: Mexican American ancestry-informative markers: examination of population structure and marker characteristics in European Americans, Mexican Americans, Amerindians and Asians. *Human Genetics*, vol. 114, no. 3, pp. 263–271, February 2004.

[125] Phillips, C., Salas, A., Sánchez, J., Fondevila, M., Gómez-Tato, A., Alvarez-Dios, J., Calaza, M., de Cal, M., Ballard, D., Lareu, M., Lareu, M.V., Carracedo, A. and SNPforID Consortium: Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Science International: Genetics*, vol. 1, no. 3-4, pp. 273–280, 2007.

[126] Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A. and Blum, M.G.B.: Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science*, vol. 338, 2012.

[127] Bensmail, C.: Efficient algorithms for selecting the best proxy ancestry in recently admixed populations: Application to infer the local ancestry in the Maghreb and South African Coloured populations. Essay 1018, African Institute for Mathematical Sciences (AIMS), 2012. `http://archive.aims.ac.za/2011-12/chamabensmail.pdf`.

[128] Schuster, S.C., Miller, W., Ratan, A., Tomsho, L.P., Giardine, B., Kasson, L.R., Harris, R.S., Petersen, D.C., Zhao, F. and Qi, J.: Complete Khoisan and Bantu genomes from southern Africa. *Nature*, vol. 463, no. 7283, pp. 943–947, 2010.

[129] Alexander, D., Novembre, J. and Lange, K.: Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, vol. 19, no. 9, pp. 1655–1664, 2009.

[130] Morin, P.A., Martien, K.K. and Taylor, B.L.: Assessing statistical power of SNPs for population structure and conservation studies. *Molecular Ecology Resources*, vol. 9, no. 1, pp. 66–73, 2009.

[131] Rayner, B.L., Owen, E.P., King, J.A., Soule, S.G., Vreede, H., Opie, L.H., Marais, D. and Davidson, J.S.: A new mutation, R563Q, of the beta subunit of the epithelial sodium channel associated with low-renin, low-aldosterone hypertension. *Journal of Hypertension*, vol. 21, no. 5, pp. 921–926, 2003.

[132] Viljoen, D.L., Carr, L.G., Foroud, T.M., Brooke, L., Ramsay, M. and Li, T.K.: Alcohol dehydrogenase-2*2 allele is associated with decreased prevalence of fetal alcohol syndrome in the mixed-ancestry population of the Western Cape province, South Africa. *Alcoholism: Clinical and Experimental Research*, vol. 25, no. 12, pp. 1719–1722, 2001.

[133] Zaahl, M.G., Winter, T., Warnich, L. and Kotze, M.J.: Analysis of the three common mutations in the CARD15 gene (R702W, G908R and 1007fs) in South African colored patients with inflammatory bowel disease. *Molecular and Cellular Probes*, vol. 19, no. 4, pp. 278–281, 2005.

[134] Fernandez, P., de Beer, P.M., van der Merwe, L. and Heyns, C.F.: COX-2 promoter polymorphisms and the association with prostate cancer risk in South African men. *Carcinogenesis*, vol. 29, no. 12, pp. 2347–2350, 2008.

[135] Dandara, C., Ballo, R. and Iqbal Parker, M.: CYP3A5 genotypes and risk of oesophageal cancer in two South African populations. *Cancer Letters*, vol. 225, no. 2, pp. 275–282, 2005.

[136] Möller, M., Nebel, A., Valentonyte, R., van Helden, P.D., Schreiber, S. and Hoal, E.G.: Investigation of chromosome 17 candidate genes in susceptibility to TB in a South African population. *Tuberculosis*, vol. 89, no. 2, pp. 189–194, 2009.

[137] Barreiro, L.B., Neyrolles, O., Babb, C.L., van Helden, P.D., Gicquel, B., Hoal, E.G. and Quintana-Murci, L.: Length variation of DC-SIGN and L-SIGN neck-region has no impact on tuberculosis susceptibility. *Human Immunology*, vol. 68, no. 2, pp. 106–112, 2007.

[138] Möller, M., Flachsbart, F., Till, A., Thye, T., Horstmann, R.D., Meyer, C.G., Osei, I., van Helden, P.D., Hoal, E.G. and Schreiber, S.: A functional haplotype in the 3' untranslated region of TNFRSF1B is associated with tuberculosis in two African populations. *American Journal of Respiratory and Critical Care Medicine*, vol. 181, no. 4, pp. 388–393, 2010.

[139] Möller, M., Nebel, A., van Helden, P.D., Schreiber, S. and Hoal, E.G.: Analysis of eight genes modulating interferon gamma and human genetic susceptibility to tuberculosis: a case-control association study. *BMC Infectious Diseases*, vol. 10, no. 1, p. 154, 2010.

[140] De Wit, E., van der Merwe, L., van Helden, P. and Hoal, E.: Gene-gene interaction between tuberculosis candidate genes in a South African population. *Mammalian Genome*, vol. 22, pp. 1–11, 2010.

[141] Adams, L.A., Möller, M., Nebel, A., Schreiber, S., van der Merwe, L., van Helden, P.D. and Hoal, E.G.: Polymorphisms in MC3R promoter and CTSZ 3' UTR are associated with tuberculosis susceptibility. *European Journal of Human Genetics*, vol. 19, no. 6, pp. 676–681, 2011.

[142] Wacholder, S., Rothman, N. and Caporaso, N.: Population stratification in epidemiologic studies of common genetic variants and cancer: Quantification of bias. *Journal of the National Cancer Institute*, vol. 92, no. 14, pp. 1151–1158, July 2000.

[143] Lehohla, P.: Census 2011 census in brief. Report 03-01-41, Statistics South Africa, 2012.

[144] Bates, D., Maechler, M. and Bolker, B.: *lme4: Linear mixed-effects models using S4 classes*, 2012. R package version 0.999999-0.
Available at: `http://CRAN.R-project.org/package=lme4`

[145] Keppel, K.G.: Ten largest racial and ethnic health disparities in the United States based on healthy people 2010 objectives. *American Journal of Epidemiology*, vol. 166, no. 1, pp. 97–103, 2007.

[146] Serpa, J.A., Teeter, L.D., Musser, J.M. and Graviss, E.A.: Tuberculosis disparity between US-born blacks and whites, Houston, Texas, USA. *Emerging Infectious Diseases*, vol. 15, no. 6, p. 899, 2009.

[147] Corbett, E.L., Watt, C.J., Walker, N., Maher, D., Williams, B.G., Raviglione, M.C. and Dye, C.: The growing burden of tuberculosis: global trends and interactions with the HIV epidemic. *Archives of Internal Medicine*, vol. 163, no. 9, p. 1009, 2003.

[148] Barreiro, L.B., Neyrolles, O., Babb, C.L., Tailleux, L., Quach, H., McElreavey, K., Van Helden, P.D., Hoal, E.G., Gicquel, B. and Quintana-Murci, L.: Promoter variation in the DC-SIGN-encoding gene CD209 is associated with tuberculosis. *PLoS Medicine*, vol. 3, no. 2, p. e20, 2006.

[149] Baldini, M., Carla Lohman, I., Halonen, M., Erickson, R.P., Holt, P.G. and Martinez, F.D.: A polymorphism in the 5' flanking region of the CD14 gene is associated with circulating soluble CD14 levels and with total serum immunoglobulin e. *American Journal of Respiratory Cell and Molecular Biology*, vol. 20, no. 5, pp. 976–983, 1999.

[150] Nyholt, D.R.: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, vol. 74, no. 4, pp. 765–769, 2004.

[151] Perneger, T.V.: What's wrong with Bonferroni adjustments. *British Medical Journal*, vol. 316, no. 7139, p. 1236, 1998.

[152] Rothman, K.J.: Six persistent research misconceptions. *Journal of General Internal Medicine*, pp. 1–5, 2014.

[153] Emigh, T.H.: A comparison of tests for Hardy-Weinberg equilibrium. *Biometrics*, pp. 627–642, 1980.

[154] Warnes, G., with contributions from Gregor Gorjanc, Leisch, F. and Man., M.: *genetics: Population Genetics*, 2012. R package version 1.3.8.
Available at: `http://CRAN.R-project.org/package=genetics`

[155] Ferguson, J.S., Martin, J.L., Azad, A.K., McCarthy, T.R., Kang, P.B., Voelker, D.R., Crouch, E.C. and Schlesinger, L.S.: Surfactant protein d increases fusion of *Mycobacterium tuberculosis*-containing phagosomes with lysosomes in human macrophages. *Infection and Immunity*, vol. 74, no. 12, p. 7005–7009, 2006.

[156] Wright, S.D., Ramos, R.A., Tobias, P.S., Ulevitch, R.J. and Mathison, J.C.: CD14, a receptor for complexes of lipopolysaccharide (LPS) and LPS binding protein. *Science*, vol. 249, no. 4975, pp. 1431–1433, 1990.

[157] Pugin, J., Heumann, D., Tomasz, A., Kravchenko, V.V., Akamatsu, Y., Nishijima, M., Glauser, M.P., Tobias, P.S. and Ulevitch, R.J.: CD14 is a pattern recognition receptor. *Immunity*, vol. 1, no. 6, pp. 509–516, 1994.

[158] Van der Spuy, G.D., Kremer, K., Ndabambi, S.L., Beyers, N., Dunbar, R., Marais, B.J., van Helden, P.D. and Warren, R.M.: Changing *Mycobacterium tuberculosis* population highlights clade-specific pathogenic characteristics. *Tuberculosis*, vol. 89, no. 2, pp. 120–125, 2009.

[159] Salie, M., van der Merwe, L., Möller, M., Daya, M., van der Spuy, G.D., van Helden, P.D., Martin, M.P., Gao, X., Warren, R.M. and Carrington, M.: Associations between human leukocyte antigen class i variants and the *Mycobacterium* tuberculosis subtypes causing disease. *Journal of Infectious Diseases*, vol. 209, no. 2, pp. 216–223, 2014.

[160] Rossouw, M., Nel, H.J., Cooke, G.S., van Helden, P.D. and Hoal, E.G.: Association between tuberculosis and a polymorphic NFκB binding site in the interferon γ gene. *The Lancet*, vol. 361, no. 9372, pp. 1871–1872, 2003.

[161] Babb, C., Keet, E.H., Helden, P.D.v. and Hoal, E.G.: SP110 polymorphisms are not associated with pulmonary tuberculosis in a South African population. *Human Genetics*, vol. 121, no. 3-4, pp. 521–522, May 2007.

[162] Möller, M.: *Human genetic susceptibility to tuberculosis: the investigation of candidate genes influencing interferon gamma levels and other candidate genes affecting immunological pathways*. Ph.D. thesis, Stellenbosch University, 2007. `http://hdl.handle.net/10019.1/1264`.

[163] Möller, M., Kwiatkowski, R., Nebel, A., van Helden, P.D., Hoal, E.G. and Schreiber, S.: Allelic variation in BTNL2 and susceptibility to tuberculosis in a South African population. *Microbes and Infection*, vol. 9, no. 4, pp. 522–528, 2007.

[164] Möller, M., Nebel, A., Kwiatkowski, R., van Helden, P.D., Hoal, E.G. and Schreiber, S.: Host susceptibility to tuberculosis: CARD15 polymorphisms in a South African population. *Molecular and Cellular Probes*, vol. 21, no. 2, pp. 148–151, 2007.

[165] Babb, C.: *Identification of candidate genes and testing for association with tuberculosis in humans*. Ph.D. thesis, Stellenbosch University, 2007. `http://hdl.handle.net/10019.1/21524`.

[166] De Wit, E.: *Analysis of host determining factors in susceptibility to tuberculosis in the South African coloured population*. Ph.D. thesis, Stellenbosch University, 2009. `http://hdl.handle.net/10019.1/4584`.

[167] Salie, M.: *Investigating candidate genes identified by genome-wide studies of granulomatous diseases in susceptibility to tuberculosis: ANXA11 and the CADM family*. Master's thesis, Stellenbosch University, 2010. `http://hdl.handle.net/10019.1/5472`.

[168] Lucas, L.: *Toll-like receptor genes and their pathway: role in susceptibility to pulmonary tuberculosis in a South African population*. Master's thesis, Stellenbosch University, 2012. `http://hdl.handle.net/10019.1/20390`.

[169] Wagman, C.: *Genetic studies on susceptibility to pulmonary tuberculosis mediated by MARCO, SP-D and CD14: molecules affecting uptake of* Mycobacterium tuberculosis *into macrophages*. Master's thesis, Stellenbosch University, 2012. `http://hdl.handle.net/10019.1/20409`.

[170] Bruiners, N.: *Investigating the Human-M. tuberculosis interactome to identify the host targets of ESAT-6 and other mycobacterial antigens*. Ph.D. thesis, Stellenbosch University, 2013. `http://hdl.handle.net/10019.1/71977`.

[171] Eichler, E.E., Flint, J., Gibson, G., Kong, A., Leal, S.M., Moore, J.H. and Nadeau, J.H.: Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, vol. 11, no. 6, pp. 446–450, 2010.

[172] Frazer, K.A., Murray, S.S., Schork, N.J. and Topol, E.J.: Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, vol. 10, no. 4, pp. 241–251, 2009.

[173] Wei, Z., Wang, K., Qu, H.-Q., Zhang, H., Bradfield, J., Kim, C., Frackleton, E., Hou, C., Glessner, J.T. and Chiavacci, R.: From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genetics*, vol. 5, no. 10, p. e1000678, 2009.

[174] Stein, C.M.: Genetic epidemiology of tuberculosis susceptibility: impact of study design. *PLoS Pathogens*, vol. 7, no. 1, p. e1001189, 2011.

[175] Velez, D.R., Hulme, W.F., Myers, J.L., Stryjewski, M.E., Abbate, E., Estevan, R., Patillo, S.G., Gilbert, J.R., Hamilton, C.D. and Scott, W.K.: Association of SLC11A1 with tuberculosis interactions with NOS2A and TLR2 in African-Americans and Caucasians. *The International Journal of Tuberculosis and Lung Disease*, vol. 13, no. 9, p. 1068, 2009.

[176] Flores-Villanueva, P.O., Ruiz-Morales, J.A., Song, C.-H., Flores, L.M., Jo, E.-K., Montaño, M., Barnes, P.F., Selman, M. and Granados, J.: A functional promoter polymorphism in monocyte chemoattractant protein-1 is associated with increased susceptibility to pulmonary tuberculosis. *The Journal of Experimental Medicine*, vol. 202, no. 12, pp. 1649–1658, 2005.

[177] Zuk, O., Hechter, E., Sunyaev, S.R. and Lander, E.S.: The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences*, vol. 109, no. 4, pp. 1193–1198, 2012.

[178] Cordell, H.J.: Detecting gene-gene interactions that underlie human diseases. *Nature Reviews Genetics*, vol. 10, no. 6, pp. 392–404, 2009.

[179] Greene, C.S., Penrod, N.M., Williams, S.M. and Moore, J.H.: Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS ONE*, vol. 4, no. 6, p. e5639, 2009.

[180] Phillips, P.C.: Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, vol. 9, no. 11, p. 855, 2008.

[181] Cordell, H.J.: Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, vol. 11, no. 20, pp. 2463–2468, 2002.

[182] Beltrao, P., Cagney, G. and Krogan, N.J.: Quantitative genetic interactions reveal biological modularity. *Cell*, vol. 141, no. 5, pp. 739–745, 2010.

[183] Brem, R.B., Storey, J.D., Whittle, J. and Kruglyak, L.: Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature*, vol. 436, no. 7051, pp. 701–703, 2005.

[184] Chou, H.-H., Chiu, H.-C., Delaney, N.F., Segrè, D. and Marx, C.J.: Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science*, vol. 332, no. 6034, pp. 1190–1192, 2011.

[185] Dixon, S.J., Costanzo, M., Baryshnikova, A., Andrews, B. and Boone, C.: Systematic mapping of genetic interaction networks. *Annual Review of Genetics*, vol. 3, pp. 601–625, 2009.

[186] Ferguson, E.L. and Horvitz, H.R.: The multivulva phenotype of certain Caenorhabditis elegans mutants results from defects in two functionally redundant pathways. *Genetics*, vol. 123, no. 1, pp. 109–121, 1989.

[187] Huang, W., Richards, S., Carbone, M.A., Zhu, D., Anholt, R.R., Ayroles, J.F., Duncan, L., Jordan, K.W., Lawrence, F., Magwire, M.M., B., W.C., Kerstin, B., Yi, H., Mehwish, J., Joy, J., N., J.S., Donna, M., Fiona, O., Lora, P., Yuan-Qing, W., Yiqing, Z., Xiaoyan, Z., A., S.E., A., G.R. and C., M.T.F.: Epistasis dominates the genetic architecture of Drosophila quantitative traits. *Proceedings of the National Academy of Sciences*, vol. 109, no. 39, pp. 15553–15559, 2012.

[188] Khan, A.I., Dinh, D.M., Schneider, D., Lenski, R.E. and Cooper, T.F.: Negative epistasis between beneficial mutations in an evolving bacterial population. *Science*, vol. 332, no. 6034, pp. 1193–1196, 2011.

[189] Lindén, R.O., Eronen, V.-P. and Aittokallio, T.: Quantitative maps of genetic interactions in yeast-comparative evaluation and integrative analysis. *BMC Systems Biology*, vol. 5, no. 1, p. 45, 2011.

[190] Okser, S., Pahikkala, T. and Aittokallio, T.: Genetic variants and their interactions in disease risk prediction-machine learning and network perspectives. *BioData Mining*, vol. 6, p. 5, 2013.

[191] Shao, H., Burrage, L.C., Sinasac, D.S., Hill, A.E., Ernest, S.R., O'Brien, W., Courtland, H.-W., Jepsen, K.J., Kirby, A., Kulbokas, E.J., Daly, M.J., Broman, K.W., Lander, E.S. and Nadeau, J.H.: Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proceedings of the National Academy of Sciences*, vol. 105, no. 50, pp. 19910–19914, 2008.

[192] Collins, R.L., Hu, T., Wejse, C., Sirugo, G., Williams, S.M. and Moore, J.H.: Multi-factor dimensionality reduction reveals a three-locus epistatic interaction associated with susceptibility to pulmonary tuberculosis. *BioData Mining*, vol. 6, no. 4, 2013.

[193] Edwards, D.R.V., Tacconelli, A., Wejse, C., Hill, P.C., Morris, G.A., Edwards, T.L., Gilbert, J.R., Myers, J.L., Park, Y.S., Stryjewski, M.E., Abbate, E., Estevan, R., Rabna, P., Novelli, G., Hamilton, C.D., Adegbola, R., Østergaard, L., Williams, S.M., Scott, W.K. and Sirugo, G.: MCP1 SNPs and pulmonary tuberculosis in cohorts from West Africa, the USA and Argentina: lack of association or epistasis with IL12B polymorphisms. *PloS ONE*, vol. 7, no. 2, p. e32275, 2012.

[194] Motsinger-Reif, A.A., Antas, P.R., Oki, N.O., Levy, S., Holland, S.M. and Sterling, T.R.: Polymorphisms in IL-1β, vitamin D receptor Fok1, and toll-like receptor 2 are associated with extrapulmonary tuberculosis. *BMC Medical Genetics*, vol. 11, no. 1, p. 37, 2010.

[195] Olesen, R., Wejse, C., Velez, D.R., Bisseye, C., Sodemann, M., Aaby, P., Rabna, P., Worwui, A., Chapman, H., Diatta, M., Hill, P.C., Østergaard, L., Williams, S.M. and Sirug, G.: DC-SIGN (CD209), pentraxin 3 and vitamin D receptor gene variants associate with pulmonary tuberculosis risk in West Africans. *Genes and Immunity*, vol. 8, no. 6, pp. 456–467, 2007.

[196] Ravikumar, M., Dheenadhayalan, V., Rajaram, K., Shanmuga Lakshmi, S., Paul Kumaran, P., Paramasivan, C.N., Balakrishnan, K. and Pitchappan, R.M.: Associations of HLA-DRB1, DQB1 and DPB1 alleles with pulmonary tuberculosis in south India. *Tubercle and Lung Disease*, vol. 79, no. 5, pp. 309–317, 1999.

[197] White, M.J., Tacconelli, A., Chen, J.S., Wejse, C., Hill, P.C., Gomes, V.F., Velez-Edwards, D.R., Østergaard, L.J., Hu, T., Moore, J.H., Novelli, G., Scott, W.K., Williams, S.M. and Sirugo, G.: Epiregulin (EREG) and human V-ATPase (TCIRG1): genetic variation, ethnicity and pulmonary tuberculosis susceptibility in Guinea-Bissau and The Gambia. *Genes and Immunity*, 2014.

[198] Brinza, D., Schultz, M., Tesler, G. and Bafna, V.: RAPID detection of gene-gene interactions in genome-wide association studies. *Bioinformatics*, vol. 26, no. 22, pp. 2856–2862, November 2010.

[199] Chen, G., Yuan, A., Zhou, J., Bentley, A.R., Adeyemo, A. and Rotimi, C.N.: Simple F test reveals gene-gene interactions in case-control studies. *Bioinformatics and Biology Insights*, vol. 6, p. 169, 2012.

[200] Hu, T., Sinnott-Armstrong, N.A., Kiralis, J.W., Andrew, A.S., Karagas, M.R. and Moore, J.H.: Characterizing genetic interactions in human disease association studies using statistical epistasis networks. *BMC Bioinformatics*, vol. 12, no. 1, p. 364, 2011.

[201] Kam-Thong, T., Czamara, D., Tsuda, K., Borgwardt, K., Lewis, C.M., Erhardt-Lehmann, A., Hemmer, B., Rieckmann, P., Daake, M., Weber, F., Wolf, C., Ziegler, A., Pütz, B., Holsboer, F., Schölkopf, B. and Müller-Myhsok, B.: EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. *European Journal of Human Genetics*, vol. 19, no. 4, pp. 465–471, April 2011.

[202] Liu, Y., Xu, H., Chen, S., Chen, X., Zhang, Z., Zhu, Z., Qin, X., Hu, L., Zhu, J. and Zhao, G.-P.: Genome-wide interaction-based association analysis identified multiple new susceptibility loci for common diseases. *PLoS Genetics*, vol. 7, no. 3, p. e1001338, 2011.

[203] McKinney, B.A., Crowe Jr, J.E., Guo, J. and Tian, D.: Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS Genetics*, vol. 5, no. 3, p. e1000432, 2009.

[204] Moore, J.H. and Williams, S.M.: New strategies for identifying gene-gene interactions in hypertension. *Annals of Medicine*, vol. 34, no. 2, pp. 88–95, 2002.

[205] Moore, J.H. and White, B.C.: Tuning ReliefF for genome-wide genetic analysis. In: *Evolutionary computation, machine learning and data mining in bioinformatics*, pp. 166–175. Springer, 2007.

[206] Motsinger-Reif, A., Dudek, S., Hahn, L. and Ritchie, M.: Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genetic Epidemiology*, vol. 32, no. 4, pp. 325–340, 2008.

[207] Robnik-Šikonja, M. and Kononenko, I.: Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, vol. 53, no. 1-2, pp. 23–69, 2003.

[208] Turner, S.D., Dudek, S.M. and Ritchie, M.D.: ATHENA: A knowledge-based hybrid backpropagation-grammatical evolution neural network algorithm for discovering epistasis among quantitative trait loci. *BioData Mining*, vol. 3, no. 1, p. 5, 2010.

[209] Ueki, M. and Cordell, H.J.: Improved statistics for genome-wide interaction analysis. *PLoS Genetics*, vol. 8, no. 4, p. e1002625, April 2012.

[210] Wellek, S. and Ziegler, A.: A genotype-based approach to assessing the association between single nucleotide polymorphisms. *Human Heredity*, vol. 67, no. 2, pp. 128–139, 2008.

[211] Wu, X., Dong, H., Luo, L., Zhu, Y., Peng, G., Reveille, J.D. and Xiong, M.: A novel statistic for genome-wide interaction analysis. *PLoS Genetics*, vol. 6, no. 9, p. e1001131, 2010.

[212] Zhang, Y. and Liu, J.S.: Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, vol. 39, no. 9, pp. 1167–1173, 2007.

[213] Zhang, X., Huang, S., Zou, F. and Wang, W.: Tools for efficient epistasis detection in genome-wide association study. *Source Code for Biology and Medicine*, vol. 6, no. 1, p. 1, 2011.

[214] Marchini, J., Donnelly, P. and Cardon, L.R.: Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genetics*, vol. 37, no. 4, pp. 413–417, 2005.

[215] Bush, W.S., Dudek, S.M. and Ritchie, M.D.: Biofilter: A knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pacific Symposium on Biocomputing*, pp. 368–379, 2009.

[216] Emily, M., Mailund, T., Hein, J., Schauser, L. and Schierup, M.: Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, vol. 17, no. 10, pp. 1231–1240, 2009.

[217] Ma, L., Brautbar, A., Boerwinkle, E., Sing, C.F., Clark, A.G. and Keinan, A.: Knowledge-driven analysis identifies a gene-gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations. *PLoS Genetics*, vol. 8, no. 5, p. e1002714, 2012.

[218] Pattin, K.A. and Moore, J.H.: Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. *Human Genetics*, vol. 124, no. 1, pp. 19–29, August 2008.

[219] Gallant, C.J., Cobat, A., Simkin, L., Black, G.F., Stanley, K., Hughes, J., Doherty, T.M., Hanekom, W.A., Eley, B., Beyers, N., Jaïs, J.P., van Helden, P., Abel, L., Alcaïs, A., Hoal, E.G. and Schurr, E.: Impact of age and sex on mycobacterial immunity in an area of high tuberculosis incidence. *The International Journal of Tuberculosis and Lung Disease*, vol. 14, no. 8, pp. 952–959, 2010.

[220] Thye, T., Owusu-Dabo, E., Vannberg, F.O., van Crevel, R., Curtis, J., Sahiratmadja, E., Balabanova, Y., Ehmen, C., Muntau, B. and Ruge, G.: Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nature Genetics*, vol. 44, no. 3, pp. 257–259, 2012.

[221] Fox, J.: Effect displays in R for generalised linear models. *Journal of Statistical Software*, vol. 8, no. 15, pp. 1–27, 2003.

[222] Gao, X., Becker, L., Becker, D., Starmer, J. and Province, M.: Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic Epidemiology*, vol. 34, no. 1, pp. 100–105, 2010.

[223] Bůžková, P., Lumley, T. and Rice, K.: Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Annals of Human Genetics*, vol. 75, no. 1, pp. 36–45, 2011.

[224] Li, J. and Ji, L.: Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, vol. 95, no. 3, pp. 221–227, September 2005.

[225] Kerr, K.: Comments on the analysis of unbalanced microarray data. *Bioinformatics*, vol. 25, no. 16, p. 2035, 2009.

[226] Shriner, D., Adeyemo, A., Gerry, N.P., Herbert, A., Chen, G., Doumatey, A., Huang, H., Zhou, J., Christman, M.F. and Rotimi, C.N.: Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. *PLoS ONE*, vol. 4, no. 12, p. e8398, 2009.

[227] Ramos, E., Chen, G., Shriner, D., Doumatey, A., Gerry, N.P., Herbert, A., Huang, H., Zhou, J., Christman, M.F., Adeyemo, A. and Rotimi, C.: Replication of genome-wide association studies (GWAS) loci for fasting plasma glucose in African-Americans. *Diabetologia*, vol. 54, no. 4, pp. 783–788, 2011.

[228] Sinnwell, J. and Schaid, D.: *haplo.stats: Statistical Analysis of Haplotypes with Traits and Covariates when Linkage Phase is Ambiguous*, 2013. R package version 1.6.8.
Available at: `http://CRAN.R-project.org/package=haplo.stats`

[229] Wickham, H.: *ggplot2: elegant graphics for data analysis*. Springer, New York, 2009.
Available at: `http://had.co.nz/ggplot2/book`

[230] Kanakry, C.G., Li, Z., Nakai, Y., Sei, Y. and Weinberger, D.R.: Neuregulin-1 regulates cell adhesion via an ErbB2/phosphoinositide-3 kinase/akt-dependent pathway: potential implications for schizophrenia and cancer. *PLoS ONE*, vol. 2, no. 12, p. e1369, 2007.

[231] Marballi, K., Quinones, M.P., Jimenez, F., Escamilla, M.A., Raventós, H., Soto-Bernardini, M.C., Ahuja, S.S. and Walss-Bass, C.: In vivo and in vitro genetic evidence of involvement of neuregulin 1 in immune system dysregulation. *Journal of Molecular Medicine*, vol. 88, no. 11, pp. 1133–1141, 2010.

[232] Benzel, I., Bansal, A., Browning, B.L., Galwey, N.W., Maycox, P.R., McGinnis, R., Smart, D., St Clair, D., Yates, P. and Purvis, I.: Interactions among genes in the ErbB-neuregulin signalling network are associated with increased susceptibility to schizophrenia. *Behavioral and Brain Functions*, vol. 3, p. 31, 2007.

[233] Hayes, M.G., Pluzhnikov, A., Miyake, K., Sun, Y., Ng, M.C., Roe, C.A., Below, J.E., Nicolae, R.I., Konkashbaev, A., Bell, G.I., Cox, N.J. and Hanis, C.L.: Identification of type 2 diabetes genes in Mexican Americans through genome-wide association studies. *Diabetes*, vol. 56, no. 12, pp. 3033–3044, 2007.

[234] Jeon, C.Y. and Murray, M.B.: Diabetes mellitus increases the risk of active tuberculosis: a systematic review of 13 observational studies. *PLoS Medicine*, vol. 5, no. 7, p. e152, 2008.

[235] Khader, S.A. and Cooper, A.M.: IL-23 and IL-17 in tuberculosis. *Cytokine*, vol. 41, no. 2, pp. 79–83, 2008.

[236] Songane, M., Kleinnijenhuis, J., Alisjahbana, B., Sahiratmadja, E., Parwati, I., Oosting, M., Plantinga, T.S., Joosten, L.A., Netea, M.G., Ottenhoff, T.H., Van de Vosse, E. and van Crevel, R.: Polymorphisms in autophagy genes and susceptibility to tuberculosis. *PloS ONE*, vol. 7, no. 8, p. e41618, 2012.

[237] Wenning, A.S., Neblung, K., Straub, B., Wolfs, M.-J., Sappok, A., Hoth, M. and Schwarz, E.C.: TRP expression pattern and the functional importance of TRPC3 in primary human T-cells. *Biochimica et Biophysica Acta (BBA)-Molecular Cell Research*, vol. 1813, no. 3, pp. 412–423, 2011.

[238] McLeod, I.X. and He, Y.-W.: Editorial: TRPV1: how thymocytes sense stress and respond with autophagy. *Journal of Leukocyte Biology*, vol. 92, no. 3, pp. 409–411, 2012.

[239] Fukami, T., Satoh, H., Williams, Y.N., Masuda, M., Fukuhara, H., Maruyama, T., Yageta, M., Kuramochi, M., Takamoto, S. and Murakami, Y.: Isolation of the mouse Tsll1 and Tsll2 genes, orthologues of the human TSLC1-like genes 1 and 2 (TSLL1 and TSLL2). *Gene*, vol. 323, pp. 11–18, 2003.

[240] Fukuhara, H., Kuramochi, M., Nobukuni, T., Fukami, T., Saino, M., Maruyama, T., Nomura, S., Sekiya, T. and Murakami, Y.: Isolation of the TSLL1 and TSLL2 genes, members of the tumor suppressor TSLC1 gene family encoding transmembrane proteins. *Oncogene*, vol. 20, no. 38, pp. 5401–5407, 2001.

[241] Galibert, L., Diemer, G.S., Liu, Z., Johnson, R.S., Smith, J.L., Walzer, T., Comeau, M.R., Rauch, C.T., Wolfson, M.F., Sorensen, R.A., Van der Vuurst de Vries, A., Branstetter, D., Koelling, R., Scholler, J., Fanslow, W., Baum, P., Derry, J. and Yan, W.: Nectin-like protein 2 defines a subset of T-cell zone dendritic cells and is a ligand for class-I-restricted T-cell-associated molecule. *Journal of Biological Chemistry*, vol. 280, no. 23, pp. 21955–21964, 2005.

[242] Luce, M.J. and Burrows, P.D.: The neuronal EGF-related genes NELL1 and NELL2 are expressed in hemopoietic cells and developmentally regulated in the B lineage. *Gene*, vol. 231, no. 1, pp. 121–126, 1999.

[243] Franke, A., Hampe, J., Rosenstiel, P., Becker, C., Wagner, F., Häsler, R., Little, R.D., Huse, K., Ruether, A., Balschun, T., Wittig, M., ElSharaw, A., May, G., Albrech, M., Prescott, N.J., Onnie, C.M., Fournie, H., Keit, T., Radelo, U., Platze, M., Mathew, C.G., Stol, M., Krawcza, M., Nürnber, P. and Schreibe, S.: Systematic association mapping identifies NELL1 as a novel IBD disease gene. *PLoS ONE*, vol. 2, no. 8, p. e691, 2007.

[244] Uhlin-Hansen, L., Eskeland, T. and Kolset, S.O.: Modulation of the expression of chondroitin sulfate proteoglycan in stimulated human monocytes. *Journal of Biological Chemistry*, vol. 264, no. 25, pp. 14916–14922, 1989.

[245] Fallahi-Sichani, M., Kirschner, D.E. and Linderman, J.J.: NF-κB signaling dynamics play a key role in infection control in tuberculosis. *Frontiers in Physiology*, vol. 3, 2012.

[246] Kingma, P.S., Zhang, L., Ikegami, M., Hartshorn, K., McCormack, F.X. and Whitsett, J.A.: Correction of pulmonary abnormalities in Sftpd-/-mice requires the collagenous domain of surfactant protein D. *Journal of Biological Chemistry*, vol. 281, no. 34, pp. 24496–24505, 2006.

[247] Kanazawa, N., Okafuji, I., Kambe, N., Nishikomori, R., Nakata-Hizume, M., Nagai, S., Fuji, A., Yuasa, T., Manki, A., Sakurai, Y., Nakajima, M., Kobayashi, H., Fujiwara, I., Tsutsumi, H., Utani, A., Nishigori, C., Heike, T., Nakahata, T. and Miyachi, Y.: Early-onset sarcoidosis and CARD15 mutations with constitutive nuclear factor-κB activation: common genetic etiology with blau syndrome. *Blood*, vol. 105, no. 3, pp. 1195–1197, 2005.

[248] D'Cunha, J., Knight, E., Haas, A.L., Truitt, R.L. and Borden, E.C.: Immunoregulatory properties of ISG15, an interferon-induced cytokine. *Proceedings of the National Academy of Sciences*, vol. 93, no. 1, pp. 211–215, 1996.

[249] Lin, Y., Jamison, S. and Lin, W.: Interferon-γ activates nuclear factor-κ B in oligodendrocytes through a process mediated by the unfolded protein response. *PloS ONE*, vol. 7, no. 5, p. e36408, 2012.

[250] Davila, S., Hibberd, M.L., Dass, R.H., Wong, H.E., Sahiratmadja, E., Bonnard, C., Alisjahbana, B., Szeszko, J.S., Balabanova, Y. and Drobniewski, F.: Genetic association and expression studies indicate a role of toll-like receptor 8 in pulmonary tuberculosis. *PLoS Genetics*, vol. 4, no. 10, p. e1000218, 2008.

[251] Rosenzweig, S.D. and Holland, S.M.: Defects in the interferon-γ and interleukin-12 pathways. *Immunological Reviews*, vol. 203, no. 1, pp. 38–47, 2005.

[252] Kuenzel, S., Till, A., Winkler, M., Häsler, R., Lipinski, S., Jung, S., Grötzinger, J., Fickenscher, H., Schreiber, S. and Rosenstiel, P.: The nucleotide-binding oligomerization domain-like receptor NLRC5 is involved in IFN-dependent antiviral immune responses. *The Journal of Immunology*, vol. 184, no. 4, pp. 1990–2000, 2010.

[253] Seldin, M., Pasaniuc, B. and Price, A.: New approaches to disease mapping in admixed populations. *Nature Reviews Genetics*, 2011.

[254] Bhangale, T.R., Rieder, M.J. and Nickerson, D.A.: Estimating coverage and power for genetic association studies using near-complete variation data. *Nature Genetics*, vol. 40, no. 7, pp. 841–843, 2008.

[255] Gravel, S.: Population genetics models of local ancestry. *Genetics*, vol. 191, no. 2, p. 607–619, 2012.

[256] Salerno, D.M., Tront, J.S., Hoffman, B. and Liebermann, D.A.: Gadd45a and Gadd45b modulate innate immune functions of granulocytes and macrophages by differential regulation of p38 and JNK signaling. *Journal of Cellular Physiology*, vol. 227, no. 11, pp. 3613–3620, November 2012.

[257] O'Kane, C.M., Elkington, P.T. and Friedland, J.S.: Monocyte-dependent oncostatin M and TNF-alpha synergize to stimulate unopposed matrix metalloproteinase-1/3 secretion from human lung fibroblasts in tuberculosis. *European Journal of Immunology*, vol. 38, no. 5, pp. 1321–1330, May 2008.

[258] Bellamy, R., Beyers, N., McAdam, K., Ruwende, C., Gie, R., Samaai, P., Bester, D., Meyer, M., Corrah, T., Collin, M., Camidge, D., Wilkinson, D., Hoal-Van Helden, E., Whittle, H., Amos, W., van Helden, P. and Hill, A.: Genetic susceptibility to tuberculosis in Africans: a genome-wide scan. *Proceedings of the National Academy of Sciences*, vol. 97, no. 14, pp. 8005–8009, July 2000.

[259] Cervino, A., Lakiss, S., Sow, O., Bellamy, R., Beyers, N., Hoal-Van Helden, E., van Helden, P., McAdam, K. and Hill, A.: Fine mapping of a putative tuberculosis-susceptibility locus on chromosome 15q11-13 in African families. *Human Molecular Genetics*, vol. 11, no. 14, pp. 1599–1603, July 2002.

[260] Zhu, Y., Yao, S., Iliopoulou, B.P., Han, X., Augustine, M.M., Xu, H., Phennicie, R.T., Flies, S.J., Broadwater, M. and Ruff, W.: B7-H5 costimulates human T cells via CD28H. *Nature Communications*, vol. 4, 2013.

[261] Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Zaitlen, N., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G. and Avila, P.C.: Analysis of Latino populations from GALA and MEC studies reveals genomic loci with biased local ancestry estimation. *Bioinformatics*, vol. 29, no. 11, pp. 1407–1415, 2013.

[262] Delaneau, O., Zagury, J.-F. and Marchini, J.: Improved whole-chromosome phasing for disease and population genetic studies. *Nature methods*, vol. 10, no. 1, pp. 5–6, 2013.

[263] Baran, Y., Pasaniuc, B., Sankararaman, S., Torgerson, D.G., Gignoux, C., Eng, C., Rodriguez-Cintron, W., Chapela, R., Ford, J.G. and Avila, P.C.: Fast and accurate inference of local ancestry in Latino populations. *Bioinformatics*, vol. 28, no. 10, pp. 1359–1367, 2012.

[264] Maples, B.K., Gravel, S., Kenny, E.E. and Bustamante, C.D.: RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *The American Journal of Human Genetics*, vol. 93, no. 2, pp. 278–288, 2013.

[265] Carr, D., ported by Nicholas Lewin-Koh, Maechler, M. and contains copies of lattice function written by Deepayan Sarkar: *hexbin: Hexagonal Binning Routines*, 2014. R package version 1.27.0.
Available at: `http://CRAN.R-project.org/package=hexbin`

[266] Goudet, J.: *hierfstat: Estimation and tests of hierarchical F-statistics*, 2013. R package version 0.04-10.
Available at: `http://CRAN.R-project.org/package=hierfstat`

[267] Price, A.L., Tandon, A., Patterson, N., Barnes, K.C., Rafaels, N., Ruczinski, I., Beaty, T.H., Mathias, R., Reich, D. and Myers, S.: Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genetics*, vol. 5, no. 6, p. e1000519, 2009.

[268] Hochberg, Y.: A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, vol. 75, no. 4, pp. 800–802, 1988.

[269] Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[270] Storey, J.D. and Tibshirani, R.: Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, vol. 100, no. 16, pp. 9440–9445, 2003.

[271] Ioannidis, J.P.: Non-replication and inconsistency in the genome-wide association setting. *Human Heredity*, vol. 64, no. 4, pp. 203–213, 2007.

[272] Van Helden, P.: Data-driven hypotheses. *EMBO Reports*, vol. 14, no. 2, pp. 104–104, 2013.

[273] Verver, S., Warren, R.M., Beyers, N., Richardson, M., van der Spuy, G.D., Borgdorff, M.W., Enarson, D.A., Behr, M.A. and van Helden, P.D.: Rate of reinfection tuberculosis after successful treatment is higher than rate of new tuberculosis. *American Journal of Respiratory and Critical Care Medicine*, vol. 171, no. 12, pp. 1430–1435, 2005.