Investigating southern African genetic diversity and its role in TB susceptibility

Caitlin Uren

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

Supervisor: Dr. Marlo Möller

Co-supervisors: Dr. Brenna Henn, Prof Eileen Hoal

December 2017

Declaration:

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

This dissertation includes 3 original papers published in a peer-reviewed journals and 1 unpublished publication. The development and writing of the papers (published and unpublished) were the principal responsibility of myself and, for each of the cases where this is not the case, a declaration is included in the dissertation indicating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.

Date: December 2017

Abstract

Recent genetic studies have established that the KhoeSan populations of southern Africa are the earliest known indigenous inhabitants of the region and distinct from all other African populations. Owing to the region's unique history, population structure in southern Africa reflects both the underlying KhoeSan genetic diversity as well as differential recent admixture. This population structure has a wide range of biomedical and sociocultural implications such as changes in disease risk profiles: there is a known correlation between ancestry and tuberculosis (TB) susceptibility.

Research presented in this thesis consolidates information from various population genetic studies that characterized admixture patterns in southern Africa with the aim to improve the understanding of differences in adverse disease phenotypes observed among populations. Further to previous studies, genome-wide polymorphism data from more than 20 southern African populations were analysed to investigate the fine-scale population structure in the area. The analyses revealed fine-scale population structure in and around the Kalahari Desert, which does not always correspond to linguistic or subsistence categories, but rather reflects the role of ecogeographic boundaries. In addition, we showed that the Khoe adopted their pastoralism through a process of largely cultural diffusion rather than demic diffusion as previously thought. The proportion and origin of KhoeSan genetic ancestry in southern African populations is of particular relevance to disease, because the KhoeSan exhibit greater variation in genetic diversity than other African populations, including unusual variation in genes with demonstrable immune function.

Utilizing data from several TB genome-wide association studies (GWAS), a bioinformatics pipeline was employed to detect regulatory polymorphisms in linkage disequilibrium with variants previously implicated in TB susceptibility. A total of 133 predicted regulatory variants were found. Association analyses were performed in TB cases and healthy controls and yielded six intronic functionally relevant variants. The post-GWAS approach, which included ancestry as a confounder, demonstrated the feasibility of combining multiple TB GWAS datasets with linkage information to identify regulatory variants associated with TB susceptibility.

In addition to classical association studies, selection scans have the ability to identify genomic regions associated with a phenotype. Signals of natural selection in southern African populations was studied using high-coverage exome sequence data. Selection signals were identified in genes associated with immune response to foreign pathogens introduced from the 12th century onwards. In addition, signals of selection were identified in pathways associated with focal adhesion and ECM receptor interaction. It is clear that there are distinct immune-related signals of positive selection present in southern African populations.

This research not only provided insight into the genetic basis and biology of human TB susceptibility, but also harnessed the unique ancestry present in southern African populations. The addition of population genetics information was shown to greatly shape and improve our investigations of TB susceptibility and may also apply to other phenotypes unique to southern Africa. Since the era of personalised medicine is imminent, more investigations of understudied southern African populations most severely affected by TB are required.

Opsomming

Onlangse genetiese studies het vasgestel dat die KhoeSan bevolking van suider-Afrika die vroegste bekende inheemse inwoners van die streek was, en dat hul kenmerkend van ander Afrika-bevolkings verskil. Op grond van die streek se unieke geskiedenis, weerspieël die bevolkingstruktuur van suider-Afrika beide die onderliggende genetiese diversiteit van die KhoeSan, asook differensiële onlangse vermenging. Hierdie bevolkingstruktuur het 'n verskeidenheid van biomediese asook sosiokulturele implikasies, soos verandering in siekterisikoprofiele: dit bekend is dat daar 'n assosiasie is tussen toenemende KhoeSan afkoms en tuberkulose (TB).

Navorsing uiteengesit in hierdie proefskrif konsolideer inligting uit verskeie bevolkingsgenetika studies, waarin vermengingspatrone in suider-Afrika omgeskryf is met die doel om verskille in siektefenotipes beter te begryp. Om die fynskaalse bevolkingstruktuur in the area te ondersoek, is genoomwye polimorfisme-data van meer as 20 bevolkings van suider-Afrika ontleed. Die analise het kleine aspekte van die bevolkingstruktuur in en rondom die Kalahari-woestyn ontbloot, wat nie altyd ooreengestem het met taalkundige en lewensbestaans kategorieë, soos voorheen voorgestel nie. Dit het eerder die rol van geografiese versperrings en die ekologie van die groter Kalahari-kom weerspieël. Ons toon aan dat die Khoe hulle pastorale bestaanstrategie eerder deur 'n proses van grotendeels kulturele diffusie aangeneem het en nie, soos wat voorheen aanvaar is, as gevolg van vermenging, vervanging of verplasing nie. Die genetiese bydrae van die KhoeSan tot die bevolkings van suider-Afrika is veral belangrik in die konteks van siekte, aangesien die KhoeSan meer variasie in genetiese diversiteit het as ander Afrika bevolkings. Dit sluit ongewone variasie in gene met bewysde immuunfunksies in.

Deur gebruik te maak van data van verskeie genoomwye assosiasiestudies (GWAS) oor TB, is 'n bioinformatikapyplyn aangestel om regulatoriese polimorfismes in koppelingsdisekwilibirum met veranderlikes voorheen aangedui in TB-kwesbaarheid op te spoor. 'n Bevolkingsgebaseerde gevallekontrole-assosiasiestudie van 133 voorspelde regulatoriese variante is in TB-gevalle asook gesonde kontroles uitgevoer. Assosiasies met ses introniese veranderlikes is waargeneem. Die post-GWAS-benadering, wat afkoms as 'n invloed ingesluit het, demonstreer die uitvoerbaarheid daarvan om meervoudige TB GWAS-datastelsels met koppelingsinligting te kombineer om regulatoriese veranderlikes geassosieerd met hierdie aansteeklike siekte te identifiseer.

Bykomend tot klassieke assosiasiestudies, kan skanderings van natuurlike seleksie areas in die genoom identifiseer wat met 'n fenotipe geassosieer is. Aanduidings van natuurlike seleksie in bevolkings van suider-Afrika is bestudeer deur die gebruik van hoë dekking eksoom-volgordebepaling data. Seleksieseine is geïdentifiseer in immuunrespons-gene geassosieer met patogene wat vanaf die 12de eeu na suider-Afrika gebring is. Positiewe seleksie is geïdentifiseer in "focal adhesion" and "ECM receptor interaction" paaie. Dit is duidelik dat daar unieke immuunverwante aanduidings van positiewe seleksie in bevolkings van suider-Afrika teenwoordig is.

Hierdie navorsing het nie net insig tot die genetiese basis en biologie van TB-vatbaarheid bygedra nie, maar het ook die unieke afkoms van bevolkings van suider-Afrika ingespan. Die toevoeging van bevolkingsgenetika verbeter nie net ons studies van TB-vatbaarheid nie, maar kan ook op ander unieke fenotipes in suider-Afrika van toepassing wees. Met die naderende era van persoonlike medisyne, is meer ondersoeke van onderbestudeerde bevolkings van suider-Afrika, wat die meeste deur TB geaffekteer word, nodig.

Acknowledgements

Over the years, there are many people that have guided and supported me. Firstly, to my supervisors (Prof Eileen Hoal, Dr. Brenna Henn and Dr. Marlo Möller), thank you for your continued guidance, advice and reviewing the endless supply of documents over the years. To Eileen, thank you for all the chats in your office and giving me endless history lessons. To Marlo, thank you for always having an open door and putting up with all my questions and concerns. To Brenna, thank you for all your input and supervision from afar, and at all hours. Prof. Paul van Helden, thank you for always believing in me.

Thank you to the HostGen group for all the chats, advice and opinions over the years. I appreciate every lab member and their contact with me while I have been part of this group.

Lastly, I wish to thank my family. Mom, you have been an inspiration to me and have always picked me up when I have fallen. Thank you for all the love and support. Dad, the university fees end now! Francois, thank you for sticking with me through thick and thin. Love you to the moon and back.

Without the input from all the communities in and around South Africa, this research would have not been possible. I therefore dedicate this thesis to each and every research participant.

Contents

Abbreviations

1000G	1000Genomes
ABCG4	ATP Binding Cassette Subfamily G Member 4
ACACA	Acetyl-coA Carboxylase Alpha
ACSM5	Acyl-CoA Synthetase Medium-Chain Family Member 5
AD	Anno Domini
AHSA2	Activator of Heat Shock 90kDa Protein ATPase Homolog 2
AIDS	Acquired Immunodeficiency Syndrome
BCG	Bacille Calmette Guerin
BWA	Burrows-Wheeler Aligner
CA	California
CBL	Cbl Proto-Oncogene
CCR5	CC-chemokine receptor 5
CDH13	Cadherin 13
CI	Confidence Interval
CTSZ	Cathepsin Z
CV	Cross validation
CXCR4	CXC-chemokine receptor 4
CYP3A4	Cytochrome P450 3A4
D6	District 6
DEIC	Dutch East India Company
ECM	Extracellular Matrix
EDARDD	EDAR Associated Death Domain
FAK	Focal Adhesion Kinase
FRAS1	Fraser Extracellular Matrix Complex Subunit 1
F_{st}	Fixation Index
GADD45A	Growth Arrest and DNA Damage Inducible Alpha
GATK	Genome Analysis Toolkit
GBGT1	Globoside Alpha-1,3-N-Acetylgalactosaminyltransferase 1
GWAS	Genome-Wide Association Study
H3Africa	Human, Heredity and Health in Africa
HGDP	Human Genome Diversity Project
HIV	Human Immunodeficiency Virus
HLA	Human Leukocyte Antigen
HWE	Hardy-Weinberg Equilibrium

IFN	Interferon
IL	Interleukin
IRGM	Immune-Related GTPase M
Kbp	Kilo base pairs
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LAI	Local Ancestry Inference
LAMP1	Lysosomal Associated Membrane Protein 1
LD	Linkage Disequilibrium
M.tb	*Mycobacterium tuberculosis*
MAF	Minor Allele Frequency
MARCO	Macrophage Receptor With Collagenous Structure
Mbp	Mega base pairs
MC3R	Melanocortin 3 Receptor
MCMC	Markov Chain Monte Carlo
MHC	Major Histocompatibility Complex
mtDNA	Mitochondrial DNA
MTOR	Mechanistic Target of Rapamycin
NCBI	National Centre for Biotechnology Information
NGS	Next Generation Sequencing
NHGRI	National Health Genome Research Institute
NISCH	Nischarin
NR4A1	Nuclear Receptor Subfamily 4 Group A Member 1
NRAC	Nutritionally-regulated adipose and cardiac enriched protein
OR	Odds Ratio
OSM	Oncostatin M
PBS	Population Branch Statistic
PCA	Principle Components Analysis
POLYPHEN	Polymorphism Phenotyping
PPD	Purified Protein Derivative
SA	South Africa
SAC	South African Coloured
SLC	Solute-Like Carrier
SNP	Single Nucleotide Polymorphism
TB	Tuberculosis
TDT	Transmission disequilibrium test
TIMM44	Translocase of Inner Mitochondrial Membrane 44
TLR	Toll-Like Receptor

TMIGD2	Transmembrane and Immunoglobulin Domain Containing 2
TNF	Tumour Necrosis Factor
TST	Tuberculin skin test
UBASH3B	Ubiquitin Associated and SH3 Domain Containing B
USA	United States of America
VEP	Variant Effect Predictor
WHO	World Health Organization
WIMSA	Working Group of Indigenous Minorities in Southern Africa
XPO1	Exportin 1
ZNF229	Zinc Finger Protein 229

List of Tables and Figures

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

Chapter 5:

Chapter 1

General Introduction

1.1 Tuberculosis- the bacteria and the disease

Tuberculosis (TB) is an airborne bacterial disease caused by various species of mycobacteria. The predominant species in humans is *Mycobacterium tuberculosis (M.tb)*. A third of the world's population is infected with *M.tb* but only 5-15% progress to active disease with severe symptoms; asymptomatic individuals are latently infected (1–3). Symptoms of active disease include coughing, blood in the sputum, weight loss as well as fever. Latently infected individuals do not display symptoms of disease and are sputum culture negative, but tuberculin skin test (TST) positive. In South Africa, ~70% of adults are TST positive once they reach adulthood (4). Due to only a fraction of individuals progressing to active disease, the question remains why some individuals are more susceptible to progress to active TB than others.

There are various factors that can affect progression to active TB. These include the virulence of the mycobacterium, environmental factors, socio-economic factors and the host's genetic make-up. In sub-Saharan Africa, a major cause of progression to active TB is HIV infection (5). Diabetes, smoking and alcohol abuse are other co-factors that influence the outcome of infection. Once immunosuppression has been ruled out as the cause of TB progression, investigations into the host genome, which underlies the immune response, may reveal the cause of disease susceptibility. Immune pathways that are associated with TB susceptibility are mainly those involved in bacterial immunity. An overview of the major proteins and pathways involved in the response to *M.tb* infection is depicted in **Figure 1**.

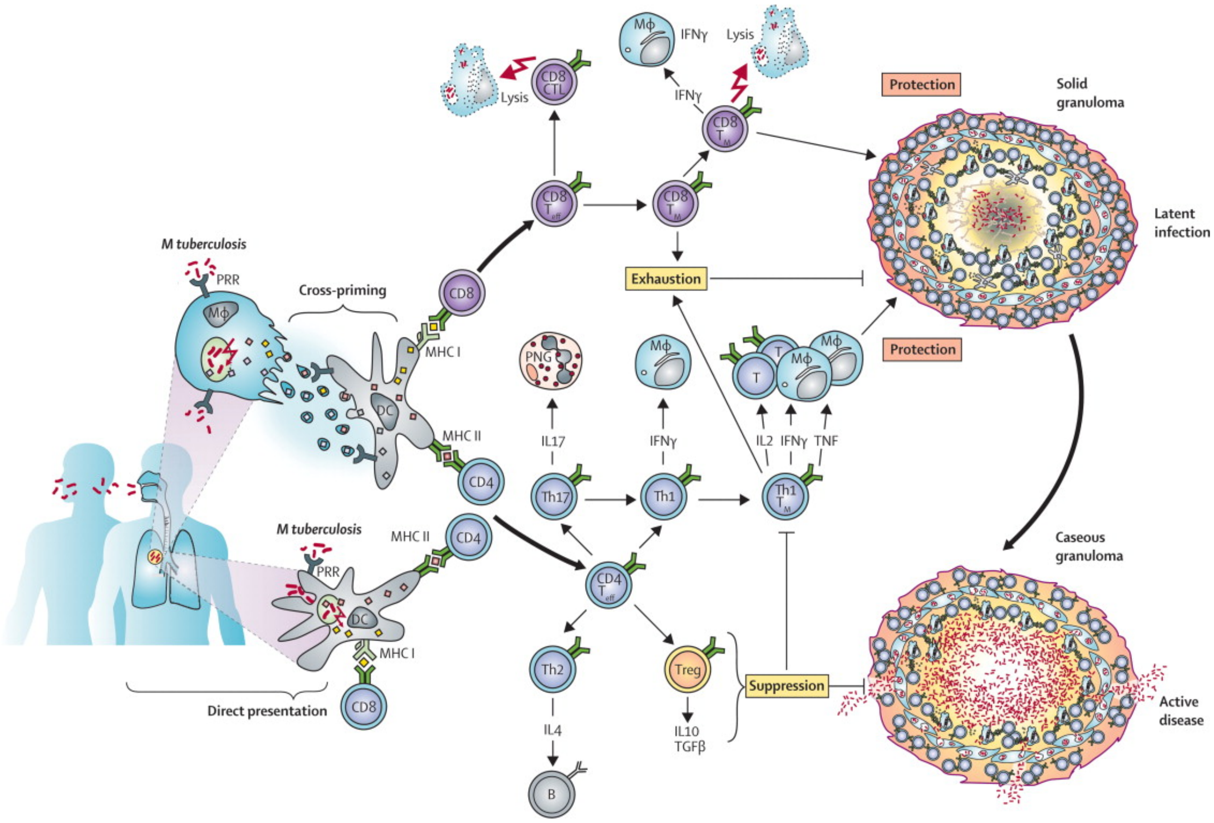


Figure 1: Immune response to infection with *Mycobacterium tuberculosis* (6).

When an individual inhales M.tb, the pathogen is engulfed by macrophages and antigens derived from these pathogens are presented to CD4/8 cells by MHC Class I/II molecules. This initiates a cascade of protein production (interleukins, other cytokines, chemokines and IFN-γ). Any disruption in these processes will lead to decreased immune functioning and thus susceptibility to the invading bacterium (7).

Research focused on identifying the genetic factors underlying susceptibility to TB is a vital addition to the fight to curb the TB incidence rate. The identification of these factors are however confounded by numerous aspects of the host's ancestral history as well as the ever adapting and complex mycobacterium. In order to progress in the combat against TB, it is necessary to consider the history of the disease and the successes and shortcomings of past strategies which identified factors influencing TB susceptibility.

1.1.1 Past

1.1.1.1 **Arrival of TB in southern Africa and its impact on resident populations**

The origin of TB is widely contested. Some studies have suggested that *M.tb* originated in Africa ~40,000-70,000 years ago (8–12), while others claim that TB was largely unknown in Africa before European arrival (the "virgin soil" hypothesis) (13,14).

The hypothesis brought forward regarding an African origin for TB and its spread to the rest of the world, suggested that it was introduced into the New World (via North America) by way of seals ~6,000 years ago (15). In contrast, earlier research suggested a date of 40,000 years ago, coinciding with the migration of anatomically modern humans out of the Horn of Africa . Archaeological data from the Eastern Mediterranean (which includes remnants of the *M.tb* in human fossils) were dated to ~9,000 years old, although this neither supports nor refutes the earlier date as this field of research is in its infancy (16). Over time, the "virgin soil" hypothesis has gained less and less attention from researchers; the African origin hypothesis is now largely accepted. A study has recently suggested an amalgamation of these two hypotheses where TB originated in Africa, but largely dissimilar European strains arrived in southern Africa ~38,300-68,300 years later (9).

Regardless of the origin, once TB arrived in modern day Africa, particularly in southern Africa, the incidence rate rapidly increased (Macvicar, 1908). Examining these incidence rates over time, especially comparing before and after its introduction, may lead researchers to conclusions regarding which populations are more or less susceptible. An example of this can be seen by studying historical records. These records indicate that TB was rife in Europe from the 17th century until the early 19th century, coinciding with the arrival of the Dutch in the southern Cape of South Africa. During the industrial revolution in Europe, the incidence of TB was high, especially in England and France (**Figure 2**). It is therefore hypothesized that virulent European TB strains were introduced to southern Africa by way of the sailors and European settlers that arrived in the 17th century (9). In contrast, the incidence of TB in southern Africa before the 17th century was very low or perhaps non-existent in some populations such as the KhoeSan (**Figure 2**). At the same time that TB was rife in the early 18th century in South Africa, so was smallpox, resulting in mortality rates to increase particularly amongst the indigenous populations. This was due to a number of smallpox epidemics and partly due to multiple TB outbreaks, although TB is seldom an acute disease which obscures its role in the mortality rate. The heightened susceptibility to TB seen in the indigenous populations from the 17th century onwards, extends to modern day populations in southern Africa today. It is thought that the KhoeSan remain susceptible to the disease (as seen from present day TB incidence in the greater Upington (1,000/100,000) area where the ≠Khomani San reside) (District TB co-ordinator, Sr. Magda Amerika, personal communication, January 2016). This susceptibility conceivably extends to populations receiving a large ancestral contribution from the KhoeSan, such as the South African Coloured (SAC) population (17,18). This is discussed in detail in **Chapter 2**.

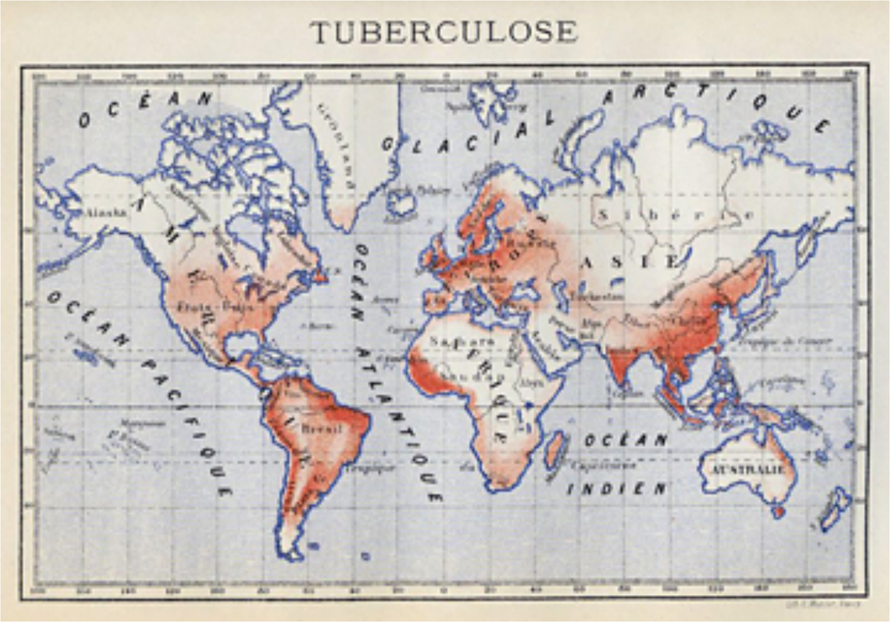


Figure 2: Incidence of TB during the industrial revolution, circa 1700 AD.

(*Center for the History of Medicine/Francis A. Countway Library of Medicine—Harvard Medical School)*

1.1.1.2 **Heritability and host genetics**

One of the most telling and notorious accidents in the scientific field, now termed the Lübeck disaster, suggested that the host's genome might play a role in TB susceptibility. Neonates (251 infants) were accidently administered the Bacille Calmette-Guérin (BCG) vaccine that was contaminated with varying levels of virulent *M.tb*. Upon follow-up, it was noted that some infants did not progress to an active, symptomatic form of TB while others died from the disease, even when a weakly virulent vaccine was used (19,20). This suggested a role of the host genome in TB susceptibility. Further studies investigating the heritability of TB susceptibility have concluded that it ranges between 36-80% (21–24). Twin and adoption studies were the main avenues of investigation in calculating heritability, yet recent studies have utilized immunological phenotypes to improve these heritability estimates. *In vitro* secretion of immunological factors (such as tumour-necrosis factor alpha and interferon-gamma) upon mycobacterial challenge suggested a heritability of > 50% (24).

Chromosomal regions involved in this heritable susceptibility have been investigated by linkage studies and later, case-control and genome-wide association studies (25–29). Originally, linkage studies suggested chromosomes 18q11.2 and 11p13 as possible regions associated with TB susceptibility (30). Later on, genome-wide association studies (GWAS) and case-control association studies identified possible causal genes. These include those involved in antigen presentation (*HLA, TLR*) (31–33), autophagy (*IRGM, LAMP1, MTOR*) (34–37) and cytokine and chemokine signalling (*IL2,IL8,IL10* amongst others) (38,39). It is hypothesized that multiple variants with a small effect size play a role in the TB susceptibility phenotype. In addition, the majority of the variants associated with TB susceptibly are intronic variants of unknown functional relevance. With the establishment of *in silico* functional effect predictors as well as the arrival of whole genome sequencing, it is now possible to fine-map the variants that are likely have the largest functional impact and thus obtain a clearer picture of the multifactorial TB susceptibility phenotype. This is one area that is addressed in this thesis (**Chapter 4**).

In addition to the classical association study, researchers have gone further and investigated the link between ethnicity and TB susceptibility and have shown that certain ethnicities have a heightened risk for the progression to active TB (17,18,40,41). This complicates the identification of possible disease loci in admixed populations as some disease causing mutations might be specific to a certain ancestral population. Therefore, understanding the link between ancestry and TB is vitally important and may help to further elucidate the genetic mechanisms behind this disease.

1.1.2 Present and future - the link between TB disease risk and a population's genetic history

There are genetic factors that contribute to TB susceptibility in southern African populations (17,40,42,43). These factors have been inherited from a wide array of ancestral populations and understanding the genetic origin and history of these ancestral populations, can help researchers to efficiently identify regions associated with the susceptibility phenotype (expanded on in Chapter 2).

All populations have been exposed to different selective pressures, the largest of which is caused by pathogens (44,45). With regards to TB, evidence of positive selection can be seen when considering the increase in resistance to TB in Europeans from the late 19th century onwards (46–48). This is in contrast to the heightened susceptibility seen in southern African populations (particularly the KhoeSan and Bantu-speaking populations) (49). This heightened susceptibility can be seen in the mortality rate of African populations in the area during the early 20th century (**Figure 3**) (49). The possible implications of the contrasting susceptibility profiles is investigated in **Chapter 5**.

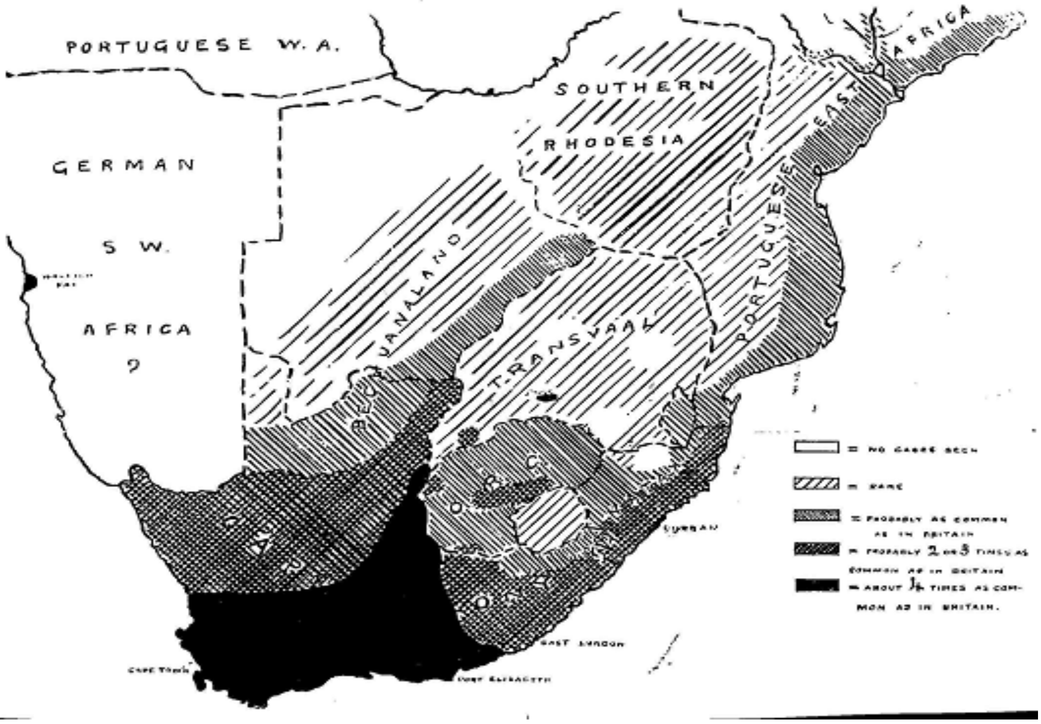


Figure 3: Death rate from phthisis 1903-1905 (50).

The KhoeSan, Bantu-speaking populations as well as Europeans contributed genetically to a number of other southern African groups (such as the SAC population) (discussed in section 1.2.1), thus perhaps transferring genetic regions that confer an increase and a decrease in TB susceptibility respectively (51–55). In the past, selection scans for these regions were confounded by the complex genetic make-up of admixed populations in southern Africa as well as the lack of truly representative ancestral populations. These caveats have since been resolved with the arrival of Next Generation Sequencing (NGS) data from a number of populations as well as the tailoring of computational and statistical tools for the identification of regions under selection in complex population scenarios. These new additions to the fields of evolutionary genetics have been harnessed in **Chapter 5**.

The addition of population genetics to any study investigating a particular disease is highly beneficial. It can not only provide insight into the history of the disease, but it can help identify possible genetic mechanisms underlying the phenotype. For this reason, understanding the genetic history and structure of southern Africa populations is vitally important and discussed in great detail in section 1.2, concluding with a discussion on a unique southern African population who are highly susceptible to TB, largely due to their genetic history (17,18,40).

1.2 Human populations in southern Africa

Deviations in allele frequencies due to gene flow influence a number of phenotypes, including infectious disease risk (56–59). It is therefore vital to understand genetic structure and diversity in southern Africa, as this will help elucidate the effect of population structure on infectious disease phenotypes. The ultimate goal is the utilization of new genetic information to combat the incidence and mortality rates of infectious diseases in southern Africa.

1.2.1 Human origins in southern Africa and the history of early population migration in the area

Southern Africa is a unique area filled with great diversity. This diversity extends to ecology, geography and the genetic composition of human populations in the area. The region's history is one of the most distinctive, reaching back at least 100,000 years to the origin of modern humans (60). An African origin for modern humans is supported by archaeological data as well as genetic analysis (60–66). Genetic diversity studies of the KhoeSan have revealed that they are the most divergent population worldwide (64,67–69).

KhoeSan is a collective term for all San and Khoe individuals living in southern Africa. It was originally thought that the San were the only indigenous inhabitants of southern Africa. However, it has now been hypothesized that the Khoe have resided in southern Africa for as long as the San (55,70). The KhoeSan have recently absorbed migrants from other African and non-African populations, starting with the migration of East African pastoralists ~2,000 -

5,000 years ago which resulted in fine-scale structure (55,71–73). However, the migration with one of the largest impacts was that of Bantu-speaking farmers which started ~5,000 – 6,000 years ago from West and Central Africa.

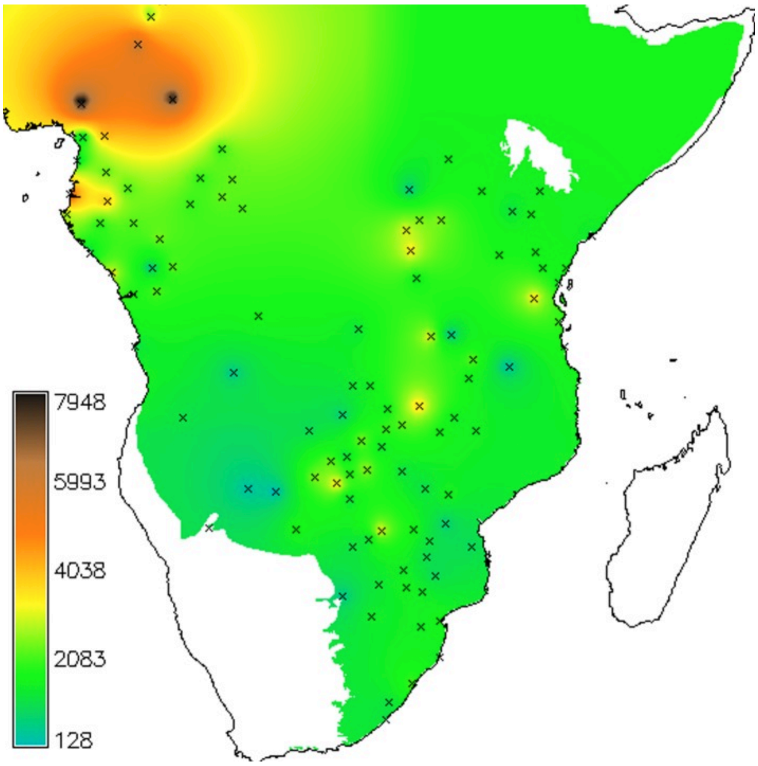


Figure 4: Bantu expansion as modelled by the accumulation of all known dates provided by archaeological data (74). The x's mark the location of the archaeological evidence. The colour key represents the scale of time from 129 years before present (BP) to 7948 years BP.

The Bantu-speaking population expansion from West and Central Africa reached southern Africa 1000-2000 years BP (74). The leading hypothesis is that this movement occurred largely along the east coast of southern Africa with some movement towards the west as seen in **Figure 4** (74,75). The western migrations have not been well documented, however the migrations along the east coast were thought to have occurred by numerous short dispersals (76,77). These individuals spread across present-day South Africa, but were limited to geographical areas in which cattle husbandry was possible. Due to climate changes in the first millennium, populations moved towards the Kalahari (west) and some populations adapted their subsistence strategy to the more arid environment (74,78). Many individuals settled in the now Eastern Cape and Kwazulu-Natal provinces of South Africa and formed the Xhosa, Zulu and Ndebele tribes (79).

The most recent documented population migration into southern Africa was by the Europeans. The Portuguese sailors passed by the tip of southern Africa in the 15th century and encountered the indigenous San and Khoikhoi. Encounters were brief and involved conflict, resulting in the movement of the Portuguese up the east coast, to Mozambique (Barthlomeu Dias, personal diary entry). The Dutch arrived in the southern Cape in 1652. A trading post and refreshment station was established by the Dutch East India Company (DEIC) for the maritime trade around the tip of Africa. Due to the amount of resources in the area, the European community flourished (80). The French and later the British were also inhabitants of the area, the latter population colonizing the area in the late 18th and early 19th Centuries.

Slave trade began in the 17th century and was initiated by the DEIC. Slaves were transported from East and South-East Asia and Madagascar by ship. In addition to the Asian slaves, the KhoeSan were used as domestic and farm workers. KhoeSan women frequently married and/or had children with European men. Due to the differing cultures of all the populations in the area and competition for resources, there was conflict. This together with the introduction of foreign pathogens resulted in the severe loss of life, particularly amongst the KhoeSan (49,81).

1.2.2 Genetic and cultural differences between and within populations in southern Africa

1.2.2.1 The KhoeSan

The KhoeSan span a broad range of subsistence strategies and language families. The majority are hunter-gatherers and speak a wide array of click languages (**Table 1**). There is no definite pattern regarding the geographical distribution of subsistence strategies or language usage across the 5 southern African countries the KhoeSan inhabit (**Table 1**) (discussed further in Chapter 2).

Table 1: Language families and subsistence strategies of the KhoeSan in southern Africa.

Population Sample	Location of Sample	Language Family	Historical Subsistence
!Xun	Namibia and Angola	Kx'a	Hunter-gatherers
//Gana	Botswana	Khoe	Hunter-gatherers
/Gui	Botswana	Khoe	Hunter-gatherers

≠Hoan	Botswana	Kx'a	Hunter-gatherers
≠Khomani	South Africa	Tuu (!Ui-Taa)	Hunter-gatherers
Damara	Northwest Namibia	Khoe	Pastoralist/ Hunter-gatherers
EastTaa	Namibia, Botswana & South Africa	Tuu (!Ui-Taa)	Hunter-gatherers
Hai||om	Namibia	Khoe	Hunter-gatherers
Ju/hoansi	Namibia, Angola & Botswana	Kx'a	Hunter-gatherers
Khwe	Namibia, Botswana & Angola	Khoe	Hunter-gatherers
Kua	Zimbabwe & Botswana	Khoe	Hunter-gatherers
Nama	Namibia & South Africa	Khoe	Pastoralist
Naro	Namibia & Botswana	Khoe	Hunter-gatherers
NorthTaa	Namibia, Botswana & South Africa	Tuu (!Ui-Taa)	Hunter-gatherers
Shua	Botswana	Khoe	Hunter-gatherers
WestTaa	Namibia, Botswana & South Africa	Tuu (!Ui-Taa)	Hunter-gatherers

Within South Africa, the ≠Khomani and Nama are the two most prominent KhoeSan populations. The ≠Khomani San are Kx'a speaking hunter-gatherers whose population history is well documented in historical records as well as present day genetic analyses (55,61,72,82,83). In contrast, the Nama are Khoe speaking pastoralists whose origin is widely contested (71,84–92). One hypothesis is that the migration of East African pastoralists into the area between 2,000 and 5,000 years ago resulted in the Khoe adopting a similar subsistence strategy to that of the migrants by a combination of cultural and demic diffusion where the latter plays the largest role (71,84,85). This area of research is one of the topics investigated later in this thesis (**Chapter 3**).

Recent studies have shown that although there are genetically homogenous KhoeSan populations in southern Africa, there is admixture present, at varying degrees, from European and Bantu-speaking populations (55,82). The origin of this ancestry can largely be attributed to population migrations into the area (as discussed in **section 1.2.1**). The pattern of admixture does not seem to correspond to language usage and/or subsistence strategies but is largely governed by the geographical location of the KhoeSan population (55).

Understanding the genetic structure within KhoeSan populations is vital to better understand the genetic history of extant populations that received genetic contributions from the KhoeSan, such as the Bantu-speaking populations in southern Africa (**section 1.2.2.2**) as well as the highly admixed SAC population (**section 1.2.2.4**).

1.2.2.2 Bantu-speaking populations

Bantu-speaking tribes in southern Africa include the Xhosa, Zulu, Sotho, Pedi, Tswana, Venda, Ndebele, Pondo, Swati and Tsonga. Each population has its own distinct history, culture, language and subsistence strategy.

Genetic structure analysis shows that the majority of Bantu-speaking populations in southern Africa originate from Central and West Africa (54,93). There is evidence however that there was some admixture with KhoeSan as well as European populations, although the latter would have mostly occurred much later (51,54). It has been suggested that the Zulu and Sotho share similar patterns of admixture with the largest proportion being hunter-gatherer ancestry (~23%) (94). This is consistent with their usage of click consonants (94).

One of the few studies characterising patterns of population structure in southern African Bantu-speaking populations investigated genetic differentiation between seven such populations. It was shown that six of the seven populations, namely Zulu, Xhosa, Southern Sotho, Pedi, Tswana and Venda cluster according to their language usage whereas the Tsonga did not (79). The Tsonga clustered with the Venda who are geographically closer. In addition, genetic distances were found to correlate with geographic distances yet linguistic distances did not correlate with either genetic or geographic distances. It was therefore concluded that language and genetic differentiation (and most likely admixture) occurred before these populations reached their present location with further genetic changes after their arrival (79).

1.2.2.3 The European-descent population

The European-descent populations in southern Africa have a diverse background. The origin of these populations predominantly represents the European power that colonized their country of residence. For example, the majority of Europeans in Namibia are of German descent due to German colonization (95) from 1884 until 1917. With the exception of Angola and Mozambique (Portuguese colonies), the British colonized the rest of southern Africa at different times until the 20th century. The Dutch (the first nation to colonize South Africa), was one of the European populations to found the Afrikaner population in South Africa (96,97). The majority of white people (60.8%) in South Africa speak Afrikaans, a dialect derived from Dutch (SA Census 2014).

There have only been a few studies investigating the population structure and ancestry of the Afrikaners (96,97); the majority are based on historical findings. A large proportion of the Afrikaner ancestry is Dutch with some influence from other European, African (both black and KhoeSan) and Asian populations (96,97). Other analyses, such as genealogy, support this conclusion (98).

The ancestors of the Afrikaner population contributed substantially to the approximately 19% European ancestry present in the SAC population (51–53).

1.2.2.4 The South African Coloured population

The SAC population is highly admixed, receiving ancestral contributions from five populations (53). The SAC population is a direct result of population migrations into southern Africa, as discussed in section 1.2.1. There is a strong maternal KhoeSan contribution as well as a strong paternal European contribution to the population, although there are regional within-population differences (99). Overall, Bantu-speaking and KhoeSan populations contributed the highest ancestral proportions (~30% each), with the remainder being made up of East and South-East Asian as well as European ancestries (51–53,99).

Within the SAC population, there is great diversity (54). Gene flow from ancestral populations was highly dependent on geographical proximity and in some cases there were barriers to gene flow. Individuals from the Northern Cape have a higher proportion of European ancestry as compared to SAC individuals from the Western Cape (~40% vs ~19% respectively) (54). Individuals from the Eastern Cape (78.8% Xhosa-speaking) and KwaZulu-Natal (77.8% Zulu-speaking) provinces have a higher proportion of Bantu-speaking ancestry (54). Lastly, the SAC in the Western Cape have both a higher proportion of Asian ancestry as well as KhoeSan ancestry. This is due to the Indian slaves that were brought into the area in the 17th century as well as the KhoeSan who are indigenous to the area (54). The large genetic diversity within the SAC can be seen from the spread of SAC individuals across a principal components analysis (**Figure 5**).

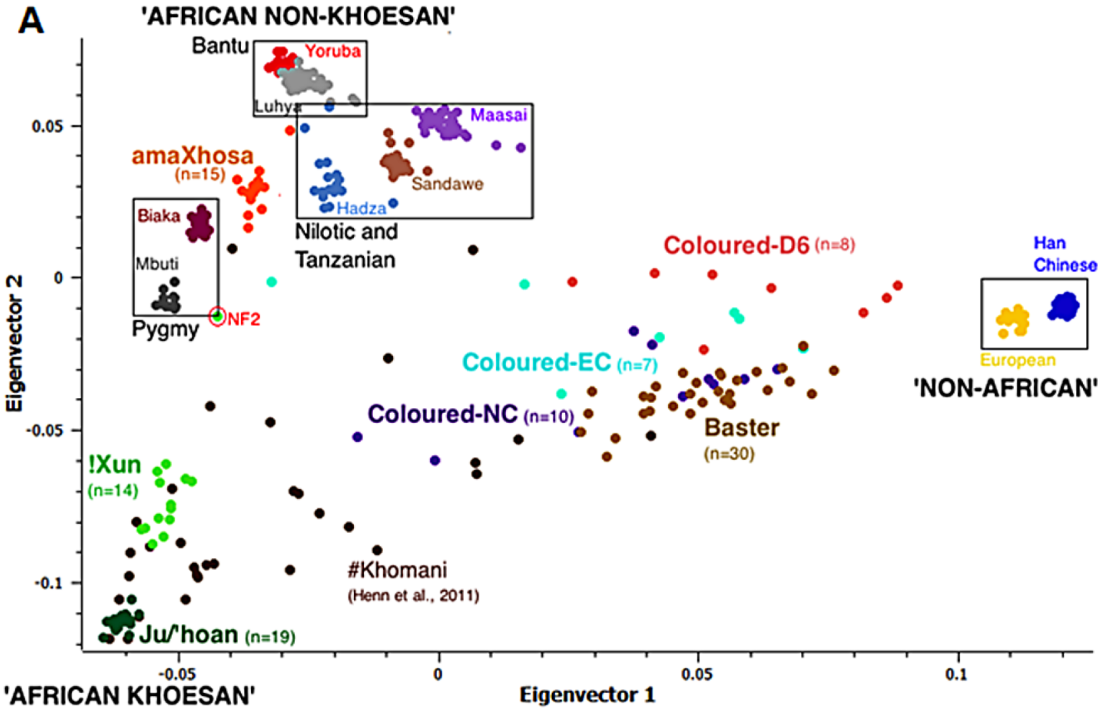


Figure 5: Principal Components Analysis depicting the clustering of SAC populations, which is highly associated with geography (54).

Due to the diverse ancestral contributions, the SAC have a unique genetic make-up. In essence, their allele frequencies, signals of selection and patterns of linkage disequilibrium are greatly different from other populations (93). This influences numerous phenotypes, of which infectious disease susceptibility is one. Promising results suggesting a link between ancestry and TB susceptibility warrant further investigation. However, a more detailed analysis of population structure is required before this can take place. Once this is complete, we can utilize the information in investigations of the precise genetic mechanisms under-pinning TB susceptibility as well as the origins of the predisposing variants.

2. Scope of the thesis

Research presented in this thesis addresses the role of population structure in susceptibility to TB by examining southern African population diversity and genetic history as well as its impact on the disease phenotype.

Studies in the field of population genetics in southern Africa have made important contributions. These include the understanding of population history, evolution, origins and genetic differences between and within populations. The extent of research in this field has increased over the past years, with the development of large international consortia. This research was summarised in a review (**Chapter 2**), with the goal of linking this information to adverse phenotypes associated with population history and genetic structure. This was done with particular reference to infectious diseases prevalent in southern Africa; HIV and TB are two such examples.

Dozens of different KhoeSan groups exist, belonging to three different language families, but very little is known about their population history. Although there have been great advances in modelling the distinct patterns of population structure in southern Africa, the explanations previously presented were one-dimensional and could not explain numerous complexities. An example of this can be seen with the "all-or-nothing" explanation researchers provide for the origins of the Khoe pastoralists in southern Africa i.e. either demic or cultural diffusion without considering that it might be an amalgamation of the two and therefore more complex than previously thought with numerous confounding factors. For this reason, we examine new genome-wide polymorphism data and whole mitochondrial genomes for 100 South Africans from the ≠Khomani San and Nama populations of the Northern Cape, analysed in conjunction with 19 additional southern African populations (**Chapter 3**). This is one of the largest southern African datasets to date and provides insight into demographic history, migration routes as well as barriers to gene flow.

Identifying variants associated with a complex phenotype is difficult and time consuming. In addition, as is the case with TB susceptibility, it is most likely a combination of exonic and intronic variants that contribute to the phenotype. Furthermore, GWAS for TB has been contradictory, perhaps due to LD patterns. For these reasons, we aimed to identify functionally relevant regulatory variants in linkage disequilibrium with SNPs previously associated with TB (**Chapter 4**). We conducted a post-GWAS analysis to determine the association of the identified putative functional variants with TB.

Signals of natural selection in southern Africa have not been explored in great detail. Due to the before mentioned genetic diversity as well as the high incidence and mortality rates of smallpox and TB from the 17th century onwards, we hypothesized that distinct signals of selection exist in the immune response genes of the SAC population. We utilized exome sequence data and adapted the Population Branch Statistic to identify these signals. We propose the functional relevance of each variant and suggest novel pathways possible associated with infectious disease susceptibility (**Chapter 5**).

Chapter 2

Population structure and infectious disease risk in southern Africa

Caitlin Uren[1], Marlo Möller[1], Paul D van Helden[1], Brenna M Henn[2], Eileen G Hoal[1§]

[1] SA MRC Centre for TB Research, DST/NRF Centre of Excellence for Biomedical Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medical and Health Sciences, Tygerberg Campus, Parow 7500
[2] Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794

Molecular Genetics and Genomics 2017 (Impact Factor 2.979)
doi:10.1007/s00438-017-1296-2

Abstract

The KhoeSan populations are the earliest known indigenous inhabitants of southern Africa. The relatively recent expansion of Bantu-speaking agropastoralists, as well as European colonial settlement along the south-west coast, dramatically changed patterns of genetic diversity in a region which had been largely isolated for thousands of years. Owing to this unique history, population structure in southern Africa reflects both the underlying KhoeSan genetic diversity as well as differential recent admixture. This population structure has a wide range of biomedical and sociocultural implications; such as changes in disease risk profiles. Here, we consolidate information from various population genetic studies that characterize admixture patterns in southern Africa with an aim to better understand differences in adverse disease phenotypes observed among groups. Our review confirms that ancestry has a direct impact on an individual's immune response to infectious diseases. In addition, we emphasize the importance of collaborative research, especially for populations in southern Africa that have a high incidence of potentially fatal infectious diseases such as HIV and tuberculosis.

Introduction

Southern Africa has a unique and complex human history reaching back at least 100,000 years (100). The region spans southern Angola, Namibia, Botswana, South Africa, Zimbabwe and Mozambique. Many diverse ethnic groups are present in the area, including KhoeSan populations, Bantu-speaking populations, European-descent groups and groups resulting from inter- and intra-continental admixture such as the South African "Coloured" population (de Wit et al. 2010; Daya et al. 2013; Chimusa et al. 2013). "Admixed" populations are the result of gene flow between distinct, historically divergent parental populations, such as those from different continents like Asia and Africa. The rate, extent and timing of gene flow between genetically distinct populations has resulted in unique genetic complexity in almost all populations in southern Africa, as well as fine-scale genetic differences between populations. Patterns of allele frequency differences among populations is described as population structure and such allele frequency differences can have subtle or profound phenotypic effects, such as differential susceptibility to infectious disease.

The genetics underlying human disease phenotype variation in African populations have been under-researched. However, the NIH and Wellcome Trust-funded initiative named the Human Heredity and Health in Africa (H3Africa) (101,102) aims to improve the health of all African populations by facilitating research in the area of genomic and environmental impacts on common diseases such as trypanosomiasis, tuberculosis, rheumatic heart disease, schizophrenia, type 2 diabetes and other cardiometabolic diseases. We focus on recent genetic investigations of the two infectious diseases with the biggest impact on health in southern Africa, viz. tuberculosis (TB) and the human immunodeficiency virus (HIV).

Tuberculosis (TB) and the human immunodeficiency virus (HIV) have high incidence and mortality rates in southern Africa (WHO, 2015). A major component of TB susceptibility is genetic, and recently, it has been established that part of this susceptibility can be attributed to a particular ancestral population, which contributed to present populations (17,18,40). Understanding the role of ancestry in infectious disease risk has manifold benefits, including the identification of the most vulnerable populations, providing effective and specialized drug therapies and/or vaccines; and in the highly probable case of the identification of novel susceptibility factors, aiding in the development of new therapies.

Here, we review the population structure and prehistory of southern African populations, as inferred from recent genetic and genomic datasets, to provide a better understanding of how population structure affects disease risk. Thorough literature searches were performed using Pubmed and Google Scholar in order to capture a wide array of the latest studies in the field using “ancestry related disease risk”, “southern Africa population genetics” and “TB and HIV in southern Africa” as keywords.

The genetic history of the KhoeSan

The KhoeSan are indigenous inhabitants of southern Africa and their ancestors may represent the earliest divergence among extant human populations (61,73,103–110). The genetic origin of the KhoeSan can be traced back to the emergence of modern humans in southern Africa (111,112). Prior to 2,500 years ago, all KhoeSan populations in southern Africa hunted game or fished, foraged for plants and gathered natural products, hence the anthropological term of hunter-gatherers. Some contemporary KhoeSan populations continue to forage while other groups have transitioned to wage labour or stock farming. The Khoekhoen are pastoralists who derive their ancestry from the original hunter-gatherer KhoeSan, but adopted sheep, goat and cattle husbandry from east African pastoralists approximately 2,000 years ago (55,70,89). Bantu-speaking farmers arrived in southern Africa from approximately AD 600 onwards, having migrated down both the west and east coasts of Africa, and subsequently impacted the Khoekhoe and San way of life. The expansion of Bantu-speaking agriculturalists was followed by Arab traders who sailed down the east coast at least as far as Sofala, Mozambique. The Portuguese (the first European visitors to South

Africa) encountered the San and Khoekhoen in Mossel Bay, South Africa in 1487, although this and subsequent encounters were brief due to conflict with these indigenous groups. The Dutch and other Europeans began a formal settlement at the Cape of Good Hope (present-day Cape Town, South Africa) in 1652. Within a short period of time, Indian and Asian slaves were brought to the area. These historical events are depicted in **Figure 1**. Over time, the Bantu-speaking, European, San and Khoekhoen, all culturally distinct groups, intermarried with one another, a fact evident not only from their genomes but also from resulting language and cultural practices (113). The European and Bantu expansion into San and Khoekhoe territory resulted in the decline of the indigenous populations due to conflict, disease and resource scarcity (49).

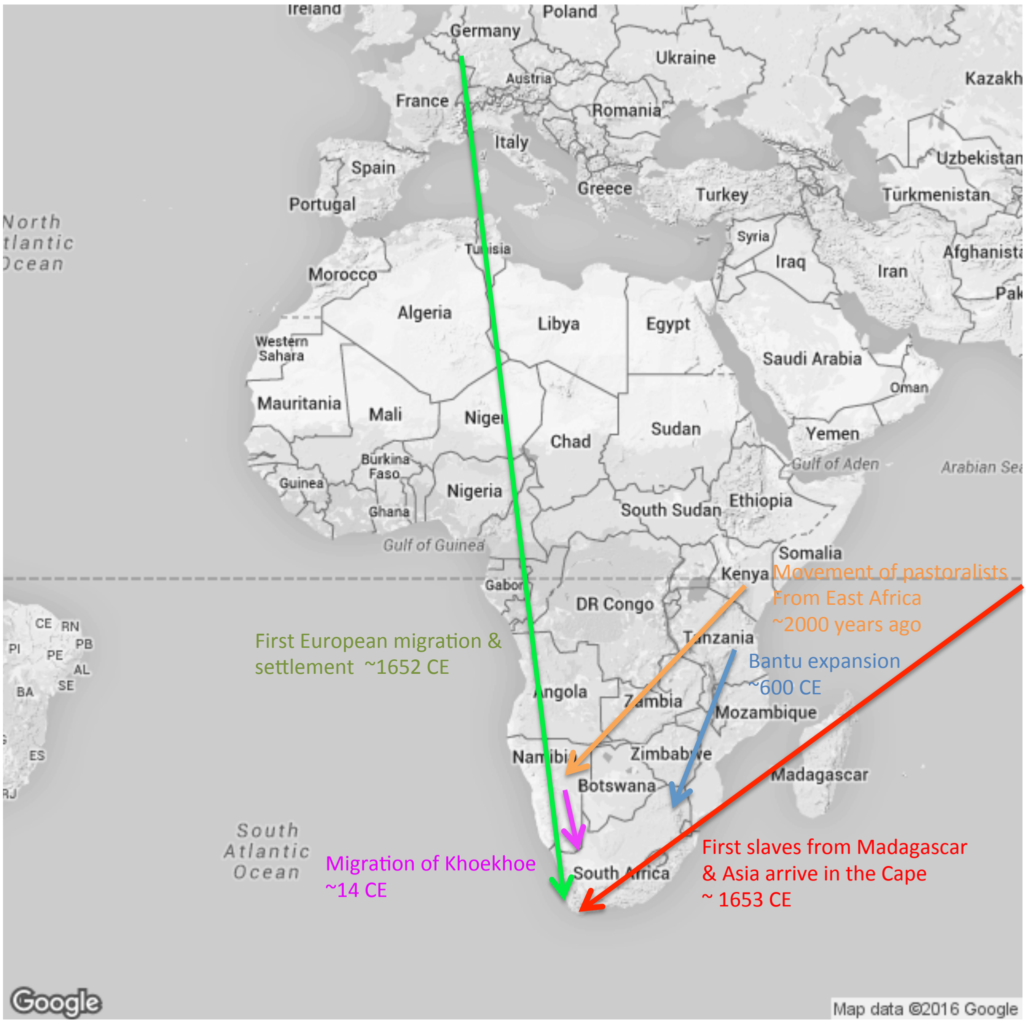


Figure 1: Southern Africa's complex and long-standing historical migrations.

Map of Africa depicting the primary population migrations into southern Africa. Geographical locations of most relevant southern African populations mentioned in this review, are depicted by black circles.

Archaeological and genetic evidence suggests that the modern human species originated within Africa, though the precise location of origin is widely contested due to the diversity of African populations and complexity of population history (114). One hypothesis proposes that the earliest population divergence among humans occurred within southern Africa based on the exceptional genetic diversity present in KhoeSan groups (61). Demographic history within KhoeSan populations has been widely investigated with particular reference to their origins and thus the origins of modern humans. One such study included click-speaking Hadza and Sandawe individuals from Tanzania, ≠Khomani San from South Africa as well as 24 other African populations. Linkage disequilibrium and heterozygosity analysis from single nucleotide polymorphism (SNP) array data demonstrated that the ≠Khomani and other KhoeSan from Namibia are two of the most genetically diverse populations in the world (61). In conjunction with F_{st} patterns (a measure of genetic distance between populations), their results suggested that humans originated in southern Africa (61). This conclusion is supported by microsatellite and indel data (64). It was found that the two southern African KhoeSan populations from this study clustered together when phylogenetic trees were constructed from genetic distances (F_{st}) between populations. These populations were also the most distinct populations worldwide (64). This is broadly consistent with studies on mitochondrial DNA and Y chromosome, which indicated divergent genetic lineages (115,116). Studies investigating the geographical origins of modern humans depend on contemporary populations which might not be truly representative of historical populations (117). Nonetheless, under standard phylogeographic inference, and supported now by whole genome analysis of effective population size (118), the oldest population divergence among humans occurred in southern Africa.

In addition to determining the origins of modern humans by inferring the geographic origin of the KhoeSan, it is important to distinguish when and how the ancestors of the KhoeSan diverged from other groups. Tishkoff et al. suggested that contemporary Central African Pygmy, KhoeSan, Hadza and Sandawe hunter-gatherer populations were remnants of a larger ancient "proto-KhoeSan-Pygmy" population with the divergence into distinct populations being estimated to have occurred >35,000 years ago (73). A deeper date of divergence between the KhoeSan populations from other African populations was estimated by Veeramah et al. (2012), who re-sequenced 40 intergenic regions in individuals from the San, Eastern and Western Pygmies as well as non-Pygmy Niger-Kordofanian populations. They concluded that the ancestors of the KhoeSan separated from a "proto-Pygmy-non-Pygmy Niger-Kordofanian group" ~100,000 years ago (65), which is consistent with (72). An even deeper date of divergence was found by analysing whole-genome sequences of six individuals from diverse ancestral backgrounds (111). The divergence of the San from all other human populations was postulated to occur 108,000-157,000 years ago whereas the divergence of Eurasians from ancestral African populations occurred 38,000-64,000 years ago (111).

After the separation between ancestors of the KhoeSan and all other populations, there has been further north-west and south-east divergence of Kalahari groups that led to deep structure within the KhoeSan (63,72,83,119). The approximate date of the divergence was estimated at ~30,000 years ago (63). This divergence can still be seen in extant KhoeSan populations as principal component analysis (PCA) reveals three main clusters of KhoeSan populations: namely non-KhoeSan, north-western Kalahari and south-eastern Kalahari (63). According to the distribution of L0d and L0k mtDNA haplogroups, which appear virtually only within the KhoeSan groups, the more northerly southern African groups (!Xun, Ju/'hoansi, and /Gui, //Gana) cluster separately from the more southerly southern African groups (≠Khomani, Karretjie and some SAC populations), who have their own distinct cluster (83).

Recent studies investigating admixture proportions and the distribution of lactase persistence alleles in extant southern African populations have shown that the KhoeSan populations are heterogeneous and some have admixture from European, Bantu-speaking and East African populations (55,71,84,85,120–122). The gene flow can largely be attributed to three migration events. The first was by East African pastoralists 2-3 kya and the second by Bantu-speaking farmers <1 kya (84,85,122,123), followed by European colonization. These migration events have contributed to the genetic diversity in a number of southern African populations, including alleles that affect phenotype. For example, lactase persistence alleles originating in East Africa were found in the Khoe-speaking Nama at a relatively high frequency, as compared to other pastoral populations such as the Himba (84,85,122,124). In addition, admixture analysis in southern and eastern Africa populations shows that the Eurasian ancestry in southern Africa originated in eastern Africa (125). These two conclusions support the hypothesis that there was admixture between East African individuals and the native Khoe populations. East African ancestry as well as Bantu-speaking ancestry arising from the later Bantu expansion could impact numerous phenotypes in southern Africa, including TB susceptibility. This has not yet been investigated in depth but initial studies have shown that Bantu-speaking ancestry in the SAC population predisposes individuals to progress to active TB (17).

Southern African Coloured populations and their link to the KhoeSan

The South Africa Coloured (SAC) population represents a highly admixed group of individuals from multiple ancestral populations (53). Within the past ~600 years, multiple non-African and African populations have moved into southern Africa and integrated with the indigenous inhabitants (the KhoeSan). The origin of the SAC population has recently been the subject of a number of studies, concerned with quantifying the number, provenance and proportions of the ancestral populations.

This complex gene flow pattern is crucial in the understanding of the origins of other populations in the region, such as the SAC population (predominantly found in the Western Cape of South Africa) which received over 30% of their ancestry from the KhoeSan (51–53,55). An early study analysing genome-wide SNP data from 20 SAC, indicated 4 ancestral contributions to this population, namely European, South Asian, Indonesian and Xhosa (126). Although these results have some similarities with later ancestral determinations, there were some significant differences. First, it is not clear whether Cape Malay individuals (admixed and similar to the SAC but with higher Asian ancestry due to the slave trade) were included in the sample, which could have biased ancestry results towards Indonesia or more broadly, South Asia. Second, it is not clear which "Bushman" population was used as a proxy ancestral KhoeSan population. Since we know that each KhoeSan population differs substantially in their ancestral admixture pattern (as shown above), interpretations of KhoeSan ancestry in the SAC can be biased or even missed. As more San reference samples became available, and the within-population structure became more evident, the understanding of admixture in the SAC and the analyses thereof improved.

In 2010, a large genome-wide analysis of 959 SAC individuals was performed, where autosomal SNPs were genotyped and combined with data from distinct populations present in the Human Genome Diversity Project (127). With the large number of putative ancestral populations available, the results fitted the historical data more accurately. It was found that ancestry proportions in the SAC were dominated by the KhoeSan ancestry which was estimated at 32-43%, followed by black African ancestry at 20-36%, European ancestry 21-28% and Asian 9-11% (53). The accuracy of any inferences made from ancestry data can be affected by the choice of reference populations, the number of individuals used in each reference population as well as the algorithm used. The analysis by Patterson et al. (2010) described only the continental admixture present in the SAC population, but the best proxy ancestral populations for the SAC were unknown (126). PROXYANC was therefore developed by Chimusa et al. (2013) and is based on two novel algorithms, namely, population genetic differentiation and optimal quadratic programming. PROXYANC identifies the most accurate and efficient ancestral populations for a multi-way admixed population. Once the most representative populations were identified using PROXYANC, ancestry proportions were calculated using ADMIXTURE. The Xhosa (black African) contributed 33%±0.226, the ≠Khomani San (KhoeSan) contributed 31%±0.195, the Europeans contributed 16%±0.118, the Gujarati Indians contributed (South Asian) 13%±0.094 and the Chinese (East Asian) contributed 7%±0.0488 (51). The combination of Bantu-speaking ancestries (East and West African) in the SAC was estimated at 33%±0.04 (55). Upon the identification of a southern KhoeSan specific ancestry, both the Nama and ≠Khomani were utilized as KhoeSan reference populations for the SAC resulting in a reported ancestral contribution of 33%±0.03 (128). In this study, European ancestry was estimated at 12%±0.02, Pathan (South Asian) ancestry at 14%±0.02 and Chinese ancestry at 7%±0.01 (128). **Table 1** summarises the ancestry proportions as determined by these studies. It clearly shows the increase in accuracy and correlation with historical data. In addition, as datasets and methodologies advanced, it is noteworthy that the estimation of the correct proxy ancestral population to be used, improved. The majority of studies investigating admixture proportions in southern African populations focused primarily on SAC individuals from the Western Cape. However, Petersen et al. (2013) investigated admixture proportions in other SAC populations from around South Africa. SAC individuals from the Eastern Cape had an increase in Bantu-speaking ancestry and individuals from District Six in Cape Town had an increase in Asian ancestry (slaves from India and Madagascar lived in District Six before slave trade was abolished) (54).

Table 1: The evolution of the South African Coloured population's ancestral proportions.

Ancestry proportions of the SAC population from studies using different ancestry determination methods and reference populations. Ancestry proportions are reported as summary means with their respective standard error.

Patterson et al. (2010)				
Xhosa		*European*	*Indonesian*	*South Asian*
37% ± 0.003		23% ± 0.008	18% ± 0.004	22% ± 0.009
de Wit et al. (2010)				
West African	*KhoeSan*	*European*	*Chinese*	*Indian*
24% ± 0.161	37% ± 0.148	18% ± 0.118	7% ± 0.0478	14% ± 0.093
Chimusa et al. (2013)				
Xhosa	*KhoeSan*	*European*	*Chinese*	*Indian*
33% ± 0.226	31% ± 0.195	16% ± 0.118	7% ± 0.0488	13% ± 0.094
Uren et al. (2016)				
Bantu-speaking	*South African KhoeSan*	*European*	*Chinese*	*Pathan*
33% ± 0.04	33% ± 0.03	12% ± 0.02	7% ± 0.01	14% ± 0.02

mtDNA can be very informative for studying sex-biased migration and formation of complex populations. Quintana-Murci et al. (2010) investigated maternal and paternal ancestral contributions to the SAC population. Sub-Saharan Africa was the origin of the greatest proportion (79%) of the SAC maternal gene pool (99). It is important to note that 60% of the SAC mtDNA was of the L0d lineage which, together with L0k, is specific to the KhoeSan (99). The SAC population therefore contains considerable maternal input from the KhoeSan. Other mtDNA haplogroups found in the SAC population derive from the Bantu expansion (19%). The remainder of mtDNA ancestry was contributed by south and south-east Asian populations, consistent with autosomal data (99). On the other hand, the paternal contribution to the SAC population was dominated by a contribution from sub-Saharan Africa, about twice that of the maternal contribution. The most dramatic difference observed was between the paternal (5.3%) and maternal (60%) KhoeSan ancestry (99). These results displayed an uneven sex-specific gene flow both between and within continents and sheds light on the admixture events and social environments that brought the modern day SAC population into being.

The modern day SAC population spans much of southern Africa, with some geographical variation in terms of genetic and cultural characteristics. The Karretjie people (officially classified as ‘Coloured’), of the Great Karoo (found in the Northern Cape and Western Cape of South Africa) were analysed in a similar way to the SAC by looking at the distinction between maternal and paternal contributions. Interestingly, the KhoeSan specific clade L0d as mentioned above was present in all the Karretjie samples (n= 31), suggesting a solely KhoeSan maternal contribution (83). In contrast, paternal contributions were more heterogeneous, similar to the SAC in the Western Cape (99). This pattern is also evident in the Rehoboth Basters, a distinct group of individuals who moved from the Cape Colony to southern Namibia 150 years ago. The paternal lineage of this group is European, while the maternal lineage is KhoeSan (54). Using a SNP array dataset, ADMIXTURE analysis indicates five ancestral components, similar to the SAC, and in the PCA, the Basters cluster with the SAC (54). Similarities in ancestry proportions were observed between the Coloured and Baster individuals but with higher European and KhoeSan ancestry in the Basters.

More recent mtDNA studies have suggested that the sex-biased admixture in southern African populations is not so straightforward as previously thought (e.g. KhoeSan maternal lineages vs. largely Bantu-speaking paternal lineages in the SAC) with analyses suggesting high levels of inter-population variance at the maternal lineage level (129,130). This complexity evident not only in the SAC population, but in other southern African populations who share such a complex genetic history such as Bantu-speaking populations (131). These conclusions are supported by Y chromosome analyses (and thus telling of paternal lineages) where an increase in mutation rate was identified in Bantu-speaking haplogroups suggesting either a population expansion and/or an older age of paternity (132).

Multiple ancestries in southern Bantu-speaking populations

The majority of southern African individuals currently belong to a large variety of Bantu-speaking populations (~70% of the population). On the whole, the manner of dispersal of southern African Bantu-speakers is largely unknown and it has been previously hypothesized that there is limited population structure between individual groups. This hypothesis arose partly due to extensive population movements during the period of civil war in the early 1800’s known as the “Mfecane” (133). Later studies have attempted to identify and characterise the putative fine-scale population structure present among southern African

Bantu-speaking populations by estimating F_{st} values determined from autosomal serogenetic, DNA and Y chromosome haplotypes. Southern Bantu-speaking populations tend to cluster in accordance with their linguistic grouping, with the exception of the Tsonga who clustered closer to the Venda, perhaps due to their close geographic proximity (79). In general, genetic distances as well as linguistic groupings correlated with geographical distances (79). The differences between these southern Bantu-speaking populations were however, small.

Although there is a relative lack of identifiable structure among Bantu-speaking populations, the demographic model that governs their dispersal into southern Africa is convoluted. Utilizing mtDNA and Y chromosome data, it has been possible to distinguish between differing models so as to explain the admixture patterns we see in modern day southern African Bantu-speaking populations. There is evidence for gene flow between the migrating farmers and the indigenous foraging KhoeSan communities (represented by the Ju/'hoansi from Namibia) (134). This signal of admixture is pronounced in Bantu-speaking populations in South-East Africa, e.g. Mozambique. This finding was supported by large migration rate estimates of 1%-2% per generation for ~900 years from the KhoeSan to Bantu-speakers, as calculated based on simulated data (134). High proportions of KhoeSan ancestry are observed in the Zulu and Sotho populations (i.e. ~23%) (94). The Xhosa (from the Eastern Cape province in South Africa), who constitute a large proportion of South Africa's Bantu-speaking population, also derive 20% of their ancestry from the southern KhoeSan and the rest from East, West and Central Africa (54,55). The admixture from the KhoeSan into Bantu-speaking South African populations is similar to the European admixture proportion in African-American populations in the USA (135); African-Americans are a canonical "admixed" population in US biomedical research. Biomedical research approaches developed for African-Americans should be considered more often when studying South African Bantu-speaking groups (e.g. admixture mapping for diseases which differ in susceptibility between the two ancestries).

Although ancestry proportions are of historical interest, they are also relevant to disease risk profiling. For example, the SAC population have arguably the highest incidence of pulmonary tuberculosis in the absence of human immunodeficiency (HIV) immunosuppression while Bantu-speaking populations have the highest incidence of HIV/AIDS (136). Knowledge of the population structure and genetic ancestry can enable the mapping of possible susceptibility causing loci (i.e. via admixture mapping). This approach can be extended to other infectious diseases prevalent in southern Africa.

Genetic perspectives of infectious disease phenotypes in southern African populations

South Africa has the second highest incidence rate of TB in the world after Lesotho, a country surrounded by South Africa (WHO, 2015). The disease is rife in all South African populations, including the SAC. A genome-wide association study (GWAS) indicated that excess KhoeSan ancestry predisposes the SAC population to TB and this effect was not confounded by socio-economic status (40). Subsequently, Daya et al. (2014b) used Ancestry Informative Markers (AIMS) in a validation study with an independent sample set of 918 cases and 507 controls. Correcting for KhoeSan ancestry affected whether a polymorphism was still significantly associated with TB or not, dependent on the frequency of the SNP in parent populations (17,52). Further investigation not only indicated that KhoeSan and African non-KhoeSan ancestries are associated with an increased risk of progression to active pulmonary TB, but that European, South Asian and East Asian ancestry are protective against TB (17). Interestingly, intergenic class II human leukocyte antigen (*HLA*) variants were recently associated with both protection against and susceptibility to active TB disease in *M. tuberculosis*–infected individuals of European ancestry (137). It is not known whether these *HLA* class II variants are at appreciable frequencies in SAC populations. From admixture mapping and the resulting ancestry correlation tests, it was estimated that every 10% increase in KhoeSan ancestry in an individual correlated with a 38% increase in the odds of progressing to active pulmonary TB (18). The regions of excess KhoeSan ancestry in TB cases include the *GADD45A* and *OSM* genes (18), which makes them candidates for further investigation.

Although only one GWAS for TB has been performed in southern African groups, linkage and numerous candidate gene case-control association studies have identified SNP variants associated with TB susceptibility (25). As an example, the Major Histocompatibility Complex (MHC) and the Leukocyte Receptor Complex (LRC) have been implicated in altering the susceptibility to infectious diseases (138–141). After investigating the interaction between the Human Leukocyte Antigen (HLA) type of the TB patient and the infecting *M. tuberculosis* strain (33), they hypothesized that three vaccines currently in clinical trials may not be effective in the SAC population as predicted from common HLA allele class I profiles. This is due to the significantly lower frequency in the SAC and Bantu-speaking populations compared to Europeans, of the HLA subtype that is required for bacterial epitope recognition (142). In addition to the MHC and LRC, African genome-wide linkage studies have shown that loci in melanocortin-3-receptor (*MC3R*) and cathepsin Z (*CTSZ*) are linked to susceptibility to TB (143), and SNPs in these genes were validated in the SAC population in a case-control association study (144).

Other variants that have been associated with TB susceptibility are involved in immune pathways in which interferon-gamma (IFN-γ) plays a crucial role (145). An intronic variant which increased IFN-γ production (therefore affecting the host immune response) displays population specific levels of positive selection with a higher level of positive selection present in African populations (Yoruba, Mandenka and Bantu-speaking individuals from Kenya) (146). Populations used in this study originated from sub-Saharan Africa, Europe and East Asia, but no southern African populations were included. An association and transmission disequilibrium test (TDT) on SAC individuals noted that a promoter polymorphism in the gene for IFN-γ (*IFNG*) was significantly associated with an increased likelihood of progression to active TB (147). Various meta-analyses have identified other possible variants in *IFNG* that are associated with TB susceptibility once stratified based on ethnicity, but no studies have investigated this association in African populations (148). In addition to IFN-γ, tumour necrosis factor (TNF) is an important inflammatory meditator and its role in TB susceptibility is well documented, especially in sub-Saharan African populations (149). Meta-analyses have shown that the association between TNF-α and TB susceptibility is stratified by population (150,151). Within the context of Africa, the well characterized -308G>A polymorphism in TNF-α was associated with pulmonary TB in African populations under a dominant model but not in the Asian or Caucasian populations (151).

A linkage analysis of the quantitative Tuberculin Skin Test (TST) reaction to injected PPD was done on 128 SAC families including 350 siblings. This study determined that a single major locus on chromosomal region 11p14 that we called *TST1,* appears to control human resistance to the bacterium, as evidenced by the lack of delayed-type hypersensitivity (152). This work was the first report of a genetic resistance factor for TB infection as opposed to disease. In the same families, they detected a major pleiotropic locus on chromosome region 11p15, termed *TNF1,* that controlled TNF production after stimulation by both BCG alone and BCG plus IFN-γ. The close proximity of these loci suggested that there is a connection between TST negativity *per se* and TNF production (153).

Toll-like receptors (TLRs) are known to play a major role in an individual’s immune response to TB. Previous association studies have led to differing results, particularly within complex populations, e.g. in southern Africa, suggesting that population substructure may be masking signals of susceptibility. For this reason, meta-analyses have looked into the relationship between TLR’s and TB, taking into account different ethnic groups as well as increasing the power to detect any statistical associations. The major TLRs (1, 2, 4, 6, and 9) play a role in

TB susceptibility in most populations (154–156), but some variants were associated with TB susceptibility in one population but not in another. For example, the AG genotype of *TLR1* r4833095 and the T allele of *TLR6* rs5743810 were associated with resistance to TB across all ethnic groups studied (Asian, African, European and Hispanic), whereas variants located in *TLR4* and *TLR2* showed associations with TB susceptibility only in the Asian and Hispanic subpopulations (157). In contrast, variants in TLR2, TLR6 and TLR8 were found to be associated with TB susceptibility in the Chinese population (155). These variants in TLR8 (rs3764879, rs3761624, rs3788935 and rs3764880) were shown to be associated with TB susceptibility in the SAC population (40). Further studies investigating ethnicity as a confounding factor have been performed but very few include African populations.

When considering susceptibility to TB, specifically the recurrence of the disease, it is relevant to discuss the effect that genetic variation has on anti-TB drug metabolism. The CYP3A4 enzyme (a human cytochrome P450) is one of the most important enzymes involved in drug metabolism. Differing allele frequencies within *CYP3A4* were found in the KhoeSan, Xhosa and SAC populations. Of the 24 SNPs detected in *CYP3A4* in these populations, one was a functional promoter polymorphism (158). Looking at drug metabolism through the lens of genetic variation leads to the prediction that although some drugs might be effective in a few populations, they might be harmful or ineffective in others (158) and in the case of TB drug metabolism, differences in metabolism could lead to the development of drug resistant TB.

TB is also the leading cause of death in HIV infected patients (WHO, 2013). A recent study hypothesized that HIV positive individuals who live in a TB endemic area and do not develop active disease could provide insight into TB resistance (159). A GWAS was done in HIV positive individuals from Tanzania and Uganda and identified a locus at chromosome 5q33.3 which may offer protection against TB (159).

Sub-Saharan Africa has the highest HIV incidence rate in the world (WHO, 2013). As the effectiveness of antiretrovirals has reached the stage where normal life expectancy is possible, research is moving towards focusing on the cause of death for infected individuals, and identifying host factors contributing to HIV infection. Viral co-receptors CCR5 and CXCR4 are the most crucial and polymorphisms in either can confer protection or susceptibility to the virus (160–162). Specifically, *CCR5*Δ32 has been shown to confer protection to HIV infection in Europeans (163). Given the high HIV incidence in southern Africa it is noteworthy that there is a statistically significant difference in the activation and expression levels of CCR5 between two South African populations, namely “South African Africans” and “South African Caucasians” (164). This could result in altered susceptibility to HIV-1 infection as well as affect the progression of the infection itself. This study did not include ancestral information to determine the origin of the phenotypes under study, which may play a pivotal role in the identification of the variant associated with the phenotype.

Conclusion

Southern African populations are unique in their culture, history, languages and may be the cradle of humankind; this uniqueness is supported by population genetic analyses. Understanding the genetic structure of these populations is not only important to reconstruct human evolutionary history, but also has implications for the study of disease risk. Significant KhoeSan ancestry in many present-day southern African populations reflects how recent migration into this region resulted in the absorption of indigenous KhoeSan groups into many ethnicities. For example, population structure analyses of the SAC population in the Western Cape indicate substantial (>30%) ancestry from the KhoeSan who were present at the Cape during initial European colonization. It is also clear that there are ancestry-linked genetic factors contributing to infectious disease susceptibility in southern Africa, particularly with regard to TB. This may also be true for other infectious and chronic diseases not mentioned in this review, but due to lack of available studies, these warrant further investigation. Research into the link between ancestry and disease risk is sparse and the lack of publically available data is one reason for this. Collaborative networks, especially with respect to sample collection and analysis, would greatly facilitate these investigations. Although there has been some progress in encouraging genetic studies of populations displaying adverse phenotypes in Africa, more research needs to be based on southern African populations, especially in the fields of HIV and TB.

Chapter 3

Fine-scale human population structure reflects ecogeographic boundaries

Caitlin Uren[1], Minju Kim[2], Alicia R. Martin[3,4], Dean Bobo[2], Christopher R. Gignoux[5], Paul D. van Helden[1], Marlo Möller[1], Eileen G. Hoal[1,6§], Brenna M. Henn[2,6§]

[1] SA MRC Centre for TB Research, DST/NRF Centre of Excellence for Biomedical Tuberculosis Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, 8000
[2] Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794
[3] Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA 02114
[4] Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA 02142
[5] Department of Genetics, Stanford University, Stanford, CA 94305
[6] Co-senior authors

Genetics 2017 (Impact Factor 5.963)
doi:10.1534/genetics.116.187369

Abstract

Recent genetic studies have established that the KhoeSan populations of southern Africa are distinct from all other African populations and have remained largely isolated during human prehistory until about 2,000 years ago. Dozens of different KhoeSan groups exist, belonging to three different language families, but very little is known about their population history. We examine new genome-wide polymorphism data and whole mitochondrial genomes for more than one hundred South Africans from the ≠Khomani San and Nama populations of the Northern Cape, analyzed in conjunction with 19 additional southern African populations. Our analyses reveal fine-scale population structure in and around the Kalahari Desert. Surprisingly, this structure does not always correspond to linguistic or subsistence categories as previously suggested, but rather reflects the role of geographic barriers and the ecology of the greater Kalahari Basin. Regardless of subsistence strategy, the indigenous Khoe-speaking Nama pastoralists and the N|u-speaking ≠Khomani (formerly hunter-gatherers) share ancestry with other Khoe-speaking forager populations that form a rim around the Kalahari Desert. We reconstruct earlier migration patterns and estimate that the southern Kalahari populations were among the last to experience gene flow from Bantu-speakers, approximately 14 generations ago. We conclude that local adoption of pastoralism, at least by the Nama, appears to have been primarily a cultural process with limited genetic impact from eastern Africa.

Introduction

The indigenous populations of southern Africa, referred to by the compound ethnicity “KhoeSan” (165), have received intense scientific interest. This interest is due both to the practice of hunter-gatherer subsistence among many groups – historically and to present-day – and genetic evidence suggesting that the ancestors of the KhoeSan diverged early on from all other African populations (61,63–65,112,129,166). Genetic data from KhoeSan groups has been extremely limited until very recently, and the primary focus has been on reconstructing early population divergence. Demographic events during the Holocene and the ancestry of the Khoekhoe-speaking pastoralists have received limited, mostly descriptive, attention in human evolutionary genetics. However, inference of past population history depends strongly on understanding recent population events and cultural transitions.

The KhoeSan comprise a widely distributed set of populations throughout southern Africa speaking, at least historically, languages from one of three different linguistic families – all of which contain click consonants rarely found elsewhere. New genetic data indicates that there is deep population divergence even among KhoeSan groups (54,61,63,72,82,83,99,112,129), with populations living in the northern Kalahari estimated to have split from southern groups 30,000-35,000 years ago (63,64,72,83). Pickrell et al. (2012) estimate a time of divergence between the northwestern Kalahari and southeastern Kalahari population dating back to 30,000 years ago; "northwestern" refers to Juu-speaking groups like the !Xun and Ju/'hoansi while "southeastern" refers to Taa-speakers. In parallel, Schlebusch et al. (2012) also estimated an ancient time of divergence among the KhoeSan (dating back to 35,000 ya), but here the southern groups include the ≠Khomani, Nama, Karretjie (multiple language families) and the northern populations refer again to the !Xun and Ju/'hoansi. Thus, KhoeSan populations are not only strikingly isolated from other African populations but they appear geographically structured amongst themselves. To contrast this with Europeans, the ≠Khomani and the Ju/'hoansi may have diverged over 30,000 *ya* but live only 1,000 km apart, roughly the equivalent distance between Switzerland and Denmark whose populations have little genetic differentiation (Novembre et al. 2008). However, it is unclear how this ancient southern African divergence maps on to current linguistic and subsistence differences among populations, which may have emerged during the Holocene. In particular, the genetic ancestry of the Khoe-speaking populations and specifically the Khoekhoe, (e.g. Nama) who practice sheep, goat and cattle pastoralism, remains a major open question. Archaeological data has been convened to argue for a demic migration of the Khoe from eastern African into southern Africa, but others have also argued that pastoralism represents cultural diffusion without significant population movement (Boonzaier 1996; K. C. MacDonald 2000; Robbins *et al.* 2005; Sadr 2008, 2015; Dunne *et al.* 2012; Pleurdeau *et al.* 2012; Jerardino *et al.* 2014). Lactase persistence alleles are present in KhoeSan groups, especially frequent in the Nama (20%), and clearly derive from eastern African pastoralist populations (84,85). This observation, in conjunction with other Y-chromosome and autosomal data (71,125), has been used to argue that pastoralism in southern Africa was another classic example of demic diffusion. However, the previous work is problematic in that it tended to focus on single loci [MCM6/LCT, Y-chromosome] subject to drift or selection. Estimates of eastern African autosomal ancestry in the KhoeSan remain minimal (<10%) and the distribution of ancestry informative markers is dispersed between both pastoralist and hunter-gatherer populations. Here, we present a comprehensive study of recent population structure in southern Africa and clarify fine-scale structure beyond "northern" and "southern" geographic descriptors. We then specifically test whether the Khoe-speaking Nama pastoralists derive their ancestry from eastern Africa, the northeastern Kalahari Basin, or far southern Africa. Our results suggest that ecological features of southern Africa, broadly speaking, are better explanatory features than either language, clinal geography or subsistence on its own.

Materials and Methods

Sample collection and ethical approval

DNA samples from the Nama, ≠Khomani San and South African Coloured populations were collected with written informed consent and approval of the Human Research Ethics Committee of Stellenbosch University (N11/07/210), South Africa, and Stanford University (Protocol 13829), USA. Community level results were returned to the communities in 2015 prior to publication. A contract for this project was approved by WIMSA (ongoing).

Autosomal data and genotyping platforms

A) ~565,000 SNP Affymetrix Human Origins SNP array dataset from Pickrell et al. (63), Lazaridis et al. (167) and additional ≠Khomani San and Hadza individuals from our collections: 33 populations and 396 individuals

B) ~320,000 SNP array dataset from the intersection of HGDP (Illumina 650Y) (168), HapMap3 (joint Illumina Human 1M and Affymetrix SNP 6.0), Illumina OmniExpressPlus and OmniExpress SNP array platforms as well as the dataset from Petersen et al. (54): 21 populations and 852 individuals.

Population structure

ADMIXTURE (169) was used to estimate the ancestry proportions in a model-based approach. Iterations through various *k* values are necessary. The *k* value is an estimate of the number of original ancestral populations. Cross-validation (CV) was performed by ADMIXTURE and these values were plotted to acquire the *k* value that was the most stable. Depiction of the Q matrix was performed in R. Ten iterations were performed for each *k* value with ten random seeds. Iterations were grouped according to admixture patterns to identify the major and minor modes by pong (170). These Q matrixes from ADMIXTURE, as well as longitude and latitude coordinates for each population were adjusted to the required format for use in an R script supplied by Prof. Ryan Raaum to generate the surface maps (Figure 2).

EEMS analysis

EEMS (171) was run on the Affymetrix Human Origins dataset. Genetic dissimilarities were calculated using the bed2diffs script and EEMS was run using the runeems_snps version of the program. A grid is constructed so as to house all demes in the data provided. Each individual is assigned to a specific deme. Using a stepping stone model, migration rates between demes are calculated. Genetic dissimilarities are calculated fitting an "isolation by distance model". In order for the MCMC iterations to converge, the number of MCMC iterations, burn iterations and thin iterations were increased. The other parameters were optimized as per the manual's recommendations, i.e. diversity and migration parameters were adjusted so as to produce 20-30% acceptance rates. The PopGPlot R package was used to visualize the data.

Association between F_{st}, geography and language

A Mantel test (F_{st} and geographic distance) and a partial Mantel test (F_{st} and language, accounting for geographic distance) was performed using the vegan package in R. Geographic distances (in kilometers) between populations were calculated using the concept of great circle distances and the latitude and longitude values as tabulated in Table S1. Weir and Cockerham genetic distances (F_{st}) were calculated from allele frequencies estimated with vcftools (172). A Jaccard phonemic distance matrix was used as formulated in Creanza et al. (173). Populations included in the analysis were the Nama, ≠Khomani, East Taa, West Taa, Naro, G|ui, G||ana, Shua, Kua, !Xuun and Khwe.

mtDNA Network

We utilized Network (ver. 4.6, copy righted by Fluxus Technology Ltd.), for a median-joining phylogenetic network analysis in order to produce Figures 5 and S6. Network Publisher (ver. 2.0.0.1, copy righted by Fluxus Technology Ltd.) was then used to draw the phylogenetic relationships among individuals.

Supplemental Information

Supplemental Information includes Supplemental Experimental Procedures (File S1), 6 figures and 3 tables.

Results

To resolve fine-scale population structure and migration events in southern Africa, we generated genome-wide data from three South African populations. We genotyped ≠Khomani San (*n*=75), Nama (*n*=13) and South African Coloured (SAC) (*n*=25) individuals on the Illumina OmniExpress and OmniExpressPlus SNP array platforms. Sampling locations are listed in Table S1, in addition to language groupings and subsistence strategies. These data were merged with HapMap3 (joint Illumina Human1M and Affymetrix SNP 6.0) (174), HGDP (Illumina 650Y) data (168) and Petersen et al. Illumina HumanOmni1-Quad (54), resulting in an intersection of ~320k SNPs for 852 individuals from 21 populations. In addition, we used the Affymetrix Human Origins SNP Array generated as part of Pickrell et al. (63) and Lazaridis et al. (167), including *n*=9 ≠Khomani San individuals from our collection and encompassing over 396 individuals from 33 populations. Whole mitochondrial genomes were generated from off-target reads from exome- and Y chromosome-capture short read Illumina sequencing. Reads were mapped to GRCh37, which uses the revised Cambridge reference sequence (rCRS). Only individuals with greater than 7x haploid coverage were included in the analysis: ≠Khomani San (*n*=64) and Nama (*n*=31); haplogroup frequencies were corrected for pedigree structure (Table S2). In this study, we address population structure among southern African KhoeSan, the genetic affinity of the Khoe, and how pastoralism diffused into southern Africa.

Population Structure in Southern African KhoeSan Populations

We first tested whether southern African populations conform to an isolation-by-distance model, or whether there is strong heterogeneity among populations relative to geographic distance. Using twenty-two southern African populations (with 560k SNPs from Affymetrix HumanOrigins array), we implemented the spatially explicit program EEMS (171) to test for effective migration patterns across the region. We observe a higher effective migration rate (*m*) in the central Kalahari basin relative to a lower migration rate that forms a rim around the Kalahari Desert (**Figure 1**). A second resistance band stretches across northern Namibia, indicating higher gene flow above northern Namibia, Angola and southern Zambia. Differences in effective migration rates can result from differences in effective population sizes. For example, a larger effective population size can result in higher effective migration rates, relative to neighboring demes with smaller N_e's. The higher *m* in the central Kalahari Basin, relative to the rim, could result from either larger N_e relative to Kalahari rim populations or simply higher migration among groups in a similar ecological area.

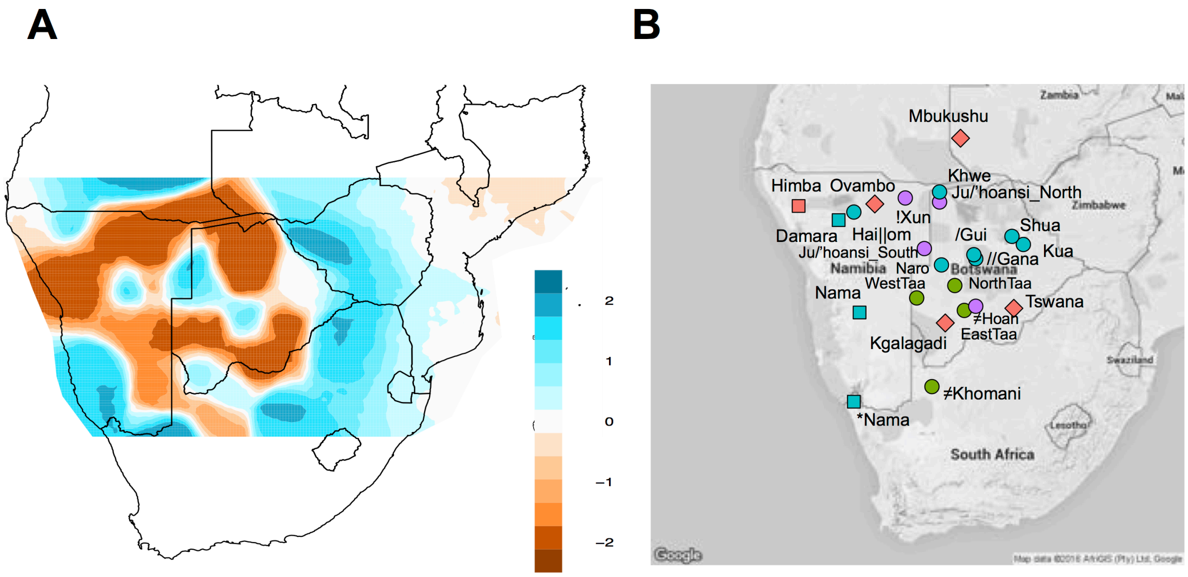


Figure 1: Effective migration rates among 22 southern African populations.

***A**) Using southern African samples from the Affymetrix HumanOrigins dataset, we estimated effective migration rates among populations using EEMS. White indicates the mean expected migration rate across the dataset, while blue indicates X-fold increase in migration among demes, and brown indicates decreased migration among demes (e.g. population structure). Effective migration rates, e_m, are plotted on a log-scale as in Petkova et al (2016). Hence, -1 e_m would indicate 10-fold decrease in the migration rate relative to the expected rate among all demes accounting for geographic distance. These results demonstrate that southern Africa is a heterogeneous environment with barriers to gene flow in northwest Namibia and the Kalahari rim, but increased gene flow within the Kalahari Basin. The grid of plotted demes was restricted to prevent unwanted extrapolation to poorly sampled areas. **B)** The topographic map indicates the subsistence strategy and language of each population sample. Colors represent language families: green= Tuu speakers, red= Niger-Congo speakers, blue= Khoe speakers and purple= Kx'a speakers. Shapes represent subsistence strategies: circle= hunter-gatherers, square= pastoralists and diamond= agropastoralists. *Nama indicates a new, second Nama sample from South Africa, which was only included in Illumina SNP array analyses.*

We then tested whether heterogeneity in population structure could be mapped to distinct genetic ancestries. Unsupervised population structure analysis identifies 5 distinct, spatially organized ancestries among the sampled twenty-two southern African populations. These ancestries were inferred from the Affymetrix Human Origins dataset using ADMIXTURE (Figure S1) (169,175). Multi-modality per *k* value was assessed using *pong* (170) and results from *k*=10 are discussed below (6/10 runs assigned to the major mode, 3/10 other runs involved cluster switching only within East Africa). Visualization of these ancestries according to geographic sampling location specifically demonstrates fine-scale structure in and around the Kalahari Desert (**Figure 2**). While prior studies have argued for a northern versus southern divergence of KhoeSan populations (63,72,83,91,129,130), the structure inferred from our dataset indicates a more geographically complex pattern of divergence and gene flow. Even recent migration events into southern Africa remain structured, consistent with ecological boundaries to gene flow (see below). The distribution of the five ancestries corresponds to: a northern Kalahari ancestry, central Kalahari ancestry, circum-Kalahari ancestry, a northwestern Namibian savannah ancestry and ancestry from eastern Bantu-speakers (**Figure 2**). This geographic patterning does not neatly correspond to linguistic or subsistence categories, in contrast to previous discussions (63,72,130).

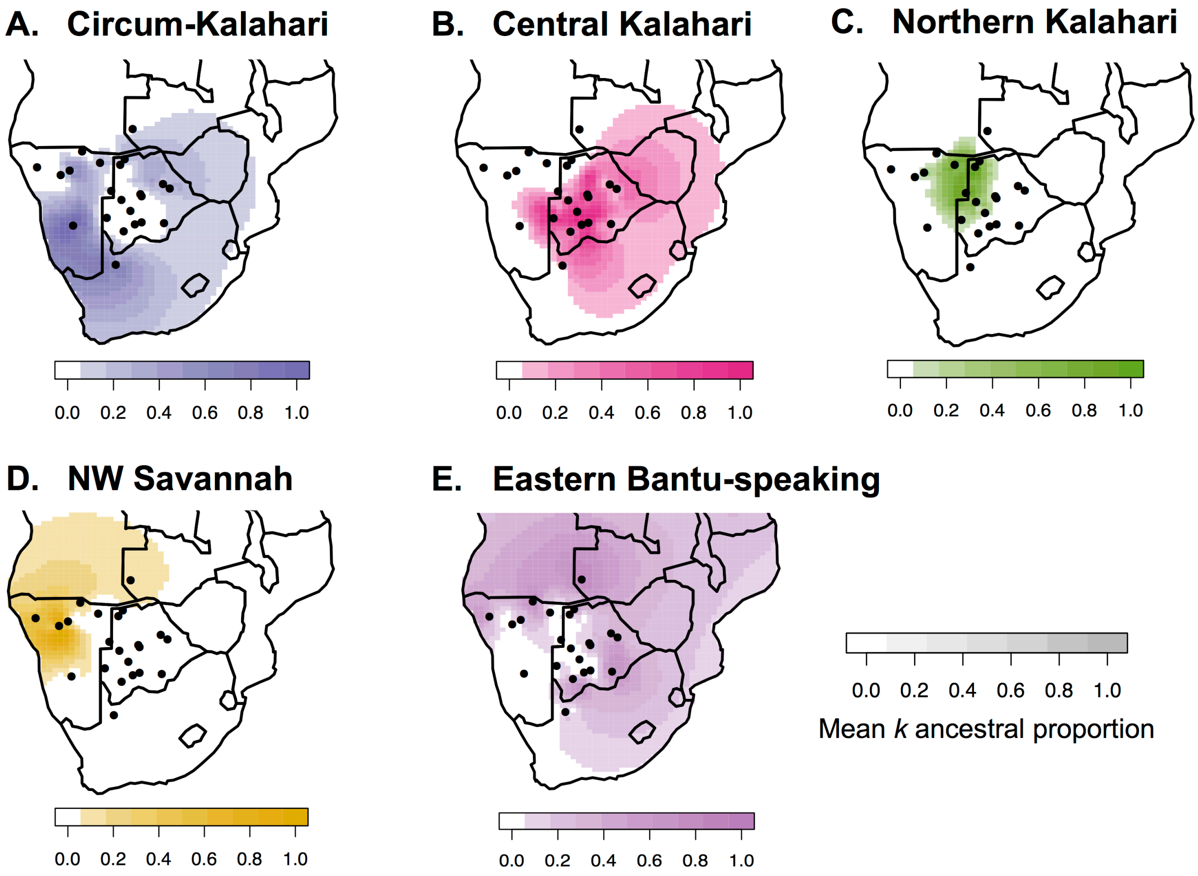


Figure 2: Five spatially distinct ancestries indicates deep population structure in southern Africa.

Using global ancestry proportions inferred from ADMIXTURE k=10, we plot the mean ancestry for each population in southern Africa. The 5 most common ancestries in southern Africa, from the Affymetrix HumanOrigins dataset, are shown separately in panels ***A)-E****). The X and Y axes for each map correspond to latitude and longitude, respectively. Black dots represent the sampling location of populations in southern Africa. The 3rd dimension in each map [depth of color] represents the mean ancestry proportion for each group for a given k ancestry, calculated from ADMIXTURE using unrelated individuals, and indicated in the color keys as 0% to 100% for five specific k ancestries. Surface plots of the ancestry proportions were interpolated across the African continent.*

The northern Kalahari ancestry is the most defined of these ancestries, encompassing several forager populations such as the Ju/'hoansi, !Xun, Khwe, Naro and to a lesser extent the Khoekhoe-speaking Hai||om. While these populations are among the best-studied KhoeSan in anthropological texts with particular reference to cultural similarities (176–179), they represent only a fraction of the diversity among Khoisan-speaking populations. We note that this cluster includes Kx'a (Juu), Khoe-Kwadi and Khoekhoe speakers, suggesting that language interacts in a complex fashion with other factors such as subsistence strategy and ecology. The Hai||om are thought to have shifted to speaking Khoekhoe from an ancestral Juu-based language (176). The second, central Kalahari ancestry occupies a larger geographical area throughout the Kalahari basin, with its highest frequency among the Taa-speakers: G|ui, G||ana, ≠Hoan and Naro. This ancestry spans all three Khoisan language families (Table S1), at considerable frequency in each; all are primarily foragers.

The third ancestry cluster is represented by southern KhoeSan populations distributed along the rim of the Kalahari Desert (**Figure 2**) – referred to here as the "circum-Kalahari ancestry". The circum-Kalahari ancestry is at its highest frequency in the Nama and ≠Khomani (see also Figure S2), with significant representation in the Hai||om, Khwe, !Xun and Shua. This ancestry spans all linguistic and subsistence strategies. We propose that the circum-Kalahari is better explained by ecology than alternative factors such as language or recent migration. Specifically, we find the Kalahari Desert is an ecological boundary to gene flow (**Figures 1,2**). The circum-Kalahari ancestry is not easily explained by a pastoralist Khoekhoe dispersal. This spatially distinct ancestry is common in both forager and pastoralist groups; indeed all of the circum-Kalahari populations were historically foragers (except for the Nama). Therefore, to support a Khoekhoe dispersal model, we would have to posit an adoption of pastoralism by a northeastern group, leading to demic expansion around the Kalahari, with subsequent reversion to foraging in the majority of the circum-Kalahari groups; this scenario seems unlikely (but see (123).

Finally, our analysis reveals two additional ancestries outside of the greater Kalahari Basin: one ancestry composed of Bantu-speakers, frequent to the north, east and southeast of the Kalahari; and a second composed of Himba, Ovambo, and Damara ancestry in northwestern Namibia distributed throughout the mopane savannah. Interestingly, the Damara are a Khoekhoe-speaking population of former foragers (later in servitude to the Nama pastoralists) whose ancestry has been unclear (*see below*).

We used our data and the Affymetrix HumanOrigins dataset containing the greatest number of KhoeSan populations to date, to test whether language or geography better explains genetic distance (see language families and subsistence strategies in Table S1). The genetic data were compared to a phonemic distance matrix (180) as well as geographic distances between each population (Table S3). In order to test whether genetic distance (F_{st}) was associated with geography or language, we performed a partial Mantel test for the relationship between F_{st} and language (173) accounting for geographic distance among 11 KhoeSan populations. This result was not significant (*r*=0.06, *p*=0.30). Although an association between F_{st} and geographical distance within Africa has been documented (64,173,181), a Mantel test for the relationship between F_{st} and pairwise geographic distance in our dataset was also null (*r*=0.021, *p*=0.38) reflecting the non-linear aspect of shared ancestry in southern Africa as seen in **Figures 1** and **2**.

Spatially distinct ancestries are also supported by principal components analysis (PCA) (**Figure 3 and S3**). The KhoeSan anchor one end of PC1 opposite to Eurasians. PC2 separates other African populations from the KhoeSan, including western Africans, as well as central and eastern African hunter-gatherers. PC3 separates the Ju/'hoansi and !Xun (northern Kalahari) from ≠Hoan, Taa-speakers and Khoe-speakers, with other KhoeSan populations intermediate. PC3 and PC4 suggest that the present language distribution may reflect recent language transitions, as genetic ancestry and linguistic structure do not neatly map onto each other (**Figure S4)**. For example, the ≠Hoan currently speak at Kx'a language but are genetically distinct from other northern Kalahari Kx'a speakers; rather, they appear to be more genetically similar to southern Kalahari Taa-speakers who cluster together. We suggest that the patterns observed here are better explained by ecogeographic patterns than either language or subsistence alone (**Figure S5**). Specifically, PC3 discriminates northern versus southern Kalahari ancestry (see below). PC4 discriminates western and eastern non-KhoeSan ancestry derived from Bantu-speakers or other populations. Finally, the intermediate position of the Nama, ≠Khomani and Hai||om on PC3 and PC4 is neither linguistic- nor subsistence-based, but represents a non-linear circum-Kalahari component featured in **Figure 2**.

A Divergent Southern KhoeSan Ancestry

This separation of northern (Ju/'hoansi) and southern (Taa and Khoe speakers) KhoeSan populations has been observed by Schlebusch et al. (72) and Pickrell et al. (63). We estimate that this trans-Kalahari genetic differentiation from the inferred ancestral allele frequencies (Figure S2) is substantial (F_{ST} = 0.05). We verify this divergence between the northern Kx'a-speakers and the shared Nama and ≠Khomani ancestry in a new, second sample of Nama, from South Africa rather than central Namibia (Table S1, Figure S3). This southern KhoeSan ancestry is also present in admixed Bantu-speaking populations from South Africa (e.g. amaXhosa) as well as the admixed Western Cape SAC populations (53), supporting a hypothesis of distinct *southern*-specific KhoeSan ancestry (Figure S1, S2) shared between indigenous and admixed groups.

Mitochondrial data support this concept of a *southern*-specific KhoeSan ancestry (82,129). Both mtDNA haplogroups L0d and L0k are at high frequency in northern KhoeSan populations (166), but L0k is absent in our sample of the Nama [*n*=31] and there is only one ≠Khomani individual [n=64] with L0k (1.56%) (**Table 1**). L0d dominates the haplogroup distribution for both the Nama and ≠Khomani (84% and 91% respectively), with L0d2a especially common in both. L0d2a, inferred to have originated in southern Africa, was also previously found at high frequencies in the Karretjie people further south in the central Karoo of South Africa, as well as the SAC population in the Western Cape (82,99). L0d2b is also common in the Nama (16%).

Minimal Population Structure Between the Nama and ≠Khomani

The ≠Khomani San are a N|u-speaking (!Ui classified language) former hunter-gatherer population that inhabit the southern Kalahari Desert in South Africa, bordering on Botswana and Namibia. The Nama, currently a primarily caprid pastoralist population, live in the Richtersveld along the northwestern coast of South Africa and up into Namibia. The ancestral geographic origin of the Nama has been widely contested over a number of years (86,182,183), but a leading hypothesis suggests that they originated further north in Botswana/Zambia and migrated into South Africa and Namibia approximately 2,000 years ago (63,86,182,183). The Nama and N|u languages are in distinct, separate Khoisan language families (Khoe and Tuu [!Ui-Taa], respectively) and these groups historically utilized different subsistence strategies. For this reason, we hypothesized that there would be strong population structure between the two populations.

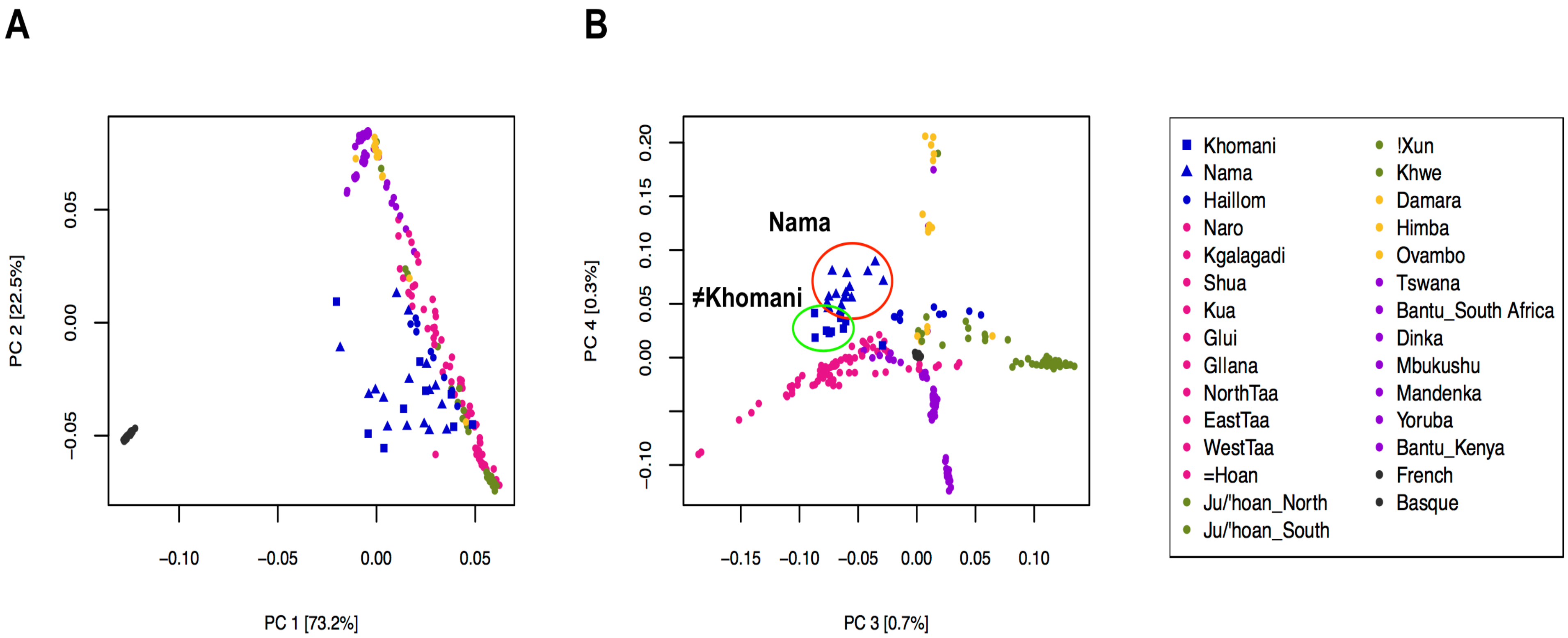


Figure 3: *Clustering of KhoeSan populations and fine-scale population structure between the Nama and ≠Khomani San*.

A PCA of the Affymetrix Human Origins dataset depicts the clustering of unrelated individuals based on the variation seen in the dataset. Colors mimic similar major ancestry colors as shown in Figure 2. Yellow denotes populations with majority northwestern Namibian ancestry; purple denotes populations with majority Bantu-speaking ancestry; pink indicates southern Kalahari majority ancestry, green indicates northern Kalahari majority ancestry and blue indicates circum-Kalahari ancestry. The red and green circles denote the fine-scale separation of the Nama and ≠Khomani populations [specified by triangles and squares, respectively]. Note that these colored ancestries and the PCs do not map onto subsistence neatly (Figure S5).

Table 1: Mitochondrial DNA haplogroup frequencies of the Nama and ≠Khomani.

		≠Khomani San			**Nama**		
Haplogroup		**n**	**Frequency**		**n**	**Frequency**	
L0d	L0d1a	8	12.50%	90.63%	3	9.68%	83.87%
	L0d1a1	2	3.13%		3	9.68%	
	L0d1b	4	6.25%		3	9.68%	
	L0d1b1	9	14.06%		2	6.45%	
	L0d1b2	2	3.13%		0		
	L0d1c1	1	1.56%		1	3.23%	
	L0d1c1a	1	1.56%		1	3.23%	
	L0d2a	25	39.06%		3	9.68%	
	L0d2a1	0			2	6.45%	
	L0d2b	0			5	16.13%	
	L0d2c	5	7.81%		2	6.45%	
	L0d3	1	1.56%		1	3.23%	
L0f1		1	1.56%		0		
L0k1		1	1.56%		0		
L3'4		0			1	3.23%	
L3d3a1a		0			1	3.23%	
L3e1a2		1	1.56%		0		
L4b2a2		1	1.56%		1	3.23%	
L5c		0			1	3.23%	
M36	M36	1	1.56%	3.13%	0		
	M36d1	1	1.56%		0		
M7c3c		0			1	3.23%	
Total (n)		64	100%		31	100%	

Our global ancestry results, inferred from ADMIXTURE, show minimal population structure between the Nama and ≠Khomani San in terms of their southern KhoeSan ancestry. The ≠Khomani share ~10% of their ancestry with the Botswana KhoeSan populations (Figure S1, S3), consistent with their closer proximity to the southern Botswana populations (Taa-speakers !Xo and ≠Hoan). Principal components analysis reveals a degree of fine-scale population structure between the Nama and ≠Khomani, with each population forming its own distinct cluster at PC4, partly due to the increase in Damara ancestry in the Nama (**Figure 3b**, Figure S1), but the two groups are clearly proximal. This increase in Damara ancestry (as depicted from *k*=9 in all modes of Figure S1) is likely due to integration of the Damara people as clients of the Nama over multiple generations. However, our second sample of Nama from South Africa do not harbor significant western African ancestry, suggesting heterogeneity in the Damara component (Figure S2).

Recent Patterns of Admixture in South Africa

Two Bantu-speaking, spatially distinct ancestries are present in southern Africa. The first is rooted in the Ovambo and Himba in northwestern Namibia; the other reflects gene flow from Bantu-speaking ancestry present in the east (**Figure 2**). We estimated the time intervals for admixture events into the southern KhoeSan via analysis of the distribution of local ancestry segments using RFMix (184) and TRACTs (185) for the ≠Khomani OmniExpress dataset (n=59 unrelated individuals) **(Figure 4, Table S2)**. The highest likelihood model suggests that there were 3 gene flow events. Approximately 14 generations ago (~443-473 years ago assuming a generation time of 30 years and accounting for the age of our sampled individuals), the ≠Khomani population received gene flow from a Bantu-speaking group, represented here by the Kenyan Luhya. Our results are consistent with Pickrell et al. (2012) who found that the southern Kalahari Taa-speakers were the last to interact with the expanding Bantu-speakers about 10-15 generations ago. Subsequently, this event was followed by admixture with Europeans between 6 and 7 generations ago (~233-263 years ago), after the arrival of the Dutch in the Cape and the resulting migrations of "trekboers" (nomadic pastoralists of Dutch, French and German descent) from the Cape into the South African interior. Lastly, we find a recent pulse of primarily KhoeSan ancestry 4-5 generations ago (~173-203 years ago). This event could be explained by gene flow into the ≠Khomani from another KhoeSan group, potentially as groups shifted local ranges in response to the expansion of European farmers in the Northern Cape, or other population movements in southern Namibia or Botswana.

We also considered the impact of recent immigration into indigenous South Africans, derived from non-African source populations. The South African Coloured (SAC) populations are a five way admixed population, deriving ancestries from Europe, eastern African, KhoeSan, and Asian populations (53). This unique, admixed ethnic population was founded by the Dutch who settled on the southern tip of South Africa by the 17th century and the importation of slaves from Indonesia, Bengal, India and Madagascar. However, within the SAC, strong differences in ancestry and admixture proportions are observed between different districts within Cape Town, the Eastern Cape and the Northern Cape Provinces. South African Coloured individuals from the Northern Cape, where historically there was a greater concentration of European settlement (186), have higher European ancestry. The SAC individuals from the Eastern Cape, which is the homeland of the Bantu-speaking Xhosa populations, have relatively more ancestry from Bantu-speaking populations (Figure S2). The "ColouredD6" population is from an area in Cape Town called District 6. Historically, this was a district where the slaves and political exiles from present day Indonesia resided, and many who were from Madagascar and India based on written documentation (187). The SAC D6 population consequently has a noticeable increase in south/eastern Asian ancestry represented by the Pathan and Han Chinese populations in our dataset (Figure S2).

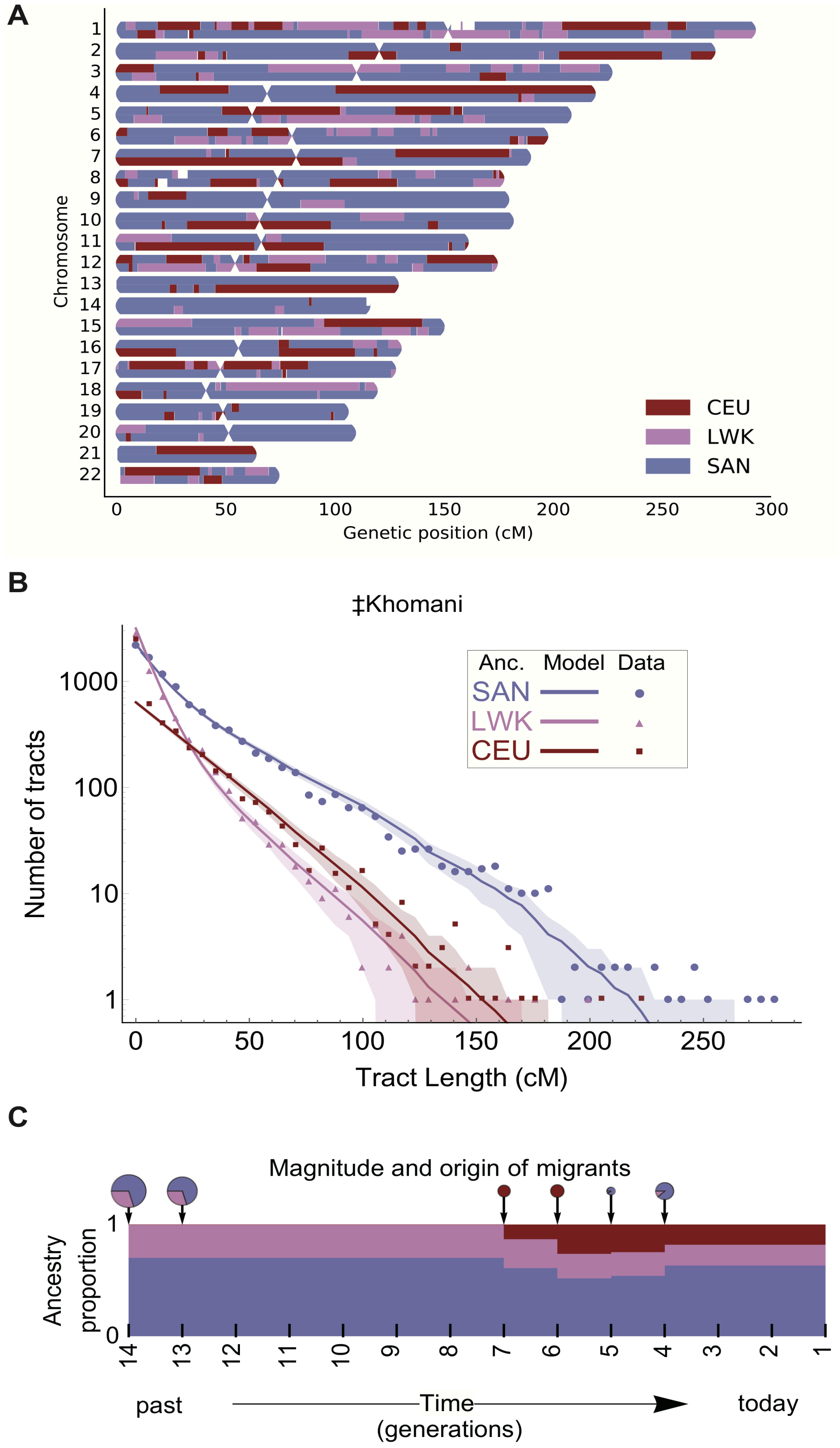


Figure 4*: Demographic reconstruction of recent admixture in the ≠Khomani San using local ancestry*.

A) A local ancestry karyogram for a representative 3-way admixed ≠Khomani San individual was constructed using RFMix. Haplotypes for admixed individuals were assigned to one of three possible ancestries: SAN [Namibian San], LWK [Bantu-speaking Luhya from Kenya], CEU [Central Europeans]. UNK indicates "unknown" ancestry (Methods). B) We employed Markov models implemented in TRACTs to test multiple demographic models and assess the best fit to the observed ≠Khomani haplotype distributions. Local ancestry tract lengths were inferred as in panel A). The tract length distribution for each ancestry across all individuals was used to estimate migration time (generations ago), volume of migrants, and ancestry proportions over time. Colored dots show the observed distribution of ancestry tracts for each ancestry, solid lines show the best fit from the most likely model, and shaded areas indicate confidence intervals corresponding to ± 1 standard deviation.

This south/eastern Asian ancestry is not confined to the SAC population, as attested by the presence of the M36 mitochondrial haplogroup. The M36 haplogroup (South Indian/Dravidian in origin) is present in two out of 64 ≠Khomani San matrilineages, (**Table 1**). The presence of M36 is likely derived from slaves of South Asian origin who escaped from Cape Town or the surrounding farms and dispersed into the northwestern region of South Africa. In addition, we observe one M7c3c lineage in the Nama (**Table 1**), which traces back to southeastern Asia but has been implicated in the Austronesian expansion of Polynesian speakers into Oceania (188,189) and Madagascar (190). The importation of Malagasy slaves to Cape Town may best explain the observation of M7c3c in the Nama.

Discussion

The KhoeSan are distinguished by their unique phenotype(s) (such as skin pigmentation (72,191) and height (192)), genetic divergence, click languages and hunter-gatherer subsistence strategy compared to other African populations; classifications of the many KhoeSan ethnic groups have primarily relied on language or subsistence strategy. Here, we generate additional genome-wide data from 3 South African populations and explore patterns of fine-scale population structure among 22 southern African groups. We find that complex geographic or "ecological" information is likely a better explanatory variable for genetic ancestry than language or subsistence. We identify 5 primary ancestries in southern Africans, each localized to a specific geographic region (**Figure 2**). In particular, we examined the "circum-Kalahari" which appears as a ring around the Kalahari Desert and accounts for the primary ancestry of the Nama, representative of the Khoekhoe-speaking pastoralists.

We observe striking ecogeographic population structure associated with the Kalahari Desert. There are two distinct ancestries segregating within the Kalahari Desert KhoeSan populations, described here as northern Kalahari and central Kalahari ancestries. Analyses of migration rates across the 22 populations indicate particularly high migration within the Kalahari Desert. This may indicate a larger effective population size for the two desert ancestries or extensive migration related to shifting ranges in response to climatic and ecological changes over time. It is worth noting that the northern Kalahari formerly supported an extensive lake (i.e. Makgadikgadi) just before and after the Last Glacial Maximum, as well as the presence of the Okavango Delta and associated river systems; archeological data may suggest high population density nearby the pans, although this likely predates the genetic structure we observe today (193,194). Our lack of samples outside of Botswana, Namibia and northern South Africa prevent precise inference of *m* in Zambia, Limpopo, and Mozambique; but **Figure 2** indicates recent extensive gene flow in the east, consistent with the expansion of Bantu-speaking agriculturalists into eastern grasslands and coastal forests. Additionally, we find a separate ancestry segregating in the far western border of Namibia and Angola, particularly frequent in the Damara and Himba, and to a lesser extent in the Ovambo and Mbukushu. This intersection of steppe and savannah along the Kunene may have facilitated recent settlement of the area during the past 500 years by Bantu-speaking pastoralists, but it is noteworthy that little Kalahari KhoeSan ancestry persists in these populations. Rather, the Damara (currently Nama-speaking) or related hunter-gatherers may have been formerly more widespread in this area and subsequently absorbed into the western Bantu-speaking pastoralists.

The practice of sheep, goat and cattle pastoralism in Africa is widespread. Within KhoeSan populations, pastoralist communities are limited to the Khoekhoe-speaking populations. Earlier hypotheses proposed that the Khoe-speaking pastoralists derived from a population originating outside of southern Africa. However, more recent genetic work supports a model of autochthonous Khoe ancestry influenced by either demic or cultural diffusion of pastoralism from East Africa ~2,500 years ago (89,125). For example, the presence of lactase persistence alleles in southern Africa indicates contact between East African herders and populations in south-central Africa, with subsequent migration into Namibia (84). This scenario is also supported by Y-chromosomal analysis that indicates a direct interaction between eastern African populations and southern African populations approximately 2,000 years ago (71). However, in both cases (i.e. MCM6/LCT and Y-chr M293), the frequency of the eastern African alleles is low in southern Africa and occurs in both pastoralist and hunter-gatherer populations. A simple model of eastern African demic diffusion into south-central

Africa, leading to the adoption of pastoralism and a Khoekhoe population expansion from this area cannot be inferred from the genetic data.

Our samples from the Khoekhoe-speaking Nama pastoralists demonstrate that their *primary* ancestry is shared with other far southern non-pastoralist KhoeSan such as the ≠Khomani San and the Karretjie (195). mtDNA also suggests that the Nama display a haplogroup frequency distribution more similar to KhoeSan south of the Kalahari than to any other population in south-central Africa. Our results indicate that the majority of the Nama ancestry has likely been present in far southern Africa for longer than previously assumed, *rather* than resulting from a recent migration from further north in Botswana where other Khoe-speakers live. The only other Khoekhoe-speaking population in our dataset is the Hai||om who share approximately 50% of the circum-Kalahari ancestry with the Nama and ≠Khomani, but are foragers rather than pastoralists. We conclude that Khoekhoe-speaking populations share a circum-Kalahari genetic ancestry with a variety of other Khoe-speaking forager populations in addition to the !Xun, Karretjie and ≠Khomani (**Figures 1** and **2**). This ancestry is divergent from central and northern Kalahari ancestries, arguing *against* a major demic expansion of Khoekhoe pastoralists from northern Botswana into South Africa. Rather, in this region, cultural transfer likely played a more important role in the diffusion of pastoralism. Of course, a demic expansion of the Khoekhoe *within* a more limited region of Namibia and South Africa may still be have occurred – but geneticists currently lack representative DNA samples from many of the now “Coloured” interior populations which may carry Khoekhoe ancestry.

This is an unusual case of cultural transmission (88). Other prehistoric economic transitions have been shown to be largely driven by demic diffusion (167,196–199). Recent analysis of Europe provides a case study of demic diffusion, which appears far more complex than initially hypothesized. The initial spread of Near Eastern agriculturalists into southern Europe clearly replaced or integrated many of the autochthonous hunter-gatherer communities. Even isolated populations such as the Basque have been shown to derive much of their ancestry from Near Eastern agriculturalists (198). The early demic diffusion of agriculture exhibits a strong south to north cline across Europe, reflecting the integration of hunter-gatherers into composite southern agriculturalist populations which then expanded northward with mixed ancestry (200). The cline of the early Near Eastern Neolithic ancestry becomes progressively diluted in far northern European populations. In contrast, we see little evidence of a clear eastern African ancestry cline within southern African KhoeSan; nor is the putative “Khoe” ancestry identified in the Nama of eastern African origin or even of clear origin from northeastern Botswana where initial pastoralist contact presumably occurred.

However, the transfer of pastoralism from eastern to southern Africa itself was not purely cultural (see above). We also report here the presence of mitochondrial L4b2 that supports limited gene flow from eastern Africa, approximately during the same timeframe as the pastoralist diffusion. L4b2, formerly known as L3g or L4g, is a mtDNA haplogroup historically found at a high frequency in eastern Africa, in addition to the Arabian Peninsula. L4b2 is at high frequency specifically in click-speaking populations such as the Hadza and Sandawe in Tanzania (sometimes described as 'Khoisan-speaking') (105). Nearly 60% of the Hadza population and 48% of Sandawe belong to L4b2 (120). Even though both Tanzanian click-speaking groups and the southern African KhoeSan share some linguistic similarities and a hunter-gatherer lifestyle, they have been isolated from each other over the past 35*ky* (120). The L4b2a2 haplogroup is present at a low frequency in both the Nama and ≠Khomani San, observed in one matriline in each population (**Table 1**). L4b2 was also formerly reported in the SAC population (0.89%) (99) but has not been discussed in the literature. We identified several additional southern L4b2 haplotypes from whole mtDNA genomes deposited in public databases (129,166) and analyzed these samples together with all L4b2 individuals available in NCBI. Median-joining phylogenetic network analysis of the mtDNA haplogroup, L4b2, supports the hypothesis that there was gene flow from eastern Africans to southern African KhoeSan groups. As shown in **Figure 5** (and in more detail in Figure S6), southern African individuals branch off in a single lineage from eastern African populations in this network (120,201,202). The mitochondrial network suggests a recent migratory scenario (estimated to be < 5,000 years before present), though the source of this gene flow, whether from eastern African click-speaking groups or others, remains unclear (125).

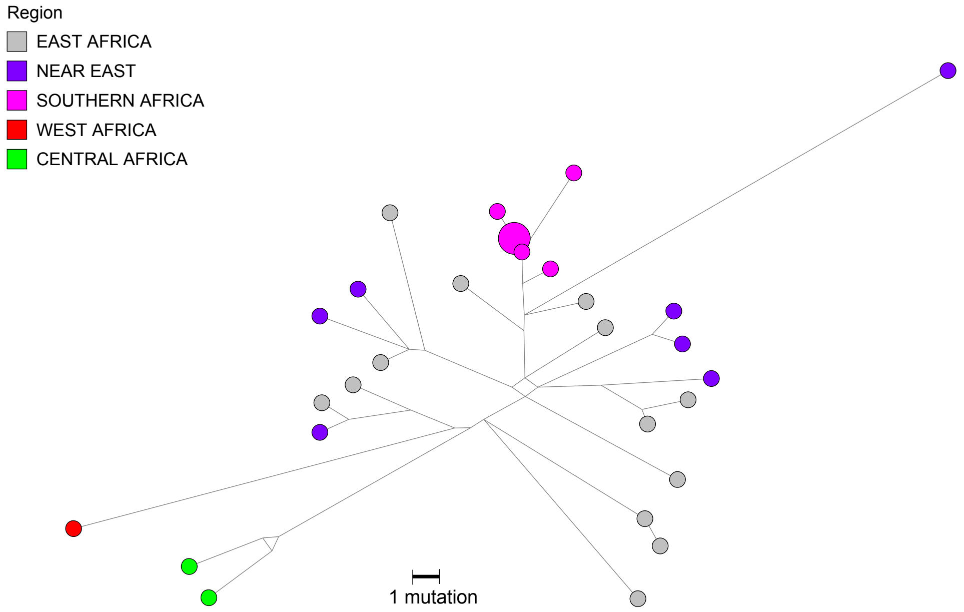


Figure 5: *L4b2 mtDNA haplogroup network.*

New L4b2 mitochondrial genomes from ≠Khomani and Nama individuals, indicated in pink as "Southern Africa", were analyzed together with publically available L4b2 mtDNA genomes from NCBI (as outlined in the Supplementary Methods). All individuals were assigned to mtDNA haplogroups using haplogrep and the haplotypes were plotted using Network Publisher.

Conclusion

Analysis of 22 southern African populations reveals that fine-scale population structure corresponds better with ecological rather than linguistic or subsistence categories. The Nama pastoralists are autochthonous to far southwestern Africa, rather than representing a recent population movement from further north. We find that the KhoeSan ancestry remains highly structured across southern Africa and suggests that cultural diffusion likely played the key role in adoption of pastoralism.

Supplemental Methods

Population structure:

chromoPainter (203) takes as input SNP data from a pre-defined recipient and donor populations as well as a genetic recombination map. The program 'paints' each recipient individual on the basis of every other individual in the dataset. fineSTRUCTURE (203) places individuals into populations based on a model for "expected variability". Software was freely available at www.paintmychromosomes.com.

Principle components analysis (PCA) was performed in R and the PC loadings were calculated from the '.chunkcounts.out' file generated from chromoPainter. These were mean transformed and plotted in the R programming environment. Three different PCA's were plotted. Figure 3 was colour and shape coded according to the majority ancestry in Figure 2. Populations in Figure S5 were plotted as different shapes according to their subsistence strategy. The language family of every population is used to colour population present in the PCA in Figure S6.

Local Ancestry Assignment and TRACTs:

We merged all ≠Khomani individuals genotyped on the OmniExpress and OmniExpressPlus arrays, the Schuster et al., (204) Namibian genotypes, along with CEU and LWK individuals genotyped in 1000 Genomes. As reference panels, we defined separate classes for European, Bantu, and KhoeSan ancestries respectively using CEU, LWK, and ≠Khomani and Schuster et al., (204) individuals with >90% KhoeSan ancestry as inferred via ADMIXTURE. We phased individuals using SHAPEIT2 with the 1000 Genomes phase 3 as a reference panel. We inferred local ancestry using RFMix (184) with a node size of 5 to reduce bias resulting from unbalanced reference panels, a minimum window size of 0.2 cM, and 1 EM iteration to better inform the small amount of admixture in the KhoeSan reference samples. We assessed the fit of 7 different models in TRACTs (185), including several two-pulse and three-pulse models. Ordering the populations as KhoeSan, Bantu, and European, we tested the following models: ppp_ppp, ppp_pxp, ppp_xxp, ppx_xxp, ppx_xxp_ppx, ppx_xxp_pxx, and ppx_xxp_xxp, where the order of each letter corresponds with the order of population given above, an underscore indicates a distinct migration event with the first event corresponding with the most generations before present, p corresponding with a pulse of the ordered ancestries, and x corresponding with no input from the ordered ancestries. We tested all 7 models preliminarily 3 times, and for all models that converged and were within the top 3 models, we subsequently fit each model with 100 starting parameters randomizations. The log-likelihood of the best fit model was -342, which provided a substantially better fit than all other models tested (next best model achieved best log-likelihood = -402).

mtDNA haplogroup frequency and networks:

Haplogroup frequency:

Coverage per individual was set at a minimum of 6.5x, therefore only 80 out of the 91 ≠Khomani and 36 Nama were used for further analysis (Table 1). To prevent oversampling of the same haplogroup in families, only one individual per matrilineage was included (Table S2). These individuals were then grouped with other publically available data. Haplotypes were assigned to haplogroups using *haplogrep* (205).

mtDNA Network:

We utilized Network (ver. 4.6, copy righted by Fluxus Technology Ltd.), for a median-joining phylogenetic network analysis in order to produce Figures 4 and S4. Network Publisher (ver. 2.0.0.1, copy righted by Fluxus Technology Ltd.) was then used to draw the phylogenetic relationships among individuals.

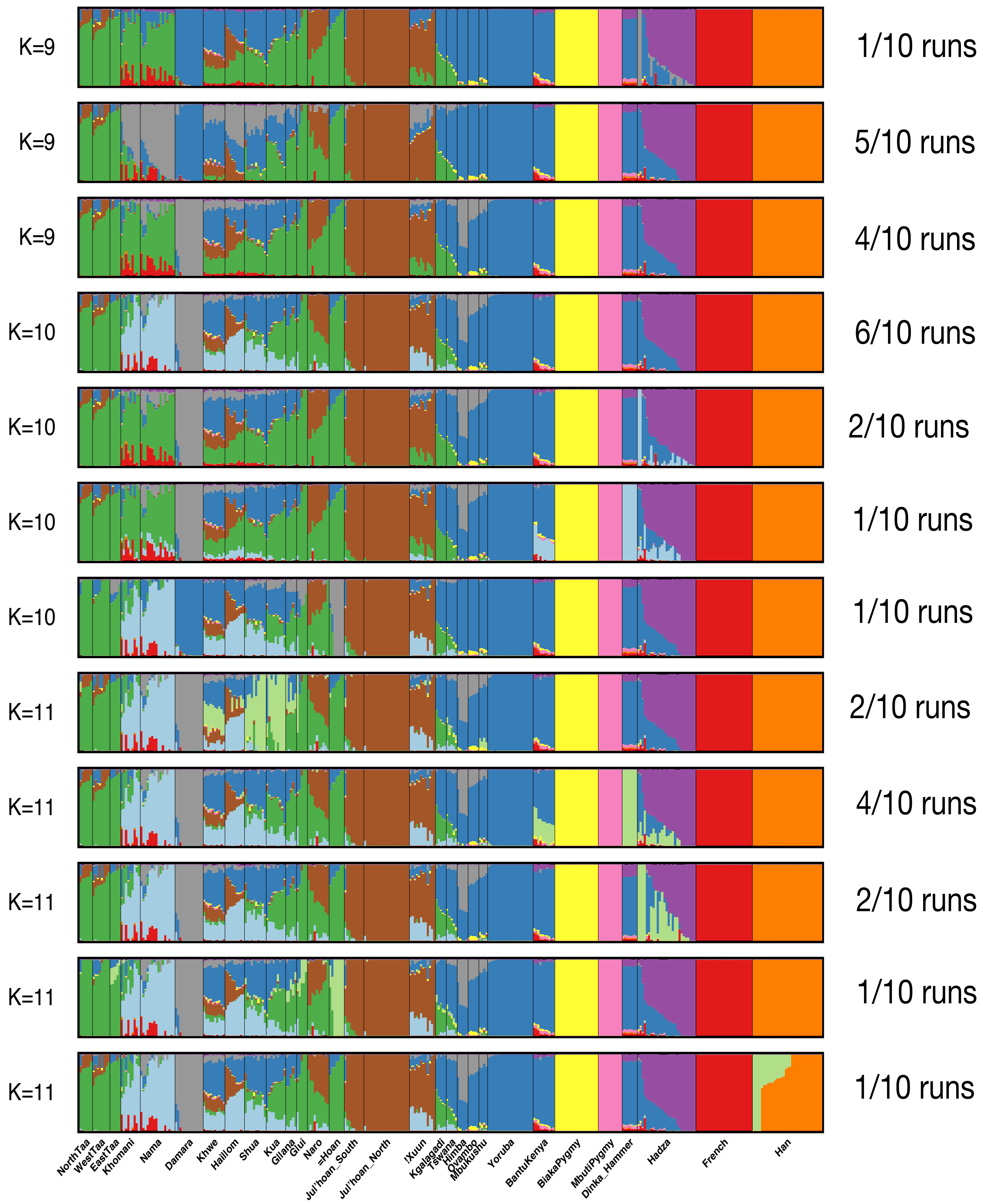


Figure S1: *Population structure in southern Africa and further evidence for a southern African specific KhoeSan ancestry.*

These diagrams display the ancestral contributions as ascertain by an unsupervised ADMIXTURE analysis. Ancestral proportions are shown as varying degrees of each color i.e. each ancestry. This is displayed for a large number of KhoeSan populations in the Affymetrix Human Origins dataset. Every hypothesis of the number of ancestral populations is taken into account (k values). As seen here due to the hypothesis of structure, multiple k values were used. Every run utilized a different random seed and thus it was necessary to pool similar results as shown, by the use of pong.

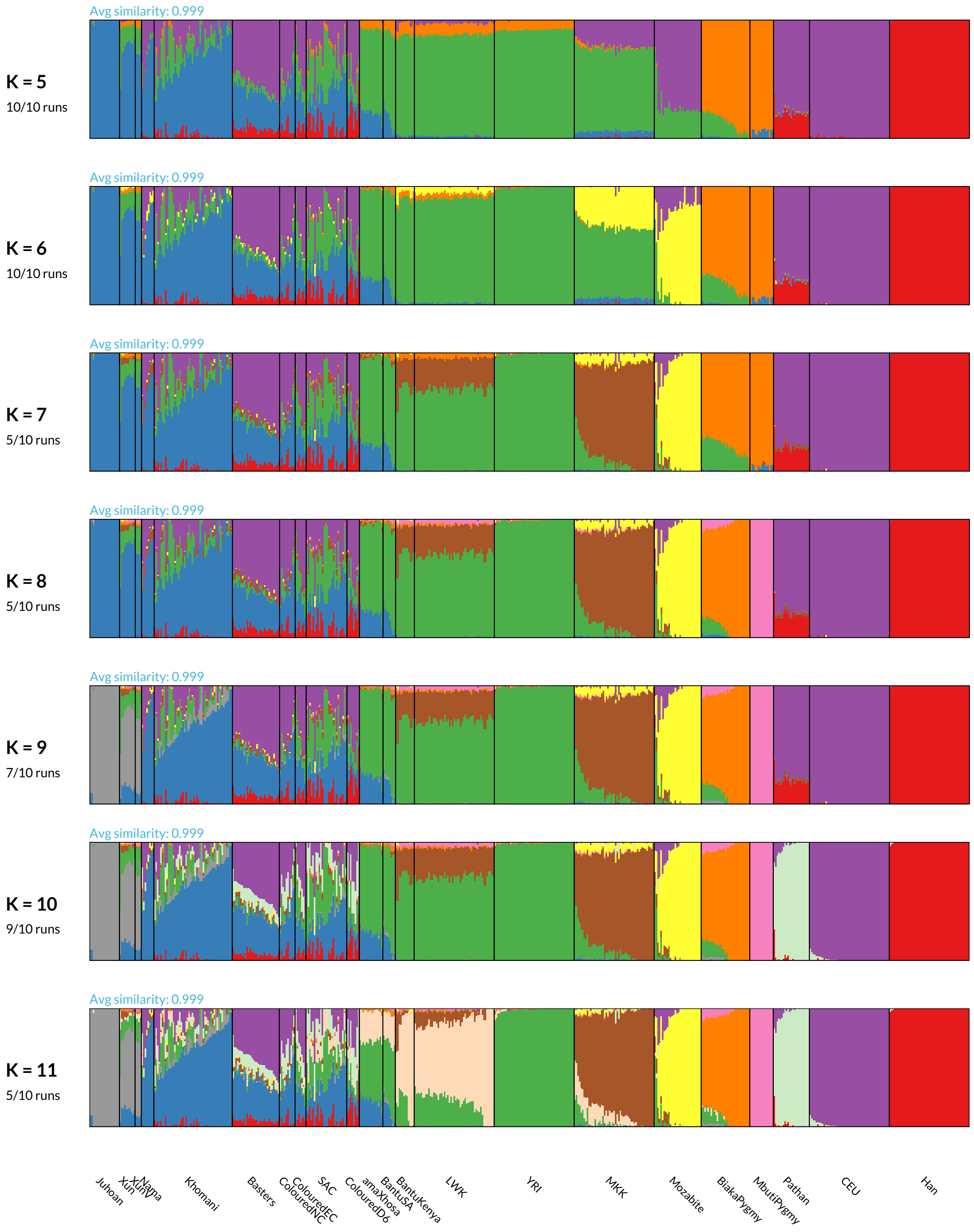


Figure S2: *Population structure in southern Africa and further evidence for a southern African specific KhoeSan ancestry, utilizing more South African specific populations.*

ADMIXTURE plots as generated from an unsupervised analysis of the 340k merged dataset. Each color represents a specific ancestry and every hypothesis of the number of ancestral populations are taken into account (k values). Multi-modularity was assessed using pong as in Figure S1, however only the major modes are displayed here. Each run utilized a different random seed and thus there were differing results. These results were grouped according to similarity using pong.

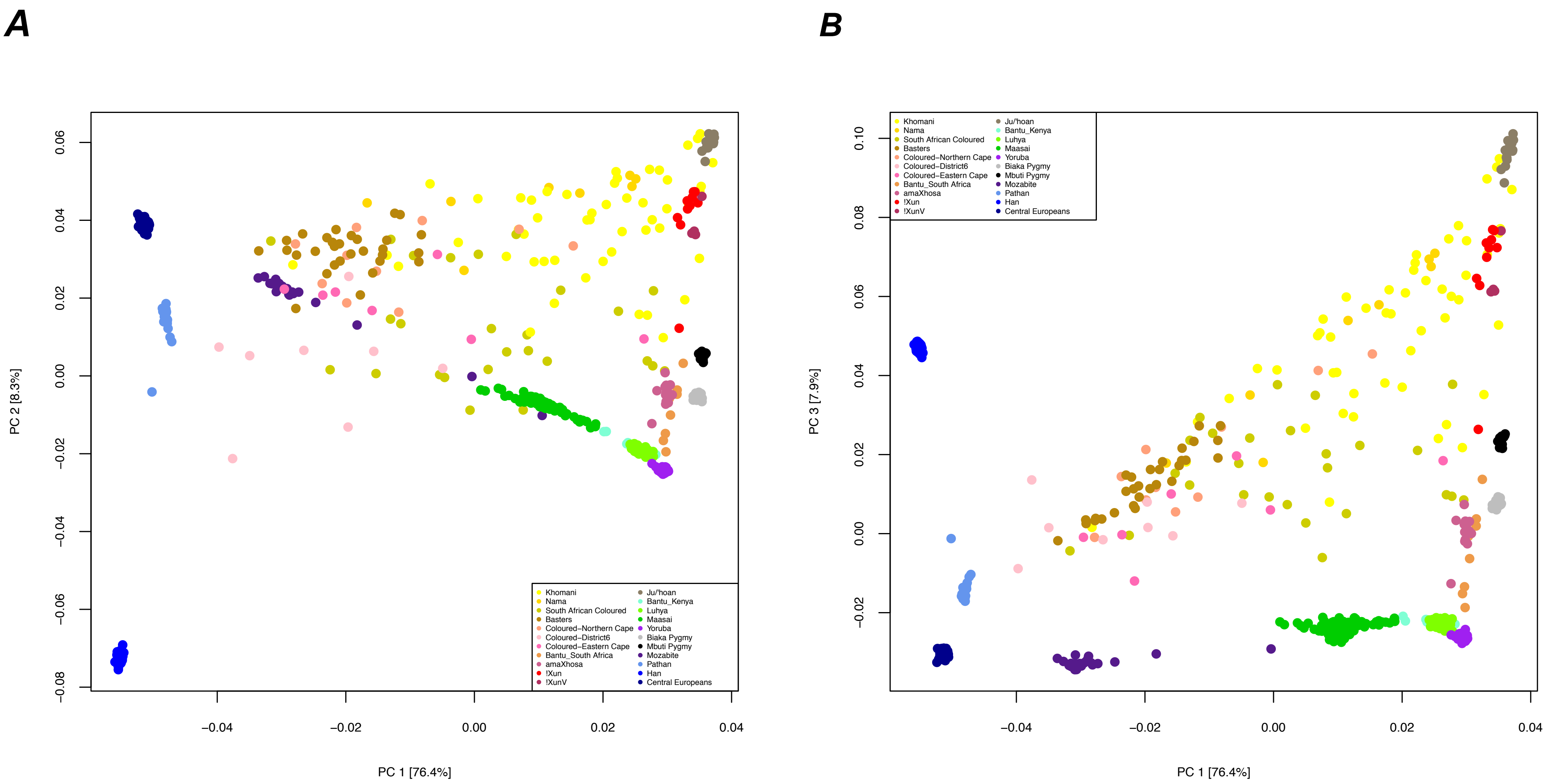


Figure S3: *Lack of clustering as well as structure related to the Nama and ≠Khomani.*

*A PCA of the merged 340k dataset depicts the clustering of unrelated individuals based on the variation seen in the dataset. PCA loadings were calculated from the *chunkcounts.out file from chromopainter using the prcomp function in R. PC 1 and 2 are depicted in A) and PC 1 and 3 are depicted in B).*

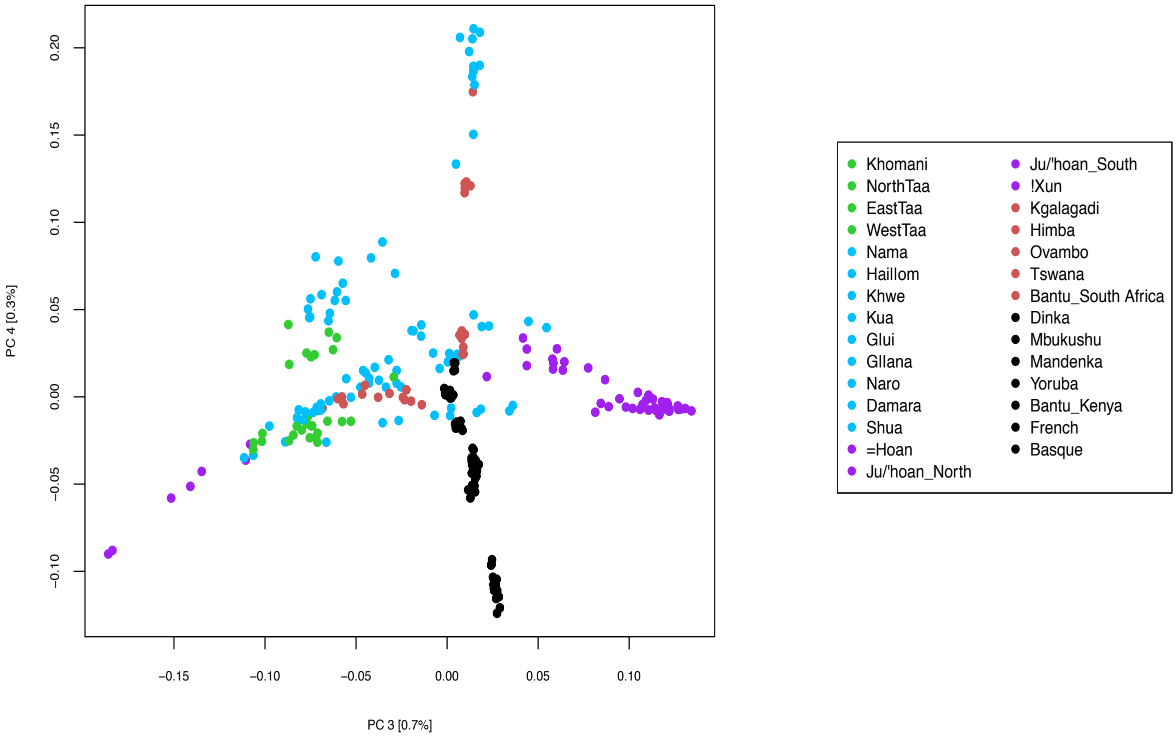


Figure S4: *Color-coding of populations based on language family shows no association between language and genetic differences.*

A PCA of the Affymetrix Human Origins dataset depicts the clustering of unrelated individuals based on the variation seen in the dataset. This PCA is identical to that in Figure 2 but is color-coded based on the language family of each population as tabulated in Table S1. Green are Tuu speaking populations. Blue are Khoe speaking populations. Purple are Kx'a speaking populations. Red are Niger-Congo speaking populations. Populations color-coded black were not included, as they did not form part of the analysis in Figure 2.

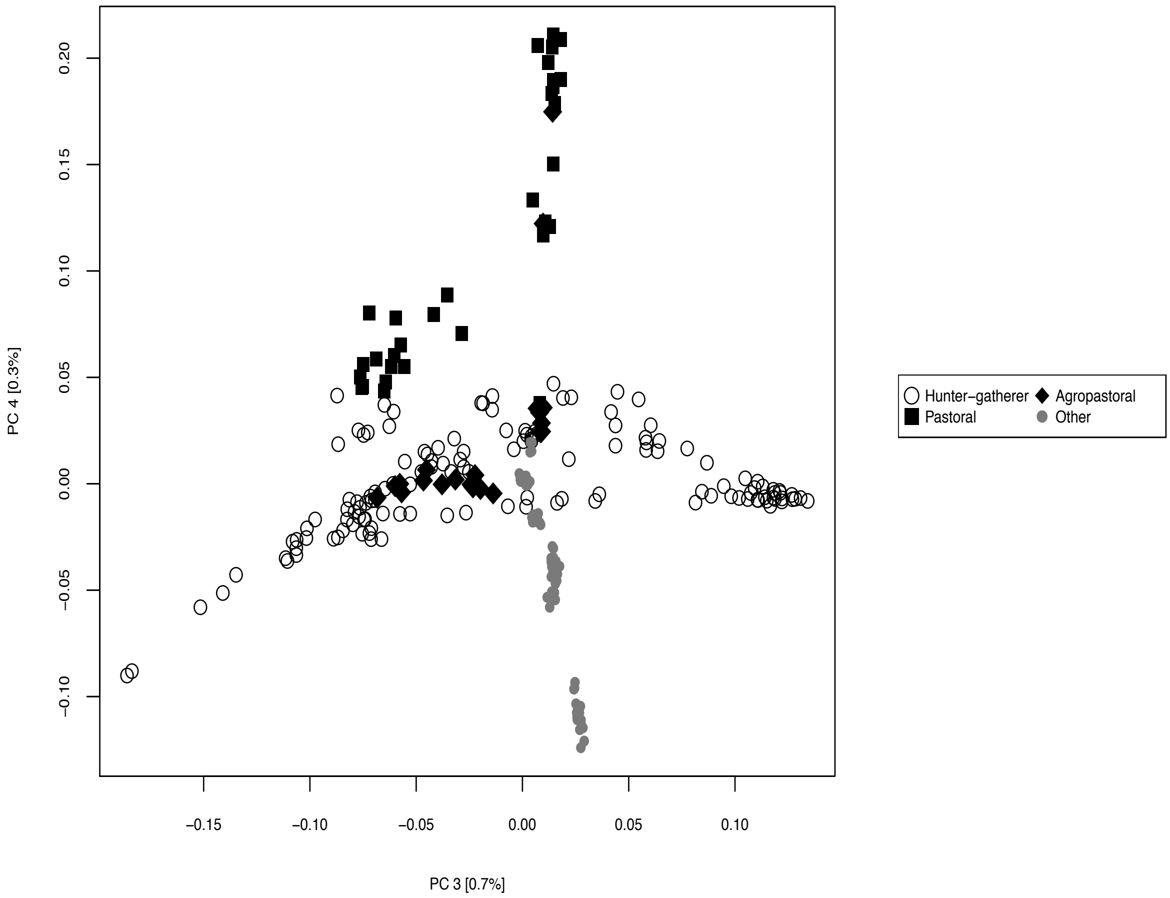


Figure S5: *Differentiation based on subsistence strategies shows some association between genetic distance and subsistence strategies.*

A PCA of the Affymetrix Human Origins dataset depicts the clustering of unrelated individuals based on the variation seen in the dataset. This PCA is identical to that in in Figure 3 but it is coded in different shapes based on the subsistence strategy of each population as tabulated in Table S1. Populations depicted by a grey circle were not included, as they did not form part of the analysis in Figure 2.

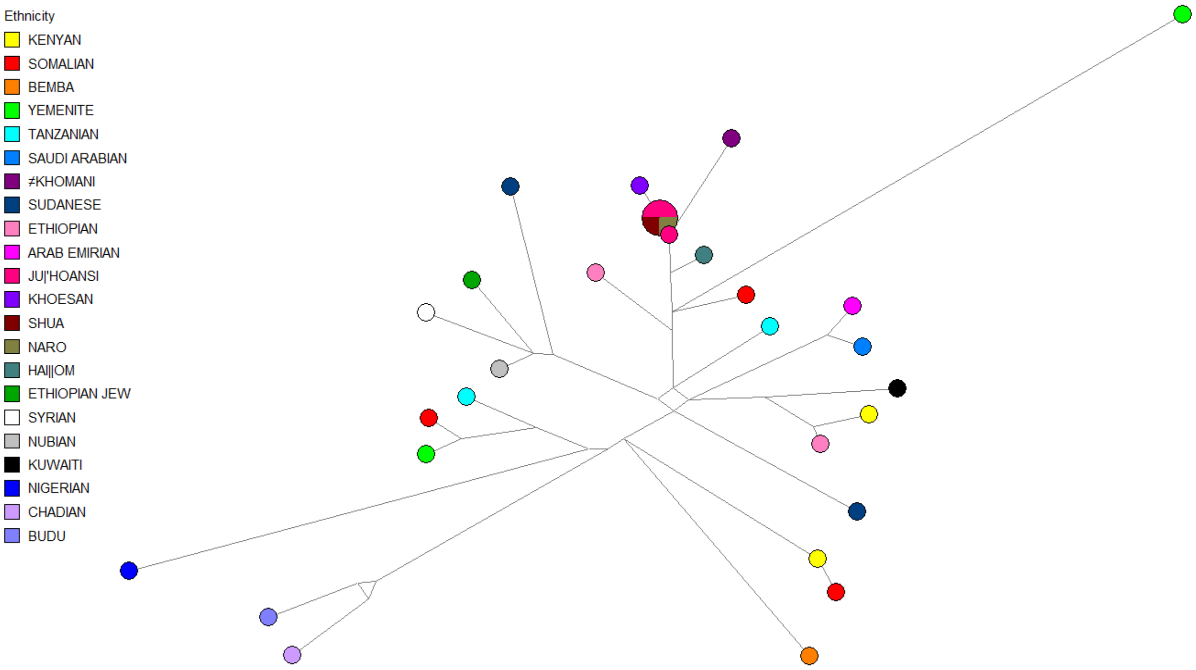


Figure S6: *L4b2 mtDNA haplogroup network- color coded per country*.

≠Khomani and Nama individuals were merged with publicly available data from NCBI (as outlined in the Supplementary Methods). All individuals were assigned mtDNA haplogroups using haplogrep and the haplotypes were plotted using Network Publisher.

Table S1: The diversity associated with the geographical location of samples populations, their language family and subsistence strategy.

Populations in bold were used to plot Figure 2. Longitude and latitude values of sampled populations were taken from Lazaridis et al (2014).

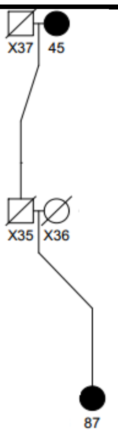
Population Sample	Location of Sample	Latitude	Longitude	Language Family	Historical Subsistence
!Xun	Namibia and Angola	**-18.7**	**19.7**	Kx'a	Hunter-gatherers
//Gana	Botswana (Central Kalahari)	**-21.7**	**23.4**	Khoe	Hunter-gatherers
/Gui	Botswana	**-21.5**	**23.3**	Khoe	Hunter-gatherers
≠Hoan	Botswana	**-24.0**	**23.4**	Kx'a	Hunter-gatherers
≠Khomani	South Africa (southern Kalahari)	**-27.8**	**21.1**	Tuu (!Ui-Taa)	Hunter-gatherers
amaXhosa	South Africa (Eastern Cape)	**-31.5**	**28.3**	Niger-Congo	Agropastoral
Bantu_Kenya	Kenyan Bantu-speakers	**-3.0**	**37.0**	Niger-Congo	Agropastoral
Bantu_SA	South African Bantu-speakers	-28.0	31.0	Niger-Congo	Agropastoral
Basque	France	43.0	0.0	Language isolate	Wage-based economy
Basters	South Africa (Northern Cape)	-23.3	17.1	Indo-European	Agropastoral
Biaka Pygmy	Southwestern Central African Republic	**4.0**	**17.0**	Niger-Congo	Hunter-gatherers
CEU	Europeans from Utah, USA	39.3	-111.1	Indo-European	Wage-based economy
ColouredD6	South Africa (District 6, Western Cape)	-33.9	18.4	Indo-European (206)	Wage-based economy
ColouredEC	South Africa (Eastern Cape)	-34.0	25.6	Indo-European (206)	Wage-based economy
ColouredNC	South Africa(Northern Cape)	-29.4	18.2	Indo-European (206)	Wage-based economy
Damara	northwest Namibia	**-19.8**	**16.2**	Khoe	Pastoral
Dinka	southern Sudan	**8.8**	**27.4**	Nilo-Saharan	Agropastoral
EastTaa	Namibia, Botswana and South Africa	**-24.2**	**22.8**	Tuu (!Ui-Taa)	Hunter-gatherers
French	France	**46.0**	**2.0**	Indo-European	Wage-based economy
Hadza	North-Central Tanzania	**-3.6**	**35.1**	Language isolate	Hunter-gatherers
Hai||om	Namibia (Etosha)	**-19.4**	**17.0**	Khoe	Hunter-gatherers
Han	China	**32.3**	**114.0**	Sino-Tibetan	Wage-based economy

Herero	Namibia, Botswana and Angola	-22	19.0	Niger-Congo	Pastoral
Himba	northern Namibia (Kunene)	**-19.1**	**14.1**	Niger-Congo	Pastoral
Ju/hoansi_North	Namibia, Angola	**-18.9**	**21.5**	Kx'a	Hunter-gatherers
Ju/hoansi_South	Namibia, Botswana and Angola	**-21.2**	**20.7**	Kx'a	Hunter-gatherers
Kgalagadi	Botswana	**-24.8**	**21.8**	Niger-Congo	Agropastoral
Khwe	Namibia, Botswana and Angola	**-18.4**	**21.5**	Khoe	Hunter-gatherers
Kua	Botswana	**-21**	**25.9**	Khoe	Hunter-gatherers
Luhya	Kenya	0.7	34.7	Niger-Congo	Agropastoral
Maasai	Southern Kenya and northern Tanzania	-1.8	36.6	Nilo-Saharan	Pastoral
Mandenka	Gambia	12.0	-12.0	Niger-Congo	Agropastoral
Mbukushu	Zambia	**-15.7**	**22.6**	Niger-Congo	Agropastoral
Mbuti Pygmy	Central Congo	**1.0**	**29.0**	Nilo-Saharan	Hunter-gatherers
Mozabite	Northern Algeria	32	3	Afro-Asiatic	Wage-based economy
Nama_AffyOrigins	Namibia	**-24.3**	**17.3**	Khoe	Pastoral
Nama_Illumina	South Africa	-28.5	17.0	Khoe	Pastoral
Naro	Namibia and Botswana (Ghanzi District)	**-22.0**	**21.6**	Khoe	Hunter-gatherers
NorthTaa	Namibia, Botswana and South Africa	**-23.0**	**22.3**	Tuu (!Ui-Taa)	Hunter-gatherers
Oroqen	China	50.4	126.5	Northern Tungusic	Wage-based economy
Ovambo	Namibia and Angola	**-19.0**	**18.1**	Niger-Congo	Agropastoral
Pathan	Pakistan	33.5	70.5	Indo-European	Wage-based economy
SAC	South Africa Coloured (Western Cape)	-33.9	18.4	Indo-European	Wage-based economy
Sandawe	Central Tanzania	-5.4	34.4	Language isolate	Hunter-gatherers
Shua	Botswana	**-20.6**	**25.3**	Khoe	Hunter-gatherers
Tswana	Botswana	-28.0	24.0	Niger-Congo	Agropastoral
WestTaa	Namibia, Botswana and South Africa	**-23.6**	**20.3**	Tuu (!Ui-Taa)	Hunter-gatherers
Yoruba	Southwestern Nigeria and southern Benin	**8.0**	**5.0**	Tonal Niger-Congo	Agropastoral

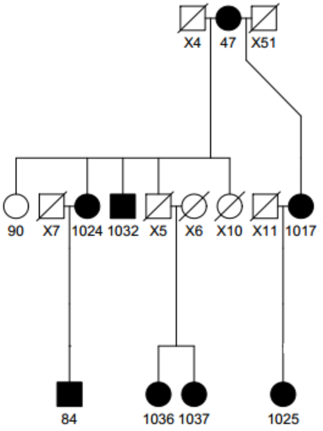
Table S2: Inferred Pedigree for ≠Khomani Samples.

***Black**: unknown (35), **Red**: known (from the 91 NGS mtDNA) (38), **Blue**: inferred (41)

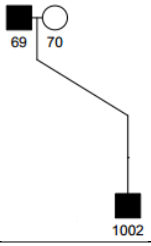
Family (total # of members)	Pedigree	Males	Females	# of matrilines	Haplogroups

F1 (5)		x37
		x35(L0d1c1)
			45(L0d1c1)
			x36(L0d2a)
			87(L0d2a)
				2 matrilines
				❶ L0d1c1
				(45 → x35(s))
				❷ L0d2a
				(x36→87(d))
					(matriline)
					L0d1c1(1)
					L0d2a(1)
					
					(individual)
					L0d1c1(2)
					L0d2a(2)
					Unknown(1)

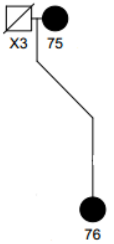
total	5 (1 + 2 + 2)	2 (1 + 1)	3 (2 + 1)	1	**2 haplogroups**

F2 (16)	 X4
	47
	X51
	90
	X7
	1024
	1032
	X5
	X6
	X10
	X11
	1017
	84
	1036 1037
	1025
		x4
		x51
		x7
		1032(L0d2a)
		x5
		x11(L0d2a)
		84(L0d2a)
			47(L0d2a)
			90(L0d2a)
			1024(L0d2a)
			x6(L0d2c)
			x10(L0d2a)
			1017(L0d2a)
			1036(L0d2c)
			1037(L0d2c)
			1025(L0d2a)
				2 matrilines
				❶ L0d2a
				(47→1024(d)→84(s))
				(47→1032(s))
				(47→1017(d)→1025(d))
				❷ L0d2c
				(x6→1036(d)&1037(d))
					(matriline)
					L0d2a (1)
					L0d2c(1)
					
					(individual)
					L0d2a(9)
					L0d2c(3)
					Unknown(4)

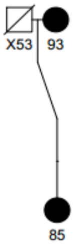
total	16 (4 + 4 + 8)	7 (4 + 1 + 2)	9 (3 + 6)	2	**2 haplogroups**

F3 (3)		69(L0d1b1)
		1002(L3e1a2)
			70(L3e1a2)	**1 matriline**
				❶ L3e1a2
				(70→1002(s))
					(matriline)
					L3e1a2(1)
					
					(individual)
					L0d1b1(1)
					L3e1a2(2)

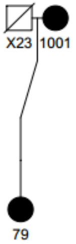
total	3 (2 + 1)	2 (2)	1(1)	1	**1 haplogroup**

F4 (3)		x3	75(L0d2a)
			76(L0d2a)
				1 matriline
				❶ L0d2a
				(75→76(d))
					(matriline)
					L0d2a(1)
					
					(individual)
					L0d2a(2)
					Unknown(1)

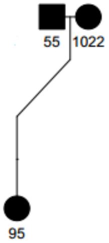
total	3 (1 + 2)	1 (1)	2 (2)	1	**1 haplogroup**

F5 (3)		x53	93
			85
				1 matriline
				❶ *unknown
				(93→85(d))
				(reason : hg of SA093 and SA085 are unknown)
					(matriline)
					*unknown(1)
					
					(individual)
					Unknown(3)

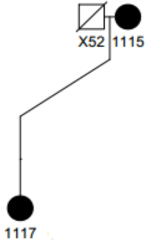
total	3 (3)	1 (1)	2 (2)	1 (1)	?

F6 (3)		x23	1001(L0d2a)
			79(L0d2a)
				1 matriline
				❶ L0d2a
				(1001→79(d))
					(matriline)
					L0d2a(1)
					
					(individual)
					L0d2a(2)
					Unknown(1)

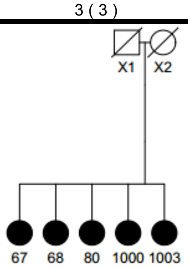
total	3 (1 + 1 + 1)	1 (1)	2 (1 + 1)	1 (1)	**1 haplogroup**

F7 (3)		55(L0d2a)	1022(L0d1b1)
			96(L0d1b1)
				1 matriline
				❶ L0d1b1
				(1022→95(d))
					(matriline)
					L0d1b1(1)
					
					(individual)
					L0d1b1(2)
					L0d2a(1)

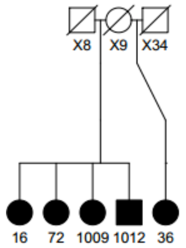
total	3 (2 + 1)	1 (1)	2 (1 + 1)	1 (1)	**1 haplogroup**

F8 (3)		x52	1115
			1117
				1 matriline
				❶ *unknown
				(1115→1117(d))
				(reason : hg of SA1115 and SA1117 are unknown.)
					(matriline)
					*unknown(1)
					
					(individual)
					Unknown(3)

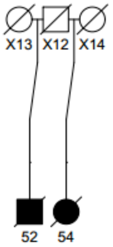
total	3 (3)	1 (1)	2 (2)	1 (1)	

F9 (7)		x1	x2(L0d2a)
			67(L0d2a)
			68(L0d2a)
			80(L0d2a)
			1000(L0d2a)
			1003(L0d2a)
				1 matriline
				❶ L0d2a
				(x2→67(d),68(d),80(d),1000(d), and 1003(d))
					(matriline)
					L0d2a(1)
					
					(individual)
					L0d2a(6)
					Unknown(1)

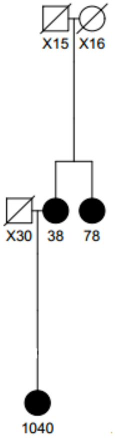
total	7 (1 + 1 + 5)	1 (1)	6 (1 + 5)	1 (1)	**1 haplogroup**

F10 (8)		x8
		x34
		1012(L0d2a)
			x9(L0d2a)
			16(L0d2a)
			72 (L0d2a)
			1009(L0d2a)
			36(L0d2a)
				1 matriline
				❶ L0d2a
				(x9→16(d),72(d),1009(d),and 1012(s))
				(x9→36(d))
					(matriline)
					L0d2a(1)
					
					(individual)
					L0d2a(6)
					Unknown(2)

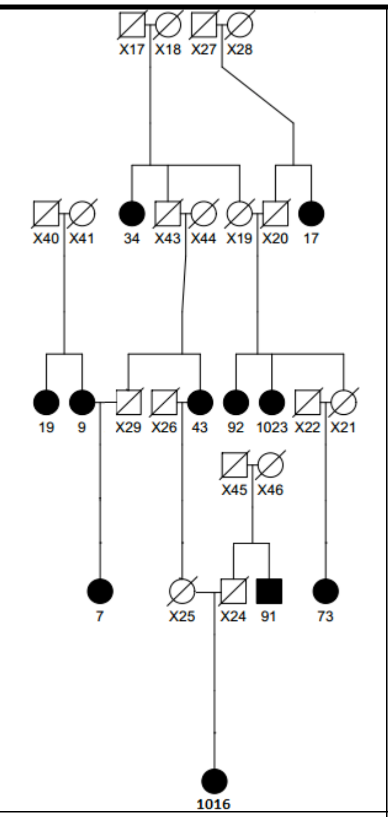
total	8 (2 + 1 + 6)	3 (2 + 1)	5 (1 + 4)	1 (1)	**1 haplogroup**

F11 (5)		x12
		52(L0d2c)
			x13(L0d2c)
			x14(L0d2c)
			54(L0d2c)
				2 matrilines
				❶ L0d2c
				(x13→52)
				❷ L0d2c
				(x14→54)
					(matriline)
					L0d2c(2)
					
					(individual)
					L0d2c(4)
					Unknown(1)

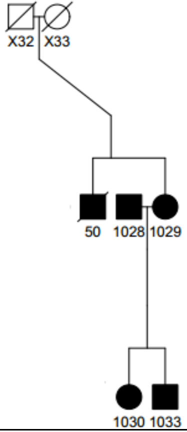
total	5 (1 + 2 + 2)	2 (1 + 1)	3 (1 + 2)	2 (2)	**2 haplogroups**

F12 (6)		x15
		x30
			x16(L0d1b1)
			38(L0d1b1)
			78(L0d1b1)
			1040(L0d1b1)
				1 matriline
				❶ L0d1b1
				(x16→78(d))
				(x16→38(d)→1040(d))
					(matriline)
					L0d1b1(1)
					
					(individual)
					L0d1b1(4)
					Unknown(2)

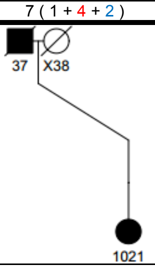
total	6 (2 + 3 + 1)	2 (2)	4 (3 + 1)	1 (1)	**1 haplogroup**

F13 (29)	 X17
	X18
	X27
	X28
	X40
	X41
	34
	X43
	X44
	X19
	X20
	17
	19
	9
	X29
	X26
	43
	92
	1023
	X22
	X21
	X45
	X46
	7
	X25
	X24
	91
	73
	1016
		x17
		x27
		x43(L0d2a)
		x20(L0d1b1)
		x29(L0d2a)
		x26
		x22
		x40
		x45
		91
		x24
			X18(L0d2a)
			x28(L0d1b1)
			34(L0d2a)
			x44(L0d2a)
			x19(L0d2a)
			17(L0d1b1)
			43(L0d2a)
			92(L0d2a)
			1023(L0d2a)
			x21(L0d2a)
			73(L0d2a)
			x41(L0d1a)
			19(L0d1a)
			9(L0d1a)
			7(L0d1a)
			x46
			x25(L0d2a)
			1016(L0d2a)
				5 matrilines
				❶ L0d2a
				(x18→34(d))
				(x18→x19(d)→x21(d)→73(d))
				❷ L0d2a
				(x44→43(d)→x25(d)→1016(d))
				❸ L0d1b1
				(x28→17(d)&x20(s))
				❹ L0d1a
				(x41→9(d)→7(d))
				(x41→19(d))
				❺ *unknown
				(x46→91(s) & x24(s))
				(reason : hg of SA091 is unknown)
					(matriline)
					L0d2a(2)
					L0d1b1(1)
					L0d1a(1)
					*unknown(1)
					
					(individual)
					L0d2a(13)
					L0d1b1(3)
					L0d1a(4)
					Unknown(9)

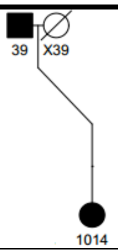
total	29 (9 + 10 + 10)	11 (8 + 3)	18 (1 + 10 + 7)	5 (4 + 1)	**4 haplogroups**

F14 (7)		x32
		50(L0d1b)
		1028(L0d1a)
		1033(L0d1b)
			x33(L0d1b)
			1029(L0d1b)
			1030(L0d1b)
				1 matriline
				❶ L0d1b
				(x33→1029(d)→1030(d)&1033(s))
				(x33→50)
					(matriline)
					L0d1b(1)
					
					(individual)
					L0d1b(5)
					L0d1a(1)
					Unknown(1)

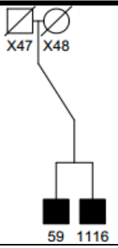
total	7 (1 + 4 + 2)	4 (1 + 2 + 1)	3 (2+ 1)	1 (1)	**1 haplogroup**

F15 (3)		37	x38(L0d1b)
			1021(L0d1b)
				1 matriline
				❶ L0d1b
				(x38→1021(d))
					(matriline)
					L0d1b(1)
					
					(individual)
					L0d1b(2)
					Unknown(1)

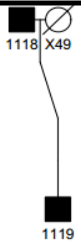
total	3 (1 + 1 + 1)	1 (1)	2 (1 + 1)	1 (1)	**1 hpalogroup**

F16 (3)		39(L0d2a)	x39(L0d2a)
			1014(L0d2a)
				1 matriline
				❶ L0d2a
				(x39→1014(d))
					(matriline)
					L0d2a(1)
					
					(individual)
					L0d2a(3)

total	3 (2 + 1)	1 (1)	2 (1 + 1)	1 (1)	**1 haplogroup**

F17 (4)		x47
		59
		1116
			x48	**1 matriline**
				❶ *unknown
				(x48→59(s)&1116(s))
				(reason : hg of SA059 and SA1116 are not known.)
					(matriline)
					*unknown(1)
					
					(individual)
					Unknown(4)

total	4 (4)	3 (3)	1 (1)	1 (1)	?

F18 (3)		1118(L0d2a)
		1119
			x49	**1 matriline**
				❶*unknown
				(x49→1119(s))
				(reason : hg of SA1119 is unknown.)
					(matriline)
					*unknown(1)
					
					(individual)
					L0d2a(1)
					Unknown(2)

Total	3 (2 + 1)	2 (1 + 1)	1 (1)	1 (1)	?

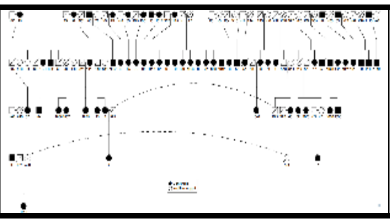
114 individuals from 18 families		46 (29 + 9 + 8)	68 (7 + 29 + 33)	25 independent matrilines	7 haplogroups +unknown

Table S3: Genetic (A), Geographic (B) and Phonemic (C) distance matrices per sampled population. Calculation of these values are described in the *Methods*.

A

	Kung	Juhoan	Nama	Kua	Shua	Khwe	Gana	Gui	Naro	EastTaa	WestTaa	Khomani
Kung	0,000	0,013	0,021	0,026	0,032	0,032	0,020	0,016	0,011	0,020	0,011	0,018
Juhoan	0,013	0,000	0,038	0,044	0,054	0,055	0,036	0,025	0,012	0,028	0,021	0,034
Nama	0,021	0,038	0,000	0,019	0,020	0,021	0,016	0,019	0,024	0,025	0,024	0,003
Kua	0,026	0,044	0,019	0,000	0,014	0,016	0,010	0,019	0,031	0,029	0,029	0,018
Shua	0,032	0,054	0,020	0,014	0,000	0,013	0,017	0,028	0,039	0,037	0,039	0,021
Khwe	0,032	0,055	0,021	0,016	0,013	0,000	0,019	0,031	0,041	0,041	0,041	0,022
Gana	0,020	0,036	0,016	0,010	0,017	0,019	0,000	0,011	0,022	0,019	0,019	0,013
Gui	0,016	0,025	0,019	0,019	0,028	0,031	0,011	0,000	0,011	0,013	0,013	0,015
Naro	0,011	0,012	0,024	0,031	0,039	0,041	0,022	0,011	0,000	0,014	0,008	0,019
EastTaa	0,020	0,028	0,025	0,029	0,037	0,041	0,019	0,013	0,014	0,000	0,011	0,021
WestTaa	0,011	0,021	0,024	0,029	0,039	0,041	0,019	0,013	0,008	0,011	0,000	0,019
Khomani	0,018	0,034	0,003	0,018	0,021	0,022	0,013	0,015	0,019	0,021	0,019	0,000

B

	Kung	Juhoan	Nama	Kua	Shua	Khwe	Gana	Gui	Naro	EastTaa	WestTaa	Khomani
Kung	0	191	671	698	624	193	511	489	417	691	549	1023
Juhoan	191	0	742	516	441	56	369	345	345	605	538	992
Nama	671	742	0	957	920	788	689	690	509	558	315	544
Kua	698	516	957	0	77	544	271	275	459	478	645	900
Shua	624	441	920	77	0	468	232	231	414	476	614	908
Khwe	193	56	788	544	468	0	418	393	401	660	592	1047
Gana	511	369	689	271	232	418	0	25	189	285	382	718
Gui	489	345	690	275	231	393	25	0	184	305	387	736
Naro	417	345	509	459	414	401	189	184	0	274	223	648
EastTaa	691	605	558	478	476	660	285	305	274	0	263	435
WestTaa	549	538	315	645	614	592	382	387	223	263	0	474
Khomani	1023	992	544	900	908	1047	718	736	648	435	474	0

C

	Kung	Juhoan	Nama	Kua	Shua	Khwe	Gana	Gui	Naro	EastTaa	WestTaa	Khomani
Kung	0,000	0,407	0,673	0,588	0,650	0,564	0,518	0,513	0,532	0,558	0,676	0,602
Juhoan	0,407	0,000	0,633	0,598	0,661	0,598	0,540	0,535	0,514	0,568	0,612	0,623
Nama	0,673	0,633	0,000	0,548	0,567	0,528	0,471	0,486	0,470	0,515	0,733	0,590
Kua	0,588	0,598	0,548	0,000	0,225	0,505	0,388	0,365	0,383	0,511	0,651	0,586
Shua	0,650	0,661	0,567	0,225	0,000	0,500	0,392	0,367	0,408	0,560	0,683	0,632
Khwe	0,564	0,598	0,528	0,505	0,500	0,000	0,266	0,263	0,363	0,477	0,660	0,557
Gana	0,518	0,540	0,471	0,388	0,392	0,266	0,000	0,028	0,141	0,429	0,616	0,516
Gui	0,513	0,535	0,486	0,365	0,367	0,263	0,028	0,000	0,139	0,424	0,612	0,510
Naro	0,532	0,514	0,470	0,383	0,408	0,363	0,141	0,139	0,000	0,405	0,658	0,500
EastTaa	0,558	0,568	0,515	0,511	0,560	0,477	0,429	0,424	0,405	0,000	0,648	0,333
WestTaa	0,676	0,612	0,733	0,651	0,683	0,660	0,616	0,612	0,658	0,648	0,000	0,686
Khomani	0,602	0,623	0,590	0,586	0,632	0,557	0,516	0,510	0,500	0,333	0,686	0,000

Chapter 4

A post-GWAS analysis of predicted regulatory variants and tuberculosis susceptibility

Caitlin Uren[1], Brenna M. Henn[2], Andre Franke[3], Michael Wittig[3], Paul D. van Helden[1], Eileen G. Hoal[1], Marlo Möller[1*]

[1]SA MRC Centre for TB Research, DST/NRF Centre of Excellence for Biomedical TB Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, 8000, South Africa
[2]Department of Ecology and Evolution, Stony Brook University, Stony Brook, New York, 11794, United States of America
[3]Institute of Clinical Molecular Biology, Christian-Albrechts-University of Kiel, Rosalind-Franklin-Strasse 12-24105 Kiel, Germany

PLoS One 2017 (Impact Factor 3.057)
doi:10.1371/journal.pone.0174738

Abstract

Utilizing data from published tuberculosis (TB) genome-wide association studies (GWAS), we use a bioinformatics pipeline to detect all polymorphisms in linkage disequilibrium (LD) with variants previously implicated in TB disease susceptibility. The probability that these variants had a predicted regulatory function was estimated using RegulomeDB and Ensembl's Variant Effect Predictor. Subsequent genotyping of these 133 predicted regulatory polymorphisms was performed in 400 admixed South African TB cases and 366 healthy controls in a population-based case-control association study to fine-map the causal variant. We detected associations between tuberculosis susceptibility and six intronic polymorphisms located in *MARCO*, *IFNGR2, ASHAS2, ACACA*, *NISCH* and *TLR10*. Our post-GWAS approach demonstrates the feasibility of combining multiple TB GWAS datasets with linkage information to identify regulatory variants associated with this infectious disease.

Introduction

Genome-wide association studies (GWAS) have advanced the investigation of complex disease genetics and identified thousands of disease-associated variants. This study design compares allele frequencies of common genetic variants across the genome with phenotypic variation in large cohorts of cases and controls. GWAS is based on the premise that causal variants will be in linkage disequilibrium (LD) with the markers present on single nucleotide polymorphism (SNP) arrays. Since 2005, when the first GWAS was published (207), associations have been detected between numerous common genetic variants and several infectious diseases including TB (159,208–210).

More than 10 GWAS investigating TB susceptibility have been published to date (**Table 1**). These studies investigated the genetic factors associated with TB susceptibility in multiple populations. Thye et al. (2010) performed the first GWAS on TB susceptibility in a case-control cohort from Ghana and the Gambia and identified a region on chromosome 18q11.2. Within this region, there are numerous immune response genes such as cadherin 13 (*CDH13)*, zinc finger protein 229 (*ZNF229)* and exportin 1 (*XPO1)*. A meta-analysis which included data from Ghana, the Gambia, Russia and Indonesia identified variants at 11p13 that were associated with TB susceptibility (30). Chimusa et al. (2014) validated several of these loci and identified novel TB associations in a South African case-control cohort (40).

Table 1: Previous TB GWAS- results.

Population	Variant/Gene	Number of Cases	Number of Controls	Reference
Ghana	rs4331426 (gene desert)	921	1740	
The Gambia		1316	1382	
Black, White, Asian from USA	rs4893980 *(PDE11A)*	48	57	(211)
	rs10488286 *(KCND2)*			
	rs2026414 *(PCDH15)*			
	rs10487416 (unknown gene)			
Thai and Japanese	Intergenic region between HSPEP1-MAFB	620	1524	(212)
Indonesia	rs1418267 *(TXNDC4)*	108	115	(213)
	rs2273061 *(JAG1)*			
	rs4461087 *(DYNLRB2)*			
	rs1051787 *(EBF1)*			
	rs10497744, rs1020941 *(TMEFF2)*			
	rs188872 *(CCL17)*			
	rs10245298 *(HAUS6)*			
	rs6985962 *(PENK)*			
Ghana	rs2057178 (*WT1*, intergenic)	2127	5636	(30)
The Gambia		1207	1349	
Russia		1025	983	
Indonesia		4441	5874	
South African Coloured	rs2057178, rs11031728 (*WT1*, intergenic)	642	91	(40)
	rs10916338,rs1925714 *(RNF187)*			
	rs6676375 *(PLD5)*			
	rs1075309 *(SOX11)*			
	rs958617 *(CNOT6L)*			
	rs1727757 *(ZFPM2)*			
	rs2505675 *(LOC100508120)*			
	rs1934954 *(CYP2C8)*			
	rs12283022,12294076 *(DYNC2H1)*			
	rs7105967,rs7947821 *(DCUN1D5)*			
	rs6538140 *(E2F7)*			
	rs1900442 *(VWA8)*			
	rs17175227 *(SMOC1)*			
	rs40363 *(NAA60)*			
	rs2837857 *(DSCAM)*			
	rs451390 *(C2CD2)*			
	rs3218255 *(IL2RB)*			
Russia	rs4733781,rs10956514,rs1017281,rs1469288, rs17285138,rs2033059,rs12680942 *(ASAP1)*	5530	5607	(209)
Morocco	rs358793 (Intergenic)	556	650	(214)
	rs17590261 (Intergenic)			
	rs6786408 *(FOXP1)*			
	rs916943 *(AGMO)*			
Uganda and Tanzania	rs4921437 *(IL-12)*	267	314	(215)
Iceland	rs557011, rs9271378 (located between *HLADQA1* and *HLA-DRB1*)	8162	277643	(216)
	rs9272785 *(HLA-DQA1)*			

The majority of TB susceptibility variants previously identified are intronic (S1 Table) and may therefore have some regulatory functions. It has recently become feasible to predict regulatory effects of variants as computational tools, such as RegulomeDB (217) and Ensembl's Variant Effect Predictor, as well as information regarding the possible impact of regulatory regions have become available (217,218). We therefore applied a post-GWAS approach to TB susceptibility to identify possible variants contributing to disease development. A post-GWAS approach entails the use of previous GWAS associations and linkage disequilibrium (LD) data to identify further variants [and possibly the causative variant] that may be associated with the phenotype. This methodology was developed as pinpointing the exact targets of these associations is a challenge (219). The post-GWAS analysis has previously been used to identify novel functional intronic variants associated with late-onset Alzheimer's disease (220), cardiovascular disease (221,222) and human aging (223). There has been no such analysis on susceptibility to TB.

Here we combine TB GWAS and candidate gene association studies and incorporate knowledge from RegulomeDB (217) and Ensembl's Variant Effect Predictor to fine-map putative regulatory variants that may predispose an individual to progress to active TB.

Methods

Study population

Sample collection was approved by the Health Research Ethics Committee of the Faculty of Health and Medical Sciences, Stellenbosch University (N95/072) and written informed consent was obtained from all study participants. Recruitment was done in two suburbs in the Western Cape, South Africa, where the incidence of TB is high (1340/100 000 population during 1996), although the HIV incidence at the time of sampling was low (~2% of population) (224). Study participants self-identified as being from the South African Coloured (SAC) population. The admixed SAC population has genetic contributions from five ancestral populations. On average, Bantu-speaking populations contribute ~30%, the KhoeSan ~30%, Europeans 12-18% South Asian ~15%, and East Asian <10% (51–53). Genotyping of ancestry informative markers (AIMS) was previously performed by Daya et al. (2013). These AIMS were used to infer admixture proportions using ADMIXTURE (52,169).

All study participants were HIV negative and unrelated. TB cases were bacteriologically confirmed (n = 400). The controls had no previous history of TB (n = 366) and were older than 18 years. Tuberculin skin tests (TST) were not performed, as the majority of adults in the communities are TST positive (> 80% of children older than 15 years) (4). DNA was extracted from blood using the Nucleon BACC3 Kit (Amersham Biosciences, Buckinghamshire, UK).

Bioinformatics Analysis

Data mining was done using the National Health Genome Research Institute's –European Bioinformatics Institute (NHGRI-EMBI) GWAS catalogue (http://www.ebi.ac.uk/gwas) and PubMed (www.ncbi.nlm.nih.gov/pubmed). Only single nucleotide polymorphisms (SNPs) that were reported to have $p < 0.05$ (after multiple testing correction during association tests) were recorded.

LD was calculated using SNAP (225). These comparisons were made against Hapmap3 (release 2) data from 4 populations that best represented the ancestral populations of the SAC; Europeans (CEU), Han Chinese (CHB), Luhya (LWK) and Gujarati Indians (GIH). The KhoeSan ancestral component was represented by 2 ≠Khomani genomes. LD between the previously published SNPs and the ancestral genomes was calculated. An r^2 threshold of 0.8 and a window size of 500 kilobases were used as filters. A per population analysis was performed and SNPs were pooled across populations.

The potential functional impact of these variants was ascertained by RegulomeDB (217) and Ensembl's Variant Effect Predictor (218). Only variants that had a RegulomeDB score of 1 (highest likelihood of a potential functional impact due to the variant) were used in further analysis due to the large number of variants. A summary of the filtering steps used are illustrated in **Fig 1**.

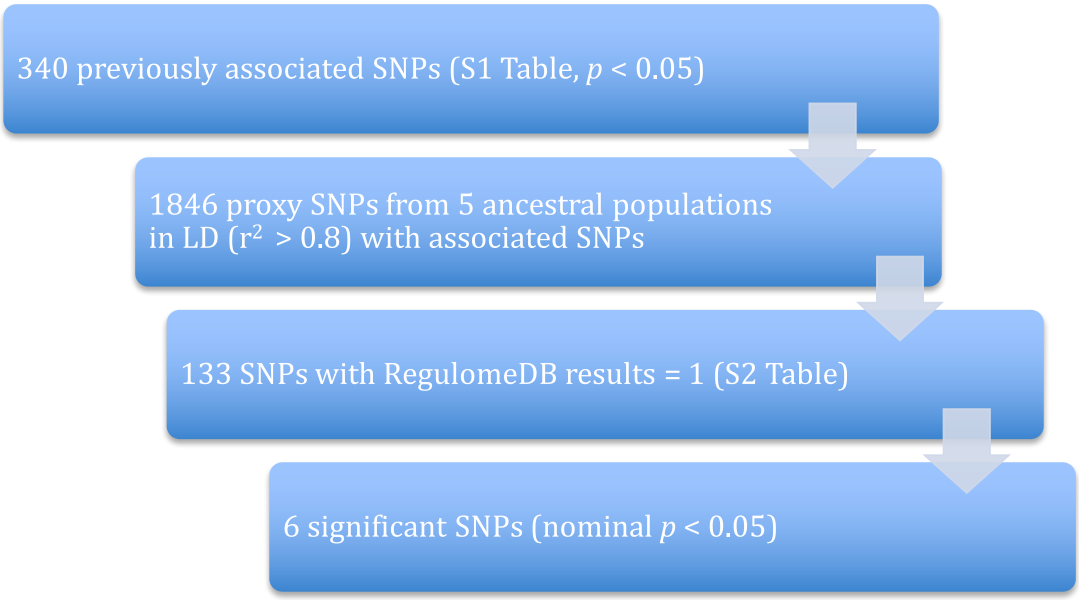

340 previously associated SNPs (S1 Table, $p < 0.05$)
1846 proxy SNPs from 5 ancestral populations in LD ($r^2 > 0.8$) with associated SNPs
133 SNPs with RegulomeDB results = 1 (S2 Table)
6 significant SNPs (nominal $p < 0.05$)

Figure 1: Bioinformatics pipeline for the prioritization of variants.

Genotyping

Genotyping was performed using the Agena MassARRAY® system (Institute of Clinical Molecular Biology, Christian-Albrecht's-University, Kiel, Germany). A total number of 133 SNPs were genotyped in 400 cases and 366 controls. Only SNPs that passed quality control (as determined by the confidence in allele call from the Typer Analyzer software package (version 4.0.20, Sequenom proprietary software)) were recorded.

Statistical Analysis

The CaTS power calculator was used to perform power calculations (226). The power to find a true deviation from the null hypothesis was calculated using a disease allele frequency of 0.06 and an alpha level (or significance level) of 0.05 to determine an odds ratio of at least 2. A TB disease prevalence of 1% was used in this analysis (227). With the sample size available, the power to detect a deviation from null using these parameters was therefore calculated to be 98%.

All statistical analyses were done in R (www.r-project.org) using functions from the base R packages. The Fisher's exact test was used to calculate Hardy-Weinberg Equilibrium (HWE) *p*-values using functions from the *genetics* R package (228). Logistic regression was used to analyse the genotypic and allelic models. All models were adjusted for the confounding factors age, ancestry and sex by including these as covariates. The allelic models (recessive, dominant and additive) were assessed and the model with the highest likelihood to correctly model the data was chosen (17). Our SNP selection strategy was based on *a priori* evidence that the genes were associated with TB and Bonferroni corrections for multiple testing would be too stringent, risking the rejection of important findings (229–231). For this reason, the Šidák step-up method was utilized (232). Nominal *p*-values were corrected for multiple testing using the *multtest* package in R (233) and the cut-off for significance was $p = 0.05$.

Results

Descriptive statistics were generated for the cases and controls (**Table 2**). Age, sex and KhoeSan ancestry differed significantly between the cases and controls and were therefore all adjusted for in the logistic regression models. Prioritization of SNPs for genotyping is shown in **Fig 1**. Data mining identified 1800 SNPs that were found to be in LD ($r^2 > 0.8$) with the 230 SNPs previously associated with TB. After filtering for RegulomeDB scores of 1 (S2 Table), 133 SNPs remained and were genotyped in 400 TB cases and 366 healthy controls. All SNPs were in Hardy-Weinberg Equilibrium ($p > 0.01$). Of the 133 SNPs, 6 variants were found to be statistically significantly associated with TB susceptibility ($p < 0.05$) after adjusting for age, sex and ancestry (**Table 3**), but not after correcting for multiple testing. In this study, we view the methods of correction for multiple testing (including Bonferonni) to be too conservative for this analysis; there is *a priori* evidence that variants in LD with those reported here were associated with TB susceptibility (**Table 1**) (229–231). For completeness however, nominal *p* values adjusted for multiple testing using the Šidák method are reported in **Table 3** (232).

Table 2: Case-control sample characteristics and TB susceptibility modelling results.

	TB Cases (n=398)	Controls (n=360)	*p* value*
Age (mean ± SD)	36.55 ± 11.26	30.69 ± 12.80	0.0001
Number of males (proportion)	211 (0.53)	111 (0.28)	< 0.0001
KhoeSan [IQR]	0.30 [0.20-0.39]	0.27 [0.18-0.36]	0.0224
West African [IQR]	0.27 [0.16-0.39]	0.25 [0.15-0.37]	0.3187
European [IQR]	0.18 [0.08-0.28]	0.19 [0.12-0.28]	0.7804
South Asian [IQR]	0.12 [0.03-0.19]	0.14 [0.06-0.22]	0.2767
East Asian [IQR]	0.09 [0.03-0.16]	0.10 [0.05-0.17]	NA[a]

[a] The East Asian component was not added to the model to avoid linear dependency.
SD standard deviation, IQR, interquartile range
*Statistic is an indication of the significance of the association between each factors with TB after adjusting for the other factors.

Table 3: Association results of statistically significant regulatory SNPS.

rsID	**Controls**		**HWE[c] *p*-val**	**TB Cases**		**HWE *p*-val**	**Association *p*-val Adjusted[d]**	**Model of penetrance**	**OR**	**2.5% CI**	**97.5% CI**	**Šidák *p*-val**
	Count[a]	**Prop[b]**		**Count**	**Prop**							
rs2284555 (*IFNGR2*)												
	352		0.457	390		0.208						
G/G	37	0.11		25	0.06		0.043					0.997
A/G	145	0.41		168	0.43			Genotypic	2.146	1.174	3.988	
A/A	170	0.48		197	0.51			Genotypic	2.012	1.108	3.716	
G	219	0.31		218	0.28							
A	485	0.69		562	0.72		0.179					
rs829161 (*ACACA*)												
	348		0.539	387		0.015						
T/T	208	0.6		236	0.61		0.037					0.990
T/C	125	0.36		121	0.31							
C/C	15	0.04		30	0.08			Genotypic	2.274	1.133	4.727	
T	541	0.78		593	0.77		0.292					
C	155	0.22		181	0.23							
rs7599352 (*MARCO*)												
	327		0.034	353		0.438						
C/C	183	0.56		213	0.6		0.021					0.941
C/T	113	0.35		126	0.36							
T/T	31	0.09		14	0.04			Genotypic	0.416	0.197	0.842	
C	479	0.73		552	0.78		0.244					
T	175	0.27		154	0.22							

rs4687614 (*NISCH*)												
	355		1	391		0.733						
G/G	7	0.02		14	0.04		0.040					
A/G	87	0.25		114	0.29							
A/A	261	0.74		263	0.67							
G	101	0.14		142	0.18							
A	609	0.86		640	0.82		0.012	Additive	1.464	1.089	1.980	0.799
rs2600665 (*AHSA2*)												
	350		0.811	383		0.691						
T/T	156	0.45		207	0.54		0.019					
T/G	154	0.44		152	0.4							
G/G	40	0.11		24	0.06							
T	466	0.67		566	0.74		0.006	Additive	0.714	0.559	0.910	0.551
G	234	0.33		200	0.26							
rs12233670 (*TLR10/1*)												
	352		0.194	390		0.023						
T/T	105	0.3		136			0.022					
T/C	186	0.53		169				Genotypic	0.699	0.491	0.994	0.948
C/C	61	0.17		85								
T	396	0.56		441			0.815					
C	308	0.44		339								

[a] Allelic and genotype counts
[b] Allelic and genotype proportions
[c] Statistic for the HWE exact test. This was stratified by TB susceptibility status.
[d] Statistic to indicate the association between genotype and TB susceptibility after adjusting for age, gender and ancestry. The allelic effect was modelled using the additive model.

The SNPs rs7599352, rs2600665, rs829161 and rs12233670 showed statistically significant associations with resistance to TB (nominal *p* = 0.021, 0.006, 0.037 and 0.022, respectively). The minor homozygote (T/T) of rs7599352 was found in fewer cases than controls and the genotype has an odds ratio of 0.416 (95% CI = 0.197 -0.842). In contrast, rs2600665 followed an additive model of penetrance and thus for every copy of the G allele, an individual is ~30% less likely to progress to active disease. The ancestral homozygote (C/C) of rs829161 was found in double the amount of cases than controls and was associated with susceptibility to TB (nominal *p* = 0.037). This genotype has an odds ratio of 2.274 (95% CI = 1.13 - 4.73). The heterozygote genotype of rs12233670 was associated with TB susceptibility and with an odds ratio of 0.699 (95% CI = 0.491 - 0.994). Two SNPs, namely rs4687614 and rs2284555, were associated with increased susceptibility to TB. For every copy of the minor G allele of rs4687614, there was an increase of ~50% in the likelihood of progressing to active TB (*p* = 0.040, 95% CI = 1.09 - 1.98). In addition, rs2284555 was associated with TB susceptibility (*p* = 0.043); both the minor homozygote and heterozygote were associated with the phenotype. The minor homozygote has an odds ratio of 2.012 (95% CI = 1.108 - 3.716) whereas the heterozygote yielded an odds ratio of 2.146 (95% CI = 1.174 - 3.988).

Discussion

The genetic factors influencing TB susceptibility have been under investigation for many years, using various study designs, with limited success (S1 Table). This is due to a number of factors, including lack of power, the inability to identify causative variants and the complexity of admixed populations. We conducted a post-GWAS analysis of predicted functional variants and investigated their associations with TB susceptibility in the admixed SAC population. This resulted in the identification of 3 variants associated with an increased risk of progressing to active TB and another 3 variants associated with resistance to active TB.

*MAR*CO (Macrophage Receptor With Collagenous Structure) is a member of the class A scavenger receptor family and has been implicated in innate antimicrobial activity, more specifically TB susceptibility, in the Gambian and Chinese Han populations (234–236). Statistically significant variants from Songane et al. (2012) and Bowdish et al. (2013) were used as query SNPs in the study presented here (235,237). The rs7599352 variant has not directly been implicated in TB susceptibility previously, but other intronic variants have been shown to have an effect on *MARCO* gene expression (235,238). This implies that intronic variants within this gene could have a greater functional effect than exonic variants.

Toll-like receptors (TLRs) are responsible for pathogen recognition and the resulting activation of an innate immune response. TLR genes, including *TLR2, TLR4* and *TLR8,* are known to contribute to TB susceptibility (237) and numerous variants within *TLR1* and *TLR10* have also been associated with TB susceptibility in previous studies (239,240). We provide evidence for a novel SNP (rs12233670) associated with a decrease in risk of TB susceptibility. This SNP was in LD with statistically significant SNPs reported in Ma et al. (2007) and Uciechowski et al. (2011) (239,241).

Due to the long-range effects of LD with variants in Dual Specificity Phosphatase 14 (*DUSP14)* and *XPO1* (S1 Table), two novel genes were associated with increased TB susceptibility and increased resistance, namely, Acetyl-CoA Carboxylase Alpha (*ACACA)* and Activator Of Heat Shock 90kDa Protein ATPase Homolog 2 (*ASHA2)* respectively. An intronic variant (rs829161) in the *ACACA* gene was highly associated with the susceptibility phenotype with an odds ratio suggesting more than a 2.2 times increased chance of the disease progressing to an active state than remaining latent. The *ACACA* gene is involved in fatty acid carboxylation and in turn mediates the removal of cholesterol. Cholesterol is used as an energy source for *Mycobacterium tuberculosis (M.tb)* therefore any disruption of the level of the fatty acid may influence the dormancy of the bacteria (242). A possible novel resistance pathway was identified in this study, involving heat shock protein (HSP) activation and resulting ATPase activity by ASHA2. In this case the variant (rs260065) is located in the 5'-untranslated region and we therefore hypothesize that it has a profound regulatory function, potentially affecting gene expression. HSPs play a pivotal role in protein folding, stabilization and degradation, and are targets of chemotherapy in cancer patients as HSP modulates tumour cell apoptosis through protein kinase B, tumour necrosis factor receptor and NF-kB functioning. Due to the interaction with these known immune response genes, we hypothesize that ASHA2 could be involved in TB immune responses.

A variant (rs4687614) in the Nischarin (*NISCH)* gene was found to be associated with an increase in TB susceptibility. *NISCH* encodes the Nischarin protein. Nischarin has recently been identified as a dual effector that interacts with members of the Rac and Rab GTPase families (243). The regulation of GTPases by Nischarin may regulate the maturation and acidification of vacuoles that are associated with phagocytosis of bacterial pathogens (243). The odds ratio associated with the variant indicated an increased risk of progressing to active TB. An association was also found for a polymorphism in the Interferon-gamma Receptor 2 (*IFNGR2)* gene. An intronic variant (rs2284555) yielded a high odds ratio of 2.01 for the A/A genotype and 2.15 for the A/G genotype. *IFNGR2* has been previously implicated in various immunodeficiencies (244–247). This study reiterates the role of IFNGR2 in antibacterial immune response and provides a novel causative SNP for TB.

Only 6 out of a total of 133 SNPs genotyped produced statistically significant associations with TB susceptibility. The predicted regulatory SNPs genotyped in our study were in strong LD with variants previously associated with TB in other populations. A comparison between the *p*-values of the original GWAS SNPs and the six variants identified here, shows that the *p*-values are lower in the original GWAS data. This could be an artefact due to the low sample size in our study, an increase in error variation or the moderate to low effect sizes associated with the six variants (248). Susceptibility to TB is a complex disease and it is possible that numerous small/moderate effect variants at a frequency less than 0.05 will play a role in this phenotype (42). It is also possible that some of the variants identified by previous studies are population-specific susceptibility variants and are therefore unlikely to be involved in disease predisposition in the SAC population. One might have expected many more associations due to the strong LD between SNPs genotyped in our study and variants previously identified as being associated with TB. The lack of associations suggest that these SNPs are not the causative SNPs that led to significant results in previous studies and that other SNPs in LD with the marker variants may play a role. Alternatively, the variants have smaller effect sizes than we could detect with our sample size. However, the power to detect a common variant with an odds ratio of 2 was 98%. Validation of these results in other case-control cohorts as well as the inclusion of recent GWAS results (159,214,216) is desirable, but complicated by the lack of available TB case-control cohorts with a similar genetic structure to that of the SAC population. In addition, since there was *a priori* evidence for an association, replication is arguably not necessary as this study attempted to fine-map the potential causal variants in loci identified by previous TB GWAS.

Age, KhoeSan ancestry and sex differed significantly between the TB cases and controls in this study, but were adjusted for in all analyses. We have previously shown that KhoeSan ancestry increases the risk of active TB (17) and old age is also a known TB risk factor (249). Based on our experience in these communities, healthy males are less likely to attend clinics, while healthy females will accompany sick children and are therefore more likely to participate as controls. However, in most countries the TB notification rate is twice as high in males as in females (5) and evidence suggests that the X chromosome does play a role in TB susceptibility (250). An investigation of sib-pair families from The Gambia indicated that chromosome Xq might be involved in TB susceptibility (250). Additionally, sex-specific TB associations for *TLR8,* an X-linked gene, has been identified in several populations (251–253), including the SAC population (33). These loci could contribute to the observed male sex-bias in this population, but will require further investigation.

Subsequent to the analysis presented here, Chimusa et al. (2016) published a post-GWAS methodology utilizing LD information and the human protein–protein interaction network, which identified novel pathways associated with breast cancer (42). The study by Chimusa et al. (2016) reiterates the need for alternative methodologies in the identification of regulatory variants associated with a phenotype.

In summary, the six predicted functional variants associated with TB susceptibility in the SAC population show that fine mapping of GWAS results can reveal candidate causal variants. Functional analyses are now required to elucidate the molecular mechanisms by which these polymorphisms may act, while well-powered TB GWAS and meta-analyses may continue to identify additional causal variants for TB susceptibility.

S1 Table: Variants associated with TB susceptibility.

This table represents the variants that have been implicated in TB susceptibility by association studies. The closest gene to the variant is reported.

Closest Gene	rsID	Reference
AHCYL2	rs7787531	(28,254)
AKT1	rs3730358 rs1130233	(255)
ALOX5	rs2228065	(256)
ANO9	rs7111432	(257)
ASAP1	rs4733781 rs10956514 rs1017281 rs1469288 rs17285138 rs2033059 rs12680942	(209)
ATG10	rs1864183 rs3734114	(237)
ATG16L1	rs2241880	(237)
ATG16L2	rs11235604	(237)
ATG2A	rs77228473 rs77833427	(237)
ATG2B	rs9323945 rs74719094	(237)
ATG4C	rs10493328 rs10493327	(237)
ATG5	rs2245214	(237)
ATG9B	rs61733329	(237)
BTNL10	rs1925714	(40)
BTNL2	rs3763313 rs9268494 rs9268492 rs9405098 rs3763317 rs2076530	(258)
C2CD2	rs451390	(40)
CCL1	rs10491110 rs3091324 rs2072069 rs159319 rs3138031 rs159291 rs159294 rs210837 rs159290	(259)
CCL2	rs1024611 rs4586 rs1799750	(260–265)
CCL5	rs2107538rs2280788 rs2280789	(266–269)
CD209	rs4804803	(270,271)

	rs735240	
CD43	rs4788172 rs17842268 rs12596308	(272)
CDH13	rs12386026	(28)
CHIT1	rs9943208 rs1065761	(273)
CISH	rs2239751 rs414171	
CNOT6L	rs958617	(40)
CTL4	rs231775 rs3087243 rs11571319	(274)
CTSZ	rs34069356 rs10369 rs9760 rs163790 rs163800 rs163801	(143,144,275)
CXCL12	rs2839693 rs1801157	(276,277)
CYP2C19	rs4986893 rs4244285	(264)
CYP2C8	rs1934954	(40)
CYP2E1	rs2031920 rs6413432	(264)
CYP3A4	rs28371759	(264)
CYP3A5	rs776746	(264)
CYP7A1	rs3808607	(264)
DCUN1D5	rs7947821	(40)
DMRTA1	rs586716	(40)
DSCAM	rs2837857	(40)
DUSP14	rs712039	(278)
DYNC2H1	rs12294076 rs12283022	(40)
E2F7	rs6538140	(40)
EBF1	rs10515787	(213)
EREG	rs7675690	(259)
ERP44	rs1418267	(213)
GLRX5	rs8005962	(28)
HAUS6	rs10245298	(213)
HLA-DQB1	rs9273665	(276,279–284)
IFNG	rs2430561 rs1861493 rs1861494	(26,147,274,285–289)
IFNGR1	rs2430561 rs2234711 rs1327474 rs7749390 rs4896243 rs1861493	(26,147,274,285–292)

	rs1861494 rs41401746 rs2248814 rs3729508 rs3729718 rs2274894 rs2314809 rs944722 rs7215373	
IFNGR2	rs2430561 rs2834213 rs1059293 rs1861493 rs1861494	(26,147,274,285–289,293)
IL10	rs1800896 rs1800872 rs1518111 rs1554286 rs1800870 rs1800871	(265,290,294–297)
IL10RA	rs1800896 rs1800872 rs1518111 rs1554286 rs1800870 rs3135932 rs1800871	(265,290,294–297)
IL12B	rs3212227 rs11574790 rs3212220 rs2853694	(26,298)
IL12RB1	rs11575934 rs375947 rs401502	(299–301)
IL12RB2	rs11810249	(302)
IL18	rs1946518	(303)
IL18R1	rs3755276	(304)
IL1A	rs1800587	(305)
IL1B	rs1143634	(306)
IL1RN	rs4252019	(26)
IL23R	rs11209026	(265,306)
IL23R	rs10889677 rs7518660	(307,308)
IL2RB	rs3218255	(40)
IL4	rs2243250 rs2070874	(26,306,309)
IL6	rs1800795 rs13306436 rs2069824 rs36215814 rs1800797 rs1800796 rs56077270	(305,310)
IL6R	rs1800795 rs13306436	(305,310)

	rs2069824 rs36215814 rs1800797 rs1800796 rs56077270 rs1552481 rs2229238 rs3887104 rs4379670	
IRF8	rs925994 rs11117415 rs10514611	(311)
IRGM	rs10065172 rs4958842 rs4958843 rs4958846 rs72553867 rs4958847	(34,37,312 ,313)
JAG1	rs2273061	(213)
KCND2	rs10488286	(211)
LAMP1	rs9577229	(34)
LAMP3	rs482912	(34)
LOC101929709	rs160441	(28)
LOC104923116	rs10005603	(28)
LOC105372021	rs4331426	(28)
LOC105373238	rs6676375	(40)
LOC105377003	rs2202157	(40)
LTA	rs361525 rs1800629 rs1041981 rs7791836 rs1399431 rs2229094	(314)
LTA4H	rs2540474 rs1978331	(315,316)
MARCO	rs17009726 rs2278588 rs17795618 rs1371562 rs6761637 rs2011839	(235,238)
MBL2	rs1800451 rs5030737 rs1800450 rs1800405	(317,318)
MC3R	rs3827103 rs6127698	(143,144)
MC4R	rs4257308	(28)
MIF	rs755622	(319,320)
MRC1	rs34039386 rs71497223 rs71497224 rs34284571	(321)
MTOR	rs6701524	(235,237)

	rs10492975 rs4491733rs13389814 rs12998782rs7559955	
NAA60	rs40363	(40)
NOD2	rs2066842 rs2066844 rs5743278 rs7194886	(322,323)
NOS2	rs2274894 rs7215373 rs2255929 rs944722 rs2314809 rs3729718 rs2248814 rs3729508 rs1327474 rs5030729 rs2274894 rs5030729 rs7234985 rs57234985 rs9282799 rs8073782	(324,325)
P2RX7	rs1653624 rs3751143 rs2393799	(237,326–331)
PARD3B	rs2335704	(28)
PCDH15	rs2026414	(211)
PDE11A	rs10488286	(211)
PENK	rs6985962	(213)
PKP3	rs10902158 rs7105848	(257)
PSMB8	rs2071543	(255)
PTPN22	rs2476601 rs33996649	(332,333)
PXDNL	rs7821565	(28)
SFTPA1	rs1136450 rs72659390 rs1136451	(334–336)
SFTPA2	rs17886395 rs17886221 rs1965708	(334–336)
SLC11A1	rs17235409 rs3731865 rs2276631 rs3731863 rs17221959 rs34448891 rs17229009 rs17235416 rs1816702	(336–346)
SLC22A5	rs274559	(347)

	rs274554 rs274553	
SLC40A1	rs10188230 rs10188680 rs10202029 rs11568344 rs11568350 rs116496357 rs11884632 rs13008848 rs13012833 rs1439814 rs1439816 rs1568351 rs2304704 rs2352262 rs3792079 rs3811621 rs61525883 rs6706281 rs994226	(348)
SMOC1	rs17175227	(40)
SNORD114-31	rs6575836	(40)
SOX11	rs1075309	(40)
SP110	rs2114592 rs3948464 rs9061 rs1135791 rs722555 rs11679983 rs7581442 rs7573954 rs13389060 rs11556887	(349)
SPON1	rs1819084	(40)
STAT1	rs7576984	(350)
STXBP5	rs9373523	(28)
TAP2	rs1135216	(255)
TERF2IP	rs1948632	(28)
TIRAP	rs7932766 rs8177374 rs8177400 rs3804099rs7932766 rs3802814 rs3802813	(351–353)
TLR1	rs4833095 rs76798247 rs5743810	(354,355)
TLR10	rs11096957	(240)
TLR2	rs5743708 rs3804099 rs3804100 rs3731865 rs1816702 rs4696480	(154,325,356,357)

	rs1898830 rs7932766	
TLR4	rs4986790 rs5030729 rs2248814 rs1399431 rs7791836 rs2274894 rs7215373	(358–362)
TLR6	rs4833095 rs76798247 rs5743810	(354)
TLR8	rs3764879 rs3788935 rs3761624 rs3764880	(251,253)
TLR9	rs164637 rs352143 rs352139 rs5743836	(363)
TMEFF2	rs10497744 rs1020941	(213)
TNF	rs361525 rs1800629 rs7791836 rs1399431	(305,360,364–367)
TNFRSF1A	rs361525 rs1800629 rs4149623 rs4149639 rs4149622 rs4149578 rs7791836 rs1399431	(305,360,364–368)
TNFRSF1B	rs361525 rs1800629 rs496888 rs1061624 rs7791836 rs1399431	(301,305,360,364–367,369)
TOLLIP	rs3750920 rs5743899	(370)
VDR	rs7975232 rs2228570 rs1544410 rs731236 rs11568820 rs4516035	(268,339,360,364,371–378)
VWA8	rs1900442	(40)
WIPI1	rs883541	(237)
WT1	rs2057178	(40)
XPO1	rs6545883	(28)
ZFPM2	rs17217757	(40)
ZNF229	rs1434579	(28)

S2 Table: The functional impact of proxy SNPs as determined by RegulomeDB.

A score of 1 is the lowest score possible as determined by RegulomeDB. This indicates strong evidence of a functional impact. This has been determined by a variety of comprehensive datatypes.

Chr	Base Pair (hg19)	Regulome DB Result	rsID	Consequence	Gene
chr1	63322743	1f	rs1981067	intron_variant	ATG4C
chr1	206944233	1f	rs1554286	intron_variant	IL10
chr1	63266810	1b	rs4409690	intron_variant	ATG4C
chr1	63313319	1f	rs12097658	intron_variant	ATG4C
chr1	63252829	1f	rs11208030	intron_variant	ATG4C
chr1	63252766	1f	rs6587988	intron_variant	ATG4C
chr1	63239831	1f	rs12080049	intergenic_variant	-
chr1	63341281	1f	rs12061691	intergenic_variant	-
chr1	63214147	1f	rs12143139	intergenic_variant	-
chr2	119731468	1f	rs17795618	intron_variant	MARCO
chr2	119726467	1f	rs17795448	intron_variant	MARCO
chr2	119740290	1f	rs7599352	intron_variant	MARCO
chr2	119732042	1f	rs11693199	intron_variant	MARCO
chr2	219249013	1f	rs2276631	synonymous_variant	SLC11A1
chr2	231050715	1f	rs3948464	missense_variant	SP110
chr2	61613261	1f	rs2593620	intron_variant	USP34
chr2	219229147	1f	rs4674297	intron_variant	C2orf62
chr2	231038403	1f	rs13026323	downstream_gene_variant	SP110
chr2	61776902	1f	rs778752	intergenic_variant	-
chr2	61659199	1f	rs2463100	intron_variant	USP34
chr2	61735446	1f	rs766448	intron_variant	XPO1
chr2	61581890	1f	rs778143	intron_variant	USP34
chr2	61604383	1f	rs2694643	intron_variant	USP34
chr2	61417617	1f	rs777591	downstream_gene_variant	AHSA2
chr2	61412559	1f	rs777585	intron_variant	AHSA2
chr2	61405795	1f	rs2600665	5_prime_UTR_variant	AHSA2
chr2	61554289	1f	rs778157	intron_variant	USP34
chr3	52252969	1d	rs352162	downstream_gene_variant	ALAS1
chr3	50693998	1d	rs17051043	intergenic_variant	-
chr3	52264907	1f	rs352143	synonymous_variant	TWF2
chr3	50647888	1f	rs2239751	intron_variant	CISH
chr3	50645158	1f	rs2239753	synonymous_variant	CISH

chr3	50689039	1f	rs12489607	downstream_gene_variant	MAPKAPK3
chr3	50694626	1f	rs17051045	intergenic_variant	-
chr3	50698155	1f	rs12492982	intergenic_variant	-
chr3	50645413	1f	rs2239752	synonymous_variant	CISH
chr3	52238677	1b	rs352166	intron_variant	ALAS1
chr3	50668532	1d	rs616689	intron_variant	MAPKAPK3
chr3	52238656	1f	rs352167	intron_variant	ALAS1
chr3	52261031	1f	rs187084	downstream_gene_variant	TWF2
chr3	52247314	1f	rs164640	intron_variant	ALAS1
chr3	52220203	1f	rs614288	intergenic_variant	-
chr3	52236762	1f	rs352169	intron_variant	ALAS1
chr3	50629978	1f	rs2239750	regulatory_region_variant	-
chr3	50686517	1f	rs2170840	3_prime_UTR_variant	MAPKAPK3
chr3	50683977	1f	rs876104	intron_variant	MAPKAPK3
chr3	50685642	1f	rs9879397	3_prime_UTR_variant	MAPKAPK3
chr3	50556581	1f	rs12491812	intergenic_variant	-
chr3	52287468	1b	rs11717383	downstream_gene_variant	WDR82
chr3	52378540	1f	rs13060192	missense_variant	DNAH1
chr3	52435860	1f	rs123598	downstream_gene_variant	DNAH1
chr3	52441606	1f	rs419604	upstream_gene_variant	PHF7
chr3	52492085	1f	rs4687614	intron_variant	NISCH
chr3	52493275	1f	rs9855470	intron_variant	NISCH
chr3	52506491	1f	rs6445486	intron_variant	NISCH
chr3	52513027	1f	rs9867823	intron_variant	NISCH
chr3	52524574	1f	rs6810027	upstream_gene_variant	STAB1
chr3	52442354	1f	rs123602	upstream_gene_variant	PHF7
chr3	52288945	1f	rs1060330	3_prime_UTR_variant	WDR82
chr3	52338852	1f	rs9311474	intergenic_variant	-
chr4	38777236	1f	rs10856839	5_prime_UTR_variant	TLR10
chr4	38777173	1f	rs10856838	synonymous_variant	TLR10
chr4	154927577	1f	rs11934607	intergenic_variant	-
chr4	38787216	1f	rs12233670	upstream_gene_vari	TLR10

				ant	
chr4	38803063	1f	rs5743592	intron_variant	TLR1
chr4	38806827	1f	rs5743557	upstream_gene_variant	TLR1
chr4	38820986	1f	rs7696175	downstream_gene_variant	TLR6
chr4	38833595	1f	rs6531668	upstream_gene_variant	TLR6
chr5	81555167	1b	rs10036937	downstream_gene_variant	ATG10
chr5	81547518	1f	rs4703876	intron_variant	ATG10
chr5	81541238	1f	rs2195448	intron_variant	ATG10
chr5	81438144	1f	rs4703535	intron_variant	ATG10
chr5	131720070	1f	rs274559	intron_variant	SLC22A5
chr5	131656517	1f	rs272842	intron_variant,non_coding_transcript_variant	AC034220.3
chr5	131742228	1f	rs6596075	upstream_gene_variant	C5orf56
chr5	81402224	1a	rs12515069	-	-
chr5	131714409	1f	rs274567	intron_variant	SLC22A5
chr5	131665378	1f	rs272889	intron_variant,non_coding_transcript_variant	AC034220.3
chr5	131719228	1f	rs274561	intron_variant	SLC22A5
chr5	131675864	1f	rs272872	intron_variant,non_coding_transcript_variant	AC034220.3
chr5	131653925	1f	rs156322	intron_variant,non_coding_transcript_variant	AC034220.3
chr5	131722951	1f	rs274555	intron_variant	SLC22A5
chr5	81356055	1f	rs2860007	intron_variant	ATG10
chr5	81327042	1f	rs1485587	intron_variant	ATG10
chr5	81388861	1f	rs324913	intron_variant	ATG10
chr7	22767433	1d	rs2069832	intron_variant	IL6
chr7	22741459	1f	rs4321884	intergenic_variant	-
chr7	22768124	1f	rs1474347	intron_variant	IL6
chr11	126135785	1b	rs11220432	intron_variant	SRPR
chr11	1296649	1f	rs3168046	3_prime_UTR_variant	TOLLIP
chr11	126148160	1f	rs11220437	upstream_gene_variant	TIRAP
chr11	126100932	1f	rs618176	intron_variant	FAM118B

chr11	126151422	1f	rs563011	upstream_gene_variant	TIRAP
chr12	96403894	1f	rs7296106	intron_variant	LTA4H
chr16	3509056	1d	rs40363	intron_variant,non_coding_transcript_variant	LA16c-306E5.3
chr16	50724789	1f	rs4785448	regulatory_region_variant	-
chr16	50714029	1f	rs7202124	intron_variant	SNX20
chr16	50755709	1d	rs10521209	intron_variant	NOD2
chr16	50751398	1f	rs748855	intron_variant	NOD2
chr16	50751787	1f	rs1861758	intron_variant	NOD2
chr16	50745583	1f	rs1861759	synonymous_variant	NOD2
chr17	34139218	1d	rs11868785	intron_variant	TAF15
chr17	34226272	1f	rs4796128	intron_variant,non_coding_transcript_variant	AC015849.16
chr17	34207003	1f	rs2280789	intron_variant	CCL5
chr17	34182341	1f	rs2306630	missense_variant	C17orf66
chr17	34164527	1f	rs4251769	intron_variant	TAF15
chr17	34188496	1f	rs9303692	intron_variant	C17orf66
chr17	34142362	1f	rs4251719	upstream_gene_variant	AC015849.13
chr17	34146818	1f	rs4251737	intron_variant,non_coding_transcript_variant	AC015849.13
chr17	66449122	1f	rs883541	intron_variant	WIPI1
chr17	34129067	1f	rs6505496	regulatory_region_variant	-
chr17	34211460	1f	rs4796123	upstream_gene_variant	CCL5
chr17	66439605	1f	rs2909207	intron_variant	WIPI1
chr17	34123597	1f	rs4795090	upstream_gene_variant	MMP28
chr17	66447073	1f	rs2302783	intron_variant	WIPI1
chr17	35848659	1f	rs853196	upstream_gene_variant	DUSP14
chr17	35763863	1f	rs829161	upstream_gene_variant	TADA2A
chr17	34256906	1f	rs2526327	intron_variant	RDM1
chr17	66427696	1f	rs6501468	intron_variant	WIPI1
chr19	7809327	1f	rs8105572	intron_variant	CD209
chr20	57571763	1f	rs9760	downstream_gene_variant	NELFCD

chr20	57587771	1d	rs4812048	intergenic_variant	-
chr20	57564070	1f	rs2273360	synonymous_variant	NELFCD
chr20	57559630	1f	rs12480262	intron_variant	NELFCD
chr21	34792910	1f	rs2834213	intron_variant	IFNGR2
chr21	34806288	1f	rs2284555	intron_variant	IFNGR2
chr21	34813562	1f	rs6517173	downstream_gene_variant	IFNGR2
chr21	34815794	1f	rs4816455	intergenic_variant	-
chr21	34821570	1f	rs1044218	3_prime_UTR_variant	TMEM50B
chr21	34837144	1f	rs2186280	intron_variant	TMEM50B
chr21	34815979	1f	rs12626735	intergenic_variant	-
chr21	34796886	1f	rs2834215	intron_variant	IFNGR2
chr22	37544486	1b	rs3218255	intron_variant	IL2RB
chr22	37544245	1b	rs3218258	intron_variant	IL2RB
chr22	37558356	1f	rs2051582	upstream_gene_variant	RP1-151B14.6
chr22	24233998	1f	rs875643	downstream_gene_variant	AP000350.4

Chapter 5

Signals of positive selection in immune response genes of an admixed southern African population

Caitlin Uren[1], Eileen G. Hoal[1], Gerard Tromp[1], Paul D. van Helden[1], Brenna M Henn[2,3], Marlo Möller[1,3]

[1]SA MRC Centre for TB Research, DST/NRF Centre of Excellence for Biomedical TB Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, 8000, South Africa
[2]Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794
[3]Co-senior authors

Proceedings of the Royal Society B (Impact Factor 4.823)

Submitted

Abstract

The recent availability of exome sequence data and improved statistical analyses has facilitated investigations into the extent of selective pressure due to pathogens in numerous populations. However, there have been very few studies investigating this in southern African populations where it is hypothesized that the selective pressure due to tuberculosis and smallpox was vast. Here, we perform a positive selection scan using the population branch statistic to identify signals of selection associated with viral or bacterial immune response in the highly admixed South African Coloured (SAC) population. Using ancestral populations from the 1000 Genomes Project for comparison, we found SAC-specific signals of selection in genes associated with focal adhesion and extra-cellular matrix receptor interactions. These signals were located, amongst other, in the *IL17RA, TIMM44, ACSM5, EDARDD, FRAS1* and several *SLC* genes. In addition, we identified a selection hotspot on chromosome 11q23.3 which is associated with abnormal immune system physiology. This study not only confirms the role that natural selection plays in shaping human immunity, it also highlights particular novel pathways associated with both viral and bacterial immune response that could be investigated further, particularly with respect to tuberculosis susceptibility in southern Africa.

Introduction:

Southern Africa's population history reaches back at least 100 000 years, beginning with the origin of modern humans (60,379,380). Genomic studies have concluded that the KhoeSan, who reside in southern Africa, descend from the earliest population divergence event amongst modern humans (64,67–69). The KhoeSan were largely isolated from other populations until ~2 000 years ago when numerous populations migrated to some parts of the region. These included East African pastoralists (~2 000 years ago), Bantu-speaking farmers (~1 500 years ago), Arab and Indian traders (~1 200 years ago), Europeans (~350 years ago) and Asians (~350 years ago) (183,381). As these human migrations into southern Africa occurred, populations may have adapted to novel environments (drought, parasites, different subsistence strategies and cultures, etc.). In addition, resident populations would need to adapt to the introduction of foreign pathogens by the migrants. Individuals from indigenous populations were not always able to mount an effective immune response to foreign pathogens such as smallpox and tuberculosis (TB), and were therefore highly susceptible to these diseases, resulting in immense loss of life (50).

Foreign infectious diseases in southern Africa

It is hypothesized that the variola virus (the causal agent of smallpox) was introduced into sub-Saharan Africa in the 12th century by Arab and Indian traders on the east coast of Africa (382,383). It then spread westwards as Bantu-speaking populations (such as the Tsonga) moved along the trade routes from the east coast (Maputo, formerly Delagoa Bay) to the interior of southern Africa (381). However, smallpox had not reached the tip of southern Africa until the Dutch settlers arrived from India in the 17th century (382,383). It then spread northwards (**Figure 1**). There were numerous smallpox epidemics in south-western Africa from that point on, with the most devastating documented in 1713 and 1755. It is estimated that between 50–90% of all indigenous people in the Cape region of South Africa died during these epidemics (50,384). As worldwide inoculations against the variola virus increased in frequency, the smallpox incidence rate decreased to the extent that there are no longer any known cases of smallpox in humans (385).

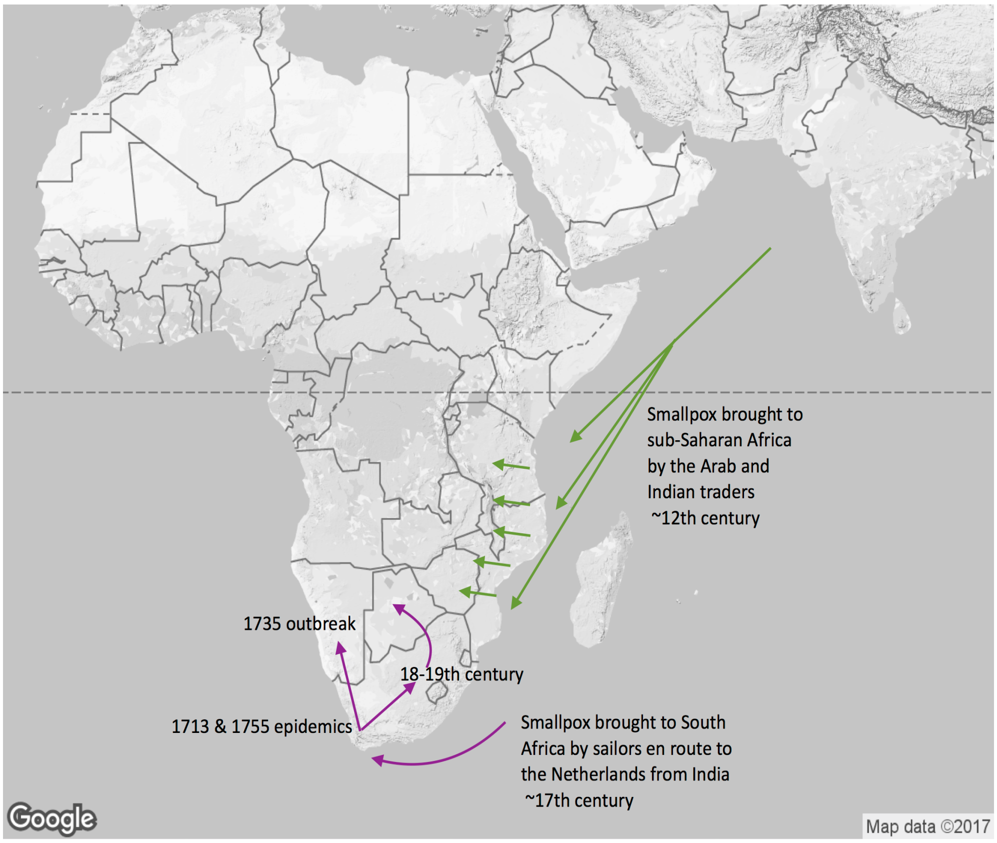


Figure 1: The arrival and spread of smallpox across Africa (based on diagram by Fenner, 1988).

Smallpox was brought to the east coast of Africa in the 12th century by Arab and Indian traders. The disease arrived at the southern tip of Africa from the 17th century onwards with resulting outbreaks in 1713 and 1755 (382).

According to historical records, TB was present in southern Africa at relatively high frequencies shortly after the Dutch arrival in the 17th century (50). Due to a lack of documentation prior to European arrival, it is not known whether TB was present in southern Africa prior to this period. Some evidence indicates that the KhoeSan and Bantu-speaking populations were not previously exposed to the disease and therefore were "virgin soil" for TB (13,14). In contrast, other studies have proposed an eastern African origin for TB and have suggested that the high incidence and mortality rates on the African continent are due to the introduction of European strains (9,10,386). Once introduced, TB (or phthisis as it was known then) resulted in catastrophic mortality in the indigenous populations, particularly those already compromised due to poor diet and living conditions (49,50). The Medical Officer of Health at the Cape Colony in the early 19th century said that "of all the diseases attacking the natives and Coloured, tuberculosis is by far the most important" (387). Over the next 200 years, TB incidence in southern Africa rapidly increased, in contrast to a significant drop in incidence rates in Europe (5,388). It has since been hypothesized that Europeans exhibit higher resistance to TB relative to southern African populations, particularly the SAC, who exhibit a high degree of non-HIV associated susceptibility (18,40,388).

Admixture in southern Africa and its effect on disease susceptibility

Although it is clear that one consequence of migrations into the area was the arrival of foreign pathogens, by far the greatest genetic consequence was admixture between migrant and resident populations. This can be seen in present day populations where it has been shown that some KhoeSan populations, particularly those in South Africa have ~15% Bantu-speaking and ~10% European ancestry (55,63,72,83,94). In contrast, South African Bantu-speaking populations have ~80% central/western African ancestry and ~20% KhoeSan ancestry (94). Although not as well characterized, the genetic makeup of European populations in southern Africa (particularly the Afrikaners) also follows this pattern of admixture, with genomes comprising approximately 10% African ancestry (97). The SAC population is a direct result of gene flow between these ethnically diverse ancestral populations (European, Bantu-speaking and KhoeSan) (51–53). In addition, they carry ancestry from southern and south-eastern Asia due to the slave trade at the Cape of Good Hope from the 17th century until the 19th century(51). The combination of these 5 ancestries has led to a uniquely admixed population (51–53).

The SAC population predominantly resides in the Western Cape of South Africa, where at the time of sampling, the TB incidence was approximately 1 000/100 000 (5) despite a relatively low prevalence of HIV (~2%) (224). The HIV prevalence has since increased to ~3% (389). Studies have shown that the SAC population are highly susceptible to progress to active TB rather than remaining infected (17,40). It has been suggested that one of the risk factors associated with this heightened susceptibility is African ancestry, in contrast to European ancestry that has a protective function (17,18). This is reflective of the relatively high level of TB susceptibility in African populations and contrasting resistance in Europeans. It has been proposed that one of the main reasons for these contrasting susceptibility profiles is selection in immune response genes (41,390,391).

Selection signals in immune response genes

Selection plays an important role in an individual's ability to correctly and efficiently respond to pathogens (388,390,392,393). Over time, advantageous alleles will be selected for in a population and deleterious alleles will be removed. These advantageous alleles confer an increase in reproductive fitness through an improved immune system. Signals of positive selection in genes influencing immune response to TB have been found in European populations (41,46,388). In addition, signals of selection were found in the Human Leukocyte Antigen (HLA) region of Native American exomes, correlating with the selective pressure exerted by European colonization in the 1800's (394). These studies demonstrate that selection on immune-related genes can take place after exposure to foreign pathogens, particularly if there is a significant loss of life, as was the case in the Cape in the 18th century. As with most scientific fields, selection studies based on African populations are sparse and this includes investigations into whether natural selection took place in southern Africa due to the selective pressure exerted by foreign pathogens.

Studies investigating selection in southern African populations have not focused on selection in immune response genes until very recently (72,93,391). Chimusa et al. (2015) identified genes under selection in the ≠Khomani San which are associated with high-risk diseases (such as malaria and common viruses) (93). These regions were identified by looking at regions with excess western African and European ancestry as compared to global ancestry proportions. Although this is one way to identify genes under selection, the lack of available Next Generation Sequencing (NGS) data in southern Africa as well as the complex genetic history of the SAC, complicates most analyses. A recent analysis using high-density SNP array data showed an increase in selection signals in immune response genes compared to genome-wide expectations, in addition to specific signals of selection in immune response genes, but largely discounted the role of fine-scale admixture in the analyses by not including it as a confounding factor (391). We have therefore utilized high-coverage exome sequence data and adapted the application of the F_{st} based population branch statistic (PBS) to allow for selection scans in highly admixed populations such as the SAC population (395).

Based on the previous research by Chimusa et al. (2015) and Owers et al. (2017), the complex genetic history of the SAC and the strong selective pressure exerted by smallpox and TB, we hypothesize that selection on immune response genes played a significant role in the SAC population's adaptation to foreign pathogens. In particular, we hypothesize the presence of multiple signals of positive selection in immune response genes related to viral and bacterial immunity.

Methods:

Population samples

Twenty individuals from the Western Cape, self-identifying as belonging to the South African Coloured (SAC) population, were included in this study. These study participants gave written informed consent for DNA extraction, and ethics approval was obtained from Stellenbosch University (Approval numbers N095/072 and N09/07/185). Twenty unrelated individuals from the British, Yoruban, Gujarati and Vietnamese populations were selected at random from the 1000 Genomes dataset (396). Twenty genetically homogenous unrelated individuals from the ≠Khomani population were selected from the dataset used by Martin et al. (2017) (under review).

DNA extraction from samples collected from the SAC population

DNA from SAC individuals was extracted from blood collected using the Nucleon BACC Genomic Extraction kit (illustra, Buckinghamshire, UK) according to the manufacturer's protocol. DNA concentration and quality was determined using the Nanodrop 2000c (Thermo Scientific, Denver, USA).

Library preparation and exome sequencing of SAC individuals

The genomic DNA was sonicated to fragment the DNA to ~150bp. Quality and size was assessed by the Bioanalyzer 2100 and DNA 1000 chip and reagent kit (Agilent Technologies Santa Clara, California, USA). Targeted enrichment of the SAC DNA samples was performed using the Nextera XT enrichment kit (> 20 000 genes at 40–60X read depth) (Illumina, San Diego, California, USA). Library preparation was completed according to standard protocols. Pair-end sequencing was performed at Christian-Albrecht's University of Kiel (CAU sequencing Kiel, Germany) using the Illumina HiSeq 2500 (Illumina, San Diego, CA, USA). A pipeline following the 1000 Genomes Consortium best-practices was used to map and pair reads using the Burrows-Wheeler Aligner (bwa) (397) against the hg19 human reference genome (http://genome.ucsc.edu/) (398) and duplicate sequence reads were identified using Picard tools (http://picard.sourceforge.net/). The Genome Analysis Tool Kit (399) (GATK) was used for the recalibration of base quantiles, indel realignment and variant calling using the variant quality score recalibration according to GATK Best Practices recommendations (400).

Data filtering

The exome data were filtered using *vcftools* (401) according to the pipeline shown in **Figure S1**. Each dataset was filtered so that only sites within the Agilent 44M capture remained. Vcf-merge was used to merge the individual .vcf files, including only those sites that passed sequencing quality control (per site quality score, GC content over all sequences etc.). Insertions, deletions and monomorphic sites were removed. No related individuals were included, as determined by Identity by Descent analysis in *plink* with an IBD cut-off of 0.1.

Statistical calculations

Minor allele frequencies and heterozygosity were calculated using *pseq (402)*. Pairwise F_{st} values per SNP were then calculated for all sites in the merged dataset (as described above) according to the formulae as specified in Weir et al. (1984) (403). All negative and missing F_{st} values were assigned to 0. Negative F_{st} values are largely a by-product of the comparatively small sample size (404). Pairwise F_{st} values per SNP were converted to a PBS value using the formula as outlined in Yi et al. (2010) (395). The PBS is computed as a product of allele frequency changes at a given site relative to the deviation from the other populations. PBS values only for comparisons including the SAC population were calculated. PBS values were calculated for every 3-way population comparison i.e. SAC-SAN-YRI, SAC-GBR-GIH etc. This was performed at a genome-wide level. All negative PBS values were assigned to 0. A 95% empirical cut-off was calculated for every population comparison and values higher than this cut-off were deemed significant. Since the purpose of the study was to identify SNPs that are under selection in the SAC after correcting for admixture, only sites with PBS values that exceed the threshold for each population comparison are reported. In this way, any false positives due to an increase in a specific ancestry are removed. The statistical workflow is summarized in **Figure S2**.

Since negative/NA F_{st} values were assigned to 0, the type I error rate was lower than expected if these values were excluded from further analysis (405). In addition, only sites with PBS values across all population comparisons are reported. For this reason, a more lenient empirical cut-off of 95% was chosen. The estimation of false discovery rates was not possible as a truly representative simulation of the SAC population cannot be performed for comparison. This is due to the lack of a SAC-specific recombination map and an incomplete genetic history.

Functional prediction and gene over-representation analysis

SNPNexus was used both to map the variants using NCBI RefSeq (406) as well as to determine the consequence to the gene or protein with the SIFT (407) and POLYPHEN (408) score output. WebGestalt with default parameters was used to identify genes (and associated pathways and phenotypes) that were over-represented in the candidate list as compared to the expectations based on the reference dataset used (409). STRING (v10) was used to identify potential protein-protein interactions (410).

Results:

In this study, we aimed to identify signals of recent positive selection in the SAC population by utilizing the PBS formula as described in Yi et al. (2010). These values were corrected for admixture by analysing SNPs with statistically significant PBS values across all 3-way ancestral population comparisons (n=10). High-coverage exome sequence data from 5 ancestral populations (British, Vietnamese, Gujarati, Luhya and ≠Khomani populations) were used in these comparisons. These specific ancestral populations were chosen as they were the best proxy ancestral populations for the SAC, as determined from recent genetic studies and historical records (51–53,55,411,412). Once the genetic data from each of these ancestral populations were merged with that of the SAC population, a total of ~310 000 SNPs were included in the PBS, over-representation and functional analysis with the final goal of identifying novel genes and pathways associated with immune response to pathogens.

Genes with the strongest signals of positive selection

PBS values of a total of 183 SNPs surpassed the 95% empirical cut-off across all 10 population comparisons. These signals are summarized in **Table S1**. The distribution of PBS values across the genome for all population comparisons are summarized in **Figure 2** and **Table S2**.

The SNP (rs12880814) which yielded the highest PBS value across all population comparisons was found in the *C14orf180* gene [or Novel Nutritionally-Regulated Adipose and Cardiac-Enriched gene (*NRAC*)], which is associated with a wide array of diseases, predominantly metabolic (Type 2 Diabetes) and cardiovascular (Coronary Artery Disease).

When the top 10 variants were ranked according to PBS value in each population comparison, rs12880814 was present in the top 10 in 3 out of the 10 population comparisons (SAC-SAN-GIH, SAC-GBR-GIH and SAC-SAN-GBR) (**Table S2**). We also observe a second gene involved in metabolic and cardiac conditions (*FMN2)*. A nonsynonymous missense variant in this gene (rs12732924) was present in the top 10 PBS values in 7 out of 10 population comparisons (PBS values ranged from 0.17–0.34) (**Table S2**).

Signals of positive selection in immune response genes

We identified signals of positive selection in several immune response genes that govern both viral and bacterial immune response; *IL17RA* (**Figure 3A**), *TIMM44, ACSM5*, *EDARDD,* multiple *SLC* genes*, LRIT3, IMPG2, LPP, FRAS1* and *CBL*. The majority of these signals were due to single nonsynonymous variants. A number of these immune genes were present in the top hits of each population comparison (**Table S2**). The immune response associated *FRAS1* gene was one of only 2 genes (the other was *LGR6*) that showed a signal of

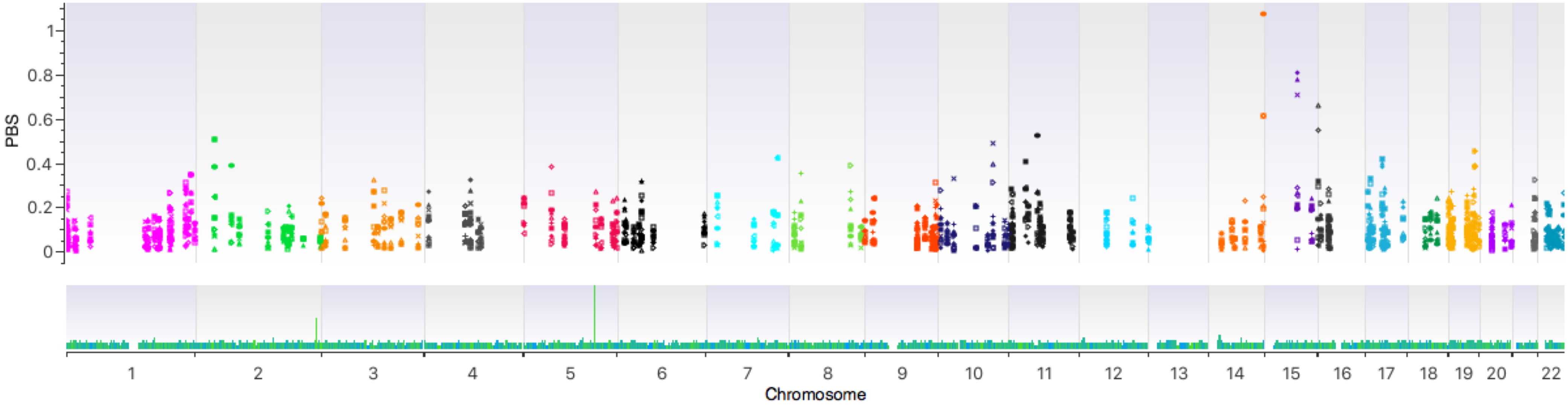


Figure 2: Distribution of PBS values across the genome with the respective gene density.

Gene density was determined by RefSeq 105v2. PBS values from all population comparisons are plotted. Only sites that surpassed the population comparison specific empirical cut-off of 95% are shown. Exact values and the associated population comparisons are available in Table S2.

selection due to more than 2 variants. The 3 variants (all present in protein-coding regions of *FRAS1*) which had significant PBS values across population comparisons were rs11933630 (a synonymous SNP), rs35933858 (a nonsynonymous SNP; the associated amino acid change was expected to be tolerated by SIFT/POLYPHEN) and rs79443837 (a nonsynonymous SNP; the associated amino acid change also expected to be tolerated by SIFT/POLYPHEN).

We analysed the predicted function of the 183 candidate variants *in silico* by looking at the possible effect of the associated amino acid changes (by using SIFT/POLYPHEN) (**Table S1**); there were 10 variants under positive selection that were expected to have a functional effect. Genes in which variants under selection were shown to be damaging as well as being associated with immune response include those in *GBGT1, NR4A1* and *TMIGD2*. The respective rsIDs of the variants as well as the SIFT/POLYPHEN scores can be seen in **Table S1**. Although SIFT/POLYPHEN predicted these variants to have a damaging effect on protein function, this does not necessarily translate to a negative impact on the patients' phenotype.

Selection hotspot on chromosome 11q23.2

Upon closer investigation of the genes under selection as well as the phenotypes that they govern, we hypothesize that there is a selection hotspot on chromosome 11q23.3 (**Figure 3B**).

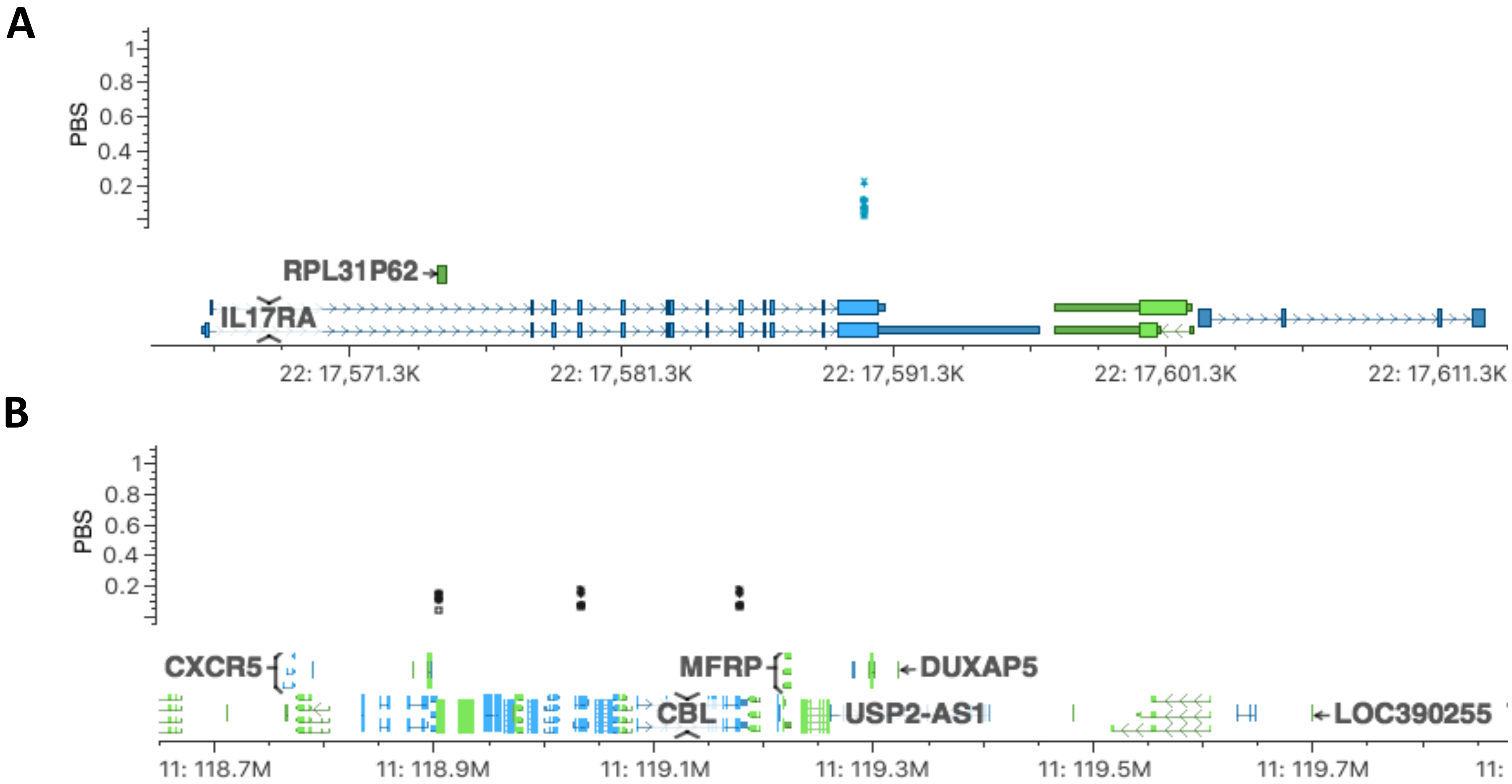


Figure 3: PBS values mapped to each value's respective gene for **A**) *IL17RA* **B**) a selection hotspot on chromosome 11q23.3.

Genomic positions and mapping was performed using RefSeq 105v2. PBS values are from the all population comparisons. Only sites that surpassed the empirical cut-off are shown.

This was identified by combining gene over-representation analysis with protein-protein interaction networks to investigate whether there were interacting genes enriched for a particular phenotype. Gene over-representation analysis suggested that this was the case for the genes associated with an "abnormal immune system" phenotype (HP:0002715) (*p* value = $6x10^{-3}$) and "abnormal immune system physiology" phenotype (*p* value = $4x10^{-3}$) (HP:0010978) (**Table 1**). The protein products of these genes on chromosome 11q23.3 which were associated with these phenotypes, were shown to directly interact (**Figure 4**). We therefore hypothesize that these genes (*SLC37A4, CBL, ABCG4* and *UBASH3B*) form part of a selection hotspot on chromosome 11q23.3 which is associated with the immune response.

Table 1: WebGestalt (409) phenotype over-representation results.

Geneset	Description	Number of genes in category	Observed	Expected	Ratio of enrichment	*p* Value	Genes
HP:0002475	Myelomeningocele	14	3	0.108	27.667	0.000	*MESP2; FRAS1; AXIN1*
HP:0011830	Abnormality of oral mucosa	161	6	1.247	4.812	0.001	*EDARADD; COL5A1; COL17A1; IL17RA; SLC37A4; SAMD9*
HP:0002435	Meningocele	33	3	0.256	11.737	0.002	*MESP2; FRAS1; AXIN1*
HP:0100585	Telangiectasia of the skin	41	3	0.318	9.447	0.004	*COL5A1; SLC37A4; POLE*
HP:0010978	Abnormality of immune system physiology	919	14	7.118	1.967	0.004	*EDARADD; COL5A1; MESP2; IL17RA; SLC37A4; LRIT3; IMPG2; POLE; SAMD9; SLCO2A1; TCF3; FRAS1; AXIN1; DNAH11*
HP:0006480	Premature loss of teeth	46	3	0.356	8.420	0.005	*EDARADD; COL5A1; IL17RA*
HP:0002715	Abnormality of the immune system	1188	16	9.201	1.739	0.006	*EDARADD; COL5A1; MESP2; IL17RA; SLC37A4; LRIT3; LPP; IMPG2; POLE; SAMD9; SLCO2A1; TCF3; FRAS1; AXIN1; CBL; DNAH11*
HP:0000704	Periodontitis	16	2	0.124	16.139	0.006	*EDARADD; COL5A1*
HP:0011355	Localized skin lesion	468	9	3.625	2.483	0.007	*CENPE; CHL1; EDARADD; COL5A1; COL17A1; IL17RA; SLC37A4; TCF3; CBL*
HP:0010651	Abnormality of the meninges	53	3	0.410	7.308	0.008	*MESP2; FRAS1; AXIN1*

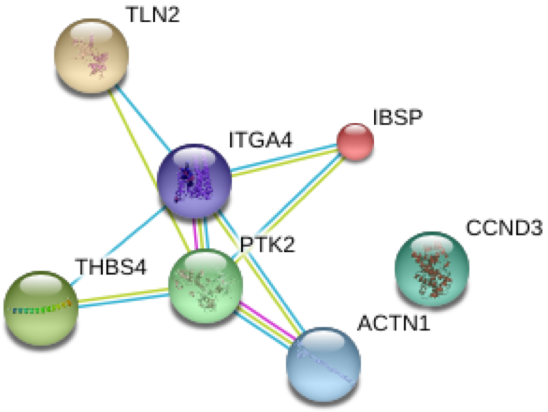


Figure 4: FAK (or PTK2) STRING (410) network showing interactions with genes under selection.

Light blue or pink lines depict a known interaction from curated databases or experimentally determined respectively. Green and black lines depict an interaction as hypothesized by text mining and co-expression respectively. Purple lines depict protein homology.

Over-representation analysis of genes associated with KEGG (413) pathways

In addition to the phenotype enrichment analysis, we investigated whether there was an over-representation of genes associated with specific KEGG (Kyoto Encyclopaedia of Genes and Genomes) (413) pathways. The analysis suggested that genes involved in the focal adhesion (p value = 6x10^{-3}), protein digestion and absorption (p value = 6x10^{-3}), ECM-receptor interaction (p value = 2.9x10^{-2}), and DNA replication (p value = 3.5x10^{-2}) KEGG pathways (413) were overrepresented in the dataset of genes under selection (**Table 2**). Genes associated with the ECM-receptor interaction pathway were also associated with the focal adhesion pathway.

Table 2: WebGestalt (409) KEGG (413) pathway over-representation results.

Geneset	Description	Number of genes in category	Observed	Expected	Ratio of enrichment	*p* value	Genes
hsa04510	Focal adhesion	203	6	1.658	3.619	0.006	*IBSP; ITGA4; THBS4; TLN2; ACTN1; CCND3*
hsa04974	Protein digestion and absorption	90	4	0.735	5.442	0.006	*COL5A1; COL17A1; COL22A1; SLC1A5*
hsa04512	ECM-receptor interaction	82	3	0.670	4.479	0.029	*IBSP; ITGA4; THBS4*
hsa03030	DNA replication	36	2	0.294	6.802	0.035	*POLA2; POLE*
hsa05166	HTLV-I infection	258	5	2.107	2.373	0.059	*NFATC4; POLE; TCF3; TLN2; CCND3*
hsa04310	Wnt signaling pathway	143	3	1.168	2.569	0.111	*NFATC4; AXIN1; CCND3*
hsa00603	Glycosphingolipid biosynthesis - globo and isoglobo series	15	1	0.123	8.163	0.116	*GBGT1*
hsa04520	Adherens junction	74	2	0.604	3.309	0.122	*ACTN1; FARP2*
hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	74	2	0.604	3.309	0.122	*ITGA4; ACTN1*
hsa04390	Hippo signaling pathway	154	3	1.258	2.385	0.131	*FRMD1; AXIN1; CCND3*

Discussion:

In this study, we aimed to detect signals of very recent positive selection in immune response genes of the SAC population. We hypothesize that selection acted on the newly founded admixed population and would have occurred over the past 400 years. The selection coefficient required for such evolutionary changes would have had to be extremely strong.

This is consistent with the severe mortality rates (due to smallpox and TB) seen in southern Africa in the 17th and 18th centuries, particularly in the Western Cape (50,81,382).

The SAC population was founded by five distinct ancestral populations, all of which were previously exposed to different selective pressures (51–53,55). Therefore, we need to consider that some of these ancestral populations may have harboured inherent adaptions to pathogens prior to their arrival in southern Africa. The European and Asian populations (in contrast to the KhoeSan and Bantu-speaking populations) were exposed to both TB and smallpox prior to their arrival in southern Africa (382,388). It is however unlikely that there would have been sufficient time for the populations to fully adapt to such a selective pressure before migration to southern Africa (382,388). This is supported by the apparent decline in TB mortality and incidence rates from the late 18th century onwards. However, *if* there was a similar selection coefficient acting on these populations, favourable alleles would have nearly reached fixation. If this was the case, we would not be able to detect such signals in the SAC, as we accounted for admixture from each of the source populations; strong positive selection in one of the source populations and in the SAC at the same locus would result in a low F_{st} value.

Therefore, this study aimed to detect recent signals of positive selection in the SAC population, which are not due to “adaptive introgression”. To do so, exome sequencing data from the SAC population was compared to that of their ancestral populations to identify regions where natural selection has acted, with a focus on regions associated with immune response to viral and bacterial pathogens. We find evidence for novel associations between this phenotype as well as genes and associated pathways.

The results presented here, as well as in previous studies, suggest that the extent of positive selection was greater in immune response genes than in other genes throughout the genome (391). In addition, several novel immune-related genes were found to be under positive selection. However, the selection signals presented here are not associated with genes classically implicated in immune response pathways (HLA, Interleukin regions and Toll-Like Receptors) (156,215,216,308). Due to the complexity associated with immune response and the lack of investigations on southern African populations, it is not surprising that the genes classically associated with immune response did not feature in the results we report here. This was also found in one of the only other studies investigating selection in immune response genes in southern African populations (391). When comparing the specific signals of selection we found in the SAC with signals found in the ≠Khomani San by Owers et al. (2017), there is little overlap. Our results did however yield a variant within 700 kbp of the start of the Fc-receptor-like cluster which also provided multiple signals of positive selection in the ≠Khomani (391). Furthermore, similar regions were consistent across both studies (e.g. chr4: 100–105 Mbp and chr6: 30–35Mbp), but the specific genes reported to be under selection were not consistent across studies.

An explanation of the apparent lack of overlap between our study and Owers et al. (2017) can be partially explicated by population history and the history of foreign pathogens in southern Africa. At the time of the Dutch arrival in the Cape of Good Hope (now Cape Town) in the 17th century, it was mainly the Khoekhoe who inhabited the area and are thought to have contributed substantially to the founding of the SAC population (51–53,183). The ≠Khomani San resided further north in the country, near the Kalahari Desert. Due to their geographic isolation, the ≠Khomani may have been less exposed to foreign pathogens than the SAC population. In contrast, the SAC population resided in the Western Cape area which bore the brunt of the high TB and smallpox incidence and mortality rates and therefore experienced a greater pressure to adapt. We therefore propose that the two populations underwent convergent evolution and that differing sets of genes were selected for in each population.

Impact of signals of positive selection on immune response

We find several immune-related functions under positive selection. It is therefore interesting that the top PBS score was for a gene previously not identified as immune-related (*C14orf180*/*NRAC*). In addition, a nonsynonymous variant in a protein coding region of *IL17RA* yielded a statistically significant PBS value suggesting the presence of positive selection. Variants in this gene have previously been implicated in both viral and bacterial susceptibility, although the variant identified here (rs41323645), is a novel candidate (240). In addition, evidence for a selection hotspot on chromosome 11q23.3 suggests that genes involved with an “abnormal immune system” and an “abnormal immune system physiology” were selected for. It is not, however, clear whether “abnormal” is an indication that the immune system is more or less effective as several diseases associated with this large phenotype ontology term are auto-inflammatory conditions as well as immunodeficiencies. As several of the genes presented here may function in inhibitory pathways, it is not known whether the variants under selection in these genes contribute to up- or down-regulation of gene expression. Using TB as an example, a heightened immune response may prove to be more effective in controlling bacterial load, but the over-stimulation may have deleterious (pathological) effects due to excess inflammatory response (414).

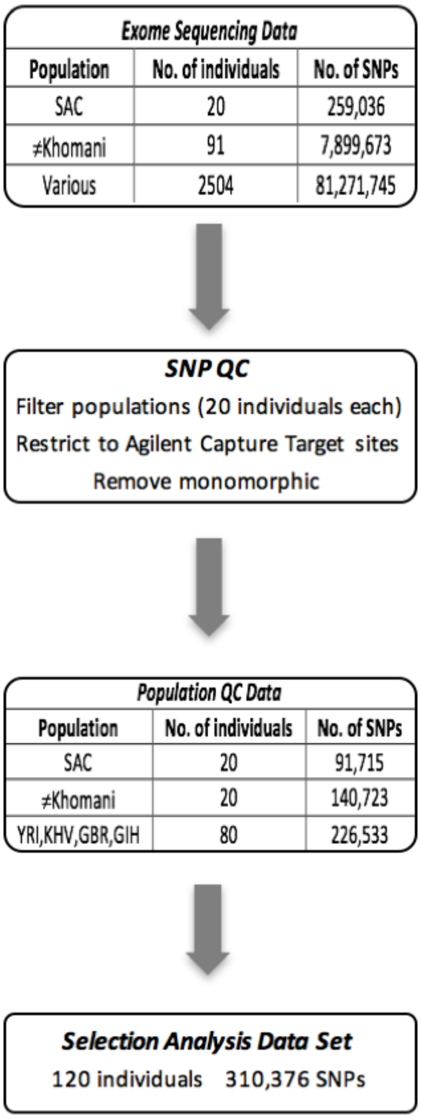
Most of the variants under selection were found to be functionally tolerated, but some were expected to be damaging. One such example is a nonsynonymous variant (rs2073924) in the 5’-untranslated region of *GBGT1*. *GBGT1* encodes a glycosyltransferase which aids in the synthesis of the Forssman glycolipid (FG). FG forms binding sites for the attachment of pathogens to cells (415). Therefore, a change in expression of FG may influence host tropism. Another example is a nonsynonymous variant (rs1882118) in *NR4A1*. It is thought that NR4A1 inhibits the NF-kappaB transactivation of IL2 (416). IL2 is a major regulator of white blood cells which are a large component of the immune system. Lastly, a nonsynonymous variant (rs28477168) in *TMIGD2* was found. TMIGD2 enhances T-cell proliferation as well as cytokine production, both of which are major role-players in viral immunity (417).

Closer inspection of the genes under selection identified an over-representation of genes in the focal adhesion KEGG pathway (413), which directly interact with focal adhesion kinase (FAK) (**Figure 4**). The FAK protein has been implicated in both viral and bacterial immune response (418–421). It appears to be involved in cytoplasmic entry and replication of viruses (particularly influenza A and the herpes simplex virus) as well as regulating polymerase activity (418,419). With regards to bacteria, the activation of FAK by *Salmonella* suppresses autophagy and promotes survival in macrophages (420). The role of FAK during *Mycobacterium tuberculosis* (*M.tb)* infection is poorly understood, but *M.tb* and *Salmonella* are similar in that they are both facultative intracellular pathogens. A selection scan based on deviations in local ancestry performed by Chimusa et al. (2015) revealed that, amongst other pathways, genes associated with the focal adhesion pathway were under positive selection in the ≠Khomani (93). Although the selection scan presented here was augmented to account for admixture in the SAC and was not based on local ancestry deviations, it is plausible that some overlap across studies would be possible, the extent of which is a direct result of differing selective pressures across populations. Additional research into selection patterns and the role of focal adhesion in both viral and bacterial immunity will provide valuable insights into TB susceptibility in both the SAC and ≠Khomani populations.

Strengths and limitations of using exome sequence data from a highly admixed population

Exome datasets provide genome-wide polymorphism data that lack ascertainment bias, and allow for the analysis of functionally relevant regions. In this study, the F_{st} based PBS was applied in such a way as to mitigate the confounding effects of admixture on the detection of signals of selection. The PBS has successfully been used to detect very recent signals of selection in genes associated with the adaptation to high altitudes (395) as well as selection in immune response genes in the ≠Khomani (391). These studies support our ability to identify recent positive selection in a relatively small sample of exome sequenced individuals. Furthermore, the methodology used to account for admixture in this study by preventing an increase in a particular ancestry producing a false positive, is novel and extends the application of the PBS to highly admixed populations. As with most methodologies however, this does not fully eliminate admixture as a potential confounder since there is a possibility that the reference populations chosen were not truly representative of the ancestral composition of the SAC and therefore an ancestral component is missing. It can be argued that the PBS was not the best statistic to use however, given the fact that exome sequencing data was used, we cannot use a statistic based on haplotype formation e.g. iHS (422). Furthermore, any statistic in which simulations are used to determine significance (based on a neutral model) may be inaccurate due to the SAC's complex demographic history and lack of a population specific recombination map.

Overall, this study provides evidence for selection in immune response genes in the SAC population with particular reference to signals associated with an abnormal immune system and the focal adhesion pathway. Our results therefore provide a stepping-stone to better understand the host immune response to pathogens, particularly in southern Africa.



Exome Sequencing Data

Population	No. of individuals	No. of SNPs
SAC	20	259,036
≠Khomani	91	7,899,673
Various	2504	81,271,745

SNP QC

Filter populations (20 individuals each)

Restrict to Agilent Capture Target sites

Remove monomorphic

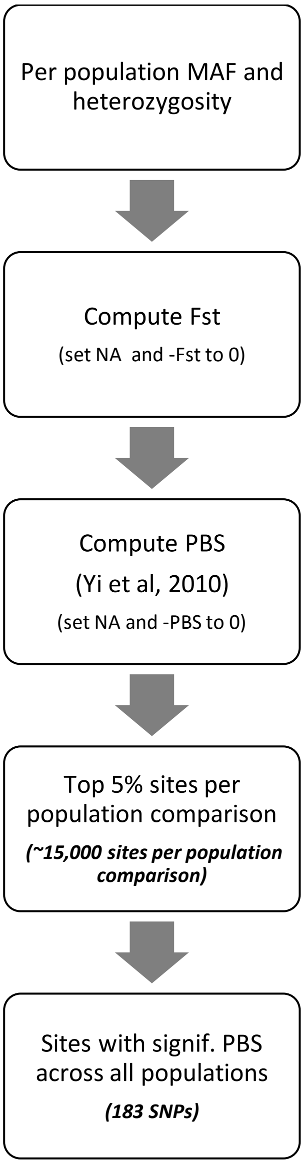
Population QC Data

Population	No. of individuals	No. of SNPs
SAC	20	91,715
≠Khomani	20	140,723
YRI,KHV,GBR,GIH	80	226,533

Selection Analysis Data Set

120 individuals 310,376 SNPs

Figure S1: Exome filtering pipeline.



Per population MAF and heterozygosity
Compute Fst
(set NA and -Fst to 0)
Compute PBS
(Yi et al, 2010)
(set NA and -PBS to 0)
Top 5% sites per population comparison
(~15,000 sites per population comparison)
Sites with signif. PBS across all populations
(183 SNPs)

Figure S2: PBS calculation pipeline.

Table S1: Annotated list of the 183 variants under selection.

TFBS, Transcription Factors Binding Site; GAD, Genetic Association Database

chr	pos	ref allele	alt allele	rsID	closest gene	SIFT	TFBS	GAD
1	897325	G	C	rs4970441	*KLHL17*			
1	1115415	C	T	rs57556493	*TTLL10*			
1	1178925	G	A	rs12093154	*FAM132 A*	TOLERATED		
1	2436404	G	C	rs113392853	*PLCH2*	TOLERATED		
1	3786245	G	A	rs12738235	*DFFB*	TOLERATED		CANCER
1	15541607	T	C	rs3820065	*TMEM51*			CHEMDEP ENDENCY
1	19634747	C	T	rs111682618	*AKR7A2*	TOLERATED		INFECTION
1	47882497	C	T	rs34082359	*FOXE3*			VISION
1	153662340	G	A	rs35240348	*NPR1*	TOLERATED		CARDIOVA SCULAR
1	158549420	C	T	rs863361	*OR10X1*			
1	158549511	A	G	rs863363	*OR10X1*	TOLERATED		
1	170965681	T	C	rs9427213	*MROH9*			
1	179057105	C	A	rs12061876	*TOR3A*			CHEMDEP ENDENCY
1	179057117	G	A	rs12092348	*TOR3A*			CHEMDEP ENDENCY
1	179887125	G	A	rs627897	*TOR1AI P1*			
1	180145088	A	G	rs2298206	*QSOX1*			
1	202287206	T	C	rs788795	*LGR6*	TOLERATED		HEMATOL OGICAL
1	202287537	G	A	rs788794	*LGR6*		Pax-5	HEMATOL OGICAL
1	202287813	T	C	rs788793	*LGR6*			HEMATOL OGICAL
1	203140671	G	A	rs3737875	*MYBPH*		Pax-5	CARDIOVA SCULAR
1	230898494	C	T	rs2282319	*CAPN9*		GR-alpha	CANCER
1	230903350	T	C	rs3828126	*CAPN9*			CANCER
1	236557742	G	A	rs79233817	*EDARA DD*			
1	240371554	A	G	rs12732924	*FMN2*	TOLERATED		CARDIOVA

								SCULAR
1	248845356	G	T	rs41311583	*OR14I1*	DAMAGING		
2	37406644	A	G	rs10182091	*SULT6B1*		Bach1	
2	37406680	C	G	rs10205833	*SULT6B1*	TOLERATED	POU2F1	
2	68546374	T	C	rs4671898	*CNRIP1*		HNF-1A	CHEMDEPENDENCY
2	85628745	C	T	rs2229669	*CAPG*			CARDIOVASCULAR
2	141707868	G	T	rs6748626	*LRP1B*			NEUROLOGICAL
2	170403030	T	C	rs2253680	*FASTKD1*	TOLERATED		
2	179666982	C	A	rs35683768	*TTN*	TOLERATED	CUTL1	CARDIOVASCULAR
2	182374534	A	G	rs1143674	*ITGA4*			IMMUNE
2	185801755	A	G	rs4667002	*ZNF804A*			PSYCH
2	209036778	A	G	rs6435421	*C2orf80*	TOLERATED		
2	242350466	A	G	rs2240479	*FARP2*		MIF-1	CANCER
3	361508	C	T	rs2272522	*CHL1*	TOLERATED		OTHER
3	8809703	G	A	rs2228485	*OXTR*			PSYCH
3	44761729	A	G	rs17076938	*ZNF502*			
3	100949842	G	A	rs348867	*IMPG2*			
3	108672514	G	A	rs6783631	*GUCA1C*			
3	122354792	G	A	rs12489170	*PARP15*	DAMAGING		PHARMACOGENOMIC
3	133653566	G	A	rs72978391	*SLCO2A1*		AP-4	METABOLIC
3	155481440	T	G	rs113890115	*C3orf33*	TOLERATED		
3	188590446	A	G	rs1136644	*LPP*		p53	IMMUNE
4	8394094	G	A	rs28627156	*ACOX3*		C/EBP alpha	INFECTION
4	79421011	G	T	rs11933630	*FRAS1*		Pax-2	DEVELOPMENTAL
4	79434685	T	G	rs35933858	*FRAS1*	TOLERATED		DEVELOPMENTAL
4	79443837	A	T	rs931605	*FRAS1*	TOLERATED		DEVELOPMENTAL
4	88732531	T	C	rs4693878	*IBSP*			METABOLIC
4	88732746	A	G	rs13144371	*IBSP*	TOLERATED		METABOLIC
4	104067195	T	C	rs35505100	*CENPE*	TOLERATED		CANCER
4	110790911	A	T	rs764205	*LRIT3*	TOLERATED		CHEMDEPENDENCY

5	678079	G	A	rs7737452	*TPPP*			IMMUNE
5	53815066	C	A	rs2548613	*SNX18*		CUTL1	
5	78752802	T	C	rs35944172	*HOMER1*			CHEMDEPENDENCY
5	79351735	C	T	rs423906	*THBS4*			CARDIOVASCULAR
5	140563173	G	T	rs28664170	*PCDHB16*	TOLERATED		
5	149216142	G	A	rs45617032	*PPARGC1B*			METABOLIC
5	172750392	G	A	rs6861827	*STC2*			NEUROLOGICAL
5	176024880	C	T	rs4868663	*GPRIN1*			
5	180166866	G	A	rs61737927	*OR2Y1*	DAMAGING		
6	13711279	A	G	rs6905991	*RANBP9*		Pax-5	
6	17541324	A	G	rs34206659	*CAP2*	TOLERATED		DEVELOPMENTAL
6	31556928	C	A	rs3179003	*NCR3*	TOLERATED		INFECTION
6	41903783	G	A	rs3218102	*CCND3*	TOLERATED		CANCER
6	46826979	A	G	rs9381487	*GPR116*			CHEMDEPENDENCY
6	47847401	A	G	rs2068006	*PTCHD4*			
6	69666684	A	G	rs1932618	*BAI3*	TOLERATED	POU2F1	IMMUNE
6	168459845	G	C	rs1548349	*FRMD1*	TOLERATED		
7	21932044	C	T	rs12537531	*DNAH11*	TOLERATED		CARDIOVASCULAR
7	92734983	A	G	rs6969691	*SAMD9*	TOLERATED		
7	130418689	A	G	rs111400400	*KLF14*	TOLERATED	MIF-1	METABOLIC
7	137150665	T	C	rs35245703	*DGKI*			PSYCH
8	10388826	G	A	rs4841367	*PRSS55*			
8	10396056	G	C	rs61743179	*PRSS55*	DAMAGING		
8	23001988	A	G	rs1133782	*TNFRSF10D*	TOLERATED		NORMALVARIATION
8	120594802	A	G	rs4871364	*ENPP2*			CHEMDEPENDENCY
8	139749799	G	T	rs10111520	*COL22A1*	TOLERATED		HEMATOLOGICAL
9	712137	G	C	rs912175	*KANK1*			IMMUNE
9	15784631	A	G	rs1539172	*CCDC171*	TOLERATED		
9	18639300	G	A	rs776755	*ADAMTSL1*	TOLERATED		IMMUNE
9	100105782	C	G	rs2061634	*CCDC180*	TOLERATED		
9	100122291	T	C	rs3747495	*CCDC180*	TOLERATED		
9	100995758	G	T	rs879368	*TBC1D2*	TOLERATED	HSF1	DEVELOP

								(long)	MENTAL
9	116060124	T	C	rs3750534	*RNF183*	TOLERATED	CREB	
9	116060221	C	T	rs3750533	*RNF183*	TOLERATED		
9	130098409	G	A	rs60971543	*GARNL3*			CHEMDEPENDENCY
9	131862805	G	A	rs1127926	*CRAT*		AP-4	INFECTION
9	136037742	G	A	rs2073924	*GBGT1*	DAMAGING		AGING
9	137707834	G	A	rs3827848	*COL5A1*			OTHER
9	139617662	T	C	rs11849	*FAM69B*			
10	5136651	C	G	rs12529	*AKR1C3*	TOLERATED		CANCER
10	18266989	G	A	rs2478568	*SLC39A12*	TOLERATED		CANCER
10	30316872	T	C	rs3739996	*KIAA1462*			NEUROLOGICAL
10	72324143	A	G	rs7093516	*PALD1*			
10	95072906	T	C	rs787666	*MYOF*			NEUROLOGICAL
10	105810400	T	C	rs805722	*COL17A1*	TOLERATED	GR-alpha	IMMUNE
10	127753478	G	A	rs1278279	*ADAM12*			METABOLIC
10	135089035	A	G	rs2275725	*ADAM8*			IMMUNE
11	5719667	T	C	rs2291842	*TRIM22*			IMMUNE
11	7324475	T	C	rs4412741	*SYT9*		Egr-2	CHEMDEPENDENCY
11	7509566	A	T	rs12805648	*OLFML1*	DAMAGING		
11	7949350	A	G	rs4758258	*OR10A6*	TOLERATED		
11	33076191	T	C	rs2273544	*TCP11L1*		AP-1	INFECTION
11	36484115	G	T	rs34680620	*PRR5L*			INFECTION
11	56113593	C	A	rs10896272	*OR8K1*	DAMAGING		
11	60673868	T	C	rs17155470	*PRPF19*			
11	60701987	G	A	rs7715	*TMEM132A*			
11	65064706	G	A	rs7123885	*POLA2*	TOLERATED		CANCER
11	65088687	C	T	rs35225270	*CDC42EP2*			
11	118895635	G	A	rs35010541	*SLC37A4*			METABOLIC
11	119025279	G	A	rs12277959	*ABCG4*			PHARMACOGENOMIC
11	119170362	C	T	rs1893177	*CBL*		Bach1	METABOLIC
11	122667665	G	A	rs111571941	*UBASH3B*			IMMUNE
12	52448188	C	G	rs1882118	*NR4A1*	DAMAGING	ARP-1	NEUROLOGICAL

12	53166603	T	C	rs61730600	*KRT76*			
12	104089586	C	T	rs17034395	*STAB2*			CHEMDEPENDENCY
12	133220526	T	C	rs5744934	*POLE*	DAMAGING		CANCER
14	24845543	T	C	rs2228232	*NFATC4*		c-Rel	CARDIOVASCULAR
14	45696037	T	G	rs34101857	*MIS18BP1*	TOLERATED		
14	50319499	G	T	rs3100887	*NEMF*			
14	69352230	G	A	rs15993	*ACTN1*			DEVELOPMENTAL
14	101004293	T	C	rs111729292	*BEGAIN*			METABOLIC
14	105054934	A	G	rs12880814	*C14orf180*			
15	63131117	T	A	rs937418	*TLN2*		YY1	CHEMDEPENDENCY
15	90320000	G	A	rs28462216	*MESP2*	TOLERATED		
16	396264	A	G	rs1805105	*AXIN1*			CANCER
16	4833970	C	T	rs759991	*SEPT12*		Pax-6	
16	20370810	C	T	rs9652588	*PDILT*	TOLERATED		
16	20423000	G	A	rs9928053	*ACSM5*	TOLERATED		INFECTION
16	20430678	G	A	rs7192210	*ACSM5*	TOLERATED		INFECTION
16	23768908	C	T	rs16972893	*CHP2*			IMMUNE
17	4442818	G	A	rs73335853	*MYBBP1A*			METABOLIC
17	6515392	T	G	rs79173884	*KIAA0753*			
17	6531552	A	C	rs16955985	*KIAA0753*	TOLERATED		
17	10633181	A	G	rs3826448	*TMEM220*			
17	16001756	G	A	rs79413281	*NCOR1*			CANCER
17	32647831	A	C	rs1133763	*CCL8*			METABOLIC
17	32904586	C	T	rs887230	*C17orf102*			
17	36474601	A	G	rs59939216	*MRPL45*			INFECTION
17	38146154	T	C	rs9916279	*PSMD3*		Cart-1	HEMATOLOGICAL
17	38948660	C	A	rs80110057	*KRT28*			
17	44077044	T	C	rs71375329	*STH*			NEUROLOGICAL
17	73016489	C	T	rs34496172	*ICT1*	TOLERATED		INFECTION
18	34324091	G	A	rs2303510	*FHOD3*	TOLERATED		PHARMACOGENOMIC
18	43795986	C	T	rs61736697	*C18orf25*	TOLERATED		
18	56149099	T	C	rs7240666	*ALPK2*	TOLERATED		RENAL

19	1110829	G	A	rs2302109	*SBNO2*		GR-alpha	IMMUNE
19	1619333	G	A	rs1140828	*TCF3*		E2F	
19	2732725	C	T	rs111940360	*SLC39A3*			
19	4292841	C	G	rs28477168	*TMIGD2*	DAMAGING		
19	4929413	T	C	rs2251520	*UHRF1*			CARDIOVASCULAR
19	4929473	A	G	rs2123731	*UHRF1*			CARDIOVASCULAR
19	7992052	A	C	rs12976850	*TIMM44*			INFECTION
19	7992126	G	A	rs11542188	*TIMM44*			INFECTION
19	35434238	A	G	rs1811	*ZNF30*	TOLERATED		
19	35449760	G	C	rs2651080	*ZNF792*			CHEMDEPENDENCY
19	35506729	G	A	rs2290647	*GRAMD1A*			
19	42408426	G	A	rs11083640	*ARHGEF1*			CARDIOVASCULAR
19	47282162	G	A	rs2070246	*SLC1A5*			OTHER
19	48525507	G	A	rs2303690	*ELSPBP1*	TOLERATED		METABOLIC
19	50017724	C	T	rs2878342	*FCGRT*		Pax-5	IMMUNE
19	50545070	A	G	rs10419911	*ZNF473*	TOLERATED		CHEMDEPENDENCY
19	57649962	C	T	rs10407445	*ZIM3*	TOLERATED		
20	24944447	G	A	rs1045772	*APMAP*			
20	24973247	C	A	rs11550623	*APMAP*			
20	24994275	G	A	rs6115001	*ACSS1*			CARDIOVASCULAR
20	48257149	C	T	rs421801	*B4GALT5*			
20	61878950	A	C	rs872808	*NKAIN4*	TOLERATED		
21	40823902	C	G	rs11575939	*SH3BGR*	TOLERATED		METABOLIC
21	43319180	T	C	rs9981024	*C2CD2*	TOLERATED		
21	43319214	C	T	rs9985096	*C2CD2*			
21	43510514	C	G	rs2839464	*UMODL1*	TOLERATED		VISION
22	17590180	G	A	rs41323645	*IL17RA*	TOLERATED		INFECTION
22	18912677	C	T	rs11913840	*PRODH*			PSYCH
22	22899234	A	G	rs1129172	*PRAME*	TOLERATED		
22	25331451	C	T	rs34889393	*TMEM211*	DAMAGING	RREB-1	CHEMDEPENDENCY
22	29533499	G	A	rs35612970	*KREMEN1*	TOLERATED		METABOLIC
22	29533572	C	G	rs34920087	*KREMEN1*	TOLERATED	AP-4	METABOLIC

22	37499386	C	T	rs11704654	*TMPRSS6*			CANCER
22	39482371	C	G	rs17496046	*APOBEC3G*	TOLERATED		INFECTION
22	51045190	C	T	rs1140555	*MAPK8IP2*			

Table S2: PBS values of the 183 candidate variants under positive selection.

Top 10 variants for each population comparison are shaded.

rsID	Gene	SAC-SAN-GIH	SAC-GIH-GBR	SAC-SAN-YRI	SAC-YRI-KHV	SAN-SAN-GBR	SAC-GIH-YRI	SAC-GIH-KHV	SAC-SAN-KHV	SAC-GBR-KHV	SAC-GBR-YRI
rs2272522	*CHL1*	0.085	0.214	0.012	0.021	0.014	0.019	0.131	0.088	0.238	0.032
rs1805105	*AXIN1*	0.143	0.315	0.026	0.105	0.026	0.052	0.545	0.292	0.658	0.059
rs7737452	*TPPP*	0.213	0.236	0.116	0.125	0.081	0.125	0.236	0.200	0.236	0.125
rs912175	*KANK1*	0.088	0.140	0.046	0.041	0.037	0.041	0.063	0.069	0.140	0.085
rs4970441	*KLHL17*	0.110	0.207	0.023	0.039	0.028	0.031	0.186	0.140	0.269	0.042
rs2302109	*SBNO2*	0.100	0.083	0.046	0.056	0.067	0.025	0.120	0.172	0.163	0.042
rs57556493	*TTLL10*	0.210	0.236	0.112	0.132	0.080	0.132	0.236	0.198	0.236	0.132
rs12093154	*FAM132A*	0.068	0.044	0.033	0.060	0.020	0.042	0.076	0.067	0.062	0.027
rs1140828	*TCF3*	0.199	0.058	0.236	0.219	0.100	0.219	0.166	0.183	0.058	0.072
rs113392853	*PLCH2*	0.062	0.081	0.017	0.006	0.017	0.024	0.031	0.023	0.048	0.029
rs111940360	*SLC39A3*	0.092	0.111	0.019	0.022	0.023	0.022	0.111	0.079	0.111	0.022
rs12738235	*DFFB*	0.090	0.045	0.020	0.023	0.037	0.019	0.097	0.098	0.064	0.009
rs28477168	*TMIGD2*	0.038	0.057	0.039	0.048	0.026	0.013	0.035	0.061	0.114	0.080
rs73335853	*MYBBP1A*	0.203	0.270	0.043	0.049	0.068	0.049	0.270	0.203	0.270	0.049
rs759991	*SEPT12*	0.070	0.096	0.161	0.150	0.051	0.096	0.052	0.092	0.149	0.217
rs2251520	*UHRF1*	0.114	0.079	0.031	0.114	0.028	0.064	0.265	0.197	0.138	0.017
rs2123731	*UHRF1*	0.125	0.079	0.040	0.114	0.037	0.064	0.265	0.216	0.138	0.017
rs12529	*AKR1C3*	0.150	0.053	0.049	0.121	0.037	0.096	0.273	0.188	0.067	0.012
rs2291842	*TRIM22*	0.105	0.122	0.066	0.059	0.121	0.013	0.095	0.222	0.282	0.077
rs79173884	*KIAA0753*	0.068	0.051	0.014	0.014	0.023	0.013	0.081	0.079	0.081	0.013
rs16955985	*KIAA0753*	0.095	0.135	0.018	0.046	0.018	0.043	0.169	0.107	0.169	0.043

rs4412741	*SYT9*	0.054	0.054	0.028	0.028	0.027	0.028	0.054	0.054	0.054	0.028
rs12805648	*OLFML1*	0.104	0.127	0.014	0.013	0.026	0.019	0.085	0.048	0.085	0.019
rs4758258	*OR10A6*	0.177	0.117	0.256	0.158	0.191	0.117	0.117	0.256	0.158	0.158
rs12976850	*TIMM44*	0.102	0.046	0.055	0.141	0.016	0.141	0.204	0.102	0.046	0.037
rs11542188	*TIMM44*	0.107	0.046	0.058	0.141	0.018	0.141	0.204	0.107	0.046	0.037
rs28627156	*ACOX3*	0.140	0.186	0.028	0.059	0.028	0.054	0.269	0.153	0.207	0.043
rs2228485	*OXTR*	0.162	0.168	0.040	0.019	0.026	0.115	0.093	0.030	0.040	0.051
rs4841367	*PRSS55*	0.084	0.074	0.030	0.074	0.013	0.074	0.141	0.084	0.074	0.034
rs61743179	*PRSS55*	0.112	0.101	0.026	0.045	0.026	0.045	0.172	0.112	0.101	0.029
rs3826448	*TMEM220*	0.324	0.300	0.094	0.043	0.093	0.145	0.197	0.122	0.115	0.087
rs6905991	*RANBP9*	0.128	0.052	0.082	0.186	0.040	0.114	0.230	0.192	0.081	0.040
rs3820065	*TMEM51*	0.102	0.060	0.020	0.027	0.037	0.021	0.139	0.133	0.100	0.015
rs1539172	*CCDC171*	0.036	0.051	0.084	0.133	0.038	0.037	0.051	0.126	0.176	0.134
rs79413281	*NCOR1*	0.092	0.111	0.029	0.048	0.023	0.048	0.111	0.079	0.111	0.048
rs34206659	*CAP2*	0.120	0.045	0.076	0.109	0.037	0.109	0.141	0.107	0.045	0.039
rs41323645	*IL17RA*	0.060	0.069	0.052	0.099	0.044	0.015	0.099	0.193	0.217	0.069
rs2478568	*SLC39A12*	0.126	0.097	0.068	0.030	0.043	0.087	0.060	0.048	0.040	0.060
rs776755	*ADAMTSL1*	0.136	0.236	0.029	0.041	0.031	0.041	0.236	0.136	0.236	0.041
rs11913840	*PRODH*	0.107	0.169	0.041	0.097	0.018	0.097	0.204	0.107	0.169	0.089
rs111682618	*AKR7A2*	0.027	0.054	0.050	0.000	0.099	0.000	0.054	0.027	0.054	0.000
rs9652588	*PDILT*	0.140	0.064	0.094	0.224	0.050	0.115	0.278	0.253	0.121	0.049
rs9928053	*ACSM5*	0.110	0.155	0.073	0.076	0.044	0.076	0.094	0.091	0.155	0.123
rs7192210	*ACSM5*	0.090	0.155	0.013	0.016	0.025	0.016	0.094	0.072	0.155	0.021
rs12537531	*DNAH11*	0.249	0.189	0.031	0.029	0.103	0.034	0.222	0.198	0.157	0.024
rs1129172	*PRAME*	0.090	0.042	0.059	0.162	0.027	0.078	0.185	0.165	0.079	0.032
rs1133782	*TNFRSF10D*	0.160	0.100	0.013	0.031	0.040	0.022	0.348	0.229	0.143	0.006

rs16972893	*CHP2*	0.084	0.141	0.011	0.015	0.013	0.015	0.141	0.084	0.141	0.015
rs2228232	*NFATC4*	0.042	0.031	0.023	0.038	0.023	0.006	0.048	0.079	0.081	0.025
rs1045772	*APMAP*	0.080	0.037	0.026	0.068	0.018	0.047	0.141	0.107	0.046	0.008
rs11550623	*APMAP*	0.098	0.047	0.042	0.068	0.034	0.047	0.141	0.133	0.068	0.016
rs6115001	*ACSS1*	0.107	0.056	0.031	0.048	0.031	0.040	0.171	0.136	0.068	0.005
rs34889393	*TMEM211*	0.065	0.082	0.023	0.054	0.010	0.054	0.082	0.052	0.082	0.054
rs35612970	*KREMEN1*	0.061	0.079	0.041	0.040	0.023	0.035	0.040	0.046	0.084	0.079
rs34920087	*KREMEN1*	0.061	0.079	0.041	0.040	0.023	0.035	0.040	0.046	0.084	0.079
rs3739996	*KIAA1462*	0.031	0.087	0.016	0.071	0.013	0.002	0.070	0.118	0.329	0.087
rs3179003	*NCR3*	0.084	0.109	0.013	0.039	0.013	0.053	0.109	0.052	0.078	0.039
rs1133763	*CCL8*	0.215	0.255	0.020	0.019	0.037	0.038	0.188	0.073	0.101	0.024
rs887230	*C17orf102*	0.414	0.202	0.163	0.163	0.190	0.167	0.396	0.379	0.196	0.097
rs2273544	*TCP11L1*	0.156	0.218	0.137	0.282	0.066	0.144	0.281	0.285	0.401	0.218
rs2303510	*FHOD3*	0.142	0.083	0.033	0.042	0.022	0.103	0.143	0.053	0.035	0.014
rs1811	*ZNF30*	0.221	0.071	0.177	0.145	0.118	0.144	0.172	0.205	0.071	0.057
rs2651080	*ZNF792*	0.108	0.102	0.076	0.063	0.040	0.088	0.078	0.066	0.078	0.088
rs2290647	*GRAMD1A*	0.194	0.162	0.033	0.059	0.028	0.109	0.247	0.088	0.082	0.027
rs59939216	*MRPL45*	0.072	0.058	0.013	0.010	0.025	0.013	0.044	0.039	0.044	0.013
rs34680620	*PRR5L*	0.141	0.204	0.107	0.169	0.039	0.169	0.204	0.141	0.204	0.169
rs10182091	*SULT6B1*	0.241	0.094	0.069	0.152	0.077	0.095	0.502	0.377	0.152	0.010
rs10205833	*SULT6B1*	0.241	0.094	0.069	0.152	0.077	0.095	0.502	0.377	0.152	0.010
rs11704654	*TMPRSS6*	0.105	0.070	0.088	0.028	0.053	0.070	0.028	0.035	0.028	0.070
rs9916279	*PSMD3*	0.189	0.099	0.034	0.049	0.072	0.037	0.287	0.261	0.133	0.014
rs80110057	*KRT28*	0.092	0.111	0.023	0.029	0.023	0.029	0.111	0.079	0.111	0.029
rs17496046	*APOBEC3G*	0.107	0.097	0.018	0.019	0.018	0.031	0.097	0.041	0.045	0.019
rs11575939	*SH3BGR*	0.049	0.042	0.025	0.045	0.028	0.009	0.093	0.153	0.146	0.025

rs3218102	*CCND3*	0.120	0.109	0.028	0.027	0.037	0.029	0.109	0.076	0.078	0.027
rs11083640	*ARHGEF1*	0.092	0.111	0.029	0.048	0.023	0.048	0.111	0.079	0.111	0.048
rs9981024	*C2CD2*	0.171	0.236	0.056	0.072	0.053	0.072	0.236	0.171	0.236	0.072
rs9985096	*C2CD2*	0.171	0.236	0.056	0.072	0.053	0.072	0.236	0.171	0.236	0.072
rs2839464	*UMODL1*	0.145	0.142	0.016	0.040	0.018	0.040	0.322	0.145	0.142	0.016
rs61736697	*C18orf25*	0.150	0.172	0.080	0.101	0.051	0.101	0.172	0.137	0.172	0.101
rs71375329	*STH*	0.082	0.082	0.028	0.028	0.041	0.028	0.082	0.082	0.082	0.028
rs17076938	*ZNF502*	0.150	0.139	0.015	0.015	0.051	0.015	0.139	0.104	0.106	0.015
rs34101857	*MIS18BP1*	0.139	0.126	0.061	0.040	0.057	0.061	0.097	0.093	0.101	0.062
rs9381487	*GPR116*	0.176	0.147	0.076	0.005	0.098	0.060	0.046	0.059	0.053	0.068
rs2070246	*SLC1A5*	0.167	0.129	0.025	0.010	0.167	0.010	0.068	0.167	0.129	0.019
rs2068006	*PTCHD4*	0.182	0.247	0.039	0.072	0.050	0.058	0.315	0.225	0.317	0.058
rs34082359	*FOXE3*	0.087	0.148	0.046	0.055	0.018	0.079	0.079	0.046	0.119	0.119
rs421801	*B4GALT5*	0.051	0.094	0.046	0.082	0.010	0.058	0.069	0.057	0.123	0.113
rs2303690	*ELSPBP1*	0.112	0.087	0.081	0.249	0.037	0.100	0.279	0.253	0.220	0.076
rs2878342	*FCGRT*	0.156	0.386	0.018	0.072	0.018	0.066	0.453	0.173	0.453	0.066
rs3100887	*NEMF*	0.054	0.054	0.014	0.014	0.027	0.014	0.054	0.054	0.054	0.014
rs10419911	*ZNF473*	0.133	0.379	0.014	0.058	0.014	0.058	0.379	0.133	0.379	0.058
rs1140555	*MAPK8IP2*	0.093	0.065	0.021	0.036	0.063	0.012	0.162	0.259	0.211	0.016
rs1882118	*NR4A1*	0.150	0.172	0.029	0.030	0.051	0.030	0.172	0.137	0.172	0.030
rs61730600	*KRT76*	0.068	0.081	0.049	0.081	0.023	0.051	0.081	0.079	0.111	0.081
rs2548613	*SNX18*	0.187	0.105	0.070	0.175	0.051	0.125	0.379	0.261	0.147	0.032
rs10896272	*OR8K1*	0.270	0.189	0.038	0.076	0.053	0.080	0.521	0.259	0.181	0.025
rs7240666	*ALPK2*	0.136	0.048	0.086	0.160	0.031	0.160	0.236	0.136	0.048	0.039
rs10407445	*ZIM3*	0.121	0.151	0.030	0.020	0.032	0.041	0.096	0.052	0.096	0.041
rs17155470	*PRPF19*	0.084	0.074	0.051	0.115	0.033	0.050	0.133	0.143	0.152	0.060

rs7715	*TMEM132A*	0.070	0.079	0.083	0.153	0.050	0.043	0.113	0.196	0.216	0.103
rs872808	*NKAIN4*	0.119	0.211	0.046	0.021	0.024	0.105	0.038	0.023	0.045	0.124
rs937418	*TLN2*	0.192	0.705	0.048	0.283	0.010	0.260	0.804	0.211	0.773	0.249
rs7123885	*POLA2*	0.092	0.111	0.029	0.048	0.023	0.048	0.111	0.079	0.111	0.048
rs35225270	*CDC42EP2*	0.092	0.111	0.029	0.048	0.023	0.048	0.111	0.079	0.111	0.048
rs4671898	*CNRIP1*	0.121	0.036	0.143	0.162	0.163	0.035	0.138	0.383	0.162	0.045
rs15993	*ACTN1*	0.055	0.130	0.035	0.098	0.013	0.043	0.098	0.090	0.224	0.130
rs1932618	*BAI3*	0.070	0.107	0.040	0.054	0.013	0.068	0.068	0.040	0.093	0.093
rs7093516	*PALD1*	0.204	0.204	0.016	0.016	0.102	0.016	0.204	0.204	0.204	0.016
rs34496172	*ICT1*	0.145	0.149	0.050	0.077	0.050	0.059	0.226	0.194	0.219	0.059
rs35944172	*HOMER1*	0.120	0.109	0.039	0.045	0.037	0.045	0.141	0.107	0.109	0.039
rs423906	*THBS4*	0.119	0.070	0.050	0.027	0.095	0.027	0.057	0.119	0.070	0.033
rs11933630	*FRAS1*	0.120	0.036	0.168	0.204	0.066	0.130	0.130	0.168	0.046	0.046
rs35933858	*FRAS1*	0.120	0.036	0.168	0.204	0.066	0.130	0.130	0.168	0.046	0.046
rs931605	*FRAS1*	0.120	0.036	0.168	0.204	0.066	0.130	0.130	0.168	0.046	0.046
rs2229669	*CAPG*	0.120	0.074	0.054	0.074	0.037	0.074	0.141	0.107	0.074	0.034
rs4693878	*IBSP*	0.140	0.212	0.023	0.052	0.025	0.043	0.320	0.175	0.275	0.037
rs13144371	*IBSP*	0.119	0.212	0.013	0.052	0.013	0.043	0.320	0.149	0.275	0.037
rs28462216	*MESP2*	0.184	0.201	0.046	0.048	0.081	0.045	0.201	0.200	0.236	0.048
rs6969691	*SAMD9*	0.123	0.144	0.020	0.025	0.020	0.052	0.116	0.045	0.066	0.034
rs787666	*MYOF*	0.072	0.031	0.025	0.027	0.025	0.029	0.056	0.051	0.028	0.002
rs2061634	*CCDC180*	0.127	0.061	0.026	0.013	0.104	0.016	0.047	0.105	0.047	0.016
rs3747495	*CCDC180*	0.103	0.061	0.035	0.036	0.037	0.036	0.091	0.080	0.057	0.014
rs348867	*IMPG2*	0.269	0.044	0.203	0.079	0.320	0.082	0.106	0.268	0.044	0.028
rs879368	*TBC1D2*	0.202	0.065	0.097	0.015	0.188	0.048	0.059	0.123	0.029	0.022
rs111729292	*BEGAIN*	0.092	0.111	0.049	0.081	0.023	0.081	0.111	0.079	0.111	0.081

rs35505100	*CENPE*	0.084	0.141	0.015	0.045	0.013	0.045	0.141	0.084	0.141	0.045
rs17034395	*STAB2*	0.084	0.125	0.026	0.037	0.031	0.025	0.125	0.136	0.236	0.037
rs12880814	*C14orf180*	0.612	1.071	0.023	0.005	0.201	0.034	0.189	0.128	0.241	0.045
rs805722	*COL17A1*	0.125	0.394	0.015	0.073	0.015	0.059	0.309	0.154	0.488	0.092
rs6783631	*GUCA1C*	0.092	0.111	0.023	0.029	0.023	0.029	0.111	0.079	0.111	0.029
rs764205	*LRIT3*	0.103	0.119	0.027	0.016	0.038	0.028	0.061	0.054	0.083	0.035
rs3750534	*RNF183*	0.149	0.073	0.116	0.110	0.054	0.139	0.125	0.102	0.054	0.066
rs3750533	*RNF183*	0.149	0.059	0.116	0.123	0.054	0.139	0.138	0.115	0.053	0.053
rs35010541	*SLC37A4*	0.095	0.109	0.107	0.141	0.037	0.109	0.109	0.107	0.141	0.141
rs12277959	*ABCG4*	0.150	0.172	0.063	0.070	0.051	0.070	0.172	0.137	0.172	0.070
rs1893177	*CBL*	0.150	0.172	0.063	0.070	0.051	0.070	0.172	0.137	0.172	0.070
rs4871364	*ENPP2*	0.115	0.234	0.032	0.108	0.027	0.065	0.268	0.189	0.383	0.094
rs12489170	*PARP15*	0.133	0.101	0.017	0.032	0.039	0.021	0.272	0.212	0.157	0.013
rs111571941	*UBASH3B*	0.065	0.082	0.023	0.054	0.010	0.054	0.082	0.052	0.082	0.054
rs1278279	*ADAM12*	0.074	0.067	0.025	0.033	0.040	0.017	0.100	0.131	0.140	0.025
rs60971543	*GARNL3*	0.065	0.082	0.023	0.054	0.010	0.054	0.082	0.052	0.082	0.054
rs111400400	*KLF14*	0.176	0.161	0.024	0.005	0.182	0.009	0.044	0.124	0.114	0.022
rs1127926	*CRAT*	0.092	0.111	0.023	0.038	0.023	0.038	0.111	0.079	0.111	0.038
rs5744934	*POLE*	0.102	0.038	0.056	0.055	0.053	0.040	0.090	0.109	0.057	0.008
rs72978391	*SLCO2A1*	0.122	0.143	0.016	0.016	0.038	0.016	0.143	0.109	0.143	0.016
rs2275725	*ADAM8*	0.102	0.068	0.033	0.027	0.086	0.026	0.061	0.120	0.075	0.029
rs2073924	*GBGT1*	0.109	0.166	0.198	0.054	0.033	0.229	0.030	0.025	0.040	0.307
rs35245703	*DGKI*	0.163	0.419	0.012	0.025	0.025	0.025	0.419	0.163	0.419	0.025
rs3827848	*COL5A1*	0.063	0.067	0.020	0.032	0.010	0.045	0.058	0.033	0.054	0.041
rs11849	*FAM69B*	0.209	0.102	0.060	0.011	0.148	0.040	0.076	0.115	0.047	0.030
rs10111520	*COL22A1*	0.069	0.056	0.084	0.109	0.013	0.109	0.078	0.052	0.056	0.074

rs28664170	*PCDHB16*	0.110	0.207	0.028	0.069	0.028	0.054	0.186	0.140	0.269	0.075
rs6748626	*LRP1B*	0.122	0.180	0.017	0.012	0.041	0.019	0.082	0.063	0.122	0.026
rs45617032	*PPARGC1B*	0.120	0.141	0.015	0.015	0.037	0.015	0.141	0.107	0.141	0.015
rs35240348	*NPR1*	0.071	0.064	0.021	0.068	0.013	0.050	0.130	0.097	0.097	0.030
rs113890115	*C3orf33*	0.150	0.172	0.029	0.030	0.051	0.030	0.172	0.137	0.172	0.030
rs863361	*OR10X1*	0.137	0.104	0.016	0.012	0.037	0.023	0.104	0.060	0.043	0.011
rs863363	*OR10X1*	0.137	0.104	0.016	0.012	0.037	0.023	0.104	0.060	0.043	0.011
rs1548349	*FRMD1*	0.094	0.072	0.085	0.168	0.027	0.108	0.157	0.133	0.109	0.080
rs2253680	*FASTKD1*	0.107	0.063	0.069	0.063	0.054	0.052	0.085	0.102	0.074	0.041
rs9427213	*MROH9*	0.079	0.051	0.061	0.153	0.020	0.082	0.153	0.132	0.098	0.051
rs6861827	*STC2*	0.078	0.122	0.035	0.047	0.037	0.025	0.104	0.122	0.225	0.055
rs4868663	*GPRIN1*	0.050	0.092	0.022	0.027	0.020	0.012	0.035	0.045	0.116	0.066
rs12061876	*TOR3A*	0.084	0.141	0.011	0.015	0.013	0.015	0.141	0.084	0.141	0.015
rs12092348	*TOR3A*	0.084	0.141	0.011	0.015	0.013	0.015	0.141	0.084	0.141	0.015
rs35683768	*TTN*	0.092	0.111	0.023	0.006	0.023	0.029	0.048	0.029	0.048	0.029
rs627897	*TOR1AIP1*	0.101	0.058	0.070	0.063	0.086	0.040	0.082	0.139	0.082	0.040
rs2298206	*QSOX1*	0.069	0.082	0.012	0.012	0.021	0.012	0.058	0.050	0.082	0.015
rs61737927	*OR2Y1*	0.083	0.171	0.045	0.132	0.013	0.085	0.171	0.104	0.236	0.132
rs1143674	*ITGA4*	0.033	0.051	0.098	0.090	0.058	0.027	0.030	0.107	0.200	0.178
rs4667002	*ZNF804A*	0.156	0.098	0.089	0.056	0.072	0.080	0.100	0.110	0.074	0.054
rs1136644	*LPP*	0.122	0.206	0.016	0.019	0.022	0.030	0.140	0.065	0.152	0.032
rs788795	*LGR6*	0.194	0.068	0.062	0.057	0.137	0.039	0.180	0.261	0.105	0.011
rs788794	*LGR6*	0.091	0.036	0.050	0.079	0.054	0.023	0.139	0.198	0.119	0.013
rs788793	*LGR6*	0.194	0.068	0.062	0.057	0.137	0.039	0.180	0.261	0.105	0.011
rs3737875	*MYBPH*	0.095	0.058	0.052	0.093	0.042	0.045	0.156	0.186	0.126	0.032
rs6435421	*C2orf80*	0.054	0.054	0.054	0.054	0.027	0.054	0.054	0.054	0.054	0.054

rs2282319	*CAPN9*	0.143	0.273	0.142	0.286	0.026	0.253	0.288	0.158	0.306	0.271
rs3828126	*CAPN9*	0.100	0.208	0.085	0.108	0.023	0.123	0.108	0.070	0.182	0.208
rs79233817	*EDARADD*	0.120	0.141	0.076	0.109	0.037	0.109	0.141	0.107	0.141	0.109
rs12732924	*FMN2*	0.259	0.177	0.342	0.342	0.171	0.259	0.259	0.342	0.219	0.219
rs2240479	*FARP2*	0.060	0.045	0.060	0.064	0.018	0.064	0.045	0.041	0.045	0.064
rs41311583	*OR14I1*	0.102	0.058	0.128	0.092	0.051	0.113	0.053	0.069	0.040	0.094

Chapter 6

Discussion

6.1 Motivation

The work presented in this thesis investigated patterns of gene flow and genetic diversity in southern Africa and how this impacts infectious disease risk, with particular reference to TB susceptibility. Investigating the association between population structure and TB susceptibility is crucial to understand why some individuals in southern Africa are more susceptible to the progression to active TB disease than others. As the genetic diversity in the area is highly complex, many avenues of research need to be explored in order to fully understand this phenotype. Understanding genetic diversity in sub-Saharan Africa will further assist research in populations of African descent, which will also be important in selecting appropriate and efficacious TB vaccines for these populations. The ultimate goal is to provide novel hypotheses regarding the genetic history of populations in southern Africa in addition to novel tools and methodologies to identify the underlying genetic mechanisms in infectious disease susceptibility. The current methodologies used to combat the high TB incidence rates are inefficient (i.e. a partially effective TB vaccine (423) and lengthy antibiotic regimes that promote drug resistance (424)) and these approaches are unlikely to be the solution to the TB crisis. These methods also do not take the ancestry of the affected individuals into account and this probably contributed to the failure of the Ag85BESAT-6, Ag85B-TB10.4 and Mtb72f TB vaccines in most South African populations(142). For this reason a multipronged approach with an emphasis on the host's genetic make-up is vital.

6.2 Research Highlights

Reviewing past research is crucial to identify areas where improvements could be made and how diverse research fields may provide contrasting or supporting data. Previous research suggested that the admixture patterns seen in modern populations is greatly affected by past migration events and that these distinct patterns of admixture affect infectious disease risk. Furthermore, the lack of biomedical studies based on southern African populations illustrates the importance of inter-continental communication and collaborations between researchers. The review presented in **Chapter 2** reiterated the importance of the inclusion of population structure data in the assessment of disease risk and provided the opportunity to identify research topics that have not yet been exploited in southern African populations. Although African populations are gravely affected by TB, there is little research currently being conducted on these groups due to the lack of available genetic data. However, recent initiatives (discussed in section 6.3.1) have started to recognize the importance of including African populations in studies (such as GWAS) (94,101,102). This can be seen by the nearly

20% increase in non-European GWAS participants (425). This increase is however largely driven by Asian populations. Additionally, the inclusion of admixed populations is gaining momentum as most modern-day civilizations experienced admixture at some stage in history. Understanding the genetic diversity associated with these populations prior to making assumptions regarding genotype-phenotype associations is vitally important so as to determine the origin of the association and to ensure that the finding is not a false positive.

For this reason, understanding the genetic history of populations in *southern African* is crucial. One area that requires further investigation in this field, is the detailed analysis of population structure in KhoeSan populations as previous studies have not investigated the subject in depth. Initial research demonstrated that there *is* population structure within KhoeSan populations, but the patterns of structure were hypothesized to follow a very simplistic model (63,72,82). In this thesis, we showed that although there are fine-scale genetic differences between populations, these differences are much more complex than previously thought. The patterns observed largely corresponded to ecogeographic boundaries rather than language usage or subsistence strategies (**Chapter 3**). This association between structure and geography is consistent with a subsequent study by Montinaro et al. (2017). However, in contrast to our findings, which identified five distinct KhoeSan ancestral components, their study only detected three (70). One consistent result across these two studies is the presence of a southern African specific KhoeSan ancestral component.

Although several conclusions can be made from results presented in structure analysis, improved resolution and de-convolution of the apparent southern African specific KhoeSan ancestry can only occur once more KhoeSan populations are sampled. This is a limiting factor of recent studies. However, the number of southern African populations included in *this* study was the largest at the time and therefore provided a unique opportunity to develop novel hypotheses. This research not only broadens our knowledge regarding southern African genetic history, but it contributes to the knowledge of how this history may affect phenotypes such as infectious disease risk. As we had improved our hypotheses regarding population history in Chapter 3, we were able to more efficiently model admixture and early population movements. This enabled the efficient inclusion of admixture as a confounder in a post-GWAS analysis and the estimation of the extent and impact of selective pressure caused by foreign pathogens in southern Africa.

A standard GWAS makes use of hundreds of thousands of markers and a large case-control cohort to identify variants associated with a phenotype. This can be costly, time-consuming and inefficient. Furthermore, the majority of these variants are intronic with unknown functional consequences (426). Since GWAS have been largely uninformative on their own, in this thesis we proposed a post-GWAS approach methodology which utilized known LD patterns in ancestral populations of the SAC, to identify functional variants associated with TB susceptibility. Three variants were shown to provide a protective function whereas another three increased the odds of progressing to active TB (**Chapter 4**). This study reaffirms that TB susceptibility is a complex phenotype affected by multiple variants each with moderate effect sizes, suggesting that a standard GWAS might not be as efficient as previously thought. Moreover, at present TB susceptibility is classified as an polygenic trait governed by multiple genes (427–431). This underlying assumption has recently been modified to suggest an omnigenic model of complex disease which proposes that all genetic variants that actively regulate gene expression in tissues relevant to disease will directly or indirectly affect the phenotype via interconnected pathways (432,433). This concept is supported by the selection scan findings presented in this thesis where multiple signals of positive selection were identified following a statistically stringent selection scan of exomes from the SAC population. However, these signals were not in immune response genes characteristically associated with infectious disease risk profiles. Pathway over-representation analysis of the candidate gene list implicated the involvement of two inter-connected KEGG pathways (413) namely, the focal adhesion and ECM receptor interaction pathways. Furthermore, the candidate gene list was enriched for genes associated with an abnormal immune system (**Chapter 5**). The identification of these interconnected pathways and a broad phenotype under positive selection once again supports an omnigenic model for TB susceptibility.

Overall, these studies provided novel candidate genes and pathways which may be of interest to researchers investigating infectious disease risk in unique populations. These findings require replication in another cohort with similar population characteristics as the one presented here, however this is not possible at present.

6.3 Ongoing and future work

6.3.1 **Further sampling**

As discussed in the previous section, despite African-specific efforts to increase genomic research in the regions, there is a lack of data from diverse populations in southern Africa. In addition, truly representative reference populations are scarce and the correct identification of these populations is difficult at best. International consortia (such as Human Heredity and

Health in Africa (H3Africa) and the African Genome Variation Project) (94,101) are now sampling further African populations with biomedical applications in mind. Although data from these projects was not available at the time of analysis, every effort was made to gather as much genetic data as possible from a variety of populations. Further sampling from other populations will improve future results.

One of the outcomes of the research presented in this thesis has been the establishment of a new TB case-control sample collection in the Northern Cape. The goal is to obtain more than 1000 samples from individuals of diverse backgrounds and to have disease-specific phenotypes available. We have trained 12 nurses in clinics in the Northern Cape to take saliva samples with informed consent from TB suspects. The saliva sample will be taken in addition to ethnicity, body measurements and other phenotype information. We are approximately halfway through data collation for approximately 400 samples and will start DNA extraction from the saliva samples soon. This genetic and phenotypic data will significantly contribute to our ancestry-specific TB susceptibility studies and will serve as a southern African replication cohort. Given the heterogeneous nature of TB, it may also be necessary to clearly define the TB phenotype in the patients before more extensive studies are conducted.

6.3.2 **Southern African specific recombination map**

As computational tools improve and software programs are perhaps tailored to unique population scenarios, the conclusions made by researchers will become more accurate. A recombination map specific to admixed southern African populations would greatly benefit the biomedical and population genetic fields. Local ancestry inference (LAI) is one of many analyses which is highly dependent on an accurate recombination map. The unsurmountable gain of knowledge brought about by a southern African specific recombination map and more accurate LAI will in turn lead to clearer hypotheses regarding the genetic history of southern African populations (by estimating migration routes and extent of admixture) and in turn perform an in-depth, multi-statistic selection scan to identify genetic regions where natural selection has acted.

6.4 Concluding remarks

To conclude, this thesis built on a review of prior knowledge regarding population structure work in southern Africa and provided further novel information regarding the genetic prehistory of these populations. Understanding southern African population diversity and structure impacted various avenues of scientific research. Firstly, in the determination of regions of the genome that are under positive selection and secondly in the identification of novel variants that were associated with TB susceptibility. These fields coalesce at the point of understanding human history, adaptation and improved biomedical hypotheses regarding infectious disease. The work presented here contributes to these fields and to the ultimate aim of improving the global health status.

References

1. Corbett EL, Watt CJ, Walker N, et al. The growing burden of tuberculosis: Global trends and interactions with the hiv epidemic. Arch Intern Med. 2003 May 12;163(9):1009–21.

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

3. Raviglione MC. The TB epidemic from 1992 to 2002. Tuberculosis. 2003 Feb;83(1–3):4–14.

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

5. WHO | Global tuberculosis report 2016 [Internet]. WHO. [cited 2016 Dec 20]. Available from: http://www.who.int/tb/publications/global_report/en/

6. Kaufmann SH, Hussey G, Lambert P-H. New vaccines for tuberculosis. The Lancet. 375(9731):2110–9.

7. Cooper AM. Cell mediated immune responses in Tuberculosis. Annu Rev Immunol. 2009;27:393–422.

8. Cave AJE, Demonstrator A. The evidence for the incidence of tuberculosis in ancient Egypt. Br J Tuberc. 1939 Jul;33(3):142–52.

9. Comas I, Hailu E, Kiros T, Bekele S, Mekonnen W, Gumi B, et al. Population Genomics of Mycobacterium tuberculosis in Ethiopia Contradicts the Virgin Soil Hypothesis for Human Tuberculosis in Sub-Saharan Africa. Curr Biol CB. 2015 Dec 21;25(24):3260–6.

10. Comas I, Coscolla M, Luo T, Borrell S, Holt KE, Kato-Maeda M, et al. Out-of-Africa migration and Neolithic coexpansion of Mycobacterium tuberculosis with modern humans. Nat Genet. 2013 Oct;45(10):1176–82.

11. Gutierrez MC, Brisse S, Brosch R, Fabre M, Omaïs B, Marmiesse M, et al. Ancient origin and gene mosaicism of the progenitor of Mycobacterium tuberculosis. PLoS Pathog. 2005 Sep;1(1):e5.

12. Wirth T, Hildebrand F, Allix-Béguec C, Wölbeling F, Kubica T, Kremer K, et al. Origin, spread and demography of the Mycobacterium tuberculosis complex. PLoS Pathog. 2008 Sep 19;4(9):e1000160.

13. Collins TF. The history of southern Africa's first tuberculosis epidemic. South Afr Med J Suid-Afr Tydskr Vir Geneeskd. 1982 Nov 13;62(21):780–8.

14. Cummins SL. "VIRGIN SOIL"-AND AFTER: A WORKING CONCEPTION OF TUBERCULOSIS IN CHILDREN, ADOLESCENTS, AND ABORIGINES. Br Med J. 1929 Jul 13;2(3575):39–41.

15. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature. 2014 Oct 23;514(7523):494–7.

16. Hershkovitz I, Donoghue HD, Minnikin DE, Besra GS, Lee OY-C, Gernaey AM, et al. Detection and Molecular Characterization of 9000-Year-Old Mycobacterium tuberculosis from a Neolithic Settlement in the Eastern Mediterranean. PLoS ONE. 2008 Oct 15;3(10):e3426.

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

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

19. Fox GJ, Orlova M, Schurr E. Tuberculosis in Newborns: The Lessons of the "Lübeck Disaster" (1929–1933). PLOS Pathog. 2016 Jan 21;12(1):e1005271.

20. Moegling A. Die „Epidemiologie" der Lübecker Säuglingstuberkulose. In: Die Säuglingstuberkulose in Lübeck [Internet]. Springer Berlin Heidelberg; 1935 [cited 2016 Aug 30]. p. 1–24. Available from: http://link.springer.com/chapter/10.1007/978-3-642-92013-4_1

21. Comstock GW. Tuberculosis in twins: a re-analysis of the Prophit survey. Am Rev Respir Dis. 1978 Apr;117(4):621–4.

22. Jepson A, Fowler A, Banya W, Singh M, Bennett S, Whittle H, et al. Genetic regulation of acquired immune responses to antigens of Mycobacterium tuberculosis: a study of twins in West Africa. Infect Immun. 2001 Jun;69(6):3989–94.

23. Newport MJ, Goetghebuer T, Weiss HA, Whittle H, Siegrist C-A, Marchant A, et al. Genetic regulation of immune responses to vaccines in early life. Genes Immun. 2004 Mar;5(2):122–9.

24. Cobat A, Gallant CJ, Simkin L, Black GF, Stanley K, Hughes J, et al. High heritability of antimycobacterial immunity in an area of hyperendemicity for tuberculosis disease. J Infect Dis. 2010 Jan 1;201(1):15–9.

25. Abel L, El-Baghdadi J, Bousfiha AA, Casanova J-L, Schurr E. Human genetics of tuberculosis: a long and winding road. Phil Trans R Soc B. 2014 Jun 19;369(1645):20130428.

26. Abhimanyu, Bose M, Jha P, Indian Genome Variation Consortium. Footprints of genetic susceptibility to pulmonary tuberculosis: cytokine gene variants in north Indians. Indian J Med Res. 2012 May;135(5):763–70.

27. Puffer R. Familial susceptibility to tuberculosis. 1946.

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

29. Thye T, Browne EN, Chinbuah MA, Gyapong J, Osei I, Owusu-Dabo E, et al. IL10 haplotype associated with tuberculin skin test response but not with pulmonary TB. PloS One. 2009;4(5):e5420.

30. Thye T, Owusu-Dabo E, Vannberg FO, van Crevel R, Curtis J, Sahiratmadja E, et al. Common variants at 11p13 are associated with susceptibility to tuberculosis. Nat Genet. 2012 Feb 5;44(3):257–9.

31. Duarte R, Carvalho C, Pereira C, Bettencourt A, Carvalho A, Villar M, et al. HLA class II alleles as markers of tuberculosis susceptibility and resistance. Rev Port Cardiol. 2011 Jan 1;17(1):15–9.

32. Wu F, Zhang W, Zhang L, Wu J, Li C, Meng X, et al. NRAMP1, VDR, HLA-DRB1, and HLA-DQB1 Gene Polymorphisms in Susceptibility to Tuberculosis among the Chinese Kazakh Population: A Case-Control Study, NRAMP1, VDR, HLA-DRB1, and HLA-DQB1 Gene Polymorphisms in Susceptibility to Tuberculosis among the Chinese Kazakh Population: A Case-Control Study. BioMed Res Int BioMed Res Int. 2013 Jun 10;2013, 2013:e484535.

33. Salie M, van der Merwe L, Möller M, Daya M, van der Spuy GD, van Helden PD, et al. Associations between human leukocyte antigen class I variants and the Mycobacterium tuberculosis subtypes causing disease. J Infect Dis. 2014 Jan 15;209(2):216–23.

34. Songane M, Kleinnijenhuis J, Alisjahbana B, Sahiratmadja E, Parwati I, Oosting M, et al. Polymorphisms in Autophagy Genes and Susceptibility to Tuberculosis. PLoS ONE [Internet]. 2012 Aug 6 [cited 2016 Jan 4];7(8). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3412843/

35. King KY, Lew JD, Ha NP, Lin JS, Ma X, Graviss EA, et al. Polymorphic allele of human IRGM1 is associated with susceptibility to tuberculosis in African Americans. PloS One. 2011;6(1):e16317.

36. Intemann CD, Thye T, Niemann S, Browne ENL, Amanua Chinbuah M, Enimil A, et al. Autophagy gene variant IRGM -261T contributes to protection from tuberculosis caused by Mycobacterium tuberculosis but not by M. africanum strains. PLoS Pathog. 2009 Sep;5(9):e1000577.

37. Che N, Li S, Gao T, Zhang Z, Han Y, Zhang X, et al. Identification of a novel IRGM promoter single nucleotide polymorphism associated with tuberculosis. Clin Chim Acta Int J Clin Chem. 2010 Nov 11;411(21–22):1645–9.

38. Zhang J, Chen Y, Nie X-B, Wu W-H, Zhang H, Zhang M, et al. Interleukin-10 polymorphisms and tuberculosis susceptibility: a meta-analysis. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2011 May;15(5):594–601.

39. Zhang J, Zheng L, Zhu D, An H, Yang Y, Liang Y, et al. Polymorphisms in the interleukin 18 receptor 1 gene and tuberculosis susceptibility among Chinese. PloS One. 2014;9(10):e110734.

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

41. Stead WW, Senner JW, Reddick WT, Lofgren JP. Racial differences in susceptibility to infection by Mycobacterium tuberculosis. N Engl J Med. 1990 Feb 15;322(7):422–7.

42. Chimusa ER, Mbiyavanga M, Mazandu GK, Mulder NJ. ancGWAS: a post genome-wide association study method for interaction, pathway and ancestry analysis in homogeneous and admixed populations. Bioinforma Oxf Engl. 2016 Feb 15;32(4):549–56.

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

44. Deschamps M, Laval G, Fagny M, Itan Y, Abel L, Casanova J-L, et al. Genomic Signatures of Selective Pressures and Introgression from Archaic Hominins at Human Innate Immunity Genes. Am J Hum Genet. 2016 Jan 7;98(1):5–21.

45. Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, et al. Identifying Recent Adaptations in Large-scale Genomic Data. Cell. 2013 Feb 14;152(4):703–13.

46. Grigg ER. The arcana of tuberculosis with a brief epidemiologic history of the disease in the U.S.A. Am Rev Tuberc. 1958 Aug;78(2):151–172 contd.

47. Stead WW. Genetics and resistance to tuberculosis. Could resistance be enhanced by genetic engineering? Ann Intern Med. 1992 Jun 1;116(11):937–41.

48. Stead WW, Senner JW, Reddick WT, Lofgren JP. Racial differences in susceptibility to infection by Mycobacterium tuberculosis. N Engl J Med. 1990 Feb 15;322(7):422–7.

49. Fourie J, van Zanden JL. GDP in the Dutch Cape Colony: The National Accounts of a Slave-Based Society. South Afr J Econ. 2013 Dec 1;81(4):467–90.

50. Macvicar N. Tuberculosis among the South African natives. South Afr Med Rec. 1903 1910;1-3-8.

51. Chimusa ER, Daya M, Möller M, Ramesar R, Henn BM, van Helden PD, et al. Determining Ancestry Proportions in Complex Admixture Scenarios in South Africa Using a Novel Proxy Ancestry Selection

Method. PLoS ONE [Internet]. 2013 Sep 16 [cited 2015 Jan 30];8(9). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3774743/

52. Daya M, van der Merwe L, Galal U, Möller M, Salie M, Chimusa ER, et al. A panel of ancestry informative markers for the complex five-way admixed South African coloured population. PloS One. 2013;8(12):e82224.

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

54. Petersen DC, Libiger O, Tindall EA, Hardie R-A, Hannick LI, Glashoff RH, et al. Complex patterns of genomic admixture within southern Africa. PLoS Genet. 2013;9(3):e1003309.

55. Uren C, Kim M, Martin AR, Bobo D, Gignoux CR, van Helden PD, et al. Fine-Scale Human Population Structure in Southern Africa Reflects Ecogeographic Boundaries. Genetics. 2016 Jul 29;

56. Cooke GS, Hill AVS. Genetics of susceptibitlity to human infectious disease. Nat Rev Genet. 2001 Dec;2(12):967–77.

57. Doeschl-Wilson AB, Davidson R, Conington J, Roughsedge T, Hutchings MR, Villanueva B. Implications of Host Genetic Variation on the Risk and Prevalence of Infectious Diseases Transmitted Through the Environment. Genetics. 2011 Jul;188(3):683–93.

58. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet. 2001 Jul;69(1):124–37.

59. Vasseur E, Quintana-Murci L. The impact of natural selection on health and disease: uses of the population genetics approach in humans. Evol Appl. 2013 Jun;6(4):596–607.

60. Bräuer G, Rimbach KW. Late archaic and modern Homo sapiens from Europe, Africa, and Southwest Asia: Craniometric comparisons and phylogenetic implications. J Hum Evol. 1990 Dec 1;19(8):789–807.

61. Henn BM, Gignoux CR, Jobin M, Granka JM, Macpherson JM, Kidd JM, et al. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. Proc Natl Acad Sci U S A. 2011 Mar 29;108(13):5154–62.

62. Leakey REF. Early Homo sapiens Remains from the Omo River Region of South-west Ethiopia: Faunal Remains from the Omo Valley. Nature. 1969 Jun 21;222(5199):1132–3.

63. Pickrell JK, Patterson N, Barbieri C, Berthold F, Gerlach L, Güldemann T, et al. The genetic prehistory of southern Africa. Nat Commun. 2012;3:1143.

64. Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, et al. The genetic structure and history of Africans and African Americans. Science. 2009 May 22;324(5930):1035–44.

65. Veeramah KR, Wegmann D, Woerner A, Mendez FL, Watkins JC, Destro-Bisol G, et al. An early divergence of KhoeSan ancestors from those of other modern humans is supported by an ABC-based analysis of autosomal resequencing data. Mol Biol Evol. 2012 Feb;29(2):617–30.

66. White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G, et al. Pleistocene Homo sapiens from Middle Awash, Ethiopia. Nature. 2003 Jun 12;423(6941):742–7.

67. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A. 2005 Nov 1;102(44):15942–7.

68. Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL, Underhill PA. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. Am J Hum Genet. 2002 Jan;70(1):265–8.

69. Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC. African populations and the evolution of human mitochondrial DNA. Science. 1991 Sep 27;253(5027):1503–7.

70. Montinaro F, Busby GBJ, Gonzalez-Santos M, Oosthuitzen O, Oosthuitzen E, Anagnostou P, et al. Complex Ancient Genetic Structure and Cultural Transitions in Southern African Populations. Genetics. 2017 Jan;205(1):303–16.

71. Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, et al. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. Proc Natl Acad Sci. 2008 Aug 5;105(31):10693–8.

72. Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 2012 Oct 19;338(6105):374–9.

73. Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, et al. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol. 2007 Oct;24(10):2180–95.

74. Russell T, Silva F, Steele J. Modelling the Spread of Farming in the Bantu-Speaking Regions of Africa: An Archaeology-Based Phylogeography. PLoS ONE [Internet]. 2014 Jan 31 [cited 2016 Aug 30];9(1). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3909244/

75. Beleza S, Gusmão L, Amorim A, Carracedo A, Salas A. The genetic legacy of western Bantu migrations. Hum Genet. 2005 Aug 1;117(4):366–75.

76. Newman JL. The Peopling of Africa: A Geographic Interpretation. Yale University Press; 1995. 260 p.

77. Phillipson DW. African Archaeology. Cambridge University Press; 2005. 407 p.

78. Grollemund R, Branford S, Bostoen K, Meade A, Venditti C, Pagel M. Bantu expansion shows that habitat alters the route and pace of human dispersals. Proc Natl Acad Sci. 2015 Oct 27;112(43):13296–301.

79. Lane A b., Soodyall H, Arndt S, Ratshikhopha M e., Jonker E, Freeman C, et al. Genetic substructure in South African Bantu-speakers: Evidence from autosomal DNA and Y-chromosome studies. Am J Phys Anthropol. 2002 Oct 1;119(2):175–85.

80. Cilliers J, Fourie J. New estimates of settler life span and other emographic trends in South Africa, 1652-1948. Stellenbosch Work Pap Ser. 2012;WP20/2012.

81. Croix SL. The Decline of the Khoikhoi Population, 1652-1780: A Review and a New Estimate [Internet]. University of Hawaii at Manoa, Department of Economics; 2016 [cited 2017 May 5]. Report No.: 201622. Available from: https://ideas.repec.org/p/hai/wpaper/201622.html

82. Schlebusch CM, Lombard M, Soodyall H. MtDNA control region variation affirms diversity and deep sub-structure in populations from southern Africa. BMC Evol Biol. 2013;13:56.

83. Schlebusch CM, Soodyall H. Extensive population structure in San, Khoe, and mixed ancestry populations from southern Africa revealed by 44 short 5-SNP haplotypes. Hum Biol. 2012 Dec;84(6):695–724.

84. Breton G, Schlebusch CM, Lombard M, Sjödin P, Soodyall H, Jakobsson M. Lactase persistence alleles reveal partial East african ancestry of southern african Khoe pastoralists. Curr Biol CB. 2014 Apr 14;24(8):852–8.

85. Macholdt E, Lede V, Barbieri C, Mpoloka SW, Chen H, Slatkin M, et al. Tracing pastoralist migrations to southern Africa with lactase persistence alleles. Curr Biol CB. 2014 Apr 14;24(8):875–9.

86. Boonzaier E. The Cape Herders: A History of the Khoikhoi of Southern Africa. New Africa Books; 1996. 158 p.

87. Dunne J, Evershed RP, Salque M, Cramp L, Bruni S, Ryan K, et al. First dairying in green Saharan Africa in the fifth millennium BC. Nature. 2012 Jun 21;486(7403):390–4.

88. Jerardino A, Fort J, Isern N, Rondelli B. Cultural diffusion was the main driving mechanism of the Neolithic transition in southern Africa. PloS One. 2014;9(12):e113672.

89. Pleurdeau D, Imalwa E, Détroit F, Lesur J, Veldman A, Bahain J-J, et al. "Of Sheep and Men": Earliest Direct Evidence of Caprine Domestication in Southern Africa at Leopard Cave (Erongo, Namibia). PLoS ONE. 2012 Jul 11;7(7):e40340.

90. Sadr K. Invisible herders? The archaeology of Khoekhoe pastoralists. South Afr Humanit. 2008;20(1):179–203.

91. Sadr K. Livestock First Reached Southern Africa in Two Separate Events. PloS One. 2015;10(8):e0134215.

92. Robbins LH, Campbell AC, Murphy ML, Brook GA, Srivastava P, Badenhorst S. The Advent of Herding in Southern Africa: Early AMS Dates on Domestic Livestock from the Kalahari Desert. Curr Anthropol. 2005;46(4):671–7.

93. Chimusa ER, Meintjies A, Tchanga M, Mulder N, Seoighe C, Soodyall H, et al. A Genomic Portrait of Haplotype Diversity and Signatures of Selection in Indigenous Southern African Populations. PLoS Genet. 2015 Mar 26;11(3):e1005052.

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

95. Weigend GG. German Settlement Patterns in Namibia. Geogr Rev. 1985;75(2):156–69.

96. Die herkoms van die Afrikaner, 1657-1867. A. A. Balkema; 1971. 335 p.

97. Greeff JM. Deconstructing Jaco: genetic heritage of an Afrikaner. Ann Hum Genet. 2007 Sep;71(Pt 5):674–88.

98. De Villiers CC PC. Geslagsregisters van di our Kaapse families. Vol. 1 A-K. Cape Town: A.A Balkema; 1966.

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

100. Rito T, Richards MB, Fernandes V, Alshamali F, Cerny V, Pereira L, et al. The First Modern Human Dispersals across Africa. PLOS ONE. 2013 Nov 13;8(11):e80031.

101. Adoga MP, Fatumo SA, Agwale SM. H3Africa: a tipping point for a revolution in bioinformatics, genomics and health research in Africa. Source Code Biol Med. 2014;9:10.

102. Ramsay M. Growing genomic research on the African continent: The H3Africa Consortium. South Afr Med J Suid-Afr Tydskr Vir Geneeskd. 2015 Dec;105(12):1016–7.

103. Chen YS, Olckers A, Schurr TG, Kogelnik AM, Huoponen K, Wallace DC. mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. Am J Hum Genet. 2000 Apr;66(4):1362–83.

104. Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, et al. Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol. 2001 Jul;18(7):1189–203.

105. Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, et al. African Y Chromosome and mtDNA Divergence Provides Insight into the History of Click Languages. Curr Biol. 2003 Mar 18;13(6):464–73.

106. Brown KS, Marean CW, Herries AIR, Jacobs Z, Tribolo C, Braun D, et al. Fire as an engineering tool of early modern humans. Science. 2009 Aug 14;325(5942):859–62.

107. Schlebusch CM, Naidoo T, Soodyall H. SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. Electrophoresis. 2009 Nov;30(21):3657–64.

108. Naidoo T, Schlebusch CM, Makkan H, Patel P, Mahabeer R, Erasmus JC, et al. Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. Investig Genet. 2010;1(1):6.

109. Marean CW. Pinnacle Point Cave 13B (Western Cape Province, South Africa) in context: The Cape Floral kingdom, shellfish, and modern human origins. J Hum Evol. 2010 Oct;59(3–4):425–43.

110. Brown KS, Marean CW, Jacobs Z, Schoville BJ, Oestmo S, Fisher EC, et al. An early and enduring advanced technology originating 71,000 years ago in South Africa. Nature. 2012 Nov 22;491(7425):590–3.

111. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet. 2011 Oct;43(10):1031–4.

112. Henn BM, Cavalli-Sforza LL, Feldman MW. The great human expansion. Proc Natl Acad Sci U S A. 2012 Oct 30;109(44):17758–64.

113. Scheinfeldt LB, Soi S, Tishkoff SA. In the Light of Evolution. National Academy of Sciences; 2010. (5. Working Towards a Synthesis of Archaeological, Linguistic, and Genetic Data for Inferring African Population History; vol. 4: The Human Condition).

114. Batini C, Jobling MA. The jigsaw puzzle of our African ancestry: unsolved, or unsolvable? Genome Biol. 2011;12(6):118.

115. Behar DM, van Oven M, Rosset S, Metspalu M, Loogväli E-L, Silva NM, et al. A “Copernican” reassessment of the human mitochondrial DNA tree from its root. Am J Hum Genet. 2012 Apr 6;90(4):675–84.

116. Poznik GD, Henn BM, Yee M-C, Sliwerska E, Euskirchen GM, Lin AA, et al. Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. Science. 2013 Aug 2;341(6145):562–5.

117. Haber M, Mezzavilla M, Xue Y, Tyler-Smith C. Ancient DNA and the rewriting of human history: be sparing with Occam's razor. Genome Biol [Internet]. 2016 [cited 2016 Nov 29];17. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4707776/

118. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016 Oct 13;538(7624):201–6.

119. Kidd JM, Sharpton TJ, Bobo D, Norman PJ, Martin AR, Carpenter ML, et al. Exome capture from saliva produces high quality genomic and metagenomic data. BMC Genomics. 2014 Apr 4;15:262.

120. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet. 2007 Jan;39(1):31–40.

121. Schlebusch CM, Skoglund P, Sjödin P, Gattepaille LM, Hernandez D, Jay F, et al. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science. 2012 Oct 19;338(6105):374–9.

122. Macholdt E, Slatkin M, Pakendorf B, Stoneking M. New insights into the history of the C-14010 lactase persistence variant in Eastern and Southern Africa. Am J Phys Anthropol. 2015 Apr;156(4):661–4.

123. Smith A. The origins of Herding in Southern Africa: Debating the "Neolithic" model. Saarbrücken: LAP LAMBERT Academic Publishing; 2014. 68 p.

124. Ranciaro A, Campbell MC, Hirbo JB, Ko W-Y, Froment A, Anagnostou P, et al. Genetic Origins of Lactase Persistence and the Spread of Pastoralism in Africa. Am J Hum Genet. 2014 Apr 3;94(4):496–510.

125. Pickrell JK, Patterson N, Loh P-R, Lipson M, Berger B, Stoneking M, et al. Ancient west Eurasian ancestry in southern and eastern Africa. Proc Natl Acad Sci. 2014 Feb 18;111(7):2632–7.

126. Patterson N, Petersen DC, van der Ross RE, Sudoyo H, Glashoff RH, Marzuki S, et al. Genetic structure of a unique admixed population: implications for medical research. Hum Mol Genet. 2010 Feb 1;19(3):411–9.

127. Cann HM. A human genome diversity cell line panel. (letter). Science. 2002;296:261.

128. Uren C, Kim M, Martin AR, Bobo D, Gignoux CR, Helden PD van, et al. Fine-scale human population structure in southern Africa reflects ecological boundaries. bioRxiv. 2016 Feb 3;38729.

129. Barbieri C, Vicente M, Rocha J, Mpoloka SW, Stoneking M, Pakendorf B. Ancient Substructure in Early mtDNA Lineages of Southern Africa. Am J Hum Genet. 2013 Feb 7;92(2):285–92.

130. Barbieri C, Güldemann T, Naumann C, Gerlach L, Berthold F, Nakagawa H, et al. Unraveling the complex maternal history of Southern African Khoisan populations. Am J Phys Anthropol. 2014 Mar;153(3):435–48.

131. Barbieri C, Vicente M, Oliveira S, Bostoen K, Rocha J, Stoneking M, et al. Migration and Interaction in a Contact Zone: mtDNA Variation among Bantu-Speakers in Southern Africa. PLOS ONE. 2014 Jun 5;9(6):e99117.

132. Barbieri C, Hübner A, Macholdt E, Ni S, Lippold S, Schröder R, et al. Refining the Y chromosome phylogeny with southern African sequences. Hum Genet. 2016 Apr 4;135(5):541–53.

133. Walker EA. A History of Southern Africa. Longmans, Green; 1928. 623 p.

134. Marks SJ, Montinaro F, Levy H, Brisighelli F, Ferri G, Bertoncini S, et al. Static and moving frontiers: the genetic landscape of Southern African Bantu-speaking populations. Mol Biol Evol. 2014 Sep 14;msu263.

135. Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. The genetic ancestry of African Americans, Latinos, and European Americans across the United States. Am J Hum Genet. 2015 Jan 8;96(1):37–53.

136. Kenyon C, Buyze J, Colebunders R. HIV Prevalence by Race Co-Varies Closely with Concurrency and Number of Sex Partners in South Africa. PLOS ONE. 2013 May 21;8(5):e64080.

137. Sveinbjornsson G, Gudbjartsson DF, Halldorsson BV, Kristinsson KG, Gottfredsson M, Barrett JC, et al. HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. Nat Genet. 2016 Mar;48(3):318–22.

138. Balamurugan A, Sharma SK, Mehra NK. Human leukocyte antigen class I supertypes influence susceptibility and severity of tuberculosis. J Infect Dis. 2004 Mar 1;189(5):805–11.

139. Kettaneh A, Seng L, Tiev KP, Tolédano C, Fabre B, Cabane J. Human leukocyte antigens and susceptibility to tuberculosis: a meta-analysis of case-control studies. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2006 Jul;10(7):717–25.

140. Kostyu DD, Hannick LI, Traweek JL, Ghanayem M, Heilpern D, Dawson DV. HLA class I polymorphism: structure and function and still questions. Hum Immunol. 1997 Sep 15;57(1):1–18.

141. Lombard Z, Brune AE, Hoal EG, Babb C, Van Helden PD, Epplen JT, et al. HLA class II disease associations in southern Africa. Tissue Antigens. 2006 Feb;67(2):97–110.

142. Salie M. The Role of the Major Histocompatibility Complex and the Leukocyte Receptor Complex Genes in Susceptibility to Tuberculosis in a South African Population. [Tygerberg]: University of Stellenbosch; 2014.

143. Cooke GS, Campbell SJ, Bennett S, Lienhardt C, McAdam KPWJ, Sirugo G, et al. Mapping of a novel susceptibility locus suggests a role for MC3R and CTSZ in human tuberculosis. Am J Respir Crit Care Med. 2008 Jul 15;178(2):203–7.

144. Adams LA, Möller M, Nebel A, Schreiber S, van der Merwe L, van Helden PD, et al. Polymorphisms in MC3R promoter and CTSZ 3'UTR are associated with tuberculosis susceptibility. Eur J Hum Genet EJHG. 2011 Jun;19(6):676–81.

145. Möller M, Nebel A, van Helden PD, Schreiber S, Hoal EG. Analysis of eight genes modulating interferon gamma and human genetic susceptibility to tuberculosis: a case-control association study. BMC Infect Dis. 2010;10:154.

146. Manry J, Laval G, Patin E, Fornarino S, Tichit M, Bouchier C, et al. Evolutionary genetics evidence of an essential, nonredundant role of the IFN-γ pathway in protective immunity. Hum Mutat. 2011 Jun;32(6):633–42.

147. Rossouw M, Nel HJ, Cooke GS, van Helden PD, Hoal EG. Association between tuberculosis and a polymorphic NFkappaB binding site in the interferon gamma gene. Lancet. 2003 May 31;361(9372):1871–2.

148. Tian C, Zhang Y, Zhang J, Deng Y, Li X, Xu D, et al. The +874T/A polymorphism in the interferon-γ gene and tuberculosis risk: an update by meta-analysis. Hum Immunol. 2011 Nov;72(11):1137–42.

149. Mabunda N, Alvarado-Arnez LE, Vubil A, Mariamo A, Pacheco AG, Jani IV, et al. Gene polymorphisms in patients with pulmonary tuberculosis from Mozambique. Mol Biol Rep. 2015 Jan;42(1):71–6.

150. Anoosheh S, Farnia P, Kargar M. Association between TNF-Alpha (-857) Gene Polymorphism and Susceptibility to Tuberculosis. Iran Red Crescent Med J. 2011 Apr;13(4):243–8.

151. Yi Y-X, Han J-B, Zhao L, Fang Y, Zhang Y-F, Zhou G-Y. Tumor necrosis factor alpha gene polymorphism contributes to pulmonary tuberculosis susceptibility: evidence from a meta-analysis. Int J Clin Exp Med. 2015;8(11):20690–700.

152. Cobat A, Gallant CJ, Simkin L, Black GF, Stanley K, Hughes J, et al. Two loci control tuberculin skin test reactivity in an area hyperendemic for tuberculosis. J Exp Med. 2009 Nov 23;206(12):2583–91.

153. Cobat A, Hoal EG, Gallant CJ, Simkin L, Black GF, Stanley K, et al. Identification of a major locus, TNF1, that controls BCG-triggered tumor necrosis factor production by leukocytes in an area hyperendemic for tuberculosis. Clin Infect Dis Off Publ Infect Dis Soc Am. 2013 Oct;57(7):963–70.

154. Chen Y-C, Hsiao C-C, Chen C-J, Chin C-H, Liu S-F, Wu C-C, et al. Toll-like receptor 2 gene polymorphisms, pulmonary tuberculosis, and natural killer cell counts. BMC Med Genet. 2010;11:17.

155. Sun Q, Zhang Q, Xiao H, Bai C. Toll-like receptor polymorphisms and tuberculosis susceptibility: A comprehensive meta-analysis. J Huazhong Univ Sci Technol Med Sci Hua Zhong Ke Ji Xue Xue Bao Yi Xue Ying Wen Ban Huazhong Keji Daxue Xuebao Yixue Yingdewen Ban. 2015 Apr;35(2):157–68.

156. Schurz H, Daya M, Möller M, Hoal EG, Salie M. TLR1, 2, 4, 6 and 9 Variants Associated with Tuberculosis Susceptibility: A Systematic Review and Meta-Analysis. PLoS ONE [Internet]. 2015 Oct 2 [cited 2016 May 3];10(10). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4592262/

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

158. Drögemöller B, Plummer M, Korkie L, Agenbag G, Dunaiski A, Niehaus D, et al. Characterization of the genetic variation present in CYP3A4 in three South African populations. Front Genet. 2013;4:17.

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

160. Feng Y, Broder CC, Kennedy PE, Berger EA. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. Science. 1996 May 10;272(5263):872–7.

161. Alkhatib G, Combadiere C, Broder CC, Feng Y, Kennedy PE, Murphy PM, et al. CC CKR5: a RANTES, MIP-1alpha, MIP-1beta receptor as a fusion cofactor for macrophage-tropic HIV-1. Science. 1996 Jun 28;272(5270):1955–8.

162. Choe H, Farzan M, Sun Y, Sullivan N, Rollins B, Ponath PD, et al. The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. Cell. 1996 Jun 28;85(7):1135–48.

163. Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, Farber CM, et al. Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature. 1996 Aug 22;382(6593):722–5.

164. Picton ACP, Shalekoff S, Paximadis M, Tiemessen CT. Marked differences in CCR5 expression and activation levels in two South African populations. Immunology. 2012 Aug;136(4):397–407.

165. Schlebusch C. Issues raised by use of ethnic-group names in genome study. Nature. 2010 Mar 25;464(7288):487; author reply 487.

166. Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, et al. The dawn of human matrilineal diversity. Am J Hum Genet. 2008 May;82(5):1130–40.

167. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014 Sep 18;513(7518):409–13.

168. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008 Feb 22;319(5866):1100–4.

169. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009 Sep 1;19(9):1655–64.

170. Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. pong: fast analysis and visualization of latent clusters in population genetic data. bioRxiv. 2015 Nov 14;31815.

171. Petkova D, Novembre J, Stephens M. Visualizing spatial population structure with estimated effective migration surfaces. Nat Genet. 2016 Jan;48(1):94–100.

172. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinforma Oxf Engl. 2011 Aug 1;27(15):2156–8.

173. Creanza N, Ruhlen M, Pemberton TJ, Rosenberg NA, Feldman MW, Ramachandran S. A comparison of worldwide phonemic and genetic variation in human populations. Proc Natl Acad Sci. 2015 Feb 3;112(5):1265–72.

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

175. Zhou H, Alexander D, Lange K. A quasi-Newton acceleration for high-dimensional optimization algorithms. Stat Comput. 2011 Jan 4;21(2):261–73.

176. Barnard A. Hunters and Herders of Southern Africa: A Comparative Ethnography of the Khoisan Peoples. Cambridge University Press; 1992. 384 p.

177. Bleek DF. The Naron: A Bushman Tribe of the Central Kalahari. CUP Archive; 1928. 84 p.

178. Dornan SS. Pygmies and Bushmen of the Kalahari: An Account of the Hunting Tribes Inhabiting the Great Arid Plateau of the Kalahari Desert ... Seeley, Service & Company; 1925. 318 p.

179. Schapera I. The Khoisan Peoples of South Africa. Routledge & Kegan Paul; 1934. 494 p.

180. Jaccard P. Nouvelles Recherches Sur La Distribution Florale. Vol. 44. Bulletin de la Société vaudoise des Sciences Naturelles; 1908. 223-270 p.

181. Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci U S A. 2005 Nov 1;102(44):15942–7.

182. Nurse GT, Jenkins T. Health and the Hunter-Gatherer. Biomedical studies on the hunting and gathering populations of Southern Africa. Monogr Hum Genet. 1977;8:1–126.

183. Barnard A. Hunters and Herders of Southern Africa: A Comparative Ethnography of the Khoisan Peoples. Cambridge University Press; 1992. 384 p.

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

185. Gravel S. Population Genetics Models of Local Ancestry. Genetics. 2012 Jun 1;191(2):607–19.

186. Theal GM. History of the Boers in South Africa, or, The wanderings and wars of the emigrant farmers [microform] : from their leaving the Cape colony to the acknowledgement of their independence by Great Britain [Internet]. London : S. Sonnenschein, Lowrey; 1887 [cited 2015 Sep 22]. 443 p. Available from: http://archive.org/details/cihm_33995

187. Plessis IDD. The Cape Malays. South African Institute of Race Relations; 1947. 152 p.

188. Delfin F, Myles S, Choi Y, Hughes D, Illek R, Oven M van, et al. Bridging Near and Remote Oceania: mtDNA and NRY Variation in the Solomon Islands. Mol Biol Evol. 2012 Feb 1;29(2):545–64.

189. Kayser M. The Human Genetic History of Oceania: Near and Remote Views of Dispersal. Curr Biol. 2010 Feb 23;20(4):R194–201.

190. Poetsch M, Wiegand A, Harder M, Blöhm R, Rakotomavo N, Freitag-Wolf S, et al. Determination of population origin: a comparison of autosomal SNPs, Y-chromosomal and mtDNA haplogroups using a Malagasy population as example. Eur J Hum Genet. 2013 Dec;21(12):1423–8.

191. Martin AR, Gignoux CR, Lin M, Granka JM, Adams A, Liu X, et al. A Complex, Polygenic Architecture for Lightened Skin Pigmentation in the Southern African KhoeSan. In 2017 [cited 2017 Aug 15]. Available from: http://meeting.physanth.org/program/2017/session33/martin-2017-a-complex-polygenic-architecture-for-lightened-skin-pigmentation-in-the-southern-african-khoesan.html

192. Lin M, Granka JM, Martin AR, Myrick J, Atkinson EG, Werely CJ, et al. High Heritability and Ancestry Dominance are behind the Genetics of Short Stature in South African KhoeSan Populations. In 2017 [cited 2017 Aug 15]. Available from: http://meeting.physanth.org/program/2017/session25/lin-2017-high-heritability-and-ancestry-dominance-are-behind-the-genetics-of-short-stature-in-south-african-khoesan-populations.html

193. Burrough SL. Late Quaternary Environmental Change and Human Occupation of the Southern African Interior. In: Jones SC, Stewart BA, editors. Africa from MIS 6-2 [Internet]. Springer Netherlands; 2016 [cited 2016 May 16]. p. 161–74. (Vertebrate Paleobiology and Paleoanthropology). Available from: http://link.springer.com/chapter/10.1007/978-94-017-7520-5_9

194. Robbins LH, Brook GA, Murphy ML, Ivester AH, Campbell AC. The Kalahari During MIS 6-2 (190–12 ka): Archaeology, Paleoenvironment, and Population Dynamics. In: Jones SC, Stewart BA, editors. Africa from MIS 6-2 [Internet]. Springer Netherlands; 2016 [cited 2016 May 16]. p. 175–93. (Vertebrate Paleobiology and Paleoanthropology). Available from: http://link.springer.com/chapter/10.1007/978-94-017-7520-5_10

195. Schlebusch CM, de Jongh M, Soodyall H. Different contributions of ancient mitochondrial and Y-chromosomal lineages in “Karretjie people” of the Great Karoo in South Africa. J Hum Genet. 2011 Sep;56(9):623–30.

196. Fort J. Synthesis between demic and cultural diffusion in the Neolithic transition in Europe. Proc Natl Acad Sci. 2012 Nov 13;109(46):18669–73.

197. Malmström H, Linderholm A, Skoglund P, Storå J, Sjödin P, Gilbert MTP, et al. Ancient mitochondrial DNA from the northern fringe of the Neolithic farming expansion in Europe sheds light on the dispersion process. Philos Trans R Soc Lond B Biol Sci. 2015 Jan 19;370(1660):20130373.

198. Skoglund P, Malmström H, Omrak A, Raghavan M, Valdiosera C, Günther T, et al. Genomic Diversity and Admixture Differs for Stone-Age Scandinavian Foragers and Farmers. Science. 2014 May 16;344(6185):747–50.

199. Gignoux CR, Henn BM, Mountain JL. Rapid, global demographic expansions after the origins of agriculture. Proc Natl Acad Sci. 2011 Apr 12;108(15):6044–9.

200. Sikora M, Carpenter ML, Moreno-Estrada A, Henn BM, Underhill PA, Sánchez-Quinto F, et al. Population genomic analysis of ancient and modern genomes yields new insights into the genetic ancestry of the Tyrolean Iceman and the genetic structure of Europe. PLoS Genet. 2014 May;10(5):e1004353.

201. Gonder MK, Mortensen HM, Reed FA, de Sousa A, Tishkoff SA. Whole-mtDNA genome sequence analysis of ancient African lineages. Mol Biol Evol. 2007 Mar;24(3):757–68.

202. Salas A, Richards M, De la Fe T, Lareu M-V, Sobrino B, Sánchez-Diz P, et al. The making of the African mtDNA landscape. Am J Hum Genet. 2002 Nov;71(5):1082–111.

203. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012 Jan;8(1):e1002453.

204. Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, et al. Complete Khoisan and Bantu genomes from southern Africa. Nature. 2010 Feb 18;463(7283):943–7.

205. Kloss-Brandstätter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, et al. HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. Hum Mutat. 2011 Jan 1;32(1):25–32.

206. Davids A. The Afrikaans of the Cape Muslims from 1815 to 1915. Protea Book House; 2011. 318 p.

207. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005 Apr 15;308(5720):385–9.

208. Wei Z, Liu Y, Xu H, Tang K, Wu H, Lu L, et al. Genome-Wide Association Studies of HIV-1 Host Control in Ethnically Diverse Chinese Populations. Sci Rep. 2015;5:10879.

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

210. DeLorenze GN, Nelson CL, Scott WK, Allen AS, Ray GT, Tsai A-L, et al. Polymorphisms in HLA Class II Genes Are Associated With Susceptibility to Staphylococcus aureus Infection in a White Population. J Infect Dis. 2016 Mar 1;213(5):816–23.

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

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

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

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

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

216. Sveinbjornsson G, Gudbjartsson DF, Halldorsson BV, Kristinsson KG, Gottfredsson M, Barrett JC, et al. HLA class II sequence variants influence tuberculosis risk in populations of European ancestry. Nat Genet. 2016 Mar;48(3):318–22.

217. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012 Sep;22(9):1790–7.

218. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.

219. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: Illuminating the Dark Road from Association to Function. Am J Hum Genet. 2013 Nov 7;93(5):779–97.

220. Rosenthal SL, Barmada MM, Wang X, Demirci FY, Kamboh MI. Connecting the dots: potential of data integration to identify regulatory SNPs in late-onset Alzheimer's disease GWAS findings. PloS One. 2014;9(4):e95152.

221. Cavalli M, Pan G, Nord H, Wadelius C. Looking beyond GWAS: allele-specific transcription factor binding drives the association of GALNT2 to HDL-C plasma levels. Lipids Health Dis. 2016;15:18.

222. Bastami M, Nariman-Saleh-Fam Z, Saadatian Z, Nariman-Saleh-Fam L, Omrani MD, Ghaderian SMH, et al. The miRNA Targetome of Coronary Artery Disease is Perturbed by Functional Polymorphisms Identified and Prioritized by in-depth Bioinformatics Analyses Exploiting Genome-wide Association Studies. Gene. 2016 Sep 2;

223. Haider SA, Faisal M. Human aging in the post-GWAS era: further insights reveal potential regulatory variants. Biogerontology. 2015 Apr 17;16(4):529–41.

224. Munch Z, Van Lill SWP, Booysen CN, Zietsman HL, Enarson DA, Beyers N. Tuberculosis transmission patterns in a high-incidence area: a spatial analysis. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2003 Mar;7(3):271–7.

225. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, Bakker PIW de. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008 Dec 15;24(24):2938–9.

226. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 2006 Feb;38(2):209–13.

227. den Boon S., Van Lill SW, Borgdorff MW, Enarson DA, Verver S, Bateman ED, et al. High prevalence of tuberculosis in previously treated patients, Cape Town, South Africa. EmergInfectDis. 2007 Aug;13(8):1189–94.

228. Warnes G, Gorjanc with contributions from G, Leisch F, Man and M. genetics: Population Genetics [Internet]. 2013. Available from: https://cran.r-project.org/web/packages/genetics/index.html

229. Perneger TV. What's wrong with Bonferroni adjustments. BMJ. 1998 Apr 18;316(7139):1236–8.

230. Campbell H, Rudan I. Interpretation of genetic association studies in complex disease. PharmacogenomicsJ. 2002;2(6):349–60.

231. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. AmJHumGenet. 2004 Apr;74(4):765–9.

232. Šidák Z. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. J Am Stat Assoc. 1967 Jun 1;62(318):626–33.

233. Pollard K, Dudoit S, Laan M van der. Multiple Testing Procedures: R multtest Package and Applications to Genomics. UC Berkeley Div Biostat Work Pap Ser [Internet]. 2004 Dec 9; Available from: http://biostats.bepress.com/ucbbiostat/paper164

234. Ma M-J, Wang H-B, Li H, Yang J-H, Yan Y, Xie L-P, et al. Genetic variants in MARCO are associated with the susceptibility to pulmonary tuberculosis in Chinese Han population. PloS One. 2011;6(8):e24069.

235. Bowdish DM, Sakamoto K, Lack NA, Hill PC, Sirugo G, Newport MJ, et al. Genetic variants of MARCO are associated with susceptibility to pulmonary tuberculosis in a Gambian population. BMC Med Genet. 2013;14:47.

236. Jing J, Yang IV, Hui L, Patel JA, Evans CM, Prikeris R, et al. Role of macrophage receptor with collagenous structure in innate immune tolerance. J Immunol Baltim Md 1950. 2013 Jun 15;190(12):6360–7.

237. Songane M, Kleinnijenhuis J, Alisjahbana B, Sahiratmadja E, Parwati I, Oosting M, et al. Polymorphisms in Autophagy Genes and Susceptibility to Tuberculosis. PLOS ONE. 2012 Aug 6;7(8):e41618.

238. Ma M-J, Wang H-B, Li H, Yang J-H, Yan Y, Xie L-P, et al. Genetic Variants in MARCO Are Associated with the Susceptibility to Pulmonary Tuberculosis in Chinese Han Population. PLOS ONE. 2011 Aug 23;6(8):e24069.

239. Ma X, Liu Y, Gowen BB. Full-exon resequencing reveals toll-like receptor variants contribute to human susceptibility to tuberculosis disease. PloS One. 2007;2:1318.

240. Bulat-Kardum LJ, Etokebe GE, Lederer P, Balen S, Dembic Z. Genetic Polymorphisms in the Toll-like Receptor 10, Interleukin (IL)17A and IL17F Genes Differently Affect the Risk for Tuberculosis in Croatian Population. Scand J Immunol. 2015 Jul;82(1):63–9.

241. Uciechowski P, Imhoff H, Lange C. Susceptibility to tuberculosis is associated with TLR1 polymorphisms resulting in a lack of TLR1 cell surface expression. J Leukoc Biol. 2011;90:377–388.

242. Brzostek A, Pawelczyk J, Rumijowska-Galewicz A, Dziadek B, Dziadek J. Mycobacterium tuberculosis Is Able To Accumulate and Utilize Cholesterol. J Bacteriol. 2009 Nov 1;191(21):6584–91.

243. Kuijl C, Pilli M, Alahari SK, Janssen H, Khoo P-S, Ervin KE, et al. Rac and Rab GTPases dual effector Nischarin regulates vesicle maturation to facilitate survival of intracellular bacteria. EMBO J. 2013 Mar 6;32(5):713–27.

244. Cooke GS, Campbell SJ, Sillah J, Gustafson P, Bah B, Sirugo G, et al. Polymorphism within the Interferon-γ/Receptor Complex Is Associated with Pulmonary Tuberculosis. Am J Respir Crit Care Med. 2006 Aug 1;174(3):339–43.

245. Al-Muhsen S, Casanova J-L. The genetic heterogeneity of mendelian susceptibility to mycobacterial diseases. J Allergy Clin Immunol. 2008 Dec;122(6):1043-1051-1053.

246. Hijikata M, Shojima J, Matsushita I, Tokunaga K, Ohashi J, Hang NTL, et al. Association of IFNGR2 gene polymorphisms with pulmonary tuberculosis among the Vietnamese. Hum Genet. 2012 May;131(5):675–82.

247. Kong X-F, Vogt G, Itan Y, Macura-Biegun A, Szaflarska A, Kowalczyk D, et al. Haploinsufficiency at the human IFNGR2 locus contributes to mycobacterial disease. Hum Mol Genet. 2013 Feb 15;22(4):769–81.

248. Bush WS, Moore JH. Chapter 11: Genome-Wide Association Studies. PLoS Comput Biol [Internet]. 2012 Dec [cited 2016 Sep 20];8(12). Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531285/

249. Rieder HL, Disease IU against T and L. Epidemiologic basis of tuberculosis control. International Union Against Tuberculosis and Lung Disease; 1999. 180 p.

250. Bellamy R, Beyers N, McAdam KP, Ruwende C, Gie R, Samaai P, et al. Genetic susceptibility to tuberculosis in Africans: a genome-wide scan. Proc Natl Acad Sci U S A. 2000 Jul 5;97(14):8005–9.

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

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

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

254. Baker AR, Zalwango S, Malone LL, Igo RP, Qiu F, Nsereko M, et al. Genetic Susceptibility to Tuberculosis Associated with Cathepsin Z Haplotype in a Ugandan Household Contact Study. Hum Immunol. 2011 May;72(5):426–30.

255. Wang D, Zhou Y, Ji L, He T, Lin F, Lin R, et al. Association of LMP/TAP gene polymorphisms with tuberculosis susceptibility in Li population in China. PloS One. 2012;7(3):e33051.

256. Herb F, Thye T, Niemann S, Browne ENL, Chinbuah MA, Gyapong J, et al. ALOX5 variants associated with susceptibility to human pulmonary tuberculosis. Hum Mol Genet. 2008 Apr 1;17(7):1052–60.

257. Horne DJ, Randhawa AK, Chau TTH, Bang ND, Yen NTB, Farrar JJ, et al. Common polymorphisms in the PKP3-SIGIRR-TMEM16J gene region are associated with susceptibility to tuberculosis. J Infect Dis. 2012 Feb 15;205(4):586–94.

258. Lian Y, Yue J, Han M, Liu J, Liu L. Analysis of the association between BTNL2 polymorphism and tuberculosis in Chinese Han population. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2010 May;10(4):517–21.

259. Thuong NTT, Dunstan SJ, Chau TTH, Thorsson V, Simmons CP, Quyen NTH, et al. Identification of Tuberculosis Susceptibility Genes with Human Macrophage Gene Expression Profiles. PLOS Pathog. 2008 Dec 5;4(12):e1000229.

260. Flores-Villanueva PO, Ruiz-Morales JA, Song C-H, Flores LM, Jo E-K, Montaño M, et al. A functional promoter polymorphism in monocyte chemoattractant protein-1 is associated with increased susceptibility to pulmonary tuberculosis. J Exp Med. 2005 Dec 19;202(12):1649–58.

261. Buijtels PC a. M, van de Sande WWJ, Parkinson S, Petit PLC, van der Sande M a. B, van Soolingen D, et al. Polymorphism in CC-chemokine ligand 2 associated with tuberculosis in Zambia. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2008 Dec;12(12):1485–8.

262. Yang L, Shi GL, Song CX, Xu SF. Relationship between genetic polymorphism of MCP-1 and non-small-cell lung cancer in the Han nationality of North China. Genet Mol Res GMR. 2010;9(2):765–71.

263. Ganachari M, Ruiz-Morales JA, Pretell JCG de la T, Dinh J, Granados J, Flores-Villanueva PO. Joint Effect of MCP-1 Genotype GG and MMP-1 Genotype 2G/2G Increases the Likelihood of Developing Pulmonary Tuberculosis in BCG-Vaccinated Individuals. PLOS ONE. 2010 Jan 25;5(1):e8881.

264. Feng W-X, Mokrousov I, Wang B-B, Nelson H, Jiao W-W, Wang J, et al. Tag SNP Polymorphism of CCL2 and Its Role in Clinical Tuberculosis in Han Chinese Pediatric Population. PLOS ONE. 2011 Feb 4;6(2):e14652.

265. Ben-Selma W, Ben-Abderrahmen Y, Boukadida J, Harizi H. IL-10R1 S138G loss-of-function polymorphism is associated with extrapulmonary tuberculosis risk development in Tunisia. Mol Biol Rep. 2012 Jan;39(1):51–6.

266. Chu S-F, Tam CM, Wong HS, Kam KM, Lau YL, Chiang AKS. Association between RANTES functional polymorphisms and tuberculosis in Hong Kong Chinese. Genes Immun. 2007 Sep;8(6):475–9.

267. Sanchez CMF, Baquero IC, Fau BI, Sanchez V, P. S, Fau VP, et al. Polymorphisms in CCL5 promoter are associated with pulmonary tuberculosis in. Int J Tuberc Lung Dis. 2009;13:480–485.

268. Sharma PR, Singh S, Jena M, Mishra G, Prakash R, Das PK, et al. Coding and non-coding polymorphisms in VDR gene and susceptibility to pulmonary tuberculosis in tribes, castes and Muslims of Central India. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2011 Aug;11(6):1456–61.

269. Ben-Selma W, Harizi H, Bougmiza I, Ben Kahla I, Letaief M, Boukadida J. Polymorphisms in the RANTES gene increase susceptibility to active tuberculosis in Tunisia. DNA Cell Biol. 2011 Oct;30(10):789–800.

270. Selvaraj P, Alagarasu K, Swaminathan S, Harishankar M, Narendran G. CD209 gene polymorphisms in South Indian HIV and HIV-TB patients. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2009 Mar;9(2):256–62.

271. Kobayashi K, Yuliwulandari R, Yanai H, Lien LT, Hang NTL, Hijikata M, et al. Association of CD209 polymorphisms with tuberculosis in an Indonesian population. Hum Immunol. 2011 Sep;72(9):741–5.

272. Campo M, Randhawa AK, Dunstan S, Farrar J, Caws M, Bang ND, et al. Common polymorphisms in the CD43 gene region are associated with tuberculosis disease and mortality. Am J Respir Cell Mol Biol. 2015 Mar;52(3):342–8.

273. Lee P, Waalen J, Crain K, Smargon A, Beutler E. Human chitotriosidase polymorphisms G354R and A442V associated with reduced enzyme activity. Blood Cells Mol Dis. 2007 Dec;39(3):353–60.

274. Tso HW, Ip WK, Chong WP, Tam CM, Chiang AKS, Lau YL. Association of interferon gamma and interleukin 10 genes with tuberculosis in Hong Kong Chinese. Genes Immun. 2005 Jun;6(4):358–63.

275. Baker AR, Zalwango S, Malone LL, Igo RP, Qiu F, Nsereko M, et al. Genetic susceptibility to tuberculosis associated with cathepsin Z haplotype in a Ugandan household contact study. Hum Immunol. 2011 May;72(5):426–30.

276. Selvaraj P, Alagarasu K, Harishankar M, Vidyarani M, Narayanan PR. Regulatory region polymorphisms of vitamin D receptor gene in pulmonary tuberculosis patients and normal healthy subjects of south India. Int J Immunogenet. 2008 Jun;35(3):251–4.

277. Selvaraj P, Alagarasu K, Singh B. Stromal cell-derived factor-1 (SDF-1/CXCL12) gene polymorphisms in pulmonary tuberculosis patients of south India. Int J Immunogenet. 2012 Feb;39(1):26–31.

278. Barreiro LB, Tailleux L, Pai AA, Gicquel B, Marioni JC, Gilad Y. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. Proc Natl Acad Sci U S A. 2012 Jan 24;109(4):1204–9.

279. Ravikumar M, Dheenadhayalan V, Rajaram K, Lakshmi SS, Kumaran PP, Paramasivan CN, et al. Associations of HLA-DRB1, DQB1 and DPB1 alleles with pulmonary tuberculosis in south India. Tuber Lung Dis Off J Int Union Tuberc Lung Dis. 1999;79(5):309–17.

280. Terán-Escandón D, Terán-Ortiz L, Camarena-Olvera A, González-Avila G, Vaca-Marín MA, Granados J, et al. Human leukocyte antigen-associated susceptibility to pulmonary tuberculosis: molecular analysis of class II alleles by DNA amplification and oligonucleotide hybridization in Mexican patients. Chest. 1999 Feb;115(2):428–33.

281. Dubaniewicz A, Moszkowska G, Szczerkowska Z. Frequency of DRB1,DQB1 two,locus haplotypes in tuberculosis: preliminary report. Tuberculosis. 2005;85:259–267.

282. Kim HS, Park MH, Song EY, Park H, Kwon SY, Han SK, et al. Association of HLA-DR and HLA-DQ genes with susceptibility to pulmonary tuberculosis in Koreans: preliminary evidence of associations with drug resistance, disease severity, and disease recurrence. Hum Immunol. 2005 Oct;66(10):1074–81.

283. Delgado JC, Baena A, Thim S, Goldfeld AE. Aspartic acid homozygosity at codon 57 of HLA-DQ beta is associated with susceptibility to pulmonary tuberculosis in Cambodia. J Immunol Baltim Md 1950. 2006 Jan 15;176(2):1090–7.

284. Figueiredo JF de C, Rodrigues M de LV, Deghaide NHS, Donadi EA. HLA profile in patients with AIDS and tuberculosis. Braz J Infect Dis Off Publ Braz Soc Infect Dis. 2008 Aug;12(4):278–80.

285. López-Maderuelo D, Arnalich F, Serantes R, González A, Codoceo R, Madero R, et al. Interferon-gamma and interleukin-10 gene polymorphisms in pulmonary tuberculosis. Am J Respir Crit Care Med. 2003 Apr 1;167(7):970–5.

286. Henao MI, Montes C, París SC, García LF. Cytokine gene polymorphisms in Colombian patients with different clinical presentations of tuberculosis. Tuberc Edinb Scotl. 2006 Jan;86(1):11–9.

287. Sallakci N, Coskun M, Berber Z, Gürkan F, Kocamaz H, Uysal G, et al. Interferon-gamma gene+874T-A polymorphism is associated with tuberculosis and gamma interferon response. Tuberc Edinb Scotl. 2007 May;87(3):225–30.

288. Amim LHLV, Pacheco AG, Fonseca-Costa J, Loredo CS, Rabahi MF, Melo MH, et al. Role of IFN-gamma +874 T/A single nucleotide polymorphism in the tuberculosis outcome among Brazilians subjects. Mol Biol Rep. 2008 Dec;35(4):563–6.

289. Ansari A, Talat N, Jamil B, Hasan Z, Razzaki T, Dawood G, et al. Cytokine gene polymorphisms across tuberculosis clinical spectrum in Pakistani patients. PloS One. 2009;4(3):e4778.

290. Stein CM, Zalwango S, Chiunda AB, Millard C, Leontiev DV, Horvath AL, et al. Linkage and association analysis of candidate genes for TB and TNFalpha cytokine expression: evidence for association with IFNGR1, IL-10, and TNF receptor 1 genes. Hum Genet. 2007 Jul;121(6):663–73.

291. Velez DR, Hulme WF, Myers JL, Weinberg JB, Levesque MC, Stryjewski ME, et al. NOS2A, TLR4, and IFNGR1 interactions influence pulmonary tuberculosis susceptibility in African-Americans. Hum Genet. 2009 Nov;126(5):643–53.

292. He J, Wang J, Lei D, Ding S. Analysis of functional SNP in ifng/ifngr1 in Chinese Han population with tuberculosis. Scand J Immunol. 2010 Jun;71(6):452–8.

293. Hijikata M, Shojima J, Matsushita I, Tokunaga K, Ohashi J, Hang NTL, et al. Association of IFNGR2 gene polymorphisms with pulmonary tuberculosis among the Vietnamese. Hum Genet. 2012 May;131(5):675–82.

294. Guwatudde D, Zalwango S, Kamya MR. Burden of tuberculosis in Kampala, Uganda. Bull World Health Organ. 2003;81:799–805.

295. Oh J-H, Yang C-S, Noh Y-K, Kweon Y-M, Jung S-S, Son JW, et al. Polymorphisms of interleukin-10 and tumour necrosis factor-alpha genes are associated with newly diagnosed and recurrent pulmonary tuberculosis. Respirol Carlton Vic. 2007 Jul;12(4):594–8.

296. Ates O, Musellim B, Ongen G, Topal-Sarikaya A. Interleukin-10 and tumor necrosis factor-alpha gene polymorphisms in tuberculosis. J Clin Immunol. 2008 May;28(3):232–6.

297. Taype CA, Shamsuzzaman S, Accinelli RA, Espinoza JR, Shaw M-A. Genetic susceptibility to different clinical forms of tuberculosis in the Peruvian population. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2010 May;10(4):495–504.

298. Morris GAJ, Edwards DRV, Hill PC, Wejse C, Bisseye C, Olesen R, et al. Interleukin 12B (IL12B) genetic variation and pulmonary tuberculosis: a study of cohorts from The Gambia, Guinea-Bissau, United States and Argentina. PloS One. 2011;6(2):e16656.

299. Akahoshi M, Nakashima H, Miyake K, Inoue Y, Shimizu S, Tanaka Y, et al. Influence of interleukin-12 receptor beta1 polymorphisms on tuberculosis. Hum Genet. 2003 Mar;112(3):237–43.

300. Remus N, El Baghdadi J, Fieschi C, Feinberg J, Quintin T, Chentoufi M, et al. Association of IL12RB1 polymorphisms with pulmonary tuberculosis in adults in Morocco. J Infect Dis. 2004 Aug 1;190(3):580–7.

301. Kusuhara K, Yamamoto K, Okada K, Mizuno Y, Hara T. Association of IL12RB1 polymorphisms with susceptibility to and severity of tuberculosis in Japanese: a gene-based association analysis of 21 candidate genes. Int J Immunogenet. 2007 Feb;34(1):35–44.

302. Verma VK, Taneja V, Jaiswal A, Sharma S, Behera D, Sreenivas V, et al. Prevalence, distribution and functional significance of the -237C to T polymorphism in the IL-12Rβ2 promoter in Indian tuberculosis patients. PloS One. 2012;7(4):e34355.

303. Han M, Yue J, Lian Y-Y, Zhao Y-L, Wang H-X, Liu L-R. Relationship between single nucleotide polymorphism of interleukin-18 and susceptibility to pulmonary tuberculosis in the Chinese Han population. Microbiol Immunol. 2011 Jun;55(6):388–93.

304. Zhang J, Zheng L, Zhu D, An H, Yang Y, Liang Y, et al. Polymorphisms in the Interleukin 18 Receptor 1 Gene and Tuberculosis Susceptibility among Chinese. PLOS ONE. 2014 Oct 31;9(10):e110734.

305. Amirzargar AA, Rezaei N, Jabbari H, Danesh A-A, Khosravi F, Hajabdolbaghi M, et al. Cytokine single nucleotide polymorphisms in Iranian patients with pulmonary tuberculosis. Eur Cytokine Netw. 2006 Jun;17(2):84–9.

306. Naslednikova IO, Urazova OI, Voronkova OV, Strelis AK, Novitsky VV, Nikulina EL, et al. Allelic polymorphism of cytokine genes during pulmonary tuberculosis. Bull Exp Biol Med. 2009 Aug;148(2):175–80.

307. Jiang D, Hu X, Li S, Julaiti A, Xia Y, Wang J, et al. [Polymorphisms of IL-23 receptor gene are associated with susceptibility to pulmonary tuberculosis and drug-resistant pulmonary tuberculosis]. Zhonghua Yi Xue Za Zhi. 2015 May 26;95(20):1576–80.

308. Jiang D, Wubuli A, Hu X, Ikramullah S, Maimaiti A, Zhang W, et al. The variations of IL-23R are associated with susceptibility and severe clinical forms of pulmonary tuberculosis in Chinese Uygurs. BMC Infect Dis [Internet]. 2015 Dec 1 [cited 2016 Aug 12];15. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4665827/

309. Vidyarani M, Selvaraj P, Prabhu Anand S, Jawahar MS, Adhilakshmi AR, Narayanan PR. Interferon gamma (IFNgamma) & interleukin-4 (IL-4) gene variants & cytokine levels in pulmonary tuberculosis. Indian J Med Res. 2006 Oct;124(4):403–10.

310. Zhang G, Zhou B, Wang W, Zhang M, Zhao Y, Wang Z, et al. A functional single-nucleotide polymorphism in the promoter of the gene encoding interleukin 6 is associated with susceptibility to tuberculosis. J Infect Dis. 2012 Jun;205(11):1697–704.

311. Ding S, Jiang T, He J, Qin B, Lin S, Li L. Tagging single nucleotide polymorphisms in the IRF1 and IRF8 genes and tuberculosis susceptibility. PloS One. 2012;7(8):e42104.

312. King KY, Lew JD, Ha NP, Lin JS, Ma X, Graviss EA, et al. Polymorphic allele of human IRGM1 is associated with susceptibility to tuberculosis in African Americans. PloS One. 2011;6(1):e16317.

313. Bahari G, Hashemi M, Taheri M, Naderi M, Eskandari-Nasab E, Atabaki M. Association of IRGM polymorphisms and susceptibility to pulmonary tuberculosis in Zahedan, Southeast Iran. ScientificWorldJournal. 2012;2012:950801.

314. Taype CA, Shamsuzzaman S, Accinelli RA, Espinoza JR, Shaw M-A. Genetic susceptibility to different clinical forms of tuberculosis in the Peruvian population. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2010 May;10(4):495–504.

315. Curtis J, Kopanitsa L, Stebbings E, Speirs A, Ignatyeva O, Balabanova Y, et al. Association analysis of the LTA4H gene polymorphisms and pulmonary tuberculosis in 9115 subjects. Tuberc Edinb Scotl. 2011 Jan;91(1):22–5.

316. Yang J, Chen J, Yue J, Liu L, Han M, Wang H. Relationship between human LTA4H polymorphisms and extra-pulmonary tuberculosis in an ethnic Han Chinese population in Eastern China. Tuberc Edinb Scotl. 2014 Dec;94(6):657–63.

317. Alagarasu K, Selvaraj P, Swaminathan S, Raghavan S, Narendran G, Narayanan PR. Mannose binding lectin gene variants and susceptibility to tuberculosis in HIV-1 infected patients of South India. Tuberc Edinb Scotl. 2007 Nov;87(6):535–43.

318. Capparelli R, Iannaccone M, Palumbo D, Medaglia C, Moscariello E, Russo A, et al. Role played by human mannose-binding lectin polymorphisms in pulmonary tuberculosis. J Infect Dis. 2009 Mar 1;199(5):666–72.

319. Gómez LM, Sánchez E, Ruiz-Narvaez EA, López-Nevot MA, Anaya J-M, Martín J. Macrophage migration inhibitory factor gene influences the risk of developing tuberculosis in northwestern Colombian population. Tissue Antigens. 2007 Jul;70(1):28–33.

320. Sadki K, Lamsyah H, Rueda B, Akil E, Sadak A, Martin J, et al. Analysis of MIF, FCGR2A and FCGR3A gene polymorphisms with susceptibility to pulmonary tuberculosis in Moroccan population. J Genet Genomics Yi Chuan Xue Bao. 2010 Apr;37(4):257–64.

321. Zhang X, Jiang F, Wei L, Li F, Liu J, Wang C, et al. Polymorphic allele of human MRC1 confer protection against tuberculosis in a Chinese population. Int J Biol Sci. 2012;8(3):375–82.

322. Austin CM, Ma X, Graviss EA. Common nonsynonymous polymorphisms in the NOD2 gene are associated with resistance or susceptibility to tuberculosis disease in African Americans. J Infect Dis. 2008 Jun 15;197(12):1713–6.

323. Pan H, Dai Y, Tang S, Wang J. Polymorphisms of NOD2 and the risk of tuberculosis: a validation study in the Chinese population. Int J Immunogenet. 2012 Jun;39(3):233–40.

324. Möller M, Nebel A, Valentonyte R, van Helden PD, Schreiber S, Hoal EG. Investigation of chromosome 17 candidate genes in susceptibility to TB in a South African population. Tuberc Edinb Scotl. 2009 Mar;89(2):189–94.

325. Velez DR, Hulme WF, Myers JL, Weinberg JB, Levesque MC, Stryjewski ME, et al. NOS2A, TLR4, and IFNGR1 interactions influence pulmonary tuberculosis susceptibility in African-Americans. Hum Genet. 2009 Nov;126(5):643–53.

326. Fernando SL, Saunders BM, Sluyter R, Skarratt KK, Goldberg H, Marks GB, et al. A polymorphism in the P2X7 gene increases susceptibility to extrapulmonary tuberculosis. Am J Respir Crit Care Med. 2007 Feb 15;175(4):360–6.

327. Niño-Moreno P, Portales-Pérez D, Hernández-Castro B, Portales-Cervantes L, Flores-Meraz V, Baranda L, et al. P2X7 and NRAMP1/SLC11 A1 gene polymorphisms in Mexican mestizo patients with pulmonary tuberculosis. Clin Exp Immunol. 2007 Jun;148(3):469–77.

328. Sambasivan V, Murthy KJR, Reddy R, Vijayalakshimi V, Hasan Q. P2X7 gene polymorphisms and risk assessment for pulmonary tuberculosis in Asian Indians. Dis Markers. 2010;28(1):43–8.

329. Sharma S, Kumar V, Khosla R, Kajal N, Sarin B, Sehajpal P. Association of P2X7 receptor +1513 (A-->C) polymorphism with tuberculosis in a Punjabi population. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2010 Sep;14(9):1159–63.

330. Ben-Selma W, Ben-Kahla I, Boukadida J, Harizi H. Contribution of the P2X7 1513A/C loss-of-function polymorphism to extrapulmonary tuberculosis susceptibility in Tunisian populations. FEMS Immunol Med Microbiol. 2011 Oct;63(1):65–72.

331. Singla N, Gupta D, Joshi A, Batra N, Singh J. Genetic polymorphisms in the P2X7 gene and its association with susceptibility to tuberculosis. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2012 Feb;16(2):224–9.

332. Gomez LM, Anaya J-M, Martin J. Genetic influence of PTPN22 R620W polymorphism in tuberculosis. Hum Immunol. 2005 Dec;66(12):1242–7.

333. Kouhpayeh H-R, Hashemi M, Hashemi S-A, Moazeni-Roodi A, Naderi M, Sharifi-Mood B, et al. R620W functional polymorphism of protein tyrosine phosphatase non-receptor type 22 is not associated with pulmonary tuberculosis in Zahedan, southeast Iran. Genet Mol Res GMR. 2012;11(2):1075–81.

334. Floros J, Lin HM, García A, Salazar MA, Guo X, DiAngelo S, et al. Surfactant protein genetic marker alleles identify a subgroup of tuberculosis in a Mexican population. J Infect Dis. 2000 Nov;182(5):1473–8.

335. Madan T, Saxena S, Murthy KJR, Muralidhar K, Sarma PU. Association of polymorphisms in the collagen region of human SP-A1 and SP-A2 genes with pulmonary tuberculosis in Indian population. Clin Chem Lab Med. 2002 Oct;40(10):1002–8.

336. Malik S, Abel L, Tooker H, Poon A, Simkin L, Girard M, et al. Alleles of the NRAMP1 gene are risk factors for pediatric tuberculosis disease. Proc Natl Acad Sci U S A. 2005 Aug 23;102(34):12183–8.

337. Bellamy R, Ruwende C, Corrah T. Variations in the NRAMP1 gene and susceptibility to tuberculosis in West Africans. N Engl J Med. 1998;338:640–644.

338. Ryu S, Park YK, Bai GH, Kim SJ, Park SN, Kang S. 3'UTR polymorphisms in the NRAMP1 gene are associated with susceptibility to tuberculosis in Koreans. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2000 Jun;4(6):577–80.

339. Liu W, Cao WC, Zhang CY, Tian L, Wu XM, Habbema JDF, et al. VDR and NRAMP1 gene polymorphisms in susceptibility to pulmonary tuberculosis among the Chinese Han population: a case-control study. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2004 Apr;8(4):428–34.

340. Zhang W, Shao L, Weng X, Hu Z, Jin A, Chen S, et al. Variants of the natural resistance-associated macrophage protein 1 gene (NRAMP1) are associated with severe forms of pulmonary tuberculosis. Clin Infect Dis Off Publ Infect Dis Soc Am. 2005 May 1;40(9):1232–6.

341. Taype CA, Castro JC, Accinelli RA, Herrera-Velit P, Shaw MA, Espinoza JR. Association between SLC11A1 polymorphisms and susceptibility to different clinical forms of tuberculosis in the Peruvian population. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2006 Sep;6(5):361–7.

342. Leung KH, Yip SP, Wong WS, Yiu LS, Chan KK, Lai WM, et al. Sex- and age-dependent association of SLC11A1 polymorphisms with tuberculosis in Chinese: a case control study. BMC Infect Dis. 2007;7:19.

343. Søborg C, Andersen AB, Range N, Malenganisho W, Friis H, Magnussen P, et al. Influence of candidate susceptibility genes on tuberculosis in a high endemic region. Mol Immunol. 2007 Mar;44(9):2213–20.

344. Discovery of novel diarylketoxime derivatives as selective and orally active. Bioorg Med Chem Lett 19 5339-5345 LID. 2009;

345. Velez DR, Hulme WF, Myers JL, Stryjewski ME, Abbate E, Estevan R, et al. Association of SLC11A1 with tuberculosis and interactions with NOS2A and TLR2 in African-Americans and Caucasians. Int J Tuberc Lung Dis Off J Int Union Tuberc Lung Dis. 2009 Sep;13(9):1068–76.

346. McDermid JM, Hennig BJ, van der Sande M, Hill AVS, Whittle HC, Jaye A, et al. Host iron redistribution as a risk factor for incident tuberculosis in HIV infection: an 11-year retrospective cohort study. BMC Infect Dis. 2013;13:48.

347. Ridruechai C, Mahasirimongkol S, Phromjai J, Yanai H, Nishida N, Matsushita I, et al. Association analysis of susceptibility candidate region on chromosome 5q31 for tuberculosis. Genes Immun. 2010 Jul;11(5):416–22.

348. Baker MA, Wilson D, Wallengren K, Sandgren A, Iartchouk O, Broodie N, et al. Polymorphisms in the gene that encodes the iron transport protein ferroportin 1 influence susceptibility to tuberculosis. J Infect Dis. 2012 Apr 1;205(7):1043–7.

349. Liang L, Zhao Y, Yue J, Liu J, Han M, Wang H, et al. Association of SP110 gene polymorphisms with susceptibility to tuberculosis in a Chinese population. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2011 Jul;11(5):934–9.

350. Hall NB, Igo RP, Malone LL, Truitt B, Schnell A, Tao L, et al. Polymorphisms in TICAM2 and IL1B are associated with TB. Genes Immun. 2015 Mar;16(2):127–33.

351. Hawn TR, Dunstan SJ, Thwaites GE, Simmons CP, Thuong NT, Lan NTN, et al. A polymorphism in Toll-interleukin 1 receptor domain containing adaptor protein is associated with susceptibility to meningeal tuberculosis. J Infect Dis. 2006 Oct 15;194(8):1127–34.

352. Thuong NTT, Hawn TR, Thwaites GE, Chau TTH, Lan NTN, Quy HT, et al. A polymorphism in human TLR2 is associated with increased susceptibility to tuberculous meningitis. Genes Immun. 2007 Jul;8(5):422–8.

353. Selvaraj P, Harishankar M, Singh B, Jawahar MS, Banurekha VV. Toll-like receptor and TIRAP gene polymorphisms in pulmonary tuberculosis patients of South India. Tuberc Edinb Scotl. 2010 Sep;90(5):306–10.

354. Ma X, Liu Y, Gowen BB, Graviss EA, Clark AG, Musser JM. Full-exon resequencing reveals toll-like receptor variants contribute to human susceptibility to tuberculosis disease. PloS One. 2007;2(12):e1318.

355. Uciechowski P, Imhoff H, Lange C. Susceptibility to tuberculosis is associated with TLR1 polymorphisms resulting in a lack of TLR1 cell surface expression. J Leukoc Biol. 2011;90:377–388.

356. Ben-Ali M, Barbouche M-R, Bousnina S, Chabbou A, Dellagi K. Toll-like receptor 2 Arg677Trp polymorphism is associated with susceptibility to tuberculosis in Tunisian patients. Clin Diagn Lab Immunol. 2004 May;11(3):625–6.

357. Thuong NTT, Hawn TR, Thwaites GE, Chau TTH, Lan NTN, Quy HT, et al. A polymorphism in human TLR2 is associated with increased susceptibility to tuberculous meningitis. Genes Immun. 2007 Jul;8(5):422–8.

358. Ferwerda B, Kibiki GS, Netea MG, Dolmans WMV, van der Ven AJ. The toll-like receptor 4 Asp299Gly variant and tuberculosis susceptibility in HIV-infected patients in Tanzania. AIDS Lond Engl. 2007 Jun 19;21(10):1375–7.

359. Velez DR, Hulme WF, Myers JL. NOS2A, TLR4, and IFNGR1 interactions influence pulmonary tuberculosis susceptibility in African-Americans. Hum Genet. 2009;126:643–653.

360. Motsinger-Reif AA, Antas PRZ, Oki NO, Levy S, Holland SM, Sterling TR. Polymorphisms in IL-1beta, vitamin D receptor Fok1, and Toll-like receptor 2 are associated with extrapulmonary tuberculosis. BMC Med Genet. 2010;11:37.

361. Pulido I, Leal M, Genebat M. The TLR4 ASP299GLY polymorphism is a risk factor for active tuberculosis in Caucasian HIV-infected patients. Curr HIV Res. 2010;8:253–258.

362. Najmi N, Kaur G, Sharma SK, Mehra NK. Human Toll-like receptor 4 polymorphisms TLR4 Asp299Gly and Thr399Ile influence susceptibility and severity of pulmonary tuberculosis in the Asian Indian population. Tissue Antigens. 2010 Aug;76(2):102–9.

363. Velez DR, Wejse C, Stryjewski ME, Abbate E, Hulme WF, Myers JL, et al. Variants in toll-like receptors 2 and 9 influence susceptibility to pulmonary tuberculosis in Caucasians, African-Americans, and West Africans. Hum Genet. 2010 Jan;127(1):65–73.

364. Merza M, Farnia P, Anoosheh S, Varahram M, Kazampour M, Pajand O, et al. The NRAMPI, VDR and TNF-alpha gene polymorphisms in Iranian tuberculosis patients: the study on host susceptibility. Braz J Infect Dis Off Publ Braz Soc Infect Dis. 2009 Aug;13(4):252–6.

365. Fan H-M, Wang Z, Feng F-M, Zhang K-L, Yuan J-X, Sui H, et al. Association of TNF-alpha-238G/A and 308 G/A gene polymorphisms with pulmonary tuberculosis among patients with coal worker's pneumoconiosis. Biomed Environ Sci BES. 2010 Apr;23(2):137–45.

366. Ma M, Xie L, Wu S, Tang F, Li H, Zhang Z, et al. Toll-like receptors, tumor necrosis factor-α, and interleukin-10 gene polymorphisms in risk of pulmonary tuberculosis and disease severity. Hum Immunol. 2010 Oct;71(10):1005–10.

367. Ben-Selma W, Harizi H, Boukadida J. Association of TNF-α and IL-10 polymorphisms with tuberculosis in Tunisian populations. Microbes Infect Inst Pasteur. 2011 Sep;13(10):837–43.

368. Stein CM, Zalwango S, Chiunda AB, Millard C, Leontiev DV, Horvath AL, et al. Linkage and association analysis of candidate genes for TB and TNFalpha cytokine expression: evidence for association with IFNGR1, IL-10, and TNF receptor 1 genes. Hum Genet. 2007 Jul;121(6):663–73.

369. Mokrousov I, Wu X-R, Vyazovaya A, Feng W-X, Sun L, Xiao J, et al. Polymorphism of 3'UTR region of TNFR2 coding gene and its role in clinical tuberculosis in Han Chinese pediatric population. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2011 Aug;11(6):1312–8.

370. Shah JA, Vary JC, Chau TTH, Bang ND, Yen NTB, Farrar JJ, et al. Human TOLLIP regulates TLR2 and TLR4 signaling and its polymorphisms are associated with susceptibility to tuberculosis. J Immunol Baltim Md 1950. 2012 Aug 15;189(4):1737–46.

371. Bornman L, Campbell SJ, Fielding K, Bah B, Sillah J, Gustafson P, et al. Vitamin D receptor polymorphisms and susceptibility to tuberculosis in West Africa: a case-control and family study. J Infect Dis. 2004 Nov 1;190(9):1631–41.

372. Olesen R, Wejse C, Velez DR, Bisseye C, Sodemann M, Aaby P, et al. DC-SIGN (CD209), pentraxin 3 and vitamin D receptor gene variants associate with pulmonary tuberculosis risk in West Africans. Genes Immun. 2007 Sep;8(6):456–67.

373. Regulatory region polymorphisms of vitamin D receptor gene in pulmonary. 28 Selvaraj P Fau. 2008;35:251–254.

374. Alagarasu K, Selvaraj P, Swaminathan S, Narendran G, Narayanan PR. 5' regulatory and 3' untranslated region polymorphisms of vitamin D receptor gene in south Indian HIV and HIV-TB patients. J Clin Immunol. 2009 Mar;29(2):196–204.

375. Banoei MM, Mirsaeidi MS, Houshmand M, Tabarsi P, Ebrahimi G, Zargari L, et al. Vitamin D receptor homozygote mutant tt and bb are associated with susceptibility to pulmonary tuberculosis in the Iranian population. Int J Infect Dis IJID Off Publ Int Soc Infect Dis. 2010 Jan;14(1):e84-85.

376. Zhang H-Q, Deng A, Guo C-F, Wang Y-X, Chen L-Q, Wang Y-F, et al. Association between FokI polymorphism in vitamin D receptor gene and susceptibility to spinal tuberculosis in Chinese Han population. Arch Med Res. 2010 Jan;41(1):46–9.

377. Ates O, Dolek B, Dalyan L, Musellim B, Ongen G, Topal-Sarikaya A. The association between BsmI variant of vitamin D receptor gene and susceptibility to tuberculosis. Mol Biol Rep. 2011 Apr;38(4):2633–6.

378. McNamara L, Takuva S, Chirwa T, MacPhail P. Prevalence of common vitamin D receptor gene polymorphisms in HIV-infected and uninfected South Africans. Int J Mol Epidemiol Genet. 2016;7(1):74–80.

379. Pickrell JK, Reich D. Toward a new history and geography of human genes informed by ancient DNA. Trends Genet. 2014 Sep;30(9):377–89.

380. Campbell MC, Tishkoff SA. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet. 2008;9:403–33.

381. Crampton H. The Side of the Sun at Noon. Jacana Media; 2014. 486 p.

382. Fenner F, Henderson DA, Arita I, Jezek Z, Ladnyi ID, Organization WH. Smallpox and its eradication. 1988 [cited 2017 Mar 17]; Available from: http://www.who.int/iris/handle/10665/39485

383. The Greatest Killer [Internet]. [cited 2017 Mar 17]. Available from: http://www.press.uchicago.edu/ucp/books/book/chicago/G/bo3647267.html

384. Hewlett B. The peoples of southern Africa and their affinities. By G.T. Nurse, J.S. Weiner, and T. Jenkins. New York: Oxford University Press. 1986. xvi + 409 pp., figures, tables, index. $ 69.00 (cloth). Am J Phys Anthropol. 1987 Sep 1;74(1):135–6.

385. WHO | Frequently asked questions and answers on smallpox [Internet]. WHO. [cited 2017 Apr 4]. Available from: http://www.who.int/csr/disease/smallpox/faq/en/

386. Takiff HE, Feo O. Clinical value of whole-genome sequencing of Mycobacterium tuberculosis. Lancet Infect Dis. 2015 Sep;15(9):1077–90.

387. Oswald NC. Pulmonary tuberculosis in African native troops. Thorax. 1946 Jun;1:100–17.

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

389. South African National HIV Prevalence, Incidence & Behaviour Survey, 2012 [Internet]. Health-e. 2014 [cited 2017 Mar 26]. Available from: https://www.health-e.org.za/2014/04/01/south-african-national-hiv-prevalence-incidence-behaviour-survey-2012/

390. Karlsson EK, Kwiatkowski DP, Sabeti PC. Natural selection and infectious disease in human populations. Nat Rev Genet. 2014 Jun;15(6):379–93.

391. Owers KA, Sjödin P, Schlebusch CM, Skoglund P, Soodyall H, Jakobsson M. Adaptation to infectious disease exposure in indigenous Southern African populations. Proc R Soc B. 2017 Apr 12;284(1852):20170226.

392. Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Ferrer-Admettla A, Pattini L, et al. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genet. 2011 Nov;7(11):e1002355.

393. Fumagalli M, Sironi M. Human genome variability, natural selection and infectious diseases. Curr Opin Immunol. 2014 Oct;30:9–16.

394. Lindo J, Huerta-Sánchez E, Nakagome S, Rasmussen M, Petzelt B, Mitchell J, et al. A time transect of exomes from a Native American population before and after European contact. Nat Commun. 2016 Nov 15;7:13175.

395. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science. 2010 Jul 2;329(5987):75–8.

396. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.

397. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio [Internet]. 2013 Mar 16 [cited 2017 Apr 17]; Available from: http://arxiv.org/abs/1303.3997

398. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001 Feb 15;409(6822):860–921.

399. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010 Sep;20(9):1297–303.

400. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al. 2013 Oct 15;11(1110):11.10.1-11.10.33.

401. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011 Aug 1;27(15):2156–8.

402. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, Roussos P, et al. A polygenic burden of rare disruptive mutations in schizophrenia. Nature. 2014 Feb 13;506(7487):185–90.

403. Weir BS, Cockerham CC. Estimating F-Statistics for the Analysis of Population Structure. Evolution. 1984;38(6):1358–70.

404. Willing E-M, Dreyer C, Oosterhout C van. Estimates of Genetic Differentiation Measured by FST Do Not Necessarily Require Large Sample Sizes When Using Many SNP Markers. PLOS ONE. 2012 Aug 14;7(8):e42649.

405. Banerjee A, Chitnis UB, Jadhav SL, Bhawalkar JS, Chaudhury S. Hypothesis testing, type I and type II errors. Ind Psychiatry J. 2009;18(2):127–31.

406. Pruitt KD, Tatusova T, Brown GR, Maglott DR. NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res. 2012 Jan;40(Database issue):D130–5.

407. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–81.

408. Adzhubei I, Jordan DM, Sunyaev SR. Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Curr Protoc Hum Genet Editor Board Jonathan Haines Al. 2013 Jan;0 7:Unit7.20.

409. Zhang B, Kirov S, Snoddy J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W741-748.

410. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015 Jan;43(Database issue):D447-452.

411. Early slavery at the Cape of Good Hope, 1652-1717 - Protea Boekhuis [Internet]. [cited 2017 May 19]. Available from: http://www.proteaboekhuis.com/site.php/early-slavery-at-the-cape-of-good-hope-1652-1717.html

412. MapSlaveRoute.pdf [Internet]. [cited 2017 May 19]. Available from: http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CLT/pdf/MapSlaveRoute.pdf

413. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017 Jan 4;45(D1):D353–61.

414. Marx J. Inflammation and Cancer: The Link Grows Stronger. Science. 2004 Nov 5;306(5698):966–8.

415. Xu H, Storch T, Yu M, Elliott SP, Haslam DB. Characterization of the human Forssman synthetase gene. An evolving association between glycolipid synthesis and host-microbial interactions. J Biol Chem. 1999 Oct 8;274(41):29390–8.

416. Harant H, Lindley IJD. Negative cross-talk between the human orphan nuclear receptor Nur77/NAK-1/TR3 and nuclear factor-kappaB. Nucleic Acids Res. 2004;32(17):5280–90.

417. Zhu Y, Yao S, Iliopoulou BP, Han X, Augustine MM, Xu H, et al. B7-H5 costimulates human T cells via CD28H. Nat Commun. 2013;4:2043.

418. Cheshenko N, Liu W, Satlin LM, Herold BC. Focal adhesion kinase plays a pivotal role in herpes simplex virus entry. J Biol Chem. 2005 Sep 2;280(35):31116–25.

419. Elbahesh H, Cline T, Baranovich T, Govorkova EA, Schultz-Cherry S, Russell CJ. Novel roles of focal adhesion kinase in cytoplasmic entry and replication of influenza A viruses. J Virol. 2014 Jun;88(12):6714–28.

420. Owen KA, Meyer CB, Bouton AH, Casanova JE. Activation of Focal Adhesion Kinase by Salmonella Suppresses Autophagy via an Akt/mTOR Signaling Pathway and Promotes Bacterial Survival in Macrophages. PLOS Pathog. 2014 Jun 5;10(6):e1004159.

421. Elbahesh H, Bergmann S, Russell CJ. Focal adhesion kinase (FAK) regulates polymerase activity of multiple influenza A virus subtypes. Virology. 2016 Dec;499:369–74.

422. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS Biol. 2006 Mar;4(3):e72.

423. Packard RM. White Plague, Black Labor: Tuberculosis and the Political Economy of Health and Disease in South Africa. University of California Press; 1989. 422 p.

424. Tiberi S, Buchanan R, Caminero JA, Centis R, Arbex MA, Salazar M, et al. The challenge of the new tuberculosis drugs. Presse Médicale. 2017 Mar;46(2, Part 2):e41–51.

425. Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nat News. 2016 Oct 13;538(7624):161.

426. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci. 2009 Jun 9;106(23):9362–7.

427. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, et al. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 2010 Jul;42(7):565–9.

428. Shi H, Kichaev G, Pasaniuc B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. Am J Hum Genet. 2016 Jul 7;99(1):139–53.

429. Pickrell JK. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. Am J Hum Genet. 2014 Apr 3;94(4):559–73.

430. Li YI, Geijn B van de, Raj A, Knowles DA, Petti AA, Golan D, et al. RNA splicing is a primary link between genetic variation and disease. Science. 2016 Apr 29;352(6285):600–4.

431. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014 Jan 1;42(D1):D1001–6.

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

433. Shaw MA, Collins A, Peacock CS, Miller EN, Black GF, Sibthorpe D, et al. Evidence that genetic susceptibility to Mycobacterium tuberculosis in a Brazilian population is under oligogenic control: linkage study of the candidate genes NRAMP1 and TNFA. Tuber Lung Dis Off J Int Union Tuberc Lung Dis. 1997;78(1):35–45.

Contribution to Publications:

Chapter 2: Population structure and infectious disease risk in southern Africa

- First author
- Writing of manuscript

Chapter 3: Fine-scale human population structure reflects ecogeographic boundaries

- First author
- Autosomal data analysis and interpretation
- Writing of manuscript

Chapter 4: A post-GWAS analysis of predicted regulatory variants and tuberculosis susceptibility

- First author
- Conceived and designed computational pipeline
- Analysis and interpretation of data
- Writing of manuscript

Chapter 5: Signals of positive selection in indigenous southern African populations

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