THE WHITE PLAGUE CAME TO AFRICA: GENETIC SUSCEPTIBILITY TO TUBERCULOSIS AND THE IMPACT OF ANCESTRY

Prof Eileen Hoal

April 2016

THE WHITE PLAGUE CAME TO AFRICA: GENETIC SUSCEPTIBILITY TO TUBERCULOSIS AND THE IMPACT OF ANCESTRY

Inaugural lecture delivered on 11 April 2016

Prof Eileen Hoal
Division of Molecular Biology and Human Genetics
Faculty of Medicine and Health Sciences
Stellenbosch University

Editor: SU Language Centre
Printing: SUN MeDIA
ISBN: 978-0-7972-1605-1

BIOGRAPHY

Professor Eileen Hoal was born in Kroonstad in the Free State. She grew up in Queenstown in the Eastern Cape, where she attended primary school, and Windhoek in the then South West Africa, where she went to high school. She obtained her BSc at Rhodes University, after which she moved to the University of Cape Town (UCT) for her Honours, and UCT Medical School for her PhD. She and her husband (Professor Paul van Helden) then left for New Jersey, USA, where she did a postdoctorate at the Roche Institute of Molecular Biology. A few years later she completed the Higher Education Diploma through UNISA, for good measure. After her postdoctoral fellowship, she joined Stellenbosch University in 1983, and was promoted to associate professor in 2004.

Professor Hoal is head of the group that focuses on the host genetics of tuberculosis in the Division of Molecular Biology and Human Genetics. She has devoted 20 years to investigating the genetic contribution of the human host to individual and population susceptibility to tuberculosis. She used association studies, genome-wide linkage studies and genome-wide association studies to identify genes and loci that have provided novel candidates to inform the variation in disease outcome between individuals and populations. The complex admixed population in South Africa provided the opportunity to find ethnic-specific risk variants using admixture mapping, and to find associations between *M. tuberculosis* strain types and host genetic variation.

Eileen has published 88 papers in peer-reviewed journals, has an h-index of 31 and was awarded a B2 rating by the National Research Foundation in 2012.

ACKNOWLEDGEMENTS

First and foremost, I wish to acknowledge my family, which makes up in quality what it lacks in quantity. I barely knew my grandfathers owing to the long generation times in my family, but I had two formidable grandmothers. My father Arthur was what I like to call a renaissance man – he was an engineer, but his knowledge on all topics was encyclopaedic, and when he needed to check on something, my brother and I would have to join in the search in the Encyclopaedia Britannica. My mother Zaideth was top of her class, and became a full-time mom to her children. She cannot be here today as she, at 92, is in Health Care at her retirement village. My two children, Stephen and Lesley, have taught me most of all, and were very patient and understanding of their often distracted mother.

I would like to thank colleagues and students too numerous to mention who have all enriched my life and work, and all my collaborators, both local and abroad, particularly my closest colleagues, Drs Marlo Möller and Craig Kinnear. Last but not least, I give credit to my husband Paul, my collaborator in life. Although we found that working at the same lab bench was not a good idea, we have happily and successfully worked in the same environment for over 35 years and shared the ups and downs of a life in science.

THE WHITE PLAGUE CAME TO AFRICA: GENETIC SUSCEPTIBILITY TO TUBERCULOSIS AND THE IMPACT OF ANCESTRY

INTRODUCTION

Tuberculosis (TB) was declared a global public health emergency by the World Health Organization (WHO) in 1993. Nearly 25 years later, after a major drive to eradicate or at least fully understand TB, the global burden of this disease is not significantly diminishing. According to the WHO report of 2015, nearly 10 million people became ill with TB in 2014 and 1.5 million died.

The second highest incidence of TB in the world is found in the Western Cape in South Africa. Although infection with *Mycobacterium tuberculosis (M. tuberculosis)* is necessary, it is not sufficient to cause tuberculosis in most people. This is substantiated by the accepted notion that only 5–10% of infected persons who are immunocompetent will ever develop the active disease, while the majority of the population control the bacterium effectively.

The idea that tuberculosis is influenced not only by the bacterium, but also by both genetic and environmental factors was formally stated 57 years ago in the book *The White Plague: Tuberculosis, Man and Society* (Dubos and Dubos, 1952). The authors were also of the opinion that medical solutions alone would not prevent or cure tuberculosis. Thus far, history has proved this opinion correct. However, the authors did not consider the possible effect of research in the field of human genetic susceptibility to tuberculosis.

Before the discovery of *M. tuberculosis*, it was observed that tuberculosis frequently occurred in several members of the same family, appearing to be hereditary. However, the discovery of the bacterium focused attention on the importance of the pathogen, while the host genome was largely ignored.

The historical impression that tuberculosis was an inherited disorder has come full circle and substantial evidence now exists for the human genetic contribution to susceptibility to tuberculosis. A classic epidemiological study on the premature death of adoptees in Denmark suggested that the genetic contribution to infectious disease is greater than that for cancer or cardiovascular disease (Sorensen et al., 1988). It has also been noted that populations not previously exposed to the bacterium, such as the Qu'Appelle Indians, at first had a high annual TB mortality rate (10%), with a subsequent decrease in deaths (to 0.2%) after years of exposure and the eradication of half of the families (Motulsky, 1960). Motulsky suggested that this could have been due to strong selection against susceptibility genes for TB. Natural selection against susceptibility genes could also be the explanation for the seemingly different TB resistance of various populations: European individuals seem to be less susceptible to TB, possibly due to many centuries of contact with the bacterium in Europe which resulted in the selection of a more resistant population. In contrast, sub-Saharan Africa was only recently exposed to *M. tuberculosis*, and modern-day TB treatment, although important for the affected individual, may have hampered the selection against susceptibility variants in African populations. This difference between populations is not due to socioeconomic factors alone, as first suggested by a study done in a USA nursing home. There, individuals with African ancestry were twice as likely to be infected with *M. tuberculosis* as individuals with European ancestry, even though they shared the same environment (Stead et al., 1990). A second study found that white US-born TB patients were more likely to be homeless than black US-born TB patients (Serpa et al., 2009), which supported the initial finding.

TB has now been established to be a multifactorial disease with host, pathogen and environmental factors all contributing to the development of active TB disease. The current failure of antibiotics to alleviate the high global incidence of the disease, despite their widespread use, suggests that additional approaches that target aspects of TB other than the causative pathogen (*M. tuberculosis*) are required to fight the disease. Heritability analysis of anti-mycobacterial activity, which measures the strength of the genetic influence on the immune phenotype, has been extended from twins to families and household contacts (Stein et al., 2003), and these studies estimate heritability at 30% to 68%.

As the variable host immune response plays an important role in the outcome of infection, studying how the host genetic make-up differs between individuals who develop active TB and those who remain healthy may well lead to the identification of genes and gene variants that could help explain this variable outcome (see Fig. 1).

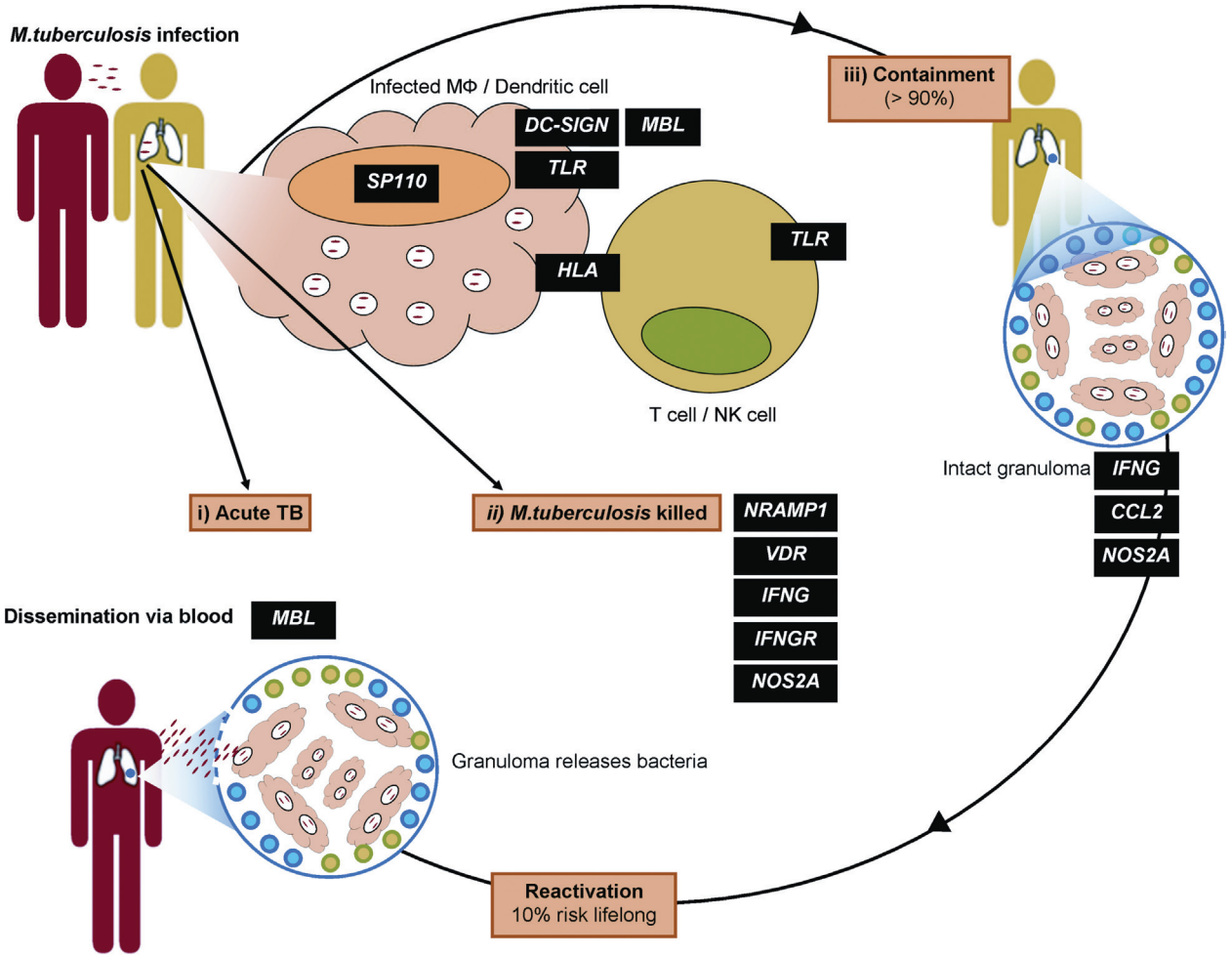

Figure 1: Genetic involvement in the tuberculosis disease process (Möller et al., 2010)

In the last 15–20 years, my research group has employed this approach to look for the genetic factors influencing the resistance or susceptibility of an individual to TB. Much evidence has come from several whole-genome linkage scans; numerous case-control association studies where the candidate genes were derived from genome screens; animal models; hypotheses pertaining to the disease pathways; and, more recently, other whole genome approaches.

ASSOCIATION STUDIES

The most widely employed approach to investigate TB susceptibility for many years was candidate gene association studies. Population-based case-control studies are classic association studies where the allele frequency of a specific marker is compared between unrelated cases (affected individuals) and controls (unaffected individuals) when Hardy–Weinberg equilibrium holds (Lander and Schork, 1994). This design has much greater power than linkage studies have (Risch and Merikangas, 1996) for detecting genes of small effect, provided the sample size is adequate.

Many candidate genes have consistently been associated with TB in multiple studies, implicating a variety of pathways in the disease process. These include genes such as HLA, NRAMP1, IFNG, NOS2A, SP110, CCL2, MBL, CD209, VDR and TLR, which code for cytokines, chemokines, transmembrane transporters, Vitamin D and its receptor, nitric oxide, the apoptosis pathway and cellular receptors for *M. tuberculosis* itself (see Fig. 2). Evidence for the involvement of these genes has been found in many different populations, and is reviewed in Möller et al. (2010) and Möller and Hoal (2010).

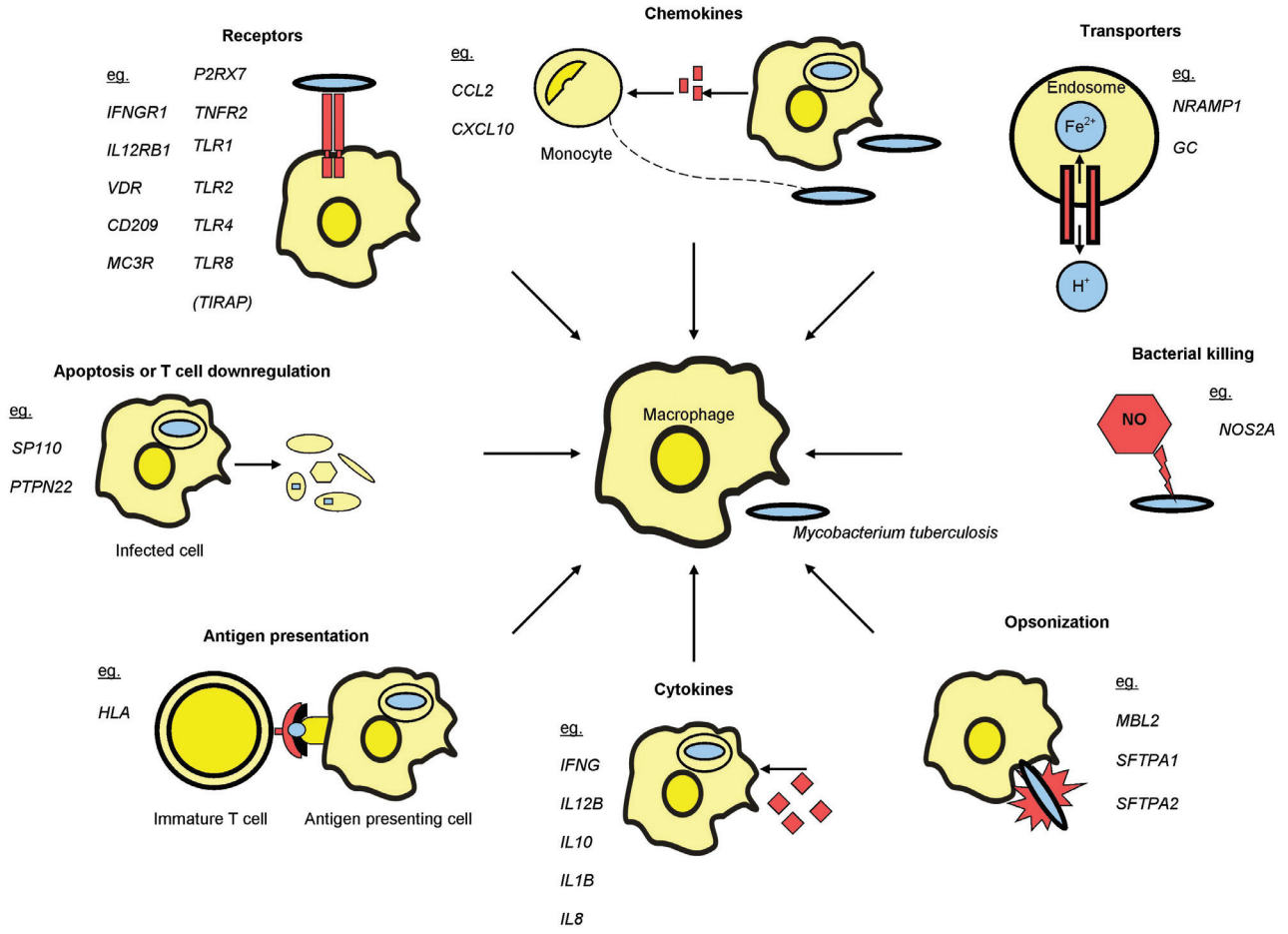

Figure 2: Human genes and pathways implicated in host resistance or susceptibility to tuberculosis (Möller and Hoal, 2010)

Toll-like receptors (TLRs) are emerging as some of the most repeatedly identified candidate genes in TB susceptibility. TLRs each have a distinct function, and are involved in the initiation of the innate immune response, an essential component of the host defence against invading pathogens (Misch and Hawn, 2008). TLRs are able to recognise pathogen-associated molecular patterns (PAMPs), conserved microbial structures necessary for survival, resulting in the initiation of signal transduction, transcription of pro-inflammatory cytokine genes and development of the adaptive immune response. We genotyped 23 polymorphisms in five TLR genes in 1 216 cases and controls, and detected some interesting sex-specific associations for TLR8 polymorphisms, with three SNPs associated in females, and only two of these associated in males, with the Odds Ratios in the opposite direction (Salie et al., 2015a).

X CHROMOSOME

The findings above suggest a partial explanation for the male bias in tuberculosis ratios, as TLR8 is located on the X chromosome at Xp22 (Salie et al., 2015a). Several infectious diseases have sex-specific differences (Van Lunzen and Altfeld, 2014), including TB (Nhamoyebonde and Leslie, 2014). Females have stronger immune responses than males and have higher CD4+ T lymphocyte counts (Dale et al., 2006), and the X chromosome contains chiefly immunomodulatory genes. A study in Indonesia (Davila et al., 2008) showed that the four TLR8 SNPs investigated above were associated with TB susceptibility in males. Females showed the same trend, but this was not significant. The association was validated in Russian males (females were not tested). The first genome-wide linkage study, including our population, suggested a locus on the Xq chromosome (Bellamy et al., 2000), and worldwide, across many cultures, the incidence of TB is skewed towards the male gender, with an approximate average male/female ratio of 2.5. However, the only other evidence for the X chromosome involvement in TB susceptibility is in MSMD patients where disease appears to be male-specific, and mutations in the gene called inhibitor of kappa light polypeptide gene enhancer in B-cells, kinase gamma (IKBKG or NEMO) and two

other candidate regions on X have been identified. Biological mechanisms may account for most of this difference, but other factors (such as the environment, including hormonal milieu, or bacterial factors) will also influence the outcome of infection. This highlights the need for investigating X chromosome-linked genes for TB susceptibility variants.

In genome-wide association studies (GWAS), discussed below, the X chromosome results were usually not included, firstly because males are haploid and females diploid for X-linked genes and secondly because females undergo X chromosome inactivation, which complicates the statistical analysis. Thirdly, software and tools for analysing the X chromosome are not as well developed as tools for analysing the autosome. Newer chips have a higher SNP density and better coverage on the X chromosome, and methods and software are being developed for analysis of X in GWAS.

M. TUBERCULOSIS STRAINS AND HLA

Many association studies have not been validated. Factors such as population stratification, inadequate sample size, incorrect phenotyping and publication bias could explain this. In addition, all human TB genetic studies are hindered by our inability to determine the degree of exposure to *M. tuberculosis*. Further variability is introduced by the number of different strains of *M. tuberculosis* causing disease, with pronounced variation by continent and country. In West Africa for example, approximately one-third of TB cases are caused by *M. africanum*, regarded as another species in the *M. tuberculosis* complex, which may influence the human genes associated with susceptibility.

Collectively, most mycobacteria causing human disease are referred to as members of the *M. tuberculosis* complex (MTBC) (Cole et al., 1998) and all share a common African ancestor about 35 000–15 000 years ago. All modern *M. tuberculosis* strains, on the other hand, are thought to be descended from a common ancestor approximately 20 000–15 000 years ago. These modern *M. tuberculosis* strains have been classified into six major lineages and are geographically structured (Gagneux et al., 2006), where the East-Asian lineage has been shown to be more dominant in many Far East countries, while the Euro-American lineage is predominant in Europe and the Americas, as illustrated in Figure 3. Human-adapted MTBC is more genetically diverse than assumed (Hershberg et al., 2008).

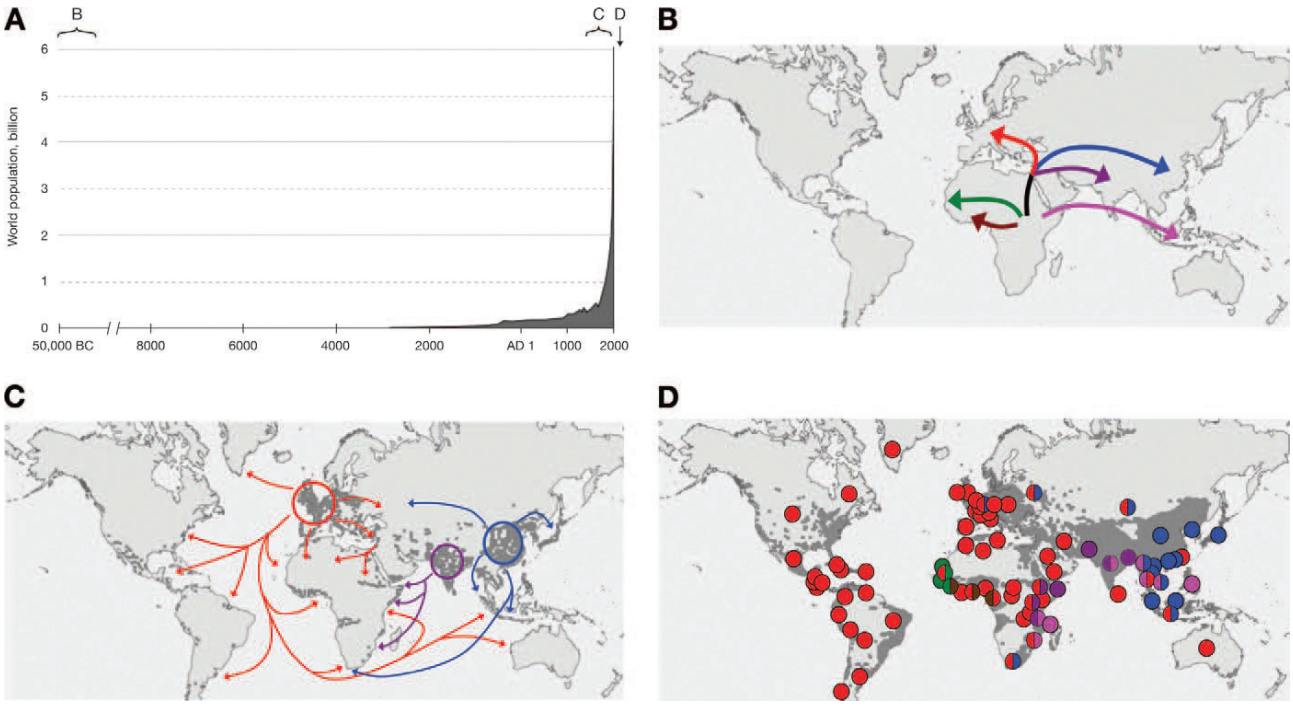

Figure 3: Out-of-and-back-to-Africa scenario for the evolutionary history of human-adapted MTBC. (A) Global human population size during the last 50 000 years. (B) Hypothesised migration out of Africa of ancient lineages of MTBC. (C) Recent increase of global human population. Each dark grey dot corresponds to 1 million people. (D) Human population at 6 billion, with the distribution of the six main MTBC lineages observed today (Hershberg et al., 2008)

In humans, the human leucocyte antigen (HLA) region consists of approximately 200 genes, many of which are involved in antigen presentation. Genes that are involved in protective immunity show greater variance than other genes, presumably owing to selection pressure (Murphy, 1993) and this is also the case for the HLA region, which varies between populations and is highly polymorphic.

We investigated the role of the HLA class I genes in the adaptation of *M. tuberculosis* strains to specific human populations. Three hundred patients with tuberculosis from the South African Coloured (SAC) population group were typed for their HLA class I alleles, and their *M. tuberculosis* strain was genotyped. We showed that the Beijing strain occurred more frequently in individuals with multiple disease episodes and the human HLA-B27 allele lowered the odds of having an additional episode (Salie et al., 2014). This supertype allele is found frequently in individuals who are able to control their HIV infections without any antiretroviral treatment and slow disease progression (Migueles and Connors, 2010). This is thought to be due to an increased CD8 T-cell response in individuals with this allele and induction of the apoptotic pathway through the increased presence of cytotoxic proteins (Migueles et al., 2009).

Beijing strains are the most dominant lineage globally: first in many Asian countries, and are emerging as the dominant strain in several other countries, including South Africa (Van der Spuy et al., 2009). We postulated a natural experiment in coevolution taking place in the Cape Town area, which has experienced a multiplicity of human visitors and their mycobacterial strains over the past 350 years. The resident population is extremely diverse (De Wit et al., 2010b), with inputs from Khoesan, Bantu, European, and Asian people and could therefore be assumed to have HLA types from all these ancestral populations. The *M. tuberculosis* strains present can be expected to have experienced intense competition and the incidence of tuberculosis is one of the highest in the world, namely 1 005 per 100 000 in 2007 (Health Systems Trust, 2009), thereby enabling us to investigate correlations between bacterial strain and HLA type in adequate numbers of patients (Salie et al., 2014).

The present-day population structure of *M. tuberculosis* can be attributed to ancient human migrations (Hershberg et al., 2008). Such long-standing host–pathogen interactions could lead to adaptive genetic changes in both the host and pathogen populations, evidenced by the selection of strains from a distinct sublineage by a human population in a defined geographical setting, as we have seen in the Western Cape (Hanekom et al., 2007).

Studies in animals and humans have shown that *M. tuberculosis* is capable of stimulating MHC class I restricted CD8+ T cells and the involvement of several different pathways for class I presentation of mycobacterial antigens via cross-presentation (Sigal et al., 1999). A clear association of the predominant MTBC lineage and the HLA class I allele frequency to fit the hypothesis of the coevolution of strains with the HLA class I genes was not always apparent (Salie et al., 2014), perhaps because HLA genes are involved in several biological processes and some could thus be under balancing selective pressures (Spurgin and Richardson, 2010).

We also identified associations between specific TLR polymorphisms and disease caused by specific strains, respectively TLR4 and TLR9 SNPs with disease caused by the LAM strain (the most frequent in the study population)(Salie et al., 2015a). It is known that different *M. tuberculosis* strains activate distinct TLRs and stimulate disparate immune responses (Carmona et al., 2013).

KIR

In a closely allied study, we investigated the role of killer immunoglobulin-like receptor (KIR) genes and HLA class-I variants in susceptibility to tuberculosis. Natural killer (NK) cells are part of the early defence mechanisms employed when cells are experiencing various forms of stress, such as infections and malignant transformations, promoting the expression of cytokines and direct cytotoxicity (Cooper et al., 2001). The primary receptors on the surface of NK cells, which allow the cells to distinguish normal host cells from allogeneic or abnormal autologous cells, are the killer cell immunoglobulin-like receptors (KIRs). NK cell activity and cytokine production is largely regulated by the KIRs and their HLA ligands. As with cytotoxic T cells, NK cells mediate programmed cell death through the release of cytotoxic granules that penetrate the cell membrane of infected cells. When detecting bacteria, NK cells secrete α-defensins, antimicrobial peptides that disrupt the bacterial cell wall. In a sample set of 759 cases and controls, we showed that the KIR3DS1 gene and KIR genotypes with five or more activating KIRs (aKIRs), and the presence of 3DS1, protect against developing active TB in the SAC population (Salie et al., 2015b). Individuals with a greater number of aKIRs have a greater NK cell response to mycobacterial infection owing to an increased interferon-gamma (IFN-γ) response (Portevin et al., 2012).

BOVINE TB

Much physiology to do with infectious diseases is common across species, particularly among mammals, and we recently embarked on a study of the African buffalo, which is in the grip of a major epidemic caused by the introduced *M. bovis*, causing bovine TB (BTB). The "White Plague", according to Wikipedia, refers to both TB and white colonists, and it is certainly true that both combined to introduce mycobacterial disease to the wildlife of South Africa. In buffalo, the species has not had time to develop resistance to the disease, and can therefore be expected to manifest high frequencies of susceptibility alleles. African buffalo are a maintenance host of BTB in southern Africa, and apart from informing the management of this important wildlife species, findings in this project may well indicate genes worthy of investigation in the human.

We performed next-generation sequencing and mapping of the African buffalo genome from a total of nine animals, on an ABI SOLiD4 sequencer. The resulting short reads were mapped to the UMD3.1 *Bos taurus* (cow) genome assembly using two different software packages, which identified over 270 000 SNPs, 173 of which were validated by fluorescent genotyping in 87 individuals (Le Roex et al., 2012).

We explored the polymorphisms in immune-related genes and their association with BTB susceptibility in the African buffalo and found SNPs located in three genes to be significantly associated (Le Roex et al., 2013). Further work using a modified killing assay for mycobacteria has implicated TLR6 diversity and demographic parameters in the ability of African buffalo to restrict mycobacterial growth (Le Roex, in preparation). In addition, we have very recently completed the first whole genome sequencing of the African buffalo, to be published shortly.

LINKAGE STUDIES

Linkage studies can identify the estimated genetic location of TB susceptibility by testing for co-segregation between a genetic marker and a possible disease locus, and are less likely to yield false positive results than association studies. If this is by a genome-wide scan, it will ensure that all major genomic regions involved in disease susceptibility are identified. This provides the opportunity to find new genes and pathways that might not have been suspected previously to contribute to the disease studied.

Ideally, linkage studies should indicate major susceptibility genes in a disease, but, since TB is a complex disease, it is not expected that this will be the case. A number of genome-wide linkage studies of TB phenotypes have been done. The majority of these have all identified different loci, perhaps owing to the particular diagnostic criteria or phenotype used, including the *M. tuberculosis* strain, or to population genetic characteristics. However, there have been some confirmed loci.

A nonparametric linkage analysis (affected sib-pair study) was done as a follow-up to our first scan (Bellamy et al., 2000). Using South African and Malawian populations, we identified areas on chromosomes 6p21–q23 and 20q13.3 (Cooke et al., 2008) and follow-up association studies validated the cathepsin Z (CTSZ) and melanocortin 3 receptor (MC3R) genes in tuberculosis susceptibility (Adams et al., 2011). The CTSZ protein is mostly expressed in immune cells such as macrophages and monocytes, and the MC3R protein is expressed in peripheral tissues such as immune cells and plays a role in several biological systems, regulating energy homeostasis, fat metabolism, and inflammation.

TB RESISTANCE LOCI

The most recent genome-wide linkage scan in the SAC population, and potentially the one most likely to impact on our understanding of the resistance phenotype, enabled us to identify the first genetic resistance factor for TB infection (Cobat et al., 2009).

In this study of 128 South African families including 350 siblings, most individuals were likely to have been exposed to *M. tuberculosis*, but 40% had no reaction to the Tuberculin Skin Test (TST), which tests for delayed-type hypersensitivity (DTH). The TST is an *in vivo* test which we have shown has low redundancy with other *in vitro* assays (Gallant et al., 2009). The chief purpose of the project was to correlate a marker of immune reactivity to the genetic make-up, and the study was exceptionally powerful as it was family-based. The heritability of the various measures of the immune reaction was high, with heritabilities from 50–75% (Cobat et al., 2010).

Linkage analysis of this quantitative immunological reaction to injected PPD determined that a single major locus on chromosomal region 11p14, which we called *TST1,* appears to control human resistance to the bacterium, as evidenced by the lack of DTH.

A second locus on chromosome 5p15, which we called *TST2*, was found to determine the intensity of the TST response. This region had previously shown evidence of linkage with persistently low TST reactivity in Uganda (Stein et al., 2008). Fine-mapping of the region identified the solute carrier family 6 gene as a promising candidate (Cobat et al., 2009).

These genetic factors might determine whether an infected individual keeps the bacterium dormant or develops the disease. The measurement of TST did not correlate with IFN-γ release, indicating that other pathways could be involved. Further fine mapping of these regions is ongoing. This work (Cobat et al., 2009) was the first report of a genetic resistance factor for TB infection as opposed to disease. One major locus determines innate resistance to *M. tuberculosis* infection in endemic areas and a second controls the extent of that response via regulators of T-cell-dependent DTH. These findings suggest that we could one day manipulate cellular mechanisms to prevent TB, and they demonstrate the pivotal role played by host genetics in quantitative measures of antimycobacterial immunity underlying immune diagnosis of TB infection.

A further phenotype measured in the study above was the extent of Tumor necrosis factor α (TNF) production after stimulation with BCG or BCG plus IFN-γ, using a whole blood assay in 392 children belonging to 135 nuclear families. TNF is a key immune regulator of tuberculosis resistance, as exemplified by the highly increased risk of TB among individuals receiving TNF-blocker therapy. We conducted classical univariate and bivariate genome-wide linkage analysis of TNF production, and its extent following both stimulation protocols was highly correlated (r = 0.81). Using a multivariate approach, we detected a major pleiotropic locus ($P < 10-5$) on chromosome region 11p15, termed TNF locus 1 (*TNF1*), that controlled TNF production after stimulation by both BCG alone and BCG plus IFN-γ. Surprisingly, the *TNF1* locus was mapped in the vicinity of the *TST1* locus we had found previously, which controls TST negativity, i.e. T-cell-independent resistance to *M. tuberculosis* infection. This suggested that there is a connection between TST negativity per se and TNF production (Cobat et al., 2013).

TNF is critical for the sequestration of *M. tuberculosis* in the granuloma during latent infection. Hence, the overlap of *TST1* with *TNF1* provides additional support for the interpretation of the *TST1* locus as conferring resistance to *M. tuberculosis* infection. In addition, the observation that the suggestive locus on chromosomal region 19q13 mapped in the vicinity of a locus also found to be linked to TNF production in the South African family sample (Cobat et al., 2013) reinforces the hypothesis of a genetic connection between TST negativity and TNF production.

The original linkage was then replicated in an independent sample from Paris, France, where the incidence is low, and the study design was a household contact study (Cobat et al., 2015). A genome-wide analysis of TST negativity identified a significant linkage signal in the close vicinity of *TST1*. A combined analysis of the samples from France and South Africa increased evidence of linkage, further implicating genetic factors located on 11p14-15. This region overlaps the *TNF1* locus controlling mycobacteria-driven TNF production (Cobat et al., 2013) (see Fig. 4). The South African sample was of the Coloured ethnic group, in a hyperendemic area, while the French individuals belonged to several ethnic groups (white, African, Asian and other), in an area where the endemicity of tuberculosis is low (annual incidence of 22.1 cases per 100 000). Replication of *TST1* in such different settings suggests an important role of this locus in the control of TST negativity in humans. Significantly, the risk of developing tuberculosis for immunocompetent persons without TST reactivity, despite sustained exposure to *M. tuberculosis,* has been shown to be extremely small (Rose et al., 1995).

GENE INTERACTION STUDIES

Since TB is influenced by multiple genetic, environmental and bacterial factors, it is likely that gene interactions could determine the outcome of infection with *M. tuberculosis*. A commonly posited explanation for the missing heritability of complex disease is gene-gene interactions, also referred to as epistasis, encompassing the idea that interactions play a far more important part in an individual's susceptibility to a complex disease than single polymorphisms alone. It has been argued (Seldin et al., 2011) that it is important to study complex disease epistasis in admixed populations, and that this may well uncover novel interactions that are not detectable in the source populations.

In our first study, we genotyped over 800 cases and controls for 11 polymorphisms in nine genes, all of which had shown evidence of association with TB. With eight instances of statistically significant gene-gene interactions, the importance of epistasis was clearly identifiable (De Wit et al., 2010a). Subsequently, we investigated the role of gene-gene interactions using a discovery data set incorporating genotypes from a large number of candidate gene studies as well as our genome-wide data. After limiting our search space to pairs of putative TB susceptibility genes, as well as pairs of genes that have been curated in online databases as potential interactors, we used statistical modelling to identify pairs of interacting SNPs. This was a more appropriate approach than data mining techniques as our data set

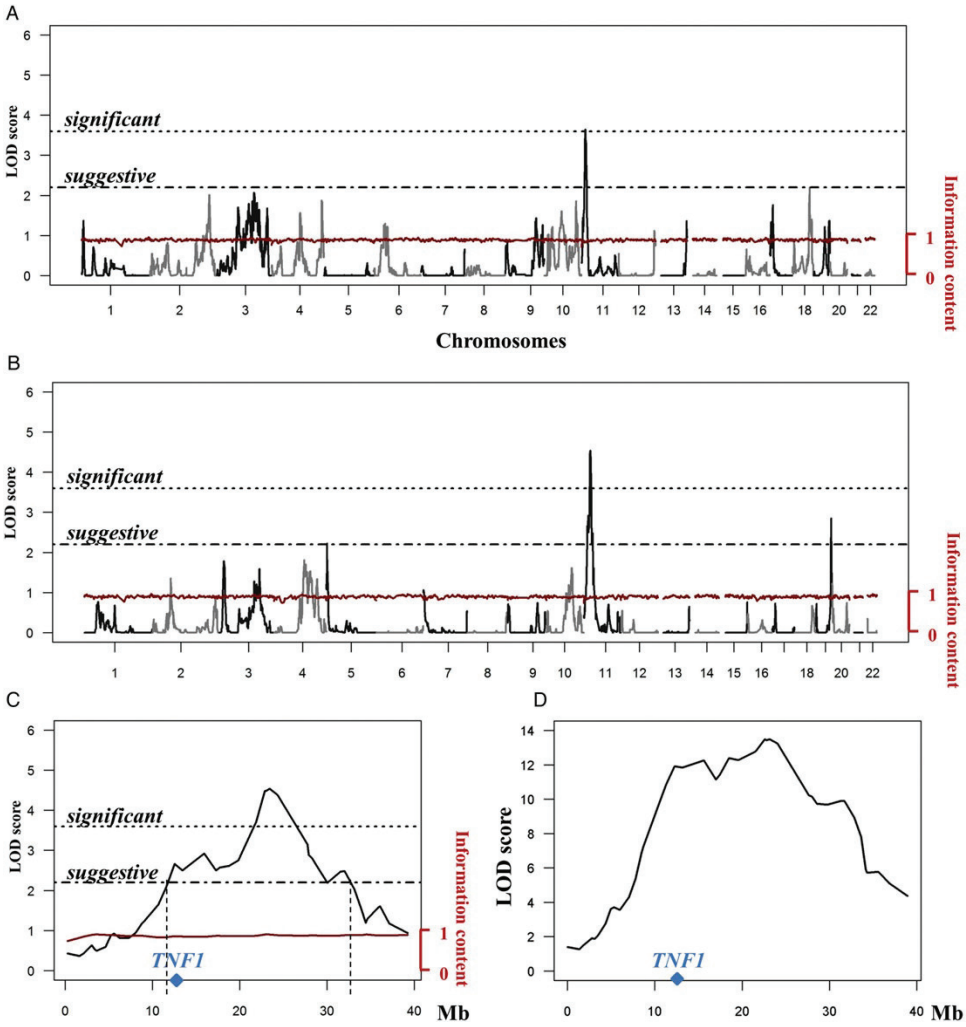

Figure 4: Genome-wide model-free linkage analysis of tuberculin skin test negativity (TST induration diameter, 0 mm vs >0 mm) in (A) the French and (B) the combined French and South African samples; (C) Expanded view of the region with the highest LOD score in the combined analysis; (D) Multipoint evidence of linkage in the combined sample restricted to 114 families contributing to the linkage peak on 11p14-15 (Cobat et al., 2015)

comprises sample sets genotyped in a number of different studies which did not always overlap well, resulting in a relatively sparse data set. Statistical modelling allowed us to adjust for known differences between cases and controls (age, sex and ancestry) and utilised all available data for each test, without requiring imputation or other complex strategies to deal with missing data. We used an independent Gambian TB case-control data set to validate the top models identified in our discovery data set. A number of models where there was overlapping data were successfully validated (Daya et al., 2015).

Genetic susceptibility studies in tuberculosis are additionally complicated by the presence of two different genomes, the bacterium and the host, and the influence their interaction can have on the disease. This includes the influence of the mycobacterial strain, discussed above under the heading '*M. tuberculosis* strains and HLA'.

GENOME-WIDE ASSOCIATION STUDIES

Since the completion of the human genome sequence and the International HapMap project, the ever-decreasing cost of genotyping and improvements in genotyping technology have led to the ever-increasing volume of data generated. 'Big data' has totally changed the face of genetics over the last 15 years, and the pace of change continues to increase. We have progressively been able to afford to do genome-wide association studies (GWAS), exome sequencing, and now whole genome sequencing (Hoal, 2011).

Many GWAS of complex diseases have been done. The most common polymorphisms in the genome, regardless of the location of possible causal variants, are genotyped. This approach is based on the 'common disease, common variant' (CDCV) hypothesis and, because most of the genome is surveyed, it eliminates the disadvantages of the single polymorphism or candidate gene approach, where only a few polymorphisms are investigated (Hirschhorn and Daly, 2005).

The underlying assumption was that a number of genes of modest effect sizes will be implicated in the disease, but it is now apparent that the number of common variants involved is large, and even positive GWAS have shown that a great deal of heritability remains unexplained. All genes found seem to have a relatively small effect size. However, it could be argued that one sees what one looks for, and GWAS (by virtue of their design) will never detect rare variants. Recently, the 'Common Disease, Rare Variant' (CDRV) hypothesis has arisen, postulating that multiple rare polymorphisms are the main cause of common diseases and could account for the hitherto unexplained heritability.

GWAS of TB have been slow and limited by small sample sizes. A GWAS in Indonesia identified four polymorphisms in the TLR8 gene on chromosome X (see above), and this was replicated in a Russian cohort (Davila et al., 2010). A study combining African populations in Ghana and Gambia identified the 18q11.2 locus (Thye et al., 2010) and later a larger analysis reported a new TB susceptibility locus on chromosome 11p13 in a Ghanaian population near the WT1 gene, which was replicated in samples from Gambia, Indonesia and Russia (Thye et al., 2012).

ADMIXTURE MAPPING I

Owing to our not having tens of thousands of samples to do a definitive GWAS, we decided on the specialised approach of admixture mapping. An admixed population is defined as one where the ancestral founders came from different genetic backgrounds. Admixture in a population is a potential confounder in association studies, since the heterogeneity of genetic backgrounds among study participants may lead to false positive or false negative results. Admixture mapping however takes advantage of admixture and specifically requires a population which developed from two or more genetically diverse parent populations, usually at the scale of continental diversity. This type of study also has the requirement that the incidence of the disease of interest, and therefore presumably of the causal risk variants, is different in the founding populations. In order to identify the loci responsible for the disease in cases, the regions of the genome inherited by cases from a specific ancestral population must be identified.

The primary tools are the genetic markers that occur with significantly different allele frequencies in different population groups. When risk alleles vary across populations, genetically mixed individuals with the disease under investigation are likely to have a higher probability of having inherited the loci near the disease loci from the population at higher risk of the disease. The study design is implemented as a genome-wide approach in admixed populations and has been used in a variety of complex diseases in African-Americans.

We hypothesised that it was feasible to use admixture mapping to study TB in the SAC population, since there are different disease rates, not based on socioeconomic differences, between European and black individuals (Stead et al., 1990). The admixed population in South Africa has major genetic contributions from both these groups, discussed below, but the first admixture mapping study with respect to TB in this population by our research group was ahead of its time, and this technique could not be used owing to spurious deviations in average local ancestry generated by the local ancestry inference methods at that time.

Instead, we conducted a GWAS on our data, followed by a meta-analysis and transethnic fine mapping (Chimusa et al., 2014). Our GWAS results confirm the WT1 chr11 susceptibility locus previously identified (Thye et al., 2012), showing association that is almost genome-wide significant, despite the limited sample size. This first ancestry-specific GWAS of TB risk in the complex admixed SAC population is not due to confounding by socioeconomic status. Another novel finding was that, with our 642 cases and 91 controls, a positive correlation existed between African San ancestry and TB susceptibility, and negative correlations existed with European and Asian ancestries.

WHAT ADMIXTURE?

In preparation for using the Affymetrix 500k chip data for admixture mapping, we needed to investigate the ancestral population contributions to the SAC population. The result was the dissection of the ancestral populations of nearly 1 000 individuals living in Cape Town, constituting a definitive study of the SAC in the Western Cape, the home of the majority of this population group (De Wit et al., 2010b). The structure of the SAC population was not quite what was expected in terms of the high Khoesan input, as Figure 5 shows.

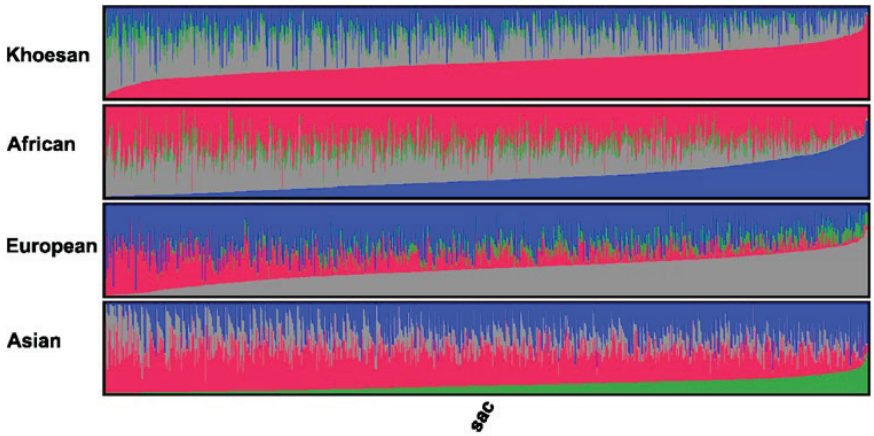

Figure 5: STRUCTURE plot. Proportion of each individual's ancestry (K = 4) sorted (in ascending order from left to right) by the proportion of ancestry for each of the major contributions to the SAC

Analysis by means of both the admixture and linkage models in STRUCTURE revealed that the major ancestral components of the SAC population are Khoesan (32–43%), Bantu-speaking Africans (20–36%), European (21–28%) and a smaller Asian contribution (9–11%). A further investigation of the SAC population to determine the relative contributions from the maternal and paternal sides of the various ancestors, again found an exceptionally high Khoesan contribution from the maternal side, providing much historical insight (Quintana-Murci et al., 2010). This is illustrated in the multidimensional scaling plots in Figure 6. In the early days of the Cape settlement by the Dutch, a small but significant number of women of Khoekhoe or of slave descent and their children were integrated into colonial households, often by marriage (Mountain, 2003; Shell, 1994). In the majority of cases, and particularly after 1700, the progeny of such mixed marriages and liaisons were assimilated into the growing group now known as the 'Coloureds' (Keegan, 1996).

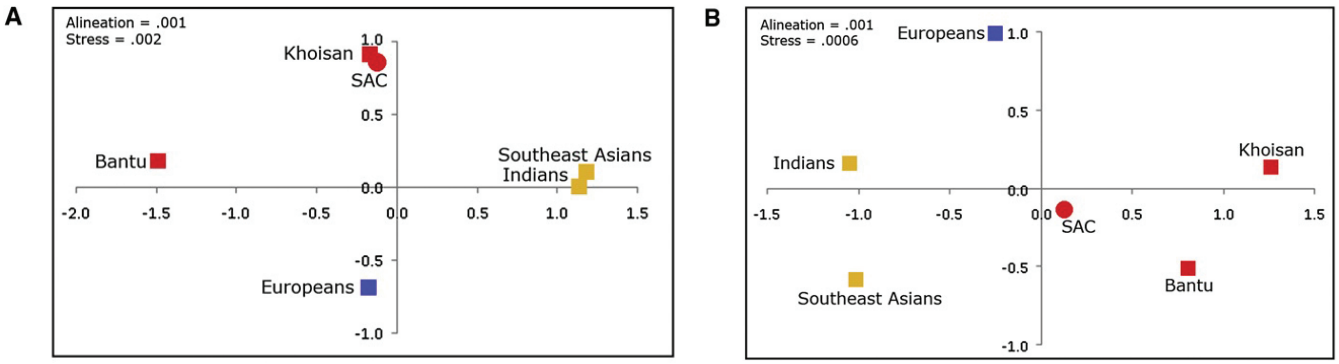

Figure 6: Multidimensional scaling (MDS) plots illustrating the relationships between the SAC and parental populations, based on (A) mtDNA and (B) NRY haplogroup frequencies (Quintana-Murci et al., 2010)

Inferences of admixture proportions and local ancestry make use of reference populations, and accuracy depends on the choice of ancestral populations. A few years later, owing to data from more individuals of putative ancestral groups, especially Khoesan, becoming available, we refined these estimates. Using a novel algorithm for proxy ancestry selection (Chimusa et al., 2013), and genome-wide SNP data from over 764 individuals, we estimated the genetic contributions from the best ancestral populations: Xhosa (33%), ǂKhomani San (31%), European (16%), Indian (13%), and Chinese (7%). It is important to note that the Xhosa themselves have Khoesan admixture, which may cause this Xhosa estimate to be slightly too high, and the San estimate to be too low. A principal component analysis showing the SAC and ancestral populations can be seen in Figure 7.

In genetic association studies, admixture is a well-known confounder. If genome-wide data is not available, as would be the case for candidate gene studies, ancestry informative markers (AIMs) are required in order to adjust for admixture. As humans migrated out of Africa, genetic drift or adaptation and selection resulted in different frequencies of genetic variants in the resultant populations. It is often possible to cluster individuals into population groups that correspond to their self-reported ancestry because of these differences. AIMs are those polymorphisms with the greatest difference in frequency between populations.

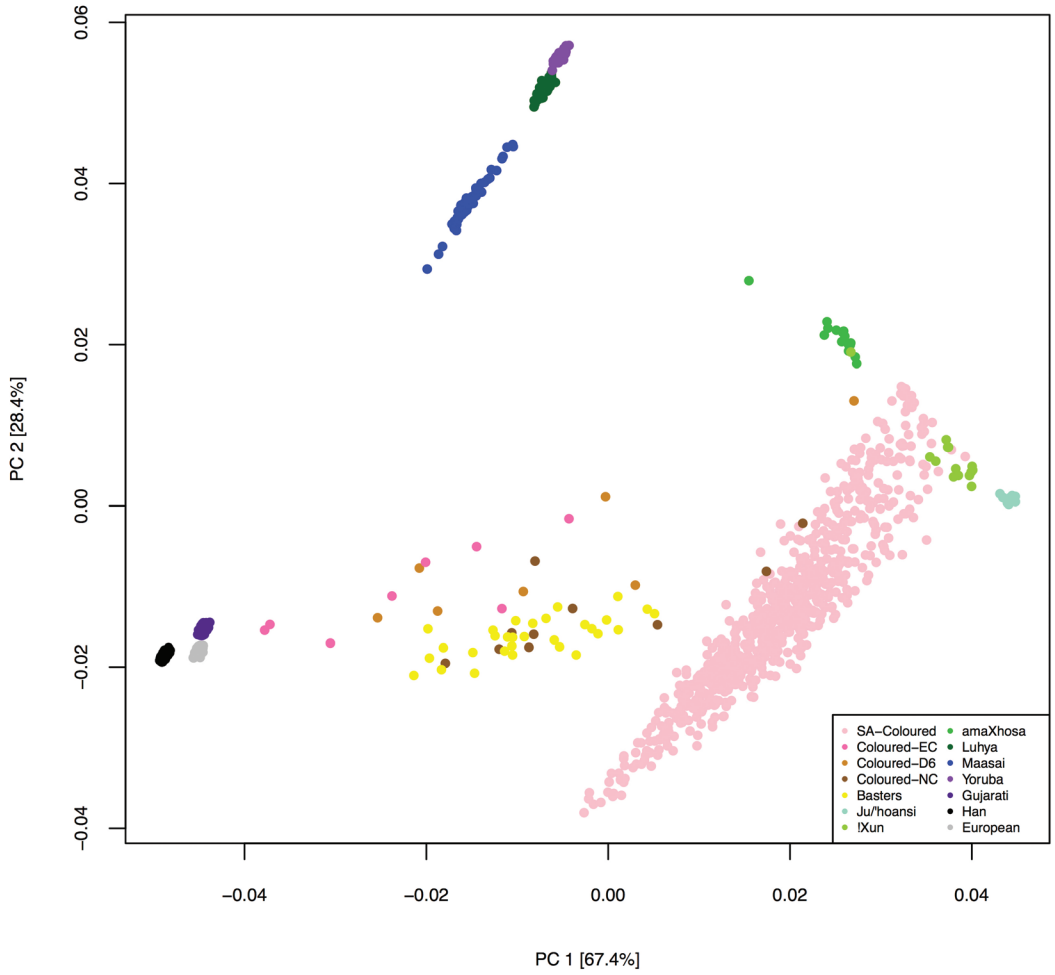

Figure 7: Principal component analysis (PCA) of SAC and populations selected as putative ancestral populations (C. Uren)

A small set of AIMs optimised to distinguish between the five source populations of the SAC population (African San, African non-San, European, South Asian, and East Asian) would enable researchers cost-effectively to reduce false positive findings resulting from ignoring admixture in genetic association studies. Using genome-wide data to find SNPs with large allele frequency differences between the source populations of the SAC, we therefore developed a panel of AIMs. We found that a panel of 96 AIMs can be used to assess ancestry proportions and to adjust for the confounding effect of the complex five-way admixture (Daya et al., 2013). The distribution of the proportions estimated using AIMs was comparable to the distribution of proportions from genome-wide data for all the groups.

We then utilised these AIMs to validate the previous finding of ancestry-specific susceptibility from the GWAS, which had only 91 controls. We genotyped a panel of AIMs in additional individuals, yielding a data set of 918 cases and 507 controls. Using logistic regression models, we confirmed the substantial effect of ancestry (Daya et al., 2014a), and showed that African San ancestry increases susceptibility to TB. African non-San ancestry may also increase TB susceptibility, although the association we found was relatively weak. European and Asian ancestry appears to be protective, but this result may simply reflect lower African ancestry, which would decrease susceptibility.

We also investigated the effect of adjusting for ancestry in candidate gene TB association studies of the SAC, and demonstrated that association results are likely to be affected by adjustment for ancestry if allele frequencies differ markedly in the source populations of the SAC.

ADMIXTURE MAPPING II

The admixed South African Coloured population is ideally suited to the discovery of tuberculosis susceptibility genetic variants and their probable ethnic origins, but previous attempts at finding such variants using genome-wide admixture mapping were hampered by the inaccuracy of local ancestry inference (LAI). When admixture occurs between two or more population

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

A novel algorithm for LAI, implemented in RFMix, was used to identify regions of excess San or Bantu ancestry, which we hypothesised may harbour TB susceptibility genes. San and Bantu ancestries are of particular interest as southern African populations were not exposed to modern strains of *M. tuberculosis*, the most prevalent in our SAC study group (Van der Spuy et al., 2009), until the recent past (Dubos and Dubos, 1952). The relative lack of exposure of the SAC and Bantu populations to modern strains of *M. tuberculosis* could possibly have resulted in decreased resistance to developing the disease, especially in densely populated areas and poor socioeconomic conditions.

Genetic regions with either excess San ancestry or excess African (San or Bantu) ancestry could be relatively accurately inferred, and we identified such regions – in cases but not in controls. A number of promising regions were found, including chromosomes 15q15 and 17q22, which are close to genomic regions previously implicated in TB. Promising immune-related susceptibility genes such as the *GADD45A*, *OSM* and *B7-H5* genes are also harboured in the identified regions (Daya et al., 2014b). Chromosome 15q may contain an African TB susceptibility locus, as it was previously identified as containing TB susceptibility genes in a linkage study using African families (Bellamy et al., 2000). Chromosome 17q22 also showed evidence of having excess African ancestry in TB cases. It is known that chromosome 17q11-q21 may contain a cluster of susceptibility genes for diseases caused by intramacrophage pathogens, such as *M. tuberculosis* (Möller et al., 2009).

Although the SAC population is exceptionally complex, we were able to develop the techniques necessary to utilise the genetic make-up of a five-way admixed population to draw conclusions regarding local ancestry and TB susceptibility.

SOUTH AFRICAN POPULATIONS AND HISTORY

South Africa has been a melting pot of diverse population groups for millennia. The result is that admixture is a fact of life today for all the population groups in this country. A brief, highly simplified history of the country begins with the emergence of anatomically modern *Homo sapiens* about 200 000 years ago, with the consensus suggesting an origin within sub-Saharan Africa. Genetic data indicates that the present-day San or Bushman people represent the oldest human lineage. The Out-of-Africa movement of humans occurred about 70 000 years ago, and genetic estimates suggest a split between north-western and south-eastern Kalahari populations up to 30 000 years ago (Pickrell et al., 2012).

The origin of the Khoe is widely contested (Boonzaaier et al., 1996; Henn et al., 2008). One hypothesis is that there was a migration into the area between 2 000 and 5 000 years ago by pastoralists from East Africa who migrated along the coast and settled in southern Africa, resulting in their influence on the San, and the diffusion of their cattle herding custom into the southern Kalahari (Uren et al., 2016).

The 'Bantu expansion' from Central West Africa (Cameroon) spread eastwards and southwards, resulting in Bantu-speaking farmers moving into southern Africa from the north-east with their cattle from 600 AD, with the settlement of Mapungubwe in 1075 AD. As they moved into the former Natal and Transkei they frequently incorporated the San inhabitants, and the Xhosa have a relatively high Khoesan admixture, evidenced also by clicks in the language.

When the Ottoman Empire blocked the overland route between Europe and India, Portuguese traders looked for a sea route and sailed round the Cape of Good Hope (then named the Cape of Storms) in the late 1400s. These traders were followed by the Dutch and the English. As the weather on the coast of South Africa is notoriously rough, with winds and currents on the aptly named Wild Coast causing the formation of 20 m high waves, many ships were wrecked, with the total averaging one per kilometre. The surviving castaways, European and Indian sailors, and passengers and their slaves were often taken in by the coastal people. Many instances are well-documented, such as a young English girl shipwrecked in about 1740 in Lambasi Bay, who became the Great Wife of the chief of the amaTshomane. Her descendants are still found in East Pondoland. Other groups of male castaways and their

local wives gave rise to the abeLungu and the amaMolo. The wreck of the Grosvenor in 1782 and numerous unnamed wrecks left more castaways who were taken in and stayed (Crampton, 2006).

Arab traders had been plying their trade and settling down the east coast as far as Sofala by 1 000 AD. Later, Delagoa Bay was the source of a highly active trade route that went westwards as far as the Orange River in the vicinity of Keimoes and Kakamas, the site of large Khoe settlements, as evidenced by the trade goods Dutch travellers found there when they got to the Orange River and, later, the headwaters of the Vaal River. These included beads and metal, and techniques for building in stone and for early smallpox vaccination.

In 1652 the occasional visits to the Cape were formalised and the Dutch East India Company set up its refreshment station at the foot of Table Mountain. The people who stayed and their visitors brought their genes and their bacteria – including the 'White Plague'. The social disturbance caused by the aggressive trading of the Dutch at the Cape to supply their ships, and the competition between the indigenous Khoe groups at least partly resulting from this, was already destabilising Khoe society before the devastating 1713 smallpox outbreak, after which the survivors had no power or land. The farmed area settled by the Dutch expanded, and trekboers moved further into the interior. In addition, renegades and deserters from the Cape settlement and runaway slaves traded and raided into the north-west of the country, and effectively caused the breakdown of the Khoe societies, even those settled along the Orange River (Crampton, 2015). The Korana Wars in the 1870s completed the destruction.

After 1750 Delagoa Bay was a major trading centre, and the source of trade-initiated hostilities between local chiefs, which destabilised the eastern part of South Africa. This, together with other factors including a very severe drought, contributed to weakening of societies, which was exacerbated by the Zulu king Shaka, who probably took advantage of the situation, and the subsequent warfare brought on the Mfecane, resulting in hundreds of thousands of deaths from war, displacement and starvation as various generals led raids south into the Transkei area in the 1820s, using the scorched earth policy, and north into Zimbabwe, and vast areas were depopulated. This was followed by numerous Frontier Wars in the Eastern Cape, and further population movement.

On the whole, the manner of dispersal of southern African Bantu-speakers is largely unknown and it has been hypothesised that there is not much population structure between individual populations. It is very difficult to reconstruct society before the 1820s, but it is certainly true to say that virtually all South African populations today are admixed to some extent.

IN AND OUT OF AFRICA: HUMANS AND TB

The South African population is unique and our situation with TB is unique. Ideally, even vaccines (Van Helden and Hoal, 2013) should be designed for a particular population, taking into account the unique combination of genomes of the host and the prevalent pathogen. We have both a very diverse population and an extraordinary variety of *M. tuberculosis* strains. Understanding and managing the interaction between bacterial and host genes is the key to slowing the TB epidemic, and we are in one of the hot spots on earth where our research can make a disproportionate difference.

REFERENCES

Adams, L.A., Möller, M., Nebel, A., Schreiber, S., Van der Merwe, L., Van Helden, P.D. and Hoal, E.G. (2011). Polymorphisms in MC3R promoter and CTSZ 3'UTR are associated with tuberculosis susceptibility. Eur. J. Hum. Genet. *19*, 676–681.

Bellamy, R., Beyers, N., McAdam, K.P., Ruwende, C., Gie, R., Samaai, P., Bester, D., Meyer, M., Corrah, T., Collin, M., et al. (2000). Genetic susceptibility to tuberculosis in Africans: A genome-wide scan. Proc. Natl. Acad. Sci. U.S.A. *97*, 8005–8009.

Boonzaaier, E., Malherbe, C., Smith, A. and Berens, P. (1996). The Cape Herders: A History of the Khoikhoi of Southern Africa (David Philip Publishers, Cape Town).

Carmona, J., Cruz, A., Moreira-Teixeira, L., Sousa, C., Sousa, J., Osorio, N.S., Saraiva, A.L., Svenson, S., Kallenius, G., Pedrosa, J., et al. (2013). *Mycobacterium tuberculosis* strains are differentially recognized by TLRs with an impact on the immune response. PloS One *8*, e67277.

Chimusa, E.R., Daya, M., Möller, M., Ramesar, R., Henn, B.M., Van Helden, P.D., Mulder, N.J. and Hoal, E.G. (2013). Determining ancestry proportions in complex admixture scenarios in South Africa using a novel proxy ancestry selection method. PloS One *8*, e73971.

Chimusa, E.R., Zaitlen, N., Daya, M., Möller, M., Van Helden, P.D., Mulder, N.J., Price, A.L. and Hoal, E.G. (2014). Genome-wide association study of ancestry-specific TB risk in the South African Coloured population. Hum. Mol. Genet. *23*, 796–809.

Cobat, A., Gallant, C.J., Simkin, L., Black, G.F., Stanley, K., Hughes, J., Doherty, T.M., Hanekom, W.A., Eley, B., Beyers, N., et al. (2010). High heritability of anti-mycobacterial immunity in a hyper-endemic area for tuberculosis disease. J.Infect.Dis *201*, 15–19.

Cobat, A., Gallant, C.J., Simkin, L., Black, G.F., Stanley, K., Hughes, J., Doherty, T.M., Hanekom, W.A., Eley, B., Jais, J.P., et al. (2009). Two loci control tuberculin skin test reactivity in an area hyperendemic for tuberculosis. J.Exp.Med. *206*, 2583–2591.

Cobat, A., Hoal, E.G., Gallant, C.J., Simkin, L., Black, G.F., Stanley, K., Jaïs, J.-P., Yu, T.-H., Boland-Auge, A., Grange, G., et al. (2013). 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. *57*, 963–970.

Cobat, A., Poirier, C., Hoal, E., Boland-Auge, A., De La Rocque, F., Corrard, F., Grange, G., Migaud, M., Bustamante, J., Boisson-Dupuis, S., et al. (2015). Tuberculin skin test negativity is under tight genetic control of chromosomal region 11p14-15 in settings with different tuberculosis endemicities. J. Infect. Dis. *211*, 317–321.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., et al. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature *393*, 537–544.

Cooke, G.S., Campbell, S.J., Bennett, S., Lienhardt, C., McAdam, K.P., Sirugo, G., Sow, O., Gustafson, P., Mwangulu, F., Van Helden, P., et al. (2008). Mapping of a novel susceptibility locus suggests a role for MC3R and CTSZ in human tuberculosis. Am J Respir Crit Care Med *178*, 203–207.

Cooper, M.A., Fehniger, T.A., and Caligiuri, M.A. (2001). The biology of human natural killer-cell subsets. Trends Immunol. *22*, 633–640.

Crampton, H. (2006). The Sunburnt Queen (Saqi Books, Johannesburg).

Crampton, H. (2015). The Side of the Sun at Noon (Jacana Media, Johannesburg).

Dale, E., Davis, M. and Faustman, D.L. (2006). A role for transcription factor NF-kappaB in autoimmunity: Possible interactions of genes, sex, and the immune response. Adv. Physiol. Educ. *30*, 152–158.

Davila, S., Hibberd, M.L., Hari, D.R., Wong, H.E., Sahiratmadja, E., Bonnard, C., Alisjahbana, B., Szeszko, J.S., Balabanova, Y., Drobniewski, F., et al. (2008). Genetic association and expression studies indicate a role of toll-like receptor 8 in pulmonary tuberculosis. PLoS.Genet. *4*, e1000218.

Davila, S., Wright, V.J., Khor, C.C., Sim, K.S., Binder, A., Breunis, W.B., Inwald, D., Nadel, S., Betts, H., Carrol, E.D., et al. (2010). Genome-wide association study identifies variants in the CFH region associated with host susceptibility to meningococcal disease. Nat. Genet. *42*, 772–776.

Daya, M., Van der Merwe, L., Galal, U., Möller, M., Salie, M., Chimusa, E.R., Galanter, J.M., Van Helden, P.D., Henn, B.M., Gignoux, C.R., et al. (2013). A panel of ancestry informative markers for the complex five-way admixed South African Coloured population. PLoS ONE *8*, e82224.

Daya, M., Van der Merwe, L., Van Helden, P.D., Möller, M. and Hoal, E.G. (2014a). The role of ancestry in TB susceptibility of an admixed South African population. Tuberc. Edinb. Scotl. *94*, 413–420.

Daya, M., Van der Merwe, L., Gignoux, C.R., Van Helden, P.D., Möller, M. and Hoal, E.G. (2014b). Using multi-way admixture mapping to elucidate TB susceptibility in the South African Coloured population. BMC Genomics *15*, 1021.

Daya, M., Van der Merwe, L., Van Helden, P.D., Möller, M. and Hoal, E.G. (2015). Investigating the role of gene-gene interactions in TB susceptibility. PLoS ONE *10*.

De Wit, E., Van der Merwe, L., Van Helden, P.D. and Hoal, E.G. (2010a). Gene-gene interaction between tuberculosis candidate genes in a South African population. Mamm. Genome 1–11.

De Wit, E., Delport, W., Rugamika, C.E., Meintjes, A., Moller, M., Van Helden, P.D., Seoighe, C. and Hoal, E.G. (2010b). Genome-wide analysis of the structure of the South African Coloured population in the Western Cape. Hum. Genet. *128*, 145–153.

Dubos, R., and Dubos, J. (1952). The White Plague: Tuberculosis, Man and Society (Little, Brown & Co., Boston).

Gagneux, S., Deriemer, K., Van, T., Kato-Maeda, M., de Jong, B.C., Narayanan, S., Nicol, M., Niemann, S., Kremer, K., Gutierrez, M.C., et al. (2006). Variable host-pathogen compatibility in *Mycobacterium tuberculosis*. Proc. Natl. Acad. Sci. U.S.A *103*, 2869–2873.

Gallant, C.J., Cobat, A., Simkin, L., Black, G.F., Stanley, K., Hughes, J., Doherty, T.M., Hanekom, W.A., Eley, B., Beyers, N., et al. (2009). Tuberculin skin test and in-vitro assays provide complementary measures of anti-mycobacterial immunity in children and adolescents. Chest *DOI:10.1378/chest.09-1852*.

Hanekom, M., Van der Spuy, G.D., Gey van Pittius, N.C., McEvoy, C.R.E., Ndabambi, S.L., Victor, T.C., Hoal, E.G., Van Helden, P.D. and Warren, R.M. (2007). Evidence that the spread of *Mycobacterium tuberculosis* strains with the Beijing genotype is human population dependent. J. Clin. Microbiol. *45*, 2263–2266.

Health Systems Trust (2009). Incidence of TB in the Provinces of South Africa.

Henn, B.M., Gignoux, C., Lin, A.A., Oefner, P.J., Shen, P., Scozzari, R., Cruciani, F., Tishkoff, S.A., Mountain, J.L. and Underhill, P.A. (2008). Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. Proc. Natl. Acad. Sci. *105*, 10693–10698.

Hershberg, R., Lipatov, M., Small, P.M., Sheffer, H., Niemann, S., Homolka, S., Roach, J.C., Kremer, K., Petrov, D.A., Feldman, M.W., et al. (2008). High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol. *6*, e311.

Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat. Rev. Genet. *6*, 95–108.

Hoal, E. (2011). Genome-sequencing anniversary. Famine in the presence of the genomic data feast. Science *331*, 874.

Keegan, T. (1996). Colonial South Africa and the Origins of the Racial Order (David Philip Publishers, Claremont).

Lander, E.S., and Schork, N.J. (1994). Genetic dissection of complex traits. Science *265*, 2037–2048.

Le Roex, N., Noyes, H., Brass, A., Bradley, D.G., Kemp, S.J., Kay, S., Van Helden, P.D. and Hoal, E.G. (2012). Novel SNP discovery in African buffalo, *Syncerus caffer*, using high-throughput sequencing. PloS One *7*, e48792.

Le Roex, N., Koets, A.P., Van Helden, P.D. and Hoal, E.G. (2013). Gene polymorphisms in African buffalo associated with susceptibility to bovine tuberculosis infection. PloS One *8*, e64494.

Migueles, S.A. and Connors, M. (2010). Long-term nonprogressive disease among untreated HIV-infected individuals: Clinical implications of understanding immune control of HIV. J. Am. Med. Assoc. *304*, 194–201.

Migueles, S.A., Weeks, K.A., Nou, E., Berkley, A.M., Rood, J.E., Osborne, C.M., Hallahan, C.W., Cogliano-Shutta, N.A., Metcalf, J.A., McLaughlin, M., et al. (2009). Defective human immunodeficiency virus-specific CD8+ T-cell polyfunctionality, proliferation, and cytotoxicity are not restored by antiretroviral therapy. J. Virol. *83*, 11876–11889.

Misch, E.A. and Hawn, T.R. (2008). Toll-like receptor polymorphisms and susceptibility to human disease. Clin. Sci. Lond. Engl. 1979 *114*, 347–360.

Möller, M., De Wit, E. and Hoal, E.G. (2010). Past, present and future directions in human genetic susceptibility to tuberculosis. FEMS Immunol Med Microbiol *58*, 3–26.

Möller, M. and Hoal, E.G. (2010). Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. Tuberculosis.(Edinb.) *90*, 71–83.

Möller, M., Nebel, A., Valentonyte, R., Van Helden, P.D., Schreiber, S. and Hoal, E.G. (2009). Investigation of chromosome 17 candidate genes in susceptibility to TB in a South African population. Tuberculosis.(Edinb.) *89*, 189–194.

Motulsky, A.G. (1960). Metabolic polymorphisms and the role of infectious diseases in human evolution. Hum. Biol. *32*, 28–62.

Mountain, A. (2003). The First People of the Cape (David Philips Publishers, Claremont).

Murphy, P.M. (1993). Molecular mimicry and the generation of host defense protein diversity. Cell *72*, 823–826.

Nhamoyebonde, S. and Leslie, A. (2014). Biological differences between the sexes and susceptibility to tuberculosis. J. Infect. Dis. *209 Suppl 3*, S100–S106.

Pickrell, J.K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Güldemann, T., Kure, B., Mpoloka, S.W., Nakagawa, H., Naumann, C., et al. (2012). The genetic prehistory of southern Africa. Nat. Commun. *3*, 1143.

Portevin, D., Via, L.E., Eum, S. and Young, D. (2012). Natural killer cells are recruited during pulmonary tuberculosis and their ex vivo responses to mycobacteria vary between healthy human donors in association with KIR haplotype. Cell. Microbiol. *14*, 1734–1744.

Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Zaporozhchenko, V., Bormans, C., Van Helden, P.D., Hoal, E.G. and Behar, D.M. (2010). Strong maternal Khoisan contribution to the South African coloured population: A case of gender-biased admixture. Am. J. Hum. Genet. *86*, 611–620.

Risch, N., and Merikangas, K. (1996). The future of genetic studies of complex human diseases. Science *273*, 1516–1517.

Rose, D.N., Schechter, C.B. and Adler, J.J. (1995). Interpretation of the tuberculin skin test. J. Gen. Intern. Med. *10*, 635–642.

Salie, M., Daya, M., Lucas, L.A., Warren, R.M., Van der Spuy, G.D., Van Helden, P.D., Hoal, E.G. and Möller, M. (2015a). Association of toll-like receptors with susceptibility to tuberculosis suggests sex-specific effects of TLR8 polymorphisms. Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis. *34*, 221–229.

Salie, M., Daya, M., Möller, M. and Hoal, E.G. (2015b). Activating KIRs alter susceptibility to pulmonary tuberculosis in a South African population. Tuberc. Edinb. Scotl. *95*, 817–821.

Salie, M., Van der Merwe, L., Möller, M., Daya, M., Van der Spuy, G.D., Van Helden, P.D., Martin, M.P., Gao, X., Warren, R.M., Carrington, M., et al. (2014). Associations between human leukocyte antigen class I variants and the *Mycobacterium tuberculosis* subtypes causing disease. J. Infect. Dis. *209*, 216–223.

Seldin, M.F., Pasaniuc, B. and Price, A.L. (2011). New approaches to disease mapping in admixed populations. Nat. Rev. Genet. *12*, 523–528.

Serpa, J.A., Teeter, L.D., Musser, J.M. and Graviss, E.A. (2009). Tuberculosis disparity between US-born blacks and whites, Houston, Texas, USA. Emerg. Infect. Dis. *15*, 899–904.

Shell, R. (1994). Children of Bondage (Witwatersrand University Press, Johannesburg).

Sigal, L.J., Crotty, S., Andino, R. and Rock, K.L. (1999). Cytotoxic T-cell immunity to virus-infected non-haematopoietic cells requires presentation of exogenous antigen. Nature *398*, 77–80.

Sorensen, T.I., Nielsen, G.G., Andersen, P.K. and Teasdale, T.W. (1988). Genetic and environmental influences on premature death in adult adoptees. N. Engl. J. Med. *318*, 727–732.

Spurgin, L.G. and Richardson, D.S. (2010). How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. Proc. Biol. Sci. *277*, 979–988.

Stead, W.W., Senner, J.W., Reddick, W.T. and Lofgren, J.P. (1990). Racial differences in susceptibility to infection by *Mycobacterium tuberculosis*. N. Engl. J. Med. *322*, 422–427.

Stein, C.M., Guwatudde, D., Nakakeeto, M., Peters, P., Elston, R.C., Tiwari, H.K., Mugerwa, R. and Whalen, C.C. (2003). Heritability analysis of cytokines as intermediate phenotypes of tuberculosis. J. Infect. Dis. *187*, 1679–1685.

Stein, C.M., Zalwango, S., Malone, L.L., Won, S., Mayanja-Kizza, H., Mugerwa, R.D., Leontiev, D.V., Thompson, C.L., Cartier, K.C., Elston, R.C., et al. (2008). Genome scan of *M. tuberculosis* infection and disease in Ugandans. PLoS. ONE. *3*, e4094.

Thye, T., Owusu-Dabo, E., Vannberg, F.O., Van Crevel, R., Curtis, J., Sahiratmadja, E., Balabanova, Y., Ehmen, C., Muntau, B., Ruge, G., et al. (2012). Common variants at 11p13 are associated with susceptibility to tuberculosis. Nat. Genet. *44*, 257–259.

Thye, T., Vannberg, F.O., Wong, S.H., Owusu-Dabo, E., Osei, I., Gyapong, J., Sirugo, G., Sisay-Joof, F., Enimil, A., Chinbuah, M.A., et al. (2010). Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. Nat. Genet. *42*, 739–741.

Uren, C., Kim, M., Martin, A.R., Bobo, D., Gignoux, C.R., Van Helden, P.D., M□ller, M., Hoal, E.G. and Henn, B.M. (2016). Fine-scale human population structure in southern Africa reflects ecological boundaries. BioRxiv doi: http://dx.doi.org/10.1101/038729

Van der Spuy, G.D., Kremer, K., Ndabambi, S.L., Beyers, N., Dunbar, R., Marais, B.J., Van Helden, P.D. and Warren, R.M. (2009). Changing *Mycobacterium tuberculosis* population highlights clade-specific pathogenic characteristics. Tuberc. Edinb *89*, 120–125.

Van Helden, P.D. and Hoal, E.G. (2013). A new TB vaccine: Fact or fiction? Comp. Immunol. Microbiol. Infect. Dis. *36*, 287–294.

Van Lunzen, J. and Altfeld, M. (2014). Sex differences in infectious diseases –Common but neglected. J. Infect. Dis. *209 Suppl 3*, S79–S80.