Mapping Genes Underlying Ethnic Differences in Tuberculosis Risk by Linkage Disequilibrium in the South African Coloured Population of the Western Cape

Emile Chimusa Rugamika (emile@cbio.uct.ac.za)

Department of Molecular and Cell Biology

University of Cape Town, South Africa

A thesis submitted for the degree of

Doctor of Philosophy in Computational Biology

Supervised by: Prof. **Nicola Mulder**, University of Cape Town, SA

Co-supervised by: Ass. Prof. **Alkes Price**, Harvard School Public Health, USA

Co-supervised by: Prof. **Eileen Hoal van Helden**, University of Stellenbosch, SA

11^{th} *of February, 2013*

I am greatly indebted to my supervisor Prof. Nicola Mulder, who gave me all the possible support, I needed to carry on my PhD study. I am grateful to my co-supervisors Prof. Alkes Price and Prof. Eileen Hoal van Helden who were enough courageous to co-supervise my PhD research and allowed together to slowly plow way through this work. From them, I learnt the importance of expressing ideas clearly, both verbally and in writing.

I would like to thank my parents and my little family -Annie and Imani Emilson and Wivina Emilson, for their faith in me and their love and support, without which none of this would ever have come to pass.

To God be all the Glory.

Thanks

Acknowledgements

I am grateful to all South African Coloured subjects who participated in this research project and would like to thank them for their contributed blood and saliva samples for DNA extraction. During my studies I was supported by the Carnegie Corporation and the National Research Foundation. Travel grants from the University of Cape Town and Carnegie Corporation allowed me to present some of this work at an international conference and to work with my co-supervisor at Harvard School of Public Health. This research was supported by grants awarded to me by the Carnegie Cooperation, University of Cape Town, Clinical Laboratory Sciences Department, Medical School. My sincere appreciation goes to associate Professor Nicola Mulder, my supervisor, for her assistance and guidance throughout this study, and for reading several drafts of this thesis. I wish to express my sincere gratitude to both my co-supervisors Prof. Alkes Price, Harvard School of Public Health and Prof. Eileen Hoal van Helden, DST/NRF Centre of Excellence for Biomedical TB Research, Department of Biomedical Sciences Faculty of Health Sciences, University of Stellenbosch. In addition, I am thankful to Stokes Prof. Cathal Seoighe and Assistant Prof. Noah Zaitlen for strategic informations and for helpful discussions during my PhD study. I am also thankful to Lynne Teixeira at African Institute for Mathematical Sciences (AIMS) for her assistance in reading this thesis.

Finally, I would like to express my deepest gratitude to my parents, family and friends, for their constant support and encouragement. Most of all, my appreciation goes to Imani Emilson, Wivina Emilson and Makasawa Mpangi who have always been an incredible source of help, love and encouragement throughout the years.

Abstract

The South Africa Coloured population of the Western Cape is the result of unions between Europeans, Africans (Bantu and Khoisan), and various other populations (Malaysian or Indonesian descent). The world-wide burden of tuberculosis remains an enormous problem, and is particularly severe in this population. In general, admixed populations that have arisen in historical times can make an important contribution to the discovery of disease susceptibility genes if the parental populations exhibit substantial variation in susceptibility. Despite numerous successful genome-wide association studies, detecting variants that have low disease risk still poses a challenge. Furthermore, admixture association studies for multi-way admixed populations pose constant challenges, including the choice of an accurate ancestral panel to infer ancestry and for imputing missing genotypes to identify possible genetic variants causing susceptibility to disease. This thesis addresses some of these challenges. We first developed PROXYANC, an approach to select the best proxy ancestral populations for admixed populations. From the simulation of a multi-way admixed population, we demonstrated the ability and accuracy of PROXYANC in selecting the best proxy ancestry and illustrated the importance of the choice of ancestries in both estimating admixture proportions and imputing missing genotypes. We applied this approach to the South African Coloured population, to refine both the choice of ancestral populations and their genetic contributions. We also demonstrated that the ancestral allele frequency differences correlated with increased linkage disequilibrium in the SAC, and that the increased LD originates from admixture events rather than population bottlenecks. Secondly, we conducted a study to determine whether ancestry-specific genetic contributions affect tuberculosis risk. We additionally conducted imputation genome-wide association studies and a meta-analysis incorporating previous genome-wide association studies of tuberculosis. Our results demonstrated significant evidence of an association (odds ratio $=1.46$, p $= 1.58e^{-05}$) between ‡Khonami (Khoisan) ancestry and tuberculosis risk that is not due to confounding by socio-economic status, and confirmed a previously identified susceptibility locus ($rs2057178$: odds ratio $= 0.62$, p $= 2.71e-06$). This provides insights into identifying disease genes and ancestry-specific disease risk in multi-way admixed populations. Because of the importance of inference of locus-specific ancestry in understanding both population history and disease scoring statistics, and in identifying the most

significant gene or pathway underlying ethnic difference in complex diseases risk, we thirdly, assessed the accuracy of current approaches to estimate local ancestry in a multi-way admixed population. Our result demonstrated the limitation of the accuracy of these methods in inferring local ancestry and highlighted the need for developing a method of accurately inferring the local ancestry along the genome of multi-way admixed individuals, which in turn may complement the disease scoring statistics and be informative in fine mapping methods for diseases for which risk differs depending on ancestry. Finally, to fully characterize the susceptibility genes in multi-way admixed populations, this work introduced an algebraic graph-based method (ancGWAS) to identify significant sub-networks underlying ethnic differences in complex disease risk in a recently admixed population by integrating the association signal from standard Genome-wide Association Study data sets, the locus-specific ancestry and pair-wise linkage disequilibrium into the human protein-protein interaction network. Through simulation of interactive disease loci in the simulation of a 4-way admixed population, we demonstrated that ancGWAS holds promise for comprehensively examining the interactions between genes underlying the pathogenesis of complex diseases and also for identifying possible signals of unusual differences in excess/deficiency of ancestry at the gene and pathway levels. We applied this approach to the imputed genome-wide association study data set of TB in the admixed South Coloured population. We were able to refine the association signal of 6 genes, including MEGF10 ($p = 2.44e-11$), PRRC1 ($p = 2.44e-11$), HNRNPK ($p = 6.28e-09$), SLC8A3 ($p = 8.99e-09$), SMOC1 ($p = 8.99e-09$) and CTXN3 ($p = 2.30e-08$). In addition, our result replicated 4 known TB associated genes, which include IL8 ($p = 0.0039$), SLC11A1 ($p = 0.0035$), WT1 ($p = 0.0015$), CCL2 ($p = 0.0015$) and IFNGR1 ($p = 0.0034$). We identified a novel central sub-network that is mostly implicated in acute and chronic myeloid leukemia signaling pathways, and includes the WT1 and IL8 genes. This result provides further insights into tuberculosis pathogenesis and is potentially relevant for further biomedical research in this field.

Contents

List of Figures

List of Tables

Chapter 1

Introduction, Background and Literature Review

1.1 Introduction

1.1.1 Population Diversity in South Africa

An extensive population diversity with groups originating from ancestral African (79%), Asian (2.5%) and European (9.6%) populations is found in South Africa (deWit *et al.*, 2010a). As reported in (deWit *et al.*, 2010a; Mountain, 2003), both multiple colonization history and South Africa's location with respect to major trade routes from the 15^{th} to the 19^{th} century are consequences of population diversity in South Africa. The contribution of these previously continentally divided population groups from Europe, Asia and the rest of Africa, to South Africa's diversity resulted in the establishment of a mixed ancestry population, based mainly in the Western Cape Province self-identifying as the South African Coloured population (SAC) (Adhikari, 2005; Nurse *et al.*, 1985; Ross, 1993). This population, which presently comprises approximately 54% of the population of the Western Cape province and 9% of the entire South African population, has a complex genetic history influenced by historical legislation.

The South African Coloureds have part of their roots in the indigenous Khoekhoen and San (Boonzaaier *et al.*, 1996; Elphick, 1985; Mountain, 2003), the former being native to a large area comprising the south-western parts of Africa including the current Western Cape Province of South Africa. During early colonization by European settlers of the Dutch East India Company (VOC) in 1652 (Davis & Dollard, 1994; Mountain, 2003), a refreshment station was established at the Cape of Good Hope, now Cape Town, and the company brought slaves from the Indian subcontinent (25.9%), and small numbers of political exiles from Indonesia and Malaysia (Mountain, 2004), the east coast of Africa (26.4%), Madagascar (25.1%) and Indonesia (22.7%) (Davis &

Dollard, 1994; Nurse *et al.*, 1985). These estimations were obtained from the records of the slave trade (Davis & Dollard, 1994). The San, in particular, were the original inhabitants of Southern Africa and one of the last remaining hunter-gatherer societies. Khoekhoen pastoralists apparently arrived in Southern Africa shortly before the Bantu (Mountain, 2004). Over time, some Khoi abandoned pastoralism and adopted the hunter-gatherer economy of the San, likely due to a drying climate, and are now considered San. Therefore the name Khoesan was introduced to name both Khoekhoen and San populations. The indigenous Khoekhoen and San were not enslaved, but frequently served as indentured labourers or serfs on the farms (Davis & Dollard, 1994; Mountain, 2003). Women from Khoekhoen or slave descent and their children were integrated into the colonial household, often by marriage (Davis & Dollard, 1994; Mountain, 2003). Mixed marriages, usually between European men and women who were either Khoekhoen, San, slaves or of mixed parentage (Keegan, 1996), and between Khoekhoen, San and slave (Mountain, 2003) were not socially forbidden in the early Cape community. Since 1700, the progeny of mixed marriages and liaisons gradually grew into a group known as the "Cape Coloured's" (Keegan, 1996; Mountain, 2003; Nurse *et al.*, 1985). The name of "Cape Coloured's" population was introduced in the mid nineteenth century (Keegan, 1996). Furthermore, these intermarriages were more common in the farming areas (Davis & Dollard, 1994; Mountain, 2003), and later on after 1806, race-based restrictions were formalised under the British administration (Mountain, 2003). Therefore, both the legislature introduced during the apartheid era $(1948 - 1994)$ and the establishment of missionary stations (from 1738) strengthen cohesion among Coloured and Khoesan populations (Mountain, 2004).

After emancipation by the British administration $(1834 - 1838)$, many ex-slaves and other indigent people settled at mission stations (Mountain, 2004), some of which formed the kernel of a "Coloured" group area (Boonzaaier *et al.*, 1996; Mountain, 2003). Many of the Khoesan at these mission stations had European or African (particularly Xhosa) ancestry (Keegan, 1996). The formalization of the racial order in society began in the late 1700's. From 1910, and particularly $1948 - 1994$, the apartheid regime introduced legislature that outlawed inter-racial marriage and predefined areas of residence (http://www.sahistory.org.za/pages/chronology/special-chrono/governance/apartheid-legislation.html). This separation of ethnic groups increased cohesion of the already established admixed SAC population in the Western Cape (Adhikari, 2005; Cilliers, 1985). In the Western Cape, 17.6% of the South African Coloureds are English-speaking, and 83.0% are Afrikaans-speaking, these figures are according to the 2011 South African's census.

1.2 Motivation and Thesis Overview

The South African Coloured individuals presented in this thesis were enrolled from Ravensmead and Uitsig, two suburbs of Cape Town, 90.1% are Afrikaans-speaking and 9.3% are English-speaking. The population of Ravensmead/Uitsig is 90.5% Christian, and only 1.7% Muslim (2011 SA census). Importantly, this mixed population has the highest incidence of tuberculosis (TB) in sub-Saharan of Africa. In addition, recent investigations indicated that tuberculosis frequently occurs in many members of the same family of the mixed Coloured population and therefore, heritable factors, including environmental and migration factors maybe involved in determining susceptibility and resistance to active tuberculosis after infection (Babb *et al.*, 2007; Hoal *et al.*, 2004). It is of interest to establish whether differences in tuberculosis risk are likely to have a genetic basis, to show the relationship between risk of tuberculosis and proportion of admixture from the higher-risk ancestral population of this admixed population and to identify possible genetic variants causing susceptibility to tuberculosis.

The central premise of this thesis is concerned with identifying the genomic loci of the South African Coloured population with possible evidence of ethnic difference and association with tuberculosis risk. This thesis aims to utilize the effects of admixture to map genes that underlie differences in tuberculosis risk based on case/control data. It aims to establish whether ancestry differences in tuberculosis risk are likely to have a genetic basis and to show the relationship between risk of tuberculosis and proportion of admixture from the higher-risk ancestral population of the Coloureds. The admixture association methods utilize the latent ancestry states in recent admixed populations at the putative genetic disease locus and tests for genetic linkage by detecting association of the genetic locus ancestry with the disease. Incorporating admixture association signals into GWAS of admixed populations has been shown to likely be informative for diseases for which risk differs depending on ancestry (Pasaniuc *et al.*, 2011). The first aim in this project was therefore to understand the genetic make-up of this population by developing approaches to examine the fine genetic characterization of ancestral components, and lastly to assess the accuracy of current approaches to estimate local ancestry in multi-way admixed populations and provide an application of local ancestry in identifying significant gene or pathway underlying ethnic difference in complex diseases risk in multi-way admixed populations, particularly in the South African Coloured population. This thesis involves six main axes of investigation:

(1) Fine Characterization of genetic ancestry of this population by developing an approach of accurately selecting the best proxy ancestral populations for a multi-way admixed population.

(2) The examination of whether the genetic contribution can increase tuberculosis incidence, and the evaluation of the contribution of socio-economic status to the ancestry-tuberculosis relationship in the SAC.

(3) Genome-wide Association Study (GWAS) with correction for genome-wide ancestry, accounting for both population stratification and hidden relatedness that can result from the genealogy.

(4) Meta analysis of a combined imputation Genome-wide Association Study of the SAC and a recent studied African TB case-control series from Ghana, Gambia and Malawi, and four polymorphisms in the TLR8 gene on chromosome X.

(5) Assessment of the accuracy of inferring local ancestry on both simulation and real data of the SAC.

(6) Because of the complex nature of the immune system and the polygenic nature of TB, the project aims to develop an algebraic graph-based method (ancGWAS) that incorporates both the association signal from Genome-wide Association Study and the available human protein-protein interaction (PPI) information for testing the combined effects of SNPs and searching for significantly enriched sub-networks associated with complex diseases, and testing for possible signals of difference in excess/deficiency of individual ancestry. Application of ancGWAS method is conducted on the imputation TB GWAS data set of the SAC.

Below is an overview of the chapters of this thesis:

Chapter 2 introduces a novel approach to choose the best proxy ancestry for multi-way admixed populations. This approach searches for a combination of reference populations that can minimize the genetic distance (using F_{ST} as an objective function through an optimal quadratic cone programming algorithm) between the admixed population and all possible synthetic populations, consisting of a linear combination of reference populations. In addition, PROXYANC also computes the proxy-ancestry score by regressing a statistic for LD between a pair of SNPs in the admixed population against a weighted allele frequency differentiation in the non-admixed reference populations. This approach is applied for downstream analysis in a uniquely admixed Coloured population from South Africa (SAC). The African, European, South and East and South Asian origins of the SAC are characterized by applying PROXYANC to a cohort of the SAC (764 unrelated individuals) and we refine both the choice of best ancestral populations and their genetic contributions.

Chapter 3 is concerned with the assessment of whether the genetic make-up and the observed linkage disequilibrium (LD) in the SAC is a result of ancestral admixture or has been influenced by founder effects or population bottlenecks. To address this, panels of ancestry informative markers for the South African Coloured population are considered by implementing two types of algorithms for selecting genetic markers that are differentiated in ancestry.

Chapter 4 examines the relationship between genetic ancestry proportions and TB status in the SAC. As the observed relationship between genetic ancestry and TB status could be a consequence of confounding due to socio-economic status (SES), this possibility are investigated by studying two SES variables, household income and self income. In addition, allele frequency differences between the SAC control and case individuals at common SNPs is computed based on a chi^2 statistic to search for unusual population differentiation that accounts for the effects of neutral genetic drift.

Chapter 5 analyses Genome Wide Association Study data, with correction for genome-wide ancestry, and accounting for both population stratification and hidden relatedness that can result from the genealogy.

Chapter 6 covers the imputation of unobserved genotypes in the study sample, which has been conducted to increase genome coverage in GWAS conducted in previous chapter. This chapter also covers the meta analysis of a combined genome-wide association study of the SAC and a recent African TB case-control series from Ghana, Gambia and Malawi, including four polymorphisms in the *TLR8* gene on chromosome X.

Chapter 7 covers the inference of local ancestry in a multi-way admixed population, with application to the SAC. The accuracy of inferring local ancestry on both simulation and real data of the SAC are assessed, and possible approaches to estimate the date of multi-way admixture events are also discussed.

Chapter 8 introduces an algebraic-graph algorithm to examine the association signal from a combined genome-wide SNP case-control and admixture analysis and the available human protein-protein interaction (PPI) information for testing the combined effects of SNPs. It searches for both significant genes and enriched sub-networks underlying ancestry difference in common disease risk, in particular, TB risk. Finally conclusions and some future directions are covered in the last chapter 9.

The current chapter continues with a review of the relevant literature.

1.3 Population Genetics of Admixture

1.3.1 Human Genetics Diversity

The most important concerns in human population genetics include on understanding the consequence of past human migrations, the causes of human diversity in the world today and the related evolutionary history that generated that diversity, through mathematical modelling of complex patterns of geographic genetic diversity. These patterns of geographic genetic diversity were caused by mutation, natural selection, genetic drift and gene flow that change within and between populations. A number of studies have examined how genetic variation is distributed geographically, and have established that human population differences are mainly due to the presence of low-frequency alleles that have not diffused far from their geographic place of origin. In addition, a recent study by Rosenberg and colleagues demonstrated that the worldwide human genetic variation within human populations is larger $(93-95\%)$ than that seen between populations $(5-7\%)$ (Rosenberg & Pritchard, 2008; Rosenberg *et al.*, 2003), suggesting that classification of the human species according to racial or continental lines appears to be inappropriate descriptors of the distribution of human genetic variation (Tishkoff & Kidd, 2004). The literature was surveyed to quantify the human genetic variation within and between human populations using Wrights F_{ST} statistic (Weir, 2008; Weir & Cockerham, 1984) as follows,

$$F_{ST} = \frac{\sum_{i=1}^{L} p_i^*(1-p_i^*) - \mathrm{F}_i}{\sum_{i=1}^{L} p_i^*(1-p_i^*)} \tag{1.1}$$

where p_i^* is the average allele frequency (over all populations) of the i^{th} allele, L is the number of alleles, and F_i is the value of F_{ST} for each allele, so for two populations we have,

$$\mathrm{F}_i = \frac{\sum_{k=1}^{2}(p_i^k - p_i^*)^2}{p_i^*(1-p_i^*)},$$

where p_i^k is the frequency of the i^{th} allele in population k. Several related measures for understanding the genetic variation, and estimating ancient admixture in human history such as f_4-ratio, 3 Population Test and 4 Population Test (Reich *et al.*, 2009) have been introduced, to account for closely related and admixed populations.

Consider a bi-allelic marker j in two given populations to be in Hardy-Weinberg equilibrium, respectively. Let variant alleles, b_1, and b_2 have population frequency p_1 and p_2 in populations 1 and 2 respectively. Setting $q_i = 1 - p_i$, for i = 1, 2 . An other measure of divergence (Pickrell *et al.*, 2012; Reich *et al.*, 2009) at a given locus j is given by,

$$F_{ST}^j = \frac{p_1(q_2-q_1) + p_2(q_1-q_2)}{p_1q_2 + q_1p_2}. \tag{1.2}$$

Let S be a set of markers m_j, $(j = 1, ...M)$, then we define population pair-wise Wrights F_{ST} by averaging the equation 1.2 over all the markers $(j = 1, ...M)$.

The 3 population Test $(f_3 - statistic)$ is utilised for testing whether a particular population has inherited a mixture of ancestries; while the 4 Population Test is a more sensitive test for detecting the admixture in populations, though it is highly model-based and a positive signal is more difficult to interpret (Reich *et al.*, 2009).

Let consider d, l, c as the allele frequencies in different populations D, L, C, respectively, at a single polymorphism (Patterson *et al.*, 2012). Assuming the population C was derived from the admixture of D and L. It follows the f_3 -statistic is given as,

$$f_3(C; D, L) = \mathbb{E}\left[(c - d)\,(c - l)\right].$$

where $\mathbb{E}$ is the expected value. The allele frequencies dont affect f_3, as choosing the alternate allele simply flips the sign of both terms in the product. Let q denote the allele frequency of a given SNP, and consider e, l, c, the allele frequencies in D, L, C (where $e, l$and c are the alleles frequencies in population D, L and C), where D, L, C are different populations.

Then,

$$\mathbb{E}\left[(c - e)(c - l)\right] = \mathbb{E}\left[(c - x + x - e)(c - x - + x - l)\right] = \mathbb{E}\left[(c - x)^2\right] \geq 0$$

since $\mathbb{E}[e|x] = x$, and $\mathbf{E}[x - l] = \mathbf{E}[q - l - (q - x)] = 0$.

If both D and L are the ancestry of C, then $\mathbf{E}\left((c - e)(c - l)\right)$ will be negative (Patterson *et al.*, 2012). The estimator of $f_3 - statistic$ defined as $(a - b)(a - c)$ with two a

$$q = (a' - b')(a' - c'),$$

$$q = \left((a' - a) - (b' - b) + (a - b)\right)\left((a' - a) - (b' - b) + (a - b)\right).$$

where $\mathbf{E}(a' - a)^2$ is the bias of q.

$$\mathbf{E}(a' - a)^2 = \frac{a(1 - a)}{n_A},$$

where $n_A = \alpha_0 + \alpha_1$, the total allele count for the population A and $h_A = a(1 - a)$. Thus

$$f_3\,(A, B, C) = (a' - b')(a' - c') - \hat{h}_A / n_A.$$

1.3.2 Genetics of Admixture

Throughout human history, contacts between two or more previously isolated populations are mostly due to different population migrations, colonization waves, or forced displacements due to many reasons such as ecology, climate, agriculture and hunting. Most of these human contacts or admixture processes have been influenced by sociocultural laws on inter-marriage in contexts of ethnic conflict or discrimination, slavery, and clan or caste systems. Nevertheless, it has been shown that the mixture of previously isolated populations, results in admixed populations that benefit from several genetic advantages such as increased genetic variation, the creation of novel genotypes and the masking of deleterious mutations (Halder & Shriver, 2003; McKeigue, 2005). These admixture benefits are thought to play an important role in biological invasions (Verhoeven *et al.*, 2010). In addition, population admixture has an important application in assessing patterns of migration and genetic structure (Pritchard *et al.*, 2002), and in detecting natural selection (Lohmueller *et al.*, 2011; Tang *et al.*, 2006). Population admixture provides valuable baseline data for subsequent analysis of disease association, particularly identifying phenotypically relevant genes through admixture-mapping strategies (Halder & Shriver, 2003; McKeigue, 2005; Reich *et al.*, 2005; Seldin *et al.*, 2011; Smith & O'Brien, 2005).

Because of differences in allele frequency between the putative ancestral populations, admixture creates linkage disequilibrium (LD) between genetic loci, even between unlinked genetic markers. Evans & Cardon (2005); Li & Stephens (2003); Schramm *et al.* (2002) reported that between unlinked genetic markers, the linkage disequilibrium rapidly decays with successive generations while between linked markers it can persist for many more generations. This type of linkage disequilibrium known, as admixture LD, which occurs when considerable chromosomal segments are transmitted from a particular ancestral population can provide the necessary basis for conducting association studies (Hoggart *et al.*, 2004; McKeigue, 2005; Patterson *et al.*, 2004; Rosenberg & Pritchard, 2008).

1.3.3 Nature and Measures of Linkage Disequilibrium

When admixture occurs between multiple populations with different prevalence for a certain disease and with different allele frequencies; the resulting hybrid chromosomes are transmitted to the offspring during meiosis, and this process continues through subsequent generations (McKeigue, 2005; Rosenberg & Pritchard, 2008). Since admixture can generate linkage disequilibrium even between unlinked genetic markers, Goldstein & Weale (2001) reported that SNP frequencies can diverge if the genetic marker influences the phenotype, but a genetic marker variant can be associated with the condition not because it is biologically causal, but because it is statistically correlated with a causal variant. This fact arises because alleles at different loci are sometimes

found together more or less often than expected according to their frequencies. Linkage disequilibrium varies across populations and genome regions and between pairs of genetic markers in close proximity (Reich *et al.*, 2005). Several factors generate variation in LD, such as genetic drift, admixture and inbreeding, and these are population specific. There are other additional contributors to the extent and distribution of disequilibrium such as recombination rate, gene conversion and natural selection, which are specific to the genomic region (Kristin *et al.*, 2002; Reich *et al.*, 2005; Weir, 2008). Kristin *et al.* (2002) indicated that most of these involve demographic aspects of a population, and tend to distort the relationship between linkage disequilibrium strength and the physical distance between genetic loci. In addition, (Chakravati & Weiss, 1998; Kristin *et al.*, 2002; Lewontin, 1964; Spielman *et al.*, 1993) indicated that it is possible to restrict the genetic interval around the disease locus by identifying linkage disequilibrium between nearby genetic markers and the disease locus, if most affected individuals in a population share the same mutant allele at a causative locus. This assumption made use of many opportunities for crossovers between genetic markers and the disease locus during the large number of generations since the first appearance of mutation (McKeigue, 2005). Thus, there has been an increase in interest in linkage disequilibrium, that is owed largely to the belief that association studies can offer substantially more power for mapping common disease genes. Linkage disequilibrium was originally defined as the difference between the observed frequency of a two-locus haplotype and the frequency it would be expected to show if the alleles were segregating at random (Evans & Cardon, 2005; Kristin *et al.*, 2002; Weir, 2008). The most popular measures of linkage disequilibrium is the r^2 in equation 1.4 below. Considering two different loci A and B, with two alleles (A, a and B, b) at each genetic locus, respectively. The measure of linkage disequilibrium was initially given by

$$D = f_{AB} - f_A f_B, \tag{1.3}$$

where the observed frequency of the haplotype that consists of alleles A and B is denoted by f_{AB}. The expected haplotype frequency in the absence of linkage disequilibrium is computed as the product of the allele frequencies $f_A f_B$ of each of the two alleles, where f_A and f_B are the allele frequency of the allele A and B, respectively.

There are several alternative measures based on the measure D. Since these measures have different properties and measure different things, it might be difficult to compare different reports on the extent of linkage disequilibrium (Conrad *et al.*, 2010; Kristin *et al.*, 2002; Shiheng *et al.*, 2001; Weir, 2008). In addition, (Falush *et al.*, 2003) distinguished three types of linkage disequilibrium in human populations:

(1) Mixture linkage disequilibrium which is due to the population admixture, between unlinked genetic markers and known to be the main source of inflated type I error in case-control association studies.

(2) Admixture linkage disequilibrium which occurs when considerable chromosomal segments are transmitted from a particular ancestral population. It provides the necessary basis for conducting association studies.

(3) Background linkage disequilibrium which exists within ancestral populations because of correlation among polymorphisms over very short distances and is the main subject of case-control association studies.

The measure D given in equation 1.3 depends on allele frequency, and is not commonly used to measure the strength of linkage disequilibrium (Evans & Cardon, 2005; Goldstein & Weale, 2001). The normalized measure, D' of D, and r^2 are known as the most popular measures of linkage disequilibrium.

(1) The normalized measure, D' is determined by dividing D by its maximum possible value, given the allele frequencies at the two genetic loci, with alleles A and B, respectively

$$\begin{cases} D' = \left| \dfrac{D}{max\left[f_A(1-f_B),(1-f_A)f_B\right]} \right| \text{ when } D < 0 \\ D' = \left| \dfrac{D}{min\left[f_A f_B,(1-f_A)(1-f_B)\right]} \right| \text{ when } D > 0, \end{cases}$$

$D' = 1$ if, and only if two SNPs have not been separated by recombination during the history of the sample and there is complete linkage disequilibrium. Values of $D' < 1$ can indicate that the complete ancestral linkage disequilibrium has been disrupted and there is no clear interpretation of the values for $D' > 1$. According to (Evans & Cardon, 2005; Goldstein & Weale, 2001), this measure is known to be strongly dependent on the sample size. More details can be found in (Goldstein & Weale, 2001; Kristin *et al.*, 2002).

(2) The measure r^2 is complementary to D', and has recently emerged as the measure of choice for quantifying and comparing linkage disequilibrium in the context of association mapping (Chakravati & Weiss, 1998; Kristin *et al.*, 2002; Patterson *et al.*, 2004). It is the Pearson correlation of alleles at the two sites, and is obtained by dividing D^2 by the product of the four allele frequencies at the two genetic loci,

$$r^2 = \frac{D^2}{f_A f_B f_a f_b}. \tag{1.4}$$

The case of $r^2 = 1$ is known as perfect linkage disequilibrium, and occurs only if the markers have not been separated by recombination and have the same allele frequency (Chakravati & Weiss, 1998; Kristin *et al.*, 2002). This expected value of the disequilibrium coefficient r^2 is generally drawn from a probability distribution that results from the evolutionary process (Magnus, 2000; Patterson *et al.*, 2004). This process is known as the coalescent (in population genetics) (Magnus, 2000). When a sample of chromosomes is drawn from a population, all the chromosomes are related by some unknown genealogy, known as a coalescent tree (Magnus, 2000; Patterson *et al.*, 2004). Genetic markers that are quite close together on a chromosome have either the same or similar genealogies, and this induces dependence between the alleles at different markers. Genetic markers that are farther apart may have different ancestral genealogies, because of recombination (Chakravati & Weiss, 1998; Kristin *et al.*, 2002; Magnus, 2000; Patterson *et al.*, 2004). For this reason, the strength of linkage disequilibrium between pairs of genetic markers decreases as a function of the genetic distance between markers. The expected value of r^2 is a function of the parameter $\rho = 4N_e c$, where c is the total recombination rate between the two genetic markers (when $\rho = 4N_e c$ is assumed for a region containing a series of genetic markers, c is normally taken to be the total recombination rate across the entire region) and N_e is the effective population size.

1.4 Population Structure and Local Ancestry

1.4.1 Genetics Ancestry Overview

The identification of the genetic variation of recently admixed populations can reveal historical population events, and can be utilized for the identification of genetic markers associated with complex human diseases through association studies and admixture mapping. Over the last 80 years, statistical models have been developed in order to detect the probable ancestral origins of chromosomal segments and to understand the mosaic structure of the genome of admixed populations (Baran *et al.*, 2012; Falush *et al.*, 2003; Hoggart *et al.*, 2004; Pasaniuc *et al.*, 2009; Patterson *et al.*, 2006; Price *et al.*, 2009b; Sankararaman *et al.*, 2008; Tang *et al.*, 2006). Several questions that have arisen in analyzing the multi-locus genetic data have been solved, including: Are the samples from a homogeneous population? Are the sample sizes sufficient to infer the ancestry or to apply admixture mapping analysis? Does the data set contain subgroups that are genetically different or is there evidence that the samples in the data set are from a structured population? Although major progresses have been made in answering these questions, challenges still remain in the accuracy of modelling the background linkage disequilibrium which is expected

to be strong at short distances and can be increased due to founder events or be increased by population dynamics. These can lead to spurious ancestry inference (Falush *et al.*, 2003).

Information about population structure is well known to be useful in admixture mapping and studies of disease genes (Montana & Pritchard, 2004; Patterson *et al.*, 2004; Rosenberg & Pritchard, 2008). Recent investigations using a variety of genetic markers, have shown that individuals sampled worldwide fall into groups, approximately along continental lines, as well as self-identifying racial groups (Zhu *et al.*, 2008). To understand the population structure and estimate the genome-wide ancestry proportion (Global ancestry), researchers have developed statistical models to identify the probable ancestral origins (Alexander *et al.*, 2009; Falush *et al.*, 2003; Hoggart *et al.*, 2004) of a sample (probabilist clustering approaches) and used analytic techniques, such as Principal Component Analysis (PCA) to determine the underlying structure of populations (Price *et al.*, 2006; Rosenberg & Nordborg, 2006). For example, to determine whether a sub-population of particular samples are more closely related to each other than they are to the population as a whole. These statistical models consider admixed populations as statistical combinations of the source of the ancestral populations, by treating allele frequencies in a hybrid population as linear combinations of allele frequencies in the source of ancestral populations.

A specific location in the genome may inherit 0, 1, or 2 copies of a particular ancestry. Inferring an individual's local ancestry, or their number of copies of each ancestry at each location in the genome, also has important applications in disease mapping and in understanding human history. As the genomes of individuals from admixed populations consist of chromosomal segments of different ancestry, a specific location in the genome may contain 0, 1, or 2 copies (local specific ancestry or local ancestry) from a particular ancestral population. It has been shown that the inference of an individual's local ancestry have a wide range of applications from disease mapping to learning about history (Price *et al.*, 2009b; Sankararaman *et al.*, 2008). Various approaches for inferring local ancestry have been developed, and these methods can be clustering into three categories:

(1) Haplotype-based inference of locus-specific ancestry includes methods such as HAPMIX (Price *et al.*, 2009b), SPECTRUM (Sohn & Xing, 2007), HAPAA (Sundquist *et al.*, 2008), and SABER (Tang *et al.*, 2006) and make use of all SNPs of the genome of the admixed populations. The Haplotype-based inference makes use of Hidden Markov models (HMMs) based on the population-specific allele frequency profiles. This approach is known to be accurate when using two-way admixed populations.

(2) Overlapping windows based inference methods which uses the whole-genome data from a multi-way admixed population (Baran *et al.*, 2012; Qin *et al.*, 2010). Examples include

LAMP (Sankararaman *et al.*, 2008), and WINPOP (Pasaniuc *et al.*, 2009). This approach infers local ancestry by partitioning the genome into overlapping, contiguous windows of SNPs. It optimizes the likelihood model over each of the windows, and combines the solutions by casting a majority vote for each SNP (Pasaniuc *et al.*, 2009; Sankararaman *et al.*, 2008).

(3) A combined Haplotype-based and overlapping window based inference method on whole-genome data leverages the structure of linkage disequilibrium in the ancestral population, and incorporates the constraint of Mendelian segregation when inferring local ancestry in families. Examples include MULTIMIX (Churchhouse & Marchini, 2012), which is is used to estimate locus-specific ancestry that involves a model on background LD which extends across windows of SNPs and has the advantage of being applicable to the complex multi-way admixed populations for which a panel of genotyped patterns is available. Moreover, the method can handle either phased or unphased genotype data on the study partens and source populations (Churchhouse & Marchini, 2012).

(4) Principal Component Analysis-based method, such as PCADMIX relies on Principal Components Analysi (PCA) to quantify the information that each SNP contributes to distinguishing the ancestry of a genomic region of an admixed population (Henn *et al.*, 2012).

(5) Imputation-based approach including ALLOY (Rodriguez *et al.*, 2012) which enables the incorporation of complex models for linkage disequilibrium in the ancestral populations. This method applies a factorial hidden Markov model to capture the parallel process producing the maternal and paternal admixed haplotypes. In addition, this method models background LD in ancestral populations via an inhomogeneous variable length Markov chain.

Today, inferring local ancestry conditional on more than two ancestral populations is considered to be unsolved (Baran *et al.*, 2012; Pasaniuc *et al.*, 2009). Methods have improved for ancient admixture events and admixture between more closely related populations based on two-way admixed populations, but challenges remain in accurately inferring local ancestry for multi-way admixed populations and accounting for admixed parental populations. Nevertheless, high-throughput genotyping or sequencing and new methods to infer local ancestry can allow for joint admixture and association analysis. This may benefit association mapping in admixed populations by eliminating the effects of confounding due to variation in ancestry (Baran *et al.*, 2012; Price *et al.*, 2009b).

1.4.2 Principal Component Analysis (PCA)

This approach focuses on the decomposition of variance and the covariance matrix for dimensionality reduction. Let C be a large rectangular matrix with rows indexed by individual and columns, indexed by polymorphic markers; for each marker we choose the reference and the variant allele, where n is a marker and m an individual, $C(i,j)$ the numbers of variant allele for marker j and individual i. We assumed that there is no missing data. Let us subtract the column means, from each column, thus (Patterson *et al.*, 2006)

$$\mu(j) = \frac{\sum_{i-1}^{m} C(i,j)}{m}, \tag{1.5}$$

and then the correct entries are:

$$C(i,j) - \mu(j). \tag{1.6}$$

Set $p(j) = \mu(j)/2$, an estimation of underlying allele frequency. Then each entry in the resulting matrix is

$$M(i,j) = \frac{C(i,j) - \mu(j)}{\sqrt{p(j)(1-p(j))}}. \tag{1.7}$$

The equation 1.7 is a normalization due to the fact that the frequency change of SNPs caused by genetic drift occurs at a rate proportional to $\sqrt{p(j)(1-p(j))}$ per generation. It is also normalized if the data is in Hardy-Weinberg equilibrium. The PCA method incorporates the Tracy-Widom Theory for finding the probability of the largest eigenvalue. Let $m \times n$ be a matrix M. Let

$$X = \frac{1}{n} MM',$$

where X is a Wishart (Wishart distribution is a generalization to multiple dimensions of the chi-squared distribution) matrix. Let $\{\lambda_i\}_{1 \leq i \leq m}$ be the eigenvalue of X. For when the m, n, are large, the distribution of the largest eigenvalue λ_1 is a Tracy-Widom distribution. Setting

$$\mu(m,n) = \frac{(\sqrt{n-1} + \sqrt{m})^2}{n}, \tag{1.8}$$

$$\sigma(m,n) = \frac{(\sqrt{n-1} + \sqrt{m})}{n} \left(\frac{1}{\sqrt{n-1}} + \frac{1}{\sqrt{m}} \right)^{\frac{1}{3}},$$

where σ is the variance of the normal distribution used for the cells of rectangular matrix M. Now setting

$$x = \frac{\lambda_1 - \mu(m-n)}{\sigma(m,n)}. \tag{1.9}$$

Let distinguished n, the actual number of columns of our data array, and n', a theoretical statistical parameter. We fit σ, n with maximum likelihood. The likelihood, has a function of two parameters, has two sufficient statistics, which are $\sum_i \lambda_i$, and $\sum_i \log \lambda_i$. In genetic applications, the maximum likelihood maybe due to $\sum_i \log \lambda_i$ which is not reliable with the small eigenvalues, thus we are concerned about large eigenvalues (Patterson *et al.*, 2006).

$$n' = \frac{(m+1)\left(\sum_i \lambda_i\right)^2}{\left((m-1)\sum_i \lambda_i^2\right) - \left(\sum_i \lambda_i\right)^2}. \tag{1.10}$$

In order to study whether the analysed population was structured in the biallelic dataset, the algorithm below was run.

Algorithm 1 A test for population structure algorithm

(1) Compute the matrix M as in Equations 1.5 and 1.6 and 1.7. M as m rows and n column;

(2) Compute $X = MM'$. X is $m \times n$;

1. Order the eigenvalues of X so that$\lambda_1 > \lambda_2, \cdots > \lambda_{m'} > 0$; where $m' = m - 1$. (on a large dataset X will always have rank m')

(3) Using the eigenvalues $\lambda_i (1 \leq i \leq m')$, estimate n' from the Equation 1.10.

(4) The largest eigenvalues of M is λ_i. Set

$$l = \frac{(m')\lambda_i}{\sum_i \lambda_i}.$$

(5) Normalize l with the Equations 1.8 and 1.9, where the effective number of markers n' replace n. This yields a test statistic $x = x(M)$.

The $x(M)$ is approximately Tracy-Widom distribution.

1.4.3 Probabilistic Approach

Based on the (Falush *et al.*, 2003; Pritchard *et al.*, 2002) model, we consider a sample of N individuals, each genotyped at L loci, assuming there are K distinct populations that contribute

to the ancestry of our study sample. Individuals have ancestors in more than one population. The ancestry of each individual can be defined as the proportion of that individuals genome inherited from each of the K populations. For instance, the ancestry of individual i is specified by a vector,

$$\begin{aligned} q^{(i)} &= \left(q_1^{(i)}, q_2^{(i)}, \ldots, q_K^{(N)}\right) \\ q_k^{(i)} &= Pr\left(z_l^{(i),j} = k|r, Q\right), l = 1, \ldots, L.i = 1, \ldots, N \qquad (1.11) \\ \sum_{k=1}^{K} q_k^{(i)} &= 1, \qquad (1.12) \end{aligned}$$

where $q_k^{(i)}$ is the ancestry proportion of individual i from population k. As the genome of recently admixed individuals is viewed as a series of chromosomal segments each of which descends as an intact unit, without recombination from one of the ancestral populations, we denote Q as the multi-dimensional vector and its components are each values of $q^{(i)}$. We assume that for each individual i, each chromosomal segment comes from population k independently with probability $q_k^{(i)}$. This is assumed to be drawn independently from the population of origin $(1, \ldots, K)$, equation 1.11. $z_l^{(i,j)}$ is the population of origin $(1, \ldots, K)$ of the j^{th} copy of genetic marker l in individual i. We denote Z as a multi-dimensional vector containing all the values of z. For the haploid data independently for each individual i, the populations of origin $(1, \ldots, K)$ along each of individual $i's$ chromosomes form independent Markov chains (Falush *et al.*, 2003; Pritchard *et al.*, 2002) satisfying,

$$Pr\left(z_{l+1}^{(i)} = k'|z_l^{(i)} = k, r, Q\right) = \begin{cases} e^{(-d_l r)} + \left[1 + e^{(-d_l r)}\right] q_{k'}^{(i)} & \text{if } k' = k \\ \left[1 - e^{(-d_l r)}\right] q_k^{(i)} & \text{otherwise,} \end{cases} \qquad (1.13)$$

where d_l is the genetic distance from locus l to locus $l+1$. (Montana & Pritchard, 2004) suggested the average size of chromosomal segments to be $\frac{100}{r}$ cM, where r is roughly viewed as the average time since admixture, and the breakpoints from one segment to the next are assumed to occur as a Poisson process, with a rate of r per Morgan (Falush *et al.*, 2003; Montana & Pritchard, 2004; Pritchard *et al.*, 2002). We can use a series of genetic markers along each chromosome, to infer the hidden pattern of chromosomal segments. Each population is characterized by a list of the allele frequencies at each of genotyped markers. We denote P as the multi-dimensional vector that contains the allele frequencies p_{klj} of allele j at each genetic marker l in each population k where the allele frequencies are unknown in advance, but will usually be

samples of non-admixed representatives from the original populations to assist in their estimation (Montana & Pritchard, 2004; Price *et al.*, 2007; Zhu *et al.*, 2006). A Bayesian framework can be performed for the purpose of the inference and it demands prior informations for P and Q (Falush *et al.*, 2003; Pritchard *et al.*, 2002):

(1) The multi-dimensional P, is the vector of allele frequencies at genetic locus l in population k and are drawn from a symmetric Dirichlet distribution parametrized by a single hyper-parameter λ, independently for each ancestral population k (Falush *et al.*, 2003).

(2) The admixture proportions $q^{(i)}$ for individual i are also drawn from a symmetric Dirichlet distribution with a hyper-parameter α. The parameter α is viewed as a vector of K values, with α_k representing the relative contribution of ancestral population k to the genetic material in the sample. More details can be found in (Falush *et al.*, 2003; Pritchard *et al.*, 2002).

1.4.3.1 Markov Chain Monte Carlo

Markov Chain Monte Carlo is commonly used to sample from the posterior distribution of P, Q, Z, λ, r, and α, given the genotype data X and the number K of ancestral populations (Falush *et al.*, 2003; Hubisz *et al.*, 2009; Pritchard *et al.*, 2002).

$$Pr\left(P, Q, Z, r, \alpha, \lambda | X, K\right)$$

Markov Chain Monte Carlo (MCMC) can arbitrarily use initial choices for each parameter and then propose updates that change a subset of these, conditional on the other parameters and the data (Sohn & Xing, 2007; Xing *et al.*, 2007). We provide a short MCMC scheme for sampling from a Markov chain with stationary distribution $P\left(P, Q, Z, r, \alpha, \lambda | X, K\right)$, a full description of which can be found in (Falush *et al.*, 2003; Hubisz *et al.*, 2009).

(1) Sample Z from $Pr(Z|P, r, Q, X)$.

(2) Sample P from $Pr(P|Z, r, Q, X) = P(P|Z, X)$.

(3) Update r,F,Q by Metropolis-Hastings update (Falush *et al.*, 2003; Pritchard *et al.*, 2002). The parameters α and λ could also be updated by Metropolis-Hastings (Falush *et al.*, 2003; Hubisz *et al.*, 2009; Montana & Pritchard, 2004; Terry, 2003; Warren & Grant, 2005).

Step 1 is done separately for each individual using the Hidden Markov model within the Forward-Backward algorithm in section 1.4.3.2.

1.4.3.2 Hidden Markov Model

Let $\{\vartheta_t\}_{t=1}^{T}$ denote T ordered, observed genotypes along a chromosome and $\{z_t\}_{t=1}^{T}$ the unobservable number of ancestral alleles at the corresponding genetic marker loci. We denote A and a the two alleles at genetic marker locus t, $x_t \in \{0,1,2\}$ denotes the genotype at the genetic marker locus t.

$$\begin{cases} 0 & \text{for genotype aa} \\ 1 & \text{for genotype Aa} \\ 2 & \text{for genotype AA.} \end{cases} \tag{1.14}$$

We assume that x_t depends not only on z_t but also on the past history. We denote $\{x_t, z_t\}_{t=1}^{T}$ as the hidden Markov model framework,

$$\begin{array}{lll} Observed & genotype & x_1 \rightarrow x_2 \ldots\ldots \rightarrow x_T \\ & & \uparrow \quad \uparrow \\ Hidden & states & z_1 \rightarrow z_2 \ldots\ldots \rightarrow z_1. \end{array}$$

The estimation of the hidden states of the Markov chain for Z is then performed independently for each individual by use of the Baum-Welch (Forward-Backward) algorithm based on the forward and backward probability quantities (Zhang *et al.*, 2004). Here, this algorithm is presented in order to compute the marginal posterior assignment probabilities at each locus for the purpose of locus-specific ancestry that could be used in admixture mapping analysis. For each chromosome from each individual, we define the forward and backward probabilities, equations 1.15 and 1.16 respectively, and these probabilities are defined for all states k and for all genetic loci from 1 to L.

$$\alpha_{lk} = Pr\left(x_1, \ldots, x_l, z_l = k | P, r, Q\right) \tag{1.15}$$

$$\beta_{lk} = Pr\left(x_{l+1}, \ldots, x_L | z_l = k, P, r, Q\right). \tag{1.16}$$

It follows that,

$$\alpha_{lk}\beta_{lk} = Pr\left(x_1, \ldots, x_L, z_l = k | P, r, Q\right).$$

Therefore, for a given locus l the likelihood can be computed as follows,

$$\sum_{k=1}^{K} \alpha_{lk}\beta_{lk} = Pr\left(x_1, \ldots, x_L | P, r, Q\right) = L_l. \tag{1.17}$$

The conditional probabilities for all loci l and all populations k is written as follows,

$$Pr\left(z_l = k | X, P, r, Q\right) = \frac{Pr\left(x_1, \ldots, x_L, z_l = k | P, r, Q\right)}{Pr\left(x_1, \ldots, x_L | P, r, Q\right)} = \frac{\alpha_{lk}\beta_{lk}}{L_l}. \tag{1.18}$$

Considering the transition probabilities of the Markov chain, equations 1.18, we denote

$$P_{kk'} = Pr\left(z_{l+1}^{(i)} = k' | z_l^{(i)} = k, r, Q\right).$$

Starting with the case of complete phase information, the Forward probabilities are,

$$\begin{cases} \alpha_{1k} = q_k p_{k1} x_1, & k = 1, \ldots, K \\ \alpha_{(l+1)k'} = \left[\sum_{k=1}^{K} \alpha_{lk} P_{kk'}\right] p_{k'(l+1)} x_{l+1}, & l = 1, \ldots, L. \end{cases} \tag{1.19}$$

The backward probabilities are,

$$\begin{cases} \beta_{1k} = 1, & k = 1, \ldots, K \\ \beta_{lk'} = \sum_{k=1}^{K} \beta_{(l+1)k} P_{kk'} x_{l+1} p_{k(l+1)}, & l = 1, \ldots, L. \end{cases} \tag{1.20}$$

When phase information is missing or only partially known, the forward and backward probabilities and the resulting joint conditional probability of the ancestral states in the two allele copies are as follows,

$$\alpha_{lk^1k^2} = Pr\left(x_1^1, x_1^2, \ldots, x_l^1, x_l^2; z_l^1 = k^1, z_l^2 = k^2 | P, r, Q\right) \tag{1.21}$$

$$\beta_{lk^1k^2} = Pr\left(x_{l+1}^1, x_{l+1}^2, \ldots, x_L^1, x_L^2 | z_l^1 = k^1, z_l^2 = k^2, P, r, Q\right) \tag{1.22}$$

$$Pr\left(z_l^1 = k^1, z_l^2 = k^2 | X, P, r, Q\right) = \frac{\alpha_{lk^1k^2}\beta_{lk^1k^2}}{L_l}, \tag{1.23}$$

where the superscripts $(^1)$ and $(^2)$ in the equations 1.21, 1.22 and 1.23 refer to the first and second allele copy at each locus, respectively. Let c_l represent the probability that the first alleles of adjacent loci l and $l+1$ are on the same chromosome. For unphased data, the order of the allele copies is random, and so c_l can be set to 0.5. Starting with the case of unphased data, the forward probabilities are,

$$\begin{cases} \alpha_{lk^1k^2} = q_{k^1} q_{k^2} p_{k^1 1} x_1^1 p_{k^2 1} x_1^2, \left(k^1, k^2 = 1, \ldots, K\right). \\ \alpha_{(l+1)k'^1k'^2} = \sum_{k^1=1}^{K} \sum_{k^2=1}^{K} \alpha_{lk^1k^2} p_{k'^1(l+1)} x_{l+1}^1 p_{k'^2(l+1)} x_{l+1}^2 \left[\right. \\ \left. c_l P_{k'^1k^1} P_{k'^2k^2} + (1 - c_l) P_{k'^1k^2} P_{k'^2k^1}\right] \\ \text{For } l = 1, \ldots, L. \end{cases} \tag{1.24}$$

The backward probabilities are,

$$\begin{cases} \beta_{1k^1k^2} = 1 \\ k^1, k^2 = 1, \ldots, K. \\ \beta_{lk'^1k'^2} = \sum_{k^1=1}^{K} \sum_{k^2=1}^{K} \beta_{(l+1)k^1k^2} p_{k^1(l+1)} x_{l+1}^1 p_{k^2(l+1)} x_{l+1}^2 [\\ c_l P_{k^1k'^1} P_{k^2k'^2} + (1-c_l) P_{k^1k'^2} P_{k^2k'^1}] \\ \text{For } l = l+1, \ldots, L. \end{cases} \tag{1.25}$$

1.4.3.3 Locus-Specific Ancestry

From the posterior estimates of P, Q and r derived at each iteration of Markov Chain Monte-Carlo, we denote the posterior mean estimates of P, Q and r by $\hat{P}$, $\hat{Q}$ and $\hat{r}$, respectively. These posterior mean estimates can be evaluated through the hidden Markov model described in subsection 1.4.3.1 on page 17. Therefore, the posterior average quantities are defined in order to estimate the average ancestry proportions of affected individuals and of the controls, in equations 1.26 and 1.27, respectively.

$$\bar{q}_d = \frac{1}{n_d} \sum_{i=1}^{n_d} E\left[q^{(i)}|X\right] \tag{1.26}$$

$$\bar{q}_c = \frac{1}{n_c} \sum_{i=1}^{n_c} E\left[q^{(i)}|X\right], \tag{1.27}$$

where n_c and n_d are respectively the number of controls and cases in the sample data. Next, let us denote $\bar{z}_l^{(i)}$ as the posterior average ancestry of individual i at locus l, evaluated at $\hat{P}$, $\hat{Q}$ and $\hat{r}$.

$$\bar{z}_l^{(i)} = \frac{1}{2} \sum_{j=1}^{2} P\left(z_l^{(i),j} = k|X, \hat{P}, \hat{Q}, \hat{r}\right). \tag{1.28}$$

Equation 1.28 is viewed as the locus-specific ancestry of an individual at locus l, and j is the genetic copy index. Thus, the posterior averages of z at locus l among cases and controls are also denoted by $\bar{z}_{l,d}$ and $\bar{z}_{l,c}$, and given in equations 1.29 and 1.30 respectively.

$$\bar{z}_{l,d} = \frac{1}{2n_d} \sum_{i=1}^{n_d} \sum_{j=1}^{2} P\left(z_l^{(i),j} = k|X, \hat{P}, \hat{Q}, \hat{r}\right) \tag{1.29}$$

$$\bar{z}_{l,c} = \frac{1}{2n_c} \sum_{i=1}^{n_c} \sum_{j=1}^{2} P\left(z_l^{(i),j} = k|X, \hat{P}, \hat{Q}, \hat{r}\right) \tag{1.30}$$

Equations 1.29 and 1.30 are viewed as the average locus-specific ancestries of an individual at locus l among cases and controls.

1.5 Genetic Diseases

1.5.1 Overview of Genetic Diseases

Genetic disease is a disease mainly caused by abnormalities in an individual's genetic material (genome). A single mutation or variant in human genome can be sufficient to cause a disease, and in other cases a variant may interact with many other genetic variants and environmental factors to lead to a disease. Genetic diseases can be classified into three different types, including (1) single-gene (in some cases mitochondrial diseases (Stopple, 1996)), (2) multi-factorial (polygenic disease), (3) chromosomal. Single-gene diseases, known as Mendelian or monogenic diseases are caused by mutations that occur in the DNA sequence of a single gene, and are inherited in recognizable patterns such as autosomal dominant, autosomal recessive, and X-linked. There are more than $6,000$ known single-gene disorders, which are known to occur in about 1 out of every 200 births, such as cystic fibrosis, sickle cell anaemia, Marfan syndrome, Huntingtons disease, and hereditary hemochromatosis (Stopple, 1996). In addition, mitochondrial disorder is classified as single-gene diseases, which is a rare type of genetic disorder caused by mutations in the non-chromosomal DNA of mitochondria.

Multi-factorial (complex or polygenic) disease is caused by a combination of environmental factors and mutations in multiple genes. Complex diseases arise as a result of genetic variation at several genetic loci in the human genome, each of low penetrance and implying that each mutation has a weak effect on its own (Stopple, 1996). Polygenic disease is caused by the combined action of more than one gene. Examples of polygenic conditions include hypertension, coronary heart disease, and diabetes. Because such disorders depend on the simultaneous presence of several genes, they are not inherited as simply as are single-gene diseases. Chromosomes are distinct structures made up of DNA and protein, are located in the nucleus of each cell. Because chromosomes are carriers of genetic material, abnormalities in chromosome structure such as missing or extra copies or gross breaks and rejoining (translocations) can result in disease (chromosomal disorder).

1.5.2 Mendelian versus Complex Diseases

By the early 1900s it became clear that many common human diseases show familial aggregation that does not follow simple Mendelian inheritance patterns, but appears to be due instead to a large and usually unknown number of genes, often with interacting environmental factors (Smith,

2007; Smith & Ebrahim, 2004). Diseases such as schizophrenia, asthma, diabetes, obesity, tuberculosis, coronary heart disease, hypertension, various cancers, Alzheimers disease and Parkinsons disease among related individuals are examples of important human complex diseases with a genetic contribution to susceptibility (Weir, 2008). In such diseases, only a very small fraction of the disease susceptibility can be attributed to any given mutant gene (Smith, 2004). Mendelian diseases are generally derived from mutations in a single nucleotide with high penetrance, and a large effect on protein function, consistent with the fact that these diseases involve single mutations with strong phenotypic effects (Magnus, 2000; Spielman *et al.*, 1993). Such mutations are rare at the population level (Chakravati & Weiss, 1998; Goldstein & Weale, 2001; Halder & Shriver, 2003), transmitted by Mendelian inheritance and have often initially been identified with characteristic patterns of transmission (X-linked, dominant and recessive). The successes in the study of Mendelian diseases owed much to the fact that the genetic diseases under investigation in humans were relatively simple, i.e. monogenic, high-penetrance disorders and obey the principles of Mendelian inheritance (Halder & Shriver, 2003; McKeigue, 2005). Most of these were identified by linkage analysis, using data collected from affected families. In addition, the regions of the genome were also identified that co-segregate with the disease in many independent families over many generations of a long pedigree (Kristin *et al.*, 2002; Patterson *et al.*, 2004). (Excoffier & Hamilton, 2003; Halder & Shriver, 2003; McKeigue, 2005), which indicated that disease genes can generally be localized only to large intervals using this method because the co-segregating piece of DNA is delineated by observed crossovers which occur at relatively low frequency. However, complex diseases are typically caused by genetic variation at several genetic loci in the human genome and influenced by several environmental factors, each of which makes only a small contribution to the final phenotype and implying that each mutation has a weak effect on its own (Halder & Shriver, 2003).

Many common human diseases and traits are believed to be influenced by several genetic and environmental factors. These diseases do not have a clear-cut pattern of inheritance. Since genes contribute to diseases with complex inheritance architecture, only a small fraction (less than $1\% \sim 7\%$ of affected individuals) owes its origin to a single mutant gene transmitted by Mendelian inheritance (Scheuner *et al.*, 2004). Initially, alleles have been assumed to be the genetic factors underlying common diseases. Allelic architecture (effect size and frequency of susceptibility variants) may differ across phenotypes, and that heritability may take a different form for different diseases. Currently, the knowledge about the nature of genetic variation underlying complex diseases in humans is limited, which makes it difficult to determine a persons risk of inheriting the disease (Manolio *et al.*, 2004). Although Genome-wide association studies (GWAS) are designed as a powerful tool for investigating the genetic architecture of complex diseases (section 1.6.2 and 1.7), a number of challenges still remain (section 1.7).

1.6 Disease-mapping Methods

The identification of genes underlying genetic disease has been a critical concern of geneticists. Historically, the first disease genes were identified by pure position-independent methods, because no relevant mapping information existed and the techniques were not developed yet (Strachan & Read, 1999). Thus, statistical and mathematical approaches have been developed to this end. In particular, studies of Mendelian disorders have been greatly enhanced over the last few decades by remarkable achievements in gene mapping and the development of rigorous statistical methods (Lee & Yen, 2003; Martin *et al.*, 2001; Schaid, 1998; Smith, 2004). Most of the progress in human genetics during that time has come from the studies of families with rare segregating high-risk alleles. Considering the limitations of pedigree studies and family-based approaches, other approaches were sought to reduce the interval in which a disease gene might lie and to use the information generated by recent admixture of populations from historically distinct geographic origins. These approaches include genetic linkage studies and population based Genome-wide case-control association studies (also admixture mapping), which model the linkage disequilibrium through classic likelihood and Bayesian statistics.

1.6.1 Pedigree and Family-based Methods

In the mid-$1990s$, the methods of choice for disease mapping became the family-based population methods and the most popular techniques for detecting linkage or association between a genetic marker locus and a disease susceptibility locus was the transmission disequilibrium test and its extensions (Zhu *et al.*, 2008). These methods focused on the transmission of alleles from heterozygous parents to their offspring. The original transmission-disequilibrium test (TDT) has been utilized to test for linkage disequilibrium in family triads, containing two parents and an affected offspring (Spielman *et al.*, 1993). For a marker locus with two alleles, the TDT compares the number of heterozygous parents who transmit one allele with the number of heterozygous parents who transmit the other allele to the affected offspring (Dinga & Lina, 2006; Zhu *et al.*, 2008). The so-called informative nuclear families contain at least one affected child, both parents genotyped at the marker and at least one parent is heterozygous. The informative discordant sibships (children produced by a pair of parents) have at least one affected and one unaffected sibling (DSP) with different genetic marker genotypes and may or may not have the parental genotype data. The informative extended pedigrees contain at least one informative nuclear family and (or) discordant sibship (Lee & Yen, 2003).

Considering a genetic marker locus with two alleles, A_1 and A_2; η_{A_1} is the number of allele A_1 transmitted and ζ_{A_1} is the number of allele A_1 not transmitted. For any family triad, there can be a pair of alleles that can be transmitted to the affected offspring and a pair of alleles that

are not transmitted. For each triad within an informative nuclear family, we can define a random variable,

$$X_T = \eta_{A_1} - \zeta_{A_1}. \tag{1.31}$$

We denote by $\bar{\eta}_{A_1}$ as the number of allele A_1 in the affected sib and $\bar{\zeta}_{A_1}$ the number of allele A_1 in the unaffected sib. We can similarly define another random variable for each DSP within an informative discordant sibship,

$$X_S = \bar{\eta}_{A_1} - \bar{\zeta}_{A_1}. \tag{1.32}$$

A summary random variable can be defined, for a pedigree that contains m_T triads from informative nuclear families and m_S DSPs from informative discordant sibships,

$$D = \frac{1}{m_T + m_S}\left(\sum_{k=1}^{m_T} X_{T_k} + \sum_{k=1}^{m_S} X_{S_k}\right), \tag{1.33}$$

D in the equation 1.32 above is the average that includes all possible triads from informative nuclear families and all possible DSPs from informative discordant sibships from the pedigree. It follows that under the null hypothesis of no linkage disequilibrium, $E(X_T) = 0$ for all triads and $E(X_S) = 0$ for all DSPs, therefore for any pedigree $E(D) = 0$. M is the total number of unrelated informative pedigrees in the sample and D_i is the summary random variable for the i^{th} pedigree. Thus, under the null hypothesis of no linkage disequilibrium,

$$E\left(\sum_{k=1}^{M} D_k\right) = 0$$

then,

$$Var\left(\sum_{k=1}^{M} D_k\right) = \sum_{k=1}^{M} Var\left(D_k\right) = E\left(\sum_{k=1}^{M} D_k^2\right).$$

The pedigree transmission disequilibrium test (PDT) is based on statistic T,

$$T = \frac{\sum_{k=1}^{M} D_k}{\sqrt{\sum_{k=1}^{M} D_k^2}} \tag{1.34}$$

The statistic in equation 1.34 is asymptotically normal, with mean 0 and variance 1, under the null hypothesis of no linkage disequilibrium and requires the genotypes of the parents in order to be computed. TDT has also been extended to allow for multiple affected offspring while remaining a valid test of linkage disequilibrium (Martin *et al.*, 2001). TDT has been extended to sibships

with at least one affected and one unaffected individual and this extension was referred to as the Sibling Transmission-Disequilibrium Test (S-TDT). For an allele of interest at a genetic marker locus, the S-TDT essentially compares the frequency of that allele among affected individuals with the frequency of the allele among unaffected individuals. It has been utilized when the data contains some missing genotypes among parents. Thus, S-TDT could use the genotypes of phenotypically discordant sibships and reconstruct parental genotypes from the genotypes of offspring (Dinga & Lina, 2006; Schaid, 1998).

(Dinga & Lina, 2006; Horvath *et al.*, 2000; Schramm *et al.*, 2002) have proposed an extension of TDT to the maximum-likelihood based on variance-components procedures and statistical selection for mapping quantitative-trait genetic loci in sib pairs. This approach allowed a joint test of both linkage and allelic association. It involved modelling of the allelic means for the test of association, with simultaneous modelling of the sib-pair covariance structure for a test of linkage (Dinga & Lina, 2006; Martin *et al.*, 2001). In fact, the maximum-likelihood variance-components controlled for spurious associations due to population structure and admixture by grouping the mean effect of a genetic locus into between and within-sibship components (Dinga & Lina, 2006; Horvath *et al.*, 2000). (Dinga & Lina, 2006; Martin *et al.*, 2001; Schaid, 1998) have suggested modelling the full starship covariance structure by maximizing the natural log of the likelihood of multivariate normal data (equation 1.34)

$$\mathcal{L} = \prod_{k=1}^{M} (2\pi)^{-(\frac{L_k}{2})} (\mid \Sigma_k \mid)^{-(\frac{1}{2})} (e)^{-(\frac{1}{2})} \left[(Y_k - \mu_k)\Sigma_k^{-1}(Y_k - \mu_k) \right], \tag{1.35}$$

where M is the number of families, and Y_k is the vector of observed scores obtained for siblings in family k. L_k is the number of variables (siblings in the single-phenotype case) measured in family k. For family k, μ_k is the vector of expected means, which is used to model the association parameters and Σ_k, the expected covariance matrix among siblings, is used to model the linkage. The elements of the covariance matrix Σ_k and the mean vector μ_k can be estimated directly and be made a function of the theoretical parameters of interest (Horvath *et al.*, 2000). Equation 1.35 is utilized for modelling of quantitative phenotypes obtained from sibships or extended families (Dinga & Lina, 2006; Martin *et al.*, 2001). These theoretical parameters are tested for statistical significance by fitting the model with the parameter of interest, and computing the $\log$ of the likelihood of the data $\log(\mathcal{L}_1)$; by refitting without these theoretical parameters (i.e. initial parameters derived from the elements of the covariance matrix Σ_k and the mean vector μ_k) and computing the $\log$ of the likelihood of the data, $\log(\mathcal{L}_0)$ (Horvath *et al.*, 2000). Thus, for a large data set,

$$2\left[\log(\mathcal{L}_1) - \log(\mathcal{L}_0)\right], \tag{1.36}$$

is asymptotically distributed as a χ^2 statistic. All parameters are estimated as a full model, compared with various sub-models which allow individual tests of association and linkage. More details can be found in (Dinga & Lina, 2006; Horvath *et al.*, 2000; Martin *et al.*, 2001).

Because of the low penetrance of complex diseases, the identification of genetic loci that contribute to the complex disease require a vast amount of information and pedigrees of adequate size are very costly (Excoffier & Hamilton, 2003; Halder & Shriver, 2003). Even if the candidate regions can be identified from pedigrees, (Excoffier & Hamilton, 2003; McKeigue, 2005) indicated that the resolution of linkage studies is generally in the order of a few centimorgans, which in terms of the human genome, may correspond to several mega-bases of DNA, and thousands of genes (Excoffier & Hamilton, 2003). Even if pedigree studies could resolve complex disease loci to the gene level, (Kristin *et al.*, 2002) mentioned that there is a strong discovery bias towards variants that cause Mendelian forms of complex disease which actually contribute relatively little to the disease phenotype on a population scale. Linkage studies are comprehensive and localize any gene that exerts a major signal on disease susceptibility, but it has relatively low power and still fails to identify genes carrying only a moderate signal of risk of genetic disease (Hoggart *et al.*, 2004; Montana & Pritchard, 2004; Zhang *et al.*, 2004). In addition, there are several factors that reduce the power and efficiency of the Transmission Disequilibrium Test and its variates. First, it demands a lot of data; at least three individuals have to be genotyped for each data point. Second, obtaining parental genotypes can be difficult. Finally, in order to be informative at a locus, parents have to be heterozygous at a genetic locus. Although efforts can be made to use genetic loci with high heterozygosity, Chakravati & Weiss (1998); Patterson *et al.* (2004); Spielman *et al.* (1993) reported that a significant fraction of affected individuals and their parents will always be uninformative. In the same vein, (Chakravati & Weiss, 1998; McKeigue, 2005) indicated that allelic heterogeneity (multiple susceptibility alleles at a disease locus), multiple contributory loci, low penetrance and environmental effects all act to reduce the power of these family-based population methods. For these reasons, McKeigue (2005) mentioned that a population view of complex disease may be preferable.

1.6.2 Population-Based Genome-Wide Association

Whole-genome association studies often use a case-control design to identify genetic variants related to a specific complex genetic disease that result, in weak genotype-phenotype correlation (Draghici, 2003; Excoffier & Hamilton, 2003; McKeigue, 2005). This compares allele frequencies between unrelated individuals that are affected to those that are unaffected. Association studies have much greater power but, as association is detectable over much smaller regions than linkage analysis (section 1.6.1), it is expected that testing the genome with dense SNPs can capture the

linkage disequilibrium and produce results that explain much of the risk. Regardless, many more markers would need to be typed to conduct a genome wide association study, which was extremely costly (Excoffier & Hamilton, 2003), but is now becoming affordable. Whole-genome association studies have been suggested, in principle, to be able to find genes of weak effect (SNPs with frequencies greater than 1% are responsible for conferring the risk of most genetically complex disorders) and to detect risk factors that may contribute to common human diseases (Rosenberg & Pritchard, 2008). In general, GWAS requires three essential elements, including large study samples from populations under study, polymorphic alleles that can be inexpensively and efficiently genotyped and cover the whole genome adequately, and analytic methods that are statistically powerful and that can be utilized to detect the genetic associations in an unbiased manner.

The substantial number of recently published GWAS are mainly conducted on European populations or populations of European descent, for which large samples of ancestrally homogeneous individuals from relatively homogeneous environments are available (Cantor *et al.*, 2010; Rosenberg *et al.*, 2010). Recent technological advances in high-throughput genotyping have allowed the expansion of human genetic studies to include diverse non-European populations in order to:

(1) Detect novel loci absent or not readily identifiable in European populations due to both low statistical power and allele frequencies (Cantor *et al.*, 2010).

(2) Find the extent to which the GWAS results from studies of European populations can be extended to non-European populations (Cantor *et al.*, 2010).

(3) Investigate possible phenotypes or diseases of high prevalence present in non-European populations such human African trypanosomiasis, known as sleeping sickness (Cantor *et al.*, 2010).

Despite these successes in European populations, for most genetic disorders, only a few common variants were found to be involved and the associated loci explain only a small fraction of the genetic risk. Moreover, the smaller extent of linkage disequilibrium (LD) between variants in African populations is an advantage for fine-scale mapping, which is still a constant challenge for GWAS (Cantor *et al.*, 2010). The risk of false-positive genotype-phenotype associations due to difference in ancestry is a major challenge for association studies in admixed populations (Rosenberg & Nordborg, 2006). To this end, several methods have been developed to control for the false positive results in samples of ancestrally homogeneous individuals, including principal components, genomic control, structured association testing, propensity scores and variance components (Epstein *et al.*, 2007; Price *et al.*, 2010; Rosenberg & Nordborg, 2006; Tiwari *et al.*, 2008). These approaches make use of inferred genome-wide ancestry proportion from individuals

as a covariate in order to control for confounding due to variation in individual ancestry (Redden *et al.*, 2006). The use of linear mixed models (LMMs) in genome-wide association studies (GWAS) is now widely accepted (Kang *et al.*, 2010; Zhou & Stephens, 2012) as LMMs have been shown to correct for several forms of confounding due to genetic relatedness, such as population structure and familial relatedness (Zhou & Stephens, 2012). Here we describe a similar approach developed in (Kang *et al.*, 2010).

1.6.2.1 An Overview of the Mixed Model in GWAS

Let us consider n measurements of phenotype of i individuals. The linear mixed model can be written in organism association mapping mode (Kang *et al.*, 2010; Zhou & Stephens, 2012) as

$$Y = X\beta + Z\mu + \epsilon, \tag{1.37}$$

where Y is an $n \times 1$ vector of observed phenotypes, X is an $n \times q$ matrix of fixed effects. β is a $q \times 1$ vector of the fixed effects coefficients. Z is an $n \times i$ incidence matrix from each observed phenotype to one of the i individuals. μ is the random effect of the mixed model with $Var(\mu) = \sigma_g^2 K$; where K is the $i \times i$ relationship matrix inferred from genotypes, and ϵ is an $n \times n$ matrix of residual effect such that $Var(\epsilon) = \sigma_\epsilon^2 I$, I is an identity matrice. Instead of solving this mixed model using the best linear unbiased prediction of random effect u, a direct estimate of dispersion parameters of a restricted maximum likelihood (REML) can be obtained. Thus, under Gauss-Markov assumptions using equation 1.37, it follows

$$\mu \sim N\left(0, \sigma_g^2 K\right) \text{ and } \epsilon \sim N\left(0, \sigma_\epsilon^2 I\right). \tag{1.38}$$

The restricted likelihood avoids a descending bias of maximum likelihood estimates of variance components by taking into account the loss in degrees of freedom associated with fixed effects. Under the null hypothesis, the full log-likelihood function can be written as

$$l_F\left(y; \beta, \sigma, \delta\right) = \frac{1}{2}\left[-n \log\left(2\pi\sigma^2\right) - \log|H| - \frac{1}{\sigma^2}\left(y - X\beta\right)' H^{-1}\left(y - X\beta\right)\right], \tag{1.39}$$

and the restricted log-likelihood function as

$$l_R\left(y; \sigma, \delta\right) = l_F\left(y : \hat{\beta}, \sigma^2, \delta\right) + \frac{1}{2}\left[-q \log\left(2\pi\sigma^2\right) + \log|X'X| - \log|X'H^{-1}X|\right],$$

assuming that $\delta = \sigma_\epsilon^2 / \sigma_g^2$ does not change appreciably in a GWAS scan. The model set $\eta = Z\mu + \epsilon$, so that equation 1.37 can be written as

$$y = X\beta + \eta$$

with

$$Var\left(\eta\right) = Var\left(Z\mu\right) + Var\left(\epsilon\right),$$

thus

$$Var\left(\eta\right) \propto \sigma_g^2 K + \sigma_\epsilon^2 I.$$

The overall phenotype variance-covariance matrix can be represented as σ_g and σ_ϵ, that maximizes the full likelihood

$$V = \sigma_g^2 ZKZ' + \sigma_\epsilon^2 I$$

To obtain the generalized least squares (GLS), equation (1.37) can be re-written as

$$y^* = X^*\beta + \epsilon^*, \tag{1.40}$$

Solving the equation (1.40) by the ordinary least squares (OLS), we have $y^* = M^{-1}y$, $X^* = M^{-1}X$, $\epsilon^* = M^{-1}\epsilon$, $\eta = MM'$.

Finally, the equation (1.40) is maximized when β is $\hat{\beta} = \left(X'H^{-1}X\right)^{-1} H'H^{-1}y$ and the optimal variance component is $\hat{\sigma}_F^2 = \frac{R}{n}$ for the full likelihood and $\hat{\sigma}_R^2 = \frac{R}{n-q}$ for the restricted likelihood, with $R = \left(y - X\hat{\beta}\right)^{-1} H^{-1}\left(y - X\hat{\beta}\right)$, a function of δ.

1.6.2.2 Genome-Wide Admixture Association

It was shown that the approaches described in subsection 1.6.2.1 above cannot control for confounding at the level of specific SNPs (Redden *et al.*, 2006). Therefore, since local-specific and genome-wide average ancestry are weakly correlated (Qin *et al.*, 2010), it was suggested to control for confounding due to admixture by conditioning on both local-specific and genome-wide average ancestry. An alternative to these approaches for low-penetrance risk variants for common human diseases is also admixture mapping (Excoffier & Hamilton, 2003; McKeigue, 2005; Zhu *et al.*, 2008). Admixture mapping extends to human populations the principles that underlie linkage analysis of an experimental cross (Hoggart *et al.*, 2004; Montana & Pritchard, 2004). It is known to currently be a low cost and powerful method for localizing disease genes in populations of recently mixed ancestry in which the ancestral populations have different genetic risk (Excoffier & Hamilton, 2003; Montana & Pritchard, 2004). It has been widely discussed as a potential strategy for localizing susceptible genes (Falush *et al.*, 2003; McKeigue, 2005; Pritchard *et al.*, 2002) based on admixture linkage disequilibrium. The theory behind admixture mapping has been outlined several years ago, its applications have been boosted by the availability of genome-wide panels of genetic markers informative for ancestry between worldwide human populations and statistical methods that combine information from these genetic markers to infer ancestry (Excoffier

& Hamilton, 2003; Hoggart *et al.*, 2004; Sankararaman *et al.*, 2008; Santafe *et al.*, 2006). The first attempt at admixture mapping was conducted with the recently admixed African-American population followed by the Mexican-American population, in which the founding populations are European, Native American and African (Excoffier & Hamilton, 2003; Patterson *et al.*, 2004; Zhu *et al.*, 2008). In addition, the information about population structure and local ancestry inference, are critically well known to be useful in admixture mapping studies of disease genes (Montana & Pritchard, 2004; Patterson *et al.*, 2004; Rosenberg & Pritchard, 2008). Current methods developed for disease scoring in admixed populations have succeeded in studying two-way admixed populations, but do not apply to multi-way admixed populations such as five-way admixed populations (Pasaniuc *et al.*, 2011; Rosenberg *et al.*, 2010).

1.7 Issues in Association Studies

Today, most of the associated SNPs resulting from both admixture and association studies explain only a small fraction of the genetic risk (small effect sizes) (Cantor *et al.*, 2010; Jia *et al.*, 2010). Many authors have pointed out that GWAS may not detect SNP with low or moderate risk that may not reach the intrinsic genome-wide significance cut-off (regions that met the statistical criteria of genome-wide association) of $P < 5 \times 10^{-8}$ (Jia *et al.*, 2010; Peng *et al.*, 2008). GWAS may fail to reveal a significant signal of a gene polymorphism, if the changing effect of a variant in another gene is not taken into account. Therefore, single discovery SNP-based analysis in GWAS may generate false negatives (Jia *et al.*, 2010; Peng *et al.*, 2008) or inconclusive results. Furthermore, the question still arises as to why so much of the heritability is apparently unexplained by GWA findings. This question is relevant because of a substantial proportion of individual differences in disease susceptibility. Understanding this genetic variation may contribute to diagnosis, treatment and prevention of disease (Manolio *et al.*, 2004). A number of explanations for this missing heritability have been suggested, including much rarer variants (possibly with larger effects) or variants of low minor allele frequency (MAF), defined roughly $0.5\% <$ MAF $< 5\%$, that are poorly detected by available genotyping arrays that focus on variants present in 5% or more of the population. A considerable number of variants of smaller effect yet to be found; structural variants poorly captured by existing arrays; low power to detect gene-gene interactions; and inadequate accounting for shared environment among relatives (Manolio *et al.*, 2004; Scheuner *et al.*, 2004).

In any case, the associated SNPs from single-SNP admixture and association studies will always provide preliminary genetic information available for additional analysis by statistical procedures that accumulate evidence (Cantor *et al.*, 2010; Peng *et al.*, 2008). In addition, considering

the multiple genetic and environmental factors that contribute to development of complex diseases, GWAS by itself, may be insufficient to examine complex genetic structure of complex diseases (Cantor *et al.*, 2010; Jia *et al.*, 2010; Peng *et al.*, 2008).

Another method such as analysis of epistasis, which uses a single GWAS study was introduced in order to identify stronger results that are revealed when genes interact (Anton *et al.*, 1998; Wu *et al.*, 2009). Moreover, suggestions have also been made to pursue sequencing studies in order to detect the contributions of rare variants to the same genetic disorders that the standard GWAS failed to detect (Dickson *et al.*, 2010). Rare variants are found in less than 1% of the population (Cantor *et al.*, 2010), however using large-scale sequencing, which is more financially feasible today, can provide additional information regarding the genetic etiology of complex disorders and may shed light on investigations of common and rare variants (Cantor *et al.*, 2010; Dickson *et al.*, 2010; Gronau *et al.*, 2011). The rare variant analyses will present a large number of statistical challenges, and should result in the development of interesting and useful methods that will reveal important results (Cantor *et al.*, 2010; Dickson *et al.*, 2010).

Post admixture and association analyses were also implemented to combine different results of GWAS to reveal larger effects in order to provide valuable information that will be useful for prioritizing the most important results (Han & Eskin, 2011; Wray *et al.*, 2010). To combine associations across different association studies, even when the original data are unavailable, meta-analysis is used. Meta-analysis pools information from multiple GWAS to increase the chances of finding true positives among the false positives (Cantor *et al.*, 2010; Han & Eskin, 2011). Examining the combined effects of genes by detecting genetic signals beyond single gene polymorphisms has the increasing benefit of fully characterizing the susceptible genes and the genetic structure of complex diseases (Jia *et al.*, 2010; Peng *et al.*, 2008). Therefore, incorporating both the association signal from GWAS and the available human protein-protein interaction (PPI) information may be helpful in testing the combined effects of SNPs and searching for significantly enriched sub-networks for a particular complex disease. This approach is proposed to present a new paradigm for GWAS (Jia *et al.*, 2010; Peng *et al.*, 2008) in order to elucidate the genetic susceptibility of disease. More details have been developed in Chapter 8.

Chapter 2

Proxy Ancestry Selection Method: Ancestral components of a South African multi-way Admixed Population

2.1 Introduction

2.1.1 Background and Motivation

Single nucleotide polymorphism data has become significantly more widespread over the last three years. The availability of the genome-wide multi-locus genotype profiles has fuelled long standing interest in analysing patterns of genetic variations to trace back the ancestry component of recently admixed human populations. Single Nucleotide Polymorphisms (SNPs) can represent a consistent class of individual differences in DNA, and high-frequency SNPs can shed light on the evolutionary history and migrations of recently admixed human populations (Rosenberg & Pritchard, 2008). In addition, the high-frequency SNPs can predict human population diversification, infer the ancestry-specific loci that can be utilized for more accurate genetic analysis of human complex diseases, and be useful for other population genetics problems (Nianjun *et al.*, 2006). In order to understand the genetic variation which could be observed at genetic marker locations within and among populations, the inference of both locus-specific ancestry (Baran *et al.*, 2012; Pasaniuc *et al.*, 2009; Patterson *et al.*, 2006; Price *et al.*, 2009b; Sankararaman *et al.*, 2008) and population structure (Alexander *et al.*, 2009; Falush *et al.*, 2003; Hoggart *et al.*, 2004; Patterson *et al.*, 2006) from the genotypes of single nucleotide polymorphisms is the crucial step. The inference of both locus-specific ancestry and genome-wide ancestry (global ancestry) and the imputation of missing genotypes in Genome-wide association studies (GWAS), utilize panels of reference ancestral populations based on place-of-origin, ethnic or continental affiliation

(Browning & Browning, 2009; Li *et al.*, 2012; Marchini & Howie, 2008). The availability of high-throughput genotype data from various populations may facilitate the choice of best proxy ancestry of a recently admixed population from a pool of reference populations. This choice is critical in both the study of population genetics and in identifying genes underlying ethnic difference in genetic diseases risk (Hoggart *et al.*, 2004; McKeigue, 2005; Seldin *et al.*, 2011; Winkler *et al.*, 2010). Furthermore, the accuracy of these inferences is in part related to the choice of reference populations. An insufficient or inaccurate ancestral proxy can weaken these inferences, resulting in erroneous inferred ancestry, and errors and uncertainty in the imputed genotypes. These issues may consequently affect the inference of ancestry and the detection power of GWAS and meta-analysis when using imputation, particularly in multi-way admixed populations.

2.1.2 Impact of Selecting Proxy Ancestry in both Estimating Ancestry and Imputing Missing Genotype in Admixed Populations

Because distinct populations exhibit substantial variation in genetic disease risk, the choice of reference populations for a multi-way admixed population may be sensitive and critical in biomedical research. Current algorithms for identifying the best proxy ancestral populations are inadequate for multi-way admixed populations. To address these challenges and the uncertainty in ancestral populations, we developed PROXYANC, an approach to select the proxy ancestry for recently admixed populations. We implemented two novel algorithms in PROXYANC, based on population genetic differentiation and optimal quadratic programming, respectively. We demonstrated through simulation of a complex multi-way admixed population that these two algorithms can select the best proxy ancestry for an admixed population given a pool of groups of related/unrelated or admixed reference populations. Our simulation demonstrated that our complementary algorithms have the advantage to precisely select the best proxy ancestry for a multi-way admixed population more accurately than the $f3$ statistic (Patterson *et al.*, 2012). We additionally demonstrated the impact of choosing the best proxy ancestral populations in both estimating admixture proportion and imputing missing genotypes in a multi-way admixed population.

2.1.3 The SAC Provides an Ideal Population to Study the Choice of Best Proxy Ancestry

The South African Coloured population (SAC) has a high level of intercontinental admixture and therefore a diverse ancestry (Davis & Dollard, 1994; Mountain, 2003; Tishkoff *et al.*, 2009). Historical sources (section 1.1.1) and a few genetic studies reported that this population is the result of unions between Europeans, African (Bantu-speaker and Click-speaker groups), and

various other population groups of Indian or Indonesian descent (Botha, 1972; deWit *et al.*, 2010a; Ross, 1993; Tishkoff *et al.*, 2009). A study conducted by (Tishkoff *et al.*, 2009) on the characterization of the genetic variation and the relationships among populations across the African continent, revealed that the ancestral components in the SAC include nearly equally high levels of southern African San, Niger-Kordofanian), Indian, European, and lower levels of East Asian ancestry (Tishkoff *et al.*, 2009). However, their study used 39 samples from a subgroup of the SAC, possibly including Cape Malays (deWit *et al.*, 2010a). Based on 20 samples from the SAC population, a study by Patterson *et al.* (2009) showed that there is substantial genetic contribution from at least four distinct population groups in the SAC including Europeans, South Asians, Indonesians and a population genetically close to the isiXhosa, the sub-Saharan Bantu. Quintana-Murci *et al.* (2010) examined the gender-specific ancestry contributions in the SAC, using mitochondrial DNA ($n = 563$) and Y-chromosome ($n = 228$) variation analysis. Recent studies that include mtDNA, Y-chromose and autosomal results of different samples of the SAC, including Pickrell *et al.* (2012); Schlebusch *et al.* (2012), have globally inferred at least five different ancestral populations (Clicking-speaker, Bantu-speakers, Europeans, Indians, and South-East Asians) (Quintana-Murci *et al.*, 2010). An early in depth investigation by deWit *et al.* (2010a) was done which had the advantage of using a very large cohort of the SAC (959 samples) and $75,000$ autosomal single nucleotide polymorphisms (SNPs) common to HapMap and Human Genome Diversity Project (HGDP) data sources. The study exploited both subsets of selected random SNPs and ancestry informative markers (AIMs) from $75,000$ autosomal SNPs, to address the question of ancestry contribution in the SAC. This early investigation used a small sample of San (5 samples obtained from HGDP), and no suitable ancestral population samples from local southern African populations, and showed four major inferred contributions to the SAC with the greatest from San (click-speaker group) Africans, followed by non-clicker-speaker Africans, Europeans and a smaller East Asian contribution (deWit *et al.*, 2010a). However, the low San sample size may have biased the estimate of the ancestry contributions. Overall, these recent investigations have documented the genome-wide average admixture proportions in the SAC to be in the range of 23% to 65% for African, 19% to 40% for European, and 7% to 10% for East Asian, with some regional variation, and also with substantial variation among individuals. These variations at genetic loci commonly exhibit geographic structure and may contribute to phenotypic differences between populations (Campbell & Tishkoff, 2008). While different authors have focused on the global admixture (continental admixture) underlying the genetic origin of the SAC, attention has not yet been paid to which specific continental populations or ethnic groups contributed to the admixture. In addition, recent studies demonstrated the existence of diversity among both African Bantu-speaker and Clicking-speaker populations (Pickrell *et al.*, 2012; Schlebusch *et al.*, 2012; Tishkoff *et al.*, 2009), which for example make sensitive the choice

of the best reference African ancestral group for the SAC. The sensitive choice of reference ancestral populations affects admixture mapping methods, the imputation of missing genotype and estimating both global and local ancestry in multi-way admixed populations.

2.1.4 Study Overview

In this chapter, we develop PROXYANC, an approach to choose the best proxy ancestry for multi-way admixed populations. PROXYANC makes use of two novel algorithms including the correlation between observed linkage disequilibrium in an admixed population and population genetic differentiation in ancestral populations, and an optimal quadratic programming based on the linear combination of population genetic distances (F_{ST}). We validate these algorithms through the simulation of a multi-way admixed population, and assess the impact of choosing the best proxy ancestral populations in both estimating admixture proportion and imputing missing genotypes in a multi-way admixed population. We applied this approach for downstream analysis in a uniquely admixed Coloured population from South Africa. We characterized the African, European, East and South Asian origins of the SAC by applying PROXYANC to a cohort of the SAC (764 unrelated individuals) and refining the contributions of genetic ancestry components. We established that the SAC has had a substantial admixture from isiXhosa, ǂKhomani, Central European, Indian (Gujarati) and Chinese populations. Using the estimated best proxy ancestral populations of the SAC, we demonstrated that the ancestral allele frequency differences correlated with increased linkage disequilibrium (LD) in the SAC, indicating that increased admixture LD is present in this population, and the observed LD has its origin from admixture events. This result supports the rejection of the evidence of founder effects or of population bottlenecks that could be due to the racial segregation of the past, formalized during the recent apartheid regime in South Africa (http://www.sahistory.org.za/pages/chronology/special-chrono/governance/apartheid-legislation.html).

2.2 Materials and Methods

2.2.1 Samples, Genotype Data and Genotype Quality Control

The South African Coloured (SAC) population under study is located in the metropolitan area of Cape Town in the Western Cape Province in South Africa (Hoal *et al.*, 2004). Since the ethnicity, socio-economic status and HIV infection may be confounders in TB association studies (Stein, 2011), this area was selected due to the high incidence of TB as well as the uniform ethnicity, socio-economic status and low prevalence of HIV (Hirschhorn & Daly, 2003). This is due to the follows reasons:

(1) Uniform ethnicity and socio-economic status is important in disease association studies as it removes some of the confounding variables.

(2) Low prevalence of HIV is important because in the presence of HIV infection, an individual has a greatly increased chance of progressing to TB disease once infected, simply because of an impaired immune system, and not necessarily because of genetic susceptibility.

The definition for TB diagnosis and recruitment of appropriate controls for infectious diseases such as TB has been shown to be important in the interpretation of GWAS results (Stein, 2011). Therefore, TB patients were identified through bacteriological confirmation (smear positive and/or culture positive). Controls were selected from the same community living under the same conditions including socio-economic status and availability of health facilities. These healthy individuals had no previous history of TB disease or treatment. Approval from the Ethics Committee of the Faculty of Health Sciences, Stellenbosch University (project number $95/072$) was obtained before blood samples were collected with informed consent, and known HIV positive individuals were excluded from the study. The collective term for people of mixed ancestry in southern Africa is Coloured, and this is officially recognized in South Africa as a census term, and for self-classification. Whilst we acknowledge that some cultures may use this term in a derogatory manner, these connotations are not present in South Africa, and are certainly not intended here.

The study samples were genotyped on the Affymetrix $500K$ chip and SNP calling was done as described by deWit *et al.* (2010a). Quality-control filters were applied to the $500K$ Affymetrix data from 797 cases and 91 controls. A total of $6,450$ SNPs failed the minor allele frequency ($MAF < 1\%$)and missingness test ($GENO > 0.05$), as well as the HardyWeinberg equilibrium (HWE) test in controls (alpha level 0.0001). Outliers, related individuals and individuals with a genotyping rate of less than 95% were then removed. We retained $390,887$ SNPs for 888 individuals (381558 autosomal SNPs; 797 cases and 91 controls; 489 males, of which 444 are cases and 45 controls) to be used in the association study in chapters 5 and 6. Further relatedness analysis using PLINK (Purcell *et al.*, 2007) was conducted and resulted in the removal of 155 related individuals, producing a data set suitable for methods that assume independent samples. It has $390,887$ SNPs for 733 individuals ($381,558$ autosomal SNPs, 642 cases and 91 controls; 406 males of which 361 are cases and 45 controls). To evaluate whether controls are genetically similar to cases except for the presence of TB, we performed PCA analysis on the resulting data set (Chapters 4). To further check the homogeneity of the samples, we additionally performed the identity-by-state (IBS) permutation test, where case-control labels were permuted, and then recalculated between group metrics based on average IBS (fixed $10,000$ permutations).

To examine the choice of best proxy ancestry in multi-way admixed populations, this chapter used the samples of 733 unrelated South African Coloured individuals. A total of 77 samples from local southern African Bantu (isiXhosa, Sotho-Tswana, Zulu and Herero), and 23 indigenous San individuals from Namibia, genotyped on Affymetrix 6.0 are used. Additionally, genome-wide SNP data from three public data sources, including the Human Genome Diversity Cell Line Panel (http://hagsc.org/hgdp/files.html) (Cann *et al.*, 2002), the International Haplotype Map (http://hapmap.ncbi.nlm.nih.gov/) Phase 3 project (Frazer & et al, 2007), and additional African populations from (Henn *et al.*, 2011) are also included. Detailed information about the number of individuals included in this analysis is provided in Table 2.1. Quality-control filters on each reference population is separately performed using PLINK (Purcell *et al.*, 2007), resulting in removal of SNPs that failed the Hardy-Weinberg exact test $P < 0.000001$ and have a call rate $> 95\%$ across all samples per population. Population outliers and unknown relatedness are assessed using the smartpca program implemented in EIGENSOFT (Patterson *et al.*, 2006; Price *et al.*, 2006). After applying the quality-control filters to each population separately, the SNPs genotyped in this chapter are reduced to a subset $(n = 49{,}930)$ shared between the SAC, the three public data sources and the local southern Bantu from South Africa (Table 2.1). Grouping each population per continent, the African, European, South Asian, East Asian and Middle East sets, were merged in one data with the data of the SAC.

Table 2.1: **List of putative ancestral populations that were included in population genetic structure analysis of the South African Coloured population.**

Pop. Label	Source	Pop Location	Individuals
Admixed South African Coloured			
sac	deWit *et al.* (2010a)	South Africa Coloured population	764
African: non Click-speaker			
moz	HGDP	Mozabite-Algeria	9
yor	HGDP	Yoruba in Ibadan-Nigeria	21
man	HGDP	Mandenka-Senegal	24
bpg	HGDP	Biaka,Pygmy-Central Africa	21
mpg	HGDP	MbutiPygmy-Congo	12
kaba	HGDP	North of the Central African	17
fang	HGDP	Equatorial-Bantu	15
fulani	HGDP	West -central Africa	2
bulala	HGDP	Central Chad	12

Continued on next page

Table 2.1 – continued from previous page

Pop. Label	Source	Pop Location	Individuals
mada	HGDP	Cameroon	12
hausa	HGDP	West Africa Niger and Nigeria	12
bamoun	HGDP	Cameroon	18
kongo	HGDP	Atlantic coast of Congo	9
brong	HGDP	Ghana	8
lwk	HapMap3	Luhya in Webuye, Kenya	104
mkk	HapMap3	Maasai in Kinyawe, Kenya	108
yri	HapMap3	Yoruba in Ibadan, Nigeria	147
Igbo	(Henn *et al.*, 2012)	Southeastern Nigeria	15
man	HapMaP3	Mandenka from Africa	22
African: Local South African Populations			
san	HGDP	Jul'huan, Namibia	5
khs	(Chimusa *et al.*, 2013)	Jul'huan, Namibia	22
kho	(Henn *et al.*, 2012)	‡Khomani, South Africa	8
zul	(Chimusa *et al.*, 2013)	Zulu-South-Africa	18
sts	(Chimusa *et al.*, 2013)	Sotho-Tswana,South Africa	24
xhs	(Chimusa *et al.*, 2013)	Xhosa-South-Africa 20	
her	(Chimusa *et al.*, 2013)	Herero, South Africa-Namibia	14
had	(Henn *et al.*, 2011)	Hadza,Tanzania	17
bus	(Henn *et al.*, 2011)	Bushmen, South Africa	16
African: Afroasiatic			
tns	(Henn *et al.*, 2011)	Berber from Tunisia	18
European			
bas	HGDP	Basque-France	24
sar	HGDP	Sardinian-Italy	27
ita	HGDP	Italian-Italy-Bergamo	13
orc	HGDP	Orkney-Islands	14
fre	HGDP	French-France	29
ady	HGDP	Adygei-Russia-Caucasus	15
rus	HGDP	Russian-Russia	24
ceu	HapMap3	Northern European	112
East Asia			
mia	HGDP	Miao-China	10
jap	HGDP	Japanese-Japan	28
nax	HGDP	Naxi-China	9

Continued on next page

Table 2.1 – continued from previous page

Pop. Label	Source	Pop Location	Individuals
dai	HGDP	Dai-China	10
yi	HGDP	Yi-China	10
tuj	HGDP	Tujia-China	10
she	HGDP	She-China	10
lah	HGDP	Lahu-China	7
oro	HGDP	Oroqen-China	10
uyg	HGDP	Uygur-China	9
hez	HGDP	Hezhen-China	9
yak	HGDP	Yakut-Siberia	19
dau	HGDP	Daur-China	9
xib	HGDP	Xibo-China	9
tuu	HGDP	Tu-China	10
mon	HGDP	Mongola-China	10
cam	HGDP	Cambodian-Cambodia	11
chb	HapMap3	Han-Chinese in Beijing	137
chd	HapMap3	Chinese in Denver,Colorado	109
jpt	HapMap3	Japanese in Tokyo	113
South Asia			
han	HGDP	Han-Chinese	43
bra	HGDP	Brahui-Pakistan	23
bal	HGDP	Balochi-Pakistan	23
mak	HGDP	Makrani-Pakistan	22
kal	HGDP	Kalash-Pakistan	25
pat	HGDP	Pathan-Pakistan	23
sin	HGDP	Sindhi-Pakistan	25
bur	HGDP	Burusho-Pakistan	23
haz	HGDP	Hazara-Pakistan	22
Gih	HapMap3	Gujarati Indians in Texas	93
Middle East			
bed	HGDP	Bedouin-Israel-Negev	35
dru	HGDP	Druze-Israel-Carmel	26
qatari	(Henn *et al.*, 2012)	Qatar	22
pal	HGDP	Palestinian-Israel-Central	40

2.2.2 PROXYANC: F_{ST}-optimal Quadratic Cone Programming

The question we want to address is, given a pool of continental affiliated (Europe, Africa, etc.) populations, which population for example can be the best European, African, etc. proxy ancestry of the admixed population under study. To limit the effect of background linkage disequilibrium, let us assume adjacent SNPs in each populations are spaced 10 Kb from each other. Let denote Z a set of pools (set) of distinct reference ancestral populations. Suppose we have SNP j, let N_j and p_j be the total variant allele count and observed population allele-frequency in the admixed population (*Mix*), and N_{jk} and p_{jk} be the total variant allele count and the population observed allele-frequency in reference populations $k = 1, 2, \ldots, K$ of unrelated individuals. Given different combinations C of $L = |Z|$ reference populations of unrelated individuals from each pool $S_i \in Z, (i = 1, \ldots, |Z|)$. Each combination C of $|Z|$ reference populations can be obtained from a set of Cartesian product $T = \prod_i^{|Z|} S_i, C \subseteq Z$. Thus, from each $C \subseteq Z$ we construct a synthetic populations consisting of L populations as the follows linear combination,

$$p_{j\alpha} = \sum_{k=1}^{L} \alpha_l p_{jl}, \tag{2.1}$$

where α_l is the ancestral proportion. A particular combination of L populations (synthetic admixed population) consists of best proxy ancestries of *Mix* if their linear combination (in equation 2.1) minimizes the $F_{ST}(Mix, p_{j\alpha})$ (in equation 2.2). This problem is related to an optimal quadratic cone programming, where the objective function (F_{ST}) is given by,

$$F_{ST}^{j}(\alpha) = \left[(p_{j\alpha} - p_j)^2 - p_j \frac{(1-p_j)}{N_j} - \sum_{l=1}^{L} \alpha_l^2 p_j \frac{(1-p_j)}{N_{jl}} \right] \times \frac{1}{p_j(1-p_j).L}, \tag{2.2}$$

at SNP j. Subject to $\sum_{l=1}^{L} \alpha_l = 1$ and

$$\alpha_l \leqslant 0, \ \forall l \in \{1, \ldots, L\}.$$

Equation 2.2 is a generalization form of the one described in (Price *et al.*, 2009a), and is a quadratic convex function with respect to α_l (ancestry proportion), therefore a global minimum can be found. To obtain a matrix representation of the optimal cone programming, equation 2.2 can be expanded. Let us denote $C_1 = \frac{1}{p_j(1-p_j)K}$, $C_2 = p_j(1-p_j)$, and $C_3 = p_j \frac{(1-p_j)}{N_j}$. Thus, equation 2.2 becomes,

$$F_{ST}^{j}(\alpha) = \left[(p_{j\alpha} - p_j)^2 - C_3 - \sum_{l=1}^{L} \frac{\alpha_l^2}{N_{jl}} C_2 \right] \times C_1. \tag{2.3}$$

It follows,

$$F_{ST}^{j}(\alpha) = \left[p_{j\alpha}^2 - 2p_{j\alpha}p_j + \underbrace{p_j^2 - C_3}_{C_4} - \sum_{l=1}^{L} \frac{\alpha_l^2}{N_{jl}} C_2 \right] \times C_1. \tag{2.4}$$

Substituting equation 2.1 into equation 2.4, we obtain,

$$F_{ST}^{j}(\alpha) = \left[(\sum_{l=1}^{L} \alpha_l p_{jk})^2 - 2\sum_{l=1}^{L} \alpha_l p_{jl} p_j + C_4 - \sum_{l=1}^{L} \frac{\alpha_l^2}{N_{jl}} C_2 \right] \times C_1. \tag{2.5}$$

Now expanding equation 2.5, using a squared finite sum, $(\sum_{l=0}^{L} x_l)^2 = \sum_{l=0}^{L} x_l^2 + \sum_{l \neq n} x_l x_n$, $s.t$ x is a variable, it follows,

$$\begin{aligned} F_{ST}^{j}(\alpha) &= \left[\sum_{l=1}^{L} \alpha_l^2 p_{jk}^2 + \sum_{l \neq n} (\alpha_l \alpha_n) p_{jl} p_{jn} - 2\sum_{l=1}^{L} \alpha_l p_{jl} p_j + C_4 - \sum_{l=1}^{L} \frac{\alpha_l^2}{N_{jl}} C_2 \right] \times C_1 \\ &= \left[\sum_{k=1}^{L} \alpha_l^2 (p_{jl}^2 - \frac{C_2}{N_{jl}}) + \sum_{l \neq n} (\alpha_l \alpha_n) p_{jl} p_{jn} - 2\sum_{l=1}^{L} \alpha_l p_{jl} p_j + C_4 \right] \times C_1. \end{aligned} \tag{2.6}$$

Knowing that the ancestral proportion must sum to 1, $\sum_{l=1}^{L} \alpha_l = 1$ then

$$\sum_{l=1}^{L} \alpha_l C_4 = C_4,$$

equation 2.6 becomes,

$$\begin{aligned} F_{ST}^{j}(\alpha) &= \left[\sum_{l=1}^{L} \alpha_l^2 (p_{jl}^2 - frac C_2 N_{jl}) C_1 \right] + \left[\sum_{l \neq n} (\alpha_l \alpha_n) p_{jl} p_{jn} C_1 \right] - 2\sum_{l=1}^{L} \alpha_l p_{jl} p_j C_1 + \sum_{l=1}^{L} \alpha_l C_4 C_1 \\ &= \left[\sum_{l=1}^{L} \alpha_l^2 (p_{jl}^2 - \frac{C_2}{N_{jl}}) C_1 \right] + \left[\sum_{l \neq n} (\alpha_l \alpha_n) p_{jl} p_{jn} C_1 \right] + \left[\sum_{l=1}^{L} \alpha_l (C_4 - 2p_{jl} p_j) C_1 \right]. \end{aligned} \tag{2.7}$$

Therefore, the matrice representation of the optimal Cone Programming can be obtained as follows,

$$min_\alpha = \left(\frac{1}{2} \alpha^T P \alpha + q^T \alpha \right) \text{ subject to } -\alpha \leqslant 0 \text{ and } \sum_{l=1}^{L} \alpha_l = 1, \tag{2.8}$$

where α is a vector of L-dimensions of unknown ancestry proportions, G is an identity vector of L-dimensions, A is a vector of allele frequencies of L-dimensions, P is a positive semi definite matrice, and its diagonal elements are all coefficients of α^2:

$$(\alpha^2)_l = 2\frac{p_{jl}^2 - \frac{p_j(1-p_j)}{N_{jl}}}{p_j(1-p_j)L}, \tag{2.9}$$

and the mixture coefficients $\alpha_l\alpha_n$ consist of its symmetric elements, and are given by:

$$(\alpha)_{ln} = 2\frac{p_{jl}p_{jn}}{p_j(1-p_j)L}, \qquad \text{for } k \neq n, \tag{2.10}$$

and the linear coefficients α_l are the elements of vector q in equation 2.8, and are represented by:

$$(\alpha)_l = \frac{(p_j^2 - p_j\frac{(1-p_j)}{N_j} - 2p_{jl}p_j)}{p_j(1-p_j)L}. \tag{2.11}$$

For the optimization of the equation (3) or (2) with respect to α_l (ancestry proportions, $l = 1, \ldots, L$), the matrix form in equation (3) is constructed by summing equations (2), (4), (5) and (6) independently across all SNPs.

2.2.3 PROXYANC: Proxy-Ancestry Score

When admixture occurs between two or more previously isolated populations with differences in allele frequency, admixture creates linkage disequilibrium (LD) between genetic loci. Accounting for this assumption, we can compute the proxy ancestry score from the data of the admixed population and pair-wise reference populations. Computing the correlation between the LD in the admixed population and allele frequency differentiation in each pair of ancestral populations, the Proxy-Ancestry Score algorithm is as follows:

(1) Given N samples from the data of the admixed population and the data of K groups of reference populations without missing genotypes data, we compute the expected squared correlation ρ^2 between diploid genotype at each pair of SNPs S_i and s_j, $(i \neq j)$.

$$\rho^2_{S_i,S_j} = \frac{\overline{COV(S_i,S_j)}}{\overline{var(S_i)} \times \overline{var(S_j)}}.$$

Taking the Fisher's transformation on ρ^2,

$$y = \frac{1}{2}\log\left(\frac{1+\rho^2}{1-rho^2}\right), \tag{2.12}$$

thus, we compute the LD for each pair of SNPs located at distances (< 0.2 Morgans),

$$L\left(s_i, s_j\right) = \frac{y}{\sqrt{N-3}}, \tag{2.13}$$

(2) For each different pair of reference populations, we compute the allele frequency difference $d(s_i)$ and $d(s_j)$, respectively.

(3) We regress $L(s_i, s_j) \sim d(s_i) \times d(s_j)$, and obtain p-value p^n, $n = 1, \ldots, N$.

(4) For $n = 1, \ldots, N$ possible combinations of each reference population (k) with other reference ancestral populations, we compute the inverse normal distribution ϕ^{-1}

$$p_k^n = \phi^{-1}\left(1 - p^n\right), \tag{2.14}$$

using the p-value obtained in the previous step. In this way, a smaller p-value corresponds to a larger p_k^n.

(5) Thus, for each reference population $k = 1, \ldots, K$, we compute the proxy ancestry score as follows,

$$p_k^{score} = \sum \frac{p_k^n}{\sqrt{K}}. \tag{2.15}$$

(6) To determine whether the proxy ancestry score in equation 2.15 is higher than expected, we normalized it. To address this we consider a vector of all proxy ancestry score $V = (p_1^{score}, \ldots, p_{k-1}^{score}, p_{k+1}^{score}, \ldots, p_K^{score})$ excluding p_k^{score}, and we compute the normalization of it as follows,

$$Z_k = \frac{p_k^{score} - mean(V)}{\sqrt{var(V)}}. \tag{2.16}$$

The algorithms in sections 2.2.2 and 2.2.3 are implemented in the PROXYANC programme (http://www.cbio.uct.ac.za/PROXYANC)

Both models described assume prior knowledge of geographical potential ancestral populations. Both models tackle the following problem: Given a pool of geographical potential ancestral populations, for example given a pool of European/African populations, which population is the best European/ African proxy ancestry of the admixed population under study.

2.2.4 Experimental Admixed Data to Evaluate PROXYANC

To start our simulation, we independently phased each putative ancestral population. From these phased putative ancestral populations using BEAGLE (Browning & Browning, 2009), we chose the following five as parental populations for the simulated population: European (CEU), isiXhosa, Khomani, East Asia (CHD) and Gujarati Indian.. To generate k diploid admixed individuals, our simulation framework uses $2k$ ancestral haplotypes, where k should be the minimun sample size among the parental populations. Therefore, we independently expanded each putative ancestral population following Rogers and Harpendings (1992) model of exponential population growth. We implemented this model using three parameters,$\theta_0 = 2 * N_0 * \mu$, $\theta_1 = 2 * N_1 * \mu$ and $\tau = 2 * \mu * t$ to a total size of 1500 plus its original size. An initial population of effective size N_0, is assumed to grow exponentially to a new size of N_1 at a time t generations back from the present. The mutation rate μ, is the per-generation probability that a mutation strikes a random nucleotide along the genome. From each expanded ancestral population, we split the resulting samples in two separate groups. 1500 samples from each of these reference populations were used to simulate admixed individuals and the remaining samples were dropped. Thus, the original population samples were used to test PROXYANC.

To simulate the genome of an admixed individual that can mimics the genetic make-up of the SAC, we sample haplotypes from European (CEU), isiXhosa, ‡Khomani, East Asia (CHD) and Gujarati Indian with probability related to a given ancestral proportion from each putative ancestral population (20%, 32%, 29%, 8% and 11%, respectively). These ancestral proportions are chosen to mimic the genetic structure of the SAC. Considering a continuous gene flow model (Price *et al.*, 2009b), in 100 generations and accounting for the Wright-Fisher model with random mating, from the beginning to the end of each chromosome, the ancestry is re-sampled using related ancestral proportion above, at each SNP in order to identify the occurrence of the admixture event. Following this process, the chromosomal segment of ancestral population is copied to the genome of the admixed individual, and record the locus-specific ancestry (the true ancestry) which will serve to assess the estimated ancestry. Using this procedure, we simulated the genomes of 750 individuals of mixed ancestry from Europeans (CEU), isiXhosa, ‡Khomani, East Asia (CHD) and Gujarati Indian.

To evaluate PROXYANC, we applied both approaches implemented in PROXYANC (FST-optimal quadratic cone programming and proxy-ancestry score) to select the best ancestral proxy for the above simulated data. Since, the true number of ancestral populations is known, one can choose closely related or geographically close populations to the true ancestral populations or do a pre-population structure analysis. Here, we use a pool of 20 reference populations geographically close to the true ancestors, including CEU, Italian, French, Russian, Gujarati, Pathan, Druze, isiXhosa, Zulu, Herero, Kongo, Yoruba, ‡Khomani, Jul'huan, San, Bushmen, dai, Chinese(CHD)

Japanese (JPT) and Daur. Particularly, for these five putative ancestral populations (CEU, isiXhosa, ‡Khomani, East CHD and Gujarati) used in our simulation framework, we used the initial samples that were not used in the simulation of the admixed population. To evaluate the impact of selecting the best proxy ancestral populations for an admixed population in estimating admixture proportions, we separately ran the ADMIXTURE software (Alexander *et al.*, 2009) on the simulated data together with the expanded and initial samples from ancestral populations (CEU, isiXhosa, ‡Khomani, CHD and Gujarati Indian), respectively (as described above). We again ran ADMIXTURE on the simulated data together with a panel that included reference populations that are geographically close to the selected proxy ancestral populations, including Russian, Japanese, Palestine, Yoruba and Jul'huan. This allowed us to assess the estimated admixture proportions versus the true proportions.

To investigate if a restricted panel of only the best chosen proxy ancestral of an admixed population can be useful in imputing accurately missing genotypes, as it is been the case for using all available reference populations, we assess the impact of selecting the best reference ancestral populations in imputing missing genotypes of an admixed population, we removed $2,044$ out of $39,064$ SNPs on chromosome 1 from the simulated data, and we imputed them using 4 different sets of reference populations, including a panel of populations (CEU, CHD, GIH, isiXhosa, ‡Khomani) used directly in the simulation. This panel was used to test PROXYANC, a panel of all 20 populations listed above, and a panel formed by the Russian, Japanese, Palestinian, Yoruban and Jul'huan populations. This allowed us to assess the genotype call rate after the imputation using these different reference panels.

2.2.5 Admixture and Principle Component Analysis

In order to identify the ancestral populations that have contributed through admixture to the SAC and simulation data, we applied the algorithm implemented in ADMIXTURE (Alexander *et al.*, 2009) to determine the ancestral population clustering on a world-wide data set, which includes African, European, South Asian, East Asia and Middle East populations merged with the SAC data. Furthermore, once the proxy ancestral populations for the SAC and simulation data are selected using PROXYANC, we construct a merged data set of the SAC and its proxy ancestral populations, then ADMIXTURE (Alexander *et al.*, 2009) is run to estimate the admixture proportions in this population (the same for the simulated data). Averaging the SAC's individual admixture proportion, we obtained the genome-wide population admixture proportion (ancestry contribution). The DISTRUCT program (Rosenberg, 2004) was applied on the resulting Q-matrices from ADMIXTURE to plot the results from real and simulation data of admixed populations. In order to perform principal component analysis (PCA) to evaluate the extent of

substructure of the South African Coloured population, the smartpca programme in the EIGENSOFT package was applied to merged data sets of the SAC and the world-wide populations (African, European, South Asian, East Asia and Middle East populations), with the proxy ancestral populations, respectively.

2.3 Results and Discussion

2.3.1 Evaluation of PROXYANC Algorithms

We developed the method PROXYANC (http://www.cbio.uct.ac.za/proxyanc), that searches for a best combination of reference populations that can minimize the genetic distance (using F_{ST} as the objective function of ancestral proportions as variables through an optimal quadratic cone programming algorithm) between the admixed population and all possible synthetic populations, consisting of a linear combination from reference populations (section 2.2.2). In the same vein, PROXYANC also computes a proxy-ancestry score by regressing a statistic for LD (at short distance < 0.25 Morgan) between a pair of SNPs in the admixed population against a weighted ancestral allele frelscapequency differentiation (section 2.2.3). To evaluate PROXYANC, we mimic a 5-way admixture scenario by simulating (see section 2.2.1) the genomes of 750 individuals of mixed ancestry through the haplotype samples from Europeans (CEU), ‡Khomani, isiXhosa, Chinese (CHD) and Gujarati Indian with probability related to a given ancestral proportion from each putative ancestral population 20%, 32%, 29%, 8% and 11%, respectively.

We performed both approaches implemented in PROXYANC to select the best ancestral proxies for the above simulated data using 5 distinct pools of reference populations, including African non-Click speaking group (isiXhosa, Zulu, Yoruba, Kongo, Herero), South Asia (Gujarati, Pathan, Druze), East Asia (CHD, Dai, Daur, Japanese), European (CEU, Russian, Italian, French) and click-speaker groups (‡Khomani, Jul'huan, Bushmen, San). From each pool, our algorithms have to select the best ancestral population for our simulated data. The result from the simulation demonstrates the highest proxy-ancestry scores (Table 2.2) are from the five reference populations that contributed to the admixture in the simulated data (Figure 2.1).

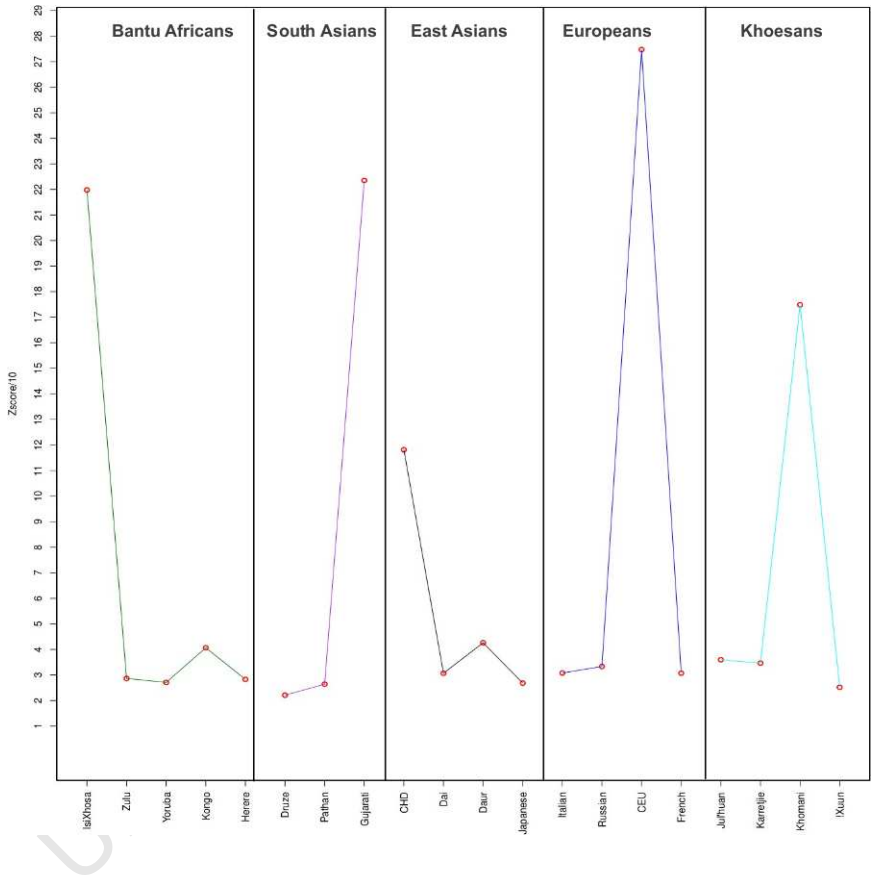

Figure 2.1: **Plot of proxy-ancestry scores of each population in each group of reference populations. All the highest peaks can be observed from the five ancestral populations that contributed to the admixture in the simulated data.**

In addition, among different linear combinations of five reference populations, the linear combination formed from the five populations used in our simulation (CEU, ‡Khomani, isiXhosa, Chinese and Gujarati) minimizes the genetic distance (F_{ST}) within the simulated data (Table 2.3).

Table 2.2: **Proxy-ancestry score for 5 distinct pools, including African (isiXhosa, Zulu, Yoruba, Kongo, Herero), South Asia (Gujarati, Pathan, Druze), East Asia (CHD, Dai, Daur, Japanese), European (CEU, Russian, Italian, French) and click-speaker groups (‡Khomani, Jul'huan, Bushmen, San) using the simulated data. The result shows that highest scores are from CEU, ‡Khomani, isiXhosa, Chinese (CHD) and Gujarati in the pools.**

Populations	PScore	Standard Error	Z
African non-Click Speakers Group			
isiXhosa	−0.124	1.138	219.793
Zulu	−0.015	0.001	28.648
Yoruba	−0.010	0.001	27.101
Kongo	−0.008	0.001	40.658
Herero	−0.008	0.001	28.306
South Asia Group			
Gujarati	0.015	0.007	223.504
Pathan	−0.007	0.001	26.427
Druze	−0.008	0.001	22.115
East Asia Group			
CHD	−0.001	0.003	118.144
Dai	−0.008	0.001	30.695
Daur	−0.007	0.001	42.628
Japanese	−0.008	0.001	26.847
European Group			
CEU	0.019	0.009	274.700
Russian	−0.008	0.001	33.347
Italian	−0.008	0.001	30.793
French	−0.008	0.001	30.716
African click-speaker Group			
‡Khomani	0.010	0.007	174.846
Jul'huan	−0.007	0.001	35.968
Bushmen	−0.007	0.001	34.664
San	−0.008	0.001	25.196

Table 2.3: **Top 16 linear combinations that minimize the F_{ST} between simulated data and a combination of 5 reference populations. The top linear combination is CEU, ‡Khomani, isiXhosa, Chinese (CHD) and Gujarati, consistent with Table 2.2 and with our simulation scheme.**

Population Linear Combination	F	Standard error	$95\%CI$
(isiXhosa, Gujarati, CHD, CEU, ‡Khomani)	-0.00075	0.0005599	$(-0.001, 0.0005)$
(isiXhosa, GIH, CHD, CEU, San)	-0.00058	0.0005599	$(-0.001, 0.0005)$
(isiXhosa, GIH, CHD, Italian, San)	-0.00057	0.0005599	$(-0.001, 0.0005)$
(isiXhosa, GIH, CHD, Italian, ‡Khomani)	-0.00054	0.0005599	$(-0.001, 0.0005)$
(isiXhosa, GIH, Japanese, Italian, San)	-0.00053	0.0005586	$(-0.001, 0.0005)$
(isiXhosa, GIH, Japanese, Italian, ‡Khomani)	-0.00054	0.0005586	$(-0.001, 0.0005$
(isiXhosa, GIH, Japanese, CEU, San)	-0.00051	0.0005585	$(-0.001, 0.0005)$
(isiXhosa, GIH, Japanese, CEU, ‡Khomani)	-0.00054	0.0005586	$(-0.001, 0.0005)$
(Yoruba, GIH, CHD, Italian, San)	-0.000371	0.0001110	$(-0.0005, -0.0001)$
(Yoruba, GIH, CHD, Italian, ‡Khomani)	-0.000361	0.0001110	$(-0.0005, -0.0001)$
(Yoruba, GIH, CHD, CEU, San)	-0.000371	0.0001110	$(-0.0005, -0.0001)$
(Yoruba, GIH, CHD, CEU, ‡Khomani)	-0.000372	0.0001110	$(-0.0005, -0.0001)$
(Yoruba, GIH, Japanese, Italian, San)	-0.000362	0.0001085	$(-0.0005, -0.0001)$
(Yoruba, GIH, Japanese, Italian, ‡Khomani)	-0.000365	0.0001085	$(-0.0006, -0.0001)$
(Yoruba, GIH, Japanese, CEU, San)	-0.000362	0.0001085	$(-0.0005, -0.0001)$
(Yoruba, GIH, Japanese, CEU, ‡Khomani)	-0.000362	0.0001085	$(-0.0005, -0.0001)$

Our result demonstrates that the selected proxy ancestries are in agreement and consistent with the ancestral populations used to generate these 750 admixed individuals (simulation data). The higher the proxy score is the more likely it is that the related reference population is a good proxy ancestry. To compare our algorithms to the $f3$ statistic (Patterson et al. 2012), which is a 3-population test for admixture given two reference populations and the admixed population (target), we applied $f3$ statistic to the same simulated data above within each pair of populations from the 5 pools from 20 reference populations described above. The results in Table 2.4 demonstrate that in many cases the $f3$ statistic fails to provide clear evidence/non-evidence of admixture in our simulated data which mimicked a multi-way admixed population. Given different pools of reference populations for a multi-way admixed population, the $f3$ statistic clearly may not enable an accurate selection of the best proxy ancestry from each pool. Although the reference populations within a given pool may be closely related, the simulation shows that both approaches developed in PROXYANC produce the highest score from the best ancestral proxy.

2.3.1.1 Impact of Selecting Proxy Ancestry in both Estimating Ancestry and Imputing Missing Genotype in Admixed Population.

To evaluate the impact of selecting the best proxy ancestral populations for an admixed population on estimating admixture proportion, we run the ADMIXTURE software on the simulated data together with the ancestral populations (CEU, isiXhosa, ‡Khomani, CHD and Gujarati Indian) obtained after expansion, each has 1500 individuals used to simulate data. Similar analysis is preformed using the best proxy ancestral populations (original samples), each contains initial sample sizes before expansion, and includes CEU, isiXhosa, ‡Khomani, CHD and Gujarati Indian, section 2.2.4) obtained from PROXYANC. In addition, we also run the same analysis using a panel of randomly selected inappropriate proxy ancestral populations.

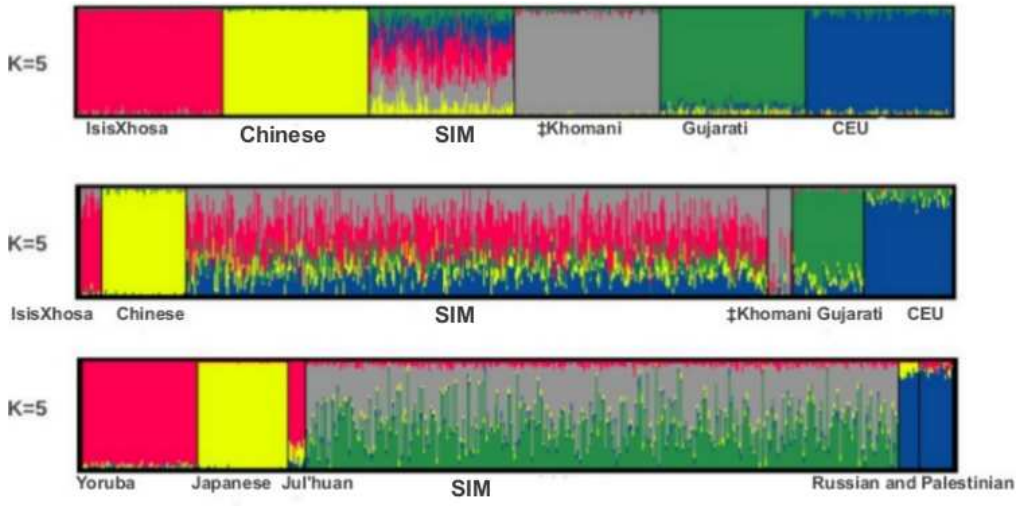

Figure 2.2: **Plot for individual's ancestry. The first two top plots are based on the combined expanded (First top figure) and original (second top figure) reference population panels (section 2.2.1) together with the simulated data, respectively. The bottom plot is based on a panel of inappropriate proxy ancestral populations of the simulation data. The admixture proportion is non-optimal in the bottom plot, and inconsistent to the true admixture proportions in our simulated data,** 2.9% **from both Russian and Palestine,** 2.6% **from Japanese,** 2.6% **from both Yoruba and Jul'huan and** 40% **and** 50% **from two unknown populations. This result demonstrates the use of inappropriate proxy ancestries for a admixed population in estimating admixture proportion may result in a non-optimal estimation of individual's ancestry.**

The ancestry proportions obtained using both panels CEU:($20\% \pm 0.0999$ and $19\% \pm 0.1039$), CHD:($8\% \pm 0.0709$ and $8\% \pm 0.0691$), Gujarati:($11\% \pm 0.0784$ and $11\% \pm 0.0839$), isiKhosa:($32\% \pm 0.1169$ and $34\% \pm 0.1545$ and ‡Khomani:($29\% \pm 0.1201$ and $27\% \pm 0.1428$), respectively are

in agreement with the ancestry proportion used in our simulation (Figure 2.2). We run the ADMIXTURE software again on the simulated data within a panel that now includes possible reference populations or populations that are more or less geographically close to the selected proxy ancestral populations, including Russian, Japanese, Palestine, Yoruba and Jul'huan (Figure 2.2). Comparing the results to the true ancestral proportions used in our simulation, we obtained a bias and inconsistent admixture proportions, $2.9\% \pm 0.2540$ from both Russian and Palestine, $2.6\% \pm 0.0229$ from Japanese, $2.6\% \pm 0.023$ from both Yoruba and Jul'huan and $40\% \pm 0.2074$ and $50\% \pm 0.2056$ from two unknown populations. Of note, a sensitive African ancestry case (isiXhosa versus Yoruba contribution in the simulated data) is displayed in Figure 2.3. In this figure, we compared the true individual admixture proportions versus those estimated from the best proxy ancestry (isiXhosa) and an inappropriate proxy ancestry (Yoruba), respectively. The estimated individual admixture proportions from isiXhosa are closer to the true individual ancestral proportion than those from Yoruba (Figure 2.3).

This result shows the impact and the sensitivity of selecting the best proxy ancestry in estimating admixture proportions which, in turn are often used in admixture association and Genome-Wide association Studies to correct for stratification. Furthermore, the sensitivity and impact are not only limited to estimating global ancestry, but have a direct impact on inferring ancestry at each locus in multi-way admixed population.

Including all available reference populations in imputing has recently been discussed to be useful in inferring accurate imputed genotypes. However, it becomes computationally expensive to the imputation engine to choose the best haplotype among several available reference populations. To address this, we assess the impact of selecting the best reference ancestral populations in imputing missing genotypes of an admixed population, we removed $2,044$ SNPs out of $39,064$ SNPs on chromosome 1 from the simulated data, and we imputed them using 4 different sets of reference populations. These four sets of reference populations include the panel of populations (CEU, CHD, Gujarati, isiXhosa, ‡Khomani) used directly in the simulation (with equal sample size of 1500 each, see Materials and Methods), a panel of populations (CEU, CHD, Gujarati, isiXhosa, ‡Khomani) used to test PROXYANC (see Materials and Methods), a panel of all populations listed in the 5 pools above and a panel formed by Russia, Japanese, Palestine, Yoruba and Jul'huan populations. The result in Figure 2.4 indicates a high call rate when imputing missing genotypes of the simulated data using the true ancestry. The imputation using the first panel of populations (True ancestry) used directly in our simulation, yielded perfectly imputed genotypes. Importantly, our simulation demonstrated that the proxy ancestral panel achieved a similar accuracy as when including all available populations in imputing missing genotype of an admixed population, suggesting the choice of an accurate ancestral panel can help in reducing the

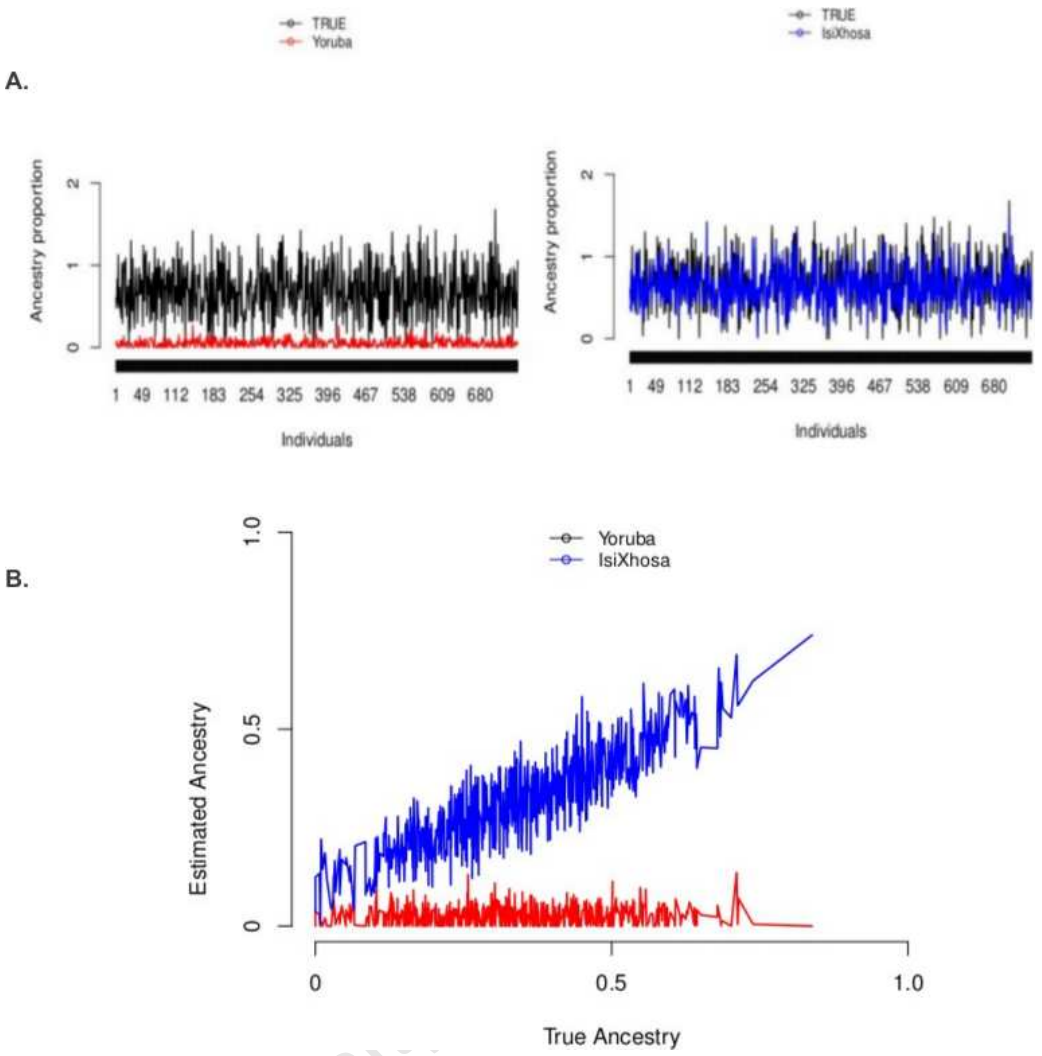

Figure 2.3: **(A) Plot of the estimated individual's ancestry from best proxy ancestry (isiXhosa) and the true individual's ancestry from the** 750 **admixed individuals obtained from the simulation. Plot of inappropriate proxy ancestry (Yoruba) estimated individuals ancestry and the true individual's ancestry from the** 750 **admixed individuals obtained from the simulation (see Materials and Methods). (B) Plot of the true ancestry versus the estimated individual's ancestry from best proxy ancestry (isiXhosa) and the estimated individuals ancestry from inappropriate proxy ancestry (Yoruba), respectively.**

computational cost of the imputation engine for searching for best haplotype among all available populations during imputation processes.

The imputation using the second and third panels also yielded a realistic imputed genotype. Because of small sample size used in second panel (Material and Methods) the imputation based on this panel (consisting of five proxy ancestral populations (see Table 2.2) with their original

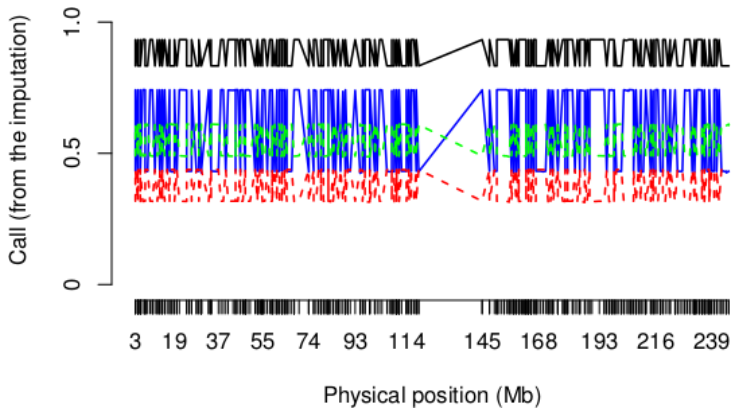

Figure 2.4: **Plot of genotype call rate in imputing** $2,044$ **SNPs on the simulated data using** 4 **sets of reference populations. Panels include Black (Expanded samples (used to simulate the data**) from CEU, CHD, GIH, isiXhosa, ǂKhomani), **Green:**(Initial samples from CEU, CHD,GIH, isiXhosa, ǂKhomani), **Blue:** All populations used to evaluate PROXYANC (see Materials and Methods) and **Red:**(Russia, Japanese, Palestine, Yoruba and Jul'huan). This plot highlights the importance of using correct proxy ancestral populations for the imputation of missing genotype in multi-way admixed populations.

samples size) does not reach the same genotype call rate as the first panel. Using the last panel of populations which does not include proxy ancestors, we obtained poor accuracy imputation of missing genotypes in our simulation data (Figure 2.4).

2.3.2 Genetic Fine Characterization of the Ancestral Components of the South African Coloured Population.

2.3.2.1 PROXYANC: Selecting Proxy Ancestry in the SAC

To select the proxy ancestral populations using the real data of the SAC, we apply PROXYANC on 5 pools of reference populations implicated by both PCA (Figure 2.5) and admixture analysis (Figures 2.9, 2.10, 2.11, 2.12, and 2.13). We first constructed African, European, South Asian and East Asian population data sets using populations described in Table 2.1, each including 764 unrelated SAC samples. The data analyzed was from four sources: The African population panel (Henn *et al.*, 2011), n = 169 samples from 11 African populations genotyped on an Illumina Beadchip 550K custom v2 chip, the Human Genome Diversity Cell Line Panel (Cann *et al.*, 2002), n = 732 samples from 54 populations genotyped on an Illumina 650K array), the International Haplotype Map (HapMap) Phase 3 data ((Frazer & et al, 2007), n = 856 samples from 10 populations genotyped on an Illumina 1M array), and samples of southern Bantu from South Africa (n = 77) and unrelated indigenous San from Namibia (n = 22, genotyped on Affymetrix

6.0). We performed admixture analysis using the ADMIXTURE software (Alexander *et al.*, 2009) and Principal Component Analysis (n = 49,930 autosomal SNPs) on each data set described above. We were able to identify the most relevant reference populations to be candidates for the proxy ancestry analysis (Figure 2.5 and Figures 2.9, 2.10, 2.11, 2.12 and 2.13).

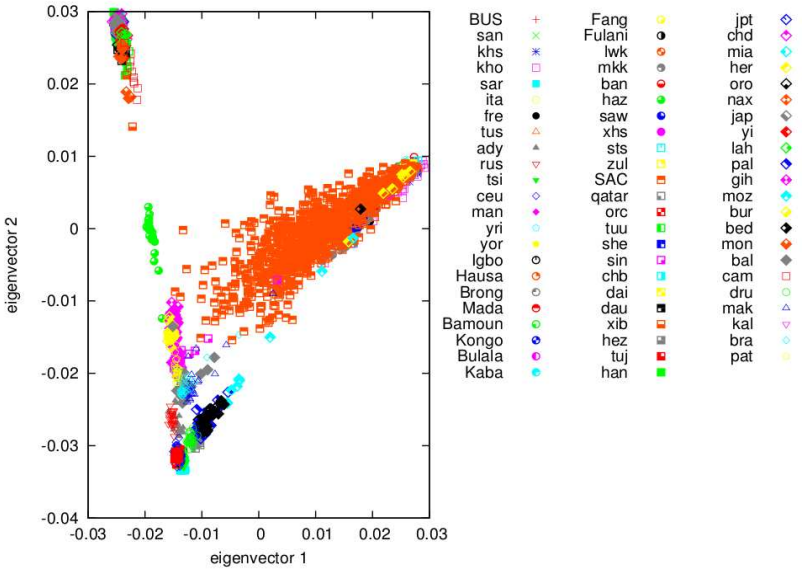

Figure 2.5: **Principal Component Analysis (PCA) of the SAC and the World-wide populations. The first and the second eigenvectors in the PCA of the combined SAC and worldwide populations are shown.**

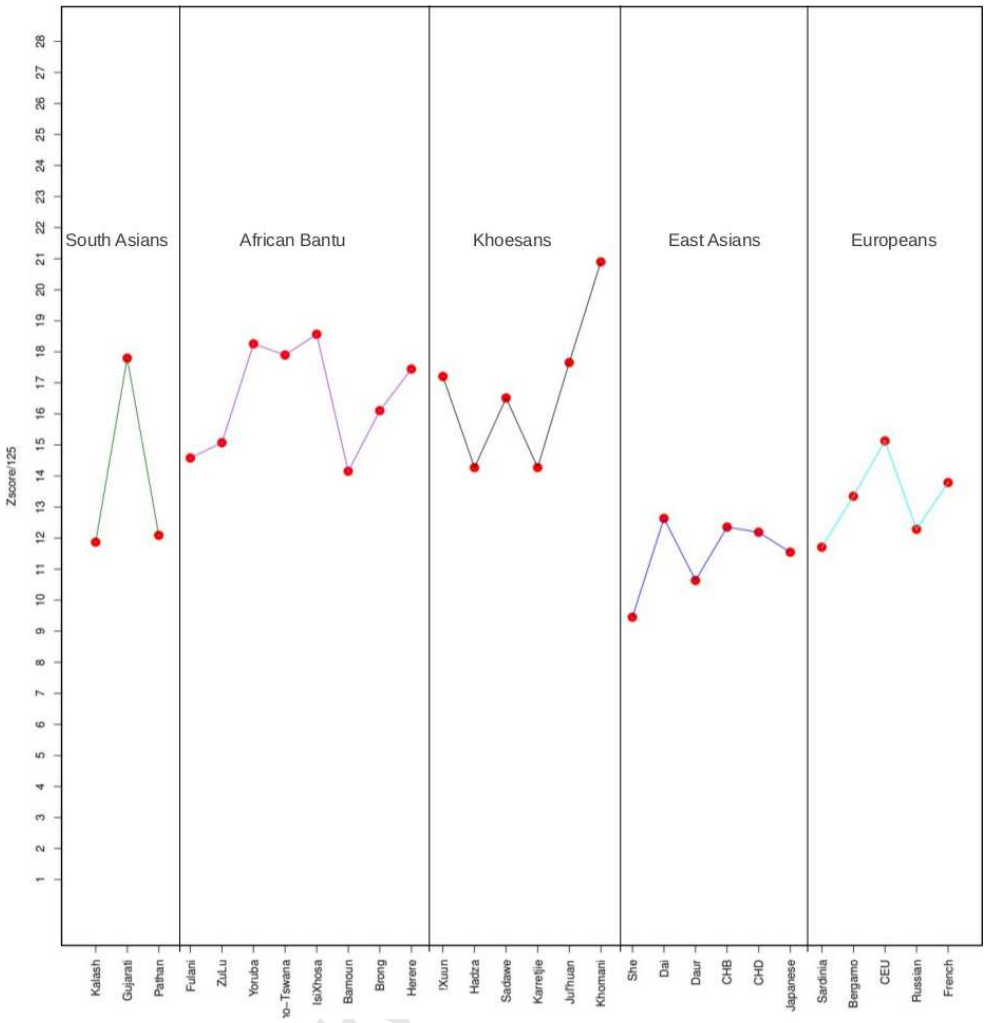

Figure 2.6: **Plot of proxy-ancestry scores for each population in each group of reference populations. The highest peak indicates the best proxy ancestry for the South African Coloured population.**

We performed both proxy ancestry score and F_{ST}-optimal quadratic cone programming on 5 distinct pools of reference populations. The results from both proxy ancestry score (Table 2.5 and Figure 2.6) and F_{ST}-optimal quadratic cone programming (Table 2.6) were in agreement and reveal that the combination of CEU, isiXhosa, Gujarati, CHD, and ‡Khomani formed the best proxy ancestry for the SAC (Table 2.5 and Table 2.6). The result in both Figure 2.6 and Table 2.5 suggest that a Southern Bantu population (isiXhosa), and a South African click-speakers (‡Khomani) are the best Bantu-speaker and click-speaker proxy ancestral populations for the SAC, compared to the more frequently used Yoruba and the Namibia San (Jul'huan) of previous studies (deWit *et al.*, 2010a; Quintana-Murci *et al.*, 2010; Tishkoff *et al.*, 2009).

Table 2.4: $f3$ **Statistic: the signal of admixture in the simulation data (simulation obtained from 5-way admixture of ǂKhomani, isiXhosa, Chinese (CHD) Gujarati Indian and CEU) using pair-wise ancestral populations. The $f3$ statistic fails to provide clear evidence/non-evidence of population admixture based on simulated data of 5-way admixed population.**

Pop 1	Pop 2	Target	f3	Standard Error	Z
CEU	San	Simulated data	−0.00827	0.00149	−5.57
CEU	CHD	Simulated data	0.01321	0.00085	15.58
CEU	Gujarati	Simulated data	0.02476	0.00079	31.33
CEU	Herero	Simulated data	−0.00586	0.00140	−4.18
CEU	isiXhosa	Simulated data	−0.01748	0.00049	−36.0
CEU	ǂKhomani	Simulated data	−0.0163	0.00051	−32.13
CEU	Pathan	Simulated data	−0.00602	0.00156	−3.86
CEU	Russian	Simulated data	−0.00451	0.00137	−3.29
CHD	San	Simulated data	−0.00289	0.00208	−1.39
CHD	Gujarati	Simulated data	0.02148	0.000794	27.134
CHD	isiXhosa	Simulated data	−0.01389	0.00057	−24.19
CHD	Italian	Simulated data	−0.00178	0.00166	−1.07
CHD	Japanese	Simulated data	−0.00352	0.00157	−2.24
CHD	ǂKhomani	Simulated data	−0.01133	0.00058	−19.53
CHD	Pathan	Simulated data	−0.00308	0.00163	−1.89
CHD	Russian	Simulated data	−0.00111	0.00167	−0.7
Gujarati	isiXhosa	Simulated data	−0.01537	0.00049	−31.34
Gujarati	ǂKhomani	Simulated data	−0.01452	0.00051	−28.27
ǂKhomani	Druze	Simulated data	−0.00139	0.00106	−1.32
ǂKhomani	French	Simulated data	−0.00151	0.00098	−1.54
ǂKhomani	Herero	Simulated data	−0.00084	0.00105	−0.80
ǂKhomani	isiXhosa	Simulated data	0.00247	0.00036	6.79
ǂKhomani	Italian	Simulated data	−0.00128	0.00103	−1.24
ǂKhomani	Japanese	Simulated data	−0.00042	0.00104	−0.40
ǂKhomani	Kongo	Simulated data	−0.00076	0.00096	−0.79
ǂKhomani	Pathan	Simulated data	−0.00023	0.00107	−0.22
ǂKhomani	Russian	Simulated data	−0.0011	0.00097	−1.1

Table 2.5: **Proxy-ancestry score for 5 distinct pools, including African non-Click speaking group, East Asian, European, click-speaker group and South Asian populations using the SAC data. The result shows that the highest scores are from CEU, ‡Khomani, isiXhosa, Chinese and Gujarati in the relevant pool.**

Populations	PScore	Standard Error	Z
South Asia Group			
Kalash	−0.003	0.001	1483.76
Gujarati	0.003	0.001	2224.43
Pathan	−0.002	0.001	1511.30
African Non-Click Speaking Group			
Fulani	0.001	0.002	1822.48
Zulu	0.001	0.001	1884.28
Yoruba	0.004	0.001	2282.03
Sotho-Tswana	0.003	0.001	2237.05
isiXhosa	0.003	0.001	2320.63
Bamoun	−0.002	0.001	1769.27
Brong	0.001	0.001	2013.24
Herero	0.002	0.001	2180.48
African Click-speak Group			
San	0.002	0.001	2150.70
Hadza	−0.003	0.001	1783.85
Sandawe	0.001	0.001	2064.319
Bushmen	−0.003	0.001	1784.10
Jul'huan	0.003	0.002	2206.76
‡Khomani	0.007	0.001	2612.07
East Asia Group			
She	−0.007	0.001	1181.64
Dai	−0.003	0.001	1579.25
Daur	−0.004	0.001	1329.53
CHB	−0.003	0.001	1523.72
CHD	−0.003	0.001	1544.38
Japanese	−0.003	0.001	1443.25
European Group			
Sardinia	−0.003	0.001	1463.5
Belgarmo	−0.001	0.001	1668.56
CEU	0.000	0.001	1891.314
Russian	−0.002	0.001	1535.53
French	−0.001	0.001	1723.62

Table 2.6: **Top 12 linear combinations that minimize the F_{ST} between SAC data and a combination of 5 pools of reference populations. The top linear combination is CEU, ‡Khomani, isiXhosa, Chinese (CHD) and Gujarati, consistent with Table 2.5 and with our simulation scheme.**

Pop Linear Combination	F	Standard error	$95\%CI$
(Gujarati, Sotho, ‡Khomani, CHB, CEU)	-0.0042	0.0010	$(-0.006, -0.0025)$
(Gujarati, Sotho, ‡Khomani, CHB, Russian)	-0.0042	0.00102	$(-0.006, -0.0023)$
(Gujarati, Sotho, ‡Khomani, CHD, CEU)	-0.0042	0.00101	$(-0.006, -0.0023)$
(Gujarati, Sotho, ‡Khomani, CHD, Russian)	-0.0042	0.00101	$(-0.006, -0.0023)$
(Gujarati, isiXhosa, ‡Khomani, CHB, CEU)	-0.00374	0.00060	$(-0.005, -0.003)$
(Gujarati, isiXhosa, ‡Khomani, CHB, Russian)	-0.00374	0.00060	$(-0.005, -0.003)$
(Gujarati, isiXhosa, ‡Khomani, CHD, CEU)	-0.00374	0.00060	$(-0.005, -0.003)$
(Gujarati, isiXhosa, ‡Khomani, CHD, Russian)	-0.00374	0.00060	$(-0.005, -0.003)$
(Gujarati, Brong, ‡Khomani, CHB, CEU)	-0.02483	0.00605	$(-0.037, -0.013)$
(Gujarati, Brong, ‡Khomani, CHB, Russian)	-0.02483	0.00605	$(-0.037, -0.013)$
(Gujarati, Brong, ‡Khomani, CHD, CEU)	-0.02483	0.00605	$(-0.037, -0.013)$
(Gujarati, Brong, ‡Khomani, CHD, Russian)	-0.02483	0.00605	$(-0.037, -0.013)$

2.3.2.2 Refinement of Admixture Proportion in the SAC

Using the result from PROXYANC on the SAC data (section 2.3.2.1), we combined the top proxy ancestral populations (CEU, CHD, Gujarati, isiXhosa, ‡Khomani) (Table 2.5 and Table 2.6), including the SAC, into one data set. We repeated both the PCA and the ancestral population clustering analysis. From these analyses, our inferred five major ancestral contributions (Table 2.7 and Figure 2.7) to the SAC population have a balanced African ancestral proportion from IsiXhosi (33%) and ‡Khomani (31%), followed by European (CEU) (16%), Gujarati Indian (12%) and a smaller admixture proportion from Chinese (8%). It is also clear from the PCA plots in Figure 2.7, that the SAC lie on a direct line with these five groups of proxy ancestors.

In addition, both the isiXhosa and ‡Khomani groups were related to the SAC, indicating their close ancestral affiliations with this population and reflecting the role of both Southern Bantu and indigenous Sub-Kalahari click-speakers in the early establishment of the SAC population (Mountain, 2003). The other putative groups of proxy ancestral populations; CEU, Gujarati Indian and Chinese, are separated from each other, and the SAC is in the convex hull of the three. These findings agree well with the result obtained from the admixture analysis with $K = 5$ in Figure 2.7. As we expected, the PCA in Figure 2.7 revealed the greatest genetic differentiation between these five proxy ancestries of the SAC, which clearly reflects the admixture of the SAC from these five proxy ancestors. In addition, we compare our estimated admixture proportions with previous estimates in (Patterson *et al.*, 2009) and we redo the admixture analysis using the ancestral populations used in deWit *et al.* (2010a), which included Yoruba, CEU, San (Jul'huan), Gujarati, and Chinese (CHB).

Table 2.7 displays the estimated admixture proportions obtained using the best proxy ancestral populations and from the previous studies (deWit *et al.*, 2010a).

Figure 2.8 indicates a large difference of African ancestry of the SAC between the two analyses (using proxy ancestries panel and the panel from deWit *et al.* (2010a), suggesting the choice of African Ancestry for the SAC is critical and sensitive in conducting ancestry inferences and admixture mapping studies. This may due to the diversity and close relatedness of most African populations. Overall, our result highlights the importance of selecting the best proxy ancestral populations for multi-way admixed populations, and demonstrates that an inaccurate reference ancestral population can result in inaccurate inferred ancestry, which is used in admixture association or admixture mapping study. This can lead to erroneous interpretation of the results when identifying genomic location underlying genetic ancestry difference in complex disease risk.

Taken together, our results above provide confidence that our inferred five ancestral components with balanced African contributions from isiXhosa and ‡Khomani populations, followed by north-western European, Gujarati Indian and a smaller Chinese contribution, are closer to the true level of ancestral contributions, and agree with the SAC's history. We believe that our result

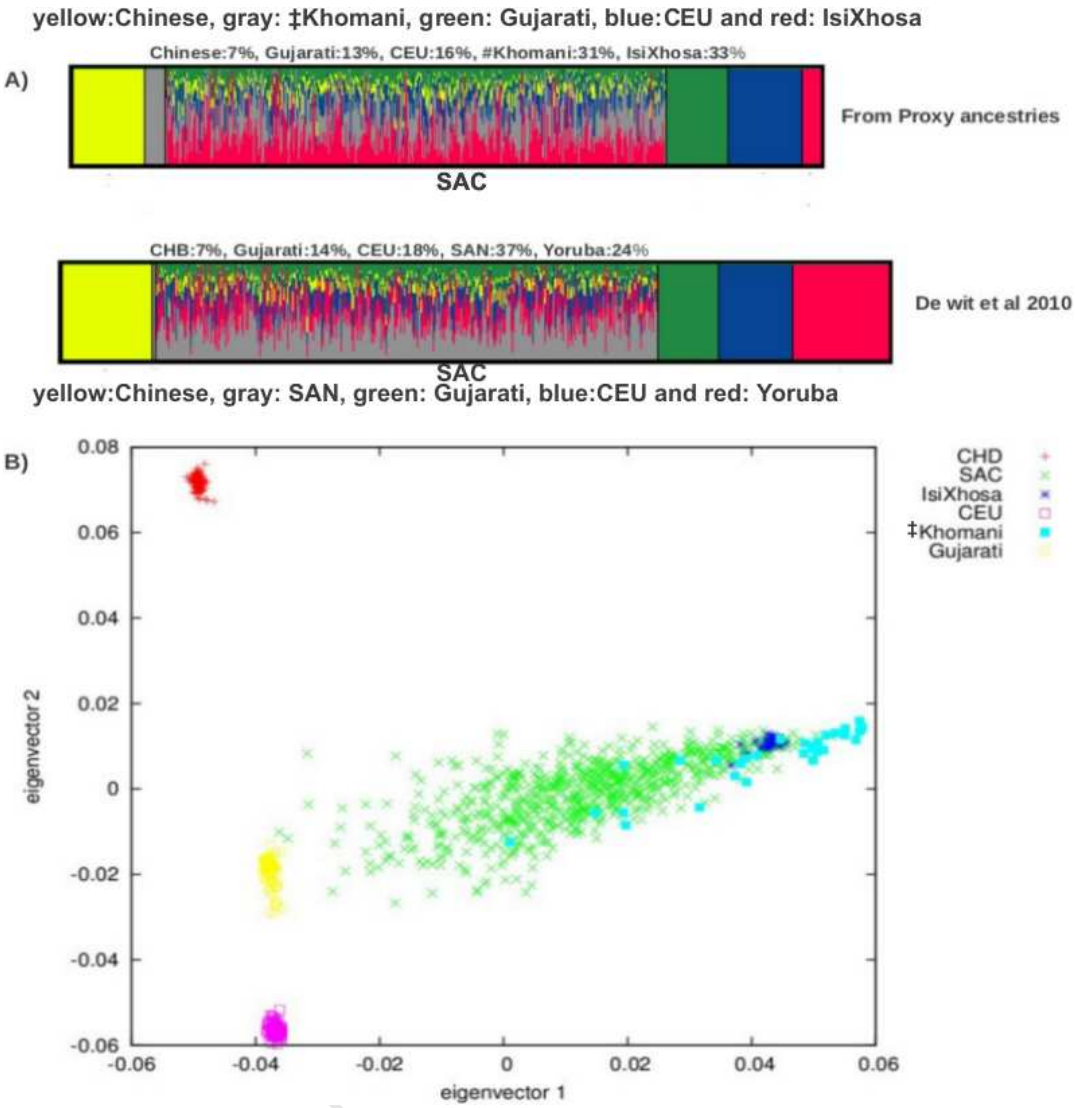

Figure 2.7: **Individual's ancestry proportion and Principal Component Analysis (PCA) based on** $49,930$ **autosomal SNPs in the SAC data: (A) Population clustering analysis of the SAC using both the current selected best proxy ancestors as a reference panel (top figure) and the reference panel used in deWit et al. (2010a). (B) Principal Component Analysis (PCA) on the merged data of the SAC with our selected best proxy ancestral populations.**

also has the advantage of handling sample size differences and using accurate proxy ancestral populations, and believe that both the number of SNPs (n $= 49,930$) and target population sample size used can provide sufficient resolution to support our inferred ancestral contribution.

Table 2.7: **Summary mean and standard error on proportion of ancestral populations contributing to the genetic make-up of the South African Coloureds. This table displays the mean and the standard errors of ancestral proportions with the best proxy ancestors obtained from PROXYANC, with the reference populations panel used in deWit et al. (2010a) and the SAC's ancestral proportions reported in (Patterson et al., 2009).**

Using the best proxy ancestral populations				
isiXhosa	‡Khomani	CEU	CHD	Gujarati
$33\% \pm 0.226$	$31\% \pm 0.195$	$16\% \pm 0.118$	$7\% \pm 0.0488$	$13\% \pm 0.094$
Using the same panel as in deWit *et al.* (2010a)				
Yoruba	San (Jul'Huan)	CEU	CHB	Gujarati
$24\% \pm 0.161$	$37\% \pm 0.148$	$18\% \pm 0.118$	$7\% \pm 0.0478$	$14\% \pm 0.093$
Reported ancestral proportions in Patterson et al. 2009				
isiXhosa	X	European	Indonesian	South Asian
$37\% \pm 0.003$	–	$23\% \pm 0.008$	$18\% \pm 0.004$	$22\% \pm 0.009$

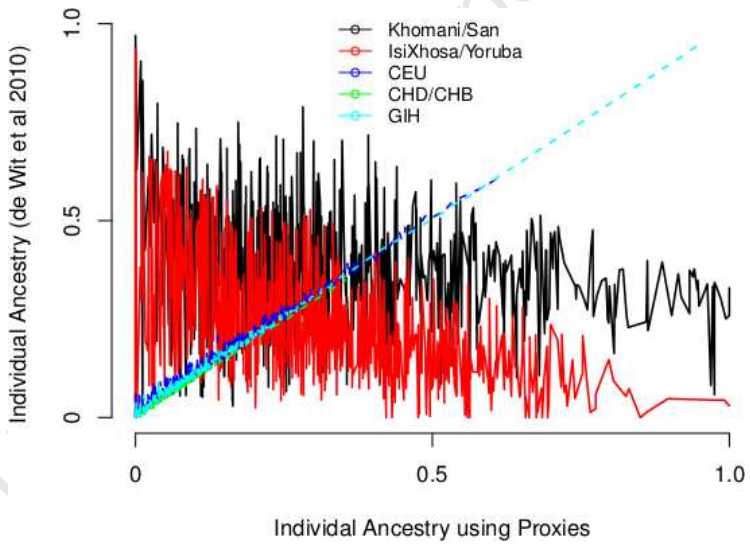

Figure 2.8: **Difference in individual's ancestry proportions between panel of selected best proxy ancestral population of the SAC and the panel of reference population used in deWit et al. (2010a).This plot indicates a large difference of African ancestry of the SAC between the two analyses, suggesting the choice of African Ancestry of the SAC is critical and sensitive due to the diversity and closely relatedness of most African populations.**

2.4 Conclusion and Remarks

We introduced PROXYANC, an approach to select the best proxy ancestry for multi-way admixed populations. We assessed its accuracy through a simulation of a multi-way mixed population, and demonstrated the impact and sensitivity of the choice of reference panel in estimating global and local ancestry and in imputing missing genotypes. To the best of our knowledge, this use of PROXYANC is the first approach to select the best reference ancestral panel given pools of reference ancestral panels. Our methods to select proxy ancestral populations in multi-way admixed populations have enabled us to characterize the genetic ancestry component of the uniquely admixed Coloured population of South Africa, that accounts for 54% of the population of the Western Cape Province. Previous studies of this historically complex population were hampered by the relatively few samples and few putative ancestral populations publicly available, particularly the very low number of San individuals. In the present study we have utilized the increased number of reference populations available from local sources, and the best proxy ancestries of the SAC obtained from PROXYANC allowed us to document a contribution of the isiXhosa, ǂKhomani, central European, Gujarati Indian and Chinese genetic materials to the SAC (with proportions 33%, 31%, 16%, 12% and 7%, respectively). We expected a southern Bantu-speakers, such as isiXhosa instead of Yoruba, to be a better proxy ancestor of the SAC. isiXhosa as a better proxy ancestor of the SAC reflects the early mixing of indigenous females from both click-speaker group and Southern Bantu-speaker groups with male settlers, mainly from the Netherlands, Britain, Germany and France, or male slaves from South East Asia (Boonzaaier *et al.*, 1996; Keegan, 1996). The substantial number of ǂKhomani (a sub-Kalahari click-speaker) individuals available for this study greatly increases our confidence in the accuracy of the ancestry estimates presented here. Our results also emphasize the point that click-speaker groups are often very different from one another, and grouping San individuals from different areas together as generic San may result in a loss of discrimination at the genetic level (Pickrell *et al.*, 2012; Schlebusch *et al.*, 2012). This was also illustrated by the deep genetic differences between individual San (Bushmen) genomes (Schuster *et al.*, 2010). In the case of the SAC in the Western Cape, it is perhaps to be expected that a click-speaker group from the southern Kalahari, including ǂKhomani, Bushmen and San, which are geographically closer to the place of origin of the SAC, would be a better proxy ancestor of this group than Jul'huan from Namibia, and this is what we have shown. This also gives credence to an earlier suggestion that only some of the click-speaker people contributed to the SAC population (Quintana-Murci *et al.*, 2010).

Furthermore, since existing methods that infer local ancestry assume that non-admixed ancestral populations are the most suitable, it may not be advisable to use the isiXhosa, which have some Khoesan ancestry as an ancestral population for admixture mapping. Until such time as

these methods are updated, the highest ranking putative non-admixed African populations listed in Tables 2.5 and 2.6, such as the Yoruba, can be used as proxy ancestral population(s) instead of the isiXhosa.

Overall, this chapter has highlighted the importance of selecting the best proxy ancestry for potential downstream analysis in a multi-way admixed population. The SAC provides a perfect population to enable the choice of best proxy ancestry. Furthermore, the obtained best proxy ancestry for this population provides opportunities to conduct downstream analysis in examining the ancestry-specific Tuberculosis risk.

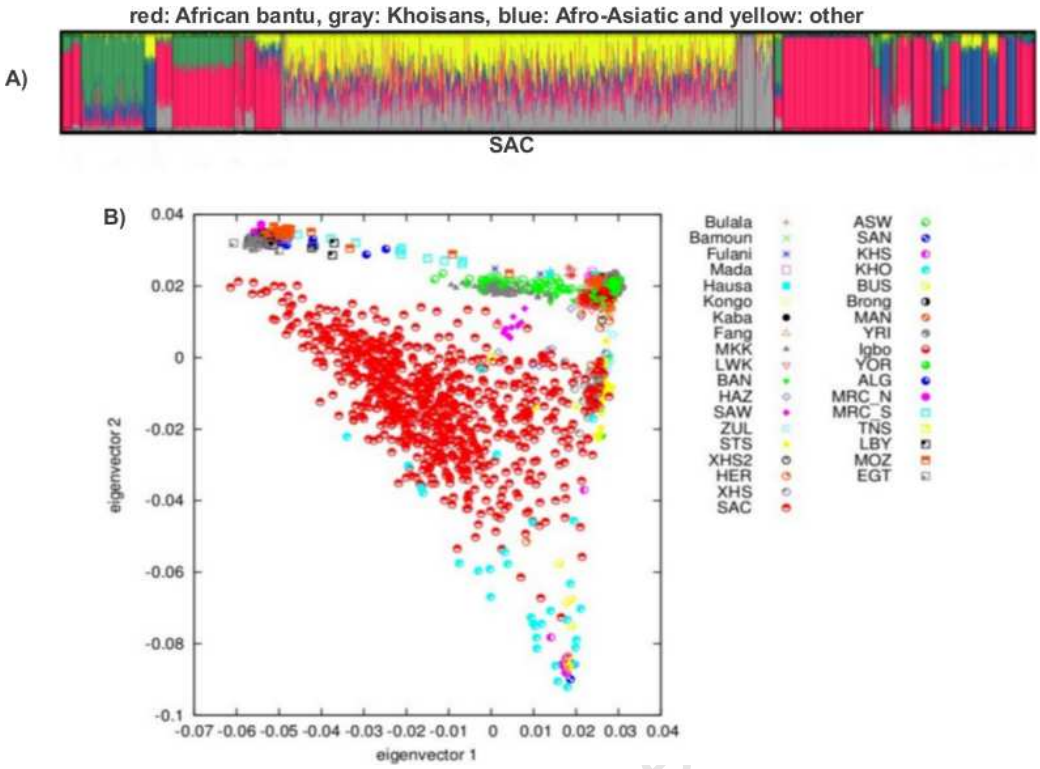

Figure 2.9: **Ancestral population clustering (A) and Principal Component Analysis (B) of the SAC and African populations. The Plot in (A) is the proportion of each individual's ancestry. (B) The plot is of the first and the second eigenvectors in the PCA of the combined populations.**

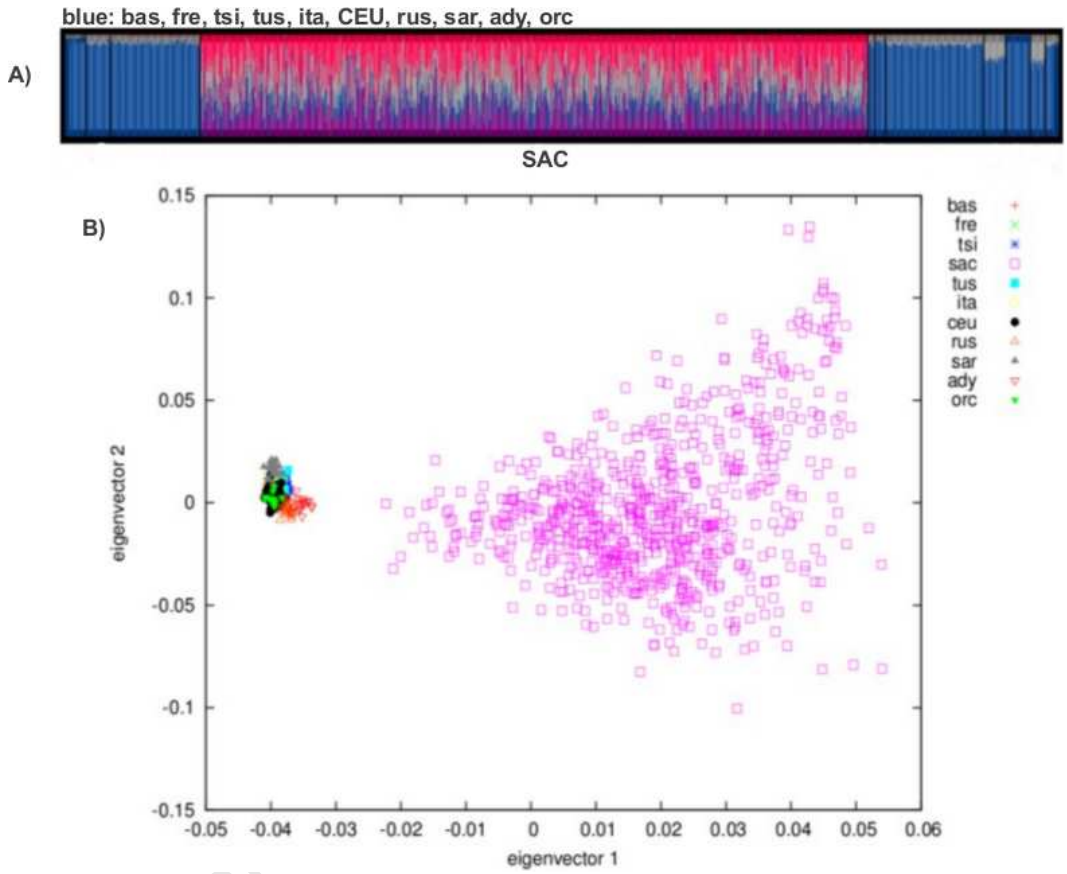

Figure 2.10: **Ancestral population clustering and Principal Component Analysis (PCA). (A) Population clustering analyses of the SAC and European populations. The Plot in (A) is the proportion of each individuals ancestry. (B) The plot of the first and the second eigenvectors in the PCA of the combined populations.**

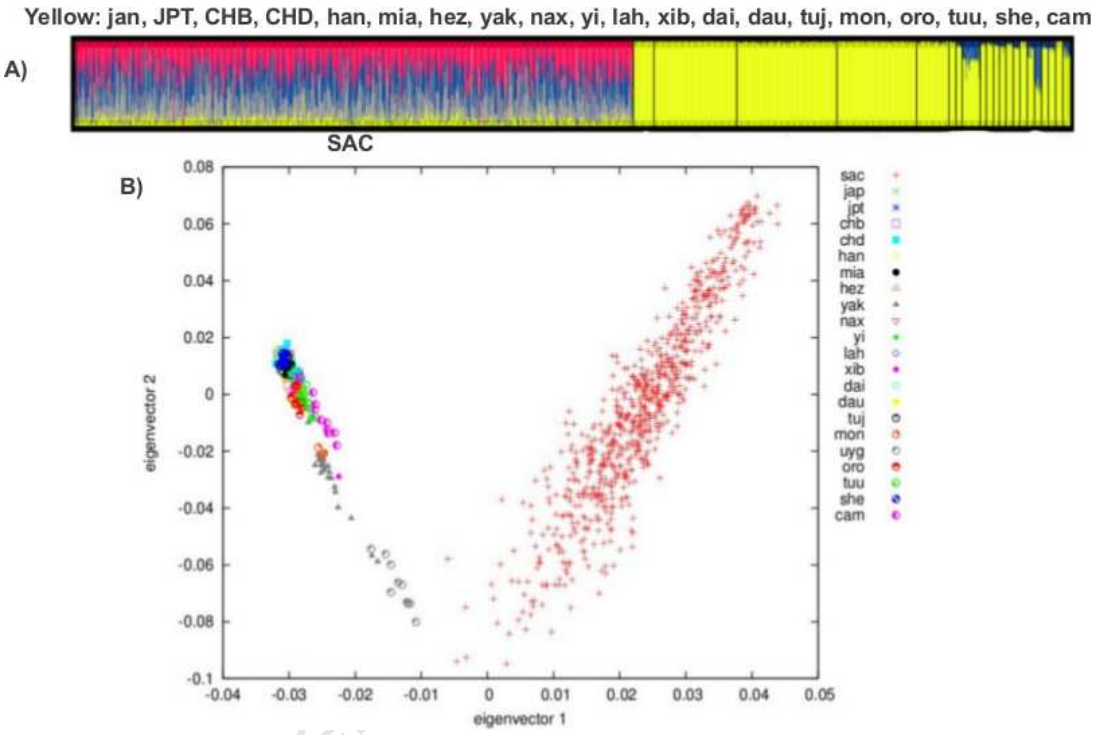

Figure 2.11: **Ancestral population clustering (A) and Principal Component Analysis (B) of the SAC and East Asian populations. (A) The Plot in (A) is the proportion of each individual's ancestry. (B) The plot is of the first and the second eigenvectors in the PCA of the combined populations.**

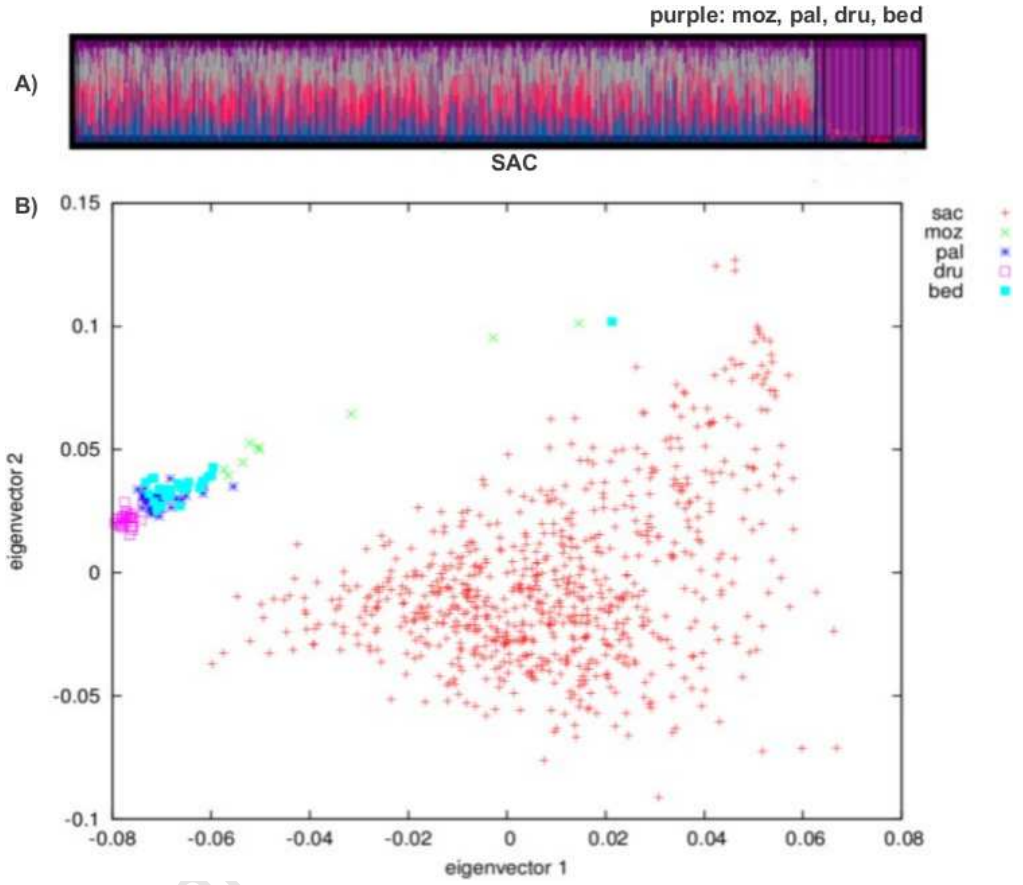

Figure 2.12: **Ancestral population clustering (A) and Principal Component Analysis (B) of the SAC and Middle East populations. (A) The Plot in (A) is the proportion of each individual's ancestry. (B) The plot is of the first and the second eigenvectors in the PCA of the combined populations.**

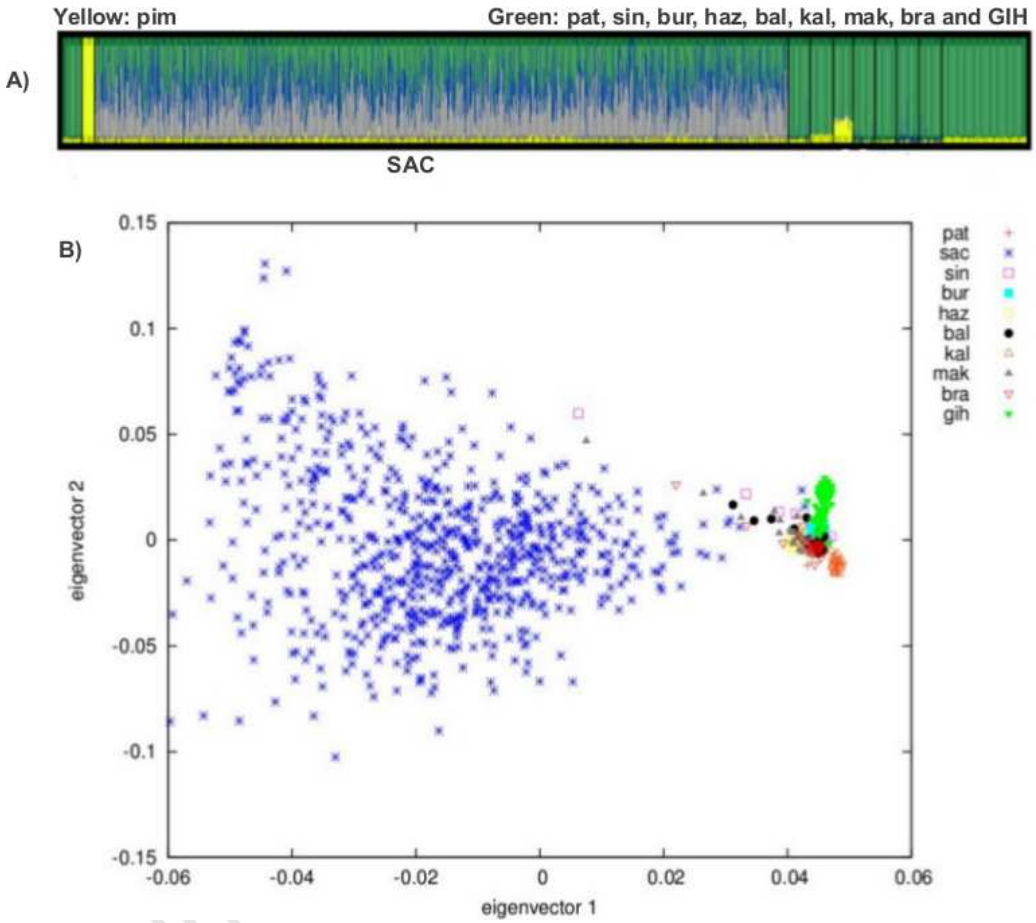

Figure 2.13: **Ancestral population clustering (A) and Principal Component Analysis (B) of the SAC and South Asian populations. (A) The Plot in (A) is the proportion of each individual's ancestry. (B) The plot is of the first and the second eigenvectors in the PCA of the combined populations.**

Chapter 3

Ancestry Informative Markers: Admixture Linkage Disequilibrium and Haplotype Diversity in the Coloured population

3.1 Introduction

Since ancestry informative markers (AIMs) are those polymorphisms with the greatest difference in frequency between populations, they can be used to examine the admixture linkage disequilibrium in a admixed population, and efficiently analyse the signal of admixture from its putative ancestral populations. Furthermore, selecting a subset of highly informative genetic markers for a particular population has a range of applications from the inference of individual ancestry to admixture association (Kosoy *et al.*, 2009; Paschou *et al.*, 2007). The ancestry informative markers (AIMs) are genetic polymorphisms with striking allele frequency differences between geographically distant populations or ancestral populations of an admixed population. While markers with strong geographic correlations are rare overall, recent genetic studies have investigated the identification of small panels of AIMs that can provide an estimate of the ancestry of individuals or estimate of the apportionment of ancestry components from admixed populations. AIMs can limit the number of tests to a subset of the genome and focus hypotheses on the subset of these genetic markers (Montana & Pritchard, 2004; Paschou *et al.*, 2007). Therefore, a subset of genetic markers that specifically differentiate chromosomes derived from suitable ancestral populations is needed (Smith & O'Brien, 2005). Three basic questions have commonly arisen in selecting efficient subsets of genetic markers:

(1) Given a set of M genetic markers, which genetic markers should constitute a desired panel of informative genetic markers?

(2) How should the number of informative genetic markers genotyped be determined?

(3) How well can these informative markers predict the remaining set of the unselected genetic markers?

A number of approaches, including (Galanter *et al.*, 2012),Kosoy *et al.* (2009), (Paschou *et al.*, 2007), Rosenberg (2005) and Rosenberg *et al.* (2003), have been used to select ancestry informative markers and these approaches were mostly applied to two or three way admixed population data. The informative genetic markers have been traditionally selected to maximize the absolute difference in allele frequency between ancestries (Lewontin, 1964; Vega *et al.*, 2006). The statistical proprieties of the absolute difference in allele frequencies are not well defined and can only be used for two or three source populations at a time. Here, we developed two different algorithms to select a subset of genetic markers, and apply these algorithms to select the most informative markers from the genome-wide data of the South African Coloured population (SAC).

The history of the Coloured population (SAC) before and during the last apartheid regime in South Africa which separated ethnic groups and outlawed inter-racial marriage (http://www.sahistory.org.za/pages/chronology/special-chrono/governance/apartheid-legislation.html) may influence its genetic make-up by exhibiting higher frequencies of recessive genetic disorders, haplotype identity-by-descent and linkage disequilibrium (LD) (Arcos-Burgos & Muenke, 2002; Peltonen *et al.*, 2000). The admixture LD in this population has not been give attention yet. In addition, the genetic signature of founder events and the possibility that bottlenecks may influence its genetic structure have also not yet been considered. Since we aim to select AIMs that explain the admixture LD in the admixed population, and account for background LD in ancestral population panel, we additionally used the related AIMs panel to compare the genome-wide haplotype diversity and the percentage haplotype sharing by IBD between the SAC and its proxy ancestral populations. To address this, and investigate the population admixture processes in the SAC, we first introduce and implement two different algorithms to select a subset of ancestry informative markers. These constructed panels can be used to examine the admixture linkage disequilibrium and efficiently analyse the signal of admixture from the putative ancestral populations of the SAC.

The first algorithm demands prihor knowledge of the ancestry of the studied samples and it uses the relationship between the observed local multi-locus linkage disequilibrium in a recently admixed population and ancestral population difference in allele frequency. The second algorithm is an unsupervised method based on the Kernel principal component analysis (Kernel-PCA),

which is the extension of the linear PCA. It allows us to learn the non-linear dependency and to find meaningful projections (i.e. the subspace of the largest variance). We apply these algorithms to select the most informative markers from the genome-wide data of the uniquely 5-way admixed South African Coloured population (SAC). We use the subset of informative markers that differentiate the best proxy ancestral populations of the SAC obtained from the PROXYANC algorithms (sections 2.2.2 and 2.2.3) to examine the pattern of LD and the level of admixture LD in the SAC as a result of ancestral admixture.

3.2 Methods

3.2.1 Genetic Marker Selection: Relationship between Population Differentiation and Admixture Linkage Disequilibrium

Given a pair of populations k and l from a pool of K ancestral populations of an admixed population, assuming the minor allele frequency at SNPs i and j are greater that 0.005. Similar in (Shiheng *et al.*, 2001), we defined the admixture linkage disequilibrium as,

$$L_{ij} = mL_{ij}^{k} + (1-m)L_{ij}^{l} + m(1-m)\delta_i^{kl} \times \delta_j^{kl}, \tag{3.1}$$

where m is the ancestral proportion, δ_i and δ_j are differences in allele frequency at SNPs i and j in population k and l, respectively.

Assuming for each pair of SNPs i and j there is not linkage disequilibrium in the ancestral populations, it thus follows,

$$L_{ij} = m(1-m)\delta_i^{kl} \times \delta_j^{kl} \tag{3.2}$$

$$1 = \frac{m(1-m)\delta_i^{kl} \times \delta_j^{kl}}{L_{ij}} \tag{3.3}$$

Equation 3.3 establishes a perfect relationship between the observed linkage disequilibrium L_{ij} in the recently admixed population and ancestral population differentiation at a given pair of SNPs i and j in the admixed population. Equation 3.3 is a total ancestry content (AC) at a pair of SNPs i and j. Assuming a uniform ancestral proportion, and summing equation 3.3 over all possible pairs of proxy ancestral populations, we can obtain the ancestry informativeness I_{ij} of each pair of SNPs i and j as follows,

$$I_{ij} = \frac{1}{4 \times K} \sum_{k \neq l}^{K} \frac{\delta_i^{kl} \times \delta_j^{kl}}{L_{ij}}. \tag{3.4}$$

Let M be the total number of SNPs. For $i \in \{1, ..., M\}$, let N_i be the total number of pair-wise LD within SNP i, we obtain the ancestry informativeness at SNP i as follows,

$$I_i = \sum_{j=1}^{N_i} \frac{I_{ij}}{\sqrt{M}}. \tag{3.5}$$

3.2.2 Principal Component Analysis (PCA) Selection-based Method

Principal component analysis is a dimensionality-reduction method, it is a procedure to rotate data such that maximum variability is projected onto orthogonal axes according to a minimum-square-error criterion (Lin & Altman, 2004; Paschou *et al.*, 2007). Essentially a set of correlated variables is transformed into a substantially smaller set of uncorrelated variables (principal components) that represent most of the variation in the original data, where the principal components are linear combinations of the original set of variables (Patterson *et al.*, 2006). One of the challenges using PCA on the genotype data is that the principal components that are defined do not correspond to actual genotypes (Lin & Altman, 2004). Thus, we need to determine out the way to map the principal components optimally to the original genotype data. Here, we make use of the Kernel PCA methods within a greedy-discard algorithm in order to select the most informative markers. Based on Kernel PCA, our algorithm is the generalization of the existing linear PCA selection-based method, add the advantage of extracting non-linear dependencies and finding meaningful projections throughout the dataset. Our Kernel PCA uses the Gaussian kernel function defined in 2 dimensions in order to map the data matrix (kernelize the data). Consider a matrix of genetic markers $\mathbf{D}$. Each pair of rows $(\mathbb{Y}_i, \mathbb{Y}_{i+1})$ represents an individual sample, $i = 1, 2, \ldots, 2N$ where N is the number of samples. Each column $\mathbb{X}_j$ corresponds to the genetic markers (diploid genotyes), $j = 1, 2, \ldots, M$ such that $(\mathbf{D}[i, j], \mathbf{D}[i+1, j])$is the genotype of the sample i at marker j. We describe the algorithm through the following six steps:

(1) Set the number of dimensions in the dimensionally reduced subspace $1 \leq L \leq M$.

(2) From the data matrix $\mathbf{D}$, use the Gaussian Kernel function defined in 2-D,

$$\kappa = \frac{1}{2\pi\sigma^2} exp\left(-\frac{|| \mathbb{X}_i - \mathbb{X}_j ||^2}{2\sigma^2}\right)$$

to construct the kernel matrix, $\mathbf{K}$,

$$K_{ij} = \kappa\left(\mathbb{X}_i, \mathbb{X}_j\right)$$

(3) From the kernel matrix $\mathbf{K}$, we compute the covariance matrix $\mathbf{C}$,

$$\mathbf{C_{ij}} = K_{ij} - \sum_{k=1}^{M} K_{ik} - \sum_{l=1}^{M} K_{lj} + \sum_{l,k=1}^{M} K_{kl}.$$

(4) From the covariance matrix $\mathbf{C}$, we compute the set of the eigenvectors $\mathbf{V} = \mathbf{V}[k,n]_{k,n=1,2,\ldots,M}$ and the eigenvalues $\Lambda = \Lambda[k,n]_{k,n=1,2,\ldots,M}$, through equation 3.7. The matrix $\mathbf{V}$ contains all eigenvectors of the covariance matrix $\mathbf{C}$, one eigenvector per column. Λ is a diagonal matrix that contains all eigenvalues of the covariance matrix $\mathbf{C}$ along its principal diagonal and 0 for all other elements. The eigenvalues and eigenvectors are ordered and paired in such a way that the m^{th} eigenvalue corresponds to the m^{th} eigenvector.

$$\mathbf{C}v_k = \lambda_k v_k, \text{ where } k = 1, 2, \ldots, M. \tag{3.6}$$

Sort the columns of the eigenvector matrix $\mathbf{V}$ and eigenvalues in order of decreasing eigenvalue Λ.

(5) As each eigenvalue is the amount of variance explained by the eigenvector, choose L eigenvectors with the largest eigenvalues. Each eigenvalue is a weighted sum of the original data as,

$$p_l = \sum_{k=1}^{M} \mathbf{V}[l,k]\mathbf{D}[k,j], \text{ where } l = 1, 2, \ldots, L, \tag{3.7}$$

where the weights are the coefficients of the eigenvector. The sum of variances of L chosen eigenvectors is equal to the sum of variances of original genetic marker data. Consequently, the proportion of the variance in the M original genetic markers that L eigenvectors account for is

$$\rho = \frac{\sum_{l=1}^{L} \lambda_l}{\sum_{m=1}^{M} \lambda_m}. \tag{3.8}$$

(6) At this stage, the chosen eigenvectors do not correspond to any subset of the original genotype data. We apply the greedy-discard approach (Lin & Altman, 2004) to map these eigenvectors to the most corresponding genetic markers.

(a) Start by the eigenvector in the eigenvectors space with the smallest eigenvalue to the $(m-l)^{th}$ eigenvector in the space of L chosen eigenvectors with the smallest eigenvalue, then reject the genetic marker that has the largest absolute coefficient value in the $(m-l)^{th}$ space of chosen eigenvectors (equation 3.7) and that has not yet been discarded.

(b) In the reverse order, map the retained L eigenvectors to the remaining L genetic markers in the original data as L Kernel-PCA markers.

3.2.3 Admixture Linkage Disequilibrium

Increased LD in a population can be due to founder events, admixture of previously isolated populations, population bottlenecks (Kruglyak, 1999) or other factors. We examine the observed LD in the SAC by comparing the significance level of increase in LD at short distances (< 0.1 cM) and long distances (> 0.2 cM), within and between the SAC and its proxy ancestors. To account for the sample size effect in computing the LD, we first scaled each population samples, including the SAC's samples to roughly equal size each. The LD-r^2 values is computed for linked and unlinked SNP-pairs along the genome using the LD statistic described in section 2.2.3. Thus, we directly compare the LD-r^2 for each SNP-pair by ranking the number of pairs that had higher LD-r^2 (> 0.5) in the SAC and in each proxy ancestral population. Furthermore, we compute the correlation between ancestral allele-frequency differences and LD-r^2 in the SAC. The allele-frequency differences is calculated on the first (δ_1) and second (δ_2) SNP based on the SNP-pair having LD-$r^2 > 0.5$ in the admixed population. The correlations is computed between $\delta_{s_1} \times \delta_{s_2}$ and LD-r^2 in the SAC, and we report the average p-values and the correlations. To see whether the level of the observed admixture in the SAC can account for the increased LD, we also estimate the maximum expected admixture LD from each pair of reference ancestral populations and we compare them with the observed LD in the SAC. Given the LD and allele-frequency from pair of unrelated ancestral populations, X and Y of the admixed population Z, the admixture LD (D_Z) is related to the LD D_X and D_Y from X and Y (Shiheng *et al.*, 2001), and is modelled as,

$$D_Z = mD_X + (1-m)D_Y + m(1-m)\delta_{s_1} \times \delta_{s_2}, \tag{3.9}$$

at SNPS, s_1 and s_2, where m is the ancestral proportion. This equation is a quadratic equation of the second order of the form $m_2 + bm + c$, where $a = -\delta_{s_1} \times \delta_{s_2}$, $b = D_X - D_Y + \delta_{s_1} \times \delta_{s_2}$ and $C = D_Y$. We denoted δ_{s_1} and δ_{s_2} as the difference in allele frequency at genetic marker s_1 and s_2 from X and Y populations. To obtain the admixture proportion m at which admixture LD reaches its maximum, we differentiate D_Z with respect to m and obtain the maximum expected admixture LD as

$$D_{exp} = D_Y + \frac{(D_X - D_Y + \delta_{s_1} \times \delta_{s_2})^2}{4\delta_{s_1} \times \delta_{s_2}}. \tag{3.10}$$

To assess the admixture LD, we compute the expected square correlation between the observed LD in a recently admixed population and D_{exp} from each pair of candidate proxy ancestral populations.

All the methods described in section 3.2.1, 3.2.2 and 3.2.3 above have been implemented in PROXYANC (http://www.cbio.uct.ac.za/proxyanc) too.

In addition, we also used a recent method that computes the weighted linkage disequilibrium (LD) statistic for making inference about population admixture, implemented in ALDER (Loh *et al.*, 2013) software. This analysis is conducted in order to validate our approach for assessing the admixture in the SAC as a consequence of admixture events from its proxy ancestral populations, but not due to population bottleneck. To infer the weighted LD decay curves in the SAC, we used the entire (all available SNPs) diploid genotype data from the SAC and each of its two proxy ancestral populations and plot the LD decay curve.

3.2.4 Genetic Diversity, Identity-by-Descent (IBD) and Haplotypes Shared IBD

Aside from the level of the observed admixture in the SAC, we computed the proportion of IBD and the pairwise population concordance (PPC) test. For the pairwise identity-by-state (IBS) test, we ran PLINK with $10,000$ permutations between populations in the same data set (SAC versus each proxy ancestral population). We coded the SAC as cases and its proxy ancestries as controls. We calculated the empirical p-values to determine whether case/case-pairs were less similar to each other compared to control/control-pairs (Purcell *et al.*, 2007). To compare the haplotypes shared IBD within and between the SAC and its proxy ancestral populations, the PLINK software package was run to infer the phased-haplotype of each population (SAC, isiXhosa, European (CEU), ‡Khomani, Gujarati Indian and Chinese (CHD)). For each population, we estimated the haplotype diversity as

$$H = N\frac{1-\sum h_i^2}{N-1}, \tag{3.11}$$

where h_i is the haplotype frequency and N is the haplotype sample size. The mean haplotype diversity was reported. The haplotype frequency was computed for each population using PLINK (Purcell *et al.*, 2007). The detection of extended haplotypes shared IBD, was done using PLINK on each population separately.

3.3 Results

3.3.1 Selection of Ancestry Informative Markers

Feasibility and sufficient power of both Genome-wide Association Studies and admixture mapping relies on the patterns and the extent of LD across chromosomal regions with a considerable marker density (Winkler *et al.*, 2010). Understanding the extent of admixture LD is useful in designing disease mapping tests in the admixed populations (Winkler *et al.*, 2010). Here, we applied the Kernel-PCA algorithm described in section 3.2.2 on SAC samples genotyped at $550K$, to select the most informative Kernel-PCA markers (unsupervised method). The algorithm was able to select 1001 Kernel-PCA markers with at least $1MB$ spacing between adjacent genetic markers along the genome. In addition, we selected 1121 AIMs with at least $1MB$ spacing between adjacent genetic markers along the genome based on the relationship between ancestral population (using the obtained best proxy ancestral populations in Tables 2.2 and 2.3) differentiation and observed LD in the SAC (algorithm described in section 3.2.1) using the SAC data. There are 48 SNPs overlap between the two sets of AIMS and 753 SNPs between the two sets of AIMS are in LD ($r2 > 0.5$). Since, these two AIMs panels produce similar individual's ancestry proportions 3.1, we used the 1121 AIMs panel to examine the pattern of linkage disequilibrium in the SAC. These panels can be downloaded from http://www.cbio.uct.ac.za/AIMs/.

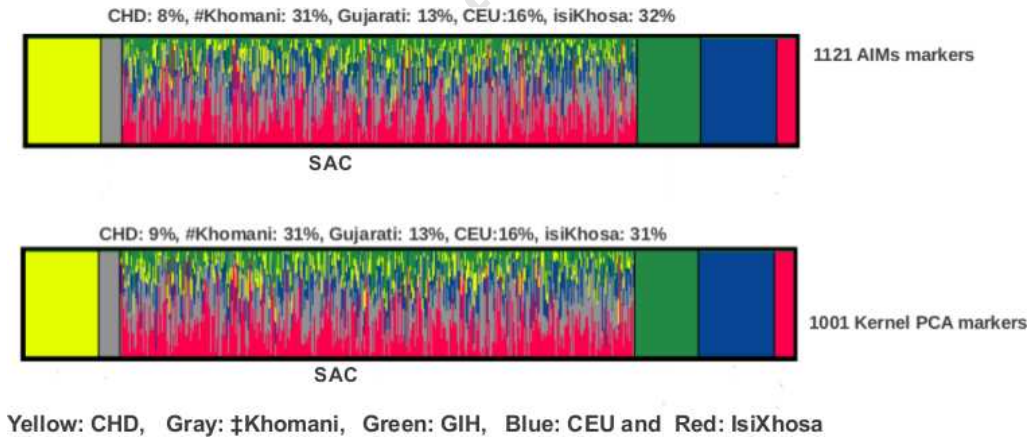

Figure 3.1: **Individual's ancestry proportion based on 1121 AIMs obtained from the method described in section 3.2.1 (Top plot) and 1001 Kernel-PCA markers obtained from the method described in section 3.2.2 (bottom plot).**

3.3.2 Assessing Admixture LD

To assess the pattern of admixture LD in the SAC as a result of ancestral admixture, we first compared LD between the SAC and its putative proxy ancestors. We calculated the LD ($r^2 > 0.2$) across the whole genome of each population and found that LD is consistently higher at very short distances in the SAC (Figure. 3.2).

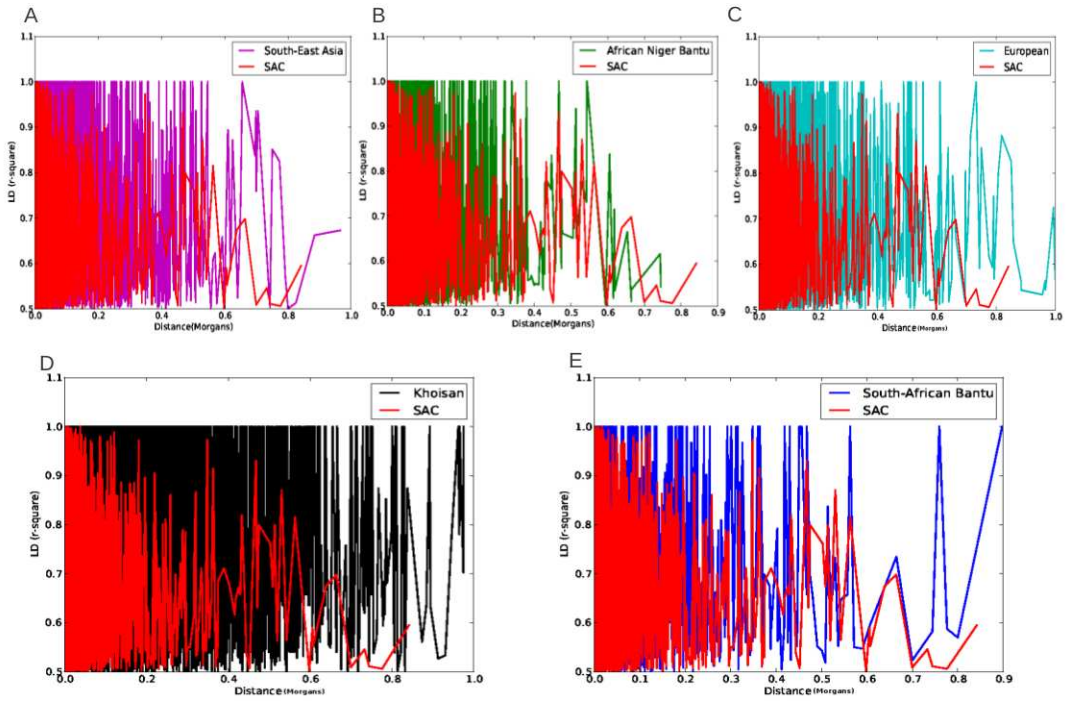

Figure 3.2: **LD across 1121 AIMs markers between the South African Coloured populations and the five proxy ancestral groups. (A-E) the LD plot ($r^2 >= 0.5$) is between pairs of SNPs (combined linked and unlinked AIMs SNPs) within 1.2 Mb from each other. In the figure, we denote ‡Khomani, CEU, CHD+Gujarati Indian, isiXhosa and Yoruba as Khoesan, European South-East Asian, South-African Bantu and African Niger Bantu populations, respectively.**

The LD in the SAC decays from regions > 0.2 Morgan (Figure 3.2), suggesting that this LD may primarily be as a result of admixture rather than founder effects. This finding is consistent with prior studies that established that the admixture LD decays within a few generations at long distances ($> 20cM$) but decays slowly at short distances ($< 10cM$) (Chakravati & Weiss, 1998; Li & Stephens, 2003). Recent admixture between genetically differentiated populations gives rise to an increase in admixture LD proportion (Winkler *et al.*, 2010).

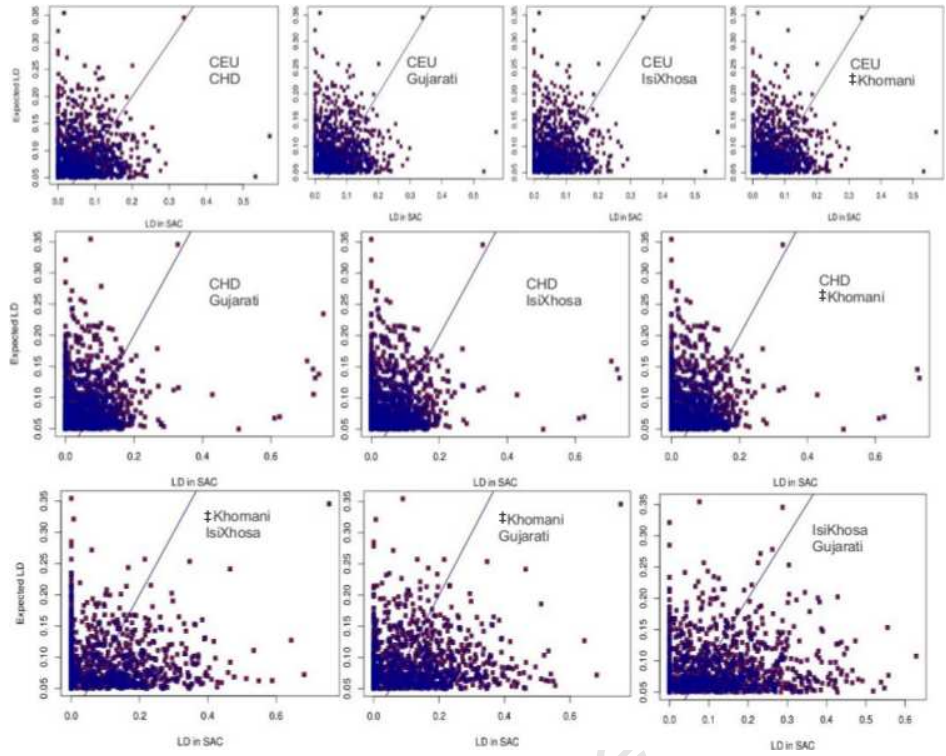

Figure 3.3: **Admixture LD in the SAC as consequence of the admixture events from proxy ancestral populations (CEU, ‡Khomani, CHD, Gujarati and isiXhosa). To generate these plots, we computed the LD between all pairs of markers in the SAC and the expected admixture from each pair of ancestral populations. The plots show the scatter of LD in the SAC (red dot) and the expected admixture LD in any two pairs of ancestral populations (blue dot).**

To test for the admixture LD due as consequence of the admixture events from the five proxy ancestral populations of the SAC, we computed the LD between all pairs of AIMs (n $= 1121$ AIMs) in the SAC, weighted by their frequency difference (see section 3.2.3) between each pair of these five proxy ancestral populations, including isiXhosa, ‡Khomani, Central European (CEU), Gujarati Indian and Chinese (CHD). Through linear regression of the allele frequency differences of each pair of proxy ancestral groups with LD in the SAC, we obtained a correlation ($R^2 = 0.74$, intercept$= 0.38$, slope $= 0.41$) with a significant p-value $= 0.000018$, indicating an association of allele frequency differences with increased LD in the SAC.

We finally estimated the maximum expected admixture LD (see section 3.2.3) from each pair of proxy ancestral populations and we compared them with the observed LD in the SAC. Table 3.1 shows the correlation between the expected admixture LD from each pair of proxy ancestral groups and the observed LD in the SAC, which is significant (Figure 3.3). Through an additive linear model, we obtained a lower p-value $= 2.2e^{-16}$ under the null hypothesis of no correlation

between LD in the SAC and these expected admixture LDs, indicating that the LD in the SAC correlated with the expected admixture LD and mainly has its origin in different admixtures from the five proxy ancestral populations. This result confirms that admixture between populations related to these five proxy ancestral groups (isiXhosa and ‡Khomani, Central European (CEU), Gujarati Indian and Chinese (CHD)) largely contributed to the admixture LD observed in the present SAC population.

Table 3.1: **P-value obtained from the correlation between expected admixture LD from each pair of proxy ancestral group with respect to the observed LD in the SAC.**

Pair-wise populations	**P-value**	$OR[95\%CI]$
(CHD, Gujarati)	$7.25e-10$	$0.99[0.99, 1.00]$
(isiXhosa, Gujarati)	$9.35e-8$	$0.98[0.97, 0.99]$
(CEU, CHD)	0.92	$0.99[0.99, 1.001]$
(CHD, ‡Khomani)	$4.34e-10$	$0.98[0.97, 0.99]$
(‡Khomani, isiXhosa)	$1.01e-08$	$0.96[0.94, 0.97]$
(‡Khomani, Gujarati)	$1.21e-8$	$0.97[0.95, 0.98]$
(CEU, Gujarati)	0.42	$0.99[0.98, 1.0]$
(CEU, ‡Khomani)	$7.16e-7$	$0.99[0.98, 1.0]$
(CHD, isiXhosa)	$8.076e-10$	$0.98[0.97, 0.998]$
(CEU, isiXhosa)	$3.79e-06$	$0.99[0.98, 1.00]$

Importantly, to support our approach for testing the admixture LD in the SAC as a consequence of admixture events resulting from ancestral populations related to the five proxy ancestral populations, we estimated the weighted LD decay curves in the SAC using ALDER using all available SNPs (see section 2.2). The results in Figure 3.4 are consistent with the result obtained in Table3.1 and Figure 3.3. All the results suggest the admixture increased its genetic diversity and that the observed LD in the SAC has mainly its origin from the admixture.

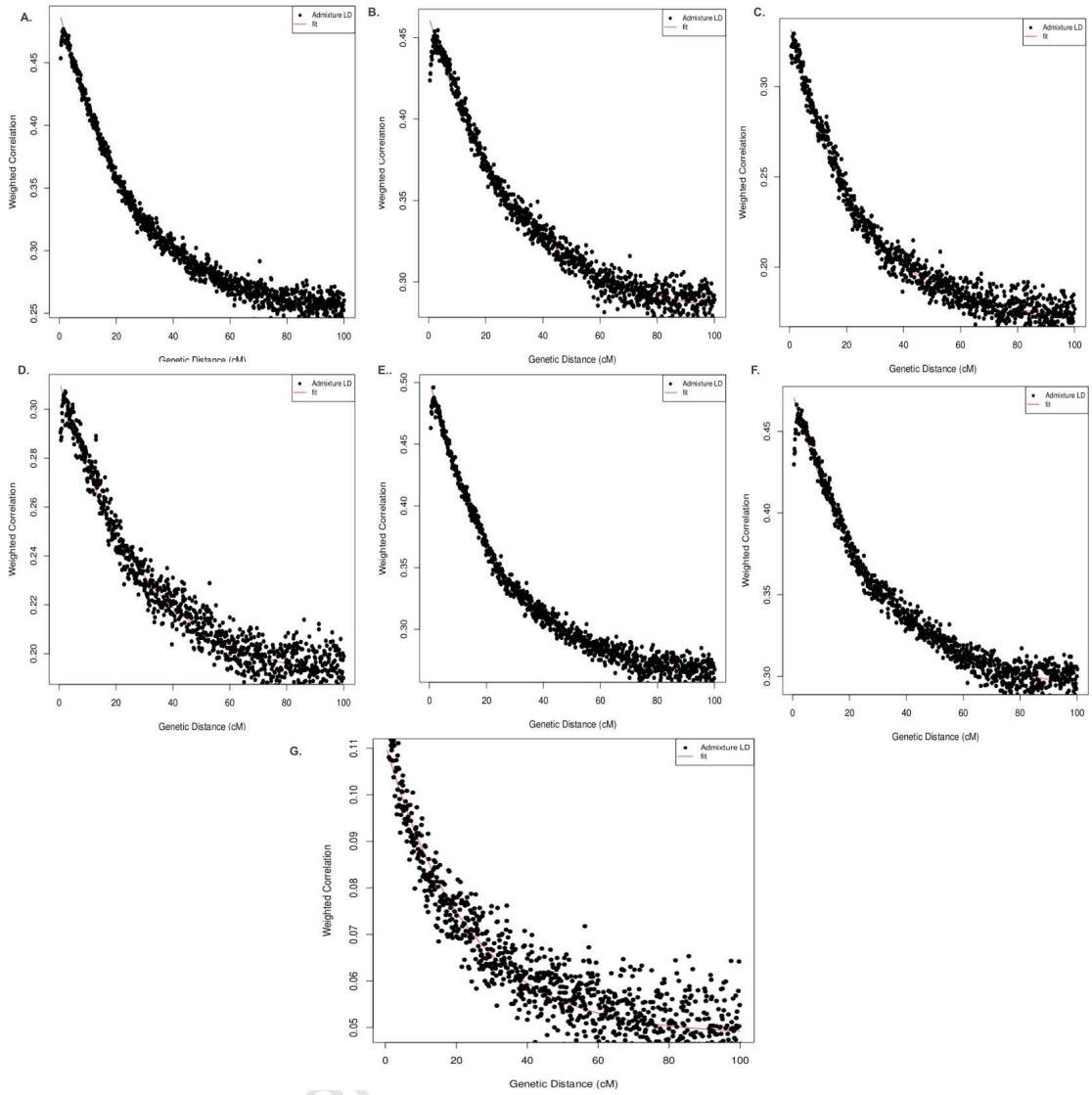

Figure 3.4: **Weighted LD decay curves in the South African Coloured population with any two pairs of ancestral populations. These plots show the decay of admixture LD in the SAC respect to each pair of its proxy ancestry as consequence of the admixture events. (A)CEU and ‡Khomani within the SAC. (B) CEU and isiXhosa within the SAC. (C) CHD and ‡Khomani with the SAC. (D) CHD and isiXhosa with the SAC. (E) GIH and ‡Khomani with the SAC. (F) GIH and isiXhosa within the SAC. (G) ‡Khomani and isiXhosa within the SAC. All SNPs were used to generate these plots.**

3.3.3 Genetic Diversity and Haplotype Identity-by-Descent.

We compared the genome-wide haplotype diversity and the percentage haplotype sharing by IBD (see Materials and Methods), and the result in Table 3.2 indicates that the SAC has a higher haplotype diversity than any of its five proxy ancestral groups. The result suggests that both the

higher diversity and higher LD at short distances observed in the SAC are the result of admixture events, and not founder effects or an extreme bottleneck. In addition, we found that the SAC has a higher percentage of shared haplotype segments by IBD at short distances (in the region $< 2.5cM$) than three of the proxy ancestral groups (Table 3.2), which is also consistent with the observed admixture LD. The pairwise IBS permutation test confirmed the greater genetic variation among the SAC samples, and indicated that the average pair of SAC individuals has significantly less genome-wide IBS sharing than pairs of each proxy ancestral groups (empirical p-value $= 0.00202$). The observed higher level of genetic diversity in the admixed SAC is likely to be the result of the geographic location of South Africa with respect to major trade routes in the past (from the 15^{th} to the 19^{th} centuries) and its history of multi-faceted colonization (Mountain, 2003).

Table 3.2: **Comparing genetic diversity between the South African Coloured population (SAC) and the five proxy ancestral groups contributing to the SAC admixture. Mean and standard error of shared haplotype segment in cM (Hap.segment), mean and standard error of haplotype diversity measure (Hap.diversity) and proportion of IBD (Prop.IBD).**

	Hap. Segment	Hap. diversity	Prop. IBD
SAC	1.022 ± 0.004	81.975 ± 0.002	(0.0018)
isiXhosa	0.9058 ± 0.042	16.860 ± 0.003	(0.0284)
‡Khomani	1.123 ± 0.033	5.214 ± 0.004	(0.1714)
CEU	1.192 ± 0.043	50.544 ± 0.003	(0.0189)
CHD	0.715 ± 0.0417	54.885 ± 0.003	(0.1051)
Gujarati	0.614 ± 0.042	57.883 ± 0.003	(0.0512)

3.4 Discussion

We implemented two complementary algorithms (supervised and unsupervised) to select ancestry informative markers in a multi-way admixed population, particularly we use these algorithms to construct two panels of AIMs for the SAC. Furthermore, these two algorithms performed as well as using all available SNPs in estimating individual's ancestry proportion in the SAC (see section 2.3.2.2). Our first algorithm has an advantage of selecting SNPs based on the relationship between ancestral population (using selected proxy ancestors of the admixed population) differentiation and the observed admixture Linkage Disequilibrium in the admixed population. The AIMs panels from this algorithm were used to examine the pattern of linkage disequilibrium in this population, in comparing it with those from its proxy parental populations.

A higher degree of LD is expected in admixed populations, and this could at certain points of its history be influenced by population bottlenecks, or only be a result of the admixture itself. We demonstrated in the SAC population that the allele frequency differences between each pair of proxy ancestral populations correlated with increased LD, suggesting that the admixture increased the genetic diversity and that the observed LD in the SAC has its origin mainly from the admixture. This study observed a weak level of founder haplotypes identical-by-descent along the genome of the SAC, which strengthens the evidence against population bottlenecks that could have been found as a consequence of the past legislated separation of ethnic groups in South Africa, including the SAC.

Chapter 4

Genome-wide Association Study of Ancestry-specific TB Risk in the South African Coloured Population.

4.1 Introduction

Tuberculosis (TB) remains a source of morbidity and mortality worldwide, particularly in developing countries. It is a leading cause of HIV-related deaths, as almost one in four deaths among people with HIV infection is due to TB (Kaufmann & McMichael, 2005; WHO, 2000). In 2010, there were 8.8 million new cases of TB, of which 1.1 million were among people living with HIV (Dye *et al.*, 1998a; WHO, 2000). TB susceptibility is well known to be a complex trait influenced by both environmental and genetic factors (Comstock, 1978). The environmental factors that influence TB susceptibility include smoking, socio-economic conditions, and acute infection (Babb *et al.*, 2007; Bellamy, 1998; Bellamy *et al.*, 2000). One-third of the world's individuals are infected with TB, but only 10% go on to develop active TB during their lifetime (www.who.int/tb/en/) (Dye *et al.*, 1998a,b). In addition, twin studies in humans and animal models also demonstrate a strong genetic influence on TB susceptibility (Comstock, 1978; Sorensen *et al.*, 1988). The differing rate of concordance of TB among monozygous compared with dizygous twins was reported from these twin studies in tuberculosis. The rate of concordance of TB among monozygotic twins (18/55, 32.7%) was more than twice (odds ratio of concordance: 2.4; 95% CI : 1.44.0) that observed among dizygotic twins (21/150, 14.0%) (Flynn, 2006; Sorensen *et al.*, 1988). These estimates suggest that genetic factors may play an important role in TB susceptibility in determining both the host response and the outcome of infection (Daniel, 1997; Kaufmann & McMichael, 2005).

Several cohort studies have demonstrated that the incidence of tuberculosis varies considerably depending on the population and region studied (Dye *et al.*, 1999; Small, 1996). Therefore, it is becoming increasingly evident that analysis of the correlation between genetic ancestry contribution and phenotype in recently admixed populations can improve the predictions of disease and provide crucial insights into medical genetics (Kumar *et al.*, 2010). Among these studies, Kumar and colleagues examined whether the genetically determined percentage of African ancestry is associated with lung function and whether its use could improve predictions of lung function in African American populations (Kumar *et al.*, 2010), their results suggested genetic ancestry exert a major influence in improving lung-function estimates and categorizing asthma severity. Overall, these results suggested that even within ethnic groups, genetic factors exert a major influence in susceptibility. Therefore, investigating ancestry-specific disease risk in multi-way admixed population may provide crucial insight for biomedical research.

The second highest incidence of TB in the world is in the Western, Eastern and Northern Cape in South Africa, particularly in the admixed South African Coloured population (Babb *et al.*, 2007; Bellamy *et al.*, 2000; Hoal *et al.*, 2004). However, ancestry-specific TB risk has not yet been considered in this population with mixed ancestry. In this chapter, the aim is to evaluate the genetic ancestry of samples of TB cases and controls from this population. Importantly, we examine whether the genetic contribution can increase tuberculosis prevalence, and evaluate the contribution of socio-economic status to the ancestry-specific TB risk. In addition, due to the land-borne immigrants of sub-Saharan (West and East) Africans originally, followed by more recent sea-borne immigrants from Europe, Asia and Indonesia to shape the genetic make-up of the SAC, and due to the observed difference of individual's ancestry proportion in TB cases and controls from both click-speakers/African and Non-African ancestry (Table 4.4), it is particularly meaningful to investigate whether there is an excess of common SNPs with large allele frequency differences between TB cases and controls samples from the South African Coloured population.

4.2 Materials and Methods

4.2.1 Genetic Ancestry and TB Risk Relationship

Socio-economic (SES) questionnaires were available for 82 cases and included information on two categories of income, per week self and per week household. These incomes were estimated based on the South African Rand (R) currency. These incomes were coded as follows: $0 =< R50$, $1 = R50$ to $R150$, $2 = R150$ to $R300$, $3 = R300$ to $R500$, $4 = R500$ to 1000, $5 => R1000$ and $9 =$ missing. We first computed the fraction of ancestry for each individual from five putative ancestral populations using the program ADMIXTURE (Alexander *et al.*, 2009). We separately

regressed TB status against genetic ancestry proportion from each ancestral population. We evaluate the correlation between pairs of ancestral populations. To control the correlation between genetic ancestry in the SAC which can potentially be confounded, we test for the difference in TB risk (conditional risk) between pairs or triple of ancestral populations.

Suppose β_k and ε_k are the effect size and standard error from the regression model of the fraction of ancestry k in the admixed population against TB binary trait, respectively. To test for the difference in TB risk between pairs of ancestral populations k and l, we have to adapt the normal test statistic under the null hypothesis of no difference in risk between two ancestral populations. Thus, we computed the Zscore of difference in risk, $Z_{kl} = (\beta_k - \beta_l)/\sqrt{+}$, $(k \neq l)$ which has a standard normal distribution $Z_{kl} \sim N(0,1)$. We computed the probability, (two-sided p-value $= 2 * (1 - P(< |Z_{kl}|))$ that the value may be less than the Zscore. To account for the correlation among ancestry proportions in the admixed population, we first conducted a permutation test whereby the above distribution of the test statistic under the null hypothesis is re-sampled 10000 times under the rearrangements of the case/control status. In addition, we adjusted for the covariance by computing $cov_{kl} = \rho_{kl} \times \varepsilon_k \times \varepsilon_l$ where ρ_{kl} is the correlation of the fraction of ancestry from ancestral population k and l. We derived the corrected test statistic by subtracting out $2 \times cov_{kl}$. Thus, the above test is applied between pairs (and triples) of ancestral populations, each African/non-African ancestral groups conditional on non-African/African ancestral groups and each ancestral group conditional on all others.

We additionally computed the correlation between TB-ancestry and ancestry-SES. Because we have socio-economic data only for TB cases, we regressed socio-economic status against genetic ancestry. Naturally this sample size may not provide sufficient power to identify correlations. Fortunately, because of the uniform ethnicity and socio-economic status where the SAC's case/control sampling was conducted (Materials and Methods), we derived the relationship between TB status and socio-economic status based on the correlations (a 95% confidence interval on the correlation of ancestry and socio-economic status) obtained from TB-ancestry and ancestry-SES models.

4.2.2 Unusual Difference in Allele Frequency

Accounting for minimization of deviation from the normality assumption, SNPs with minor allele frequencies < 0.05 are excluded. Thus, at a given locus i, the difference $(p_i^k\text{-}p_i^l)$ between observed variant allele frequencies of two populations, k and l, can be approximated as a normal distribution under neutral drift with mean 0 and variance $p(1-p)(2F_{ST} + \frac{1}{N_k} + \frac{1}{N_l})$ (Price *et al.*, 2009a); where F_{ST} is the genetic distance between populations k and l. N_k and N_l are total variant allele counts in each population, and p is the ancestral allele frequency that is commonly approximated

as the average of the two observed variant allele frequencies (Price *et al.*, 2009a). As in (Price *et al.*, 2009a), it follows that

$$U_{kl}^1 = \frac{(p_i^k - p_i^l)^2}{[p(1-p)(2F_{ST} + \frac{1}{N_k} + \frac{1}{N_l}]}, \tag{4.1a}$$

$$U_{kl}^2 = \frac{(p_i^k - p_i^l)^2}{p(1-p)}. \tag{4.1b}$$

Equations (4.1a) and (4.1b) above are χ^2 distributed with 1 degree of freedom (d.o.f.), and can be applied to unrelated and related samples, respectively. An excess of large values of the χ^2 statistic indicates deviations from the null model (equations (4.1a) and (4.1b)), suggesting the action of natural selection (Price *et al.*, 2009a).

4.3 Results and Discussion

4.3.1 Relationship between TB Risk and Genetic Ancestry

To examine the relationship between genetic ancestry and TB status in this SAC data set, we regressed case-control status against the estimated fraction of isiXhosa, ‡Khomani, Gujarati, CHD and CEU ancestry, respectively, in 733 unrelated SAC individuals (section 2.2.1). We observed a statistically significant correlation ($r = 0.165$, OR $95\%CI = 1.46[1.23, 1.79]$, $p = 1.58e-05$) between ‡Khomani ancestry and TB status. The CEU ($r = -0.122$, OR $95\% = 0.71[0.58, 0.86]$, $p = 0.000657$), CHD ($r = -0.13$, OR $95\%CI = 0.42[0.26, 0.68]$, $p = 0.000489$) and Gujarati ($r = -0.011$, OR $95\% = 0.65[0.50, 0.85]$, $p = 0.00192$) ancestry in the SAC were negatively correlated with TB status (Table 4.1).

Table 4.1: **Association of genetic ancestry with TB risk in the South African Coloured population, with nominal p-values before correcting for hypotheses tested.**

Model	(TB-ancestry)	(ancestry-self incomes)	(ancestry-household incomes)
POP	Correlation, OR 95% CI, p-value	Correlation, OR 95% CI, p-value	Correlation, OR 95% CI, p-value
‡Khomani	0.165, $1.46[1.23, 1.79], 1.58e-05$	-0.013, $1.00[0.99, 1.02]$, 0.741	-0.011, $1.01[0.99, 1.04]$,0.399
isiXhosa	0.06,$1.11[0.97, 1.30]$,0.10	-0.012, $0.99[0.98, 1.02]$, 0.86	-0.027,$0.99[0.97, 1.02]$,0.34
CEU	-0.122,$0.71[0.58, 0.86]$,0.0007	-0.006, $0.99[0.98, 1.01]$,0.459	-0.037,$1.01[0.98, 1.03]$,0.689
Gujarati	-0.111,$0.65[0.50, 0.85]$,0.002	-0.006,$1.00[0.99, 1.01]$,0.437	0.036,$0.99[0.97, 1.00]$,0.0185
CHB+JPT	-0.123,$0.42[0.26, 0.68]$,0.0005	-0.014,$1.00[0.99, 1.01]$,0.916	-0.041,$1.00[0.99, 1.01]$, 0.779

Table 4.2: **The correlation between the fraction of ancestry from five putative ancestral populations (isiXhosa, ‡Khomani, CEU, CHD and Gujarati, respectively) of the South African Coloured population. The table displays the OR[**$95\%CI$**] and p-value of the ancestry correlation. There is correlation between all ancestral groups.**

	isiXhosa	CEU	Gujarati	CHD
‡Khomani	$0.9[0.81, 0.91]$, $8.9e-07$	$0.7[0.63, 0.74]$, $2e-16$	$0.5[0.42, 0.52]$,$2e-16$	$0.2[0.19, 0.28]$, $2e-16$
isiXhosa	-	$0.4[0.38, 0.45]$, $2e-16$	$0.4[0.33, 0.43]$, $2.2e-16$	$0.3[0.19, 0.31]$, $2.2e-16$
CEU	-	-	$1.4[1.24, 1.53]$, $2.9e-09$	$2.0[1.68, 2.4]$,$4.9e-14$
Gujarati	-	-	-	$3.2[2.8, 3.48]$, $2.2e-16$

isiXhosa ancestry proportion was not significantly correlated ($r = 0.06$, OR $95\%CI = 1.11[0.97, 1.30]$, $p = 0.10$) with TB. Similar results were obtained when including age and gender as covariates in the analysis. Furthermore, we observed a statistically significant correlation of age ($r = 0.165$, $p = 1.01e-05$, mean age 37 in cases and 31 in controls) with risk of TB and no evidence of correlation between sex and TB risk ($r = -0.039$, $p = 0.597$). We computed the correlation between the fraction of ancestry from these five putative ancestral populations, and found a correlation between all ancestral groups (Table 4.2).

Due to the correlation between the individual ancestry fractions (Table 4.2), we additionally checked if the above test can be potentially confounded by testing for the difference in TB risk (conditional risk test) between pairs/triples of ancestral populations and each ancestral group conditional on all others (see Material and Methods). Our results demonstrate that African ancestry (‡Khomani, isiXhosa) related TB risk in the SAC is not significantly conditional on non-African ancestry (CEU and CHD) risk. With the exception of Indian (Gujarati) ancestry, non-African ancestry (CEU and CHD) risk is significantly conditional on African ancestry risk (Table 4.3). What we have shown is isiXhosa and ‡Khomani are differently correlated with risk than CEU, CHD and Gujarati, and are not significantly conditional on either, respectively (Table 4.1 and Table 4.3). CHD, Gujarati and CEU are not significantly conditional on each other and all correlated with TB risk (Tables 4.1 and Table 4.3). We see that ‡Khomani confers risk, CEU, CHD and Gujarati confer protection, and isiXhosa shows no evidence of correlation (Table 4.1).

4.3.2 Relation between TB Risk and Socio-economic Status

A potential concern was that the observed relationship between genetic ancestry and TB status could be a consequence of confounding due to socio-economic status (SES), as described in a recent study of type 2 diabetes in Latinos (Florez *et al.*, 2009). We investigated this possibility by studying two SES variables (see Materials and Method), household income and individual income. These variables were available in only a subset of 82 SAC cases. When testing for correlations between each of these variables and each of the five ancestries, none of the results were statistically significant after correcting for 10 hypotheses tested (Table 4.1). However, ‡Khomani ancestry had a non-significant (after correction) trend towards positive correlation ($95\%r = -0.013[-0.018, -0.008]$, OR $95\%CI = 1.00[0.99, 1.02]$ and nominal $p = 0.741$) with SES. This would not explain the correlation ($95\%r = 0.165[0.046, 0.283]$) and (OR $95\%CI = 1.46[1.23, 1.79]$, nominal $p = 1.58e-05$) between ‡Khomani ancestry and TB status, for two reasons. Firstly, the correlation with SES was smaller than the correlation with TB status, so that even if TB status was 100% determined by SES status, which is highly unlikely, the correlation with TB status could still not be explained. Secondly, the correlation with SES is in the wrong

direction to explain the correlation between ‡Khomani ancestry and TB status, since TB status is usually associated with low SES (deWit *et al.*, 2010b; Hudelson, 1996; WHO, 2004, 2005). Given the obtained 95% confidence interval from the correlation between SES and ancestry based on 82 samples analysed and that between ancestry and TB based on 733 unrelated samples, this provides evidence that a negative correlation does not exist between ‡Khomani ancestry and SES that would be sufficient to explain the correlation between ‡Khomani ancestry and TB status in this population. Therefore, the observed ancestry difference between cases and controls (Table 4.4) is unlikely to be a direct consequence of socio-economic status in this population.

Table 4.3: **Ancestral population pair-wise conditional risk test. The values in table are the p-value, $OR[95\%CI]$ from the corrected test and adjusted for the covariance.**

	isiXhosa	‡Khomani	CEU	CHD	Gujarati
isiXhosa	-	0.001, 0.90[0.86, 0.96]	0.0047, 0.91[0.87, 0.97]	0.0003, 0.91[0.87, 0.96]	0.005,0.75[0.63, 0.88]
‡Khomani	-	-	0.0001, 0.99[0.99, 1.0]	0.0003, 0.9[0.99, 1.0]	0.0002,0.75[0.64, 0.87]
CEU	-	-	-	0.0001, 0.9[0.99, 1.0]	$6.4e-05$, 0.73[0.62, 0.85]
CHD	-	-	-	-	0.0002, 0.74[0.63, 0.86]
(isiXhosa,‡Khomani)	-	-	0.098, 0.9[0.8, 1.01]	0.16,0.92[0.81, 1.03]	0.001,0.72[0.59, 0.88]
(CEU, CHD)	0.0015, 0.89[0.84, 0.96]	0.84,0.99[0.95, 1.03]-	-	0.0006,0.75[0.63, 0.88]	
(Gujarati, CHD)	$3.5e-21$,0.7[0.65, 0.75]	$1.1e-27$,0.72[0.68, 0.88]	$5.4e-25$, 0.75[0.71, 0.79]	-	-
(CEU, Gujarati)	0.03,0.92[0.85, 0.99]	0.88,0.99[0.94, 1.1]	0.97,0.99[0.94, 1.1]		
(CEU,Gujarati,CHD)	0.002,0.92[0.87, 0.97]	0.003,1.0[0.99, 1.0]	-	-	-
All Other	0.0006,0.92[0.87, 0.96]	0.0007,0.95[0.9, 0.96]	1.02,0.97[0.9, 0.99]	1.007,0.97[0.91, 0.99]	0.0003,0.74[0.63, 0.87]

4.3.3 Unusual Difference in Allele Frequency from TB Case-control Study in the SAC

We compute the differences between ancestry fractions in the TB cases and the controls from each of these five putative ancestral populations, Table 4.4 displays these results. We observed that the TB cases have slightly higher African components (‡Khomani and isiXhosa), while the controls have greater non-African (CEU, Gujarati, and CHD) contributions.

Table 4.4: **Mean and standard error of ancestry proportion from each of five populations contributing to the admixture in the South African Coloured (using 90 controls and 623 cases) population.**

	IsiXhosa	‡Khomani	CEU	CHD	GIH
Control	0.29 ± 0.16	0.24 ± 0.11	0.22 ± 0.12	0.09 ± 0.04	0.15 ± 0.07
Case	0.32 ± 0.18	0.31 ± 0.13	0.18 ± 0.11	0.07 ± 0.04	0.12 ± 0.07
Overall	0.31 ± 0.18	0.30 ± 0.14	0.18 ± 0.11	0.08 ± 0.049	0.13 ± 0.08

We examine whether there is an excess of common SNPs with large allele frequency differences between the SAC case and control individuals. We computed the distribution of allele frequency differences between between the SAC 761 case and 91 control individuals, expressed as a χ^2 (1 d.o.f.) statistic under a model of neutral genetic drift (see section 4.2.2). The most significant P-value was $e-04$, a value that is not statistically significant after correcting for the number of SNPs and regions tested. This result is consistent with the hypothesis that the date of the admixture event to produce the SAC is recent and has been too short for differential selective forces to have had a significant impact on allele frequencies.

4.4 Conclusion

In summary, we used a combination of two complementary methods to examine whether the genetic contribution from particular ancestral population can increase tuberculosis risk, and evaluated the contribution of socio-economic status (SES) to the ancestry-tuberculosis relationship in the SAC. Our results demonstrated significant evidence of an association between Khoesan ancestry (‡Khomani) and TB status that is not confounded by SES. This an important epidemiological result and illustrates the value of the inclusion of admixture association methods in the set of methods used to conduct TB association studies in this population. When the extremely high incidence of TB in the SAC population is considered, together with our finding that a significant percentage of their ancestry is derived from the ‡Khomani and other African populations,

it appears possible that there may be an element of population level genetic susceptibility to this disease. Our study is the first investigation of ancestry-specific TB risk in this population. In addition, the model introduced for assessing possible evidence of an excess of common SNPs with large allele frequency differences can be applied to any pair of populations in order to detect signatures of natural selection.

Chapter 5

Genome-wide Scan for TB Risk in the Admixed South African Coloured Population.

5.1 Introduction

As mentioned in the previous chapter, the second highest incidence of TB in the world is in the Western, Eastern and Northern Cape in South Africa, particularly in the admixed South African Coloured population (SAC). Investigations based on candidate genes studies and genome-wide linkage scans on the data of the admixed South African Coloured population were previously conducted (Hoal *et al.*, 2004; Moller & Hoal, 2010a,b; Moller *et al.*, 2009). Babb *et al.* (2007) investigated a cohort of pulmonary TB patients in South African populations to determine whether three polymorphisms of the vitamin D receptor gene (VDR), namely polymorphisms FokI, known to be a functional polymorphism, ApaI, known to be in intron vIII, and TaqI, known as a silent polymorphism (T/C) located in exon IX, were associated with TB susceptibility. From their analysis, they reported no significant association between pulmonary TB and the VDR polymorphisms. However, the Fat haplotype was reported to possibly be protective against TB as it was unusually over-represented in controls compared to cases. In the same vein, Hoal *et al.* (2004) investigated the association between *SLC11A1* (*NRAMP1*) polymorphisms and susceptibility to TB, and whether polymorphisms in *SLC11A2* are associated with TB. Their case-control study design was based on the data from the Western Cape region of South Africa and certain suburbs of metropolitan Cape Town. They reported that the $5(GT)9$ allele in the promoter of *SLC11A1* was associated with protection against TB in the majority of the populations studied. Surprisingly, the *SLC11A2* (*NRAMP2*) polymorphism was not associated with susceptibility to TB in this high-incident community of South Africa, which includes the SAC. Although, these early TB

genetic studies on the SAC were restricted to well-characterized markers within genes, most of them failed to observe a statistical association with the markers that were examined and resulted in inconclusive results (Moller & Hoal, 2010a,b). Moreover, the use of too few genomic control markers to correct for potential population substructure in most these studies such as deWit *et al.* (2010b) and Barreiro *et al.* (2006), may result in not correcting the bias (false positive/negative) in results, as mentioned in Marchini & Howie (2008). Not using enough SNPs to capture the linkage disequilibrium in the admixed SAC may also substantially affect the power to detect significant association in these analyses. Despite some failures, a few genetic association studies have identified candidate genes for tuberculosis susceptibility using data from the admixed South African Coloured population (Moller & Hoal, 2010a), but with recently, GWAS for TB had not yet been considered in this population.

Genotyping techniques and genome-wide advanced statistical approaches have resulted in moving from the candidate gene-based association analysis approach to genome-wide association studies (GWAS). GWAS does not require a prior hypothesis related to disease associated genes or knowledge of susceptibility genes or gene functions (Hirschhorn & Daly, 2003; Kennedy *et al.*, 2003; Risch, 2000). From a recent review on GWAS in Rosenberg *et al.* (2010), GWAS have successfully identified genetic variants that contribute to complex human diseases mainly in European populations. Dispite these successes, possible technical challenges with using non-Europeans populations, in particular African populations for GWAS, was recently debated in Rosenberg *et al.* (2010). These challenges include: the smaller extent of linkage disequilibrium (LD) between variants in African populations, resulting in a limited coverage of their common variation panels; and genotype-imputation and tag-SNP portability commonly based on the HapMap populations may be reduced due to the level of the population structure and the genetic diversity across African populations. In spite of these limitations in using non-Europeans populations for GWAS, recent waves of GWA studies in non-European populations began to gain success. Non-European GWAS successes include investigations on Japanese (Unoki & et.al, 2008; Yasuda & et.al, 2008), Korean (Cho *et al.*, 2009; Kim & et.al, 2009), Chinese (Garcia-Barceloa *et al.*, 2009; Zhang *et al.*, 2009) and recently on combined Ghana, Gambia and Malawi (Thye *et al.*, 2010, 2012) populations.

The progress in identifying new contributing genetic variants through GWAS in African host susceptibility to infectious disease, such as TB, has so far been slow and weakened due to study design (Moller & Hoal, 2010a,b; Stein, 2011), small sample size of the population under study, and the small number of genotyped SNPs ($300K$-$500K$). However, Thye and colleagues conducted a combined GWAS to investigate the host susceptibility to pulmonary tuberculosis, using $2,100$ cases and $3,000$ controls from African populations in Ghana and Gambia, with replication in a combination of $11,425$ individuals from both Ghana and Malawi (Thye *et al.*, 2010). A single SNP on chromosome $18q11$ was found to be associated with disease. Recently, Thye *et al.* (2012)

reported a new TB susceptibility locus on chromosome $11p13$ after imputation of genome-wide data from Ghana. This finding was replicated in samples from Gambia, Indonesia and Russia (Thye *et al.*, 2012). Furthermore, Davila *et al.* (2008) identified four polymorphisms in the *TLR8* gene on chromosome X, including $rs3764880$, $rs3764879$, $rs3761624$ and $rs3788935$ to be associated with TB susceptibility; the association was replicated in males from a follow up cohort from Russia (Davila *et al.*, 2008). Recently, a study by Dai *et al.* (2011) used a cohort of over one thousand Chinese TB patients and $1,280$ healthy controls using melting temperature shift allele-specic genotyping analysis to determine whether the identified SNP in Thye *et al.* (2010) are associated with TB in the Chinese population. Importantly, SNP $rs4331426$ in chromosome $18q11$ was signicantly associated with TB in the Chinese population, but the effect was opposite to the finding in Thye *et al.* (2010).

As mentioned above, few genetic association studies have implicated candidate genes in tuberculosis susceptibility from the data of the admixed South African Coloured population, but until recently tuberculosis genome-wide association studies had not yet been performed in this population. The SAC, has a mixed ancestry traced back over 350 years from various populations (see chapter 2). This variation among admixed individuals in their proportions of ancestry could result in spurious associations between genotypes and phenotypes (Marchini & Howie, 2008; Rosenberg *et al.*, 2010). Some authors argue that using admixed populations in a GWAs is the same as using different sub-populations in a larger population (Marchini & Howie, 2008; Rosenberg *et al.*, 2010). Fortunately, a well-designed GWAS and statistical tool can control false-positive/negative associations due to both population structure and local ancestry (Qin *et al.*, 2010; Redden *et al.*, 2006; Rosenberg & Nordborg, 2006; Setakis *et al.*, 2006; Zhu *et al.*, 2008). In addition, although the LD between SNPs in recently admixed populations can differ, they have a much greater LD as a new population compared to other, more ancient Africa populations (such as Yoruba, Ghana, Gambia), therefore both GWAS and genotyping imputation in recently admixed populations are feasible.

GWAS of admixed populations was recently proposed to be informative for diseases for which risk differs depending on ancestry prevalence (Pasaniuc *et al.*, 2011; Seldin *et al.*, 2011). These recent methods involve joint modelling of the admixture (accounting for local ancestry) and SNP-association signals. Recent methods have shown, in simulation and real data (Pasaniuc *et al.*, 2011), increased statistical power compared to using SNP case-control and admixture association separately (Pasaniuc *et al.*, 2011). Furthermore, an accurate and unbiased estimation of the ancestry at every SNP in multi-way admixed populations was suggested to potentially provide crucial insights into identifying disease genes in these populations (Baran *et al.*, 2012; Pasaniuc *et al.*, 2011; Seldin *et al.*, 2011). However, the accuracy of most inference of local ancestry approaches, which is one of the first steps in these admixture association studies, is limited when

using multi-way admixed populations such as the SAC. The joint modelling of the admixture and SNP association signals are only successful when applying them to two-way admixed populations such as African-Americans (Pasaniuc *et al.*, 2011; Seldin *et al.*, 2011). In addition, methods developed for disease scoring in admixed populations have successfully been applied to two or three-way admixed populations such as African Americans and Hispanic Americans, but do not apply to multi-way admixed populations (Kang *et al.*, 2010; Pasaniuc *et al.*, 2011). Here, our main focus is to identify possible association signal in the multi-way admixed Coloured population. To address this, we conduct GWAS with correction for genome-wide ancestry, accounting for both population stratification and hidden relatedness that can result from the genealogy.

5.2 Materials and Methods

5.2.1 Population Study, Quality Control

Because of the high incidence of tuberculosis in the metropolitan area of Cape Town in the Western Cape Province in South Africa as well as the uniform ethnicity, socio-economic status and low prevalence of HIV, this area was selected for sampling (Hirschhorn & Daly, 2003). This is also due to the following reasons:

(1) Uniform ethnicity and socio-economic status is important in disease association studies as it removes some of the confounding variables.

(2) Low prevalence of HIV is important because in the presence of HIV infection, an individual has a greatly increased chance of progressing to TB disease once infected, simply because of an impaired immune system, and not necessarily because of genetic susceptibility.

To conduct the GWAS, we used the data set obtained from the quality control filter descibed in chapter 2 (in section 2.2.1).

5.2.2 Association Analysis

The association testing was performed on the full data set of 888 individuals which contained related individuals. To account for both population stratification and hidden relatedness that can result from the genealogy, we applied EMMAX (Kang *et al.*, 2010), which corrects for these relationships during the association mapping. We first applied EMMAX-kin to compute a pair-wise relatedness matrix from our data set which represents the structure of our samples. EMMAX estimated the contribution of the sample structure to the TB phenotype using a variance component model, resulting in an estimated covariance matrix of phenotype that models the effect

of genetic relatedness on the TB phenotype. We ran EMMAX on TB phenotype data using the estimated covariance matrix to detect possible association. To account for rare variants that EMMAX could not address adequately, we separately performed the Fisher's Exact test, which is known to be appropriate for rare SNPs (Purcell *et al.*, 2007). To adjust our association study by gender and age, we additionally ran EMMAX with both sex and age as covariates. To report on the most significant SNP associated with TB, the p-values from the obtained GWAS dataset were assessed and given m SNPs for association with TB, we expected around $m \times 0.05$ to have p-value less than 0.05 in each data set. We thus, for genotype data, considered the genome-wide significance level at $\alpha = \frac{0.05}{2 \times m}$.

5.3 Result: Association Study in South African Coloured population

The difference in genome-wide ancestry between SAC cases and controls (Table 4.4) implies that correction for genome-wide ancestry is critical when performing a GWAS (Price *et al.*, 2010). Accordingly, we conducted a PCA analysis of the 888 SAC samples together with samples from the 5 ancestral populations (Figure 5.1). By regressing the first and second eigenvectors against case/control TB status, we obtained significant p-values $= 3.7e^{-06}$ and 0.002, respectively. As we expected, the PCA in Figure 5.1 reveals the greatest genetic differentiation between the five proxy ancestral and the SAC is in the convex hull of the three (GIH, CEU and JPT-CHB) and dispersed along a line joining African (SAN, YRI) and GIH populations. Of note, the first principal component differentiates the SAC's TB cases and controls, where most of the TB cases are pooled toward African ancestry and controls toward the non-African ancestry. This provides evidence of a significant difference in genetic ancestry between cases and controls, consistent with the result in Table 4.4, and suggesting the need to account for stratification when performing a GWAS in this population.

After quality-control filters (described in section 2.2.1), we performed the association mapping for TB using EMMAX (Kang *et al.*, 2010), the Genomic Control lambda from the obtained GWAS dataset was $\lambda_{GC} = 1.05$ (Figure 5.2).

As shown in Figure 5.3, a SNP on chromosome $14q24.2$, $rs17175227$ ($p = 8.99e-09$ and $OR = 0.141$) appears to be a genome-wide significant association signal. The SNP $rs17175227$ has a low minor allele frequency of 0.01642. We performed a well-calibrated test for rare SNPs, the Fisher's Exact test, to see if the specific SNP would still be genome-wide significant. The result suggested that $rs17175227$ was not genome-wide significant (p $= 2.77e-06$, OR $= 0.141$) (Figure 5.4). There is no tower of other linked SNPs associated with $rs17175227$ which would

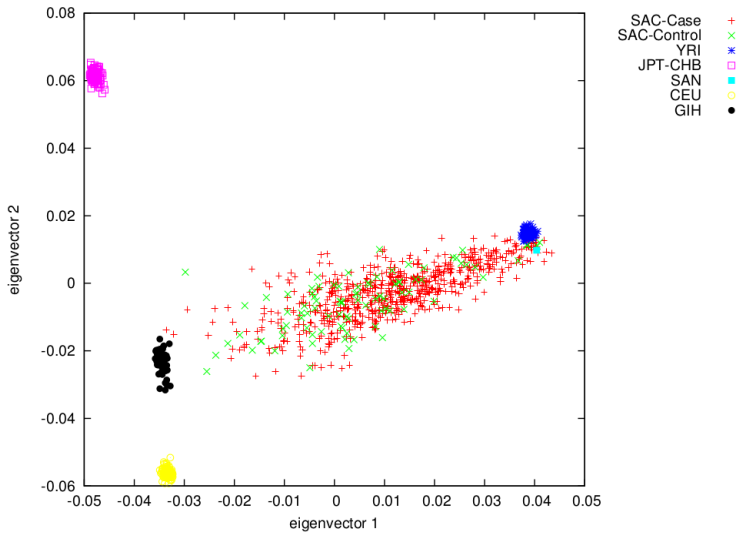

Figure 5.1: **PCA analysis of the SAC's 797 case and 91 control individuals as distinct groups within five putative ancestral populations. The first principal component differentiates the SAC's TB cases and controls, where most of the TB cases are pooled toward African ancestry and controls to non-African ancestry. The second principal component shows great genetic differentiation between the five proxy ancestral populations, and the SAC lies in their convex hull.**

be expected for true associations in GWAS. In addition, this highlights an important challenge in association analysis of low-frequency $(1-5\%)$ variants, which may often attain genome-wide significance in standard tests such as mixed model association or logistic regression due to the imperfect asymptotic distribution of those tests in the case of low-frequency variants.

Here, we have addressed this challenge by computing Fishers exact test p-values for variants that achieve the most significant mixed model association p-values.

From the GeneCard database (http://www.genecards.org/), the SNP $rs17175227$ is associated with the *SMOC1* and *SLC8A3* genes. The *SMOC1* gene is known to encode a protein that may have a crucial role in limb development and the mutations in this gene are associated with microphthalmia and limb anomalies. However, *SLC8A3* encodes a member of the sodium/calcium exchanger integral membrane protein family. Mutations in *SLC8A3* cause both progressive external ophthalmoplegia (type of eye movement disorder) and infantile onset spinocerebellar ataxia (http://www.labome.com/), and are also associated with several mitochondrial depletion syndromes, which is an autosomal inherited disease associated with grossly reduced cellular levels of mitochondrial DNA in infancy (Blake *et al.*, 1999). An additional 36 genetic markers with suggestive p-values (10^{-05} to 10^{-06}) that did not survive genome-wide significance, are listed in Tables 5.1.

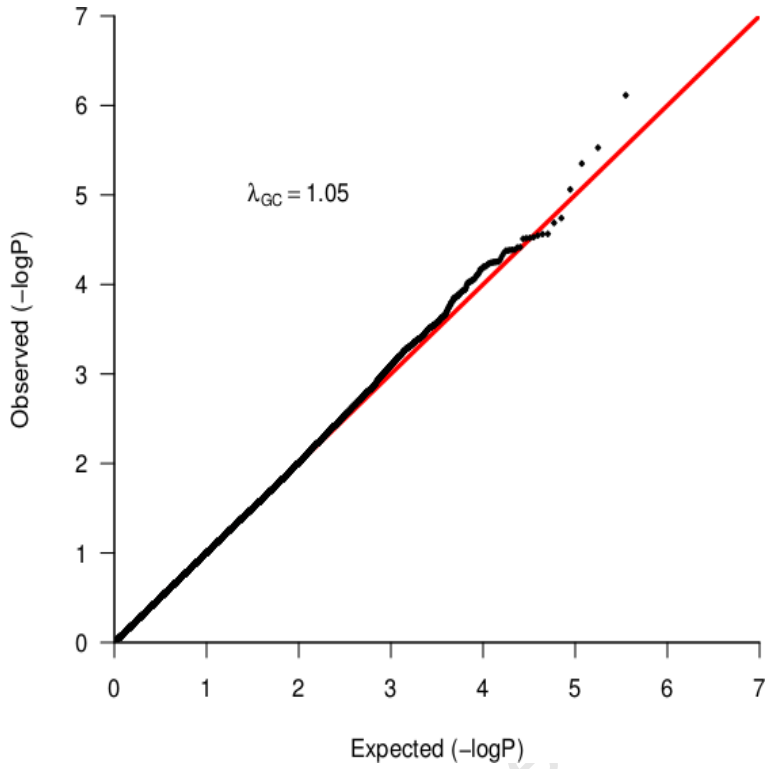

Figure 5.2: **Q-Q Plot of population stratification effects to compare the distribution of observed p-values with the expected distribution**: The lower red line shows the 90^{th} percentile, while the upper one denotes the point where the p-values diverge from the expected line. The λ_{GC} values indicate the residual population stratification effects (after correction) which are minimal.

5.4 Discussion and Conclusion

We conducted genome-wide association analysis of TB case-controls from the admixed South African Coloured population, resulting in the identification of a low-frequency variant at SNP $rs17175227$. Similar results were obtained when including age and gender as covariates in the analysis (Table 5.1). Because of the imperfect asymptotic distribution of mixed model association or logistic regression in the specific case of low-frequency variants, which may often reach genome-wide significance; we computed Fishers exact test values for variants that achieved the most significant mixed model association p-values. This resulted in $rs17175227$ not reaching the genome-wide cut-off. Power to detect association is a function of allele frequency and rare variants are underpowered when sample sizes are limited. However, because current mixed models or logistic regression association do not account for rare variants, we have addressed this challenge by computing Fishers exact test p-values for variants that achieve the most significant mixed model association p-values. Importantly, Fisher's exact test allowed us to demonstrate that a rare variant

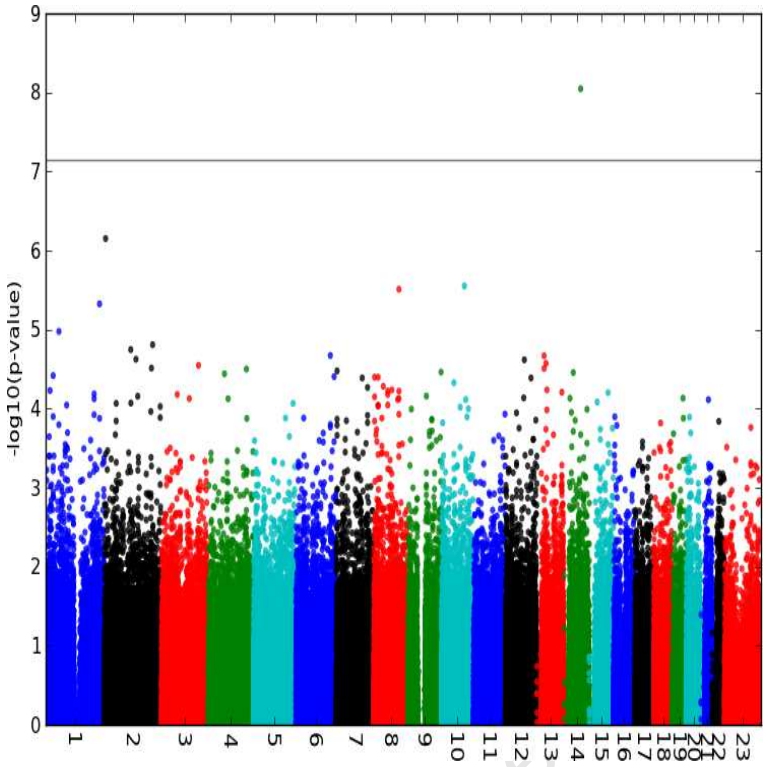

Figure 5.3: **Manhattan plot of genome-wide association analyses of TB in the South African Coloureds from typed dataset only.**

is not genome-wide significant although it achieved significant mixed model association p-values. Our study is the first typed and imputation GWAS of this complex admixed population, and it confirmed loci identified previously. Some limitations should be noted in association analyses. Firstly, the present study is underpowered to detect risk variants of more modest effect size, because of our modest sample size. Secondly, despite applying Fisher's Exact test to correct the imperfection of the mixed model for association used in our study, particularly in the case of rare variants, the implementation of newer sequencing technologies is still required to search for rare risk variants. This may potentially provide crucial insights into identifying TB susceptibility genes and, therefore, inform the development of novel interventions. In addition, our results suggest that we should conduct a genotype imputation and a meta-analysis of genome-wide association studies (see next chapter 6), by combining data from different studies, in particular, by combining our study with previously reported TB case-control studies such as in (Davila *et al.*, 2008; Thye *et al.*, 2010, 2012) in order to improve the ability to detect disease variants with small to moderate effects (see next chapter).

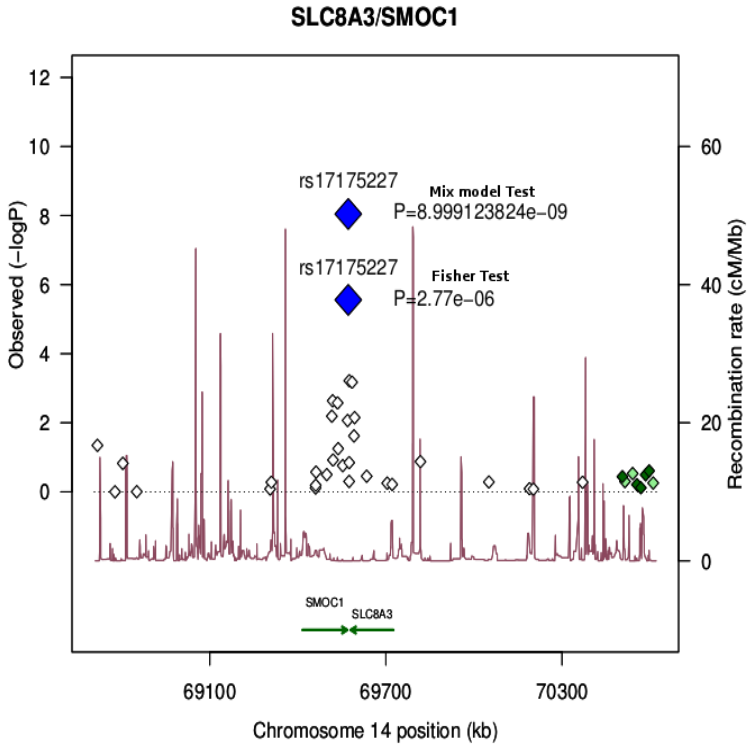

Figure 5.4: **Regional plot of SNP with the lowest p-value in TB association analysis in the South African Coloured population. Blue diamonds represent the typed-SNP with its lowest p-value from both Fisher and Mix model Test based on EMMAX. Estimated recombination rates (taken from HapMap) are plotted to show the local LD structure around the associated SNPs and their correlated proxies. White points denotes typed SNPs around** $rs17175227$ **SNP and other colour points denote imputed SNPs in the region. All genotyped SNPs in the TB genome scan are plotted with their p-values (as** $-log10$**) as a function of genomic position (with NCBIBuild** 37**).**

Table 5.1: 36 **genetic markers with moderate p-values obtained from the association analysis with the tuberculosis phenotype on the typed dataset. POS and CHR denotes chromosome, and physical position, respectively.** $A1/A2$ **are reference/derived alleles.** MAF **is minor allele frequency and CALL is genotype call rate.**

CHR	SNPs	Position	Region	A1/A2	MAF	P	P.Adj.Sex	P.Adj.Age	P.Fisher	OR	Gene
1	*rs*16861827	18550757	p36.13	C/T	0.072	$5.91e^{-05}$	$6.11e^{-05}$	$2.43e^{-05}$	0.00013	0.37	*IGSF21*
1	*rs*6694316	56197709	p32.3	G/T	0.076	$1.06e^{-05}$	$1.48e^{-05}$	$1.38e^{-05}$	$6.90e^{-06}$	0.32	*PPAP2B*
1	*rs*823122	203991651	q32.1	C/T	0.243	$7.51e^{-05}$	$4.67e^{-05}$	$6.09e^{-05}$	0.00064	0.55	*NUCKS1*
1	*rs*823123	203991969	q32.1	A/G	0.193	$6.53e^{-05}$	$3.62e^{-05}$	$3.53e^{-05}$	0.0003	0.51	*NUCKS1*
2	*rs*12328060	49824910	p16.3	C/T	0.125	$8.60e^{-05}$	$8.71e^{-05}$	$3.10e^{-05}$	0.00128	0.49	*RPL7*
2	*rs*12691834	133668510	q21.2	C/T	0.366	$2.38e^{-05}$	$2.23e^{-05}$	$5.08e^{-05}$	$1.08e^{-05}$	2.26	*NCKAP5*
2	*rs*16844441	141140892	q22.1	C/T	0.199	$6.99e^{-05}$	$5.67e^{-05}$	$4.03e^{-05}$	0.00014	0.49	*LRP1B*
2	*rs*17040773	112216506	q13	A/C	0.117	$8.53e^{-05}$	$4.77e^{-05}$	$8.59e^{-05}$	0.00019	0.44	*ANAPC1*
2	*rs*17826270	199266424	q33.1	C/T	0.351	$3.08e^{-05}$	$3.62e^{-05}$	$4.30e^{-05}$	0.00013	2.03	*PLCL1*
2	*rs*231802	204416524	q33.2	C/T	0.029	$1.55e^{-05}$	$1.19e^{-05}$	$8.58e^{-05}$	0.00129	0.29	*CTLA4*
2	*rs*724710	111624162	q13	C/T	0.198	$1.79e^{-05}$	$2.24e^{-05}$	$3.23e^{-05}$	$6.49e^{-05}$	0.48	*RGPD5*
3	*rs*816546	157630062	q25.31	C/G	0.033	$2.84e^{-05}$	$2.01e^{-05}$	$4.31e^{-05}$	0.00024	0.28	*KCNAB1*
3	*rs*880167	65770008	p14.1	A/G	0.122	$6.63e^{-05}$	$8.58e^{-05}$	$8.98e^{-05}$	$5.85e^{-06}$	0.38	*MAGI1*
4	*rs*12640159	161586073	q32.2	C/T	0.157	$3.16e^{-05}$	$3.12e^{-05}$	$4.21e^{-05}$	$1.04e^{-05}$	0.42	*FSTL5*
4	*rs*13151552	68214198	q13.2	G/T	0.033	$3.63e^{-05}$	$4.69e^{-05}$	$6.05e^{-05}$	0.00094	0.31	*UBA6*
4	*rs*17006173	83866646	q21.22	C/T	0.013	$7.51e^{-05}$	$7.23e^{-05}$	$6.63e^{-05}$	0.00383	0.22	*SCD5*
5	*rs*12658168	168290298	q35.1	A/G	0.028	$8.59e^{-05}$	$8.05e^{-05}$	$6.83e^{-05}$	0.00058	0.28	*SLIT3*

Continued on next page

Table 5.1 – continued from previous page

CHR	SNPs	Position	Region	A1/A2	MAF	P	*P.Adj.Sex*	*P.Adj.Age*	P.Fisher	OR	Gene
6	*rs*449377	145894130	q24.3	C/G	0.461	$2.13e^{-05}$	$1.55e^{-05}$	$6.66e^{-05}$	0.00016	1.91	*ZNF131*
7	*rs*17133300	3422220	p22.2	A/G	0.156	$3.34e^{-05}$	$2.51e^{-05}$	$2.18e^{-05}$	0.00026	0.47	*SDK1*
7	*rs*7783665	109826432	q31.1	A/G	0.303	$4.10e^{-05}$	$4.91e^{-05}$	$7.34e^{-05}$	0.00055	0.56	*IMMP2L*
8	*rs*1449546	76747441	q21.11	A/G	0.157	$5.81e^{-05}$	$6.48e^{-05}$	$5.77e^{-05}$	0.00047	2.72	*HNF4G*
8	*rs*16889079	40269078	p11.21	A/G	0.033	$5.23e^{-05}$	$6.27e^{-05}$	$7.11e^{-05}$	0.00098	0.31	*C8orf4*
8	*rs*1817023	106698141	q23.1	A/C	0.233	$7.47e^{-05}$	$8.62e^{-05}$	$4.02e^{-05}$	$7.55e^{-05}$	0.49	*ZFPM2*
8	*rs*895695	3232222	p23.2	A/G	0.478	$4.00e^{-05}$	$3.88e^{-05}$	$2.25e^{-05}$	0.0001306	0.53	*CSMD1*
8	*rs*895696	3232022	p23.2	A/G	0.485	$7.10e^{-05}$	$7.02e^{-05}$	$5.31e^{-05}$	0.0001354	0.53	*CSMD1*
9	*rs*11103291	138087620	q34.3	A/G	0.155	$3.46e^{-05}$	$3.76e^{-05}$	$4.92e^{-05}$	$2.65e^{-05}$	0.43	*NACC2*
9	*rs*4745272	75765361	q21.13	C/T	0.041	$6.95e^{-05}$	$8.11e^{-05}$	$4.90e^{-05}$	0.003629	0.38	*RORB*
10	*rs*2144861	51979762	q11.23	C/G	0.15	$4.71e^{-05}$	$4.55e^{-05}$	$1.23e^{-05}$	$8.15e-06$	0.4	*SGMS1*
12	*rs*1245016	79097100	q21.31	A/G	0.229	$2.42e^{-05}$	$1.94e^{-05}$	$6.93e^{-05}$	$2.66e^{-05}$	0.46	*RPL7*
12	*rs*41489249	107314707	q23.3	C/T	0.04	$4.10e^{-05}$	$4.47e^{-05}$	$2.14e^{-05}$	0.001292	0.34	*CMKLR1*
13	*rs*17503526	29415041	q12.3	G/T	0.023	$2.15e^{-05}$	$2.24e^{-05}$	$2.74e^{-05}$	0.001253	0.27	*UBL3*
13	*rs*17587770	29407009	q12.3	A/G	0.024	$3.11e^{-05}$	$3.09e^{-05}$	$4.79e^{-05}$	0.001235	0.27	*UBL3*
13	*rs*683479	37651676	q13.3	C/T	0.218	$2.68e^{-05}$	$3.53e^{-05}$	$6.61e^{-05}$	$8.20e-06$	0.45	*LINC00571*
14	*rs*854406	24274191	q12	C/T	0.096	$7.40e^{-05}$	$7.28e^{-05}$	$5.55e^{-05}$	0.00323	0.49	*STXBP6*
19	*rs*16979659	50286498	q13.32	C/G	0.028	$7.38e^{-05}$	$8.32e^{-05}$	$1.38e^{-05}$	0.002265	0.32	*GEMIN7*
21	*rs*2832542	30327668	q21.3	A/G	0.024	$7.72e^{-05}$	$7.91e^{-05}$	$8.76e^{-05}$	0.0004053	0.2	*GRIK1*

Chapter 6

Genome-wide Imputation for TB Risk in the Admixed South African Coloured Population and Comparison with Previous TB Studies.

6.1 Introduction

Imputation is a useful tool in genome-wide association studies (GWAS), and often used in the meta-analysis of GWAS, for combining data from different studies, in order to improve the ability for detecting disease variants with small to moderate effects (Li *et al.*, 2012). Since most of the susceptibility loci that remain undiscovered are believed to have small effects (Ferreira *et al.*, 2008; Han & Eskin, 2011; Li *et al.*, 2012), large sample sizes are usually required to achieve sufficient statistical detection powers. However, such as sample size requirement can be beyond the capacity of a single GWA study. Meta-analysis has been suggested to be an alternative solution to this matter. This approach combines standard GWAS data sets from multiple studies of relatively small sample sizes, in order to detect genes underlying susceptibility loci with greater power, and has shown to produce more precise estimation of genetic effects and more convincing conclusions than each individual study does (Han & Eskin, 2011; Li *et al.*, 2012). Furthermore, Meta-analysis has been applied to and improved the understanding of a number of complex traits, including type 2 diabetes (Sanghera *et al.*, 2009; Staiger *et al.*, 2008), bipolar disorder (Ferreira *et al.*, 2008) and Parkinson's disease (Evangelou *et al.*, 2008), demonstrating the usefulness of meta-analysis of GWAS.

To achieve sufficient power in our limited sample size of 888 SAC samples in detecting associations at a level of genome-wide significance and identifying shared risk loci with previously

reported TB case-control studies, this chapter covers GWAS imputation and meta-analysis of our study and previously reported TB studies, including Thye *et al.* (2010), Thye *et al.* (2012) and Davila *et al.* (2008).

6.2 Materials and Methods

6.2.1 Quality Control and Imputation Procedures

To account for the population structure in the admixed SAC in imputing the untyped genotypes, we consider the imputation model based on population genetic parameters in the coalescent framework implemented in IMPUTE2 (Marchini & Howie, 2008). Exploring the advantage of the model in IMPUTE2, we combined all available reference phased haplotype data from both release 2 of the HapMap 3 dataset (NCBI Build 36, includes: Utah residents (CEPH) with Northern and Western European ancestry (CEU), Japanese in Toyko (JPT), Chinese in Denver (CHD), Maasai in Kinyawa (MKK), Toscani in Italia (TSI), Gujarati Indian in Houston (GIH), African Ancestry in Southwest (ASW), Luhya in Webuye (LWK), Mexican Ancestry in Los Angeles (MEX), Han Chinese in Beijing (CHB) and Yoruba in Ibadan(YRI)) and 1000 Genomes Project (includes CEU, YRI, British from England and Scotland (GBR), Finnish from Finland (FIN), Han Chinese South (CHS), Puerto Rican (PUR), Chinese in Denver (CHD), JPT, LWK, Mexican Ancestry in Los Angeles (MXL), ASW, TSI, Colombian in Medellin (CLM) and Iberian populations in Spain (IBS)). We decided to impute SNPs by splitting each chromosome into 5 Mb regions for analysis by IMPUTE2. For resulting imputed datasets, post-imputation quality controls were similarly conducted as described in section 2.2.1 in order to account for imputation uncertainty.

6.2.2 Association and Meta Analyses

The association testing was performed on the two obtained imputed data sets (section 6.2.1) of the SAC using EMMAX software as in section 5.2.2. To identify associations with small effect sizes which the standard single GWAS could not identify, we combined two African TB genome-wide association studies including our GWAS and the recently combined TB study of Ghanaian, Gambian and Malawian populations in a single GWAS analysis. A random effects model (Han & Eskin, 2011) based on inverse-variance-weighted effect size was used to combine the results (log-odds ratio and standard error) from typed GWAS (obtained from section 5.2.2) and two imputation GWAS. The imputation was separately based on the data from both HapMap 3 and the 1000 Genomes Project, including the non-pseudoautosomal region (nonPAR) and two pseudoautosomal regions (PAR1 and PAR2) of X chromosome. We additionally applied random and binary effects models described in the MetaSoft program (Han & Eskin, 2011) and we used

the study p-values, the M-values (the posterior probability that the effect exists in the study), the mean effect and I-square heterogeneity statistics to interpret the association results showing high heterogeneity (Han & Eskin, 2011).

6.3 Results: Imputation Association Study in South African Coloured Population

Using IMPUTE2 (Marchini & Howie, 2008), we imputed the SAC's untyped genotypes using both HapMap3 release 2 and 1000 Genomes project populations. After post-imputation quality control on the genome-wide imputation, there were $1,453,294$ and $4,467,279$ genetic variants retained from each imputation panel, respectively. To account for both population stratification and hidden relatedness, we applied the mixed model approach from EMMAX (Kang *et al.*, 2010) to these data sets, the Quantile-Quantile (QQ) plots are shown in Figure 6.1. The Genomic Control lambda from the imputed dataset based on HapMap3 $\lambda_{GC} = 1.05$ and from the imputed dataset from 1000 Genomes $\lambda_{GC} = 1.09$, and from the combined GWAS datasets (typed and two imputed GWAS) $\lambda_{GC} = 1.08$.

As shown in Figure 6.2, the imputed SNP $rs12294076$ ($p = 9.56e^{08}$) on chromosome $11q21 - q22.1$ narrowly misses the threshold of genome-wide significance, which we define as $1.7e^{-08}$ and $5.5e^{-09}$ based on $1,453,294$ and $4,467,279$ SNPs tested (Figure 6.2) from imputed data using both HapMap3 and 1000 genome data, respectively (see section 6.2.2). The genetic variant $rs12294076$ has a minor allele frequency of 0.16 in the SAC, 0.22 in Yoruba and 0.0 in other HapMap populations, and is likely to be an Africa-specific SNP.

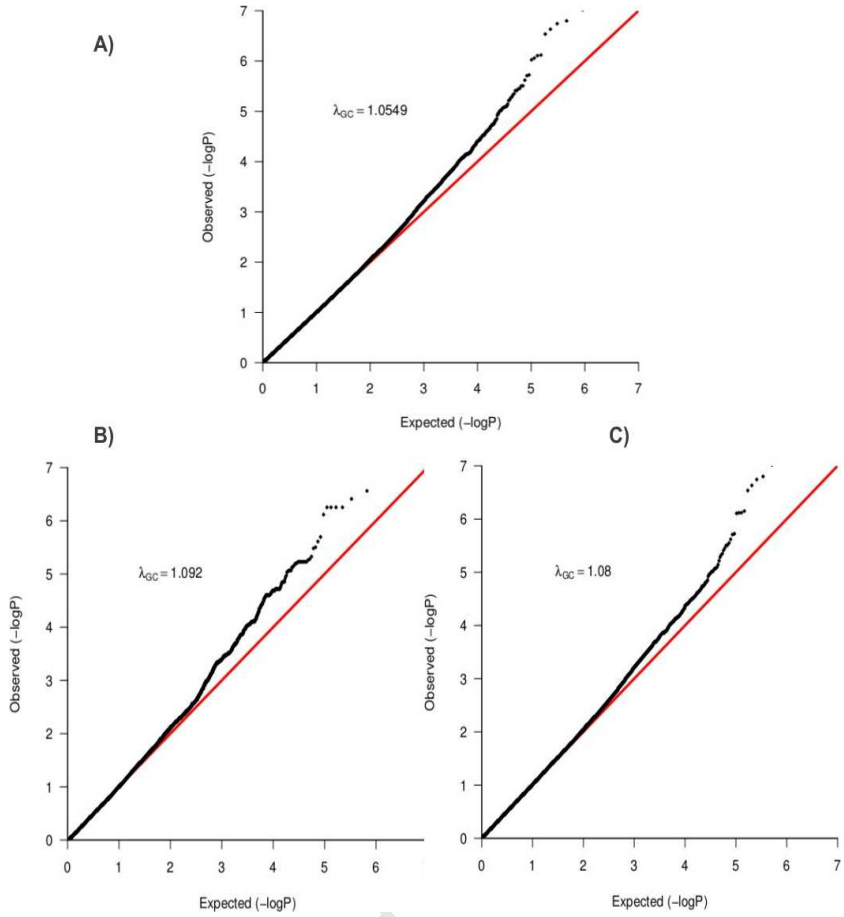

Figure 6.1: **Q-Q Plot of population stratification effects to compare the distribution of observed p-values with the expected distribution**: The lower red line shows the 90^{th} percentile, while the upper one denotes the point where the p-values diverge from the expected line. The λ_{GC} values indicate the residual population stratification effects (after correction) which are minimal. The Q-Q plot obtained from GWAS using imputed genotype from HapMap3 (A), imputed genotype from the 1000 genomes project (B) and the combined GWAS datasets (typed and two imputed GWAS) (C).

The SNP $rs12294076$ is associated with the *DYNC2H1* gene. This gene encodes a large cytoplasmic dynein protein known to be involved in retrograde transport in the cilium with a major role in intraflagellar transport (Hokayem *et al.*, 2012). Mutations in *DYNC2H1* cause a heterogeneous spectrum of conditions related to altered primary cilium function. The sub-cellular distribution of dynein shows specific association with elements of the late endocytic pathway (Hokayem *et al.*, 2012). An additional genetic markers with suggestive p-values (10^{-05} to 10^{-06}) that did not survive genome-wide significance, are listed in Table 6.3 for the imputed datasets.

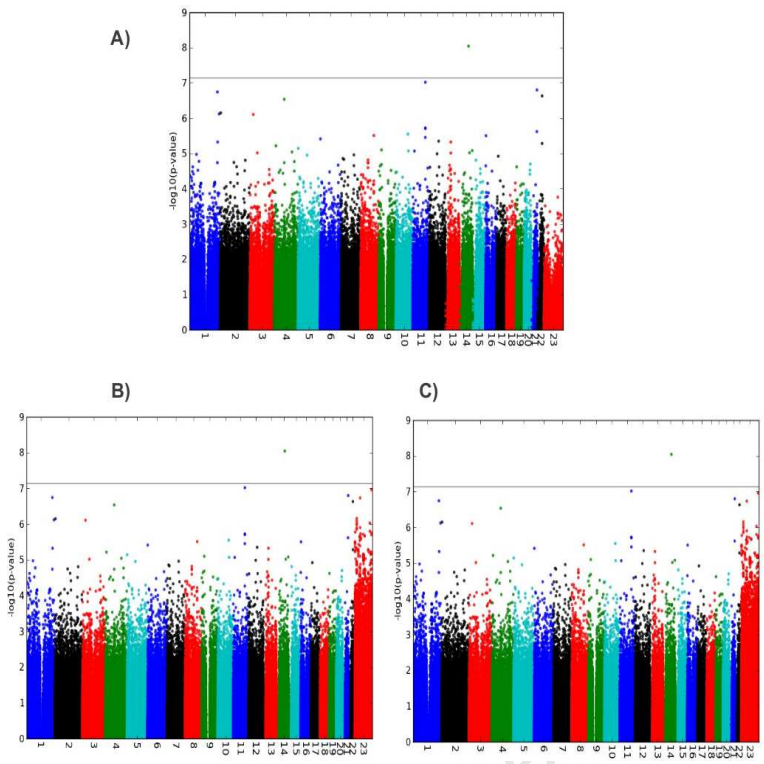

Figure 6.2: **Manhattan plot of genome-wide association analyses of TB in the South African Coloureds from imputed dataset based on HapMap3 (A), from imputed dataset based on** 1000 **Genomes project populations (B) and from the combined datasets (typed and two imputed GWAS) (C). The horizontal line indicates significance cut-off.**

6.3.1 Replication of SNPs Reported in Previous Studies

Comparing our TB GWAS to a recently combined study of African TB case-control series from Ghana, Gambia, Indonesia and Russia in Thye *et al.* (2012), we found that the associated SNP, $rs2057178$ (p $= 2.63e^{-09}$, OR $= 0.77$ and MAF $= 0.33$) on 11p13 reported in (Thye *et al.*, 2012), is on the boundary of genome-wide significance (p $= 2.71e^{-06}$, OR $= 0.62$ and MAF $= 0.08$) in the SAC-TB imputation GWAS (Table 6.1). A second reported significant SNP in the Ghanaian study group, $rs11031728$ (p $= 5.25e-09$, MAF $= 0.32$ and OR $= 0.77$), yielded a moderate association in our imputation GWAS study (p $= 2.86e-06$, MAF $= 0.08$ and OR $= 0.61$). The third most significant SNP in their study was $rs11031731$ (p $= 7.01e^{-09}$, MAF $= 0.31$ and OR $= 0.78$), which was poorly imputed in our study (CALL $= 0.70$), therefore did not provide convincing association evidence. The $rs2057178$, $rs11031728$ and $rs11031731$ SNPs are not covered in GIH and SAN data, therefore accounting for the linkage disequilibrium in the admixed SAC, we computed the $r2$ LD between these three SNPs and other SNPs in the *WT1* locus using the SAC data, YRI, CEU and JPT+CHB data from the 1000 Genomes

project.*WT1* is a tumor suppressor gene located on chromosome 11p13. *WT1* is known as Wilms's Tumor Protein, which provides instructions for making a protein that is involved in the development of the kidneys and gonads (ovaries in females and testes in males) before birth (Sum *et al.*, 2002). Furthermore, it is also known as a transcription factor, since it regulates the activity of other genes by binding to specific regions of DNA. Querying a comprehensive human Protein-Protein Interaction (PPI) network (http://cbg.garvan.unsw.edu.au/pina/), *WT1* has known direct interactions (Sum *et al.*, 2002) with *UBE2I*, *AREG*, *WTAP*, *AREGB*, *U2AF2*, *TP73*, *SDGF*, *PRKACA* and *P53* genes (Figure 6.3 shows the related sub-network). In particular, this gene is unusually expressed in certain types of lung and prostate cancer, and is seen in some cancers of blood-forming cells (leukemias), such as acute lymphoblastic leukemia, chronic myeloid leukemia, and childhood acute myeloid leukemia (Sum *et al.*, 2002).

Previous results in (Thye *et al.*, 2012), reported that $rs2057178$, $rs11031728$ and $rs11031731$ SNPs are in strong LD in the Ghanaian data. We obtained $r2(rs2057178,\ rs11031728) = 0.90,\ 0.90,\ 1$ and 0.8; $r2(rs2057178,\ rs11031731) = 0.70$,$0.90,\ 1$ and 1; and $r2(rs11031728,\ rs11031731) = 0.70,\ 1,\ 1$ and 0.90 in SAC, CEU, YRI and JPT+CHB, respectively. The SNPs $rs2057178$, $rs11031728$ and $rs11031731$ are associated with *WT1*.

Figure 6.3: **Biological network of genes interacting with WT1 (11p13), TLR8 (Xp22.2) and RBBP8 (18q11.2). The interactions were obtained from the comprehensive human PPI network downloaded from the Protein Interaction Network Analysis platform (PINA) (Wu et al., 2009). The plot shows that the sub-networks of interactions with WT1, TLR8 and RBBP8 do not overlap each other, consistent with the fact that the SNPs in each of these loci (WT1, TLR8 and RBBP8) were not in LD.**

The identified susceptibility locus $rs4331426$ on chromosome $18q11.2$ in Thye *et al.* (2010) (MAF= 0.48, Gambia: p $= 0.003$ and OR $= 1.18$, Ghana: p $=0.004$ and OR $= 1.19$ and Combined data: p $= 6.8e^{-09}$ and OR $= 1.19$), for TB in the study of combined Gambia and Ghana populations (Thye *et al.*, 2010), did not yield any convincing evidence of association with TB in our study samples (Table 6.1). In our study, we obtained a p $= 0.83$, MAF $= 0.19$ and OR $= 1.00$, and no suggestive signals in the SAC data located near the variant. Similarly to the above, we computed $r2$ LD in the region of 18q11.2 in the data of the SAC, CEU, YRI, JPT+CHB, GIH and SAN. Four SNPs, including $rs4264496$, $rs4331426$, $rs4239431$ and $rs4239432$ in the entire region of 18q11.2 have $r2 >= 0.5$, but all have weak p-values from the association study with TB in the SAC data. In addition, the $rs4331426$ SNP is not in LD with any SNPs in the *WT1* locus in the data of the SAC, CEU, YRI and JPT+CHB. $rs4331426$ is associated with *RBBP8* gene. This gene is known to interact with *LMO4*, Retinoblastoma-like protein 2, Retinoblastoma-like protein 1, Ataxia telangiectasia mutated, Retinoblastoma protein, *CTBP1*, SIAH1 and *BRCA1* (Rauscher, 1993). Figure 6.3 shows no overlap between the *WT1* and *RBBP8* sub-networks. The susceptibility locus of $rs4331426$ discovered in the African populations (Ghana, Gambia and Malawi) in Thye *et al.* (2010) could not be validated in the SAC population, and recently it could not be validated in the Chinese population either (Dai *et al.*, 2011).

To compare our study to previous findings of association with TB susceptibility at four polymorphisms in the $TLR8$ gene on X chromosome from Davila *et al.* (2008), we conducted an additional imputation GWAS on the non-pseudoautosomal region (nonPAR) and two pseudoautosomal regions (PAR1 and PAR2) of the X chromosome in the SAC. The results displayed in Table 6.1 compare our results and those from Davila *et al.* (2008). Our imputation GWAS suggests a weak association with TB of these four polymorphisms in the *TLR8* gene on the X chromosome, which include $rs3764880$, $rs3764879$, $rs3761624$ and $rs3788935$ (Table 6.1). These four SNPs are in LD with each other ($r2 >= 0.5$) in the data of the SAC. The *TLR* gene plays a fundamental role in pathogen recognition, activation of innate immunity and is predominantly expressed in lung and peripheral blood leukocytes (Peng *et al.*, 2011). However, these four SNPs do not yield any convincing evidence of association in the SAC, and thus, could not be validated in this admixed population.

We additionally examined whether the genes interacting with *WT1*, *TLR8* and *RBBP8*, form a network of sub-networks that overlap each other. We used a comprehensive human PPI network downloaded from the Protein Interaction Network Analysis platform (PINA) (Wu *et al.*, 2009) which collected and annotated data from six public PPI databases (MINT, IntAct, DIP, BioGRID, HPRD, and MIPS/MPact), queried these interactions with respect to *WT1, TLR8 and RBBP8*, and plot their combined interactive sub-networks. The plot in Figure 6.3 shows that sub-networks of genes interacting with *WT1, TLR8 and RBBP8* are disconnected and do not overlap each

other, this is consistent with the fact that no SNPs between loci *WT1, TLR8* and *RBBP8* were found to be in LD ($r2 > 0.5$) with each other in the SAC, CEU, YRI, JPT+CHB, GIH and SAN populations.

Table 6.1: **Investigating replication of SNPs reported in previous studies.**

				SAC TB Study			Thye et al. 2012		
SNP	**CHR**	**POS**	**A1/A2**	**MAF**	**P-value**	**OR(95%CI)**	**MAF**	**P-value**	**OR(95%CI)**
*rs*2057178	11	32364187	G/A	0.08	$2.70e-07$	$0.62(0.50-0.75)$	0.33	$2.63e-09$	$0.77(0.71-0.84)$
*rs*11031728	11	32363616	C/G	0.08	$2.86e-06$	$0.61(0.50-0.75)$	0.32	$7.01e-09$	$0.78(0.71-0.8)$
								Thye et al. 2010	
*rs*4331426	18	196761760	G/A	0.19	0.83	$1.00(0.95-1.04)$	0.48	$6.8e-09$	$1.19(1.1-1.3)$
								Davila et al. 2008	
*rs*3788935	X	12922659	A/C	0.386	0.1465	$1.30(0.91-1.85)$	-	0.014	$1.4(1.07-1.8)$
*rs*3761624	X	12923681	A/C	0.382	0.1844	$1.27(0.89-1.81)$	-	0.016	$1.4(1.06-1.8)$
*rs*3764879	X	12924697	A/C	0.386	0.2854	$1.23(0.87-1.80)$	-	0.01	$1.4(1.06-1.8)$
*rs*3764880	X	12924826	A/C	0.383	0.2278	$1.25(0.95-0.99)$	-	0.016	$1.4(1.006-1.8)$

Table 6.2: **Meta-analysis of two TB case-control studies, SAC-TB, WTCCC-TB and 4 polymorphisms on chromosome X previously identified by Davila et al. 2008.** $p.RAN$ **is the p-value of fixed effect,** $p.BE$ **is the p-value of binary-effect, ST1 and ST2 are the statistic mean effect and heterogeneity, respectively. Mvalue is the posterior probability that the effect exists in each study.**

SAC-TB + WTCCC-TB						SAC TB Study		Thye et al. 2012	
SNP	**CHR**	$p.RAN$	$p.BE$	**ST1**	**ST2**	**p-value**	**Mvalue**	**p-value**	**Mvalue**
*rs*2057178	11	$3.26e^{-3}$	$9.83e^{-13}$	53.05	2.91	$2.75e^{-06}$	1.0	$2.52e^{-09}$	1.0
*rs*11031728	11	$4.73e^{-07}$	$4.08e^{-10}$	41.19	0.0	$2.98e^{-06}$	0.98	$7.03859e^{-09}$	1.0
								They et al. 2010	
*rs*4331426	18	0.28	$1.90e^{-08}$	1.15	32.6	0.002	0.0	$6.83e^{-09}$	1.0
								Davila et al. 2008	
*rs*3788935	*X*	0.00457	0.012	8.039	0.0	0.15	0.78	0.014	0.778
*rs*3761624	*X*	0.0066	0.014	7.382	0.0	0.18	0.74	0.016	0.743
*rs*3764879	*X*	0.0063	0.014	7.469	0.0	0.28	0.70	0.01	0.886
*rs*3764880	*X*	0.0080	0.018	7.018	0.0	0.23	0.72	0.016	0.858

6.3.2 Meta-analysis with SAC and WTCCC Data

Identifying common variants of modest and weak effect is still a challenge, and large sample size has been suggested in order to increase the power. The sample sizes of both TB cases and controls in this study do not provide sufficient power to obtain associations at a stringent level of statistical significance. However, one of the proposed solutions to this problem is to combine analyses of several clinically close phenotypes from different studies (Bhattacharjee *et al.*, 2012; Han & Eskin, 2011). To increase the power to detect common variants, we did a meta-analysis by combining our study with the previously published GWAS from WTCCC-TB (Thye *et al.*, 2010, 2012) and four polymorphisms in the TLR8 gene on chromosome X which was previously identified by Davila et al. 2008 (Davila *et al.*, 2008). To address, this, we first independently combined the results (odds ratio and its standard error) from typed GWAS and two imputation GWAS (imputation based on both the data from HapMap 3 and 1000 Genomes Project, including the non-pseudoautosomal region (nonPAR) and two pseudoautosomal regions (PAR1 and PAR2) of X chromosome) from the SAC and WTCCC-TB data. The results obtained from both typed GWAS and imputation GWAS based on the WTCCC-TB data are not shown, to avoid replication of the results from WTCCC-TB (Thye *et al.*, 2010, 2012). Merging the two resulting GWAS data sets, a total of $1,009,364$ autosomal SNPs were meta-analyzed across the two studies. We applied the random and binary-effects methods implemented in MetaSoft program (Han & Eskin, 2011) to the combined studies and report results meta-analyses (Table 6.2). We obtained reasonable inflation rates from the fixed-effect ($\lambda_{GC} = 1.062$), binary-effect ($\lambda_{GC} = 1.05$) and from each individual study, SAC-TB ($\lambda_{GC} = 1.094$) and WTCCC-TB ($\lambda_{GC} = 1.0495$), respectively (Figure 6.4).

In addition to standard p-values we also examined the posterior probability (m-value) that the effect exists in each study (Han & Eskin, 2011). Using a threshold m-value > 0.7, we observed two genetic variants, $rs2057178$ and $rs11031728$ (Figure 6.5 and Table 6.2) with similar p-values to the standard GWAS, that resulted in a significant association with risk of TB and had effects in our study and the Thye et al. study (Thye *et al.*, 2012). These SNPs are both on chromosome region $11p13$ and replicate the recent findings of (Thye *et al.*, 2012) in our imputation GWAS. Other variants (Figure 6.5 and Table 6.2) yielded a weak effect in the SAC-TB study.

Although Metasoft provides a slightly different p-value (Table 6.2) at SNP $rs4331426$ than that from the standard GWAS (Table 6.1), which may due to high heterogeneity (ST2 $= 32.6$, see Table 6.2), this susceptibility locus reported in Thye *et al.* (2010) does not survive genome-wide significance in the TB meta-analysis of the SAC and WTCCC (Thye *et al.*, 2010). Moreover, the TB SAC TB meta-analysis of the SAC and four polymorphisms in the *TLR8* gene on the X chromosome reported in an Indonesia population from Davila *et al.* (2008) studies did not yield

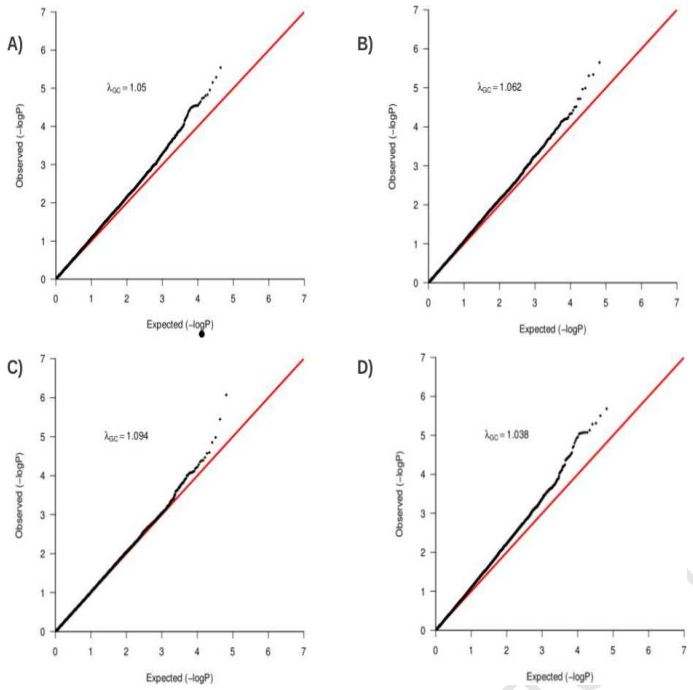

Figure 6.4: **Meta analyses Q-Q Plot of genomic control factors effects: The lower red line shows the 90^{th} percentile, while the upper one denotes the point where the p-values diverge from the expected line. The λ_{GC} values indicate the residual population stratification effects (after correction), which are minimal. The plots are from the fixed-effect (A), binary-effect models (B), the SAC-TB study (C) and WTCCC-TB study (D), respectively.**

any convincing evidence of association with risk of TB. This suggests no replication observed at the *TLR8* locus in the admixed SAC.

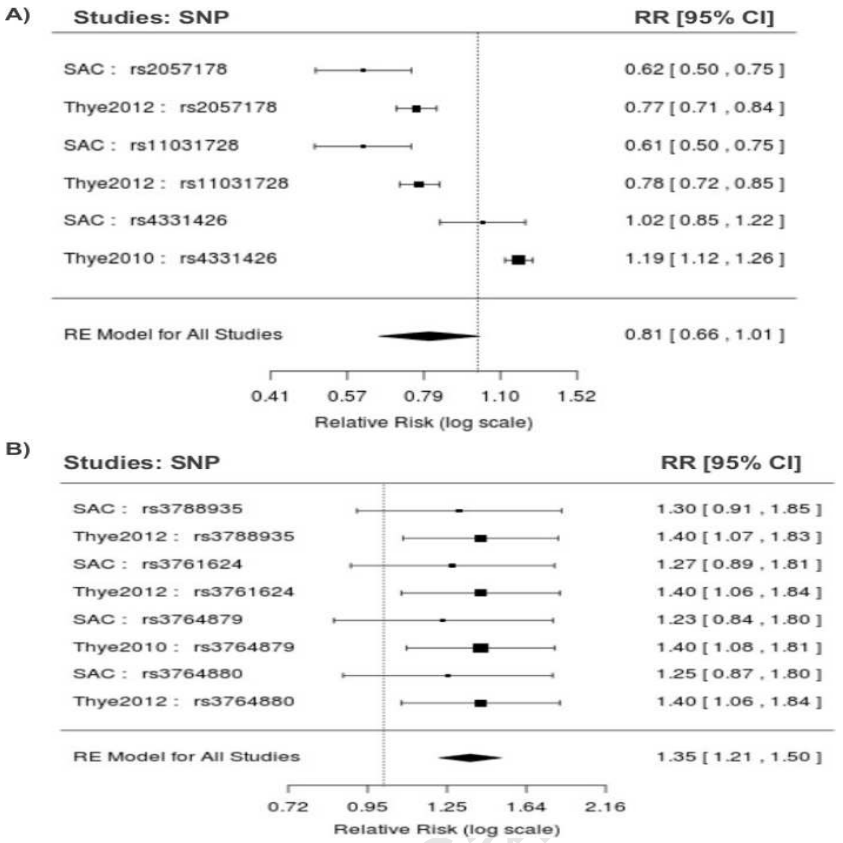

Figure 6.5: **(A) Forest plot of relative risk from genome-wide meta-analysis of TB in the South African Coloured and WTCCC-TB studies based on findings in (Thye et al., 2010, 2012). (B) Plot of relative risk from binary and random effect model from both the South African Coloured and four polymorphisms in the TLR8 gene on the X chromosome reported in an Indonesia population from Davila et al. (2008) studies.**

6.4 Discussion and Conclusion

To achieve sufficient power to detect associations at a level of genome-wide significance and identify shared risk loci with a previously reported African TB case-control study (Thye *et al.*, 2010, 2012) and four polymorphisms in the *TLR8* gene on chromosome X previously identified by Davila et al. 2008, the GWAS meta-analysis was performed under fixed-effect and binary-effect models. In combining GWAS data across these studies, two loci ($rs2057178$ and $rs11031728$) had an association result with genome-wide significance, and showed strong effect in both our study and the previous study (Thye *et al.*, 2010, 2012).

Our study is the first imputation GWAS of this complex admixed population, as well as the meta-analysis with a previous GWAS on African populations, which confirmed loci identified previously. A major limitation in this study is that, imputing missing genotype data of a complex

admixed population is still an important challenge based on the choice and size of haplotype of existing reference panels. In particular, the imputation of missing genotype data of this complex admixed SAC population was suboptimal, suggesting a challenge in imputation of missing genotypes of a such multi-way admixed population as is the case for inferring the locus-specific ancestry along the genome of such a population (Baran *et al.*, 2012; Marchini & Howie, 2008; Rodriguez *et al.*, 2012). Nonetheless, the increased number of SNPs generated by imputation analyses was useful in this study, yielding the replication of TB susceptibility loci (Thye *et al.*, 2012).

Table 6.3: 62 **genetic markers with moderated p-values obtained from the association analysis with the tuberculosis phenotype on an imputed dataset. POS and CHR denotes chromosome, and physical position, respectively. A1/A2 are reference/derived alleles. MAF is minor allele frequency and CALL is genotype call rate.**

SNP	CHR	POS	A1/A2	Call	INFO	MAF	P	Fisher	OR	Gene
*rs*10917420	1	23935574	C/T	0.89	0.71	0.417	$3.25e^{-05}$	$2.15e^{-06}$	0.27	*TCEB3*
*rs*16851354	1	15368207	C/T	0.76	0.52	0.268	$4.84e^{-05}$	0.0005	0.29	*TMEM51*
*rs*17739539	1	216121310	C/T	0.97	0.86	0.083	0.00027	0.0002	0.36	*LINC00210*
*rs*1926278	1	68226447	C/T	0.96	0.93	0.356	$2.99e^{-05}$	$2.78e^{-07}$	0.41	*GNG12-AS1*
*rs*2182200	1	226782789	A/C	0.99	0.97	0.115	$1.82e^{-05}$	$1.06e^{-06}$	0.37	*RHOU*
*rs*315087	1	76761852	C/T	0.75	0.47	0.248	$1.67e^{-05}$	$1.39e^{-05}$	0.13	*ST6GALNAC3*
*rs*7541416	1	26495498	A/G	0.82	0.73	0.492	$2.40e^{-05}$	$6.90e^{-06}$	4.41	*UBXN11*
*rs*1032044	3	158190024	A/C	0.93	0.88	0.38	$3.99e^{-05}$	$8.60e^{-08}$	0.35	*LEKR1*
*rs*1385715	3	59716555	A/C	0.9	0.85	0.496	$9.65e^{-06}$	$1.78e^{-06}$	2.84	*FHIT*
*rs*1595665	4	161630317	C/T	0.99	0.97	0.18	$1.69e^{-05}$	$1.91e^{-06}$	4.94	*FSTL5*
*rs*17493657	4	35753287	C/T	0.59	0.3	0.442	$3.44e^{-05}$	0.0014	0.04	*ARAP2*
*rs*17653240	4	46972055	C/T	0.84	0.66	0.234	$2.95e^{-05}$	$9.81e^{-05}$	0.28	*GABRB1*
*rs*16898876	5	13263200	C/T	0.91	0.86	0.463	$4.69e^{-05}$	$1.06e^{-05}$	0.42	*RPS23P5*
*rs*240727	6	75900080	A/G	0.82	0.48	0.162	$3.33e^{-05}$	$7.09e^{-05}$	0.16	*COL12A1*
*rs*2505675	6	2300674	C/T	0.87	0.61	0.154	$3.87e^{-06}$	$1.12e^{-06}$	0.22	*LOC100508120*
*rs*2286182	7	26590917	A/C	0.95	0.84	0.145	$3.36e^{-05}$	$5.08e^{-06}$	0.35	*KIAA0087*
*rs*2576507	7	54586437	A/G	0.77	0.62	0.413	$4.68e^{-05}$	0.0007	0.34	*VSTM2A*

Continued on next page

Table 6.3 – continued from previous page

SNP	CHR	POS	A1/A2	Call	INFO	MAF	P	Fisher	OR	Gene
*rs*8764215	7	103371937	C/T	0.68	0.34	0.267	$2.22e^{-05}$	0.0003	0.08	-
*rs*9639391	7	21737815	G/T	0.83	0.72	0.405	$1.47e^{-05}$	$6.93e^{-05}$	0.35	*DNAH11*
*rs*4738654	8	59315323	A/G	0.9	0.59	0.103	$1.52e^{-05}$	$8.82e^{-06}$	0.27	*FAM110B*
*rs*6995423	8	59330138	A/G	0.85	0.53	0.134	$2.53e^{-05}$	$7.99e^{-06}$	0.24	*FAM110B*
*rs*10809117	9	10531177	G/T	0.79	0.64	0.335	$2.27e^{-05}$	$7.41e^{-06}$	0.27	*PTPRD*
*rs*10816229	9	9902827	A/C	0.82	0.67	0.31	$4.24e^{-05}$	0.000764	0.4	*PTPRD*
*rs*1410978	9	22394681	C/T	0.93	0.89	0.402	$2.06e^{-05}$	$6.29e^{-06}$	0.43	*DMRTA1*
*rs*586716	9	22478678	A/G	0.83	0.51	0.154	$7.93e^{-06}$	$4.17e^{-05}$	0.21	*DMRTA1*
*rs*7901781	10	5109544	C/T	0.89	0.55	0.121	$3.35e^{-05}$	$9.44e^{-06}$	0.2	*AKR1C3*
*rs*12283022	11	102485804	A/G	0.76	0.48	0.245	$1.88e^{-06}$	$8.51e^{-07}$	0.14	-
*rs*1819084	11	13952731	A/C	0.75	0.54	0.288	$8.53e^{-06}$	$2.90e^{-07}$	0.16	*SPON1*
*rs*7104341	11	122086148	G/T	0.84	0.62	0.194	$2.60e^{-05}$	$4.49e^{-05}$	0.26	*UBASH3B*
*rs*7105967	11	102434653	C/T	0.75	0.47	0.254	$3.51e^{-06}$	$1.89e^{-06}$	0.15	*DCUN1D5*
*rs*7947821	11	102452675	C/T	0.75	0.47	0.252	$1.95e^{-06}$	$9.92e^{-07}$	0.14	*DCUN1D5*
*rs*12426185	12	5579896	C/G	0.89	0.78	0.25	$2.45e^{-05}$	$1.01e^{-05}$	0.38	*ANO2*
*rs*6538140	12	76262136	A/G	0.81	0.63	0.259	$4.46e^{-06}$	$1.83e^{-06}$	0.23	*E2F7*
*rs*1886235	13	73233391	A/C	0.94	0.81	0.129	$3.69e^{-05}$	$5.91e^{-05}$	0.37	*KLF12*
*rs*1900442	13	41403674	C/T	0.97	0.91	0.146	$4.72e^{-06}$	$3.68e^{-07}$	0.37	*VWA8*
*rs*28493371	13	41387501	C/T	0.95	0.86	0.201	$3.92e^{-05}$	$4.66e^{-06}$	0.39	*KIAA0564*
*rs*7318112	13	41423876	C/T	0.96	0.86	0.151	$2.87e^{-05}$	$1.80e^{-06}$	0.37	*VWA8*
*rs*7318638	13	41399654	C/T	0.97	0.91	0.145	$9.73e^{-06}$	$5.47e^{-07}$	0.37	*VWA8*
*rs*11844457	14	86255183	A/C	0.93	0.83	0.198	$4.96e^{-05}$	0.00059	0.48	-

Continued on next page

Table 6.3 – continued from previous page

SNP	CHR	POS	A1/A2	Call	INFO	MAF	P	Fisher	OR	Gene
*rs*1948724	14	32907418	G/T	0.86	0.63	0.176	$3.95e^{-05}$	0.00011	0.34	*NPAS3*
*rs*6575836	14	100749008	A/G	0.82	0.58	0.211	$8.30e^{-06}$	$7.35e^{-05}$	0.25	*SNORD114-31*
*rs*7163165	15	59541650	G/T	0.85	0.75	0.384	$2.73e^{-05}$	$1.26e^{-06}$	0.34	-
*rs*7171652	15	59497604	C/T	0.89	0.82	0.381	$3.73e^{-05}$	$3.83e^{-06}$	0.4	*RORA*
*rs*1074182	16	52028858	G/T	0.92	0.8	0.262	$3.17e^{-05}$	$3.61e^{-05}$	0.34	*RBL2*
*rs*40363	16	3449057	A/G	0.76	0.51	0.275	$3.13e^{-06}$	$1.60e^{-07}$	0.09	*NAA60*
*rs*582998	20	47827998	C/T	0.86	0.66	0.196	$3.87e^{-05}$	$1.21e^{-05}$	0.28	*SLC9A8*
*rs*6126645	20	50745422	C/T	0.88	0.67	0.166	$1.10e^{-05}$	$2.43e^{-06}$	0.23	*TSHZ2*
*rs*681074	20	47814889	A/G	0.85	0.66	0.207	$2.95e^{-05}$	$3.71e^{-05}$	0.31	*SLC9A8*
*rs*2837857	21	41138825	C/T	0.8	0.65	0.299	$2.40e^{-06}$	$1.46e^{-05}$	0.3	*DSCAM*
*rs*3218258	22	35874191	A/G	0.8	0.65	0.309	$5.18e^{-06}$	$8.42e^{-06}$	0.27	*IL2RB*
*rs*11797250	*X*	17167482	A/C	1	1	0.072	$4.61e^{-05}$	$2.33e^{-05}$	0.23	*REPS2*
*rs*138067008	*X*	142838848	A/C	1	1	0.02	$6.20e^{-06}$	$2.62e^{-05}$	0.12	-
*rs*139956886	*X*	142842119	A/C	1	1	0.02	$5.96e^{-06}$	$2.66e^{-05}$	0.12	-
*rs*141261373	*X*	142827897	A/C	1	1	0.02	$6.91e^{-06}$	$4.09e^{-05}$	0.13	-
*rs*142513793	*X*	47906480	A/C	1	1	0.031	$1.84e^{-07}$	$3.14e^{-05}$	0.2	*ZNF630*
*rs*145189928	*X*	142830316	A/C	1	1	0.02	$6.91e^{-06}$	$4.09e^{-05}$	0.13	-
*rs*149912409	*X*	142832475	A/C	1	1	0.02	$6.91e^{-06}$	$4.09e^{-05}$	0.13	-
*rs*190796883	*X*	142827026	A/C	1	1	0.02	$6.91e^{-06}$	$4.09e^{-05}$	0.13	-
*rs*192138826	*X*	142823406	A/C	1	1	0.02	$6.91e^{-06}$	$4.09e^{-05}$	0.13	-
*rs*5924599	*X*	17139624	A/C	1	1	0.077	$8.16e^{-05}$	$3.72e^{-05}$	0.24	*REPS2*
*rs*5924602	*X*	17151123	A/C	1	1	0.075	$8.57e^{-05}$	$5.75e^{-05}$	0.25	*REPS2*

Continued on next page

Table 6.3 – continued from previous page

SNP	CHR	POS	A1/A2	Call	INFO	MAF	P	Fisher	OR	Gene
*rs*5928363	*X*	33784063	A/C	1	1	0.021	$3.72e^{-06}$	$2.01e^{-05}$	0.12	-

Chapter 7

Locus-specific Ancestry: Block Length distribution in multi-way Admixed Populations.

7.1 Introduction

Examining the genetic make-up of an admixed population has been suggested to be useful for understanding differences in disease prevalence and drug response among different populations. The analysis of the pattern of shared chromosomal segments between populations has provided critical insights into human colonization history, including multiple migration waves across continents, and the complex movement of people around the world (Price *et al.*, 2009b). Studying the admixture patterns in human populations has a wide range of critical applications from identifying both local selection and genetic variants underlying ethnic difference in disease risk, to an understanding of history (Seldin *et al.*, 2011). Methods have been developed to study the local genetic ancestry at the level of individuals within admixed populations (Baran *et al.*, 2012; Churchhouse & Marchini, 2012; Falush *et al.*, 2003; Henn *et al.*, 2012; Hoggart *et al.*, 2004; Pasaniuc *et al.*, 2009; Patterson *et al.*, 2006; Price *et al.*, 2009b; Rodriguez *et al.*, 2012). Most of these approaches have proven to be successful when using two-way or three-way admixed populations, such as African-Americans (Baran *et al.*, 2012). However, the accuracy of these methods have yet to be proven when using a multi-way admixed population such as the unique South African Coloured population. In addition, even locus-specific ancestry methods introduced recently, including ALLOY (Rodriguez *et al.*, 2012), PCAdmix (Henn *et al.*, 2012), MULTIMIX (Churchhouse & Marchini, 2012) and (Lawson *et al.*, 2010) could not achieved superior accuracy to LampLD (Baran *et al.*, 2012). All the approaches demonstrated equivalent accuracy to WinPOP.

The distribution of ancestry proportions of admixed individuals may be used to estimate distinct time of admixture events, to make inferences about population history, to complement case-control SNP association statistics in improving power in disease association studies (Pasaniuc *et al.*, 2011) and to identify the most significant sub-network underlying ethnic difference in complex diseases risk (sections 8.2.2 and 8.2.3 of the next chapter). Although the date of admixture events can be estimated from a direct estimation of the number of breakpoints Price *et al.* (2009b), new methods have been developed to date the admixture events in recently admixed populations which include:

(1) a likelihood-based method (HAPMIX) from the haplotype block information (Price *et al.*, 2009b).

(2) a PCA-based genome scan approach (StepPCO), that applies the wavelet decomposition of the estimated admixture signal to estimate the date of the admixture events (Pugach *et al.*, 2011).

(3) ROLLOFF method based on the rate of exponential decline of admixture linkage disequilibrium (LD). ROLLOFF fits an exponential distribution to the correlation between the LD of pairs of SNPs and a weighted function describing their allele frequency differentiation in the ancestral populations, with respect to a pre-identified admixed population (Moorjani *et al.*, 2011).

The above mentioned methods to estimate the date of admixture events are also limited to two-way admixture populations and recent admixture events. The timing of admixture events estimated from these methods show no simple relationship (Table 7.1). The estimation of date of admixture events is still in its infancy, and different approaches provide different results even with a simple two-way admixture population model (Table 7.1).

Table 7.1: **Example of comparing the estimated of date of admixture events (number of generations) for two-way admixed populations using the results from HAPMIX (Price et al., 2009b), StepPCO (Pugach et al., 2011) and ROLLOFF (Moorjani et al., 2011) methods.**

Populations	HAPMIX	StepPCO	ROLLOFF
Bedouin	90	83	31.3
Palestinian	75	72	33
Druze	60	90	44

An accurate and unbiased estimation of the ancestry at every SNP in multi-way admixed populations may potentially provide crucial insights into identifying disease genes, and provide information on the timing of the ancient or recent admixture event itself in any admixed populations (Seldin *et al.*, 2011). Because of the importance of the inference of locus-specific ancestry in both understanding population history and disease scoring statistics, this chapter assesses the accuracy of inferring local ancestry on a simulated multi-way admixed population using the most popular methods, including LampLD and WinPOP. We then aim to apply the most accurate method to real data of the SAC. We discuss another possibility of dating distinct admixture events in a multi-way admixed population such as the SAC using an exponential distribution of the ancestry block length along the genome of admixed individuals.

7.2 Materials and Methods

7.2.1 Assessment of Local Ancestry Inference in Multi-way Admixed Populations

Similarly to the simulation framework described in section 2.2.4, here we used two main parameters, the mixture proportion, that represents the probability that a particular sampled haplotype comes from an ancestry gene pool, and distinct dates of the admixture event as the number of generations since admixture occurred. These two parameters were used in terms of recombination breakpoints within the ancestral population chromosomes to generate samples at each generation. At each generation the ancestry information and breakpoint locations for a particular sample were stored. Each putative proxy ancestral population was independently phased as in section 2.2.4. For 165 CEU, 101 GIH, 203 YRI, 250 CHB+JPT and 22 SAN, BEAGLE created an ancestral haplotype pool of 330, 202, 406, 500 and 44 haploid (CEU, GIH, YRI, CHB+JPT and SAN) genomes, respectively. To generate n diploid admixed individuals, the simulation framework uses $2n$ ancestral haplotypes. We aimed to mix large populations to avoid elevated linkage disequilibrium (LD) caused by founder effects so that we can control the levels of true LD and admixture LD. Therefore, each ancestral population was independently expanded to a total size of 1500 plus its original size. From each expanded ancestral population, we split the resulting samples in two separate sizes. 1500 samples were eventually used to simulate diploid admixed individuals and the remaining simulated data of the original size was used to run two commonly used local ancestry methods, WINPOP Pasaniuc *et al.* (2009) and LampLD Baran *et al.* (2012).

Our aim to assess the accuracy of both WINPOP and LampLD in inferring local ancestry was achieved by looking at the correlation between inferred Y and the true ancestry Z. To estimate this correlation, we similarly computed an estimate of the expected squared correlation between

Y and Z as in Price *et al.* (2009b). Given an ancestral population k, the expected squared correlation between Y and Z is a ratio of the expected covariance of Y and Z and the product of the expected variance of both Y and Z taken over loci and individuals:

$$r^2_{yz} = \frac{\bar{c}ov(y,z)}{var(y).var(z)}. \tag{7.1}$$

In addition, based on the true Z and inferred Y local ancestry, we compute the rate of calling true ancestry among different populations. Given a true ancestral segment of length N_k ($2N_k$ is the total number of true ancestral alleles) derived from population $k \in K$ along the simulated genome of an admixed population, we computed the distribution of the rate of calling true ancestry $k \in K$ and the error rate of calling $k \neq j \in K$ ancestral populations instead of k as:

$$\tilde{e}rr = \frac{\tilde{\tau}}{2N_k}, j \in K, \tag{7.2}$$

where $\tilde{\tau}$ is the number of inferred ancestral alleles from ancestral population $j \in K$, the rate of calling true ancestry by summing over all loci and averaging over all individuals can thus be obtained.

7.2.2 Ancestry Block Size Distribution in Multi-way Admixed Populations

From the inferred ancestry at each location of the genome of an admixed population, we estimated ancestral block sizes at each interval of $1cM$ along the genome of each admixed individual as sets of contiguous SNPs for which either 1 or 2 alleles were assigned to each of the respective proxy ancestral groups. This approach is similar to estimating haplotype blocks using linkage disequilibrium in pure populations, as has been done for the HapMap populations (Frazer & et al, 2007). Given the ancestry block size b^k_{ij} derived from ancestral population k in the admixed individual i at interval j, we fitted a likelihood model to estimate the time since admixed occurred. We assumed each ancestry block size to be independent and identically distributed according to the Poisson distribution with parameter g (referred to as the number of generations since admixture occurred). Thus, for individual i, the joint probability density function of ancestry block size from ancestral population k, b^k_{ij} is given as

$$P(g|b_1,\ldots,b_J) = \prod_{j=1}^{J} P(b_j|g), \tag{7.3}$$

From Bayes theorem, the posterior distribution is known to be proportional to the product of Gamma prior (with α and β, the shape and scale parameters, respectively) $P(g)$ for the g and the likelihood function $L(g|b_1,\ldots,b_J)$. It follows,

$$\begin{aligned} Posterior &\propto L(g|b_1,\ldots,b_J)\frac{\beta^{\alpha}}{\Gamma(\alpha)}g^{\alpha^{-1}}exp(-g\beta) \\ &\propto g\left[\sum_{J}^{j=1} b_j exp(-Jg)\right]\left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}g^{\alpha^{-1}}exp(-g\beta)\right) \\ &\propto g\sum_{j=1}^{J} b_j + \alpha^{-1}exp\left[-g(\bar{b}_j+\beta)\right]. \end{aligned} \tag{7.4}$$

Equation 7.4 is a gamma distribution with $\alpha^* = \sum_{j=1}^{J} b_j + \alpha$ and $\beta^* = \bar{b}_j + \beta$, therefore,

$$g = \Gamma(\alpha^*, \beta^*) \tag{7.5}$$

7.3 Results and Discussions

7.3.1 Accuracy of Local Ancestry Inference in Simulated Data

As shown in previous results in section 2.3.2.2, the SAC has a complex admixture formed by the mixture of mostly five ancestral populations, namely Europeans, Southern Bantu, SAN, and South and East Asians. In order to see if the inferred local ancestry in the SAC can accurately be inferred, we first assessed the accuracy of two recent methods for inferring local ancestry in multi-way admixed populations, LampLD (Baran *et al.*, 2012) and WINPOP (Pasaniuc *et al.*, 2009). We simulated 749 individuals of mixed European (CEU), Chinese and Japanese (CHB+JPT), Bantu (YRI) and SAN ancestry (section 7.2.1). The simulation algorithm generated related information on ancestry and breakpoint locations for each simulated sample.

Each admixed individual was designed to be a mosaic of haplotypes from the above putative ancestral populations and was reflected in the admixture in the SAC. From the inferred local ancestry from both LampLD and WINPOP, we assessed the accuracy. We compared an estimate of correlation between the inferred local ancestry and the true ancestry information by computing the r^2 (section 7.2.1). LampLD reached a more similar magnitude to the true average of locus-specific ancestry across the genome than WINPOP (Figure 7.1). The r^2 in Table 7.2 suggests that LampLD provides greater accuracy for local ancestry inference in five-way simulated data than WINPOP.

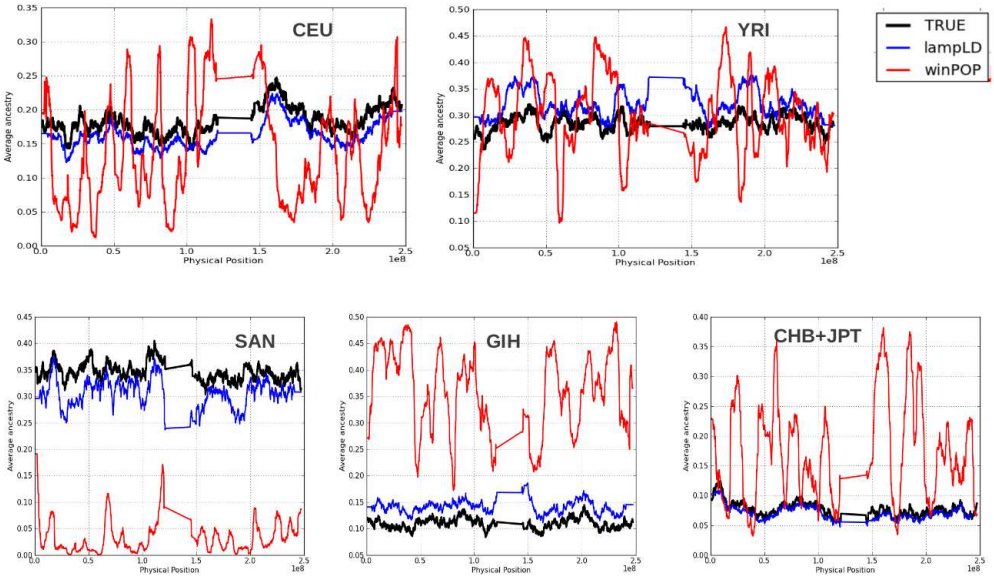

Figure 7.1: **The average of local ancestry across the genome of 749 diploid admixed individuals of mixed European (CEU), Chinese-Japanese (CHB+JPT), Bantu (YRI) and SAN ancestry. The plots Compare the true and the inferred average of local ancestry across the genome of a simulated multi-way admixed population.**

Table 7.2: **The average r^2 value (as described in section7.2.1) comparing the accuracy of WINPOP and LampLD in inferring the local ancestry on simulated data of 749 admixed individuals of mixed European (CEU), Chinese-Japanese (CHB+JPT), Bantu (YRI) and SAN ancestry.**

	CEU	YRI	GIH	CHB+JPT	SAN
LAMLD	0.89	0.87	0.88	0.92	0.92
WINPOP	0.51	0.69	0.49	0.49	0.67

The results in Figure 7.2 also show that LampLD inferred the true ancestral allele better than WINPOP. The superior accuracy of LampLD to WinPOP was expected, and supported by the results in Table 7.2.

In both simulation data and a real population of Latinos (Baran *et al.*, 2012) (known to be the result of a mixture of three ancestral populations), LampLD also demonstrated its superior accuracy to other existing algorithms.

As LampLD provides greater accuracy than WINPOP, we then deeply assessed the ability of LampLD to correctly infer the local ancestry in a multi-way admixed population by computing

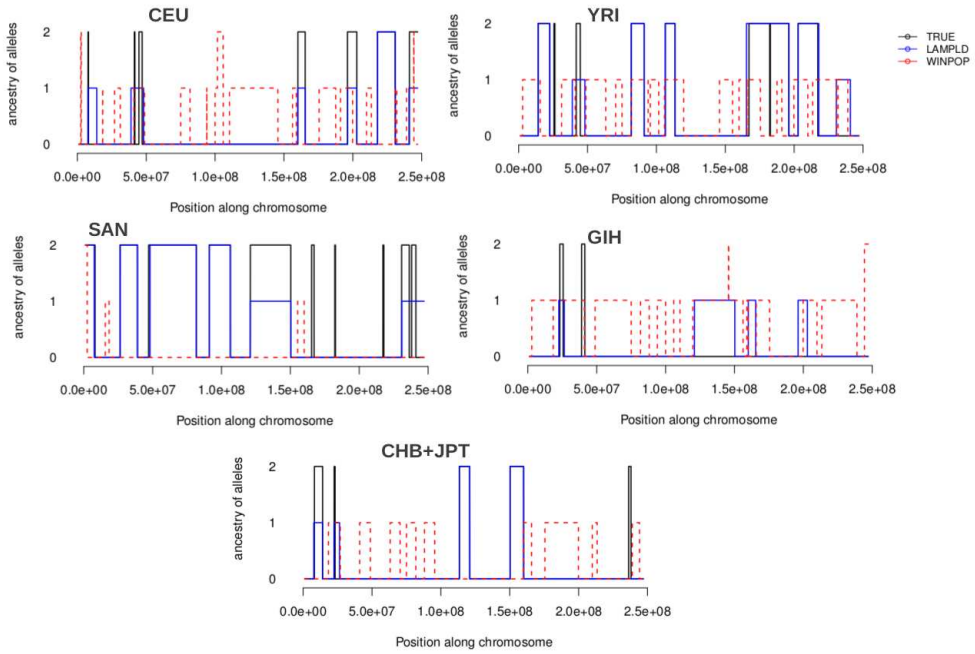

Figure 7.2: **The ancestral allele across the genome of one of the individuals of mixed CEU, CHB+JPT, YRI and SAN ancestry. Comparison of the true versus the inferred alleles across the genome of one individual picked randomly among the simulated samples.**

the rate of calling true ancestral and the error of calling other ancestral populations instead of the true ones. Table 7.3 demonstrates that even LampLD has still not correctly estimated the true local ancestry in a multi-way admixed population. Table 7.3 shows that the true CEU ancestry in the admixed population (simulation data) is miscalled as GIH ancestry (17%) more often than true GIH ancestry is miscalled as CEU ancestry (8.4%). The true SAN ancestry in the admixed population (simulation data) is miscalled as YRI ancestry (14.8%) slightly more often than true YRI ancestry is miscalled as SAN ancestry (14.2%). Many approaches of estimating the population of origin along the genome of an individual with a mixed ancestry, including HAPMIX (Price *et al.*, 2009b), LAMP (Baran *et al.*, 2012; Sankararaman *et al.*, 2008), WINPOP (Pasaniuc *et al.*, 2009), MULTIMIX (Churchhouse & Marchini, 2012) have been able to accurately estimate the local ancestry in 2-way or 3-way admixed populations (single point admixture event), such as African-Americans, Latinos, but their accuracies are still limited or not tested when using multi-way admixed populations (multi point admixture events). This is supported by our result in Table 7.3, which shows the limitation of LampLD, a current method for inferring local ancestry along the genome of multi-way admixed individuals which is known to achieve a reasonable accuracy.

Table 7.3: **Error rates in LampLD local ancestry inference in simulated data. In the first row, we list the probability of inferring each ancestry when the true ancestry is CEU. Other rows are analogous. We note that the table is asymmetric: for example, true CEU ancestry is miscalled as GIH ancestry more often than true GIH ancestry is miscalled as CEU ancestry.**

	CEU	YRI	GIH	CHB+JPT	SAN
CEU	79%	3%	17%	0.7%	0.3%
YRI	1.4%	72%	2%	0.4%	14.2%
GIH	8.4%	3.4%	85.8%	1.4%	0.8%
CHB+JPT	2.1%	3.2%	8%	86%	0.7%
SAN	0.9%	14.8%	1%	0.3%	74%

7.3.2 The SAC: Locus-Specific Ancestry and Ancestry Block Size Distribution

To maximize the genotype coverage in inferring local ancestry, the locus-specific ancestry was inferred using LampLD on the SAC data within the ancestral haplotypes from IsiXhosa, European (CEU), ‡Khomani (KHO) and Gujarati (GIH) and East Asian (CHD). Figure 7.3 displays the average ancestry at each genetic locus along the genome of the SAC.

We estimated the length of ancestry blocks in the SAC using the inferred locus-specific ancestry from LampLD Baran *et al.* (2012). The length of ancestry blocks contributed by each of the putative ancestral population (Bantu, European, Khoesan and South-East Asian) were estimated at each interval of $1cM$ along the genome of each admixed individual of the SAC. Ancestral blocks were identified by sets of contiguous SNPs at which at least 1 of the two alleles were assigned to a particular ancestral proxy (section 7.2.2). From the estimated ancestry block sizes, we fitted a likelihood model to estimate different date of admixture events (Figure 7.4) from different proxy ancestral groups. Overall, from the different admixing times from different ancestral populations, our result in Figure 7.4 shows that the genetic make-up of the SAC started 9 to 11 generations (385 years) ago, if we consider 35 years for one generation. This result suggests an early admixture that started between populations related to current African Bantu-speakers and click-speakers populations (as well as GIH), then followed complex admixture to result in the current SAC.

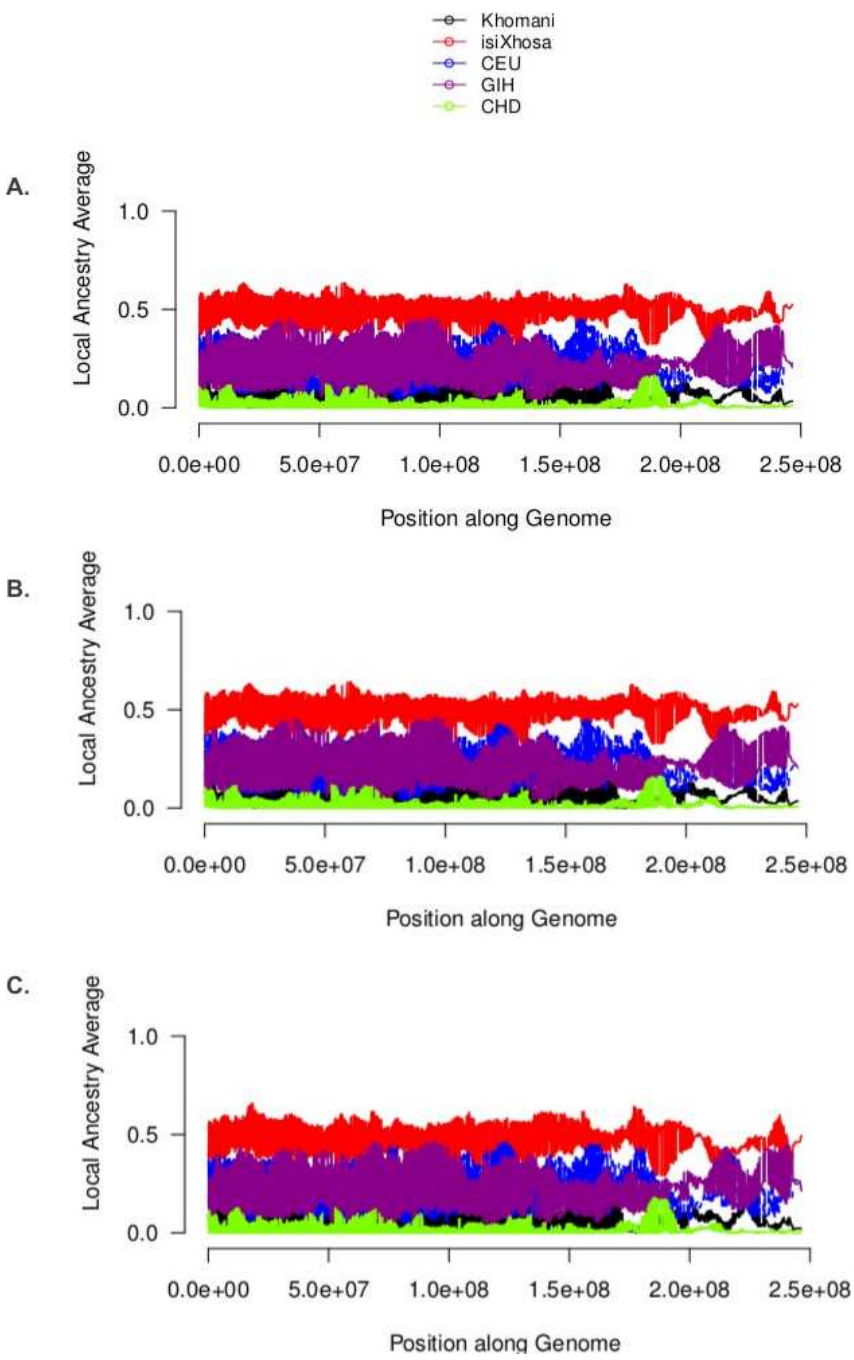

Figure 7.3: **The average of local ancestry across the genome of the SAC using all samples, cases and controls. The plot shows different ancestry segments from CEU, CHD, GIH, IsiXhosa and ‡Khomani in the SAC, (A) using all samples, (B) using only case samples and (C) using only control samples.**

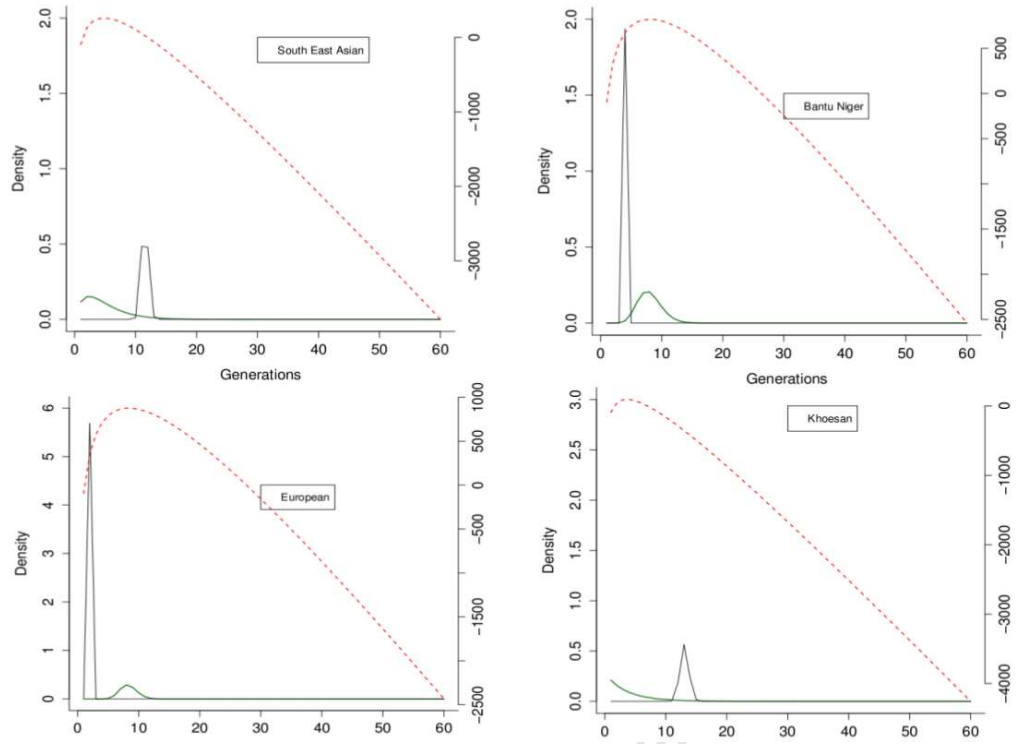

Figure 7.4: **Plots are generated from our likelihood model to estimate the number of generations (g) since admixture occurred based on the ancestry block length distributions. The dotted red line is the likelihood of g with its y-axis on the right of the graph, the green line is the prior parameter of g and the black solid line is the posterior of g.**

7.4 Concluding Remarks

Through simulation of a complex 5-way admixed population, we assessed the accuracy of current approaches to estimate local ancestry in multi-way admixed populations. Our result demonstrates the limitation in accuracy of these methods in inferring local ancestry in multi-way admixed populations. In addition, although we were able to estimate date of admixture events in the SAC by fitting a likelihood model on the distribution of ancestry block length from the local ancestry along the genome of admixed individuals, the accuracy of both dating admixture events and local ancestry in multi-way admixed population are still open questions. In addition, an accurate inferred local ancestry may complement the disease scoring statistics (Pasaniuc *et al.*, 2011) in admixed populations and be informative in fine mapping methods for diseases for which risk differs depending on ancestry.

Chapter 8

Genes and Sub-networks Underlying Ethnic Difference in Complex Disease Risk in a Recently Admixed Population.

8.1 Introduction

Despite numerous successes of Genome-wide Association Studies (GWAS) based on single discovery SNP methods, many authors have pointed out that GWAS may not detect genetic variants having low or moderate risk that do not reach the intrinsic genome-wide significance cut-off of $P < 5 \times 10^{-8}$ (Peng *et al.*, 2008). Moreover, only a few common variants have presently been found to be involved and the associated loci explain only a small fraction of the genetic risk (Cantor *et al.*, 2010). Because the effect of a gene polymorphism, is viewed in isolation, GWAS may fail to reveal a significant signal if the effect of a variant on another gene is not taken into account. Therefore, single discovery SNP based analysis in GWAS may generate false negative results (Jia *et al.*, 2010; Peng *et al.*, 2008), and in many cases, an inconclusive result. One of the remaining challenges of GWAS is the translation of associated loci into biological hypotheses suitable for further investigation in the laboratory. Another critical challenge improving our understanding of how multiple, modestly-associated loci within genes interact to influence a phenotype (Cantor *et al.*, 2010; Jia *et al.*, 2010; Peng *et al.*, 2008).

Recent investigations have demonstrated that there is a relationship between gene function and phenotype, and that functionally-related genes are more likely to interact (Jia *et al.*, 2010; Peng *et al.*, 2008). Genes can influence each other, e.g through enhancement or hindrance. This can occur directly at the genomic level, where a gene could code for regulator gene preventing transcription of the other gene. Alternatively, the effect can occur at the phenotypic level, where a pair of gene products can work together to produce a specific phenotype. Thus, pathways have

critical roles in aiding in the understanding of the cause of disease. In addition, risk-associated genes may differ in different individuals, but may be in the same pathway. Identifying pathways associated with a disease may, therefore, allow us to more easily discover the pathogenesis of the disease. Furthermore, considering the multiple genetic and environmental factors contributing to development of a complex disease, such as infectious diseases, in particular TB, GWAS alone is insufficient to elucidate the complex genetic structure of complex diseases. Thus, examining the combined effects of genes by detecting genetic signals beyond single gene polymorphisms provides increased potential to fully characterize the susceptible genes and the genetic structure of complex diseases (Jia *et al.*, 2010; Moller & Hoal, 2010b; Peng *et al.*, 2008). Inspired by this insight, researchers have suggested conducting a post GWAS analysis that combines different association studies to reveal larger effects and to provide valuable information that will be useful for prioritizing the most important results (Han & Eskin, 2011; Wray *et al.*, 2010). This approach is known as Meta Analysis (as we conducted in section 6.3.2), and it aims to pool information from multiple GWAS to increase the chances of finding associations with small effect sizes (Cantor *et al.*, 2010; Han & Eskin, 2011), it has already successfully identified susceptibility loci (Han & Eskin, 2011). Another post association analysis was recently suggested as a new paradigm for GWAS (Cantor *et al.*, 2010; Jia *et al.*, 2010; Peng *et al.*, 2008), i.e to elucidate genetic susceptibility by incorporating both the association signal from GWAS and the human protein-protein interaction (PPI) network for testing the combined effects of SNPs and searching for significantly enriched sub-networks for a particular complex disease. This approach is based on combining p-values from standard GWAS for correlated SNPs into an overall significance level to represent a gene, and using the combined p-values to investigate the association of a pathway with the disease (Jia *et al.*, 2010). However, in many cases SNPs within a gene, and genes within a pathway are correlated, but most of these methods do not account for this dependency of p-values, which are assumed to be independent and uniformly distributed under a null hypothesis. The violation of independent assumptions in these methods may generate erroneous results.

In this chapter, we present a new algebraic graph-based method (ancGWAS) to identify the most significant sub-network underlying ethnic difference in complex diseases risk in a recently admixed population. This is done by integrating the association signal from a GWAS data set, the local ancestry, and SNP pair-wise linkage disequilibrium from the admixed population into the PPI network. The ancGWAS method accounts for the correlation that exists between SNPs within a gene and genes within a pathway. ancGWAS is based on graph-based centrality measures, considers linkage disequilibrium, and applies a statistical score to the resulting sub-graphs to identify the most significant sub-graphs associated with complex disease risk, and also tests for possible signals of unusual differences in an excess/deficiency of particular ancestry. In addition, this method introduces flexibility in estimating gene and sub-network-specific ancestry.

Through simulation of interactive disease loci in a simulated 4-way admixed population, we evaluated ancGWAS. The results from our simulation demonstrated that ancGWAS holds promise for comprehensively examining the interactions between genes underlying the pathogenesis of complex diseases and also underlying ethnic difference in disease risk. We applied ancGWAS to the imputation TB GWAS data of the South African Coloured population. Our results replicate previous tuberculosis loci and introduce novel genes and sub-networks predominately with African-specific ancestry.

8.2 Development of ancGWAS

8.2.1 Assignment of Ancestry, P-values and LD from SNPs to Gene Level

We constructed a pair-wise PPI dataset, by adding $35,671$ human PPIs to the $64,000$ interactions in the comprehensive pair-wise human PPI network downloaded from the Protein Interaction Network Analysis platform (PINA) (Wu *et al.*, 2009). The PINA data were collected and annotated from six public PPI databases (MINT, IntAct, DIP, BioGRID, HPRD, and MIPS/MPact). Our updates were based on the same six databases, and we manually included TB related PPIs from published papers (Costa *et al.*, 2012; deWit *et al.*, 2010b). We finally generated a total of $99,671$ PPIs for a network. We merged the TB GWAS data set of the admixed Coloured population from South Africa with its estimated local ancestry data from LAMPLD (Baran *et al.*, 2012) into one data set. The merged data set and the PPI data set were used as inputs for the ancGWAS method.

SNPs and their local ancestry together with the associated p-values were assigned to a gene if the SNPS were located within a gene is primary transcript or 40 kilobases (kb) downstream or upstream. If a SNP was assigned to multiple genes due to overlapping flanking windows, the closer gene was chosen according to a specified boundary cut-off. To achieve this, we downloaded genomic coordinates for all genes from the NCBI ftp-server (ftp://ftp.ncbi.nih.gov/), retaining only entries for the human reference sequence and protein-coding genes. We updated genomic coordinates to the latest assembly using the Lift-Over tool on GALAXY (https://main.g2.bx.psu.edu/). We made use of four statistical methods (Peng *et al.*, 2008) for assigning both the association p-values and local ancestry information to genes, including Fisher's method (section 8.2.3), Simes, the Smallest (section 8.2.3) and the Smallest gene-wise FDR methods, as was done previously in (Jia *et al.*, 2010; Peng *et al.*, 2008).

(1) **simes:** Let $p_1 \leqslant p_2 \leqslant \ldots \leqslant p_m$, be m ordered p-values from SNPs associated with a gene g_k. The combined p-value in a gene g_k is calculated as

$$p_{g_k} = min_j\{\frac{mp_i}{j}\}$$

(2) **FDR method:** Let us denote π to be the proportion of tests with a true null hypothesis and $H(\beta)$ be the expected proportion of tests yielding a p-value less than or equal to β, and let us denote $Z(\beta)$ to be the expected proportion of tests giving a false positive result with significance level β.

Now assuming there are m distinct p-values among $p = \{p_1, \ldots, p_k\}$. Let us also assume that $\tilde{p}_1 < \tilde{p}_2 < \ldots < \tilde{p}_m$. And denote n_i to be the number of p-values among p that are equal to $\tilde{p}_i$. It follows,

$$\tilde{H}(\beta) = \frac{1}{m} \times \sum_{i=1}^{m} I(\tilde{p}_i \leqslant \beta) \times n_i$$

where *I* is an indicator function. For a two-sided test define $\pi = min(1, 2\bar{p})$, and for a one-sided test (χ^2-test, trend test) define $\pi = min(1, 2\bar{\beta})$, where $\bar{p} = \frac{1}{2} \times \sum_{j=1}^{m} p_j$, $\bar{\beta} = \frac{1}{2}\sum_{j=1}^{m} \beta_j$ and $\beta_j = 2 \times min(p_j, 1 - p_j)$. $Z(\beta)$ is estimated by $Z(\beta) = \pi\beta$. It follows, the test for association at the gene or network level is given by

$$T_j = \frac{Z(p_j)}{H(p_j)}$$

To incorporate the strength of correlation (LD) between two genes into the human PPI network, we compute pair-wise linkage disequilibrium (LD) between SNPs in each pair of interacting genes. Given SNPs s_i and s_j ($s_i \neq s_j$) among M and N SNPs associated to the first and second gene, respectively, the pair-wise SNP-LD is computed using the $r2$ measure. We provide three approaches for weighting these interactions.

(1) **closestLD:**
Considering SNPs s_j are assigned to their closest genes G_j, we immediately assign the SNP-LD $LD_{s_i s_j}$ to gene-LD $r_{G_i G_j}$,

$$r_{G_i G_j} = LD_{s_i s_j} \tag{8.1}$$

(2) **ZscoreLD:**
Assuming multiple SNPs are assigned to genes G_j and SNPs between pair of genes, G_k and G_l are independent and uniformly distributed under the null hypothesis, we consider the Z

score of LD from all possible N pairs of SNPs within a pair of genes, G_k and G_l $(k \neq l)$ with multiple assigned SNPs $i = 1, 2, ..., n$ and $j = 1, 2, ..., m$, respectively.

$$r_{G_k G_l} = \frac{\sum_{i \neq j}^{N} LD_{s_i s_j}}{\sqrt{N}}. \tag{8.2}$$

(3) **maxLD:** Alternatively to the case above, if SNPs between a given pair of genes are dependent or correlated, we consider the maximum $LD_{s_i s_j}$ among all possible N pair of SNPs between the pair of genes.

$$r_{G_k G_l} = \max(L_{ij}), \tag{8.3}$$

Equations 8.1, 8.2 and 8.3 are used as the weight of the edge between G_k and G_l genes in the PPI network.

8.2.2 Searching for Sub-networks Using Centrality Measures

Here, we discuss graph-based measures to quantify the relevancy of nodes (genes in our case) in our LD-weighted PPI network. Genes are interacting in a large networks of genes, RNA, DNA, metabolites and other molecules in every single living cell. These interactions are generally described as networks, and some nodes in the network are more important or central than others. For instance, highly connected nodes in PPI networks can be functionally important and the removal of such nodes is related to lethality. We consider our weighted PPI network as an undirected network, $G = (V, E)$, with n nodes defined as genes and edges as interactions found between genes, weighted using LDs. To cluster G into sub-networks, we analyse the general properties of G and quantify the usefulness of each gene in G using their centrality scores; closeness, betweenness, degree or eigenvector. Let us first define the follows centrality measures:

(1) **Degree Centrality C_d:** The degree centrality C_d of a node in an undirected graph is given by $C_d = deg(u)$, In terms of adjacency matrix $\mathcal{A}$ [28], the degree centrality of a node $u \in V(G)$ is simply the sum of components in the row or the column corresponding to the node u, and is given by

$$deg(u) = \sum_{v=1}^{n} a_{uv}, \tag{8.4}$$

where v is any other node in $V(G)$. The degree centrality provides an indicator of the influence of a gene on the biological system, and can indicate that the gene plays a key role in the functioning of the system. C_d is also used, for instance, to correlate the degree of a gene in the network with the lethality of its removal.

(2) **Closeness Centrality Measure C_c:** The closeness centrality C_c is given by

$$C_c = \frac{1}{\sum_{u \in V} dist\,(u,v)}, \tag{8.5}$$

we interpret equation 8.5 as the probability of a gene being functionally relevant for several other genes, with the possibility of being irrelevant for a few other genes. Thus, the gene with high closeness, compared to the closeness of the whole network, may be central to the regulation of other genes.

(3) **Shortest Path Betweenness Centrality Measure C_{spb}:** Let us denote γ_{uv}, the number of shortest paths between u and v, and $\gamma_{uv}(t)$ the number of shortest paths between u and v in the network G using t as an interior node, for $t, u, v \in V(G)$. The rate of communication between u and v, Δ_{uv} that can be controlled by an interior node t, is given by

$$\Delta_{uv} = \frac{\gamma_{uv}(t)}{\gamma_{uv}},$$

if $\gamma_{uv} = 0$, then we set $\Delta_{uv} := 0$. The shortest path betweenness centrality $C_{spb}(t)$ is given by

$$C_{spb} = \sum_{u \in V \wedge u \neq t} \sum_{v \in V \wedge v \neq t} \delta_{uv}(t).$$

In protein signalling networks, the shortest path betweenness centrality of a protein can determine its relevance as functionally able to hold together communicating genes and also can indicate the capability of a protein to facilitate communication between distant genes.

(4) **Eigenvector Centrality Measure C_{ev}:**

The eigenvector centrality measure concerns the usefulness or weight of functional connections of genes and can only be considered as a measure of centrality if nodes are ranked with regard to their participation in different sub-networks. The eigenvector centrality measure assigns relative weights to all genes in the network based on the fact that connections to high-weighted genes contribute more to the weight of the protein target.

Let us denote $A = (a_{uv})$ the adjacency matrix of $G = (V, E)$, for any $u, v \in V(G)$. For each node u, let the centrality score x_i be proportional to the sum of the scores of all nodes v connected to u. It follows that,

$$x_u = \frac{1}{\lambda} \sum_{uv \in E(G) \wedge u \neq v} x_v, \tag{8.6}$$

$$x_u = \frac{1}{\lambda} \sum_{v=1}^{n} a_{uv} x_v,$$

where v is any other node connected to u, n is the number of nodes of the network G and λ is a constant.

It is believed that a gene associated with human complex disease susceptibility, may be central nodes of a particular biological sub-networks, whereby other genes within that sub-network or other sub-networks are linked to it via few steps (path or edges in the network) (Jia *et al.*, 2010). These centres are structural hubs with centrality scores beyond a certain threshold value. Biological topological property tests of a biological network confirm this. Let us denote $o(G)$ the order and $s(G)$ the size of G, respectively. We denote SP_{mean}, the shortest path mean from every node to every destination within the network G, we perform the following steps to identify sub-networks using centrality scores of each gene:

Algorithm 2 : Sub-network Searching Algorithm (SSA)

(1) Given network G, find structural hubs and connected components;

(2) For each gene, compute the betweenness score, the closeness and the eigenvector score;

(3) For each centrality score, compute the cut-off for central genes of sub-graphs BetOf, ClosOf, DegOf and EigOf;

(4) Consider a gene as a hub if its score is greater than or equal to the corresponding cut-off;

(5) Consider a gene as a central gene only if the gene is a hub for all four scoring measures in step (3);

(6) For each central gene, search for its neighbours given a step n or the mean shortest path. The central gene and its neighbours constitute a sub-network of G.

8.2.3 Scoring Gene and Sub-network Ancestry

(1) **Fisher's Method**

Let $M = \{m_1, ..., m_K\}$ be the set of sub-networks, each with a hub generated from our clustering approach described above. For $k = 1, ..., K$, let $m_k^g = (g_1, ..., g_{N_k})$ be k^{th} sub-network ($|m_k^g| = N_k$ genes), and $m_k^p = (p_1, ..., p_{N_k})$ be the N_k-dimension vector of p-values associated with the gene within m_k^g. It implies that

$$T = -2 \times \sum_{i=1}^{N_k} log(p_i), \tag{8.7}$$

is $\mathcal{X}^2$ with a degree of freedom $2 \times N_k$.

Therefore, the p-value of m_k^p is obtained as follows,

$$p_{m_k} = 1 - \mathcal{X}^2_{cdf}(T, dof), \tag{8.8}$$

(2) **Stouffer Z'score Method**

Let ϕ^{-1} be the inverse normal distribution. It follows that the Z'score of a sub-network m with $|m| = N$ genes,

$$Z_m = (\sum_{i=1}^{N} \phi^{-1}(1 - p_i)) \frac{1}{\sqrt{N}} \tag{8.9}$$

is a normal distribution. Therefore, we can obtain the p-value of m_k^p as follows,

$$p_{m_k} = 1 - N_{cdf}(Z_m) \tag{8.10}$$

Let $M^p = \{p_{m_1}, ..., p_{m_k}\}$ be a set of p-values associated with sub-networks. Let $H = M^p - \{p_{m_k}\}$, thus we obtain the normalized score for sub-network m_k as follows,

$$S_k = \frac{p_{m_k} - mean(H)}{var(H)} \tag{8.11}$$

Given a sub-network m with $|m| = N$ genes, we expected around $N \times 0.05$ to have p-value less than 0.05 in each sub-network, respectively. We thus, estimate for each sub-network the genome-wide significance level as $\alpha = \frac{0.05}{\sqrt{N}}$.

(3) **Gene and Sub-network Ancestry-specific Method**

Given the genome-wide ancestral proportion α_k from ancestral populations $k \in \{1, ..., K\}$ in I samples of an admixed population. Let $\phi_k^{i,m}$ be the estimated locus-specific ancestry of individual i at genetic marker $m \in \{1, 2, ..., M\}$ associated with a particular gene, from the k^{th} ancestral population. We compute the deficiency or excess of ancestry, at each SNP using the estimated admixture proportion (that may be obtained from a programme such as ADMIXTURE, STRUCTURE as a baseline). We thus define, under a null hypothesis, the deficiency/excess of ancestry from ancestral population k at marker m as,

$$\delta_k^m = (\frac{1}{N}\sum_{i=1}^{N}\phi_k^{i,m}) - \alpha_k = \overline{\phi_k^m} - \alpha_k, \tag{8.12}$$

where $\overline{\phi_k^m}$ is the average locus-specific ancestry at SNP m. δ_k^m can be approximated as a normal distribution under neutral drift with mean 0 and empirical variance, derived from the distribution of $\phi_k^{i,m}$ values among the N individuals. It thus follows that,

$$Z_k^m = \frac{(\delta_k^m)^2}{\sqrt{\hat{v}ar(\phi_k^{i,m})}} \tag{8.13}$$

is a χ^2 with 1 degree of freedom. Summing-up equation 8.12 over all SNPs assigned to a gene, we can obtain the deficiency/excess of ancestry at the gene level. Summing-up equation 8.13 over all SNPs assigned to a gene, equation 8.13 will be a χ^2 with $M-1$ degrees of freedom. This allows us to assess the statistical significance of a deficiency/excess of ancestry at the gene level. To assess unusual difference in a deficiency/excess of ancestry between a pair of ancestral populations given SNP $m \in \{1,2,...,M\}$ within a gene, we compute

$$\hat{t}_{lk} = \frac{\sum_{m=1}^{M}(\delta_k^m - \delta_l^m)^2}{\sqrt{\frac{\hat{v}ar(\phi_k^{i,m}) + \hat{v}ar(\phi_l^{i,m})}{M}}} \tag{8.14}$$

which is a two-sample t-statistic with $M-2$ degrees of freedom. For a pair of populations, $k \neq l \in \{1,2,...,K\}$, we compute the overall unusual difference in a deficiency/excess of ancestry,

$$\hat{t} = \frac{\sum_{m=1}^{M}\sum_{l\neq k}^{K}(\delta_k^m - \delta_l^m)^2}{\sum_{l\neq k}^{K}\sqrt{\frac{\hat{v}ar(\phi_k^{i,m}) + \hat{v}ar(\phi_l^{i,m})}{M}}} \tag{8.15}$$

Thus, given the deficiency/excess of ancestry at the gene level, the above statistical analysis can be replicated at the sub-network level. For each method described above, the bootstrap approach has been used to compute the overall score (or p-values) and their 95% confidence interval for a single gene and sub-network of genes.

8.2.4 Evaluation of the ancGWAS Approach

To simulate case and control data of a non-admixed population, we use the simulation method implemented in Hapgen2 (Zhan *et al.*, 2011). This method resamples known haplotypes and produces samples with patterns of linkage disequilibrium (LD), which mimic those in real data. In order to capture the patterns of linkage disequilibrium (LD) in a dense real dataset, we first simulated a non-admixed population, with 1000 cases and 1000 controls, using the Yoruba (YRI) HapMap3 population with 2 disease SNPs on two polymorphisms, $rs2297977$ and $rs841404$. These are associated with the SLC2A1 gene, with heterozgyote risks 1.5 and 2, homozygote risks 2.25 and 4, and risk alleles set to 1 and 0 at each SNP, respectively. The resulting simulated population (SIM) can now be used as a new reference population in a panel also including European (CEU), Gujarati Indian (GIH) and Chinese (CHB) from HapMap3 data. After expanding each of these four ancestral populations to an additional 2000 samples, we sampled haplotypes from CEU, GIH, CHB, SIM (the simulated homogenous population from YRI) with probability related to a given ancestral proportion. To simulate n diploid admixed individuals, we sample the haplotypes from SIM, European (CEU), Gujarati Indian (GIH) and Chinese (CHB) with probability related to a given ancestral proportion from each putative ancestral population (**60**%, **20**%, **12**% and **8**%, respectively). Considering a continuous gene flow model in ten generations and accounting for the Wright-Fisher model with random mating, we simulated the genomes of 1000 cases and 1000 controls of mixed ancestry from SIM, CEU, GIH and CHB. Using the obtained admixed population we simulated four disease SNPs (including the previous two SNPs), at $rs2297977$, $rs841404$, $rs790633$ and $rs6664119$ with heterozgyote risks 1.5, 2, 1.5 and 2, homozygote risks 2.25, 4, 2.25 and 2.25 and risk alleles set to 1, 0, 1 and 0 at each SNP, respectively. These four SNPS are in linkage disequilibrium. Our simulation was based on chromosome 1 (n $= 116,415$ SNPs) and the simulated disease loci were on region $1p31.3$ (IL23R gene) for $rs2297977$ and $rs841404$ SNPs (transmitted from parental population) and $1p34.2$ (*SLC2A1* gene) for $rs790633$ and $rs6664119$ SNPs (simulated in resulting admixed population while expanding it). Of note, *IL23R* and *SLC2A1* are interacting genes. We conducted standard GWAS on the final simulation data set by applying EMMAX (Kang *et al.*, 2010), which accounts for both population stratification and hidden relatedness. To account for interacting disease SNPs and moderate risk that may not reach the intrinsic genome-wide significance cut-off of $P < 5 \times 10^{-08}$ in the standard GWAS above, we applied ancGWAS to the simulation GWAS result and previous imputation TB GWAS dat set from the SAC.

8.3 Results and Discussion

We implemented the algorithms described in sections 8.2.2 and 8.2.3 in ancGWAS, which is available at http://www.cbio.uct.ac.za/ancGWAS. ancGWAS has the advantage of not only using a linkage disequilibrium weighted network, but also the flexibility to test for possible signals of unusual differences in an excess/deficiency of ancestry and ancestry proportions at the gene and sub-network level. ancGWAS achieves these by integrating the association signal from GWAS data, the local ancestry and SNP pair-wise linkage disequilibrium from the admixed population into the human protein-protein interaction (PPI) network (Figure 8.1).

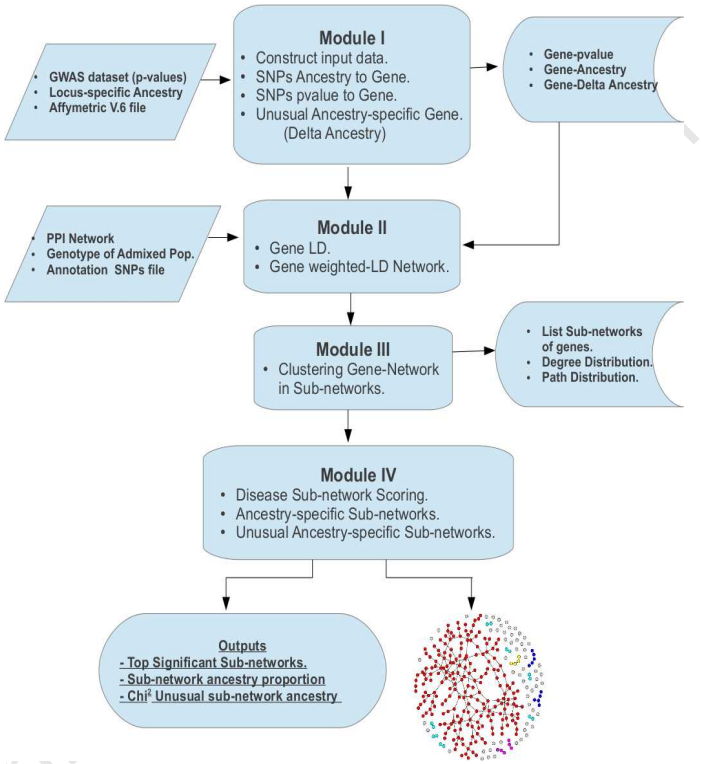

Figure 8.1: **Work-flow of ancGWAS approach, describing the functioning of the program and providing an overview of the inputs, modules and outputs.**

8.3.1 Evaluation of ancGWAS on Simulated Data

We evaluated ancGWAS using the simulation data of a 4-way admixed population within four disease loci in the regions $1p31.3$ (*SLC2A1* gene) and $1p34.2$ (*IL23R* gene) (see section 8.2.4). We first conducted the association analysis on this simulation data by applying EMMAX, which accounts for both population stratification and hidden relatedness. Table 8.1 lists the top 18 most significant SNPs obtained from EMMAX, including the four simulated disease loci. Of

note, EMMAX partially failed to significantly identify simulated disease loci at the $rs841404$ and $rs790633$ SNPs and other related SNPs under linkage disequilibrium with the simulated disease loci, including $rs841856$, $rs790633$ and $rs1385129$. These SNPs are below genome-wide significance (Table 8.1).

Table 8.1: **Top 18 genetic markers with moderate/significant p-values obtained from the association analysis with simulated disease loci on the simulation data of the admixed population.**

GENE	Closest SNP	True Disease SNP	P
SLC2A1	$rs3738514$	NO	$2.53e^{-08}$
SLC2A1	$rs841404$	YES	$1.26e^{-05}$
NLRP3	$rs10157521$	NO	$7.23e^{-05}$
PTGER3	$rs2300177$	NO	$4.73e^{-05}$
IL23R	$rs6664119$	YES	$1.32e^{-30}$
RPE65	$rs4313431$	NO	$4.69e^{-05}$
RPS7	$rs4926338$	NO	$8.96e^{-05}$
SGIP1	$rs17492182$	NO	$3.49e^{-05}$
IL23R	$rs790633$	YES	$5.00e^{-08}$
SLC2A1	$rs3806401$	NO	$1.06e^{-08}$
SLC2A1-AS1	$rs1385129$	NO	$4.98e^{-07}$
PLD5	$rs7554715$	NO	$3.46e^{-05}$
NUP133	$rs16849788$	NO	$7.17e^{-05}$
SLC2A1	$rs2297977$	YES	$8.40e^{-09}$
MIR101-1	$rs555146$	NO	$7.93e^{-05}$
GNG12-AS1	$rs12239301$	NO	$6.70e^{-07}$
SLC2A1	$rs841856$	NO	$2.80e^{-06}$
SLC2A1-AS1	$rs844501$	NO	$3.34e^{-08}$

To identify the moderate risk that did not reach the intrinsic genome-wide significance cut-off p-value $< 5 \times 10^{-8}$ in the above GWAS based on simulation data (Table 8.1), we combine the effects of all SNPs in a particular gene, and at the pathway level using ancGWAS. According to the work-flow in Figure 8.1, we first combined the obtained GWAS data set and the true locus-specific ancestry obtained from the simulation of mixed ancestral populations. We computed the summary p-value of each gene from multiple SNPs using the statistical Fisher's method. Since all the methods described in sections 8.2.2 and 8.2.3 produced similar summary p-values at the gene level, to simplify the presentation of results, we report on just one method. The results in Table 8.2 display top the 29 moderate/significant genes from the ancGWAS analysis using the combined effect from several SNPs for each gene to refine the association signal. Interestingly, the simulated disease genes *SLC2A1*, including *SLC2A1-AS1* and *FAM183A* genes, which are in

LD with *SLC2A1*, which were on the boundary of genome-wide significance from standard GWAS (Table 8.1), are now significant (Table 8.2) after combining effects of different SNPs within each gene. This result demonstrates the power of examining the combined effects of genes by detecting genetic signals beyond single SNPs. We tested for possible signal of unusual difference of a deficiency/excess of ancestry under a null hypothesis and the reported χ^2 values in Table 8.2 indicate no significant signal of unusual difference of a deficiency/excess of ancestry, which is consistent with our simulation framework which did not account for a model of differentiate ancestral allele frequency. This result can also be explained by the fact the simulated time of single admixture event was too short to have an impact of unusual deficiency/excess of ancestry in the simulated. The gene ancestry-specific information from the mixed ancestral populations in Table 8.2 is proportional to the true ancestry proportion used to simulate the admixed population.

To gain the benefit of fully characterizing the susceptible genes and the genetic structure of the simulated disease, we then conducted sub-network association analysis using ancGWAS (see method in section 8.2.3 and 8.2.2). To this end, three methods are available, including closestLD, maxLD and ZscoreLD (section 8.2.1). These three methods have similar results, therefore we only report the simulation result from the closestLD method. The LD-weighted network was constructed using the closestLD method described in section 8.2.1. A topological test was performed on the constructed LD-weighted network of $1,742$ pair-wise gene-gene interactions. We wanted to assess whether there is realistically an opportunity to use topological properties of the network as factor for clustering. Figure 8.2 reveals that the network exhibits scale-free topology, which means the degree distribution of genes approximates a power law $P(k) = k^{-\gamma}$, where $\gamma \approx 2.19$ is the degree exponent obtained by fitting the model using the least-square approach. This indicates that most of the genes have few interacting partners, but some have many.

Figure 8.3.1, shows that the network has a small world property, suggesting that the spread of information in the network is achieved through 7.01 steps, which corresponds to the average shortest path length in the network. We used the topological properties of nodes to break down our network in sub-networks, applying clustering algorithm described in algorithm 2. First, we found all the hubs of the networks and successively, the betweenness centrality, the closeness centrality and the eigenvector centrality measures for each node were computed. We computed the cut-offs for each centrality measure, and the intersection of the resulting sets were considered as the set of centre nodes. For simplicity of presentation, we limited our sub-network search at $step = 1$. We assessed the significance of each sub-network using the Fisher's method in ancGWAS.

Table 8.2: **Top 29 genes with moderate/significant p-values obtained from the ancGWAS method of combined SNP association analysis with simulated disease on the simulation data of an admixed population. The table also displays ancestry-specific information from each ancestral population at the gene level. The header chi^2_D denotes the χ^2 of unusual difference in an excess/deficiency of ancestry.**

GENE	CEU	CHB	GIH	SIM	chi^2_D	P
IL23R	0.198	0.074	0.125	0.603	0.003	$1.32e^{-30}$
SLC2A1	0.225	0.088	0.106	0.581	0.02	$8.4e^{-09}$
SLC2A1-AS1	0.225	0.088	0.106	0.581	0.02	$8.4e^{-09}$
ZNF691	0.223	0.088	0.106	0.583	0.024	$8.4e^{-09}$
FAM183A	0.22	0.095	0.107	0.578	0.043	$1.06e^{-08}$
GNG12-AS1	0.21	0.072	0.12	0.599	0.002	$6.73e^{-07}$
ERMAP	0.222	0.088	0.106	0.583	0.024	$4.98e^{-07}$
GNG12	0.206	0.072	0.121	0.601	0.002	$6.7e^{-07}$
RPS7	0.213	0.072	0.116	0.599	0.004	$4.69e^{-05}$
NUP133	0.195	0.078	0.125	0.602	0.002	$7.17e^{-05}$
PTGER3	0.221	0.068	0.11	0.6	0.008	$4.73e^{-05}$
JAK1	0.195	0.076	0.135	0.593	0.007	$7.93e^{-05}$
DEPDC1	0.212	0.073	0.118	0.597	0.004	$9e^{-05}$
MIR186	0.221	0.068	0.111	0.6	0.007	$4.73e^{-05}$
ZRANB2	0.221	0.068	0.111	0.6	0.007	$4.73e^{-05}$
ABCB10	0.194	0.078	0.125	0.603	0.003	$7.17e^{-05}$
MIR101-1	0.194	0.075	0.133	0.598	0.005	$7.93e^{-05}$
ACTA1	0.195	0.078	0.125	0.602	0.002	$7.17e^{-05}$
PLD5	0.21	0.075	0.122	0.594	0.003	$3.46e^{-05}$
AK3L1	0.194	0.076	0.133	0.598	0.004	$7.93e^{-05}$
RPS29	0.19	0.076	0.127	0.607	0.005	$7.93e^{-05}$
SGIP1	0.197	0.077	0.124	0.602	0.001	$3.49e^{-05}$
MIR3671	0.194	0.075	0.133	0.598	0.005	$7.93e^{-05}$
AK4	0.194	0.076	0.132	0.598	0.004	$7.93e^{-05}$
NLRP3	0.197	0.085	0.121	0.597	0.002	$7.23e^{-05}$
ZRANB2-AS1	0.221	0.068	0.11	0.6	0.008	$4.73e^{-05}$
ZRANB2-AS2	0.221	0.069	0.112	0.599	0.008	$4.73e^{-05}$
GPR177	0.212	0.071	0.119	0.598	0.003	$4.69e^{-05}$
RPE65	0.214	0.073	0.117	0.596	0.004	$4.69e^{-05}$

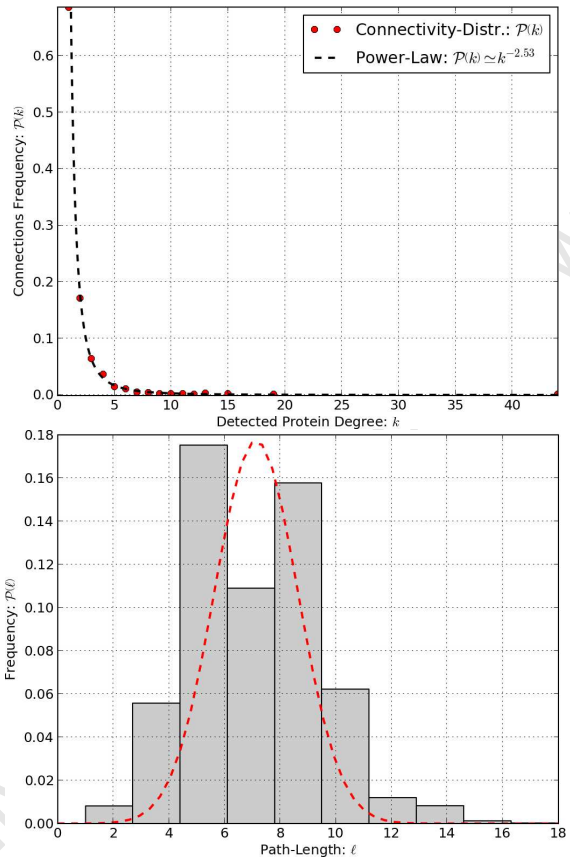

Figure 8.2: **A topological analysis of properties of the network, showing the probability distribution of the connectivity in the network and distribution of path lengths.**

Table 8.3: 20 **top significant sub-networks obtained from the simulation data of a 4-way admixed population using ancGWAS.**

95%CI	Score	chi^2_D	CEU	CHB	GIH	SIM	Sub-network List
(0.04, 0.08)	91.324	−0.003	0.211	0.078	0.114	0.597	*DISC1, CEP170, MACF1, GNB1, CCDC24, SRGAP2, DISC1, CCDC141, KIFAP3, PDE4B*
(0.04, 0.09)	92.952	0.001	0.203	0.081	0.114	0.601	*HSPA8, STMN1, PPP1R12B, CCT3, HSPA8, HSP90AA1, RGS2, IKBKE, GOT2, TNFRSF1B, HSPBP1*
(0.05, 0.10)	99.015	0.001	0.206	0.083	0.11	0.601	*PTPRC, RNF11, PLK3, FCGR3A, LSM1, LEPR, CD247, PTPRC, TIE1, NTRK1, SLAMF1, LCK*
(0.05, 0.10)	102.114	0.006	0.198	0.083	0.113	0.606	*GNAI3, RGS16, PTPRU, CD48, S1PR1, RGS18, RGS19, RGS5, RGS7, RGS2, GPSM2, GNAI3*
(0.05, 0.10)	102.665	−0.005	0.212	0.079	0.114	0.595	*TNFRSF14, EIF3I, TRAF3, TRAF5, SPCS2, DHX9, TNFRSF14, PFDN2, ST13, CNIH4, SSB, GCLM, TARDBP*
(0.05, 0.11)	103.803	−0.004	0.206	0.085	0.113	0.596	*HNRNPA1, MRPL37, MOV10, PABPC4, HNRNPR, HNRNPA1, RPL21, YTHDF2, CAPN2, SUFU, TTF2, IGF2BP2, TARDBP*
(0.06, 0.12)	110.472	−0.007	0.211	0.083	0.113	0.593	*UBQLN4, STMN1, RNF11,NOTCH2NL, EEF1A1, QSOX1, CYB5R1, UBQLN4, GPX7, SCMH1, GABRD, MDM2, ATPIF1, PBXIP1, NPPA*
(0.06, 0.12)	110.975	−0.008	0.21	0.083	0.115	0.592	*EEF1A1, KIF1B, TMSB4X, EEF1A1, NRAS, PABPC4, UBQLN4, SFN, MYOC, CRCT1, HBXIP, TP53BP2, SULT1E1, ACTB*
(0.06, 0.13)	112.805	−0.004	0.206	0.08	0.118	0.596	*EPB41, DHX9, VAMP3, S100A11, SCP2, ATP6V1E1, SRP9, EPB41, CACYBP, RPS3A, AK2, GOT2, TAGLN2, ACTB, NPPA*
(0.06, 0.13)	116.106	−0.004	0.206	0.082	0.116	0.596	*MYOC, PKLR, FUBP1, EEF1A1, OLFML3, CAP1, NOTCH2, C1QB, OLFM3, ENO1, ECE1, MYOC, ACTB*
(0.06, 0.14)	116.975	−0.004	0.206	0.084	0.114	0.596	*TNFRSF1B, RPS27, TNFRSF1B, PHGDH, RPS27L, HSPA8, HAX1, HNRNPU, DDOST, ATP1A1, ATAD3A, KRT18, HSPA6, DBT, HIVEP3*
(0.07, 0.15)	122.634	−0.006	0.211	0.082	0.113	0.594	*LCK,PTPN22,CD48,CD55,KHDRBS1,NFKBIA, FCGR3A,PTPRF,SH2D2A,CD247,PTPRC,ADAM15,CSF3R,FASLG,LCK*
(0.07, 0.15)	125.201	−0.005	0.207	0.08	0.117	0.595	*SFN,ERRFI1,ILDR2,PI4KB,ARHGEF16,CGN,RALGPS2,EEF1A1,HNRNPU,SFN,PIK3C2B,PKP3,MARK1,LAD1,MDM4*
(0.08, 0.17)	134.555	−0.007	0.213	0.081	0.113	0.593	*SETDB1,HDAC1,HIST3H3,SNIP1,OLFML3,PABPC4,PPP1R8,SETDB1,TPI1,HIST2H3D,HIST2H3C,S100A10,GIPC2,PRKRA,CLSTN1,KDM1,TARDBP*
(0.08, 0.18)	139.946	−0.002	0.206	0.082	0.114	0.598	*ACTB,NCF2,CLIC4,RAB4A,TMSB4X,EEF1A1,TPM3,HNRNPU,CAP1,PFN1,CAPZA1,S100A11,MYOC,ACTB,EPB41,LMOD1,LMNA*
(0.09, 0.20)	151.471	−0.009	0.207	0.085	0.116	0.591	*HDAC1,HDAC1,RERE,HDAC3,TAL1,PIAS3,MIER1,PEX14,RAP1A,RBBP4,SPEN,RUNX3,KDM1,H3F3A,NR0B2,GATAD2B,TXNIP,ARID4B,CDC20,NPM1,SETDB1*
(0.10, 0.21)	156.821	−0.004	0.212	0.079	0.113	0.596	*ACTA1,ACTA1,KLHL20,TMSB4X,MACF1,TPM3,MIB2,SPTA1,PFN1,NEXN,MINPP1,TNNI1,S100A4,TRIM63,S100A1,TNNI3K,ADSS*
(0.12, 0.26)	181.453	−0.001	0.209	0.082	0.11	0.599	*SHC1, MAPKAPK2, ITGB3, PPAP2B, DDR2,FCGR2B,MPL,PEAR1,EPHA2,NTRK1,PIK3C2B,CD247,TPR,FCGR1A,CSF3R,FCGR3A,FCGR2A,NPM1,VAV3,SHC1,KRT18*
(0.305, 0.76)	338.988	−0.007	0.209	0.081	0.116	0.593	*IKBKE, CAPZB, CTPS, ADSS, RPL23A, DSTYK, HSPA8, RPL18A, PSMD2, FH, MRPS14, ST13, IKBKE, CACYBP, VAMP3, PGD, RBM8A, TPM3, RPL22, YARS, EPRS, ATP5F1, PFDN2, CRYZ, SIKE1, PABPC4, NCDN, NASP, PARP1, TPD52L2, RHOC, AKR1B1, SRM,NPM1,TAGLN2,SEC22B,CAPZA1,SDHB,BPNT1,PTGES3,AK2,RPL31,RPS3A,DLST,PSMB4,SSB*

The overlapping of each sub-network was computed, and these scored sub-networks were subjected to a permutation over 1000 using Gaussian noisy data generated through a bootstrap method, to assess the confidence, and to make sure that the score of a module did not occur by chance. Finally, 20 sub-networks (containing 295 genes) were significant and ranked by score and confidence interval (Table 8.3). Table 8.3 also provides the ancestral proportions per sub-network, which is still consistent with the ancestral proportion used in the simulation. The chi^2 statistic displayed in Table 8.3 also shows no evidence of unusual difference in a deficiency/excess of ancestry for each those top 20 sub-networks. In Figure 8.3, we display the 20 top sub-networks, but excluding those genes with less than two edges.

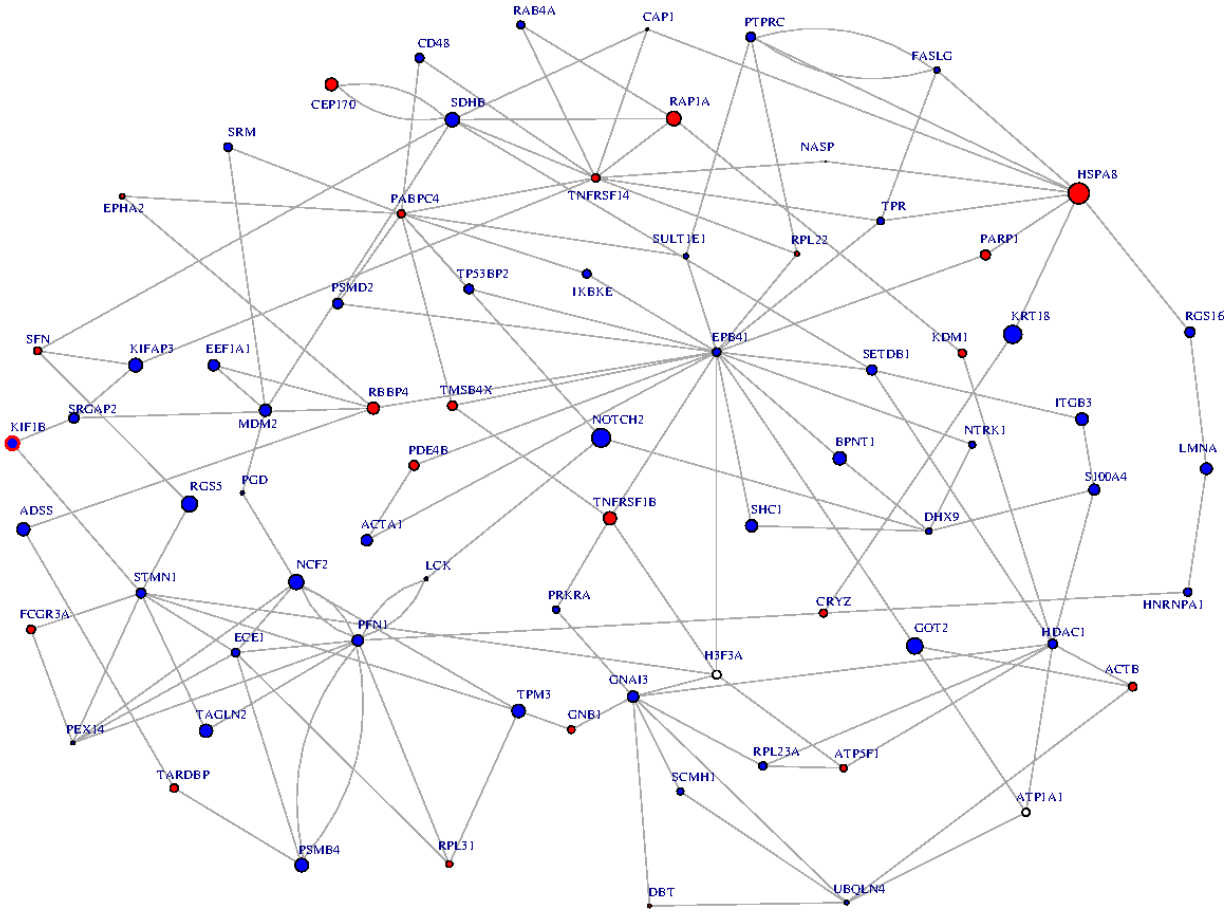

Figure 8.3: **Top** 20 **ranked sub-networks from the simulation data, enriched for disease risk in the simulated data and highly connected sub-networks of** < 295 **connected genes. The size of a node denotes its significance from small to large. The blue nodes show no signal of unusual difference in an excess/deficiency of ancestry, while and the red nodes have moderate signal.**

Importantly, when applying EnrichNet-Network-based enrichment analysis (ENRICH-NET) (Glaab *et al.*, 2012) to the top 20 sub-networks (Table 8.3), the annotations for pathway/process of these 20 top sub-networks clusters them into signaling pathways. The Adipocytokine signaling pathway is associated with our simulated disease genes (*IL23R*, *SLC2A1*). This result highlights the benefit of fully characterizing the susceptible genes beyond standard GWAS for analysis of the genetic structure of diseases. Taken together, through the simulation of a 4-way admixed population, we demonstrated the accuracy of ancGWAS and its ability to examine the interactions between genes underlying the pathogenesis of complex diseases from a standard GWAS, as well as gene or sub-network-specific ancestry and to detect possible unusual differences in a deficiency/excess of ancestry of SNPs and at both gene and pathway level.

8.3.2 Application of ancGWAS to the TB GWAS Dataset from the South African Coloured Population

Taking into consideration the GWAS of TB in the SAC using typed and imputed SNPs conducted in sections 5.3 and 6.3, here we aim to address the moderate risk SNPs that did not reach the intrinsic genome-wide significance cut-off of p-value $< 5 \times 10^{-8}$. To address this we combine the effects of all SNPs within a particular gene, and of all genes at the pathway level using ancGWAS in order to characterize the susceptible genes and the genetic structure of TB risk. Similarly to the simulated data shown in section 8.2.4 above, we accounted for the advantage of linkage disequilibrium in the SAC, and combined the TB imputation GWAS data set with the estimated locus-specific ancestry in the SAC into a comprehensive human PPI network weighted by linkage disequilibrium. The estimation of locus-specific ancestry in the SAC was conducted in 5-way admixture using SAN (all merged Khoesan populations), CEU, YRI, GIH and CHB in order to increase the ancestral haplotype samples and to account for the current limitation of LampLD in inferring local ancestry in multi-way admixture. Using the method described in ancGWAS, in particular the Fisher's method, we computed the summary p-value of multiple SNPs assigned to a gene. Combining the signal of SNPs within a gene and accounting for linkage disequilibrium that exists within and between genes, the results in Table 8.4 display 95 moderate/significant genes from the ancGWAS analysis. Six of the genes, including *MEGF10* ($p = 2.44e^{-11}$), *PRRC1* ($p = 2.44e^{-11}$), *HNRNPK* ($p = 6.28e^{-09}$), *SLC8A3* ($p = 8.99e^{-09}$), *SMOC1* ($p = 8.99e^{-09}$) and *CTXN3* ($p = 2.30e^{-08}$) are significantly associated with TB (Table 8.4). Interestingly, our results also (Table 8.4) replicated known associated TB genes such as *IL8* ($p = 0.0039$), *SLC11A1* ($p = 0.0035$), *WT1* ($p = 0.0015$), *CCL2* ($p = 0.0015$) and *IFNGR1* ($p = 0.0034$).

Table 8.4: 95 **genes with significant/moderate p-values obtained from the ancGWAS method of combined GWAS based SNPs association analysis. The table displays gene-specific ancestry from each ancestral population. The header χ^2_D denotes the χ^2 of unusual difference in an excess/deficiency of ancestry.**

GENE	SAN	YRI	CEU	GIH	CHD	χ^2_D	P
MEGF10	0.885	0.023	0.082	0.044	0.001	0.071	$2.44e^{-11}$
PRRC1	0.981	0.012	0.002	0.005	0.0	0.013	$2.44e^{-11}$
HNRNPK	0.0	0.245	0.376	0.215	0.02	0.024	$6.28e^{-09}$
SLC8A3	0.959	0.012	0.017	0.012	0.001	0.01	$8.99e^{-09}$
SMOC1	0.952	0.012	0.024	0.012	0.001	0.01	$8.99e^{-09}$
CTXN3	0.862	0.064	0.057	0.031	0.0	0.013	$2.30e^{-08}$
C2CD2	0.928	0.034	0.021	0.013	0.005	0.018	$1.58e^{-07}$
RHOU	0.0	0.016	0.5	0.484	0.0	0.052	$1.8e^{-07}$
RNF187	0.496	0.059	0.424	0.244	0.0	0.052	$1.8e^{-07}$
TRIM17	0.0	0.016	0.5	0.484	0.0	0.052	$1.8e^{-07}$
CNOT6L	0.48	0.049	0.13	0.337	0.003	0.027	$2.9e^{-07}$
CXCL13	0.427	0.049	0.116	0.3	0.003	0.027	$2.9e^{-07}$
ALLC	0.0	0.02	0.5	0.48	0.0	0.05	$7.07e^{-07}$
SOX11	0.0	0.022	0.5	0.478	0.0	0.05	$7.07e^{-07}$
CEP170	0.344	0.035	0.384	0.27	0.002	0.055	$7.58e^{-07}$
PLD5	0.0	0.009	0.5	0.491	0.0	0.055	$7.58e^{-07}$
RPL41	0.983	0.016	0.074	0.002	0.0	0.013	$7.58e^{-07}$
DSCAM	0.922	0.031	0.028	0.015	0.006	0.018	$2.4e^{-06}$
CYP2C19	0.993	0.005	0.0	0.001	0.001	0.012	$2.81e^{-06}$
CYP2C8	0.984	0.005	0.001	0.002	0.001	0.012	$2.81e^{-06}$
ZFPM2	0.776	0.038	0.088	0.058	0.003	0.01	$3.09e^{-06}$
CLUAP1	0.927	0.044	0.016	0.012	0.002	0.017	$3.13e^{-06}$
NAA60	0.944	0.033	0.018	0.013	0.002	0.017	$3.13e^{-06}$
NLRC3	0.927	0.044	0.016	0.012	0.002	0.017	$3.13e^{-06}$
ZNF174	0.927	0.044	0.016	0.012	0.002	0.017	$3.13e^{-06}$
ZNF434	0.927	0.044	0.016	0.012	0.002	0.017	$3.13e^{-06}$
ZNF597	0.937	0.038	0.017	0.012	0.002	0.017	$3.13e^{-06}$
C6orf195	0.982	0.014	0.001	0.001	0.001	0.009	$3.87e^{-06}$
GMDS	0.982	0.012	0.001	0.002	0.003	0.01	$3.87e^{-06}$
LOC100508120	0.983	0.015	0.073	0.055	0.003	0.01	$3.87e^{-06}$

Continued on next page

Table 8.4 – continued from previous page

GENE	SAN	YRI	CEU	GIH	CHD	χ^2_D	P
ADAMTS19	0.934	0.023	0.103	0.087	0.002	0.014	$1.87e^{-05}$
E2F7	0.994	0.038	0.0	0.034	0.0	0.008	$4.46e^{-06}$
MIR4435-1	0.004	0.165	0.498	0.326	0.007	0.041	$1.79e^{-05}$
MIR4435-2	0.004	0.165	0.498	0.326	0.007	0.041	$1.79e^{-05}$
RGPD5	0.0	0.19	0.5	0.301	0.009	0.043	$1.79e^{-05}$
NAV3	0.995	0.002	0.001	0.0	0.002	0.011	$4.46e^{-06}$
VWA8	0.935	0.008	0.0	0.0	0.0	0.008	$4.72e^{-06}$
VWA8-AS1	0.993	0.007	0.0	0.0	0.0	0.008	$4.72e^{-06}$
GYG1	0.133	0.058	0.429	0.367	0.013	0.053	$1.06e^{-05}$
USP24	0.0	0.004	0.5	0.468	0.028	0.053	$1.056e^{-05}$
ACOXL	0.0	0.177	0.5	0.313	0.01	0.041	$1.79e^{-05}$
BCL2L11	0.0	0.188	0.5	0.302	0.01	0.045	$1.79e^{-05}$
NCKAP5	0.016	0.301	0.492	0.184	0.0	0.094	$2.37e^{-05}$
PTPRQ	0.992	0.003	0.001	0.001	0.003	0.008	$2.42e^{-05}$
RPL7	0.835	0.031	0.061	0.043	0.013	0.014	$2.42e^{-05}$
LINC00571	0.963	0.039	0.145	0.05	0.001	0.014	$2.68e^{-05}$
TRPC4	0.988	0.01	0.001	0.0	0.001	0.01	$2.68e^{-05}$
UFM1	0.936	0.031	0.003	0.03	0.0	0.017	$2.68e^{-05}$
PLCL1	0.0	0.274	0.5	0.226	0.0	0.077	$3.079e^{-05}$
SATB2	0.0	0.265	0.5	0.235	0.0	0.074	$3.079e^{-05}$
FSTL5	0.484	0.078	0.349	0.089	0.0	0.1	$3.16e^{-05}$
RAPGEF2	0.479	0.083	0.343	0.096	0.0	0.098	$3.16e^{-05}$
CLEC14A	0.973	0.024	0.001	0.0	0.001	0.011	$3.50e^{-05}$
SEC23A	0.975	0.023	0.001	0.0	0.001	0.013	$3.50e^{-05}$
UBA6	0.484	0.037	0.091	0.388	0.0	0.032	$3.63e^{-05}$
FABP3	0.0	0.001	0.5	0.465	0.033	0.053	$3.83e^{-05}$
SERINC2	0.0	0.001	0.5	0.465	0.033	0.053	$3.83e^{-05}$
TINAGL1	0.0	0.001	0.5	0.465	0.033	0.053	$3.83e^{-05}$
CSMD1	0.926	0.018	0.028	0.028	0.001	0.012	$4.00e^{-05}$
NAT1	0.98	0.012	0.002	0.006	0.0	0.013	$4.01e^{-05}$
NAT2	0.981	0.011	0.002	0.006	0.0	0.013	$4.01e^{-05}$
IMMP2L	0.988	0.008	0.002	0.003	0.0	0.012	$4.1e^{-05}$
PPP6C	0.932	0.049	0.004	0.014	0.001	0.017	$4.83e^{-05}$
SCAI	0.932	0.049	0.004	0.015	0.004	0.018	$4.83e^{-05}$
UBXN2B	0.942	0.034	0.001	0.021	0.001	0.014	$6.62e^{-05}$

Continued on next page

Table 8.4 – continued from previous page

GENE	SAN	YRI	CEU	GIH	CHD	χ^2_D	P
KIAA0564	0.993	0.007	0.0	0.0	0.0	0.008	$5.80e^{-05}$
HNF4G	0.921	0.07	0.001	0.008	0.0	0.018	$5.81e^{-05}$
FAM110B	0.942	0.034	0.001	0.021	0.001	0.014	$6.62e^{-05}$
ZFHX4	0.924	0.065	0.002	0.009	0.0	0.018	$5.81e^{-05}$
IGSF21	0.131	0.006	0.402	0.382	0.027	0.053	$5.91e^{-05}$
DAOA	0.971	0.027	0.0	0.001	0.001	0.015	$6.21e^{-05}$
SLC10A2	0.988	0.012	0.0	0.0	0.001	0.01	$6.21e^{-05}$
NUCKS1	0.0	0.045	0.5	0.445	0.01	0.045	$6.53e^{-05}$
RAB7L1	0.0	0.045	0.5	0.445	0.01	0.045	$6.53e^{-05}$
LRP1B	0.0	0.458	0.5	0.043	0.0	0.132	$6.98e^{-05}$
GZMB	0.991	0.004	0.0	0.003	0.002	0.012	$7.40e^{-05}$
STXBP6	0.988	0.006	0.0	0.003	0.003	0.013	$7.40e^{-05}$
PM20D1	0.0	0.045	0.5	0.445	0.01	0.045	$7.51e^{-05}$
SLC41A1	0.0	0.045	0.5	0.445	0.01	0.045	$7.51e^{-05}$
SLC45A3	0.0	0.045	0.5	0.445	0.01	0.045	$7.51e^{-05}$
PABPC1	0.96	0.025	0.002	0.01	0.003	0.018	$7.67e^{-05}$
SNX31	0.916	0.045	0.005	0.025	0.008	0.018	$7.67e^{-05}$
FAM178A	0.968	0.023	0.0	0.01	0.0	0.012	$7.68e^{-05}$
PAX2	0.967	0.023	0.0	0.01	0.001	0.014	$7.68e^{-05}$
FMN1	0.977	0.021	0.0	0.002	0.0	0.011	$8.26e^{-05}$
ANAPC1	0.0	0.201	0.5	0.292	0.007	0.041	$8.52e^{-05}$
LOC541471	0.0	0.211	0.5	0.28	0.01	0.052	$8.52e^{-05}$
PAFAH1B1	0.852	0.048	0.112	0.043	0.001	0.019	$8.52e^{-05}$
KCNMA1	0.984	0.01	0.001	0.006	0.0	0.012	$9.61e^{-05}$
CCL2	0.936	0.031	0.005	0.027	0.001	0.017	0.0015
WT1	0.934	0.032	0.019	0.012	0.003	0.017	0.0015
IFNGR1	0.975	0.01	0.001	0.005	0.01	0.011	0.0034
SLC11A1	0.0	0.113	0.5	0.387	0.0	0.036	0.0035
IL8	0.486	0.063	0.097	0.352	0.001	0.028	0.0039

We examined the signal of unusual difference in a deficiency/excess of ancestry, but the reported χ^2 values in Table 8.4 indicate no significant signals, which is consistent with the hypothesis that the admixture events to create the SAC have occurred too recently for differential

deficiency/excess of ancestry to have had a significant impact on its ancestry proportions. In addition, associated gene-specific ancestry proportions from each ancestral population are displayed in Table 8.4 and plotted in Figure 8.4 for the significant/moderately associated genes in the SAC. The results indicate high ancestry proportion from African ancestral populations associated with the susceptibility genes, despite the fact that the χ^2 yielded a weak signal of unusual difference in an excess of ancestry at the susceptibility genes (Table 8.4).

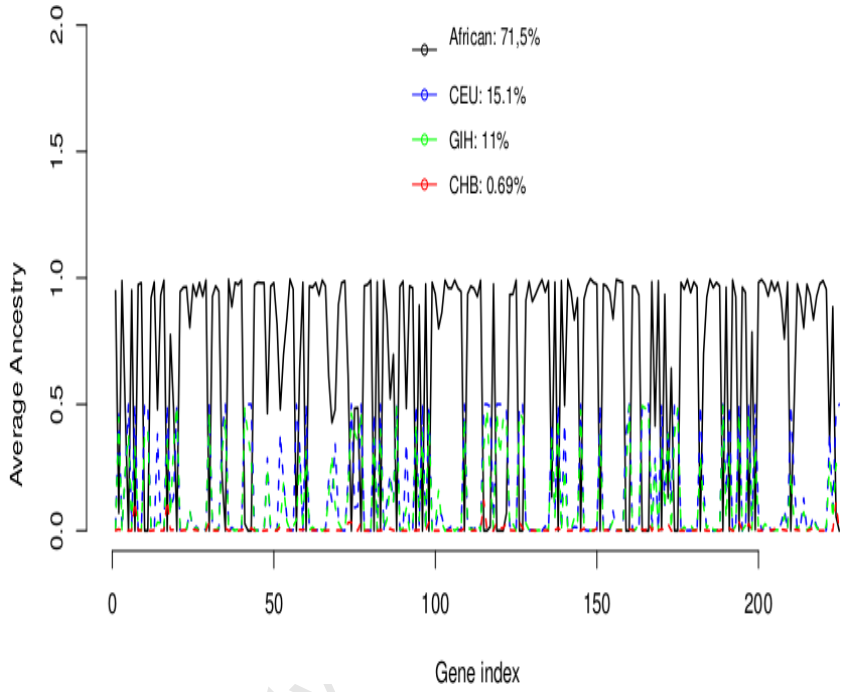

Figure 8.4: **Admixture proportions for significant/moderately associated genes. The genome-wide average of gene-specific ancestry in the SAC is predominately African. The average ancestral population proportions are African (71.5%), European (15.1%), Indian (11%) and Asian (0.69%) related ancestral population, respectively.**

We mapped genes with TB-associated p-values and their ancestry proportions, into a network weighted by linkage disequilibrium (see method in section 8.2.3). After analysing the resulting network of $46,955$ pair-wise gene interactions, we determined that the spread of information can be achieved through 4.01 steps, which corresponds to the average shortest path lengths in this network. Following our clustering algorithm 2, ancGWAS analyses all topological properties to break down the constructed weighted network into sub-networks. We determined all the hubs of the networks, and the betweenness centrality, closeness centrality and eigenvector centrality measures for each gene. We computed the cut-offs for each centrality measure, and the intersection of the resulting sets that were above the cut-off were considered to be the set of central genes. Using the first step in searching for sub-networks, a total of 525 sub-networks were obtained with

13 gene hubs. We assessed the significance of each sub-network using the sub-network statistical Fisher's scoring method in ancGWAS, and retained the 20 top highly scoring sub-networks (Table 8.5). Table 8.5 provides the ancestral proportions per sub-network, showing a dominant African ancestry proportion, but the χ^2_D statistic displays in Table 8.5 shows no significant evidence of unusual difference in a deficiency/excess of ancestry at the sub-network level.

Table 8.5: **Top 20 sub-networks associated with moderate/significant statistical score obtained using ancGWAS method by combining the gene associated p-values. The table displays ancestry-specific interaction for sub-networks from each ancestral population. The header χ^2_D denotes the χ^2 of unusual difference in an excess/deficiency of ancestry. The final column displays top annotation pathway obtained from EnrichNet-Network-based enrichment analysis (ENRICH-NET) (Glaab et al., 2012) .**

95%CI	nScore	Zscore	χ^2_D	African	CHB	GIH	CEU	Sub-network lists	Pathway
(0.009, 0.01)	0.01	951.25	−0.207	0.708	0.043	0.135	0.104	*MBP, TEP1, TRIM29, AKAP5, RPL10, ATP2B1, ADRA1B, NFATC1, HMGN2, PRKG1, BTG2, MARCKS, GRIA1, DGKZ, FAS, RGS7, RGS2, ANXA2, PRKCE*	Salivary secretion
(0.009, 0.01)	0.01	952.508	−0.226	0.772	0.043	0.103	0.073	*CSNK2A1, APEX1, GPI, PAFAH1B1, CSNK2A1, TCF7L2, XRCC4, HSPH1, FAF1, SET, HDAC2, HMGA2, HMGA1, ABCA1, MME, HSP90AA1, IL8, HNRNPA2B1, EEF1B2, PTPRC*	NOD-like receptor signaling pathway
(0.01, 0.011)	0.01	988.872	−0.224	0.77	0.039	0.104	0.082	*SMAD3, ARHGEF7, MAP3K7, MAGI2, RUNX2, RUNX3, PARD3B, RASD2, RUNX1, EPAS1, SMAD3, ZBTB16, HMGA2, PARD3, RPLP0, HIVEP1, FOXO1, GLI3, RGS3, DACH1*	Acute myeloid leukemia
(0.011, 0.012)	0.011	1047.554	−0.232	0.778	0.037	0.095	0.074	*ISL1, PTMA, GRIP1, PAK6, RGS3, CHD9, UBE3A, RNF4, PRDM2, CCND1, SMAD3, GNAI1, ZBTB16, FHL2, NFKB1, RXRA, FOXO1, SOS1, TDG, PSMB9, HMGB1, SMARCA2, XBP1*	Acute myeloid leukemia
(0.011, 0.012)	0.011	1048.017	−0.196	0.673	0.046	0.152	0.114	*LRPPRC, TSFM, GOT2, KCTD12,* UBA2, MTHFD1, STRN3, EEF1B2, *DDOST, CUTA, RPL23A, PFKP, APEX1, ANXA2, ESD, NPM1*	Phenylalanine metabolism
(0.011, 0.012)	0.011	1051.657	−0.204	0.697	0.039	0.14	0.109	*MIF, OTUD7A, PTMA, FKBP1A, CCT3,* RNF139, EPAS1, HDAC2, ACP1, HINT1, *SET, RCC2, CUTA, RPL23A, DGKZ, DGKI, HNRNPA2B1, ANXA2, MTPN*	Phenylalanine metabolism
(0.011, 0.012)	0.011	1063.969	−0.202	0.682	0.048	0.15	0.114	*ARHGEF7, DCC, CBLB, MYRIP, NCKAP5, PKN2, RHOU, FLNB, SNX7, CAST, CELSR2, SOS1, ID4, KDR, CYFIP2, P2RX7, SASH1*	Chronic myeloid leukemia
(0.011, 0.012)	0.011	1070.501	−0.215	0.704	0.048	0.14	0.104	*CBLB, FGFR2, MYRIP, CD2AP, TULP4, SHB, FLNB, RET, CAST, SOS1, ID4, IRS2, MME*	Bacterial invasion of epithelial cells
(0.011, 0.012)	0.011	1076.847	−0.196	0.719	0.044	0.127	0.105	*OSBPL3, YWHAE, RPS2, RASSF8, TAF15, RPL3, FOXO1, KCNK15, CEP170, SMCR7L, SAMD4A, TBC1D4, KRTAP19-5, RAB11FIP2, IRS2, CYFIP2, EEF1A1*	Insulin signaling pathway
(0.011, 0.012)	0.012	1092.394	−0.225	0.761	0.032	0.115	0.092	*FHIT, PKP2, RAPGEF2, CDH9, CDH8, MAGI2, CDH5, CDH7, CSNK2A1, TCF7L2, RUNX3, AJAP1, CCND1, ACP1, SMAD3, CTNND2, FHL2, NFKB1, KDR, SPN, FER, CDH11, CDH18, FOXO1, PYGO1, PTPRC,*	Adherens junction

Continued on next page

Table 8.5 – continued from previous page

95%CI	nScore	Zscore	χ^2_D	African	CHB	GIH	CEU	Sub-network lists	Pathway
								RXRA,CTNNA3,PARD3,PTPRG	
(0.011, 0.012)	0.012	1097.011	−0.231	0.789	0.042	0.094	0.076	*IFNAR2,CAMK4,KLF4,WT1,* *DACH1,GATA2,CSNK2A1,PTMA,* *RUNX1,ACTA2,TDG,DAXX,PAX5,* *ING1,MAF,ONECUT1,SMAD3,FHL2,* *CITED2,SND1,ABCA1,FOXO1,* *KLF13,ETS2,GLI3,ZBTB2,* *SMARCB1,HMGA1,SERTAD2,E2F3*	Chronic myeloid leukemia
(0.011, 0.012)	0.012	1105.579	−0.195	0.683	0.042	0.149	0.121	*GAPDH,KRT8,VAV2,MAP2K1,* *CBLB,KRT7,SH3GL2,ALCAM,* *ITGA5,FER,SOS1,CD59,FAS,* *DOK5,NRG1,PTPRC,NCK2*	T cell receptor signaling pathway
(0.012, 0.014)	0.013	1177.896	−0.213	0.711	0.042	0.133	0.106	*ST5,HCN4,NCKAP5,CD2AP,* TULP4,TERF1,LRBA,GPX1, *CTNND2,CAST,EFNA5,SOS1,* *AHSG,PRDX1,P2RX7,ROBO1,* *DAAM1,MBP,YWHAE*	Dorso-ventral axis formation
(0.014, 0.015)	0.014	1280.19	−0.21	0.747	0.043	0.111	0.089	*IFNAR2,APEX1,ZFPM2,PTMA,* *CCND1,TCF7L2,RUNX2,RUNX3,* *RUNX1,NFATC1,HMGN2,TDG,MN1,* *PAX6,ZBTB16,MAF,ACTA2,* *SMAD3,NEDD1,FHL2,ING1,* *CITED2,NR2F2,SET,MAP2K1,* *ETS2,MRE11A,EPAS1,MEF2D*	Acute myeloid leukemia
(0.014, 0.016)	0.015	1318.33	−0.201	0.705	0.044	0.136	0.105	*DYNLL1,LRPPRC,GOT2,PABPC1,* *PFKP,KCTD12,MIF,RPL3,* *HINT1,MTPN,UBA2,MTHFD1,* *PREP,ADSS,SET,SEC23A,* *RPL23A,SND1,MAP2K1,DDOST,* *DAD1,APEX1,PSMC1,ANXA2,* *ESD,EEF1B2,NPM1,RCC2*	Phenylalanine metabolism
(0.016, 0.018)	0.017	1459.521	−0.198	0.684	0.048	0.148	0.112	*LRPPRC,GLRX3,ADSS,ZC3H15,* *EEF1B2,MIF,CYLD,PABPC1,* *RPL3,UBA2,HDAC2,ACP1,* *SET,SND1,RPL23A,DAD1,* *ANXA2,MTPN,NPM1,RPL36*	Phenylalanine metabolism
(0.017, 0.018)	0.017	1475.243	−0.201	0.684	0.044	0.146	0.116	*CD36,SPN,ITK,HCN4,CD2AP,* *SKAP2,NCKAP5,CD48,TULP4,* *HSP90AA1,LRBA,SLAMF1,ACP1,* *CTNND2,CAST,HNRNPK,KDR,* *SOS1,CBLB,FAS,PRKCE,* *PTPRC,RPL10*	T cell receptor signaling pathway
(0.021, 0.023)	0.022	1735.384	−0.209	0.701	0.046	0.137	0.101	*ASXL2,MIF,LRPPRC,MARCKS,* MBP,KCTD12,VTA1,CYLD, *PABPC1,MAP3K7,RUNX1,* *BUB3,UBE2E1,UBA2,FLNB,* *SET,RPL23A,RCC2,FHL2,* *MTHFD1,MAP2K1,SEPT9,PREP,* *IRF8,PFKP,APEX1,PSMC1,* *ANXA2,MTPN,ESD,NPM1,*	Phenylalanine metabolism

Continued on next page

Table 8.5 – continued from previous page

95%CI	nScore	Zscore	χ^2_D	African	CHB	GIH	CEU	Sub-network lists	Pathway
								MAP3K7IP2,RPL36	
(0.023, 0.026)	0.025	1908.168	−0.202	0.697	0.046	0.144	0.108	*ASS1,CCT3,KCNK15,ING1, NUFIP1,EEF1A1,PARD3,TSFM, ADRA2A,NFATC1,CEP170,RFC4, RPLP0,SET,HNRNPA1,IRS2, RPL10A,PFKP,YWHAE,PANK1, PRKCE,RAPGEF2,WDR61,RPL19, GAPDH,HSPH1,HSP90AA1,PDE3A, PRDX1,ATL2,PPIA,FOXO1,RGS3, ANXA2,TBC1D4,LRPPRC,WWC1, HSP90AB1,RPL6,HNRNPK,ARL6IP1, EEF1B2,CAND1,LDHA,NPM1*	Pantothenate and CoA biosynthesis
(0.025, 0.027)	0.026	2004.695	−0.203	0.69	0.048	0.143	0.111	*RNF10,AGT,CBLB,WDR1,HMGN2, SHB,FLNB,RET,SMAD3,AHSG,ST5, CUGBP2,SNX7,MYRIP,CCL5,ITK, IRS2,DNAJB11,KRT8,CAST,KDR, CUTA,MAP2K5,IK,KRT7,NCKAP5, RHOU,VAV2,KCNB2,SOS1,ID4, CD59,P2RX7,ESD,NPM1*	Chronic myeloid leukemia

Using EnrichNet-Network-based enrichment analysis (ENRICH-NET) (Glaab *et al.*, 2012), the most common pathway/process annotations of the top 20 sub-networks are acute or Chronic myeloid leukemia. Considering only genes with p-value < 0.0004, we plotted the 20 top sub-networks in Figure 8.5. The following genes are the central hubs *HNRNPK* ($p = 6.283310622e-09$), *RHOU* ($p = 1.8e-07$), *GRIA1* ($p = 0.0002$), *PAFAH1B1* ($p = 8.56e-05$), *PABPC1* ($p = 7.67e-05$), *NPM1* ($p = 0.0001$), *PRDX1* ($p = 0.0001$), *GLI3* ($p = 0.00014$), *WT1* ($p = 0.0015$), *EPAS1* ($p = 0.0002$), *HNRNPA1* (p=0.0002), *CDH5* ($p = 0.0002$) and *YWHAZ* ($p = 0.0071$). Since these 20 sub-networks overlap and the hubs are connected to each other, we searched for the most important and central sub-network within the network in Figure 8.6 by excluding those genes with less than three edges. Figure 8.6 is the most important sub-network found and contains relevant novel and previously associated TB genes, such as *WT1* and *IL8*.

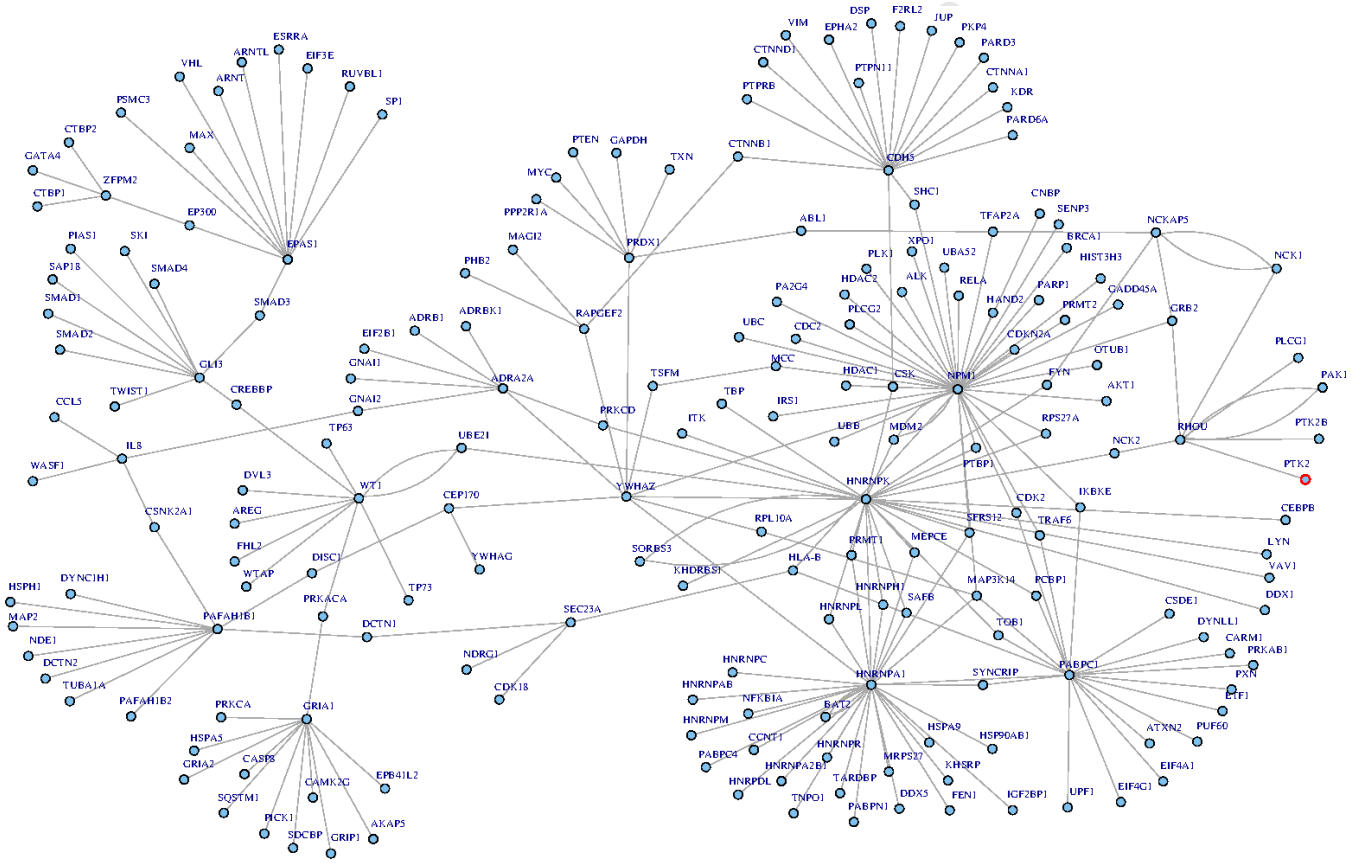

Figure 8.5: **Relevant sub-networks from TB imputation GWAS of South African Coloured population, including enriched and highly connected sub-networks of moderate or significant genes.**

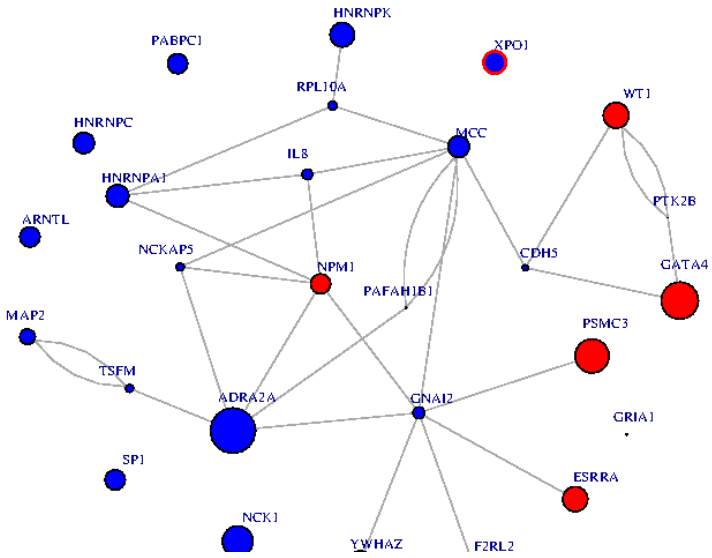

Figure 8.6: **Central sub-network from TB imputation GWAS of South African Coloured population. In Figure 8.6, the size of a node denotes its significance from small to big size, while the blue colour denotes no signal of unusual difference in excess/deficiency of ancestry and the red colour is a moderate signal.**

8.3.3 Summary

In summary, we introduced ancGWAS, a post GWAS method for recently admixed or non-admixed populations, that integrates the association signal from GWAS data sets, the local ancestry and gene pair-wise linkage disequilibrium into the human protein-protein interaction network. In addition, our method accounts for the correlation that exists between SNPs within a gene and genes within pathways and introduces flexibility in estimating gene-specific and sub-network-specific ancestry, and tests for signals of unusual difference in an excess/deficiency of ancestry. To our knowledge these new present contributions to post-GWAS methods. We validated ancGWAS through simulating interactive disease loci in an admixed population, and showed that ancGWAS holds promise for comprehensively examining the interactions between genes underlying the pathogenesis of genetic diseases and also underlying ethnic differences. Importantly, ancGWAS was able to recover and refine the signal of a simulated disease gene *SLC2A1* that was scoring on the boundary of genome-wide significance from standard GWAS (Table 8.1). We applied ancGWAS to the imputation TB GWAS data of the admixed South African Coloured populations. Our results yielded the top 20 sub-networks that are not only significantly enriched, but suggested to have a role in TB immunopathogenesis, and were predominantly African specific, although they had no statistical evidence of unusual difference in an excess of ancestry. The

enrichment-test revealed that the significant sub-networks are mostly implicated in acute and chronic myeloid leukemia pathways. Interestingly, both our gene-based and pathway-based results demonstrated the convergence of SNP signal to gene signal and from the gene signal to the 20 significant sub-networks (and to a novel central TB sub-network) of the human interactome that are enriched with interesting TB biological pathways, including genes previously identified to be associated with TB. The most notable, finding of a central sub-network in Figure 8.6 may provide further insights into TB pathogenesis and could thus facilitate drug development. In addition, the convergence of SNP signal to related TB sub-networks and candidate genes supports our hypothesis of finding significant association based on post-GWAS analysis. In particular the finding of 6 genes, including *MEGF10*, *PRRC1*, *HNRNPK*, *SLC8A3*, *SMOC1* and *CTXN3* came from combining the effect of SNPs assigned to each gene. Importantly, we were able to replicate 4 known TB associated genes, including *IL8*, *SLC11A1*, *WT1*, *CCL2* and *IFNGR1*. These genes have a lower significance thatn other listed genes Overall, although the accuracy of inference of local ancestry in multi-way admixed populations is still a challenge, here ancGWAS highlights the value of identifying the ancestry proportions of pathways associated with a disease which may allow us to discover the pathogenesis of genetic diseases and the link to ethnic differences.

Chapter 9

Discussion and Conclusion

9.1 Discussion

9.1.1 Genetic Variation in the South African Coloured Population

We introduce PROXYANC, an approach to select the best proxy ancestry for complex multi-way admixed populations. We assessed its accuracy through a simulation of a multi-way admixed population and demonstrated the impact and sensitivity of the choice of reference panel in estimating global and local ancestry and in imputing missing genotypes. Our methods to select proxy ancestral populations in a multi-way admixed population have enabled us to characterize the genetic ancestry component of the uniquely admixed Coloured population of South Africa that accounts for 49% of the population of the Western Cape Province (Statistics South Africa, Census 2011). Previous studies of this historically complex population were hampered by the relatively small sample size and few publicly available putative ancestral populations, and particularly the very low number of San individuals. In the present study we have utilized the increased number of reference populations available, and the best proxy ancestries of the South African Coloured population obtained from PROXYANC. These allowed us to document a contribution of the isiXhosa, ‡Khomani, European, Gujarati Indian, and Chinese genetic material to the South African Coloured population (33%, 31%, 16%, 12% and 7%, respectively). We expected a southern Bantu-speaking group such as isiXhosa instead of a West African group such as the Yoruba to be a better proxy ancestor of the South African Coloured population. The isiXhosa as the best proxy ancestor of the South African Coloured population reflects the early mixing of mainly indigenous San females with the Southern Bantu groups, and subsequently with male settlers, mainly from the Netherlands, Britain, Germany and France, or male slaves from South Asia (Boonzaaier *et al.*, 1996; Keegan, 1996; Mountain, 2003). The substantial number of ‡Khomani (sub-Kalahari San) individuals available for this study greatly increases our confidence in the ac-

curacy of the ancestry estimates presented here. Our results also emphasize the point that San clans are often very different from one another, and grouping San individuals from different areas together as generic San may result in a loss of discrimination at the genetic level. This was also illustrated by the deep genetic differences between individual San (Bushmen) genomes (Pickrell *et al.*, 2012; Schlebusch *et al.*, 2012; Schuster *et al.*, 2010). In the case of the South African Coloured population in the Western Cape, it is perhaps to be expected that San groups from the southern Kalahari, including ǂKhomani, Bushmen and San, which are geographically closer to the place of origin of the South African Coloured population, would be better proxy ancestors of this group than the Jul'huan from Namibia, and this is what we have shown. This also gives credence to an earlier suggestion that only some of the San peoples contributed to the South African Coloured population population (Quintana-Murci *et al.*, 2010).

A higher degree of linkage disequilibrium is expected in admixed populations, and this could at certain points of its history be influenced by population bottlenecks, or only be a result of the admixture itself. To address this, we first implemented two different algorithms to select a subset of informative markers. We used the obtained subsets of informative markers that differentiate the best proxy ancestral populations of the South African Coloured population obtained from PROXYANC algorithms to examine the pattern of linkage disequilibrium and the level of admixture linkage disequilibrium in the South African Coloured population as a result of ancestral admixture. We demonstrated that the allele frequency differences between each pair of proxy ancestral populations correlated with the degree of linkage disequilibrium in the South African Coloured population, suggesting that the admixture increased genetic diversity and that the observed linkage disequilibrium in the South African Coloured population has its origin mainly in the admixture. This study observed a weak level of founder haplotypes identical-by-descent along the genome of the South African Coloured population, which strengthens the evidence against population bottlenecks that could have been found as a consequence of the past legislated separation of ethnic groups in South Africa, including the South African Coloured population. However, in spite of this isolation the original admixed population was large and a population bottleneck is therefore unlikely. Although the accuracy of estimating both local ancestry and ancient dates of different admixture events in multi-way admixed populations is still in the exploratory stage, we estimated the length of ancestry blocks in the South African Coloured population using the inferred locus-specific ancestry from its proxy ancestral populations and we fitted a likelihood model on the length of ancestry block distribution to estimate different dates of admixture events in this population. Our result suggested the genetic make-up of the South African Coloured population arose 9 to 11 generations (385 years) ago, if we consider 35 years for one generation.

9.1.2 Genome-wide Association Study

We used a combination of two complementary methods to examine whether the genetic contribution can increase tuberculosis risk, and evaluated the contribution of socio-economic status to the ancestry-tuberculosis relationship in the South African Coloured population. Our results demonstrated significant evidence of an association between ‡Khomani ancestry and tuberculosis status that is not confounded by socio-economic status. This an important epidemiological result and illustrates the value of the inclusion of admixture association methods in the set of methods used to conduct tuberculosis association studies in this population. When the extremely high incidence of tuberculosis in the South African Coloured population population is considered, together with our finding that a significant percentage of their ancestry is derived from the San and other African populations, it appears possible that there may be an element of population level genetic susceptibility to this disease.

We conducted genome-wide association analysis of tuberculosis case-controls from the admixed South African Coloured population, resulting in the identification of a low-frequency variant at SNP $rs17175227$. After imputation we also identified a rare variant at SNP $rs12294076$ at the borderline of genome-wide significance and we moderately replicate a recently reported susceptibility locus, $rs2057178$. Because of the imperfect asymptotic distribution of mixed model association or logistic regression in the specific case of low-frequency variants, which may often reach genome-wide significance; we computed Fishers exact test values for variants that achieved the most significant mixed model association p-values. This resulted in $rs17175227$ not reaching the genome-wide cut-off. Power to detect association is a function of allele frequency and rare variants are underpowered when sample sizes are limited. However, because current mixed models or logistic regression association do not account for rare variants, we have addressed this challenge by computing Fishers exact test p-values for variants that achieve the most significant mixed model association p-values. Importantly, Fisher's exact test allowed us to demonstrate that a rare variant is not genome-wide significant although it achieved significant mixed model association p-values.

Some limitations should be noted in association analyses. Firstly, the present study is underpowered to detect risk variants of more modest effect size, because of our modest sample size. Secondly, imputing missing genotype data of a complex admixed population is an important challenge based on the choice and size of haplotype of existing reference panels. In particular, the imputation of missing genotype data of this complex admixed South African Coloured population population was suboptimal. Nonetheless, the increased number of SNPs generated by imputation analyses was useful in this study, yielding the replication of tuberculosis susceptibility loci (Thye *et al.*, 2012). Third, despite applying Fisher's Exact test to correct the imperfection of the mixed model for association used in our study, particularly in the case of rare variant, the

implementation of newer sequencing technologies is still required to search for rare risk variants. This may potentially provide crucial insights into identifying tuberculosis susceptibility genes and, therefore, inform the development of novel interventions.

9.1.3 Post Genome-wide Association Study Analysis

To achieve sufficient power to detect associations at a level of genome-wide significance and identify shared risk loci with a previously reported African tuberculosis case-control study (Thye *et al.*, 2010, 2012), a genome-wide meta-analysis was performed under random-effect and binary-effect models. In combining Genome-wide association studies data across these studies, two loci ($rs2057178$ and $rs11031728$) had an association result with genome-wide significance, and showed strong effect in both our study and the previous African tuberculosis case-control study (Thye *et al.*, 2012).

In order to examine the combined effects of genes by detecting genetic signals beyond single SNPs in Genome-wide Association Studies and fully characterize the susceptible genes and the genetic structure of complex diseases, we developed ancGWAS. ancGWAS is a post Genome-wide Association Study analysis tool for both recently admixed and non-admixed populations, which is based on a graph-based centrality measure within linkage disequilibrium and applies a statistical score to the resulting sub-graphs to identify the significant genes and networks associated with complex disease risk and to test for possible signals of unusual deficiency/excess of particular ancestry. Through a simulation of interactive disease loci in a simulation of an admixed population, we demonstrated the power of ancGWAS to significantly refine the signal of a disease gene that the standard Genome-wide Association analysis could not. We applied ancGWAS to the imputation Genome-wide Association Study data set of tuberculosis in the South African admixed Coloured population. Our results yielded 6 candidate genes, which are genome-wide significantly associated with tuberculosis, and moderately replicate 4 previously identified tuberculosis associated genes. We identified a novel central sub-network implicated mostly in acute and chronic myeloid leukemia signaling pathways, which potentially provides further insights into tuberculosis pathogenesis relevant to biomedical studies. All these genes were African ancestry-specific, i.e had predominately African ancestry, which supports the finding from chapter 4 that TB risk correlates with ‡Khomani ancestry. However, we observed no statistical evidence of unusual difference in an excess/deficiency of a ancestry in this unique admixed population, which may be explained by the fact the admixture event to create the SAC is too recent for selective forces to have had a significant impact on allele frequencies.

9.2 Conclusion

In conclusion, this PhD research has highlighted the importance of selecting the best proxy ancestry for potential downstream analysis in a multi-way admixed population by developing PROXYANC, a novel method for selecting the best proxy ancestral populations for a multi-way admixed population. This research demonstrated the benefit of refining standard genome-wide association studies signals, to fully characterize the susceptible genes and the genetic structure of complex disease by developing an algebraic graph-based method (ancGWAS) that identifies the most significant sub-network in complex diseases risk in recently admixed or non-admixed populations. It does this by integrating the association signal from genome-wide association study (GWAS), the local ancestry and SNP pair-wise linkage disequilibrium into the human protein-protein interaction (PPI) network. This research applied these newly developed approaches to understanding the genetic structure, and mapping possible disease genes in the uniquely 5-way admixed South African Coloured population which has unusually high rates of tuberculosis.

We refined both the choice of ancestral populations and their genetic contributions in the South African Coloureds. The investigation of admixture linkage disequilibrium and the identification of source populations for the South African Coloured population has not only deepened our understanding of its evolutionary history, but also provided opportunities for designing a method to account for a combined genome-wide SNP case-control study and admixture mapping in a multi-way admixed population such as the South African Coloured population. Importantly, our findings of the ancestral contributions of the South Africa Coloured populations may be regional specific, it will be important to generalize the results by analysing different dataset of Coloured population across the South Africa. PROXYANC also provides a useful tool for the investigation of other multi-way admixed populations.

We conducted the first ancestry-specific tuberculosis risk, typed and imputation GWAS of this complex admixed population, as well as a meta-analysis with a previous genome-wide association studies on African populations, which confirmed loci identified previously. Our results demonstrated significant evidence of an association between ǂKhomani ancestry and tuberculosis status that is not confounded by socio-economic status. Of note, the WT1 chr11 locus identified by Thye *et al.* (2012) is close to genome-wide significance in our standard GWAS. This provides crucial insights into identifying ancestry-specific tuberculosis risk in this multi-way admixed population. Combining the effect of SNPs for each gene from SNP signals from the GWAS using ancGWAS, revealed no signal of unusual difference in an excess/deficiency of ancestry at both the gene and pathway level in this population. However, we identified 6 novel candidate genes associated with tuberculosis and moderately replicate 4 known tuberculosis genes. Importantly, our results provide a novel significantly enriched central sub-network that may have a role in

acute and chronic myeloid leukemia signaling pathways. Future work will be to examine an accurate, unbiased estimation of the ancestry at every SNP in a multi-way admixed population to potentially provide crucial insights into identifying disease genes. This will provide a method to account for a combined genome-wide SNP case-control and admixture analysis in a multi-way admixed population such as the South African Coloured population.

Bibliography

Adhikari, M. (2005). Not white enough, not black enough: Racial identity in the south african coloured community. *Ohio University Press. H-SAfrica ISBN 978-0-89680-244-5*. (pages 1, 2).

Alexander, D., Novembre, J. & Lange, L. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research. 19, 1655-1664*. (pages 12, 32, 45, 45, 45, 55, 85).

Anton, I., Wange, L., Mayer, B., Ramesh, N. & Geha, R. (1998). The wiskott-aldrich syndrome protein-interacting protein (wip) binds to the adaptor protein nck. *J Biol Chem. 273, 20992-20995*. (page 31).

Arcos-Burgos, M. & Muenke, M. (2002). Genetics of population isolates. *Clin Genet. 61, 233-247*. (page 71).

Babb, O., van der Merwe, Beyers, N., Pheiffer, P., Walzla, W., Duncand, D. & Hoal, E. (2007). Vitamin d receptor gene polymorphisms and sputum conversion time in pulmonary tuberculosis patients. *Tuberculosis. 87(4), 295-302*. (pages 3, 84, 85, 94).

Baran, Y., Bogdan, P., Sankararaman, S., Dara, G., Gignoux, C., Celeste, C., Torgerson, W., Chapela, R., JeanFord, G., Avila, C.P., Rodriguez-Santana, J., Burchard, E.G. & Eran, E. (2012). Fast and accurate inference of local ancestry in latino populations. *Bioinformatics. 28, 1359-1367*. (pages 11, 12, 13, 13, 32, 96, 119, 124, 124, 124, 126, 128, 129, 130, 131, 136).

Barreiro, L., Neyrolles, L., Babb, O., Tailleux, L., Quach, L., McElreavey, H., Helden, K., Hoal, E., Gicquel, E. & Quintana-Murci, L. (2006). Promoter variation in the dc-sign encoding gene cd209 is associated with tuberculosis. *PLoS Med. 3, e20*. (page 95).

Bellamy, R. (1998). Genetic susceptibility to tuberculosis in human. *Thorax. 53, 588-593*. (page 84).

Bellamy, R., Beyers, N., McAdam, K., Ruwende, C., Gie, R., Samaai, P., Bester, D., Meyer, M., Torrah, Collin, M., Camidge, D., Wilkinson, D., Hoal, E., Whittle, H., Amos, W., Helden, V. & Hill, A. (2000). Genetic susceptibility to tuberculosis in africans:a genome-wide scan. *PNAS. 97, 8005-8009*. (pages 84, 85).

Bhattacharjee, S., Rajaraman, P., Jacobs, K., Wheeler, W., Melin, B., Hartge, P., Consortium, G., Yeager, M., Chung, C., Chanock, S. & Chatterjee, N. (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am J. of Hum Genet. 90, 821-835*. (page 116).

Blake, J., Taanman, J., Morris, A., Gray, R., Cooper, J., McKiernan, P., Leonard, J. & Schapira, A. (1999). Mitochondrial dna depletion syndrome is expressed in amniotic fluid cell cultures. *Am J. Pathol. 155(1), 67-70*. (page 99).

Boonzaaier, E., Malherbe, C., Smith, A. & Berens, P. (1996). The cape herders: A history of the khoikhoi of southern africa. *David Philip publishers, Cape Town*. (pages 1, 2, 63, 163).

Botha, M. (1972). Blood group gene frequencies. an indication of the genetic constitution of population samples in cape town. *S Afr Med J. 46, Suppl 1-26*. (page 34).

Browning, B. & Browning, B. (2009). A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Pam J Hum Genet. 84, 210-223*. (pages 33, 44).

Campbell, M. & Tishkoff, S. (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet. 9, 403-433*. (page 34).

Cann, H., de Toma, Cazes, L., Legrand, M., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W., Bonne-Tamir, B., Cambon-Thomsen, A., Chen, Z., Chu, J., Carcassi, C., Contu, L., Du, R., Excoffier, L., Ferrara, G., Friedlaender, J., Groot, H., Gurwitz, D., Jenkins, T., Herrera, R., Huang, X., Kidd, J., Kidd, K., Langaney, A., Lin, A., Mehdi, S., Parham, P., Piazza, A., Pistillo, M., Qian, Y., hu, Q., Xu, J., Zhu, S., Weber, J., Greely, H., Feldman, M., Thomas, G., Dausset, J. & Cavalli-Sforza, L. (2002). A human genome diversity cell line panel. *Science. 296, 261-262*. (pages 37, 54).

Cantor, R., Lange, K. & Sinsheimer, J. (2010). Prioritizing gwas results: A review of statistical methods and recommendations for their application. *Am J of Hum Genet. 86, 6-22*. (pages 27, 27, 27, 27, 27, 30, 30, 31, 31, 31, 31, 31, 134, 134, 135, 135).

Chakravati & Weiss (1998). Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Nat l.Acad.Science. 85, 9119-9123*. (pages 9, 10, 11, 11, 22, 26, 26, 78).

Chimusa, E., Meintjes, A., Tchanga, M., Mulder, N., Soodyall, H. & Ramesar, R. (2013). Genome-wide haplotype and signature of selection in indigenous southern african populations. *(Preparation)*.. (page 38, 38, 38, 38, 38).

Cho, Y., Go, M., Kim, Y., Heo, J., Oh, J., Ban, H., Yoon, D., Lee, M., Kim, D., Park, M., Cha, S., Kim, J., Han, B., Min, H., Ahn, Y., Park, M., Han, H., Jang, H., EY, C., Lee, J., Cho, N., Shin, C., Park, T., Park, J., Lee, J., Cardon, L., Clarke, G., McCarthy, M., Lee, J., Lee, J., Oh, B. & Kim, H. (2009). A large-scale genome-wide association study of asian populations uncovers genetic factors influencing eight quantitative traits. *Nature Genet. 41, 527-534*. (page 95).

Churchhouse, C. & Marchini, J. (2012). Multiway admixture deconvolution using phased or unphased ancestral panels. *Genet. Epidemiology. 37, 1-12*. (pages 13, 13, 124, 124, 130).

Cilliers, S. (1985). The coloureds of south africa; a factual survey. *Banier Publishers (Pty) Ltd*. (page 2).

Comstock, G. (1978). Tuberculosis in twins, a re-analysis of the prophit survey. *Am Rev Respir. 117, 621-624*. (page 84, 84).

Conrad, D., Jakobsson, M., Coop, G., Wen, X., Wall, J., Rosenberg, N. & Pritchard, J. (2010). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat. Genet. 38, 1251-1260*. (page 9).

Costa, G., Magno, L., Santana, C., C, Saito, S., Machado, M., Pietro, D., Bastos-Rodrigues, L., Miranda, D., Marco, L.D., Romano-Silva, M. & Rios-Santos, F. (2012). Genetic interaction between nat2, gstm1, gstt1, cyp2e1, and environmental factors is associated with adverse reactions to anti-tuberculosis drugs. *Mol Diagn Ther. 16(4), 241-350*. (page 136).

Dai, Y., Zhang, X., Pan, H., Tang, S., Shen, H., & Wang, J. (2011). Fine mapping of genetic polymorphisms of pulmonary tuberculosis within chromosome 18q11.2 in the chinese population: a case-control study. *BMC Infect Dis. 11, 211-282*. (pages 96, 112).

Daniel, T. (1997). Captain of death, the story of tuberculosis. *University of Rochester Press, Rochester, New York. 15, 131-142*. (page 84).

Davila, S., Hibberd, M., Dass, R., Wong, E.H., Sahiratmadja, E., Bonnard, C., Alisjahbana, B., Szeszko, J., Balabanova, Y., Drobniewski, F., Crevel, R., van Vosse, E., Nejentsev, S., Ottenhoff, T. & Seielstad, M. (2008). Genetic association and expression studies indicate a role of toll-like receptor 8 in pulmonary tuberculosis. *PLoS Genet. 4, e1000218*. (pages 96, 96, 101, 106, 112, 112, 116, 116, 118).

Davis, A. & Dollard, J. (1994). Children of bondage: the personality development of negro youth in the urban south. *American Council on Education. (1940). xxviii 299 pp*. (pages 1, 1, 2, 2, 2, 2, 33).

deWit, E., Delport, W., Chimusa, E., Meintjes, A., Moller, M., Helden, P., Seoighe, C. & Hoal, E. (2010a). Genome-wide analysis of the structure of the south african coloured population in the western cape. *Hum Genet. 128(2), 145-53*. (pages ix, 1, 1, 34, 34, 34, 34, 36, 37, 56, 60, 60, 60, 61, 62, 62, 62).

deWit, E., der Merwe, L., Helden, P.V. & Hoal, E. (2010b). Gene-gene interaction between tuberculosis candidate genes in a south african population. *Mammalian Genome. 22, 100-110*. (pages 90, 95, 136).

Dickson, P.S., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, B. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biology. 8(1), e1000294*. (page 31, 31, 31).

Dinga, J. & Lina, S. (2006). Monte carlo pedigree disequilibrium test for markers on the x chromosome. *Am J Hum Genet. 79(3), 567-573*. (pages 23, 25, 25, 25, 25, 25, 25, 26).

Draghici, S. (2003). *Data Analysis Tools For DNA MicroArrays*. Chapman, Hall/CRC, Boca Raton Londre New York, Revised Second ISBN:1584883154. (page 26).

Dye, C., Garnett, G., Sleeman, K. & Williams, B. (1998a). Directly observed short-course therapy. *Lancet. 352, 1886-1891*. (page 84, 84).

Dye, C., Garnett, G., Sleeman, K. & Williams, B. (1998b). Prospects for worldwide tuberculosis control under the who dots strategy. *Lancet. 352(9144), 1886-91*. (page 84).

Dye, C., Scheele, S., Dolin, P., Pathania, V. & Vaviglione, M. (1999). Global burden of tuberculosis: Estimated incidence, prevalence, and mortality by country. *JAMA. 282, 677-686*. (page 85).

Elphick, R. (1985). Khoikhoi and the founding of white south africa. *UWC Printing Dept. Ravan Press, Johannesburg*. (page 1).

Epstein, M., Allen, A. & Satten, G. (2007). A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet. 80, 921-930*. (page 27).

Evangelou, E., Maraganore, D. & Ioannidis, J. (2008). Meta-analysis in genome-wide association datasets: Strategies and application in parkinson disease. *Plos One. 2, e196*. (page 105).

Evans, D. & Cardon, L. (2005). A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am J Hum Genet. 76, 681-687*. (pages 8, 9, 10, 10).

Excoffier, L. & Hamilton, G. (2003). Comment on genetic structure of human populations. *Science. 300, 5627-1877*. (pages 22, 26, 26, 26, 26, 27, 29, 29, 29, 30).

Falush, D., Stephens, A. & Pritchard (2003). Inference of population structure: Extensions to linked loci and correlated allele frequencies. *Am J Hum Genet. 164, 1567-1587*. (pages 9, 11, 12, 12, 15, 16, 16, 17, 17, 17, 17, 17, 17, 17, 29, 32, 124).

Ferreira, M., O'Donovan, M., Meng, Y., Jones, I., Ruderfer, D., Jones, L., Fan, J., Kirov, G., Perlis, R., Green, E., Smoller, J., Grozeva, D., Stone, J., Nikolov, I., Chambert, K., Hamshere, M., Nimgaonkar, V., Moskvina, V., Thase, M., Caesar, S., Sachs, G., Franklin, J., Gordon-Smith, K., Ardlie, K., Gabriel, S., Fraser, C., Blumenstiel, B., Defelice, M., Breen, G., Gill, M., Morris, D., Elkin, A., Muir, W., McGhee, K., Williamson, R., MacIntyre, D., MacLean, A., CD, S., Robinson, M., Beck, M.V., Pereira, A., Kandaswamy, R., McQuillin, A., Collier, D., Bass, N., Young, A., Lawrence, J., Ferrier, I., Anjorin, A., Farmer, A., Curtis, D., Scolnick, E., McGuffin, P., Daly, M., Corvin, A., Holmans, P., Blackwood, D., Gurling, H., Owen, M., Purcell, S., Sklar, P., Craddock, N. & WTCCC (2008). Collaborative genome-wide association analysis supports a role for ank3 and cacna1c in bipolar disorder. *Nature Genet. 40, 1056-1058*. (page 105, 105).

Florez, J., Price, A., Campbell, D., Riba, L., Parra, M., Yu, F., Duque, C., Saxena, R., Gallego, N., Tello-Ruiz, M., Franco, L., Rodrguez-Torres, M., Villegas, A., Bedoya, G., Aguilar-Salinas, C., Tusi-Luna, M., Ruiz-Linares, A. & Reich, D. (2009). Strong association of socioeconomic status with genetic ancestry

in latinos: implications for admixture studies of type 2 diabetes. *Diabetologia. 52(8), 1528-36*. (page 89).

Flynn, J. (2006). Lessons from experimental mycobacterium tuberculosis infections. *Microbes Infect. 8, 1179-1188*. (page 84).

Frazer, K. & et al (2007). A second generation human haplotype map of over 3.1 million snps. *Nature. 449, 851-861*. (pages 37, 54, 127).

Galanter, J., Fernandez-Lopez, J., Gignoux, C., Barnholtz-Sloan, J., Fernandez-Rozadilla, C., Via, M., Hidalgo-Miranda, A., Contreras, A., Figueroa, L., Raska, P., Jimenez-Sanchez, G., Zolezzi, I., Torres, M., Ponte, C., Ruiz, Y., Salas, A., Nguyen, E., Eng, C., Borjas, L., Zabala, W., Barreto, G., Gonzlez, F., Ibarra, A., Taboada, P., Porras, L., Moreno, F., Bigham, A., Gutierrez, G., Brutsaert, T., Len-Velarde, F., Moore, L., Vargas, E., Cruz, M., Escobedo, J., Rodriguez-Santana, J., Rodriguez-Cintrn, W., Chapela, R., Ford, J., Bustamante, C., Seminara, D., Shriver, M., Ziv, E. & Burchard, E. (2012). Development of a panel of genome-wide ancestry informative markers to study admixture throughout the americas. *PLOS Genet. 8(3), e1002554*. (page 71).

Garcia-Barceloa, M., Tangc, C., Ngana, E., Luia, V., Chena, Y., Soa, M., Yuk-yu, T., Miao, X., Shuma, C., Liua, F., Yeung, M., Yuane, Z., Guof, W., Liuc, L., Sung, X., Huang, L., Tou, J., Song, Y., han, D., Cheung, K., Wong, K., Chernyc, S., Shamb, P. & Tam, P. (2009). Genome-wide association study identifies nrg1 as a susceptibility locus for hirschsprungs disease. *Proc. Natl Acad. Science. 106, 2694-2699*. (page 95).

Glaab, E., Baudot, A., Krasnogor, N., Schneider, R. & Valencia, A. (2012). Enrichnet: network-based gene set enrichment analysis. *Bioinformatics. 28 (18), i451-i457*. (pages 151, 157, 160).

Goldstein, D. & Weale, M. (2001). Population genomics: linkage disequilibrium holds the key. *Curr Biol. 11(14), R576-9*. (pages 8, 10, 10, 10, 22).

Gronau, I., Hubisz, M., Gulko, B., Danko, C. & Siepel, A. (2011). Bayesian inference of ancient human demography from individual genome sequences. *Nature Genet. 43, 1031-1034*. (page 31).

Halder, H. & Shriver, S. (2003). Measuring and using admixture to study the genetics of complex diseases. *Hum Genomics. 1, 52-62*. (pages 8, 8, 22, 22, 22, 22, 26).

Han, B. & Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet. 88, 586-598*. (pages 31, 31, 105, 105, 106, 106, 107, 116, 116, 116, 135, 135, 135).

Henn, B., Gignouxb, C., Jobinc, M., Grankae, J., Macphersonf, Kidda, J., Rodrguez-Botigug, L., Ramachandranh, S., Honf, L., Brisbini, A., Linj, A., Underhilllj, P., Comasg, D., Kiddk, K., Normanl, P., Parhaml, P., Bustamantea, C., Mountainf, J. & Feldmane, M. (2011). Hunter-gatherer genomic diversity suggests a southern african origin for modern humans. *PNAS. 108, 5154-5162*. (pages 37, 38, 38, 38, 54).

Henn, B., Botigue, L., Gravel, S., Wang, W., Brisbin, A., Byrnes, J., Fadhlaoui-Zid, K., Zalloua, P., AMoreno, Bertranpetit, J., Bustamante, C. & Comas, D. (2012). Genomic ancestry of north africans supports back-to-africa migrations. *Nat Comm. 3 (1143) 2140*. (pages 13, 38, 38, 39, 124, 124).

Hirschhorn, J. & Daly, M. (2003). Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet. 6(2), 95-108*. (pages 35, 95, 97).

Hoal, E., Lewis, L., Jamieson, J., Tanzer, T., Rossouw, R., Victor, V. & el.al (2004). Slc11a1 (nramp1) but not slc11a2 (nramp2) polymorphisms are associated with susceptibility to tuberculosis in a high-incidence community in south africa. *Stellenbosch University Faculty of Health Sciences, and Metropolitan Cape Town, Western Cape,South Africa*. (pages 3, 35, 85, 94, 94).

Hoggart, H., Shiver, S. & McKeigue, P. (2004). Design and analysis of admixture mapping studies. *Am J Hum Genet. 74(5), 965-978*. (pages 8, 11, 12, 26, 29, 30, 32, 33, 124).

Hokayem, J., Huber, C., Couv, A., Aziza, J., Baujat, G., Bouvier, R., Cavalcanti, D., Collins, F., Cordier, M., Delezoide, A., Gonzales, M., Johnson, D., Merrer, M., Levy-Mozziconacci, A., Loget, P., Martin-Coignard, D., Martinovic, J., Mortier, G., Marie-Jos, P., Roume, J., Scarano, G., Munnich, A. & Cormier-Daire, V. (2012). Nek1 and dync2h1 are both involved in short rib polydactyly majewski type but not in beemer langer cases. *Am J. Med Genet. 49, 227-233*. (page 108, 108).

Horvath, S., Windemuth, C. & Knapp, M. (2000). The disequilibrium maximum-likelihood binomial test does not replace the transmission and disequilibrium test. *Am J Hum Genet. 67(2), 531-534*. (pages 25, 25, 25, 25, 26).

Hubisz, M., Falush, D., Stephens, M. & Pritchard, J. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources. 9, 1322-1332*. (page 17, 17, 17).

Hudelson, P. (1996). Gender differentials in tuberculosis and lung disease, management of tuberculosis: a guide for low income countrie. *International Union Against Tuberculosis and Lung Disease, 5th ed. Paris*. (page 90).

Jia, P., Zheng, S., Long, J., Zheng, W. & Zhao, Z. (2010). dmgwas: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics. 27, 95-102*. (pages 30, 30, 30, 31, 31, 31, 134, 134, 134, 135, 135, 135, 136, 140).

Kang, H., Sul, J., Service, S., Zaitlen, N., Sit-yee, K., Freimer, N., Sabatti, C. & Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nat Genet. 42, 348-354*. (pages 28, 28, 28, 97, 97, 98, 107, 143).

Kaufmann, S. & McMichael, A. (2005). Annulling a dangerous liaison: vaccination strategies against aids and tuberculosis. *Nat Med. 11, S33-S44*. (page 84, 84).

Keegan, T. (1996). Colonial south africa and the origins of the racial order. *David Philip publishers. Claremont, South Afica*. (pages 2, 2, 2, 2, 63, 163).

Kennedy, G., Matsuzaki, H., Dong, S., Liu, W., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J., Liu, W., Yang, G., Di, X., Ryder, T., He, Z., Surti, U., Phillips, M., Boyce-Jacino, M., Fodor, S. & Jones, K. (2003). Large-scale genotyping of complex dna. *Nat Biotechnol. 21, 1233-7*. (page 95).

Kim, A. & et.al, H.A. (2009). Alpha-t-catenin (ctnna3) gene was identified as a risk variant for toluene diisocyanate-induced asthma by genome-wide association analysis. *Clin. Exp. Allergy. 39, 203-212*. (page 95).

Kosoy, R., Nassir, R., Tian, C., White, P., Butler, L., Silva, G., Kittles, R., Alarcon-Riquelme, M., Gregersen, P., Belmont, J., DeLaVega, F. & Seldin, M. (2009). Ancestry informative marker sets for determining continental origin and

admixture proportions in common populations in america. *Hum Mutat. 30(1), 69-78*. (pages 70, 71).

Kristin, C., Kruglyak, L. & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genet. 3, 299-309*. (pages 9, 9, 9, 9, 9, 10, 10, 11, 11, 22, 26).

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet. 22, 139-144*. (page 75).

Kumar, R., Seibold, A.M., Aldrich, C.M., Williams, L.K., Reiner, P.A., Colangelo, L., Galanter, J., Gignoux, C., Hu, D., Sen, S., Choudhry, S., Peterson, L.E., Rodriguez-Santana, J., Rodriguez-Cintron, W., Nalls, M., Leak, T., Meara, E., Meibohm, B., Kritchevsky, S., Li, R., Harris, T., Nickerson, D., Fornage, M., Enright, P., Ziv, E., Smith, L., Liu, K. & Gonzlez-Burchard, E. (2010). Genetic ancestry in lung-function predictions. *N Engl J Med. 363, 321-330*. (page 85, 85).

Lawson, D., Hellenthal, G., Myers, S. & Falush, D. (2010). Inference of population structure using dense haplotype data. *PLoS Genet. 8(1):e1002453*. (page 124).

Lee, W. & Yen, Y. (2003). Admixture mapping using interval transmission/disequilibrium tests. *Ann Hum Genet. 67, 580-8*. (page 23, 23).

Lewontin, L. (1964). The interaction of selection and linkage, general consideration; heterotic models. *Genet. 49(1), 49-67*. (pages 9, 71).

Li, J., Guo, Y., Pei, Y. & Hong-Wen, D. (2012). The impact of imputation on meta-analysis of genome-wide association studies. *PLoS ONE. 7(4), e34486*. (pages 33, 105, 105, 105).

Li, N. & Stephens, M. (2003). Modelling linkage disequilibrium, and identifying recombination hotspots using snp data. *Genetics. 165, 2213-2233*. (pages 8, 78).

Lin, Z. & Altman, R. (2004). Finding haplotype tagging snps by use of principal components analysis. *Am Soc of Hum Genet. 75(5), 850-61*. (pages 73, 73, 74).

Loh, P., Lipson, M., Patterson, N., Moorjani, P., Pickrell, J., Reich, D. & Berger, B. (2013). Inferring admixture histories of human populations using linkage disequilibrium. *Nat Genetics (in press)*.. (page 76).

Lohmueller, K., Albrechtsen, A., Li, Y., Kim, S.Y., Korneliussen, T., Vinckenbosch, N., Tian, G., Huerta-Sanchez, E., Feder, A.F., Grarup, N., Jrgensen, T., Jiang, T., Witte, D.R., Sandbk, A., Hellmann, I., Lauritzen, T., Hansen, T., Pedersen, O., Wang, J. & Nielsen, R. (2011). Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. *PLoS Genet 7(10), e1002326*. (page 8).

Magnus, N. (2000). Coalescent theory. *Am J Hum Genet. 52, 506-16*. (pages 11, 11, 11, 11, 22).

Manolio, A., Francis, S., Collins, Nancy, J., Cox, David, B., Goldstein, Lucia, A., Hunter, D., McCarthy, M., Ramos, E., Cardon, L., Chakravarti, A., Cho, J., Guttmacher, A., Kong, A., Kruglyak, L., Mardis, E., i, C.R., Slatkin, M., Valle, D., Whittemore, A., Boehnke, M., Clark, A., Eichler, E., Gibson, G., Haines, J., Mackay, T., McCarroll, S. & Visscher, P. (2004). Finding the missing heritability of complex diseases. *Nature. 461(7265), 747-753*. (pages 22, 30, 30).

Marchini, J. & Howie, B. (2008). Comparing algorithms for genotype imputation. *Am J Hum Genet. 83, 535-539*. (pages 33, 95, 96, 96, 106, 107, 119).

Martin, D., Bass, M. & Kaplan, N. (2001). Correcting for a potential bias in the pedigree disequilibrium test. *Am J of Hum Genet. 68(4), 1065-1067*. (pages 23, 24, 25, 25, 25, 26).

McKeigue, P. (2005). Prospects for admixture mapping of complex traits. *Am J Hum Genet. 76(1), 1-7*. (pages 8, 8, 8, 8, 9, 22, 22, 26, 26, 26, 26, 29, 29, 33).

Moller, M. & Hoal, E. (2010a). Current findings, challenges and novel approaches in human genetic susceptibility to tuberculosis. *Tuberculosis. 90, 71-83*. (pages 94, 95, 95, 95).

Moller, M. & Hoal, E. (2010b). Past, present and future directions in human genetic susceptibility to tuberculosis. *FEMS Immunol Med Microbiol. 58, 3-26*. (pages 94, 95, 95, 135).

Moller, M., Nebel, A., Valentonyte, R., Helden van, S.S. & Hoal, E. (2009). Investigation of chromosome 17 candidate genes in susceptibility to tb in a south african population. *Tuberculosis. 89, 189-194*. (page 94).

Montana, G. & Pritchard, J. (2004). Statistical tests for admixture mapping with case-control and case-only data. *Am J Hum Genet. 75(5), 771-789*. (pages 12, 16, 16, 17, 17, 26, 29, 29, 30, 70).

Moorjani, P., Patterson, N., Hirschhorn, J., Keinan, A., Hao, L., Atzmon, G., Burns, E., Ostrer, H., Price, A. & Reich, D. (2011). The history of african gene flow into southern europeans, levantines, and jews. *PLoS Genet. 7(4), e1001373*. (page 125, 125).

Mountain, A. (2003). The first people of the cape: A look at their history and the impact of colonialism on the cape's indigenous people. *David Philips Publishers. ISBN:0-86486-623-2*. (pages 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 33, 60, 82, 163).

Mountain, A. (2004). An unsung heritage: Perspectives on slavery. *David Philips Publishers. ISBN:0-86486-622-4*. (pages 1, 2, 2, 2).

Nianjun, L., Wu, B. & Zhao, H. (2006). Inference of population structure using mixture model. *Department of Epidemiology and Public Health and Yale University School of Medicine, USA*. (page 32).

Nurse, G., Weiner, J. & Jenkins, T. (1985). The peoples of southern africa and their affinities. *Clarendon Press, ISBN:0-19-857541-6*. (pages 1, 2, 2).

Pasaniuc, B., Sankararaman, S., Kimmel, G. & Halperin, E. (2009). Inference of locus-specific ancestry in closely related population. *Bioinformatics. 25, i213-i221*. (pages 11, 13, 13, 13, 32, 124, 126, 128, 130).

Pasaniuc, B., Zaitlen, N., Lettre, G., Chen, G., Tandon, A., Kao, W.L., Ruczinski, I., Fornage, M., Siscovick, D., Zhu, X., Larkin, E., Lange, L., Cupples, A., Yang, Q., Akylbekova, E., Musani, S., Divers, J., Mychaleckyj, J., Li, M., Papanicolaou, G., Millikan, R., Ambrosone, C., John, E., Bernstein, L., Zheng, W., Hu, J., Ziegler, R., Nyante, S., Bandera, E., Ingles, S., Press, M., Chanock, S., Deming, S., Rodriguez-Gil, J., Palmer, C., Buxbaum, S., Ekunwe, L., Hirschhorn, J., Henderson, B., Myers, S., Haiman, C., Reich, D., Patterson, N., Wilson, J. & Price, A. (2011). Enhanced statistical tests for gwas in admixed populations: Assessment using african americans from care and a breast cancer consortium. *PLoS Genet.7(4), e1001371*. (pages 3, 30, 96, 96, 96, 96, 97, 97, 125, 133).

Paschou, P., Ziv, E., Burchard, E., Choudhry, S., Rodriguez-Cintron, W., Mahoney, M. & Drineas, P. (2007). Pca-correlated snps for structure identification in world wide human population. *Plos Genet. 3(9), e160*. (pages 70, 70, 71, 73).

Patterson, N., Hattangadi, N. & Lane, B. (2004). Methods for high-density admixture mapping of disease genes. *Am J Hum Genet. 74(5), 979-1000*. (pages 8, 10, 11, 11, 11, 12, 22, 26, 30, 30).

Patterson, N., Price, A. & Reich, D. (2006). Population structure and eigenanlysis. *PLoS Genet. 2(12), e190*. (pages 11, 14, 15, 32, 32, 37, 73, 124).

Patterson, N., Petersen, D., van.der.Ross, R., Sudoyo, H., Glashoff, R., Marzuki, S., Reich, D. & Hayes, V. (2009). Genetic structure of a unique admixed population: implications for medical research. *Huma Molecular Genet. 19, 411-419*. (pages 34, 60, 62).

Patterson, N., Moorjani, P. & Yontao Luo, e.a. (2012). Ancient admixture in human history. *Genet. Society of Am. 10, 112.145037*. (pages 7, 7, 33).

Peltonen, L., Palotie, A. & Lange, K. (2000). Use of population isolates for mapping complex traits. *Nat Rev Genet. 1, 182-19*. (page 71).

Peng, G., Luo, L., Hoicheong, S., Zhu, Y., Pengfei, H., Hong, S., Jinying, Zhao, X., Xiaodong, Z., Reveille, D.J., Jin, L., Amos, C. & Xiong, M. (2008). Gene and pathway-based analysis: Second wave of genome-wide association studies. *European J of Hum Genet. 18, 111-117*. (pages 30, 30, 30, 31, 31, 31, 134, 134, 134, 134, 135, 135, 136, 136).

Peng, G., Guo, Z., Kiniwa, Y., Voo, K., Peng, W., Fu, T., Wang, D., Li, Y., Wang, H. & Wang, R. (2011). Toll-like receptor 8-mediated reversal of cd4+ regulatory t cell function. *Science. 309 (5739), 1380-4*. (page 112).

Pickrell, K., Patterson, N., Barbieri, C., Berthold, F., Gerlach, L., Lipson, M., Loh, L.P.R., Guldemann, T., Kure, B., Mpoloka, W., Nakagawa, H., Naumann, C., Mountain, J., Bustamante, C., Berger, B., Henn, B., Stoneking, M., Reich, D. & Pakendorf, B. (2012). The genetic prehistory of southern africa. *Nature Communications 3 (1143) doi:10.1038/ncomms2140*. (pages 6, 34, 34, 63, 164).

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N. & Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet. 38, 904-909*. (pages 12, 37).

Price, A., Patterson, N. & Fuli, Y. (2007). A genomewide admixture map for latino populations. *Am J Hum Genet. 80(6), 1024-1036*. (page 17).

Price, A., Helgason, A., Palsson, S., Stefansson, H., Clair, D., Andreassen, O., Reich, D., Kong, A. & Stefansson, K. (2009a). The impact of divergence time on the nature of population structure: An example from iceland. *PLoS Genet. 5(6), e1000505*. (pages 40, 86, 87, 87, 87).

Price, A., Tandon, A., Patterson, N., Barnes, K., Rafaels, N., Ruczinski, I., Beaty, T., Mathias, R., Reich, D. & Myers, S. (2009b). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *Plos Genet. 5, e1000519*. (pages 11, 12, 12, 13, 32, 44, 124, 124, 125, 125, 125, 127, 130).

Price, A., Zaitlen, N., Reich, D. & Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genet. 11, 459-463*. (pages 27, 98).

Pritchard, S., Stephens, M. & Donnelly, M. (2002). Inference of population structure using multi-locus genotype data. *Am J Hum Genet. 155, 945-959*. (pages 8, 15, 16, 16, 17, 17, 17, 17, 29).

Pugach, I., Matveyev, R., Wollstein, A., Kayser, M. & Stoneking, M. (2011). Dating the age of admixture via wavelet transform analysis of genome-wide data. *Genome Biology. 12, R19*. (page 125, 125).

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M., Bender, D., Maller, J., Sklar, P., de Bakker, Daly, M. & Sham, P. (2007). Plink: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet. 81, 559-575*. (pages 36, 37, 76, 76, 98).

Qin, H., Morris, N., Kang, S., Li, M., Tayo, B., Lyon, H., Hirschhorn, J., Cooper, R. & Zhu, X. (2010). Interrogating local population structure for fine mapping in genome-wide association studies. *Bioinformatics. 26, 2961-2968*. (pages 12, 29, 96).

Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Bormans, Z., van.Helden, P., Hoal, E. & Behar, M.D. (2010). Strong maternal khoesan contribution to the south african coloured population:a case of gender-biased admixture. *Am Soc of Hum Genet. 86, 611-620*. (pages 34, 34, 56, 63, 164).

Rauscher, F. (1993). The wt1 wilms tumor gene product: a developmentally regulated transcription factor in the kidney that functions as a tumor suppressor. *FASEB J. 896-903, PMID 8393820*. (page 112).

Redden, D., Divers, J., Vaughan, L., Tiwari, H., Beasley, T., Fernndez, J., Kimberly, R., Feng, Padilla, M., Liu, N., Miller, M. & Allison, D. (2006). Regional admixture mapping and structured association testing: conceptual unification and an extensible general linear model. *PLoS Genet. 2(8), e137*. (pages 28, 29, 96).

Reich, D., Patterson, N., Jager, P. & McDonald, G. (2005). A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nature Genet. 37, 1113-1118*. (pages 8, 9, 9).

Reich, D., Thangaraj, K., Patterson, N., Price, A. & Singh, L. (2009). Reconstructing indian population history. *Nature. 461, 489-494*. (pages 6, 6, 7).

Risch, N. (2000). Searching for genetic determinants in the new millennium. *Nature. 405, 847-56*. (page 95).

Rodriguez, J., Bercovici, S., Elmore, M. & Batzoglou, S. (2012). Ancestry inference in complex admixtures via variable-length markov chain linkage models. *J. Comput Biol. 20(3):199-211*. (pages 13, 119, 124, 124).

Rosenberg, N. (2004). Distruct: a program for the graphical display of population structure. *Molecular Ecology Notes. 4, 137-138*. (page 45).

Rosenberg, N. (2005). Algorithms for selecting informative marker panels for population assignment. *J Computational Biology. 12(9), 1183-1201*. (page 71).

Rosenberg, N. & Nordborg, M. (2006). A general population-genetic model for the production by population structure of spurious genotype-phenotype associations in discrete, admixed or spatially distributed populations. *Genet. Society of Am. 173, 1665-1678*. (pages 12, 27, 27, 96).

Rosenberg, N. & Pritchard, J. (2008). Genetics structure of human populations. *Science. 298, 2381-2385*. (pages 6, 8, 8, 12, 27, 30, 32).

Rosenberg, N., Li, L., Ward, R. & Pritchard, J. (2003). Informativeness of genetic markers for inference of ancestry. *Am J Hum Genet. 73, 1402-1422*. (pages 6, 71).

Rosenberg, N., Huang, L., Jewett, E., Szpiech, Z., Jankovic, I. & Boehnke, M. (2010). Genomewide association studies in diverse populations. *Nature Reviews Genet. 11, 356-366*. (pages 27, 30, 95, 95, 96, 96).

Ross, V. (1993). 100 questions about coloured south africans. *UWC Printing Department, Cape Town*. (pages 1, 34).

Sanghera, D., Been, L., Ortega, L., Wander, G., Mehra, N., Aston, C., Mulvihill, J. & Ralhan, S. (2009). Testing the association of novel meta-analysis-derived diabetes risk genes with type ii diabetes and related metabolic traits in asian indian sikhs. *J of Hum Genet. 54, 162-168*. (page 105).

Sankararaman, S., Kimmel, G., Halperin, E. & Jordan, M. (2008). On the inference of ancestries in admixed populations. *Genome Res. 18(4), 668-675*. (pages 11, 12, 13, 13, 30, 32, 130).

Santafe, G., Lozano, J. & Larranaga, P. (2006). Bayesian model averaging of naive bayes for clustering. *EEE Trans Syst Man Cybern B Cybern. 36(5), 1149-6*. (page 30).

Schaid, A. (1998). Transmission disequilibrium, family controls, and great expectations. *Am J of Hum Genet. 63, 935-941*. (pages 23, 25, 25).

Scheuner, M., Yoon, P. & Khoury, M. (2004). Contribution of mendelian disorders to common chronic disease; opportunities for recognition, intervention, and prevention. *Am J. Med Genet. 125C(1), 50-65*. (pages 22, 30).

Schlebusch, C., Skoglund, P., Sjdin, P., Gattepaille, L., Hernandez, D., Jay, F., Li, S., Jongh, M., Singleton, A., Blum, M., Soodyall, H. & Jakobsson, M. (2012). Genomic variation in seven khoe-san groups reveals adaptation and complex african history. *Science. 338, 374-379*. (pages 34, 34, 63, 164).

Schramm, C., Phillips, H., Operario, C., Lee & Weber, J. (2002). Ethnic-difference markers for use in mapping by admixture linkage disequilibrium. *Am J of Hum Genet. 70(3), 737-50*. (pages 8, 25).

Schuster, S., Miller, W., Ratan, A., Tomsho, L., Giardine, B., Kasson, L., Harris, R., Petersen, D., Zhao, F., Qi, J., Alkan, C., Kidd, J., Sun, Y., Drautz, D., Bouffard, P., Muzny, D., Reid, J., Nazareth, L., Wang, Q., Burhans, R., Riemer, C., Wittekindt, N., Moorjani, P., Tindall, E., Danko, C., Teo, W., Buboltz, A., Zhang, Z., Ma, Q., Oosthuysen, A., Steenkamp, A., Oostuisen, H., Venter, P., Gajewski, J., Zhang, Y., Pugh, B., Makova, K., Nekrutenko, A., Mardis, E., Patterson, N., Pringle, T., Chiaromonte, F., Mullikin, J., Eichler, E., Hardison, R., Gibbs, R., Harkins, T. & Hayes,

V. (2010). Complete khoisan and bantu genomes from southern africa. *Nature, 463, 943-947*. (pages 63, 164).

Seldin, M., Pasaniuc, B. & Price, A. (2011). New approaches to disease mapping in admixed populations. *Nat Rev Genet. 36, S21-S27*. (pages 8, 33, 96, 96, 97, 124, 126).

Setakis, E., Stirnadel, H. & Balding, D. (2006). Logistic regression protections against population structure in genetic association studies. *Genome Research*. (page 96).

Shiheng, T., Rongmei, Z., Jianhua, C., Xiaoming, L., Liping, D., Qingyuan, Q. & Zewei, L. (2001). A population genetics model of linkage disequilibrium in admixed populations. *Chinese Science Bullin. 46, 193-197*. (pages 9, 72, 75).

Small, P. (1996). Tuberculosis research. balancing the portfolio. *JAMA. 276, 1512-1513*. (page 85).

Smith, D. (2004). Contribution of mendelian disorders to common chronic disease; opportunities for recognition, intervention, and prevention. *Am J of Med Genet. 15, 125C(1):50-65*. (pages 22, 23).

Smith, D. (2007). Capitalising on mendelian randomization to assess the effects of treatments. *J R Soc Med. 100, 432-435*. (page 21).

Smith, D. & Ebrahim, E. (2004). Mendelian randomization: prospects, potentials, and limitations. *International J. of Epidemiology. 33, 30-42*. (page 22).

Smith, S. & O'Brien, B. (2005). Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. *Nature Rev. Genet. 6, 623-632*. (pages 8, 70).

Sohn, K.A. & Xing, E. (2007). Spectrum: Joint bayesian inference of population structure and recombination events. *Bioinformatics. 23, i479-i489*. (pages 12, 17).

Sorensen, T., Nielsen, G., Andersen, P. & Teasdale, T. (1988). Genetic and environmental influences on premature death in adult adoptees. *New Engl J. Med. 318, 727-732*. (page 84, 84).

Spielman, R., McGinnis, R. & Ewens, W. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am J Hum Genet. 52, 506-16*. (pages 9, 22, 23, 26).

Staiger, H., Machicao, F., Kantartzis, K., Schafer, S., Kirchhoff, K., Guthoff, M., Silbernagel, G., Stefan, N., Fritsche, A. & Hring, H. (2008). Novel meta-analysis-derived type 2 diabetes risk loci do not determine prediabetic phenotypes. *Plos One. 3, e3019*. (page 105).

Stein, C. (2011). Genetic epidemiology of tuberculosis susceptibility:impact of study design. *PLoS Pathog. 7(1), e1001189*. (pages 35, 36, 95).

Stopple, M. (1996). *Genetic diseases overview*. (page 21, 21, 21).

Strachan, P. & Read, A. (1999). *Human Molecular Genetics*. (page 23).

Sum, S., Eleanor, Y., Peng, B., Yu, X., Chen, J., Byrne, J., Lindeman, G. & Visvader, J. (2002). The lim domain protein lmo4 interacts with the cofactor ctip and the tumor suppressor brca1 and inhibits brca1 activity. *J.Biol.Chem. 277 (10), 7849-56, PMID 11751867*. (page 110, 110, 110).

Sundquist, A., Fratkin, E., Do, B.C. & Batzoglou, S. (2008). Effect of genetic divergence in identifying ancestral origin using hapaa. *Genome Res. 18, 676-682*. (page 12).

Tang, H., Coram, M., Wang, P., Zhu, X. & Risch, N. (2006). Reconstructing genetic ancestry blocks in admixed individuals. *Am J Hum Genet. 79, 1-12*. (pages 8, 11, 12).

Terry, S. (2003). *Statistical Analysis of Gene Expression Microarray Data*. Genet. research. 77, 123-128. (page 17).

Thye, T., Vannberg, F., Wong, S., Owusu-Dabo, E., Osei, I., Gyapong, J., Sirugo, G., Sisay-Joof, F., Enimil, A., Chinbuah, M., Floyd, S., Warndorff, D., Sichali, L., Malema, S., Crampin, A., Ngwira, B., Teo, Y., Small, K., Rockett, K., Kwiatkowski, D., Fine, P., Hill, P., Newport, M., Lienhardt, C., Adegbola, R., Corrah, T., Ziegler, A., WTCCC, W., Morris, A., Meyer, C., Horstmann, R. & Hill, A. (2010). Genome-wide association analyses identifies a susceptibility locus for tuberculosis on chromosome 18q11.2. *Nat. Genet. 42, 739-741*. (pages 95, 95, 96, 96, 101, 106, 112, 112, 112, 116, 116, 116, 116, 118, 118, 118, 166).

Thye, T., Owusu-Dabo, E., Vannberg, F., Crevel, R., Curtis, J., Sahiratmadja, E., Balabanova, Y., Ehmen, C., Muntau, B., Ruge, G., Sievertsen, J., Gyapong, J., Nikolayevskyy, V., Hill, P., Sirugo, G., Drobniewski, F., de.Vosse, E., Newport, M., Alisjahbana, B., Nejentsev, S., Ottenhoff, T.,

Hill, A., Horstmann, R. & Meyer, C. (2012). Common variants at 11p13 are associated with susceptibility to tuberculosis. *Nat. Genet. 44, 257-259*. (pages 95, 95, 96, 101, 106, 109, 109, 110, 116, 116, 116, 116, 118, 118, 118, 119, 165, 166, 166, 167).

Tishkoff, S. & Kidd, K. (2004). Implications of biogeography of human populations for 'race' and medicine. *Nat Rev Genet. 36, S21-S27*. (page 6).

Tishkoff, S., Reed, F., Friendlaender, F., Ehret, C. & Ranciaro, A. (2009). The genetic structure and history of africans and african americans. *Sciences. 324, 1035-1044*. (pages 33, 34, 34, 34, 34, 56).

Tiwari, H., Barnholtz-Sloan, J., Wineinger, N., Padilla, M., Vaughan, L. & Allison, D. (2008). Review and evaluation of methods correcting for population stratification with a focus on underlying statistical principles. *Hum Hered. 66, 67-86*. (page 27).

Unoki, A. & et.al, H.A. (2008). Snps in kcnq1 are associated with susceptibility to type 2 diabetes in east asian and european populations. *Nature Genet. 40, 1098-1102*. (page 95).

Vega, F., Isaac, H. & Scafe, C. (2006). A tool for selecting snps for association studies based on observed linkage disequilibrium patterns. *Pacific Symposium on Biocomputing*. (page 71).

Verhoeven, K., Macel, M., Wolfe, M.L. & Biere, A. (2010). Population admixture, biological invasions and the balance between local adaptation and inbreeding depression. *Proc. R. Soc. B. 278, 2-8*. (page 8).

Warren, E. & Grant, G. (2005). *Statistical Methods in Bioinformatics*. Springer, New York, 10013,ISBN:0-387-9529-2. (page 17).

Weir, B. (2008). linkage disequilibrium and association mapping. *Ann Rev of Genomics and Hum Genet. 9, 129-142*. (pages 6, 9, 9, 9, 22).

Weir, B. & Cockerham, C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution. 38, 1358-1370*. (page 6).

WHO (2000). World health organization. *Bulletin of WHO*. (page 84, 84).

WHO (2004). Reaching the poor: challenges for tuberculosis programmes in the western pacific region. *Manila: WHO Regional Office for the Western Pacific*. (page 90).

WHO (2005). 30 gender-based analysis of tuberculosis-related data and other information. *WHO Representative in Viet Nam*. (page 90).

Winkler, C., Nelson, G. & Smith, M. (2010). Admixture mapping comes of age. *Annu.Rev.Genomics Hum. Genet. 11, 65-89*. (pages 33, 77, 77, 78).

Wray, N., Pergadia, M., Blackwood, D., Penninx, B., Gordon, B., Nyholt, D., Ripke, S., MacIntyre, D., McGhee, K., Maclean, A., SmitJH., J., Hottenga, J., Willemsen, G., Middeldorp, C., Geus, D., Lewis, C., McGuffin, P., Hickie, I., van.den, E., Liu, J., Macgregor, S., McEvoy, B., Byrne, E., Medland, S., Statham, D., Henders, A., Heath, A., Montgomery, G., Martin, N., Boomsma, D., Madden, P. & Sullivan, P. (2010). Genome-wide association study of major depressive disorder: new results, meta-analysis, and lessons learned. *Molecular Psychiatry. 17, 36-48*. (pages 31, 135).

Wu, J., Vallenius, T., Ovaska, K., Westermarck, J., Mkel, T. & Hautaniemi, F. (2009). Integrated network analysis platform for protein-protein interactions. *Nat. Meth. 6, 75-77*. (pages 31, 111, 112, 136).

Xing, E., Jordan, M. & Sharan, R. (2007). Bayesian haplotype inference via the dirichlet process. *J. of Computational Biology. 14, UCB/CSD 3/1275*. (page 17).

Yasuda, A. & et.al, H.A. (2008). Variants in kcnq1 are associated with susceptibility to type 2 diabetes mellitus. *Nature Genet. 40, 1092-1097*. (page 95).

Zhan, S., Marchini, J. & Donnelly, P. (2011). Hapgen2: simulation of multiple disease snps. *Bioinformatics. 7 (16), 2304-2305*. (page 143).

Zhang, C., Chen, K., Seldin, M. & Hongzhe, L. (2004). A hidden markov modeling approach for admixture mapping based on case-control haplotype data. *Genet. Epidemiol. 27(3), 225-39*. (pages 18, 26).

Zhang, X., Huang, W., Yang, S., Sun, L., Zhang, F., Zhu, Q., Zhang, F., Zhang, C., Du, W., Pu, X., Li, H., Xiao, F., Wang, Z., Cui, Y., Hao, F., Zheng, J., Yang, X., Cheng, H., He, C., Liu, X., Xu, L., Zheng, H., Zhang, S., Zhang, J., Wang, H., Cheng, Y., Ji, B., Fang, Q., Li, Y., Zhou, F., Han, J., Quan, C., Chen, B., Liu, J., Lin, D., Fan, L., Zhang, A., Liu, S., Yang, C., Wang, P., Zhou, W., Lin, G., Wu, W., Fan, X., Gao, M., Yang, B., Lu, W., Zhang, Z., Zhu, K., Shen, S., Li, M., Zhang, X., Cao, T., Ren, W., Zhang, X., He, J., Tang, X., Lu, S., Yang, J., Zhang, L., Wang, D., Yuan, F., Yin, X., Huang, H., Wang, H., Lin, X. & Liu, J. (2009). Psoriasis genome-wide association study identifies susceptibility variants within lce gene cluster at 1q21. *Nature Genet. 41, 205-210*. (page 95).

Zhou, X. & Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. *Nat Genet. 44(7), 821-4*. (page 28, 28, 28).

Zhu, X., Zhang, S., Tang, H. & Cooper, R. (2006). A classical likelihood based approach for admixture mapping using em algorithm. *Hum Genet. 120, 431-445*. (page 17).

Zhu, X., Tang, H. & Risch, N. (2008). Admixture mapping and the role of population structure for localizing disease genes. *Adv Genet. 60, 547-69*. (pages 12, 23, 23, 29, 30, 96).